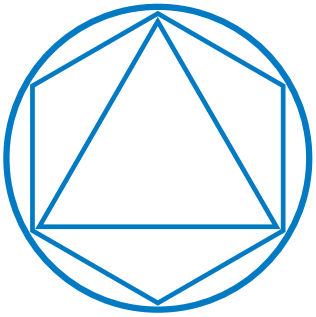
HELMHOLTZ ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE UNIVERSITÄT MÜNCHEN

Statistical learning for prediction of Type 1 Diabetes using clinical risk factors and omics data

Dissertationsschrift
an der Fakultät für Mathematik
der Technischen Universität
München



vorgelegt von
Michael Laimighofer

München, August 2017

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik — Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Statistical learning for prediction of Type 1 Diabetes using clinical risk factors and omics data

Michael Laimighofer

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr. Silke Rolles

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. Fabian J. Theis
2. Univ.-Prof. Claudia Czado, Ph.D.
3. Univ.-Prof. Dr. Matthias Schulze, Universität Potsdam

Die Dissertation wurde am 11.08.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 18.07.2018 angenommen.

Danksagung

An dieser Stelle möchte ich all jenen danken, die mich im Rahmen meiner Doktorarbeit begleitet haben.

Zuallererst möchte ich mich bei Fabian Theis besonders bedanken. Vielen Dank für die uneingeschränkte Unterstützung und Förderung, während der letzten vier Jahre.

Ein besonderer Dank gilt dir Jan. Deine großartige Betreuung während der gesamten Promotionszeit möchte ich dankend hervorheben. Auch vielen Dank für die gemeinsamen fachlichen Diskussionen und deine uneingeschränkte Unterstützung in allen Projekten.

Bedanken möchte ich mich auch bei Prof. Silke Rolles, der Vorsitzenden meiner Prüfungskommission, bei Prof. Claudia Czado für die Zweitbegutachtung und bei Prof. Matthias Schulze als externer Gutachter meiner Doktorarbeit.

Vielen Dank auch an euch beide, Anette und Ezio, für die spannende, tolle und äußerst interessante Zusammenarbeit während vieler gemeinsamer Projekte. Damit verbunden, ein Dankeschön an das gesamte Institut für Diabetesforschung für die produktive Kollaboration. Außerdem möchte ich mich bei Christine vT für die sehr gute Zusammenarbeit bedanken.

Großer Dank an alle vergangenen und aktuellen 'Systems Medicine of Diabetes' Teammitglieder, es war immer eine große Freude mit euch zusammenzuarbeiten. Danke auch Alida fürs Korrekturlesen der Doktorarbeit, danke auch an meine Bürokollegen Laleh, Atefeh und Thomas und Flo für die tolle Betreuung zu Beginn.

Danke auch ans ICB für die tolle Arbeitsatmosphäre, dem 'Bavarian Dream'-Team, sowie den unvergesslichen Tischtennisspielen mit Steffen, Ferdi, Jörg, Michi Str., Norbert und Ivan.

Besonderer Dank gilt auch meinen restlichen Freunden - besonders die Mittwochsgruppe -

Monika, Tina, Linda, Laura, Wolfgang und Verena für die moralische Unterstützung und gemeinsamen Urlaube in Italien im Sommer und Österreich im Winter. Ganz besonders bedanken möchte ich mich bei dir, Gunther, für die Freundschaft, Urlaube und alles sonst.

Meiner gesamten Familie möchte ich auch danken für die grenzenlose Unterstützung und Liebe nicht nur während dieser Promotion.

Ingrid, vielen lieben Dank auch an dich, dein Verständnis, deine Unterstützung und deine Liebe!

Abstract

Type 1 Diabetes (T1D) is an autoimmune disease, where the human immune system destroys insulin producing beta cells. This continuous decrease of beta cells leads to chronically high blood glucose levels and finally to the onset of T1D. To date, this process is irreversible and affected patients have to deal with lifelong insulin treatment and suffer from complications of hypo- or hyperglycemic situations. Various risk factors for T1D have been identified; however, the actual molecular mechanisms leading to T1D are not fully understood. Therefore, statistical models may help to discover new risk factors of T1D and further mechanistical studies of these risk factors may provide a better understanding of the molecular mechanisms leading to T1D.

In recent years, with the rise of modern high-throughput technologies, so-called 'omics' data provide insights into different molecular layers. These individual layers serve as a promising resource to obtain a holistic picture of the biological system and the functions therein. However, high-throughput omics measurements are characterized by a high-dimensionality, i.e. a large number of features p and a smaller number of observations n. Novel statistical methods are needed which are able to deal with such high-dimensional datasets, leading to a sparse set of features with reliable prediction accuracy.

In this thesis, my focus was to detect novel biomarkers for the prediction of T1D, with the goal of developing improved risk models based on high-throughput omics data and to investigate the early effects of T1D risk factors and the resulting molecular mechanisms.

First, we identified the need for the development of new methods in high-dimensional survival data. To this end, we proposed a novel method providing an unbiased estimation of the prediction accuracy and performing feature selection in high-dimensional datasets using a repeated nested cross-validation approach.

Second, we sought to identify sets of biomarkers to predict future T1D onset or progression time to T1D based on prospectively collected samples measured before any symptoms

have been detected. Thereby, we extended our method to classification and applied it to a proteomics dataset to predict autoantibody status (a prediabetic stage) and progression time.

Third, we evaluated a previously published genetic risk model on a new cohort. Specifically, we tested the discrimination ability of the risk model and subsets thereof on a general population cohort and children having a first-degree relative diagnosed with T1D. We confirmed the discrimination performance of the genetic risk model within both subgroups of the new cohort.

Finally, we investigated the effects of the type of delivery - Cesarean section (CS) vs. vaginal delivery - on the transcriptome in the first year of life, to identify molecular changes induced by CS and its connection to the development of autoantibodies. To this end, we compared the effects of CS with the effects after autoantibody development on the transcriptome. In addition, we investigated the molecular mechanisms using pathway enrichment analysis, in order to obtain insights in the early molecular effects of T1D risk factors.

Zusammenfassung

Typ 1 Diabetes (T1D) ist eine Autoimmunerkrankung, bei der das menschliche Immunsystem die Insulin produzierenden Betazellen zerstört. Der stetige Verlust dieser Betazellen führt bei den Betroffenen zu dauerhaft hohen Blutzuckerwerten und letztlich zur Diagnose von T1D. Mit den aktuellen Möglichkeiten ist dieser Prozess irreversibel und T1D-Patienten benötigen eine lebenslange Insulintherapie und leiden unter den durch Hypo- bzw. Hyperzucker hervorgerufenen Komplikationen. Bisher wurden verschiedene Faktoren, die das Risiko für eine T1D Erkrankung erhöhen, identifiziert, jedoch ist der exakte molekulare Mechanismus, der T1D zugrunde liegt, noch nicht geklärt. Aus diesem Grund werden statistische Modelle benötigt, um damit neue T1D Risikofaktoren zu entdecken. Diese neuen Faktoren können dann in mechanistischen Studien ein besseres Verständnis dieser molekularen Mechanismen liefern.

Durch die Entwicklung von modernen Hochdurchsatzverfahren in den letzten Jahren, geben so genannte 'omics'-Daten Einblicke in verschiedene biologische Ebenen. Diese einzelnen Ebenen bilden eine vielversprechende Grundlage, um damit ein ganzheitliches Bild des biologischen Systems sowie der Funktionen darin zu gewinnen. Die genannten Hochdurchsatzverfahren erzeugen jedoch hochdimensionale Datensätze, das heißt Datensätze mit einer großen Zahl von gemessenen Features p und einer kleineren Zahl von Beobachtungen n. Um solch hochdimensionalen Datensätzen zu bearbeiten, werden neue statistische Methoden benötigt, die zu einem kleinen Set von Features führen, und die eine verlässliche Einschätzung der Prädiktionsgüte erlauben.

Der Fokus dieser Thesis lag bei der Entdeckung neuer biologischer Marker zur Vorhersage von T1D mit dem Ziel verbesserte Risikomodelle basierend auf Hochdurchsatzdaten zu entwickeln. Ein weiteres Augenmerk war die frühen Effekte von T1D Risikofaktoren zu untersuchen, sowie die daraus resultierenden molekularen Mechanismen genauer zu bestimmen.

Zuerst sahen wir die Notwendigkeit neue Methoden für hochdimensionale Überlebenszeitdaten zu entwickeln. Dazu stellten wir eine neue Methode basierend auf wiederholten verschachtelten Kreuvalidierungen vor, die eine erwartungstreue Abschätzung der Prädiktionsgüte liefert und zugleich eine Variablenselektion durchführt.

Zweitens, identifizierten wir Kombinationen von Biomarkern um T1D und die Progressionszeit bis zum Beginn von T1D vorherzusagen. Dazu verwendeten wir prospektiv gesammelte Proben bevor noch jegliche Symptome bei den Patienten festgestellt werden konnten. Dabei erweiterten wir unsere vorgestellte Methode um Klassifikation und verwendeten sie in einem proteomischen Datensatz zur Vorhersage des Antikörperstatus (einer Vorstufe von T1D) und der Progressionszeit.

Drittens, evaluierten wir ein zuvor publiziertes genetisches Risikomodell anhand einer neuen Kohorte. Hier testeten wir die Diskriminierungsgüte dieses Risikomodells, sowie von Submodellen an einer Gruppe der Allgemeinbevölkerung und einer Gruppe von Kindern, die einen erstgradig Verwandten mit T1D Diagnose hatten. Dabei konnten wir die Diskriminierungsgüte des genetischen Risikomodells in beiden Subgruppen erfolgreich validieren.

Zuletzt untersuchten wir die Auswirkungen von Kaiserschnitt im Gegensatz zu einer herkömmlichen Geburt auf das Transkriptom im ersten Lebensjahr. Von besonderem Interesse waren die durch die Kaiserschnittgeburt hervorgerufenen molekularen Veränderungen und damit eine mögliche Verbindung zur Enstehung von Autoantikörpern herzustellen. Dazu verglichen wir die Effekte von Kaiserschnitt mit den Effekten von der Entwicklung von Autoantikörper, jeweils auf das Transkriptom. Um die molekularen Mechanismen im Detail zu untersuchen, führten wir eine Analyse der funktionalen und biologischen Prozesse durch, um damit Einblicke in frühe Effekte von T1D Risikofaktoren zu gewinnen.

List of contributed articles

i) **Michael Laimighofer**, Jan Krumsiek, Florian Buettner, and Fabian J Theis. **Unbiased prediction and feature selection in high-dimensional survival regression**. Journal of Computational Biology, 23(4):279–290, 2016.

ii) Christine von Toerne*, **Michael Laimighofer***, Peter Achenbach, Andreas Beyerlein, Tonia de las Heras Gala, Jan Krumsiek, Fabian J. Theis, Anette G Ziegler, and Stefanie M Hauck. **Peptide serum markers in islet autoantibody-positive children**. Diabetologia, 60 (2):287–295, 2017.

iii) Brigitte I. Frohnert*, **Michael Laimighofer***, Jan Krumsiek, Fabian J. Theis, Christiane Winkler, Jill M. Norris, Anette-Gabriele Ziegler, Marian J. Rewers, Andrea K. Steck. **Prediction of type 1 diabetes using a genetic risk model in the Diabetes Autoimmunity Study in the Young**. Pediatr Diabetes. 2017;0:1–7. doi: 10.1111/pedi.12543

iv) **Michael Laimighofer**, Ramona Lickert, Rainer Fürst, Fabian J. Theis, Ezio Bonifacio, Anette-Gabriele Ziegler, and Jan Krumsiek. **Common patterns of gene regulation associated with Cesarean section and the development of islet autoimmunity - indications of immune cell activation**. bioRxiv, Cold Spring Harbor Labs Journals, 2017. doi: 10.1101/167676

* Authors contributed equally

See chapter 3 for detailed contribution to each article

Contents

Chapter 1

Introduction

The central aim of this dissertation was to detect novel markers with the goal of improved risk models based on high-throughput omics data for the pathogenesis of Type 1 Diabetes (T1D).

1.1 Type 1 Diabetes

Type 1 Diabetes (T1D) is an autoimmune disease, where an endogenous lack of insulin leads to chronically high blood glucose levels. Untreated T1D would lead to hyperglycemic coma and death [1]. Affected patients suffer not only from lifelong insulin injections, but also from risks and complications involved in hypo- or hyperglycemic situations, such as cognitive dysfunction, cardiovascular events or kidney failure [2–4]. Therefore, therapeutically establishing a tight glycemic control is vital for the patients [5]. Notably, the incidence of newly diagnosed T1D has been increasing in the last decades, in particular in children and adolescents [6]. Currently, in Germany about 400,000 people suffer from diagnosed T1D and are dependent on daily injections of insulin. These insulin treatments cause high costs to the medical health care system [7]. Therefore, risk models predicting T1D and strategies for a deeper understanding of T1D pathogenesis are urgently needed, ideally leading to new interventions to prevent, delay or reverse T1D.

T1D pathogenesis is characterized by a humoral immune system response, destroying beta cells in the islets of Langerhans of the pancreas [8]. These beta cells are responsible for the production of insulin, and therefore essential for blood glucose balance. CD4+ and CD8+ T-cells of the adaptive immune system are mainly responsible for the destruction

process, induced by different sets of antibodies: islet cell antigen (ICA) antibodies, insulin and pro-insulin (INS) antibodies, glutamic acid decarboxylase (GAD) antibodies, protein tyrosine phosphatase (IA-2) antibodies, and Zinc transporter 8 (ZnT8A) antibodies [9]. The repeated detection of two or more autoantibodies (multiple autoantibodies) in blood in consecutive samples is called seroconversion and marks a necessary early stage in T1D development [10], see also Figure 1.1. Time from seroconversion to overt T1D is called progression time and varies greatly between affected children [11]. The progressive loss of beta cells leads to a reduced insulin production, in turn leading to elevated blood glucose levels which can be diagnosed with an oral glucose tolerance test. Pancreatic biopsies showed a loss of around 80% of beta cell mass at the time of diagnosis [12].

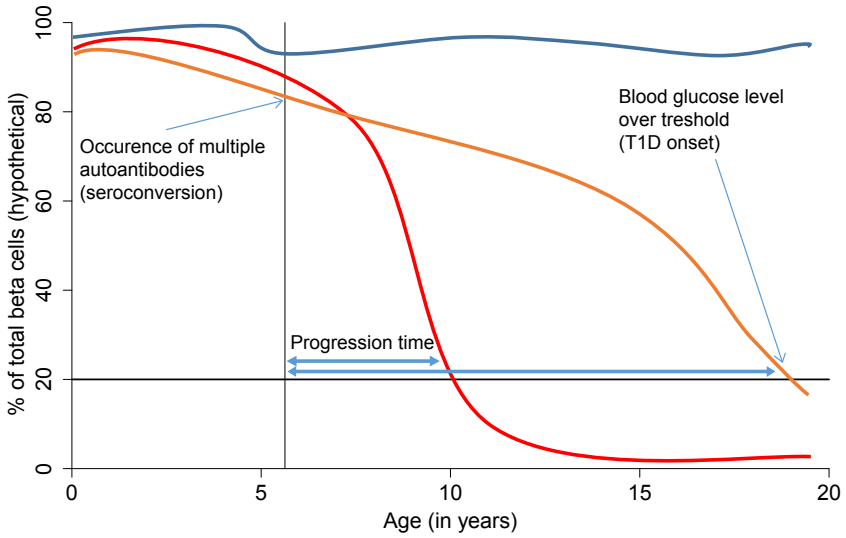


Figure 1.1: Three schematic trajectories of beta cell function: the blue line indicates an autoantibody negative child, not progressing to T1D. The red trajectory indicates a rapid progressor with a short progression time from seroconversion to T1D onset. The orange trajectory indicates a slow progressor with a longer duration between seroconversion and T1D onset. Of note, progression time varies greatly between children.

A variety of risk factors for the development of T1D and multiple autoantibodies have been identified. First, genetic susceptibility for T1D has been described for a number of genes [13] in genome-wide association studies (GWAS). GWAS have been used to examine, whether genetic variants, typically single-nucleotide polymorphisms (SNPs), are associated with a trait (such as T1D). In particular, genes within the human leukocyte antigen (HLA) class II region on chromosome 6 encoding the major histocompatibility complex, have the highest impact on T1D risk - especially combined polymorphisms in two loci of HLA-

DR and HLA-DQ (i.e. in particular HLA DRB1*03 DQA1*0501 DQB1*0201 and HLA DRB1*04 DQA1*0301 DQB1*0302) [14]. The entire genomic HLA region confers about 40-50% of total T1D risk [15]. In addition, genetic variation in more than 40 additional loci has been reported to increase the risk of T1D [13]. Similar genetic risk loci have been identified for the development of multiple autoantibodies before the onset of T1D [16]. Notably, the genetic risk factors for progression from multiple autoantibodies to T1D are distinct from those of overall disease development. Most importantly, the HLA genotype is considerably less predictive for progression time than for T1D [11].

Besides direct genetic influences, a number of familial and environmental risk factors have been identified. One important risk factor is the family history of T1D, also used in screening as inclusion criteria in most of the large T1D cohorts. Children who are born into a family with an already diagnosed T1D first-degree family member have a 5% risk of developing T1D, compared to the baseline risk of 0.3% [17]. Interestingly, the risk factor 'first-degree-relatives' consists of a genetic and an environmental part. Genetic studies of monozygotic twins have shown that about 65% of identical twins developed concordant T1D [18]. At the same time, 'first-degree-relatives' accounts for the lifestyle, family-habitues and spatial effects.

In the following, several exemplary, purely environmental risk factors for T1D will be discussed, emphasizing the diversity and complexity of T1D risk factors. First, an early exposure to gluten in the first 3 month of life has been shown to increase the risk for T1D (odds-ratio = 4.00) [19] - for details of the interpretation of odds-ratios 2.2.1 and hazard ratios see 2.2.3. As another example, in a pooled meta-analysis of 43 studies, breastfeeding has been identified as an environmental risk factor with a weak protective effect (odds-ratio = 0.75) [20] for the infants. In addition, recurrent respiratory virus infections in the first 6 month of life have been associated with higher T1D rates with an odds-ratio of 1.2 [21]. As an example of an important very early factor, children born by Cesarean section (CS) have a 20% increased risk for getting T1D later in life [22]. An interesting interaction effect between CS and a specific SNP has been reported with a hazard ratio of 2.40 [23]. Similar to genetics, progression is considered to be influenced by a different set of factors than the development of T1D and autoantibodies in general. For example, an important risk factor for progression time is the age of seroconversion. Children who develop multiple autoantibodies before the age of three progress considerable faster to T1D with a hazard ratio of 1.65 [11].

In summary, T1D is a complex disease with various genetic and environmental risk factors, and interactions thereof. The actual molecular mechanisms leading to T1D and autoimmu-

nity are not fully explored, and therefore, a more integrative and molecular understanding of the disease onset is needed.

1.2 High-throughput omics profiling for T1D

In recent years with the development of high-throughput technologies, so called omics data have emerged in biological and medical research [24]. Such omics data are now fast and cheap to generate, and cover most important layers of the molecular biological system. Specifically, mass spectrometry is mainly used to generate measurements of protein expression, metabolite profiles, and lipids [25]. Moreover, chip-based and sequencing-based methods quantify expression of transcripts, capture methylation changes of the DNA and identify variations of the DNA sequence. Typical studies nowadays cover up to hundreds to millions of molecular markers of those different layers. Multi-omics datasets aim to build a holistic picture of the biological system and the functions therein, leading to the development of novel research areas called 'systems genetics' and 'systems medicine' [26]. The promise is that by integrating the different biological layers, multi-omics datasets will help to develop deeper knowledge and a better understanding of the pathogenesis, and aid in the identification of novel biomarkers. Figure 1.2 shows the complex interplay between omics layers and environment, resulting in the complex phenotype T1D. Starting with an inherited genetic background risk, environmental factors change DNA methylation and gene expression, which are then interacting with the proteome. The metabolome - considered to be an endpoint of biological processes - is influenced by all other omics layers and is often referred to as the link between genotype and phenotype (T1D onset) [27].

Importantly, the omics technologies are now also well-established in most major T1D study cohorts. One of this large cohorts is the 'Type 1 Diabetes Genetics Consortium' (T1DGC), an international multicenter program with the aim to identify genes associated with T1D [29]. In T1DGC, data of 1,307 subjects of a genome wide linkage scan and 9,976 subjects with HLA genotyping and about 3,000 SNPs within the major histocompatibility complex are available. In addition, 'The Diabetes Autoimmunity Study in the Young' (DAISY) is a birth-cohort study, which includes 2,542 children [30]. Study participants are first-degree-relatives and an HLA high risk-selected general population, both of which show a higher risk for T1D. In DAISY, data is available on SNP genotyping, environmental exposures, and longitudinal islet-cell autoimmunity. As our main local data sets from Munich, BABYDIAB consists of 1,650 children from parents with T1D, and BABYDIET contains 150 high risk first-degree relatives of T1D patients, constituting the largest German cohorts for T1D [31] [32]. In BABYDIAB/DIET, high-risk SNP genotyping, methylation, tran-

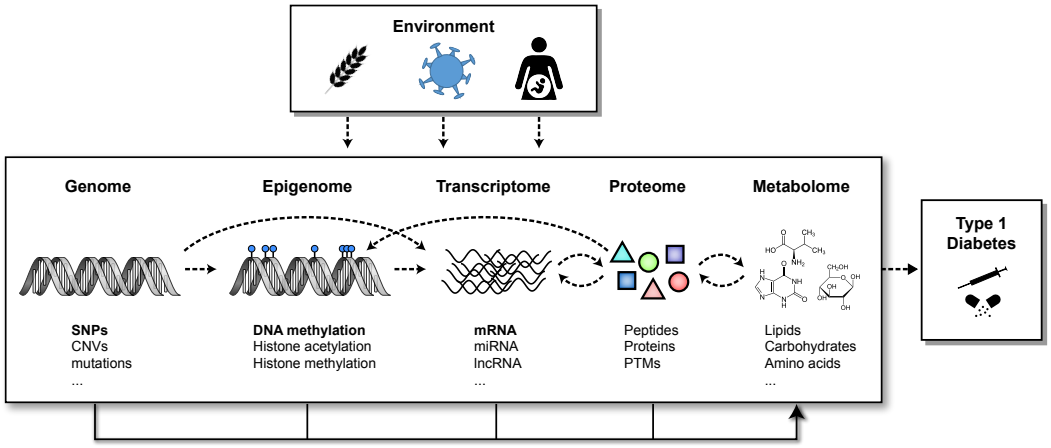


Figure 1.2: Complex interactions between different omics layers, environment, and phenotype, in T1D (adapted from [28]).

scriptomics, proteomics, metabolomics data, and environmental covariates have been measured. Moreover, 'The Environmental Determinants of Diabetes in the Young' (TEDDY) has been established, representing the largest multi-omics T1D population cohort of 8,676 study participants [33]. For TEDDY, data has been recorded of environmental variables, such as food nutrient information and birth sizes, proteomics, high risk SNP genotyping or marker on inflammation. Notably in DAISY, BABYDIAB/DIET and TEDDY high-risk children have been longitudinally observed before T1D onset. Table 1.1 describes all datasets used in the publications related to this thesis.

Cohort	Omics data type	Sample size n	Features p
BABYDIAB/DIET	Proteomics (untargeted)	45	2,021 peptides
BABYDIAB/DIET	Proteomics (targeted)	140	82 peptides
DAISY	Genetics	1941	9 SNPs + HLA
BABYDIAB/DIET	Transcriptomics (longitudinal)	454 (109 children)	18,720 genes

Table 1.1: Characterisation of the datasets used in the publications related to this thesis.

1.3 High-dimensional data analysis

All omics datasets are characterized by a large number of measured features p, readily reaching thousands (e.g. in proteins or metabolites) to hundreds of thousands (for methylation data) to millions of features (e.g. SNPs). In contrast to p, the number of samples n

is rising less steeply, usually with only up to a few thousands observations available even in the biggest studies, and more commonly a few hundreds or even less. Therefore, we are dealing with high-dimensional datasets, that is $p > n$, see also Table 1.1 above showing the datasets we used. In the following part, we give an overview of selected methods for high-dimensional data. In particular, we present methods used in our publications, and two other approaches used with T1D datasets by other colleagues, namely a Bayesian model selection approach [34] and a dimension reduction approach [35]. Mathematical details on used models can be found in chapter 2.

1.3.1 Methods for high-dimensional data analysis

Univariate statistics often serve as a starting point for most omics-applications. To this end, the associations between all individual predictors and the response of interest are calculated. However, univariate approaches fail to simultaneously aggregate the information of all features, as the correlation structure within the data is neglected. Hence, multivariable approaches are favored in order to combine the information into one joint model. Since classical models are not uniquely solvable in $p > n$, and since subsets of features might be sufficient for predictive modeling, regularization and feature selection methods have been a major focus of methods development.

A commonly used variant of regularized models are penalized regression approaches. In these models, a penalization factor is put on parameter estimates, in order to allow for unique model identification. First, putting the L_2 norm on the squared coefficients is known as *ridge* regression [36]. It is defined as a Euclidean metric on the parameters. All coefficients are proportionally shrunken and non-zero. This also means, ridge regression does not perform feature selection. The optimization problem for the L_2 norm can be analytically solved. Second, the least absolute shrinkage selection operator (LASSO) incorporates a L_1 regularization on the parameter estimates [37]. Hereby, the method also shrinks the coefficients towards zero and sets some coefficients to exactly zero, leading to an intrinsic feature selection. Moreover, LASSO states a convex optimization problem, which is computationally favorable to solve, in contrast to L_d norms with $d < 1$. In regularization approaches, the parameter estimates (such as odds-ratios or relative risks) remain interpretable compared to other approaches, for example support vector machines or dimension reduction methods which are mentioned in the following.

Support vector machines (SVM) are another popular statistical learning tool able to deal with high dimensionality. SVMs aim to derive class separation by finding a decision boundary, represented by so called support vectors, which denote observations closest

to the decision boundary [38]. The complexity of a SVM is determined by these support vectors, rather than the high dimensionality of the feature space, and therefore able to deal with $p > n$. Moreover, SVMs may include non-linearities of features by using the so called 'kernel-trick'. However, in SVMs parameter estimates are not interpretable compared to regularized models.

Dimension reduction methods define another model class which is able to deal with $p > n$ datasets. In these approaches, the original feature space is transformed into a lower dimensional space of p' features with $p' < n$, ideally still capturing a large extent of the original variability. These lower dimensional features are then used to fit classical models and predict the outcome. As examples, principal components regression (PCR) [39], partial least squares (PLS) [40], and object oriented regression (OOR) [41] use this strategy to model high-dimensional data. One drawback of these dimension reduction methods is again the non-interpretability of derived features. For example, in principal component analysis, a transformed feature (principal component) is a linear combination of the original feature space.

Finally, Bayesian models are another important class of high-dimensional learners. Here, prior probability distributions are used to infer knowledge or beliefs about the data structure, before data itself is taken into account [42]. Using a suitable set of prior distributions, a posterior distribution can be derived by applying Bayes' theorem. Thus, all prior probability distributions multiplicatively add to the posterior distribution, which is often not available in closed form. In order to obtain samples from the posterior distribution, Markov chain Monte Carlo (MCMC) methods are applied, which are computationally demanding, especially for high-dimensional data. In addition, checks of the convergence and mixing property of different MCMC chains have to be performed [43]. Approaches dealing with high-dimensional data and performing feature selection are reversible jump MCMC, Bayesian subset regression, and Bayesian model averaging [34] [44] [45]. Bayesian models remain interpretable, and are intrinsically able to deal with missing data, compared to all methods described before.

1.3.2 Bias-variance tradeoff in high-dimensional data analysis

In general, any modeling approach has to deal with the so called bias-variance tradeoff. *Bias* estimates the concordance of model predictions to the true outcome, such as T1D. *Variance* describes the sensitivity of the model predictions to changes in the underlying data. Minimizing the bias would lead to perfect predictions, whereas the generated prediction model is not able to capture the variability of new data, leading to bad generalization.

This issue is also known as overfitting. All of the above described methods attempt to deal with overfitting and the bias-variance trade-off. To avoid overfitting, statistical models aim to estimate a sparser representation of input features p, based only on a subset of features, also known as '*bet on sparsity*' [46]. More details on the bias-variance tradeoff are given in section 2.1.1. In the Munich T1D cohort we also had high-dimensional data with $p > n$ (see Table 1.1). Thus, we intend to avoid overfitting and aim to identify a set of biomarkers with high discrimination accuracy which can be easily applied to other cohorts or even in patient screening in clinical practice.

1.4 Motivation of the thesis

During the last years, various omics layers have been measured in T1D research facing the issues of high-dimensional data analysis as described above. We had ready access to high-dimensional omics data of the largest German T1D cohort available and identified the need for method development in general and for tailored analysis in T1D in particular. This combination put us into the unique position to develop high-dimensional models for T1D biomarker discovery and mechanistical research. To the best of our knowledge, we have been the first developing a nested cross-validation approach performing both feature selection and estimation of the generalization performance in multivariable high-dimensional survival data. Our models and analysis were developed and applied in close collaboration with medical and biological experts from T1D research to meet the needs and characteristics of T1D data, and to provide a link between biological systems and complex mathematical and statistical modeling.

Chapter 2

Methodology

Statistical learning of high-dimensional data lays the basis for this work. This chapter begins with a general definition of statistical learning. Then, we define classification models and time-to-event models, which are used to infer prediction rules. We moreover define measures that estimate the predictive accuracy of a risk prediction model. As a novel approach, a repeated, nested cross-validation algorithm is described, which we developed in our publication [47], extended, and applied it in a second publication [48]. In the last section, we describe an approach for complex datasets including longitudinal measurements, random effects, and fixed effects [49].

2.1 Statistical learning

Throughout this thesis, let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ be a multivariate set of random variables, typically in $\mathbb{R}^p$, and defined as a $n \times p$ input matrix with real valued, independent and identically distributed (iid), random realizations x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, such that

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} \tag{2.1}$$

with p being the number of features and n the number of observations. We define $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ as the column vector of observations for sample i. Moreover, in this section we define $\mathbf{y} = (y_1, \ldots, y_n)$ being a real valued random outcome vector of length n with quantitative measurements, such as body mass index or blood pressure ($\mathbf{y} \in \mathbb{R}$). Then, in statistical learning, we aim to infer the unknown functional relationship $f(\cdot)$ between $\mathbf{y}$ and $\mathbf{X}$ described by

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \forall i \tag{2.2}$$

with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ being an error term, which is independent of $\mathbf{X}$ and $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. For example, in linear regression this functional relationship is defined as $f(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ for sample i with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ being the vector of regression coefficients to be estimated. Another example are generalized additive models (GAMs), where the functional relation is defined as $f(\mathbf{x}_i) = f(x_{i1}) + \ldots + f(x_{ip})$, which is far more flexible than the linear model described above.

2.1.1 Bias-variance tradeoff

By using the 'training data' $\mathbf{X}$ and $\mathbf{y}$, a prediction rule $\hat{f}$ is estimated. In order to estimate the generalization error, we apply $\hat{f}$ on new unseen data $\mathbf{X}'$ and $\mathbf{y}'$ ('test data'). We calculate the *bias* defined as $\text{Bias}\big[\hat{f}(\mathbf{X}')\big] = \mathbb{E}\big[\hat{f}(\mathbf{X}') - f(\mathbf{X}')\big]$ as the concordance of model predictions to the real test outcome, emphasizing that $\hat{f}$ is estimated and f denotes the true relationship. In addition, the *variance* is defined as $\text{Var}\big[\hat{f}(\mathbf{X}')\big] = \mathbb{E}\Big[\big(\hat{f}(\mathbf{X}') - \mathbb{E}[\hat{f}(\mathbf{X}')]\big)^2\Big]$ of $\hat{f}$. Moreover, we define the generalization (or expected) error of $\hat{f}$ for the test data $\mathbf{X}'$ as

$$\begin{aligned} \mathbb{E}\Big[(\mathbf{y}' - \hat{f}(\mathbf{X}'))^2\Big] &= \mathbb{E}\Big[(\mathbf{y}')^2 + \hat{f}(\mathbf{X}')^2 - 2\mathbf{y}'\hat{f}(\mathbf{X}')\Big] \\ &= \text{Bias}\big[\hat{f}(\mathbf{X}')\big]^2 + \text{Var}\big[\hat{f}(\mathbf{X}')\big] + \sigma^2. \end{aligned} \tag{2.3}$$

The second line of equation 2.3 shows the bias-variance decomposition (also known as bias-variance trade-off) plus the - so called 'irreducible' - error σ^2. A prediction rule which perfectly fits the training data ($\hat{f}(\mathbf{X}) - \mathbf{y} = \mathbf{0}$) also models the error term $\boldsymbol{\epsilon}$ and is therefore

unlikely to perform well on unseen new data. A high bias indicates systematic deviations from the true relation (underfitting). Strong sensitivity to small changes in the underlying training data indicates high variance and is known as overfitting. In general, linear models exhibit high bias and low variance, while more flexible models (e.g. generalized additive models (GAM)) show low bias, but high variance (see Figure 2.1). Thus, one aims to derive a prediction rule exhibiting a balance between bias and variance and therefore reliable generalization.

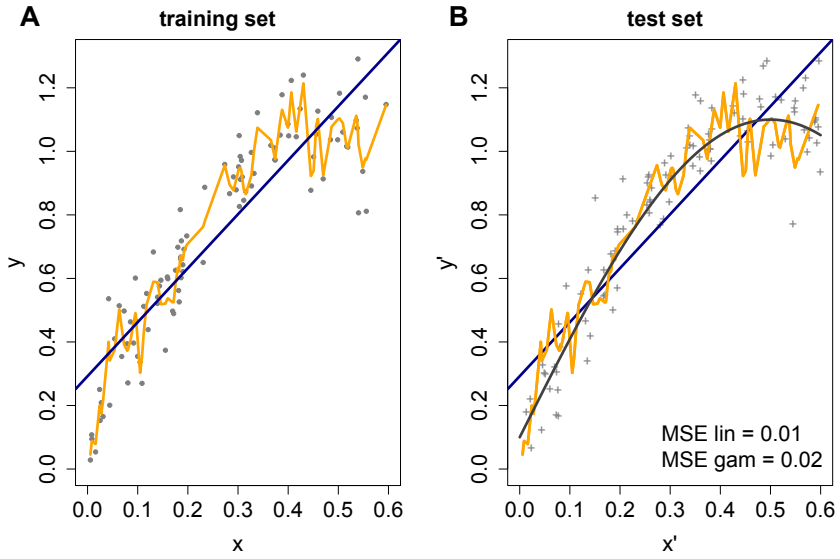


Figure 2.1: **A:** Observed training data (grey dots) is fitted with a linear approximation (blue line) and a more flexible GAM (orange line). **B:** New test data (grey crosses) are used to evaluate the estimated linear and GAM based $\hat{f}$. In addition, the true data generating function $f(x_i) = 0.1 + \sin(\pi x_i)$ is shown in black. The effect of overfitting is illustrated with the GAM function (orange line), where the trained function is too noisy compared to the truth and therefore exhibits higher variance. The linear function (blue line) shows less variance, but systematic deviations form the true model can be identified (underfitting).

2.1.2 Resampling methods

In order to balance under- and overfitting and to obtain high generalization, resampling methods have been proposed [50]. Generally, these resampling methods split a dataset into training and test data to first estimate the prediction rule on the training data and then apply it onto the test data. Here, we give details on the most important resampling methods. As defined above our data set consists of n samples.

Bootstrapping: The training data is obtained by drawing n samples each with probability $1/n$ from the dataset with replacement [51]. Since we draw with replacement, samples in the bootstrap set will not be unique. The selected samples are used for estimating the prediction rule, whereas the samples not drawn serve as test set. This sampling procedure is repeated several times.

Cross-validation: In k-fold cross-validation the full data set is randomly divided into k subsets of equal size [52]. Since the dataset is split into subsets, it is sampling without replacement. The training data consists of $k-1$ subsets to estimate the prediction rule, whereas the k-th fold is used for testing. This procedure is repeated for all k subsets, whereby each fold k serves once as test set.

Subsampling: In subsampling a fraction r of the data set is drawn without replacement [53]. This subset is used to estimate the prediction rule and the hold-out samples are used for testing. Similar to bootstrapping, this procedure is repeated several times for different data splits.

All described resampling methods are also used to estimate unknown variances of point estimates [51].

2.2 Generalized linear models

Up to now, we assumed that the response vector $\mathbf{y}$ is a quantitative measurement. By assuming a normal error distribution of iid $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ a linear regression is obtained and we can define the relationship between $\mathbf{y}$ and $\mathbf{X}$ probabilistically as

$$\mathbb{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = N(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}). \tag{2.4}$$

In order to estimate the model parameters $\boldsymbol{\beta}$, maximum likelihood (ML) estimation of 2.4 or equivalently residual least-squares methods are used resulting in the parameter estimates $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

In T1D research, we often deal with a binary outcome, such as T1D onset ('yes/no') or status of multiple autoantibodies ('AB+/AB-'). Instead of using an error distribution for $\boldsymbol{\epsilon}$, in classification the probability $\mathbb{P}(y_i = 1|\mathbf{x}_i)$ is directly modeled using a binomial

distribution. As another type of outcome, we investigate the progression time from development of multiple autoantibodies to T1D. Such observations are subject to censoring, i.e. for children who have not developed T1D so far, and represent time-to-event data ('survival data'). The concepts of overfitting and bias-variance tradeoff presented so far, apply as well for binary or survival response.

In the following, we present generalized linear models (GLMs), in which linear regression, logistic regression models and parametric survival models can be summarized. These models are characterized by a linear predictor η_i for sample i, which is given by

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{2.5}$$

with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ the vector of regression coefficients to be estimated and observed features $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. Specifically, in GLMs the linear predictor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ is related through a link function $h(\cdot)$ with the conditional expectation value of the response $\mathbb{E}(\mathbf{y}|\mathbf{X}) = h(\boldsymbol{\eta})$ given $\mathbf{X}$ from 2.1. E.g. in linear regression the identity link for $h(\cdot)$ is used. Another characteristic of GLMs is that distributions of the response are members of the exponential family. Thus, the density function in the univariate case of the exponential family can be rewritten as

$$f(y_i|\theta) = \exp\Big(\frac{y_i\theta - b(\theta)}{\phi} + c(y_i, \phi)\Big) \tag{2.6}$$

with ϕ being a dispersion parameter, θ the canonical (or natural) parameter, and $b(\cdot)$ such that the first and second order derivative exist and that $b(\cdot)$ normalizes $f(y_i|\theta)$. Moreover, in GLMs parameter estimation is based on ML approaches [54].

2.2.1 Logistic regression

In logistic regression we estimate the probability $\pi_i = \mathbb{P}(y_i = 1|\mathbf{x}_i) \in [0, 1]$ of a binary outcome $\mathbf{y} = (y_1, \ldots, y_n)$ with $y_i \in \{0, 1\}$. The link function $h(\cdot)$ relates $\pi_i = h(\eta_i) \in [0, 1]$, and therefore, with the observed features $\mathbf{x}_i$. A commonly used link function $h(\cdot)$ for logistic regression is the *logit* link defined by:

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i. \tag{2.7}$$

Hence, the odds ratio $\frac{\pi_i}{1-\pi_i} = \exp(\eta_i)$ or log-odds ratio from equation 2.7 allow for useful interpretations: an increase of 1 unit of a covariate changes the logarithmic odds by the corresponding β. The likelihood can be written as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{2.8}$$

using the binomial distribution of $\mathbf{y}$ with $y_i \sim B(1, \pi_i)$. In order to obtain ML estimates, the likelihood is logarithmized $l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$ and the score function $s(\boldsymbol{\beta})$ is calculated as the first derivative of the log-likelihood according to $\boldsymbol{\beta}$

$$s(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i(y_i - \pi_i). \tag{2.9}$$

Equating the score function to zero and solving this set of equations, we obtain the ML estimates ($\hat{\boldsymbol{\beta}}$) for the regression coefficients. Optimization is iteratively performed using Newton-Raphson method or the Fisher scoring algorithm [54]. In case of $p > n$, this system of equations in logistic regression is not uniquely solvable. One approach dealing with this $p > n$ setup is to penalize the parameters $\boldsymbol{\beta}$.

2.2.2 Penalized logistic regression

In penalized logistic regression, the regression coefficients are penalized using an L_1-norm (LASSO), L_2-norm (ridge) or a combination of the two (elastic net) [55]. This regularization can be formulated as a Bayesian prior distribution on the regression coefficients. In particular, for ridge regression it corresponds to a multivariate Gaussian prior, whereas the LASSO type of penalization is obtained assuming a Laplace prior on the regression coefficients. The log-likelihood of the LASSO type of penalization has the form

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\big(y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\big) + \lambda||\boldsymbol{\beta}||_1 \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(y_i \log h(\eta_i) + (1-y_i)\log(1-h(\eta_i))\big) + \lambda||\boldsymbol{\beta}||_1
\end{aligned}
\tag{2.10}
$$

with λ being the penalization or shrinkage parameter and $||\boldsymbol{\beta}||_1 := \sum_{j=1}^{p} |\beta_j|$. An advantage of the L_1-norm compared to ridge regression ($||\boldsymbol{\beta}||_2 = \sum_{j=1}^{p} \beta_j^2$) is that it leads to sparse solutions, since some of the β's will be set to 0, leading to an intrinsic model selection. Moreover, each value of λ leads to a different model with a different number of non-zero parameters. Optimization of $\boldsymbol{\beta}$ along the λ-path is performed using coordinate descent and can be computationally performed very efficiently [56]. This computational efficiency is obtained, since the optimization problem is still convex, in contrast to approaches with a $L_d,\ d < 1$ norm. In order to derive an optimal λ, cross-validation is usually applied [36]. The λ value corresponding to the smallest cross-validated test error gives an optimal model choice.

2.2.3 Cox proportional hazards model

The Cox proportional hazards model is a commonly used method to describe time-to-event data. The observations per sample i are defined as in equation 2.1 with $\mathbf{x}_i$ denoted as the observed covariates of sample i. In survival models the response variable is defined as a two dimensional outcome vector with the observation time T_i and a censoring indicator $\delta_i \in \{0, 1\}$ for a failure event ($\delta_i = 1$, such as T1D diagnosis or death), or censoring event ($\delta_i = 0$, e.g. end of observation period or end of trial). In particular, T_i is called 'censoring time' (if $\delta_i = 0$) and 'event time' (if $\delta_i = 1$). Let $t_1 < \ldots < t_m$ be defined as the ordered unique (without ties) event times (set of unique T_i with $\delta_i = 1$), at time t_k with $k = 1, \ldots, m$ and m the number of events, and $x_{(i)j}$ be the jth-covariate of the individual with event time t_i. Furthermore, all individuals with longer observation time $T_i > t_k$ constitute the risk set $R(t_k)$ at time t_k.

In the Cox proportional hazards model the covariates $\mathbf{x}_i$ of an individual are related with survival in the hazard function defined as

$$h_i(t|\mathbf{x}_i) = h_0(t)\exp(\sum_{j=1}^{p} x_{ij}\beta_j) \tag{2.11}$$

where h_0 is the common baseline hazard and $\boldsymbol{\beta}$ is a vector of regression coefficients of length p. Inference of $\boldsymbol{\beta}$ is performed by maximizing the partial likelihood, defined as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp(\sum_{j=1}^{p} x_{(i)j}\beta_j)}{\sum_{k\in R(t_i)} \exp(\sum_{j=1}^{p} x_{kj}\beta_j)}, \tag{2.12}$$

where the baseline hazard $h_0(t)$ has already been canceled out. Similar to logistic regression, the Cox proportional hazards model is not uniquely solvable in case of $p > n$.

In order to calculate the survival probability $S_i(t)$ for one subject at time t, defined as $S_i(t) = \exp(-\int_0^t h_i(u|\mathbf{x}_i)du)$ with estimated $\hat{\boldsymbol{\beta}}$, we need to specify the baseline hazard $h_0(t)$ from equation 2.11. To this end, the Nelson-Aalen estimator of the cumulative hazard $H_0(t)$ can be used [57], which is defined as

$$\hat{H}_0(t) = \sum_{t_i \le t} \frac{d_i}{r_i} \tag{2.13}$$

with d_i the number of events at t_i and r_i the number of individual at risk.

Since the Cox survival model is a proportional hazard model, the ratio of two individual hazard rates can be interpreted, similar to the log-odds ratio in logistic regression in equation 2.7. The interpretation of an estimated $\hat{\beta}_j$ is that a unit increase of covariate j leads to an increase of $\exp(\hat{\beta}_j)$ in the hazard ratio (relative risk), keeping all other covariates constant.

2.2.4 Penalized Cox proportional hazards model

In order to deal with high-dimensional datasets with $p > n$ and a time-to-event response, a L_1 norm on the regression coefficients can be added [58] to (2.12), leading to the log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{m} \log \big(\frac{\exp(\sum_{j=1}^{p} x_{(i)j}\beta_j)}{\sum_{k \in R(t_i)} \exp(\sum_{j=1}^{p} x_{kj}\beta_j)} \big) + \lambda ||\boldsymbol{\beta}||_1 \tag{2.14}$$

and has to be optimized with respect to $\boldsymbol{\beta}$, which is strongly related to the logistic regression formulation from equation 2.10. Again, the complexity parameter λ determines the amount of shrinkage. In Simon et al. [59] an efficient implementation of the regularization path has been described using a coordinate-descent approach. Details for the optimal λ choice have been given in section 2.2.2.

2.3 Support Vector Machines

Another approach in classification which is able to deal with high-dimensional data are support vector machines (SVM) [38]. As above, for sample i we have a set of observations $\mathbf{x}_i$ which are linked to a binary outcome $y_i \in \{-1, 1\}$. The aim of SVMs is to find a decision boundary between these two classes. In order to derive this decision boundary, we try to maximize the margin M defined as $M = 1/||\boldsymbol{\beta}||$ with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ between two hyperplanes, separating the two classes. A hyperplane is defined as $G(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} \mathbf{x}_j\beta_j = 0$. Using a hyperplane $G(\mathbf{x}_i)$, we obtain a classification rule $f(\mathbf{x}_i)$ with $f(\mathbf{x}_i) = \text{sign}\left[\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right]$. Thus, the SVM optimization problem is defined as

$$\begin{aligned} \max_{\beta_0, \boldsymbol{\beta}} M \iff \min_{\beta_0, \boldsymbol{\beta}} ||\boldsymbol{\beta}|| \\ \text{subject to } y_i(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j) \geq 1 - \xi_i, \forall i \text{ and } y_i \in \{-1, 1\} \\ \xi_i \geq 0, \sum_{i=1}^{n} \xi_i \leq C, \end{aligned} \tag{2.15}$$

where ξ_i are error terms (also called slack variables) for non separable cases and C is a tuning parameter. This tuning parameter C can be interpreted as a budget for violating the margin M. In order to obtain a solution for the set of equations in 2.15, we can rewrite this convex optimization problem as a Lagrange primal function

$$L_P = \frac{1}{2}||\boldsymbol{\beta}||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i(\beta_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) - (1-\xi_i)\right] - \sum_{i=1}^{n}\mu_i\xi_i \tag{2.16}$$

with α_i and μ_i being Lagrange multipliers. The function L_P is minimized with respect to $\boldsymbol{\beta}, \beta_0$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ and setting the derivatives to zero, we obtain $\boldsymbol{\beta} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{n}\alpha_i y_i = 0, \alpha_i = C - \mu_i \ \forall i$, and $\alpha_i, \mu_i, \xi_i \geq 0 \ \forall i$. Applying this set of equations to (2.16) we get the Lagrange dual objective function

$$L_D = \sum_{i=1}^{n}\alpha_i y_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j, \tag{2.17}$$

which is maximized under constraints of $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{n}\alpha_i y_i = 0$. Additionally, the constraints of the Karush-Kuhn-Tucker conditions apply

$$\begin{aligned} \alpha_i[y_i(\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0) - (1-\xi_i)] &= 0, \\ \mu_i\xi_i &= 0, \\ y_i(\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0) - (1-\xi_i) &\geq 0. \end{aligned} \tag{2.18}$$

Equations 2.17 and 2.18 uniquely constitute the solution to the primal and dual problem. The solution for $\boldsymbol{\beta}$ is obtained by

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}\hat{\alpha}_i y_i \mathbf{x}_i \tag{2.19}$$

with some or all coefficients $\hat{\alpha}_i$ are intrinsically estimated to be non-zero. The samples with $\alpha_i \neq 0$ are called *support vectors*. Finally, we classify the observations according to $f(\mathbf{x}_i) = \text{sign}\left[\hat{\beta}_0 + \sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right]$.

From (2.17), we see that the input features only occur as inner (scalar) products. Thus, we can use a function $h(.)$ to transform the feature vectors and compute their inner product $\langle h(\mathbf{x}_i), h(\mathbf{x}_k)\rangle$ [55]. Instead of specifying $h(.)$, we use the so-called 'kernel trick' by defining a kernel function $K(\mathbf{x}_i, \mathbf{x}_k) = \langle h(\mathbf{x}_i), h(\mathbf{x}_k)\rangle$. Such a function K requires symmetricity $K(\mathbf{x}_i, \mathbf{x}_k) = K(\mathbf{x}_k, \mathbf{x}_i)$, and the kernel matrix $K_{i,k} := K(\mathbf{x}_i, \mathbf{x}_k)$ needs to be positive (semi-) definite $\forall \mathbf{x}_i, i = 1, \ldots, n$ [60]. Moreover, the space of functions spanned by K is called reproducing kernel Hilbert space (RKHS) [60]. For every kernel function as defined above, a unique RKHS exists and vice versa [61]. Of note, depending on the kernel function we derive a non-linear generalization of the input features.

In (2.17) we use a linear kernel, but several other kernel functions are often used - utilizing non-linearities in feature space. Example kernels include the radial basis function with $K(\mathbf{x}_i, \mathbf{x}_k) = \exp(-\nu||\mathbf{x}_i - \mathbf{x}_k||^2)$ and neural network kernels $K(\mathbf{x}_i, \mathbf{x}_k) = \tanh(\kappa_1 \langle \mathbf{x}_i, \mathbf{x}_k\rangle + \kappa_2)$ with ν, κ_1, and κ_2 being additional tuning parameters. In case of the linear kernel the tuning parameter C is derived by cross-validation on the data set.

2.4 Validation of a prediction rule

In order to obtain an estimator of prediction accuracy, measures of discrimination are applied to internal and external validation. First, internal validation assesses the prediction accuracy within one data set. It is commonly performed using cross-validation, bootstrapping or subsampling approaches 2.1.2. Such concepts of estimating the performance on hold-out data are particularly useful if new data sets are not available. Second, external validation aims to apply a prediction model on a new population. External validation is generally considered as a stronger evidence for generalizability of prediction rules.

In order to validate a prediction rule $\hat{f}(\mathbf{X})$ on new data $\mathbf{X'}$, we first apply the estimated $\hat{f}$ on $\mathbf{X'}$. In GLMs, the prediction rule (or risk score) is defined as a linear predictor $\hat{\eta}_i$ of input features per subject i:

$$\text{risk score}_i = \hat{\eta}_i = \hat{\beta}_0 + \sum_{j=1}^{p^*} \hat{\beta}_j x'_{ij} \tag{2.20}$$

with $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p^*)$ being the estimated effect estimates and p^* being the selected features - usually a sparse set of features compared to the input space $\mathbb{R}^p$. In order to determine whether the scores are concordant to the true classes, a concordance statistic c can be estimated with $\hat{\eta}_{i,y'=1}$ being a score of a positive instance, $\hat{\eta}_{i,y'=0}$ for a negative instance. Therefore, we first calculate a confusion matrix for a given cut-off k with true positives ($TP_k = \sum_i I(\hat{\eta}_{i,y'=1} \geq k)$), false positives ($FP_k = \sum_i I(\hat{\eta}_{i,y'=1} < k)$), true negatives ($TN_k = \sum_i I(\hat{\eta}_{i,y'=0} < k)$), and false negatives ($TP_k = \sum_i I(\hat{\eta}_{i,y'=0} \geq k)$). Based on these measures, the true positive rate ($TPR_k = (TP_k/(TP_k + FN_k))$) and false positive rate ($FPR_k = (FP_k)/(FP_k + TN_k)$) are computed. By varying over all possible cut-offs k, we obtain pairs of TPR and FPR, which, if plotted against each other, result in the so called receiver operating characteristic (ROC) curve. To calculate a summary statistic of all pairs of (TPR, FPR), the area under the ROC curve (AUC) has been widely established [62]. An AUC is defined as the integral over all cut-offs k:

$$AUC = \int_{\text{inf}}^{-\text{inf}} TPR_k FPR_k' dk = \mathbb{P}(\hat{\eta}_{i,y'=1} > \hat{\eta}_{j,y'=0}). \tag{2.21}$$

For a binary outcome the AUC is identical to the concordance statistic c. Intuitively, the AUC can be interpreted as the probability that any pair of cases and controls are correctly ordered. An AUC of 1 indicates perfect separation, whereas an AUC of 0.5 means random class assignment.

In a survival setting, the observations are subject to censoring and, therefore, not all pairs of observations are comparable. To overcome this shortcoming, several authors proposed concordance statistics for survival data, also called survival AUC [63], [64], or [65]. In our publication [47], we used the survival AUC definition of Uno et al. [63], where they accounted for the fact that the distribution of censored times are usually shorter than the event times (censoring bias). It is defined as

$$c_\tau = \frac{\sum_{j=1}^n \sum_{k=1}^n \hat{S}(T_j)^{-2} I(T_j < T_k, T_j < \tau) I(\hat{\eta_j} > \hat{\eta_k}) \delta_j}{\sum_{j=1}^n \sum_{k=1}^n \hat{S}(T_j)^{-2} I(T_j < T_k, T_j < \tau) \delta_k} \in [0, 1] \tag{2.22}$$

with τ being a pre-specified point in time and $I(\cdot)$ the indicator function. $\hat{S}(T_j)$ denotes the Kaplan-Meier estimator of the unconditional survival function, which is estimated from the data and is defined as

$$\hat{S}(t) = \prod_{t_j \leq t} 1 - \frac{d_j}{R(t_j)}, \tag{2.23}$$

with d_j the number of events at time t_j. c_τ is estimated non-parametrically, thereby adjusting for the censoring bias via inverse probability weighting.

2.5 Repeated nested cross-validation algorithm

In order to have a unified model for estimating an unbiased prediction accuracy and performing feature selection, we proposed a novel algorithm [47] as a contribution of the thesis, which is also summarized in chapter 3. In this section, we give a description of our repeated nested cross-validation approach for classification and the final feature selection using a weighting approach. To obtain a sparser model and estimate generalizability, we first perform feature selection within an inner cross-validation loop and then estimate the prediction performance in an outer cross-validation, see also Figure 2.2. To select important features, we used a ranking approach based on SVMs for classification, which has been performed in our second publication [48].

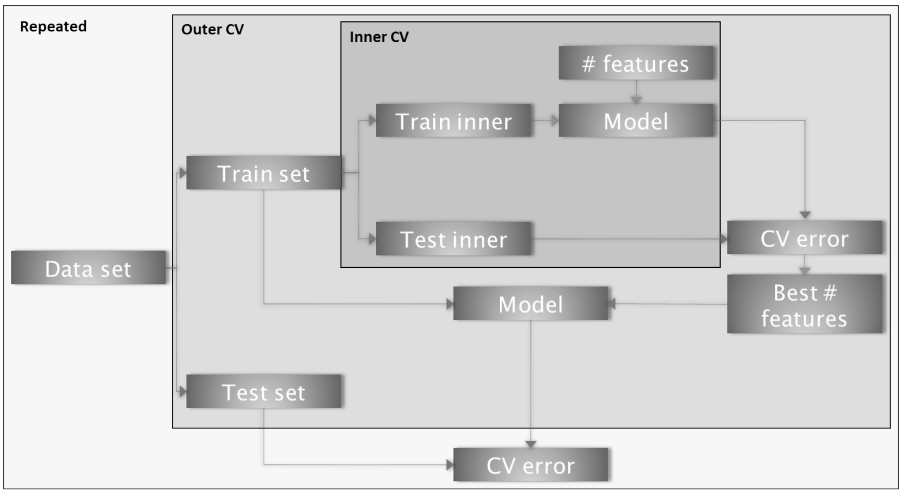


Figure 2.2: Schematic overview of the repeated nested cross-validation approach (taken from [47]). The inner cross-validation is used to determine the optimal number of parameters. In the outer cross-validation loop the unbiased prediction accuracy is estimated. By repetition of the entire procedure the variance of the prediction accuracy is estimated.

Feature ranking SVM

We define $\mathbf{x}^j = (x_{1j}, \ldots, x_{nj})$ with $j = 1, \ldots, p$ as the vector of observations for feature j. In order to generate a ranking on features, we applied recursive feature elimination in a linear kernel SVM approach [38]. This approach uses weights $\hat{\boldsymbol{\beta}}$ from equation 2.19 of a trained SVM to derive a ranking using $r_j = \hat{\beta}_j^2 \ \forall j$ of the features. The feature with the lowest r_j is ranked last and excluded. Then, another SVM is trained on the reduced data, features are ranked and the least important feature is excluded. This procedure is recursively performed until a complete list of ranked features is created.

Repeated nested cross-validation for estimating generalizability

As defined above in 2.1.2 in cross-validation we split the full dataset D with $D_i := (\mathbf{y}_i, \mathbf{x}_i)$ of sample i into subsets of equal size. Since we apply a nested cross-validation procedure, the (outer) training set is denoted as $D^{\overline{cv_{out}}}$ and the (outer) test set $D^{cv_{out}}$ with index set cv_{out} and its complement.

For the *inner* cross-validation we perform a second (nested) cross-validation stratification on the training set $D^{\overline{cv_{out}}}$, obtaining an inner training $D^{(\overline{cv_{out}}, \overline{cv_{in}})}$ and inner test set $D^{(\overline{cv_{out}}, cv_{in})}$. Then, the above-mentioned SVM based ranking function is applied to the inner training set. By adding one feature in a stepwise fashion (according to the ranking), a logistic regression model is fitted using the inner training data and its performance is evaluated with the AUC from equation 2.21 for the inner test data. This stepwise procedure is performed up to a predefined maximum number of features and repeated for all inner cross-validation folds. By averaging over these inner cross-validations the optimal number of features is chosen corresponding to the maximum mean AUC value.

In the *outer* cross-validation a feature ranking is created for the entire training set $D^{-cv_{out}}$. Then, using the best number of features (derived in the inner cross-validation), a logistic regression model is fitted to the training set, thereby obtaining effect estimates $\hat{\boldsymbol{\beta}}$ per selected features. Using these effect estimates, the unbiased prediction performance with the unseen test set $D^{cv_{out}}$ is quantified. We apply this entire procedure (including the inner cross-validation) to all outer cross-validation folds.

To estimate the variability of prediction accuracy, the nested cross-validation approach is repeated q times for different cross-validation splits of the dataset. A pseudo code of the algorithm for survival is shown in [47].

Final model

A ranked set of selected features is obtained for each outer cross-validation run. Additionally, the performance on the unseen test set for each run is recorded (total number of runs $L = cv_{out}q$). Importantly, these ranked lists of selected features are not necessarily the same. In Laimighofer et al. [47], we established a weighted approach that uses the information from all individual cross-validation runs to determine a final combined set of features for which a final model can be fit.

Briefly, this weighted approach uses information from the outer cross-validation test performance c_l corresponding to each run l. The weight w_l of each run l is calculated as follows:

$$w_l = \begin{cases} \frac{1}{L}\exp(\log(2)\frac{devAUC_l}{0.1}), & \text{if } c_l \geq 0.5 \\ 0, & \text{if } c_l < 0.5 \end{cases} \tag{2.24}$$

where $devAUC_l = (c_l - \sum_l c_l L^{-1})/\sum_l c_l L^{-1}$ denotes the relative AUC-statistic of one cross-validation run compared to the averaged performance of all runs. These weights w_l are further normalized to sum to one ($w'_l = w_l/\sum w_l$) [47]. Using majority voting the final set of predictors is given as follows:

$$I(p_j) = \begin{cases} 1 & \text{if } p_j > 0.5 \\ 0, & \text{if } p_j \leq 0.5 \end{cases} \text{ with } p_j = \sum_{l=1}^{L} II(j,l)w'_l \tag{2.25}$$

where the indicator function $II(j,l)$ is 1 if the feature p_j was selected in run l and 0 otherwise.

A final model can be computed with the selected features using the whole dataset, thereby obtaining effect estimates $\hat{\boldsymbol{\beta}}$ and risk scores as in 2.20 for each observation. Moreover the regression estimates $\hat{\boldsymbol{\beta}}$ are used to predict probabilities with unseen data. In our publication, we showed that a similar predictive power on the new data is obtained as estimated in the nested cross-validation.

2.6 Modeling of longitudinal transcriptomics data

In [49], we were interested in differences of gene expression levels of children born by Cesarean section vs. vaginal delivery. Our dataset included longitudinally collected samples of children in the first year of life, where moreover the number of samples varied between children. For each sample, microarray measurements were performed - for details on the dataset see Table 1.1 and [49]. Instead of building a prediction model as described above, we here aimed to gain deeper insights into the pathogenesis of T1D, using functional data analysis [66]. To this end, we estimated for each gene a statistical model described below. Then we used functional gene annotations to increase the statistical power and allow for functional interpretability. In the next part, we will give an introduction to B-splines used in our application.

2.6.1 B-spline functions

Splines are a flexible modeling approach which can account for non-linear effects [67]. The principle idea behind splines is to put piecewise polynomials of certain degrees onto each other, in order to obtain a smooth function. Places where two polynomial functions are connected are known as knots κ with $\kappa_0 \leq x_i \leq \kappa_m \forall i$ with $\mathbf{x} = (x_1, \ldots, x_n)$ of n observations. Usually, knots are equidistantly distributed across the domain of $\mathbf{x}$. Moreover, any polynomial function can be expressed as a linear combination of basis functions B_k (or B-spline basis functions), which guarantee a sufficiently smooth function at the knots and a continuous differentiability [54]. A spline function $g(x_i)$ is defined as a sum of basis functions

$$g(x_i) = \sum_{k=1}^{K} \gamma_k B_k(x_i), \tag{2.26}$$

with γ_k being the coefficient of the basis function B_k. B-splines of degree 0 (a piecewise constant function) are defined as

$$B_k^0(x_i) = I_{\kappa_k, \kappa_{k+1}}(x_i) = \begin{cases} 1, \text{if } \kappa_k \leq x_i < \kappa_{k+1} \\ 0, \text{otherwise} \end{cases} \tag{2.27}$$

for knots κ_k with $k = 1, \ldots, K-1$. Moreover, a general B-spline basis function of degree $p \geq 1$ is defined as

$$B_k^p(x_i) = \frac{x_i - \kappa_k}{\kappa_{k+p} - \kappa_k} B_k^{p-1}(x_i) + \frac{\kappa_{k+p+1} - x_i}{\kappa_{k+p+1} - \kappa_{k+1}} B_{k+1}^{p-1}(x_i). \tag{2.28}$$

From equation 2.28, it can be seen that B-splines are recursively defined and a B-spline of degree p can be expressed as a B-spline of degree $p-1$. In addition, the number of knots indicates the flexibility of the spline function, see also Figure 2.3. Notably, a trade-off between bias (higher number of knots) and variance (smoothness of function g) needs to be fixed. To this end, a ridge penalization factor is typically added to the regression coefficients $\boldsymbol{\gamma}$, yielding a penalized residual sum of squares to be minimized

$$penRSS(\lambda) = \sum_{i=1}^{n} \Big(y_i - \sum_{k=1}^{K} \gamma_k B_k(x_i) \Big)^2 + \lambda \sum_k \gamma_k^2 \tag{2.29}$$

with λ being the smoothing parameter and therefore the trade-off between bias and variance.

2.6.2 Generalized additive mixed model

A 'mixed' model is defined as the combination of a fixed effect and a random effect in one statistical model [68]. We applied this type of model to analyze the gene expression differences between children born by Cesarean section vs. vaginal delivery including three main components into the model: an age effect of the gene expression measurements, multiple measurements per child and the type of delivery. Specifically, the model at single gene level for gene j is defined as

$$\mathbb{E}(\mathbf{y}_{itj} | \mathbf{b}_i j, \mathbf{x}_i t j) = \beta_{0j} + \beta_j^{CS} x_{itj}^{CS} + g(x_{itj}^{age}, \lambda_j) + b_{ij}^{ID} x_{itj}^{ID} + \epsilon_{itj}, \tag{2.30}$$

where $\mathbf{y}_{itj}$ are the gene expression values of children i with $t = 1, \ldots, T_i$ and T_i is the number of samples of children i, β_{0j} is the intercept or gene-wise average expression of

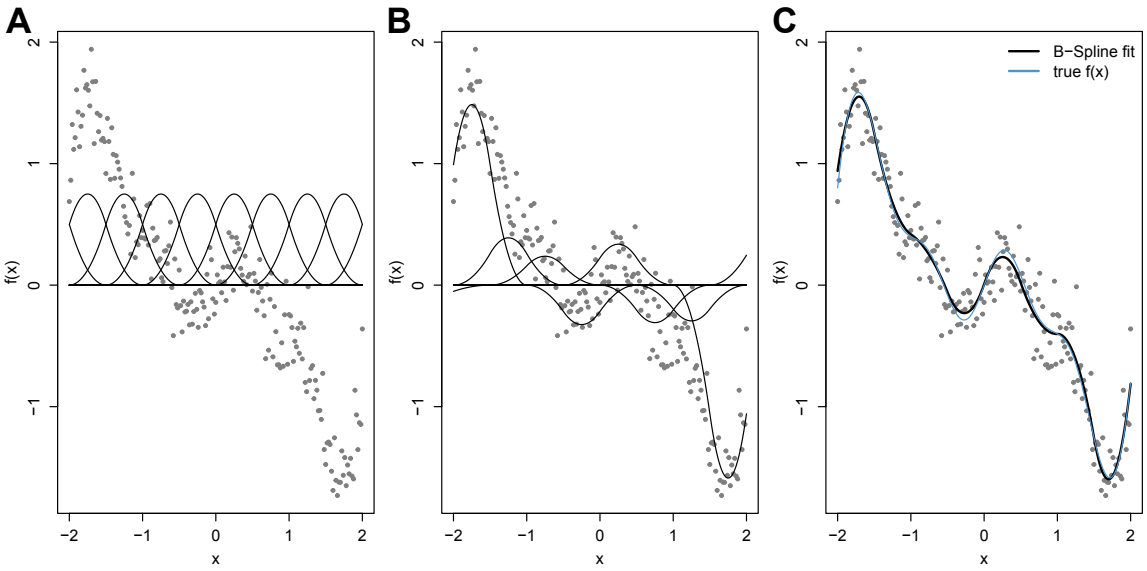


Figure 2.3: Schematic construction of a B-spline. A: B-spline basis functions of degree 2, with equidistant knots of 0.5 steps. B: Scaled B-spline basis function according to fitted regression parameters $\boldsymbol{\gamma}$. C: Sum of scaled B-spline basis functions, resulting in the B-spline fit. Blue line indicates data generating function ($f(x) = (x^5 - 5x^3 + 2x + \sin(2\pi x))/5 + \epsilon$ with $\epsilon \sim N(0, 0.2)$).

gene j and an error term ϵ_{itj} as independent and identical normally distributed noise $\epsilon_{itj} \sim N(0, \sigma_\epsilon \mathbf{I})$. In addition:

i The time dependency $\mathbf{x}^{age}$ of gene expression measurements is modeled using spline functions described above.

ii Multiple measurements per child x_{itj}^{ID} are modeled using a random effect b_{ij}^{ID}. This random effect (random intercept) explains the variance of gene expression within each child as the deviance from the global level β_{0j}. It is defined as $b_{ij}^{ID} \sim N(0, \tau \mathbf{I})$ with unknown variance τ to be estimated. The individual noise of ϵ_{itj} and the noise of the subject specific random effect τ are assumed to be independent.

iii The type of delivery $\mathbf{x}^{CS}$ is modeled using a fixed effect β^{CS} for group differences between children born by CS vs. vaginally delivered children.

Dropping the index per gene j, the log likelihood is defined as

$$l(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{b}, \tau) = \sum_{i=1}^{n} \log \Big(\int f(\mathbf{y}_i | \beta_0, \beta^{CS}) p(\mathbf{b}_i | \tau) d\mathbf{b}_i \Big) - \frac{1}{2} \sum_{k=1}^{m} \lambda_k \boldsymbol{\gamma}_k' \mathbf{K}_k \boldsymbol{\gamma}_k \qquad (2.31)$$

with f being the conditional density of $\mathbf{y}$ in the form of the exponential family (see equation 2.6), for which we assume a Gaussian distribution for gene expression values. $p(\mathbf{b}_i|\tau)$ denotes the density of the random effects and the penalty term of $\boldsymbol{\gamma}_k' \mathbf{K}_k \boldsymbol{\gamma}_k$ on the spline coefficients in included to reduce overfitting. The estimation of the model was performed using a restricted maximum likelihood (REML) approach [69]. For generalized additive mixed models REML based approaches are more appropriate than ML estimation, since ML estimates of the variance are downward biased, as they do not properly take into account the degrees of freedom lost during estimation of fixed effects [70]. Optimization of equation 2.31 is performed using a penalized iterative reweighted least square algorithm [69].

2.6.3 Assessment of functional gene annotations

In order to further analyze the results from the single gene model described above, we aimed to increase statistical power, and biological interpretation by including information of functionally related sets of genes (pathways). Therefore, we used the KEGG database, in which biological, cellular, and molecular related sets of genes are defined [71]. Specifically, we have been interested, if a pathway is differentially expressed in case of CS or vaginal delivery - in particular, if the genes in a pathway are together differentially expressed compared to a random background set of genes. To this end, estimation of differential expression of pathways was performed using the R package 'SAFE' [72]. This method uses a permutation approach in order to assess the statistical significance of differential expression. This permutation procedure accounts for unknown gene correlations, which can substantially influence the pathway associations [73].

Chapter 3

Summary of contributed articles

In this chapter, I provide a detailed summary of the articles that constitute this publication-based dissertation. I am the author in charge of all of these articles. Detailed descriptions of the contributions for each publication are included in the summaries below. Articles i)-iii) have been peer-reviewed and published in international established journals. Moreover, they are not used in any other publication-based dissertation. Article iv) is currently submitted to a peer-reviewed, international journal. A preprint is available online at [https://doi.org/10.1101/167676]. This paper has also not been used in any other publication-based dissertation. The articles are sorted in chronological order. Full texts of these articles can be found in the Appendix.

i) **Michael Laimighofer**, Jan Krumsiek, Florian Buettner, and Fabian J Theis. **Unbiased prediction and feature selection in high-dimensional survival regression**. Journal of Computational Biology, 23(4):279–290, 2016.

   Summary: In this article, I proposed a novel, unified approach for high-dimensional survival regression with two important goals for prediction. First, a selected set of features must obtain high generalizability, that is a predictor of selected features must perform reliably on unseen observations (see Section 2.1.1). Second, I want to derive a sparse set of important features. To this end, a repeated nested cross-validation approach is applied to estimate an accurate prediction accuracy within one data set and to select the most predictive features with a novel feature combination heuristic, see also Figure 2.2 and Section 2.5. Specifically, the dataset is split by an outer cross-validation into an outer training and an outer test set (see Section 2.1.2). The outer training set is further divided by an inner cross-validation into an inner training and inner test set. The aim of this inner cross-validation procedure

is to determine an optimal number of features for predictions on the inner test set. For this purpose, a ranking of features on the inner training data is generated, by ordering the features either uni- and multivariatly according to their association to the survival outcome. Features are then subsequently added in the order of this ranking, and survival Cox proportional hazard models are estimated on each subset. Using these Cox regression models, the prediction performance on the inner test set can be calculated. By averaging over all inner cross-validation folds, the number of features with highest prediction accuracy is determined. Then, the ranking function is applied on the entire outer training data, and a Cox regression model with the derived number of features is fitted on the outer training data. This estimated model is used to predict the outcome of the outer test set such that an unbiased estimate of the prediction accuracy can be obtained. This procedure is repeated for all outer cross-validation folds. Moreover, by repeating the entire nested CV with different random data splits, an estimate of the variance of generalization accuracy is obtained. In order to quantify prediction accuracy, I use a survival AUC accounting for censoring of the data. To select the most predictive features, I aggregate the results from all cross-validation runs. Specifically, a weighting approach of features is applied utilizing the prediction accuracy of all outer cross-validation runs. As the last step, a final set of features is obtained by majority voting.

In a simulation study, I compared our approach with different ranking functions to a standard Cox LASSO model. I could show that the prediction accuracy was reliably measured by our algorithm in an internal validation as well as on external datasets. In the standard Cox LASSO model, I observed a drop of survival AUC for the external datasets, probably due to the fact that the Cox LASSO model selected too many noisy features. In addition, I checked that the 'true' features have been selected. Again, the algorithm obtained higher scores compared to the standard Cox LASSO model.

Finally, I applied our model to three publicly available breast cancer data sets of transcriptomics measurements, predicting overall survival of breast cancer patients after surgery. I used one dataset to internally validate the prediction accuracy and to select the set of features for prediction. Then, I tested the performance of the derived marker set on the other two datasets. Compared to the LASSO survival model, our approach was again able to estimate a reliable prediction accuracy in the internal validation, as well as in the external validation datasets.

In addition to the scientific contributions, I was the author in charge of this publication. I wrote the first complete draft of the paper, and iterated it with Jan Krumsiek,

Florian Buettner and Fabian Theis. Moreover, I implemented the algorithm in R, which is publicly available as an R package on CRAN (*SurvRank*).

ii) Christine von Toerne*[1], **Michael Laimighofer***[1], Peter Achenbach, Andreas Beyerlein, Tonia de las Heras Gala, Jan Krumsiek, Fabian J Theis, Anette G Ziegler, and Stefanie M Hauck. **Peptide serum markers in islet autoantibody-positive children**. Diabetologia, 60 (2):287–295, 2017.

Summary: In this article, I extended our approach of repeated nested cross-validation to classification and applied it on proteomics data from two large German T1D cohorts (BABYDIAB/DIET) (see Section 2.5). A discovery set included prospectively collected samples of autoantibody-negative and autoantibody-positive children - slow and fast progressors, defined by fast ($\leq$3.5 years) or slow ($\geq$9.5 years) progression from autoantibody positivity to T1D onset. This discovery set was used to prioritize peptides, measured in an untargeted shotgun proteomics approach. I used our nested cross-validation approach adapted for a classification setting, in order to identify discriminating peptides between the three different groups (see Section 2.5). Specifically, we applied a support vector machine model which ranks the features with a recursive feature elimination algorithm (see Section 2.3). The goal of this first discovery phase was to define a set of predictive peptides used in the following targeted application phase.

The application dataset consisted of a larger sample set of autoantibody-negative and autoantibody-positive children. To this end, the previously selected peptides were measured in a more sensitive, targeted proteomics approach. First, I compared the peptide abundances of autoantibody-positive and -negative samples using the repeated nested cross-validation for classification with an SVM (similar to the discovery phase), to derive a sparser biomarker predicting autoantibody positivity. Importantly, I obtained two peptides (APOM and APOC4) which discriminate between autoantibody positives and negatives with an unbiased AUC of 0.77, compared to an AUC of 0.75 in the discovery phase. In addition, I investigated the peptide abundances within the autoantibody-positive children to identify a marker predicting disease progression time until T1D onset. To this end, I applied the repeated nested cross-validation approach for survival data (see Section 2.5). Strikingly, a predictive combination consisting of three peptides (HGFAC, CP, and CFH) and age of seroconversion, was identified with an unbiased survival AUC of 0.72. The obtained peptide marker significantly improved prediction of progression time over age alone and the selected peptides did not correlate with age of seroconversion. Based on

these peptides and age of seroconversion, I finally built a risk score identifying high, medium and low risk groups for fast disease progression.

In addition to the scientific contributions, I was the author in charge of this publication, leading statistics, method developments, implementation, and result interpretation. I wrote the first complete draft of the paper, and iterated it with Jan Krumsiek and Fabian Theis.

iii) Brigitte I. Frohnert*[2], **Michael Laimighofer***[2], Jan Krumsiek, Fabian J. Theis, Christiane Winkler, Jill M. Norris, Anette-Gabriele Ziegler, Marian J. Rewers, Andrea K. Steck. **Prediction of type 1 diabetes using a genetic risk model in the Diabetes Autoimmunity Study in the Young (DAISY)**. Pediatric Diabetes, 2017;0:1–7. doi: 10.1111/pedi.12543

Summary: In this article, I applied an existing, previously published [34] genetic risk model for T1D derived from the T1DGC dataset to a new dataset from the DAISY cohort [30], see also Section 2.4. DAISY consists of two subgroups, a general population of HLA high-risk individuals and a second high-risk population of 'first-degree relatives'. We investigated (1) if the genetic risk score is applicable in the DAISY subgroups 'first-degree relatives' and 'general population', (2) if a reduced model with less SNPs is sufficient for prediction of T1D, and (3) if the risk score predicts time to T1D onset.

To examine (1), we used the previously identified weighted genetic risk model, consisting of 9 SNPs and the HLA genotype (10-factor risk model), to estimate the discrimination ability (AUC) within these two subgroups. We showed that the genetic risk model can be applied in DAISY with an AUC of 0.75 for the 'first-degree relatives' and 0.77 for the 'general population. To answer (2), we compared the prediction accuracies of the 10-factor risk model, to a 3-factor model, consisting of HLA, PTPN22, and INS, and to a HLA-only model. To this end, we calculated the 'integrated discrimination improvement' introduced by [74]. For both subgroups, the 3-factor model indicated a significant improvement over the HLA-only model. Interestingly, the 10-factor model only showed a significant improvement for the 'first-degree relatives' subgroup, but not in the 'general population'. We elaborated (3) by using the estimated risk score to predict time from birth to T1D in a survival model. The resulting survival curves indicated a significant difference, and, therefore, predictive ability of time from birth to T1D onset.

In conclusion, the genetic risk score performed well on an independent validation cohort, such as DAISY.

In addition to the scientific contributions, I was the author in charge of this publication, leading statistics, method developments, implementation, and result interpretation. I wrote the first complete draft of the paper, and iterated it with Jan Krumsiek and Fabian Theis.

iv) **Michael Laimighofer**, Ramona Lickert, Rainer Fürst, Fabian J. Theis, Ezio Bonifacio, Anette-Gabriele Ziegler, and Jan Krumsiek. **Common patterns of gene regulation associated with Cesarean section and the development of islet autoimmunity - indications of immune cell activation**. bioRxiv, Cold Spring Harbor Labs Journals, 2017. doi: 10.1101/167676

Summary: In this article, I first investigated differences in gene expression of children born by Cesarean section (CS) compared to vaginally delivered children. To this end, I examined the effect of CS on gene expression differences in the first year of life. Specifically, I estimated a generalized additive mixed model per gene to account for the longitudinally measured samples, different numbers of observations per child, and the effect of CS, see also Section 2.6. After multiple testing, no differentially expressed single genes could be identified. However, using a pathway analysis approach I found two significantly differentially expressed pathways, in particular the 'Pentose phosphate pathway' and the 'Pyrimidine metabolism'.

Second, I analyzed gene expression differences between children shortly after autoantibody development and age-matched controls. In this analysis, I observed numerous differentially expressed genes and various differentially expressed pathways. Interestingly, we observed an overlap for both pathways described above.

Finally, we related the effects of CS and autoantibody positivity, and found a strong correlation (r=0.61) on transcript level. To assess the significance of this association, we performed a permutation-based procedure, resulting in a significant empirical p-value. We also investigated additional covariates, such as gender, maternal diabetes, and multiple 'first-degree relatives with T1D', but the effects of those variables showed no significant correlation with the effects of CS. Moreover, we identified a similar correlation on pathway level between effects of CS and autoantibody positivity (r=0.49). We further investigated the roles of these pathways in the immune activation, and suspected that CS and autoantibody positivity induce an activation of immune cells. This hypothesis is supported by the significant enrichment of innate immune genes [75] of differentially regulated genes in both analyses. In addition, we found empirical evidence of correlation between transcriptomics data of naive vs. activated CD4+ T-cells and the effects from CS and autoantibody status, both on single gene and pathway level.

Taken together, we found a remarkably coherent transcriptional link between CS and autoantibody positivity on single gene and pathway level, further indicating an activation of immune cells.

In addition to the scientific contributions, I was the author in charge of this publication. I wrote the first complete draft of the paper, and iterated it with Jan Krumsiek and Fabian Theis.

*[1] Authors contributed equally. Christine von Toerne performed the experimental part and generated the data. I led the theoretical statistical parts, method developments, implementation, and result interpretation.

*[2] Authors contributed equally. Brigitte Frohnert was responsible for generating the data and data acquisition. I led the theoretical statistical parts, method developments, implementation, and result interpretation.

Chapter 4

Discussion and perspectives

4.1 Summary

The primary goal of my PhD project was to identify novel markers for the pathogenesis of T1D based on high-throughput omics data and to develop novel risk models with reliable prediction accuracy and feature selection.

In the first part of this thesis, we showed the implementation and application of an algorithm for high-dimensional survival data. In the proposed approach, a repeated, nested cross-validation was used to obtain a sparse set of features along with an unbiased estimate of prediction accuracy in a unified fashion. It was published in Laimighofer et al. [47], where we showed that the method has been able to reliably estimate the prediction accuracy. Furthermore, it selects a sparse set of the most predictive features for survival modeling. In a simulation study, we compared our repeated, nested cross-validation to a standard survival LASSO, and revealed that our approach identifies the 'true' features, whereas the LASSO approach produced too many features. In addition, our approach obtained a similar prediction accuracy in the training and test dataset. The method was also applied to three datasets on breast cancer survival, and again a similar prediction performance was obtained for the training breast cancer data, compared to the independent test datasets. Taken together, we developed and published a novel approach for biomarker discovery in high-dimensional survival regression, which is available as a free-to-use R package.

By adapting our repeated nested cross-validation approach to a classification problem in a T1D proteomics dataset, we established a set of biomarkers consisting of two peptides to

distinguish between autoantibody-positive and -negative children. Moreover, we derived a combination of only three peptides and age which were selected to predict progression time from development of autoantibodies to T1D [48]. This combination of features could act as a marker in medical practice where children who recently tested positive for autoantibodies could be classified into slow or fast progressors.

In another publication, we evaluated the discrimination performance of a previously published genetic risk model on the DAISY cohort. For this purpose, we estimated the performance of the weighted SNP model on a genetically high-risk general population of DAISY and samples with 'first-degree relatives' diagnosed with T1D [76]. We confirmed the applicability of the genetic risk score on new cohorts in both subpopulations. In addition, the genetic risk score was used to determine the time until T1D onset. Again, the genetic risk score showed significant prediction accuracy with respect to T1D onset. Of note, the genetic risk score originally derived from a European cohort can be successfully applied to an American cohort.

In the second part of this dissertation focusing on transcriptomics, we presented an approach to identify gene expression differences in a heterogeneous, longitudinally measured, high-dimensional dataset [49]. We established a generalized additive mixed model including spline functions to model the time-dependent measurements, a random effect to account for the random fluctuation of measurement per child, and an effect term for Cesarean section (CS) on transcript expression. In order to aid functional interpretation and increase statistical power, we performed a pathway enrichment analysis of functionally related sets of genes. In a second modeling approach, we investigated the effect of seroconversion on the transcriptome for single genes and for pathways. Finally, we combined these two results of CS and autoantibody status. To this end, we calculated the correlation between the two effects and identified a significant association between CS and autoantibody status using a permutation-based approach [72]. Interestingly, we also detected this correlation on pathway level. The differentially expressed pathways belong to the immune system and indicate an activation of immune cells after CS and after autoantibody development. We further investigated this hypothesis of activation of immune cells, using transcriptomics data of naive and activated human CD4+ T-cells. We observed significant correlations of effects on single gene and on pathway level compared to CS and autoantibody status, providing empirical evidence of immune activation. In conclusion, we found a transcriptional relation between CS and autoantibody status, which may indicate an activation of immune cells.

Taken together, we performed model development for high-dimensional data, detected

novel biomarkers for progression and autoantibody positivity, and identified a possible proliferation signal for both Cesarean section and autoantibody development at transcriptome level.

4.2 Perspectives

In this section, I will provide an outlook on T1D research in the context of high-dimensional data. These perspectives and ideas stem from considerations and experiences gained throughout the duration of my PhD. First, I will examine on the discovery of multi-omics biomarkers in T1D, which questions arise and how to tackle those. Then, I will discuss increasing sample sizes accompanied with statistical challenges and model advancements for 'Big Data'. Finally, I will present an idea for additional data acquisition to further investigate T1D etiology.

4.2.1 Multi-omics biomarker discovery in T1D

All previous studies in T1D research performed biomarker detection at single omics level. For example, in Winkler et al. [34] and in Oram et al. [77], genetic risk scores have been developed in order to predict T1D onset. In other studies, peptide markers have been identified for the discrimination of autoantibody-positive and -negative children [78]. In addition, metabolomics profiles have been investigated to analyze T1D development [79]. We also used single omics data to predict autoantibody status and progression time from proteomics data [48] and T1D onset from genetic data [76]. However, to the best of our knowledge, studies on combined multi-omics measurements have not been performed yet in T1D. Integrating different layers of omics data could give a more comprehensive picture of T1D pathogenesis and allow for the identification of biomarkers from different layers. Ideally, we would concatenate the single omics layer into one big data matrix and apply data analysis algorithms onto this dataset.

To date, various methods for multi-omics integration have been proposed, such as multiple partial least squares (mPLS) [80] [81], sparse canonical correlation analysis (sCCA) [82] [83], and sparse group lasso (sGL) [84] [85]. mPLS and sCCA are related since both methods aim to find linear combinations of input features by identifying latent variables for each omics layer (mPLS) and by making use of the cross-covariance matrices of the individual omics layers (sCCA). Both mPLS and sCCA include a L_1 penalty term to obtain a sparse representation of features. In contrast, the sGL aims to identify a sparse

combination of features per group, i.e. the individual omics layer. However, we observed that simply applying these methods is not applicable in practice, since several issues arose which are discussed in the following.

First, each individual omics layer requires a unique procedure of preprocessing, consisting of quality control, normalization, and imputation techniques, accounting for the different measurement technologies. In order to incorporate the preprocessing steps for each omics layer into a multi-omics framework, the expertise on the different omics measurements needs to be gained and collected, also across research groups. For instance, in our proteomics dataset [48], we had to correct for a decreasing signal strength after the mass spectrograph was cleaned. In addition, we adjusted the measurements for batch effects and normalized for control peptides. In another application on a dataset of measured cytokine expressions, we observed that non-measured values correspond to expressions below the detection limit of the electrochemiluminescence array. In order to account for this missingness pattern, we used this information by applying an appropriate imputation method. In the same cytokine dataset, we observed a seasonal trend of cytokine expression where higher values of cytokine expression have been correlated with samples taken in summer. Thus, as a preprocessing step we corrected the data for seasonal variation. Bundling the preprocessing and data-handling knowledge across omics is a first important step for discovering a multi-omics biomarker, rather than merely appending data matrices.

Second, a large set of overlapping samples where all multi-omics measurements are performed is needed to derive and to compute a multi-omics biomarker. This statement seems trivial, however, in our Munich cohort, we observed only a small overlap of measured samples across omics levels. The individual omics datasets have been measured with different outcome definitions, such as T1D onset [31], autoantibody status, or progression rate [32]. For instance, to investigate progression time, a sample of the child is needed close to seroconversion, otherwise we cannot use it in this analysis. The change of interest for different outcomes, has also been driven by discoveries made by the T1D research community. For example, risk factors of T1D onset have been identified, such as CS. This discovery has led to further data acquisitions of the transcriptome, in order to investigate the effects of CS on the gene expression. With the rise of high-throughput technologies generating sufficient sample sizes with overlapping measurements across omics is getting cheaper and easier, therefore, aiming for multi-omics biomarker discovery in T1D.

Third, concatenating individual omics layers increases the high-dimensionality ($p >> n$) of the dataset even further. Even strong regularization methods are not able to detect the true underlying features shown in simulation studies for $p >> n$ [55]. In order to deal with

these methodological shortcomings, the developments in high-throughput technologies allow for fast and cheap to generate omics measurements of new samples. This progress in technology has led to an increasing number of measured samples. Moreover, these efforts of having both p and n large lead to so called 'Big data' which is discussed in the following section.

Taken together, to discover a multi-omics biomarker in T1D we could apply one of the aforementioned methods taking into consideration the prerequisites of bundling the preprocessing and data-handling knowledge per omics layer, of generating sufficient sample sizes with overlapping omics measurements, and of increasing the overall sample size with further development in high-throughput technologies.

4.2.2 Promises and challenges of 'Big data'

The increase of sample sizes leads to both large p and large n, denoted as 'Big data' [86]. In recent years 'Big data'-sets have emerged in medical and biological research, since high-throughput measurements are fast and cheap to generate. However, these 'Big data'-sets are accompanied by promises and challenges, more details in Fan et al. [86], Rossell [87], and Alyass et al. [88].

One promise of 'Big data' is that such datasets allow for a better understanding of similarity and heterogeneity in populations and therefore allow for identification of new subgroups [86] [89]. In small datasets, samples might be denoted as outliers if they show strong dissimilarities compared to all other samples; e.g. in our transcriptomics data, we had to exclude several samples, since these samples showed distinct correlation of their gene expression patterns. In large cohorts, these previously identified outliers might form a novel subpopulation with similar gene expression characteristics. This identification of subpopulations may lead to targeted treatment strategies[90].

Another promise of 'Big data' is to detect statistical interactions between and within omics layers and to obtain reliable estimates thereof. These interaction effects include effects between genes and environmental factors, e.g. in T1D research an interaction has been reported between Cesarean section and a SNP [23]. In this publication, they showed that children born by CS and having a homozygous (GG vs. GA and AA) genotype of the interferon-induced helicase 1 gene have a significant increased risk of T1D compared to children without this combination. In order to systematically identify such interaction effects, larger sample sizes are needed, since the interaction term usually exhibits higher

variance than the main effects [91]. In other words, to detect interaction effects more statistical power is required and, therefore, the generation of 'Big data' is necessary.

One major challenge of 'Big data' is the increased computational burden with large datasets. On the one hand, the huge amount of measured data has led to large data storage systems and appropriate computing infrastructures. On the other hand, algorithms and suitable statistical methods have to be developed and applied which are able to process and analyze 'Big data' in suitable time. This implies that algorithms have to upscale with increased sample size and feature space. Routinely used methods, such as algorithms for matrix decompositions, do not work in 'Big data' models, and therefore, scalability is not given [92]. The scalability of algorithms to large scale problems is closely related with the necessity of parallelization [93] [94]. If the optimization of an algorithm can be performed in parallel on numerous computer cores, computation time can be significantly decreased. For instance, in our 'SurvRank'-package, we implemented parallelization strategies, in order to speed up computation time. To tackle the computational challenge of 'Big data', we need an appropriate computer infrastructure to store the data and algorithmic development including parallelization [93], to spread the computational workload of large datasets.

As an outlook for the increasing sample sizes in T1D, the 'Fr1da' study has been launched in Bavaria with the aim to screen all newborns for islet autoantibodies as precursors for T1D onset [95] to prevent hypoglycemic events, to educate the affected families, and to develop new preventive therapies. At-risk identified children may be further enrolled in the follow-up study 'Pre-POINT', investigating the effect of oral insulin dose - a potential candidate of vaccination for T1D [96]. Similar studies will be started in other federal states of Germany. These efforts may lead to large n samples also in T1D research.

4.2.3 Deep learning models for 'Big data'

Progress has been also made in the field of machine learning, where novel methods for classification and prediction models are developed for the big data context.

One particular class of models, deep neural networks [97], has gained a lot of attention, since it outperforms classical methods and showed high prediction accuracy with unseen data [98]. Typically, these deep neural networks consist of an input layer, the data matrix, a number of hidden layers and an output layer, relating the input with the response variable. Within the hidden layers, linear combinations of features are created. In addition, non-linear transformation functions, so called activation functions, are applied onto these

linear combinations. By design, these activation functions build non-linear interactions of the input features, which we do not cover with linear models. Deep neural networks are usually applied in problems where prediction without interpretation is the main goal [55], where these models exploit the non-linearity in features and outperform other methods [99] [100]. This strong non-linearity of deep neural networks may help to reveal multifaceted interaction effects between individual omics layers and environmental factors, as described in Filippi and von Herrath [101] and Biros et al. [102] for T1D. Moreover, deep neural networks need a large number of observations n in order to be accurately trained.

For currently available T1D datasets, the number of observations is too small to apply deep neural networks. In the future with additional measurements and combined cohorts, this approach may be a promising alternative which could lead to improved predictions of T1D onset, for instance, prediction of autoantibody development long before seroconversion.

4.2.4 Requirements for data acquisition in T1D research

Having multiple omics and large sample sizes, still the puzzle of T1D might not be solved. Since it is a complex disease, we need more longitudinal data and a different data acquisition, in order to understand the etiology of T1D. Here, we want to describe a possible data acquisition procedure to ideally identify the mechanisms, which lead to development of autoantibodies and furthermore to T1D onset.

As described in the introduction, T1D is caused by a combination of a genetic background and environmental factors. One major theory about T1D pathogenesis is based on the hypothesis that events trigger the immune system leading to incorrectly attacking the beta cells in the pancreas. Respiratory viral infections, early childhood nutrition, or environmental changes are hypothesized to be such trigger events. In order to investigate the effects of such trigger events on the immune system, a longitudinal gene expression dataset in addition to a time dependent questionnaire, asking for recent viral infections, nutritional details, and changes in the environment, may help to detect changes of the gene expression induced by trigger events. An additional data resource might be mobile apps to directly gather data from T1D-risk children, recording their lifestyle, habitues, and nutrition details.

4.3 Conclusions

Type 1 Diabetes is a complex autoimmune disease with genetic and environmental components. The rise of high-throughput measurement methods of omics data led to the widespread generation of high-dimensional datasets. In this thesis, we identified biomarkers for T1D disease progression, proposed a novel algorithm in high-dimensional survival regression, and detected a genetic link between CS and autoantibody development. We thereby obtained novel markers and new insights into the pathogenesis of T1D.

Bibliography

[1] Dyanne P Westerberg. Diabetic ketoacidosis: evaluation and treatment. *American family physician*, 87(5), 2013.

[2] Marcus Lind, Ann-Marie Svensson, Mikhail Kosiborod, Soffia Gudbjörnsdottir, Aldina Pivodic, Hans Wedel, Sofia Dahlqvist, Mark Clements, and Annika Rosengren. Glycemic control and excess mortality in type 1 diabetes. *New England Journal of Medicine*, 371(21):1972–1982, 2014.

[3] AM Jacobson, CM Ryan, PA Cleary, BH Waberski, K Weinger, G Musen, W Dahms, DCCT/EDIC Research Group, et al. Biomedical risk factors for decreased cognitive functioning in type 1 diabetes: an 18 year follow-up of the diabetes control and complications trial (dcct) cohort. *Diabetologia*, 54(2):245–255, 2011.

[4] B Milton, P Holland, and M Whitehead. The social and economic consequences of childhood-onset type 1 diabetes mellitus across the lifecourse: a systematic review. *Diabetic Medicine*, 23(8):821–829, 2006.

[5] Andrew Tomlin, Susan Dovey, and Murray Tilyard. Health outcomes for diabetes patients returning for three annual general practice checks. *The New Zealand Medical Journal (Online)*, 120(1252), 2007.

[6] Jean M Lawrence, Giuseppina Imperatore, Dana Dabelea, Elizabeth J Mayer-Davis, Barbara Linder, Sharon Saydah, Georgeanna J Klingensmith, Lawrence Dolan, Debra A Standiford, Catherine Pihoker, et al. Trends in incidence of type 1 diabetes among non-hispanic white youth in the united states, 2002-2009. *Diabetes*, page DB_131891, 2014.

[7] Christopher F Jasinski, Rosa Rodriguez-Monguio, Ksenia Tonyushkina, and Holley Allen. Healthcare cost of type 1 diabetes mellitus in new-onset children in a hospital compared to an outpatient setting. *BMC pediatrics*, 13(1):1, 2013.

[8] Jeffrey A Bluestone, Kevan Herold, and George Eisenbarth. Genetics, pathogenesis and clinical interventions in type [thinsp] 1 diabetes. *Nature*, 464(7293):1293–1300, 2010.

[9] P Narendran, E Estella, and S Fourlanos. Immunology of type 1 diabetes. *Qjm*, 98 (8):547–556, 2005.

[10] Ezio Bonifacio. Predicting type 1 diabetes using biomarkers. *Diabetes Care*, 38(6): 989–996, 2015.

[11] Anette G Ziegler, Marian Rewers, Olli Simell, Tuula Simell, Johanna Lempainen, Andrea Steck, Christiane Winkler, Jorma Ilonen, Riitta Veijola, Mikael Knip, et al. Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. *Jama*, 309(23):2473–2479, 2013.

[12] AE Butler, R Galasso, JJ Meier, Rita Basu, Robert Allan Rizza, and PC Butler. Modestly increased beta cell apoptosis but no increased beta cell replication in recent-onset type 1 diabetic patients who died of diabetic ketoacidosis. *Diabetologia*, 50(11):2323–2331, 2007.

[13] Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Cécile Julier, Grant Morahan, Jørn Nerup, Concepcion Nierras, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics*, 41(6):703–707, 2009.

[14] Kendra Vehik, Richard F Hamman, Dennis Lezotte, Jill M Norris, Georgeanna J Klingensmith, Marian Rewers, and Dana Dabelea. Trends in high-risk hla susceptibility genes among colorado youth with type 1 diabetes. *Diabetes Care*, 31(7): 1392–1396, 2008.

[15] A Paul Lambert, Kathleen M Gillespie, Glenys Thomson, Heather J Cordell, John A Todd, Edwin AM Gale, and Polly J Bingley. Absolute risk of childhood-onset type 1 diabetes defined by human leukocyte antigen class ii genotype: a population-based study in the united kingdom. *The Journal of Clinical Endocrinology & Metabolism*, 89(8):4037–4043, 2004.

[16] Caroline A Brorsson, Flemming Pociot, Type 1 Diabetes Genetics Consortium, et al. Shared genetic basis for type 1 diabetes, islet autoantibodies, and autoantibodies associated with other immune-mediated diseases in families with type 1 diabetes. *Diabetes Care*, 38(Supplement 2):S8–S13, 2015.

[17] Ezio Bonifacio and Anette G Ziegler. Advances in the prediction and natural history of type 1 diabetes. *Endocrinology and metabolism clinics of North America*, 39(3): 513–525, 2010.

[18] Maria J Redondo, Joy Jeffrey, Pamela R Fain, George S Eisenbarth, and Tihamer Orban. Concordance for islet autoimmunity among monozygotic twins. *New England Journal of Medicine*, 359(26):2849–2850, 2008.

[19] Anette-G Ziegler, Sandra Schmid, Doris Huber, Michael Hummel, and Ezio Bonifacio. Early infant feeding and risk of developing type 1 diabetes–associated autoantibodies. *Jama*, 290(13):1721–1728, 2003.

[20] Chris R Cardwell, Lars C Stene, Johnny Ludvigsson, Joachim Rosenbauer, Ondrej Cinek, Jannet Svensson, Francisco Perez-Bravo, Anjum Memon, Suely G Gimeno, Emma JK Wadsworth, et al. Breast-feeding and childhood-onset type 1 diabetes. *Diabetes care*, page DC_120438, 2012.

[21] Andreas Beyerlein, Ewan Donnachie, Sibille Jergens, and Anette-Gabriele Ziegler. Infections in early life and development of type 1 diabetes. *JAMA*, 315(17):1899–1901, 2016.

[22] CR Cardwell, LC Stene, G Joner, O Cinek, J Svensson, MJ Goldacre, RC Parslow, P Pozzilli, G Brigis, D Stoyanov, et al. Caesarean section is associated with an increased risk of childhood-onset type 1 diabetes mellitus: a meta-analysis of observational studies. *Diabetologia*, 51(5):726–735, 2008.

[23] Ezio Bonifacio, Katharina Warncke, Christiane Winkler, Maike Wallner, and Anette-G Ziegler. Cesarean section and interferon-induced helicase gene polymorphisms combine to increase childhood type 1 diabetes risk. *Diabetes*, 60(12):3300–3306, 2011.

[24] Nils Gehlenborg, Seán I O'Donoghue, Nitin S Baliga, Alexander Goesmann, Matthew A Hibbs, Hiroaki Kitano, Oliver Kohlbacher, Heiko Neuweger, Reinhard Schneider, Dan Tenenbaum, et al. Visualization of omics data for systems biology. *Nature methods*, 7:S56–S68, 2010.

[25] Francesco D Girolamo, Isabella Lante, Maurizio Muraca, and Lorenza Putignani. The role of mass spectrometry in the "omics" era. *Current organic chemistry*, 17 (23):2891–2905, 2013.

[26] Mete Civelek and Aldons J Lusis. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–48, 2014.

[27] Oliver Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology*, 48(1-2):155–171, 2002.

[28] Jan Krumsiek, Jörg Bartel, and Fabian J Theis. Computational approaches for systems metabolomics. *Current opinion in biotechnology*, 39:198–206, 2016.

[29] Stephen S Rich, Patrick Concannon, Henry Erlich, Cecile Julier, Grant Morahan, Jorn Nerup, Flemming Pociot, and John A Todd. The type 1 diabetes genetics consortium. *Annals of the New York Academy of Sciences*, 1079(1):1–8, 2006.

[30] M Rewers, TL Bugawan, JM Norris, A Blair, B Beaty, M Hoffman, RS McDuffie, RF Hamman, G Klingensmith, GS Eisenbarth, et al. Newborn screening for hla markers associated with iddm: diabetes autoimmunity study in the young (daisy). *Diabetologia*, 39(7):807–812, 1996.

[31] Anette-G Ziegler, Michael Hummel, Michael Schenker, and Ezio Bonifacio. Autoantibody appearance and risk for development of childhood diabetes in offspring of parents with type 1 diabetes: the 2-year analysis of the german babydiab study. *Diabetes*, 48(3):460–468, 1999.

[32] Sandra Hummel, Maren Pflüger, Michael Hummel, Ezio Bonifacio, and Anette-G Ziegler. Primary dietary intervention study to reduce the risk of islet autoimmunity in children at increased risk for type 1 diabetes the babydiet study. *Diabetes care*, 34(6):1301–1305, 2011.

[33] William A Hagopian, Åke Lernmark, Marian J Rewers, Olli G Simell, JIN-XIONG SHE, Anette G Ziegler, Jeffrey P Krischer, and Beena Akolkar. Teddy–the environmental determinants of diabetes in the young. *Annals of the New York Academy of Sciences*, 1079(1):320–326, 2006.

[34] Christiane Winkler, Jan Krumsiek, Florian Buettner, Christof Angermüller, Eleni Z Giannopoulou, Fabian J Theis, Anette-Gabriele Ziegler, and Ezio Bonifacio. Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes. *Diabetologia*, 57(12):2521–2529, 2014.

[35] Lue Ping Zhao, Hamid Bolouri, Michael Zhao, Daniel E Geraghty, and Åke Lernmark. An object-oriented regression for building disease predictive models with multiallelic hla genes. *Genetic epidemiology*, 40(4):315–332, 2016.

[36] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

[37] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[38] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46 (1-3):389–422, 2002.

[39] Anne-Laure Boulesteix and Korbinian Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1):32–44, 2007.

[40] Donghwan Lee, Woojoo Lee, Youngjo Lee, and Yudi Pawitan. Super-sparse principal component analyses for high-throughput genomic data. *BMC bioinformatics*, 11(1): 1, 2010.

[41] Lue Ping Zhao and Hamid Bolouri. Object-oriented regression for building predictive models with high dimensional omics data from translational studies. *Journal of biomedical informatics*, 60:431–445, 2016.

[42] Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

[43] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[44] Faming Liang, Qifan Song, and Kai Yu. Bayesian subset modeling for high-dimensional generalized linear models. *Journal of the American Statistical Association*, 108(502):589–606, 2013.

[45] Paul Yau, Robert Kohn, and Sally Wood. Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics*, 2012.

[46] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.

[47] Michael Laimighofer, Jan Krumsiek, Florian Buettner, and Fabian J Theis. Unbiased prediction and feature selection in high-dimensional survival regression. *Journal of Computational Biology*, 23(4):279–290, 2016.

[48] Christine von Toerne, Michael Laimighofer, Peter Achenbach, Andreas Beyerlein, Tonia de las Heras Gala, Jan Krumsiek, Fabian J Theis, Anette G Ziegler, and

Stefanie M Hauck. Peptide serum markers in islet autoantibody-positive children. *Diabetologia*, 60(2):287–295, 2017.

[49] Michael Laimighofer, Ramona Lickert, Rainer Fürst, Fabian Theis, Christiane Winkler, Ezio Bonifacio, Anette-Gabriele Ziegler, and Jan Krumsiek. Common patterns of gene regulation associated with cesarean section and the development of islet autoimmunity — indications of immune cell activation. *bioRxiv*, 2017. doi: 10.1101/167676. URL `http://www.biorxiv.org/content/early/2017/07/24/167676`.

[50] Chong Ho Yu. Resampling methods: concepts, applications, and justification. *Practical Assessment, Research & Evaluation*, 8(19):1–23, 2003.

[51] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[52] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[53] Patrice Bertail, Dimitris N Politis, and Joseph P Romano. On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association*, 94(446):569–579, 1999.

[54] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression: models, methods and applications*. Springer Science & Business Media, 2013.

[55] Trevor J.. Hastie, Robert John Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2011.

[56] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33 (1):1, 2010.

[57] Melvin L Moeschberger and J Klein. *Survival analysis: Techniques for censored and truncated data: Statistics for Biology and Health*. Springer, 2003.

[58] Robert Tibshirani and others. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[59] Noah Simon, Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. ISSN 1548-7660. URL `http://www.jstatsoft.org/v39/i05`.

[60] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[61] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[62] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[63] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and LJ Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[64] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

[65] Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[66] Daniel J Levitin, Regina L Nuzzo, Bradley W Vines, and JO Ramsay. Introduction to functional data analysis. *Canadian Psychology/Psychologie canadienne*, 48(3): 135, 2007.

[67] Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

[68] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.

[69] Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011.

[70] Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica*, pages 731–761, 2004.

[71] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[72] William T Barry, Andrew B Nobel, and Fred A Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.

[73] Daniel M Gatti, William T Barry, Andrew B Nobel, Ivan Rusyn, and Fred A Wright. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC genomics*, 11(1):574, 2010.

[74] Michael J Pencina, Ralph B D'Agostino, and Olga V Demler. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Statistics in medicine*, 31(2):101–113, 2012.

[75] Karin Breuer, Amir K Foroushani, Matthew R Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L Winsor, Robert EW Hancock, Fiona SL Brinkman, and David J Lynn. Innatedb: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic acids research*, page gks1147, 2012.

[76] Brigitte I Frohnert, Michael Laimighofer, Jan Krumsiek, Fabian J Theis, Christiane Winkler, Jill M Norris, Anette-Gabriele Ziegler, Marian J Rewers, and Andrea K Steck. Prediction of type 1 diabetes using a genetic risk model in the diabetes autoimmunity study in the young. *Pediatric Diabetes*, 2017.

[77] Richard A Oram, Kashyap Patel, Anita Hill, Beverley Shields, Timothy J McDonald, Angus Jones, Andrew T Hattersley, and Michael N Weedon. A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. *Diabetes care*, 39(3):337–344, 2016.

[78] Robert Moulder, Santosh D Bhosale, Timo Erkkilä, Essi Laajala, Jussi Salmi, Elizabeth V Nguyen, Henna Kallionpää, Juha Mykkänen, Mari Vähä-Mäkilä, Heikki Hyöty, et al. Serum proteomes distinguish children developing type 1 diabetes in a cohort with hla-conferred susceptibility. *Diabetes*, page db140983, 2015.

[79] Brigitte I Frohnert and Marian J Rewers. Metabolomics in childhood diabetes. *Pediatric diabetes*, 17(1):3–14, 2016.

[80] Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1):35–2, 2008.

[81] Wenyuan Li, Shihua Zhang, Chun-Chi Liu, and Xianghong Jasmine Zhou. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, 28(19):2458–2466, 2012.

[82] Dongdong Lin, Jigang Zhang, Jingyao Li, Vince D Calhoun, Hong-Wen Deng, and Yu-Ping Wang. Group sparse canonical correlation analysis for genomic data integration. *BMC bioinformatics*, 14(1):245, 2013.

[83] Samuel M Gross and Robert Tibshirani. Collaborative regression. *Biostatistics*, 16 (2):326–338, 2015.

[84] Silvia Pineda, Francisco X Real, Manolis Kogevinas, Alfredo Carrato, Stephen J Chanock, Núria Malats, and Kristel Van Steen. Integration analysis of three omics data using penalized regression methods: An application to bladder cancer. *PLoS Genet*, 11(12):e1005689, 2015.

[85] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[86] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[87] David Rossell. Big data and statistics: a statistician's perspective. *Metode science studies journal: annual review*, 5:143, 2015.

[88] Akram Alyass, Michelle Turcotte, and David Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics*, 8(1):33, 2015.

[89] Walid M Abdelmoula, Benjamin Balluff, Sonja Englert, Jouke Dijkstra, Marcel JT Reinders, Axel Walch, Liam A McDonnell, and Boudewijn PF Lelieveldt. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences*, page 201510227, 2016.

[90] Nirupa Murugaesu, Su Kit Chew, and Charles Swanton. Adapting clinical paradigms to the challenges of cancer clonal evolution. *The American journal of pathology*, 182 (6):1962–1971, 2013.

[91] Andrew C Leon and Moonseong Heo. Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Computational statistics & data analysis*, 53(3):603–608, 2009.

[92] Volkan Cevher, Stephen Becker, and Mark Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Processing Magazine*, 31(5):32–43, 2014.

[93] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.

[94] Konstantinos Slavakis, Georgios B Giannakis, and Gonzalo Mateos. Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. *IEEE Signal Processing Magazine*, 31(5):18–31, 2014.

[95] Fr1da, 2016. URL `https://www.typ1diabetes-frueherkennung.de/fileadmin/FRIEDA/PDF/Fr1da_Standard_24.02.2016_fuer_Homepage.pdf`.

[96] AG Ziegler, T Danne, DB Dunger, R Berner, R Puff, W Kiess, G Agiostratidou, JA Todd, and E Bonifacio. Primary prevention of beta-cell autoimmunity and type 1 diabetes–the global platform for the prevention of autoimmune diabetes (gppad) perspectives. *Molecular metabolism*, 5(4):255–262, 2016.

[97] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[98] Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7):2524–2530, 2016.

[99] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015.

[100] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.

[101] Christophe Filippi and Matthias von Herrath. How viral infections affect the autoimmune process leading to type 1 diabetes. *Cellular immunology*, 233(2):125–132, 2005.

[102] Erik Biros, Margaret A Jordan, and Alan G Baxter. Genes mediating environment interactions in type 1 diabetes. *Rev Diabet Stud*, 2(4):192–207, 2005.

Unbiased prediction and feature selection in high-dimensional survival regression.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 23, Number 4, 2016
Mary Ann Liebert, Inc.
Pp. 279–290
DOI: 10.1089/cmb.2015.0192

Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression

MICHAEL LAIMIGHOFER,[1,2] JAN KRUMSIEK,[1,3]
FLORIAN BUETTNER,[1,4] and FABIAN J. THEIS[1,2]

ABSTRACT

With widespread availability of omics profiling techniques, the analysis and interpretation of high-dimensional omics data, for example, for biomarkers, is becoming an increasingly important part of clinical medicine because such datasets constitute a promising resource for predicting survival outcomes. However, early experience has shown that biomarkers often generalize poorly. Thus, it is crucial that models are not overfitted and give accurate results with new data. In addition, reliable detection of multivariate biomarkers with high predictive power (feature selection) is of particular interest in clinical settings. We present an approach that addresses both aspects in high-dimensional survival models. Within a nested cross-validation (CV), we fit a survival model, evaluate a dataset in an unbiased fashion, and select features with the best predictive power by applying a weighted combination of CV runs. We evaluate our approach using simulated toy data, as well as three breast cancer datasets, to predict the survival of breast cancer patients after treatment. In all datasets, we achieve more reliable estimation of predictive power for unseen cases and better predictive performance compared to the standard CoxLasso model. Taken together, we present a comprehensive and flexible framework for survival models, including performance estimation, final feature selection, and final model construction. The proposed algorithm is implemented in an open source R package (SurvRank) available on CRAN.

Key words: feature selection, high-dimensional survival regression, repeated nested cross validation.

1. INTRODUCTION

In past years, new experimental technologies that allow measurement of tens of thousands of SNPs, transcripts, peptides, and metabolites in a cost-effective, high-throughput fashion have been developed. Consequently, omics measurements in patient samples are increasingly becoming part of clinical trials (McShane et al., 2013), because they promise to serve diagnostic purposes and accurately model patient

[1]Institute of Computational Biology, Helmholtz-Zentrum München, Neuherberg, Germany.
[2]Department of Mathematics, TU München, Garching, Germany.
[3]German Center for Diabetes Research (DZD), München-Neuherberg, Germany.
[4]European Bioinformatics Institute, European Molecular Biology Laboratory Hinxton, Cambridge, United Kingdom.

survival times. However, for such survival models to be adopted in clinical practice and diagnosis, it is crucial to accurately estimate the generalizability of these models (i.e., how well they perform with new patient cohorts). In addition, identification of a small set of highly predictive features in a high-dimensional survival setting is of particular clinical interest as it can facilitate large-scale screening of large patient cohorts. Example applications include identification of genetic marker sets to predict survival times after surgery in cancer research (Desmedt et al., 2007; van de Vijver et al., 2002) and the prediction of time to diabetes onset (Abbasi et al., 2012).

In high-dimensional medical datasets, the number of features p usually far exceeds the number of observations n ($n << p$). Several previous studies have addressed the $n << p$ problem in survival settings using regularization or feature selection approaches. Some authors have combined test statistics from univariate analyses into risk scores, for example, for lung cancer (Beer et al., 2002) and colorectal cancer (Eschrich et al., 2005). A drawback of these approaches is that each feature is individually associated with survival; however, joint information across features is not used. With polygenic risk scores or multivariate biomarkers, interest in full multivariable models has increased. As standard regression-based models are prone to overfitting in the $n << p$ scenario, shrinkage-based models, which regularize the effect estimates, are commonly used (Gui and Li, 2005; Wu et al., 2011; Gong et al., 2014; Datta et al., 2007). Alternatively, dimensionality reduction (e.g., PCA or clustering) can be performed prior to survival modeling (Alizadeh et al., 2000; Takamizawa et al., 2004; Zhao et al., 2005).

Here, we propose an approach that tackles two major challenges for predictive survival models in a single unified algorithm. TASK 1: A predictor must show good generalizability, that is it must correctly predict an outcome using unseen observations. Here, we aim to obtain unbiased predictions using only training data, that is in the absence of a validation dataset. The generalizability of this type of prediction model can be quantified using measures such as the concordance index (C-index) within a cross-validation (CV) framework for survival data (Harrell et al., 1982). For applicability in clinical settings, it is crucial to estimate this predictive power for new, unseen patients in an unbiased fashion. TASK 2: We aim to select a reduced set of informative features that retains high predictive accuracy. While different approaches to address these tasks in binary classification settings exist, to the best of our knowledge, there is no unified framework for high-dimensional survival settings.

We use a repeated nested CV strategy to tackle both tasks (Fig. 1). Specifically, we use a feature ranking-based approach to perform model selection followed by determination of the optimal number of features in the inner CV loop. The outer CV is used to estimate the prediction accuracy with the C-index with unseen data. By repeating the entire procedure, we quantify the intrinsic variation in the prediction accuracy. As

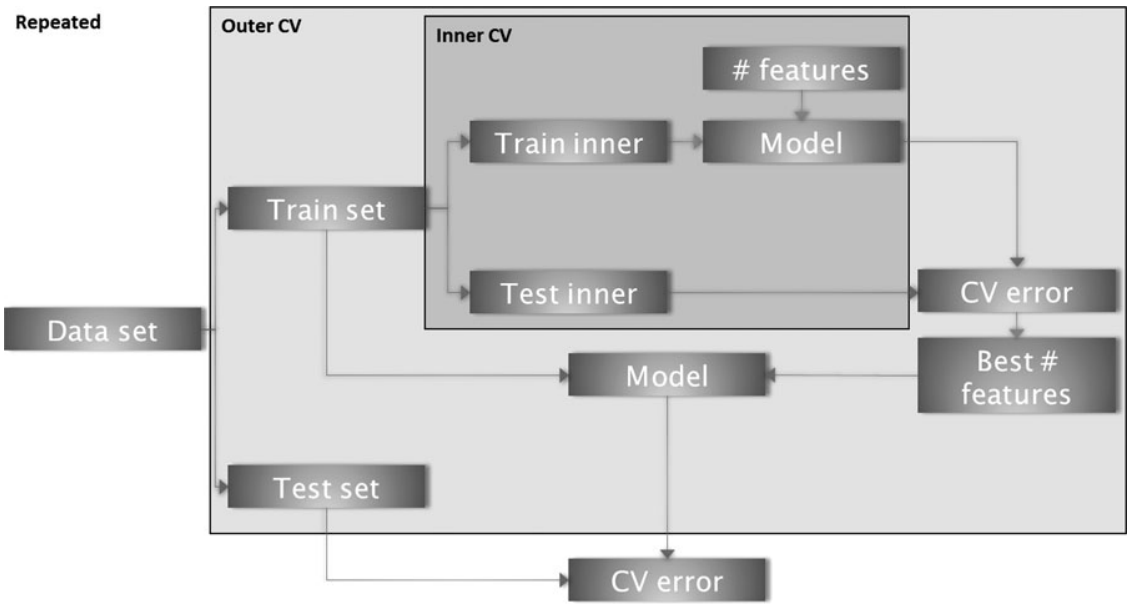


FIG. 1. Overview of the repeated nested Cross-validation scheme. In the inner CV, the optimal number of parameters is determined. The outer CV loop estimates unbiased prediction accuracy. The variance of the prediction accuracy is estimated by repeating the entire procedure.

different CV folds will produce different lists of feature rankings, we propose an algorithm to combine results. We weight the features according to their performance in the CV. TASK 1 is addressed by our method due to the strict separation of the training and test sets. We solve TASK 2 using our proposed approach to aggregate CV information into a final set of features.

We evaluate our approach with simulated data with a fixed set of features and show that existing methods (a regularized survival Cox model) exhibit strong bias. In addition, we test performance with three publicly available breast cancer datasets. These microarray-based datasets contain gene expression data from patients with lymph node-negative breast cancer after surgery or radiotherapy. Our pipeline is available as an R package (R Core, Team, 2014) SurvRank online.

2. METHODS

A survival dataset is defined by the triple $(T_i, \delta_i, \mathbf{x}_i)$ $i=1,\ldots,n$ subjects, where T_i is the observed time (either failure time or censoring time), $\delta_i \in \{0, 1\}$ denotes the censoring indicator for a failure event (e.g., $\delta_i = 1$ in the case of relapse or death) or censoring information ($\delta_i = 0$), and the p-dimensional vector $\mathbf{x}_i$ defines the observed covariates of subject i. A subject is at risk if it undergoes an event or is censored. With $t_1 < \ldots < t_m$ being the ordered unique event times (with $\delta_i = 1$), at time t_j, all at-risk individuals constitute the risk set $R(t_j)$, which is defined as the set of all observations with longer observation time $T_i > t_j$.

In order to relate survival and observed covariates in our algorithm, we use the Cox proportional hazards model (Cox, 1972). In this model, the hazard for subject i is defined in semi-parametric form:

$$h_i(t|\mathbf{x_i}) = h_0(t) \exp\left(\sum_{k=1}^{p} x_{i,k}\beta_k\right) \tag{2.1}$$

where h_0 is a common baseline hazard and $\boldsymbol{\beta}$ is a vector of regression coefficients of length p. Inference on $\boldsymbol{\beta}$ is performed by maximizing the partial likelihood, defined as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left(\sum_{k=1}^{p} x_{i,k}\beta_k\right)}{\sum_{j \in R_i} \exp\left(\sum_{k=1}^{p} x_{j,k}\beta_k\right)} \tag{2.2}$$

where the baseline hazard $h_0(t)$ cancels out. The estimated risk score per subject is summarized by $\hat{\eta}_i = \sum_{k=1}^{p} x_{k,i}\hat{\beta}_k$, which expresses the linear combination of covariates with an estimated vector of coefficients $\hat{\boldsymbol{\beta}}$.

Investigation of the prediction accuracy and feature selection in our framework is performed with the C-index definition (Uno et al., 2011), denoted as C_{Uno}. The C-index of Uno for a prespecified point in time τ is defined as follows:

$$C_{Uno,\tau} = \frac{\sum_{j,k} \hat{G}(T_j)^{-2} I(T_j < T_k) I(\hat{\eta}_j < \hat{\eta}_k) \delta_j}{\sum_{j,k} \hat{G}(T_j)^{-2} I(T_j < T_k) \delta_j} \in [0, 1] \tag{2.3}$$

where $I()$ is an indicator function. Here, $\hat{G}(T_j)$ is estimated from the training data and is defined as the Kaplan–Meier estimator of the unconditional survival function:

$$\hat{G}(t) = \prod_{t_j \leq t} 1 - \frac{d_j}{R(t_j)} \tag{2.4}$$

with d_i being the number of events at t_j. The C_{Uno} index is estimated nonparametrically, thereby adjusting for the censoring bias via inverse probability weighting. A risk score η_i is estimated for the selected features with new data $\mathbf{x}_{test}$ for each individual in the test set. This score is used as input for the C_{Uno} function. To obtain the variation in C_{Uno} with an independent test set, we calculated prediction performance with different random subsamples (of 90%).

An advantage of the C_{Uno} approach compared to other C-index definitions (Heagerty and Zheng, 2005; Antolini et al., 2005) lies in its independence of the Cox proportional hazard assumption. The C-index can be interpreted as the probability of concordance between the predicted and observed survival times over all pairs of observations at a given time point. Similar to the standard binary AUC, a value of 0.5 indicates that the marker is not better than random guessing and a value of 1 represents perfect separation. In contrast to the standard area under the ROC curve, models with C-index of relatively low values (between 0.6 and 0.7) are often considered as

having satisfactory predictive power. For example, a C-index of 0.67 was achieved (Tice et al., 2005) in a model predicting breast cancer based on genetic information, known as the Gail model (Gail et al., 1989). In cancer research, the absolute discrimination power is often not required; however, separation and classification of patients into groups of high and low risk is the primary goal. Therefore, this C-index is a favorable choice.

2.1. SurvRank

A schematic overview of the algorithm is shown in Figure 1, and further details are given in Algorithm 1. To fit a survival model and estimate generalizability, a repeated nested CV approach is used to first estimate the best number of features within an inner CV loop and then to estimate the performance of the model containing these features in an outer CV loop. Note that the identification of important features within the CV is based on different ranking methods.

2.1.1. Feature ranking methods. Three approaches to generate ranked output lists of features were considered, that is, an approach based on the log-rank statistic (survCox), a Lasso-based approach for survival data (survLasso), and a randomized Cox model (survRand).

Cox score ranking - survCox The Cox-based ranking approach sorts covariates according to their association with the survival response based on the Wald score test. For each feature, a univariate Cox model is fitted, and the obtained log-rank statistic is used as the ranking criterion (Moeschberger and Klein, 2003). A high test statistic indicates stronger association with the outcome. Note that this Cox score ranking is univariate in contrast to the other two approaches.

L_1 norm (Lasso) ranking - survLasso In this approach, ranking is generated using a penalized L_1 Cox regression (Tibshirani and others, 1997). Briefly, the L_1 penalty (Lasso) in the Cox regression case seeks to find a solution for the following:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{2}{n} \left(\sum_{i=1}^{m} \mathbf{x}_{j(i)}^T \boldsymbol{\beta} - \log \left(\sum_{j \in R_i} \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right) \right) - \lambda \sum_{i=1}^{p} |\beta_i| \right) \right). \tag{2.5}$$

An efficient implementation of the regularization path has been demonstrated (Simon et al., 2011). The complextity parameter λ determines the amount of shrinkage. The ranks of features are then obtained according to their appearance in the regularization path. All covariates not selected in the model obtain a rank that corresponds to the number of features p.

Randomized Cox ranking - survRand This ranking method consists of a two-step procedure. In the first step, L_1 penalization is used to preselect a smaller number of features ($p_{\text{pre}} < p$). The cut-off criterion in the Lasso is defined such that at least 95% of the deviance is explained at the end of the lambda sequence. In the second step, a sub-sampling approach (500 times) randomly chooses without replacement a smaller number of features and estimates a multivariate Cox model. To avoid convergence issues in the fitting procedure of the multivariate Cox model, the number of features in each subsampling step n_{sub} is limited to the number of observations ($n_{sub} = n/3$). Each feature in one subsampled Cox model yields a Z-statistic. The number of selections per feature is controlled by p_{pre}, thereby leading to $p_{pre}/3$ number of Z-statistics for each feature on average. Finally, by calculating the mean over all Z-score subsamples, a final feature score is derived and used for ranking in survRand.

2.1.2. Nested CV for estimating generalizability—TASK 1. The full dataset $D := D_i$ with $D_i := (T_i, \delta_i, x_i)$ is split into training set $D^{-cv_{out}}$ and test set $D^{cv_{out}}$ with index set cv_{out} and its complement.

Inner CV Inner CV is applied to only the training set $D^{-cv_{out}}$ by performing a second CV stratification, thereby yielding inner training $D^{(-cv_{out}, -cv_{in})}$ and inner test set $D^{(-cv_{out}, cv_{in})}$. Then, one of the described ranking functions is applied to the inner training set. By adding one feature at a time (following the ranking), a Cox model is estimated using the inner training data and evaluated with C_{Uno} for the inner test data. This procedure is performed until a predefined maximum number of features is achieved and is repeated for all inner CV folds. The best number of features is determined by averaging over all inner CVs and selecting the number of features that corresponds to the maximum mean C_{Uno}.

Outer CV For the outer CV, one feature ranking is performed for the whole training set. Then, using the best number of features derived in the inner CV, a Cox model is estimated using the training set, thereby yielding effect estimates for the selected features. Using these estimates, the unbiased prediction performance with the unseen test set is quantified by C_{Uno}, corresponding to TASK 1. Note that the entire procedure (including the inner CV) is applied to all outer CV folds.

Repeated CV To obtain an estimate of the variance of prediction accuracy, this approach is repeated *t_times* for different splits of the dataset.

2.1.3. Final model—TASK 2. The repeated nested CV combined with stepwise feature selection based on the ranking function yields a ranked set of features for each CV run. In addition, the performance on the test set for each run is recorded (number of runs $K = cv_{out} \times t\_times$). As these ranked lists of selected features are not necessarily the same, it is not clear how to aggregate them to a final set of features that can be used for predicting risk scores for new patients. Here, we propose an approach that leverages the information from all individual CV runs to determine a final set of features for which a final model can be fit.

Our weighted approach uses information from the outer CV performance corresponding to each run, thereby addressing TASK 2. The weight of run i is defined as follows:

$$w_i = \begin{cases} \frac{1}{K}\exp\left(\log(2)\frac{devAUC_i}{0.1}\right), & \text{if } C_{Uno,i} \geq 0.5 \\ 0, & \text{if } C_{Uno,i} < 0.5 \end{cases}. \tag{2.6}$$

Here, $devAUC_i$ denotes the relative C_{Uno} of an individual CV run compared to the average performance of all runs. The weights w_i' are further normalized to sum to one ($w_i' = w_i / \Sigma w_i$). The final set of predictors is determined by majority voting as follows:

$$I(p_j) = \begin{cases} 1 & \text{if } p_j > 0.5 \\ & \text{with } p_j = \sum_{i=1}^{K} I(j,i) w_i' \\ 0, & \text{if } p_j \leq 0.5 \end{cases} \tag{2.7}$$

where $I(j, i)$ is 1 if the feature p_j was selected in run i.

Algorithm 1: SurvRank algorithm with repeated nested CV

Data: survival data set (T_j, δ_j, $\mathbf{x_j}$);
parameters of rep CV: repetition *t_times*, outer CV *cv_out*, inner CV *cv_in*;
maximum number of features *max_var*;
ranking_fct (survLasso, survCox, survRand);
coxmodel function estimates $\hat{\beta}$ on a data set;
final_feature_fct (weighted);
Result: final set of selected features of the nested CV approach
for $t = 1 : t\_times$ **do**
  for $j = 1 : cv_{out}$ **do**
    train_outer ← $(T_j^{(-cv_{out})}, \delta_j^{(-cv_{out})}, \mathbf{X_j}^{(-cv_{out})})$;
    test_outer ← $(T_j^{(cv_{out})}, \delta_j^{(cv_{out})}, \mathbf{X_j}^{(cv_{out})})$;
    for $k = 1 : cv_{in}$ **do**
      train_inner ← $(T_j^{(-cv_{out}, -cv_{in})}, \delta_j^{(-cv_{out}, -cv_{in})}, \mathbf{X_j}^{(-cv_{out}, -cv_{in})})$;
      test_inner ← $(T_j^{(-cv_{out}, cv_{in})}, \delta_j^{(-cv_{out}, cv_{in})}, \mathbf{X_j}^{(-cv_{out}, cv_{in})})$;
      ranking_in ← ranking_fct(train_inner);
      for $i = 1 : max\_var$ **do**
        coxmodel_in ← coxmodel(ranking_in[1 : i], train_inner);
        surv_in[i, k] ← Cindex(coxmodel_in, test_inner);
      end
    end
    meanCurve ← mean(surv_in,k);
    maxFeature ← which_max(meanCurve);
    ranking_out ← ranking_fct(train_outer)[1:maxFeature];
    coxmodel_out ← coxmodel(ranking_out, train_outer);
    surv_out[j,t] ← Cindex(coxmodel_out, test_outer);
  end
end
sel_features ← final_feature_fct(surv_out, ranking_out);
final_model ← coxmodel(sel_features, (T_j, δ_j, $\mathbf{x_j}$));

Finally, one survival model can be calculated with the selected features using the entire dataset, thereby leading to effect estimates $\hat{\boldsymbol{\beta}}_{train}$ and risk scores for each subject. This is used to predict survival probabilities with unseen data with similar predictive power as estimated in the nested CV.

2.2. Comparison method

To compare this approach to existing methods, a commonly used regularized survival model based on Cox-Lasso was selected (coxLasso). For coxLasso, the same unbiased approach was performed to estimate the prediction accuracy with CV by applying the same repeated CV parameters. One CV step consists of separation into different folds and optimizing the penalization parameter $\hat{\lambda}$ by the inner CV of one fold. This optimized $\hat{\lambda}$ was used to predict the unseen test fold, thereby measuring performance with C_{Uno}. For coxLasso, the final selection of covariates, which are used for prediction with the test set, was estimated by applying CV to the entire training dataset once. By optimizing the partial likelihood in the Cox regression, the number of features was obtained with cross-validated minimum deviance for coxLasso.

3. RESULTS

3.1. Simulation and validation setup

To evaluate our algorithm, we generated a high-dimensional, multivariate normally distributed dataset with $n=100$ observations and $p=500$ features. The survival times T_i followed an exponential distribution with mean $\eta_i = 1/(\lambda_T \sum_{i_1}^{4} x_i \beta_i)$, which we set to $\lambda_t = 0.5$ and $\beta_1 = 1.5$, $\beta_2 = -1.5$, $\beta_3 = -1$, and $\beta_4 = 1$ for our framework. An independent random censoring time T_{cens} was simulated such that it followed an exponential distribution, which we fixed to mean 2. The observed survival times T_{obs} are expressed by $T_{obs} = min(T_{cens}, T_i)$, which leads to independent censoring of approximately 50%. The maximum number of features was set to 30. Partitioning into training and test sets was applied in all configurations with the same parameters ($cv_{in}=10$, $cv_{out}=10$, $t\_times=10$). To calculate C_{Uno}, we fixed τ at the last observed survival time.

We first used the simulated data to estimate generalizability accurately, which is directly related to TASK 1. By applying a final model fit on the training set and estimating the performance with 10 simulated test sets, we retrieved the performance of our model selection with new unseen data. Ideally, the performance difference between the training and test data should be small. Otherwise, we would have a classical overfitting situation with the training data, where generalization accuracy to new unseen test data is not fulfilled. This procedure was repeated for 100 different training datasets. Furthermore, we calculated the true C_{Uno} for the training set and the test sets using only the true effects $\beta_1,\ldots,\beta_4$.

We then attempted to retrieve the correct set of features, thereby addressing TASK 2 (feature selection). To achieve this, we calculated the F_1 score, which is defined as the harmonic mean of precision and recall, that is, $F_1 = 2 \times \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. Here, the F_1 score was calculated to compare the selected features with respect to the four true features.

We then compared our approach with a commonly used regularized survival model. Here, we estimated a penalized survival Cox model with Lasso (coxLasso based on the R package glmnet).

3.2. Simulated dataset results

We observed good performance with the test data and comparable results for accuracy with the training set compared to the test sets (Fig. 2), thereby addressing TASK 1. The coxLasso approach performed similarly with the training data compared to survLasso from our package; however, as expected, prediction with unseen new data shows substantial overfitting. The survRand ranking function demonstrated higher variance of C_{Uno} with the training set. survCox ranking performed worse with the training data; however, the final feature selection results showed comparable prediction accuracy with new test data. The overall worse performance of survCox illustrates the advantage of the multivariate ranking function of survLasso and survRand compared to survCox with univariate ranking.

Compared to standard coxLasso, the sparser set of selected features represents an advantage of our ranking and final feature selection approach (Fig. 3A), thereby addressing TASK 2. This illustrates that selecting features according to the data fit (deviance), as used in the standard coxLasso approach, produces too many selected features. In addition, we investigated whether the correct covariates were selected. We

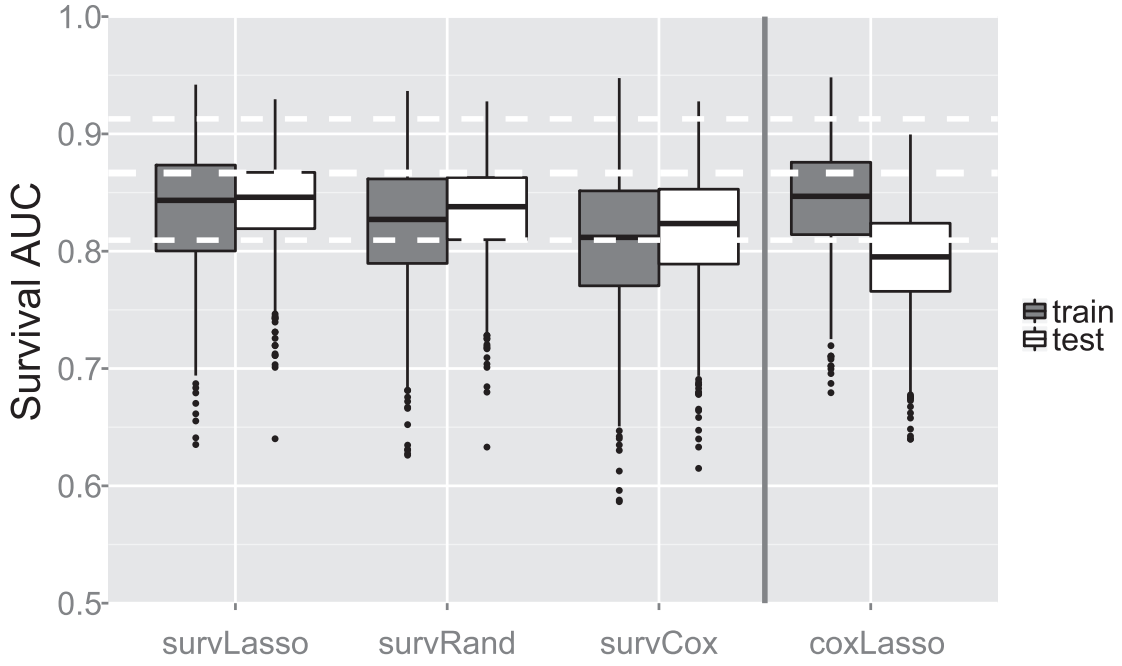


FIG. 2. Prediction performance with simulated data. A total of 100 training datasets were simulated, and unbiased C_{Uno}s were obtained for each repetition of the nested CV. For each of the 100 training datasets, 10 test sets were created to test prediction performance with new data. White dashed lines indicate the average of the true C_{Uno} with the simulated datasets with an empirical 95% quantile range.

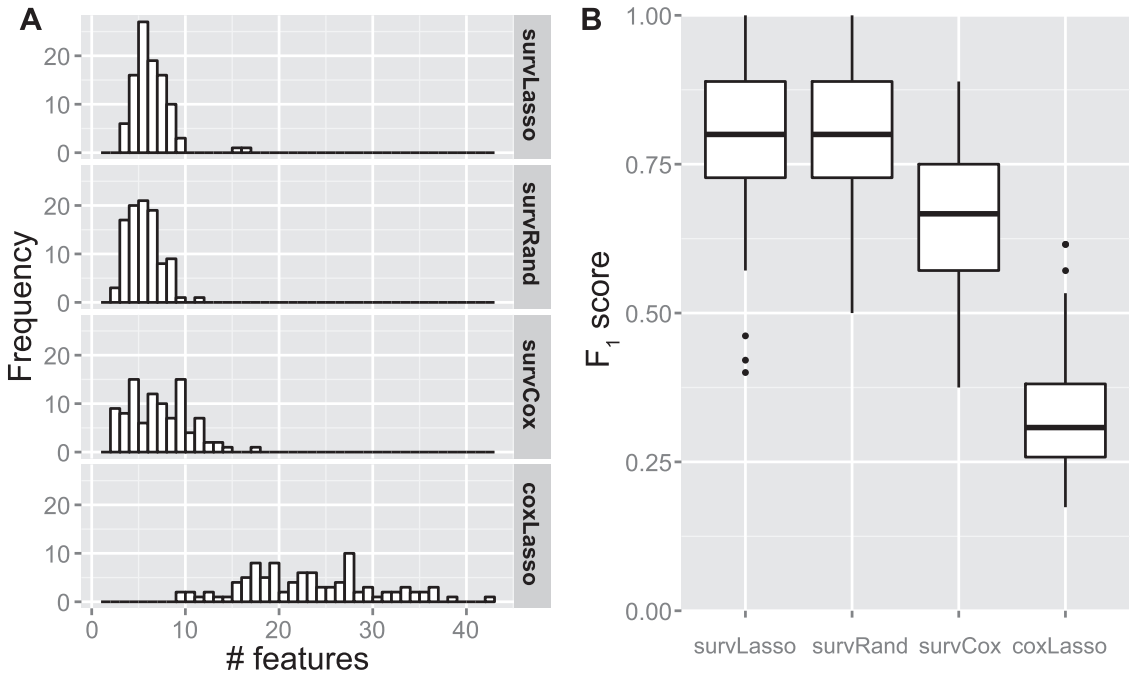


FIG. 3. **(A)** Number of selected features across simulated training datasets in the weighted approach. **(B)** F_1 scores for selected features to compare the selected features with the set of four true features.

Table 1. Computation Time in Minutes for Different $p \times n$ Setups

p	100	200	200	500
n	100	100	200	100
survLasso	9.00	9.43	20.33	19.00
survCox	10.03	16.43	16.48	27.58
survRand	77.07	82.70	186.57	86.28
coxLasso	3.70	3.98	14.38	5.42

Parameters set to $t\_times = 10$, $cv_{out} = 10$, and $cv_{in} = 10$.

observed higher F_1 scores (Fig. 3B) with our approach compared to coxLasso. These results illustrate the overfitting of the coxLasso approach, that is, it selects several random, noninformative features (resulting in a high FPR) and considerably overestimates predictive power with training sets (reduction of C_{Uno} on average of 0.05 or 6% from training to test).

3.2.1. Runtime evaluation. An important aspect for nested CV approaches is the required computation time. The SurvRank package inherently supports parallelization across multiple cores on the same machine. Table 1 shows the runtimes for different variable settings for the three ranking functions using a single core of an Intel Core i5 2.6 GHz CPU. Here, we observed that the number of features p scaled approximately linearly with computation time for survLasso, survRand, and coxLasso. survLasso was slower than coxLasso in the first two settings by a factor of approximately 2.5, taking the additional stepwise selection into account. Doubling the number of observations n increased computation time by a factor of 2.2 for survLasso and survRand and by a factor of 3.6 for coxLasso. In contrast, the computation time of survCox scaled approximately linearly with the number of features due to the univariate ranking procedure. For survCox, an increasing sample size increased computation time only slightly.

3.3. Application to three breast cancer gene expression datasets

To evaluate our approach with real clinical data, we applied the pipeline to microarray datasets from breast cancer patients with survival information (relapse time) after surgery (mastectomy) or radiotherapy. We used two independent datasets to estimate the prediction accuracy with unseen data to assess how well our method performs with TASK 1. To identify a predictive subset of features, we used our approach with different ranking functions, thereby addressing TASK 2. In addition, we compared the performance of our approach to a standard CoxLasso model and a set of 76 marker genes identified in the primary publication (referred to as geneMarker). This geneMarker was derived by ranking the features according to an averaged Cox score (using bootstrap samples).

The first dataset contained 286 patients with lymph node-negative breast cancer. For each patient, information about estrogen receptor status positive (ER+) and estrogen receptor status negative (ER-) was recorded, assuming that disease progression differs for these subgroups. This first dataset served as the training set [accession number GSE2034 (Wang et al., 2005)]. Wang et al. identified a predictive set of 76 genes (geneMarker) composed of 60 genes for the ER+ group and 16 genes for the ER- group. We attempted to obtain an alternative sparse set of genes with better generalizability to evaluate the performance of our approach with two independent validation sets, that is, accession numbers GSE7390 (Desmedt et al., 2007) and GSE1456 (Pawitan et al., 2005). There was an overlap of 18,842 features across the three datasets. In the training data, there were 209 patient samples in the ER+ group and 77 observations with ER- status. The first test dataset (test set 1) consisted of 134 samples in the ER+ group and 64 in the ER- group. The second test set (test set 2) contained 125 subjects in the ER+ group and 27 in the ER- group. Due to the larger number of observations, we focused on the ER+ subgroup for our evaluation.

We applied our different ranking algorithms to the dedicated training set and obtained a final marker. Furthermore, the selected genes were evaluated with the new and unseen test sets. The parameters of the repeated nested CV were determined as $t\_times = 20$, $cv_{out} = 10$, and $cv_{in} = 10$. The maximum number of features was set to 75, and τ in C_{Uno} was set to 10 years.

The geneMarker and the coxLasso approach served as comparison models for our ranking algorithms. The results of geneMarker were calculated by applying ridge regression to the training data and then

evaluating performance with the two test sets. For coxLasso, we repeated the final feature selection ten times to determine the optimal penalization parameter, because coxLasso depends on the sampling of CV folds.

3.4. Breast cancer data results

For our approach, performance with the unseen test dataset showed similar prediction accuracy compared to the training data (Fig. 4). This indicates that our nested CV strategy was able to estimate the generalizability of the predictor correctly, thereby solving TASK 1. The number of selected features varied slightly between the three approaches of our package (24, 19, and 29 for survLasso, survRand, and survCox, respectively), thereby addressing TASK 2. survLasso and survCox showed larger overlap of selected genes compared to survRand (Fig. 5). As in the simulation study, survLasso performed considerably better than survCox (on average C_{Uno} decreased by 0.03 or 5%), again illustrating the advantages of a multivariate ranking approach compared to univariate ranking. Similar to the results of the simulation study, coxLasso selected 53 features with too many false positives, resulting in a reduced performance with the test data sets. geneMarker resulted in clear overfitting of this marker set with the training dataset (as expected), where geneMarker was derived. Therefore, these results can be interpreted as training performance. Consequently, the predictive power decreased strongly with the test sets. Comparing the geneMarker set with the selected markers in survLasso, survRand, and survCox yielded a small overlap, that is, survLasso 2 genes, survRand 0, and survCox 5 (details in Supplementary Fig. S1, available online at www.liebertpub.com/cmb).

4. DISCUSSION

We have proposed a new framework to reliably estimate prediction accuracy and generalizability and to select the most predictive features in a high-dimensional survival prediction setting. To avoid overfitting

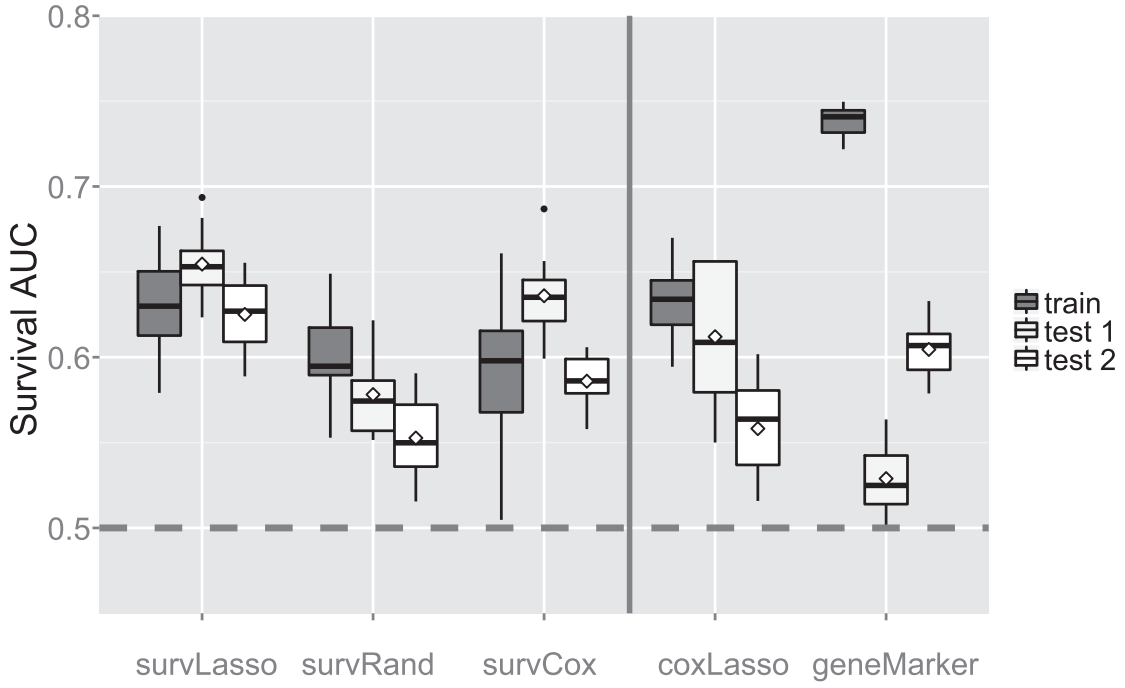


FIG. 4. Prediction accuracy with three breast cancer data sets. The performance of the training data set was compared to two independent test sets for the ER+ group. Feature selection was based on the weighted approach. Diamonds show performance with the whole test set, whereas variation in the boxplots was obtained by subsampling the test data sets.

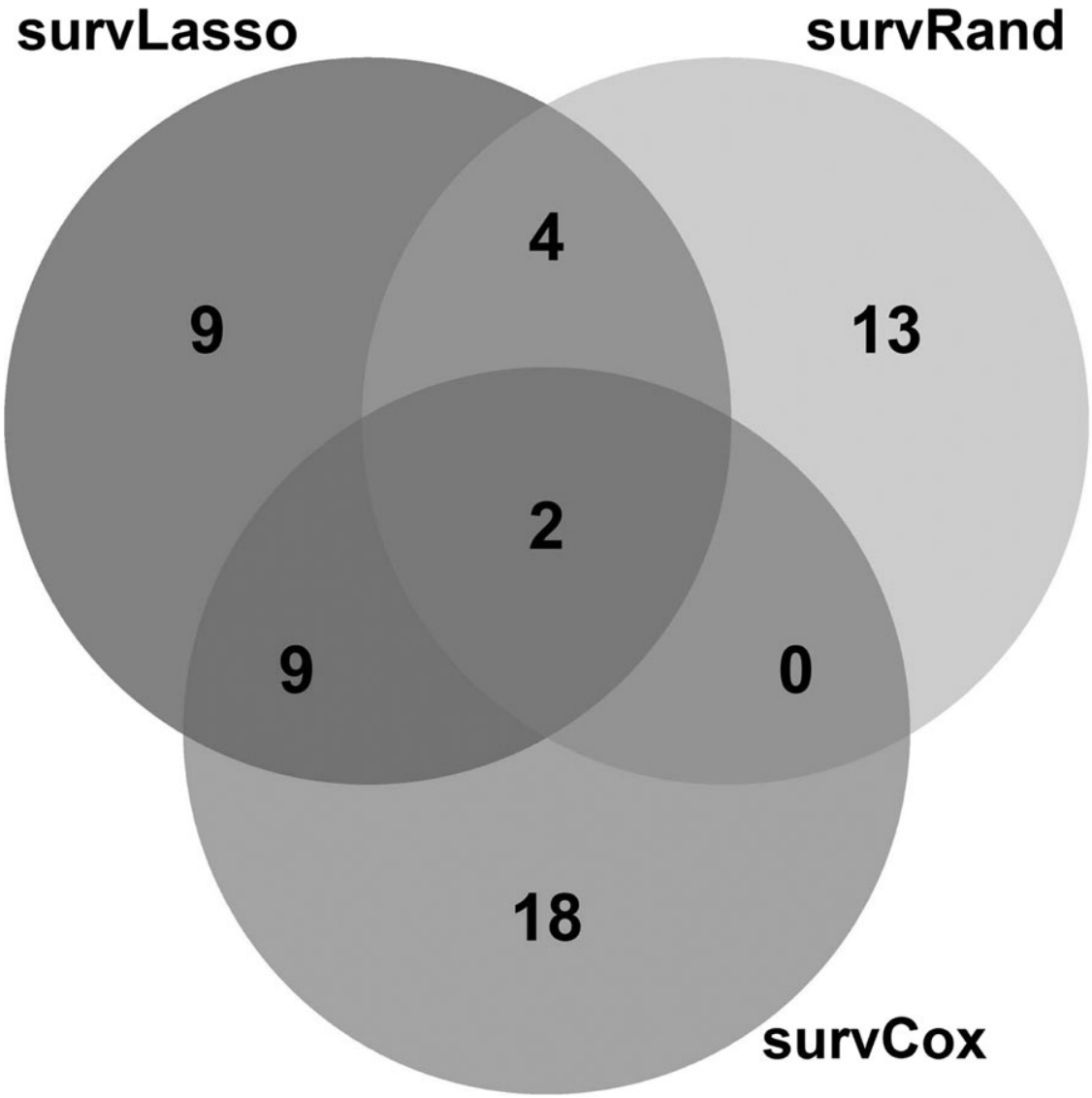


FIG. 5. Overlap of selected genes of the different ranking functions.

while selecting features with high predictive power, the proposed approach estimates accuracy and performs feature selection using repeated nested CV with novel feature combination heuristics.

Our approach differs from standard approaches, such as the CoxLasso approach, in two ways. First, the selection of features is determined by the best predictive feature combination (using C_{Uno}) rather than the best data fitting combination, thereby reducing the risk of overfitting. Second, for final feature selection, our approach leverages information from different CV runs. The CoxLasso approach uses the minimum cross-validated deviance of the whole dataset, while the proposed approach aggregates the results of different CV runs and applies a weighting scheme to select only predictive features. This combination of aggregating CV runs by weighting results in sparser feature selection with more accurate estimation of predictive power.

Using simulated data, we demonstrated that the proposed method can identify true features and can correctly estimate prediction accuracy with new data without overfitting. By comparing the results of different methods in this simulation setup, we observed that survLasso dominates survCox with training and test data. This effect can be explained by the multivariable ranking procedure of survLasso (considering all features) in contrast to the univariate ranking of survCox, which treats features independently.

With breast cancer data, our pipeline based on two of our ranking approaches was able to estimate similar prediction performance with the test datasets compared to the training data. However, the survRand

approach showed a drop in prediction performance with the breast cancer test data. This effect is illustrated in Figure 5, where we observe that this ranking approach has only small overlap compared to survLasso and survCox. The 19 selected features in this approach lead to lower prediction performance. By comparing coxLasso and survRand, we observed an overlap of six features that are only picked by these methods (Supplementary Fig. S1), thereby introducing noise to the model. In addition, the sampling strategy of survRand might introduce some noise to the selection process. This again confirms the robust performance of survLasso compared to the other ranking methods.

Our approach can be extended in several directions. (1) In clinical applications, variables such as age, gender, height, and BMI are collected routinely. Therefore, it would be desirable to force such features into the model and evaluate the additional benefit of omics data. (2) Our framework uses the Cox proportional hazards model. Extending the approach to accelerated failure time models or frailty models may improve the baseline hazard estimation, such as time-varying hazards or random effects. (3) Applying repeated nested CV to classification tasks may also be an interesting extension.

Importantly, our approach as a biomarker discovery method focuses on identifying a predictive biomarker combination and does not provide functional interpretation of the selected features (e.g., genes and transcripts). Therefore, we recommend using the SurvRank package with the survLasso approach and weighted final feature selection, due to the low computational demands and best results from both the simulation study and the clinical data.

In summary, we provide a flexible, ready-to-use toolbox for survival data that allows for unbiased estimation of prediction accuracy for survival models and extracts the most predictive features from high-dimensional survival datasets.

ACKNOWLEDGMENTS

This work was funded in part by grants from the German Federal Ministry of Education and Research (BMBF), grant No. 01ZX1313C (project e:Athero-MED) and 01ZX1314G (project IntegraMent), and from the European Union's Seventh Framework Programme [FP7-Health-F5-2012] under grant agreement number 305280 (MIMOmics). F.B. was supported by a UK Medical Research Council Career Development Award (Biostatistics).

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abbasi, A., Peelen, L.M., Corpeleijn, E., et al. 2012. Prediction models for risk of developing type 2 diabetes: Systematic literature search and independent external validation study. *BMJ* 345, e5900.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.

Antolini, L., Boracchi, P., and Biganzoli, E. 2005. A time-dependent discrimination index for survival data. *Stat. Med.* 24, 3927–3944.

Beer, D.G., Kardia, S.L., Huang, C.-C., et al. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8, 816–824.

Cox, D.R. 1972. Regression models and life-tables. *J. R. Stat. Soc. B* 34, 187–220.

Datta, S., Le-Rademacher, J., and Datta, S. 2007. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* 63, 259–271.

Desmedt, C., Piette, F., Loi, S., et al. 2007. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* 13, 3207–3214.

Eschrich, S., Yang, I., Bloom, G., et al. 2005. Molecular staging for survival prediction of colorectal cancer patients. *J. Clin. Oncol.* 23, 3526–3535.

Gail, M.H., Brinton, L.A., Byar, D.P., et al. 1989. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* 81, 1879–1886.

Gong, H., Wu, T.T., and Clarke, E.M. 2014. Pathway-gene identification for pancreatic cancer survival via doubly regularized Cox regression. *BMC Syst. Biol.* 8, 1–9.

Gui, J., and Li, H. 2005. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 21, 3001–3008.

Harrell, F.E., Califf, R.M., Pryor, D.B., et al. 1982. Evaluating the yield of medical tests. *JAMA* 247, 2543–2546.

Heagerty, P.J., and Zheng, Y. 2005. Survival model predictive accuracy and ROC curves. *Biometrics* 61, 92–105.

McShane, L.M., Cavenagh, M.M., Lively, T.G., et al. 2013. Criteria for the use of omics-based predictors in clinical trials. *Nature* 502, 317–320.

Moeschberger, M.L., and Klein, J. 2003. *Survival Analysis: Techniques for Censored and Truncated Data: Statistics for Biology and Health*. Springer, New York.

Pawitan, Y., Bjhle, J., Amler, L., et al. 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. *Breast Cancer Res.* 7, R953.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Simon, N., Friedman, J.H., Hastie, T., and Tibshirani, R. 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1–13.

Takamizawa, J., Konishi, H., Yanagisawa, K., et al. 2004. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res*. 64, 3753–3756.

Tibshirani, R., et al. 1997. The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395.

Tice, J.A., Cummings, S.R., Ziv, E., and Kerlikowske, K. 2005. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast Cancer Res. Treat*. 94, 115–122.

Uno, H., Cai, T., Pencina, M.J., et al. 2011. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med*. 30, 1105–1117.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., et al. 2002. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347, 1999–2009.

Wang, Y., Klijn, J.G.M., Zhang, Y., et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365, 671–679.

Wu, T.T., Gong, H., and Clarke, E.M. 2011. A transcriptome analysis by lasso penalized Cox regression for pancreatic cancer survival. *J. Bioinform. Comput. Biol.* 9 Suppl 1, 63–73.

Zhao, H., Ljungberg, B., Grankvist, K., et al. 2005. Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLoS Med* 3, e13.

Address correspondence to:
Florian Buettner or Fabian Theis
Institute of Computational Biology
Helmholtz-Zentrum München
Ingolstädter Landstraße 1
85764 Neuherberg
Germany

E-mail: buettner@ebi.ac.uk
or
fabian.theis@helmholtz-muenchen.de

Laimighofer, Michael

Von: Ballen, Karen <KBallen@liebertpub.com>
Gesendet: Freitag, 24. März 2017 21:53
An: michael.laimighofer@helmholtz-muenchen.de
Betreff: RE: LiebertPub Website Customer Question

Dear Michael:

As an author, you may use the article in your doctoral thesis, but not for commercial purposes.

Kind regards,

Karen Ballen
Manager, Reprints, Permissions and Open Access

From: michael.laimighofer@helmholtz-muenchen.de [mailto:michael.laimighofer@helmholtz-muenchen.de]
Sent: Friday, March 24, 2017 5:53 AM
To: Ballen, Karen <KBallen@liebertpub.com>
Subject: LiebertPub Website Customer Question

Name - Michael Laimighofer
Position -
Department - ICB
Institution/affiliation - Helmholtz Zentrum Muenchen
Address Line1 -
City - Munich
State -
Country - GER
Zip -
Email - michael.laimighofer@helmholtz-muenchen.de
Phone -
Regarding - Reprints and permissions
For Publication - Journal of Computational Biology
Questions/Comments - Dear Sir or Madam, I would like to kindly ask about a permission to use our publication "Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression" published in the Journal of Computational Biology in my doctoral thesis? Thanks in advance, Michael Laimighofer

Peptide serum markers in islet autoantibody-positive children.

Diabetologia
DOI 10.1007/s00125-016-4150-x

ARTICLE

Peptide serum markers in islet autoantibody-positive children

Christine von Toerne[1] · Michael Laimighofer[2,3] · Peter Achenbach[4,5,6] · Andreas Beyerlein[4,5] · Tonia de las Heras Gala[7,8] · Jan Krumsiek[2,7] · Fabian J. Theis[2,3] · Anette G. Ziegler[4,5,6] · Stefanie M. Hauck[1]

Received: 10 August 2016 / Accepted: 5 October 2016

Abstract
Aims/hypothesis We sought to identify minimal sets of serum peptide signatures as markers for islet autoimmunity and predictors of progression rates to clinical type 1 diabetes in a case–control study.
Methods A double cross-validation approach was applied to first prioritise peptides from a shotgun proteomic approach in 45 islet autoantibody-positive and -negative children from the BABYDIAB/BABYDIET birth cohorts. Targeted proteomics for 82 discriminating peptides were then applied to samples from another 140 children from these cohorts.
Results A total of 41 peptides (26 proteins) enriched for the functional category lipid metabolism were significantly different between islet autoantibody-positive and autoantibody-negative children. Two peptides (from apolipoprotein M and apolipoprotein C-IV) were sufficient to discriminate autoantibody-positive from autoantibody-negative children. Hepatocyte growth factor activator, complement factor H, ceruloplasmin and age predicted progression time to type 1 diabetes with a significant improvement compared with age alone.
Conclusion/interpretation Distinct peptide signatures indicate islet autoimmunity prior to the clinical manifestation of type 1 diabetes and enable refined staging of the presymptomatic disease period.

Keywords Autoantibody-positive · Autoimmunity · BABYDIAB/BABYDIET · LC-MS/MS · Progression time · Risk score · Selected reaction monitoring · Targeted proteomic · Type 1 diabetes

Christine von Toerne and Michael Laimighofer contributed equally to this study.

Electronic supplementary material The online version of this article (doi:10.1007/s00125-016-4150-x) contains peer-reviewed but unedited supplementary material, which is available to authorised users.

✉ Anette G. Ziegler
anette-g.ziegler@helmholtz-muenchen.de

✉ Stefanie M. Hauck
hauck@helmholtz-muenchen.de

[1] Research Unit Protein Science, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, D-85764 München, Germany

[2] Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

[3] Department of Mathematics, Technische Universität München, Garching, Germany

[4] Institute of Diabetes Research, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, D-85764 München, Germany

[5] Forschergruppe Diabetes, Klinikum rechts der Isar, Technische Universität München, Munich, Germany

[6] Forschergruppe Diabetes e.V., Neuherberg, Germany

[7] German Center for Diabetes Research (DZD), Neuherberg, Germany

[8] Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Published online: 04 November 2016

Abbreviations

APO	Apolipoprotein
CF	Complement factor
CP	Ceruloplasmin
dCV	Double cross-validation
DIPP study	Type 1 Diabetes Prediction and Prevention study
FDR	False discovery rate
HGFAC	Hepatocyte growth factor activator
HNF1A	Hepatocyte nuclear factor 1α
IQR	Interquartile range
LC-MS/MS	Liquid chromatography tandem MS
SRM	Selected reaction monitoring

Introduction

The development of type 1 diabetes includes an asymptomatic period of autoimmunity identified by the presence of islet autoantibodies, with subsequent progression to dysglycaemia and clinical diabetes [1]. While the development of islet autoantibodies is most prominent around 1–2 years of age [2–4], the incidence of clinical diabetes appears to be relatively constant in multiple islet autoantibody-positive children and adolescents [5]. Biomarkers and genetics that are associated with islet autoimmunity are of interest for elucidating pathogenesis, and biomarkers that predict the rate of progression [6–10] may improve staging the presymptomatic disease period of type 1 diabetes.

Proteomics has been used to identify biomarkers in diverse diseases such as cardiovascular diseases [11], prostrate and other cancers [12, 13], Parkinson's disease [14] and metabolic disorders [11, 15]. In type 1 diabetes, previous proteomic biomarker screening studies have compared patients with type 1 diabetes to autoantibody-negative control participants [16–18] and identified protein signatures correlated with clinical disease. A recent longitudinal study in Finland compared islet autoantibody-positive children with autoantibody-negative children, and identified a protein signature that distinguished between healthy children and those with autoimmunity [19].

Here, we applied proteomics to our cohorts of children followed from birth to islet autoimmunity and clinical diabetes in order to search for signatures associated with islet autoimmunity, and which could help predict the progression rate to clinical diabetes in multiple autoantibody-positive children.

Methods

This study was performed using sera from children participating in either the BABYDIAB [20] or BABYDIET [21] studies. These birth cohort studies enrolled children with a family history of type 1 diabetes and are prospectively monitoring the natural history of islet autoimmunity and type 1 diabetes. Together, they have enrolled 2441 children [20, 21]. By November 2014, 124 children had developed multiple islet autoantibodies and 82 of these children had progressed to clinical type 1 diabetes [22].

Islet autoantibodies were measured using radiobinding assays as previously described [8, 20]. The antibody assays were evaluated in the Diabetes Autoantibody Standardization Program (Laboratory 121) [23–25]. Diabetes was diagnosed according to the ADA Expert Committee criteria [26]. Both studies were approved by the ethics committee of Bavaria, Germany (Bayerische Landesärztekammer No. 95357 and Ludwig-Maximilians University No. 329/00, respectively), and adhered to the principles of the Declaration of Helsinki.

Sample selection and study design The analysis was performed in two phases: a peptide-selection phase in which shotgun proteomics was performed to identify peptides of potential interest, which were then measured by targeted proteomics in a second application phase (Fig. 1 and electronic supplementary material [ESM] Fig.1). For the selection phase, we applied shotgun proteomics to samples from children who developed islet autoantibodies and progressed to clinical diabetes within 3.5 years ('rapid' progression: 15 children; median follow-up from seroconversion 1.9 years, interquartile range [IQR] 1.0–2.9 years, range 0.5–3.3 years) or ≥9.5 years ('slow' progression: 15 children; median follow-up from seroconversion 14.5 years, IQR 12.9–15.5 years, range 9.5–17.4 years), and from 15 children who remained islet autoantibody-negative (median follow-up from birth 15.9 years, IQR 14.2–17.4 years, range 5.9–21.7 years) matched for sex and age (Fig. 1). Two sample times were separately analysed. Specifically, one sample from each child was obtained shortly after seroconversion to the first islet autoantibody (median 0.8 years, IQR 0.3–1.4 years; sample set 1) or at the corresponding age in islet autoantibody-negative children, while the other sample was obtained at a later time (median 1.2 years after the first sample, IQR 0.8–2.9 years; sample set 2). Four children were excluded from sample set 2 in the selection phase because they had already progressed to overt diabetes by the time the second sample had been collected after seroconversion.

For the application phase, we randomly selected 70 of the remaining children who developed islet autoantibodies (median age 3.2 years, median follow-up time 12.8 years, IQR 9.6–16.6 years) and 70 sex- and age-matched islet autoantibody-negative children (median age 3.1 years, median follow-up time 10.8 years, IQR 7.2–14.4 years) (Fig. 1).

We performed targeted proteomics on the peptides that discriminated between groups in the selection phase (see detailed description below). Samples from the 70 islet autoantibody-positive children were obtained shortly after

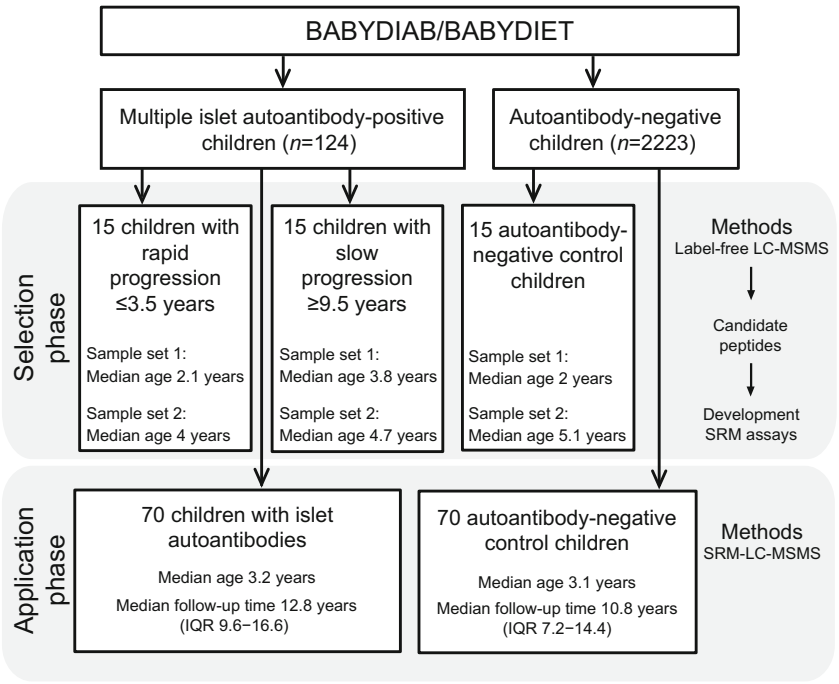


Fig. 1 Study design and analytical workflow

seroconversion (median 1.0 years, IQR 0.5–1.3 years; Fig. 1) and 60 children were multiple islet autoantibody-positive at the time of proteomics measurement.

Sample preparation for MS Plasma samples were depleted from highly abundant proteins and proteolysed with trypsin as previously described [27]. All samples were randomly distributed into one of three batches for processing, and the experimenters were blinded to the sample-group allocation during the experiment. For quality control of depletion, digestion and MS measurements, each sample was spiked with ribulose-1,5-bisphosphate carboxylase oxygenase (Sigma Aldrich, Taufkirchen, Germany) at a final amount of 50 fmol in each 10 µl serum sample. After digestion, samples were stored at −80°C until further use.

Non-targeted liquid chromatography tandem MS (LC-MS/MS) and label-free quantification LC-MS/MS analyses were performed as previously described [28] on an LTQ-Orbitrap XL instrument (Thermo Fisher Scientific, Dreieich, Germany) operated with an RSLC system (Ultimate 3000, Thermo Fisher Scientific). The RAW files (Thermo Fisher Scientific) were analysed using the Progenesis LC-MS software (version 4.0; Nonlinear Dynamics, Waters, Eschborn, Germany), as previously described [27, 29].

Targeted LC-MS/MS using selected reaction monitoring (SRM) Skyline software (MacCoss Lab Software, Seattle, WA, USA) was used to create the SRM assays [30]. We developed and optimised an SRM assay if at least one peptide per protein satisfied the quality criteria defined using the AuDIT algorithm [31] for reproducible and reliable SRM measurement. Isotope-labelled, synthetic peptides (heavy peptides; PEPotec; Thermo Fisher Scientific, Ulm, Germany) were used as internal controls for correct signal integration and relative quantification. The heavy peptide mix was added to the digested sample before the MS measurement.

SRM-MS analyses were performed on a Tempo Nano MDLC system (Eksigent Technologies, Dublin, OH, USA) coupled online to a triple quadrupole QTrap4000 (AB SCIEX, Framingham, MA, USA) MS equipped with a nanospray ion source [27]. During the MS measurements, the preselected proteotypic peptides were fragmented and the areas under the chromatographic curves of the resulting transitions formed the basis of the SRM quantifications.

Processing of SRM data SRM data were processed using the Skyline software as previously described [15]. Briefly, after manual quality control, heavy to light peptide ratios were calculated on fragment levels, $\log_2$ transformed and corrected for batch effects by linear regression, followed by averaging fragment values to peptides. The peptide values were normalised against control protein peptides and are referred to as adjusted intensities. Peptides with unreliable signals (>20% of measurements below the limits of quantification per peptide) were removed, resulting in robust SRM assays for 82 peptides covering 50 proteins (ESM Table 1).

Statistical analysis in the selection phase In the selection phase, using a univariate non-parametric test (Wilcoxon

rank-sum test), we assessed group differences in both sample sets (one collected shortly after seroconversion and one collected at a later time point) between: (1) islet autoantibody-positive vs autoantibody-negative children; (2) autoantibody-negative children vs slow progressors; (3) autoantibody-negative children vs rapid progressors; and (4) slow vs rapid progressors. Multiple hypothesis testing was corrected for by controlling the false discovery rate (FDR) at 0.05.

A double cross-validation (dCV) approach was then used to identify multivariable predictive protein and peptide signatures for the same eight comparisons (two sample sets and four group comparisons each). This approach selected a minimal combination of peptides that provided high discriminative accuracy, and estimated an unbiased, non-over-fitted AUC [32]. A detailed explanation of the approach and the parameter settings used in our study can be found in the ESM Method.

Peptides occurring with at least 75% selection frequency in at least one of the eight comparisons were compiled into a candidate 'selection' list. To maximise our coverage, this list was extended by 14 peptides that were reported in a recent proteomics study [17].

Statistical analysis in the application phase In the application phase, we tested for differences in peptide levels between islet autoantibody-positive and autoantibody-negative children using Wilcoxon rank-sum tests. To model the time from seroconversion to type 1 diabetes, we fitted univariate Cox regression models within the islet autoantibody-positive samples. Multiple hypothesis testing was corrected for by controlling the FDR at 0.05. Highly correlated peptides were identified using Pearson's correlation coefficient.

We again applied the dCV algorithm to find multivariable peptide signatures discriminating between islet autoantibody-positive and autoantibody-negative samples. A modified version of this algorithm that used Cox models instead of classification models was then applied to identify a predictive signature of progression time within the autoantibody-positive children. For the dCV analyses in the application phase, we also included age as an explanatory variable. Details on the dCV approach in the application phase can be found in the ESM Method.

Peptides with a selection frequency of at least 50% were used to fit a final Cox model, yielding progression time risk scores for each autoantibody-positive individual in the application set. These scores were divided into low-, medium- and high-risk tertiles. Differences in the survival curves between the tertiles were assessed using logrank tests. In order to investigate the improvement in discrimination conferred by the selected peptides in addition to age, a Cox model containing only age was compared with the combined model by ANOVA. In addition, the discrimination performance over time of the combined model and of age alone was evaluated using the survival AUC measure [33]. As an overall measure of discrimination, an integrated AUC was calculated.

All analyses were performed using R version 3.2.0 (www.r-project.org).

Enrichment analysis GeneRanker software (Genomatix software suite V3.5; Genomatix, Munich, Germany) was used to evaluate protein enrichment. Gene symbols for the respective proteins were used as identifiers. Gene ontology enrichment was calculated by comparing all significantly different proteins identified in the application phase as discriminating between islet autoantibody-positive and autoantibody-negative children against all proteins identified in plasma in the discovery phase. Redundancies in enriched terms for biological processes were curated manually.

Results

Shotgun proteomics identified tryptic peptides, which discriminated between autoantibody statuses and progression rates Shotgun proteomics of serum samples from the selection group resulted in the quantification of 2021 tryptic peptides (covering 204 proteins) in the first sample set and 2996 tryptic peptides (243 proteins) in the second sample set. A total of 215 peptides (covering 106 proteins) were selected by the dCV approach for discrimination in at least one between-group comparison (islet autoantibody-positive vs autoantibody-negative; slow vs rapid; autoantibody-negative vs slow; and autoantibody-negative vs rapid). Of these, 169 peptides overlapped between the first and second sample sets and were evaluated for SRM development.

Robust SRM assays were developed for 82 peptides (covering 50 proteins; ESM Table 1). These included 14 peptides that were added from a previous study [17] but were not selected as significant in the selection phase of this study (ESM Table 1).

Application phase: targeted proteomic analyses for discriminating between islet autoantibody-positive and autoantibody-negative children In univariate analysis, the abundance of 26 proteins (represented by 41 peptides) differed significantly between autoantibody-positive and autoantibody-negative children (Table 1; ESM Fig. 2). Eight of those proteins (represented by 14 peptides) overlapped with findings from previous studies [16–19] (Table 1). This included four of the 14 peptides that were tested in our study because they had been identified in a previous study [17] (Table 1). Pearson's correlation test revealed several correlated peptides. As expected, the highest correlations were observed for peptides belonging to the same protein, indicating a high reliability of SRM measurements (ESM Fig. 3). Peptides representing proteins belonging to the same protein family, such as apolipoproteins (APOs), also

Table 1 Univariate comparison of peptide abundance between islet autoantibody-positive and autoantibody-negative children

Protein	Sequence	p value[a]	Effect size[b]
APOA4[c]	ISASAEELR	1.73×10^{-4}	−1.14
	LGEVNTYAGDLQK	5.95×10^{-5}	−0.71
APOE	LEEQAQQIR	1.10×10^{-3}	−1.07
FN1	YQCYCYGR	5.67×10^{-4}	−1.01
	WLPSSSPVTGYR	6.69×10^{-4}	−0.82
	WCHDNGVNYK	3.89×10^{-3}	−0.78
APOC4[c,d]	ELLETVVNR	2.45×10^{-5}	−0.94
C4A[d,e]	GLEEELQFSLGSK	5.67×10^{-4}	−0.92
	ITQVLHFTK[f]	4.73×10^{-2}	−0.22
BTD	VDLITFDTPFAGR	3.04×10^{-4}	0.67
	LSSGLVTAALYGR	1.23×10^{-3}	0.44
ITIH1	QAVDTAVDGVFIR	3.94×10^{-4}	−0.65
CP	QSEDSTFYLGER	7.88×10^{-4}	0.6
	HYYIGIIETTWDYASDHGEK	1.58×10^{-2}	0.32
C8B	LPLEYSYGEYR	1.18×10^{-4}	0.51
HPX	GECQAEGVLFFQGDR	4.28×10^{-4}	0.5
	NFPSPVDAAFR	6.34×10^{-3}	0.42
KNG1[d,e]	YFIDFVAR	3.04×10^{-4}	−0.49
TTR[e]	GSPAINVAVHVFR[f]	2.45×10^{-5}	−0.48
	AADDTWEPFASGK[f]	2.45×10^{-5}	−0.42
ALB	HPDYSVVLLLR	1.18×10^{-4}	0.43
	QNCELFEQLGEYK	1.06×10^{-3}	0.3
C3[d,e]	QELSEAEQATR	1.92×10^{-2}	−0.42
	SGSDEVQVGQQR	1.92×10^{-2}	−0.42
CPN1	IVQLIQDTR	1.48×10^{-3}	0.42
C5	TSTSEEVCSFYLK	5.87×10^{-4}	0.41
	FQNSAILTIQPK	3.85×10^{-3}	0.35
APOM	SLTSCLDSK	2.45×10^{-5}	0.41
	WIYHLTEGSTDLR	1.80×10^{-3}	0.27
C9[c]	TSNFNAAISLK	3.73×10^{-3}	−0.40
PROZ	DFAEHLLIPR	3.29×10^{-2}	−0.39
CLU[e,g]	TLLSNLEEAK	1.18×10^{-4}	−0.38
	LFDSDPITVTVPVEVSR	1.18×10^{-4}	−0.33
	ELDESLQVAER	5.87×10^{-4}	−0.28
QSOX1	AHFSPSNIILDFPAAGSAAR	7.83×10^{-3}	−0.36
APOA2	SPELQAEAK	1.01×10^{-3}	0.33
ITIH2	TEVNVLPGAK	1.42×10^{-3}	0.33
	FLHVPDTFEGHFDGVPVISK	3.85×10^{-3}	0.25
SERPINF2	LGNQEPGGQTALK[f,h]	7.24×10^{-3}	−0.32
EFEMP1	LNCEDIDECR	1.42×10^{-2}	0.31
APOD	NILTSNNIDVK	1.37×10^{-2}	0.19

[a] Univariate FDR-adjusted p values were obtained by Wilcoxon rank-sum analysis of islet autoantibody-positive vs autoantibody-negative groups

[b] Positive effect sizes represent higher abundance and negative effect sizes represent lower abundance in islet autoantibody-positive vs autoantibody-negative children

[c] Significant differences were reported by Moulder et al [19]

[d] Significant differences were reported by Zhi et al [18]

[e] Significant differences were reported by Zhang et al [17]

[f] Peptides originating from Zhang et al [17], but not selected by dCV in the selection phase of this study

[g] Significant differences were reported by Metz et al [16]

[h] Peptide levels were not significantly different in the Zhang et al [17] validation cohorts

ALB, albumin; BTD, biotinidase; C, complement component; CLU, clusterin; CPN1, carboxypeptidase N; EFEMP, epidermal growth factor-containing fibulin-like extracellular matrix protein 1; FN1, fibronectin 1; HPX, haemopexin; ITIH, inter-α-trypsin inhibitor heavy chain; KNG1, kininogen 1; PROZ, vitamin K-dependent protein Z; QSOX1, quiescin Q6 sulfhydryl oxidase 1; SERPINF2, α2-antiplasmin; TTR, transthyretin

showed highly correlated abundance patterns (ESM Fig. 3). Gene ontology enrichment analysis recovered a significant accumulation of differentially abundant proteins in terms associated with lipid metabolic processes and homeostasis, indicative of changes in lipid metabolism (ESM Table 2).

The multivariable dCV method selected two peptides, SLTSCLDSK from APOM and ELLETVVNR from APOC4, to discriminate between islet autoantibody-positive and autoantibody-negative children (Fig. 2, Table 2) and yielded an unbiased median AUC of 0.77 (IQR 0.75–0.78). Using the logistic regression coefficients (Table 2) as weights, we calculated a combined risk score to discriminate between islet autoantibody-positive and autoantibody-negative children based solely on these two peptides (Fig. 2). The AUC of 0.83 for this combined model was significantly higher than that for APOM alone (AUC 0.75) and for APOC4 alone (AUC 0.74) at $p = 2.5 \times 10^{-5}$.

Targeted proteomics to predict disease progression There were no significant univariate associations of individual peptides with progression time.

When we applied the survival dCV approach, we found that three peptides (representing hepatocyte growth factor activator [HGFAC], complement factor [CF]H and ceruloplasmin [CP]) and age at measurement were predictive covariates for progression time (Fig. 3a–c, Table 2). The median survival AUC was 0.72 (IQR 0.69–0.75). In order to investigate the improvement in discrimination conferred by the three peptides in addition to age, we compared the AUC of the combined model with that of age alone (ESM Fig. 4). The combined model displayed a significant improvement in discrimination ($p = 0.001$), mainly due to an improvement after 4 years of follow-up. Importantly, the abundance levels of most peptides, including the three peptides predictive for progression time, were not correlated with age (ESM Fig. 5). Only the levels of both peptides representing carnosine dipeptidase 1 significantly increased with age (ESM Fig. 5) but these peptides were not, however, selected by the dCV for either progression rates or autoimmunity status.

Using these peptides and age, we calculated risk scores by including the weights from Table 2 in a multivariable Cox model, and separated the children into tertiles of high, medium and low risk (Fig. 3d; ESM Table 3). Children in the low-risk group progressed to type 1 diabetes with a probability of <10% within 5 years after seroconversion (95% CI 2.2%, 29%). The corresponding rate in the high risk group was 78% (95% CI 60%, 92%) (Fig. 3d). As expected, children in the high-risk group were younger; however, they did not differ from the other risk groups in islet autoantibody status or HLA genotype (ESM Table 3).

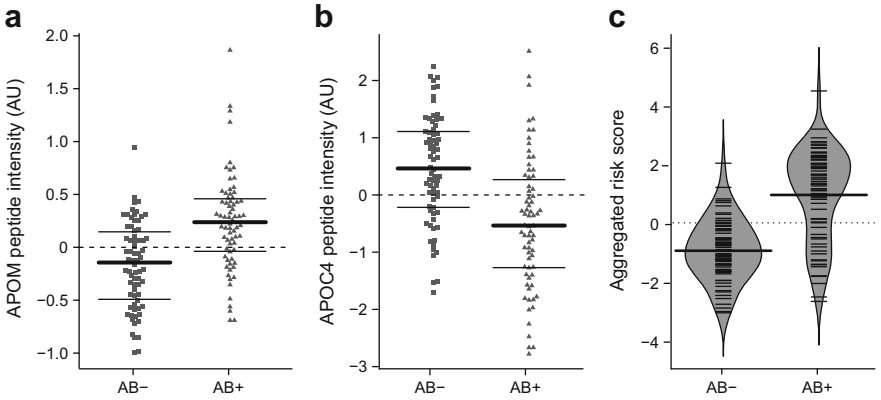


Fig. 2 Quantitative differences of the best discriminating peptides in islet autoantibody-positive (AB+) and autoantibody-negative (AB−) children. Quantifications are based on SRM measurements in the application sample set. (**a**) Adjusted peptide (SLTSCLDSK) intensities of APOM (Wilcoxon test, $p = 2.5 \times 10^{-5}$) in arbitrary units (AU). (**b**) Adjusted peptide (ELLETVVNR) intensities of APOC4 ($p = 2.5 \times 10^{-5}$). (**c**) Risk score for the final model. The risk score was calculated using a logistic regression model with the selected peptides using the weights shown in Table 2

Discussion

Using a proteomics strategy, we determined the protein expression profiles of 185 children from the BABYDIAB/BABYDIET birth cohorts with high genetic risk for type 1 diabetes. We found that 26 proteins, represented by 41 peptides, could discriminate between islet autoantibody-positive and autoantibody-negative children. The 26 proteins were enriched for pathways involved in lipid-associated metabolic processes and homeostasis, suggesting that changes in lipid metabolism occur early in the autoimmunity process. We also identified a proteomic signature that, together with age, was able to discriminate fast and slow progression to clinical diabetes in islet autoantibody-positive children.

Previous studies have used LC-MS/MS-based proteomics approaches and applied extensive prefractionation techniques on pooled samples [17, 18], followed by applying selected candidate proteins using ELISA [18], LC-SRM-MS [18] or other methods [19]. We designed our study in two phases, capitalising on the high analytical depth of a shotgun proteomics approach for selecting interesting peptides followed by an application using sensitive targeted proteomics specifically developed for the subset of potentially relevant peptides. The technical advantages of the targeted proteomics approach include high accuracy and robustness of quantifications, and that all peptides are consistently measured across all LC-MS runs, thus avoiding the occurrence of missing values.

Consistent with previous studies in children with overt type 1 diabetes, we found lower levels of APOA4 [19], APOC4 [19], CF3 [17, 18], CF4 [17–19], clusterin [16, 17], kininogen [17] and transthyretin [17] in children with islet autoantibodies. We also found lower levels of CF9 in autoantibody-positive children, while others [19] have reported slightly increased levels. The peptide of APOM that was selected for discrimination in the risk score has not been identified in previous studies. In addition, we identified changes in the levels of 18 proteins (represented by 27 peptides) that have not been previously described.

In order to prioritise the peptide signatures, we used the dCV method for feature selection. This method aims to derive a minimal, predictive combination of peptides, and to estimate

Table 2 Results of the dCV in the application phase

Protein	Sequence	Selection frequency (%)[a]	Weight
Islet autoantibody-positive vs autoantibody-negative			
APOM	SLTSCLDSK	98	2.495
APOC4	ELLETVVNR	92	−0.939
Progression to type 1 diabetes			
Age		100	–0.303
CFH	SSIDIENGFISESQYTYALK	70	0.648
HGFAC	VANYVDWINDR	68	0.443
CP	HYYIGIIETTWDYASDHGEK	52	–0.378

[a] The selection frequencies of peptides or age were calculated by applying the dCV method to SRM-adjusted intensities in the application phase (70 islet autoantibody-positive and 70 autoantibody-negative children). Peptides with selection frequencies of >50% are listed, with high selection frequencies indicating higher importance of the single peptide

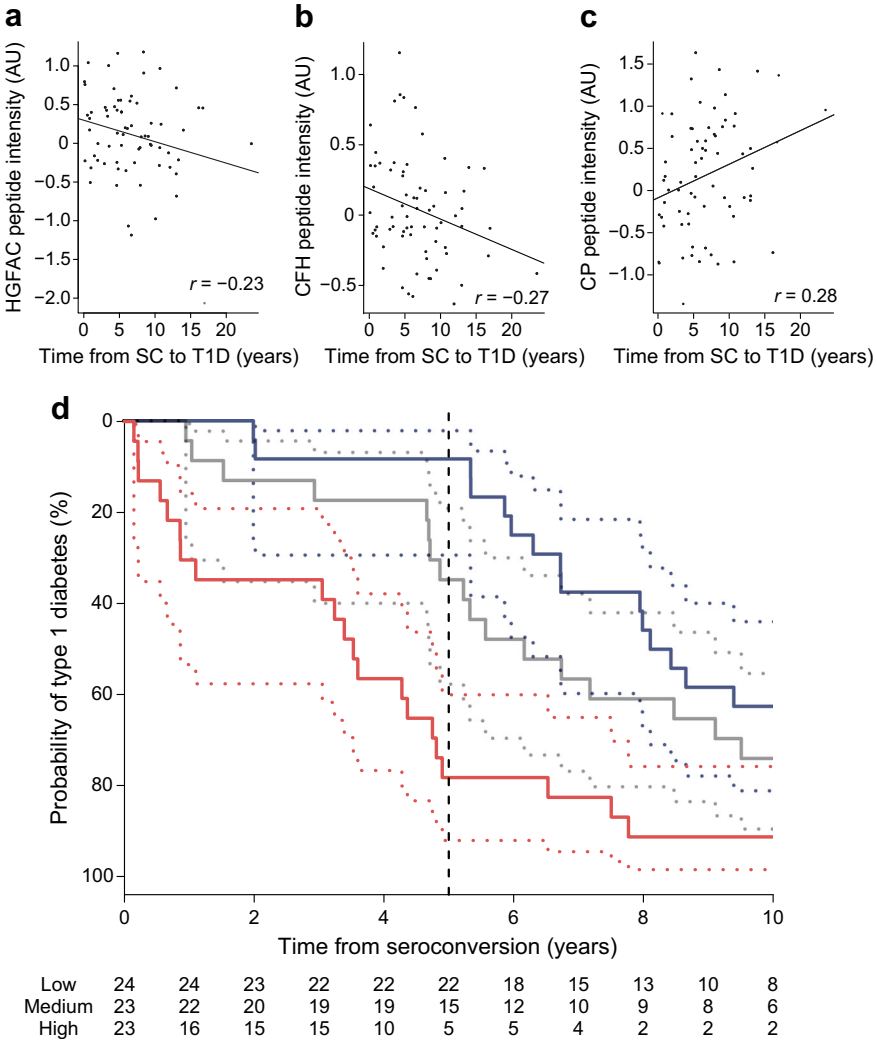


Fig. 3 Progression time analysis. Adjusted intensities of the selected peptides of (**a**) HGFAC (VANYVDWINDR), (**b**) CFH (SSIDIENGFISESQYTYALK) and (**c**) CP (HYYIGIIETTWDYASDHGEK) in arbitrary units (AU) and the corresponding time from seroconversion to type 1 diabetes in the application cohort. (**d**) Kaplan–Meier curves of the high-, medium- and low-risk score groups (defined by age, HGFAC, CFH and CP) for the time from seroconversion to type 1 diabetes. Blue line, low-risk group; grey line, medium-risk group; red line, high-risk group; dotted lines, CIs; dashed line, 5 year interval. The low- and high-risk survival curves were significantly different ($p = 1.6 \times 10^{-5}$). The numbers of children remaining at risk at a given time are shown below the time axis. SC, seroconversion; T1D, type 1 diabetes

the predictive power within a dataset in an unbiased fashion, without substantial overfitting effects [32]. Two peptides, one from APOM and one from APOC4, were deemed to be sufficient for between-group discrimination with a median AUC of 0.83. Both peptides were also among the top hits for discrimination in the univariate analysis. APOM levels were higher and APOC4 levels lower in the children with islet autoantibodies in our study. APOM is a member of the lipocalin protein family involved in lipid transport [34]. Polymorphisms in the promoter region of APOM that increase promoter activity have also been reported to increase susceptibility to the development of type 1 diabetes in two different cohorts [35]. Because the *APOM* gene is regulated by hepatocyte nuclear factor 1α (HNF1A), APOM is also considered to be a marker of HNF1A-dependent MODY. However, APOM levels have been found to be significantly lower in individuals with MODY than in those with type 1 diabetes [36].

APOC4, the other major marker for discriminating between islet autoantibody-positive and autoantibody-negative children in this study, is also a member of the APO family. The lower levels of APOC4 in autoantibody-positive children confirm previous findings reported in the Finnish Type 1 Diabetes Prediction and Prevention (DIPP) study [19], in which APOC4 levels were decreased even before seroconversion in children who eventually progressed to type 1 diabetes. Lower APOA1 levels have been reported to be associated with viral infections [37], and Moulder et al have suggested an association between viral infections and the development of type 1 diabetes [19]. So far, to the best of our knowledge, APOC4 has not been described to play a role in the immune response. However, recent studies have discussed APOs such as APOM [38] in the context of autoimmunity [38–40], and future studies might unravel as-yet unidentified roles for APOC4.

The combined discriminative power of the candidate proteins APOM and APOC4 (median AUC 0.83) is comparable with the results reported for APOC4 and afamin (AUC 0.85) in the DIPP study [19].

Another aim was to explore whether proteomic signatures could predict the progression time to type 1 diabetes in children with islet autoantibodies. We identified a set of three

peptides representing three proteins, CFH, HGFAC and CP, in addition to age, as predictive covariates for progression time with a median survival AUC of 0.72. Predictions including these peptides were slightly but significantly superior to those using age alone. Higher levels of CFH and HGFAC and lower levels of CP in combination with young age were associated with faster progression in later follow-up. CFH, HGFAC and CP have previously been discussed in relation to insulin resistance [41], type 1 diabetes [17] and type 2 diabetes [42, 43], respectively.

The strengths and novelties of our study included the multivariate statistical approach for extracting relevant peptide signatures, minimising false-positive associations; the exclusive investigation of samples from patients close to seroconversion without overt diabetes, thus reducing the confounding effect of hyperglycaemia on proteomic signatures; and the large cohort of children with islet autoimmunity. A limitation of our study is that we did not validate our signature of progression rate in a separate cohort. Other limitations include the lack of repeated longitudinal measurements and the relatively small contribution of the peptide signature to the progression risk score, as compared with age alone.

In conclusion, we found that serum proteomics signatures of islet autoantibody-positive children close to the date of seroconversion were dominated by proteins involved in lipid metabolism. Some of these protein markers have been previously identified in studies of patients with overt diabetes, and the changes in their levels close to the onset of autoimmunity suggest they are early markers. In addition, the peptide signatures significantly improved the categorisation of islet autoantibody-positive children into high- or low-risk groups for rapid progression to type 1 diabetes over age alone.

Acknowledgements The authors thank S. Becker (Research Unit Protein Science, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany) for her excellent technical assistance.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Funding This work was funded by JDRF grant 17-2012-617 and the iMED programme of the Helmholtz Association. The research programmes have also received funding from the EU's Seventh Framework Program [FP7-Health-F5-2012] under grant agreement no. 305280 (MIMOmics). This work was further supported by grants from the German Federal Ministry of Education and Research to the German Center for Diabetes Research (DZD e.V.).

Duality of interest statement The authors declare that there is no duality of interest associated with this manuscript.

Author contributions CvT designed the study, performed and analysed the MS experiments, and wrote the manuscript. ML designed and performed statistical analyses and wrote the manuscript. PA performed sample selection and reviewed/edited the manuscript. AB performed statistical analyses and reviewed/edited the manuscript. TdlHG developed the procedure to process the SRM data. JK and FJT supervised the statistical analysis and reviewed/edited the manuscript. AGZ designed the study, supervised sample collection and selection, and wrote the manuscript. SMH designed the study, supervised the MS experiments and wrote the manuscript. CvT, ML and PA are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors contributed to data interpretation, critically reviewed and edited the manuscript, and approved the version to be published.

References

1. Insel RA, Dunne JL, Atkinson MA et al (2015) Staging presymptomatic type 1 diabetes: a scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association. Diabetes Care 38:1964–1974
2. Ziegler AG, Bonifacio E (2012) Age-related islet autoantibody incidence in offspring of patients with type 1 diabetes. Diabetologia 55:1937–1943
3. Parikka V, Nanto-Salonen K, Saarinen M et al (2012) Early seroconversion and rapidly increasing autoantibody concentrations predict prepubertal manifestation of type 1 diabetes in children at genetic risk. Diabetologia 55:1926–1936
4. Krischer JP, Lynch KF, Schatz DA et al (2015) The 6 year incidence of diabetes-associated autoantibodies in genetically at-risk children: the TEDDY study. Diabetologia 58:980–987
5. Ziegler AG, Rewers M, Simell O et al (2013) Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. JAMA 309:2473–2479
6. Achenbach P, Bonifacio E, Williams AJ, Ziegler AG, Gale EA, Bingley PJ (2008) Autoantibodies to IA-2β improve diabetes risk assessment in high-risk relatives. Diabetologia 51:488–492
7. Achenbach P, Warncke K, Reiter J et al (2004) Stratification of type 1 diabetes risk on the basis of islet autoantibody characteristics. Diabetes 53:384–392
8. Achenbach P, Lampasona V, Landherr U et al (2009) Autoantibodies to zinc transporter 8 and SLC30A8 genotype stratify type 1 diabetes risk. Diabetologia 52:1881–1888
9. Achenbach P, Koczwara K, Knopff A, Naserke H, Ziegler AG, Bonifacio E (2004) Mature high-affinity immune responses to (pro)insulin anticipate the autoimmune cascade that leads to type 1 diabetes. J Clin Invest 114:589–597
10. Mayr A, Schlosser M, Grober N et al (2007) GAD autoantibody affinity and epitope specificity identify distinct immunization profiles in children at risk for type 1 diabetes. Diabetes 56:1527–1533
11. Yassine H, Borges CR, Schaab MR et al (2013) Mass spectrometric immunoassay and MRM as targeted MS-based quantitative approaches in biomarker development: potential applications to cardiovascular disease and diabetes. Proteomics Clin Appl 7:528–540
12. Pin E, Fredolini C, Petricoin EF 3rd (2013) The role of proteomics in prostate cancer research: biomarker discovery and validation. Clin Biochem 46:524–538
13. Chambers AG, Percy AJ, Simon R, Borchers CH (2014) MRM for the verification of cancer biomarker proteins: recent applications to human plasma and serum. Expert Rev Proteomics 11:137–148

14. Alberio T, Bucci EM, Natale M et al (2013) Parkinson's disease plasma biomarkers: an automated literature analysis followed by experimental validation. J Proteomics 90:107–114
15. von Toerne C, Huth C, de Las Heras Gala T et al (2016) MASP1, THBS1, GPLD1 and ApoA-IV are novel biomarkers associated with prediabetes: the KORA F4 study. Diabetologia 59:1882–1892
16. Metz TO, Qian WJ, Jacobs JM et al (2008) Application of proteomics in the discovery of candidate protein biomarkers in a diabetes autoantibody standardization program sample subset. J Proteome Res 7:698–707
17. Zhang Q, Fillmore TL, Schepmoes AA et al (2013) Serum proteomics reveals systemic dysregulation of innate immunity in type 1 diabetes. J Exp Med 210:191–203
18. Zhi W, Sharma A, Purohit S et al (2011) Discovery and validation of serum protein changes in type 1 diabetes patients using high throughput two dimensional liquid chromatography-mass spectrometry and immunoassays. Mol Cell Proteomics 10: M111.012203
19. Moulder R, Bhosale SD, Erkkila T et al (2015) Serum proteomes distinguish children developing type 1 diabetes in a cohort with HLA-conferred susceptibility. Diabetes 64:2265–2278
20. Ziegler AG, Hummel M, Schenker M, Bonifacio E (1999) Autoantibody appearance and risk for development of childhood diabetes in offspring of parents with type 1 diabetes: the 2-year analysis of the German BABYDIAB Study. Diabetes 48:460–468
21. Hummel S, Pfluger M, Hummel M, Bonifacio E, Ziegler AG (2011) Primary dietary intervention study to reduce the risk of islet autoimmunity in children at increased risk for type 1 diabetes: the BABYDIET study. Diabetes Care 34:1301–1305
22. Giannopoulou EZ, Winkler C, Chmiel R et al (2015) Islet autoantibody phenotypes and incidence in children at increased risk for type 1 diabetes. Diabetologia 58:2317–2323
23. Torn C, Mueller PW, Schlosser M, Bonifacio E, Bingley PJ (2008) Diabetes Antibody Standardization Program: evaluation of assays for autoantibodies to glutamic acid decarboxylase and islet antigen-2. Diabetologia 51:846–852
24. Schlosser M, Mueller PW, Torn C, Bonifacio E, Bingley PJ (2010) Diabetes Antibody Standardization Program: evaluation of assays for insulin autoantibodies. Diabetologia 53:2611–2620
25. Lampasona V, Schlosser M, Mueller PW et al (2011) Diabetes antibody standardization program: first proficiency evaluation of assays for autoantibodies to zinc transporter 8. Clin Chem 57:1693–1702
26. Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (2003) Report of the expert committee on the diagnosis and classification of diabetes mellitus. Diabetes Care 26(Suppl 1): S5–20
27. von Toerne C, Kahle M, Schafer A et al (2013) Apoe, Mbl2, and Psp plasma protein levels correlate with diabetic phenotype in NZO mice—an optimized rapid workflow for SRM-based quantification. J Proteome Res 12:1331–1343
28. Graessel A, Hauck SM, von Toerne C et al (2015) A combined omics approach to generate the surface atlas of human naive $CD4^+$ T cells during early T-cell receptor activation. Mol Cell Proteomics 14:2085–2102
29. Hauck SM, Dietter J, Kramer RL et al (2010) Deciphering membrane-associated molecular processes in target tissue of autoimmune uveitis by label-free quantitative mass spectrometry. Mol Cell Proteomics 9:2292–2305
30. MacLean B, Tomazela DM, Shulman N et al (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26:966–968
31. Abbatiello SE, Mani DR, Keshishian H, Carr SA (2010) Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. Clin Chem 56:291–305
32. Laimighofer M, Krumsiek J, Buettner F, Theis FJ (2016) Unbiased prediction and feature selection in high-dimensional survival regression. J Comput Biol 23:279–290
33. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30:1105–1117
34. Xu N, Dahlback B (1999) A novel human apolipoprotein (apoM). J Biol Chem 274:31286–31290
35. Wu X, Niu N, Brismar K et al (2009) Apolipoprotein M promoter polymorphisms alter promoter activity and confer the susceptibility to the development of type 1 diabetes. Clin Biochem 42:17–21
36. Mughal SA, Park R, Nowak N et al (2013) Apolipoprotein M can discriminate HNF1A-MODY from type 1 diabetes. Diabet Med 30: 246–250
37. Singh IP, Chopra AK, Coppenhaver DH, Ananatharamaiah GM, Baron S (1999) Lipoproteins account for part of the broad nonspecific antiviral activity of human serum. Antiviral Res 42:211–218
38. Tsai HC, Han MH (2016) Sphingosine-1-phosphate (S1P) and S1P signaling pathway: therapeutic targets in autoimmunity and inflammation. Drugs 76:1067–1079
39. Ley K (2016) 2015 Russell Ross Memorial Lecture in Vascular Biology: protective autoimmunity in atherosclerosis. Arterioscler Thromb Vasc Biol 36:429–438
40. Black LL, Srivastava R, Schoeb TR, Moore RD, Barnes S, Kabarowski JH (2015) Cholesterol-independent suppression of lymphocyte activation, autoimmunity, and glomerulonephritis by apolipoprotein A-I in normocholesterolemic lupus-prone mice. J Immunol 195:4685–4698
41. Moreno-Navarrete JM, Martinez-Barricarte R, Catalan V et al (2010) Complement factor H is expressed in adipose tissue in association with insulin resistance. Diabetes 59:200–209
42. Cunningham J, Leffell M, Mearkle P, Harmatz P (1995) Elevated plasma ceruloplasmin in insulin-dependent diabetes mellitus: evidence for increased oxidative stress as a variable complication. Metabolism 44:996–999
43. Memisogullari R, Bakan E (2004) Levels of ceruloplasmin, transferrin, and lipid peroxidation in the serum of patients with type 2 diabetes mellitus. J Diabetes Complications 18:193–197

SPRINGER LICENSE TERMS AND CONDITIONS

Mar 27, 2017

This Agreement between Michael Laimighofer ("You") and Springer ("Springer") consists of your license details and the terms and conditions provided by Springer and Copyright Clearance Center.

License Number	4077081403083
License date	Mar 27, 2017
Licensed Content Publisher	Springer
Licensed Content Publication	Diabetologia
Licensed Content Title	Peptide serum markers in islet autoantibody-positive children
Licensed Content Author	Christine von Toerne
Licensed Content Date	Jan 1, 2016
Licensed Content Volume	60
Licensed Content Issue	2
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	1
Author of this Springer article	Yes and you are a contributor of the new work
Order reference number	
Title of your thesis / dissertation	Statistical learning models for prediction of Type 1 Diabetes risk factors using clinical data and omics data
Expected completion date	Apr 2017
Estimated size(pages)	112
Requestor Location	Michael Laimighofer Ingolstaedter Landstr 1 Neuherberg, 85764 Germany Attn: Michael Laimighofer
Billing Type	Invoice
Billing Address	Michael Laimighofer Ingolstaedter Landstr 1 Neuherberg, Germany 85764 Attn: Michael Laimighofer
Total	0.00 EUR
Terms and Conditions	

Introduction

The publisher for this copyrighted material is Springer. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

Limited License

With reference to your request to reuse material on which Springer controls the copyright, permission is granted for the use indicated in your enquiry under the following conditions:

- Licenses are for one-time use only with a maximum distribution equal to the number stated in your request.
- Springer material represents original material which does not carry references to other sources. If the material in question appears with a credit to another source, this permission is not valid and authorization has to be obtained from the original copyright holder.
- This permission
• is non-exclusive
• is only valid if no personal rights, trademarks, or competitive products are infringed.
• explicitly excludes the right for derivatives.
- Springer does not supply original artwork or content.
- According to the format which you have selected, the following conditions apply accordingly:
• **Print and Electronic:** This License include use in electronic form provided it is password protected, on intranet, or CD-Rom/DVD or E-book/E-journal. It may not be republished in electronic open access.
• **Print:** This License excludes use in electronic form.
• **Electronic:** This License only pertains to use in electronic form provided it is password protected, on intranet, or CD-Rom/DVD or E-book/E-journal. It may not be republished in electronic open access.

For any electronic use not mentioned, please contact Springer at permissions.springer@spi-global.com.

- Although Springer controls the copyright to the material and is entitled to negotiate on rights, this license is only valid subject to courtesy information to the author (address is given in the article/chapter).
- If you are an STM Signatory or your work will be published by an STM Signatory and you are requesting to reuse figures/tables/illustrations or single text extracts, permission is granted according to STM Permissions Guidelines: http://www.stm-assoc.org/permissions-guidelines/

For any electronic use not mentioned in the Guidelines, please contact Springer at permissions.springer@spi-global.com. If you request to reuse more content than stipulated in the STM Permissions Guidelines, you will be charged a permission fee for the excess content.

Permission is valid upon payment of the fee as indicated in the licensing process. If permission is granted free of charge on this occasion, that does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

-If your request is for reuse in a Thesis, permission is granted free of charge under the following conditions:

This license is valid for one-time use only for the purpose of defending your thesis and with a maximum of 100 extra copies in paper. If the thesis is going to be published, permission needs to be reobtained.

- includes use in an electronic form, provided it is an author-created version of the thesis on his/her own website and his/her university's repository, including UMI (according to the definition on the Sherpa website: http://www.sherpa.ac.uk/romeo/);
- is subject to courtesy information to the co-author or corresponding author.

Geographic Rights: Scope

Licenses may be exercised anywhere in the world.

Altering/Modifying Material: Not Permitted

Figures, tables, and illustrations may be altered minimally to serve your work. You may not alter or modify text in any manner. Abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s).

Reservation of Rights

Springer reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction and (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

License Contingent on Payment
While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Springer or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received by the date due, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Springer reserves the right to take any and all action to protect its copyright in the materials.
Copyright Notice: Disclaimer
You must include the following copyright and permission notice in connection with any reproduction of the licensed material:
"Springer book/journal title, chapter/article title, volume, year of publication, page, name(s) of author(s), (original copyright notice as given in the publication in which the material was originally published) "With permission of Springer"
In case of use of a graph or illustration, the caption of the graph or illustration must be included, as it is indicated in the original publication.
Warranties: None
Springer makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.
Indemnity
You hereby indemnify and agree to hold harmless Springer and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
No Transfer of License
This license is personal to you and may not be sublicensed, assigned, or transferred by you without Springer's written permission.
No Amendment Except in Writing
This license may not be amended except in a writing signed by both parties (or, in the case of Springer, by CCC on Springer's behalf).
Objection to Contrary Terms
Springer hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
Jurisdiction
All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in the Federal Republic of Germany, in accordance with German law.
Other conditions:
V 12AUG2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Prediction of type 1 diabetes using a genetic risk model in the Diabetes Autoimmunity Study in the Young (DAISY).

Received: 23 November 2016 | Revised: 24 April 2017 | Accepted: 25 April 2017

DOI: 10.1111/pedi.12543

WILEY ISPAD

ORIGINAL ARTICLE

Prediction of type 1 diabetes using a genetic risk model in the Diabetes Autoimmunity Study in the Young

Brigitte I Frohnert[1†] | Michael Laimighofer[2†] | Jan Krumsiek[2,3] | Fabian J Theis[2] | Christiane Winkler[4] | Jill M Norris[5] | Anette-Gabriele Ziegler[4] | Marian J Rewers[1] | Andrea K Steck[1]

[1]Barbara Davis Center for Childhood Diabetes, School of Medicine, University of Colorado, Aurora, Colorado

[2]Institute of Computational Biology, Helmholtz Zentrum München, München-Neuherberg, Germany

[3]German Center for Diabetes Research (DZD), München-Neuherberg, Germany

[4]Institute of Diabetes Research, Helmholtz Zentrum München and Forschergruppe Diabetes, Klinikum rechts der Isar, Technische, Universität München, Neuherberg, Germany

[5]Department of Epidemiology, Colorado School of Public Health, University of Colorado, Aurora, Colorado

Correspondence
Brigitte I Frohnert, MD, PhD, Barbara Davis Center, 1775 Aurora Court, A140, Aurora, Colorado 80045.
Email: brigitte.frohnert@ucdenver.edu

Funding Information
JDRF, Grant/Award numbers: 17-2013-535, 11-2010-206, 2-SRA-2015-13-Q-R; National Institutes of Health, Grant/Award numbers: R01 DK32083, DK32493, DK049654, 5K12DK094712, P30 DK57516; Leona M. and Harry B. Helmsley Charitable Trust, Grant/Award number: 2015PG-T1D072; Helmholtz-Gemeinschaft Grant/Award number: HIRG-0018; German BMBF-funded iMed Helmholtz Alliance.

Background: Genetic predisposition for type 1 diabetes (T1D) is largely determined by human leukocyte antigen (HLA) genes; however, over 50 other genetic regions confer susceptibility. We evaluated a previously reported 10-factor weighted model derived from the Type 1 Diabetes Genetics Consortium to predict the development of diabetes in the Diabetes Autoimmunity Study in the Young (DAISY) prospective cohort. Performance of the model, derived from individuals with first-degree relatives (FDR) with T1D, was evaluated in DAISY general population (GP) participants as well as FDR subjects.

Methods: The 10-factor weighted risk model (HLA, *PTPN22*, *INS*, *IL2RA*, *ERBB3*, *ORMDL3*, *BACH2*, *IL27*, *GLIS3*, *RNLS*), 3-factor model (HLA, *PTPN22*, *INS*), and HLA alone were compared for the prediction of diabetes in children with complete SNP data (n = 1941).

Results: Stratification by risk score significantly predicted progression to diabetes by Kaplan-Meier analysis (GP: P = .00006; FDR: P = .0022). The 10-factor model performed better in discriminating diabetes outcome than HLA alone (GP, P = .03; FDR, P = .01). In GP, the restricted 3-factor model was superior to HLA (P = .03), but not different from the 10-factor model (P = .22). In contrast, for FDR the 3-factor model did not show improvement over HLA (P = .12) and performed worse than the 10-factor model (P = .02)

Conclusions: We have shown a 10-factor risk model predicts development of diabetes in both GP and FDR children. While this model was superior to a minimal model in FDR, it did not confer improvement in GP.

Differences in model performance in FDR vs GP children may lead to important insights into screening strategies specific to these groups.

KEYWORDS

child, diabetes mellitus, epidemiology, prospective study, risk factors, type 1

1 | INTRODUCTION

The increasing incidence of type 1 diabetes (T1D) in children can be attributed to the action of environmental factors in a context of genetic predisposition. Current screening strategies for at-risk individuals are limited by the low specificity of genetic screening in the general population (GP), who make up about 85% of those who develop T1D,[1] and low sensitivity in those with a family history of T1D. More precise identification of children at high risk of developing T1D is important for recruitment into natural history studies to better understand the etiologic factors contributing to islet autoimmunity (IA) and T1D. Moving forward, these strategies will also allow for better identification of individuals who could benefit from both primary and secondary prevention trials. Further, as population screening for

[†]These authors contributed equally to this work.

IA is being explored,[2] better definition of genetic risk could serve as a second line of screening.

The human leukocyte antigen (HLA) region of chromosome 6p21 plays a significant role, conferring up to 50% of the genetic risk for diabetes.[3] In addition, however, more than 50 other genetic susceptibility markers have been associated with development of T1D.[1,4–9] While any 1 non-HLA gene may not confer significant risk increase alone, an improvement in prediction strategy can be achieved by giving weight to varying gene contributions.[10] We have previously used multivariable logistic regression and Bayesian feature selection to generate a weighted risk model with single nucleotide polymorphisms (SNPs) selected from 41 genetic susceptibility markers included in data from the Type 1 Diabetes Genetics Consortium (T1DGC) dataset.[10] This 10-factor model included HLA genotype plus 9 SNPs from the *PTPN22*, *INS*, *IL2RA*, *ERBB3*, *ORMDL3*, *BACH2*, *IL27*, *GLIS3*, and *RNLS* genes and was used to predict progression to T1D or multiple islet autoantibody positive status.[10] The model was validated in a group of children and young adults less than 20 years of age with new onset diabetes,[11] and in a group of T1D parents of children with T1D from the German BABYDIAB study, with the non-T1D parents as the control group.[12] In this previous analysis, the 10-factor model showed improved discrimination of those at risk for the development of T1D when compared with HLA genotype alone. Further, the risk of development of T1D or multiple IA status was significantly higher in the portion of the prospectively followed BABYDIAB cohort with risk scores in the upper quintile when compared to the lower quintile, using Kaplan-Meier analysis.[10]

The Diabetes Autoimmunity Study in the Young (DAISY) is a cohort consisting of first-degree relatives (FDR) of T1D patients, similar to the BABYDIAB study. However, unlike BABYDIAB, the DAISY cohort also includes individuals recruited from the GP based on high-risk HLA status. The purpose of the current study was to validate the weighted 10-factor model of HLA plus 9 other SNPs for prediction of development of T1D in participants of the DAISY cohort. Validation in this novel cohort allows examination of the performance of a model developed from a T1D FDR cohort in a group of GP individuals. Further, the performance of this 10-factor model was compared with a more parsimonious 3-factor model or HLA alone.

2 | METHODS

2.1 | Study participants

DAISY is a prospective cohort study that has followed 2547 children at increased risk of T1D for a median of 9 years. The details of screening and follow-up have been previously published.[13,14] Recruitment began in 1993 and included 2 groups of children: FDR of T1D patients, enrolled between birth and 7 years of age; and GP subjects born at a Denver, Colorado hospital. While FDR subjects were enrolled regardless of HLA type, GP subjects were enriched for high-risk HLA-DR,DQ susceptibility genotypes for T1D. Specifically, of the 31 881 newborns screened, all children with DR3/4,DQB1*0302, DR3/3, and DR4/4,DQB1*0302 and a sample of those with DR4/DRx, DQB1*0302, or DR3/DRx (where DRx $\neq$ DR3 or DR4) were invited to participate in DAISY. Distribution of HLA types for the GP and FDR subjects is shown in Figure S1, Supporting Information. Characteristics of FDR and GP participants are shown in Table S1. Follow-up results are available through July of 2015. At this point, 94 participants had developed T1D. Written informed consent was obtained from the parents of study participants. The Colorado Multiple Institutional Review Board approved all study protocols.

2.2 | Outcome measures

Children in DAISY were tested for islet autoantibodies during the prospective follow-up, beginning at 9 months, 15 months, 24 months and, if negative, annually thereafter. DAISY subjects were tested for glutamic acid decarboxylase (GAD65) autoantibody (GADA), insulin autoantibody (IAA), and islet antigen 2 autoantibody (IA-2A), which were performed in the Clinical Immunology Laboratory at the Barbara Davis Center using radio-immunoassays as previously described.[15] Since January 2010, GADA and IA-2A have been measured with a harmonized assay.[16] Positive antibody tests are confirmed by blinded quality control. If positive for any of these 3 antibodies, subjects are tested more frequently (every 3-6 months). After the identification of zinc transporter 8 (ZnT8) as a T1D-associated antigen,[17] all subjects who had ever been antibody positive were retrospectively tested for ZnT8 antibodies, and this is now carried out prospectively. ZnT8A antibodies were measured in the Clinical Immunology Laboratory at the Barbara Davis Center, as previously described.[17,18] Study subjects were considered persistently islet autoantibody positive if they had at least 2 confirmed positive samples that were not because of maternal islet autoantibody transfer or if they had 1 confirmed positive sample and developed diabetes prior to the next sample collection. Subjects were considered multiple antibody positive when they were persistently antibody positive and had tested positive for more than 1 autoantibody. T1D was diagnosed according to American Diabetes Association criteria. Time to development of T1D was defined as time from birth to T1D diagnosis.

2.3 | Genotyping

Typing for HLA class II alleles at HLA-DRB1, HLA-DQA1, and HLA-DQB1 was previously described.[19,20] In the DAISY study, genotyping of 9 non-HLA SNPs was as follows: *INS*-23Hph1 (rs689), *PTPN22* R620W (rs2476601) polymorphisms were genotyped using a linear array (immobilized probe) method essentially as described in Mirel et al. The following SNPs were genotyped in the laboratory of Dr. Cisca Wijmenga using Illumina GoldenGate Beadexpress assays (veracode 48-plex): *IL2RA* (rs12251307) and *BACH2* (rs11755527). Taqman SNP genotyping assays (Applied Biosystems, Foster City, California) were utilized to obtain genotype information for *GLIS3* (rs7020673), *GSDM* (rs2290400), *ERBB3* (rs2292239), and *IL27* (rs4788084) as described previously.[19] All but one of the SNPs used in the Winkler et al risk model[10] were present in the DAISY dataset. The SNP for *IL2RA* measured in the DAISY cohort differed from the SNP in the T1DGC study used for development of the Winkler et al 10-factor risk model.[10] The *IL2RA* SNP from the DAISY dataset (rs12251307) is located in an intergenic region, while the SNP used in the Winkler et al model (rs12722495)[10] is from an intron of *IL2RA*. The 2 SNPs were queried for linkage disequilibrium using the Broad

Institute SNP Annotation and Proxy Search Pairwise LD tool on the 1000 Genomes Pilot 1 dataset,[21] which resulted in an R^2 of .543 and a D′ of .843.

HLA risk genotypes were categorized as: 6 = DR3/DR4-DQ8; 5 = DR4-DQ8/DR4-DQ8; 4 = DR3/DR3; 3 = DR4-DQ8/x; 2 = DR3/DRx; 1 = DRx/DRx (where x represents the non-DR3 and non-DR4-DQ8 alleles). For other SNPs, a score of 2 was given to persons homozygous for the susceptibility allele, 1 when heterozygous, and 0 when homozygous for the non-susceptibility allele.

2.4 | Statistical analysis

Demographic characteristics were compared using chi-square analysis for categorical variables and Wilcoxon rank sum tests for continuous variables. Time of follow-up was calculated from birth to age at last visit or diagnosis of T1D. The weights generated by previous analysis[10] were used to calculate risk scores in the DAISY children. Specifically, the risk score per patient *i* in DAISY was calculated as:

$$\text{risk score}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \text{SNP}_{i,j} \quad (1)$$

using the weights $\hat{\beta}_j$ and the intercept $\hat{\beta}_0$ derived in Winkler et al[10] and *p* is the number of SNPs plus the HLA risk categories. The 9 non-HLA SNPs used in the model were: *PTPN22* R620W (rs2476601), *INS*-23Hph1 (rs689), *IL2RA* (rs12722495), *ERBB3* (rs2292239), *ORMDL3/GSDM* (rs2290400), *BACH2* (rs11755527), *IL27* (rs4788084), *GLIS3* (rs7020673), and *RNLS/C10orf59* (rs10509540) genes.[10]

The discriminative power of the model was assessed using receiver operating characteristic (ROC) analysis and the area under the ROC curve (ROC AUC). Model refits on the DAISY GP and FDR group were performed using the abovementioned multivariable logistic regression model. Improvement in prediction by additional markers was quantified using the integrated discrimination improvement, IDI.[22] Kaplan-Meier survival curves were obtained based on the resulting risk score and time from birth to development of T1D. Individuals lost to follow-up were treated as censored data. Differences in survival curves were assessed using log-rank tests. All analyses were performed using R version 3.2.0.[23]

3 | RESULTS

Using the 10-factor genetic risk model, risk scores were calculated for the 1941 individuals from the DAISY cohort with SNP data available. The genetic risk score distributions from DAISY participants, both GP and FDR, who were T1D cases vs those who did not develop T1D are shown in Figure 1. For both GP and FDR, the children who developed T1D had a risk score distribution that was shifted to higher risk scores compared with those who did not progress to T1D. The FDR participants showed a much broader distribution of risk scores than the GP individuals, whose risk scores were more tightly clustered. Interestingly, amongst the FDR children, both T1D cases and controls showed lower risk score distributions relative to the GP group, which can be explained by the HLA selection of the GP group.

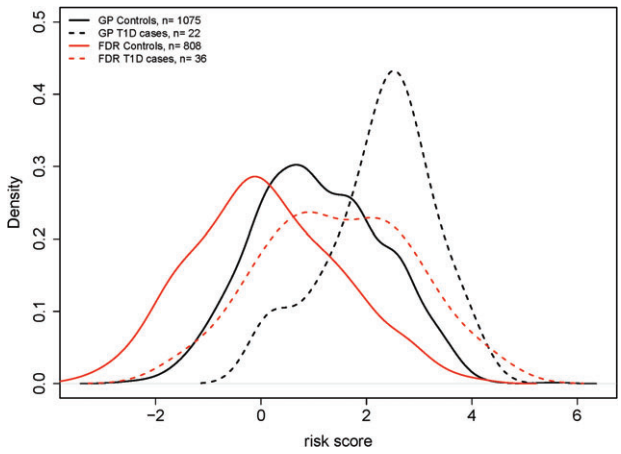


FIGURE 1 SNP Risk Score Distribution for type 1 diabetes (T1D) vs non-T1D controls in Diabetes Autoimmunity Study in the Young (DAISY) subjects recruited from first-degree relatives (FDR) of T1D patients or the general population (GP). The risk score (x-axis) is calculated using the 10-factor weighted risk model. A shift to the right indicates a higher risk score. Density (y-axis) represents the portion of the subjects with a risk score at each level. Red lines represent children who have a FDR with T1D. Black lines represented children recruited from the GP. In each case, the risk score distribution of the T1D children is represented by dashed lines, while the risk score distribution of the non-T1D children is represented by solid lines

In order to examine the discriminative power of the 10-factor model in DAISY GP and FDR subjects, these 2 subgroups were examined separately using the ROC AUC. Given the importance of HLA genotype, the ROC AUC was first calculated using HLA genotype as a sole predictor. A minimal risk model of 3 factors (HLA plus the top 2 weighted SNPs, *PTPN22* and *INS*) and the full 10-factor model were also examined for discrimination between T1D and non-T1D outcomes (Figure 2).

For both the GP and FDR groups, the probability of T1D outcome was calculated using HLA alone, the 3-factor model or the 10-factor model. In all cases, the mean risk score was higher in those who did indeed develop T1D than in those who did not. Thus, the discrimination slope, an estimation of the difference in probabilities of outcome, was positive for all 3 risk score calculations in both GP and FDR (Figure 3), The improvement in prediction by new or additional information can be quantified using the integrated discrimination improvement, IDI, an estimation in the difference in discrimination slopes.[22] Examination of the performance of the 10-factor model and the simpler 3-factor model relative to HLA risk group prediction, as well as to each other, was calculated in the DAISY GP and FDR groups. For the DAISY GP participants, both the 10-factor (P = .03) and the 3-factor (P = .03) models showed improvement over HLA alone (Figure 3A). Of note, in the GP group, comparison of the 10-factor model to the 3-factor model showed no significant improvement (P = .22), indicating that the adding 7 more SNPs provided no additional information compared with the more minimal 3-factor model. In the DAISY FDR group, the 10-factor model showed improvement over both HLA alone (P = .01) and the 3-factor model (P = .02). However, there was no significant difference between the 3-factor model and HLA alone (P = .12, Figure 3B).

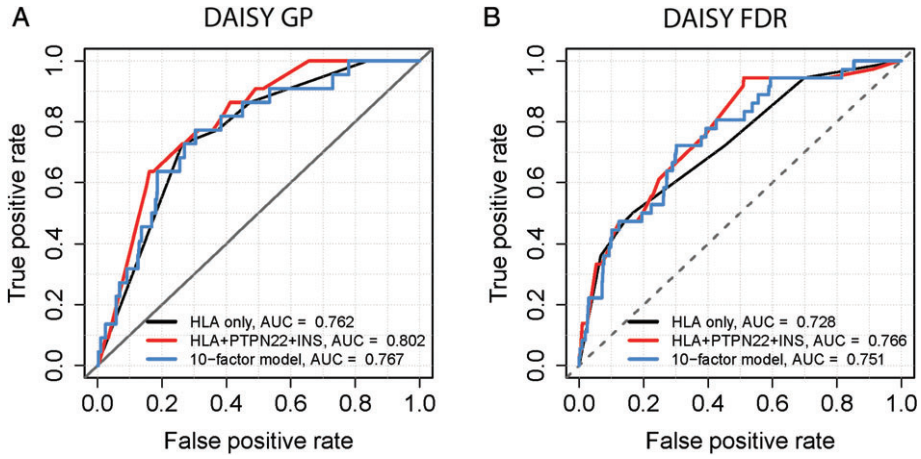


FIGURE 2 Receiver operator curve (ROC) for the prediction of type 1 diabetes (T1D) outcome using human leukocyte antigen (HLA) only, 3-factor model (HLA plus *PTPN22* and *INS* SNPs) and 10-factor model for (A) Diabetes Autoimmunity Study in the Young (DAISY) first-degree relatives (FDR), (B) DAISY subjects from the general population (GP)

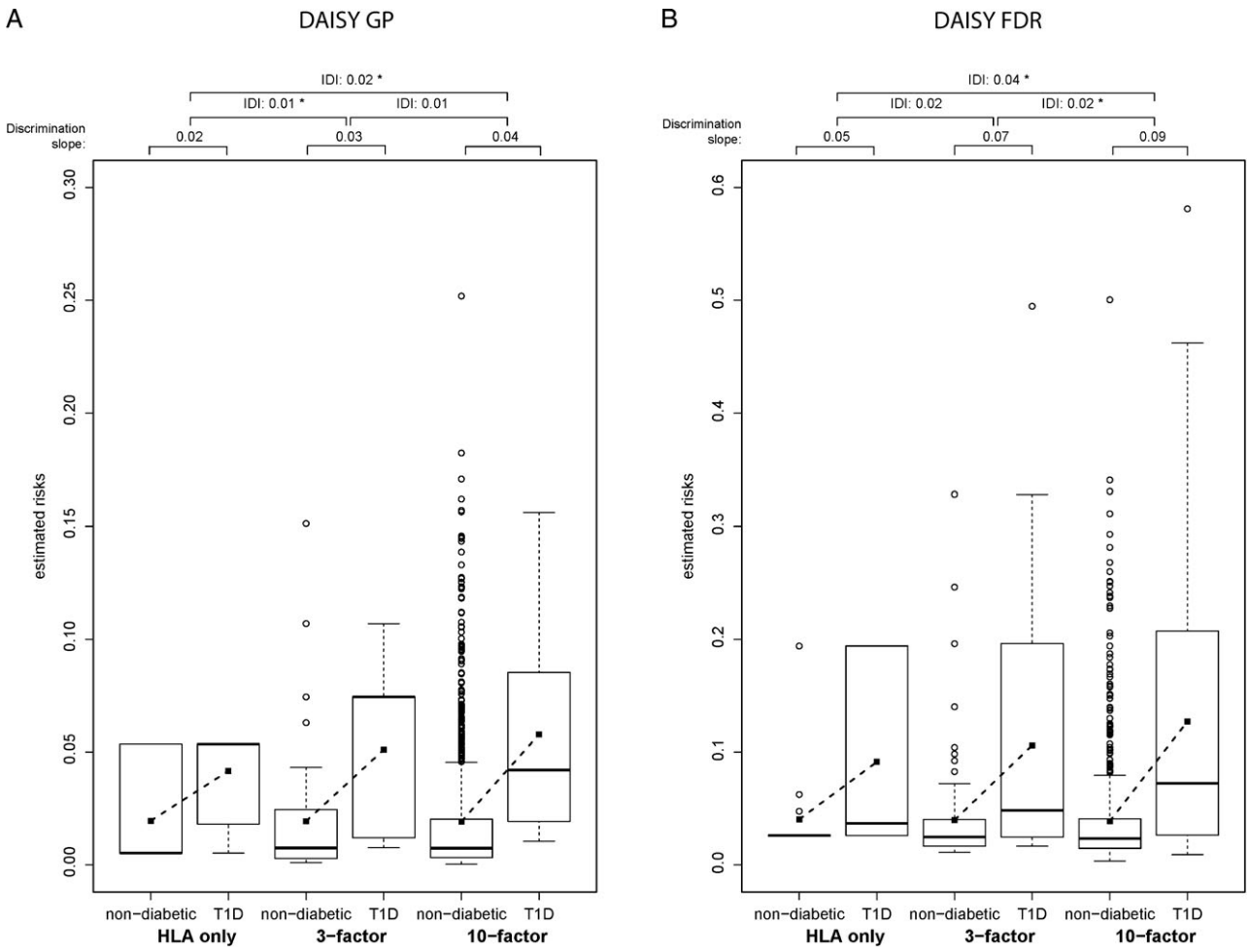


FIGURE 3 Comparison of classification performance using human leukocyte antigen (HLA) prediction alone vs 3-factor or 10-factor models in Diabetes Autoimmunity Study in the Young (DAISY) subjects recruited from (A) general population (GP) or (B) first-degree relatives (FDR) of type 1 diabetes (T1D) patients. Discrimination slope measures (difference in mean probability of T1D outcome) are given for risk score calculated by HLA, 3-factor or 10-factor model. Reclassification index integrated discrimination improvement (IDI) measures the improvement in classification performance for 1 model relative to another. *$P < .05$

In order to demonstrate the impact of the various models on prediction, the 3 models were used to classify the GP (Table 1) and FDR (Table 2) study participants using a fixed sensitivity cutoff of .5. For the GP individuals, the 3-factor model was able to identify 12 true positives, compared with the HLA only model which did not identify any true positives at this cutoff. The 10-factor model showed no improvement in categorization over the 3-factor model, paralleling the findings from the reclassification index analysis. For the FDR individuals, there were no changes from HLA alone to the 3-factor model; however, the 10-factor model identified 54 additional true negatives by reducing the number of false positives, illustrating the superior performance of the 10-factor model demonstrated in the reclassification analysis.

TABLE 1 Reclassification in Diabetes Autoimmunity Study in the Young (DAISY) general population screened participants for a sensitivity cutoff of .5. ND, nondiabetic; T1D, type 1 diabetes (n = 1097)

	HLA only		3-Factor model		10-Factor model	
	True outcome		True outcome		True outcome	
Model prediction	ND	T1D	ND	T1D	ND	T1D
ND	1075	22	932	10	940	11
T1D	0	0	143	12	135	11

TABLE 2 Reclassification in Diabetes Autoimmunity Study in the Young (DAISY) first-degree relatives for a sensitivity cutoff of .5. ND, nondiabetic; T1D, type 1 diabetes (n = 844)

	HLA only		3-Factor model		10-Factor model	
	True outcome		True outcome		True outcome	
Model prediction	ND	T1D	ND	T1D	ND	T1D
ND	674	18	677	19	731	18
T1D	134	18	131	17	77	18

The performance of the 10-factor risk model for predicting time to development of T1D was examined in the DAISY GP and FDR subgroups (Figure 4). Children from each group were stratified by risk score into the upper quintile, middle 3 quintiles, and lower quintile. Kaplan-Meier analysis was used to determine cumulative risk of development of T1D over follow-up from birth. The risk model showed significant discrimination for risk of development of T1D between the highest and lowest quintiles by risk score for both the GP (Figure 4A, P = .00006) and FDR (Figure 4B, P = .00022) groups. Of note, amongst the GP group, those with risk scores in the bottom quintile showed no incidence of T1D over the first 15 years (Figure 4B) while in the highest quintile, T1D-free survival probability at 15 years was .88 (95% CI: .93, .81). For the DAISY FDR group, disease-free survival at 15 years in the highest vs lowest quintiles by risk score was .87 (95% CI: .92, .80) vs .98 (95% CI: .99, .92).

4 | DISCUSSION

While we[10,19,24,25] and others[26–28] have previously examined the role of minor genetic susceptibility genes to improve prediction of autoimmunity and progression to T1D, this is the first study to evaluate the performance of a risk model derived from a group of FDR with T1D in a prospective cohort of individuals recruited from the GP as well as FDR. We have shown that the Winkler et al 10-factor model[10] was able to effectively predict T1D outcome and showed improvement in classification over HLA class alone in both GP and FDR children.

A more minimalist 3-factor model using HLA type plus *PTPN22* and *INS* SNPs was able to improve outcome classification over HLA alone in GP, but did not reach significance for FDR children. The differences between the 3-factor model's performance in the FDR and GP subgroups may stem partly from the enrichment of the GP group for high-risk HLA types (Figure S1). Amongst the enriched population, the role of *PTPN22* and *INS* in prediction of outcome becomes more pronounced. In contrast, in the more diversely distributed HLA background of the FDR group, HLA plays a more prominent role in prediction of outcome and therefore the 3-factor model does not provide enough new discriminating information to show improvement over HLA alone, but the 10-factor model does.

While the 10-factor model was significantly better at prediction of outcome than either HLA alone or the minimal 3-factor model in the FDR group, it showed no improvement over the 3-factor in the GP group. These findings may imply that children from the GP without a close family history of T1D may have a different profile of predictive minor genetic susceptibility genes from those with a T1D FDR. The 10-factor model was developed in the T1DGC dataset, a cohort composed of families with multiple members diagnosed with

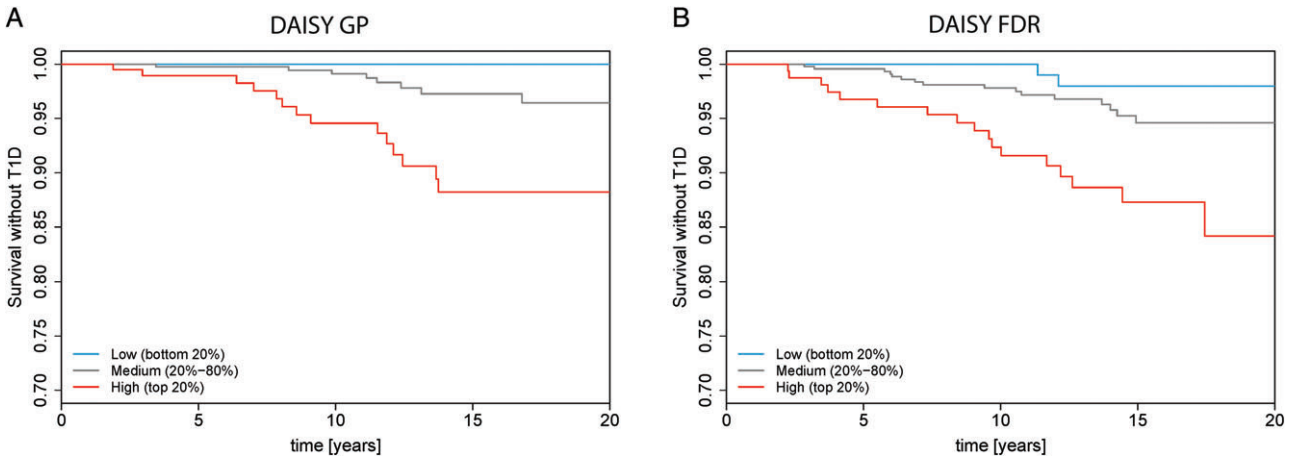


FIGURE 4 Disease-free survival without type 1 diabetes (T1D) for Diabetes Autoimmunity Study in the Young (DAISY) general population (GP) and first-degree relatives (FDR) children, separated by highest risk score quintile (red), middle 3 risk quintiles (gray) and lowest risk quintile (blue) as determined using the 10-factor risk model. Time measured from birth to event. (A) GP subjects in DAISY cohort (highest vs lowest quintile, P-value = .00006) (B) First-degree relatives in DAISY cohort (highest vs lowest quintile, P-value = .00022)

T1D. Of the 10 factors in the risk model, most of the SNPs with smaller effect sizes would not have been selected had the risk model been fit on either DAISY group, based on their performance in the model refitting (Figure S2). Indeed, when features are ranked in DAISY GP and FDR separately and added in a stepwise fashion, the feature ranking algorithm selects a different subset of features for GP (*IL2RA* not selected) and FDR groups (*IL27* and *ORMDL3* not selected) (Figure S4A,B). This could potentially reflect differences in populations. The DAISY study is composed of subjects from 1 geographic region. The T1DGC dataset, in contrast, was derived from multiple countries. Another potential effect is the role of age. The T1DGC dataset includes older adults with T1D among the parents of the family units, therefore including individuals who could have been diagnosed well into adulthood, while the DAISY population is younger. Only some of the risk SNPs identified by Winkler et al's 10-factor model have been identified in previous multivariate analysis of genetic risk markers in the DAISY cohort. Specifically, both *PTPN22* and *UBASH3A* (rs11203203 AA) were identified as being significantly associated with development of IA, while *GLIS3* and *IL2RA* showed borderline association.[19] In addition, *INS* and *UBASH3A* (rs11203203 AA) were significantly associated with progression from IA to T1D, while *PTPN22* showed borderline association.[19]

We have shown that the 10-factor model derived by Winkler et al[10] is also able to effectively discriminate between children who are likely to progress to T1D over the next 15 years and those who are not by survival analysis of longitudinal data. Of children from the DAISY cohort recruited from the GP, none of the children with a low risk score progressed to T1D over the follow-up time, indicating that the stratification by risk score is useful in GP children as well as FDR children, as previously described.[10] It is interesting to note that the GP children in the lowest risk quintile were largely from the lowest risk HLA group (DRx/DRx, where x represents the non-DR3 and non-DR4-DQ8 alleles) (Figure S3), underlining the importance of HLA genotype for prediction of T1D risk.

One limitation of this study is the difference between the *IL2RA* SNP used for the DAISY cohort from the SNP identified in the risk modeling from the T1DGC cohort. While these SNPs are associated by linkage disequilibrium, the difference may be enough to affect the performance of the risk model. While the performance of the risk model in GP vs FDR groups is intriguing, the enrichment of the GP group for moderate and high-risk HLA types may limit the generalizability to the population as a whole. Another limitation of this study is the characteristics of the non-T1D group as it represents a group already selected for T1D risk. Overall, the ROC AUCs in the DAISY cohort were lower than the values in the T1DGC dataset and the German validation set described previously (.87 and .84, respectively),[10] although ROC AUC would be expected to be lower in a validation set. Also, both of the previously studied datasets contain a large group of control subjects who were from the GP and had only background risk for T1D. In contrast, the DAISY "control" groups consist of study subjects who were recruited because of their high-risk status as either a FDR or as having a high-risk HLA type and may still eventually progress to T1D in later life.

Although the majority of new onset patients with T1D have no family history,[1] the prevalence of T1D is relatively low in the GP, about .5% by age 20, use of these prediction models for GP screening would inevitably result in many false positives. As with the selection of the DAISY study population, an initial selection for higher risk HLA types or FDR status would improve the performance. Another strategy would be to apply the risk model with high sensitivity and then follow with a secondary screening. To date, this has meant serial islet autoantibody measurements. Future research may identify other layered screening strategies, for instance other prognostic markers in the blood, urine or microbiome. Alternately, identification of a constellation of pathogenic environmental exposures could yield an exposure risk assessment or score. A recent analysis found that the use of HLA with antibody measurement every 6 months until age 5 was not cost-effective if the outcome was prevention of DKA. However, less frequent autoantibody testing and/or advancements in laboratory technology enabling cost reduction in HLA and autoantibody testing to less than $1 and $.03, respectively, would render this strategy cost-effective.[29]

In conclusion, we have shown that a 10-factor risk model, previously validated in FDR of T1D patients, is also effective in prediction of T1D outcome in a cohort of children including both those with and without family history of T1D. While the 10-factor model is superior to a more minimal 3-factor model including only HLA, *PTPN22* and *INS* in FDR children, its performance does not differ from the 3-factor model in GP children. Children from the GP with low risk scores, as determined by the 10-factor model, did not progress to T1D over up to 20 years of follow-up. This may indicate the utility of a genetic risk model for selection of children for future prevention trials or population screening programs. Further, the differences between model performances in the GP relative to the FDR groups indicate that there may be important genetic risk factors to be discovered by differentiating these populations in future analysis. The identification of subgroup differences in the performance of risk modeling may lead to increased insight into genetic factors which play a role in epidemiologic variation across regions and ethnic groups as well as FDR compared with those without family history of T1D.

ACKNOWLEDGEMENTS

This research utilizes resources provided by the T1DGC, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and the Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418. The UK case series collection was additionally funded by the JDRF and Wellcome Trust and the National Institute for Health Research Cambridge Biomedical Centre, at the Cambridge Institute for Medical Research, UK (CIMR), which is in receipt of a Wellcome Trust Strategic Award (079895). The SNP data from the T1DGC study were supplied by the NIDDK Central Repositories. This manuscript was not prepared in collaboration with Investigators of the T1DGC study and does not necessarily reflect the opinions or views of the T1DGC study, the NIDDK Central Repositories or the study sponsors.

This work was supported by JDRF grants 17-2013-535, 11-2010-206, 2-SRA-2015-13-Q-R and the National Institutes of Health grants R01 DK32083, DK32493, DK049654, and 5K12DK094712 and NIH NIDDK grant number "P30 KD57516". Also supported by Leona M. and Harry B. Helmsley Charitable Trust 2015PG-T1D072, Helmholtz HIRG-0018. This work was supported by iMed–the Helmholtz Initiative on Personalized Medicine.

REFERENCES

1. Michels A, Zhang L, Khadra A, Kushner JA, Redondo MJ, Pietropaolo M. Prediction and prevention of type 1 diabetes: update on success of prediction and struggles at prevention. *Pediatr Diabetes*. 2015;16:465-484.
2. Insel RA, Dunne JL, Ziegler A-G. General population screening for type 1 diabetes: has its time come? *Curr Opin Endocrinol Diabetes Obes*. 2015;22:270-276.
3. Noble JA, Valdes AM, Cook M, Klitz W, Thomson G, Erlich HA. The role of HLA class II genes in insulin-dependent diabetes mellitus: molecular analysis of 180 Caucasian, multiplex families. *Am J Hum Genet*. 1996;59:1134-1148.
4. Ilonen J, Kiviniemi M, Lempainen J, et al. Genetic susceptibility to type 1 diabetes in childhood – estimation of HLA class II associated disease risk and class II effect in various phases of islet autoimmunity. *Pediatr Diabetes*. 2016;17:8-16.
5. Concannon P, Erlich HA, Julier C, et al. Type 1 diabetes: evidence for susceptibility loci from four genome-wide linkage scans in 1,435 multiplex families. *Diabetes*. 2005;54:2995-3001.
6. Cooper JD, Smyth DJ, Smiles AM, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet*. 2008;40:1399-1401.
7. Barrett JC, Clayton DG, Concannon P, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet*. 2009;41:703-707.
8. Concannon P, Chen W-M, Julier C, et al. Genome-wide scan for linkage to type 1 diabetes in 2,496 multiplex families from the Type 1 Diabetes Genetics Consortium. *Diabetes*. 2009;58:1018-1022.
9. Burren OS, Adlem EC, Achuthan P, Christensen M, Coulson RMR, Todd JA. T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research. *Nucleic Acids Res*. 2011;39:D997-D1001.
10. Winkler C, Krumsiek J, Buettner F, et al. Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes. *Diabetologia*. 2014;57:2521-2529.
11. Thümer L, Adler K, Bonifacio E, et al. German new onset diabetes in the young incident cohort study: DiMelli study design and first-year results. *Rev Diabet Stud*. 2010;7:202-208.
12. Ziegler AG, Hummel M, Schenker M, Bonifacio E. Autoantibody appearance and risk for development of childhood diabetes in offspring of parents with type 1 diabetes: the 2-year analysis of the German BABYDIAB Study. *Diabetes*. 1999;48:460-468.
13. Rewers M, Norris JM, Eisenbarth GS, et al. Beta-cell autoantibodies in infants and toddlers without IDDM relatives: Diabetes Autoimmunity Study in the Young (DAISY). *J Autoimmun*. 1996;9:405-410.
14. Rewers M, Bugawan TL, Norris JM, et al. Newborn screening for HLA markers associated with IDDM: Diabetes Autoimmunity Study in the Young (DAISY). *Diabetologia*. 1996;39:807-812.
15. Yu L, Rewers M, Gianani R, et al. Antiislet autoantibodies usually develop sequentially rather than simultaneously. *J Clin Endocrinol Metab*. 1996;81:4264-4267.
16. Bonifacio E, Yu L, Williams AK, et al. Harmonization of glutamic acid decarboxylase and islet antigen-2 autoantibody assays for national institute of diabetes and digestive and kidney diseases consortia. *J Clin Endocrinol Metab*. 2010;95:3360-3367.
17. Wenzlau JM, Juhl K, Yu L, et al. The cation efflux transporter ZnT8 (Slc30A8) is a major autoantigen in human type 1 diabetes. *Proc Natl Acad Sci U S A*. 2007;104:17040-17045.
18. Wenzlau JM, Moua O, Sarkar SA, et al. SlC30A8 is a major target of humoral autoimmunity in type 1 diabetes and a predictive marker in prediabetes. *Ann N Y Acad Sci*. 2008;1150:256-259.
19. Steck AK, Dong F, Wong R, et al. Improving prediction of type 1 diabetes by testing non-HLA genetic variants in addition to HLA markers. *Pediatr Diabetes*. 2014;15:355-362.
20. Ziegler A-G, Bonifacio E, BABYDIAB-BABYDIET Study Group. Age-related islet autoantibody incidence in offspring of patients with type 1 diabetes. *Diabetologia*. 2012;55:1937-1943.
21. SNAP Pairwise LD [Internet]. https://www.broadinstitute.org/mpg/snap/ldsearchpw.php. Accessed October 30, 2015.
22. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;30:157-172.
23. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
24. Winkler C, Krumsiek J, Lempainen J, et al. A strategy for combining minor genetic susceptibility genes to improve prediction of disease in type 1 diabetes. *Genes Immun*. 2012;13:549-555.
25. Achenbach P, Hummel M, Thümer L, Boerschmann H, Höfelmann D, Ziegler AG. Characteristics of rapid vs slow progression to type 1 diabetes in multiple islet autoantibody-positive children. *Diabetologia*. 2013;56:1615-1622.
26. Lempainen J, Härkönen T, Laine A, Knip M, Ilonen J, Register Finnish Pediatric Diabetes. Associations of polymorphisms in non-HLA loci with autoantibodies at the diagnosis of type 1 diabetes: INS and IKZF4 associate with insulin autoantibodies. *Pediatr Diabetes*. 2013;14:490-496.
27. Lempainen J, Laine A-P, Hammais A, et al. Non-HLA gene effects on the disease process of type 1 diabetes: from HLA susceptibility to overt disease. *J Autoimmun*. 2015;61:45-53.
28. Lempainen J, Hermann R, Veijola R, Simell O, Knip M, Ilonen J. Effect of the PTPN22 and INS risk genotypes on the progression to clinical type 1 diabetes after the initiation of β-cell autoimmunity. *Diabetes*. 2012;61:963-966.
29. Meehan C, Fout B, Ashcraft J, Schatz DA, Haller MJ. Screening for T1D risk to reduce DKA is not economically viable. *Pediatr Diabetes*. 2015;16:565-572.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Frohnert BI, Laimighofer M, Krumsiek J, Theis FJ, Winkler C, Norris JM, Ziegler A-G, Rewers MJ, Steck AK. Prediction of type 1 diabetes using a genetic risk model in the Diabetes Autoimmunity Study in the Young. *Pediatr Diabetes*. 2017;0:1–7. https://doi.org/10.1111/pedi.12543

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Jul 14, 2017

This Agreement between Michael Laimighofer ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4147651433114
License date	Jul 14, 2017
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Pediatric Diabetes
Licensed Content Title	Prediction of type 1 diabetes using a genetic risk model in the Diabetes Autoimmunity Study in the Young
Licensed Content Author	Brigitte I Frohnert,Michael Laimighofer,Jan Krumsiek,Fabian J Theis,Christiane Winkler,Jill M Norris,Anette-Gabriele Ziegler,Marian J Rewers,Andrea K Steck
Licensed Content Date	Jul 11, 2017
Licensed Content Pages	1
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	Statistical learning models for prediction of Type 1 Diabetes risk factors using clinical data and omics data
Expected completion date	Apr 2017
Expected size (number of pages)	112
Requestor Location	Michael Laimighofer Ingolstaedter Landstr 1 Neuherberg, 85764 Germany Attn: Michael Laimighofer
Publisher Tax ID	EU826007151
Billing Type	Invoice
Billing Address	Michael Laimighofer Ingolstaedter Landstr 1 Neuherberg, Germany 85764 Attn: Michael Laimighofer
Total	0.00 EUR
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing

transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED

WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library

http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Common patterns of gene regulation associated with Cesarean section and the development of islet autoimmunity - indications of immune cell activation.

Common patterns of gene regulation associated with Cesarean section and the development of islet autoimmunity – indications of immune cell activation

M. Laimighofer [1,2], R. Lickert [3], R. Fürst [3], F. J. Theis [1,2], C. Winkler [3], E. Bonifacio [4,5], A.-G. Ziegler [3], and J. Krumsiek [1,6,#]

[1] Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

[2] Department of Mathematics, Technische Universität München, Garching, Germany

[3] Institute of Diabetes Research, Helmholtz Zentrum München, and Forschergruppe Diabetes, Klinikum rechts der Isar, Technische Universität München, Germany

[4] DFG Center for Regenerative Therapies Dresden, Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

[5] Paul Langerhans Institute Dresden, German Center for Diabetes Research (DZD), Technische Universität Dresden, Dresden, Germany

[6] German Center for Diabetes Research (DZD), Neuherberg, Germany

[#] corresponding author, jan.krumsiek@helmholtz-muenchen.de

Abstract

Background: Birth by Cesarean section increases the risk of developing type 1 diabetes later in life; however, the underlying molecular mechanisms of this effect remain unclear. We aimed to elucidate common regulatory processes observed after Cesarean section and the development of islet autoimmunity, which precedes type 1 diabetes, by investigating the transcriptome of blood cells in the developing immune system.

Methods: We analyzed gene expression of peripheral blood mononuclear cells taken at several time points from children with increased familial and genetic risk for type 1 diabetes (n = 109). We investigated effects of Cesarean section on gene expression profiles of children in the first year of life using a generalized additive mixed model to account for the longitudinal data structure. To investigate the effect of islet autoimmunity, we compared gene expression differences between children after initiation of islet autoimmunity and age-matched children who did not develop islet autoantibodies. Finally, we compared both results to identify common regulatory patterns of Cesarean section and islet autoimmunity at the gene expression level.

Results: We identified two differentially expressed pathways in children born by Cesarean section: the pentose phosphate pathway and pyrimidine metabolism, both involved in nucleotide synthesis and cell proliferation. Islet autoantibody analysis revealed multiple differentially expressed pathways generally involved in immune processes, including both of the above-mentioned nucleotide synthesis pathways. Comparison of global gene expression signatures showed that transcriptomic changes were systematically and significantly correlated between Cesarean section and islet autoimmunity. In addition, signatures of both Cesarean section and islet autoimmunity correlated with transcriptional changes observed during activation of isolated CD4+ T lymphocytes.

Conclusions: We identified coherent gene expression signatures for Cesarean section, an early risk factor for type 1 diabetes, and islet autoantibodies positivity, an obligatory stage of autoimmune response prior to the development of type 1 diabetes. Both transcriptional signatures were correlated with changes in gene expression during the activation of CD4+ T lymphocytes, reflecting common molecular changes in immune cell activation.

Background

Type 1 diabetes is an autoimmune disease in which immune cells destroy insulin-producing beta cells in the pancreas. Loss of beta-cell mass leads to insulin deficiency, impaired glucose tolerance, and ultimately the onset of type 1 diabetes [1]. This process is marked by the appearance of islet autoantibodies to beta-cell-related antigens [2]. Autoantibody seroconversion usually occurs during childhood and early adolescence, with a peak conversion rate between 9 months and two years [3]. Notably, an increasing incidence of type 1 diabetes has been observed within the last few decades, especially in children [4]. Several previous studies indicated transcriptional changes in immune cells caused by the development of islet autoantibodies [5, 6], indicating substantial molecular changes long before the onset of type 1 diabetes.

Children born by Cesarean section have an odds ratio of 1.23 for the development of type 1 diabetes compared to vaginally delivered children [7]. Moreover, Cesarean section has been reported to be associated with a faster progression to type 1 diabetes, but not with an increase the risk for islet autoimmunity [8]. The underlying mechanisms of these associations are not yet fully understood. Some reports indicate that the mode of delivery affects colonization of microbiota in the intestinal tract [9], which in turn affects the developing immune system of infants [10]. Such differences in the gut microbiome and its interaction with the immune system may lead to increased risk of asthma, childhood allergies [11], and autoimmune diseases, such as type 1 diabetes.

Here we investigated the impact of Cesarean section and islet autoimmunity on the immune system by analyzing gene expression data in peripheral blood mononuclear cells (PBMCs) as a readout (Figure 1A). The overall goal of the analysis was to elucidate common immune modulation patterns between an early risk factor, Cesarean section, and the development of autoimmunity. We first investigated the effects of Cesarean section on gene expression profiles of children in the first year of life (Figure 1B). To this end, we analyzed data from several time points in this early period in children with an increased familial risk for type 1 diabetes. A generalized additive mixed model was used to account for the longitudinal information when extracting gene expression differences between children born by Cesarean section and vaginal delivery (Figure 1C). In a second analysis, we compared gene expression levels of children with increased risk of familial type 1 diabetes after initiation of islet autoimmunity with age-matched children who did not develop islet autoantibodies. We then combined both results to identify genes and pathways that were differentially regulated in both analyses, to link Cesarean section and islet autoimmunity at the level of gene expression. Finally, we compared the patterns identified for

Cesarean section and islet autoantibodies with gene expression data from activated human CD4+ T lymphocytes to identify candidates for common molecular mechanisms of immune activation.

Results

Effects of Cesarean section on the transcriptome in the first year of life

To focus on early postpartum effects on genetic regulation after delivery by Cesarean section, we restricted our analysis to samples from children in their first year of life (Supplemental file 1: Figure S1). This resulted in a total of 154 PBMC gene expression samples (71 Cesarean section, 83 vaginal delivery) from 87 children (39 Cesarean section, 48 vaginal delivery), Figure 1A+B. The number of samples per child ranged between 1 and 4 (Supplemental file 1: Figure S1). Details on the dataset and preprocessing steps are provided in the Methods section.

We used a generalized additive mixed effect model (GAMM) to account for the longitudinal nature of the dataset. The model consisted of three major parts (Figure 1C): (1) the change of expression profiles over time was modeled using a spline-based function, (2) the overall trend of expression values per child was captured by a random mixed effect, and (3) a term representing the Cesarean section effect extracted the association in which we were primarily interested. Applying this model to the data set, we observed an enrichment of low p-values before adjusting for multiple testing (Figure 2A and Supplemental file 2). However, after multiple testing correction by controlling the false discovery rate (FDR) at 0.1, no significant single genes could be identified in this first step (Figure 2B).

To improve statistical power and facilitate biological interpretability, we performed a pathway enrichment analysis based on the KEGG pathway database [12]. From this analysis, we obtained two significantly differentially expressed pathways after multiple testing correction (FDR <= 0.1): “Pentose phosphate pathway” and “Pyrimidine metabolism” (Figure 2C and Supplemental file 2). The “Pentose phosphate pathway” included 27 genes (20 up- and 7 down-regulated genes in children born by Cesarean section) and the “Pyrimidine metabolism” pathway contained 92 genes (57 up-regulated and 35 down-regulated).

Both pathways are well known to be involved in the synthesis of nucleotides during cell proliferation [13, 14]. This indicated that there is a change in expression profiles of proliferation-related nucleotide synthesis pathways as an early consequence of delivery by Cesarean section.

Effects of islet autoimmunity on the transcriptome

To identify gene expression signatures associated with seroconversion, we compared the earliest sample of children after the development of islet autoantibodies (up to 6 months post-seroconversion; 15 children) with all available age-matched samples of children who did not develop islet autoantibodies (74 children). We applied linear regression to explain gene expression differences induced by islet autoantibody status, corrected for age (see Methods). We detected a strong accumulation of differentially expressed genes (Figure 3A), of which 3,867 were significant after multiple testing correction (FDR ≤ 0.1, Figure 3B). The majority of these genes were found to be up-regulated in children with islet autoimmunity (64% up-regulated vs. 36% down-regulated, Supplemental file 2).

Pathway enrichment analysis on KEGG pathways identified 20 as significantly regulated (FDR <= 0.1, Figure 3C and Supplemental file 2). The top-ranked pathways were “p53 signaling pathway”, “Ubiquitin mediated proteolysis”, and "RIG-I-like receptor signaling pathway.” The two significant pathways from the Cesarean section analysis, “Pyrimidine metabolism” and “Pentose phosphate pathway”, also appeared as significant in the islet autoantibody analysis. Notably, comparing gene expression samples of children before seroconversion (up to 6 months before) and age-matched children without islet autoantibodies yielded no significant genes or pathways (Supplemental file 3). In addition, we investigated children that had gene expression samples both before and after seroconversion (up to 6 month before and after) in a paired analysis, leaving only seven children. No genes or pathways were found to be differentially expressed (Supplemental file 3).

Taken together, we observed several immune system-related and nucleotide synthesis pathways associated with islet autoimmunity, which included the pathways found in our Cesarean section analysis.

Coherent gene expression changes between Cesarean section and islet autoimmunity

We further investigated the similarities of transcriptional changes between the two risk factors. First, we found that the individual gene regulation of both “Pyrimidine metabolism” and the “Pentose phosphate pathway” pathway showed similar patterns of up- and down-regulation between Cesarean section and islet autoimmunity (Figure 4A, Supplemental file 4).

Extending this analysis, we quantified the relationship of gene regulation between the two factors at a systematic level. We correlated the standardized effects from Cesarean section and development of islet autoantibodies across all genes (Figure 4B), revealing a striking correlation of 0.606 ($p = 0.0062$, Figure 4C). In other words, genes regulated by Cesarean section, despite the rather weak signal strength in our study, were also regulated in the same direction by the initiation of islet autoimmunity. A similar correlation between the effects of Cesarean section and islet autoimmunity was observed at the pathway level, with a correlation of 0.49 ($p = 0.02$, Figure 4D+E), supporting the functional agreement of transcriptional changes between the two risk factors. Importantly, we did not observe that children born by Cesarean section developed islet autoantibodies more frequently than children born by vaginal delivery, ruling out a confounding effect not related to gene expression (Supplemental file 1: Table S1).

In contrast to the profound Cesarean section to islet autoantibodies correlation at transcript level, we found the effects of gender, maternal diabetes and multiple first-degree relatives to be randomly correlated with Cesarean section at the single gene level (maternal diabetes, $r = -0.22$, $p = 0.53$; gender, $r = 0.09$, $p = 0.78$; multiple first-degree relatives, $r = -0.24$, $p = 0.49$) and at the pathway level (maternal diabetes, $r = 0.01$, $p = 0.97$; gender, $r = -0.01$, $p = 0.97$; multiple first-degree relatives, $r = -0.23$, $p = 0.34$), as shown in Figure 4 C+E and Supplemental file 5. This demonstrates the specificity of the Cesarean section–islet autoantibody effect correlation.

In summary, this analysis showed that the gene expression changes in these two risk factors, Cesarean section and islet autoimmunity, are remarkably coherent.

Signatures of immune cell activation

The pentose phosphate pathway is a universal, central metabolic pathway in the cytosol, which supports cell proliferation and survival [15]. The non-oxidative branch of the pentose phosphate pathway branches off glycolysis and generates ribose 5-phosphate as a precursor for the synthesis of nucleotides and amino acids necessary for cell growth and division. Moreover, the pyrimidine metabolism pathway is directly related to the pentose phosphate pathway in its role in nucleotide synthesis. Since we investigated PMBCs, these differentially regulated pathways in Cesarean section and islet autoimmunity point toward a general activation of immune cells. A direct proof of this hypothesis in the same children was unfeasible in the context of this study. Instead, we collected evidence from several secondary analysis steps.

First, we investigated whether regulated genes from both analyses enriched immune genes annotated in innateDB [16]. Indeed, there was a significant accumulation of higher standardized effects for immune genes compared to non-immune genes, for both Cesarean section (Wilcoxon rank sum test: $p = 0.0002$) and autoimmunity ($p = 2.6 \times 10^{-15}$); see Supplemental file 6.

Second, we compared results from the mixture of PBMC cells with published data on activation in isolated immune cells. For the analysis, we used transcriptomics data from isolated naïve and activated human CD4+ T cells [17]. We calculated gene expression differences before and after activation, and applied enrichment analysis to identify differentially expressed pathways. In particular, the "Pentose phosphate pathway" ($p = 0.049$) and "Pyrimidine metabolism" ($p = 0.035$) pathways were significantly differentially expressed in activated CD4+ T cells (Supplemental file 2). To compare the effects of CD4+ T cell activation with effects from the Cesarean section and islet autoantibody analyses, we calculated correlations of the standardized effects. We observed borderline significant correlations between changes in activated CD4+ T cell and Cesarean section ($r = 0.20$, $p = 0.049$, Figure 5 A+B) and islet autoimmunity ($r = 0.24$, $p = 0.052$, Figure 5 C+D). Remarkably, the association was substantially stronger at pathway level for both Cesarean section ($r = 0.45$, $p = 0.021$, Figure 5 E+F) and islet autoimmunity ($r = 0.57$, $p = 0.008$, Figure 5 G+H). The pathway association was replicated in a second transcriptomics dataset from naïve and activated CD4+ T cells, monocytes, and natural killer cells, published in [18]. Detailed results are shown in Supplemental file 7.

In summary, we observed a significant correlation between the functional effects of human lymphocyte activation and the effects of Cesarean section and islet autoimmunity on gene expression in PBMCs.

Discussion

We identified coherent gene expression signatures for Cesarean section, an early risk factor for type 1 diabetes, and islet autoantibody positivity, an obligatory stage of autoimmune response prior to the development of autoimmune type 1 diabetes. Specifically, at the transcriptome level, we identified two pathways involved in nucleotide synthesis and cell proliferation that were regulated in PBMCs of children born by Cesarean section. This analysis required an extended statistical model to incorporate the complex time information in our present dataset. Islet autoantibody analysis revealed various pathways generally involved in immune processes, including the aforementioned nucleotide synthesis pathways. Comparing global gene expression signatures, we found that transcriptomic changes were

systematically and significantly correlated between the Cesarean section and islet autoantibody-positive analyses. Importantly, transcriptional signatures of Cesarean section did not correlate with gender, maternal diabetes, or multiple first-degree relatives, demonstrating that the correlation is specific to Cesarean section and islet autoimmunity. In a functional follow-up analysis, both Cesarean section and islet autoimmunity signatures were significantly correlated with gene expression changes observed during activation of isolated CD4+ T lymphocytes. At pathway level, this correlation was also observed for monocytes and natural killer cells in a second dataset.

We can speculate on the biological basis of our statistical observations. The coherent regulation of proliferation pathways in blood PBMCs may indicate a general activation of the immune system and immune cells for both Cesarean section and seroconversion. In the case of Cesarean section, this activation might indirectly reflect different microbial exposures during birth. This idea is indirectly supported by findings that the microbiome is affected by the mode of delivery, and in turn affects the developing immune system [9, 10]. For the autoimmunity results, the precise interplay and timing of the occurrence of environmental stimuli, immune response and the development of islet autoantibodies still need to be elucidated. Regarding a hypothetical disease trajectory leading to type 1 diabetes, it is conceivable that the observed gene expression signatures may reflect a transient deflection of the immune system and that this primes a child for subsequent progression to type 1 diabetes in the presence of other risk factors.

An interesting general observation in our analysis was the increased correlation of gene expression signatures when performing pathway analysis instead of single gene analysis. This pattern was observed both for the comparison of Cesarean section and islet autoantibody-positive signatures, and for comparisons of these two signatures with the isolated immune cells. These findings indicate that pathway analysis substantially reduced the noise compared to single gene analysis, which allowed us to identify common patterns at a functional level.

Our study could be extended and improved in several directions. (1) The present dataset has rather low statistical power for the islet autoantibody analysis, with only 15 positive samples available for at most 6 months post-seroconversion. While differential expression changes were remarkably significant, the analysis should be validated with a larger sample size. (2) The hypothesis of coherent immune system activation should ideally be confirmed in an isolated primary T cell population from children after Cesarean section or after seroconversion. We took an indirect route using PBMCs rather than a more general activation experiment with isolated immune cells. (3) The incorporation of further

environmental factors, such as nutrition and medical parameters, in combination with the children's genetic background, is expected to give a more complete picture of the parameters leading to autoimmunity and type 1 diabetes.

Conclusions

In summary, we found a transcriptional link between Cesarean section and islet autoimmunity, pointing toward a transiently altered immune system in the susceptible period of islet autoimmunity generated by Cesarean section, which was remarkably coherent with the changes observed after islet autoimmunity.

Methods

PBMC gene expression data

We used the BABYDIET PBMC gene expression data deposited in ArrayExpress (accession number: E-MTAB-1724) [6], in combination with non-public data on Cesarean section, islet autoimmunity, family history, gender and age. In this data set, 454 samples from 109 children were available. One individual of the original 109 children was removed in the Cesarean section analysis, since no information about the type of delivery was recorded. Raw gene expression data of 33,297 probes were normalized using the Robust Multi-Array Average (RMA) method [19]. Probes without annotation in the Affymetrix hugene11 data and duplicates were removed using the R package *genefilter*, leaving 18,720 genes for further analysis. Age at sampling ranged from 0.21 years to 9.15 years, with a median of 1.53 years.

Generalized additive mixed model for time-resolved data

Transcriptomics samples were available for multiple time points per child. To model gene expression from multiple, non-matching timepoints in relation to Cesarean section in a joint approach, we employed a generalized additive mixed model (GAMM) [20]. In this GAMM, the model structure is defined as

$$y_{i,j} = \beta_{0,i} + \beta_{CS,i}\, \boldsymbol{x_{CS,j}} + \sum_k b_{k,i}(\boldsymbol{t_{i,j}})\beta_{k,i} + b_{\mathrm{ID},i}\, \boldsymbol{x_{ID,j}} + \epsilon_{i,j}$$

where $y_{i,j}$ is the gene expression of sample *j* for gene *i*, $\beta_{0,i}$ is the intercept or gene-wise average expression, $\beta_{CS,i}$ describes the effect of Cesarean section on gene expression with $\boldsymbol{x_{CS,j}} = 1$ for Cesarean section and $\boldsymbol{x_{CS,j}} = 0$ for vaginal delivery. Adjusting for the multiple measurements in time $\boldsymbol{t_{i,j}}$, a spline function was included as $\sum_k b_{k,i}(\boldsymbol{t_{i,j}})\beta_{k,i}$ per gene over all samples, where k is the estimated number of basis functions, $b_{k,i}$ are the basis functions of the spline and $\beta_{k,i}$ their regression coefficients. In addition, we included a random effect as $b_{\mathrm{ID},i} \sim N(0, \sigma_i)$, with σ_i estimated per gene, correcting for the dependence of several measurements per child, denoted as $\boldsymbol{x_{ID,j}}$, and an error term $\epsilon_{i,j}$ as i.i.d. normally distributed noise. The noise $\epsilon_{i,j}$ and the noise of the subject specific random effect σ_i were assumed to be independent. Inference of the GAMM was performed using a restricted maximum likelihood approach [21] using the R package *mgcv*.

Linear model for seroconversion analysis

To investigate differences in gene expression between children after development of islet autoantibodies and age-matched children who did not develop islet autoantibodies, we selected the first sample up to six months after development of islet autoimmunity for each child, when such a sample was available (see Figure 1B). These children were compared to all available age-matched children who did not develop autoantibodies, by applying a linear regression model per gene:

$$y_{i,j} = \beta_{0,i} + \beta_{AB,i}\, \boldsymbol{x_{AB,j}} + \beta_{age,i}\, \boldsymbol{x_{age,j}} + \epsilon_{i,j},$$

where $y_{i,j}$ is the gene expression of child *j* for gene *i*, $\beta_{0,i}$ the intercept, $\beta_{AB,i}$ describes the effect on gene expression of islet autoimmunity, with $\boldsymbol{x_{AB,j}}$ = 1 for islet autoantibody-positives and $\boldsymbol{x_{AB,j}}$ = 0 for islet autoantibody-negatives, and $\beta_{age,i}$ the effect of age $\boldsymbol{x_{age,j}}$ on gene expression. The residual term $\epsilon_{i,j}$ was defined as i.i.d. normally distributed noise.

Pathway enrichment analysis

We performed pathway enrichment analysis based on the KEGG database [12]. We identified differentially regulated pathways using the "Significance Analysis of Function and Expression" (SAFE) algorithm [22], R package *safe*. In this approach, p-values from local gene tests are computed and Wilcoxon rank sum statistics assess whether local statistics are systematically increased in the pathway, compared to the background (global test). Local gene tests were calculated on the residuals of models from equations (1) and (2), but excluding the term for the factor of interest (Cesarean section or islet autoimmunity). SAFE uses a sample permutation-based approach for p-value calculation, which avoids false positive results due to correlating transcripts. The number of permutations was set to 5,000.

For node coloring of pathways (Figure 4A+B), we used a "directed p-value", defined as (-log10(p-value) * sign(regression coefficient)). For nodes in the pathway that had multiple genes annotated, the highest effect was shown.

Comparing Cesarean section and islet autoimmunity results

To compare the results obtained from the Cesarean section and islet autoimmunity (Ab+) analyses, we calculated the Pearson correlation between the standardized effect estimates of both analyses. Standardized effect estimates were represented by the t-statistic in the single gene analysis, and a "directed p-value" (analogously to previous section) in the pathway analysis. To assess the statistical significance of the correlation, we calculated

$$p_{perm} = \frac{1}{B}\sum_{i=1}^{B} I(abs\ (\rho_{i\ (CS_{Ab+})}) > \rho_{obs\ (CS_{Ab+})})$$

with B being the number of permutations, $\rho_{obs\ (CS_{Ab+})}$ the observed "true" correlation of effects between both analyses, *I()* the indicator function and $\rho_{i\ (CS_{Ab+})}$ the correlation between the effects of Cesarean section and permuted islet autoimmunity status. The number of permutations was fixed at B = 5,000 for the single gene analysis and 1,000 for the pathway analyses. The same analysis with 1,000 permutations was repeated for the factors maternal diabetes, gender, and multiple first-degree relatives.

Immune cell activation analysis

To investigate the enrichment of genes within annotated immune genes from innateDB [16], we calculated Wilcoxon rank sum tests using the t-statistics from the single gene level analysis (see above) to identify differences between the distribution of immune genes and non-immune genes. For comparison of standardized effects, we used two published datasets, one from isolated CD4+ T cells before and after activation (GEOD-33272, processed data downloaded) and one containing different subsets of human leukocytes before and after activation (GEOD-22886, processed data downloaded). Single gene differential analysis was performed on log2-transformed expression values using linear regression, with gene expression as the response, and activation status as the explaining variables. For SAFE pathway enrichment, the number of permutations was set to 1,000. To calculate permutation-based p-values of Pearson correlations, we used 1,000 permutations.

All statistical analyses were performed using the computing environment R version 3.3.2 [23].

List of abbreviations

PBMC: peripheral blood mononuclear cells,

GAMM: generalized linear mixed model,

Funding

This work was supported by grants from Deutsche Forschungsgemeinschaft ZI-310/14-1, ZI-310/14-2, ZI-310/14-3, and ZI-310/14-4 (BABYDIET), by a grant from the European Union's Seventh Framework Programme [FP7-Health-F5-2012] under grant agreement 305280 (MIMOmics), and by the Helmholtz Cross-Program Initiative Personalized Medicine 'iMED'.

Acknowledgments

We thank Dr. Nikola Müller and Dr. Steffen Sass for sharing their expertise on transcriptomics data.

Declarations

Ethics approval and consent to participate

The BABYDIET study was approved by the ethics committee of Ludwig-Maximilians University (Protocol No. 329/00) and is registered at ClinicalTrials.gov NCT01115621. The parents or guardians of each child provided informed consent for participation in the BABYDIET study. All samples and information were collected after obtaining signed informed consent.

Consent for publication

Not applicable.

Availability of data and materials

This publication uses publicly available gene expression datasets, available via ArrayExpress (accession number E-MTAB-1724) and the Gene Expression Omnibus (accession numbers GSE33272 and GSE22886). The informed consent given by study participants does not cover data posting of the clinical covariates in public databases.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Study concept and design: ML, AGZ, JK. Statistical analysis: ML. Acquisition of data: RP, RF, CW, AGZ. Analysis and interpretation of data: ML, EB, JK. Drafting of the manuscript: ML, EB, AGZ, JK. Critical revision of the manuscript for important intellectual content: ML, RP, RF, FJT, CW, EB, AGZ, JK. All authors read and approved the final manuscript.

References:

1. Donath MY, Halban PA: **Decreased beta-cell mass in diabetes: significance, mechanisms and therapeutic implications**. *Diabetologia* 2004, **47**(3):581-589.
2. Krischer JP, Lynch KF, Schatz DA, Ilonen J, Lernmark A, Hagopian WA, Rewers MJ, She JX, Simell OG, Toppari J *et al*: **The 6 year incidence of diabetes-associated autoantibodies in genetically at-risk children: the TEDDY study**. *Diabetologia* 2015, **58**(5):980-987.
3. Giannopoulou EZ, Winkler C, Chmiel R, Matzke C, Scholz M, Beyerlein A, Achenbach P, Bonifacio E, Ziegler AG: **Islet autoantibody phenotypes and incidence in children at increased risk for type 1 diabetes**. *Diabetologia* 2015, **58**(10):2317-2323.
4. Lipman TH, Levitt Katz LE, Ratcliffe SJ, Murphy KM, Aguilar A, Rezvani I, Howe CJ, Fadia S, Suarez E: **Increasing incidence of type 1 diabetes in youth: twenty years of the Philadelphia Pediatric Diabetes Registry**. *Diabetes care* 2013, **36**(6):1597-1603.
5. Kallionpaa H, Elo LL, Laajala E, Mykkanen J, Ricano-Ponce I, Vaarma M, Laajala TD, Hyoty H, Ilonen J, Veijola R *et al*: **Innate immune activity is detected prior to seroconversion in children with HLA-conferred type 1 diabetes susceptibility**. *Diabetes* 2014, **63**(7):2402-2414.
6. Ferreira RC, Guo H, Coulson RM, Smyth DJ, Pekalski ML, Burren OS, Cutler AJ, Doecke JD, Flint S, McKinney EF *et al*: **A type I interferon transcriptional signature precedes autoimmunity in children genetically at risk for type 1 diabetes**. *Diabetes* 2014, **63**(7):2538-2550.
7. Cardwell CR, Stene LC, Joner G, Cinek O, Svensson J, Goldacre MJ, Parslow RC, Pozzilli P, Brigis G, Stoyanov D *et al*: **Caesarean section is associated with an increased risk of childhood-onset type 1 diabetes mellitus: a meta-analysis of observational studies**. *Diabetologia* 2008, **51**(5):726-735.
8. Bonifacio E, Warncke K, Winkler C, Wallner M, Ziegler AG: **Cesarean section and interferon-induced helicase gene polymorphisms combine to increase childhood type 1 diabetes risk**. *Diabetes* 2011, **60**(12):3300-3306.
9. Biasucci G, Benenati B, Morelli L, Bessi E, Boehm G: **Cesarean delivery may affect the early biodiversity of intestinal bacteria**. *The Journal of nutrition* 2008, **138**(9):1796S-1800S.
10. Caicedo RA, Schanler RJ, Li N, Neu J: **The developing intestinal ecosystem: implications for the neonate**. *Pediatric research* 2005, **58**(4):625-628.
11. Neu J, Rushing J: **Cesarean versus vaginal delivery: long-term infant outcomes and the hygiene hypothesis**. *Clinics in perinatology* 2011, **38**(2):321-331.
12. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic acids research* 2000, **28**(1):27-30.

13. Jiang P, Du W, Wu M: **Regulation of the pentose phosphate pathway in cancer**. *Protein & cell* 2014, **5**(8):592-602.
14. Lane AN, Fan TW: **Regulation of mammalian nucleotide metabolism and biosynthesis**. *Nucleic acids research* 2015, **43**(4):2466-2485.
15. O'Neill LA, Kishton RJ, Rathmell J: **A guide to immunometabolism for immunologists**. *Nature reviews Immunology* 2016, **16**(9):553-565.
16. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock RE, Brinkman FS, Lynn DJ: **InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation**. *Nucleic acids research* 2013, **41**(Database issue):D1228-1233.
17. Maliga Z, Junqueira M, Toyoda Y, Ettinger A, Mora-Bermudez F, Klemm RW, Vasilj A, Guhr E, Ibarlucea-Benitez I, Poser I *et al*: **A genomic toolkit to investigate kinesin and myosin motor function in cells**. *Nature cell biology* 2013, **15**(3):325-334.
18. Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, Fong S, van Lookeren Campagne M, Godowski P, Williams PM *et al*: **Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data**. *Genes and immunity* 2005, **6**(4):319-331.
19. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.
20. Wood SN: **Generalized additive models: an introduction with R**: CRC press; 2017.
21. Wood SN: **Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011, **73**(1):3-36.
22. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach**. *Bioinformatics* 2005, **21**(9):1943-1949.
23. Team RC: **R: a language and environment for statistical computing. R Development Core Team, Vienna**. In.; 2016.

Figures

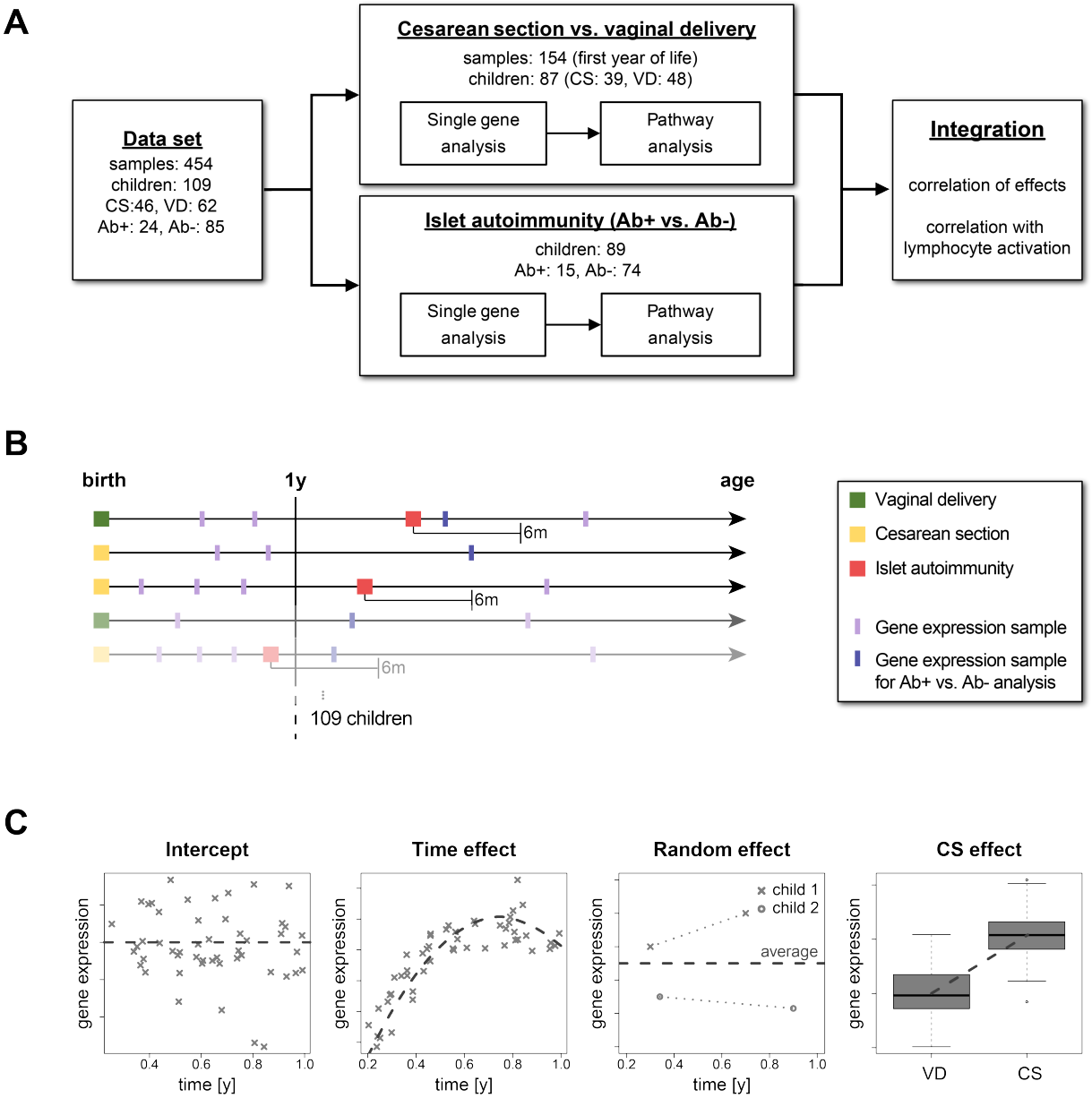


Figure 1: Overview. A: Study workflow: Parallel analyses were performed to detect differential gene expression and pathway enrichment for Cesarean section and islet autoimmunity. The results were then compared in a combined analysis and related to expression patterns of lymphocyte activation. **B:** Schematic overview of the longitudinal study design for Cesarean section and islet autoimmunity analyses. **C**: Schematic illustration of the generalized additive mixed effect model (GAMM) to analyze the longitudinal dataset, including intercept, a time effect, a random effect for multiple measurements, and the investigated Cesarean section vs. vaginal delivery effect. Abbreviations: CS = Cesarean section, VD = Vaginal delivery, Ab+ = Islet autoantibody-positive, Ab- = Islet autoantibody-negative.

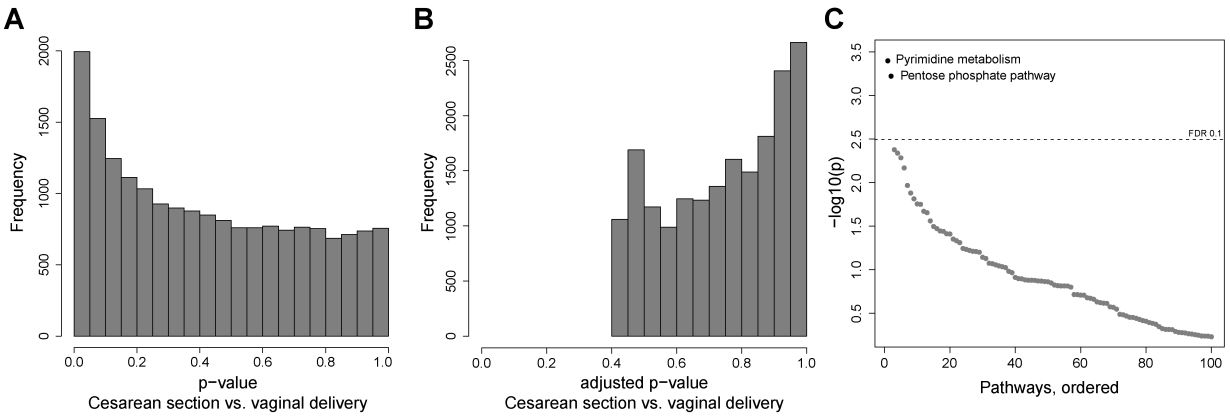


Figure 2: Cesarean section analysis. A: Histogram of unadjusted p-values for CS association per gene. **B**: Histogram of p-values after multiple testing adjustment by controlling the false discovery rate. **C:** Sorted -log10(p-values) of pathway enrichment. Dashed line indicates the multiple testing threshold at an FDR of 0.1. Abbreviations: CS = Cesarean section, VD = Vaginal delivery.

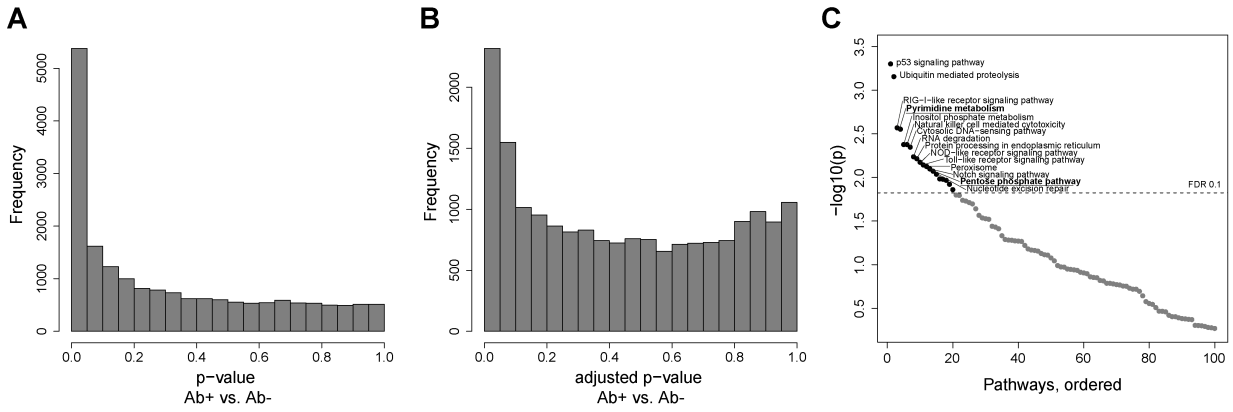


Figure 3: Ab+ analysis. A: Histogram of unadjusted p-values from single gene analysis of islet autoimmunity positive vs. age-matched children who did not develop islet autoimmunity. **B:** Histogram of p-values after multiple testing, controlling the false discovery rate. **C:** Sorted -log10(p-values) of pathway enrichment for islet autoimmunity status. Dashed line indicates the multiple testing threshold at an FDR of 0.1. Abbreviations: Ab+ = Islet autoantibody-positive, Ab- = Islet autoantibody negative.

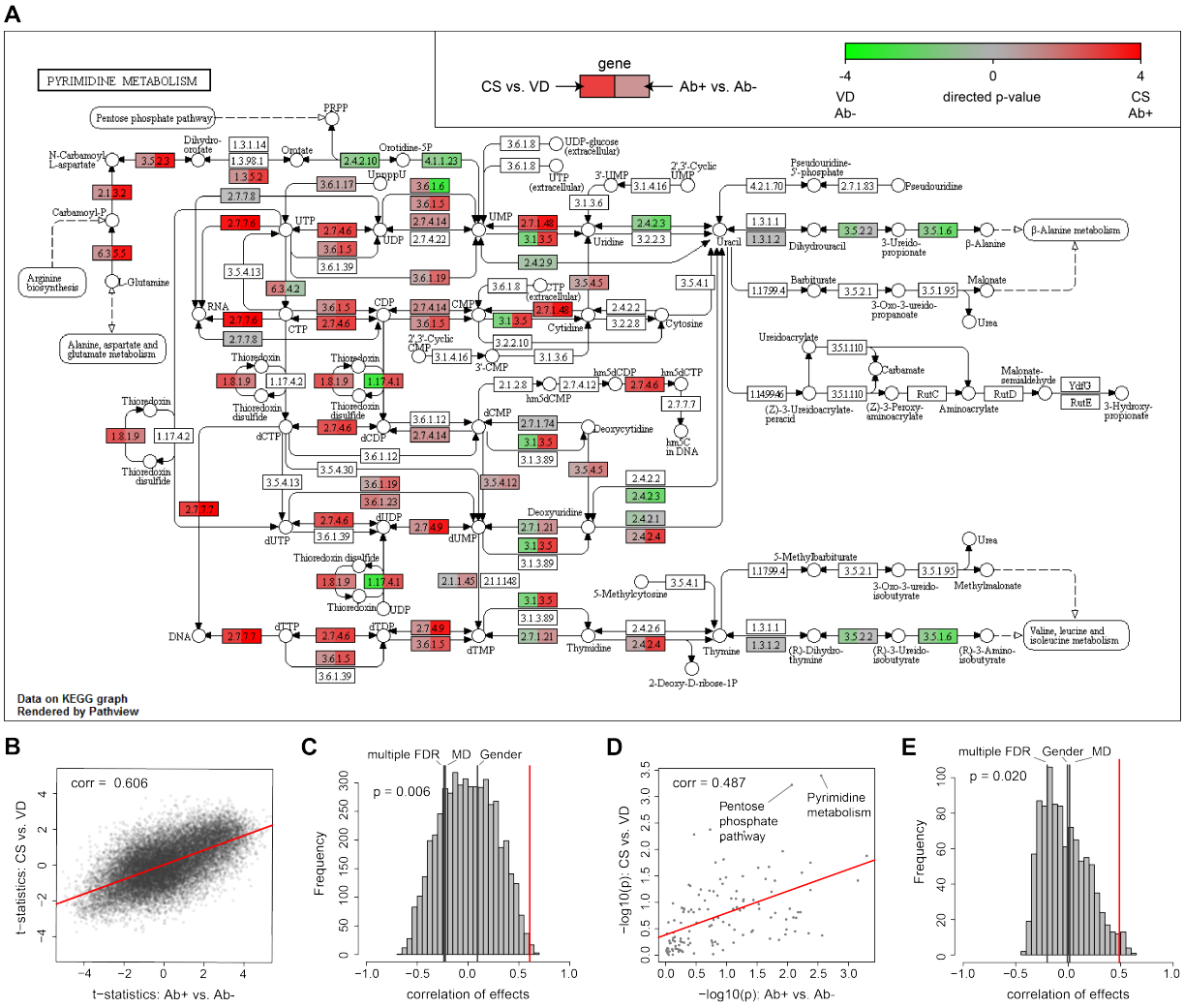


Figure 4: Comparison of Cesarean section and islet autoimmunity signatures. A: The pyrimidine metabolism pathway is shown as an example, with two-sided node coloring according to directed p-values (-log10(p) * sign(t-statistic)). **B:** T-statistics per gene from both analyses; the x-axis indicates results from islet autoimmunity status and the y-axis the results from Cesarean section vs. vaginal delivery. **C:** Empirical distribution of correlation coefficients between Cesarean section and permuted class labels of islet autoimmunity status for 5,000 permutations. The red line indicates the 'true' correlation between the results from the analysis of Cesarean section and islet autoimmunity status. Black lines indicate the correlation between Cesarean section and multiple first-degree relatives (multiple FDR), maternal diabetes (MD), and gender. **D:** Correlation of pathway directed p-values in Cesarean section analysis and islet autoimmunity status analysis. **E:** Empirical distribution of correlation coefficients between Cesarean section and permuted class labels of islet autoimmunity status at pathway level for 1,000 permutations. The red line indicates the 'true' correlation between Cesarean section pathways and islet autoimmunity status pathways. Abbreviations: CS = Cesarean section, VD = Vaginal delivery, Ab+ = Islet autoantibody-positive, Ab- = Islet autoantibody negative.

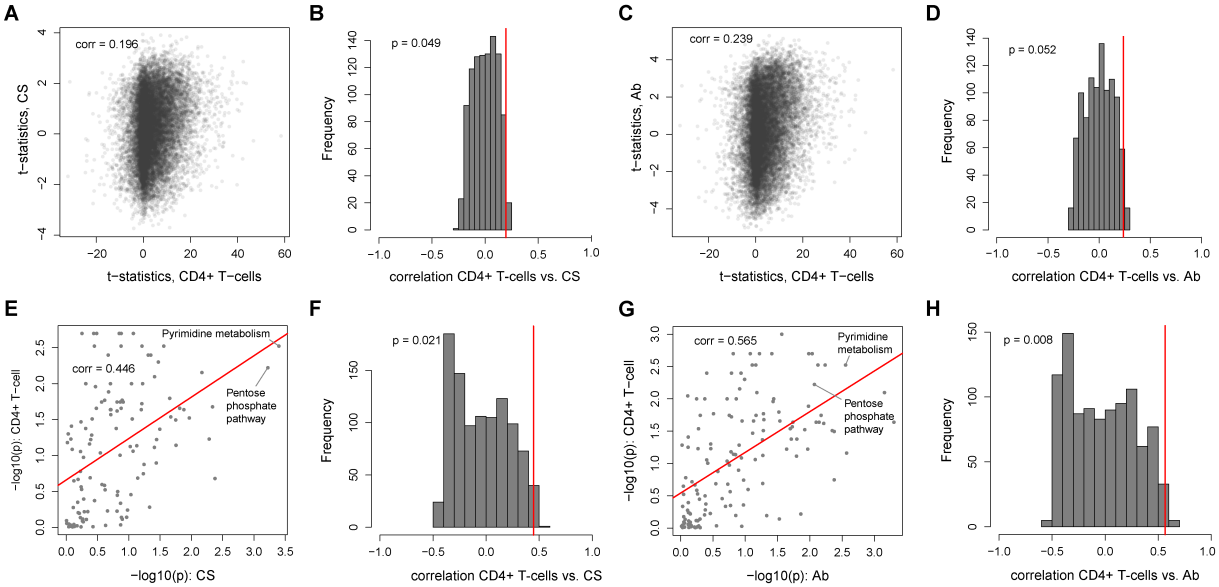


Figure 5: Signatures of immune cell activation- A: Correlation between single gene effects in Cesarean section (CS) compared to the association between naïve and activated CD4+ cells. **B:** Histogram of correlation of permuted class labels of CD4 T cells and Cesarean section at the single gene level and the "true" correlation effect. **C:** Correlation between single gene effects in islet autoimmunity status (Ab) compared to the association between naïve and activated CD4+ cells. **D:** Histogram of correlation of permuted class labels of CD4 T cells and islet autoimmunity status at the single gene level and the "true" effect. **E:** Correlation between pathway effects in Cesarean section compared to the association between naïve and activated pathways for CD4+ cells. **F:** Histogram of correlation of permuted class labels of CD4 T cells and Cesarean section at the pathway level and the "true" effect. **G:** Correlation between pathway effects in islet autoimmunity status compared to the association between naïve and activated pathways for CD4+ cells. **H:** Histogram of correlation of permuted class labels of CD4 T cells and islet autoimmunity status at the pathway level and the "true" effect.

Supplemental files

Supplemental file 1: Detailed information on longitudinal sampling and on Cesarean section in the dataset.

Supplemental file 2: Statistical results of all gene and pathway analyses.

Supplemental file 3: Results of analysis for samples up to 6 months before islet autoimmunity compared to age matched islet autoantibody negatives children. Results of paired analysis of samples before and after seroconversion.

Supplemental file 4: Pentose phosphate pathway with two-sided node colouring according to directed p-values.

Supplemental file 5: Permutation results for maternal diabetes, gender and multiple first-degree relative.

Supplemental file 6: Gene list downloaded from innateDB filtered for human genes. T-statistics of annotated and not annotated genes are shown in a violin plot.

Supplemental file 7: Detailed single gene and pathway results of transcriptomics dataset GEO-22886 of activated immune cells