Identifying novel phenotype profiles of diabetic complications and their genetic components using machine learning approaches

Iiro Toppila

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 1.5.2016

Thesis supervisor:

Prof. Harri Lähdesmäki

Thesis advisor:

D.Sc. (Tech.) Niina Sandholm

Aalto University
School of Electrical Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Iiro Toppila

Title: Identifying novel phenotype profiles of diabetic complications and their genetic components using machine learning approaches

Date: 1.5.2016 Language: English Number of pages: 8+89

Department of Computer Science

Professorship: Computational and Cognitive Biosciences

Supervisor: Prof. Harri Lähdesmäki

Advisor: D.Sc. (Tech.) Niina Sandholm

Patients with Type 1 diabetes (T1D) may develop a wide variety of additional slowly progressing complications, which have been shown to be partly heritable and to correlate with each other. However, the genetic and biological mechanisms behind them are still mostly unknown. The goal of this work was to use machine learning and data mining approaches that could capture the progressive nature of multiple complications simultaneously, and create novel phenotype classes that could help to solve the pathogenesis and genetics of diabetic complications.

To achieve this, a dual-layer self-organizing map (SOM) was trained using clinical and environmental patient data from the FinnDiane study, and the trained SOM node prototypes were clustered to classes using agglomerative hierarchical clustering. The genetic differences between the created classes were evaluated using heritability estimates, and the genetic markers associated with the class assignments showing significant heritability were analysed in genome-wide association study (GWAS). The created class assignments were biologically plausible, and were estimated to be up to 42% genetically determined. The GWAS analyses detected a genetic marker (rs202095311, located in the last intron of the gene *NRIP1*) genome-wide significantly ($p < 5 \times 10^{-8}$) associated with one of the created class assignments. In addition, GWAS detected multiple other genetic regions with suggestive p-values that contained mostly genes and processes previously linked to diabetic complications or their risk factors.

Overall, the new approach to study the genetics of complex diseases was found to perform well in case of T1D and its complications, and could be used to study also other complex traits and diseases.

Keywords: complications of type 1 diabetes, FinnDiane, multi-layer self-organizing map (SOM), clustering, heritability, genome-wide association study (GWAS)

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Iiro Toppila

Työn nimi: Diabeettisten komplikaatioiden uusien fenotyyppiprofiilien etsintä, sekä ryhmien välisten geneettisten komponenttien tunnistus koneoppimismenetelmiä hyödyntäen

Päivämäärä: 1.5.2016 Kieli: Englanti Sivumäärä: 8+89

Tietotekniikan laitos

Professuuri: Laskennallinen ja kognitiivinen biotiede

Työn valvoja: Prof. Harri Lähdesmäki

Työn ohjaaja: TkT Niina Sandholm

Tyypin 1 diabetikoille saattaa kehittyä useita hitaasti eteneviä lisäsairauksia, jotka ovat osittain perinnöllisiä sekä keskenään korreloivia. Sekä geneettiset että biologiset mekanismit näiden taustalla ovat kuitenkin pääasiassa vielä tuntemattomia. Tämän työn tarkoituksena oli hyödyntää koneoppimis- ja tiedonlouhintamenetelmiä, joiden avulla pystyttäisiin vangitsemaan samanaikaisesti useiden diabeettisten komplikaatioden etenevä luonne, sekä muodostamaan uusia fenotyyppiluokkia diabeettisten komplikaatioiden ja niiden genetiikan tutkimuksen avuksi.

Työssä opetettiin monitasoinen itseorganisoituva kartta (SOM) käyttäen FinnDiane tutimuksessa kliinisistä muuttujista sekä ympäristötekijöistä kerättyä potilasdataa. Uusien fenotyyppiluokkien luomiseksi opetetun kartan prototyyppialkiot klusteroitiin kokoavalla hierarkkisella klusteroinnilla. Luokkien välisiä geneettisiä eroja vertailtiin heritabiliteettiestimaateilla. Lisäksi luokkajakoon assosioituvien geneettisten markkereiden vaikutusta tutkittiin perimänlaajuisessa assosiaatiotutkimuksessa (GWAS) niiden luokkien välillä, jotka saavuttivat merkitseviä estimaatteja heritabiliteeteille.

Muodostetut potilasluokat olivat biologisesti mielekkäitä ja muodostetun luokkajaon estimoitiin olevan jopa 42% geneettisesti määräytyvä. Perimänlaajuisissa assosiaatiotutkimuksissa geneettinen variantti (rs202095311 *NRIP1* geenin viimeisessä intronissa) assosioitui yhteen muodostetuista luokkajaoista genominlaajuisella merkitsevyystasolla ($p < 5 \times 10^{-8}$). Lisäksi analyyseissa havaittiin viitteellisillä p-arvoilla useita muita geneettisiä alueita, joilla sijaitsee aiemmin diabeettisiin komplikaatioihin tai niiden riskitekijöihin yhdistettyjä geenejä ja prosesseja.

Yleisesti, uusi lähestymistapa kompleksisten sairauksien genetiikan tutkimukseen suoriutui sille asetetuista haasteista tyypin 1 diabeteksen ja sen komplikaatioiden tutkimuksessa ja vastaava lähestymistapa voisi olla hyödynnettävissä myös muiden kompleksisten sairauksien tutkimuksessa.

Avainsanat: tyypin 1 diabeteksen komplikaatiot, FinnDiane, monitasoinen itseorganisoituva kartta, klusterointi, heritabiliteetti, perimänlaajuinen assosiaatiotutkimus

Preface

I started as a part time summer trainee in the FinnDiane Study Group while I was still working with my Bachelor's thesis, long before I even dreamt of the Master's thesis and degree. Now it's been more than three years since, and lots have happened during my time in the FinnDiane. However, now my Master's is finally ready. Oh' the time really flies!

I would like to express my sincere gratitude to my advisor Niina Sandholm for everything you have done during my FinnDiane years. Thank you for the valuable interactive feedback and making me to push this thesis higher in the long priority list of various other projects and tasks I was working with. I'm really grateful for you for sparing your time for this thesis, even during your maternity leave. Congratulations and the best of luck to your firstborn daughter and the whole family! I would also like to thank my supervisor Prof. Harri Lähdesmäki for all the feedback that guided my thesis to its final form. I also want to add special thanks to Carol Forsblom for taking time to carefully read the thesis through without sparing a red marker, as without it, I would have ended up using methods detonating some parts of my results and conclusions. Last, I want to thank the whole FinnDiane Study Group and its leader Per-Henrik Groop.

Of course, I can't forget my family and friends supporting me; thank you for being there. In addition, I want to thank people and activities at Yliopiston Taido and Suomen Taidon valmennusryhmä for taking my mind off from this thesis now and then. Thanks for all the laugh, sweat and tears; for creating the (im)balance between physical and mental training in my life; and basically eating up all the rest of my free time.

Finally, I want to thank my beloved wife Minna for making it all worth the effort.

Otaniemi, 1.5.2016

Iiro Toppila

Contents

Symbols and abbreviations

Symbols and Operators

$\mathbb{R}^d$	d-dimensional real-valued coordinate space
$\sum$	Sum over elements
$\arg\min\limits_{i}$	Index i of element minimizing the argument
$\lvert \cdot \rvert$	Cardinality, "number of elements in a set"
$\lVert \cdot \rVert$	Euclidean norm
$\lfloor \cdot \rfloor$	Floor function, rounding down to nearest integer
$O(\cdot)$	Big O notation, "the complexity is dominated by the function"

Abbreviations

2-AG	2-arachidonoyl glycerol
AC	Agglomerative coefficient
ACR	Albumin-to-creatinine ratio
AER	Albumin excretion rate
AGE	Advanced glycation end-products
AHT	Antihypertensive (medication)
AMI	Acute myocardial infarction
ApoA-I	Apolipoprotein A-I
ApoB	Apolipoprotein B
bp	Base pair (nucleotides in DNA)
BL	Baseline, refers to data measures obtained from patients at the time when they entered the study
BMU	Best matching unit
CAD	Coronary artery disease
CHD	Coronary heart disease
chr	Chromosome
CVD	Cardiovascular disease
DAG	Diacylglycerol
DBP	Diastolic blood pressure
DN	Diabetic nephropathy
DPN	Diabetic polyneuropathy
DR	Diabetic retinopathy
EDNSG	European Diabetic Nephropathy Study Group
ESRD	End-stage renal disease, the most severe form of diabetic nephropathy
FinnDiane	Finnish Diabetic Nephropathy (Study)

FU	Follow-up, refers to data measures obtained from patients at the second visit during the study
GCTA	Genome-wide Complex Trait Analysis (software)
GRM	Genetic relationship matrix
GWAS	Genome-wide association study
HbA_{1c}	Glycoted haemoglobin A1c
HDL	High-density lipoprotein
IHD	Ischemic heart disease
IMT	Intima media thickness
IQR	Interquartile range
kb	Kilobases, genetic distance of 1,000 nucleotide base pairs
LADA	Latent autoimmune diabetes of the adult
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAP	Mean arterial pressure
MATLAB	Matrix Laboratory, computational software and programming language by MathWorks
MODY	Maturity onset diabetes of the young
OR	Odds ratio
PC	Principal component
PDR	Proliferative diabetic retinopathy
PKC	Protein kinase C
PVD	Peripheral vascular disease
QC	Quality control
QE	Quantization error of SOM
QQ-plot	Quantile-quantile plot
REML	Restricted maximum likelihood (analysis)
ROS	Reactive oxygen species
SBP	Systolic blood pressure
SE	Standard error
SNP	Single-nucleotide polymorphism
SOM	Self-organizing map
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TG	Triglyceride
TE	Topographical error of SOM
V-ATPase	Vacuolar-type H^{+}-ATPase
WHR	Waist-to-hip ratio

1 Introduction

Finland has the largest prevalence of type 1 diabetes (T1D) in the world [1], and the incidence of T1D is steadily increasing both in Finland and worldwide [2, 3]. The diagnosis of T1D is usually followed by a variety of different complications in the following years, and these complications are the major cause of the decrease in life quality and increase in premature mortality observed in diabetes. Whereas complication-free patients with T1D have the same life expectancy compared with their non-diabetic peers, for example the T1D patients with end-stage renal disease (ESRD) have an 18-fold risk of dying early [4]. However, the most common cause of death for patients with T1D is not nephropathy itself, but cardiovascular disease (CVD) [5]. At the same time diabetic retinopathy (DR) is the leading cause of acquired visual disability among people of working age in all industrialized countries [6].

The development and progression of diabetic complications is not fully understood but many clinical and environmental risk factors have been identified for these traits. It seems that some of the complications share common risk factors, mechanisms and pathways and many of the clinical phenotypes correlate with each other. For example almost all patients developing diabetic nephropathy (DN), a complication affecting the kidneys, have also complications of DR.

Many of the complications of T1D have been shown to be partly heritable [7, 8, 9], however, only few genetic markers affecting them have been found regardless of large efforts [10, 11, 12]. This might be partly due to the dependencies between different complications, cluttering the analyses focusing only on one trait at a time. Our previous studies have shown that some of the genetic associations can only be detected in smaller subsets of patients within a certain trait, for example we have identified a genetic marker affecting the risk of ESRD in women, but not in men [13]. Similarly, analysing a set of extreme patients having both ESRD and severe DR amplified the strength of genetic associations compared with analysing the complications separately (regardless of the smaller patient sample) [14]. Thus there is a need for new approaches that take into account multiple traits simultaneously, if we want to identify the genetic components affecting diabetic complications.

Due to recent advancements in computer sciences, the amounts of stored data and the computational power are increasing at an exploding rate. New tools in fields of data mining and machine learning are necessary in order to handle and process these massive datasets. These tools are able to extract the important and relevant information among all the data available. Most of us encounter these tools almost daily without even noticing it. They are the base for example in the search engines and make it possible for Google to find the internet page of your interest using only a couple of search terms, or in marketing when the YouTube video you are watching is interrupted by an advertisement of an item of your interest.

This rapid increase in available data applies also to medical sciences, and the goals in them are generally not so different from the examples above; to use the patterns in vast data available to find an important phenomena, and further transform this information into something concrete for the people. Instead of providing the user with a desired internet page, the ultimate goal in medical sciences could be for example better care for the patients affected by a disease, or even medicine and cure for it. However, the basic principle is the same: to go from data to knowledge and even beyond. Thus the idea of applying these machine learning and data mining approaches also in medical sciences seems generally appealing.

The traditional analyses used to solve the genetics of the diabetic complications have shown thus far only small success, possibly due to the slowly progressing nature of the complications and/or the limited number of patients that have been included in the studies. The role of this work is to try alternative approaches that could tackle these limitations and move the study of diabetic complications and their genetics forward using machine learning approaches on the massive amounts of data already available.

1.1 Aims of the study

This work combines multiple methods from the areas of machine learning, data mining and statistics with a goal to create novel phenotype classifications for patients with T1D. The created classes are further evaluated by comparing their genetic background. The goal is to find the markers having an effect on the profile differences, which can be used to better understand the pathogenesis of the diabetic complications. These new approaches are needed in order to capture the progressive nature of the diabetic complications and to find the associations that might be missed by traditional approaches studying only one of the complications at a time.

The specific aims of this study are:

1. Create novel phenotype classes of patients with Type 1 diabetes by clustering the prototypes of a multi-layer SOM, which has been trained using clinical and environmental patient data.

2. Evaluate whether the created classes show genetically divergent background and find the genetic variants associated with this class assignment.

In order to achieve these goals, first a multi-layer SOM is trained using data from the Finnish Diabetic Nephropathy (FinnDiane) Study that contains thousands of patients with T1D and hundreds of variables describing them. The multi-layer SOM is trained using measures from different time points in different layers in order to capture the progressive nature of the diabetic complications, which can easily be missed with traditional approaches. After this, multiple different clustering methods are applied to classify the trained SOM prototypes in order to find the method that is most suitable for the task of creating novel phenotype profiles.

Finally the most suitable clustering method is applied to create the novel phenotype classes by clustering the SOM prototypes into differing number of clusters. When the number of clusters is increased, more specified patient subtypes can be identified. As a proof of concept, these groups are evaluated by intergroup heritabilities to detect the classes that have genetically divergent background. Ultimately, millions of genetic markers are tested for their association with these group assignments in a

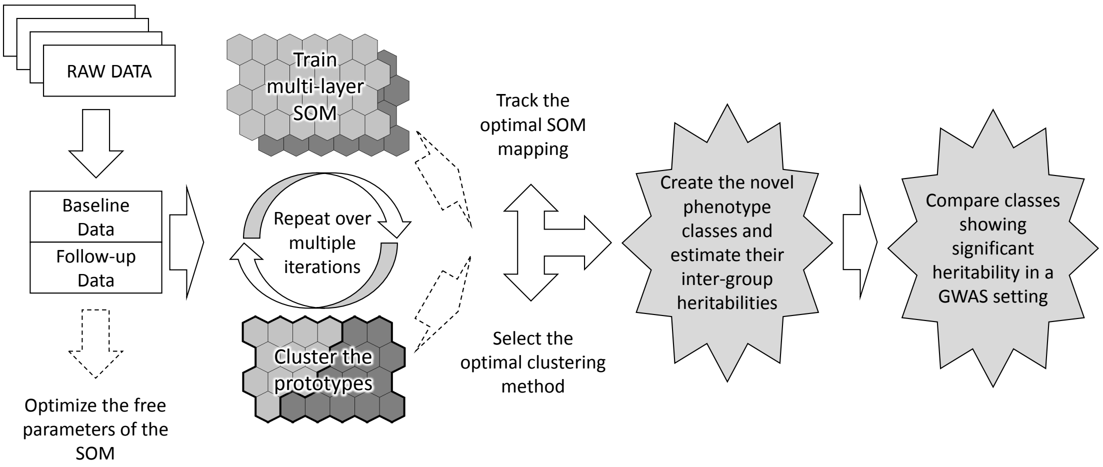

RAW DATA
Baseline Data
Follow-up Data
Optimize the free parameters of the SOM
Train multi-layer SOM
Repeat over multiple iterations
Cluster the prototypes
Track the optimal SOM mapping
Select the optimal clustering method
Create the novel phenotype classes and estimate their inter-group heritabilities
Compare classes showing significant heritability in a GWAS setting

Figure 1: General overview of the analysis pipeline. Refer to sections 3 and 4 for more detailed information.

genome-wide association study (GWAS) setting to pinpoint the genetic component creating the differences in complication prevalence and/or rate of progression to the next level. The general overview of the analysis pipeline is presented in Figure 1.

On top of this potentially clinically important information, also the methodologically interesting questions are addressed. The secondary goals of this work are:

1. Evaluate whether the multi-layer SOM is able to capture the progressive aspect of the diabetic complications.
2. Evaluate which of the clustering methods is the most suitable to create the novel phenotype classifications, when combined with the multi-layer SOM approach.
3. Evaluate the performance of the approach as whole, and whether similar pipeline could be used to study the genetics of other progressive complex diseases.

2 Physiology of diabetes

Diabetes is a general class of metabolic disorders, which are all characterized by elevated blood glucose concentrations. The rise in blood glucose levels can be caused by a decreased secretion of insulin by pancreatic β-cells, an impaired effect of insulin in target tissues or by combination of both. On top of affecting the carbohydrate metabolism, diabetes also disturbs the normal lipid and protein metabolism of the body.

Diabetes can be roughly categorized into two major forms, type 1 diabetes (T1D) and type 2 diabetes (T2D), even though the distinction between the two is not always straightforward and usually patients have features of both types. In T1D an autoimmune reaction destroys the pancreatic β-cells in the insulin secreting islets of Langerhans, and results in complete insulin deficiency. It usually occurs at a young age (less than 35 years), and requires a lifelong insulin therapy. T2D is mainly caused by decreased insulin sensitivity and/or decreased insulin secretion, and it is usually prevalent in older age groups (over 40 years) or in the presence of obesity. In addition, there exist rarer subtypes of diabetes: latent autoimmune diabetes of the adult (LADA) which is autoimmune in origin as T1D, but clinically resembles T2D; maturity onset diabetes of the young (MODY), which is a class of highly prevalent monogenic diabetes; and gestational diabetes, which occurs during pregnancy. This work concentrates on T1D, and later when the term diabetes is used, it refers to T1D if not otherwise specified.

T1D requires a lifelong insulin therapy, where the patients are required to monitor their blood glucose levels and inject insulin accordingly to keep it stable. A modern insulin therapy consists of multiple daily injections, or continuous subcutaneous insulin infusion using an insulin pump. In the more common form of the therapy where the insulin injections are used, a long-lasting basal insulin is injected once or twice a day accompanied by additional rapid acting insulin bolus injections during the mealtimes. In pump therapy, a rapid acting insulin is administered with a constant rate with addition of patient-activated bolus doses during the mealtimes. The amount of insulin in the boluses (both in pump therapy and when using multiple daily insulin injections) has to be determined separately each time based on the pre-meal glucose levels of the patient and the carbohydrate content of the meal. The insulin dose also need to be adjusted according to daily factors such as diet and physical activity as well as patient-dependent factors such as insulin sensitivity, stress, and pubertal status. The monitoring of glucose levels and determining the suitable insulin doses are the responsibility of the patients themselves creating an additional burden to the daily activities.

Keeping the blood glucose on a suitable and stable level is important as the glycaemic control is one of the most important risk factors for the development of long-term diabetic complications as later discussed in section 2.2. In addition, too high or too low glucose levels can also cause acute symptoms. When the glucose levels rise to a high level, i.e. during acute hyperglycaemia, the normal body metabolism is disturbed. The first signs include for example thirst, hunger, increased urinary volume, sleepiness, and blurred vision. If the condition is left untreated it can lead to a state called diabetic ketoacidosis, where the body starts to burn fatty acids resulting in acidic ketone bodies. Ketoacidosis will further cause symptoms such as deep rapid breathing, confusion, decreased level of consciousness and impaired cognitive function. Ketoacidosis is a medical emergency and requires hospital treatment. If the ketoasidosis is left untreated it can result in coma or even the death of the patient. Correspondingly, hypoglycaemia, i.e. too low blood glucose level, cause acute symptoms which typically occur more rapidly compared with hyperglycaemia. This state can occur if the patient takes too much insulin decreasing the blood glucose to too low levels. First the body releases natural counter-regulatory hormones (adrenaline and glucagon) which cause symptoms such as anxiety, nervousness, stronger and faster heart beat, sweating, nausea, vomiting and headache. These hormones can postpone the drop in the blood glucose for a while, however, the lack of glucose will soon start to affect the brain. At this point additional symptoms may occur including impaired judgement, irritation and mood changes, confusion, dizziness, blurred vision, flashes of light in the field of vision, difficulty of speaking, temporal paralysis and seizures. As in the case of hyperglycaemia, in the extreme cases also hypoglycaemia will result in coma and in the worst case the death of the patient. As the symptoms of impaired glucose management are so severe, monitoring and keeping the blood glucose levels stable may cause an unnecessary fear and additional stress to some patients with T1D.

Some of the patients with T1D may also develop long-term diabetic complications that take years to appear. Next in section 2.1 the most common of these are introduced after which section 2.2 will go through the known risk factors affecting these traits.

2.1 Diabetic complications

Although diabetes can be managed by modern insulin therapy and by monitoring of blood glucose levels, the treatment cannot fully mimic the human pancreas and return the body metabolism to normal. This is evident as a subset of patients with diabetes develop a wide variety of long-term diabetic complications [15]. These complications are the major causes of premature mortality and decrease in the life quality [4]. The most common long-term diabetic complications can roughly be divided into micro- and macrovascular complications.

The microvascular complications are the result of cellular damage in the small blood vessels and tissues surrounding them. The most common of these is diabetic retinopathy (DR) [16], where the vessels in the retina of the eye get weaker and might leak to the surrounding tissues. If advanced, the small arteries become stiffer and new abnormal weak blood vessels start to grow around the damaged areas in response to ischemia. This advanced form of DR is called proliferative diabetic retinopathy (PDR). Usually the first signs of DR are harmless, but even the smallest changes can affect the patients vision and in the worst case DR can lead to total blindness. In fact, DR is the leading cause of blindness in the working age population in the western world [6]. Still, almost all of the patients with T1D will get some degree of DR during their lifetime [17]. Another microvascular complication affecting many patients with diabetes is diabetic (poly)neuropathy (DPN), a heterogeneous class of microvascular complications affecting the nerves of the patient. It is thought to result from injury in the small blood vessels supplying the nerves. It may affect all peripheral nerves including pain fibres, motor neurons and the autonomic nervous system. Different forms of DPN can cause for example unpleasant sensory symptoms such as pain, numbness, burning sensation or tingling as well as exercise intolerance, diarrhoea, erectile dysfunction and loss of bladder control to name few of the possible symptoms.

While these two complications are mainly not life threatening and cause mainly symptoms decreasing the life quality, the third microvascular complication, diabetic nephropathy (DN), makes an exception. The small vessels in the kidney form special structures, nephrons, that are responsible for the filtration of the waste products from blood to urine and the overall fluid balance of the body. In DN, when these microvascular structures are damaged, the filtration rate of the kidney gets worse, and the substances that should not be filtered from the blood start to leak to the urine through the kidneys. In order to confirm DN, an invasive kidney biopsy and histological studies would be required, however, DN is clinically diagnosed in Finland and around the Europe by measuring the protein (albumin) levels of the urine and classified into three groups based on the level: normal albumin excretion rate (AER), microalbuminuria and macroalbuminuria. The fourth and the most severe stage, end-stage renal disease (ESRD), is referred to when the kidney function is lost to a level where a regular dialysis treatment or a kidney transplantation is required for the survival of the patient. Whereas the early symptoms of DN might be undetectable for the patient, it is a significant risk factor for increased mortality in T1D. In the absence of DN, the long-term survival rates of patients with T1D do not differ from the general background population without diabetes, but patients with ESRD have an 18-fold risk of the premature death [4].

All of the microvascular complications share common pathogenetic mechanisms, pathways and risk factors and they are strongly associated with each other. They are all slowly progressing diseases and take usually years to develop. After the duration of 15 years almost all patients have at least the first signs of DR [18], and approximately half of the patients have DPN after 15 to 20 years of diabetes [19]. The largest increase in incidence of PDR is seen after 10 years of diabetes [20], whereas in DN the peak incidence occurs after 15 to 20 years of diabetes, after which approximately one third of the patients have signs of DN [21].

The second large group of complications, macrovascular complications, can be divided into different subclasses of cardiovascular diseases (CVD): coronary artery disease/coronary heart disease/ischemic heart disease (CAD/CHD/IHD), stroke/cerebrovascular disease, and peripheral vascular disease (PVD). They are all caused by atherosclerosis in the large blood vessels, and the different disorders are classified by the location of the damage (similarly as in diabetic microvascular complications). CAD includes forms affecting the arteries supplying the heart, cerebrovascular disease includes forms affecting the arteries supplying the brain and PVD generally the lower limbs.

These macrovascular complications are common in patients with T1D, and the risk for developing them is much higher compared with the background population without diabetes. It has been shown that the intima media thickness (IMT, a surrogate marker for atherosclerosis) in patients with T1D is comparable to 20 years older peers without diabetes [22]. This is directly reflected in the risk of CVD events and mortality. Patients with T1D have 10 to 15-fold increased risk of lower-extremity amputation due to PVD compared with the background population [23]. Correspondingly, the risk for stroke is reported to be approximately 18-fold in men and 26-fold in women with T1D [24], and the risk for CHD mortality has been reported to be 9-fold for men and over 40-fold for women [25], compared with non-diabetic peers. Importantly, the incidence rate of CVD events in women with T1D is comparable to the incidence rate in men with T1D, suggesting a loss of sex specific protection seen in the background population. Thus the relative risks of CVD events for women are much higher compared with men. Furthermore, the risk for CVD events is highly correlated with the severity of DN [24], and this is notably seen also in the all-cause mortality risk [4], as CVD events are the most common cause of death in patients with T1D [5]. However, the causes for the strong association between DN and CVD are still unknown.

In addition to the micro- and macrovascular complications, patients with T1D are also more susceptible to other autoimmune diseases, when compared with the background population without diabetes, such as autoimmune thyroid disease, coeliac disease, multiple sclerosis, rheumatoid arthritis and asthma, as reviewed in [26]. Psychiatric disorders are also common including depression, anxiety, diabetes-related stress, eating disorders and other mental health symptoms [26]. Additionally, patients with T1D suffer from various other health related problems, for example they are generally more susceptible to bacterial infections compared with the diabetes-free background population [27].

2.2 Genetic and environmental risk factors of diabetic complications

Many different risk factors, both environmental and genetic, have been identified for the long-term diabetic complications. The environmental risk factors usually overlap for many of the complications, as illustrated in Table 1, that summarises some of the most common risk factors and their association with different complications. Among these risk factors, glycaemic control can be considered the most important for microvascular complications, as it affects multiple damaging biochemical pathways common for all of them. Hyperglycaemic conditions cause overproduction of advanced glycation end-products (AGEs) and activation of the polyol, hexosamine and diacylglycerol-protein kinase C pathways, which all have been associated with microvascular damage [28, 29]. For these pathways, the overproduction of reactive oxygen species (ROS) have been suggested as "the unifying mechanism" [29]. Glycaemic control affects also the risk for macrovascular complications, but other markers such as the lipid profile, age and blood pressure have a larger effect on them. On top of the glycaemic control also different time related variables (such as age, diabetes duration and age at diabetes onset) as well as blood pressure, smoking and

Table 1: Most common environmental risk factors for micro- and macrovascular diabetic complications.

	DN	DR	DPN	CVD
Blood pressure	*	*	*	*
Glycaemic control	*	*	*	*
Insulin resistance	*	*	*	*
Lipid profile	*	(*)	*	*
Obesity (/WHR/BMI)	*	*	*	*
Smoking	*	*	*	*
Time related variables	*	*	*	*
Anemia	*	*		
Gender	*		*	*
Height	*		*	
Puberty	*	*		
Adiponectin	*			
Birth weight	*			
Heavy alcohol consumption		*		
High protein diet	*			
Inflammatory markers	*			
Pregnancy		(**)		
Recent cataract surgery		*		

* = associated risk factor, (*) = debated, (**) = transient, long-term risk not affected

lipid profile (to some degree) seem to be common for both micro- and macrovascular complications. Additionally the different diabetic complications are considered as risk factors for each other.

These environmental risk factors do not fully explain all the attributable risk for any of these complications, suggesting the involvement of genetic risk factors. These assumptions are also supported by studies showing familial clustering of many of these complications [7, 8, 9]. However, despite large efforts only few genetic markers have been found for them. Where environmental risk factors are more general affecting many complications simultaneously, genetic factors are mainly found associated with one of the complications at a time. However, there is also some studies showing pleiotropic effects of certain markers [14].

Of the diabetic microvascular complications, DN is the most widely studied in terms of genetics [12]. Multiple smaller candidate gene studies have been performed, and the results have been combined in systematic meta-analysis. Mooyart *et al.* reported multiple variants associated with DN based on this meta-analysis, located near genes such as *ACE*, *AKR1B1*, *APOC1*, *APOE*, *EPO*, *NOS3*, *VEGFA*, *CARS*, and *GREM1* to name a few [30]. However, these associations are under debate, as the level of significances did not reach genome-wide level even in the meta-analyses [12]. For DN, also large consortia-based GWAS analyses have been performed, which have yielded associations reaching the genome-wide significant threshold. Two variants near the gene *AFF3*, and between genes *RGMA* and *MCTP2* were associated with ESRD, and a variant in an intronic region of *ERRB4* was suggestively associated with DN [10]. In addition, a marker that is associated with ESRD in women only was found between the genes *SP3* and *CDCA7* [13].

For DR, multiple candidate gene studies showing associations to different genetic markers have been performed, but the reported results have been mainly conflicting (with a significance far away from the genome-wide level), possibly due to small sample sizes and varying phenotype definitions. Of these, the genes *AKR1B1*, *AGER*, *VEGF/VEGFA*, *NOS3*, and *ACE* are among the most studied [31, 32], but there are also many other candidate regions tested and even GWASs performed, however, none of which have yielded genome-wide significant associations [32]. Also systemic efforts to replicate the previous DR associations in independent cohorts have been made [11], however, they have not been able to confirm any of the associations affecting the risk of DR.

Of microvascular complications, DPN has shown least results with genetic studies, possibly because it is a more heterogeneous complication compared with DN and DR. Until now there are no GWASs performed on DPN, and no genetic associations worth mentioning have been found, excluding variants in the genes *SCN9A* and *TRPA1*, which are found to be causal of certain monogenic neuropathic conditions [12].

The macrovascular complications have been widely studied in healthy populations without diabetes, and the performed GWASs have been magnitude(s) larger compared with the diabetes specific complications. For example, a GWAS of almost 200,000 patients was performed for CHD by CARDIoGRAMplusC4D Consortium, and after their analyses the number of genetic loci reaching genome-wide significance for CHD risk was increased to 46 [33]. However, as the risk for CVD events is clearly elevated in diabetes, the interesting study question is whether there are diabetes specific genetic markers affecting CVD, or is there an interaction between diabetes and known genetic CVD risk markers. Such interactions have been found with different CVD outcomes and T2D as reviewed by Ahlqvist *et al.* in [12], however, there is little if any such results in patients with T1D.

3 Computational methods

In this work multiple different methods from the areas of machine learning, data mining and statistics are used to better understand the mechanisms behind diabetic complications. In this section, the theoretical background of the most relevant methods used in this work are discussed. First, section 3.1 presents the theory of traditional and multi-layer SOM going through the algorithm and the properties of this approach. Next, section 3.2 introduces the problem of clustering and the methods used to group SOM prototypes together in order to create the novel phenotypes. Finally, section 3.3 introduces the methods used to estimate the heritability and genetic associations with the novel phenotypes.

3.1 Self-organizing map

The SOM is an unsupervised neural networks algorithm originally based on a model of the nervous system and competitive learning. The algorithm was developed by Teuvo Kohonen and introduced in the 1980s [34]. Briefly, the algorithm performs mapping of high dimensional data to a low dimensional ordered grid (usually two-dimensional), while preserving the most important local relationships between data points. While effectively reducing the dimensionality of the data, allowing effective visualization, the algorithm also creates abstractions of the data in form of model prototypes. Both of these properties will be useful when studying large multidimensional datasets with underlying hidden patterns.

In competitive learning, the basic principle is that neurons in neural network adapt to the data in order to become sensitive for certain patterns. In the easiest form, the neuron that best matches the presented input data "wins the competition" and is allowed to learn and adapt closer to presented data. This same idea applies also in SOM, but the speciality it introduces is the connections between these nodes. Not only the winning node is learning, but also its neighbours close by are gaining the information and adapting based on it.

The basis for SOM is a grid of nodes that have well defined spatial positions or connections between them. Possibly the easiest form is a two dimensional regular square lattice (see Figure 3 in section 3.1.1). The nodes and the spatial relations (distances) are presented in a two-dimensional space $\mathbb{R}^2$, and each of these nodes have an initial built-in prototype in the d-dimensional space $\mathbb{R}^d$, where d is the number of the features in the used training samples. In the SOM training algorithm, the input data vectors (the training data) are presented to the adapting neural network randomly one at the time. At every iteration step t, the node indexed with c whose current prototype $\boldsymbol{m}_c(t)$ is closest to the presented data vector $\boldsymbol{x}(t)$ (in $\mathbb{R}^d$) among all possible prototype vectors is selected as a winner of the round:

$$c = c(\boldsymbol{x}(t)) = \arg\min_i ||\boldsymbol{x}(t) - \boldsymbol{m}_i(t)||^2. \tag{1}$$

Generally the Euclidean distance is used, but other forms can be implemented as well. After the winner is selected, all the node prototypes are moved closer to the presented data vector (in $\mathbb{R}^d$) according to the following rule:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{c,i}(t)(\boldsymbol{x}(t) - \boldsymbol{m}_i(t)), \tag{2}$$

where $h_{c,i}(t)$ is the neighbourhood function of the winning node. In basic formulation [35, 36], it has a Gaussian form of:

$$h_{c,i}(t) = \alpha(t)\text{exp}\left(-\frac{||\boldsymbol{r}_i - \boldsymbol{r}_c||^2}{2\sigma^2(t)}\right), \tag{3}$$

where $0 < \alpha(t) < 1$ is the learning-rate factor, which decreases monotonically with the regression steps. $\boldsymbol{r}_i \in \mathbb{R}^2$ and $\boldsymbol{r}_c \in \mathbb{R}^2$ are the vectorial locations of the nodes in the ordered two dimensional grid and $\sigma(t)$ is the width of the neighbourhood function which also decreases monotonically with the regression steps. In this form all the nodes are adjusted, but the strength of the adjustment depends from their position on the grid in relation to the winning unit. Only the winner itself is moved with the factor of learning rate $\alpha(t)$, and all the other nodes are adjusted less. The immediate neighbourhood will also be affected notably, but as the neighbourhood function decreases rapidly as a function of the distance in the grid lattice, nodes far away from the winner are left virtually unchanged. Also other forms of neighbourhood function can be implemented, which are also used in practice. A simpler form of the neighbourhood function can be implemented as follows:

$$h_{c,i}(t) = \begin{cases} \alpha(t), & \text{if } ||\boldsymbol{r}_i - \boldsymbol{r}_c|| < r^*(t) \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $r^*(t)$ is the monotonically decreasing radius, which defines the size of the neighbourhood. In this form, the winning unit and the close by neighbours are all moved towards the presented data vector with the same multiplier and nodes not belonging to the neighbourhood set are left as is, as visualized in Figure 2. Regardless of the form of the neighbourhood function, this iterative training is repeated for a predetermined t times and the same data vectors are usually presented multiple times to the SOM at different steps when the neighbourhood radius and learning rate have decreased in order to achieve a large enough number of iterations [37].

In the beginning of the SOM training algorithm, when both the neighbourhood radius and learning rate are at their maximum, the map is quickly changing and adapting to the data as multiple nodes are moved at once relatively strongly towards the presented data vectors. In this phase, the grid will quickly unfold and order itself in the d-dimensional data space so that nearby units are in close segments of the space matching the overall form of the training data cloud. As both the neighbourhood radius and the learning rate decreases as the function of the iteration steps, the nodes gradually start to stabilize and resemble the data vectors located in close by regions. Usually (depending on the parameters) in the very end of the training phase only the winning unit is moved at each iteration, thus, allowing the fine tuning of the prototypes.

One of the advantages of the SOM algorithm, which makes it appealing to apply to biological data, is that it can handle also data vectors with missing values during the training. When a sample vector with missing values is presented, the winning node

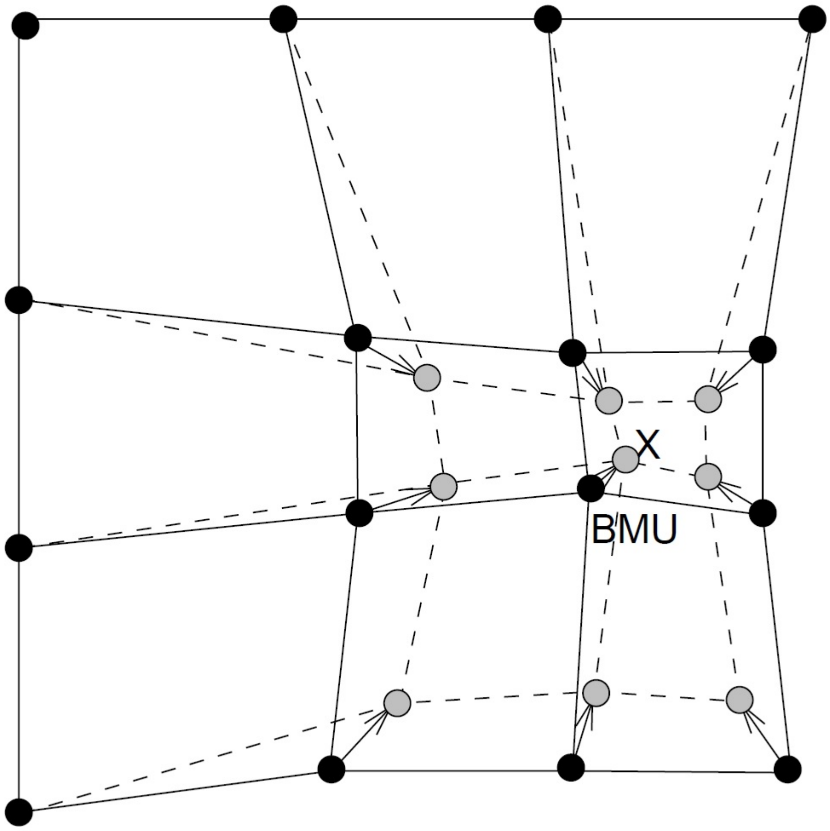


Figure 2: The adaptation of nodes during SOM training, visualized by using the radial form of the neighbourhood function given in Equation (4). Picture from [38].

can be selected by computing the distances using only the subspace of non-missing features of the presented data vector. If the proportion of missing elements is small, the result will be "statistically fairly accurate". However, this assumption is no longer valid when a majority of the features of the presented data vector are missing. When updating the node prototypes using a vector with partly missing values, only the dimensions present in the sample are to be altered [39].

After the SOM training phase the approach starts to really show its power when the visualization and abstraction properties are taken advantage of. The positions of the nodes stay unchanged in the two-dimensional grid during the training, but each of them have learned a unique prototype that works as an abstraction describing the data. Moreover, as these prototypes have adapted gradually together to their final configuration and values, each of the node prototypes resemble quite closely the neighbouring units. Thus the grid map will be full of local regions that share common properties. If the nodes of the grid are coloured in the sense of a heat map according to the values of the prototypes one feature (dimension) of the original data at a time, one can detect that close by nodes share similar values and that the colouring (height of the map) usually has a sub-continuous smooth surface. Based on these surfaces it is possible to visually detect regions that have high/low values for certain interesting features. Comparing multiple surfaces can help to find the features that have similar or contradicting patterns, which helps hypothesis generation and understanding of the possible interplay between the features. Correspondingly, one might possibly detect features that show close to no "heat patterns" in the grid, only a random fluctuation, and thus, find variables that have close to no role in the organization of the map. Furthermore, these prototypes can be used in clustering in order to find classes within the data [40].

The trained SOM can be used to visualize the trained prototypes, but more importantly, also new data vectors can be mapped to its units. This can be done by computing the distances between the presented data vector and all the SOM prototypes (in the feature space $\mathbb{R}^d$), and mapping the new data to the node corresponding to the closest prototype. These new vectors can be mapped to their best matching unit (BMU) regardless of the possibly missing features, as the distance can be minimized also in the subspace of the present features in the same way as when training the SOM using data vectors with a missing values.

Even though being only one of the approaches among many provided by the field of machine learning and data mining, the SOM gives promising attributes to achieve the goals of this work. The multidimensional data with quite a lot of missing values can be used to train the SOM and the resulting mapping can be used to gain additional knowledge of the interplay between the measured variables. Furthermore, the trained prototypes can be also clustered into novel phenotype classifications and they can be further used to classify new patients.

3.1.1 Free parameters of SOM

When implementing a SOM, decisions regarding multiple free parameters are to be made, regarding for example the size and form of the used grid, initialization of the nodes, and the decay rates of the learning algorithm. There is no optimal decision for any of these that would always outperform the others as the optimal parameters are always problem dependent. Fortunately, there are multiple "rules of thumb" and different methods to test the performance of the selected parameters to help the researcher make these decisions.

The SOM topology, i.e. the shape and connections of node grid, can be in principle selected arbitrarily, but there are two generally used topologies: a regular square or a regular hexagonal lattice. In the square lattice, each of the nodes (expect the ones in the very edges of the grid) have four neighbours connected to them. Correspondingly, in the hexagonal topology each of the nodes are connected to six neighbours. Advantages of these lattices are that they are regular and easy to visualize in a two-dimensional plane, which makes the human interpretation of the maps easier. However, these lattices have a set of special nodes which have less neighbours (the nodes at the edges and corners of the map) that tend to have special properties. Moving these nodes in the training iterations have less effect on the overall map (as there are less neighbours) and more extreme values tend to locate into these edges and corners of the SOM map. This phenomenon can be seen to occur by definition, as similar nodes seek to move closer to each other making distant nodes dissimilar. Therefore, the opposite edges and corners should be most distinct by definition and

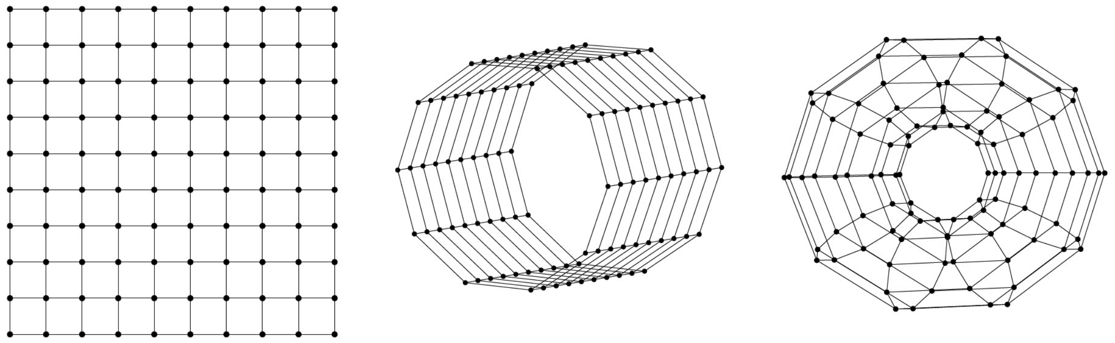


Figure 3: Different grid topologies used in SOM. From left, a regular square grid, a regular square grid folded into a cylinder and a regular square clynider folded into a torus. Figure editted from [38].

thus present the most extreme profiles of the data. An easy way of fixing this possibly problem causing property is to connect the edges of the lattice, so that the topmost nodes are connected as neighbours to the bottommost nodes, creating a cylinder. If the leftmost nodes are further connected to the rightmost nodes, the grid creates a torus. This way the grid can still be visualized in two-dimensional plane, but there is no corners or edges that would have special properties and each of the nodes have exactly the same number of neighbours and the same form of the neighbourhood. These different grid topologies are presented in Figure 3.

Another free parameter of the SOM is the size of the used grid. Both the number of nodes, and what is the proportion of height and width of the grid can be freely chosen. However, both of these selections have impact on the SOM performance and the selection is not straightforward to make. The number of nodes should not be too small in order to avoid too general prototypes i.e. nodes being representative of too many data vectors. Correspondingly, the number of nodes should not be too large, so that the resulting prototypes indeed would summarize the close by data vectors, and that there would not be too many empty nodes where no data vectors are mapped at all. For similar reasons the height/width proportion of the grid should not be too large or small. Multiple different methods to select these parameters have been proposed, but there is no guarantee that any of these would yield better results than the others. One of the methods to determine the grid size is based on the general "rule of thumb", implemented for example in the SOM-toolbox for MATLAB [38]. By this rule the goal number of nodes in the SOM grid, m, is set at

$$m = 5\sqrt{n}, \tag{5}$$

where n is the number of training samples. The ratio between side lengths is set to be the ratio between the two largest eigenvalues of the covariance matrix of the training data.

Also, the initial positions of the grid nodes in the data space $\mathbb{R}^d$ is one of the free parameters to be selected. Generally for this, there are two different approaches: a random initialization and a data analysis based initialization. The latter can be implemented based on a certain projection of the data, for example projection to the subspace spanned by the top two eigenvectors. However, none of these methods guarantee a better solution compared with the others [41].

Finally the shape and form of the neighbourhood function, its learning rates and decaying parameters, as well as the number of training iterations are to be selected. The neighbourhood function can in principle be of any form and arbitrarily selected, however, the most used forms were introduced in chapter 3.1 in Equations (3) and (4). The corresponding monotonically decreasing decay parameters including the learning rate are usually implemented as linearly decreasing, exponential or inversely proportional to the time, but the exact rule is not important [36]. The learning rate can start from close to a unity and the neighbourhood should initially include a large proportion of the nodes in order to allow the global ordering. Ultimately, the number of iteration steps should be large enough to allow a good statistical accuracy, and a value of "at least 500 times the number of network units" has been suggested by Kohonen himself [36].

3.1.2 Multi-layer SOM

The basic model of the SOM can also be extended to contain multiple layers of SOM grids that are trained simultaneously, using different datasets (for the same samples) in each of the layers. These datasets can contain even a different number of totally different features, as the training of the layers is done independently in their own $\mathbb{R}^{d_j}$ dimensional feature spaces. In principle, the only difference occurs during the process where the index of the winning node, c, is selected during the training. Whereas in a traditional single layer SOM the selection is made according to the distance in d-dimensional space (see Equation (1)), in multi-layer SOM the decision is made according to the sum of weighted distances throughout multiple d_j-dimensional spaces, each corresponding to one of the layers j and therefore one of the datasets, i.e.

$$c = c(\boldsymbol{x}(t)) = \arg\min_i \sum_{j=1}^{l} w_j ||\boldsymbol{x}_j(t) - \boldsymbol{m}_{i,j}(t)||^2, \tag{6}$$

where l is the number of layers, w_j is the weight associated to layer j, $\boldsymbol{x}_j(t)$ is the data vector of presented input in the dataset corresponding to the layer j and $\boldsymbol{m}_{i,j}(t)$ is the current prototype of the i-th node in the j-th layer.

This selected node therefore describes best the presented data point "on average" throughout the layers. To clarify, the grids in different layers are of the same size and shape, and the nodes in the same positions corresponds to each other. At every iteration, the selected winning node is the same in every layer, and that node is moved closer to the presented data point (according to its position in each of the d_j-dimensional data spaces of the corresponding layer) in every layer individually in the same fashion as in the single layer SOM. The weights of the layers can be adjusted to give more importance to certain datasets, but the intuitive approach is to set them all equal or corresponding to the number of features of the dataset in the corresponding layer. In this study, both of these approaches would yield the same outcome as the different layers used are of the same size corresponding to the cross-sectional measures in the different time points.

The training data presented for the SOM needs to contain data for each of the layers in order to train them. However, samples with totally missing data for a certain layer can still be mapped back to the SOM after the prototypes are trained, in a similar way as in the traditional single layer SOM the samples with missing data can be mapped back to the grid. An implementation of the multi-layer SOM described above can be found in the `kohonen` R package implemented as `supersom` [42], which was used in the analyses of this work.

3.1.3 Evaluation of the SOM mapping

The form of the trained SOM is heavily affected by the selected free parameters, and more importantly, it is random by definition as the training vectors are presented in randomized order during the training iterations. Thus, different measures have been developed to test whether the resulting mapping is good, which can later be used to evaluate the outcome when the free parameters are tuned, or to select the best fit of the SOM throughout different iterations.

Probably the most used measure is the quantization error (QE), which measures how closely the trained prototypes resemble the actual data vectors. QE is defined as the average distance between introduced sample vectors and their BMU of the SOM prototypes in the feature space $\mathbb{R}^d$. If the prototypes resemble the samples very closely, the distance between the vector and its BMU in the feature space is small. Controversially, if the trained prototypes end up presenting some arbitrary subsection of the feature space, the distance for the mapped (relatively different) samples can become large. Thus small QE will denote good mapping and larger ones worse mapping.

A similar measure, describing the smoothness of the prototype grids, is called topographical error (TE). If the mapping is smooth, the mapped samples should be located approximately in between two adjacent prototypes, and thus the closest prototype (BMU) and the second closest prototype should be connected in the SOM grid. The basic form of TE computes the proportion of sample vectors, whose first BMU and second BMU are not adjacent. The value for TE is limited to the interval $[0, 1]$, and smaller values denote smooth mapping and larger values poorer mapping. However, this measure does not capture the whole picture, as it does not consider how far the second BMU is located if it is not adjacent. Thus, this proportion based TE can also be extended to distance based TE, where the error is defined as the average distance (in the grid plane $\mathbb{R}^2$) between the first and second BMU of multiple sample vectors. This measure will penalize the samples whose non-adjacent first and second BMU are far away in the grid plane, but will suffer only small penalties if they are still relatively close. The magnitude of the error is not limited to a certain interval, but is dependent on the grid size. However, smaller values again denote smoother mapping.

All three of these measures were used in this work when the performance of the trained SOM was evaluated.

3.2 Data clustering

Clustering is a general problem setting, where the goal is to divide the given data into a (possibly unknown) number of groups, so that the elements within the same group resemble each other more than the elements between different groups. The field of clustering is very broad and there exists multiple different approaches to tackle the problem. In this work the clustering is used to group the trained prototypes of the multi-layer SOM into different classes that later can be used to classify the individual patients accordingly.

The clustering approaches can be divided into classes by using multiple different criteria. One methodological difference having a major impact on the clustering problem at hand is the type of the input that the given method requires. Where some of the clustering methods need the exact vector coordinates for the data points to be grouped, some can manage using only the (dis)similarity matrix presenting the relations between the data points.

As the multi-layer SOM prototypes to be clustered have position and coordinates in multiple d-dimensional spaces, applying the basic implementations of most of the clustering methods requiring the coordinate presentation of the data is not possible. Some of these methods could also be hand tailored to suit the problem at hand, but this is out of the scope of this work. There are also some workarounds to avoid this problem, but they are not generalizable in all settings. Thus, if one wishes to expand the pipeline of this work to other problems, diseases or traits, these workarounds cannot necessarily be reasonably applied any more.

Luckily, the similarities between the multi-layer SOM prototypes are straightforward to determine, if one applies the same rules that are used when finding the BMU in the training or mapping phase of the SOM as shown in the Equation (6). Applying the same principle it is possible to determine the distance between any two prototypes by computing the weighted average of euclidean distances throughout the layers. Thus, it is easy to compute the distance matrix for all pairs of prototypes. This distance matrix can be further negated to create a similarity matrix and either of these is sufficient to perform the clustering using multiple different clustering methods.

Therefore, this work concentrates only on the clustering methods that can be applied to cluster the data based on the (dis)similarity matrix. Three different methods were selected that all have well documented R implementation available. The selected methods are widely used and/or have been shown to perform well when applied on biological data before. The selected methods will be further introduced in the following sections. First, in section 3.2.1 one of the the simplest forms of clustering, the hierarchical agglomerative clustering, is introduced with the Ward's criteria for cluster merging. Next, section 3.2.2 introduces the concept of spectral clustering, an approach which is based on clustering a spectral decomposition of a similarity matrix. Finally, section 3.2.3 highlights one of the recently suggested clustering methods, an affinity propagation, that is based on an iterative "message passing" between the data points to be clustered.

3.2.1 Agglomerative hierarchical clustering and Ward's criteria

Agglomerative hierarchical clustering is a relatively simple class of traditional clustering methods. Common for all the agglomerative clustering methods is the general approach, where clusters are grown step by step by merging the previous clusters by optimizing a certain criteria. At first, every data point belongs to its own cluster and these clusters are merged step by step. Usually this iterative cluster merging is continued all the way until only one cluster remains. By keeping track of the values of the optimized criteria during the merges, it is possible to build a dendrogram, a certain type of graph showing how the data points were merged into clusters (for example see Figure 12 in section 5.4). Based on this dendrogram, it is possible to gain insight to the structure of the data, and possibly the number of underlying clusters. The final cluster assignments are made by "cutting the dendrogram" from a certain level creating k clusters, which corresponds to stopping the cluster merging after $(n-k)$ merge steps. Therefore, the number of clusters does not have to be predetermined, but it can be selected *post hoc* based on the dendrogram in order to create meaningful cluster assignments.

There exist multiple different possibilities to select the cost function to be optimized, and the selection heavily affects the resulting dendrogram and clusters. All of the approaches have their own advantages and they are suited for certain problems. However, it is hard to tell in advance which of the cost functions is most suitable for the given data. Fortunately, there exists a measure, the agglomerative coefficient (AC), that can be used to determine the quality of the resulting dendrogram as introduced by Kaufman and Rousseeuw in [43]. To compute AC, first for each data point i the ratio between cost function value at the first cluster merging and the cost function value of cluster merging in the final step, denoted by $m(i)$, is computed. Finally the AC is defined as:

$$AC = \frac{1}{n}\sum_{i=1}^{n} 1 - m(i), \tag{7}$$

where n is the number of elements to be clustered, i.e. as the average of $1 - m(i)$ over all the initial clusters. The AC is limited to the interval $[0, 1]$, where higher values denote a good clustering. AC tends to have higher values when the number of data points n increases, thus it cannot be used to compare clusterings of datasets of very different sizes. However, there is no such bias when comparing different clusterings of the same dataset. Later in this work, AC is used when selecting the most suitable optimization criteria for the agglomerative clustering.

The traditional selection of the cluster merging criteria is based on single linkage or complete linkage. In single linkage, the cluster distances are determined by the pair of points that have the shortest between-cluster distance. On the contrary, in complete linkage the between-cluster distance is the maximum distance between all pairs of points between the clusters. Similarly, it is easy to set the between cluster distance to the average or median of all the distances between clusters or as a distance between cluster centroids. For example, the R package `stats` [44] offers multiple of these cost functions implemented in function `hclust`. In this work, five of these criteria (single-linkage, complete-linkage, average-linkage, McQuitty's, and Ward's criteria) were used. All of these approaches yield slightly different clusters, and they prefer clusters of a different shape. Next, one of these, the Ward's criteria, will be introduced in more detail.

Among the agglomerative hierarchical clustering methods, Ward's clustering is the only one that is based on the classical sum-of-squares criterion [45]. Thus it minimizes the within group dispersion at each cluster fusion. The cost function to be minimized at each merge step is:

$$D(c_i, c_j) = \frac{|c_i||c_j|}{|c_i| + |c_j|}||c_i - c_j||^2, \tag{8}$$

where $|c_i|$ is the size of cluster i, and $||c_i - c_j||^2$ is the squared euclidean distance between the cluster centers. Although the value of cost function is computed in the euclidean space, an effective implementation based on the Lance-Williams dissimilarity update formula allows the optimization of merges relying purely on the dissimilarity matrix of the elements.

Two different implementations of the Ward's criteria are found in literature, which give slightly different results. In this work the algorithm implemented in R package `stats` [44] in the function `hclust` as `method='ward'` in R v.3.0.2 is used with a squared dissimilarity matrix, but the resulting dendrogram is rescaled using a square root transformation in order to yield results corresponding to the original Ward's criteria, as discussed by Murthag *et al.* in [45]. Thus the results fully correspond to using `method='ward.D2'` in R v.3.0.3 onwards. Scaling the height of the dendrogram becomes important when computing the AC and comparing different clustering methods.

3.2.2 Spectral clustering

Spectral clustering is yet another large family of clustering techniques. There exist multiple different implementations and algorithms for them, however, the general approach is mostly the same. First, the affinity matrix (or simply similarity matrix) of the data is computed, and the graph Laplacian is constructed from this matrix. Next, the eigenvalue problem is solved, and the k eigenvectors corresponding to the k largest eigenvalues (where k is the number of desired clusters) are chosen. Finally, the data is clustered in this subspace with a traditional clustering method, for example by using the k-means algorithm [46].

In this work the method proposed by Ng *et al.* [47], implemented in R package `kernlab` [48] as the function `specc`, is used and the technical details for this approach are presented next.

The implementation allows either a precomputed similarity matrix (used directly as affinity matrix $\boldsymbol{K}$) or the raw data as an input. If the latter is used, the affinity matrix $\boldsymbol{K}$ is first computed using a desired (user specified) kernel function. Next the normalized Laplacian is computed as:

$$\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{K}\boldsymbol{D}^{-1/2}, \tag{9}$$

where $\boldsymbol{D}$ is a diagonal matrix of form $\boldsymbol{D}_{ii} = \sum_{j=1}^{m} \boldsymbol{K}_{ij}$. Thus each diagonal element is the sum of the corresponding row of the similarity matrix $\boldsymbol{K}$. Then the first k largest eigenvectors of the Laplacian $\boldsymbol{L}$ are computed, where k is the desired number of clusters. These eigenvectors are used to create a $n \times k$ matrix $\boldsymbol{Y}$, where each column corresponds to one of the top eigenvectors in a decreasing order. The rows of this matrix are further scaled to unity length. Finally each of these rows are used as a coordinate representation of the original data points (row i corresponding to the i-th data point in the original distance matrix) and clustered to k clusters by using the k-means approach correspondingly. If the data contains clusters, they are clearly separated in this spectral presentation, and the traditional clustering methods applied in the last phase perform well in the task.

The clusters that spectral clustering can find are not restricted by shape or form, and it can successfully cluster groups of an arbitrary shape and size opposed to using only the traditional clustering methods. A good example of this is two intertwined spirals (as presented by Karatzoglou *et al.* in [48]), circles within each other and clearly connected shapes of an arbitrary form (as presented by Ng *et al.* in [47]). Therefore, it is a good candidate for the clustering approach to be used when the SOM prototypes are clustered, as no presumptions of the resulting clusters are to be made.

3.2.3 Affinity propagation

The third and final clustering approach used in this work is the newest among the different approaches examined. Affinity propagation was proposed in 2007 by Frey and Dueck [49], and it has already been successfully used in many tasks in the field of bioinformatics [50]. Affinity propagation is a prototype-based clustering method, where the cluster centers are selected among the data points, and all the other points are assigned to these "examplars" to form the clusters. The special properties which the affinity propagation offers are that it simultaneously considers all the points as potential examplars by "message passing" (thus avoiding randomness). It can also find the examplars relying purely onto the similarity matrix without the exact vector coordinates, opposed to most prototype-based clustering approaches. Furthermore, the affinity propagation can determine the most suitable number of clusters during the clustering, or it can be predefined by the user.

The algorithm by Frey and Dueck [49] has been implemented in R package `apcluster` [50], with minor improvements on speed and flexibility compared with the original implementation. Next the algorithm based on this implementation is introduced.

First, the algorithm takes as an input the similarity matrix $\boldsymbol{S}$ of the elements to be clustered. On the abstract level, each value $\boldsymbol{S}_{ik}$ defines how well the data point with index k is suited to be an examplar for data point i. This similarity matrix can be arbitrarily defined real valued similarity (where also negative values are allowed), and is not required to satisfy the properties of a metric or even to be symmetric. In this similarity matrix, the diagonal elements $\boldsymbol{S}_{kk}$ (referred in the original publication [49] as "preferences") have a special role, as they affect how likely each of the points are selected as examplars. If each data point to be clustered is *a priori* as likely to be an examplar, the diagonal values are to be set equal. However, the magnitude of the values still affect the number of resulting clusters. Frey and Deuck suggest that the input preferences are to be set to the median of the input similarities, or as the minimum of them if a smaller number of resulting clusters is desired [49]. In the R implementation the median is used as a default, but also a new optional argument `q` is introduced. It defines the quantile of similarities to be used as an input preference and thus allows a smooth transition between the two approaches originally suggested.

The clustering algorithm itself consists of iteratively updating two matrices, the "responsibility" $\boldsymbol{R}$ and the "availability" $\boldsymbol{A}$. Each element $\boldsymbol{R}_{ik}$ of the responsibility matrix presents the cumulated evidence for element k to serve as an examplar for an element i. Correspondingly, each element of the availability matrix $\boldsymbol{A}_{ik}$ presents how suitable it would be for point i to select k for its exemplar, taking in account how well the other points support k to be an exemplar at all.

First, the availabilities $\boldsymbol{A}_{ik}$ are all initialized to zero. After this both matrices are updated one after another, starting from the responsibility matrix as:

$$\boldsymbol{R}_{ik} = \boldsymbol{S}_{ik} - \max_{k' \neq k} \{\boldsymbol{A}_{ik'} + \boldsymbol{S}_{ik'}\}. \tag{10}$$

Next, the availabilities are updated as:

$$\boldsymbol{A}_{ik} = \min \left\{ 0, \boldsymbol{R}_{kk} + \sum_{i' \notin \{i,k\}} \max \{0, \boldsymbol{R}_{i'k}\} \right\}, \tag{11}$$

with a special update rule for the diagonal elements:

$$\boldsymbol{A}_{kk} = \sum_{i' \neq k} \max \{0, \boldsymbol{R}_{i'k}\}. \tag{12}$$

The message passing can be stopped after a predefined number of iterations, after the magnitude of passed messages fall below certain threshold, or after the local changes (examplars and examplar assignments) stay unchanged for a number of iterations. For any step, point i can be identified to belong to cluster c_k defined by examplar k by:

$$c_k = \arg\max_{k} \{\boldsymbol{A}_{ik} + \boldsymbol{R}_{ik}\}, \tag{13}$$

or being the examplar for the cluster c_k if $k = i$. For further discussion, interpretation of the message passing procedure, small alterations to avoid numerical oscillation, and further comparisons with other clustering methods, refer to the original publication by Frey and Deuck [49].

3.3 Genetic data and analyses

The ultimate goal of this work is to compare the created novel phenotypes with each other and find the exact genetic markers differing between the groups. Additionally, the different clustering methods will be evaluated by their ability to distinguish groups with different genetic background at a wider scale. Both of these tasks require different methods to analyse the genetics of the novel phenotypes, which will be further described in the following section. First, section 3.3.1 will introduce the concept of array-based genome-wide genotyping, and the data that can be created using this approach, which will be the base for the rest of the genetic approaches in this work. Next, section 3.3.2 will introduce the statistical view of the heritability measure, which will later be used to evaluate the performance of the clustering methods. Thereafter, section 3.3.3 briefly will review the linkage structure of the human genome and how it can be used to impute additional markers based on the ones directly genotyped. Finally, section 3.3.4 will introduce the statistical tests used to evaluate the genetic associations between case-control phenotypes.

3.3.1 Array-based genome-wide genotyping

The human DNA consists of approximately three billion nucleotides of genetic code [51], that is mostly shared between individuals. However, many forms of genetic variation exist. The most common and smallest form is a single-nucleotide polymorphism (SNP), where one base pair of the DNA is changed to another. These mutations are among the most studied, and are the target of GWASs.

For GWAS analyses, the genetic variant carried by each individual is to be solved for hundreds of thousands to even millions of SNPs, in thousands of patient samples. This can be achieved by using array-based genotyping approaches. One of the providers of commercial genotyping arrays is Illumina, whose approach for genotyping will be described in more detail next.

The base for the Illuminas approach is the BeadChip platform, where carefully designed oligonucleotides (single strand DNA probes) are attached to small "beads" (few µm in diameter) from their 5' end. The different beads each contain probes of only one type that will detect the DNA sequence next to the SNP to be genotyped. These beads are then arrayed on a silicon surface on a regular grid, so that there are multiple copies of each bead type across the array. Next the DNA sample to be genotyped is preprocessed, fragmented, and added to the array as single strand DNA. The DNA sequences in the bead probes will hybridize with these fragmented sample strands only if the complementary DNA sequences match exactly. After this, the fragments not hybridized are washed away from the array leaving only the fused DNA fragments to the array, attached to the beads. Next, the bead probes are extended by one nucleotide, which will match the SNP in the fused sequence. The extended nucleotides are stained with fluorescent colors (A and T with one, C and G with another), to differ between the variants. In addition, the chips contain pairs of special probes with parallel detection method, that can also distinguish A/T and G/C SNPs by design, where only one of the probes is extended in presence of genotyped variant. After the extension, the array is imaged, and the signals from each bead are recorded. In case of a homozygous genotype, the signal will contain only one of the wavelengths, and in case of a heterozygous genotype both of them. Finally, these intensity signals are normalized and preprocessed, mapped to the rs-codes of the SNPs, and the genotypes of the individual ("AA", "AB" or "BB") will be called for each chip based on the data [52, 53].

The array technologies are constantly developing and improving in speed, throughput and variety of SNPs captured. The array used to create the genotype data in this work was the Illuminas Infinum CoreExome BeadChip, that can genotype more than 550,000 SNPs, with an estimated throughput of 2,800 samples per week [54].

3.3.2 Heritability

From a technical point of a view, heritability is a population parameter describing the proportion of the phenotypic variance attributable to the genetic variance. Even though widely used in common speech to describe whether a trait can be passed from a generation to another, in technical means it has a wider interpretation and specific definitions as reviewed by Visscher *et al.* in [55].

The base for the following definitions is a model, according to which the phenotype (P) of an individual can be expressed as a sum of unobservable genetic components (G) and unobservable environmental components (E) both affecting the phenotype outcome: $P = G + E$. Thus, the variance in the observable phenotype σ_P^2 can be partitioned to variances by the genetic component σ_G^2 and the environmental component σ_E^2:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2. \tag{14}$$

The broad-sense heritability of the phenotype is defined as the ratio of the genetic and phenotypic variance:

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}, \tag{15}$$

thus, expressing the phenotypic variance attributable to the genetic variance. However, the genetic variance can be further divided into components describing the additive genetic effect (σ_A^2), the dominant genetic effect (σ_D^2) and epistatic genetic effects (σ_I^2, i.e. the interaction between alleles at different loci) as $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$. Usually, instead of the broad-sense heritability, a narrow-sense heritability (also referred as strict sense heritability, or just as heritability) is used, describing the variance due to additive genetic factors:

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}. \tag{16}$$

As the matter of fact, the partitioned phenotypic variance should also contain the covariance of the genetic and environmental factors as well as the variance due to G-E interaction. However, these are usually ignored as they cannot be estimated as discussed in [55]. Later in this work, when the term heritability is used, narrow-sense heritability, i.e. the proportion of phenotypic variance attributable to additive genetic factors, is referred.

Using genetic data from genome-wide genotyping, it is possible to estimate the additive genetic variance by multiple different methods. One of the tools used to achieve this is the Genome-wide Complex Trait Analysis (GCTA) software. It uses the genetic relationship matrix (GRM) of unrelated individuals in a mixed linear model to estimate the variance explained by the hundreds of thousands of genotyped single-nucleotide polymorphisms (SNPs) via the restricted maximum likelihood (REML) approach [56].

In the GRM, $\boldsymbol{A}$, each element $\boldsymbol{A}_{jk}$ describes genetic relationship between individuals j and k, defined as:

$$\boldsymbol{A}_{jk} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}, \tag{17}$$

where, x_{ij} is the number of copies of the reference allele of the i-th SNP for the j-th individual, p_i is the frequency of the reference allele, and the summation goes over all the N genotyped SNPs.

The GRM defined this way is used in a mixed linear model of form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \epsilon, \text{ with,} \tag{18}$$

$$\text{var}(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{A}\sigma_g^2 + \boldsymbol{I}\sigma_\epsilon^2, \tag{19}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of phenotypes of the n individuals, $\boldsymbol{\beta}$ is the vector of fixed effects for the covariates $\boldsymbol{X}$, and $\boldsymbol{g}$ is an $n \times 1$ vector of total genetic affects of the individuals with $\boldsymbol{g} \sim N(0, \boldsymbol{A}\sigma_g^2)$. $\boldsymbol{I}$ is an $n \times n$ identity matrix, ϵ is the vector of residual effects with $\epsilon \sim N(0, \boldsymbol{I}\sigma_\epsilon^2)$, and σ_g^2 is the (additive) genetic variance explained by all the SNPs, which is to be estimated via the iterative REML approach. For the details of the implementation of the estimation method relying on an average information algorithm, and further possibilities to partition the heritability to separate chromosomes, refer to the original publication by Yang *et al.* [56].

Using this approach, GCTA can estimate the additive genetic variance and therefore the narrow-sense heritability captured by the genotyped SNPs (i.e. observable with the used genotyping platform, thus, sometimes referred as array-heritability). In this work, the GCTA (v.1.24.4) was used when computing heritabilities between the novel phenotype classifications.

3.3.3 Genetic linkage and genotype imputation

Half of the genetic code in the DNA of an individual is inherited from the mother, and the other half from the father. When the egg and sperm are created in meiosis, parts of the DNA may change between the paternal and maternal chromosomes and this mixture of DNA will be later inherited to the next generation. This crossover of DNA can happen almost all around the genome, but the DNA sequences between two of these crossovers are inherited as a whole. However, the occurrence of these crossovers are not totally random. The human DNA contains multiple so called recombination hotspots, where these crossovers happen more frequently than in the surrounding regions. Thus, the long DNA sequences between these hotspots are usually inherited directly from a generation to the next in a single haplotype. Thus genetic variants in same haplotype and close-by regions tend to correlate with each other. This phenomenon is called linkage disequilibrium (LD), and it is taken advantage of in genotype imputation.

The SNPs selected in the array-based genotyping methods are based on the relatively well known haplotype structure of the human genome, so that each of the genotyped SNPs will also capture the structure of the nearby genetic region. However, this property can also be used to predict genetic markers not being directly genotyped in the samples. The basic idea is to observe the successfully genotyped SNPs of an individual, and compare them to reference genomes with a tighter SNP panel genotyped (for example from the HapMap [57] or 1000Genomes [58] studies) to solve the haplotype of the individual. In many of the cases, it is possible to find with high confidence the corresponding haplotype, and thus predict the most likely variant of nearby SNPs with high confidence, i.e. impute the missing variants. Imputing the genotype data can help both the fine mapping of regions around the genotyped SNPs (i.e. increase resolution), and by increasing the overall power to detect the genetic associations [59, 60].

In most of the approaches, the methods yield three different probabilities (summed to one) for every variant: the probability of a patient to be homozygous (p_{aa} for being homozygous for the minor variant and p_{AA} for being homozygous for the major variant) and the probability of the patient to be heterozygous (p_{Aa}). These genotype probabilities can be used directly in certain analysis methods, or sometimes the most likely genotype is selected to be used as such. Another widely used approach is to transform the genotype probabilities to genetic dosages as:

$$D = 0 \times p_{AA} + 1 \times p_{Aa} + 2 \times p_{aa}. \tag{20}$$

This dosage is a continuous extension of the count of minor alleles, and it is restricted to the interval $[0, 2]$, but captures also the possible uncertainty of the imputation (compared with using most likely genotypes directly). It can be used in many analysis settings similarly as the counts of directly genotyped SNPs, as next presented in section 3.3.4.

3.3.4 Genome-wide association analyses for case-control setting

The goal of Genome-wide association study (GWAS) analyses is to test multiple (up to tens of millions) genetic markers for their association with the phenotype of interest in a large population of unrelated individuals. There exist multiple different tests and models to evaluate genetic association in different settings. However, in the GWAS setting the analysis is usually performed by testing every genetic marker independently, in order to make the analyses scalable. The number of possible genotype combinations increase rapidly if more genetic markers are considered simultaneously, and therefore the required time for computation would become quickly infeasible in a genome-wide setting using other approaches.

Even when the approaches are limited to single marker tests, the selection of available tools is wide. In this work we consider only one of the most widely used tools, PLINK [61, 62], and the test models it implements. We further restrict the theory to testing binary traits, as all the phenotypes created in this work are treated as such.

The simplest form of association testing implemented in PLINK performs a chi-squared test with one degree of freedom to the allele frequencies between cases and controls for every genetic marker separately (initiated with `--assoc` flag).

PLINK has also models implementing Fischer's exact test, Cochran-Armitage trend test (dividing the table to three based on genotype) and different models of inheritance (dominant or recessive, further grouping the genotype columns of count tables). However, all of these approaches suffer from the limitation that they cannot include any covariates to the model. Thus it would be impossible to include the effect of the known risk factors, or different confounding factors, and correct these analyses for their effects. Thus, if an additive genetic model is desired, a logistic regression is used instead of the models mentioned previously due to its flexibility and ability to include covariates into the model. In this approach, a logistic regression line,

$$F(x) = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x)}}, \tag{21}$$

is fitted to the data, where β_0 is the intercept, β_1 is the regression coefficient for the genetic effect and x is the genotype (count of minor alleles). Now $F(x)$ can be interpreted as a probability of an individual being a case given the genotype x, as the logistic function maps all the real values to the interval $[0, 1]$. If the equation is solved as follows:

$$\frac{F(x)}{1 - F(x)} = e^{\beta_0 + \beta_1 x}, \tag{22}$$

the left hand side of the equation is the odds (proportion of probabilities) of the patient being a case (versus not being a case). Using this idea it is possible to define the odds ratio (OR) for a genotype containing one additional copy of the effect allele (compared with a genotype with one copy less):

$$\mathrm{OR} = \frac{\frac{F(x+1)}{1-F(x+1)}}{\frac{F(x)}{1-F(x)}} = \frac{e^{\beta_0 + \beta_1 (x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}. \tag{23}$$

Thus, the resulting exponentiated fit coefficient β_1 is to be interpreted as an odds ratio for the disease risk, i.e. a multiplier for risk for each copy of the genetic variant. The statistical significance for genetic association is the two sided p-value from a hypothesis $\beta_1 = 0$.

Using this model it is possible to add the effect of other factors by:

$$F(x_1, x_2, ..., x_n) = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} = \frac{1}{1 - e^{-\boldsymbol{\beta x}}} = F(\boldsymbol{x}), \tag{24}$$

where $x_2, ..., x_n$ are the additional covariates included in the model, and $\beta_2, ..., \beta_n$ are the corresponding regression coefficients. Corresponding to the model without additional covariates, β_1 and x_1 are the genotypes and regression coefficient for the genetic effect. The interpretation of the results and the statistical significance for genetic association are also the same. This model is implemented in PLINK and can be initiated with the `--logistic` flag. This approach can be applied directly also to continuous genetic dosages resulting from genotype imputation, and such an analysis can be initiated with `--dosage` flag in PLINK.

As this widely used GWAS analysis approach is based on single marker testing, the results should take into account the problem of multiple testing. However, this is not straightforward in genetic association testing, as nearby markers correlate with each other due to genetic linkage (for more details see section 3.3.3), and thus the tests performed for two SNPs in LD are not completely independent. Thus, the GWAS analyses have widely accepted p-value of $p < 5 \times 10^{-8}$ as a significance threshold for association, corresponding to a Bonferroni correction of an approximated one million independent SNPs in the European population ($p < 0.05/1,000,000 = 5 \times 10^{-8}$) [63].

4 Materials and methods

The base for this work is the comprehensive data collected from patients with T1D during the FinnDiane study. In this work, different computational methods are applied to this large dataset in order to create novel patient classifications. These new classes are further used in genetic analyses, with the goal to find genetic variants that could help to explain the pathophysiology and causes behind the diabetic complications. The following section describes in more detail the data and how the different methods were applied.

First, section 4.1 describes the FinnDiane Study, and the collected datasets that were used in the analyses of this work. Next, section 4.2 describes how the data was preprocessed, how the multi-layer SOM was trained, and how its parameters were selected. Then, section 4.3 describes how the prototypes of the trained SOM were clustered, how the different clustering methods were evaluated, and how the most suitable method was selected to be used to create the novel phenotypes. After this, section 4.4 goes through how the novel phenotypes of diabetic complication profiles were created, and finally section 4.5 describes the GWAS analyses performed for them.

4.1 The FinnDiane Study

This work is a part of the FinnDiane Study (1997–), which is an ongoing nationwide multicenter study of T1D and its long-term complications with the aim to define the clinical, environmental and genetic risk factors of diabetic complications with special emphasis on DN. The number of volunteer patients with T1D in the study is currently more than 7,000 and continuously increasing, which represents approximately 15-20% of the population with T1D in Finland (approximated to be currently roughly 40,000 patients). The comprehensive study protocol and patient recruitment have been described in more detail before [64]. Briefly, adult patients with T1D have been recruited by their attending physicians through 77 hospitals and primary healthcare centers (all five university hospitals, all 16 central hospitals, 26 other hospitals and 30 primary healthcare units) across Finland. During the recruitment process, multiple questionnaires are completed by the physician and the patients themselves, and both blood and urine samples are collected. Currently, more than 5,000 patients have been recruited this way. In addition, the FinnDiane study includes close to 2,000 patients recruited through the Finnish National Institute of Health and Welfare (THL) across Finland. The overall distribution of the recruited patients in the study follows closely the general population distribution of Finland as shown in Figure 4.

On top of the still ongoing patient recruitment, the prospective phase of the study started in 2004 calling the patients back for a follow-up (FU) visit. Thus far, more than 1,800 patients have also FU data collected. For this follow-up, samples and questionnaires corresponding to the baseline visit are collected. Additionally, the FinnDiane study uses the medical records and multiple national registries, linked via the personal identity code of the patients, to continuously update the clinical phenotypes of the participants.

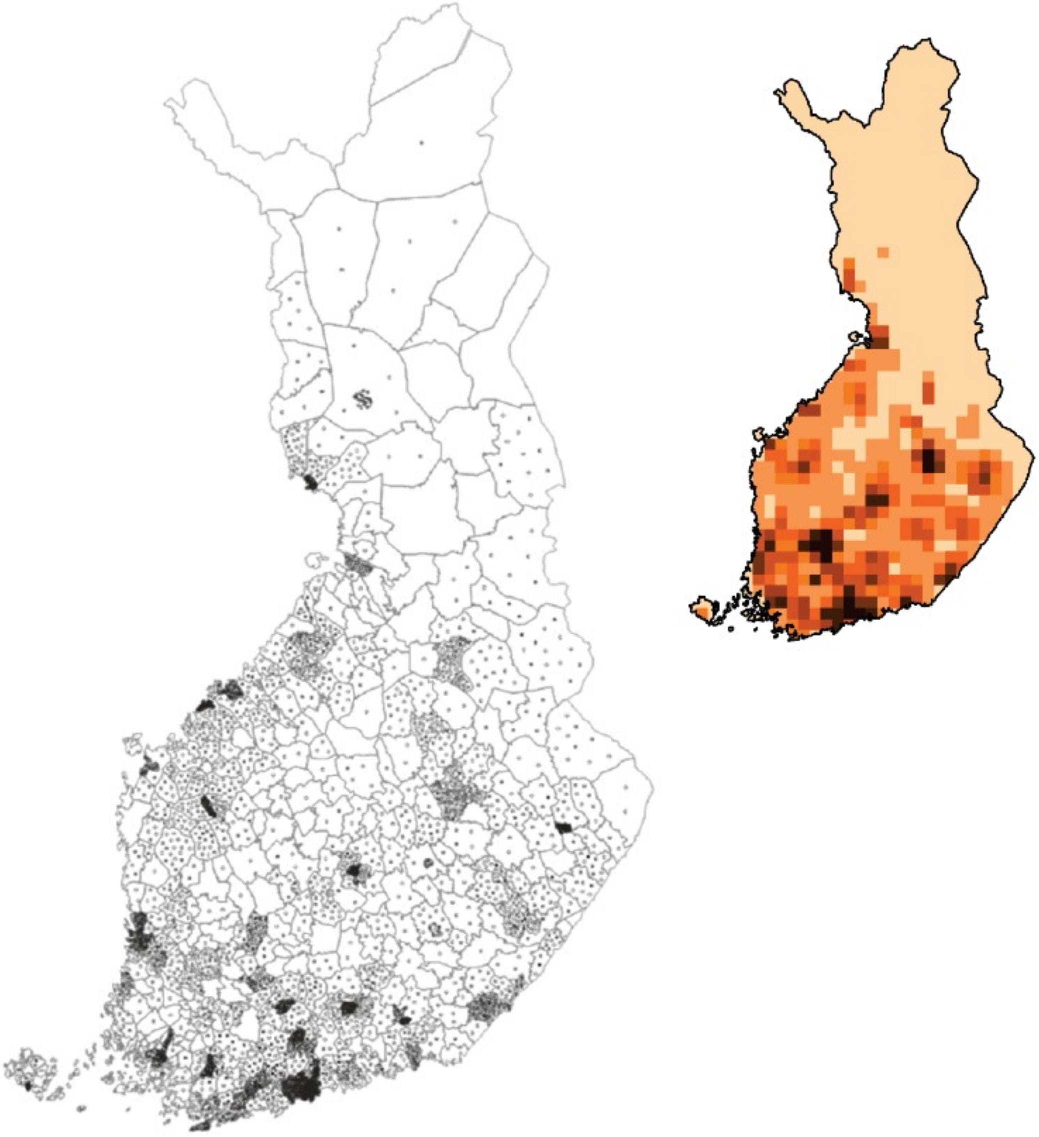


Figure 4: The distribution of the patients in the FinnDiane Study (left) and the distribution of the general population in Finland (right). Figures from [65] and [66], correspondingly.

Furthermore, a large proportion of the FinnDiane patients, including the patients recruited through THL, have an array based genome-wide genotyping performed on them. This data is further described in section 4.1.2.

The study protocol of the FinnDiane Study has been approved by the local ethnics committee, all the studies are conducted in accordance of the Declaration of Helsinki, and all the patients gave their informed consent prior to enrolment.

4.1.1 Clinical variables and patient inclusion

For this study, only the patients recruited through the FinnDiane recruitment process described above were selected, and the patients recruited through THL were excluded due to limited data on variables required for the analyses of this study. The patients were further excluded unless they fulfilled the generally used criteria for T1D:

- Age at diabetes onset < 40 years
- Insulin treatment initiated within one year of the diagnosis

This way, the patients with plausible T2D were excluded from the further analyses. Additionally, patients without genotype data were excluded in the initial screening. For the remaining patients, the clinical data collected throughout the FinnDiane Study was extracted from the FinnDiane database. As the data is constantly collected and new patients are recruited, the data was frozen at the 21st of October, 2015, and this data was used throughout the rest of the analyses. The data included measures from both the baseline (BL) visit and the follow-up (FU) visit as well as the latest available clinical phenotypes derived from registries and other clinical sources.

Within the extracted data, the form filled out by the attending physician included variables measured at routine clinical examinations (for example anthropometric measures and blood pressure levels), variables describing different clinical complication phenotypes at the time of the visit, variables describing medication and insulin treatment of the patient, and some family and pedigree related questions. Correspondingly, the self-report questionnaire contained some complication related questions, a wide range of questions related to diet and life style (such as smoking, alcohol usage, education and employment), and a wide panel of questions related to family history of diabetes and its complications.

The datasets of the laboratory measures of the blood contains variables describing lipid profiles of the blood, serum concentrations of different general markers, a few inflammatory markers, as well as some diabetes specific markers (such as glycated haemoglobin, $\mathrm{HbA_{1c}}$, describing the long-term glucose management of the patient). Correspondingly, the laboratory measures from the urine samples consisted of variables describing electrolyte and protein concentrations of the urine samples, collection specific variables (time and volume), as well as some diabetes and DN specific markers (such as urine KIM-1 concentrations).

Finally, the registry based data contained variables derived from Statistics Finland (latest vital status of the patient until 31st of December 2014), and the Hospital Discharge Register (HILMO, phenotypes for cardio- and cerebrovascular events until 31st of December 2013). Additionally, the data included the latest clinical phenotype definitions for DN collected and combined from multiple sources by other members of the FinnDiane group (sources such as Finnish Kidney Disease Registry, HILMO, other clinical patient data, dates for latest data varying).

All this data was combined and pruned as described later in section 4.2.1.

4.1.2 Genome-wide genotyping and genotype imputation

A total of 6,255 patient samples in the FinnDiane cohort have genome-wide genotyping performed for them at the University of Virginia, using Illuminas Infinum® CoreExome BeadChip. The used chip contains all the highly informative tagSNPs from Illuminas HumanCore BeadChip, as well as a wide panel of exome markers from Illuminas HumanExome BeadChip. Overall, the used chip contains probes for approximately 550,000 SNPs [54].

The genotyping of the FinnDiane samples was performed in three different batches. Genotype calling and comprehensive quality control (QC) was performed for each batch separately at the university of Virginia by a third party member. This QC filtered both poor patient samples (testing multiple quality criteria such as low genotyping rate, possible sample mix-ups, extreme heterogeneity, and gender error) and SNPs with poor quality (such as low genotyping rate, markers not in Hardy–Weinberg equilibrium, and markers associated with gender issues). These three batches were merged in FinnDiane, including only SNPs and patients that passed QC in each batch separately. In this merging the SNPs were further filtered based on batch association and finally all monomorphic SNPs were removed from the data. This merged data contains a total of 6,010 patients passing the QC and more than 300,000 SNPs with a higher than 99.95% average call rate. This data from direct genotyping was used to later compute the heritabilities of the defined novel phenotype classes.

This data was also imputed by a third party member using Minimac3 software (version 1.0.13) with 1000 Genomes Phase 3 version 5 genotypes (updated on 20th of October 2015) as a reference genotype panel. The imputed data contains approximately 8.7 million SNPs with sufficient imputation quality (PLINK INFO information criteria > 0.8) for the patients, of which approximately one million are rare (minor allele frequency MAF $< 1\%$), 1.7 million relatively uncommon ($1\% <$ MAF $< 5\%$) while the rest (6 million) are common. The imputed data was used in the GWASs to pinpointed the genetic markers associated with class assignments.

4.2 SOM of patients with T1D

The comprehensive clinical data collected in the FinnDiane study was used to train the dual-layer SOMs that created the core for this work. Data from two different time points was used in the two different layers of the SOM, in order to capture also the progressive nature of the diabetic complications. The resulting prototype profiles of the trained SOM were clustered into varying number of classes. The trained prototypes were also used to map the individual patients back to the SOM, and to solve in which of the created classes each individual belongs. The following subsections will present the preprocessing of the clinical data, how the free parameters of the SOM were selected, and how the SOM was trained.

4.2.1 Variable pruning and data normalization

First, all the data was combined from different sources: questionnaires filled out by the attending physicians, self-report questionnaires, laboratory measures from blood and urine samples, and data from national registries. After this, the variables describing latest clinical complication status were separated as outcome variables and pruned later by their missingness in the remaining patients. Meanwhile, the rest of the variables (the input variables) were further filtered by including only numerical and class variables (either binary, categorical, discrete or continuous) before further pruning.

In the next step, both the remaining input variables (data matrix columns, m) and the patient visits (data matrix rows, n) were pruned based on the missingness without further consideration of the possible clinical relevance of the variables in order to avoid possible selection biases. The patient visits, BL and FU, and the corresponding measured variables were pruned in separate batches simultaneously. An iterative approach was applied, where the pruning threshold was slowly increased in ten steps starting from 1%, until a patient-wise non-missing fraction of 1/2 and a variable-wise non-missing fraction of 1/3 were reached. For each step, if a variable would have been pruned only in either the BL or FU dataset, it was removed from both datasets instead, in order to keep the included variables the same at the different time points. The iterative increase in pruning threshold was applied as the raw data contained both patient visits and variables with almost completely missing data, that would cause overestimated missingness and thus prune variables and patient visits close to the desired cut-off levels if direct cut-off threshold was used. At this point, the categorical variables were transformed into sets of binary variables.

Next, all the variables were tested for their linear independence using a subsample of the patient visits with complete data for all the remaining input variables passing the missingness threshold. The testing was performed by fitting a linear regression model, where each of the variables were used as a dependent variable one at a time and all the other variables were used as explanatory variables. If the model explained the dependent variable using the other variables with a large coefficient of determination ($R^2 > 0.95$), it was considered to be linearly dependent. After the linearly dependent variables were identified, they were manually pruned until no such dependencies were observed for any of the variables. In this selection, biologically redundant sets of variables were pruned, preferring original measures over derived values. For example the systolic blood pressure (SBP) and the diastolic blood pressure (DBP) were preferred over the mean arterial pressure (MAP), which all have a linear dependency of the form $\mathrm{MAP} = \frac{1}{3}(\mathrm{SBP} + 2 \times \mathrm{DBP})$. Excluding the directly derived variables eliminated most of the linear dependencies, and for the rest, within the sets of variables creating the dependencies, the ones with more missing values in the full dataset were pruned off. Furthermore, all variables that were known as non-linear transformations of the included variables were further filtered (for example 24h urine excretion rate for any substance is computed by: concentration × 24h urine collection volume / collection time). In addition, some of the urinary markers were transformed to ratios compared with urinary creatinine as these ratios are clinically more relevant than the pure concentrations of these markers which are heavily affected by the urine volume.

Finally, all the outcome variables describing the latest clinical phenotypes were pruned by the missingness, requiring more than 50% non-missing values for the patients passing the input variable pruning.

This way, three sets of variables were created:

- Baseline data of variables corresponding to the measures of the BL visit when the patient entered the study (the input variables for 1st SOM layer)
- Follow-up data of the patients several years after the BL visit (median 6.3 years, IQR 5.0-7.6 years) with variables corresponding to the BL visit (the input variables for 2nd SOM layer)
- Latest available data of clinically defined diabetic complication diagnoses and mortality of the patients (the outcome variables)

The first two datasets were used as an input data in the training of the SOM and mapping of the patients, whereas the third dataset was used to evaluate the performance of the patient classification for the prognosis of mortality and complication status.

Before introducing the data to SOM, it is important to perform a meaningful scaling and preprocessing for the whole data. It has been shown that when SOM approaches are applied to clinical measures, such as in this work, the clear differences between genders in certain variables can direct the training of the SOM, suppressing clinically more relevant phenomena [67]. For example a woman's body contains on average more fat compared with a man's, which affects the distribution of lipid profiles, and men are on average taller than women. Thus, these gender differences were first masked as described in [67] by rank transforming and scaling the continuous variables. To achieve this the BL and FU datasets were handled separately. First, the dataset at hand was divided into males and females. Then, for each continuous variable, the values were ranked (in gender specific subsets), and the ranks were scaled to the interval $[-1, 1]$. Next, these scaled ranks were transformed as $x = (z^3 + z)/2$ after which the male/female subsets were merged again. The transformation of scaled ranks mimics the Gaussian probability density making it compatible with the Euclidean distance metric used in the computation of the distances to BMU in the SOM algorithm [66]. On top of masking the gender related differences, this approach also normalizes the shape of skewed distributions as well as handles extreme outliers so that they do not disturb the SOM training. Additionally, all the variables (binary, discrete, and already rank transformed continuous) were standardized to have a zero mean and an unit variance in the combined dataset of both genders.

Of this data, all the patients containing both BL and FU visits were used as training data for the SOM, and all the rest (containing either BL or FU data only) were used as test data when the SOM performance was evaluated.

4.2.2 Selection of SOM parameters

In this work, a toroidal-hexagonal grid was selected as the SOM topology, as the toroidal topology should create a smoother distribution of the resulting prototypes as there are no edges and corners for extreme profiles to escape. Indeed, empirical pilot analyses show that using a non-toroidal grid seems to create scattered (disconnected) clustering of the trained SOM-prototypes more easily, which is undesirable in the scope of the goals of this work (data not shown).

The size of the selected SOM was determined based on the "rule of thumb" presented in section 3.1.1. When the pair of rules,

$$m = 5\sqrt{n} = wh, \tag{25}$$
$$\frac{w}{h} = \frac{\lambda_1}{\lambda_2}, \tag{26}$$

are solved for h and w, the height and width of the map are set according to:

$$h = \sqrt{\frac{5\sqrt{n}\lambda_2}{\lambda_1}}, \tag{27}$$
$$w = \frac{5\sqrt{n}}{h}, \tag{28}$$

where λ_1 and λ_2 are the two largest eigenvalues of the covariance matrix of the input data, and n is the number of training vectors. When computing this covariance matrix, only the pairwise complete observations for pairs of the variables were used due to the missing data. However, in order for the toroidal hexagonal grid to fold correctly, an even height for the grid is required. Thus the rule for height h and width w were slightly altered and set according to:

$$h = 2\left\lfloor \frac{1}{2}\sqrt{\frac{5\sqrt{n}\lambda_2}{\lambda_1}} \right\rfloor, \tag{29}$$
$$w = \left\lfloor \frac{5\sqrt{n}}{h} \right\rfloor. \tag{30}$$

As a dual-layer SOM was used, the height and width were computed based on eigenvalues from both the BL and FU datasets corresponding to the different layers (and thus different data matrices), and the resulting height and width values closest to the desired number of nodes $m = 5\sqrt{n}$ were selected.

The prototype positions were also initialized randomly in order to test the robustness of the approach. The random initialization implemented as a default in the used `kohonen` package was selected. In this approach, the starting points of the grid nodes are drawn randomly without replacement from the training data [42].

For the neighbourhood function, the radial function presented in Equation (4) was selected, implemented as default in the `kohonen` package. For the learning rate, the default value of the `kohonen` package was used. Therefore, it will decrease linearly from 0.05 to 0.01 through the training iterations. The number of training iterations was optimized, during which the decrease of the neighbourhood function was kept in default values, i.e. the size of the neighbourhood started as containing two thirds of the map units and decayed linearly so that after approximately half of the iterations only the winning unit was adjusted. However, after the most suitable values for the training iterations were selected, the proportion of fine tuning (i.e. the rate of decrease in the neighbourhood function) was altered as described later.

To avoid under and/or over fitting, the number of training iterations was altered from the default values of the `kohonen` package. The default implementation introduces every training vector 100 times to the SOM to be trained, but this can be altered by the `rlen` parameter. First, a reasonable range of times that the training vectors would be presented to SOM was selected $\texttt{rlen} = \{50, 100, ..., 750\}$. Then for each of these, $n = 100$ SOM maps were trained (using the default values for the neighbourhood radius decrease), using patient data with both BL and FU measures as training data and the rest as testing data. Both the QE and the TE of the testing data and the used training data was stored for each of the iterations. The lower bound of the range was selected as 50, because this resulted in a slightly smaller than recommended number of training steps, at least 500 times the number of nodes in the SOM, as suggested by Kohonen (see section 3.1.1). The number of maps to be trained, step size for `rlen` and upper bound were selected as such, as the time required for the computation has a complexity of $O(n \times \sum \texttt{rlen})$. Thus, shortening the step size or adding large values to `rlen` would have a major effect on the computation time, with only minor improvements to interpretation. Using these values took approximately 4 days to compute using an average computer with Intel Core2 Quad CPU Q9500 processor running at 2.38GHz, 4.00 GB of RAM and the Windows 7 operating system. The most suitable number of training iterations was selected based on the QE and TE values. As the `kohonen` package has not implemented TE for the multi-layer SOM, it was computed using custom scripts (both fraction and distance based measures). This approach to select the number of training iterations was motivated by the assumption that increasing the number of training iterations should decrease the errors in the training data, and for the testing data the values should first decrease, until an increase (or no further decrease) is observed, indicating that further training is not required.

As the TE was observed to increase if the number of steps during fine tuning phase (when only the winning node is moved) was increased (see section 5.1 for details), also the decrease in the neighbourhood function radius was optimized as follows. After the optimal number of training iterations was selected, different proportions of steps used for the fine tuning phase were selected (95%, ... 10%, 5% and 0% of the overall iterations steps) and another 100 SOM maps corresponding to each of these were trained using the same training and test data. Similarly as optimizing the number of training steps, the TE and QE of the resulting mappings were recorded and averaged over the runs. The final proportion of steps used in the fine tuning phase was selected based on these values.

4.3 Prototype clustering

In this work, the prototypes of the trained SOM mapping were clustered using three different methods: agglomerative hierarchical clustering, spectral clustering, and clustering by affinity propagation.

For agglomerative hierarchical clustering, first all the different methods creating dendrograms without possible reversals offered by the R package `stats` [44] as a function `hclust` were tested for their performance. These included five different cost functions: single-linkage, complete-linkage, average-linkage, McQuitty's, and Ward's criteria. In order to find the most suitable criteria among these, 100 SOM maps were trained using the training data, and the resulting prototypes were clustered using each of the aforementioned criteria. The resulting values of AC were stored for every iteration, and the merging criteria with the largest AC on average was selected to be used in agglomerative hierarchical clustering. The AC was computed using the function `coefHier` from the R package `clusters` [68], and since R version 3.0.2 was used, the height of the dendrogram by Ward's clustering was square root transformed before the AC was computed. The input distance matrix for every clustering method was defined as the distance between prototype nodes averaged throughout the different SOM layers (except for Ward's criteria, where the implementation requires this distance matrix to be squared in R version 3.0.2 or earlier). For additional information on the special treatment of the Ward's clustering, see section 3.2.1.

For the second method, spectral clustering, the implementation in the R package `kernlab` [48] as a function `specc` was used. The similarity matrix was defined by first computing the average distance throughout the layers as in agglomerative hierarchical clustering, and then negating the distance matrix creating a similarity matrix. As the used implementation of spectral clustering requires an input with non-negative values, the matrix was further scaled by subtracting the minimum (negative) value of the matrix from all the elements, thus creating a matrix with non-negative values only. This similarity matrix was used directly as an input for the method and the number of clusters were defined using the `clusters` parameter. Otherwise default parameters were used.

Finally, for the affinity propagation, the implementation in R package `apcluster` [49, 50] and function `apclusterK` were used, in order to allow the number of desired clusters to be predefined. The negated average distance matrix throughout the layers was used as an input, as affinity propagation can also handle matrices with negative values. The number of clusters were defined by the parameter `K`, and otherwise the default parameters of the implementation were used.

4.3.1 Evaluation of the clustering performance

The performance of the different methods used to cluster the prototypes of the trained SOM were evaluated by both their robustness to classify patients into the same clusters, and by their ability to distinguish patient classes that show genetic differences between the resulting classes.

As the resulting SOM prototypes depend on the random order of the presented training vectors, the resulting mapping is random by definition. However, if the parameters of the SOM are correctly selected and the data cloud indeed shows some structure, the resulting SOM mapping should be similar throughout different iterations of the algorithm and capture the global structure relatively well. Still, the small differences in the resulting prototypes can have notable effects on the resulting clusters and therefore on the groups where the mapped patients finally end up. This is not desirable and thus the clustering methods were evaluated by their robustness. In this work also the initialization of the SOM was random, and thus the differences in the robustness of the clustering methods can be detected more easily compared with a fixed (for example PCA projection based) initialization of the SOM.

In order to test robustness, 1,000 SOM maps were trained using the available training data. For each iteration, the prototypes of the trained SOM were clustered into two classes using the three clustering methods presented in section 3.2. After this, all the patients (both the training and test data) were mapped to the trained SOM, and the resulting class assignments of each patient were tracked throughout the iterations. As the clustering methods are unsupervised and the resulting clusters do not have labels, for each iteration the labels of the resulting group assignments were inverted if this resulted into a larger proportion of matching assignments compared with the classes of the first iteration. This way, it was possible to determine for every patient how often that patient was classified into the same group during the iterations. The robustness of the methods were visualized by plotting the cumulative curves of the proportion of patients (including training and test data) robustly classified as a function of the proportion of assignments to the same class. Robustness testing was done using only two clusters as altering and keeping track the corresponding cluster assignments in case of more than two clusters is not straightforward.

As one of the main goals of this work is to find the genetic differences of the created novel phenotypes, it is important that the selected clustering method can indeed distinguish patients with a possibly different genetic background. This was evaluated by computing the heritability estimates of the resulting two-class assignments using the GCTA software. For this analysis, the genetic relationship matrix (GRM) was computed for the given set of patients using autosomal genotyped SNPs with a MAF greater than 1%, after which it was pruned to contain unrelated individuals only (using flag `--grm-cutoff 0.025`). This pruned GRM was then used in the restricted maximum likelihood (REML) analysis to estimate the narrow-sense heritability of the class assignment. These evaluations of the heritability were performed for the two-class phenotype assignments defined by two different approaches described below.

In the first approach, the patients ending up in the same cluster more than 80%, 90%, 95%, and 99% of the time during the robustness iterations described above were used as cases and controls. The heritability of the case-control groups defined this way was computed for each of the three clustering methods. This approach was motivated by the assumption that the most extreme patients would have the largest genetic differences and further be most robustly distinguished by the SOM. Thus excluding the intermediate patients that could not be robustly classified by the two classes should make the possible genetic differences more clear, therefore helping to detect the clustering method which creates patient classes with the largest genetic differences.

To complement the first approach, also the best fit of the SOM was selected throughout the iterations. This was done by computing the QE and TE of each mapping during the iterations and selecting all the SOM mappings having both QE and TE below the average over the iterations (thus removing bad mapping outliers) for the second phase. In the second phase, the QE and TE values for the remaining SOM mappings were scaled to the interval $[0, 1]$ and the mapping with smallest distance from origin in this distance plane was selected as the optimal mapping. Next, the prototypes of this optimal mapping were clustered into two classes using all three clustering methods. Finally, these case-control phenotypes of the best fit of the SOM (thus now including all the patients) were evaluated for their heritability using GCTA with corresponding parameters.

Finally the most suitable method to create the final novel phenotype definitions was selected according to its robustness, the ability to distinguish genetically differing patient groups, and overall performance.

4.4 Defining the novel phenotypes and their differences

After the most suitable clustering method and the optimal SOM mapping were selected during the iterations described above, they were used to create the novel phenotype classifications. The prototypes of the optimal SOM mapping were clustered by the selected method by specifying a different number of clusters ($k = \{2, 3, 4, ...\}$). For each k, all the patients included in the study (both training and test data) were mapped back to the trained SOM map, and the phenotype of each patient was assigned according to the BMU cluster assignment. After this, the heritabilities between each pair of defined patient groups were evaluated, once again using the GCTA software and the approach and parameters described above, in order to detect clusters that show genetically divergent patient populations. The number of clusters was increased until there was no further motive to increase k. All the pairs of clusters showing significant heritability (using a Bonferroni corrected significance threshold corresponding to the number of tests for the current cluster number, i.e. $p < 0.05/[\frac{1}{2}k(k-1)]$) were further analysed as a case-control phenotype in the GWAS setting described in section 4.5.

The performance of the resulting clustering was also evaluated by visual inspection of the SOM-layers and statistical testing of input and response variable distribution differences in the created classes.

The visual inspection of variable components was achieved by mapping all the patients to their BMUs and computing for each node the mean (or prevalence for binary traits) of each untransformed input and outcome feature in both the BL and FU layers. After this, both layers were coloured according to the trait magnitude from a 1% winsorized distribution across all nodes (computed using function `Winsorize` from R package `DescTools` [69]) in sense of a heat map. Inspecting the heat maps helped to detect for example variables whose distribution changes between BL and FU layers, variables that seemed to follow the created cluster borders, variables whose distribution seemed parallel to created cluster borders, and variables that did not seem to have a significant structure in the component planes at all.

These differences were also tested by fitting generalized linear models (using R function `glm` from the package `stats` [44]) for each variable (V_j) separately, where both the created phenotype class (C_i, set of dummy variables corresponding to the categorical variable) and the SOM layer indicator (L, binary variable indicating FU layer) were included in the model with their interaction terms for each input variable, i.e.

$$V_j \sim \beta_0 + \sum_{i=2}^{k} \beta_{1i} C_i + \beta_2 L + \sum_{i=2}^{k} \beta_{3i} C_i L, \tag{31}$$

where β_{ij} are the corresponding fit coefficients of each component. Thus the group 1, BL layer, and the interaction in change from BL to FU layer in group 1 were the references. For binary variables, the modelling was done using the logit link function and for other variables using the gaussian identity link function. The data for continuous variables were 1% winsorized to constrain the effect of outliers. After the modelling, the coefficients for the groups were tested for hypothesis,

$$\beta_{12} = \beta_{13} = ... = \beta_{1k} = 0, \tag{32}$$

using Wald's test (from R package `aod` [70] using fuction `wald.test`) to evaluate whether there are group(s) that differ(s) among the created classes, similarly to ANOVA (but now applicable also for binary variables). Correspondingly, the interaction terms were combined and their effect was tested using Wald's test for hypothesis

$$\beta_{32} = \beta_{33} = ... = \beta_{3k} = 0, \tag{33}$$

to test whether the change in variable distribution from the BL to the FU layer behaves differently for any group compared with the rest. For $k = 2$ the only resulting coefficients for group and interaction were interpreted directly for their significance. The significance threshold was Bonferroni corrected for the total number of variables modelled, i.e. set to $p = 0.05/(m|k|)$, where m is the number of input features in the SOM, and $|k|$ is the number of different clusterings created.

4.5 Genetic analyses

Finally after the pairs of clusters showing a significant heritability were identified, the unrelated patients used in the heritability computations were used in a case-control GWAS analysis with the defined cluster assignment as a binary phenotype. For these analyses, PLINK v.1.07 was used with the imputed genotype data (described in more details in section 4.1.2) with logistic regression models (described in more details in section 3.3.4) for continuous genotype dosages. The models were adjusted for sex, the genotyping batch of the patient, and the first ten genetic principal components (PCs) of the genetic data, computed using the Eigenstrat software (EIGENSOFT v.3.0) [71]. Only SNPs with MAF above 1%, and imputation quality (PLINK INFO criteria) above 0.8 were analysed.

5 Results

The following section will present the results of the analyses in this work. First, section 5.1 presents the variables used to train the multi-layer SOM, and how the SOM parameters were selected. Next, section 5.2 visualizes the distribution of multiple input and output variables in the optimal SOM mapping, and evaluates the ability of the multi-layer SOM to capture the progressive nature of the complications. Section 5.3 presents the results of performance for different clustering methods, both in their robustness and ability to distinguish groups with a genetically divergent background, and how the most suitable method was selected. Then, section 5.4 illustrates the resulting cluster assignments in the optimal SOM mapping and the evaluated intergroup heritabilities between the defined classes. After this, section 5.5 illustrates the profile differences of the patients mapped to the created classes. Finally, section 5.6 summarize the results of GWAS analyses between the novel phenotype classes.

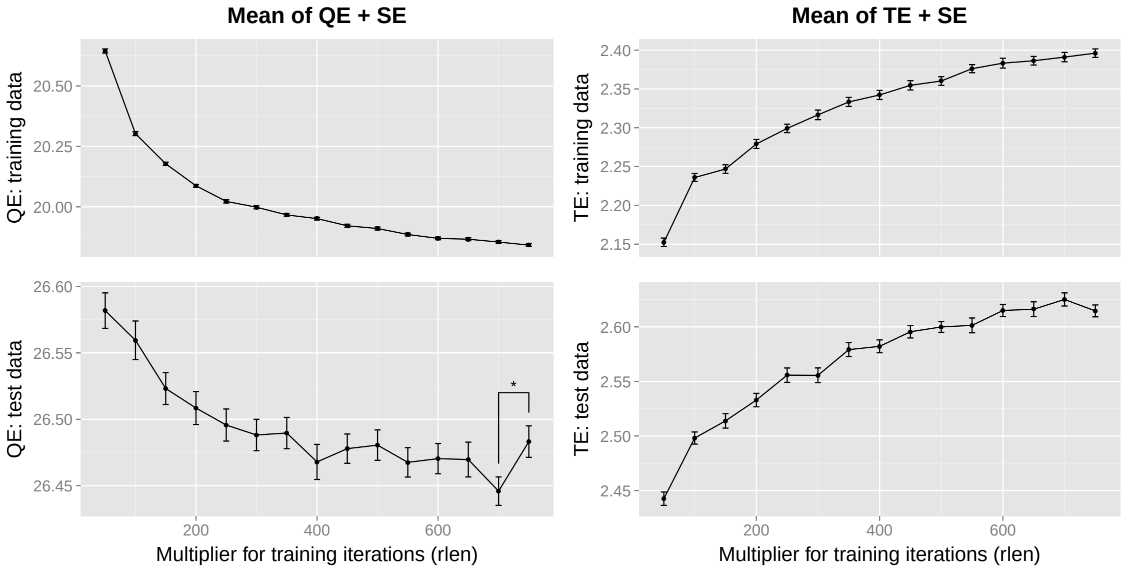
5.1 Training variables and optimized SOM parameters

After all the raw data from different sources were combined, the iterative process to prune the variables and patient visits with too much missing data was performed. After this, a total of 97 input variables remained for a total of 6,013 patient visits (BL: 4,372, FU: 1,630) for a total of 4,409 unique patients. Of these 1,600 patients had both BL and FU data present, 2,779 patients had only the BL data and a further 30 only the FU data (BL visit excluded due to missingness constraints).

After this, the subset of patient visits having complete data for all of the variables ($n = 728$) was used to prune the linearly dependent variables, and the variables derived using other variables included in the set. After this pruning, a total of 18 variables were removed from the data, and thus $n = 79$ variables remained, and were later used as input variables for the SOM. In the combined set of BL and FU data, the variable with the most missing data contained 51% non-missing values, and the overall proportion of non-missing entries in the data matrix was 92%. All of the selected variables and their missingness values are presented in Appendix A.

After the data was pruned successfully, it was further transformed and standardized as described in section 4.2.1. Next, the size of the used SOM grid was defined from the standardized input data matrices as described in section 4.2.2, and set as 10×18 nodes (height × width). After this the number of training iterations was optimized as described before. The QE and TE across the optimization iterations are illustrated in Figures 5a and 5b.

As Figure 5 illustrates, increasing the number of training iterations decreases the QE of both training and test data. Furthermore, when setting `rlen` = 700 the SOM seems to yield the smallest QE on average, and this error increases to significantly worse when `rlen` is increased to 750 ($p = 0.021$, Welch Two Sample t-test). However, it seems that increasing the number of training iterations will increase the TE of the SOM (both in training and testing data) from the very beginning. The phenomenon of increasing TE is very likely due to the fact that the default



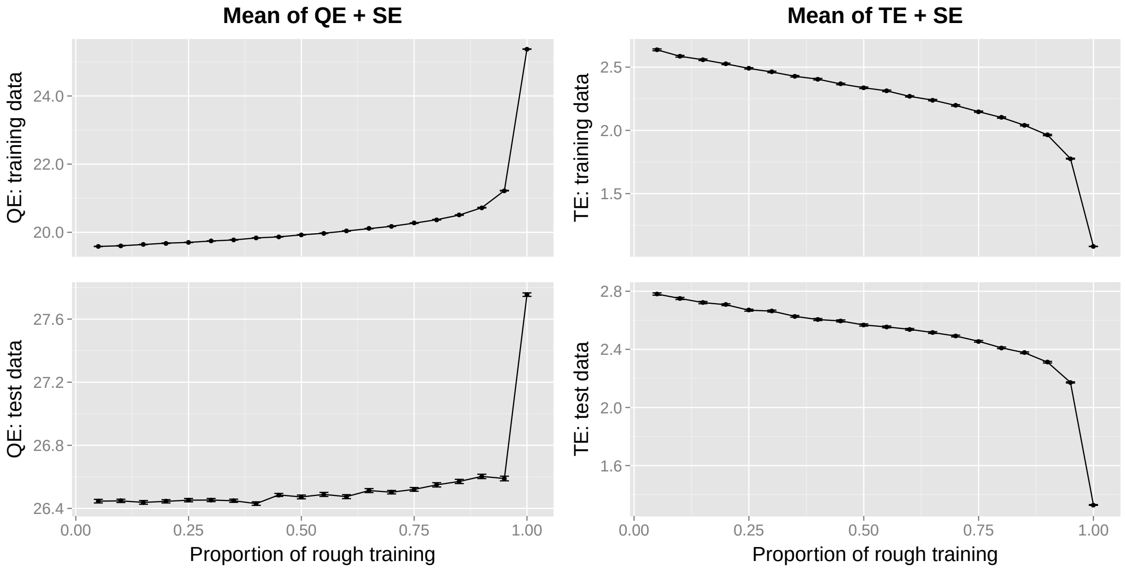
(a) Mean QE across iterations

(b) Mean TE across iterations

Figure 5: Average QE (a) and distance based TE (b) as a function of the number of training iterations during the the `rlen` parameter optimization. * $p = 0.021$

decrease in the neighbourhood function in the `kohonen` package is implemented so that the number of fine tuning iteration steps (when only one of the nodes is moved) is always approximately half of the training iterations, and thus the absolute number of such iterations increase when the number of training steps is increased. Moving a single node just slightly towards the presented sample vector in the fine tuning will improve the QE of the map (slightly). However, as the input data is high dimensional, the distances in the vector space $\mathbb{R}^d$ are getting more and more similar due to the "curse of dimensionality", and thus, moving just a single node can alter the grid structure so that the first and second BMUs for sample vectors will map to non-distinct nodes (thus increasing the TE) more easily. As the absolute number of these easily topography breaking iterations increase, the resulting TE for the map will increase. Thus the optimal value for `rlen` was set to 700 (based on the evaluated QE). The decrease rate in the neighbourhood function was later altered, as this way it was possible to improve the quality of the resulting mapping also in terms of TE. The optimization was performed as described in section 4.2.2, and the results of the QE and TE from iterations having different proportion of fine tuning out of the total number of iteration steps are illustrated in the Figures 6a and 6b.

As can be seen, decreasing the amount of fine tuning at the end of SOM training have only a minor effect on the QE (expect when totally omitted), but the decrease is clear in terms of TE from the very beginning. Thus the final decrease in the neighbourhood function was selected based on the "elbow criterion" so that it will start as containing 2/3 of the nodes (the default starting value of the `kohonen` implementation), and decrease linearly so that only the last 5% of the iteration steps will be performed as fine tuning (moving a single node at a time), as this way the trained SOM would also retain the local topology of the input data better.



(a) Mean QE across iterations

(b) Mean TE across iterations

Figure 6: Average QE (a) and distance based TE (b) as a function of the number of training iterations during the neighbourhood function radius decrease rate optimization

To summarize, the final input data and free parameters of the SOM are as follows:

- Input data features: 79 input variables, one of which was a categorical variable with 6 levels (converted to 5 dummy binary variables), and another categorical with 4 levels (converted to 3 dummy binary variables), i.e. 85 features
- Training samples: 1,600 (all patients with both BL and FU data present)
- Test samples: 2,809 (patients with data on single visit)
- Grid topology: dual-layer toroidal hexagonal (expert opinion)
- Grid size: 10 × 18 nodes (height × width) (optimized)
- Prototype initialization: random selection among training vectors without replacement (default)
- Neighbourhood function: radial boundary as described in Equation (4) (default)
- Learning rate: linearly decreasing from 0.05 to 0.01 (default)
- Number of training iterations: 700 × number of training samples, i.e. 1,120,000 rounds (optimized)
- Neighbourhood radius decrease: linear so that initially 2/3 of the nodes belong to neighbourhood, and only the last 5% of the training is performed as fine tuning (optimized)

5.2 Optimal SOM fit and visual evaluation of the approach

After the parameters of the SOM were optimized, multiple (1,000) SOMs were trained with random initializations in order to find the optimal SOM fit. After the optimal mapping was selected, the patient data was visualized on the grid plane to allow for visual evaluation of the mapping performance.

In this work, a multi-layer SOM was selected as a underlying tool to create the novel patient classes, as we hypothesized that using measures from different time points in different layers of the SOM could help to capture the progressive nature of diabetic complications. Based on the visual inspection of input data, the approach seems to reach this target (see Figure 7). When the variables in the BL layer of the SOM are illustrated in the node grid (as averages of the mapped patients), it seems that many of the traits are centered into relatively well restricted regions, or the border between high and low magnitude nodes is generally visible. When these variables are illustrated in the FU layer correspondingly, it seems that most of the variables that have progressive nature (i.e. can "only get worse" almost in all cases, for example use of medication and diabetic complications themselves) show phenomenon where the "high risk region" spreads on the grid, but mainly only to nearby nodes. Thus these "low magnitude nodes" close to "high magnitude nodes" on BL layer have a special role: the magnitude of the variables are more likely to increase compared with low magnitude nodes further away, when the BL and FU layers are compared (see Figure 7 comparing middle and right plots). On the contrary, the variables that can fluctuate to either directions between the visits (for example BMI, blood pressure markers and lipid profiles) do not show the same phenomenon, i.e. a node with intermediate magnitudes near high magnitude nodes does not necessary imply an increase from the BL to FU layer, as illustrated in Figure 8 (compare middle and right plots). Thus the SOM seems to be able to map the patients progressing between the BL and FU visits in terms of a progressive trait to the same map region, located approximately between the high and low prevalence regions of the BL layer, which therefore represents patients with a high risk of progression.

Furthermore, it seems that the baseline layer surface is very similar when the training and testing patients are compared (as illustrated in Figures 7 and 8 between the left and the middle plots), despite the highly dimensional input data, and possible small overlearning of the SOM. Thus the patients having their cross-sectional single time point profiles similar to the training patients, whose complications progressed between the BL and FU visits, are mapped to the same regions. Therefore these patients are expected to show complication progression in the near future. This also supports the assumption that mapping patients to the multi-layer SOM based on a single layer is meaningful and well motivated.

When the latest outcome variables are visualized (see Figure 9), it seems that the map also has a predictive value years beyond the FU layer. All three outcome variables, mortality, macrovascular events, and DN progression, showed notable structures in the SOM grid, which reflects also the effects of known risk factors for the complications and mortality. The regions showing the largest mortality are mapped to the same regions that had patients with the most severe complications (especially

DN and CVD), widest medication usage, and that were the oldest. Correspondingly, a high risk for incident macrovascular events was found around regions that contained patients that had macrovascular events already prior to the BL visit. Finally, the risk of progression of DN was found in the region with the patients having already more advanced DN and/or worse glycaemic control (higher blood HbA_{1c} values).

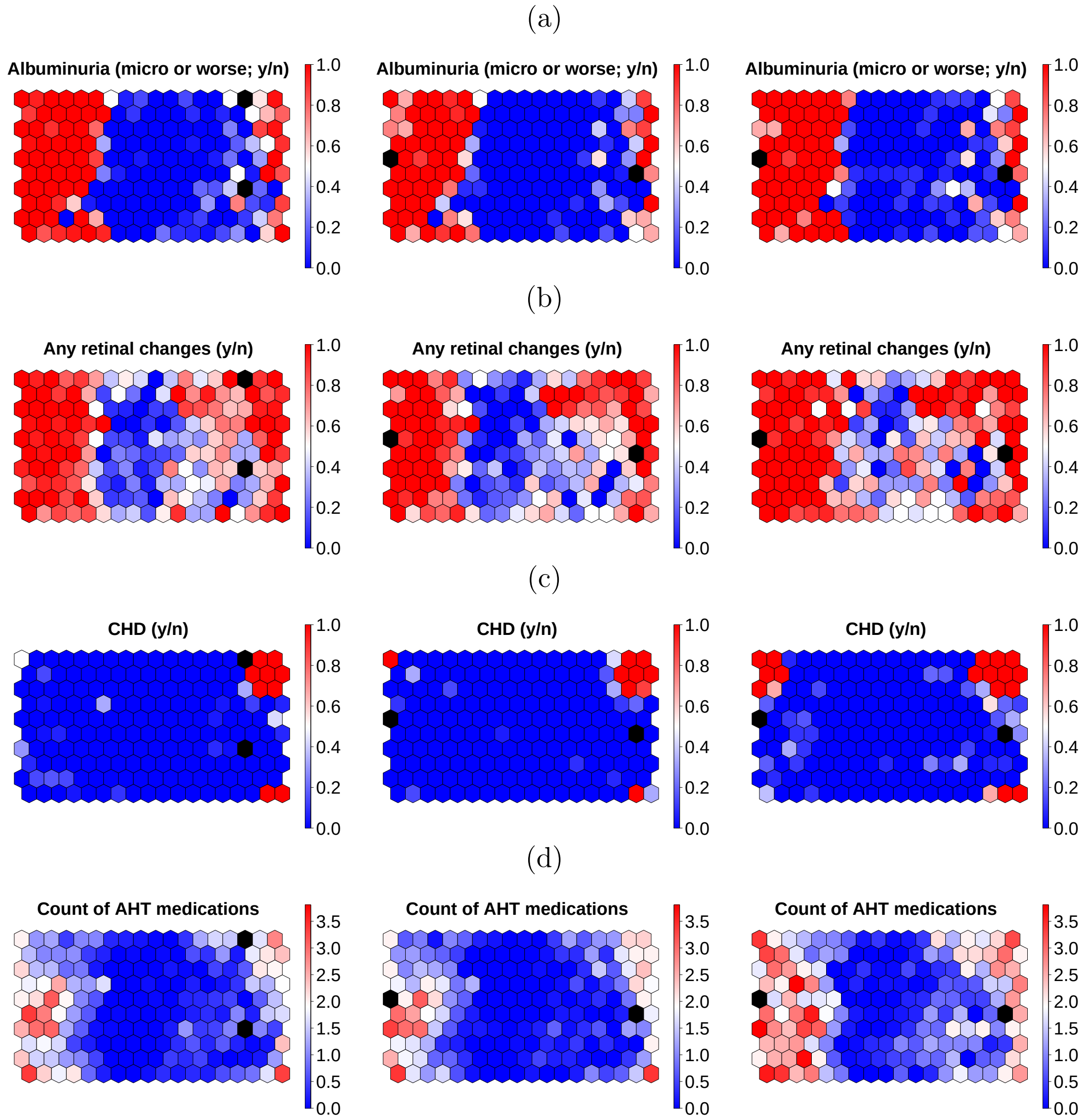
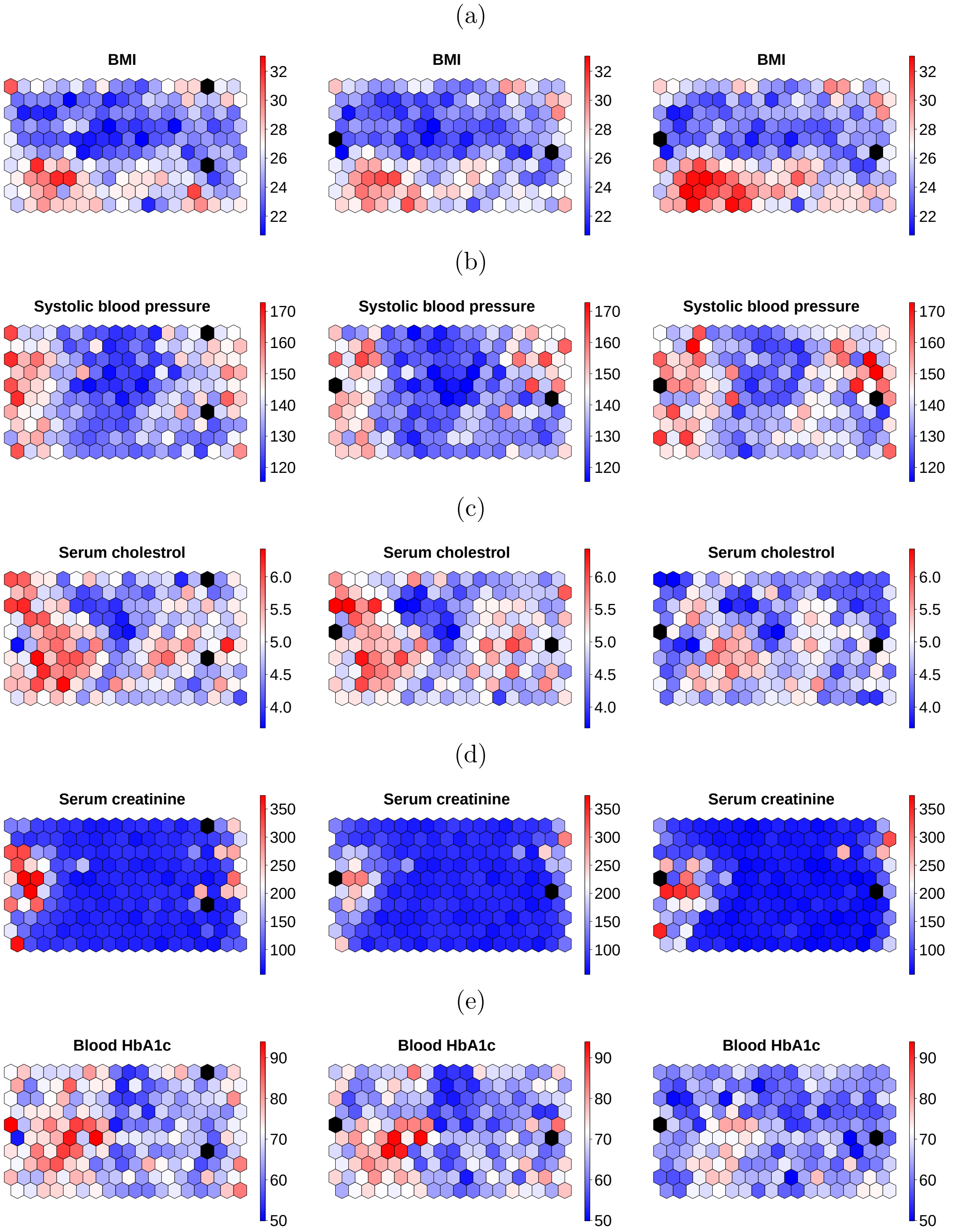


Figure 7: Distribution of selected variables with a progressive nature in the SOM layers as averages of patients mapped to each node. For each trait, plots from left to right represent: BL layer of test patients, BL layer of training patients, and FU layer of training patients. Nodes having missing data for all of the mapped patients (or no patients were mapped to them) are coloured black. The colouring scheme across datasets and layers is harmonized to help interpretation. All illustrated variables were used as input data for the SOM, except for the number of AHT medication (7d), which was excluded due to linear dependencies. Other medication related binary variables used as input variables behave similarly.


(a)

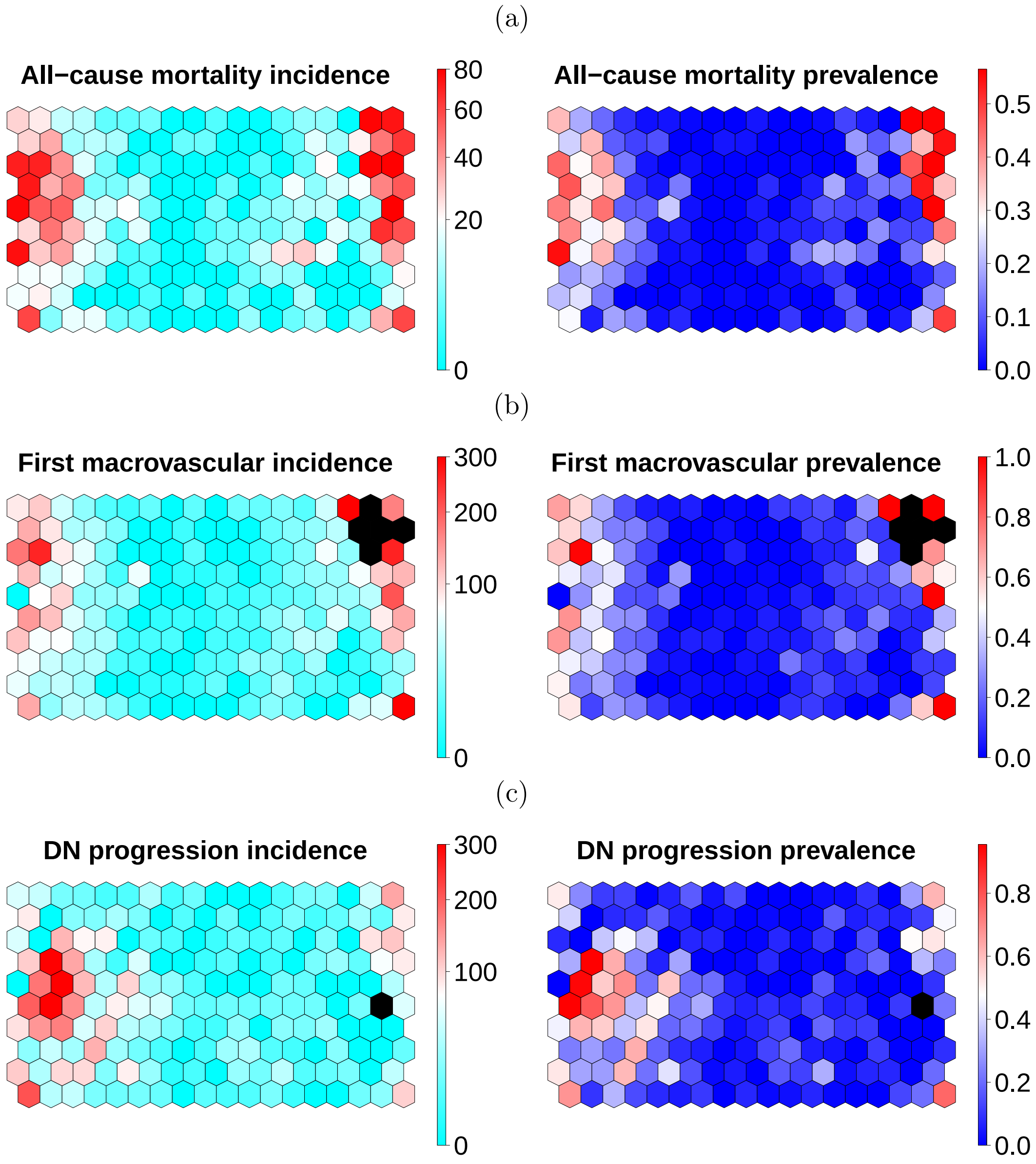
(b)

(c)

(d)

(e)

Figure 8: Distribution of selected variables in the SOM layers as averages of patients mapped to each node. For each trait, plots from left to right represent: BL layer of test patients, BL layer of training patients, and FU layer of training patients. Nodes having missing data for all of the mapped patients (or no patients are mapped to these nodes) are coloured black. The colouring scheme across datasets and layers is harmonized to help interpretation. All illustrated variables were used as input data for the SOM, except for the BMI (8a), which was excluded due to known non-linear dependencies.


(a)

(b)

(c)

Figure 9: The incidence (estimated annual events per 1000 patients; left) and follow-up time weighted prevalence (right) of traits since the BL visit estimated from latest available data. The colour scheme for incidence was square root transformed to highlight differences on nodes with lower incidence levels, and thus different colouring was used. New macrovascular events (9b) include all hospitalizations from cardoiovascular and cerebrovascular events. Only patients without macrovascular events prior to BL visit were considered, and thus the top right part of the plot, containing only patients with prior macrovascular events, is left black. For DN progression, patients with BL ESRD were excluded, as it is considered the most severe state of DN, and thus cannot progress to worse.

5.3 Selection of the clustering method

When selecting the optimal method for prototype clustering, in the first step the most suitable cost function for agglomerative hierarchical clustering was evaluated by training 100 SOM maps, and clustering the resulting prototypes using agglomerative hierarchical clustering with different cost functions as described in section 4.3. The performance of clustering was evaluated using AC, and the results are presented in Figure 10.

As Figure 10 clearly illustrates, the Ward's criteria yield the largest AC on average and the good performance is very robust compared with other cluster merging criteria. Thus it was selected as the most suitable cost function for agglomerative hierarchical clustering, and it was next compared with other more sophisticated clustering methods.

In order to do this, the 1,000 SOMs trained to find the optimal fit were clustered using the three clustering methods of interest (agglomerative hierarchical clustering using Ward's criteria, spectral clustering, and clustering by affinity propagation). The resulting cluster assignments for each patient were tracked throughout the iteration steps, and the proportions of robustly classified patients (i.e. patients classified into the same cluster more often than the current threshold) as a function of the robustness threshold are presented in Figure 11.

As Figure 11 illustrates, among the three tested clustering methods, spectral clustering can classify the patients most robustly in case of a two class problem, whereas Ward's clustering and clustering by affinity propagation seem to perform

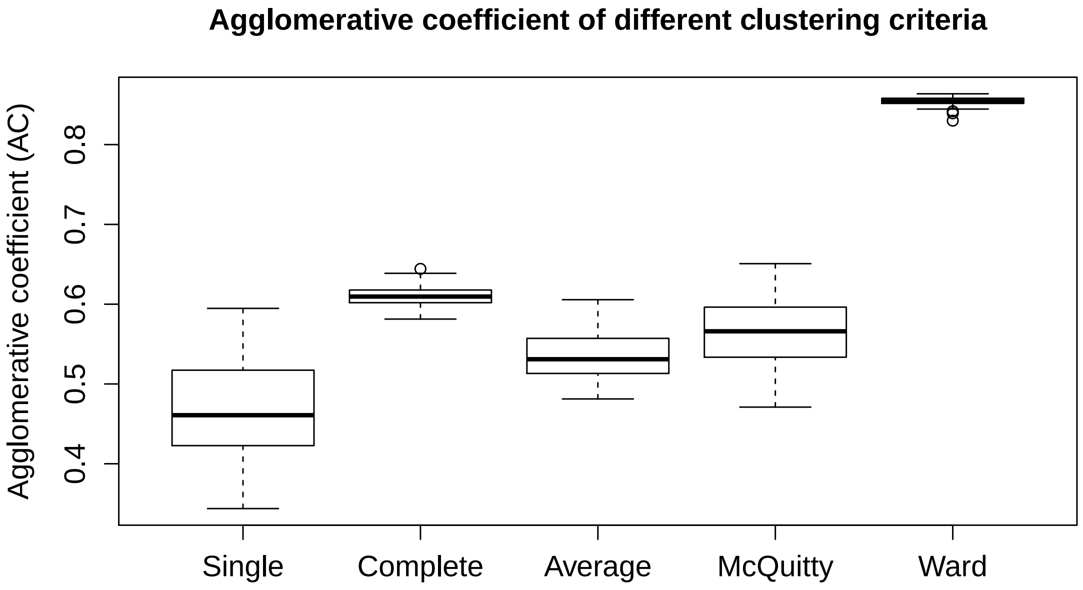


Figure 10: Box-plot of AC for different cluster merging criteria in agglomerative hierarchical clustering.

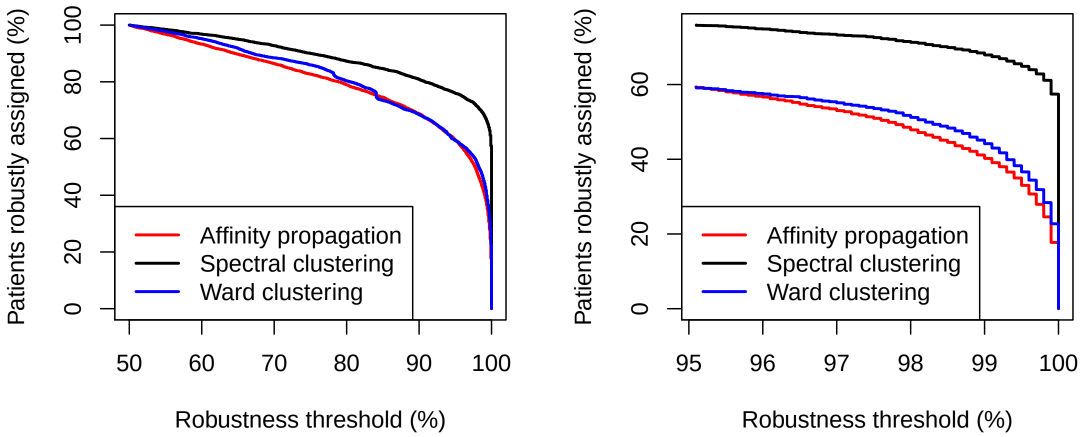


Figure 11: Proportion of patients robustly assigned to the same class throughout iterations (left) and closer zoom to higher robustness requirements (>95%, right)

relatively similar. Using spectral clustering, more than 75% of the patients are assigned to the same cluster at least 95% of the time. The corresponding proportion for both affinity propagation and Ward's clustering is 59%. Correspondingly, if the robustness threshold is increased all the way to 100%, spectral clustering can still classify approximately 57% of the patients into the same cluster regardless of the random initialization and by definition random nature of the underlying SOM. The corresponding proportion for Ward's clustering is 23%, and only 18% for affinity propagation.

Next the heritabilities between the robustly classified patient classes were estimated for the three clustering methods. The estimates were computed for four different robustness requirements (80%, 90%, 95% and 99% of the cluster assignments being same). Additionally, the heritabilities were estimated from the two-class clustering assignments of the optimal SOM mapping. The results of these analyses are presented in Table 2.

Based on these results, it seems that all the clustering methods can separate robustly classified patient classes that show significant ($p < 0.05$) heritability. However, Ward's clustering failed to repeat this outcome, when only the most robustly classified patients (99% of the class assignments same) were used, whereas the heritability of classes resulting from either spectral clustering or affinity propagation still showed significant heritabilities ($p < 0.05$, no significance threshold correction). On the contrary, Ward's clustering was the only one to create classes with significant heritability when applied to the optimal fit of the SOM. At the same time, affinity propagation was the only clustering method that showed increasing point estimates for heritability when the robustness threshold was increased (and thus more specified patient groups with genetically divergent background were identified), which is a

Table 2: Heritability estimates and their statistical significance in patients robustly classified throughout the SOM iterations, and in the optimal SOM mapping in two-class clustering problem. h^2: Estimated narrow-sense heritability; SE: Standard error of the heritability estimate; p: p-value for the hypothesis $h^2 = 0$; N: Number of unrelated individuals used to compute the estimates, pruned with GCTA using `--grm-cutoff 0.025` flag.

		Ward's clustering			**Spectral clustering**			**Affinity propagation**		
		$h^2[SE]$	N	p	$h^2[SE]$	N	p	$h^2[SE]$	N	p
Robustness	80%	0.16 [0.09]	2,845	0.033	0.15 [0.09]	3,063	0.031	0.19 [0.09]	2,820	0.013
	90%	0.20 [0.11]	2,498	0.024	0.13 [0.09]	2,879	0.063	0.19 [0.10]	2,521	0.017
	95%	0.19 [0.12]	2,213	0.038	0.17 [0.10]	2,754	0.029	0.25 [0.16]	2,222	0.008
	99%	0.15 [0.14]	1,740	0.136	0.16 [0.10]	2,520	0.046	0.34 [0.17]	1,591	0.020
Optimal fit		0.20 [0.08]	3,384	0.009	0.10 [0.08]	3,384	0.081	0.09 [0.08]	3,384	0.118

desired outcome when even tighter subgroups of patients are formed by increasing the number of clusters in the next step. Combined with the results of robustness analysis, where spectral clustering performed best among the methods, the decision of the most suitable clustering method is not straightforward.

Thus the selection was also based on the default properties of the approaches. As the affinity propagation implementation of clustering with a predefined number of clusters is based on iteratively adjusting the initial diagonal elements of the affinity matrix until the algorithm converges to a desired number of clusters, it can fail in the task in certain cases. Correspondingly, in the used implementation of spectral clustering, the final step (i.e. creating the clusters in the spectral subspace) is based on the k-means algorithm, which causes a small degree of additional randomness to the resulting clusters. The pilot analyses suggest that with small k the created clusters are robust, but increasing k above five already causes a random fluctuation for some nodes close to the cluster borders (data not shown). Thus agglomerative clustering using Ward's criteria was selected to be used to create the final phenotype classes, as the approach does not have additional degree of randomness, and it will always succeed in the clustering task. In addition, due to its hierarchical nature, it is possible to visualize the created clusters also by using a dendrogram, which adds additional insight to the number of clusters in the data. Finally, Ward's clustering was the only method to show significant heritability in the optimal SOM mapping, and the robustness requirements of the two class clustering are not vital for the main goal of this work, as the phenotypes are created using the optimal SOM mapping.

5.4 Cluster borders and novel phenotype classifications

When the patients were divided into $k = \{2, 3, ...\}$ classes using Ward's clustering, at first the magnitudes of significant heritabilities were increased (both point estimate for the heritability and its significance) with increasing k. However, increasing $k > 4$ made the heritability estimates decrease in magnitude and become less significant,

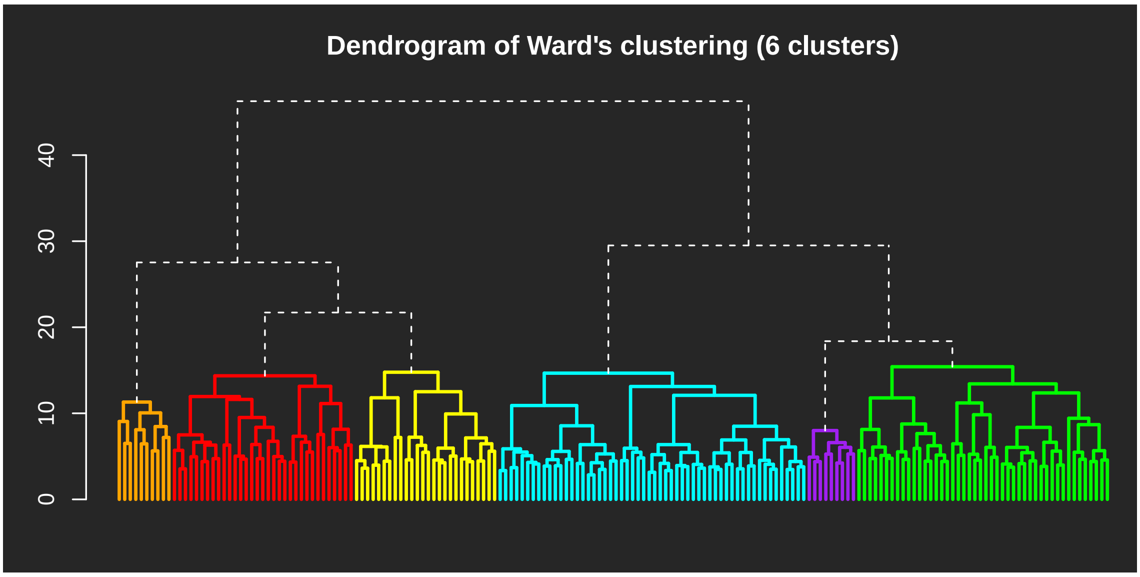


Figure 12: The dendrogram from Ward's clustering, cut to create 6 clusters. Clusters colored according to the mapping presented in Figure 13e

possibly due to the continuously decreasing size of the clusters (causing larger standard errors (SE), and thus, smaller p-values for the estimates). Still the point estimates for heritability were surprisingly large for certain pairs of patient groups, even though the estimates were non-significant after Bonferroni correction. All the heritability estimates and their significance between pairs of groups for each k are presented in Appendix B. For group labels regarding the SOM grid positions, refer to Figure 13. The significant heritabilities between pairs of groups further analysed in the GWAS setting are highlighted in green (and the results of these analyses are presented in section 5.6).

The most significant heritability was observed when the patients were divided into four classes, between groups 1 and 2 ($h^2 = 42\%$, $p = 1.8 \times 10^{-4}$). The resulting estimate is approximately of the same magnitude compared with previously reported heritability estimates of various diabetic complications [8, 72]. When the dendrogram is cut to create six clusters, an interesting small group of patients can be identified, which shows large point estimates for heritability when compared with the other groups created in the same clustering (5 vs. 6: $h^2 = 69\%$, $p = 0.11$; 5 vs. 2: $h^2 = 61\%$, $p = 0.039$; 5 vs. 1: $h^2 = 51\%$, $p = 0.041$). However, these heritabilities are non-significant (required Bonferroni corrected p-value for 15 pairings of 6 clusters $p < 0.05/15 = 0.003$), and are thus considered as suggestive.

The splitting of the clusters was stopped after $k = 6$, as the dendrogram, illustrated in Figure 12, shows that if the cut threshold would be lowered to increase k, next there would be multiple clusters merged at approximately the same level. Thus these new clusters are not as clearly separated and there is no motivation to further increase k. In addition, the heritabilities were mostly non-significant already at $k = 6$, and number of patients in additional sub-clusters would continuously become smaller causing more fluctuation into heritability estimates (and thus more likely

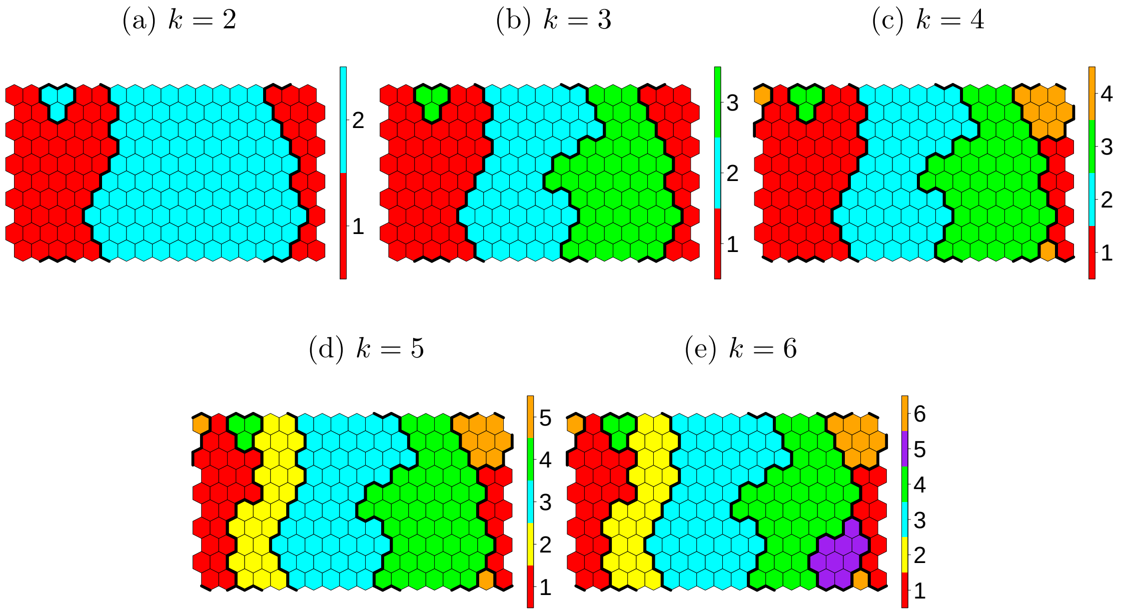

(a) $k = 2$
(b) $k = 3$
(c) $k = 4$
(d) $k = 5$
(e) $k = 6$

Figure 13: Created clusters and cluster boundaries visualized in the grid-plane of the SOM for $k = \{2, 3, 4, 5, 6\}$

non-significant estimates). The corresponding cluster borders and assignments for each $k = \{2, 3, 4, 5, 6\}$ in the SOM grid-plane are presented in Figure 13.

The resulting clusters are connected in the SOMs' grid-plane (the sides are connected as the actual plane was folded to a torus), except for a small disconnected component of three nodes visible at the top left corner of the plots. Thus the clusters are somewhat smooth, and intuitively the node prototypes (and therefore the patients mapped to them) should be relatively similar within each cluster.

5.5 Clinical profiles of the clusters

The following subsections will present the distribution and differences of clinical variables in in contrast to the created patient classes resulting from cutting the dendrogram from different resolution levels. For numerical cluster labels, refer back to Figure 13.

5.5.1 Two clusters: high and low complication risks

When the SOM map was divided into two clusters, the resulting patient division seemed to follow severe microvascular complications, and high medication usage as illustrated in Figures 14a to 14c. Thus these classes are further referred as "high risk cluster" (group 1) and "low risk cluster" (group 2).

The most striking difference was seen in terms of normal AER status, which almost perfectly follows the cluster assignments (in high risk cluster, 1,306/1,444 had abnormal AER at BL and in low risk cluster only 221/2,753 had abnormal AER at

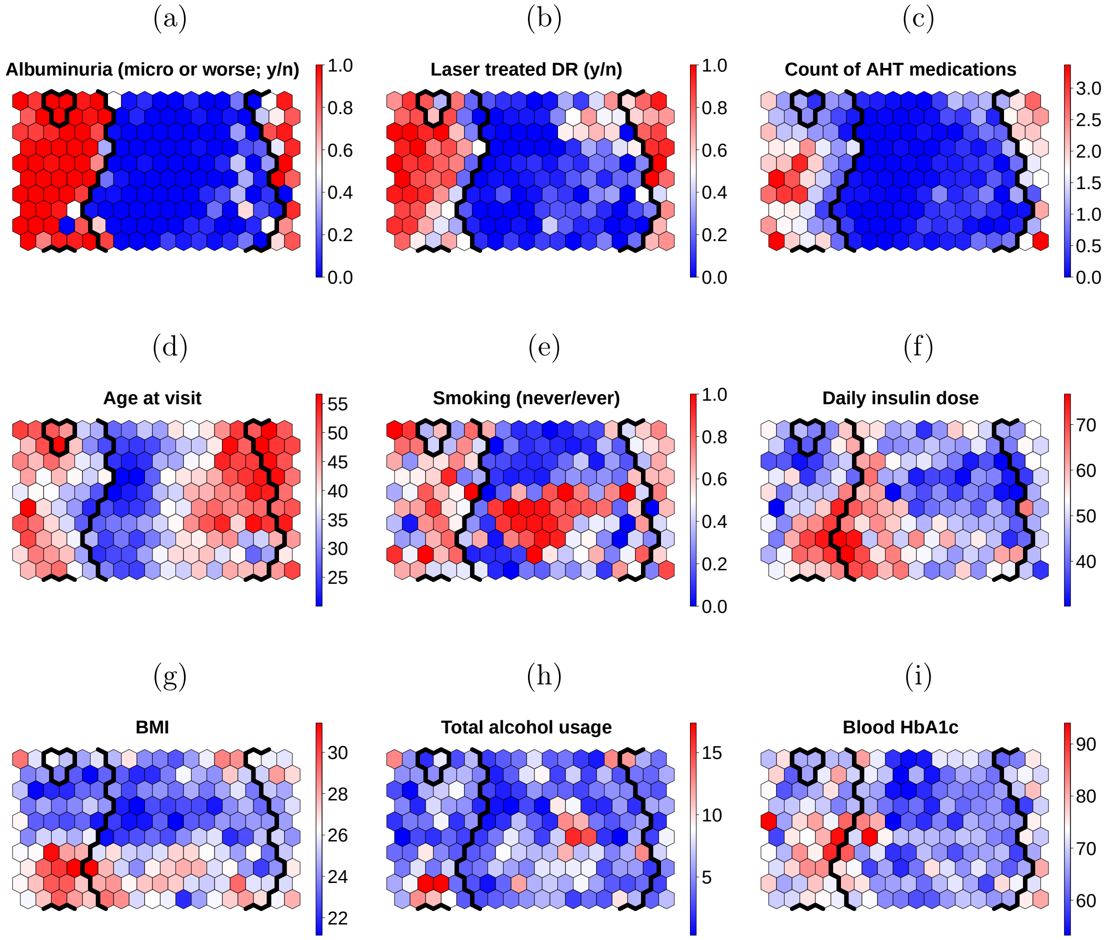


Figure 14: Selection of variables illustrated in the BL layer of the SOM using all mapped patients. Cluster borders according to assignments $k = 2$. Of these variables, the count of different antihypertensive (AHT) medication, BMI and total alcohol intake were not used as input variables (pruned due to dependencies) and are used as summarizing visualizations.

BL). Also the macrovascular complications are clearly limited mainly to regions with albuminuria, but form generally smaller sub-regions (see Figure 16). On the contrary, many clinical variables, variables related to the treatment of diabetes, lifestyle related variables, and laboratory measures did not follow the cluster borders (see Figures 14d to 14i), suggesting that the classification is capturing the interesting complication component instead of trivial differences (for example the age of the patient). However, the distribution differences between clusters for most of the input variables were still statistically significant. Among the 84 input variables, 26 did not show significant differences between the classes (Bonferroni corrected for all 84 input features tested for each k, $p < 0.05/(84 \times 5) = 1.19 \times 10^{-4}$). They included anthropometric measures (height, weight, and hip width), diabetes treatment related variables (use of tablets in diabetes treatment, insulin treatment type, insulin dose, and time from diabetes onset to initiation of insulin therapy), self-reported lesser complications and medication variables (use of thyroxin, hormone replacement and rarer types of AHT medication, self-reported asthma, thyroid disease and rheuma), lifestyle related (weekly beer and wine dose, and social classes 3 to 5), family complication related (incidence of stroke, use of AHT medication or prevalence of diabetes in father, incidence of stroke in mother, count of siblings and prevalence of diabetes in them), serum apolipoprotein A-I (ApoA-I), and volume of the urine sample. It is worth noting that all the micro- and macrovascular complications, and variables related to the medication commonly used to treat them, were highly significantly differentially distributed among the clusters. Thus the cluster separation could be roughly described as "patients with multiple severe complications versus patients without them", and the genetic differences found between these two classes can be assumed to represent mainly the microvascular (and partly macrovascular) complications, especially DN.

The difference in latest available outcome data was also significant between the clusters. Mortality in high risk cluster was higher than in low risk cluster (519/1,495 = 35% vs. 138/2,860 = 5%), incident CVD events were more frequent in high risk cluster than in low risk cluster (468/1,165 = 40% vs. 276/2,819 = 10%), and finally DN progression was also more common in high risk cluster compared with the low risk cluster (384/1,023 = 38% vs. 186/2,391 = 8%).

When the variables were inspected regarding their change between the BL and FU layers, 9 showed a different behaviour between the groups when moved from BL to FU. These included both SBP and DBP, which became more similar between the groups; CHD, which increased more rapidly in high risk cluster; beta blockers whose usage increased more in high risk cluster; ACE inhibitors, whose usage became more similar (even though there still was a clear difference in FU) between the clusters; serum apolipoprotein B (ApoB), total cholestrol and blood HbA_{1c} that were lowered to approximately similar levels in both clusters; and finally serum urate that had a stronger increase in high risk cluster.

5.5.2 Three clusters: low risk cluster split

When the number of clusters was increased from two to three, the previous low risk cluster was split into two. The other half, illustrated in the middle of the plots in Figure 15, seems to contain more strictly defined "control patients" with lower prevalence of milder degrees of complications and use of medication (Figure 15b) compared with the other half. Therefore the middle cluster (cluster 2) will further be referred as "(remaining) low risk cluster", and the other half (cluster 3) as "intermediate risk cluster". Among the complications, the difference in the prevalence of milder levels of DR changes seems to be the most distinguishing between the created sub-clusters based on the visual inspection (Figure 15c). Furthermore, patients remaining in low risk cluster seem to have also earlier diabetes onset. At this cluster division, also surprising variables show differences between the just created sub-clusters as the count of siblings seems to be higher in the intermediate risk cluster.

When the variable distributions in different clusters and between layers were

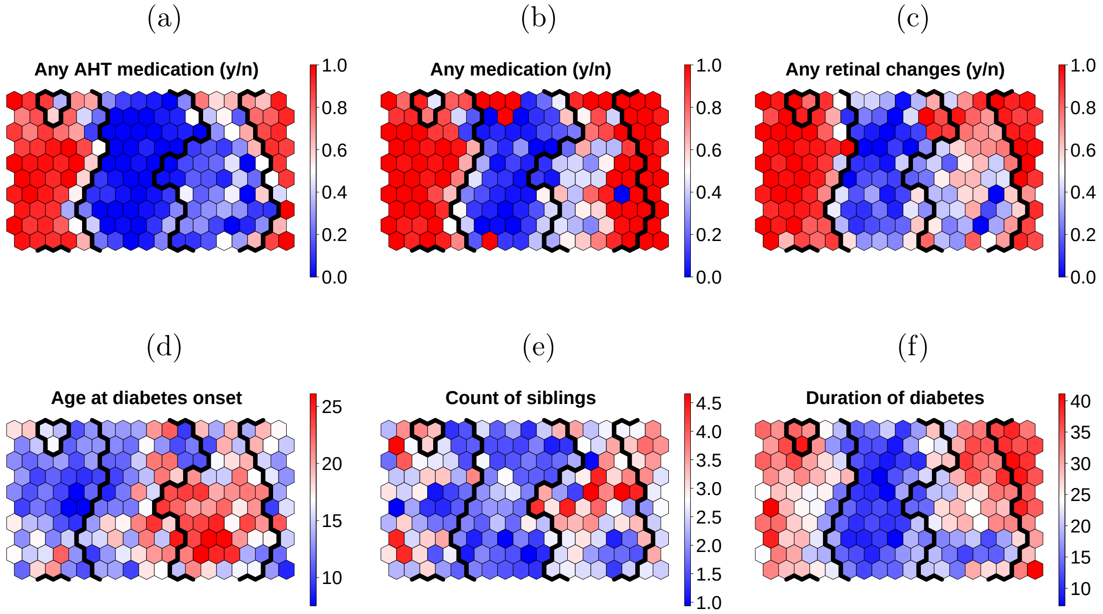


Figure 15: Selection of variables illustrated in the BL layer of the SOM using all mapped patients. Cluster borders according to assignments $k = 3$. Of these variables, any AHT medication and age at diabetes onset were not used as input variables (pruned due to dependencies).

further compared using generalized linear modelling, among the 84 input features only 6 did not show significant differences between the clusters. These variables were a smaller subset of variables not showing differences in previous group assignments ($k = 2$): hip circumference, insulin treatment type, use of rarer AHT medication, use of warfarin, weekly beer dose and social class 3.

When the changes between the layers were further inspected, the same variables showing differences in change between layers in $k = 2$ showed different behaviour also in the case of $k = 3$ with a similar interpretation. However in addition to the 9 previously found variables, now also the use of thyroxin, and the prevalence of self reported thyroid disease show a different behaviour, as the low risk cluster shows higher increase in the prevalence for both (due to very low BL prevalence).

Based on the heritability results, the most significant genetic differences can be found between high risk cluster and remaining low risk cluster. Furthermore, the cluster border between them follows relatively closely the mortality, incident CVD events, and the DN progression rates derived from the latest available registry based data (see Figure 9). This is seen also as significant differences in the cluster-wise prevalence of these traits: all three were more frequent in the high risk cluster compared with the remaining low risk cluster (mortality $519/1{,}495 = 35\%$ vs. $24/1{,}609 = 1\%$; incident CVD event $468/1{,}165 = 40\%$ vs. $57/1{,}602 = 4\%$; and DN progression $384/1{,}023 = 38\%$ vs. $114/1{,}377 = 8\%$).

5.5.3 Four clusters: high risk patients with CHD

When k was further increased to create four clusters, a small subgroup of patients was separated from the high risk cluster of the previous assignments. This group had a clearly higher prevalence of CHD related traits as illustrated in Figures 16a to 16c. Thus this small subgroup of high risk cluster will later be referred as "CHD cluster". However, the separation did not follow all macrovascular complications, as nodes with a high prevalence of stroke and PVD related traits prior to the BL visit were still included in the high risk cluster (see Figures 16d to 16f).

The cluster-wise distributions were significantly different between all variables, except for the same six variables evenly distributed in the case with $k = 3$. The set of variables showing significant cluster-layer interaction was almost the same as in the $k = 3$ assignment. Splitting of the high risk cluster removed the interaction with CHD and use of beta blockers, but introduced an interaction in terms of acute myocardial infarction (AMI), as the new CHD cluster showed a small decrease in prevalence of AMI from BL to FU data. This decrease can be explained by sampling bias, as the region is associated with high all cause mortality, and some of the patients with AMI prior to BL died before FU visit lowering the proportion of patients with AMI in FU layer.

The heritability between the newly created CHD cluster and other clusters could not be estimated, as the genetic variance component escaped the parameter space during the iterative model optimization when the heritability was estimated using GCTA. However, the remaining high risk cluster (cluster 1) showed the most significant heritability among any cluster assignment for any k tested, when it was

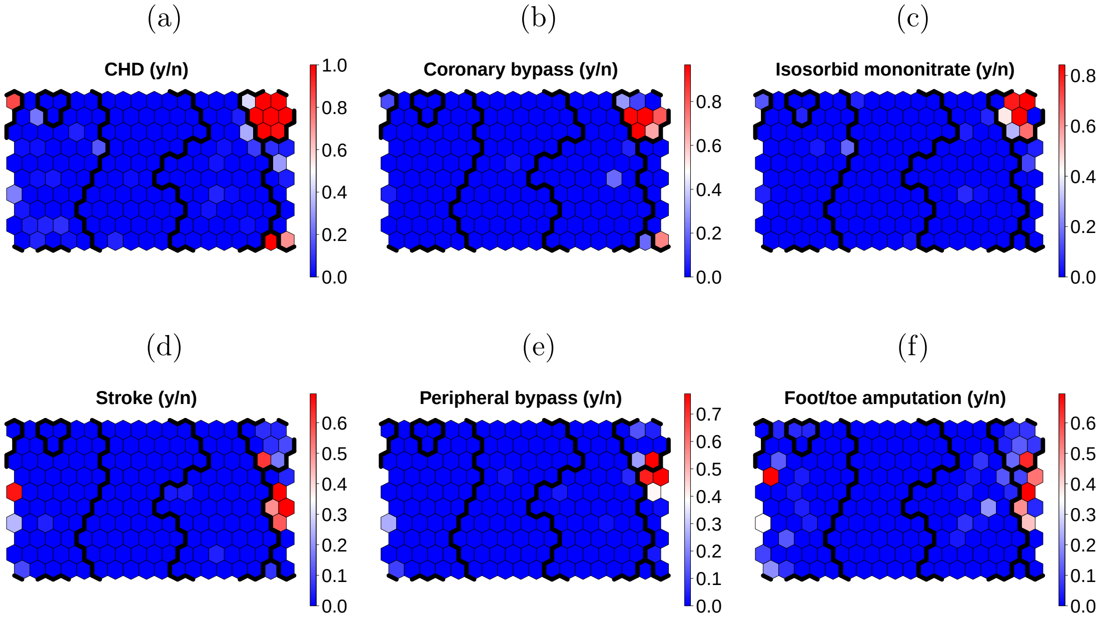


Figure 16: Selection of variables illustrated in the BL layer of the SOM using all mapped patients. Cluster borders according to assignments $k = 4$.

compared with low risk cluster (cluster 2). The difference in mortality between these clusters was slightly smaller than in case $k = 3$ (high vs. low risk clusters: 389/1,270 = 30% vs. 24/1,609 = 1%). Still, the difference of incident CVD and DN progression were of the same magnitudes compared with assignment $k = 3$ (CVD: 451/1,145 = 39% vs. 57/1,602 = 4%; DN progression: 340/901 = 38% vs. 114/1,377 = 8%)

5.5.4 Five clusters: rapid DN progressors among high risk patients

With $k = 5$, the larger high risk cluster of patients from previous assignments was split in two, mainly based on the degree of DN, as illustrated in Figure 17. Patients with microalbuminuria and ESRD were clearly separated into different clusters, however, the cluster border splits the region containing patients with a high prevalence of macroalbuminuria. The separation in this region is notable in terms of serum creatinine and cystatin C, both of which are markers of kidney function (see Figures 17a to 17e). The half with more severe complications is still referred as "(remaining) high risk cluster", and the separated half is referred as "rapid progressor cluster" due to the findings of regression modelling (further described later in this section). In addition, the remaining high risk cluster had a more frequent use of calcium channel blockers (a type of AHT medication), and yet higher blood pressure values. Finally, the separation was also seen in terms of additional medication (more common in high risk cluster), not belonging to the general classes of common medication (see Appendix A for used classes).

When the distribution of input variables in these classes were further tested, clearer differences between the clusters emerged, as the smaller subgroups began to represent more extreme ends of the population spectrum. Among the input variables, only three (insulin treatment type, other AHT medication, and social class 3) were still evenly distributed among the created classes, and all the other variables showed significant differences in distribution. The more specified clusters created showed also additional differences between the the BL and FU layers as the number of variables showing significant cluster-layer interaction increased to 19. These variables contained all the ten variables from the previous cluster assignment ($k = 4$), and the additional variables reflect mainly the more rapid progression of DN and related traits in rapid progressor cluster (cluster 2). Markers of kidney function, serum creatinine, serum cystatin-C, 24h urine collection albumin-to-creatinine ratio (ACR), liver fatty acid binding protein to creatinine ratio, and overnight urine AER, all increased more rapidly in rapid progressor cluster, whereas they stayed approximately at the same levels or decreased to lower levels in the other clusters. The prevalence of microalbuminuria, decreased in rapid progressor and CHD clusters (due to DN progression to worse levels), whereas it increased in other clusters (progression from normal AER to microalbuminuria, or survival bias in high risk cluster where some patients with more severe stages of DN deceased between BL and FU visits making the non-microalbuminuric sample in FU layer smaller, thus increasing the proportion). Other variables showing interaction included use of calcium channel blockers, that increased more rapidly in rapid progressor cluster; use of non-steroidal anti-inflammatory drugs, which became significantly lower in CHD cluster compared

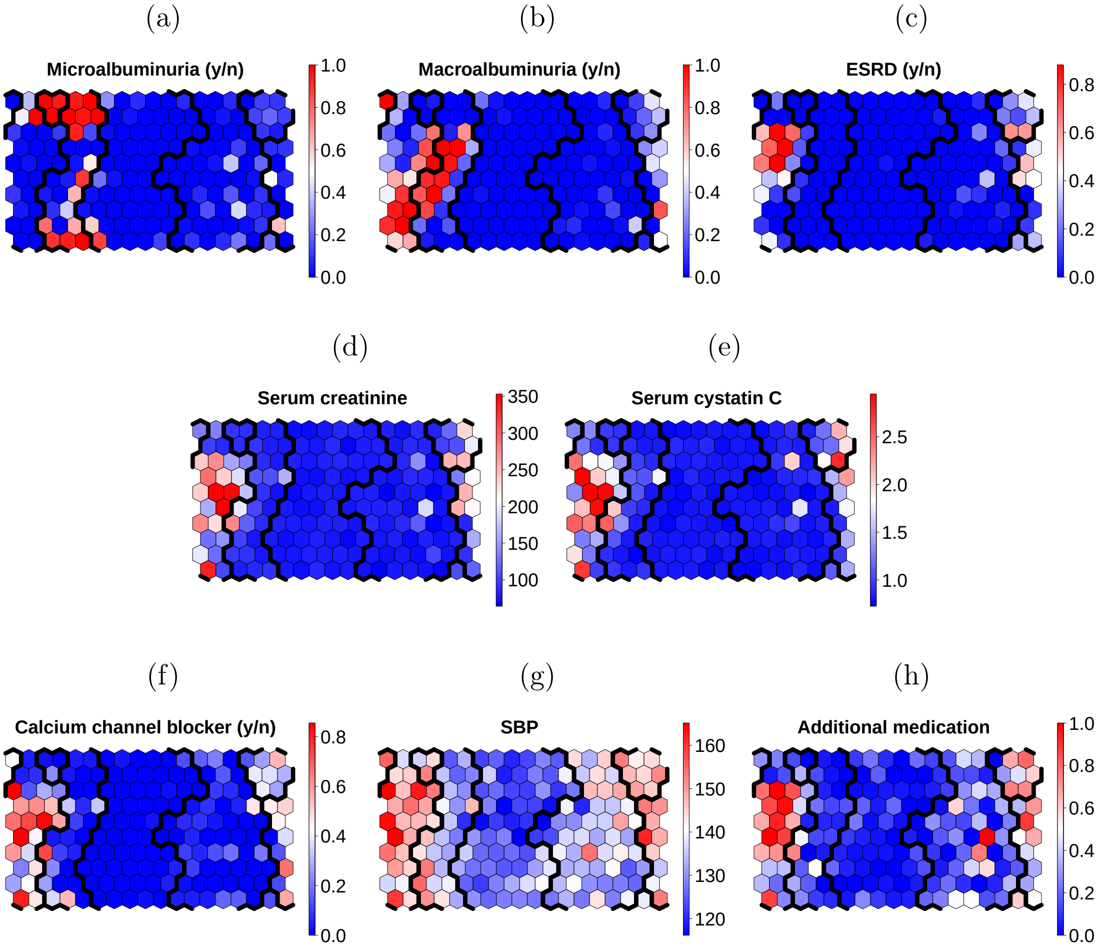


Figure 17: Selection of variables illustrated in the BL layer of the SOM using all mapped patients. Cluster borders according to assignments $k = 5$.

with the rest; and finally, number of cigarettes smoked, which increased more rapidly in rapid progressor cluster.

In terms of heritability, the point estimates became smaller and less significant, compared with $k = 4$. The only significant heritability was observed between the remaining high risk cluster and the low risk cluster. Still, this suggests that the genetic differences are most detectable between the extreme ends of complication spectrum. Splitting the high risk cluster also highlights the difference in outcome variables by increasing the mortality, DN progression and incident CVD events in the high risk cluster (mortality: $304/707 = 43\%$; incident CVD event: $311/593{=}52\%$; DN progression: $194/412{=}47\%$).

5.5.5 Six clusters: thyroid disease

In the final cluster division, a small subgroup of patients was separated from the large intermediate risk cluster. This group had a high prevalences of self-reported thyroid disease and use of thyroxin medication as illustrated in Figure 18, and is

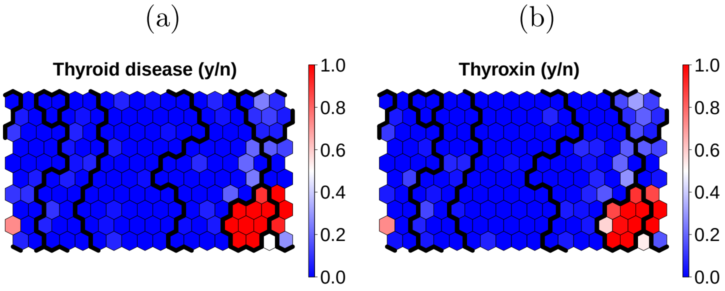


Figure 18: Selection of variables illustrated in the BL layer of the SOM using all mapped patients. Cluster borders according to assignments $k = 6$.

thus further referred as "thyroid cluster".

The separation of this small group of patients affected the cluster-wise distribution differences so that use of warfarin was not differentially distributed any more, but social class 3 became significant compared with $k = 5$. For the cluster-layer interaction, almost the same variables as in $k = 5$ showed significance, with the exception that the use of non-steroidal anti-inflammatory drugs, and 24h urine sample liver fatty acid binding protein to creatinine ratio became non-significant again.

The new thyroid cluster separated from the intermediate risk cluster showed surprisingly large point estimates of heritability between multiple groups, however, none of which were significant after Bonferroni correction (e.g. groups 2 vs. 5: $h^2 = 61\%$, $p = 0.039$). Thus it might be possible that the large heritability estimates reflect the genetic component responsible for thyroid disorders, some of which have been previously shown to be highly heritable [73, 74].

5.5.6 Summary of the created clusters

To summarize the created clusters, at the lowest resolution level when only two clusters were created, the cluster assignment could roughly be described as "patients with multiple complications versus patients without them". When the resolution was increased and more clusters were created, both the initial high and low risk clusters were separated into a smaller sub-clusters. From the initial low risk cluster, first patients with an intermediate complication risk/prevalence were separated, and later, a special group of patients with high prevalence of thyroid disorders was identified within the intermediate risk cluster. Correspondingly, within the high risk cluster, first a group of patients with high CHD prevalence was separated, after which, the remaining high risk cluster was split into two. One of the halves contained patients with already severe complications and the other contained patients with mainly microvascular complications that demonstrated more rapid progression of DN related traits compared with any of the other clusters. Generally, when the number of clusters was increased, the number of input variables evenly distributed among the classes decreased. Correspondingly, increasing the number of clusters introduced additional cluster-layer interactions for groups. The most significant features of the created clusters in terms of the clinical phenotypes are summarized in Figure 19.

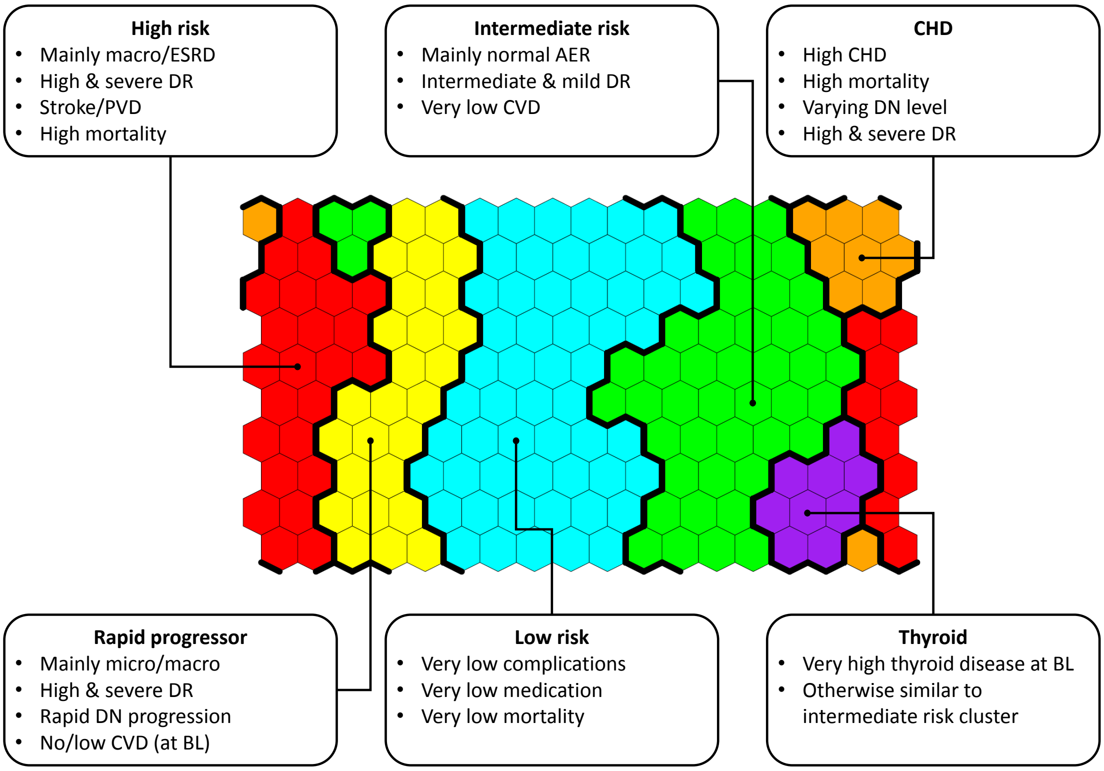


Figure 19: Summary of the most distinguishing complication profiles of the created clusters. Abbreviations micro and macro refer to microalbuminuria and macroalbuminuria correspondingly. Specifications low/intermediate/high refer to the prevalence of a complication, and mild/severe to the severity of the complication.

5.6 Genetic components of the novel phenotypes

The following section will present the results of the GWAS analyses between patient classes showing significant heritabilities, as presented in Appendix B. Each subsection will present the Manhattan and QQ-plots summarizing the GWAS analysis of the created phenotype at hand, and will then concentrate on the few top associated loci, the genes around these regions, and possible previously reported connections to diabetic complications and related traits.

5.6.1 Two clusters

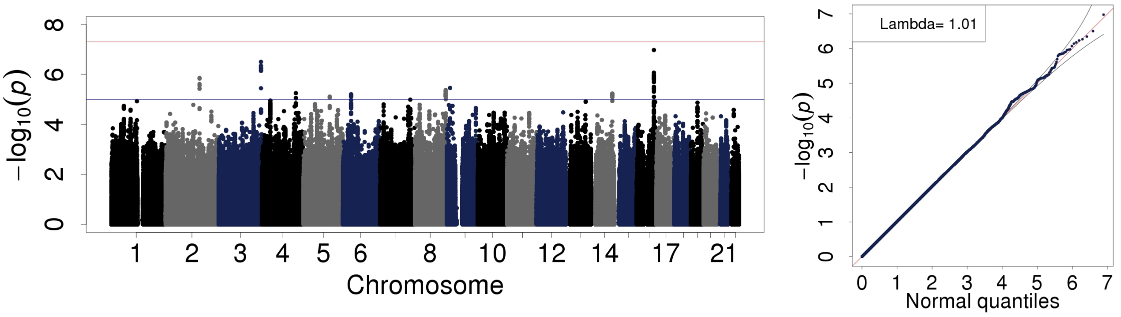
When the patients were divided into two, in the case of $k = 2$, the division could be roughly described as "patients with diabetic complications versus patients without them", as presented in section 5.5. Already these rough patient classes showed significant heritability, and were thus analysed in the GWAS setting. The results of these analyses are summarized in Figure 20.

As the QQ-plot presented in Figure 20b illustrates, there is no inflation of type I error in the analyses ($\lambda_{QC} = 1.01$). The associations show some loci with suggestive p-values ($p < 1 \times 10^{-5}$), the strongest of which on chromosomes 16, 3 and 2 (as illustrated in Figure 20a). However, no association reached genome-wide significance in these analyses ($p_{min} = 1.06 \times 10^{-7} > 5 \times 10^{-8}$). The strongest associated SNPs in the top four loci (lead SNP and two following SNPs) are presented in Table 3, including their p-values for association and effect sizes.

Even though the associations did not reach genome-wide significance, the topmost associated loci are located near genes that show biological links to the development of diabetic complications. First, the region with the most significant association (rs72803939 on chr16, $p = 1.06 \times 10^{-7}$) is located in the last intron (8/8) of the *WWOX* gene. Variants in this same intron have previously been associated with plasma high-density lipoprotein (HDL) -cholesterol and triglyceride (TG) levels [75]. Interestingly, abnormal lipid profiles are risk factors for both micro- and macrovascular diabetic complications as presented in section 2.2.

The second strongest association (rs55995864 on chr3, $p = 3.14 \times 10^{-7}$) is located near the *HES1* gene (approximately 100kb upstream), which belongs to the Notch signaling pathway, being a downstream effector of the Notch-1 receptor ligand Jagged1 encoded by *JAG1*. Upregulation of Notch-1 signaling has been previously associated with diabetic nephropathy [76], and more importantly, increased levels of Hes1 and Jagged1 were observed in renal biopsies from patients with DN compared with biopsies from healthy non-diabetic controls [77]. However, there are also studies suggesting that genetic markers in genes belonging to the Notch-1 signaling pathway are not associated with DN [78], and thus the link is under debate.

The third strongest association detected (rs707098 on chr2, $p = 1.38 \times 10^{-6}$) is located approximately 5kb upstream from the *GALNT13* gene, which is an enzyme responsible for the synthesis of O-glycan [79]. Previously, changes in O-glycans have been reported in renal tissues of alloxan diabetic rats, and other animal studies



(a) Manhattan plot of associations

(b) QQ-plot

Figure 20: Manhattan plot (a) and QQ-plot (b) of the GWAS results between high and low risk clusters of $k = 2$. In the Manhattan plot, the x-axis is the chromosome and bp-position of the tested SNP, while the y-axis is the strength of association. Each of the plotted points represents one of the SNPs tested. In the QQ-plot, the x-axis is the expected random p-values, and the y-axis represents the observed p-values of the analysis. Large deviation from the diagonal (throughout all points) suggests inflation of type I error.

Table 3: Top loci of GWAS results in the clustering $k = 2$. A1: Reference allele; A2: Risk allele, FRQ: Allele frequency of A2; INFO: Imputation quality information criteria; OR: Multiplier for risk to belong into group with higher complication prevalence for each copy of A2; SE: Standard error for beta (log[OR])

SNP	chr	bp	A1	A2	FRQ	INFO	OR	SE	p-value
rs72803939	16	78,577,080	C	G	0.12	0.93	1.53	0.08	1.06×10^{-7}
rs72801987	16	78,556,270	A	G	0.11	0.90	1.53	0.09	8.49×10^{-7}
rs72803915	16	78,563,954	C	T	0.11	0.95	1.51	0.08	1.06×10^{-6}
rs55995864	3	193,784,165	T	A	0.29	0.91	1.35	0.06	3.14×10^{-7}
rs56280518	3	193,785,348	C	A	0.29	0.95	1.34	0.06	4.51×10^{-7}
rs12487368	3	193,787,775	C	T	0.29	0.99	1.33	0.06	5.38×10^{-7}
rs707098	2	155,318,908	G	C	0.52	0.99	1.29	0.05	1.38×10^{-6}
rs741602	2	155,315,938	C	T	0.48	1.00	0.78	0.05	1.46×10^{-6}
rs10932050	2	155,316,574	C	T	0.47	0.99	0.78	0.05	2.42×10^{-6}
rs62534516	9	14,511,929	T	C	0.02	0.81	0.32	0.24	3.48×10^{-6}
rs62534485	9	14,470,752	G	T	0.03	0.91	0.42	0.20	2.00×10^{-5}
rs62534481	9	14,466,021	A	T	0.03	0.92	0.43	0.20	2.49×10^{-5}

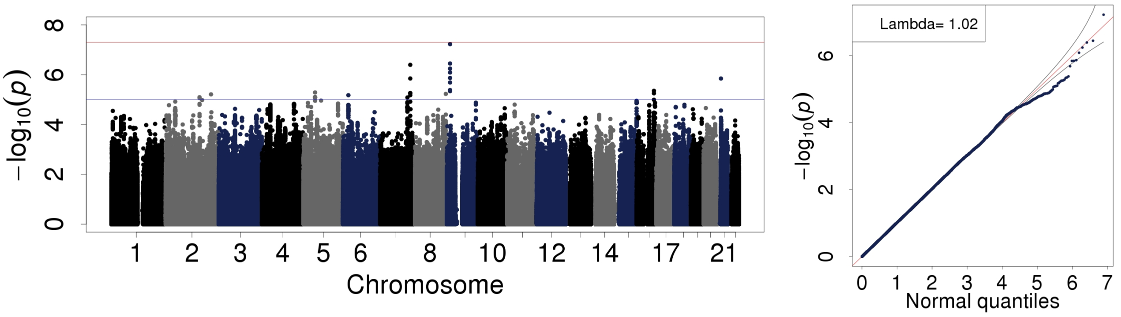
suggest that changes in O-glycans have functional relevance to the pathogenesis of diabetic complications [80]. In addition, variants in *GALNT13* have been suggestively associated with pediatric BMI [79].

The fourth most significant locus (rs62534516 on chr9, $p = 3.48 \times 10^{-6}$) is located between the genes *ZDHHC21* (approximately 40kb downstream) and *NFIB* (approximately 100kb upstream). Of these *ZDHHC21* interacts with *NOS3* (also known as eNOS) palmitoylating the enzyme encoded by *NOS3*, and thus it is related to nitric oxide (NO) metabolism [81]. *NOS3* itself has been previously linked to DN [82, 83, 30] and to DR [31]. In general, impaired NO metabolism affects blood pressure through vasodilation, and increases oxidative stress which is a common denominator of diabetic microvascular complications [29].

5.6.2 Three clusters

When the number of clusters was increased from two to three, the low risk cluster of previous cluster assignment was split into two, and only the half with lower prevalence of any complications still showed significant heritability compared with the high risk cluster. The results of these analyses are presented in Figure 21.

The analyses of these patient clusters yield an almost genome-wide significant association ($p_{min} = 5.95 \times 10^{-8}$), but it remains still above the threshold. Corresponding to previous analysis with $k = 2$, the results show suggestive loci that have connections to complications and related traits, and the inflation in type I error is



(a) Manhattan plot of associations

(b) QQ-plot

Figure 21: Manhattan plot (a) and QQ-plot (b) of the GWAS results between high and low risk clusters of $k = 3$.

still constrained ($\lambda_{QC} = 1.02$, see Figure 21b). The most significant SNPs for four of the most significant loci are presented in Table 4.

In these analyses, the most significant association (rs62534516 on chr9, $p = 5.95 \times 10^{-8}$) was the same SNP detected as the fourth most significant locus of the analyses with $k = 2$, between the genes *NFIB* and *ZDHHC21*, which has a plausible link to diabetic complications via NO metabolism, vasodilation, and oxidative stress. The second most significant hit was seen on chr7 (rs79477588, $p = 4.03 \times 10^{-7}$) in the first intron (1/33) of the *DGKI* gene. It belongs to a family of diacylglycerol kinases, which catalyze the conversion of diacylglycerol (DAG) to phosphatidic acid. DAG is known to activate the protein kinase C (PKC) pathway [84], which in turn is one of

Table 4: Top loci of GWAS results in the clustering $k = 3$.

SNP	chr	bp	A1	A2	FRQ	INFO	OR	SE	p-value
rs62534516	9	14,511,929	T	C	0.02	0.81	0.26	0.25	5.95×10^{-8}
rs62534485	9	14,470,752	G	T	0.03	0.90	0.35	0.21	3.57×10^{-7}
rs62534481	9	14,466,021	A	T	0.03	0.91	0.35	0.21	5.70×10^{-7}
rs79477588	7	137,468,992	C	T	0.04	0.81	0.43	0.16	4.03×10^{-7}
rs79020676	7	137,486,254	T	C	0.05	0.83	0.49	0.15	1.43×10^{-6}
rs74482009	7	137,480,792	C	T	0.10	0.87	0.62	0.10	5.44×10^{-6}
rs202095311	21	16,368,833	GCAAA	G	0.03	0.83	0.35	0.22	1.43×10^{-6}
rs142323171	21	16,404,339	G	A	0.02	0.85	0.32	0.27	2.73×10^{-5}
rs144965913	21	16,397,200	A	G	0.04	0.92	0.51	0.17	5.35×10^{-5}
rs1079323	16	78,667,846	C	T	0.76	0.85	0.72	0.07	4.44×10^{-6}
rs2550615	16	78,668,315	G	C	0.76	0.86	0.72	0.07	5.58×10^{-6}
rs72803939	16	78,577,080	C	G	0.13	0.92	1.48	0.09	1.02×10^{-5}

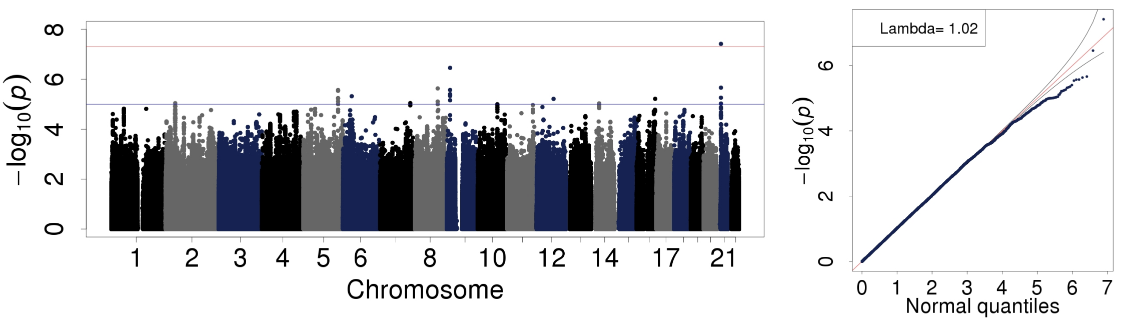
the four known damaging pathways in the development of DR [28]. Furthermore, the enzyme in rat, corresponding to the one coded by human *DGKI*, has been shown to be one of the dominant diacylglycerol kinases in the rat retina [85]. The third most significant hit (rs202095311 on chr21, $p = 1.43 \times 10^{-6}$) is located in the last intron (3/3) of the *NRIP1* gene (also known as *RIP140*), which has been associated with inflammation, as well as lipid and glucose metabolism in patients with T2D [86]. The fourth most significant locus (rs1079323, $p = 4.44 \times 10^{-6}$) was the same that was found around the lead association in the previous analyses with $k = 2$, i.e. variants in the last intron of *WWOX*. However, the lead SNP was not the same as in the previous analyses.

5.6.3 Four clusters

When the number of clusters was further increased to $k = 4$, a small subgroup of nodes was separated from the previous high risk cluster. The patients in this small subgroup showed a high prevalence of CHD and related traits, however, the only significant heritabilities were observed between the remaining high risk cluster and the previous low risk cluster. This estimated heritability was most significant among the created classes for any k. Results of the GWAS analysis using this phenotype are presented in Figure 22.

In the analysis, one genome-wide significant association was detected (rs202095311, $p = 3.81 \times 10^{-8}$). In addition, the analysis results show also other suggestive associations. However, the QQ-plot suggests that in general the detected p-values of associations start to fall below the diagonal, suggesting that the number of patients is starting to become too small for the GWAS analyses. Still the topmost associations rise above this trend, and reach smaller (more significant) p-values than in the previous analysis with $k = 3$. These and other top most significantly associated loci are presented in Table 5.

The top association reaching a genome-wide significance ($p = 3.81 \times 10^{-8}$) was located in a locus associated also in previous analyses ($k = 3$), in the last



(a) Manhattan plot of associations

(b) QQ-plot

Figure 22: Manhattan plot (a) and QQ-plot (b) of the GWAS results between high and low risk clusters of $k = 4$.

Table 5: Top loci of GWAS results in the clustering $k = 4$.

SNP	chr	bp	A1	A2	FRQ	INFO	OR	SE	p-value
rs202095311	21	16,368,833	GCAAA	G	0.03	0.83	0.24	0.26	3.81×10^{-8}
rs142323171	21	16,404,339	G	A	0.02	0.84	0.20	0.33	2.17×10^{-6}
rs182526753	21	16,334,615	T	C	0.01	0.98	0.21	0.35	5.45×10^{-6}
rs62534516	9	14,511,929	T	C	0.02	0.81	0.26	0.27	3.48×10^{-7}
rs62534485	9	14,470,752	G	T	0.03	0.90	0.36	0.22	2.68×10^{-6}
rs75697152	9	14,533,977	G	A	0.02	0.81	0.27	0.28	3.95×10^{-6}
rs10111377	8	104,121,210	A	G	0.38	0.91	1.35	0.06	2.33×10^{-6}
rs13261053	8	104,107,265	C	T	0.30	0.94	1.34	0.07	7.78×10^{-6}
rs6468859	8	104,120,370	G	C	0.29	0.93	1.34	0.07	9.96×10^{-6}
rs75669230	5	158,229,364	T	C	0.05	1.00	0.48	0.16	2.69×10^{-6}
rs12657410	5	158,204,347	A	G	0.05	1.02	0.49	0.15	2.89×10^{-6}
rs12651861	5	158,238,388	T	C	0.04	0.96	0.48	0.16	5.65×10^{-6}

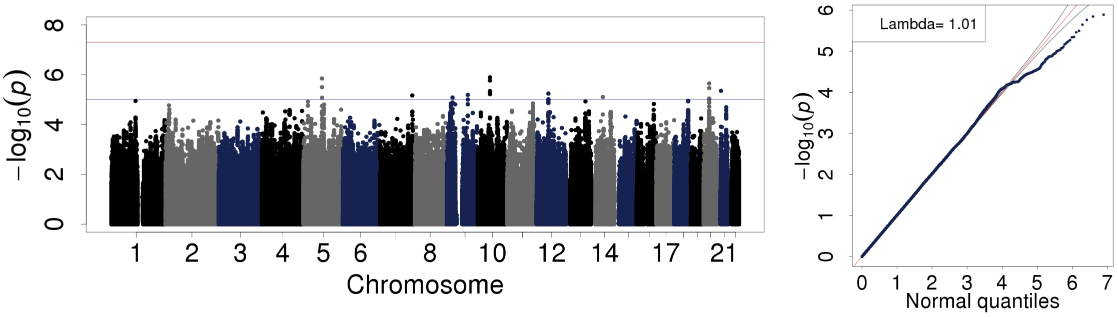
intron of *NRIP1* gene (rs202095311, chr21), which has a link to metabolism and inflammation. The second most significant association ($p = 3.48 \times 10^{-7}$) was also seen in both of the previous analyses ($k = \{2, 3\}$): it was located in the region between the genes *ZDHHC21* and *NFIB* (the same SNP rs62534516 that was leading the association in previous analyses). The third most significant association (rs10111377, $p = 2.33 \times 10^{-6}$) was seen on chr8 between the genes *ATP6V1C1* (approximately 35kb downstream) and *BAALC* (approximately 30kb upstream). Of these, *ATP6V1C1* encodes a component of the multisubunit enzyme, vacuolar ATPase (V-ATPase), which is involved in ROS production in response to bacteria [87]. Generally, increased ROS production has been hypothesised to be the unifying mechanism of the diabetic microvascular complications [29]. In addition, Gene Ontology terms link *ATP6V1C1* itself to the large insulin receptor signaling pathway (GO:0008286). The fourth most significant association (rs75669230 on chr5, $p = 2.69 \times 10^{-6}$) was located in the middle of the *EBF1* gene, in intron 8/15. The gene has been previously associated with metabolic and cardiovascular risk [88, 89] and *ebf1* knock-out mice have shown renal defects [90].

5.6.4 Five clusters

During the next cluster division, when k was increased to five, the previous high risk cluster was split into two, roughly separating patients rapid DN progression from the high risk cluster. The heritability between the remaining extreme cases (high risk cluster) and the previous control group was still significant, even though both the point estimate and significance were smaller than in the case $k = 4$, possibly due to a continuously decreasing number of patients. The GWAS results between these classes are summarized in Figure 23.

As the QQ-plot in Figure 23b illustrates, the resulting p-values are falling below the diagonal, i.e. they are generally worse than expected by random. As all the performed tests are not independent (due to the LD structure between the SNPs) this does not imply that the association results would be useless, but suggests that the number of patients is already too small to detect the underlying genetic associations, if present. In these analyses the lead associations were almost two magnitudes larger (worse) compared with the lead associations from $k = 3$ or $k = 4$. Still the most significant regions from this analysis show interesting genetic regions, and are presented in Table 6.

The most significant hit was seen on chr10 (rs1578671, $p = 1.28 \times 10^{-6}$) on top of the *PCDH15* gene (intron 4/34). Nonsynonymous variants in this gene have been previously linked to lipid abnormalities (especially TG, ApoB and total cholesterol levels) [91], and also a common variant in the region has been reported to be suggestively associated with lipid profiles [92]. The second most significant locus was found on chr5 (rs200163200, $p = 1.42 \times 10^{-6}$), around a region where there are no known genes within a distance of more than 500kb in either direction of the lead SNP. The third region is found on chr20 on top of gene *ABHD12* (intron 4/12, rs35288907, $p = 2.25 \times 10^{-6}$). The encoded enzyme catalyses the hydrolysis of 2-arachidonoyl glycerol (2-AG), one of the two most important endocannabinoids. Levels of 2-AG have been reported to correlate with body fat, visceral fat mass, fasting plasma insulin, and glucose infusion rate during a glucose clamp [93]. Furthermore, the endocannabinoid system has been shown to generally have complex effects on the bodyweight and metabolic regulation in general, possibly leading to accumulation of energy as fat, and thus, it is linked also to T2D [94]. More importantly, the endocannabinoid system is hypothesised to play a role also in the progression of DR [95]. Finally, the fourth region was the one detected previously with genome-wide significant p-value with $k = 4$: locus on top of the *NRIP1* gene on chr21 (rs202095311, $p = 4.47 \times 10^{-6}$). However, with $k = 5$, the p-values are two magnitudes smaller compared with the previous analyses, and no other SNPs in this region around the



(a) Manhattan plot of associations

(b) QQ-plot

Figure 23: Manhattan plot (a) and QQ-plot (b) of the GWAS results between high and low risk clusters of $k = 5$.

Table 6: Top loci of GWAS results in the clustering $k = 5$

SNP	chr	bp	A1	A2	FRQ	INFO	OR	SE	p-value
rs1578671	10	56,216,271	C	T	0.14	0.96	1.60	0.10	1.28×10^{-6}
rs11004315	10	56,215,605	G	T	0.14	0.97	1.59	0.10	1.72×10^{-6}
rs3070712	10	56,214,186	CATT	C	0.13	1.00	1.58	0.10	4.55×10^{-6}
rs200163200	5	84,677,581	G	GA	0.03	0.91	2.73	0.21	1.42×10^{-6}
rs188674266	5	84,764,920	G	T	0.03	0.91	2.59	0.20	3.17×10^{-6}
rs190418551	5	85,060,481	G	A	0.03	0.96	2.36	0.19	8.41×10^{-6}
rs35288907	20	25,299,299	CT	C	0.60	0.84	0.69	0.08	2.25×10^{-6}
rs5841058	20	25,433,447	C	CT	0.41	0.80	1.45	0.08	3.45×10^{-6}
rs1044573	20	25,206,654	A	G	0.51	1.00	1.37	0.07	9.11×10^{-6}
rs202095311	21	16,368,833	GCAAA	G	0.03	0.83	0.18	0.37	4.47×10^{-6}

lead SNP reached even nominal p-values ($p < 1 \times 10^{-4}$).

For the next clustering ($k = 6$) the cluster division did not affect the high and low risk clusters showing significant heritability with $k = 5$. These same groups were the only pairing showing significant heritability in the case $k = 6$, and thus it was not re-analysed. As previously noted, the dendrogram of Ward's clustering and the continuously decreasing heritability estimates suggested that no further increase in k should be done after $k = 6$. Furthermore, already using the larger clusters of $k = 5$ clustering suggested that the resulting number of patients is getting too small for the GWAS analyses. Thus no further GWAS analyses were performed.

6 Discussion and conclusions

The primary goals of this work were to use machine learning approaches to subdivide patients with T1D into novel phenotype classes (based on multiple clinical and environmental variables), that would simultaneously capture the risk for multiple diabetic complications; to evaluate whether the resulting class assignments show genetically differing profiles; and to pinpoint the exact genetic markers associated with these novel group assignments.

It has been previously shown that SOM-based approaches can map the patients with T1D successfully by their cross-sectional complication profile and mortality, using their biochemical profiles as an input [67, 96], which has greatly inspired this work. Here the traditional SOM was extended to a multi-layer approach, as we hypothesised that this approach could also capture the progressive nature of the diabetic complications more efficiently. In this work, a dual-layer SOM was trained using data from thousands of patients in the FinnDiane study, measured at the baseline visit (BL) when the patients entered the study, and during a follow-up visits (FU) years later. After the SOM was trained, the resulting node prototypes of the SOM were clustered using Ward's agglomerative hierarchical clustering in order to create the phenotype classes. These classes were further evaluated for their intergroup heritabilities, and pairs of classes showing genetically divergent profiles were used as case-control phenotypes in the GWAS setting.

The created phenotype classes showed meaningful class assignments, as multiple diabetic complications and their risk factors were differentially distributed between them. The approach seem to be able also to capture the progressive nature of the complications, as some of the classes showed more rapid progression of DN and other complication related variables between the BL and FU layers of the SOM. Among the complications, abnormal AER of the patients, rate of DN progression and biomarkers of DN showed the most notable differences between the created classes. However, this phenomenon might be slightly biased, as the FinnDiane study has a special emphasis on the study of DN, and thus the data might be partly enriched for variables reflecting the status and progression of DN. Still, also other complications showed clustering in the SOM, and some of the created phenotype classes followed these patterns well. Furthermore, the borders created to separate the different patient classes followed the mortality and complication progression patterns of the latest available data, and thus SOM and the created phenotype classes seem to have a predictive value also beyond the time points of the input data.

The created patient classes also showed differing genetic profiles for multiple pairs of groups created with different "resolution levels" of the hierarchical clustering, as evaluated by narrow-sense heritability estimates. The most significant estimates suggested that certain class assignments could be up to 42% genetically determined. For some special groups the point estimates for heritability were even larger, but non-significant possibly due to the small patient samples assigned to these groups. The pairs of groups showing significant heritability estimates were analysed in the GWAS setting, where the class assignment was used as a case-control phenotype.

In the GWAS analyses, the group pairing showing the most significant heritability highlighted a genetic marker (rs202095311, chr21) in the last intron of the *NRIP1* gene associated with the class assignment with genome-wide significance ($p < 5 \times 10^{-8}$). The gene has previously been associated with inflammation, lipid metabolism, and glucose metabolism in patients with T2D, and thus it has a biologically plausible link also to diabetic complications in T1D. The association was achieved with only 2,439 patients included in the GWAS analysis, whereas previous GWAS analyses on diabetic complications have required more than ten thousand patients to achieve genome-wide significance [10]. In addition, the analyses of this work highlighted multiple other regions that showed suggestive p-values ($p < 1 \times 10^{-5}$) for genetic association around genetic regions and genes previously associated with diabetic complications, related risk factors and/or relevant pathways, further supporting the power of the approach.

In addition to a genome-wide significant p-value, replication in an independent cohort is often required for the genetic associations from the GWAS approach before they are widely accepted. However, replicating the results of this work is very challenging, as the data in the FinnDiane study used to train the SOMs is unique in its comprehensiveness and variety, and thus there are no other cohorts with matching data that could be used as replication cohorts. However, if such a cohort would exist, it would be relatively straightforward to map the patients into the optimal SOM trained, assign them to clusters, and perform the corresponding GWAS analyses. Still, interpreting the GWAS results in this approach is generally harder compared with traditional phenotypes, as the class assignments associated with the genetic markers are created using a wide spectrum of different variables. Thus, it is possible that the observed association(s) reflect only one of the components having differences between the phenotype classes. Still, based on the associated loci, genes in nearby regions, and the traits they have been previously associated with, this is unlikely the case. However, drawing any final conclusions is still challenging, and thus, the genetic associations of this work are best suited for future hypothesis generation. They could be later used for example in pathway enrichment analyses to capture the wider spectrum of the pathogenesis of diabetic complications, however, this is out of the scope of this thesis.

To summarize, the selected approach was able to create meaningful patient classes that showed genetically divergent backgrounds. A genetic variant associated with the class assignment could be identified, and it and the additional suggestively associated loci had mostly previously known biologically plausible links to diabetic complications and/or related traits. Thus the work reached all the primary goals set for it.

The secondary goals of this work were to evaluate the used methodology: whether the multi-layer SOM approach could capture the progressive nature of the complications, which of the clustering methods would be most suitable to create the phenotype classes, and finally, whether a corresponding approach could be used also in other complex traits and diseases.

In this work, clustering of the multi-layer SOM prototypes was selected as the tool to create the novel phenotype classes. Still, machine learning and data mining are broad fields and there exist multiple approaches to group data and create classifiers in order to achieve the same goal of creating novel phenotype classes of patients.

Selecting SOM as the core tool limits the spectrum of available approaches, and among the available tools, SOM indeed is a "black box" approach that does not yield specific classification rules. However, using it does not require complex models that require pre-assumptions of the data. As SOM has been previously shown to be a good tool to map patients with T1D, and as the traditional SOM approach had room for extension into a multi-layer approach that was hypothesised to be able to capture the progressive nature of the diabetic complications, it was selected to be used in this work.

As the SOM prototype nodes to be clustered have coordinates in multiple vector spaces simultaneously, most of the clustering tools are not directly suited for the clustering task. Many of them could be hand tailored to work on such data, but creating new implementations of clustering approaches for a specific task was out of the scope of this work. Thus, the selection of tools was limited to existing, well documented R implementations of methods that can perform the clustering based on a (dis)similarity matrix alone. The selection for the most suited tool to create the phenotype classes was not straightforward. All of the tested methods had certain features where they outperformed the others, and the final selection had to be based on the core properties of the used implementations. Still, the results suggest that having the multi-layer SOM as an abstraction level of the raw data can make even very simple clustering methods perform well, and it is not necessary to increase the complexity of the approach by applying more complex clustering methods. The selected method, agglomerative hierarchical clustering using Ward's criteria, is still only suboptimal, and if the corresponding pipeline is applied in other complex traits and diseases, the selection of clustering method should be given an additional emphasis. However, the preliminary results suggest that the selection of clustering method does not have major effects on the final results (see Appendix C).

Regardless of the limited selection of tested methods, the multi-layer SOM based approach was shown to perform well, and seems to capture the progressive nature of the diabetic complications. The results of this work look generally very interesting and can be considered as "a proof of concept". Thus, a similar approaches could be helpful in hypothesis generation, and for the detection of novel genetic variants and pathways associated with other complex diseases with multiple sub-phenotypes.

If one wishes to apply a corresponding approach to other complex diseases and traits, the replication issues of GWAS analyses should also be further considered. Thus, it would be recommended to select the variables used to train the SOM and map the patients so that corresponding measures would exist in the possible replication cohorts. Alternatively, if the initial cohort is large enough (possibly even tens to hundreds of thousands of patients), separating an independent replication cohort large enough for a GWAS could be considered.

In this work, the dual-layer SOM design was used due to the structure of available data, but including additional layers should in principle improve the mapping of the patients. All of the methods used in this work scale also to additional layers of the SOM, which could be included if the data is sufficient. Thus applying the approach to other complex traits with possibly larger patient cohorts, possibly stronger genetic effects compared with diabetic complications, and possibly longer and tighter follow-up data to fill additional layers could show an even better performance compared with the results of this work.

In conclusion, using multi-layer SOM to create novel phenotype classes could help the hypothesis generation in complex diseases with a progressive nature. The approach can highlight the baseline and progression differences of multiple variables between the created patient classes, and thus, give an additional insight into the pathogenesis of the disease. The classes can be further tested for a genetically divergent background to confirm the presence of a genetic effects, and ultimately, the exact genetic markers, genes and/or pathways associated with the group assignments can be pinpointed using the GWAS setting.

References

[1] DIAMOND Project Group, “Incidence and trends of childhood type 1 diabetes worldwide 1990-1999.” *Diabetic medicine: a journal of the British Diabetic Association*, vol. 23, no. 8, p. 857, 2006.

[2] V. Harjutsalo, L. Sjöberg, and J. Tuomilehto, “Time trends in the incidence of type 1 diabetes in Finnish children: a cohort study,” *The Lancet*, vol. 371, no. 9626, pp. 1777–1782, 2008.

[3] P. Onkamo, S. Väänänen, M. Karvonen, and J. Tuomilehto, “Worldwide increase in incidence of type I diabetes–the analysis of the data on published incidence trends,” *Diabetologia*, vol. 42, no. 12, pp. 1395–1403, 1999.

[4] P.-H. Groop, M. C. Thomas, J. L. Moran, J. Wadèn, L. M. Thorn, V.-P. Mäkinen, M. Rosengård-Bärlund, M. Saraheimo, K. Hietala, O. Heikkilä *et al.*, “The presence and severity of chronic kidney disease predicts all-cause mortality in type 1 diabetes,” *Diabetes*, vol. 58, no. 7, pp. 1651–1658, 2009.

[5] S. Laing, A. Swerdlow, S. Slater, J. Botha, A. Burden, N. Waugh, A. Smith, R. Hill, P. Bingley, C. Patterson *et al.*, “The British diabetic association cohort study, II: cause-specific mortality in patients with insulin-treated diabetes mellitus,” *Diabetic Medicine*, vol. 16, no. 6, pp. 466–471, 1999.

[6] R. Klein and B. E. Klein, “Vision disorders in diabetes,” *Diabetes in America*, vol. 1, p. 293, 1995.

[7] Diabetes Control and Complications Trial Research Group *et al.*, “Clustering of long-term complications in families with diabetes in the diabetes control and complications trial,” *Diabetes*, vol. 46, no. 11, p. 1829, 1997.

[8] K. Hietala, C. Forsblom, P. Summanen, and P.-H. Groop, “Heritability of proliferative diabetic retinopathy,” *Diabetes*, vol. 57, no. 8, pp. 2176–2180, 2008.

[9] V. Harjutsalo, S. Katoh, C. Sarti, N. Tajima, and J. Tuomilehto, “Population-based assessment of familial clustering of diabetic nephropathy in type 1 diabetes,” *Diabetes*, vol. 53, no. 9, pp. 2449–2454, 2004.

[10] N. Sandholm, R. M. Salem, A. J. McKnight, E. P. Brennan, C. Forsblom, T. Isakova, G. J. McKay, W. W. Williams, D. M. Sadlier, V.-P. Mäkinen *et al.*, “New susceptibility loci associated with kidney disease in type 1 diabetes,” *PLoS Genet*, vol. 8, no. 9, p. e1002921, 2012.

[11] S. M. Hosseini, A. P. Boright, L. Sun, A. J. Canty, S. B. Bull, B. E. Klein, R. Klein, A. D. Paterson, DCCT/EDIC Research Group *et al.*, “The association of previously reported polymorphisms for microvascular complications in a meta-analysis of diabetic retinopathy,” *Human genetics*, vol. 134, no. 2, pp. 247–257, 2015.

[12] E. Ahlqvist, N. R. van Zuydam, L. C. Groop, and M. I. McCarthy, "The genetics of diabetic complications," *Nature Reviews Nephrology*, vol. 11, no. 5, pp. 277–287, 2015.

[13] N. Sandholm, A. J. McKnight, R. M. Salem, E. P. Brennan, C. Forsblom, V. Harjutsalo, V.-P. Mäkinen, G. J. McKay, D. M. Sadlier, W. W. Williams *et al.*, "Chromosome 2q31.1 associates with ESRD in women with type 1 diabetes," *Journal of the American Society of Nephrology*, pp. ASN–2 012 111 122, 2013.

[14] M. Porta, I. Toppila, N. Sandholm, S. M. Hosseini, C. Forsblom, K. Hietala, L. Borio, V. Harjutsalo, B. E. Klein, R. Klein *et al.*, "Variation in *SLC19A3* and protection from microvascular damage in type 1 diabetes," *Diabetes*, p. db151247, 2015.

[15] D. M. Nathan, "Long-term complications of diabetes mellitus," *New England Journal of Medicine*, vol. 328, no. 23, pp. 1676–1685, 1993.

[16] S. M. Marshall and A. Flyvbjerg, "Clinical review-prevention and early detection of vascular complications of diabetes," *BMJ-British Medical Journal-International Edition*, vol. 333, no. 7566, pp. 475–480, 2006.

[17] R. Klein, M. D. Knudtson, K. E. Lee, R. Gangnon, and B. E. Klein, "The Wisconsin epidemiologic study of diabetic retinopathy XXII: the twenty-five-year progression of retinopathy in persons with type 1 diabetes," *Ophthalmology*, vol. 115, no. 11, pp. 1859–1868, 2008.

[18] R. Klein, B. E. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets, "The Wisconsin epidemiologic study of diabetic retinopathy: II. prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years," *Archives of ophthalmology*, vol. 102, no. 4, pp. 520–526, 1984.

[19] S. Tesfaye, L. Stevens, J. Stephenson, J. Fuller, M. Plater, C. Ionescu-Tirgoviste, A. Nuber, G. Pozza, J. Ward, EURODIAB IDDM Complications Study Group *et al.*, "Prevalence of diabetic peripheral neuropathy and its relation to glycaemic control and potential risk factors: the EURODIAB IDDM Complications Study," *Diabetologia*, vol. 39, no. 11, pp. 1377–1384, 1996.

[20] M. Henricsson, A. Nilsson, L. Groop, A. Heijl, and L. Janzon, "Prevalence of diabetic retinopathy in relation to age at onset of the diabetes, treatment, duration and glycemic control," *Acta Ophthalmologica Scandinavica*, vol. 74, no. 6, pp. 523–527, 1996.

[21] P. Hovind, L. Tarnow, P. Rossing, M. Graae, I. Torp, C. Binder, and H.-H. Parving, "Predictors for the development of microalbuminuria and macroalbuminuria in patients with type 1 diabetes: inception cohort study," *Bmj*, vol. 328, no. 7448, p. 1105, 2004.

[22] J. Larsen, M. Brekke, L. Bergengen, L. Sandvik, H. Arnesen, K. Hanssen, and K. Dahl-Jorgensen, "Mean HbA$_1$c over 18 years predicts carotid intima media thickness in women with type 1 diabetes," *Diabetologia*, vol. 48, no. 4, pp. 776–779, 2005.

[23] P. Dickinson, A. Carrington, G. Frost, and A. Boulton, "Neurovascular disease, antioxidants and glycation in diabetes," *Diabetes Metab Res Rev*, vol. 18, no. 4, pp. 260–272, 2002.

[24] K. Sundquist and X. Li, "Type 1 diabetes as a risk factor for stroke in men and women aged 15–49: a nationwide study from Sweden," *Diabetic medicine*, vol. 23, no. 11, pp. 1261–1267, 2006.

[25] S. Laing, A. Swerdlow, S. Slater, A. Burden, A. Morris, N. Waugh, W. Gatling, P. Bingley, and C. Patterson, "Mortality from heart disease in a cohort of 23,000 patients with insulin-treated diabetes," *Diabetologia*, vol. 46, no. 6, pp. 760–765, 2003.

[26] R. Lithovius *et al.*, "Utilization and costs of prescription medication in patients with type 1 diabetes: Impact of diabetic kidney disease," Ph.D. dissertation, University of Helsinki, Faculty of Medicine, 3 2015.

[27] J. R. Simonsen, V. Harjutsalo, A. Järvinen, J. Kirveskari, C. Forsblom, P.-H. Groop, M. Lehto *et al.*, "Bacterial infections in patients with type 1 diabetes: a 14-year follow-up study," *BMJ open diabetes research & care*, vol. 3, no. 1, p. e000067, 2015.

[28] K. Hietala, "Risk factors for retinopathy in type 1 diabetes," Ph.D. dissertation, University of Helsinki, Faculty of Medicine, 6 2013.

[29] M. Brownlee, "The pathobiology of diabetic complications a unifying mechanism," *Diabetes*, vol. 54, no. 6, pp. 1615–1625, 2005.

[30] A. Mooyaart, E. Valk, L. Van Es, J. Bruijn, E. De Heer, B. Freedman, O. Dekkers, and H. Baelde, "Genetic associations in diabetic nephropathy: a meta-analysis," *Diabetologia*, vol. 54, no. 3, pp. 544–553, 2011.

[31] B. M. Hampton, S. G. Schwartz, M. A. Brantley Jr, and H. W. Flynn Jr, "Update on genetics and diabetic retinopathy," *Clinical ophthalmology (Auckland, NZ)*, vol. 9, p. 2175, 2015.

[32] D. Peng, J. Wang, R. Zhang, F. Jiang, S. Tang, M. Chen, J. Yan, X. Sun, S. Wang, T. Wang *et al.*, "Common variants in or near *ZNRF1*, *COLEC12*, *SCYL1BP1* and *API5* are associated with diabetic retinopathy in Chinese patients with type 2 diabetes," *Diabetologia*, vol. 58, no. 6, pp. 1231–1238, 2015.

[33] P. Deloukas, S. Kanoni, C. Willenborg, M. Farrall, T. L. Assimes, J. R. Thompson, E. Ingelsson, D. Saleheen, J. Erdmann, B. A. Goldstein *et al.*,

"Large-scale association analysis identifies new risk loci for coronary artery disease," *Nature genetics*, vol. 45, no. 1, pp. 25–33, 2013.

[34] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[35] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[36] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[37] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, "SOM PAK: The self-organizing map program package," *Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science*, 1996.

[38] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, *SOM toolbox for Matlab 5.* Citeseer, 2000.

[39] S. Kaski, "Data exploration using self-organizing maps," in *ACTA POLYTECHNICA SCANDINAVICA: MATHEMATICS, COMPUTING AND MANAGEMENT IN ENGINEERING SERIES NO. 82.* Citeseer, 1997.

[40] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *Neural Networks, IEEE Transactions on*, vol. 11, no. 3, pp. 586–600, 2000.

[41] M. Attik, L. Bougrain, and F. Alexandre, "Self-organizing map initialization," in *Artificial Neural Networks: Biological Inspirations–ICANN 2005.* Springer, 2005, pp. 357–362.

[42] R. Wehrens, L. M. Buydens *et al.*, "Self-and super-organizing maps in R: the Kohonen package," *Journal of Statistical Software*, vol. 21, no. 5, pp. 1–19, 2007.

[43] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis.* John Wiley & Sons, 2009, vol. 344.

[44] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[45] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?" *Journal of Classification*, vol. 31, no. 3, pp. 274–295, 2014.

[46] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[47] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[48] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab – an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004. [Online]. Available: http://www.jstatsoft.org/v11/i09/

[49] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[50] U. Bodenhofer, A. Kothmeier, and S. Hochreiter, "APCluster: an R package for affinity propagation clustering," *Bioinformatics*, vol. 27, no. 17, pp. 2463–2464, 2011.

[51] D. R. Bentley, "The human genome project—an overview," *Medicinal research reviews*, vol. 20, no. 3, pp. 189–196, 2000.

[52] F. J. Steemers and K. L. Gunderson, "Whole genome genotyping technologies on the BeadArray^{TM} platform," *Biotechnology journal*, vol. 2, no. 1, pp. 41–49, 2007.

[53] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances," *Nucleic acids research*, p. gkp552, 2009.

[54] "Infinum® CoreExome-24 v1.1 BeadChip," http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_human_core_exome_beadchip.pdf, accessed: 2016-03-10.

[55] P. M. Visscher, W. G. Hill, and N. R. Wray, "Heritability in the genomics era—concepts and misconceptions," *Nature Reviews Genetics*, vol. 9, no. 4, pp. 255–266, 2008.

[56] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis," *The American Journal of Human Genetics*, vol. 88, no. 1, pp. 76–82, 2011.

[57] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen *et al.*, "The international HapMap project," *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[58] 1000 Genomes Project Consortium *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[59] Y. Li, C. Willer, S. Sanna, and G. Abecasis, "Genotype imputation," *Annual review of genomics and human genetics*, vol. 10, p. 387, 2009.

[60] B. Howie, J. Marchini, and M. Stephens, "Genotype imputation with thousands of genomes," *G3: Genes, Genomes, Genetics*, vol. 1, no. 6, pp. 457–470, 2011.

[61] S. Purcell, "PLINK v. 1.07," http://pngu.mgh.harvard.edu/purcell/plink/.

[62] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly *et al.*, “PLINK: a tool set for whole-genome association and population-based linkage analyses,” *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[63] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, “Genome-wide association studies for complex traits: consensus, uncertainty and challenges,” *Nature Reviews Genetics*, vol. 9, no. 5, pp. 356–369, 2008.

[64] L. M. Thorn, C. Forsblom, J. Fagerudd, M. C. Thomas, K. Pettersson-Fernholm, M. Saraheimo, J. Wadén, M. Rönnback, M. Rosengård-Bärlund, C.-G. Af Björkesten *et al.*, “Metabolic syndrome in type 1 diabetes association with diabetic nephropathy and glycemic control (the FinnDiane Study),” *Diabetes Care*, vol. 28, no. 8, pp. 2019–2024, 2005.

[65] N. Tolonen *et al.*, “Lipid profile and micro- and macro vascular complications in type 1 diabetes,” Ph.D. dissertation, University of Helsinki, Faculty of Medicine, 1 2015.

[66] V.-P. Mäkinen *et al.*, “Computational analysis of the metabolic phenotypes in type 1 diabetes and their associations with mortality and diabetic complications,” Ph.D. dissertation, Helsinki University of Technology, Aalto University School of Science and Technology, P.O.Box 1100, 00076 AALTO, 2 2010.

[67] V.-P. Mäkinen, C. Forsblom, L. M. Thorn, J. Wadén, D. Gordin, O. Heikkilä, K. Hietala, L. Kyllönen, J. Kytö, M. Rosengård-Bärlund *et al.*, “Metabolic phenotypes, vascular complications, and premature deaths in a population of 4,197 patients with type 1 diabetes,” *Diabetes*, vol. 57, no. 9, pp. 2480–2487, 2008.

[68] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, *cluster: Cluster Analysis Basics and Extensions*, 2015, R package version 2.0.3 — For new features, see the ’Changelog’ file (in the package source).

[69] A. Signorell *et al.*, *DescTools: Tools for Descriptive Statistics*, 2016, R package version 0.99.16. [Online]. Available: http://CRAN.R-project.org/package=DescTools

[70] Lesnoff, M., Lancelot, and R., *aod: Analysis of Overdispersed Data*, 2012, R package version 1.3. [Online]. Available: http://cran.r-project.org/package=aod

[71] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, “Principal components analysis corrects for stratification in genome-wide association studies,” *Nature genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[72] N. Sandholm, C. Forsblom, V.-P. Mäkinen, A. J. McKnight, A.-M. Österholm, B. He, V. Harjutsalo, R. Lithovius, D. Gordin, M. Parkkonen *et al.*, "Genome-wide association study of urinary albumin excretion rate in patients with type 1 diabetes," *Diabetologia*, vol. 57, no. 6, pp. 1143–1153, 2014.

[73] B. Vaidya, P. Kendall-Taylor, and S. H. Pearce, "The genetics of autoimmune thyroid disease," *The Journal of Clinical Endocrinology & Metabolism*, vol. 87, no. 12, pp. 5385–5397, 2002.

[74] T. H. Brix, K. O. Kyvik, K. Christensen, and L. Hegedüs, "Evidence for a major role of heredity in Graves' disease: A population-based study of two Danish twin cohorts 1," *The Journal of Clinical Endocrinology & Metabolism*, vol. 86, no. 2, pp. 930–934, 2001.

[75] M. E. Sáez, A. González-Pérez, M. T. Martínez-Larrad, J. Gayán, L. M. Real, M. Serrano-Ríos, and A. Ruiz, " *WWOX* gene is associated with HDL cholesterol and triglyceride levels," *BMC medical genetics*, vol. 11, no. 1, p. 148, 2010.

[76] C.-L. Lin, F.-S. Wang, Y.-C. Hsu, C.-N. Chen, M.-J. Tseng, M. A. Saleem, P.-J. Chang, and J.-Y. Wang, "Modulation of Notch-1 signaling alleviates vascular endothelial growth factor–mediated diabetic nephropathy," *Diabetes*, vol. 59, no. 8, pp. 1915–1925, 2010.

[77] D. W. Walsh, S. A. Roxburgh, P. McGettigan, C. C. Berthier, D. G. Higgins, M. Kretzler, C. D. Cohen, S. Mezzano, D. P. Brazil, and F. Martin, "Co-regulation of Gremlin and Notch signalling in diabetic nephropathy," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1782, no. 1, pp. 10–21, 2008.

[78] D. Kavanagh, G. McKay, C. Patterson, A. McKnight, A. Maxwell, D. Savage, Warren 3/UK GoKinD Study Group *et al.*, "Association analysis of Notch pathway signalling genes in diabetic nephropathy," *Diabetologia*, vol. 54, no. 2, pp. 334–338, 2011.

[79] B. Namjou, M. Keddache, K. Marsolo, M. Wagner, T. Lingren, B. Cobb, C. Perry, S. Kennebeck, I. A. Holm, R. Li *et al.*, "EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children," *Front. Genet.*, vol. 4, no. 268, 2013.

[80] Y. Ihara, S. Manabe, M. Kanda, H. Kawano, T. Nakayama, I. Sekine, T. Kondo, and Y. Ito, "Increased expression of protein C-mannosylation in the aortic vessels of diabetic Zucker rats," *Glycobiology*, vol. 15, no. 4, pp. 383–392, 2005.

[81] C. Fernández-Hernando, M. Fukata, P. N. Bernatchez, Y. Fukata, M. I. Lin, D. S. Bredt, and W. C. Sessa, "Identification of golgi-localized acyl transferases that palmitoylate and regulate endothelial nitric oxide synthase," *The Journal of cell biology*, vol. 174, no. 3, pp. 369–377, 2006.

[82] A. McKnight, C. Patterson, N. Sandholm, J. Kilner, T. Buckham, M. Parkkonen, C. Forsblom, D. Sadlier, P.-H. Groop, A. Maxwell *et al.*, "Genetic polymorphisms in nitric oxide synthase 3 gene and implications for kidney disease: a meta-analysis," *American journal of nephrology*, vol. 32, no. 5, pp. 476–481, 2010.

[83] A. Zanchi, D. K. Moczulski, L. S. Hanna, M. Wantman, J. H. Warram, and A. S. Krolewski, "Risk of advanced diabetic nephropathy in type 1 diabetes is associated with endothelial nitric oxide synthase gene polymorphism," *Kidney international*, vol. 57, no. 2, pp. 405–413, 2000.

[84] I. Idris, S. Gray, and R. Donnelly, "Protein kinase C activation: isozyme-specific effects on metabolism and cardiovascular complications in diabetes," *Diabetologia*, vol. 44, no. 6, pp. 659–673, 2001.

[85] S. Sato, Y. Hozumi, S. Saino-Saito, H. Yamashita, and K. Goto, "Enzymatic activity and gene expression of diacylglycerol kinase isozymes in developing retina of rats," *Biomedical Research*, vol. 32, no. 5, pp. 329–336, 2011.

[86] J. Xue, H. Zhao, G. Shang, R. Zou, Z. Dai, D. Zhou, Q. Huang, and Y. Xu, "*RIP140* is associated with subclinical inflammation in type 2 diabetic patients." *Experimental and clinical endocrinology & diabetes: official journal, German Society of Endocrinology and German Diabetes Association*, vol. 121, no. 1, pp. 37–42, 2013.

[87] R. S. Flannagan, G. Cosío, and S. Grinstein, "Antimicrobial mechanisms of phagocytes and bacterial evasion strategies," *Nature Reviews Microbiology*, vol. 7, no. 5, pp. 355–366, 2009.

[88] A. Singh, M. A. Babyak, D. K. Nolan, B. H. Brummett, R. Jiang, I. C. Siegler, W. E. Kraus, S. H. Shah, R. B. Williams, and E. R. Hauser, "Gene by stress genome-wide interaction analysis and path analysis identify *EBF1* as a cardiovascular and metabolic risk gene," *European Journal of Human Genetics*, vol. 23, no. 6, pp. 854–862, 2015.

[89] P. Petrus, N. Mejhert, H. Gao, J. Bäckdahl, E. Arner, P. Arner, and M. Rydén, "Low early B-cell factor 1 (*EBF1*) activity in human subcutaneous adipose tissue is linked to a pernicious metabolic profile," *Diabetes & metabolism*, vol. 41, no. 6, pp. 509–512, 2015.

[90] K. M. Schmidt-Ott, "The *Ebf1* knockout mouse and glomerular maturation," *Kidney international*, vol. 85, no. 5, pp. 1014–1016, 2014.

[91] A. Huertas-Vazquez, C. L. Plaisier, R. Geng, B. E. Haas, J. Lee, M. M. Greevenbroek, C. van der Kallen, T. W. de Bruin, M.-R. Taskinen, K. N. Alagramam *et al.*, "A nonsynonymous SNP within *PCDH15* is associated with lipid traits in familial combined hyperlipidemia," *Human genetics*, vol. 127, no. 1, pp. 83–89, 2010.

[92] D. K. Sanghera, L. F. Been, S. Ralhan, G. S. Wander, N. K. Mehra, J. R. Singh, R. E. Ferrell, M. I. Kamboh, and C. E. Aston, "Genome-wide linkage scan to identify loci associated with type 2 diabetes and blood lipid phenotypes in the sikh diabetes study," *PLoS One*, vol. 6, no. 6, p. e21188, 2011.

[93] M. Blüher, S. Engeli, N. Klöting, J. Berndt, M. Fasshauer, S. Bátkai, P. Pacher, M. R. Schön, J. Jordan, and M. Stumvoll, "Dysregulation of the peripheral and adipose tissue endocannabinoid system in human abdominal obesity," *Diabetes*, vol. 55, no. 11, pp. 3053–3060, 2006.

[94] V. Di Marzo, "The endocannabinoid system in obesity and type 2 diabetes," *Diabetologia*, vol. 51, no. 8, pp. 1356–1367, 2008.

[95] T. Behl, I. Kaur, and A. Kotwani, "Role of endocannabinoids in the progression of diabetic retinopathy," *Diabetes Metab Res Rev*, vol. 32, no. 3, pp. 251–259, 2015.

[96] V.-P. Mäkinen, P. Soininen, C. Forsblom, M. Parkkonen, P. Ingman, K. Kaski, P.-H. Groop, and M. Ala-Korpela, "^{1}H NMR metabonomics approach to the disease continuum of diabetic complications and premature death," *Molecular systems biology*, vol. 4, no. 1, p. 167, 2008.

A Full list of the SOM input variables

This appendix contains complete list of the input variables used to train the SOMs in the analyses of this work, including short summaries of the variables and their data types. Only the 79 variables passing the missingness threshold pruning (further described in sections 4.1.1 and 5.1) are presented.

Table A1: All input variables included in the training of the SOMs in this work.

Variable	Non-missing	Variable type	Description
DNGROUP	0.96	categorical (4 levels)	Nephropathy class [normal AER, microalbuminuria, macroalbuminuria, ESRD]
AGE	1.00	continuous	Age [years]
DURATION	1.00	continuous	Duration of the diabetes [years]
HEIGHT	0.99	continuous	Height [cm]
WEIGHT	0.99	continuous	Weight [kg]
WAIST	0.97	continuous	Waist circumference [cm]
HIP	0.97	continuous	Hip circumference [cm]
SBP	0.99	continuous	Systolic blood pressure [mmHg]
DBP	0.99	continuous	Diastolic blood pressure [mmHg]
OHATREAT	1.00	binary	Diabetes treated with tablets [y/n]
ANYRETIN	0.97	binary	Any retinal changes [y/n]
LASER	0.99	binary	Eyes laser treated [y/n]
CHD	1.00	binary	Coronary heart disease [y/n]
AMI	1.00	binary	Acute myocardial infarction [y/n]
CORBYPASS	1.00	binary	Coronary bypass [y/n]
STROKE	1.00	binary	Stroke [y/n]
AMPUTATION	1.00	binary	Toe/foot amputation [y/n]
PVDBYPASS	1.00	binary	Peripheral vein bypass (foot) [y/n]
INSTREAT	0.99	binary	Insulin therapy type [injection/pump]
INSDOSE	1.00	continuous	Daily insulin dosage [IU]
ANYMED	1.00	binary	Any medication [y/n]

Continued on next page

Continued from previous page

Variable	Non-missing	Variable type	Description
ACEINHIBIT	0.99	binary	Use of ACE inhibitor (medication) [y/n]
AT2RBLOCK	0.99	binary	Use of anginotensin II receptor blocker (medication) [y/n]
BETABLOCK	0.99	binary	Use of betablocker (medication) [y/n]
CABLOCK	0.99	binary	Use of calcium channel blocker (medication) [y/n]
DIURETICS	0.99	binary	Use of diuretics (medication) [y/n]
OTHERAHTMED	0.99	binary	Use of other antihypertensive medication [y/n]
NITRO	0.99	binary	Use of isosorbid mononitrate (long acting) medication [y/n]
NSAID	0.99	binary	Use of non-steroidal anti-inflammatory drug (medication) [y/n]
LIPIDLOWMED	0.99	binary	Use of lipid lowering medication [y/n]
THYROXIN	0.99	binary	Use of use of thyroxin (medication) [y/n]
HORMONES	0.99	binary	Use of hormone replacement therapy [y/n]
PPILLS	0.99	binary	Use of P-pills (medication) [y/n]
WARFARIN	0.99	binary	Use of warfarin (medication) [y/n]
OTHERMED	0.99	binary	Any other medication not mentioned before [y/n]
ONSET-TO-INS	1.00	discrete	Time from diabetes diagnosis to insulin treatment [years]
ASTHMA	0.91	binary	Self-reported asthma [y/n]
THYROIDDIS	0.91	binary	Self-reported thyroid disease [y/n]
RHEUMA	0.91	binary	Self-reported rheumatoid arthritis [y/n]

Continued on next page

Continued from previous page

Variable	Non-missing	Variable type	Description
BEERDOSE	0.83	continuous	Self-reported beer consumed per week [bottles 1/3 l]
WINEDOSE	0.78	continuous	Self-reported wine consumed per week [glasses]
BOOZEDOSE	0.76	continuous	Self-reported spirits consumed per week [dl]
SOCIALCLASS	0.84	categorical (6 levels)	Self-reported profession [working personel without professional education, trained working personel, lower level employee, higher level employee, farmer, not classified]
MOTBIRTHYEAR	0.86	discrete	Self-reported birth year of mother [year]
MOTALIVE	0.95	binary	Self-reported mother alive [y/n]
MOTAHT	0.81	binary	Self-reported mother has/had antihypertensive medication [y/n]
MOTAMI	0.81	binary	Self-reported mother has/had acute myocardial infarction [y/n]
MOTSTROKE	0.81	binary	Self-reported mother has/had stroke [y/n]
FATBIRTHYEAR	0.84	discrete	Self-reported birth year of father [year]
FATALIVE	0.94	binary	Self-reported father alive [y/n]
FATAHT	0.74	binary	Self-reported father has/had antihypertensive medication [y/n]
FATAMI	0.76	binary	Self-reported father has/had acute myocardial infarction [y/n]
FATSTROKE	0.75	binary	Self-reported father has/had stroke [y/n]
SIBLINGCOUNT	0.90	discrete	Self-reported number of siblings

Continued on next page

Continued from previous page

Variable	Non-missing	Variable type	Description
TOTAL-SMOKES	0.92	continuous	Self-reported total smokes (derived from self-reported daily smoking amount and smoking periods)
NEVER-EVER-SMOKED	0.92	binary	Self-reported has ever smoked [y/n]
SIBLING-DM	0.97	binary	Self-reported any sibling has diabetes [y/n]
MOT-DM	0.90	binary	Self-reported mother had diabetes [y/n]
FAT-DM	0.86	binary	Self-reported father had diabetes [y/n]
S-APOA1	0.92	continuous	Serum apolipoprotein A-I [mg/dl]
S-APOB	0.92	continuous	Serum apolipoprotein B-100 [mg/dl]
S-CHOL	0.99	continuous	Serum total cholesterol [mmol/l]
S-HDLC	0.99	continuous	Serum HDL cholesterol [mmol/l]
S-HDL2C	0.96	continuous	Serum HDL2 cholesterol [mmol/l]
S-TG	0.99	continuous	Serum triglycerides [mmol/l]
B-HBA1C	0.98	continuous	Blood haemoglobin A1c [mmol/mol]
S-CREAT	0.99	continuous	Serum creatinine [µmol/l]
S-CYSTATINC	0.86	continuous	Serum cystatin C [mg/l]
S-HSCRP	0.95	continuous	Serum high sensitive C-reactive protein [mg/l]
S-SRAGE	0.90	continuous	Serum soluble receptor for AGE [pg/ml]
S-URATE	0.85	continuous	Serum urate [µmol/l]
DU-TIME	0.78	continuous	24h urine collection time [min]
DU-VOLUME	0.78	continuous	24h urine volume [ml]
DU-CREAT	0.74	continuous	24h urine creatinine concentration [mmol/l]

Continued on next page

Continued from previous page

Variable	Non-missing	Variable type	Description
DU-ACR	0.72	continuous	24h urine albumin to creatinine ratio [mg/mmol]
NU-AER	0.55	continuous	Night urine albumin excretion rate [µg/l]
DU-ADIPONCT-CREA-RTIO	0.55	continuous	24h urine adiponectin to creatinine ratio
DU-KIM1-CREA-RTIO	0.64	continuous	24h urine KIM-1 (T cell-immunoglobulin-mucin-1) to creatinine ratio
DU-LFABP-CREA-RTIO	0.51	continuous	24h urine LFABP (liver fatty acid binding protein) to creatinine ratio

B Heritability estimates between defined phenotype classes

This appendix presents the estimated intergroup heritabilities between the patient classes, created by clustering the node prototypes of the optimal SOM fit into different number of clusters ($k = \{3, 4, 5, 6\}$) using agglomerative hierarchical clustering and the Ward's criteria for the cluster merging.

For some of the class pairs, the genetic variance component escaped from the parameter space during the iterative model optimization of GCTA run, and thus, heritability could not be estimated. In these cases, $*$ is presented in the following tables. Pairs of classes showing significant heritability (Bonferroni corrected for the number of pairings for current k, i.e. reaching $p < 0.05/[\frac{1}{2}k(k-1)]$) are highlighted in green, and these pairs were further used as a case-control phenotype in the GWAS setting. For class labels for each k in contrast to the SOM grid position and the hierarchical cluster structure, refer to Figures 12 and 13 in section 5.4.

Table B1: $k = 3$

		1	2
2	h^2 [SE]	0.29 [0.11]	-
	p	0.0048	-
	N	2600	-
3	h^2 [SE]	0.18 [0.12]	0.07 [0.11]
	p	0.063	0.22
	N	2267	2380

Table B2: $k = 4$

		1	2	3
2	h^2 [SE]	0.42 [0.12]	-	-
	p	1.8×10^{-4}	-	-
	N	2439	-	-
3	h^2 [SE]	0.22 [0.13]	0.08 [0.11]	-
	p	0.058	0.23	-
	N	2113	2380	-
4	h^2 [SE]	$*$	$*$	$*$
	p	$*$	$*$	$*$
	N	1350	1636	1315

Table B3: $k = 5$

		1	2	3	4
2	h^2 [SE]	0.03 [0.21]	-	-	-
	p	0.43	-	-	-
	N	1165	-	-	-
3	h^2 [SE]	0.37 [0.13]	0.14 [0.14]	-	-
	p	0.0010	0.17	-	-
	N	2017	1904	-	-
4	h^2 [SE]	0.36 [0.16]	$*$	0.08 [0.11]	-
	p	0.01	$*$	0.23	-
	N	1690	1590	2380	-
5	h^2 [SE]	$*$	0.06 [0.34]	$*$	$*$
	p	$*$	0.43	$*$	$*$
	N	876	750	1636	1315

Table B4: $k = 6$

		1	2	3	4	5
2	h^2 [SE]	0.03 [0.21]	-	-	-	-
	p	0.43	-	-	-	-
	N	1165	-	-	-	-
3	h^2 [SE]	0.37 [0.13]	0.14 [0.14]	-	-	-
	p	0.0010	0.16	-	-	-
	N	2017	1904	-	-	-
4	h^2 [SE]	0.36 [0.16]	$*$	$*$	-	-
	p	0.023	$*$	$*$	-	-
	N	1543	1437	2253	-	-
5	h^2 [SE]	0.51 [0.31]	0.61 [0.36]	0.23 [0.16]	0.26 [0.23]	-
	p	0.041	0.039	0.057	0.12	-
	N	842	715	1605	1133	-
6	h^2 [SE]	$*$	0.06 [0.34]	$*$	$*$	0.69 [0.61]
	p	$*$	0.43	$*$	$*$	0.11
	N	876	750	1636	1162	407

C Abstract for EDNSG annual meeting 2016

This appendix contains an abstract of the preliminary results of this work. It was submitted to the 29th Annual General Meeting of the European Diabetic Nephropahty Study Group (EDNSG; Palazzo Blu, Pisa, Italy, 20th to 21st of May, 2016) and accepted to be presented as an oral presentation.

At the time, spectral clustering was used to create the phenotype classes, but these initial results of class-wise progression of DN and most significant heritability estimates correspond closely to the final results presented in this work, suggesting that the selected clustering method does not have major effects on the general results. GWAS analyses of the classes were not yet performed at the time when the abstract was written.

Identifying novel phenotype profiles of patients with T1D using machine learning approaches

Iiro Toppila[1-3], Niina Sandholm[1-3], Carol Forsblom[1-3] and Per-Henrik Groop[1-4] on behalf of the FinnDiane Study group

[1]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland, [2]Abdominal Center Nephrology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland, [3]Diabetes and Obesity Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland, [4] Baker IDI Heart and Diabetes Institute, Melbourne, Australia

Objective: Many diabetic complications correlate with multiple biological markers and each other, and this interplay might limit traditional analyses studying one trait at a time. Therefore, the aim was to create novel classes of patients with Type 1 diabetes (T1D) considering multiple traits simultaneously by using a multi-layer self-organizing map (SOM) and clustering methods, to estimate inter-group heritabilities of the created patient classes, and to characterize their profiles.

Design: A nationally representative prospective cohort study.

Setting and patients: The study includes 4,409 patients with T1D (insulin initiated within one year of diagnosis, age at diabetes onset < 40 years) from the FinnDiane Study, of which 1,600 were used to train the SOMs. We used 79 different cross-sectional measures of clinical variables or complications from two different time points (corresponding to the SOM layers) to train 1,000 dual-layer SOMs with random initialization. The optimal SOM mapping was selected, and its node prototypes were clustered using spectral clustering varying the number of clusters. The heritability between each pair of the created groups was evaluated by the GCTA-software using data from genome-wide genotyping (Illumina Human Core+ExomeChip).

Main Outcome Measurements: Narrow-sense inter-group heritability (GCTA), and the differences in the progression rate of DN and input-data profiles between defined patient groups (continuous variables tested using Welch two sample t-test and binary variables using Fischer's exact test).

Results: The most significant heritability between two of the created classes was observed when all 4,409 subjects were mapped to the optimal SOM divided into 5 clusters. The difference in DN progression was significant (high risk group: 320/714=44.8%; low risk group: 66/957=6.8%; patients with baseline ESRD excluded; $p<2.2\times10^{-16}$) between these groups during a median follow-up of 7.2 years (IQR: 4.8-11.0). The patients in the high risk group showed worse profiles for most of the input variables including age, diabetes duration, anthropometric measures, blood pressure, medication, severity of baseline complications (DN, retinopathy, cardiovascular disease), familial history of CVD, and common blood and urinary markers. However, there were no differences in gender, insulin dose and treatment type (injection vs. pump), beer and wine doses, or occurrence of diabetes in father or siblings. Heritability (i.e. proportion of phenotypic variance explained by genetic factors) between the high and low risk groups was 43% ($p=9.13\times10^{-4}$).

Conclusions: Multilayer SOM can capture the progressive nature of DN, and create novel phenotypic profiles showing different progression rates and divergent genetic background.