TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

MULTI-OMICS ANALYSIS OF EARLY MOLECULAR MECHANISMS OF TYPE 1 DIABETES

Essi Laajala

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1602 | MEDICA – ODONTOLOGICA | TURKU 2021

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

MULTI-OMICS ANALYSIS OF EARLY MOLECULAR MECHANISMS OF TYPE 1 DIABETES

Essi Laajala

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1602 | MEDICA – ODONTOLOGICA | TURKU 2021

University of Turku

Faculty of Medicine
Department of Medical Microbiology and Immunology
Turku Doctoral Programme of Molecular Medicine
Turku Bioscience Centre, University of Turku and Åbo Akademi University

Supervised by

Professor Riitta Lahesmaa
Faculty of Medicine
University of Turku
Turku, Finland

Associate Professor Harri Lähdesmäki
Department of Computer Science
Aalto University
Espoo, Finland

Reviewed by

Professor Ville Mustonen
Department of Computer Science
University of Helsinki
Helsinki, Finland

Adjunct Professor Niina Sandholm
Folkhälsan Research Center
University of Helsinki and
Helsinki University Hospital
Helsinki, Finland

Opponent

Professor Lude Franke
Department of Genetics
University Medical Centre Groningen
Groningen, The Netherlands

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-951-29-8738-2 (PRINT)
ISBN 978-951-29-8739-9 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)
Painosalama, Turku, Finland 2021

UNIVERSITY OF TURKU
Faculty of Medicine
Department of Medical Microbiology and Immunology
ESSI LAAJALA: Multi-omics Analysis of Early Molecular Mechanisms of Type 1 Diabetes
Doctoral Dissertation, 79 pp.
Turku Doctoral Programme of Molecular Medicine
November 2021

ABSTRACT

Type 1 diabetes (T1D) is a complicated autoimmune disease with largely unknown disease mechanisms. The diagnosis is preceded by a long asymptomatic period of autoimmune activity in the insulin-producing pancreatic islets. Currently the only clinical markers used for T1D prediction are islet autoantibodies, which are a sign of already-broken immune tolerance. The focus of this dissertation is on the early asymptomatic period preceding seroconversion to islet autoantibody positivity.

The genetic risk of type 1 diabetes has been thoroughly mapped in genome-wide association studies, but environmental factors and molecular mechanisms that mediate the risk are less well understood. According to the hygiene hypothesis, the risk of immune-mediated disorders is increased by the lack of exposure to pathogens in modern environments. Within a study on the hygiene hypothesis, we compared umbilical cord blood gene expression patterns between children born in environments with contrasting standards of living and type 1 diabetes incidences (Finland, Russia, and Estonia). The differentially expressed genes were associated with innate immunity and immune maturation. Our results suggest that the environment influences the immune system development already in-utero.

Furthermore, we analyzed genome-wide DNA methylation and gene expression profiles in samples collected prospectively from Finnish children and newborn infants at risk of type 1 diabetes. Bisulfite sequencing analysis did not show any association of neonatal DNA methylation with later progression to T1D. However, antiviral type I interferon response in early childhood was found to be a risk factor of T1D. This transcriptomic signature was detectable in the peripheral blood already before islet autoantibodies, and the main observations were confirmed in an independent German study. These results contributed to the hypothesis that virus infections might play a role in T1D.

Additionally, this dissertation contributed to transcriptomic and epigenomic data analysis workflows. Simple probe-level analysis of exon array data was shown to improve the reproducibility, specificity, and sensitivity of detected differential exon inclusion events. Type 1 error rate was markedly reduced by permutation-based significance assessment of differential methylation in bisulfite sequencing studies.

KEYWORDS: Bioinformatics, type 1 diabetes, transcriptomics, DNA methylation, alternative splicing, microarrays, bisulfite sequencing, RRBS

TURUN YLIOPISTO
Lääketieteellinen tiedekunta
Lääketieteellinen mikrobiologia ja immunologia
ESSI LAAJALA: Tyypin 1 diabeteksen varhaisten molekulaaristen mekanismien multiomiikka-analyysi
Väitöskirja, 79 s.
Molekyylilääketieteen tohtoriohjelma
Marraskuu 2021

TIIVISTELMÄ

Tyypin 1 diabetes (T1D) on autoimmuunitauti, jonka taustalla olevista mekanismeista tiedetään vähän. Diagnoosia edeltää pitkä oireeton jakso, jonka aikana insuliinia tuottaviin beetasoluihin kohdistuva autoimmuunireaktio etenee haiman saarekkeissa. Tämä väitöskirjatutkimus keskittyy T1D:n varhaiseen oireettomaan ajanjaksoon, joka edeltää serokonversiota autovasta-ainepositiiviseksi.

Tyypin 1 diabeteksen geneettiset riskitekijät on kartoitettu perusteellisesti genominlaajuisissa assosiaatiotutkimuksissa, mutta ympäristön riskitekijöistä ja riskiä välittävistä molekyylimekanismeista tiedetään vähemmän. Hygieniahypoteesin mukaan vähäinen altistuminen taudinaiheuttajille lisää immuunijärjestelmän häiriöiden riskiä. Hygieniahypoteesiin liittyvässä osatyössä vertasimme hygienian ja T1D:n ilmaantuvuuden suhteen erilaisissa ympäristöissä (Suomi, Venäjä ja Viro) syntyneiden lasten napaveren geeniekpressioprofiileja. Erilaisesti ekspressoituneet geenit liittyivät synnynnäiseen immuniteettiin ja immuunijärjestelmän maturaatioon. Näiden tulosten perusteella ympäristö saattaa vaikuttaa immuunijärjestelmän kehitykseen jo raskauden aikana.

Genominlaajuista DNA-metylaatiota ja geeniekspressiota analysoitiin näytteistä, jotka oli kerätty laajassa suomalaisessa seurantatutkimuksessa T1D:n riskiryhmään kuuluvilta lapsilta ja vastasyntyneiltä. Bisulfiittisekvensointianalyysin perusteella vastasyntyneen DNA-metylaation ja lapsuuden aikana kehittyvän T1D:n välillä ei ollut yhteyttä. Sen sijaan RNA:n tasolla havaittava viruksiin kohdistuva tyypin 1 interferonivaste varhaislapsuudessa todettiin T1D:n riskitekijäksi. Tämä havainto tehtiin perifeerisestä verestä jo ennen saarekevasta-aineiden ilmaantumista, ja päähavainnot vahvistettiin saksalaisessa tutkimuksessa. Nämä tulokset vahvistivat hypoteesia, jonka mukaan virukset voivat vaikuttaa T1D:n puhkeamiseen.

T1D-tutkimuksen ohella tämä väitöskirjatyö kehitti transkriptomiikkaan ja epigenomiikkaan sopivia analyysimenetelmiä. Eksonimikrosirujen koetintasoisen analyysin todettiin parantavan toistettavuutta, sensitiivisyyttä ja tarkkuutta vaihtoehtoisen silmukoinnin kartoittamisessa. Tilastollisen merkitsevyyden permutaatiopohjainen analyysi vähensi tyypin 1 virhettä bisulfiittisekvensointidatan analyysissa.

AVAINSANAT: Bioinformatiikka, tyypin 1 diabetes, transkriptomiikka, DNA-metylaatio, vaihtoehtoinen silmukointi, mikrosirut, bisulfiittisekvensointi, RRBS

Table of Contents

Abbreviations

A	Adenine (one of four bases of DNA)
ANOVA	Analysis of Variance
ANOSVA	Analysis of Splice Variation
APC	Antigen Presenting Cell
ASCII	American Standard Code for Information Interchange
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATP	Adenosine Triphosphate
AUROC	Area Under the Receiver Operating Characteristics
BMI	Body Mass Index
bp	Base Pair
C	Cytosine (one of four bases of DNA)
cDNA	Complementary Deoxyribonucleic Acid
CHG	C, followed by any other base but G, followed by G
CHH	C, followed by two bases other than G
ChIP-seq	Chromatin Immunoprecipitation Sequencing
C-peptide	Connecting Peptide (a marker of insulin secretion)
C-section	Caesarean section
CpG	C followed by G (p stands for the phosphate that connects two adjacent bases in DNA)
CVB	Coxsackievirus B
DIPP	Diabetes Prediction and Prevention Study: a large Finnish follow-up cohort of children at risk of type 1 diabetes
DMC	Differentially Methylated Cytosine
DMR	Differentially Methylated Region
DNA	Deoxyribonucleic Acid
EM	Expectation Maximization
eQTL	Expressed Quantitative Trait Locus/Loci
eQTM	Expressed Quantitative Trait Methylation
FDR	False Discovery Rate
FIRMA	Finding Isoforms Using Robust Multichip Analysis
FWER	Family-wise Error Rate

GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed-effects Model
H3K9	Lysine (K) 9 on Histone H3
HLA	Human Leukocyte Antigen
IA2A	Insulinoma-Associated Protein 2 Antibody, also known as Islet Antigen 2 Antibody
IAA	Insulin Antibody
ICA	Islet Cell Antibodies
IRLS	Iteratively Reweighted Least Squares
LC/MS	Liquid Chromatography Mass Spectrometry
LM	Linear Model
MCMC	Markov Chain Monte Carlo
MIDAS	Microarray Detection of Alternative Splicing
MM	Mismatch (probe on Affymetrix microarrays)
G	Guanine (one of four bases of DNA)
GADA	Glutamic Acid Decarboxylase Antibody
GSEA	Gene Set Enrichment Analysis
GWAM	Genome-Wide Average Methylation
GWAS	Genome-Wide Association Study/Studies
GWS	Genome-Wide Significance
meQTL	Methylation Quantitative Trait Locus/Loci
mRNA	Messenger RNA
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PECA	Probe-level Expression Change Averaging
PLIER	Probe Logarithmic Intensity Error Model
PM	Perfect match (probe on Affymetrix microarrays)
RMA	Robust Multichip Average
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
PBMC	Peripheral Blood Mononuclear Cell
ROC	Receiver Operating Characteristics
RRBS	Reduced Representation Bisulfite Sequencing
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SI	Splicing Index
SNP	Single Nucleotide Polymorphism
T	Thymine (one of four bases of DNA)
Th2	T helper type 2 (lymphocyte)
T1D	Type 1 Diabetes
U	Uracil

WGBS	Whole Genome Bisulfite Sequencing
ZnT8A	Zinc Transporter 8 Antibody

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Laajala E, Aittokallio T, Riitta Lahesmaa R, and Elo LL. 2009. Probe-Level Estimation Improves the Detection of Differential Splicing in Affymetrix Exon Array Studies. *Genome Biology*, 2009; 10 (7): R77.

II Kallionpää H*, Elo LL*, Laajala E*, Mykkänen J, Ricaño-Ponce I, Vaarma M, Laajala TD, Hyöty H, Ilonen J, Veijola R, Simell T, Wijmenga C, Knip M, Lähdesmäki H, Simell O, Lahesmaa R. Innate Immune Activity Is Detected Prior to Seroconversion in Children with HLA-conferred T1D Susceptibility. *Diabetes*, 2014; 63(7):2402–2414. *equal contribution

III Kallionpää H*, Laajala E*, Öling V, Härkönen T, Tillmann V, Dorshakova NV, Ilonen J, Lähdesmäki H, Knip M#, Lahesmaa R#, and the DIABIMMUNE study group. Standard of Hygiene and Immune Adaptation in Newborn Infants. *Clinical Immunology*, 2014; 155(1):136-147. *equal contribution

IV Laajala E, Ullah U, Grönroos T, Rasool O, Halla-aho V, Konki M, Kattelus R, Mykkänen J, Nurmio M, Vähä-Mäkilä M, Kallionpää H, Lietzén N, Ghimire BR, Laiho A, Hyöty H, Elo LL, Ilonen J#, Knip, M#, Lund RJ#, Orešič M#, Veijola R#, Lähdesmäki H, Toppari J, Lahesmaa R. DNA Methylation of Umbilical Cord Blood Samples in Children Who Later Develop Type 1 Diabetes. *medRxiv*, 2021; https://doi.org/10.1101/2021.05.21.21257593 #equal contribution

V Laajala E, Halla-aho V, Grönroos T, Ullah U, Vähä-Mäkilä M, Nurmio M, Kallionpää H, Lietzén N, Mykkänen J, Rasool O, Toppari J, Orešič M, Knip M, Lund RJ, Lahesmaa R, Lähdesmäki H. Permutation-Based Significance Analysis Reduces The Type 1 Error Rate in Bisulfite Sequencing Data Analysis of Human Umbilical Cord Blood Samples. *bioRxiv*, 2021; https://doi.org/10.1101/2021.05.18.444359

The original publications have been reproduced with the permission of the copyright holders.

1 Introduction

The immune system is a complex network of mechanisms that enable the recognition and elimination of threats, and on the other hand the tolerance towards self-antigens and other harmless substances. Autoimmune diseases are conditions, where tolerance mechanisms have failed, and some self-antigens are systematically recognized as pathogenic.

Type 1 diabetes is a common autoimmune disease with serious complications and a world-widely increasing incidence. The symptoms of type 1 diabetes can be kept under control, but currently the disease cannot be prevented or reversed. By the time of diagnosis, most of the insulin producing pancreatic beta cells have become dysfunctional and the individual is dependent on insulin injections. This is preceded by a long asymptomatic period of autoimmune activity in the pancreatic islets. Islet autoantibodies can typically be detected in the peripheral blood some months or years before the diagnosis. Earlier molecular markers of type 1 diabetes are an active area of research to which we have contributed. Our study material includes blood samples collected before seroconversion to autoantibody positivity, which are extremely hard to obtain but most valuable to improve our understanding of mechanisms that lead to immune tolerance failure.

Both genetic and environmental factors contribute to the risk of type 1 diabetes. Their effects can be mediated by epigenetics and gene expression, which are at the center of focus of this thesis. We explored transcriptomic patterns in a longitudinal sample series (Study II) and DNA methylation in umbilical cord blood samples (Study IV) of children at risk of type 1 diabetes. We also compared the transcriptomes of children born in environments with contrasting standards of living and type 1 diabetes incidences (Study III). This was motivated by the hygiene hypothesis, according to which the risk of immune-mediated disorders is increased by the lack of exposure to pathogens in modern environments.

Such explorative studies have been enabled by the rapid development of high-throughput ’omics technologies during the past two decades. The field of bioinformatics emerged and keeps developing alongside these technologies. Studies II and III utilized gene expression microarrays and Study IV quantified DNA methylation through reduced representation bisulfite sequencing (RRBS). To better answer the biological questions within this thesis, I evaluated and improved the existing bioinformatics methodology in transcriptomic and epigenomic studies. More specifically, this thesis discusses the analysis of alternative splicing events in exon microarray data (Study I) and the estimation of statistical significance in spatially correlated bisulfite sequencing data (Study V).

2 Review of the Literature

2.1 Aetiology of type 1 diabetes

2.1.1 Type 1 diabetes disease model

Type 1 diabetes, also known as insulin dependent diabetes mellitus or juvenile diabetes, is an autoimmune disease of the insulin-producing beta cells which are located in the islets of Langerhans of the pancreas. Insulin is needed to activate glucose intake and metabolism by binding to the insulin receptor, which is present on all mammalian cells (1). Impaired insulin production leads to elevated blood glucose levels (hyperglycemia), which in the long-run causes endothelial cell dysfunction through glyco-oxidation (2). Consequently, the risk of microvascular complications such as retinopathy, nephropathy, and neuropathy are elevated among diabetic individuals, and their average life expectancy is approximately 12 years shorter compared to the general population (3,4). Type 1 diabetes symptoms can be kept under control by insulin injections and careful monitoring of blood glucose levels, but the disease cannot currently be prevented or reversed.

The JDRF (Juvenile Diabetes Research Fund), the Endocrine Society, and the American Diabetes Association have proposed the following classification of type 1 diabetes stages: 1) seroconversion to two or more islet autoantibodies (details below) 2) asymptomatic dysglycemia 3) symptomatic type 1 diabetes (5). The dysglycemic state typically fluctuates and even one incident of abnormal glucose tolerance after seroconversion is highly predictive of progression to symptomatic type 1 diabetes within a few years (6). The observed time between seroconversion and diagnosis has ranged from months to decades with median values of for example 4 or 9 years, depending on the studied population (7–9).

Type 1 diabetes can be diagnosed based on clinical symptoms, when 80–90 % of insulin-producing beta cells have been destroyed or become dysfunctional. An oral glucose tolerance test can typically reveal the impaired insulin production several months before clinical symptoms (10). Currently the earliest clinical markers used to predict type 1 diabetes onset are islet autoantibodies, which unfortunately arise at a relatively late stage of the pathogenesis. According to the traditional Eisenbarth disease model for type 1 diabetes (11) the emergence of auto-antibodies (seroconversion) is preceded by a long asymptomatic period, during which functional beta cell mass slowly decreases (Figure 1).

Several aspects of the Eisenbarth disease model for type 1 diabetes have been questioned in light of more recent observations (12,13). The insulin-producing beta cell mass does not always decline until none remains. In fact, some remaining beta cells were found in the pancreata of 37 out of 42 diabetic organ donors still 4-67 years after the diagnosis with type 1 diabetes (14) and other studies have reported similar observations (15,16). Either these remaining beta cells differ from other beta cells to escape the autoimmune destruction altogether or beta cells are being continually regenerated and destroyed in some individuals with long-standing type 1 diabetes. In individuals with newly-onset type 1 diabetes (less than 1 year from diagnosis), the residual amount of insulin-positive pancreatic islets is larger than was earlier thought, especially in those with disease-onset after age 15. The observed proportions of insulin-positive islets have been on average 56 % and 38 % in newly-diagnosed type 1 diabetes cases, with disease onset after 15 and before 15 years of age, respectively (13). However, the number of insulin-positive pancreatic islets provides little information on the remaining beta cell mass and says nothing about the beta cells' ability to respond to glucose, which can be poor in pancreatic islets with normal beta cell counts and insulin levels (17,18).

Furthermore, the idea of constant immune cell activity in pancreatic islets and gradual decline of functional beta cell mass over some years before type 1 diabetes diagnosis, has been challenged. Insulitis (the immune cell infiltration in pancreatic islets) has been a very rare observation in the pancreata of pre-diabetic organ donors positive for islet autoantibodies (19) and is not always present at the time of diagnosis either (20). Autoimmune activity in pancreatic islets could be occurring in a relapsing-remitting fashion, which is typical for some other autoimmune diseases and is somewhat supported by the frequent observation of improved insulin production a few months after type 1 diabetes is diagnosed (21). The original Eisenbarth disease model from 1986 and the modern version are illustrated in Figure 1.

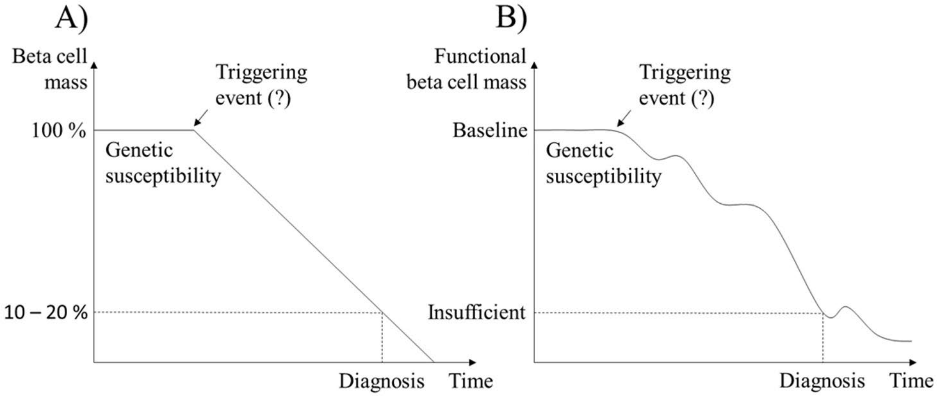


Figure 1. An illustration of A) the original and B) the updated Eisenbarth disease model of type 1 diabetes, modified from A) Eisenbarth 1986: "Type I Diabetes Mellitus", The New England Journal of Medicine 314 (21): 1360–68 and B) Herrath, Sanda, and Herold 2007: "Type 1 Diabetes as a Relapsing-Remitting Disease?" Nature Reviews. Immunology 7 (12): 988–94.

2.1.2 Roles of different cell types in type 1 diabetes

Immune cell infiltration in the pancreatic islets (insulitis) was first observed in 1902 in a child who died of ketoacidosis and has later been confirmed by several studies on organ donors with type 1 diabetes (13). Cytotoxic CD8+ T lymphocytes have been the most abundant immune cells in such pancreatic lesions, followed by macrophages, B lymphocytes and CD4+ T lymphocytes (22). The autoimmune nature of type 1 diabetes is a long-standing hypothesis, which has only relatively recently been confirmed by the observation of islet autoantigen reactive CD8+ T cells by in situ HLA tetramer staining in the pancreatic islets of organ donors with type 1 diabetes (23). Their presence was specific to type 1 diabetes and was not detected in non-diabetic individuals or individuals with type 2 diabetes.

The mechanisms that lead to this immune tolerance failure remain largely unknown. What are the most important factors that make immune cells prone to attack beta cells or the beta cells prone to be attacked? The active role of beta cells has been emphasized for example by studies on endoplasmic reticulum stress that can promote a pro-apoptotic feedback loop in the beta cells (24). While most studies have focused on T cells, which are part of the adaptive immune system, others have emphasized the role of innate immune mechanisms in triggering the inflammation that causes beta cell stress and promotes T cell autoreactivity (25). A comprehensive survey on cell-type-specific regulatory elements that might mediate the development of type 1 diabetes was recently published by Chiou et al. (26). They tested the enrichment of genomic type 1 diabetes risk loci from genome-wide association studies (GWAS) on candidate cis regulatory elements with cell-type specific accessibility in single-cell ATAC-seq data on human blood and pancreatic tissues. Significant enrichment was found on regulatory elements specifically accessible in CD4+ and CD8+ T cells.

2.1.3 Islet autoantibodies

Islet cell antibodies (ICA) target cytoplasmic proteins in beta cells and are typically measured with immunofluoresence detecting blood serum ICA binding on human pancreatic cells from organ donors (27). The specific islet autoantigens include insulin, insulinoma-associated protein 2 (IA2, also known as islet antigen 2), glutamic acid decarboxylase 65 (GAD65), and zinc transporter-8 (ZnT8). The autoantibodies for these (IAA, IA2A, GADA, and ZnT8A) are routinely measured with specific radiobinding assays. The less specific ICA immunofluoresence test can be positive for individuals negative for IAA, IA2A, GADA, and ZnT8A, indicating that all islet autoantigens have not yet been identified (27).

The median age of seroconversion to autoantibody positivity in children at risk of type 1 diabetes has been approximately 2 years (8). GADA or IAA typically appear first (28). Individuals with GADA as first vs. IAA as first-appearing autoantibody are characterized by distinct HLA-DR-DQ-haplotypes and IAA was associated with early seroconversion (29).

The risk of progression to type 1 diabetes within 10 years of seroconversion has been estimated to be 10–30 % for individuals positive for only one islet autoantibody and 60–90 % for individuals positive for multiple autoantibodies (8,30), depending on the study population and other risk factors. The risk depends also largely on the antibody or the combination of antibodies. For example the combination of IAA and IA2A confers a significantly larger risk of type 1 diabetes than any other combination of two antibodies studied in the Finnish-German-American meta-analysis (8) and IA2A was found to increase the risk of type 1 diabetes more than any other second autoantibody appearing after IAA or GADA (28).

Ongoing beta cell death and/or insulitis have only been confirmed in a small fraction of islet antibody positive individuals (19). Nevertheless, autoantibodies are a sign of autoimmunity and the above-mentioned type 1 diabetes associated autoantibodies are important predictive and diagnostic markers, for example to distinguish between type 1 and type 2 diabetes in newly diagnosed adults.

2.1.4 Disease subtypes

Type 1 diabetes is a heterogenous disease, which develops with highly variable time schedules and characteristics. Several studies have suggested the existence of different disease subtypes. For example, the above-described GADA-first and IAA-first profiles might represent different subtypes of type 1 diabetes.

The pancreata with some remaining beta cells after type 1 diabetes diagnosis differ in their patterns: either the remaining beta cells are only found in few pancreatic lobes and expressing an inhibitor of apoptosis or 100 % of pancreatic islets contain small numbers of normal-appearing beta cells (31). Individuals with childhood-onset type 1 diabetes present more often with insulitic lesions and have less remaining beta cells at diagnosis, as compared to those with disease onset after age 15 (13). Other histological studies have suggested type 1 diabetes subtypes characterized by high vs. low frequencies of CD20+ B-cells present in pancreatic islets (32). Pancreatic islets from individuals with early type 1 diabetes onset (age < 7 years) were characterized by high frequencies of CD20+ B-cells, aberrant proinsulin processing and low C-peptide levels, as compared to individuals with disease onset after 13 years of age (33). Altogether, these observations suggest a more aggressive autoimmune process in individuals with early-onset type 1 diabetes.

2.1.5 Environmental risk factors of type 1 diabetes

Even though type 1 diabetes is a highly inheritable common disease, only 15 % of newly-diagnosed individuals have a family history of type 1 diabetes (34,35). The risk of type 1 diabetes among the monozygotic twins of individuals with type 1 diabetes is 50–70 %, and the time of onset between twins can differ by decades (36,37). This partial discordance in monozygotic twins, as well as the worldwide increase in the incidence of type 1 diabetes (38,39), prove that the disease risk is a combination of hereditary and environmental factors. The role of the environment is further supported by the increase in type 1 diabetes incidence among people, who have migrated to a country with a higher incidence of type 1 diabetes (40,41). The genomic risk loci have been thoroughly mapped in genome-wide association studies (GWAS), based on data collected from hundreds of thousands of individuals (26,42,43), but non-genetic risk factors are less understood.

Environmental/behavioral factors that have been observed to correlate with type 1 diabetes include the level of hygiene (44), early exposure to cow's milk (45,46), especially A1 β-casein in cow's milk (47), vitamin D deficiency (48,49), high birth weight and rapid growth in early childhood (50–52), certain viral infections (53), and early exposure to perfluoroalkyl substances (54). Some of these observations have led to intervention studies, none of which have confirmed causality or lead to altered recommendations (12,55). The field has suffered from lack of reproducibility, which might reflect the heterogeneous nature of type 1 diabetes, complex and unknown interactions between risk factors, as well as slightly varying goals and designs of different studies (53). For example, opposite correlations between Coxsackievirus B

(CVB) infections and the risk of type 1 diabetes have been observed, depending on the virus strain and the time of infection–an early infection with some virus strain might protect from a later infection with a more diabetogenic strain (56).

Exposure to pathogens is one of the most extensively studied environmental factor that may have a role in the development of type 1 diabetes. While some pathogens, such as the common parasite Schistosoma Mansoni, might have a protective effect (57), others have been associated with an increased risk of type 1 diabetes. For example Encefalomyocarditis-D viruses (EMC-D) have been shown to selectively target pancreatic beta cells in animal models for type 1 diabetes (58). Especially in the event of a persistent viral infection, permanent immune tolerance failure might develop through increased presentation of antigens of damaged beta cells by antigen presenting cells and HLA class I hyperexpression by beta cells (23,56). Viruses might also induce autoreactivity of T cells, if the virus proteins closely resemble autoantigens (59). For example the viral glycoprotein VP7 of rotavirus is recognized by T cell receptors that also bind to the islet autoantigen IA-2 (60).

In humans, the association between type 1 diabetes and rota- and enteroviruses, especially CVB, is supported by growing evidence (61–63). Increased incidence of type 1 diabetes was associated with a CVB epidemic already in the 1980s (64). The frequency of CVB1 has been slightly greater (odds ratio 1.7) among children with type 1 diabetes compared to matched control children, based on serological evidence from 249 children per group (65). Much larger odds ratios were observed in a meta-analysis of studies investigating the correlation between enteroviral infections and type 1 diabetes with molecular virological methods (66).

The association between type 1 diabetes and enteroviral infections has been especially strong in studies that have focused on the time of type 1 diabetes onset (17,56,67,68). For example, CVB4 has been found in the beta cells of some organ donors newly diagnosed with type 1 diabetes and is able to impair insulin-production when introduced to healthy pancreatic islets (17). The presence of enteroviral capsid protein (VP1) was detected in the pancreatic islets of all six newly diagnosed living diabetic individuals, who donated pancreatic biopsy samples, whereas the same observation was made in only two out of nine control samples from non-diabetic organ donors (69).

To investigate the possible causality between CVB infection and the development of type 1 diabetes, Gallagher and others engrafted mice with human islets and infected some mice with CVB4. Out of 15 CVB-infected mice, 7 developed diabetes within five weeks, whereas all 5 mock-infected control mice remained normoglycemic (70). Enterovirus vaccination trials for risk groups of type 1 diabetes have been suggested (63). A significant decrease in type 1 diabetes incidence has been reported in Australia after the introduction of rotavirus vaccine in the national vaccination program (71).

2.1.6 DNA methylation and type 1 diabetes

DNA methylation is the addition of a methyl group to cytosine (C) that occurs almost without exception in the CpG (cytosine followed by guanine) context in mammals (72,73). It is a mitotically inheritable epigenetic mechanism that typically silences gene expression when present at the promoter (74). DNA methylation is required for processes that are essential for normal development, such as cellular differentiation and genomic imprinting (75). In mammals, the majority of the CpG sites are methylated but some genomic regions, such as promoters of expressed genes, are maintained in an unmethylated state (76). The methylation states can be spatially correlated between CpG sites up to a distance of approximately 2 kilobases, depending on the genomic context (77,78).

DNA methylation patterns are established in-utero (79) and have been suggested to mediate the impacts of the in-utero environment on later health (80–82). Compared to for example the transcriptome or the proteome, the temporal within-individual variation of DNA methylation is small (83). If epigenomic differences in early life reflect the risk of later type 1 diabetes, the differences may be observable in cross-sectional data, such as the one presented in Study IV.

Genome-wide average methylation (GWAM) on promoter regions has been observed to be highly correlated between pairs of both monozygotic (r=0.82) and dizygotic (r=0.85) twins at the time of birth, indicating that GWAM is strongly influenced by the in-utero environment but not necessarily by genetics (84). However, genetic polymorphisms known as methylation quantitative trait loci (meQTL) affect DNA methylation at specific locations (74,85), indicating some degree of genetic heritability of DNA methylation patterns. Cis-acting meQTL effects have been observed in a large proportion of type 1 diabetes associated GWAS loci, suggesting that DNA methylation might mediate the genetic risk of type 1 diabetes (86).

Associations between DNA methylation patterns and type 1 diabetes have mainly been explored between already-diagnosed individuals and healthy controls. Hypothesis-driven studies on specific genomic loci have identified type 1 diabetes associated methylation for example at the promoters of the insulin gene and interleukin-2 receptor alpha chain (IL2RA) (87,88). The most extensive observational study identified thousands of CpG sites with differentially variable methylation proportions between 52 pairs of monozygotic twins discordant for type 1 diabetes, but only one CpG site was identified as differentially methylated at genome-wide significance (89).

The only published prospective study on the association between DNA methylation and later progression to type 1 diabetes reported two differentially methylated CpG sites (DMCs) and 28 differentially methylated regions (DMRs), some of which were discovered already before the case individuals' seroconversion to islet autoantibody positivity (90). Both these studies (89,90) included some umbilical cord blood samples but they were only utilized to test, whether the above-mentioned findings could be confirmed, and genome-wide cord blood DNA methylation measurements were not published. To our knowledge, other published results on the possible association between neonatal DNA methylation patterns and later progression to type 1 diabetes are not yet available.

2.1.7 RNA-level gene expression and type 1 diabetes

The study of genome-wide gene expression on the level of RNA is referred to as transcriptomics. There are numerous examples of useful clinical applications that have been enabled by transcriptomics, such as non-invasive prediction of cardiac allograft rejection from peripheral blood mononuclear cell RNA (91) or molecular tumor profiling for the identification of the tissue of origin in cancer of unknown primary (92).

Since most type 1 diabetes associated GWAS loci reside outside protein-coding regions, transcriptomic studies are necessary to understand the mechanisms behind their impact on disease risk (43). Genetic variants that correlate with gene expression are called expression quantitative trait loci (eQTL). Human tissue-specific eQTL have been mapped for example by the Genotype-Tissue Expression (GTEx) consortium (93), and a human whole blood eQTL database is available as part of the BIOS QTL browser (94).

The association between transcriptomic patterns and type 1 diabetes have been studied especially in animal models and in human pancreatic tissues from organ donors (95). For example, gene expression microarray profiling of pancreatic islets and pancreatic lymph nodes of NOD (non-obese diabetic) mice identified two genes that were differentially expressed and alternatively spliced between NOD and nondiabetic NOD.B10 mice and showed similar patterns between diabetic and non-diabetic human organ donors (96). More recent advances include a single-cell-level transcriptomic atlas of the human pancreas (97) and an inflammation-specific beta cell regulatory landscape (98).

Given the invasiveness of pancreatic biopsies, pancreatic gene expression patterns cannot be utilized in clinical applications, unless they are reflected in peripheral blood. Therefore, blood-based signatures of type 1 diabetes have been an active area of research, reviewed for example by Cabrera et al. (99). Before Study II that was conducted within this thesis and published in 2014 together with a similar but independent German study (100), all peripheral blood transcriptomic studies on type 1 diabetes had focused on people with a clinical type 1 diabetes diagnosis at the time of sample collection (99). All these studies reported associations between gene expression and type 1 diabetes, especially among inflammatory genes (101–107). However, it was not known whether any associations could be found in a prospective study setting before disease onset or even before the seroconversion to islet autoantibody positivity.

After Study II, pre-seroconversion transcriptomic studies have been conducted in a larger whole blood data set with microarrays (108), isolated CD4+, CD8+, and CD4-CD8- cell fractions and unfractionated PBMCs with RNA sequencing (109) and within islet autoantigen responsive CD4+ T cells with single-cell PCR (110). These three recent studies conclude that already during the first year of life, gene expression patterns are associated with and/or can even be used to predict later progression to islet autoimmunity.

2.2 High-throughput technologies for transcriptomic and epigenomic studies

2.2.1 Gene expression microarrays

Gene expression microarray technology was preceded by Northern blotting, which was used to probe the expression level of one transcript at the time. The Northern blotting workflow included 1) RNA isolation from the cells/tissues, 2) size-separation of RNA molecules by gel electrophoresis, 3) transfer and attachment of the RNA from the gel on a paper or membrane, 4) hybridization of radioactively labeled probes complementary to the RNA of interest, and 5) X-ray detection of hybridization (111). In gene expression microarray technology, introduced in 1995, the probes are readily attached on an array and the expression levels of tens of thousands of mRNA molecules can be simultaneously detected (112).

Gene expression microarray study protocols include 1) isolation, fragmentation, and purification of poly-A-tailed mRNA from cells/tissue, 2) production of either cDNA or cRNA molecules complementary to the sample mRNA, 3) PCR amplification of the target cDNA or cRNA if needed, 4) hybridization of the target on the array of readily-attached probes, 5) washing of the arrays to remove unhybridized material and 6) detection of hybridization by fluorescent labels (113).

Different gene expression microarray technologies exist in two main categories: two-color and one-color microarrays. Two-color microarrays can only be used with paired study-designs: one fluorescent dye is used for the case cDNA sample and a different dye for the control cDNA sample, and each pair of samples is hybridized on a single array to detect the relative gene expression between the case and the control. The two-color technology has mostly been used in the context of spotted microarrays, where the array-to-array-variability is large (114). Spotted arrays are glass or nylon surfaces lined with spots of (typically) in-house designed cDNA probes that are attached after being generated (115).

One-color technology can be used in the context of well-standardized oligonucleotide arrays, which are produced either by in-situ-synthetization or self-assembled synthetization of probes/probesets (113). The in-situ-synthetization is done directly on the array surface and has been used for example by the manufacturer Affymetrix. Illumina uses self-assembled synthetization, which takes place on labeled silica or polystryrene beads, deposited on an array of micro-wells. The production of synthetic oligonucleotides requires pre-existing DNA sequence information but is very efficient compared to cloning probes from a DNA library for the spotted arrays (112). The usage of spotted microarrays was largely replaced by in situ synthetized oligonucleotide arrays in the end of 1990s, as publicly available DNA sequence information was rapidly increasing (113).

Short oligonucleotide microarrays suffer from significant amounts of unspecific hybridization, and probe sequences as short as 25 bases (which used to be the standard length) can match more than one transcript (116). Although these issues have been alleviated by the increasing probe sequence length, further challenges in probe design include the presence of genomic variations and alternative splicing events. Manufacturers have kept updating their probe sets to target the sequences they are intended to target, and therefore transcript-level measurements from different platform versions are not always comparable (117). Probes/probesets from different platforms of the same manufacturer might target completely non-overlapping genomic regions, even if they are labelled with the same identifier (118). The main limitation of microarrays is that they can only measure the expression of known targets. During the last decade, gene expression microarrays have been largely replaced by RNA sequencing.

2.2.2 Exon microarrays

Exon arrays have been designed for the detection of alternative splicing events. Two major classes of array designs have been implemented for genome-wide alternative splicing studies: 1) exon junction arrays and 2) arrays with exon-specific probes/probesets (119). Affymetrix human exon arrays include 4 perfect-match probes for each of 1.4 million known or predicted exons in the human genome (120). Compared to exon junction arrays, they are more easily applicable to gene expression studies, and might in fact quantify gene expression similarly or even more accurately than gene expression microarrays (121,122). The main limitation of exon arrays is that they can only detect alternative exon inclusion/exclusion events among exons targeted by the designed probes. RNA sequencing is much more flexible in detecting previously unknown alternative splicing, including events such as intron retention and alternative 3’ or 5’ splice sites in addition to exon inclusion/exclusion (123).

2.2.3 Next-generation sequencing

High-throughput sequencing is nowadays a relatively cost-efficient method to for example quantify gene expression on the RNA level, to obtain the genomic sequence of a novel pathogen or to identify DNA methylation patterns in a biological sample. Modern day short-read sequencing was preceded by Sanger sequencing, which was used in the Human Genome Project, the cost of which has been estimated to have been 0.5–1 billion dollars (124). Sanger sequencing was based on introducing the fragmented and denaturated DNA sample of interest to a mixture of DNA polymerase, primers and all four types of nucleotides, some of which include a fluorescently labeled 3’ block that irreversibly terminates the polymerization (125). The result is a mixture of double stranded DNA fragments of different lengths, each fluorescently labelled with one of the four different colors indicating the type of the terminator base. The sequence can then be read by size separation.

During the last two decades, innovations in sequencing protocols have brought the cost of a sequenced genome down to some hundreds of dollars. All protocols include: 1) Sample fragmentation and size selection 2) Adapter ligation. The adapters typically contain three segments: a short sequence that gets attached to the flow cell, a sample identifier to enable multiplexing (several samples sequenced on one lane) and a sequence complementary to the primer to initiate amplification and sequencing. 3) Library amplification by polymerase chain reaction (PCR) 4) Sequencing.

The most commonly used protocol is Illumina’s reversible terminator short read sequencing (124,126). After the library preparation steps, common to all protocols, the prepared DNA fragments are washed across a flow cell, lined with sequences complementary to the beginning of the adapter sequence. The fragments captured by the flow cell are then bridge amplified to generate a cluster of identical sequences from each fragment (each cluster corresponding to one read). Fluorecently labelled terminator bases, each base with a different colored label, are then added to the flow cell, together with DNA polymerase and primers. At each round, unattached bases are washed away and the color of the fluorescence of attached bases detected, after which the 3' block is removed to start the next round.

2.2.4 Reduced representation bisulfite sequencing

Bisulfite sequencing is a technology to detect DNA methylation at a single-nucleotide resolution. The detection is based on bisulfite treatment, which converts unmethylated cytosine (C) to uracil (U), which is then read as thymine (T). Instead of whole-genome bisulfite sequencing (WGBS), a common practice is to use reduced representation bisulfite sequencing (RRBS), which captures CpG-rich genomic regions and is therefore more cost-efficient than WGBS. In the human genome, the small fraction (1%) of DNA captured by MspI enzyme digestion typically covers 2.5–3 million CpG sites, which is approximately 10 % of the total number of human CpG sites (127,128). In comparison, the most common technology used in epigenome-wide association studies is the Illumina 450k DNA methylation microarray which targets approximately 450 000 CpG sites (129).

The RRBS protocol (127) includes the following steps (Figure 2): 1) MspI enzyme recognizes CCGG sequences and cleaves them asymmetrically, leaving CGG in the 5’ end of the top strand and only a C at the 3’ end of the bottom strand. This is followed by end repair, which completes the 3’ ends with CG. 2) Both strands are A-tailed (a single A added to the 3’ ends) and adapters are ligated to these A-tails. 3) The fragments are size selected in order to enrich for promoter regions and CpG islands. Typically, a fragment size range of 40–220 bp is selected. 4) Bisulfite conversion 5) Library amplification by PCR. If proofreading is used, it needs to be done with an enzyme that does not stall at Uracil. 6) Sequencing, for example with the above-described Illumina protocol.

The fragment length is an important property to consider in planning RRBS experiments. In MspI digested human genome the fragment length distribution is skewed towards the shorter end of the spectrum (130). Therefore, increasing read length does not linearly increase the number of detected CpG sites. Furthermore, paired-end sequencing is not as cost-efficient as it would be in the context of e.g. whole genome bisulfite sequencing (WGBS) or RNA sequencing (RNA-seq), where the fragment length is typically size-selected to be 200-400 bp (131,132). Paired-end RRBS often includes a substantial amount of overlapping pairs of reads. This leads to the rejection of some read 2 data to avoid double-calling the methylation statuses of cytosines within short fragments (133).

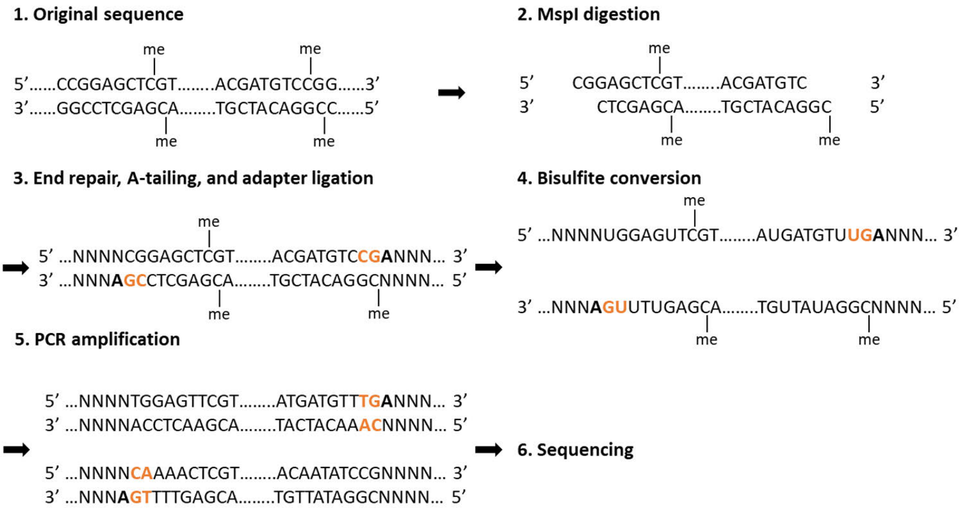

1. Original sequence
me
me
5’……CCGGAGCTCGT……..ACGATGTCCGG……3’
3’……GGCCTCGAGCA……..TGCTACAGGCC……5’
me
me
2. MspI digestion
me
5’ CGGAGCTCGT……..ACGATGTC 3’
3’ CTCGAGCA……..TGCTACAGGC 5’
me
me
3. End repair, A-tailing, and adapter ligation
me
5’ …NNNNCGGAGCTCGT……..ACGATGTCCGANNN… 3’
3’ …NNNAGCCTCGAGCA……..TGCTACAGGCNNNN… 5’
me
me
4. Bisulfite conversion
me
5’ …NNNNUGGAGUTCGT……..AUGATGTUUGANNN… 3’
3’ …NNNAGUUTUGAGCA……..TGUTAUAGGCNNNN… 5’
me
me
5. PCR amplification
5’ …NNNNTGGAGTTCGT……..ATGATGTTTGANNN… 3’
3’ …NNNNACCTCAAGCA……..TACTACAAACNNNN… 5’
5’ …NNNNCAAAACTCGT……..ACAATATCCGNNNN… 3’
3’ …NNNAGTTTTGAGCA……..TGTTATAGGCNNNN… 5’
6. Sequencing

Figure 2. An illustration of the RRBS protocol. Orange color highlights bases that are added at the end repair step and do not necessarily reflect the methylation status of the original sequence. The removal of end repair biases is described in section 2.3.1.5.

Coverage is the main limitation of whole-genome and reduced representation sequencing. Some genomic regions of interest might only be covered by a few reads, in which case the methylation proportion measurement uncertainty is high. A common practice is to perform technical validation by a targeted method. Targeted pyrosequencing was able to accurately quantify the true methylation proportion of a target sequence in a study that compared different DNA methylation assays (134). The genomic region of interest is first captured by a specifically designed assay, after which the fragments are bisulfite treated and PCR amplified. The number of PCR rounds is typically larger than in WGBS or RRBS, since the starting material is very small.

Pyrosequencing is a relatively old method (135) based on detecting DNA polymerase activity as light, emitted when ATP sulfurylase and luciferase act on pyrophosphate, which is produced only if a base is added to the template DNA of interest. Different types of bases are iteratively introduced to the template hybridized on a primer. The greater the number of added bases, the greater amount of pyrophosphate and the higher the light intensity. One limitation of this sequencing method is that the light intensity increases linearly only up to the addition of third base. If the DNA sequence of interest for example includes several thymines next to a CpG site, the methylation status of that CpG site cannot be reliably determined.

2.3 Data analysis strategies in transcriptomics and epigenomics

2.3.1 Technical biases and preprocessing

2.3.1.1 Preprocessing workflows for gene expression microarray data

The amount of hybridization of the target material at each probe is observed as fluorescence intensity and encoded as numeric values in either binary CEL-files or human-readable text files. One-color oligonucleotide microarray data preprocessing workflows typically include the following steps: 1) background correction to remove optical noise and non-specific signal 2) normalization 3) summarization of probe-level data to probeset-level or gene-level data 4) present/absent calling.

Affymetrix microarrays measure each transcript with a set of 16–20 probe pairs: each perfect match (PM) probe is paired with a mismatch (MM) probe, which is identical to the PM probe for all bases except one in the middle. The purpose of the mismatch probes was to quantify non-specific binding, which could be interpreted as background noise. Early expression measures were based on either the difference or the quotient between pairs of PM and MM intensity values. For example, Affymetrix's software MAS 5.0 summarizes each probeset with a robust average log(PM-MM) (136). However, spike-in experiments (performed with hybridization solutions of known concentrations of RNA fragments that perfectly match certain PM probes) showed that MM probes capture some signal, as well as noise (137).

The robust multi-array average (RMA) preprocessing workflow for Affymetrix microarray data includes the following steps (138): 1) Background correction, which utilizes the mode of all log-scale MM intensity values on the array. 2) Quantile normalization between arrays, followed by $\log_2$-transformation, which removes most of the correlation between intensity and variation (137). 3) The expression values μ_i are inferred for each transcript from the following linear model:

$$X_{ij} = \mu_i + a_j + \varepsilon_{ij}, \tag{1}$$

where X_{ij} is the background corrected normalized $\log_2$ transformed PM probe intensity value for probe j in sample i, μ_i is the sample-specific ($\log_2$) expression value, a_j is the probe-specific affinity and ε_{ij} is the i.i.d. (independent and identically distributed) noise. The fitting is done with Tukey's median polish method, which is robust to outliers.

The underlying assumption behind present/absent calling is that in the studied biological sample, only some genes are expressed, and some threshold value can distinguish between present and absent transcripts. Hebenstreit and others demonstrated that gene expression distributions in RNA sequencing experiments (at least in humans, mice and Drosophila) are bimodal and the two modes correspond to highly expressed and lowly expressed genes (139). For example, in mouse T helper 2 (Th2) cell RNA-seq data, the highly expressed genes corresponded to genes with activating histone marks (H3K9/14 acetylation, according to ChIP-seq data from the same Th2 samples) and included known Th2-specific transcripts, whereas the lower peak corresponded to transcripts lacking H3K9 acetylation and included transcripts specific to other cell types. They showed that the lower peak in the gene expression distribution is neither explained by technical noise nor by contamination with other cell types, but repressed genes are expressed in very small quantities. This gene expression bimodality is in line with later observations in single cells (140). The vast majority of human genes seem to be expressed in episodic bursts and can be assumed to be either in state OFF (low-level random transcription) or state ON (orders of magnitude higher expression than at the OFF-state) at a given time point (141–143).

In microarray data, expression distribution bimodality is often observed, although the lower peak is likely to include unspecific hybridization, as well as lowly expressed genes (139). A threshold value for present/absent transcripts can be determined for example by fitting a two-component Gaussian mixture model for the gene expression distribution (144,145). Other options include selecting an arbitrary threshold or estimating a threshold based on some negative and positive controls. Transcripts that are virtually absent in almost all samples (in both groups if the goal is to compare gene expression between two groups) are often excluded from further analysis.

2.3.1.2 Preprocessing of Affymetrix exon microarray data

Just as for gene expression microarray data, preprocessing steps for exon microarray data include background-correction, normalization, $\log_2$-transformation, and optionally summarization to exon-level expression values and filtering based on present/absent calls. The background correction is usually implemented by the manufacturer. For example in the case of Affymetrix human exon arrays, background noise is estimated based on pools of probes (different pools corresponding to different GC-contents) targeting such non-human genomic regions that are not expected to cross-hybridize with the human exon targeting probes (120).

Often the goal is to detect differentially spliced exons between two groups of samples, such as normal vs. tumor. At the (optional) present/absent call filtering step, exons are typically required to be expressed in at least one of the study groups within genes that are expressed in both study groups. To quantify alternative splicing, each exon expression level needs to be normalized to the expression level of the gene it belongs to (146). Since microarray measurements are noisy at the lower end of the expression range, limiting the study to highly expressed genes is likely to decrease the proportion of false positives among the detected differentially spliced exons (119). Other recommended filtering steps include removing probes with extremely low between-sample variation (which might indicate saturation) and excluding exons with very high (such as 5-fold) expression values compared to the median of other exons within the same gene and sample, since such outliers are likely to originate from cross-hybridization (120).

2.3.1.3 PCR biases in bisulfite sequencing

GC bias is a well-known phenomenon in sequencing data: sequences with a higher GC-content tend to have a lower coverage. Although the sequence content can affect several stages of the sequencing protocol, such as size selection and sequencing errors, library amplification by polymerase chain reaction (PCR) is considered the major mechanism behind the GC bias (147,148). This bias is especially relevant if the GC abundance correlates with the studied phenomenon, such as DNA methylation in bisulfite sequencing studies. After the bisulfite conversion, highly methylated regions have a higher GC-content than regions with low methylation levels and are consequently amplified less by the PCR.

A recent systematic analysis of technical biases (149) in DNA methylation data compared methylation levels estimated by bisulfite sequencing after different library preparation protocols to those quantified with LC/MS. The methylation percentage of heat-denaturated bisulfite converted PCR amplified DNA was over-estimated to be double compared to the true value (6 % vs. 3 %) at a lowly methylated genome, and most of this was ascribed to PCR amplification (hardly any over-estimation was observed with an amplification-free protocol in the same setting). The over-estimation was more modest in the context of higher methylation percentage (22 % amplified protocol vs. 15 % LC/MS). Another source of methylation percentage over-estimation is the DNA denaturation step. The decreased GC content due to the bisulfite treatment increases DNA degradation in high temperatures (150). However, according to Olova et al. (149) this only leads to a modest methylation percentage over-estimation (3.3 % vs. 3 %).

The amount of PCR duplication can obviously be decreased by decreasing the number of PCR rounds needed. This is relevant when the amount of starting material is considered. Fragment size selection is important, since there is a well-known inverse correlation between fragment size and amplification (151). If the variation of fragment size is high, shorter sequences will be overrepresented in the data.

PCR biases are especially difficult to take into account in reduced representation bisulfite sequencing (RRBS) data analysis. In the context of other types of sequencing data, a common practice is to exclude reads that map to exactly the same positions and are therefore more likely to arise from PCR duplication than to originate from different fragments (152). This is not possible for RRBS data, which includes identical fragments due to the Msp1 enzyme digestion. In order to exclude most PCR duplicates, we and others have removed CpG sites with extreme coverages. Most CpG sites are covered with 0–200 reads but a small fraction can have extreme coverages, up to approx. 100 000 reads. The authors of MethylKit have recommended removing CpG sites with coverage above the 99.9th percentile in each sample (153).

2.3.1.4 Preprocessing and alignment of high-throughput sequencing data

Sequencing read data is stored in the FASTQ format, which includes the sequence identifier, the sequence, and the quality scores for each base in ASCII format (only one character for each score). A common practice is to trim the reads before aligning them on the genome of interest. Trimming by quality scores has been shown to markedly increase the total number of aligned reads and the concordance between different alignment tools (154). The following steps are common to most preprocessing workflows of bisulfite sequencing and other types of high-throughput sequencing data:

- Reads are trimmed based on the quality scores. In protocols such as Illumina reversible terminator sequencing, the probability of sequencing errors increases towards the 3' end of the read. Quality scores are generated during the sequencing based on properties such as light intensity profiles at each sequence cluster (each read). The quality of each base within each read is typically reported as a Phred score, which is $-10\log_{10}$(error probability).
- Adapter sequences are removed.
- After the above steps, reads are excluded based on a length cutoff, such as 20 bp.

Adapter sequence and quality trimming are especially important in genetic variant and DNA methylation studies, where the exact base at each position matters. The above-mentioned trimming steps are less critical in gene expression studies (RNA sequencing), where the purpose is to quantify the RNA, in which case adapter sequence contamination and sequencing errors are harmful only if they lead to an alignment failure. A good practice is to observe quality control plots before and after trimming the reads. Quality control tools such a fastQC (155) are useful to detect e.g. adapter contamination or other sequence overrepresentation.

After read trimming and quality control steps, the next step is the alignment of sequenced reads to a reference genome. Dozens of sequencing alignment algorithms have been developed, only few of which are being actively used by the scientific community (156). Bowtie (157) and BWA (158) and their successors (such as Bowtie2 and BWA-SW) remain among the most popular short read alignment tools probably due to their efficiency and active maintenance (156).

2.3.1.5 Bismark workflow for the preprocessing and alignment of reduced representation bisulfite sequencing data

Please refer to Figure 4, Section 4.10. for an example RRBS data analysis workflow. The Bismark workflow for RRBS data starts with observing quality control plots and performing at least the above mentioned trimming steps, which are implemented for RRBS data in the Trim Galore tool (133):

- Quality trimming: The default setting in Trim Galore is to exclude bases with Phred scores below 20 (sequencing error probabilities above 1 %) at the 3' end of each read.
- Any remaining sequence that does not originate from the DNA fragment of interest could potentially lead to alignment failure and errors in methylation calls. By default, the Bismark workflow trims any adapter sequence overlap from the 3' ends of reads. In the case of Illumina adapter AGATCGGAAGAGC, even a single A observed at the 3' end of the read (overlapping the adapter sequence by one base) is removed.

In addition to the above-mentioned steps, the removal of end-repair biases is important for RRBS data. End-repair is the addition of CG to the 3' ends of the DNA fragments after the digestion with the MspI enzyme, which cuts each CCGG site, such that CGG remains at the 5' end of each fragment but only one C remains at the 3' end (Figure 2). The methylation statuses of the cytosines filled in during the end-repair step do not represent the true methylation statuses of the cytosines in the original sequence. The Bismark workflow removes this bias by excluding additional 2 bases from 3' ends of sequences that were adapter-trimmed and 2 bases from the 5' end of read 2 in the context of paired-end sequencing (133).

Alignment algorithms for bisulfite sequencing reads need to account for the conversion of unmethylated cytosines to thymines, which are observed as guanines transformed to adenines on the complementary strand. Bismark creates fully C-to-T-converted and fully A-to-G-converted versions of each read and applies Bowtie2 (159) to align them to similarly converted versions of the genome (160). This obviously requires 4 times the computational resources needed for Bowtie2 alignment in non-bisulfite context, since both versions of each read need to be aligned to both versions of each genome.

Sun and others evaluated the performances of different bisulfite sequencing read alignment methods with respect to their ability to detect CpG sites and to accurately estimate methylation proportions in simulated data (161). The best-performing methods Bismark (160), BS-Seeker2-bowtie2 (162), GSNAP (163) and bwa-meth (164) were almost identical with respect to these criteria but Bismark and GSNAP were slower than BS-Seeker2-Bowtie2 and bwa-meth (161). Although several more recent alignment tools have been developed after Bismark, the main improvements have been made in computing time, and comparisons have revealed very little difference between the tools with respect to the end results (165).

The final step in the Bismark workflow is to extract the numbers of methylated and unmethylated reads at each cytosine (or more commonly just each cytosine in the CpG context). This is a relatively fast and trivial task once the alignment has been completed. Each uniquely aligned read is compared to the corresponding reference genome sequence. An important issue to consider at this step is the possible overlap of pairs of reads in paired-end sequencing data. Such overlaps are especially common in RRBS data, where the fragment lengths are size selected to be between 40–220 bp, and typically at least half of the selected fragments are shorter than 100 bp in the MspI digested human genome (127,130). By default, Bismark methylation extractor excludes any sequence from read 2 that overlaps with read 1 to avoid redundant methylation calls. The rationale behind keeping read 1 and excluding (part of) read 2 in the event of overlaps is the higher sequencing error rate at read 2 in Illumina sequencing (132).

2.3.1.6 Estimation of bisulfite conversion efficiency

Bisulfite conversion efficiency is usually estimated by spiking in some DNA that is known to be completely unmethylated and can be easily distinguished from the genome of interest (166). For example in Studies IV and V, fully unmethylated lambda phage DNA was added to each human DNA sample and bisulfite conversion efficiency was estimated as the proportion of lambda phage cytosines that were read as thymines. Additionally, since DNA methylation occurs almost exclusively in the CpG context in mammalian genomes (72,73), the proportion of converted cytosines in CHH (C followed by any two bases other than G) and CHG (C followed by any base other than G, followed by G) contexts can be utilized as another estimate of the bisulfite conversion efficiency. Samples with low (for example < 98 %) conversion efficiency need to be excluded, unless the differential methylation analysis corrects for it (167).

2.3.1.7 M-biases in bisulfite sequencing data

For quality control purposes, the Bismark methylation extractor quantifies average methylation percentages per sequencing read position, separately for CpG, CHG and CHH contexts. Ideally, at a given context (such as CpG) the average methylation percentage should be identical across read positions; hence any deviations are called M-biases. Since M-biases are typically enriched at the ends of sequencing reads, the developers of the quality control tool BSeQC recommend quantifying normal variation based on middle positions and trimming away 3’ and 5’ end bases that exceed it (168). In RRBS data this normal variation is much larger than in WGBS, which might be explained by the presence of some transcripts with extremely high coverages in RRBS (see section 2.3.1.3 on PCR biases). Furthermore, RRBS reads should not be trimmed at the 5’ end, which always starts with a CpG site (127). Average methylation at the RRBS 5’ CpG is usually higher than at other read positions. CpG sites at middle positions are more likely to originate from CpG islands than those in the 5’ end. CpG islands are characterized by hypomethylation (169).

2.3.1.8 SNP detection in bisulfite sequencing data

A comparative population study revealed that the majority of CpG sites with differential methylation between populations contained common SNPs (single nucleotide polymorphisms) with different allele frequencies in the studied populations (170). This suggests that many of CpG sites that were observed as differentially methylated, were probably not differentially methylated. Since bisulfite treatment converts unmethylated C to U (read as T), a C to T SNP can be misinterpreted as a completely unmethylated CpG site. The removal of SNPs from bisulfite sequencing data is therefore an important preprocessing step. Most bisulfite sequencing protocols, are strand-specific (171). That is, only cytosine gets converted to uracil, whereas guanine on the opposite strand is unaffected. If a T is observed opposite to G, the T is most likely a result of the bisulfite conversion and not a SNP.

SNP detection software such as Bis-SNP (171) are based on Bayesian inference: The probability of each possible underlying genotype, given the observed reads, is proportional to the prior probability of each genotype (genotype frequencies in the population) and the likelihood of observing the reads, given each genotype. Bisulfite conversion efficiency and base calling error rates are taken into account in the likelihood term. BS-SNPer (172) is essentially a slightly simpler and more efficient version of the Bis-SNP algorithm for the detection of SNPs in bisulfite sequencing data. It does not, for example, perform base quality score recalibration, which is an important step in the Bis-SNP algorithm.

2.3.2 Confounding effects

When multiple explanatory variables affect the response variable, their effects can confound each other, and failing to deal with this issue can lead to both false negative and false positive findings. The greater the number of independent samples in a study, the smaller the likelihood of an unknown confounding factor being unbalanced with respect to the variables of interest. Since the number of available samples is often very limited in biological studies, the following four aspects need careful consideration:

1. Balanced study design: The effects of two mutually correlated explanatory variables are hard to tell apart. Studies need to be designed such that potential confounding factors are either balanced with respect to the variable(s) of interest (for example equal proportions of males and females selected for the patient group and the control group) or avoided (for example only males within age range 20–25 included in a study). Technical variables, such as sample preparation batches, deserve special attention in 'omics studies. Since the number of simultaneously processed samples is limited, matched study designs can be useful, even if the downstream statistical inference is unmatched. For example, each case individual might be matched with a control individual with similar characteristics (such as sex and time of birth). If matched pairs (or small groups) of samples are processed within the same batch throughout the study, the case and control groups are likely to be comparable even in the presence of small batches. Some algorithms have been published for the purpose of multivariate study design optimization (173,174).
2. Randomization is applicable to intervention studies. It is the process of, for example, using a random number generator to allocate individuals to drug and placebo groups in clinical trials. Compared to arbitrary human decisions, a random allocation procedure reduces the risk of confounding effects but is not sufficient in studies with small sample numbers or batch sizes. Matched study designs have been found to outperform unmatched randomized study designs, as measured with sensitivity and specificity to detect true effects in simulated data, even if the matching is done based on several irrelevant covariates in addition to some relevant ones (175). However, it must be noted that randomization and matching are by no means mutually exclusive options. In randomized trials, the random allocation to study groups can be done within each pair/group of matched individuals. Such an approach has been taken for example in the above-mentioned non-bipartite matching algorithm (174).

3. Blinding is an important principle in both intervention and observational studies: the investigator should not know the study group the sample/individual belongs to, while processing the samples or interacting with the study participants (if applicable). Just as randomization, also blinding can be combined with matching. For example, a laboratory technician might be instructed to process samples 1A and 1B together, 2A and 2B together, 3A and 3B together etc. but the labeling of cases and controls as ”A” or ”B” might have been decided with a coin flip within each pair, such that the technician is blind to the sample groups.
4. Even if the study design is carefully optimized, the explanatory variable of interest rarely has zero correlation with other explanatory variables. Two alternative approaches can be taken to deal with potential confounding effects in the data analysis phase: 1) the potentially confounding covariates can be included in the model or 2) the data can be adjusted with respect to some variable(s) before the modeling. Examples of these approaches are described in sections 2.3.2.1 (approach 1) and 2.3.2.2 (approach 2).

2.3.2.1 Multiple linear regression

In multiple linear regression, a response variable (such as the expression value of a gene) is modeled as a linear combination of multiple explanatory variables, which can be binary or continuous (such as treatment group, sex, body mass index and technical batch):

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{y}$ is the dependent variable (a vector of length n, n being the number of observations), X is the design matrix of dimensions n × p including a column for the intercept and p-1 explanatory variables, $\boldsymbol{\beta}$ is a p-dimensional vector for the coefficients to be estimated from the data, and $\boldsymbol{\varepsilon}$ is the error term. If the distribution of the i.i.d. error ε_i (for observations i=1,2,...,n) is Gaussian, the model is an ordinary linear regression model, for which the maximum likelihood solution can be computed in a closed form. Since the model is additive, each estimated coefficient $\hat{\beta}_j$ of covariate j represents the effect of $\mathbf{x}_j$ on $\mathbf{y}$ that is observed, while the (estimated) effects of other covariates are cancelled out.

2.3.2.2 Batch effect adjustment methods

If including a covariate in the model is for some reason not possible, one might try to adjust the data to remove its effects before the modeling. This approach has traditionally been taken to account for batch effects in ’omics studies. The most naïve approach is to zero-center the data within each batch. As one can easily imagine, this can lead to both removing true effects and adding false effects to the data, since other covariates are not taken into account. A slightly more sophisticated approach is to estimate the batch effect from a linear model, which includes at least one covariate effect of interest, in addition to the batch effect, and then adjust for the batch effect:

$$\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \mathrm{H}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{3}$$

$$\mathbf{y}_{\text{adjusted}} = \mathbf{y} - \mathrm{H}\hat{\boldsymbol{\gamma}}\,, \tag{4}$$

where $\mathbf{y}$ is the n-dimensional vector of e.g. gene expression values for one gene (for samples i=1,2,...,n), X is the design-matrix for some covariate(s) of interest, including a column for the intercept term, $\boldsymbol{\beta}$ is a vector of coefficients for the covariates in X, H is the design-matrix for the batches, γ is a vector of coefficients for the batches, $\boldsymbol{\varepsilon}$ is an n-dimensional vector of error terms, $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2)$, and $\hat{\boldsymbol{\gamma}}$ are the estimated batch effects. Often the models also include a multiplicative batch effect ("scale effect"), in addition to the additive effect γ ("location effect") (176,177).

However, as Nygaard and others point out, the batch-adjusted data is not free from batch effects (178). The original batch effect has just been replaced by a batch-specific estimation error $\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}$:

$$\mathbf{y}_{\text{adjusted}} = \mathrm{X}\boldsymbol{\beta} + \mathrm{H}\boldsymbol{\gamma} - \mathrm{H}\hat{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}\,. \tag{5}$$

In case the covariate of interest correlates with the batches, the conclusions on the covariate's effect on $\mathbf{y}_{\text{adjusted}}$ can be heavily influenced by the batch effect estimation error (unless the batch effects are again included in the model). The most popular batch effect correction method is ComBat, which has been cited more than 2000 times in 13 years and is still being actively used by the scientific community (90,179). It uses an empirical Bayes procedure to estimate the batch effects (176). This procedure, which utilizes information across genes to obtain robust batch effect estimates, can be expected to decrease the estimation error, as discussed in Section 2.3.3.2. However, in practice even ComBat has been discovered to increase the type 1 error rate in study settings, where the covariate of interest correlates with the batches (178).

2.3.3 Strategies to address high dimensionality

In 'omics studies the sample numbers are often very small compared to the number of features measured from each sample. For example, the number of features in a typical gene expression study is approximately 20000 (genes), while the number of independent samples is rarely more than 100, often just a handful. The multiple testing problem is introduced in Section 2.3.3.1, some strategies to address the challenge of inference with a limited number of observations are outlined in Sections 2.3.3.2 and 2.3.3.3, and Section 2.3.3.4 discusses some strategies to group biological features to larger functional entities.

2.3.3.1 Multiple testing correction in ’omics studies

The typical key question in ’omics studies is, which features, such as proteins or genomic locations, are important for a given phenomenon, such as a disease. If the association of each feature is tested separately, multiple testing correction becomes vital. For example, the expected number of nominal p-values that fall under threshold 0.05 is 1000 out of 20000 independent tests, if the data is random noise modeled with the correct distribution. The most commonly used multiple-testing correction methods are the Bonferroni correction (180) and Benjamini-Hochberg correction (181). Bonferroni correction multiplies the p-value by the number of tests to estimate the family-wise error rate (FWER), or equivalently divides the significance threshold with the number of tests. Benjamini-Hochberg correction divides the p-value by its rank and multiplies it by the total number of tests to compute the false discovery rate (FDR) i.e. the expected proportion of false discoveries among the discoveries at a given p-value threshold.

Genome-wide significance (GWS) is a concept originally developed in genome-wide association studies (GWAS), where genotype-phenotype-associations are investigated across the genome, typically including data from thousands of individuals. GWS is a Bonferroni-corrected significance threshold for an estimated number of independent tests in GWAS. The scientific community has set the threshold at $p \leq 0.5 \times 10^{-8}$, based on an estimated number of 1 million common SNPs that are independent of each other (not in linkage disequilibrium) in the European human genome (182). The estimated GWS thresholds based on different approaches have been remarkably close to each other, and genotype-phenotype-associations reaching it ($p \leq 0.5 \times 10^{-8}$) have been highly reproducible (182,183). A relaxation of this threshold has been suggested based on the observation that a large proportion of genotype-phenotype-associations with p-values close to GWS (between 0.5×10^{-8} and 10^{-7}) have been reproduced i.e. the threshold $p \leq 0.5 \times 10^{-8}$ has been reached by combining more data with the originally published data (184).

The rationale behind estimating a GWS threshold for a population is to achieve fair FWER control based on the total number of independent hypotheses, instead of the limited number of hypotheses that can be tested with a given technology. Since the number of tested associations varies from study to study, reporting the nominal (uncorrected) p-values and using a standardized significance threshold improves the comparability of results from different studies. A similar approach has later been adopted in epigenome-wide association studies (185).

FWER control is an extremely conservative approach to multiple testing correction in biological studies. Since the effect sizes are typically small, genome-wide significance can only be reached with a large number of samples. If the number of independent samples is less than 1000, typically the goal is to generate rather than to test hypotheses. The common practice is to report all potential associations with a relaxed threshold for the nominal p-value, such as 10^{-5}, required by the GWAS Catalog (186). Reported potential associations that have not reached genome-wide significance are useful as hypotheses that remain to be tested.

2.3.3.2 Bayesian and empirical Bayes methods

Bayesian approaches have some important advantages for studies with modest numbers of independent observations. The goal in Bayesian inference is to update prior knowledge of the studied phenomenon based on available data. Instead of maximum likelihood estimates, posterior probability distributions are evaluated for the parameters. If very little data is available to estimate a parameter value, this is reflected in the posterior variance; and outliers are not as likely to dominate the conclusions as in maximum likelihood inference. Furthermore, useful prior knowledge might be incorporated in the model before seeing the data. For example, if the goal is to infer the average weight gain in a group of adults within 1 month of starting a new antidepressant, a slightly informative prior, such as $N(\mu = 0, \sigma=20 \text{ kg})$ could be chosen to limit the search space to humanly possible values, while making no prior assumptions on the direction of change.

In contrast to standard Bayesian methods, empirical Bayes approaches utilize data to estimate the parameters of the prior distribution (hyperparameters). The difference between empirical Bayes and Bayesian hierarchical models is that empirical Bayes methods find maximum likelihood estimates for the hyperparameters, whereas Bayesian hierarchical models evaluate their distributions.

In bioinformatics, the most widely-adopted empirical Bayes method is implemented in the R package limma (187), originally developed for gene expression microarray studies and useful for any high-dimensional data with Gaussian noise. Limma has also become popular in RNA sequencing analysis combined with a transformation, such as voom (188), to meet the Gaussian noise assumption. Combined with voom, limma outperformed many negative binomial models in RNA sequencing data analysis in the context of modest (< 20) sample numbers (189). In another differential methylation method comparison study for RNA sequencing data, it was more robust to outliers than any other parametric model evaluated (190).

Limma is a linear model with an empirical Bayes technique to estimate the residual variances for the t- and F-statistics. The prior for the variances is estimated from all genes on the microarray. The moderated residual variance estimates for gene g become:

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}, \tag{6}$$

where s_g^2 is the residual sample variance of gene g, s_0^2 is close to the mean of all s_g^2, and d_0 and d_g are the corresponding degrees of freedom, reflecting the numbers of independent observations available to estimate s_g^2 and s_0^2 (187). If the number of samples is small (such as 2–3 per group, which was common in 2004), the variance estimates for a given gene are heavily shrunk towards the common mean of all genes.

A similar approach has later been applied to estimate the dispersion parameter of a negative-binomial model for RNA sequencing data, implemented in the R package edgeR, first for two-group comparisons (191) and later as a more flexible generalized linear model (192). The main criticism towards this approach is that in case the data contains outliers at gene g, the variance of gene g will be underestimated, since it is shrunk towards the common mean. One strategy to address this problem is to downweigh outliers during the iterative maximum likelihood estimation of parameter values. Such an extension to the GLM version of edgeR has been implemented in the R package edgeR-robust (193).

2.3.3.3 Rank-based inference

Rank-based strategies are nonparametric methods that avoid making assumptions about the distribution of the data. Instead of estimating distribution parameters, the test statistic is computed based on ranks and the significance is typically estimated with a permutation test (unless the test statistic follows a known distribution). Rank-based methods can outperform parametric methods if the data does not follow the distribution assumptions of the parametric models. This has been demonstrated for example in RNA-seq data, where extreme outliers are relatively common for both technical and biological reasons. The introduction of even a single outlier to simulated RNA-seq data sets heavily increased the type 1 error rates of some negative binomial models, such as edgeR, in the context of small sample numbers ($n \leq 10$), whereas the non-parametric SAMseq was unaffected (190). SAMseq performs a permutation test on a slightly modified Wilcoxon rank sum test statistic after a resampling procedure to normalize the expression values to account for differences in sample coverages (194).

Rank product (195) is a classical nonparametric method for two-group comparisons in gene expression microarray studies. For paired study designs, it ranks genes by pairwise fold changes, normalizes each rank by the total number of genes measured for the pair of samples, and calculates the product of these normalized ranks *RPg* for each gene (or some other feature) *g*:

$$RP_g = \prod_{i=1}^{k} r_{ig}/n_i \, , \qquad (7)$$

where r_{ig} is the rank of gene g within the i^{th} pair of samples and n_i is the total number of genes measured for the i^{th} pair of samples. For unpaired study designs, the procedure is the same, except the rank products are computed over the ranks of all pairwise fold changes between the cases and controls (or some other study groups). RP_g is the probability of the observed or better ranks, assuming the null hypothesis was true for all genes, and is hence akin to a nominal p-value. The false discovery rate is estimated with a permutation test.

Rank product was developed during the time, when two-group comparisons in microarray studies were typically based on either Welch's t-tests or just observed fold changes without any significance estimation. Estimating within-group variation based on 2–5 biological replicates or ignoring statistical significance altogether are both obviously problematic approaches. T-tests call small between-group differences significant, as long as the within-group differences are even smaller. The authors of rank product argue that small gene expression differences, such as 1.1-fold, are rarely biologically relevant, even if they are statistically significant (195). Furthermore, the within-group variation can easily be underestimated, given the limited sample numbers. The empirical Bayes approach limma (187), which was published around the same time with rank product, successfully addressed this problem by utilizing the parallel nature of microarray studies for more stable variance estimation (see Section 2.3.3.2). Rank product avoids the need to estimate variation altogether. Compared to other methods for differential expression detection, available during the microarray era (including limma), rank product has been observed to perform especially well if sample numbers are small (196,197).

Other rank based approaches in bioinformatics include various forms of transforming the data to ranks, plugging them in analysis of variance formulae and evaluating the significance of the test statistic either based on a known distribution under asymptotics or based on a permutation test, in case the number of observations is small (198). Top scoring pairs (k-TSP) (199,200) is a simple and widely-used nonparametric classification method, which has been useful as such and combined with other machine learning algorithms (201). Top scoring pairs is based on identifying pairs of features with consistent relative ranks in one group and the opposite pattern in the other. A more recent rank-based personalized medicine tool for cancer studies is based on a similar idea: Pairs of genes with stable relative ranks are identified in normal samples (minimum 99 % consistency is considered stable), after which genes or pathways with opposite patterns are identified for each diseased individual (202).

The main limitation of rank-based methods, such as the ones described above, is that they are less powerful compared to parametric alternatives, in case the data meets the distribution assumptions (194,198). Also, permutation tests can be time-consuming, and huge numbers of permutations are needed, if the goal is to reliably estimate small p-values (203). In the presence of confounding covariates, special care needs to be taken to meet the exchangeability assumption of permutation tests. That is, the empirical null distribution should reflect the heteroscedasticity structure of the original data. If an explanatory variable is permuted, its joint distribution with other explanatory variables should remain unchanged. In some cases this can be achieved with simple block-wise permutations (for example permuting case- and control-labels of individuals instead of samples, when several samples have been collected from each individual), and several strategies have been implemented for permutation testing in the presence of more than one explanatory variable (204). In practice, tools with permutation-based significance testing are often limited with respect to their ability to deal with covariate effects.

2.3.3.4 Gene set analysis

Features in 'omics studies can be grouped to functional entities, such as signaling pathways or co-expressed genes. Observations on individual features in biological high-throughput data are rarely informative unless other related features show similar patterns. If associations between features and explanatory variables are first evaluated on the level of individual features, enrichment of functional entities (e.g. pathways or gene ontologies) on the list of selected features (e.g. differentially expressed genes) is typically evaluated through Fisher's exact test. Alternative strategies achieve better statistical power by directly modeling this higher level instead of individual features. Some methods that focus on differentially methylated genomic regions instead of individual CpG sites are discussed in Section 2.3.4.3, while this section focuses on higher-level functional grouping of genes.

Most functional annotations of genes are based on evidence from scientific literature that is either automatically searched, collected from authors, and/or expert-curated. For example, Gene Ontologies (205,206) are a large scientific consortium project that aims to provide the latest scientific consensus knowledge on gene functionalities. Literature-based annotations have the obvious limitation that literature is not available for all genes. GeneNetwork.nl (207) addresses this issue by providing gene annotations that are based on principal component analysis (PCA) on publicly available RNA-seq data from more than 31 000 samples. They computed enrichment scores of functionally annotated gene sets from public databases, such as KEGG pathways and human phenotype ontology (HPO) terms, for each eigenvector. The association of any gene to terms/pathways can be estimated based on the correlation between the gene's eigenvector coefficients and the eigenvector enrichment scores for each term/pathway.

Chaussabel and others developed a functional annotation resource for the field of immunology by grouping consistently co-expressed genes to modules that were then annotated by automatic literature search (208). Specifically, they used Affymetrix gene expression microarrays to analyze peripheral blood mononuclear cells (PBMCs) from 239 individuals with one of eight different diseases, such as systemic lupus erythematosus, influenza A or melanoma. K-means clustering of gene expression profiles was done separately within each subset of samples representing different diseases, and a module was defined as a group of genes that clustered together in at least 6 out of 8 diseases. The modules were functionally annotated through an automatic literature search (209) on publication abstracts from the Medline database. Altogether 14 out of 28 modules were annotated to cell types or immune processes, based on frequent co-appearance of terms and gene names, normalized to the overall frequency of each term and the total number of abstracts containing each gene name.

Strategies that directly model associations between explanatory variables and sets of features include Gene Set Enrichment Analysis (GSEA) (210), which was developed in the early microarray era but is still being actively used by the scientific community. The enrichment score in GSEA is a Kolmogorov-Smirnov-like statistic based on the overrepresentation of features belonging to a pathway (or some other gene set of interest) on the extremes of a ranked list of features. The ranking can be based on any metric that represents the features' association to the biological phenomenon of interest, for example the average expression difference between two groups of samples. The basic principles of GSEA have been further developed by several more recent gene set analysis strategies (211,212).

2.3.4 Models for spatially correlated count data

Sequencing technology quantifies DNA/RNA sequences in read counts. In the case of bisulfite sequencing, DNA methylation at each cytosine is quantified as numbers of methylated reads (cytosines) and numbers of unmethylated reads (thymines). The total number of methylated and unmethylated reads is referred to as coverage.

Flexibility and the availability of a closed-form maximum likelihood solution are attractive properties of ordinary linear models with Gaussian noise assumptions. Although such models can not directly be applied to sequencing count data, they have been used together with some simple variance stabilizing transformations. Examples include the log transformation for RNA-seq counts and the logit transformation for methylation proportions. Unlike gene expression microarray data, sequencing count data meet the Gaussian noise assumption poorly even after such transformations. For both technical and biological reasons, bisulfite sequencing and RNA-seq data are zero-inflated (213–215).

Moreover, such strategies ignore the technical variation at each feature. If for example bisulfite sequencing counts are transformed to percentages, the coverage information is lost. That is, 10 out of 50 reads methylated at some CpG site will be observed as being no different from 1 out of 5 reads methylated (20 % methylation in both cases), although the technical uncertainty is much greater in the latter case. Some models address this issue by including coverage-weights in the linear modeling strategy (188,216). The other option is to directly model the data generating process with a suitable distribution. Poisson and negative binomial models are applicable to RNA sequencing data (191,217–220). Sections 2.3.4.1 and 2.3.4.2 introduce beta-binomial and binomial mixed effect models for bisulfite sequencing data. Section 2.3.4.3 discusses the spatial correlation issue.

2.3.4.1 Beta-binomial regression

The technical variation in bisulfite sequencing data is most naturally captured with a model, where the number of methylated reads y_{ij} in each sample i and cytosine j is binomially distributed with parameters π_{ij} and n_{ij}, where π_{ij} is the true underlying methylation proportion and n_{ij} is the total number of reads. The between-sample (biological) variation of the methylation proportion can then be modeled with a beta distribution, which is a conjugate prior of the binomial likelihood. The methylation proportion is modeled as a regression problem through a link function, such as logit. Maximum likelihood estimation of regression coefficients requires iterative optimization, but is relatively straightforward, due to the tractability of the beta-binomial distribution.

Beta-binomial regression was first applied to bisulfite sequencing data already in 2011 (221) and later implemented as differential methylation analysis tools, such as RADMeth (222), DSS (223,224), and MOABS (225). The beta-binomial model RADMeth has outperformed other types of approaches, such as ordinary linear models, Poisson regression and logistic regression (167,226,227)

2.3.4.2 Generalized linear mixed effect models for bisulfite sequencing data

The main limitation in modeling the between-sample variation with a beta distribution is that only binary covariate effects can be included. To increase flexibility, binomial mixed effect models have been developed for the detection of differential methylation (167,220,227). They can include both binary and continuous fixed covariate effects and model arbitrary covariance structures as random effects. A generalized linear mixed effect model (GLMM) for bisulfite sequencing data is formulated for one CpG site as:

$$y_i \sim Bin(n_i, \pi_i), \tag{8}$$

$$log \frac{\boldsymbol{\pi_i}}{\mathbf{1}-\boldsymbol{\pi_i}} = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{z}_i^{\mathrm{T}} \mathbf{u} + \mathrm{e_i}, \tag{9}$$

where y_i is the number of methylated reads for sample i, n_i is the total number of reads, and π_i is the underlying methylaton proportion, which is modeled through the logit link function as a sum of linear mixed effects and Gaussian noise. Here $\boldsymbol{x_i}$ is the design vector (for sample i) of length p, p is the number of covariates modeled as fixed effects, $\boldsymbol{\beta}$ is the vector of fixed effect coefficients (length p), $\boldsymbol{z_i}$ is the design vector of random effects of length k, $\mathbf{u}$ is the vector of random effect coefficients of length k, and $\mathrm{e_i}$ is the residual. The random effects and residuals are both normally distributed with mean 0. The residuals are independent and identically distributed, whereas the random effects $\mathbf{u}$ can have any covariance structure.

The logit link function is optimal for interpretability of binomial regression models: each model coefficient represents the increase in log odds when the explanatory variable value increases with 1 unit (if other variables were kept constant). Extreme values 0 and 1 are problematic in this context, since they are infinite (–∞ and ∞, respectively) in the logit-transformed space. The convergence of such GLMs can be markedly improved by adding 1 to the observed numbers of successes (methylated reads) and 2 to the numbers of trials (total number of reads), as recommended for example by the authors of PQLseq (220). Similar pseudo-count transformations are commonly used for other types of count data, such as RNA sequencing reads or microbial abundance counts (228,229).

The flexibility of GLMMs comes at the cost of having to deal with intractable marginal likelihoods. Unlike linear (Gaussian noise) models, GLMs do not have closed form solutions, and unlike GLMs, GLMMs cannot be solved with relatively simple iteratively reweighted least squares (IRLS) techniques. The options include approximating the posterior distribution with techniques such as Markov Chain Monte Carlo (MCMC) sampling or finding maximum likelihood estimates with iterative techniques tailored for GLMMs. Of the above-mentioned methods for differential methylation detection, MACAU (227) implemented a Gibb's sampler to approximate the posterior distribution, even though the goal was to find maximum likelihood estimates and standard errors for the regression coefficients to compute approximate (frequentist) Wald test p-values. LuxUS (167) is a Bayesian method that used Hamiltonian Monte Carlo sampling to approximate the posterior distributions of the model parameters and to test hypotheses by using Bayes factors. PQLseq (220) used iterative optimization to find maximum quasi-likelihood estimates for the coefficients.

Briefly, the optimization algorithm implemented in PQLseq maximizes a Laplace-approximated joint quasi-likelihood (quasi-, since it is not an actual probability distribution that would integrate to 1) with an expectation maximization

(EM) algorithm that iterates between updating pseudo-data and updating parameters until convergence. At each iteration, pseudo-data is generated based on the current estimates of the parameters using a second order Taylor expansion. The next parameter estimates can then be inferred using a standard algorithm designed for linear mixed effect models. Just as in the case of linear mixed effect model fitting (230), this process alternates between updating the fixed and the random effect parameters.

2.3.4.3 Strategies to account for spatial correlation in bisulfite sequencing data

Methylation-determining regulatory elements can establish and maintain the hypomethylated state of regions, such as promoters (231). Therefore it is not surprising that strong spatial correlation between methylation statuses of CpG sites have been observed up to a distance of 1–2 kb (77). Accounting for the spatial correlation between methylation statuses of CpG sites and studying differential methylation on the level of regions (rather than individual CpG sites) is of interest in DNA methylation array studies (232), and even more important in bisulfite sequencing studies, where the resolution is higher.

Different strategies have been developed to account for the spatial correlation either 1) before and/or during, or 2) after the model fitting and hypothesis testing.

Type 1) tools, include for example dmrseq (216), which first combines CpG sites to candidate differentially methylated regions (DMRs) by identifying adjacent sites, whose methylation difference is to the same direction (with a liberal cutoff), and fits a generalized least squares model including a region-level spatial correlation structure. BiSeq (233) utilizes information from adjacent CpG sites to improve the stability of methylation proportion estimates through kernel smoothing, performs a naïve t-test at each individual CpG site and combines differentially methylated CpG sites (DMCs) to DMRs based on a t-statistic cutoff. LuxUS (167) fits a GLMM on pre-defined candidate regions and models the spatial correlation between CpG sites directly through random effects. A pre-analysis method is implemented for candidate region formation, based on preliminary CpG-level differential methylation analysis and other criteria, such as a 2 kb maximum window length.

Type 2) approaches perform hypothesis testing on the level of individual CpG sites first, and adjust p-values for their spatial correlation afterwards for example with an autocorrelation-adjusted Z-test (222,234) or a sliding linear model SLIM (235), which is the default multiple testing correction method in MethylKit (153). The autocorrelation-adjusted Z-test is also known as a Stouffer-Lipták-Kechris correction (234), which is described and discussed in Study V. Briefly, it performs a sliding-window meta-p-value analysis, where it computes the weighted sum of probit-transformed p-values (up to a user-defined distance) and corrects for the lack of independence between the p-values by adjusting for their autocorrelation. The autocorrelation is estimated beforehand through a sliding window analysis across the genome.

3 Aims

The aim of this thesis was to contribute to a better understanding of early-life risk factors of type 1 diabetes and to improve existing analysis workflows in transcriptomic and epigenomic studies. The more specific goals were:

I. Evaluation of methods for the detection of differential exon inclusion/exclusion events between two groups in Affymetrix exon array data

II. Comparison of early longitudinal gene expression patterns between children who later develop type 1 diabetes and control children who remain healthy

III. Exploration of transcriptomic differences between neonates born in Finland, Estonia, and Russian Karelia with contrasting standards of living

IV. Comparison of DNA methylation patterns between neonates who later develop type 1 diabetes and control children who remain healthy

V. Empirical estimation of the false discovery rate of common workflows for the detection of differentially methylated regions in bisulfite sequencing data

4 Materials and Methods

4.1 Detection of differential splicing between two groups of samples (I)

Study I compared methods for the detection of differential splicing in publicly available Affymetrix Human ST 1.0 Exon Array data, where each exon is targeted with a set of probes (usually 4 probes per exon). Splicing index (SI) is a simple model for comparing exon inclusion between two samples (146). Each exon-level expression value is first normalized to the expression of the gene it belongs to. Splicing index is the log-ratio of these normalized exon-level values between two samples:

$$SI_{uveg} = (\mu_{ue} - \mu_{ug}) - (\mu_{ve} - \mu_{vg}), \tag{10}$$

where SI_{uveg} is the splicing index of exon e within gene g between samples u and v, μ_{ue} and μ_{ug} are the normalized $\log_2$-transformed exon and gene expression estimates of sample u, and μ_{ve} and μ_{vg} are those of sample v. The exon-level and gene-level summary expression values can be calculated with a procedure such RMA (138), as described in Section 2.3.1.1. We developed a probe-level splicing index:

$$SI_{uvk} = x_{uk} - x_{vk} - \mu_{uvg}, \tag{11}$$

where SI_{uvk} is the splicing index of probe k (within gene g) between samples u and v, x_{uk} and x_{vk} are probe intensities, and μ_{uvg} is the estimated gene expression change between samples u and v, which can be for example the difference between median $\log_2$ probe intensities over all probes targeting gene g, as suggested earlier in the context of gene expression studies (236). Between-group differences can then be summarized for each probe for example by calculating a moderated t-statistic over pairwise splicing indexes with the empirical Bayes method implemented in package limma (Smyth 2004, Section 2.3.3.2). These statistics are summarized for each exon as the median over the probes targeting the exon. Since this pairwise differential splicing analysis strategy was inspired by probe-level expression change averaging (PECA) (236) and the splicing index, we call it PECA-SI.

We compared the ability of PECA-SI to rank differential splicing events to seven other strategies: RMA-SI, PLIER-SI, RMA-MIDAS, PLIER-MIDAS, RMA-LM, PLIER-LM, and FIRMA. The splicing index methods RMA-SI and PLIER-SI summarize the probe level intensities to exon-level and gene-level expression values before computing the exon-level SI for each pair of samples, as in Equation (10). The summarization is done either with the RMA procedure (138) or the probe logarithmic intensity error model PLIER, which is part of commercial microarray data preprocessing software provided by Affymetrix. Similarly to RMA, it accounts for probe affinities that are assumed to have an additive effect on the total signal on the log scale, and aims at robust parameter estimation. PLIER (by default) uses the difference between each perfect match (PM) probe and its paired mismatch (MM) probe intensity to model background-corrected signal, but improves its predecessor MAS 5.0 by accounting for the dependency of the error on the intensity and does not assume the errors of the MM and PM probe intensities to be equal within each pair of probes (237). Since Affymetrix Exon arrays lack mismatch probes, the background correction is based on the median of probes that target a different genome but have the same GC-content (238).

Affymetrix's exon-level Microarray Detection of Alternative Splicing (MIDAS) approach (238) is based on the splicing index. It adds a small constant to the exon-level and gene-level intensities (before the $\log_2$-transformations) to stabilize the variance of the splicing index, which would otherwise be large close to the detection limit. The splicing index of each exon is then modeled as a sum of group effects and i.i.d. noise, and the significance of the group effect(s) of interest are evaluated by analysis of variance (ANOVA).

The linear model approaches RMA-LM and PLIER-LM were implemented as described by the authors of the two-way Analysis of Splice Variation (ANOSVA) method (239). ANOSVA models the observed exon intensities as a sum of baseline gene expression, an exon effect, a sample group effect, an interaction effect between the exon and the sample group, and the i.i.d. noise. Exons that have a significant interaction effect with the sample group are considered differentially spliced.

Finding Isoforms using Robust Multichip Analysis (FIRMA) (240) was considered the state-of-the art method to compare to. It models alternative splicing as an outlier detection problem. As in Equation (1) in Section 2.3.1.1, the normalized, background-corrected, and $\log_2$-transformed intensity value of each probe is modeled as a sum of sample-specific gene expression, probe affinity effect, and the error term. The parameters for the model are evaluated with iteratively reweighted least squares (IRLS) method. The final alternative splicing score of each exon is the median probe-level residual over the probes targeting the exon divided by their median absolute deviation. This FIRMA score is calculated separately for each sample and summarized to a t-statistic for a two-group comparison.

4.2 Evaluation of methods for the detection of differential splicing in synthetic and real data (I)

The methods for differential splicing detection were evaluated with respect to reproducibility, specificity, and sensitivity in real and simulated data. The data sets are described in Table 1, and further details can be found in the Methods section of Study I and the original publications associated with each data set.

Table 1. Data sets and criteria used in the evaluation of methods for the detection of differential splicing between two groups of samples.

Evaluation with respect to:	**Description**	**Data set**
AUROC	Synthetic data with characteristics similar to Affymetrix exon microarray data for 10000 genes with two levels of noise and 1-5 exons per gene simulated as differentially spliced between two groups of 10 samples	Simulated as described in (240)
Reproduciblility	Total RNA from two tissues (brain and heart) mixed in different proportions, 3 biological replicates of each mixture. Ideally, differences that are detected between pure brain vs. pure heart samples should also be detected between samples with unequal proportions of brain and heart tissue. Reproducibility was defined as the overlap of top k[1] findings in mixed vs. pure data.	Example tissue mixture data provided by Affymetrix[2]
Reproducibility	Tissue pool and brain RNA samples hybridized in two different laboratories (2 technical replicates of each of the 5 brain and 5 pool samples). Reproducibility was defined as the overlap of top k findings (differentially spliced exons in heart vs. tissue pool) between the two laboratories. This was evaluated with all combinations of 2–4 pairs of technical replicates per group.	Laboratory comparison data, Gene Expression Omnibus (GEO) accession code GSE13072 (241)
Sensitivity	RT-PCR-confirmed brain-specific exons in an independent set of samples from (146) were considered true positives. False positive rate at different significance thresholds was evaluated with a permutation analysis, where heart and pool sample labels were permuted 100 times	Laboratory comparison data, Gene Expression Omnibus (GEO) accession code GSE13072 (241)
Reproducibility	10 matched pairs of tumor and normal tissue samples. Reproducibility was defined as the overlap of top k findings (differentially spliced exons between tumor and normal tissue) detected in independent subsets of 2 - 4 sample pairs	Colon cancer data (242)
AUROC	Positive and negative findings in colon cancer data, as determined by RT-PCR in the same samples (242)	Colon cancer data (242)

[1] k = 100, 500, 1000, 1500, 2000

[2] Downloaded from www.affymetrix.com/support/technical/sample_data/gene_1_0_array_data.affx which nowadays (02/2021) re-directs to www.thermofisher.com/fi/en/home/life-science/microarray-analysis/microarray-data-analysis/microarray-analysis-sample-data/gene-st-array-data-set.html

4.3 Sample collection and study design (II, III, IV and V)

The samples for these studies came from two prospective type 1 diabetes study cohorts: the Finnish Diabetes Prediction and Prevention Project (DIPP, Studies II, IV and V) and DIABIMMUNE (Study III). DIPP has been collecting follow-up data on children at risk of type 1 diabetes since 1994. The genetic type 1 diabetes risk of each participant is evaluated at birth based on their HLA-DR/DQ haplotype (243,244), determined from a dried spot of umbilical cord blood, and newborn infants with high or moderate risk are invited to the follow-up, which includes islet autoantibody measurements 1–4 times per year until type 1 diabetes diagnosis or age 15. The case individuals had a type 1 diabetes diagnosis and/or were persistently positive for at least two islet autoantibodies (GADA, mIAA, IA2A or ZnT8A) measured with radiobinding assays, whereas the control individuals remained healthy throughout the follow-up. Study II, however, included two (out of 28) individuals, whose inclusion as cases was based on less specific ICA measurements by immunofluorescence (Section 2.1.3), according to an older DIPP protocol.

For the longitudinal gene expression study (II), peripheral blood samples were collected within the DIPP study in PAXgene blood RNA tubes 1–4 times per year from 28 case-control-pairs matched by date and place of birth, sex, and HLA risk class. Altogether, 356 RNA samples were included, spanning the time from before seroconversion through the time of type 1 diabetes diagnosis (Figure 3).

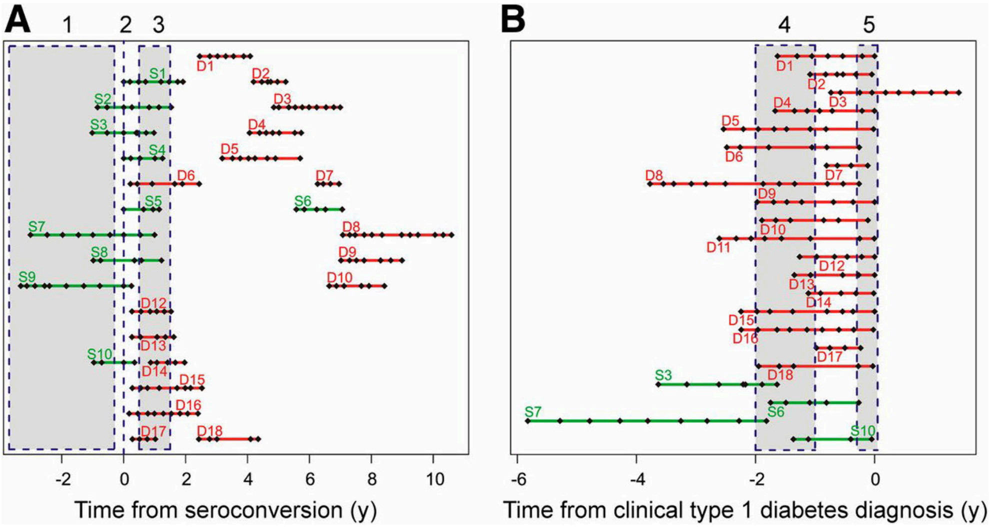


Figure 3. Reproduced from Study II with the permission of the copyright holder (American Diabetes Association, Diabetes journal): Each case-control-pair is visualized as a green or red line segment. The green and red colors correspond to sample sets S and D, respectively. Sample set D was collected closer to the time of diagnosis and hybridized on gene expression microarrays later than sample set S. The black diamonds represent the timing of sample collection relative to A) the time of the case individual's seroconversion to islet autoantibody positivity and B) the time of the case individual's diagnosis with type 1 diabetes. The dotted lines and grey boxes show five analysis time windows: 1. time before seroconversion 2. seroconversion 3. 6–18 months after seroconversion 4. 1–2 years before diagnosis 5. diagnosis.

The DNA methylation studies (IV and V) were based on DNA extracted by salting out procedure (245) from 200 umbilical cord blood samples collected from DIPP study participants in 3ml EDTA tubes immediately after birth and stored in –20°C. The metadata included all information recorded by the hospital on the pregnancy, delivery, the mother, and the neonate. Out of 200 samples, 20 were later rejected due to low (< 97 %) bisulfite conversion efficiency, 2 were excluded due to completely missing hospital metadata, and 5 were rejected due to inadequate amount or quality of DNA. Out of the 173 remaining DNA samples, 43 are from DIPP case individuals, and 79 from DIPP control individuals. Samples from 51 individuals, who did not qualify as DIPP cases or controls (for example due to transient autoantibody positivity), were excluded from the case-control-comparison in Study IV but included in the study on other covariates (Study V).

Study III is part of the DIABIMMUNE project which investigates risk factors of allergy and autoimmunity, especially type 1 diabetes and the hygiene hypothesis. According to the hygiene hypothesis, exposure to pathogens in early life improves later immune tolerance (44,246). DIABIMMUNE data collection started in 2008 in Finland, Russian Karelia, and Estonia. Compared to Russian Karelia, Finland has a higher standard of living, much lower incidence of some infections, such as Helicobacter Pylori, and a higher incidence of immune-mediated diseases and allergies (247). For example, the incidence of type 1 diabetes is six times higher in Finland as compared to Russian Karelia, even though the frequencies of predisposing and protective HLA genotypes are similar in these two areas (248). DIABIMMUNE has collected follow-up data for example on celiac disease and type 1 diabetes autoantibodies, gut microbiome, and diet from approximately 800 participants with an increased genetic (HLA) risk for autoimmunity. Another goal of DIABIMMUNE is to study immunologically relevant differences between the general populations in these three locations.

Study III compared gene expression in umbilical cord blood samples collected in Tempus RNA tubes (Applied Biosystems) from three locations: Jorvi maternity ward, Espoo, Finland (N = 48), maternity units of Tartu and Põlva, Estonia (N = 25), and two maternity wards in Petrozavodsk, Russian Karelia (N = 40). The samples were selected based on place and time of birth. Since the goal was to explore differences between the general populations of three locations, the HLA risk class information was not used for sample selection, even though it was collected and later included as a covariate in the analysis. All samples within Study III were collected between January and May 2010 according to the manufacturer's (Applied Biosystems) protocol and stored in –70°C until RNA extraction in Turku centre for Biotechnology (Finland). The compared study groups (Finland, Estonia, and Russian Karelia) had similar sex and HLA risk class distributions (III, Table 1), and all individuals within the study were born vaginally at a minimum gestational age of 37 weeks (full-term).

4.4 RNA extraction and hybridization on oligonucleotide microarrays (II and III)

Genome-wide gene expression measurements were obtained using Affymetrix U219 microarrays in Study III. Study II included two sets of samples: the first set was collected close to the time of seroconversion from 10 case-control-pairs and hybridized on Illumina HumanWG-6 version 2 arrays (later referred to as WG-6), whereas the second set was collected close to the time of type 1 diabetes diagnosis from 18 case-control-pairs and hybridized on Illumina HumanHT-12 version 3 arrays (later referred to as HT-12). These sample sets are visualized as green (the first set) and red (the second set) lines in Figure 3 and referred to as case-control-pairs S1-10 and D1-18. All 356 samples from 28 case-control-pairs were later hybridized on Affymetrix U219 arrays for technical replication.

Total whole-blood RNA was extracted from the samples using PAXgene Blood RNA kit (Qiagen) in Study II and Tempus Spin RNA isolation kit (Applied

Biosystems) in Study III. In both projects (II, III) the RNA samples were processed with the Ovation RNA amplification system v2 including the Ovation whole-blood reagent (NuGEN Technologies) and hybridized on GeneChip Human Genome U219 array plate (Affymetrix). GeneTitan Multi-Channel instrument (Affymetrix) was used for automatized hybridization, washing, staining, and scanning of the arrays. All samples within Study II were also hybridized on Illumina arrays. Illumina Sentrix human WG6 v2 expression bead chips were used for case-control pairs S1–S10, and Illumina Human HT-12 Expression BeadChips version 3 were used for case-control pairs D1–D18. Samples from case-control pairs S1–S6 were amplified with an older kit (RiboAmp OA 1 Applied Biosystems/Arcturus) for the Illumina array hybridizations (the above-mentioned Ovation kit was used for all other samples).

4.5 Preprocessing of gene expression microarray data (II and III)

For Illumina data, the background correction and the summarization of signals from individual beads to probe intensities were done using software provided by the manufacturer. The probe-level intensities were then quantile normalized and $\log_2$-transformed. For Affymetrix data (Studies II and III), the Robust Multi-array Average (RMA) preprocessing pipeline (137,138) was applied. Just as typical preprocessing workflows for Illumina data, it includes background correction, quantile normalization and log2-transformation of the measured intensities, but the RMA summarization step (Section 2.3.1.1) is specific to Affymetrix microarray data, which typically contains 15–20 partially overlapping probes per probeset, and each transcript is targeted with 1–4 probesets, whereas Illumina WG-6 and HT-12 arrays target each transcript with 1–4 unique probes (mostly 1).

Some analyses, such as the detection of transcripts that were differentially expressed between cases and controls 1–2 years before the cases' type 1 diabetes diagnosis, included samples from Illumina WG-6 and HT-12. Whenever data from different platforms needed to be combined, we only included probes that had remained similar between the platforms (challenges in probe design are discussed in Section 2.2.1). A minimum probe sequence overlap of 25 out of 50 consecutive bases (1 mismatch allowed) was required. Unless the sequences were completely identical, the probes were also required to target the same gene according to Illumina's annotations. In case the gene name had been updated, while the probe sequence had remained completely unchanged, we adopted the newer gene name. In total, 70 % of the probe sequences had remained completely identical, and another 6 % had been slightly updated but fulfilled our similarity criteria.

The present/absent calling (Section 2.3.1.1) was done using detection p-values for the Illumina data and by fitting a two-component Gaussian mixture model for the

Affymetrix data (II and III) with an expectation maximization (EM) algorithm using R package mixtools (249). We excluded probes/probesets that were called absent in all samples of at least 50 % of the individuals in all study groups (cases and controls in Study II; Finland, Espoo, and Russian Karelia in Study III).

In study III, the transportation of Russian Karelian samples was delayed. Finnish and Estonian samples were hybridized on array plates in December 2011–January 2012, whereas the Russian Karelian samples were hybridized in March 2013. To be able to estimate the batch effect, some samples (7 from Russian Karelia, 4 from Finland, and 4 from Estonia; hybridization batch 4 in Supplementary Table 1, Study III) were re-hybridized. Batch correction (Section 2.3.2.2) was done with an empirical Bayes method implemented in package ComBat (176) such that the individual and the place of birth were included as covariates. Batch effects were not an issue in Study II, where the case-control pairs were kept together during all sample processing steps and the analysis was done in a paired way.

4.6 Differential expression analysis (II and III)

In Study III, the effect of the in-utero environment (Finland, Estonia or Russian Karelia) on each log2-transformed gene expression value was estimated with a linear model implemented in the R package limma (187), which uses an empirical Bayes approach to estimate residual variances (Section 2.3.3.2). Sex, gestational age, month of birth (numeric values 1–5, representing January–May), and HLA risk class (ordinal variable with levels low, moderate, high and very high) were included as covariates. The effects of three contrasts were estimated: Finland vs. Russian Karelia, Estonia vs. Russian Karelia and Finland vs. Estonia. Probe sets with FDR < 0.01 (Benjamini-Hochberg) were considered differentially expressed.

In Study II, the non-parametric rank product method (195) was applied to detect pairwise transcriptomic differences between 28 case individuals, who developed islet autoantibodies or were diagnosed with type 1 diabetes, and 28 matched controls, who remained autoantibody-negative throughout the follow-up. Data set 1 included 10 case-control-pairs, whose samples were mostly collected close to the time of seroconversion, and data set 2 included 18 case-control-pairs, whose samples were mostly collected close to the time of diagnosis. The differential expression analysis was done A) across the longitudinal profiles, separately in data set 1 and data set 2 and B) in specific time-windows relative to the disease progression (including samples from both data sets in most time windows).

The objective of analysis A) was to find genes that showed a between-group (case vs. control) expression difference in at least one time point during the follow up but that had little within-individual variation in the control profiles. For each transcript, the expression value at each case sample was transformed to a z-score (x–

m)/s, where m is the sample mean and s is the sample standard deviation over the samples of the matched control. Such an approach had been earlier taken in similar studies where the case and control time points are not exactly matched (104,250). The maximum/minimum z-scores of each case individual were used in the rank product analysis. Transcripts with FDR < 0.05 were considered differentially expressed. We excluded transcripts that were detected as differentially expressed with FDR < 0.05 in a swapped analysis (where control-case differences were penalized for within-case variation).

For analysis B), each control sample series was matched to the time points of the corresponding case sample series to calculate pairwise expression differences. This was done by approximating the control gene expression with inter-/extrapolation at each case sample. More specifically, linear interpolation was applied for the time points that were inside the range of real control time points. For the time points needed outside this range, the expression values were approximated by constant extrapolation (set equal to the closest real measurement). Rank product was then applied for pairwise maximum/minimum differences within 5 distinct time windows (Figure 3): 1. before seroconversion to islet autoantibody positivity 2. at seroconversion (the first autoantibody-positive sample from each individual) 3. 6–18 months after seroconversion 4. 1–2 years before type 1 diabetes diagnosis, and 5. at diagnosis (the sample closest to diagnosis from each diagnosed individual)

For technical replication, all the above-described analysis steps were performed separately for Affymetrix and Illumina data (all samples within Study II had been hybridized to both platforms), and only genes detected as differentially expressed with both platforms (concordant up-/downregulation and FDR < 0.05 in both analyses) were considered differentially expressed. Affymetrix and Illumina data were combined on the level of gene symbols using Ingenuity Pathway Analysis (251) annotations.

4.7 eQTL analysis (II)

Whole blood DNA was genotyped with the Immunochip (252) array, according to Illumina protocol at the Department of Genetics, University Medical Centre Groningen (the Netherlands). Single nucleotide polymorphisms (SNPs) were mapped to National Center for Biotechnology Information build 36 (hg18). The pre-processing was done using the default pipeline within software PLINK (253): Samples with a call rate < 95 % were excluded. SNPs were excluded if they deviated from the Hardy-Weinberg equilibrium within the control group (p-value < 0.0001), had a minor allele frequency < 10 % (in our data) or a call rate < 98 %. The data set was then pruned based on linkage disequilibrium between the markers ($r^2 > 0.8$). Our final data set included 30463 SNPs.

To estimate possible cis eQTL (expressed quantitative trait loci) effects on the differentially expressed genes, a linear model for the effect of genotype (the number of alternative alleles) on gene expression was fit for each SNP-gene-pair, including all SNPs within 250 kb to both directions from the genomic coordinates of each differentially expressed gene. SNPs with significant (FDR < 0.05) correlations with gene expression in our data and their proxies ($r^2 > 0.8$) were then searched for associations with autoimmune diseases from published GWAS (254,255). The proxies of the SNPs were found based on HapMap3 release 2 and 1000 Genomes in the CEU population panel by using the Broad Institute's SNP annotation and proxy search tool, nowadays known as SNPsnap (256)

4.8 Functional interpretation of differential expression (II and III)

4.8.1 Overrepresented pathways, transcription factor binding sites and functional modules (II and III)

All the enrichment analyses were based on the Fisher's exact test. The enrichment of the differentially expressed genes on pathways was explored through the Ingenuity Pathway analysis (IPA) tool from QIAGEN (www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis, (251) and the Database for Annotation, Visualization and Integrated Discovery (DAVID) (257). Transcription factor binding site enrichment was tested in the promoter regions (62 kb around the transcription start site) of each differentially expressed gene using the transcription factor targets in the Molecular Signatures Database (210). A curated list of genes with functions in human innate immunity was retrieved from the InnateDB (258). Modules of human blood co-expressed genes, annotated with automatic literature search, were downloaded from www.biir.net/modules (208).

4.8.2 Association with immune system maturity (III)

To estimate the immune system maturity reflected in the gene expression patterns of the umbilical cord blood samples in Study III, we compared our results to two published transcritome-wide data sets that included samples from neonates and 1-year-old infants. One of these studies reported 549 genes that were differentially expressed between these age groups in activated macrophages (259). The other study (260) compared the whole blood transcriptome of children under bacterial sepsis to that of healthy control children. We downloaded preprocessed and normalized microarray data sets of (260) from Gene Expression Omnibus (GSE26440 and

GSE26378) and selected data from neonates (age < 1 month) and older infants (age 0.5–1.5 years). Approximately 1/3 of the selected data were from healthy control individuals and 2/3 from sepsis patients in both age groups. A linear model was fit for each gene expression value as a function of age (binary variable with categories "newborn" and "infant"), health status and survival. The association between age group and gene expression was evaluated with an F-test, moderated such that the residual variances were estimated with an empirical Bayes procedure implemented in R package limma (187). Altogether 2205 genes were identified as upregulated and 1432 genes as downregulated in older infants, as compared to the neonates (FDR < 0.01). The enrichment of these genes, as well as the differentially expressed genes from (259), among our differentially expressed genes were calculated using Fisher's exact test and the p-values were Benjamini-Hochberg corrected.

4.9 Bisulfite sequencing of umbilical cord blood samples (IV and V)

The Library preparation was started from 200 ng of genomic DNA and performed according to the reduced representation bisulfite sequencing (RRBS) protocol from (261). Briefly, the protocol includes MspI digestion to capture CpG rich areas, followed by end repair, A-tailing, and adapter ligation of the DNA fragments. A lower concentration of adapters (1:10 dilution) was used than recommended by the manufacturer to reduce the occurrence of adapter dimers. Bisulfite conversion and sample purification were done according to Invitrogen MethylCode Bisulfite Conversion Kit protocol. Aliquots of converted DNA were amplified by 18 cycles of PCR with a proofreading enzyme that does not stall at uracil. Library qualities were analyzed with either Advanced Analytical Fragment Analyzer or Bioanalyzer (depending on library size) and only high-quality libraries were sequenced. The sequencing was done in 32 lanes, 4–7 samples per lane, with Illumina HiSeq 2500 instrument using TruSeq v3 sequencing chemistry. Paired-end sequencing with 2 x 100 bp read length was used with 6 bp index run. Technical quality of the HiSeq 2500 run was good and the cluster amount was as expected. The yields were 18 - 37 million raw paired-end reads per sample.

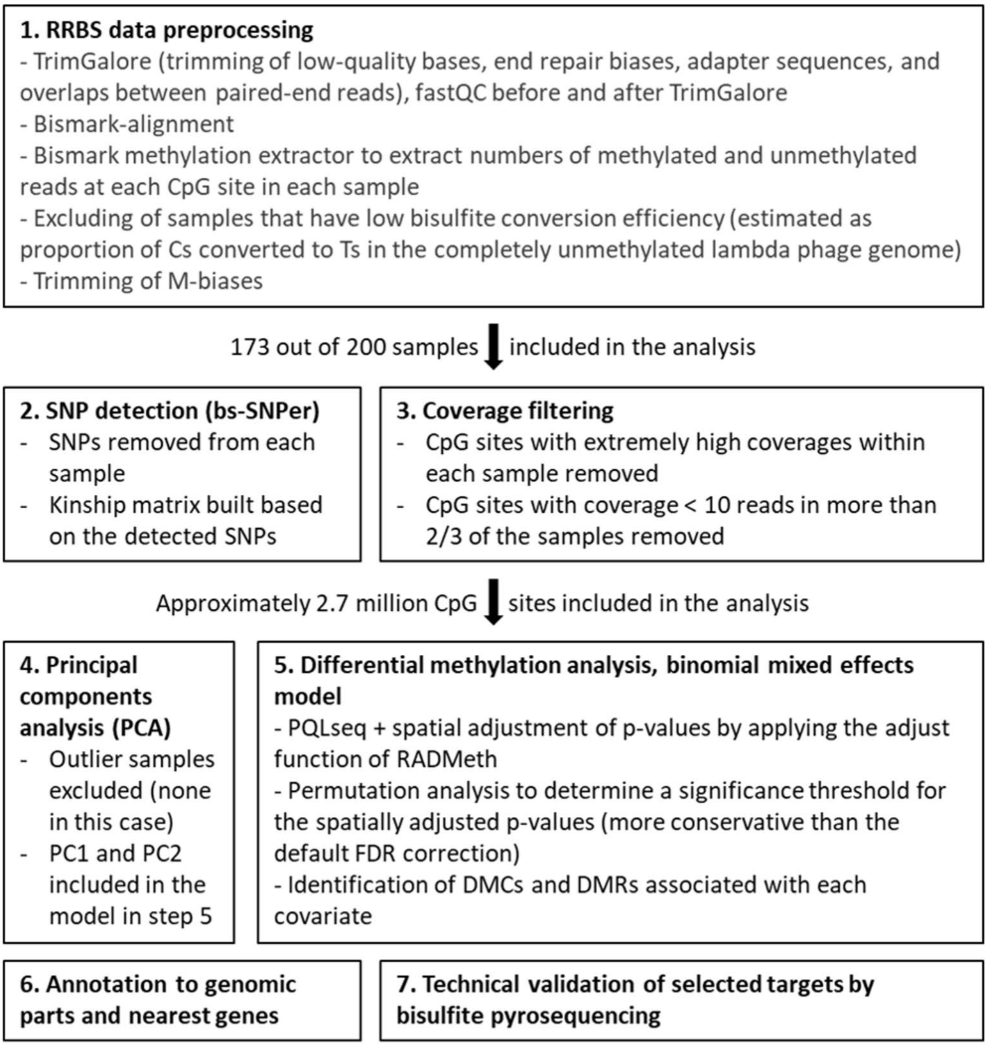
4.10 Preprocessing and filtering of bisulfite sequencing data (IV and V)

The RRBS data preprocessing and analysis workflow is reviewed in Sections 2.3.1.3–2.3.1.8, 2.3.4.2, and 2.3.4.3 and summarized in Figure 4. The reads were trimmed by running TrimGalore version 0.4.3 (133) with default parameters in paired end RRBS mode on each fastq-file to remove 1) end repair biases, 2) adapter sequences with a minimum sequence overlap of 1 base, 3) bases with base call error rate above 1%, and 4) reads shorter than 20 bp after trimming. This step removed 2 - 8 % of the raw reads from all samples except one sample, from which 24 % of raw reads were discarded.

Read alignment was done on the human GRCh37 (hg19) genome assembly (262,263) and the lambda phage genome simultaneously with Bowtie2 version 2.3.1 within Bismark version 0.17.0 (160) with default parameters in paired-end mode. Numbers of methylated and unmethylated reads at each CpG site in each sample were extracted with Bismark methylation extractor version 0.22.3 (160), which excludes read 2 bases that overlap with read 1 to avoid redundant methylation calls. Bisulfite conversion efficiency was estimated as the sum of observed unmethylated CpG counts divided by the total sum of methylated and unmethylated CpG counts within the lambda genome. The conversion efficiencies were above 97 % (median 99.4 %) for all except 20 samples, which were excluded.

To remove M-biases, positions at the 3' end of read 1 and both ends of read 2 were excluded if their CpG methylation proportions were more than 3 standard deviations below or above the mean methylation proportion at positions 10–91 (middle 80 %). In practice, Bismark methylation extractor was re-run with additional ignore-parameters to exclude biased positions after extracting the needed information from the output M-bias files of the first run. The information from both strands was merged for each CpG site by running Bismark function coverage2cytosine with parameter merge_CpG for each .cov-file produced by Bismark methylation extractor. The numbers of methylated and unmethylated reads were extracted from the output of coverage2cytosine and organized into count matrices with total and methylated numbers of reads for each CpG site and each sample.

SNP detection was done by applying bsSNPer (172) with its default parameters on bam-files sorted by genomic coordinates after excluding the lambda phage genome. The SNPs (flagged "PASS") detected for each individual were removed from the data (the individual's read counts set to NA at the SNP). In order to remove most PCR duplication biases, CpG sites with coverage above the 99.9th percentile were removed from each sample. CpG sites were completely excluded if they had a low-coverage value (total number of reads < 10) or a missing value (NA due to a potential SNP or extremely high coverage) in at least two thirds of the samples. The below-described differential methylation analysis was run for all 2 752 981 CpG sites passing these criteria. However, further covariate-specific filtering of CpG sites was done before calling differential methylation associated with any binary covariate. Minimum coverage 10 in at least one third of the samples in each group was required and further, a minimum coverage 10 was required in at least five samples per group (the second criterion is relevant only for binary covariates with less than 15 samples in one group).



1. RRBS data preprocessing

- TrimGalore (trimming of low-quality bases, end repair biases, adapter sequences, and overlaps between paired-end reads), fastQC before and after TrimGalore
- Bismark-alignment
- Bismark methylation extractor to extract numbers of methylated and unmethylated reads at each CpG site in each sample
- Excluding of samples that have low bisulfite conversion efficiency (estimated as proportion of Cs converted to Ts in the completely unmethylated lambda phage genome)
- Trimming of M-biases

173 out of 200 samples included in the analysis

2. SNP detection (bs-SNPer)

- SNPs removed from each sample
- Kinship matrix built based on the detected SNPs

3. Coverage filtering

- CpG sites with extremely high coverages within each sample removed
- CpG sites with coverage < 10 reads in more than 2/3 of the samples removed

Approximately 2.7 million CpG sites included in the analysis

4. Principal components analysis (PCA)

- Outlier samples excluded (none in this case)
- PC1 and PC2 included in the model in step 5

5. Differential methylation analysis, binomial mixed effects model

- PQLseq + spatial adjustment of p-values by applying the adjust function of RADMeth
- Permutation analysis to determine a significance threshold for the spatially adjusted p-values (more conservative than the default FDR correction)
- Identification of DMCs and DMRs associated with each covariate

6. Annotation to genomic parts and nearest genes

7. Technical validation of selected targets by bisulfite pyrosequencing

Figure 4. modified from Study V: The outline of the reduced representation bisulfite sequencing (RRBS) data analysis workflow in Studies IV and V.

4.11 Differential methylation analysis (IV and V)

4.11.1 GLMM for each CpG site

The differential methylation analysis was done by applying a binomial mixed effects model (Section 2.3.4.2) implemented in R package PQLseq version 1.1. (220) within R version 3.6.1. on the methylated and total read counts at each high-coverage CpG site on chromosomes 1–22. This was done after adding + 1 to the numbers of methylated reads and +2 to the total numbers of reads to avoid modeling methylation proportions that are exactly 0 or 1. This pseudo-count transformation was only applied to non-missing values (coverage > 0). In Study V the goal was to evaluate differential methylation associated with several covariates. Therefore, the source code of PQLseq was modified to output the coefficients, standard errors and Wald test p-values for all covariates included in the model. The modified version is available in Github (264).

From all available information collected by the Turku University Hospital maternity ward and the DIPP clinic, clinical covariates were selected to be included in the regression model, such that within each group of mutually correlating covariates (Pearson correlation coefficient > 0.3 and p-value < 0.05, or alternatively Fisher’s exact test p < 0.05, depending on the types of the covariates), the most reliably measured covariate was included (V, Supplementary Table 1 and IV, Supplementary Table 1). Table 2 lists the covariates that were part of the final design matrix, modeled as fixed effects. Missing covariate values were median-imputed, and continuous covariates were Z-transformed (divided by the standard deviation after subtracting the mean).

Table 2. Modified from Study IV and Study V: Covariates included in the differential methylation analysis, modeled as fixed effects by PQLseq (220) in Studies IV and V. Study IV focused on differential methylation associated with later progression to type 1 diabetes ("Class") but included all the other covariates to account for potential confounding effects. Study V evaluated differential methylation associated with each clinical covariate listed here. Study IV included a subset (N=122) of the individuals included in Study V (N=173).

Covariate	**Description**
Age, mother	Z-transformed continuous covariate
Apgar, low	The 1-minute Apgar points simplified to low/normal, such that values 8–10 were encoded as 0 (N=149) and values below 8 were encoded as 1 (N=24)
Birth weight	Z-transformed continuous covariate
BMI, mother (pre-pregnancy)	Z-transformed continuous covariate
Caesarean section	The mode of delivery simplified to a binary variable, such that 0=vaginal (N=152), 1=C-section (N=21)
Class	The covariate of interest in Study IV: 0=Completely autoantibody-negative control (N=79), 1=Case persistently positive for multiple autoantibodies and/or diagnosed with type 1 diabetes (N=43). Study V included class as a confounding covariate, such that 1=persistent autoantibody positivity for 1 or more autoantibodies (N=47), 0=control (N=126, including 79 totally autoantibody-negative individuals and 47 individuals with transient autoantibody-positivity).
Epidural anaesthetic	A binary covariate, 0=not used (N=87), 1=used (N=86). This value was corrected to 0 for two individuals with an elective C-section
Gestational weight gain, mother	Z-transformed continuous covariate
Height, mother	Z-transformed continuous covariate
HLA risk class	The HLA risk class was defined as described earlier (243,244) and included in the regression model in Study IV but not in Study V. This ordinal variable was divided into two binary variables ”HLA neutral” (N=31) and ”HLA high” (N=45). Moderate risk (N=46) was included in the intercept.
Induced labor	A binary covariate, 0=not induced (N=145), 1=induced (N=28).
Insulin-treated diabetes, mother	This can be gestational or any other type of diabetes treated by insulin during the pregnancy. 0=no (N=165), 1=yes (N=8)
Library preparation batch	A technical categorical covariate with 7 binray levels (intercept and 6 others). The number of samples per batch ranged from 5 to 48, median 23
Number of earlier miscarriages	The number of miscarriages was simplified to a binary variable, such that 1=one or more earlier miscarriages (N=31)
PC1 and PC2	Principal components 1 and 2 of the full median-imputed methylation proportion matrix were included to account for technical variation
Smoking during pregnancy	Since the data only included two examples of smoking only during the first trimester, and 12 examples of smoking throughout the pregnancy, this variable was simplified to smoking (N=14)/no smoking (N=159)
Month of birth	The month of birth was transformed as $\cos(2\pi m/12)$ to account for its cyclic nature (m = month, originally encoded as 1–12)
Year of birth	Z-transformed continuous covariate

Since PQLseq was originally designed to model differential methylation in the presence of population structures, we included the relatedness of the individuals as a random effect. To our knowledge, these 173 individuals are unrelated, but we estimated their genetic similarity by utilizing the SNPs detected as described above. The relatedness matrix is a correlation matrix of the samples' SNP profiles, which include all detected (flagged "PASS") SNPs with minor allele frequency > 5%, encoded as the number of reference alleles (0,1,2). This is calculated as XX^T/N_{SNPs}, where X is a $N_{samples}$ x N_{SNPs} matrix (173 x 189985) containing numbers of reference alleles, standardized to z-scores within each sample.

4.11.2 Spatial adjustment and FDR-correction

The Wald test p-values computed for each CpG site within PQLseq were spatially adjusted with an autocorrelation-corrected Z-test implemented in RADMeth (222) within Methpipe version 3.4.3. after sorting the CpG sites by chromosome and location. By default, RADMeth performs a Benjamini-Hochberg-correction on the adjusted p-values, but we found this procedure to be insufficient for FDR control. Instead, we estimated an empirical threshold for the spatially adjusted p-values through a permutation analysis. Specifically, we re-ran PQLseq and the spatial adjustment 48 times (3 times for each covariate) such that one covariate was permuted at each run. A threshold for the spatially adjusted p-value was set such that the number of differentially methylated cytosines (DMCs) associated with a permuted covariate (false discoveries) would be less than 5 % of the number of DMCs associated with the corresponding real covariate using that threshold. The median threshold value over three repeats was used.

CpG sites were considered differentially methylated if their Benjamini-Hochberg-corrected p-value was significant (< 0.05) already before any spatial adjustment, or if they were part of a differentially methylated region (DMR). To detect DMRs, we considered all candidate CpG sites with an empirically determined FDR < 0.05 after the spatial adjustment, in addition to the cytosines with evidence of differential methylation already before spatial adjustment. A differentially methylated region (DMR) was defined as a genomic region with two or more such candidate CpG sites that were within a window of 2 kb and had the same direction of methylation difference in at least 90 % of the candidate CpGs. Further, at least one of the CpGs was required to have coverage-corrected mean methylation difference > 5 %. Coverage-corrected mean methylation difference is calculated as sum(number of methylated reads in cases)/sum(number of total reads in cases) - sum(number of methylated reads in controls)/sum(number of total reads in controls).

Annotation of differentially methylated CpG sites to genomic parts (promoter, intron, exon, intergenic) and nearest UCSC known genes was done through R package genomation version 1.16.0 (265) using Genome Reference Consortium Human Build 37 (GRCh37, hg19).

4.12 Pyrosequencing validation of selected targets (IV and V)

For technical validation of candidate differentially methylated regions with targeted pyrosequencing, 60 individuals were chosen with the following criteria: even number of male and female cases and male and female controls, pregnancy duration > 37 weeks, normal birth weight (2.5–4.5 kg), no multiple pregnancies, normal Apgar points (8–10), no perinatal asphyxia, vaginal birth, and no maternal smoking. The regions of interest were captured with a targeted assay and amplified by 45 rounds of PCR. The Pyrosequencing was done with PyroMark Q24, and methylation percentages were extracted from the light intensity values at each CpG site using the manufacturer's software. Since technical uncertainty due to limited coverage is not an issue in targeted pyrosequencing, a standard linear model was applied after transforming each methylation proportion with: arcsin(2 × proportion − 1). All explanatory variables in Table 2 were included in the model, except for the above-listed variables that had zero variation among these individuals (such as maternal smoking). The model was fit with and without the covariate of interest, and the models were compared with a likelihood ratio test to assess the significance of the covariate’s effect on DNA methylation at each CpG site. These steps were done using functions lm and anova, R version 4.0.4. (266).

5 Results

5.1 Probe-level estimation improves the detection of alternative splicing events in Affymetrix exon array studies (I)

A novel method was developed for the detection of differential splicing between two groups and is currently available as function PECASI under R/Bioconductor package PECA version 1.26.0 (267). The proposed probe-level splicing index method (PECA-SI) was compared to seven other differential splicing detection methods in terms of reproducibility, specificity, and sensitivity in both real and simulated data.

The performance of all the evaluated methods was excellent (AUROC 0.94–0.99) in simulated data at a typical noise level $\sigma = 0.7$ (240), when only one exon per gene was simulated as differentially spliced between two groups of samples, but PECA-SI outperformed the other methods, when multiple exons per gene were differentially spliced (I, Table 1). At 5 differentially spliced exons per gene, the AUROC-values of the other methods dropped to 0.79–0.88, while the performance of PECA-SI decreased only slightly, as compared to the simpler simulation setting (AUROC 0.92 vs. 0.99 at 5 vs. 1 exon per gene simulated as differentially spliced). As expected, the performance of all methods decreased, as the noise level was increased to $\sigma = 1$. At the most challenging settings ($\sigma = 1$ and 5 differentially spliced exons per gene) the AUROC-values ranged from 0.66 (FIRMA) to 0.74 (PECA-SI).

Reproducibility was evaluated in the context of technical replication (ability to detect similar results in data replicated in different laboratories), controlled biological noise (the ability to reproduce pure brain vs. pure heart tissue differences in samples with varying mixtures of brain and heart tissue) and biological replication in heterogeneous colon cancer data (ability to detect similar results from independent subsets of sample pairs). The between-laboratory overlap of top 500 differentially spliced exons detected with PECA-SI was on average approx. 60 % already with as few as 2 biological replicates per group and improved to approx. 70 %, when the sample number was increased to 4 biological replicates per group (I, Figure 2). The reproducibility of the other methods also improved with increasing sample size but remained at 50 % or lower even with 4 biological replicates per group.

Out of top 500 exons differentially spliced between pure heart and pure brain tissue, PECA-SI reproduced one third between tissue mixtures with a 20 % mixture difference (95% heart/5% brain vs. 75% heart/25% brain) (Figure 5). The method with the second highest reproducibility in this context was FIRMA, which was able to reproduce 13 % of the findings already at mixture difference 20 %. As expected, the proportion of reproduced findings increased with increasing mixture difference, regardless of the differential splicing detection method. The reproducibility of PECA-SI was higher compared to the other methods across different mixture differences and top list sizes except for top list size 100, where the FIRMA method performed similarly as PECA-SI (I, Additional file 1).

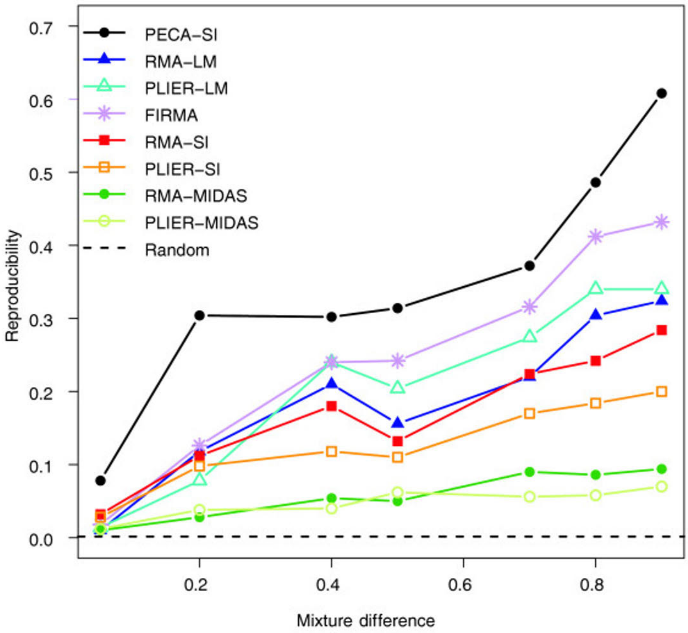


Figure 5. From Study I, reproduced with a permission from the copyright holder (BMC, Genome Biology). Reproducibility in the context of controlled biological noise: the ability of differential splicing detection methods to reproduce pure brain vs. pure heart tissue detections in samples with varying mixtures of brain and heart tissue. Reproducibility was defined as the overlap between 500 top-ranked splicing events in pure vs. mixed data sets.

To assess reproducibility in a more challenging context, we compared top ranked differentially spliced exons using independent subsets of sample pairs from a colon cancer study that collected tumor and adjacent normal tissue samples (242). The reproducibility between independent sets of tumor vs. normal samples was comparable to that in the most challenging tissue mixture setting (mixture difference

5 %). Even with the best-performing method PECA-SI, the reproducibility of such results was on average only 10 % at 2–4 biological replicates, which is not surprising, given the biological variation between tumors and individuals. Again, the other methods produced less reproducible results at all sample sizes and top list sizes (I, Figure 3 and Additional file 3).

Two of the evaluated methods (RMA-MIDAS and PLIER-MIDAS) performed poorly with respect to reproducibility in all comparisons. Their reproducibility was less than 20 % in the context of technical replication, and less than 10 % in the context of biological replication.

Finally, the methods were evaluated in terms of specificity and sensitivity to detect RT-PCR-confirmed differentially spliced exons. Out of 51 exons that were confirmed to be brain-specific based on RT-PCR measurements (146) and that matched the probesets of the brain vs. tissue pool exon array comparison (241), 44 were differentially spliced at FDR < 0.01 between brain and tissue pool, based on ranking by PECA-SI and estimating the FDR with a permutation test. All the methods were able to detect at least 37 out of 51 exons at this FDR threshold (I, Figure 4a). For the colon cancer data, we performed an ROC analysis based on 17 true and 14 false findings, according to RT-PCR measurements within the same study (242). When the exons array measurements were ranked by PECA-SI after filtering out low-intensity probes, we observed a perfect separation of RT-PCR-confirmed (10) and non-confirmed findings (8). The performance of all methods was improved by filtering out low-intensity probes (solid vs. dotted lines in I, Figure 4b). However, even without the filtering, PECA-SI performed at least as well as the other methods in this comparison.

5.2 Longitudinal gene expression patterns associated with type 1 diabetes (II)

The transcriptome-wide microarray data of longitudinal samples collected from children at risk of type 1 diabetes (Study II) is available in Gene Expression Omnibus under accession code GSE30211. We identified 109 genes at FDR < 0.05 (II, Supplementary Data 3) that were differentially expressed across samples collected close to the time of seroconversion to islet autoantibody positivity, compared to samples from matched autoantibody-negative control individuals. Interferon regulatory factor binding sites were enriched in the promoters of these genes (II, Figure 2A). The upregulated genes included two interconnected interferon regulatory factors 5 and 7 (IRF5 and IRF7) and 15 other genes that interact with IRF5 and IRF7 (II, Figure 2B), which are central mediators of the innate response to viral infections.

In the time-window-specific analysis of these samples, 124 genes were differentially expressed already before seroconversion, 60 of which were also differentially expressed at the time of clinical diagnosis (II, Figure 3), even though the data within these time windows were collected from almost independent sets of individuals (Figure 3). Human innate immune genes from innateDB (258) were enriched among the genes that were differentially expressed before and after seroconversion, as well as close to the time of diagnosis (II, Supplementary Data 3,4 and 5). Accordantly, out of co-expressed modules of genes reported in human blood (208), modules associated with interferon and neutrophils were differentially expressed across the time-line in our data (II, Figure 4, Supplementary Data 6).

To explore the possible genomic background of the observed differential gene expression patterns, genotype data (Immunochip) from the study participants was collected and all potential cis-eQTL effects within 250 kb from each differentially expressed gene were reported (II, Supplementary Data). Significant correlations between genotype(s) and gene expression (FDR < 0.05) were observed at 118 differentially expressed genes, and 27 % of these potential eQTL effects had been reported in a human blood eQTL meta-analysis, based on data from 5311 individuals with replication in 2775 individuals (268). Altogether 14 differentially expressed genes had one or more potential eQTL SNP that was previously associated with autoimmune diseases in genome-wide association studies (254,255).

5.3 Gene expression patterns associated with in-utero environment (III)

Study III explored the transcriptomic differences in umbilical cord blood samples collected from children born in Finland (N = 48), Estonia (N = 25), and Russian Karelia (N = 40). At FDR < 0.01, 3442 probesets were differentially expressed in these data between Finland and Russian Karelia, 1655 probesets between Russian Karelia and Estonia, and 130 probesets between Finland and Estonia (III, Supplementary Table 2). The probesets that were upregulated in Finland, as compared to Russian Karelia, were often also upregulated in Estonia, as compared to Russian Karelia (overlap 475, as visualized in a Venn diagram, III Figure 2). Similarly, the probesets that were downregulated in Finland, as compared to Russian Karelia, were often also downregulated in Estonia, as compared to Russian Karelia (overlap 424). Opposite patterns (a probeset being downregulated in Finland vs. Russian Karelia but upregulated in Estonia vs. Russian Karelia or vice versa) were not observed.

Genes with functions in human innate immunity (258) were enriched (Benjamini-Hochberg corrected Fisher's exact test p-value < 0.05) among genes that were upregulated in Russian Karelia, as compared to Finland and/or Estonia, and among genes that were downregulated in Russian Karelia, as compared to Finland. For example, toll-like receptor 2 (TLR2) and several other receptors responsible for recognizing microbial patterns were upregulated in Russian Karelia, as compared to Finland and/or Estonia.

To estimate whether the observed differences reflect immune maturation, we compared our results to earlier published gene expression patterns in the context of immune defense between different age groups. One of these studies (259) explored gene expression in macrophages that were extracted from blood samples of neonates and 1-year-old children and activated by lipopolysaccharides (characteristic of Gram-negative bacteria). Another study reported whole blood gene expression patterns in children under bacterial sepsis compared to healthy control children in different age groups (260). We re-analyzed the data for differential expression between 1-year-old children and neonates. We observed a highly significant overlap between transcripts that were upregulated in Russian Karelia, as compared to Finland and/or Estonia, and transcripts that were upregulated in 1-year-old children, as compared to neonates. The significances of the overlaps are visualized in III, Figure 4, and the details are in III, Supplementary Table 3.

5.4 Spatial adjustment by autocorrelation-corrected Z-test inflates p-values in RRBS data analysis (IV and V)

Differential methylation analysis is typically done separately for each CpG site, and the p-values are combined afterwards to account for spatial correlation between methylation statuses of adjacent CpG sites (Section 2.3.4.3). One of the most commonly used methods for this is the autocorrelation-adjusted Z-test implemented in RADMeth (222). If several relatively small raw p-values are associated with CpG sites that are close to each other in the genome, the spatially adjusted p-values are often several orders of magnitude smaller than any of the original ones. By default, RADMeth estimates FDR by Benjamini-Hochberg-correcting the adjusted p-values.

In Studies IV and V the goal was to locate any CpG sites, where explanatory variables such as later progression to type 1 diabetes, birth weight or maternal age have a significant effect on umbilical cord blood DNA methylation. Using the standard workflow for spatial adjustment, hundreds of CpG sites were associated even with covariates, such as the month of birth that were expected to be irrelevant. As a sanity check, we re-ran the differential methylation analysis for a permuted covariate, such that the input files remained otherwise unchanged. Ideally, no differential methylation should have been associated with the permuted covariate (especially since we made sure it did not correlate with any real clinical or technical variable). However, hundreds of false positives were observed using a standard threshold (Benjamini-Hochberg corrected spatially adjusted p-value < 0.05). After replicating this sanity check 3 times for each of 16 clinical covariates, we concluded that the spatially adjusted p-values were inflated (hundreds of false positives were detected every time, as visualized in V, Figure 3) but there was no such problem with the raw output p-values of PQLseq (220). The number of differentially methylated CpG sites associated with a permuted covariate before the spatial adjustment (Benjamini-Hochberg corrected PQLseq p-value < 0.05) was nearly always 0. Typically, the smallest raw p-values were in the order of 10^{-7} (as expected among 2.6 million tests) whereas the smallest spatially adjusted p-values were in the order of 10^{-10}-10^{-16} in the permuted analyses.

The spatial adjustment approach used here was originally intended for the beta-binomial regression p-values computed as implemented within RADMeth. To check, whether the p-value inflation could be caused by the application of RADMeth spatial adjustment on output p-values of PQLseq, we re-ran the analysis with the RADMeth pipeline. However, this did not change the conclusions. In fact, thousands of false discoveries were detected with this approach (V, Table 2). We also re-ran the analysis with comb-p (234), which is another implementation of the autocorrelation-adjusted Z-test with a different definition of a differentially methylated region. The numbers of false discoveries were close to the numbers observed for RADMeth.

Benjamini-Hochberg-correction on raw output p-values associated with individual CpG sites (such as those computed by PQLseq) assumes independence of tests and is too conservative in the analysis of bisulfite sequencing data, where measurements can be arbitrarily close to each other and high spatial correlation is present (77). Furthermore, efficient spatial adjustment such as the one implemented within RADMeth, is necessary for the study of differentially methylated regions, which can be expected to be biologically more relevant than individual CpG sites. Therefore, we kept the autocorrelation-adjusted Z-test as a convenient tool to find candidate differentially methylated regions but determined the significance threshold empirically. With this approach, few differentially methylated CpG sites or regions were associated with some covariates, but the number of findings was typically two orders of magnitude smaller than it would have been before this sanity check.

5.5 Association of umbilical cord blood DNA methylation with later progression to type 1 diabetes and other variables (IV and V)

Altogether, 297 differentially methylated regions (DMRs) were associated with sex, and in addition to the DMCs within these regions, 261 CpG sites outside DMRs were detected as differentially methylated based on Benjamini-Hochberg-corrected p-values (FDR < 0.05) already before any spatial adjustment (V, Table 1 and V, Supplementary Table 2). Our results confirmed a large proportion (overlap 221 CpG sites, Fisher's exact test p-value < 10^{-15}) of differentially methylated cytosines that had been reported in two earlier studies on sex-associated umbilical cord blood DNA methylation marks (269,270).

Promoter regions with most significant differential methylation associated with sex (among top 5, ranked by p-value) included the promoters of zona-pellucida binding protein 2 (ZPBP2) and developmental pluripotency associated 5 (DPPA5), which are overexpressed in the testis tissue and have very little or no expression in any other tissue type, according to the Genotype-Tissue Expression (GTEx) Portal on 30/11/20 (271). These promoter regions were hypomethylated in males, as compared to females. Hence, ZPBP2 and DPPA5 can be expected to be more highly expressed in males. The differentially methylated region on the promoter of ZPBP2 was selected for technical validation with targeted pyrosequencing. The coverage on the region was already excellent in RRBS data (median 78 reads per sample, range 4–298) but we selected this validation target as a positive control to confirm that DNA methylation can be reliably quantified by the pyrosequencing protocol that was newly established in our laboratory. The hypomethylation of all 6 CpG sites in males (compared to females) was confirmed with p-values in orders of 10^{-6}–10^{-9} (V, Table 4).

Very little differential methylation was associated (FDR < 0.05) with any covariate, other than sex. A couple DMRs (1–2) and/or a few DMCs outside DMRs (1–10) were associated with maternal age and height, maternal smoking during the pregnancy, the newborn infant's Apgar points, the usage of epidural anesthetic during delivery, the year of birth, maternal insulin-treated diabetes, and gestational weight gain (V, Table 1). No such associations were found for maternal pre-pregnancy BMI, earlier miscarriages, the mode of delivery, the newborn infant's birth weight, labor induction or the cosine transformed month of birth.

The goal of Study IV was to investigate, whether any differential methylation in umbilical cord blood would be associated with the child's later progression to type 1 diabetes. Before the inflation of spatially adjusted p-values was discovered, 28 genomic regions were thought to be differentially methylated between the cases, who became persistent for at least two islet autoantibodies during the follow-up, and the controls, who remained healthy. We had already tried to validate these observations technically by targeted pyrosequencing at five selected regions, but the pyrosequencing results showed no evidence for differential methylation (IV, Supplementary Results). This lack of technical validation further confirms that discoveries based on spatially adjusted p-values (with standard FDR control) are often false discoveries.

After the more conservative empirical FDR control was applied, we were left with one region that had some weak evidence of differential methylation. The region is on an intron of gene PKP3 and contains two CpG sites (chr11:400288 and chr11:400295) with spatially adjusted p-values in the order of 10^{-13}. Re-runs of PQLseq and spatial adjustment with three different permuted versions of the class covariate yielded some p-values that were of the same order of magnitude but not smaller than the p-values that indicated an association between this region and the real class covariate. Technical validation by pyrosequencing showed that this difference was not significant.

6 Discussion

Alternative splicing events, which enable the production of a variety of proteins from each gene, have been estimated to take place in at least 95 % of human genes (272). Exon microarray technology enabled the study of genome-wide alternative splicing. Although the later-developed RNA sequencing technology is more flexible, exon arrays still remain a popular choice due to their lower cost and simpler data analysis workflow (273). This is reflected by further method development for differential splicing analysis, as well as the usage of the probe-level splicing index, developed 12 years ago within Study I, in relatively recent publications (274,275).

An important limitation of Study I is that the methods were evaluated with respect to their ability to rank differentially spliced exons, and significance thresholds were not discussed. Evaluation criteria included, for example, overlaps of top k=100, 500, 1000, 1500, and 2000 top ranked exons detected using independent subsets of samples. Also, AUROC values are measures of the ability to rank features. To evaluate the sensitivity to detect RT-PCR confirmed brain-specific exon inclusion/exclusion events in an independent data set at FDR ≤ 0.01, the FDR thresholds were estimated empirically through permutations of sample labels. This procedure was the same for all compared methods and did not evaluate sensitivities at method-specific significance thresholds. The reason for this is that the primary goal was to compare the novel probe-level estimate to approaches that summarize probe intensities to exon-level values already at the preprocessing stage. The current implementation of PECA-SI (267) uses the ordinary t-test or the moderated t-test (through limma (187)) for significance testing, but in itself, probe-level splicing index is not a test.

Later-developed differential splicing detection methods for exon array data include information theoretic approaches, such as Alternative splicing Robust Entropy (ARH) (276). To compute ARH, differential splicing probabilities of exons are first estimated based on deviations of exon-specific log fold expression changes (between the compared groups) from the median log fold change over all exons within the same gene. The Shannon entropy of such a probability distribution for a given gene is large, if all exons within the gene have similar splicing probabilities, and small if only one exon deviates. To be a useful metric for alternative splicing, the entropy needs to be normalized to the number of exons within the gene, as well as the overall range of exon deviations. ARH was the best-performing method in a comparison study that attempted to include all exon-level differential splicing detection methods available at that time for exon array data (273). Although this comparison did not include PECA-SI (since the authors decided to limit the study to exon-level methods), our observations on the other methods received further confirmation. Despite being among the oldest methods in the comparison, FIRMA (240) was still among the best performers. The authors denote that methods based on classical statistical testing, such as the ANOVA approaches MIDAS (238) and ANOSVA (239) (Section 4.1), have poor sensitivity compared to approaches that estimate significance with a permutation test (273).

Study II explored longitudinal gene expression profiles of children at risk of type 1 diabetes, half of which developed multiple islet autoantibodies and/or clinical type 1 diabetes during the follow-up within the DIPP project. The long-term follow-up of thousands of DIPP study participants enabled us to include some samples that were collected already before the case individuals' seroconversion to islet autoantibody positivity. Together with a simultaneously published gene expression study of a German type 1 diabetes cohort (100), ours was the first study to report type 1 diabetes associated gene expression patterns that were already present before seroconversion.

Our study provided further support for a possible role of innate immune pathways in the initiation of type 1 diabetes, reviewed for example by (25). Importantly, a similar type I interferon signature preceding seroconversion was observed in an independent prospective type 1 diabetes study, which also associated the innate immune activity temporally with upper respiratory infections in the studied individuals (100). Together these results have contributed to the hypothesis that some virus infections might trigger the autoimmune process that leads to type 1 diabetes (277), reviewed in Section 2.1.5.

We identified an upregulated interconnected network of genes related to the innate viral response, induced by toll-like receptor 5 and regulated by interferon regulatory factors IRF5 and IRF7. Out of 12 target genes of IRF7 that were upregulated in our data, 8 were also identified as being part of an IRF7-driven regulatory network that was associated with type 1 diabetes through an enrichment analysis of GWAS SNPs (278).

Study III was part of the DIABIMMUNE project, which is motivated by the hygiene hypothesis (247,248). Gene expression microarrays were used to identify transcriptome-wide differences that are already present at the time of birth in three locations with contrasting standards of living: Finland (representing a modern environment), Estonia (a country in rapid transition) and Russian Karelia (traditional environment). Genes with functions in innate immune responses were upregulated in Russian Karelia (as compared to Finland and/or Estonia), including toll-like receptor 2 (TLR2), which is highly expressed in leukocytes responding to Gram-positive bacteria already at the time of birth (279). TLR2 has been earlier observed to be upregulated in blood samples from children living in environments with higher microbial abundance (280). Downregulation of TLR2 has also been observed in umbilical cord blood samples of children with allergic mothers, as compared to children of non-allergic mothers (281,282).

We observed a large and highly significant overlap between the genes upregulated in Petrozavodsk vs. Finland and/or Estonia and genes upregulated in 1-year-old children vs. neonates in two independent studies (259,260). Our results suggest that the immune system of Russian Karelian neonates might be more mature than the immune system of neonates born in environments with higher standards of living and lower maternal exposure to pathogens (Finland and Estonia). Based on earlier household dust and drinking water sample analyses, Finland and Russian Karelia are strikingly different living environments, Russian Karelian residents being exposed to a much higher bacterial load dominated by Gram-positive bacteria (283,284). We hypothesize that this bacterial exposure could affect the immune system development already in-utero. Similar conclusions were made in a study that compared neonatal antigen presenting cells (APCs) between Australia (modern environment) and Papua New Guinea (more traditional environment), where the Papua New Guinean neonatal APCs were characterized by markers of maturation, presumably explained by in-utero exposure to microbial antigens (285). A positive correlation between neonatal immune system maturity and maternal exposure to pathogens has also been suggested in a study that compared T and B cell maturation markers between European and African umbilical cord blood samples (286).

Other than the HLA haplotype, we did not account for genetic markers that might explain some of the observed differences in gene expression. Based on a genome-wide SNP analysis of 1564 European samples, the Estonian and Russian populations are genetically close to each other, whereas the Finnish population is fundamentally different from the rest of Europe (287). Conversely, we observed very few transcriptomic differences between the Finnish and Estonian samples compared to the extent of differential expression observed between Estonia and Russian Karelia (III, Figure 2).

Batch effects were kept at minimum by immediate sample collection in RNA stabilizing tubes and storage in –70°C after birth. Furthermore, the RNA isolation and amplification steps were centralized (to one person), and the array hybridizations were done by a robot in the Finnish Microarray and Sequencing Centre. Unfortunately, the much later arrival of Russian Karelian samples and the consequent sample processing and hybridization on a different date than the Estonian and Finnish samples, might have confounded our observations. We addressed this issue by re-hybridizing a subset of the samples from each location in one batch, estimating the batch effect with an empirical Bayes method, and adjusting the data accordingly, as implemented in R package ComBat (176). However, such batch effect correction approaches have been later shown to increase false discoveries (178), as discussed in Section 2.3.2.2.

The scope of our study (III) was limited to overall transcriptomic differences between the three populations at one time point (at birth in children born between January and May 2010). Environmental/lifestyle characteristics, other than the place of birth, were not recorded within this study, and neither did we explore associations between transcriptomic patterns and later allergies or immune-mediated diseases. Both of these aspects have later been studied within the DIABIMMUNE project (109,288–290). In line with our observations, a gut microbiome analysis of 74 infants from each location (Finland, Estonia, and Russian Karelia) revealed major differences between Russian Karelia and the other two locations but little difference between Finland and Estonia (291). The early lipopolysaccharide (LPS) exposures of Russian Karelian infants were dominated by *Esterichia coli*, which was shown to decrease the incidence and delay the onset of diabetes in susceptible mice (non-obese diabetic mice), unlike the structurally distinct LPS of *Bacteroides dorei*, which prevailed in Finnish and Estonian stool samples (291).

The data sets released in Gene Expression Omnibus (accession numbers GSE30211 and GSE53473) within this thesis have been valuable resources for further study of biological questions related to type 1 diabetes (292,293), as well as for method development in machine learning and statistics (294). We have emphasized the value of real data for example in Study I, where – in addition to more traditional criteria, such AUROC in simulated data – the exon array data analysis methods were evaluated with respect to their sensitivity to detect exon skipping events that were confirmed by RT-PCR in independent data. Simulations are important but they often over-simplify the real world. For example, the spatial adjustment through an autocorrelation corrected Z-test implemented within RADMeth has performed well in simulated data (167,222,226,227) but turned out to strongly inflate p-values in the real data presented in IV and V. Some characteristics of real DNA methylation data, such as missing values and the bi-modal distribution (high peaks at both extremes, since CpG sites are often totally unmethylated or totally methylated) are often absent in simulated data. Moreover, usually only one covariate is simulated to have a true effect, whereas real DNA methylation can be influenced by several factors. Empirical FDR control, which decreased type 1 error rate and completely changed our conclusions in Studies IV and V, was only possible through the bisulfite sequencing of a relatively large number of independent samples (N=173).

As reviewed in Section 2.1.6, very little is known of the possible association between early-childhood DNA methylation and later progression to type 1 diabetes.

Associations of umbilical cord blood DNA methylation patterns with other variables such as maternal smoking, maternal BMI, birth weight and gestational age have been explored in large meta-analyses of published DNA methylation microarray data (295–298). Most published studies on DNA methylation have been performed with microarrays and are hence limited to the CpG sites that are targeted by the probes on the array. Bisulfite sequencing data usually covers a much larger number of CpG sites and is better suited to find methylation patterns across regions. The unique bisulfite sequencing data produced within Studies IV and V have potential to be useful for future studies on umbilical cord blood DNA methylation. Since sequencing data is sensitive by nature, the raw data could not be published. However, we released the processed counts (numbers of methylated reads and total coverage at each CpG site in each sample), technical covariates, sex, and class labels in ArrayExpress (accession code E-MTAB-10530, available as soon as IV is accepted for publication). Data on other clinical variables could not be published for privacy reasons but is available from the corresponding author of IV upon reasonable request. We expect the published data to be valuable for further testing and development of differential methylation analysis methods and for future studies on the epigenomics of type 1 diabetes.

The transcriptomic and epigenomic analysis of early molecular mechanisms of type 1 diabetes presented within this thesis (II, III, and IV), as well as the epigenomic study on newborn infants and pregnancy-related variables (V), are limited to whole blood. To gain insight into the roles of different cell types and other functional entities that might contribute to the type 1 diabetes associated transcriptomic signature in Study II, we performed an enrichment analysis on literature-annotated modules of co-expressed genes in human blood (208). The module map of our differentially expressed genes in different time windows (II, Figure 4) further emphasizes the presence of an interferon signature throughout the follow-up. Also, genes associated with neutrophils and cytotoxic cells were differentially expressed from the time before seroconversion until the diagnosis with type 1 diabetes. Cytotoxic T cells have a key role in the pancreatic lesions of people newly diagnosed with type 1 diabetes (23) but this activity was not expected to be reflected in peripheral blood. The observed downregulation of neutrophil-associated genes at the time of diagnosis is concordant with an earlier publication reporting decreased circulating neutrophil counts in children and adults with newly-onset type 1 diabetes (299).

Later studies on type 1 diabetes have further explored cell-type-specific gene expression patterns even on the level of single cells (109). Given the cost of single-cell ’omics or cell fractionation followed by sequencing of several fractions of each sample, such studies typically include only a limited number of samples. The only published study on DNA methylation changes that precede type 1 diabetes is limited to peripheral whole blood (90). DNA methylation patterns in cell fractions isolated from early samples in prospective type 1 diabetes study cohorts should be explored in future studies.

The studies presented within this thesis are observational and have focused on hypothesis generation rather than testing. The expression patterns associated with type 1 diabetes in Study II and with in-utero environment in Study III are yet to be replicated in larger data sets. However, as discussed above, the main conclusions of these studies have been presented by other independent studies as well. The virus hypothesis (reviewed in Section 2.1.5) remains to be confirmed by vaccination trials, which are currently at phase 1 in Finland. An impact of in-utero environment on the neonatal immune maturity is nowadays supported by increasing evidence (300). Significant associations between umbilical cord blood DNA methylation and later progression to type 1 diabetes were not discovered in Study IV but might be discoverable in a larger or a more homogeneous data set.

7 Summary/Conclusions

Our studies on type 1 diabetes have contributed to the understanding of early mechanisms behind the immune tolerance failure, which develops before any clinical signs of autoimmunity. The prospective study on global gene expression patterns in diabetic children and their matched controls (II) provided further support for the hypothesis that virus infections could play a role in the pathogenesis of type 1 diabetes. The longitudinal RNA profiles were characterized by a type I interferon signature, which was also observed in an independent German study (100). Importantly, these were the first studies to report gene expression patterns associated with later type 1 diabetes that were observable in peripheral blood already before the islet autoantibody positivity. We also reported that immunologically relevant differences were present already at the time of birth between children born in different environments with contrasting incidences of type 1 diabetes (III). However, type 1 diabetes associated epigenetic differences were not yet present at the time of birth in our data (IV).

In addition to providing resources for the further study of biological questions related to type 1 diabetes, this thesis has contributed to the analysis workflows in transcriptomics and epigenomics. A probe-level splicing index was shown to be a useful metric to rank differential splicing events between two groups of samples, as compared to methods that summarize probe intensities to exon intensities already at the preprocessing stage. This method was developed for Affymetrix exon array technology, which is still widely used, despite the availability of RNA sequencing.
Bisulfite sequencing of a relatively large number of samples (IV and V) enabled a permutation-based significance analysis, which revealed a serious inflation of p-values caused by a commonly used spatial adjustment strategy. Adjustment of p-values by a weighted Z-test is a quick way to account for the spatial correlation in bisulfite sequencing data. However, for future studies we recommend either estimating the significance threshold empirically, as done here, or not using spatially adjusted p-values for significance assessment at all.

Acknowledgements

This work was conducted within Riitta Lahesmaa’s Molecular Systems Immunology group in Turku Bioscience Centre (University of Turku) and Harri Lähdesmäki’s Computational Systems Biology group in Aalto University. Words are not enough to thank you Riitta and Harri for supervising this work. You have listened to my detailed updates, provided valuable insight, resources, and encouragement throughout the years, and been everything a supervisor should be.

I am grateful to Ville Mustonen and Niina Sandholm for reviewing this thesis very efficiently but thoroughly. I thank Lude Franke for accepting the invitation to be the opponent and look forward to your visit.

I acknowledge Turku University Doctoral Programme of Molecular Medicine (TuDMM) for funding this thesis and for organizing the doctoral training in Turku. I have enjoyed several training and networking events organized by TuDMM. I thank my thesis committee member Jukka Heikkonen for career tips and discussions in our yearly meetings.

I am thankful for personal grants received from the Finnish Cultural Foundation and Kyllikki and Uolevi Lehikoinen Foundation. The studies within this thesis were also funded by the Academy of Finland and the Sigrid Jusélius Foundation. I acknowledge the Aalto Science-IT project and the Finnish Centre for Scientific Computing (CSC) for computational resources. I thank Turku Bioscience Centre's core facility the Finnish Functional Genomics Centre (FFGC), supported by Biocenter Finland, for the microarray and sequencing services.

The studies conducted within this thesis are interdisciplinary and would not have been possible without the expertise of my co-authors. I would especially like to thank Henna Kallionpää for all the time we spent together, your vital contribution to Studies II and III, and for supervising my M. Sc. thesis. I am thankful to Laura Elo for supervising me as an undergraduate student, enabling my participation in Study I, and being a co-author in several of my publications. I thank Viivi Halla-aho for a valuable introduction to bisulfite sequencing data analysis in the beginning of Studies IV and V and for active discussions throughout the years. I thank Ubaid Ullah for the biological insight and Toni Grönroos for the pyrosequencing in Studies IV and V. I am grateful to Riikka Lund for enabling the RRBS and bisulfite pyrosequencing in Studies IV and V and being an active co-author. I have also enjoyed lunch breaks including both non-scientific and scientific conversations with you!

I thank Riitta Veijola, Olli Simell, Tuula Simell, Jorma Toppari, Mikael Knip, Heikki Hyöty, and Jorma Ilonen for their long-term commitment to the DIPP study and for giving us the opportunity to be part of it. I am also grateful to Mari Vähä-Mäkilä, Mirja Nurmio, and Juha Mykkänen for providing the clinical information for the DIPP projects within this thesis (Study II, IV, and V). I thank Mikael Knip for being in charge of the DIABIMMUNE study. It has been a privilege to work with you! I acknowledge the whole DIABIMMUNE study group, especially Natalya V. Dorshakova, Vallo Tillman, Taina Härkönen, and Jorma Ilonen for your contribution to Study III. I acknowledge Tero Aittokallio for his contribution to Study I, and Matej Orešič, Mikko Konki, Asta Laiho, and Bishwa R. Ghimire for Studies IV and V. I am grateful to Cisca Wijmenga for her insight into Study II, for being part of my thesis committee, and for hosting us twice in Groningen. I enjoyed it very much! I thank Isis Ricaño-Ponce for a very helpful introduction to the eQTL analysis.

I thank all the present and former members of Lahesmaa lab and Lähdesmäki group, especially Syed Bilal Ahmad Andrabi, Ilona Arnkil, Kanchan Bala, Kedar Batkulwar, Santosh Bhosale, Rahul Biradar, Tanja Buchacher, Zhi Chen, Sanna Edelman, Marjo Hakkarainen, Sarita Heinonen, Mirkka Heinonen, Karoliina Hirvonen, Saara Hämälistö, Terhi Jokilehto, Päivi Junni, Roosa Kattelus, Meraj Hasan Khan, Moin Khan, Minna Kyläniemi, Anne Lahdenperä, Johanna Lammela, Sari Lehtimäki, Niina Lietzén, Maritta Löytömäki, Tapio Lönnberg, Robert Moulder, Elisa Närvä, Elina Pietilä, Nelly Rahkonen, Omid Rasool, Maheswara Emani Reddy, Emilie Rydgren, Roosa Sahla, Jussi Salmi, Verna Salo, Alexey Sarapulov, Ankitha Shetty, Inna Starskaia, Lea Toikka, Subhash Tripathi, Soile Tuomela, and Viveka Öling from Lahesmaa lab and Lu Cheng, Charles Gadd, Viivi Halla-Aho, Markus Heinonen, Jani Huuhtanen, Jukka Intosalmi, Emmi Jokinen, Kartiek Kanduri, Lingjia Kong, Juhani Kähärä, Antti Lankinen, Antti Larjo, Maia Malonzo, Henrik Mannerström, Kari Nousiainen, Maria Osmala, Siddharth Ramchandran, Sini Rautio, Emmi Rehn, Juhi Somani, Gleb Tikhonov, Juho Timonen, Matti Vaarma, Tommi Vatanen, Cagatay Yildiz, Tarmo Äijö, and Mine Ögretir from Lähdesmäki group. I learned a lot from you and enjoyed the time we spent together!

I am grateful to Satu Mustjoki for giving me the opportunity to start a post-doc while finalizing the PhD and to all University of Helsinki Hematology Research Unit members for a warm welcome and the work we have done together so far! I've been impressed by our projects, your expertise, and the friendly atmosphere.

I thank Ansa and the workers of Touhula, especially Sanna, Aino, Sanni, Taru, Annu, Petra, Miia, Sussu, Tiia, Jenni, Jani, Jatta, and Susan, and Irene from school, for the great job you have done taking care of our children! Being able to leave them in good hands every morning is a great privilege and absolutely vital for completing a thesis.

I have been blessed with an amazing childhood family mum, dad, Isla, and Antero, and extended family Emilia, Edith, Axel, Robert, Eino, Sini, Erkki, Aada, Jooa, Visa, Tiina, Pekka, Tilde, Jari, and Leena. I love you, and your support is a huge strength! You (especially mummi, tuta and mami) have very concretely contributed to this thesis by taking care of Elea and Vilja when we needed to focus on work. I would also like to thank Olli, Kati, Krista, Juhani, Osmi, Matti, Elina, Teija, and Ansku for sharing the ups and downs of life and parenting during the past seven years. You mean a lot to me!

I will never cease to praise the creator of heaven and earth for the miracle of life and the ability of the human brain to be conscious of it and to study it. I am thankful for every day I got to live as a PhD student, for the people I met along the way and for every bit of knowledge and understanding I gained. I am especially thankful for being interrupted halfway through by the birth of my two favorite people on earth: Elea and Vilja.

Rakkaat Elea ja Villi Villiäinen, maailman ihanimmat tytöt, kiitos kun olette olemassa ja kiitos kaikista haleista ja pusuista! Rakastan teitä 84000 maapallon kokoisen paljon! Rakas pikku Peukku, sinä olet vielä niin pieni, että ollaan nähty sinut vasta ultraäänikuvissa, mutta odotetaan kovin, että päästään tutustumaan sinuun. Kiitos kun tulit meille!

Most importantly, I am thankful to my favorite colleague, co-author, study-time friend and husband, Daniel, for sharing your whole life with me for the past 13 years. I admire your wisdom and patience as a parent, your intelligence and creativity as a data scientist / game developer, and the way you care for people around you. I wish I could be the kind of support to you, that you have been to me and many others. I love you Dee!

November 1st 2021
Essi Laajala

References

1. White MF, Kahn CR. The insulin signaling system. J Biol Chem. 1994 Jan 7;269(1):1–4.
2. Kolluru GK, Bir SC, Kevil CG. Endothelial dysfunction and diabetes: effects on angiogenesis, vascular remodeling, and wound healing. Int J Vasc Med. 2012 Feb 12;2012:918267.
3. Livingstone SJ, Levin D, Looker HC, Lindsay RS, Wild SH, Joss N, et al. Estimated life expectancy in a Scottish cohort with type 1 diabetes, 2008-2010. JAMA. 2015 Jan 6;313(1):37.
4. Huo L, Harding JL, Peeters A, Shaw JE, Magliano DJ. Life expectancy of type 1 diabetic patients during 1997–2010: a national Australian registry-based cohort study. Diabetologia. 2016 Jun;59(6):1177–85.
5. Insel RA, Dunne JL, Atkinson MA, Chiang JL, Dabelea D, Gottlieb PA, et al. Staging presymptomatic type 1 diabetes: a scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association. Diabetes Care. 2015 Oct;38(10):1964–74.
6. Sosenko JM, Palmer JP, Rafkin-Mervis L, Krischer JP, Cuthbertson D, Mahon J, et al. Incident dysglycemia and progression to type 1 diabetes among participants in the Diabetes Prevention Trial-Type 1. Diabetes Care. 2009 Sep;32(9):1603–7.
7. Knip M, Korhonen S, Kulmala P, Veijola R, Reunanen A, Raitakari OT, et al. Prediction of type 1 diabetes in the general population. Diabetes Care. 2010 Jun;33(6):1206–12.
8. Ziegler AG, Rewers M, Simell O, Simell T, Lempainen J, Steck A, et al. Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. JAMA. 2013 Jun 19;309(23):2473.
9. Pöllänen PM, Lempainen J, Laine A-P, Toppari J, Veijola R, Vähäsalo P, et al. Characterisation of rapid progressors to type 1 diabetes among children with HLA-conferred disease susceptibility. Diabetologia. 2017 Jul;60(7):1284–93.
10. Simmons K, Sosenko J, Ismail HM, Larsson HE, Steck A. One-hour oral glucose tolerance tests (OGTTs) for the prediction and diagnostic surveillance of type 1 diabetes (T1D). Diabetes. 2018 May;67(Supplement 1):1500-P.
11. Flier JS, Underhill LH, Eisenbarth GS. Type I diabetes mellitus. N Engl J Med. 1986 May 22;314(21):1360–8.
12. Atkinson MA, von Herrath M, Powers AC, Clare-Salzler M. Current concepts on the pathogenesis of type 1 diabetes--considerations for attempts to prevent and reverse the disease. Diabetes Care. 2015 Jun;38(6):979–88.
13. In't Veld P. Insulitis in human type 1 diabetes: a comparison between patients and animal models. Semin Immunopathol. 2014 Sep;36(5):569–79.
14. Meier JJ, Bhushan A, Butler AE, Rizza RA, Butler PC. Sustained beta cell apoptosis in patients with long-standing type 1 diabetes: indirect evidence for islet regeneration? Diabetologia. 2005 Nov;48(11):2221–8.
15. Keenan HA, Sun JK, Levine J, Doria A, Aiello LP, Eisenbarth G, et al. Residual insulin production and pancreatic ß-cell turnover after 50 years of diabetes: Joslin Medalist Study. Diabetes. 2010 Nov;59(11):2846–53.

16. Löhr M, Klöppel G. Residual insulin positivity and pancreatic atrophy in relation to duration of chronic type 1 (insulin-dependent) diabetes mellitus and microangiopathy. Diabetologia. 1987 Oct;30(10):757–62.
17. Dotta F, Censini S, van Halteren AGS, Marselli L, Masini M, Dionisi S, et al. Coxsackie B4 virus infection of beta cells and natural killer cell insulitis in recent-onset type 1 diabetic patients. Proc Natl Acad Sci U S A. 2007 Mar 20;104(12):5115–20.
18. Yin H, Berg A-K, Westman J, Hellerström C, Frisk G. Complete nucleotide sequence of a Coxsackievirus B-4 strain capable of establishing persistent infection in human pancreatic islet cells: effects on insulin release, proinsulin synthesis, and cell morphology. J Med Virol. 2002 Dec;68(4):544–57.
19. In't Veld P, Lievens D, De Grijse J, Ling Z, Van der Auwera B, Pipeleers-Marichal M, et al. Screening for insulitis in adult autoantibody-positive organ donors. Diabetes. 2007 Sep;56(9):2400–4.
20. Imagawa A, Hanafusa T, Itoh N, Waguri M, Yamamoto K, Miyagawa J, et al. Immunological abnormalities in islets at diagnosis paralleled further deterioration of glycaemic control in patients with recent-onset Type I (insulin-dependent) diabetes mellitus. Diabetologia. 1999 May;42(5):574–8.
21. von Herrath M, Sanda S, Herold K. Type 1 diabetes as a relapsing-remitting disease? Nat Rev Immunol. 2007 Dec;7(12):988–94.
22. Willcox A, Richardson SJ, Bone AJ, Foulis AK, Morgan NG. Analysis of islet inflammation in human type 1 diabetes. Clin Exp Immunol. 2009 Feb;155(2):173–81.
23. Coppieters KT, Dotta F, Amirian N, Campbell PD, Kay TWH, Atkinson MA, et al. Demonstration of islet-autoreactive CD8 T cells in insulitic lesions from recent onset and long-term type 1 diabetes patients. J Exp Med. 2012 Jan 16;209(1):51–60.
24. Vig S, Buitinga M, Rondas D, Crèvecoeur I, van Zandvoort M, Waelkens E, et al. Cytokine-induced translocation of GRP78 to the plasma membrane triggers a pro-apoptotic feedback loop in pancreatic beta cells. Cell Death Dis. 2019 Apr 5;10(4):309.
25. Needell JC, Zipris D. Targeting innate immunity for type 1 diabetes prevention. Curr Diab Rep [Internet]. 2017 Nov;17(11). Available from: http://dx.doi.org/10.1007/s11892-017-0930-z
26. Chiou J, Geusz RJ, Okino M-L, Han JY, Miller M, Melton R, et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. Nature. 2021 Jun;594(7863):398–402.
27. Andersson C, Kolmodin M, Ivarsson S-A, Carlsson A, Forsander G, Lindblad B, et al. Islet cell antibodies (ICA) identify autoimmunity in children with new onset diabetes mellitus negative for other islet cell antibodies. Pediatr Diabetes. 2014 Aug;15(5):336–44.
28. Vehik K, Bonifacio E, Lernmark A, Yu L, Williams A, Schatz D, et al. Hierarchical order of distinct autoantibody spreading and progression to type 1 diabetes in the TEDDY study. Diabetes Care. 2020 Jul 8;43(9):2066–73.
29. Krischer JP, Lynch KF, Schatz DA, Ilonen J, Lernmark Å, Hagopian WA, et al. The 6 year incidence of diabetes-associated autoantibodies in genetically at-risk children: the TEDDY study. Diabetologia. 2015 May;58(5):980–7.
30. Maclaren N, Lan M, Coutant R, Schatz D, Silverstein J, Muir A, et al. Only multiple autoantibodies to islet cells (ICA), insulin, GAD65, IA-2 and IA-2beta predict immune-mediated (Type 1) diabetes in relatives. J Autoimmun. 1999 Jun;12(4):279–87.
31. Gianani R, Campbell-Thompson M, Sarkar SA, Wasserfall C, Pugliese A, Solis JM, et al. Dimorphic histopathology of long-standing childhood-onset diabetes. Diabetologia. 2010 Apr;53(4):690–8.
32. Leete P, Willcox A, Krogvold L, Dahl-Jørgensen K, Foulis AK, Richardson SJ, et al. Differential insulitic profiles determine the extent of β-cell destruction and the age at onset of type 1 diabetes. Diabetes. 2016 May;65(5):1362–9.

33. Leete P, Oram RA, McDonald TJ, Shields BM, Ziller C, TIGI study team, et al. Studies of insulin and proinsulin in pancreas and serum support the existence of aetiopathological endotypes of type 1 diabetes associated with age at diagnosis. Diabetologia. 2020 Jun;63(6):1258–67.
34. Familial risk of type I diabetes in European children. The Eurodiab Ace Study Group and The Eurodiab Ace Substudy 2 Study Group. Diabetologia. 1998 Oct;41(10):1151–6.
35. Kostraba JN, Gay EC, Cai Y, Cruickshanks KJ, Rewers MJ, Klingensmith GJ, et al. Incidence of insulin-dependent diabetes mellitus in Colorado. Epidemiology. 1992 May;3(3):232–8.
36. Redondo MJ, Jeffrey J, Fain PR, Eisenbarth GS, Orban T. Concordance for islet autoimmunity among monozygotic twins. N Engl J Med. 2008 Dec 25;359(26):2849–50.
37. Hyttinen V, Kaprio J, Kinnunen L, Koskenvuo M, Tuomilehto J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. Diabetes. 2003 Apr;52(4):1052–5.
38. DIAMOND Project Group. Incidence and trends of childhood Type 1 diabetes worldwide 1990-1999. Diabet Med. 2006 Aug;23(8):857–66.
39. Patterson CC, Dahlquist GG, Gyürüs E, Green A, Soltész G, EURODIAB Study Group. Incidence trends for childhood type 1 diabetes in Europe during 1989-2003 and predicted new cases 2005-20: a multicentre prospective registration study. Lancet. 2009 Jun 13;373(9680):2027–33.
40. Bodansky HJ, Staines A, Stephenson C, Haigh D, Cartwright R. Evidence for an environmental effect in the aetiology of insulin dependent diabetes in a transmigratory population. BMJ. 1992 Apr 18;304(6833):1020–2.
41. Söderström U, Aman J, Hjern A. Being born in Sweden increases the risk for type 1 diabetes - a study of migration of children to Sweden as a natural experiment. Acta Paediatr. 2012 Jan;101(1):73–7.
42. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009 Jun;41(6):703–7.
43. Onengut-Gumuscu S, Type 1 Diabetes Genetics Consortium, Chen W-M, Burren O, Cooper NJ, Quinlan AR, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. 2015 Apr;47(4):381–6.
44. Bach J-F, Chatenoud L. The hygiene hypothesis: an explanation for the increased frequency of insulin-dependent diabetes. Cold Spring Harb Perspect Med. 2012 Feb;2(2):a007799.
45. Vaarala O, Knip M, Paronen J, Hämäläinen AM, Muona P, Väätäinen M, et al. Cow's milk formula feeding induces primary immunization to insulin in infants at genetic risk for type 1 diabetes. Diabetes. 1999 Jul;48(7):1389–94.
46. Knip M, Akerblom HK. Early nutrition and later diabetes risk. Adv Exp Med Biol. 2005;569:142–50.
47. Chia JSJ, McRae JL, Kukuljan S, Woodford K, Elliott RB, Swinburn B, et al. A1 beta-casein milk protein and other environmental pre-disposing factors for type 1 diabetes. Nutr Diabetes. 2017 May;7(5):e274–e274.
48. Zipitis CS, Akobeng AK. Vitamin D supplementation in early childhood and risk of type 1 diabetes: a systematic review and meta-analysis. Arch Dis Child. 2008 Jun;93(6):512–7.
49. Svoren BM, Volkening LK, Wood JR, Laffel LMB. Significant vitamin D deficiency in youth with type 1 diabetes mellitus. J Pediatr. 2009 Jan;154(1):132–4.
50. Dahlquist GG, Pundziūte-Lyckå A, Nyström L, Swedish Childhood Diabetes Study Group, Diabetes Incidence Study in Sweden (DISS) Group. Birthweight and risk of type 1 diabetes in children and young adults: a population-based register study. Diabetologia. 2005 Jun;48(6):1114–7.

51. Lamb MM, Yin X, Zerbe GO, Klingensmith GJ, Dabelea D, Fingerlin TE, et al. Height growth velocity, islet autoimmunity and type 1 diabetes development: the Diabetes Autoimmunity Study in the Young. Diabetologia. 2009 Oct;52(10):2064–71.
52. Liu X, Vehik K, Huang Y, Elding Larsson H, Toppari J, Ziegler AG, et al. Distinct growth phases in early life associated with the risk of type 1 diabetes: The TEDDY study. Diabetes Care. 2020 Mar;43(3):556–62.
53. Knip M, Simell O. Environmental triggers of type 1 diabetes. Cold Spring Harb Perspect Med. 2012 Jul;2(7):a007690.
54. McGlinchey A, Sinioja T, Lamichhane S, Sen P, Bodin J, Siljander H, et al. Prenatal exposure to perfluoroalkyl substances modulates neonatal serum phospholipids, increasing risk of type 1 diabetes. Environ Int. 2020 Oct;143(105935):105935.
55. Writing Group for the TRIGR Study Group, Knip M, Åkerblom HK, Al Taji E, Becker D, Bruining J, et al. Effect of hydrolyzed infant formula vs conventional formula on risk of type 1 diabetes: The TRIGR randomized clinical trial. JAMA. 2018 Jan 2;319(1):38–48.
56. Petzold A, Solimena M, Knoch K-P. Mechanisms of beta cell dysfunction associated with viral infection. Curr Diab Rep. 2015 Oct;15(10):73.
57. Cooke A, Tonks P, Jones FM, O'Shea H, Hutchings P, Fulford AJ, et al. Infection with Schistosoma mansoni prevents insulin dependent diabetes mellitus in non-obese diabetic mice. Parasite Immunol. 1999 Apr;21(4):169–76.
58. Jun HS, Yoon JW. The role of viruses in type I diabetes: two distinct cellular and molecular pathogenic mechanisms of virus-induced diabetes in animals. Diabetologia. 2001 Mar;44(3):271–85.
59. Härkönen T, Lankinen H, Davydova B, Hovi T, Roivainen M. Enterovirus infection can induce immune responses that cross-react with β-cell autoantigen tyrosine phosphatase IA-2/IAR. J Med Virol. 2002 Mar;66(3):340–50.
60. Honeyman MC, Stone NL, Falk BA, Nepom G, Harrison LC. Evidence for molecular mimicry between human T cell epitopes in rotavirus and pancreatic islet autoantigens. J Immunol. 2010 Feb 15;184(4):2204–10.
61. Honeyman MC, Coulson BS, Stone NL, Gellert SA, Goldwater PN, Steele CE, et al. Association between rotavirus infection and pancreatic islet autoimmunity in children at risk of developing type 1 diabetes. Diabetes. 2000 Aug;49(8):1319–24.
62. Craig ME, Nair S, Stein H, Rawlinson WD. Viruses and type 1 diabetes: a new look at an old story. Pediatr Diabetes. 2013 May;14(3):149–58.
63. Dunne JL, Richardson SJ, Atkinson MA, Craig ME, Dahl-Jørgensen K, Flodström-Tullberg M, et al. Large enteroviral vaccination studies to prevent type 1 diabetes should be well founded and rely on scientific evidence. Reply to Skog O, Klingel K, Roivainen M et al [letter]. Diabetologia. Springer Science and Business Media LLC; 2019 Jun;62(6):1100–3.
64. Wagenknecht LE, Roseman JM, Herman WH. Increased incidence of insulin-dependent diabetes mellitus following an epidemic of Coxsackievirus B5. Am J Epidemiol. 1991 May 15;133(10):1024–31.
65. Oikarinen S, Tauriainen S, Hober D, Lucas B, Vazeou A, Sioofy-Khojine A, et al. Virus antibody survey in different European populations indicates risk association between coxsackievirus B1 and type 1 diabetes. Diabetes. 2014 Feb;63(2):655–62.
66. Yeung W-CG, Rawlinson WD, Craig ME. Enterovirus infection and type 1 diabetes mellitus: systematic review and meta-analysis of observational molecular studies. BMJ. 2011 Feb 3;342(feb03 1):d35.
67. Clements GB, Galbraith DN, Taylor KW. Coxsackie B virus infection and onset of childhood diabetes. Lancet. 1995 Jul 22;346(8969):221–3.
68. Salvatoni A, Baj A, Bianchi G, Federico G, Colombo M, Toniolo A. Intrafamilial spread of enterovirus infections at the clinical onset of type 1 diabetes. Pediatr Diabetes. 2013 Sep;14(6):407–16.

69. Krogvold L, Edwin B, Buanes T, Frisk G, Skog O, Anagandula M, et al. Detection of a low-grade enteroviral infection in the islets of langerhans of living patients newly diagnosed with type 1 diabetes. Diabetes. 2015 May;64(5):1682–7.

70. Gallagher GR, Brehm MA, Finberg RW, Barton BA, Shultz LD, Greiner DL, et al. Viral infection of engrafted human islets leads to diabetes. Diabetes. 2015 Apr;64(4):1358–69.

71. Perrett KP, Jachno K, Nolan TM, Harrison LC. Association of Rotavirus vaccination with the incidence of type 1 diabetes in children. JAMA Pediatr. 2019 Mar 1;173(3):280–2.

72. Gruenbaum Y, Stein R, Cedar H, Razin A. Methylation of CpG sequences in eukaryotic DNA. FEBS Lett. 1981 Feb 9;124(1):67–71.

73. Xin Y, O'Donnell AH, Ge Y, Chanrion B, Milekic M, Rosoklija G, et al. Role of CpG context and content in evolutionary signatures of brain DNA methylation. Epigenetics. 2011 Nov;6(11):1308–18.

74. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017 Jan;49(1):131–8.

75. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009 Nov 19;462(7271):315–22.

76. Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 2013 Sep;41(16):e155.

77. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006 Dec;38(12):1378–85.

78. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. 2016 Jun 20;44(11):5123–32.

79. Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. Science. 2001 Aug 10;293(5532):1089–93.

80. Thornburg KL, Shannon J, Thuillier P, Turker MS. In utero life and epigenetic predisposition for disease. Adv Genet. 2010;71:57–78.

81. Lillycrop KA, Burdge GC. Epigenetic changes in early life and future risk of obesity. Int J Obes (Lond). 2011 Jan;35(1):72–83.

82. Küpers LK, Xu X, Jankipersadsing SA, Vaez A, la Bastide-van Gemert S, Scholtens S, et al. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. Int J Epidemiol. 2015 Aug;44(4):1224–37.

83. Edgar R, Tan PPC, Portales-Casamar E, Pavlidis P. Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. Epigenetics Chromatin. 2014 Oct 23;7(1):28.

84. Li S, Wong EM, Dugué P-A, McRae AF, Kim E, Joo J-HE, et al. Genome-wide average DNA methylation is determined in utero. Int J Epidemiol. 2018 Jun 1;47(3):908–16.

85. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, et al. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics. 2014 Feb 21;15:145.

86. Kindt ASD, Fuerst RW, Knoop J, Laimighofer M, Telieps T, Hippich M, et al. Allele-specific methylation of type 1 diabetes susceptibility genes. J Autoimmun. 2018 May;89:63–74.

87. Fradin D, Le Fur S, Mille C, Naoui N, Groves C, Zelenika D, et al. Association of the CpG methylation pattern of the proximal insulin gene promoter with type 1 diabetes. PLoS One. 2012 May 2;7(5):e36278.

88. Belot M-P, Fradin D, Mai N, Le Fur S, Zélénika D, Kerr-Conte J, et al. CpG methylation changes within the IL2RA promoter in type 1 diabetes of childhood onset. PLoS One. 2013 Jul 12;8(7):e68093.

89. Paul DS, Teschendorff AE, Dang MAN, Lowe R, Hawa MI, Ecker S, et al. Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. Nat Commun. 2016 Nov 29;7:13555.
90. Johnson RK, Vanderlinden LA, Dong F, Carry PM, Seifert J, Waugh K, et al. Longitudinal DNA methylation differences precede type 1 diabetes. Sci Rep. 2020 Feb 28;10(1):3721.
91. Deng MC, Eisen HJ, Mehra MR, Billingham M, Marboe CC, Berry G, et al. Noninvasive discrimination of rejection in cardiac allograft recipients using gene expression profiling. Am J Transplant. 2006 Jan;6(1):150–60.
92. Greco FA, Lennington WJ, Spigel DR, Hainsworth JD. Molecular profiling diagnosis in unknown primary cancer: accuracy and ability to complement standard pathology. J Natl Cancer Inst. 2013 Jun 5;105(11):782–90.
93. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)-Analysis Working Group, Statistical Methods groups-Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017 Oct 11;550(7675):204–13.
94. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017 Jan;49(1):139–45.
95. Yip L, Fuhlbrigge R, Alkhataybeh R, Fathman CG. Gene expression analysis of the pre-diabetic pancreas to identify pathogenic mechanisms and biomarkers of type 1 Diabetes. Front Endocrinol (Lausanne). 2020 Dec 23;11:609271.
96. Yip L, Fathman CG. Type 1 diabetes in mice and men: gene expression profiling to investigate disease pathogenesis. Immunol Res. 2014 May;58(2–3):340–50.
97. Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, et al. A single-cell transcriptome atlas of the human pancreas. Cell Syst. 2016 Oct;3(4):385-394.e3.
98. Ramos-Rodríguez M, Raurell-Vila H, Colli ML, Alvelos MI, Subirana-Granés M, Juan-Mateu J, et al. The impact of proinflammatory cytokines on the β-cell regulatory landscape provides insights into the genetics of type 1 diabetes. Nat Genet. 2019 Nov;51(11):1588–95.
99. Cabrera SM, Chen Y-G, Hagopian WA, Hessner MJ. Blood-based signatures in type 1 diabetes. Diabetologia. 2016 Mar;59(3):414–25.
100. Ferreira RC, Guo H, Coulson RMR, Smyth DJ, Pekalski ML, Burren OS, et al. A type I interferon transcriptional signature precedes autoimmunity in children genetically at risk for type 1 diabetes. Diabetes. 2014 Jul;63(7):2538–50.
101. Rassi DM, Junta CM, Fachin AL, Sandrin-Garcia P, Mello S, Marques MMC, et al. Metabolism genes are among the differentially expressed ones observed in lymphomononuclear cells of recently diagnosed type 1 diabetes mellitus patients. Ann N Y Acad Sci. 2006 Oct;1079(1):171–6.
102. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC. Gene expression in peripheral blood mononuclear cells from children with diabetes. J Clin Endocrinol Metab. 2007 Sep;92(9):3705–11.
103. Reynier F, Pachot A, Paye M, Xu Q, Turrel-Davin F, Petit F, et al. Specific gene expression signature associated with development of autoimmune type-I diabetes using whole-blood microarray analysis. Genes Immun. 2010 Apr;11(3):269–78.
104. Elo LL, Mykkänen J, Nikula T, Järvenpää H, Simell S, Aittokallio T, et al. Early suppression of immune response pathways characterizes children with prediabetes in genome-wide gene expression profiling. J Autoimmun. 2010 Aug;35(1):70–6.
105. Stechova K, Kolar M, Blatny R, Halbhuber Z, Vcelakova J, Hubackova M, et al. Healthy first-degree relatives of patients with type 1 diabetes exhibit significant differences in basal gene expression pattern of immunocompetent cells compared to controls: expression pattern as predeterminant of autoimmune diabetes. Scand J Immunol. 2012 Feb;75(2):210–9.

106. Jin Y, Sharma A, Carey C, Hopkins D, Wang X, Robertson DG, et al. The expression of inflammatory genes is upregulated in peripheral blood of patients with type 1 diabetes. Diabetes Care. 2013 Sep;36(9):2794–802.
107. Evangelista AF, Collares CVA, Xavier DJ, Macedo C, Manoel-Caetano FS, Rassi DM, et al. Integrative analysis of the transcriptome profiles observed in type 1, type 2 and gestational diabetes mellitus reveals the role of inflammation. BMC Med Genomics. 2014 May 23;7(1):28.
108. Mehdi AM, Hamilton-Williams EE, Cristino A, Ziegler A, Bonifacio E, Le Cao K-A, et al. A peripheral blood transcriptomic signature predicts autoantibody development in infants at risk of type 1 diabetes. JCI Insight [Internet]. 2018 Mar 8;3(5). Available from: http://dx.doi.org/10.1172/jci.insight.98212
109. Kallionpää H, Somani J, Tuomela S, Ullah U, de Albuquerque R, Lönnberg T, et al. Early Detection of Peripheral Blood Cell Signature in Children Developing β-Cell Autoimmunity at a Young Age. Diabetes. 2019 Oct;68(10):2024–34.
110. Heninger A-K, Eugster A, Kuehn D, Buettner F, Kuhn M, Lindner A, et al. A divergent population of autoantigen-responsive CD4 + T cells in infants prior to β cell autoimmunity. Sci Transl Med. 2017 Feb 22;9(378):eaaf8848.
111. Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5350–4.
112. Schulze A, Downward J. Navigating gene expression using microarrays--a technology review. Nat Cell Biol. 2001 Aug;3(8):E190-5.
113. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol. 2013 Jan;Chapter 22:Unit 22.1.
114. Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, et al. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. J Biomol Tech. 2004 Dec;15(4):276–84.
115. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467–70.
116. Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics. 2006 Jun 2;7(1):276.
117. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JFJ, Ritchie ME, Lynch AG, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res. 2010 Jan;38(3):e17.
118. Du P, Kibbe WA, Lin SM. nuID: a universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays. Biol Direct. 2007 May 31;2(1):16.
119. Srinivasan K, Shiue L, Hayes JD, Centers R, Fitzwater S, Loewen R, et al. Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods. 2005 Dec;37(4):345–59.
120. Affymetrix. Identifying and validating alternative splicing events. Affymetrix Technical Notes. 2006.
121. Kapur K, Xing Y, Ouyang Z, Wong WH. Exon arrays provide accurate assessments of gene expression. Genome Biol. 2007;8(5):R82.
122. Ha KC, Coulombe-Huntington J, Majewski J. Comparison of Affymetrix Gene Array with the Exon Array shows potential application for detection of transcript isoform variation. BMC Genomics. 2009 Nov 12;10:519.
123. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27;456(7221):470–6.
124. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015 May 21;58(4):586–97.
125. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. In: Molecular Biology. Elsevier; 1989. p. 595–604.

126. Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, et al. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci U S A. 2008 Jul 8;105(27):9145–50.
127. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011 Apr;6(4):468–81.
128. Hsu F-M, Yen M-R, Wang C-T, Lin C-Y, Wang C-JR, Chen P-Y. Optimized reduced representation bisulfite sequencing reveals tissue-specific mCHH islands in maize. Epigenetics Chromatin [Internet]. 2017 Dec;10(1). Available from: http://dx.doi.org/10.1186/s13072-017-0148-y
129. Wei S, Tao J, Xu J, Chen X, Wang Z, Zhang N, et al. Ten years of EWAS. Adv Sci (Weinh). 2021 Aug 11;e2100727.
130. Oda M, Glass JL, Thompson RF, Mo Y, Olivier EN, Figueroa ME, et al. High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. Nucleic Acids Res. 2009 Jul;37(12):3829–39.
131. Grehl C, Kuhlmann M, Becker C, Glaser B, Grosse I. How to design a whole-genome bisulfite sequencing experiment. Epigenomes. 2018 Dec 11;2(4):21.
132. Tan G, Opitz L, Schlapbach R, Rehrauer H. Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci Rep. 2019 Feb 27;9(1):2856.
133. Krueger F. TrimGalore. A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. TrimGalore (accessed on 27 August 2019). 2016;
134. BLUEPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat Biotechnol. 2016 Jul;34(7):726–37.
135. Ronaghi M. DNA SEQUENCING:A sequencing method based on real-time pyrophosphate. Science. 1998 Jul 17;281(5375):363–5.
136. Hubbell E, Liu W-M, Mei R. Robust estimators for expression analysis. Bioinformatics. 2002 Dec;18(12):1585–92.
137. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 Apr;4(2):249–64.
138. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003 Feb 15;31(4):e15.
139. Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol Syst Biol. 2011 Jun 7;7(1):497.
140. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013 Jun 13;498(7453):236–40.
141. Dar RD, Razooky BS, Singh A, Trimeloni TV, McCollum JM, Cox CD, et al. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proc Natl Acad Sci U S A. 2012 Oct 23;109(43):17454–9.
142. Wu Z, Zhang Y, Stitzel ML, Wu H. Two-phase differential expression analysis for single cell RNA-seq. Bioinformatics. 2018 Oct 1;34(19):3340–8.
143. Tiberi S, Walsh M, Cavallaro M, Hebenstreit D, Finkenstädt B. Bayesian inference on stochastic gene transcription from flow cytometry data. Bioinformatics. 2018 Sep 1;34(17):i647–55.
144. Lee H-J, Suk J-E, Patrick C, Bae E-J, Cho J-H, Rho S, et al. Direct transfer of alpha-synuclein from neuron to astroglia causes inflammatory responses in synucleinopathies. J Biol Chem. 2010 Mar 19;285(12):9262–72.

145. Thompson A, May MR, Moore BR, Kopp A. A hierarchical Bayesian mixture model for inferring the expression state of genes in transcriptomes. Proc Natl Acad Sci U S A. 2020 Aug 11;117(32):19339–46.
146. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, et al. Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol. 2007;8(4):R64.
147. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012 May;40(10):e72.
148. Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One. 2013 Apr 29;8(4):e62856.
149. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol [Internet]. 2018 Dec;19(1). Available from: http://dx.doi.org/10.1186/s13059-018-1408-2
150. Tanaka K, Okamoto A. Degradation of DNA by bisulfite treatment. Bioorg Med Chem Lett. 2007 Apr 1;17(7):1912–5.
151. Shagin DA, Lukyanov KA, Vagner LL, Matz MV. Regulation of average length of complex PCR product. Nucleic Acids Res. 1999 Sep 1;27(18):i–iii.
152. Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics. 2016 Jul 25;17 Suppl 7(S7):239.
153. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012 Oct 3;13(10):R87.
154. Yu X, Guda K, Willis J, Veigl M, Wang Z, Markowitz S, et al. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? BioData Min [Internet]. 2012 Dec;5(1). Available from: http://dx.doi.org/10.1186/1756-0381-5-6
155. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
156. Canzar S, Salzberg SL. Short read mapping: An algorithmic tour. Proc IEEE Inst Electr Electron Eng. 2017 Mar;105(3):436–58.
157. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009 Mar 4;10(3):R25.
158. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.
159. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.
160. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011 Jun 1;27(11):1571–2.
161. Sun X, Han Y, Zhou L, Chen E, Lu B, Liu Y, et al. A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. Bioinformatics. 2018 Aug 15;34(16):2715–23.
162. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics. 2013 Nov 10;14:774.
163. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010 Apr 1;26(7):873–81.
164. Pedersen BS, Eyring K, De S, Yang IV, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads [Internet]. arXiv [q-bio.GN]. 2014. Available from: http://arxiv.org/abs/1401.1129
165. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. J Biotechnol. 2017 Nov;261:105–15.

166. Leontiou CA, Hadjidaniel MD, Mina P, Antoniou P, Ioannides M, Patsalis PC. Bisulfite conversion of DNA: Performance comparison of different kits and methylation quantitation of epigenetic biomarkers that have the potential to be used in non-invasive prenatal testing. PLoS One. 2015 Aug 6;10(8):e0135058.
167. Halla-aho V, Lähdesmäki H. LuxUS: DNA methylation analysis using generalized linear mixed model with spatial correlation. Bioinformatics [Internet]. 2020 Jun 2; Available from: http://dx.doi.org/10.1093/bioinformatics/btaa539
168. Lin X, Sun D, Rodriguez B, Zhao Q, Sun H, Zhang Y, et al. BSeQC: quality control of bisulfite sequencing experiments. Bioinformatics. 2013 Dec 15;29(24):3227–9.
169. Hoelzer K, Shackelton LA, Parrish CR. Presence and role of cytosine methylation in DNA viruses of animals. Nucleic Acids Res. 2008 May;36(9):2825–37.
170. Daca-Roszak P, Pfeifer A, Żebracka-Gala J, Rusinek D, Szybińska A, Jarząb B, et al. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies. BMC Genomics. 2015 Nov 25;16(1):1003.
171. Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol. 2012 Jul 11;13(7):R61.
172. Gao S, Zou D, Mao L, Liu H, Song P, Chen Y, et al. BS-SNPer: SNP calling in bisulfite-seq data. Bioinformatics. 2015 Dec 15;31(24):4006–8.
173. Su Z. Optimal allocation of prognostic factors in randomized preclinical animal studies. Drug Inf J. 2011 Nov;45(6):725–9.
174. Laajala TD, Jumppanen M, Huhtaniemi R, Fey V, Kaur A, Knuuttila M, et al. Optimized design and analysis of preclinical intervention studies in vivo. Sci Rep [Internet]. 2016 Aug;6(1). Available from: http://dx.doi.org/10.1038/srep30723
175. Greevy R, Lu B, Silber JH, Rosenbaum P. Optimal multivariate matching before randomization. Biostatistics. 2004 Apr;5(2):263–75.
176. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118–27.
177. Giordan M. A two-stage procedure for the removal of batch effects in microarray studies. Stat Biosci. 2014 May;6(1):73–84.
178. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics. 2016 Jan;17(1):29–39.
179. Zhu T, Sun R, Zhang F, Chen G-B, Yi X, Ruan G, et al. BatchServer: A web server for batch effect evaluation, visualization, and correction. J Proteome Res [Internet]. 2020 Dec 18;(acs.jproteome.0c00488). Available from: http://dx.doi.org/10.1021/acs.jproteome.0c00488
180. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936;
181. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc. 1995 Jan;57(1):289–300.
182. Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ. Genome-wide significance for dense SNP and resequencing data. Genet Epidemiol. 2008 Feb;32(2):179–85.
183. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. Genet Epidemiol. 2008 Apr;32(3):227–34.
184. Panagiotou OA, Ioannidis JPA, Genome-Wide Significance Project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol. 2012 Feb;41(1):273–86.
185. Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, et al. Estimation of a significance threshold for epigenome-wide association studies. Genet Epidemiol. 2018 Feb;42(1):20–33.

186. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan 4;45(D1):D896–901.
187. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004 Feb 12;3(1):Article3.
188. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014 Feb 3;15(2):R29.
189. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015 Jan;16(1):59–70.
190. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013 Mar 9;14(1):91.
191. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.
192. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012 May;40(10):4288–97.
193. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014 Jun;42(11):e91.
194. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013 Oct;22(5):519–36.
195. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004 Aug 27;573(1–3):83–92.
196. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. J Bioinform Comput Biol. 2005 Oct;3(5):1171–89.
197. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics. 2006 Jul 26;7(1):359.
198. Gao X, Song PXK. Nonparametric tests for differential gene expression and interaction effects in multi-factorial microarray experiments. BMC Bioinformatics. 2005 Jul 21;6:186.
199. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol. 2004 Aug 30;3(1):Article19.
200. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics. 2005 Oct 15;21(20):3896–904.
201. Shi P, Ray S, Zhu Q, Kon MA. Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. BMC Bioinformatics. 2011 Sep 23;12(1):375.
202. Wang H, Sun Q, Zhao W, Qi L, Gu Y, Li P, et al. Individual-level analysis of differential expression of genes and pathways for personalized medicine. Bioinformatics. 2015 Jan 1;31(1):62–8.
203. Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010 Oct 31;9(1):Article39.
204. Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. Neuroimage. 2014 May;92:381–97.
205. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9.
206. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021 Jan 8;49(D1):D325–34.
207. Deelen P, van Dam S, Herkert JC, Karjalainen JM, Brugge H, Abbott KM, et al. Improving the diagnostic yield of exome- sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. Nat Commun. 2019 Jun 28;10(1):2837.

208. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity. 2008 Jul 18;29(1):150–64.

209. Chaussabel D, Sher A. Mining microarray expression data by literature profiling. Genome Biol. 2002 Sep 13;3(10):RESEARCH0055.

210. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545–50.

211. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013 Jan 16;14:7.

212. Lauria A, Peirone S, Giudice MD, Priante F, Rajan P, Caselle M, et al. Identification of altered biological processes in heterogeneous RNA-sequencing data by discretization of expression profiles. Nucleic Acids Res. 2020 Feb 28;48(4):1730–47.

213. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005 Oct 13;33(18):5868–77.

214. Falckenhayn C, Boerjan B, Raddatz G, Frohme M, Schoofs L, Lyko F. Characterization of genome methylation patterns in the desert locust Schistocerca gregaria. J Exp Biol. 2013 Apr 15;216(Pt 8):1423–9.

215. Alam M, Mahi NA, Begum M. Zero-inflated models for RNA-seq count data. J biomed anal. 2018 Sep 21;1(2):55–70.

216. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics. 2019 Jul 1;20(3):367–83.

217. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010 Oct 27;11(10):R106.

218. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010 Aug 10;11(1):422.

219. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013 Apr;14(2):232–43.

220. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Bioinformatics. 2019 Feb 1;35(3):487–96.

221. Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, et al. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell. 2011 Sep 16;146(6):1029–41.

222. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics. 2014 Jun 24;15:215.

223. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014 Apr;42(8):e69.

224. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016 May 15;32(10):1446–53.

225. Sun D, Xi Y, Rodriguez B, Park H, Tong P, Meong M, et al. MOABS: model based analysis of bisulfite sequencing data. Genome Biol. 2014;15(2):R38.

226. Klein H-U, Hebestreit K. An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. Brief Bioinform. 2016 Sep;17(5):796–807.

227. Lea AJ, Tung J, Zhou X. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. PLoS Genet. 2015 Nov;11(11):e1005650.

228. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

229. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. Front Microbiol. 2017 Nov 7;8:2114.
230. Bates DM, DebRoy S. Linear mixed models and penalized least squares. J Multivar Anal. 2004 Oct;91(1):1–17.
231. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D. Identification of genetic elements that autonomously determine DNA methylation states. Nat Genet. 2011 Oct 2;43(11):1091–7.
232. Kolde R, Märtens K, Lokk K, Laur S, Vilo J. seqlm: an MDL based method for identifying differentially methylated regions in high density methylation array data. Bioinformatics. 2016 Sep 1;32(17):2604–10.
233. Hebestreit K, Dugas M, Klein H-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2013 Jul 1;29(13):1647–53.
234. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. Bioinformatics. 2012 Nov 15;28(22):2986–8.
235. Wang H-Q, Tuominen LK, Tsai C-J. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. Bioinformatics. 2011 Jan 15;27(2):225–31.
236. Elo LL, Lahti L, Skottman H, Kyläniemi M, Lahesmaa R, Aittokallio T. Integrating probe-level expression changes across generations of Affymetrix arrays. Nucleic Acids Res. 2005 Dec 14;33(22):e193.
237. Therneau TM, Ballman KV. What does PLIER really do? Cancer Inform. 2008 Aug 27;6:423–31.
238. Affymetrix. Alternative Transcript Analysis Methods for Exon Arrays. Affymetrix White Paper [Internet]. 2005 [cited 2021]. Available from: https://assets.thermofisher.com/TFS-Assets/LSG/brochures/exon_alt_transcript_analysis_whitepaper.pdf
239. Cline MS, Blume J, Cawley S, Clark TA, Hu J-S, Lu G, et al. ANOSVA: a statistical method for detecting splice variation from expression data. Bioinformatics. 2005 Jun;21 Suppl 1(Suppl 1):i107-15.
240. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP. FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics. 2008 Aug 1;24(15):1707–14.
241. Bemmo A, Benovoy D, Kwan T, Gaffney DJ, Jensen RV, Majewski J. Gene expression and isoform variation analysis using Affymetrix Exon Arrays. BMC Genomics. 2008 Nov 7;9:529.
242. Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics. 2006 Dec 27;7(1):325.
243. Mikk M-L, Kiviniemi M, Laine A-P, Härkönen T, Veijola R, Simell O, et al. The HLA-B*39 allele increases type 1 diabetes risk conferred by HLA-DRB1*04:04-DQB1*03:02 and HLA-DRB1*08-DQB1*04 class II haplotypes. Hum Immunol. 2014 Jan;75(1):65–70.
244. Ilonen J, Kiviniemi M, Lempainen J, Simell O, Toppari J, Veijola R, 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 [Internet]. Vol. 17, Pediatric Diabetes. 2016. p. 8–16. Available from: http://dx.doi.org/10.1111/pedi.12327
245. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988 Feb 11;16(3):1215.
246. Strachan DP. Family size, infection and atopy: the first decade of the “hygiene hypothesis.” Thorax. 2000 Aug;55 Suppl 1:S2-10.
247. Seiskari T, Kondrashova A, Viskari H, Kaila M, Haapala A-M, Aittoniemi J, et al. Allergic sensitization and microbial load--a comparison between Finland and Russian Karelia. Clin Exp Immunol. 2007 Apr;148(1):47–52.

248. Kondrashova A, Reunanen A, Romanov A, Karvonen A, Viskari H, Vesikari T, et al. A six-fold gradient in the incidence of type 1 diabetes at the eastern border of Finland. Ann Med. 2005;37(1):67–72.

249. Benaglia T, Chauveau D, Hunter DR, Young D. mixtools: An R Package for Analyzing Finite Mixture Models [Internet]. Vol. 32, Journal of Statistical Software. 2009. p. 1–29. Available from: http://www.jstatsoft.org/v32/i06/

250. Huang Q, Liu D, Majewski P, Schulte LC, Korn JM, Young RA, et al. The plasticity of dendritic cell responses to pathogens and their components. Science. 2001 Oct 26;294(5543):870–5.

251. Krämer A, Green J, Pollard J Jr, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014 Feb 15;30(4):523–30.

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

253. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

254. 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 Jan;39(Database issue):D997-1001.

255. Ricaño-Ponce I, Wijmenga C. Mapping of immune-mediated disease genes. Annu Rev Genomics Hum Genet. 2013 Jul 3;14(1):325–53.

256. Pers TH, Timshel P, Hirschhorn JN. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. Bioinformatics. 2015 Feb 1;31(3):418–20.

257. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.

258. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. Nucleic Acids Res. 2013 Jan;41(Database issue):D1228-33.

259. Martino D, Holt P, Prescott S. A novel role for interleukin-1 receptor signaling in the developmental regulation of immune responses to endotoxin. Pediatr Allergy Immunol. 2012 Sep;23(6):567–72.

260. Wynn JL, Cvijanovich NZ, Allen GL, Thomas NJ, Freishtat RJ, Anas N, et al. The influence of developmental age on the early transcriptomic response of children with septic shock. Mol Med. 2011 Nov;17(11–12):1146–56.

261. Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, et al. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol. 2012 Oct 3;13(10):R92.

262. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011 Jul;9(7):e1001091.

263. Genome Reference Consortium. NCBI downloads. 2018 [cited 2019 Feb 10]. Available from: https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz

264. Laajala E. RRBS workflow. 2021 [cited 2021 May 11]. Available from: https://github.com/EssiLaajala/RRBS_workflow

265. Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics. 2015 Apr 1;31(7):1127–9.

266. R Core Team. R: A language and environment for statistical computing [Internet]. R Foundation for Statistical Computing, Vienna, Austria; 2020. Available from: https://www.R-project.org/

267. Suomi T, Hiissa J, Elo LL. PECA: Probe-level Expression Change Averaging. 2020.

268. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013 Oct;45(10):1238–43.

269. Yousefi P, Huen K, Davé V, Barcellos L, Eskenazi B, Holland N. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. BMC Genomics. 2015 Nov 9;16(1):911.
270. Maschietto M, Bastos LC, Tahira AC, Bastos EP, Euclydes VLV, Brentani A, et al. Sex differences in DNA methylation of the cord blood are related to sex-bias psychiatric diseases. Sci Rep. 2017 Mar 17;7:44547.
271. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318–30.
272. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008 Dec;40(12):1413–5.
273. Zimmermann K, Jentsch M, Rasche A, Hummel M, Leser U. Algorithms for differential splicing detection using exon arrays: a comparative assessment. BMC Genomics. 2015 Feb 27;16(1):136.
274. Van Moerbeke M, Kasim A, Talloen W, Reumers J, Göhlmann HWH, Shkedy Z. A random effects model for the identification of differential splicing (REIDS) using exon and HTA arrays. BMC Bioinformatics. 2017 May 25;18(1):273.
275. Chapman RM, Tinsley CL, Hill MJ, Forrest MP, Tansey KE, Pardiñas AF, et al. Convergent evidence that ZNF804A is a regulator of pre-messenger RNA processing and gene expression. Schizophr Bull. 2019 Oct 24;45(6):1267–78.
276. Rasche A, Herwig R. ARH: predicting splice variants from genome-wide data with modified entropy. Bioinformatics. 2010 Jan 1;26(1):84–90.
277. Richardson SJ, Horwitz MS. Is type 1 diabetes "going viral"? Diabetes. American Diabetes Association; 2014 Jul;63(7):2203–5.
278. Heinig M, Cardiogenics Consortium, Petretto E, Wallace C, Bottolo L, Rotival M, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature. 2010 Sep;467(7314):460–4.
279. Zhang J-P, Yang Y, Levy O, Chen C. Human neonatal peripheral blood leukocytes demonstrate pathogen-specific coordinate expression of TLR2, TLR4/MD2, and MyD88 during bacterial infection in vivo. Pediatr Res. 2010 Dec;68(6):479–83.
280. Lauener RP, Birchler T, Adamski J, Braun-Fahrländer C, Bufe A, Herz U, et al. Expression of CD14 and Toll-like receptor 2 in farmers' and non-farmers' children. Lancet. 2002 Aug 10;360(9331):465–6.
281. Krauss-Etschmann S, Hartl D, Heinrich J, Thaqi A, Prell C, Campoy C, et al. Association between levels of Toll-like receptors 2 and 4 and CD14 mRNA and allergy in pregnant women and their offspring. Clin Immunol. 2006 Feb;118(2–3):292–9.
282. Reece P, Thanendran A, Crawford L, Tulic MK, Thabane L, Prescott SL, et al. Maternal allergy modulates cord blood hematopoietic progenitor Toll-like receptor expression and function. J Allergy Clin Immunol. 2011 Feb;127(2):447–53.
283. von Hertzen L, Laatikainen T, Pitkänen T, Vlasoff T, Mäkelä MJ, Vartiainen E, et al. Microbial content of drinking water in Finnish and Russian Karelia - implications for atopy prevalence. Allergy. 2007 Mar;62(3):288–92.
284. Pakarinen J, Hyvärinen A, Salkinoja-Salonen M, Laitinen S, Nevalainen A, Mäkelä MJ, et al. Predominance of Gram-positive bacteria in house dust in the low-allergy risk Russian Karelia. Environ Microbiol. 2008 Dec;10(12):3317–25.
285. Lisciandro JG, Prescott SL, Nadal-Sims MG, Devitt CJ, Richmond PC, Pomat W, et al. Neonatal antigen-presenting cells are functionally more quiescent in children born under traditional compared with modern environmental conditions. J Allergy Clin Immunol. 2012 Nov;130(5):1167-1174.e10.
286. Köhler C, Adegnika AA, Van der Linden R, Agnandji ST, Chai SK, Luty AJF, et al. Comparison of immunological status of African and European cord blood mononuclear cells. Pediatr Res. 2008 Dec;64(6):631–6.

287. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, et al. Genetic structure of Europeans: a view from the North-East. PLoS One. 2009 May 8;4(5):e5472.

288. Reinert-Hartwall L, Honkanen J, Salo HM, Nieminen JK, Luopajärvi K, Härkönen T, et al. Th1/Th17 plasticity is a marker of advanced β cell autoimmunity and impaired glucose tolerance in humans. J Immunol. 2015 Jan 1;194(1):68–75.

289. Mustonen N, Siljander H, Peet A, Tillmann V, Härkönen T, Ilonen J, et al. Early childhood infections precede development of beta-cell autoimmunity and type 1 diabetes in children with HLA-conferred disease risk. Pediatr Diabetes. 2018 Mar;19(2):293–9.

290. Mustonen N, Siljander H, Peet A, Tillmann V, Härkönen T, Ilonen J, et al. Early childhood infections and the use of antibiotics and antipyretic-analgesics in Finland, Estonia and Russian Karelia. Acta Paediatr. 2019 Nov;108(11):2075–82.

291. Vatanen T, Kostic AD, d'Hennezel E, Siljander H, Franzosa EA, Yassour M, et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. Cell. 2016 Jun 2;165(6):1551.

292. Lietzen N, An LTT, Jaakkola MK, Kallionpää H, Oikarinen S, Mykkänen J, et al. Enterovirus-associated changes in blood transcriptomic profiles of children with genetic susceptibility to type 1 diabetes. Diabetologia. 2018 Feb;61(2):381–8.

293. Turturice BA, Theorell J, Koenig MD, Tussing-Humphreys L, Gold DR, Litonjua AA, et al. Perinatal granulopoiesis and risk of pediatric asthma. Elife [Internet]. 2021 Feb 10;10. Available from: http://dx.doi.org/10.7554/eLife.63745

294. Somani J, Ramchandran S, Lähdesmäki H. A personalised approach for identifying disease-relevant pathways in heterogeneous diseases. NPJ Syst Biol Appl. 2020 Jun 9;6(1):17.

295. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. Am J Hum Genet. 2016 Apr;98(4):680–96.

296. Sharp GC, Salas LA, Monnereau C, Allard C, Yousefi P, Everson TM, et al. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium. Hum Mol Genet. 2017 Oct 15;26(20):4067–85.

297. Küpers LK, Monnereau C, Sharp GC, Yousefi P, Salas LA, Ghantous A, et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. Nat Commun. 2019 Apr 23;10(1):1893.

298. Merid SK, Novoloaca A, Sharp GC, Küpers LK, Kho AT, Roy R, et al. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. Genome Med. 2020 Mar 2;12(1):25.

299. Valle A, Giamporcaro GM, Scavini M, Stabilini A, Grogan P, Bianconi E, et al. Reduction of circulating neutrophils precedes and accompanies type 1 diabetes. Diabetes. 2013 Jun;62(6):2072–7.

300. Apostol AC, Jensen KDC, Beaudin AE. Training the fetal immune system through maternal inflammation-A layered hygiene hypothesis. Front Immunol. 2020 Feb 11;11:123.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2021

ISBN 978-951-29-8738-2 (PRINT)
ISBN 978-951-29-8739-9 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)