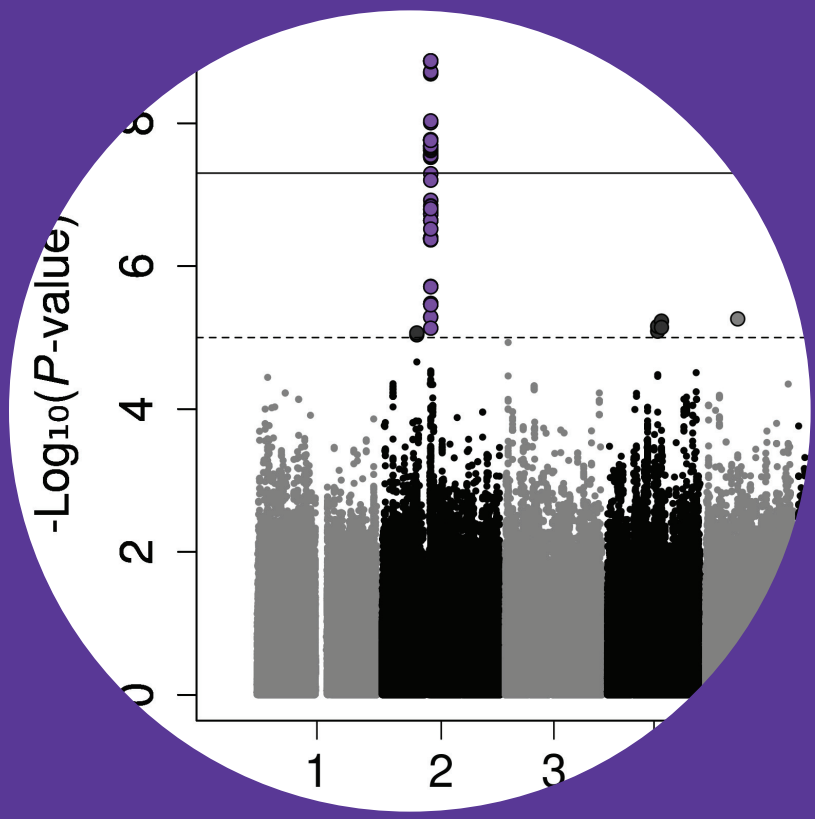
Dept. of Biomedical Engineering and Computational Science

Genome-wide associations and computational search for the genetic risk factors for diabetic nephropathy

Niina Sandholm



A!

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 104/2014

Genome-wide associations and computational search for the genetic risk factors for diabetic nephropathy

Niina Sandholm

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall F239a of the school on 15 August 2014 at 12.

Aalto University
School of Science
Dept. of Biomedical Engineering and Computational Science

Supervising professor
Prof. Kimmo Kaski

Thesis advisors
Prof. Per-Henrik Groop,
Folkhälsan Research Center, Helsinki, Finland
Division of Nephrology, Helsinki University Central Hospital, Finland

Dr. Ville-Petteri Mäkinen
South Australian Health and Medical Research Institute, Adelaide, Australia

Preliminary examiners
Prof. Sampsa Hautaniemi, University of Helsinki, Helsinki, Finland
Dr. Christophe Roos, Tampere University of Technology, Tampere, Finland

Opponent
Prof. Stephen S. Rich, University of Virginia, USA

Aalto University publication series
DOCTORAL DISSERTATIONS 104/2014

ISBN 978-952-60-5771-2
ISBN 978-952-60-5772-9 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5772-9

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Niina Sandholm

Name of the doctoral dissertation
Genome-wide associations and computational search for the genetic risk factors for diabetic nephropathy

Publisher School of Science

Unit Dept. of Biomedical Engineering and Computational Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 104/2014

Field of research Computational Systems Biology

Manuscript submitted 15 April 2014 **Date of the defence** 15 August 2014

Permission to publish granted (date) 4 June 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Type 1 diabetes (T1D) is an autoimmune form of diabetes where the patient's own immune system attacks the insulin producing islets of Langerhans in the pancreas. The long-term complications of diabetes reduce quality of life, lead to premature deaths and place a burden on the health care system. Diabetic kidney disease, known as diabetic nephropathy, is a major diabetic complication affecting one third of the patients with T1D. In some cases, diabetic nephropathy may lead to end stage renal disease (ESRD), a condition characterized by the inability of the kidneys to function at the level needed for day-to-day life. Patients with ESRD require regular dialysis treatment or kidney transplantation to survive.

While the pathogenesis of diabetic nephropathy is poorly understood, it is known that diabetic nephropathy clusters in families, suggesting that genetic risk factors affect the susceptibility to this complex disease. However, the genetic risk factors are not well known. Identification of the genetic risk factors would help to understand the biological processes causing the disease, paving the way for novel pharmacological target molecules and better biochemical risk markers.

The aim of this dissertation was to identify genetic risk factors for diabetic nephropathy by applying a range of computational methods to high-throughput genetic data.

This dissertation is mainly based on genome-wide data of ~550,000 single nucleotide polymorphisms (SNPs) genotyped in 3,650 Finnish patients with T1D. Similar genetic data were available for two other studies. Using computational methods and a European reference population, the number of SNPs for each patient was increased to 2.4 million.

With this large genomic data set, we first reassessed the previously suggested genetic risk factors for diabetic nephropathy. We then performed genome-wide association studies (GWASs) in the three cohorts. Combining our results with other studies, the resulting analysis included data from over 12,000 patients with T1D. In this larger cohort, we identified variants in the *AFF3* gene and between the *RGMA* and *MCTP2* genes associated with ESRD. Additionally, we identified variants that were only associated with the risk of ESRD in women with T1D. Furthermore, we identified risk variants for increased urinary albumin excretion, an important marker of diabetic kidney disease. Finally, using data mining methods, we identified the previously reported *RGMA – MCTP2* locus and two novel putative genetic risk factors for ESRD. All in all, this thesis reports the first genetic risk factors for diabetic nephropathy in T1D with strong statistical evidence of association.

Keywords computational genetics, genome-wide association study, GWAS, SNP, diabetic nephropathy

ISBN (printed) 978-952-60-5771-2 **ISBN (pdf)** 978-952-60-5772-9

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 100 **urn** http://urn.fi/URN:ISBN:978-952-60-5772-9

Tekijä
Niina Sandholm

Väitöskirjan nimi
Diabeettisen nefropatian geneettisten riskitekijöiden genominlaajuinen etsintä laskennallisin menetelmin

Julkaisija Perustieteiden korkeakoulu

Yksikkö Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 104/2014

Tutkimusala Laskennallinen systeemibiologia

Käsikirjoituksen pvm 15.04.2014 **Väitöspäivä** 15.08.2014

Julkaisuluvan myöntämispäivä 04.06.2014 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Ykköstyypin diabetes on yleensä nuorella iällä puhkeava krooninen sairaus joka puhkeaa kun kehon oma immuunijärjestelmä tuhoaa haiman insuliinia tuottavat solut ja normaali sokeri-aineenvaihdunta häiriintyy. Diabeteksen krooniset liitännäissairaudet ovat merkittävä uhka potilaiden hyvinvoinnille ja aiheuttavat huomattavia terveydenhoidon kustannuksia. Arviolta kolmannes ykköstyypin diabeetikoista sairastuu diabeettiseen munuaistautiin, niin kutsuttuun diabeettiseen nefropatiaan. Pahimmillaan se voi johtaa loppuvaiheen munuaissairauteen jolloin potilaan munuaiset evät enää toimi riittävän hyvin, vaan potilas tarvitsee selviytyäkseen säännöllistä dialyysihoitoa tai munuaissiirteen.

Vaikka diabeettisen munuaistaudin syntymekanismeja ei vielä tunneta tarkasti, perhetutkimusten perusteella diabeettinen munuaistauti näyttäisi olevan osittain perinnöllinen. Taudille altistavien perintötekijöiden tunnistaminen on tärkeää taudin syiden ymmärtämiseksi sekä uusien lääkkeiden ja parempien bio-markkereiden kehittämiseksi. Tämän väitöskirjan tavoitteena on etsiä perimästä laskennallisin menetelmin sellaisia muutoksia jotka lisäävät diabeettisen munuaistaudin riskiä ykköstyypin diabeetikoilla.

Väitöskirjan tutkimukset perustuvat pääasiassa suureen suomalaiseen tutkimusaineistoon joka koostui 3 650 ykköstyypin diabeetikosta. Tutkimukseen osallistuneiden genomi kartoitettiin noin 550 000 emäsparin määrityksellä, ja emäsparien lukumäärä kasvatettiin laskennallisilla menetelmillä 2,4 miljoonaan lopullisia analyyseja varten. Tutkimus sisälsi lisäksi kaksi muuta samankaltaista genominlaajuista aineistoa.

Tutkimme ensin kohdistetusti niitä emäspareja jotka on aiemmin yhdistetty diabeettiseen munuaistautiin. Etsimme sen jälkeen uusia geneettisiä riskitekijöitä koko genomin alueelta kolmessa genominlaajuisessa aineistossa. Kun yhdistimme tulokset yhdeksän muun osatutkimuksen kanssa, löysimme loppuvaiheen munuaistaudille altistavia geenimuutoksia *AFF3* geenistä sekä *RGMA* ja *MCTP2* geenien väliseltä alueelta. Lisäksi havaitsimme emäspareja joiden vaihtelu vaikuttaa loppuvaiheen munuaistaudin riskiin ainoastaan naisissa, sekä geenimuunnoksia, jotka vaikuttavat virtsan albumiinin määrään, joka on tärkeä munuaistaudin mittari. Lopuksi sovelsimme uutta laskennallista lähestymistapaa, jonka avulla havaitsimme jälleen loppuvaiheen munuaistaudille altistavia geenimuunnoksia *RGMA – MCTP2* alueella sekä kaksi kokonaan uutta aluetta joiden geenimuunnokset saattavat altistaa loppuvaiheen munuaistaudille. Kaiken kaikkiaan tämä väitöskirja esittelee ensimmäiset vahvasti diabeettiseen munuaistautiin liittyvät geneettiset riskitekijät ykköstyypin diabeetikoilla.

Avainsanat Laskennallinen genetiikka, genominlaajuinen assosiaatioanalyysi, GWAS, diabeettinen mumuaistauti, SNP

ISBN (painettu) 978-952-60-5771-2 **ISBN (pdf)** 978-952-60-5772-9

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2014

Sivumäärä 100 **urn** http://urn.fi/URN:ISBN:978-952-60-5772-9

Contents

List of original publications

This thesis consists of an overview and the following Publications and manuscript, which are referred to in the text by their Roman numerals.

I. Williams WW*, Salem RM*, McKnight AJ*, Sandholm N*, Forsblom C, Taylor A, Guiducci C, McAteer JB, McKay GJ, Isakova T, Brennan EP, Sadlier DM, Palmer C, Soderlund J, Fagerholm E, Harjutsalo V, Lithovius R, Gordin D, Hietala K, Kyto J, Parkkonen M, Rosengard-Barlund M, Thorn L, Syreeni A, Tolonen N, Saraheimo M, Waden J, Pitkaniemi J, Sarti C, Tuomilehto J, Tryggvason K, Osterholm AM, He B, Bain S, Martin F, Godson C, Hirschhorn JN, Maxwell AP, Groop PH, Florez JC, GENIE Consortium: Association testing of previously reported variants in a large case-control meta-analysis of diabetic nephropathy. *Diabetes* 2012;61(8):2187-2194.

II. Sandholm N*, Salem RM*, McKnight AJ*, Brennan EP*, Forsblom C, Isakova T, McKay GJ, Williams WW, Sadlier DM, Makinen VP, Swan EJ, Palmer C, Boright AP, Ahlqvist E, Deshmukh HA, Keller BJ, Huang H, Ahola AJ, Fagerholm E, Gordin D, Harjutsalo V, He B, Heikkila O, Hietala K, Kyto J, Lahermo P, Lehto M, Lithovius R, Osterholm AM, Parkkonen M, Pitkaniemi J, Rosengard-Barlund M, Saraheimo M, Sarti C, Soderlund J, Soro-Paavonen A, Syreeni A, Thorn LM, Tikkanen H, Tolonen N, Tryggvason K, Tuomilehto J, Waden J, Gill GV, Prior S, Guiducci C, Mirel DB, Taylor A, Hosseini SM, DCCT/EDIC Research Group, Parving HH, Rossing P, Tarnow L, Ladenvall C, Alhenc-Gelas F, Lefebvre P, Rigalleau V, Roussel R, Tregouet DA, Maestroni A, Maestroni S, Falhammar H, Gu T, Mollsten A, Cimponeriu D, Ioana M, Mota M, Mota E, Serafinceanu C, Stavarachi M, Hanson RL, Nelson RG, Kretzler M, Colhoun HM, Panduru NM, Gu HF, Brismar K, Zerbini G, Hadjadj S, Marre M, Groop L, Lajer M, Bull SB, Waggott D, Paterson AD, Savage DA, Bain SC, Martin F, Hirschhorn JN, Godson C, Florez JC, Groop PH, Maxwell AP: New susceptibility loci associated with kidney disease in type 1 diabetes. *PLoS Genet*. 2012;8(9):e1002921, 13 pp.

III. Sandholm N, McKnight AJ, Salem RM, Brennan EP, Forsblom C, Harjutsalo V, Makinen VP, McKay GJ, Sadlier DM, Williams WW, Martin F, Panduru NM, Tarnow L, Tuomilehto J, Tryggvason K, Zerbini G, Comeau ME, Langefeld CD, FIND Consortium, Godson C, Hirschhorn JN, Maxwell AP, Florez JC, Groop

PH, FinnDiane Study Group and the GENIE Consortium: Chromosome 2q31.1 associates with ESRD in women with type 1 diabetes. *J Am Soc Nephrol*. 2013;24(10):1537-1543.

IV. Sandholm N, Forsblom C, Makinen VP, McKnight AJ, Osterholm AM, He B, Harjutsalo V, Lithovius R, Gordin D, Parkkonen M, Saraheimo M, Thorn LM, Tolonen N, Waden J, Tuomilehto J, Lajer M, Ahlqvist E, Mollsten A, Marcovecchio ML, Cooper J, Dunger D, Paterson AD, Zerbini G, Groop L, on behalf of The SUMMIT Consortium, Tarnow L, Maxwell AP, Tryggvason K, Groop PH, on behalf of the FinnDiane Study Group: Genome-wide association study of urinary albumin excretion rate in patients with type 1 diabetes. *Diabetologia* 2014;57(6):1143-53.

V. Sambo F*, Malovini A*, Sandholm N*, Stavarachi M*, Forsblom C, Mäkinen V-P, Harjutsalo V, Lithovius R, Gordin D, Parkkonen M, Saraheimo M, Thorn LM, Tolonen N, Wadén J, He B, Österholm A-M, Tuomilehto J, Lajer M, Salem RM, McKnight AJ for the GENIE Consortium, Tarnow L, Panduru NM, Barbarini N, Di Camillo B, Toffolo GM, Tryggvason K, Bellazzi R, Cobelli C, Groop P-H: Novel genetic susceptibility loci for diabetic end stage renal disease identified through robust Naïve Bayes classification. *Diabetologia*. Published online 29 May 2014, 12pp.

* These authors contributed equally to the work

Author's contribution

In Publication I the author performed the statistical analyses in the FinnDiane study and contributed to the writing of the manuscript together with W.W. Williams, R.M. Salem and A.J. McKnight. In Publication II, the author performed all quality control measures, computational data preparation and computational analyses of the genome-wide genotype data in the largest of the three discovery studies, the FinnDiane Study. The author was one of the four main analysts and contributed to the analyses of the nine replication studies, meta-analyses including all 12 studies, and additional computational analyses and wrote the manuscript together with R.M. Salem, A.J. McKnight and E. Brennan. In Publication III, the author performed all computational analyses of the discovery phase, the meta-analyses of the four studies and additional computational analyses, and wrote the manuscript. In Publication IV, the author performed all computational analyses of the discovery phase, contributed to the analysis of the replication studies, performed the meta-analyses of the eight studies, contributed to the analysis of the sequencing results with A.J. McKnight, performed additional computational analyses and was the main writer of the

manuscript. In Manuscript V, the author was one of the four main analysts and writers of the manuscript together with F. Sambo, M. Stavarachi and A. Malovini. The author contributed to the association analyses in the discovery study (FinnDiane) and the Steno replication study, the meta-analyses of the four studies and supporting computational analyses. In all Publications, acquisition of the patient material, targeted genotyping, sequencing, and other laboratory analyses were performed by the co-authors. Genome-wide genotyping was outsourced to laboratories with the necessary infrastructure and expertise, and the authors performed a careful quality control to ensure the high quality of the data.

1 Introduction

A total of 347 million people worldwide have diabetes [Danaei 2011]. Type 1 diabetes (T1D) is an autoimmune variant of the disease affecting 40,000 patients in Finland. In fact, Finland has had the highest incidence of the disease in the world during the last four decades [Diabetes Epidemiology Research International Group 1988, The DIAMOND Project Group 2006], and the annual number of new T1D patients has more than doubled from 1980 to 2005 [Harjutsalo 2008]. After this enormous upsurge, the incidence of T1D in Finland seems to have settled to roughly 60 new T1D diagnoses per 100,000 persons per year [Harjutsalo 2013].

Even though diabetes is no longer the life-threatening disease it used to be a century ago, it still remains an enormous health care problem. Although the high blood glucose levels can be treated acutely with insulin injections, the long-term complications cause human suffering, premature deaths, and a significant burden to the health care system. In particular, diabetic kidney disease, also known as diabetic nephropathy (DN), is a severe late complication that in its most severe form leads to renal failure (end stage renal disease, ESRD) that requires kidney transplantation or regular dialysis for survival. DN is a common complication, as one third of the patients with T1D develop DN within 25 years of the diagnosis of diabetes in Finland [Harjutsalo 2004]. In fact, diabetes is the major cause of ESRD in the Western world [Finne 2010, U.S. Renal Data System 2011]. DN is also strongly associated with the risk of cardiovascular disease [Borch-Johnsen and Kreiner 1987, Tuomilehto 1998] and all-cause mortality [Groop 2009].

The treatment of diabetes and its complications imposes substantial demands on the health care system. In Finland the annual health care costs for diabetes exceed 12% of Finland's health care expenditure, and these costs are to a large extent due to the costs generated by the complications [Kangas 2001]. In the USA, the estimated direct medical costs of diabetes were $176 billion in 2012, of which the largest components were hospital care (43%) and medication to treat diabetic complications (18%). The diabetes medication and other supplies for diabetes care contributed only to 12% of the direct costs [American Diabetes Association 2013]. The indirect costs, i.e. the loss of productivity due to mortality and morbidity, are also high, estimated $69 billion per year and are the consequence of the diabetic complications [American Diabetes Association 2013].

Despite the severity and high prevalence of DN, the pathogenesis of the disease remains poorly understood. DN is likely to have complex environmental and genetic

origins [Thorn 2007]. Known risk factors include high blood pressure, poor blood glucose control, dyslipidemia, male gender and long duration of diabetes [Parving and Smidt 1986, The DCCT Research Group 1995, Tarnow 2008, Tolonen 2009]. In addition, sibling studies have shown that DN clusters in families. It is estimated that genetic factors increase the risk of DN by a factor of two [Seaquist 1989, Borch-Johnsen 1992, Quinn 1996, Harjutsalo 2004], but the actual genes involved remain largely unknown. Genetic risk factors for DN have been sought in multiple populations [Maeda 2007, Tarnow 2008, Pezzolesi 2009a]. Unfortunately, most of the initial associations have not been robustly replicated [Conway and Maxwell 2009, Mooyaart 2011]. Discovery of the susceptibility genes, as well as the genetic pathways they affect, could help understand the development and different phases of the disease. A better understanding of the causal factors for DN and its pathogenesis could pave the way to development of clinical applications such as new therapeutic target molecules and better biochemical and genetic risk markers and ultimately to new strategies to preemptively treat the disorder to attenuate morbidity and mortality.

This doctoral thesis investigates the genetic factors that carry propensity to DN by applying a range of computational methods to genome-wide genetic data from a unique population-based T1D cohort in Finland, the FinnDiane Study. In Publication I, we examined previously suggested genetic associations with DN in a large set of 6,366 patients with T1D. In Publications II-III and V, we describe a genome-wide search for novel genetic risk factors for DN and ESRD, whereas in Publication IV we examined the genetic risk factors for urinary albumin excretion, which is a continuous marker of incipient and established DN. Publications II-IV follow the standard methodology for the genome-wide search of risk factors, whereas in Publication V we applied a novel data mining algorithm in order to explore if additional susceptibility loci could be identified with a Bayesian approach.

2 Diabetes and its complications

2.1 Diabetes

Diabetes is a group of metabolic diseases characterized by elevated blood glucose levels, caused by the body's partial or complete inability to transfer glucose from blood to the cells. Type 1 diabetes (T1D) is usually diagnosed in children and young adults when the body initiates an autoimmune attack against its own insulin producing beta cells within the islets of Langerhans in the pancreas. Insulin is a vital hormone that is required to transfer glucose from the blood to the muscles, liver and fat tissue cells. The important role of insulin was first discovered in 1921 by Nicolae Paulescu, and insulin injections became the life-saving treatment for diabetes already in 1922 [Banting 1922]. New-onset T1D results in insulin deficiency, and if untreated, leads to increasing blood glucose concentrations and ketoacidosis, and finally coma. On the other hand, too much insulin results in the opposite condition of low blood glucose – hypoglycemia – which may also lead to unconsciousness. Therefore careful regulation of the blood glucose levels is essential for the diabetic patients.

Type 2 diabetes (T2D) is the most common form of diabetes, accounting for more than 90% of all diagnosed cases [Skyler and Oddo 2002]. T2D is sometimes called adulthood onset diabetes, as it is strongly affected by aging, life-style factors and obesity. In T2D, the body becomes increasingly resistant to the insulin action, leading to a mounting demand for insulin production. Eventually, the pancreatic beta cells cannot produce enough insulin, and the blood glucose levels start to rise. In the final phase, the adverse metabolic milieu further impairs insulin production and the beta cell function begins to decline, thus triggering a vicious cycle of diabetic feedback. Depending on the severity of the disease, the disease may be treated with a healthier diet and/or insulin sensitizing medication. Eventually insulin injections may also be required.

Even though the age at diabetes onset, severity of the symptoms, treatment, and the disease mechanisms behind T1D and T2D differ, the separation between the two forms of diabetes is not always clear. The gold standard for the diagnosis of T1D is based on the measurement of autoantibodies against islet cell antigens, and additional information may be obtained by measuring autoantibodies to insulin, glutamic acid decarboxylase, tyrosine phosphatase IA-2 and cytoplasmic islet cell antibodies [Seissler and Scherbaum 2006]. For diagnostic purposes, the amount of pancreatic insulin production may be assessed by the measurement of the serum C-

peptide concentrations. C-peptide is a part of an insulin precursor molecule, and is cleaved to form insulin [Steiner 1967]. As the patients with T1D do not have any insulin production at all, they do not have circulating C-peptide either. On the contrary, patients with T2D initially have normal or even elevated levels of C-peptide, even though the production of insulin and C-peptide may decline in advanced T2D. Furthermore, the distribution of the major genetic factors for T1D, so called human leukocyte antigen (HLA) haplotypes, differs between the patients with T1D, and patents with latent autoimmune diabetes of adults (LADA; a slow-onset type 1 autoimmune diabetes in adults) or with T2D. In research, T1D is often defined based on the age at diabetes onset and commencing permanent insulin injections within 1 year of the diagnosis [Mueller 2006, Thorn 2007, Tarnow 2008].

2.2 Diabetic nephropathy

The acute symptoms of diabetes – unconsciousness due to extremely high or low blood glucose levels – can be avoided by careful regulation of blood glucose levels. However, and despite the constantly improving care, one third of the patients with T1D develop severe long term micro- and macrovascular complications affecting the kidneys ("nephropathy"), eyes ("retinopathy"), nervous system ("neuropathy") and the cardiovascular system [Pambianco 2006]. Diabetic nephropathy (DN) is the most devastating microvascular complication. DN also increases the risk of cardiovascular disease with poor prognosis [Borch-Johnsen and Kreiner 1987, Tuomilehto 1998] and the most severe form of DN, the end stage renal disease (ESRD), is associated with an 18-fold risk of dying compared with the non-diabetic individuals of the same age [Groop 2009]. Healthy kidneys filter more than one liter of blood per minute. Their main function is to filter toxins from the blood and to maintain the body homeostasis by regulating the level of water, electrolytes and products of protein metabolism that must be eliminated from the body (Figure 1). Water and small molecules are filtered from the blood into the primary urine in the cortical part of the nephrons, in the glomeruli. The surface of the glomerular vessels has a three-layer glomerular filtration barrier that consists of endothelial cells, the glomerular basement membrane, and podocyte foot processes. The filtration barrier blocks nearly all negatively charged proteins or other macromolecules from entering the urinary space in healthy individuals. Water and important micro- and macromolecules are then reabsorbed from the primary urine into the blood in the tubuli [Sircar 2008].

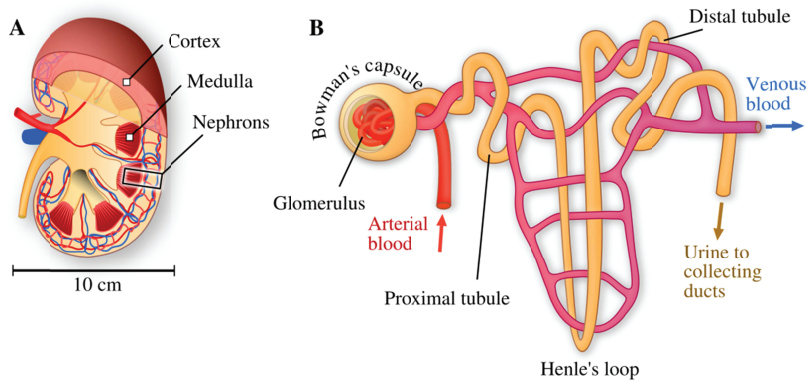


Figure 1: A: The structure of a kidney. The filtering takes place in the nephrons. B: A nephron. Primary urine is filtered from the blood in the glomeruli, and reabsorbed to the blood vessels in the tubuli. Figure is modified from [Mäkinen 2010].

DN is characterized by an increased excretion of albumin into the urine and a relentless decline in kidney function. At the earlier stages of DN, proteins – especially albumin as the most abundant protein in the circulation – start to leak through the glomerular filtration barrier. Urinary albumin excretion rate (AER) gradually increases from normal range first to microalbuminuria and then to macroalbuminuria (Figure 2). Pathogenically, this is seen as thickening of the glomerular basement membrane and the loss of podocytes. Kidney function, measured as estimated glomerular filtration rate (eGFR) may first even increase, but starts to decline as the disease leads to glomerulosclerosis i.e. scarring and overt structural lesions of the glomeruli until the filtration rate of the kidneys is close to

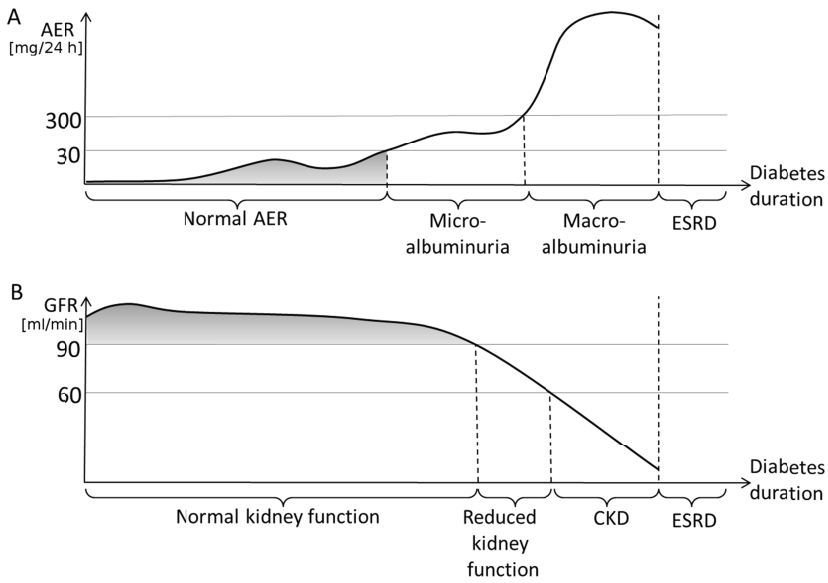


Figure 2: Schematic illustration of DN that typically starts with microalbuminuria. A) DN classification is based on AER. B) Chronic kidney disease (CKD) is diagnosed based on the kidneyfunction, measured withglomerular filtration rate (GFR).

zero [Forbes and Cooper 2013]. Finally, ESRD is reached when the kidneys can no longer filter the blood at sufficient level, but the patient requires regular dialysis treatment or a kidney transplantation for survival.

Early studies from the 80's reported that approximately 40% of the patients with T1D develop DN with a peak incidence after 13-18 years of diabetes duration [Andersen 1983, Borch-Johnsen 1985]. More recent studies have reported lower prevalence of DN between 13-32% after 20-25 years of diabetes duration [Hovind 2003, Nordwall 2004, Pambianco 2006], and the previously evident incidence peak has been changed to a relatively constant incidence reported between 20 and 34 years of diabetes duration [Pambianco 2006]. Nevertheless, recent studies in the Finnish population report that 8-23% of the patients with T1D still develop ESRD after 30-40 years of diabetes duration [Finne 2005, Harjutsalo 2011]. Therefore, it remains unclear if the decrease in the cumulative incidence of DN and ESRD is due to successful prevention of DN, or merely a delay of the disease onset [Marshall 2012].

In Europe, the clinical diagnosis of DN is often based on AER, which is an earlier marker of kidney damage than the eGFR (Figure 2; [Mogensen 1985]). Because of the relatively large daily variability of the urinary albumin excretion, the clinical definition of diabetic nephropathy requires that two out of three consecutive measurements surpass a given threshold. AER measured from an overnight timed urine collection is considered the gold standard, but albuminuria can also be defined based on a 24-hour sample or albumin-creatinine-ratio (ACR). The main variables employed for the diagnosis are summarized in Table 1 [Viberti 1994].

Persistent hyperglycemia – reflected by a high proportion of glycosylated hemoglobin (HbA_{1C}) – plays a central role in the development of DN. Strict glycemic control has been shown to decrease the occurrence of DN by 54% in the Diabetes Control and Complications Trial (DCCT) [Reichard 1993]. Importantly, the subsequent follow-up study (The Epidemiology of Diabetes Interventions and Complications, EDIC) showed a persistent effect of the intensive diabetes treatment

Table 1: Diagnostic thresholds for different stages of DN [Viberti 1994]

Diagnosis	24-h AER	overnight AER	ACR
Normal AER, no kidney disease	<30 mg/24 h	<20 µg/min	ACR <2.5 mg/mmol for men and <3.5 mg/mmol for women
Microalbuminuria	≥30 but <300 mg/24 h	≥20 but <200 µg/min	2.5-25 mg/mmol for men and 3.5-35 mg/mmol for women
Macroalbuminuria	≥300 mg/24 h	≥200 µg/min	≥25 mg/mmol for men and ≥35 mg/mmol for women
ESRD	Diagnosis based on commencing chronic dialysis treatment and/or subsequent kidney transplantation.		

on the risk of albuminuria eight years after ending the DCCT intervention [Writing Team for the DCCT/EDIC Research Group 2003]. However, even in the intensively treated group the cumulative incidence of DN was 9% after 30 years of T1D [Nathan 2009], suggesting that sustained hyperglycemia is not the only risk factor for DN.

High blood pressure is another strong risk factor for DN. The strong correlation between blood pressure and urinary AER was documented in the 1980's [Wiseman 1984, Berglund 1987]. Randomized controlled trials in subjects with T1D indicated that anti-hypertensive (AHT) medication lowers the level of AER and prevents progression from incipient to overt DN [Marre 1988, Mathiesen 1991]. On the contrary, the decline in GFR was reduced but not halted [Parving 1988]. Later studies have suggested that the AHT medication that affect the renin – angiotensin – aldosterone system, especially the angiotensin converting enzyme (ACE) inhibitors [Lewis 1993] and angiotensin II receptor blockers (ARBs) [Brenner 2001], are superior in treatment of DN compared with other groups of AHT medication, and may have renoprotective effects beyond lowering blood pressure.

In the healthy non-diabetic subjects, the male gender is a strong risk factor for ESRD, whereas women seem to be protected from ESRD at least until their menopause [U.S. Renal Data System 2011]. Similarly, the male gender is a risk factor for DN and ESRD among patients with T1D [Harjutsalo 2011]. However, among the men and women who have diabetes diagnosed before 10 years of age, no gender difference is seen in the incidence of ESRD [Harjutsalo 2011], suggesting that the usual female protection from ESRD is attenuated in women with an early onset of T1D. Some of the risk factors for DN are different for men and women [Coonrod 1993], which may indicate gender-specific mechanisms for DN or may reflect differences in the typical metabolic or hormonal profiles. Of note, estrogen has renoprotective effects in animal models [Silbiger and Neugarten 2008], whereas women with T1D have lower estradiol concentrations than non-diabetic women [Salonia 2006]. However, the role of estrogen in the progression of DN still remains unclear [Doublier 2011]. Overall, the loss of female protection in diabetes, seen as the absence of gender difference, remains controversial [Maric and Sullivan 2008, Silbiger and Neugarten 2008].

The first evidence that genes can affect the risk of DN was found in the late 80's by Seaquist *et al.* when they detected that diabetic nephropathy clustered in families [Seaquist 1989]. They studied sibling pairs where both had T1D, and compared the probability of the second sibling having DN when the first studied sibling ("proband") either had or did not have DN. In the families where the proband had ESRD, 24 out of 26 diabetic siblings (83%) had DN. In contrast, if the proband did not have DN, only two out of 11 diabetic siblings (17%) had DN. Borch-Johnsen reported similar familial clustering in Danish families [Borch-Johnsen 1992]. The difficulty of the sibling studies is reflected in the fact that out of the 619 T1D patients that they considered for the study, they identified only 24 patients with and 34 patients without DN having diabetic siblings. Familial clustering was reported also

for microalbuminuria in the DCCT study with 114 probands of which 13 had microalbuminuria: microalbuminuria was more than twice as common in the diabetic relatives if the proband had microalbuminuria (61%) versus if the proband did not have microalbuminuria (28%) [The DCCT Research Group 1997].

Later epidemiological study of a population-based cohort of Finnish T1D patients identified 537 T1D probands with 616 T1D siblings and showed that the presence of DN in one sibling more than doubles the risk of DN in the other diabetic siblings [Harjutsalo 2004]. Further epidemiological evidence of familial clustering comes from the studies showing that parental T2D, hypertension and cardiovascular disease are associated with increased susceptibility to DN in patients with T1D [Fagerudd 1999, Thorn 2007]. Even though the observed familial clustering may to some extent be due to more similar lifestyles, eating and smoking habits, and glycemic control among the diabetic siblings than among unrelated patients, the researchers conclude that genetic factors are likely involved in the development of DN.

3 Genetics

3.1 Human genome

The genetic material containing the instructions to build all the living organisms is stored and passed forward in deoxyribose nucleic acid (DNA). DNA consists of two long chains that run in opposite directions and form a double-helical structure. Each chain is formed by alternating deoxyribose (sugar) groups and phosphate diester groups. Attached to each deoxyribose group, there is either an adenine (A), guanine (G), thymine (T) or cytosine (C) nucleobase group. Together, a nucleobase, deoxyribose, and a phosphate group compose a nucleotide. The nucleobases are located in the middle of the helical structure and they pair with the nucleobases of the opposite chain in a special manner: adenine with thymine, and guanine with cytosine. The order of these four nucleobases contains the genetic code [Watson and Crick 1953].

Most of the DNA is packed into chromosomes in cell nuclei. In humans there are 22 autosomal chromosome pairs and one sex-determining pair. These 23 chromosomes contain approximately 3 billion base pairs [Bentley 2000]. Each individual carries two copies of each chromosome, one inherited from the mother, and one from the father.

The human genome contains 20,000 – 25,000 genes, which can be transcribed into messenger ribonucleic acid (mRNA) that is further translated into proteins. The genes consist of exons that encode the amino acid chains of the proteins and of introns and other untranslated regions that are removed in the splicing process of the mRNA. The protein coding exons compose only ~1.2% of the human genome, whereas 24% is spanned by introns. The remaining 75% of the genome is intergenic and its function remains poorly understood – although, at least part of it is assumed to have a regulatory role [Venter 2001, International Human Genome Sequencing Consortium 2004, ENCODE Project Consortium 2012].

3.1.1 Genetic variation

The vast majority of the genome is identical between any two individuals, but many forms of genetic variation exist. The changes may be small mutations that affect only one or a few base pairs, or they may be large structural changes in the chromosomal structure, observable with a microscope.

While one copy of each chromosome is inherited from the father, and one form the mother, these chromosomes are not passed from one generation to the next as

such. The contents of the maternal and the paternal chromosome pairs are shuffled during the meiosis (the formation of egg and sperm cells) in the so called recombination event, when similar DNA sequences from the paired chromosomes cross over each other. Errors in the recombination may result in large structural changes such as deletions, duplications, inversions, substitutions and translocations (Figure 3).

Small mutations may be introduced to the genetic code at any time when DNA is replicated or otherwise processed. These mutations include substitution of a base pair with another, or insertions or deletions of one or more base pairs (Figure 3). The mutations affect the following generation only when they happen in the germline.

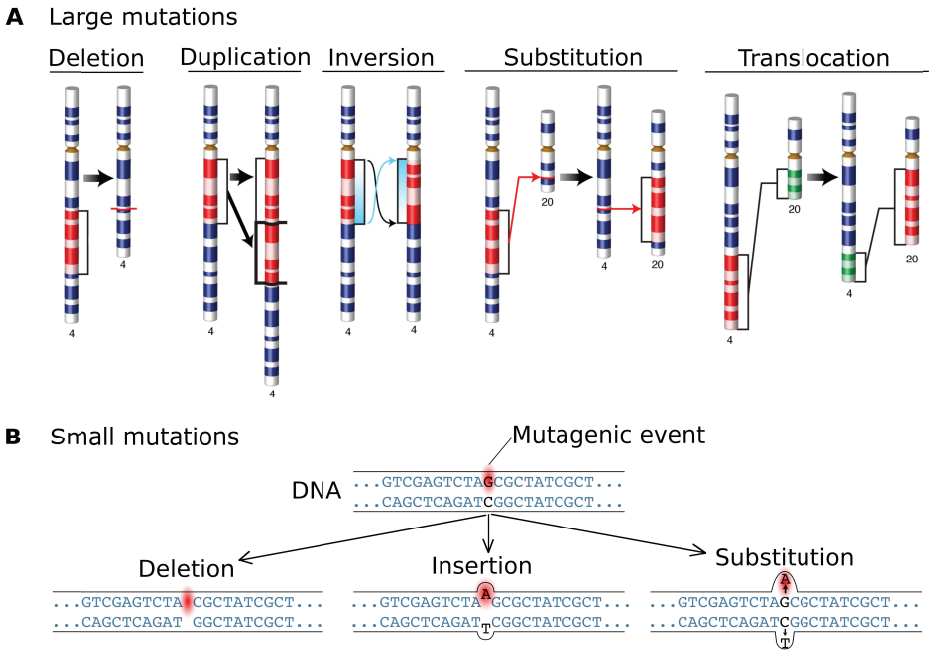


Figure 3: Different types of possible large (A) and small (B) mutations. Modified from www.genome.gov/glossary/

The most studied mutations are the single nucleotide polymorphisms (SNPs), i.e. the substitutions of one base pair in the DNA sequence (Figure 4). The different possible versions of the genotype on the SNP locus are called alleles. Mutations of one nucleotide are estimated to occur at an average rate of 10^{-8} per base pair per generation [International HapMap Consortium 2005]. The International HapMap Consortium was designed to study the common variation, and in particular the SNPs in the human genome. The first phase of the effort, published in 2005, identified 1.2 million SNPs found within 269 DNA samples from 90 individuals from Yoruba in Ibadan, Nigeria (abbreviation YRI), 90 individuals from Utah, USA, from the Centre d'Etude du Polymorphisme Humain collection (abbreviation CEU), 45 Han Chinese individuals from Beijing, China (abbreviation CHB), and 44 Japanese individuals

from Tokyo, Japan (abbreviation JPT). Among the 1.2 million SNPs, 880,000 SNPs were considered common, with the minor allele frequency (MAF) ≥ 5% [International HapMap Consortium 2005]. Two years later, the phase II of the HapMap project reported an improved map of the human genome, with over 3.1 million SNPs identified and genotyped in the same samples [International HapMap Consortium 2007].

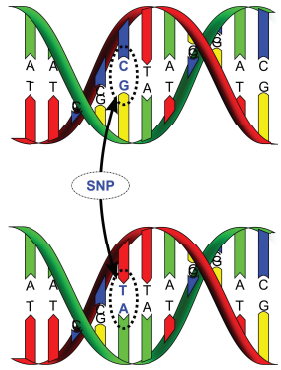


Figure 4: Single nucleotide polymorphism (SNP). An individual has otherwise two identical copies of the chromosomal region (one from the father, one from the mother), but the two DNA sequences differ by one base pair, indicated with the arrows. If the green strand is the forward strand, then the SNP is said to have either A or G allele, and each individual can carry either the AA, AG, or GG genotype for that SNP. Modified from http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

3.1.2 Linkage disequilibrium

When new mutations arise, they initially stay with the same surrounding DNA sequence from one generation to another, until a recombination event happens in the sequence region and the correlation between the variants starts to decay. The further away the two variants are from each other, the more likely it is that there has been a recombination event somewhere between the two loci after the introduction of the mutation. However, the recombination events do not occur at uniform rate across the genome, and thus, the occurrence of variants that are located close to each other is often correlated between individuals. This statistical association between alleles of two or more SNPs is known as linkage disequilibrium (LD). The sequence region that is inherited in one piece without recombination events is called a haplotype. For two variants with alleles a/A and b/B, the disequilibrium D is defined as the departure from the expected frequency distribution assuming independence of the two variants:

$$D = f_{AB} - f_A \times f_B$$

where f_{AB} is the frequency of individuals carrying the alleles A and B, and f_A and f_B are the frequencies of those alleles.

In practice, D is not often used as a measure of LD because the maximum value of D depends on the allele frequencies. In addition, the sign of D is arbitrary depending on which pair of alleles is studied. The two main measures of pairwise LD are D' and

r^2. D' ("D prime") is the absolute value of D divided by the maximum value that D can have given the allele frequencies: [Hedrick 1987]

$$D' = |D|/D_{max}$$

$$D_{max} = \begin{cases} \min\{f_A \times f_B, (1-f_A) \times (1-f_B)\}, & when\ D < 0 \\ \min\{f_A \times (1-f_B), f_B \times (1-f_A)\}, & when\ D > 0 \end{cases}$$

D' = 1 indicates complete LD, i.e. no recombination or recurrent mutation has occurred between the SNPs. D' = 0 indicates that the two SNPs are in complete linkage equilibrium, e.g. they segregate independently of each other. Another commonly used measure of LD is r^2, which is the squared correlation coefficient between two SNPs:

$$r^2 = \frac{D^2}{f_A(1-f_A) \times f_B(1-f_B)}$$

Unlike D, both D' and r^2 can reach values near one even if one or both variants have low allele frequency [Reich 2001, Slatkin 2008].

Figure 5 illustrates a genealogical tree on the left and five possible haplotypes on the right side, encompassing 13 SNPs (colored dots) that are colored according to the ancestral branch where the mutations arose. No recombination has occurred on the depicted region, and thus, all the SNP pairs (for example, the yellow SNP 6 and the purple SNPs 5 and 12) have D' = 1. r^2 = 1 only when the SNPs are on the same branch of the genealogy (of the same color in Figure 5) and undisrupted by recombination. Apart from the SNPs on the same branch, the pairwise r^2 in the illustration ranges from 0.05 between the yellow and blue SNPs, to 0.82 between the red and purple SNPs. Of note, r^2 does not depend on the physical distance on the short range. Using $r^2 \geq 0.8$ threshold for tagging SNPs, the genetic variation of the 13 illustrated SNPs can be captured with four SNPs, for example SNP1, SNP2, SNP3 and SNP6 [International HapMap Consortium 2005].

One important finding from the HapMap project was the realization how

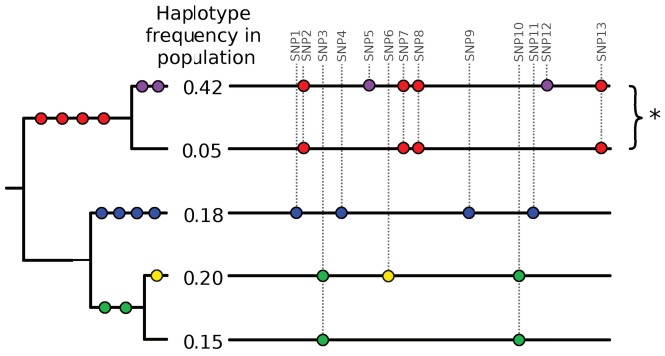


*Figure 5: A genealogical tree on the left and the five corresponding haplotypes on the right side. The sequence region contains 13 SNPs (colored dots), colored according to the ancestral branch where the mutations arose. D' = 1 between all the 13 SNPs. The SNPs that arose on the same branch of genealogy have r^2 = 1.*The purple and red SNPs have r^2 > 0.8: frequency of purple SNPs = 0.42, frequency of red SNPs = 0.42+0.05 = 0.47; $r^2 = (0.42-0.42\times0.47)^2/(0.42\times(1-0.47)\times0.47\times(1-0.42))=0.82$.*

extensive the LD is in the human genome. Based on computer simulations, the LD was expected to extend only a few kilobases (kb, 1,000 bases) [Kruglyak 1999]. However, the recombination events are concentrated in recombination hot spots (typically spanning 2 kb) where the recombination occurs much more often than in the surrounding region. The LD thus remains high between the recombination hotspots, and LD may extend hundreds of kb [Reich 2001, International HapMap Consortium 2005].

3.2 Genetic variation of diseases

The differences between individuals are due to differences in the genetic background or to environmental exposure. Similarly, the diseases may be caused by genetic factors, the environment, or both. The different modes of genetic background are illustrated in Figure 6. On the top left corner of the illustration, the monogenic "Mendelian diseases", named after Gregor Johann Mendel, are fully inherited. In these rare diseases one mutation is not only required but also sufficient to cause the disease. An example of such a monogenic disease is cystic fibrosis, with 5% of the population in Europe and 1% in Finland carrying one of the many disease causing mutations in the *CFTR* gene [Halme and Kajosaari 2006].

However, genetic variation also contributes to the risk of many common diseases, such as diabetes (both T1D and T2D), coronary artery disease, Crohn's disease, rheumatoid arthritis, bipolar disorder, and many more. These complex diseases have many common genetic susceptibility variants, each of which moderately increase the risk of the disease in addition to the effect imposed by the environmental factors (bottom right corner in Figure 6) [Wellcome Trust Case Control Consortium 2007]. Thus, an unfavorable collection of inherited high-risk variants does not alone cause a common disease, and conversely, one can develop the disease even without the genetic risk variants.

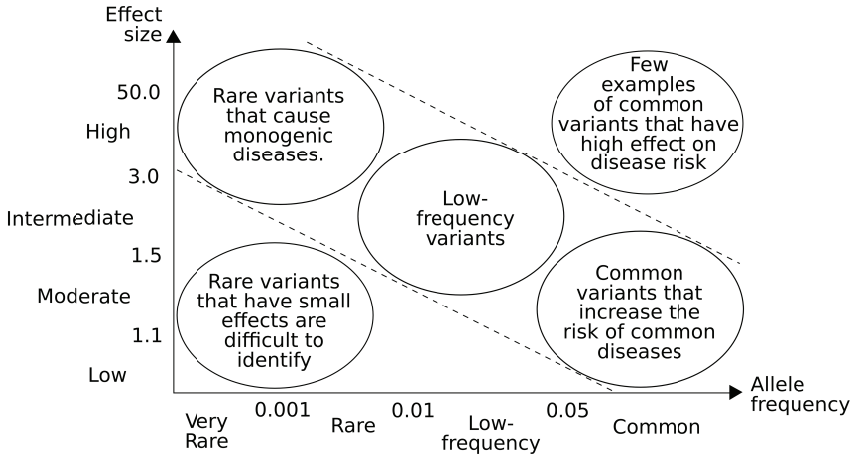


Figure 6: The monogenic diseases are often caused by rare genetic variants, whereas the common diseases have multiple common SNPs that moderately increase the risk of disease. Modified from [McCarthy 2008].

Rare variants with only a minor effect on disease development may exist, but these are difficult to identify as large number of samples would be required to detect such factors. On the other hand, common variants with large effect size are highly unlikely for any disease, as that would lead to a large proportion of the population affected by the disease [McCarthy 2008]. Outside the disease domain, an example of such case is the common variant (MAF=0.35 in the 1000 Genomes project [1000 Genomes Project Consortium 2012]) that causes the "O" blood group in the ABO blood histo-group system.

3.2.1 Disease heritability

The heritability of a disease or a phenotype (i.e. the observable characteristics or traits of an organism) refers to the proportion of the phenotype variance that is due to the genetic variation. Formally, the phenotype (P) of interest can be partitioned to contribution of genetic (G) and environmental (E) factors, and the variance of the phenotype is the sum of the genetic and environmental variances:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

The genetic variance can be further divided into additive genetic effects σ_A^2, dominance genetic effects σ_D^2 (i.e. interactions between the alleles at the same locus), and epistatic genetic effects σ_I^2 (i.e. interactions between alleles between different loci). The broad-sense heritability (H^2) takes into consideration the total genetic effects, whereas the more commonly used narrow-sense heritability (h^2, often called heritability) takes only the additive genetic effects into account, defined as the proportion of additive genetic variance from the total phenotypic variance: [Visscher 2008]

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}$$

Heritability has been traditionally estimated by comparison of phenotypes in families, for example with regression analysis of parental and offspring phenotypes, or inferring the phenotype differences in monozygotic and dizygotic twin pairs.

The level of heritability can be assessed also for dichotomous disease traits [Visscher 2008]. Sibling recurrence-risk ratio λ_S is a commonly used epidemiological parameter to determine the familial aggregation of a complex disease. λ_S is defined as the ratio of disease manifestation (D), given that one's sibling is affected, compared with disease prevalence (p) in the general population:

$$\lambda_S = \frac{P(D|sib(D))}{p}$$

The population prevalence p is often determined as the rate of disease manifestation given that one's sibling is not affected.

With the current genome-wide genotyping methods, the heritability of a trait can be inferred from a large group of individuals: Even though most genome-wide genotyping efforts include non-related individuals (close relatives removed), the genetic data can be used to estimate the residual relatedness between each pair of

individuals. This relatedness structure can then be correlated with the phenotypic variation, resulting in an estimate of the narrow-sense heritability observable with the applied genotyping platform [Yang 2011].

3.3 Search for genetic factors behind the common diseases

The assumed mode of heritability is a key factor affecting the study design when searching for genetic factors behind a disease. For rare Mendelian diseases and rare monogenic forms of common diseases, the best option is to study the affected families and narrow down the possible genetic region by searching for the regions that co-segregate with the disease in the families. For complex diseases that have both a genetic and an environmental background (i.e. $h^2<1$), this is often not feasible, since not all carriers of a specific genetic risk factor will develop the disease, and not all subjects affected by the disease carry the same genetic risk factors.

3.3.1 Association studies of candidate genes

The first step to detect the genes that predispose to complex diseases was to perform candidate gene association studies by genotyping SNPs and other genetic variants (i.e. defining the genotype of a given variant from a DNA sample) in patients with and without the disease and by comparing the genotype frequencies between the affected and non-affected subjects. As genotyping was relatively expensive, time-consuming and laborious, these genetic association studies were concentrated on the genes that were assumed to participate in the biological processes and genetic pathways affecting the disease process. A good example of a successful candidate gene study was the identification of a common amino acid substitution in the PPARγ gene, associated with T2D and extensively replicated in other studies [Beamer 1998].

With more and more performed genetic association studies, many variants were found associated with various common diseases. However, conclusive and repeated replication of the candidate genes was challenging, with some of the other study populations supporting and the others refuting the findings. The lack of replication may have been due to a small number of patients and low effect sizes of the associations that result in low statistical power to replicate, or discrepancies in the studied clinical traits. However, many of the reported but non-replicated initial findings may have been false positive findings due to poor study design or over-estimation of the association [NCI-NHGRI Working Group on Replication in Association Studies 2007]. Another drawback of the candidate gene approach is that the research efforts were limited to the existing hypotheses of the biology, and thus they restrain from finding truly new genetic pathways [Doria 2008].

3.3.2 Family-based linkage analysis

As the candidate gene studies did not explain the expected heritability of the common diseases, the question arose whether the susceptibility genes would be

genes with an unknown function, or genes that were not expected to affect the disease. Genome-wide family based linkage studies were launched with the aim to detect new, unforeseen chromosomal regions harboring genes for common diseases. By genotyping hundreds of markers across all the chromosomes in the family-based linkage studies, one can detect chromosomal regions that are inherited from one generation to the next one, together with the disease. Similarly, comparison of markers that are shared by siblings aims to identify chromosomal regions that are co-segregated with the disease. The strength of the linkage is typically assessed with the logarithm of odds (LOD) score for the co-segregation taking place due to true linkage between the locus and the disease, versus due to chance. LOD score >3 is generally considered significant [Strachan and Read 1999].

However, the family based linkage studies have not revealed much of the genetic background of the common and complex diseases with heterogeneous genetic effects. The linkage studies often suffer from low regional resolution as only a few cross-over events take place within each studied family, and thus, large chunks of genome around the causal variant are co-segregated in the families. Moreover, the recruitment of an adequate number of families with multiple disease cases is challenging. The linkage studies are best suited for diseases with very strong familial segregation and rare genetic variants (Figure 6).

3.4 Genome-wide association studies (GWASs)

The association studies of unrelated individuals are better suited for the analysis of complex diseases than the linkage studies. The GWASs are based on the assumption that many common variants contribute to the common disease, but the effect of each individual variant on the phenotype is moderate [Balding 2006, McCarthy 2008]. Many successful examples exist where GWASs have identified novel genetic susceptibility variants for common diseases such as diabetes (both T1D and T2D), coronary artery disease, Crohn's disease, rheumatoid arthritis and bipolar disorder [Wellcome Trust Case Control Consortium 2007]. The association tests have higher statistical power to detect disease predisposing variants with moderate relative risk than the family based methods [Risch and Merikangas 1996, Botstein and Risch 2003]. The drawback of the association studies is that they require a substantial number of patients, normally several thousands, to detect the modest effect sizes. Nevertheless, the recruitment of unrelated individuals is easier than the identification of families with many patients. In addition, association studies allow higher regional resolution than family based linkage studies, thanks to numerous cross-over events in the population level that break the genetic linkage to smaller parts [Balding 2006].

Characterization of the common variation in the human genome by the HapMap project [Frazer 2007] and recent advances in genotyping technology made the association studies feasible on a genome-wide scale. Whereas the price of

sequencing one base pair was 20-30$ in the 1990's [Carlson 2003] and the first association studies included hundreds of patients [Groop 1993], currently the genome-wide association studies (GWASs) include thousands of patients genotyped at 300,000 to 1,000,000 genetic markers that cover the majority of the human genome, with the price per marker less than 0.001€.

Compared with the candidate gene association testing, the GWAS allows searching without prior knowledge of the underlying genes. One consequence of not limiting the search space is that the vast majority of the findings are located outside the genes, and only 5% of them are on protein coding exon regions [Schaub 2012]. In addition, the nearest – or even the underlying gene – is not necessarily the causal gene [Smemo 2014], and regulatory regions may affect expression of genes that are located far away [Sanyal 2012]. Furthermore, the associated LD blocks may span hundreds of kilobases making it difficult to decide which of the associated SNPs, if any, is the causal one [Reich 2001, International HapMap Consortium 2005]. For convenience, the associated loci are often named according to the nearest gene.

3.4.1 Association testing

In a typical GWAS the significance of each SNP is evaluated separately by computing a statistical association model that correlates the SNP genotypes to the observed phenotype. Commonly used models are for example Pearson goodness-of-fit test (often known as χ^2 test) or logistic regression. In the same way, GWASs can be used to detect genetic risk factors for increasing continuous traits, typically using linear regression [Balding 2006]. The statistical power to detect associations can be further improved by combining the data from multiple GWASs in meta-analyses that can include millions of SNPs analyzed in tens of thousands of individuals as for T2D [Voight 2010]. Among the most studied continuous traits are the obesity-related traits such as body mass index (BMI), for which 32 susceptibility loci were identified in a meta-analysis of nearly 250,000 individuals [Speliotes 2010].

3.4.2 Multiple testing and interpretation of *P*-values

The analysis of millions of markers and thus millions of separate study hypotheses creates a fundamental problem of multiple testing: The significance level α can be interpreted as the greatest tolerated probability of type I error (false positive). In practice, using α=5% corresponding to a *P*-value threshold of 0.05 would allow five false positive associations if 100 hypotheses (SNPs in this case) were tested. In a GWAS of 1,000,000 tested SNPs, the nominally significant *P*-value threshold of 0.05 can be expected to return 0.05×1,000,000 = 50,000 false positive associations. Therefore, much stricter *P*-value thresholds are commonly used for the GWASs. Bonferroni correction ($\alpha_{\text{test-wise}} = \alpha_{\text{study-wise}}/n_{\text{tests}}$) is often performed to account for multiple testing, but it may be overly conservative especially when the markers are in high LD with each other [Balding 2006]. In the GWAS setting, a *P*-value threshold of 5×10^{-8} has become a commonly accepted limit for genome-wide statistical significance, corresponding to Bonferroni adjustment for the rough estimate of 1

million independent SNPs in European population ($0.05/1{,}000{,}000 = 5\times10^{-8}$) [McCarthy 2008].

3.4.3 Quality control of genome-wide genotyping

Given the vast number of SNPs, the genotyping of the high-throughput data has to be automated, and careful quality control is essential to avoid any spurious findings due to genotyping errors or systematic bias. Any suspicious SNPs or patient samples should be discarded. The SNPs and samples should have high genotyping success rate and the genotype calls of each SNP should be clearly separable from each other. Hardy-Weinberg equilibrium (HWE) refers to the state where the two alleles of a SNP segregate independently of each other in the genotypes. In the evolution theory, deviation from HWE indicates inbreeding, population stratification or selection, but in the genotyping context, it is more often a sign of failed genotype clustering where the heterozygous genotypes are miscalled as homozygous, or *vice versa* [Balding 2006]. Other typical filters include removal of closely related individuals and testing for differences in genotyping that are correlated with the genotyping plates.

The population structure is an important source of bias and has to be carefully taken into account. The allele frequencies vary between different populations. If the cases and controls show different proportions of subjects from different populations, any SNP with a different allele frequency between these populations would show evidence of association. For example, variants in the lactase gene (*LCT*) and variants on the HLA region have been shown to correlate with the north - south and west - east axes of the place of origin for the European subjects [Heath 2008]. Furthermore, the first two principal components (PCs) based on GWAS genotypes have been shown to correlate with the geographical location of the European subjects [Novembre 2008]. Even the Finns that are commonly considered a homogenous population show a clear East – West difference when using a similar approach [Salmela 2008]. While the best practice is to include in the analysis only subjects from one population and to remove any samples with different population background, the principal components can be used to adjust the analysis for any remaining population substructure [Price 2006].

3.4.4 Genotype imputation

Different commercial platforms for the genotyping of the GWASs typically address 300,000 to 1,000,000 SNPs. The number of the examined SNPs can be computationally enhanced by *in silico* prediction of additional genotypes that are not assayed on the original platform, based on a more densely genotyped reference population. This process is called genotype imputation. The aim of imputation is to boost the statistical power, to fine-map associated regions, and to facilitate meta-analysis of GWASs that are often genotyped using different genotyping platforms with different set of assayed SNPs [Marchini and Howie 2010]. Currently the most commonly used reference panel is based on the HapMap II population that identified more than three million SNPs [International HapMap Consortium 2007]

in 120, 120 and 180 CEU, YRI, and JPT+CHB subjects, respectively. However, even larger reference panels are now available with genotype information for more than 1,000 individuals of diverse origin [1000 Genomes Project Consortium 2012].

Most of the currently used imputation methods are based on hidden Markov models (HMMs) that model the sequence of genotypes G_i in each individual based on the known haploid genotype sequences in the reference population, so called phased haplotypes. In essence, given the reference haplotypes *H,* the genotype probability can be given as

$$P(G_i|\, H, \mu, \theta) = \sum_Z P(G_i|Z, \mu) \times P(Z|H, \theta)$$

where the hidden stages Z can be thought as pairs of haplotypes from the reference panel. $P(Z \mid H, \theta)$ allows switching from one haplotype to another according to the local cross-over parameter θ, and $P(G_i \mid Z, \mu)$ allows differences from the reference based on the mutation parameter μ [Marchini and Howie 2010]. Different software use various implementations to fit the models, for example Markov chains applied in MaCH 1.0 [Li 2009, Li 2010] , Markov chain Monte Carlo (MCMC) approach used in IMPUTE v2 [Howie 2009], and expectation maximization algorithm in fastPHASE [Scheet and Stephens 2006].

The imputation accuracy is affected by multiple factors, such as the SNP allele frequency, size of the reference panel and its similarity with the study population, and the selection of the SNPs on the genotyping chip. The error rates for the most likely genotypes ("best guess genotypes") are typically 5-6% depending on the method and data [Marchini and Howie 2010], but the direct use of the most likely genotypes is not advised. Instead, the uncertainty of the imputation quality should be taken into account in the analysis, either by weighting the contribution of each possible genotype by its imputation probability, implemented for example with a score test in the SNPTEST software [Marchini and Howie 2010], or by using the expected genotype counts, also called posterior mean genotypes or allele dosages [Guan and Stephens 2008].

3.4.5 Data mining methods for genome-wide studies

Owing to the vast number of SNPs in a typical GWAS, the commonly used GWAS methods rely on single-SNP tests performed with simple statistical models. However, some steps have been taken to apply more sophisticated statistical techniques.

Bayesian single-SNP approaches: The common ("frequentist") way of assessing association for each SNP is based on calculating a *P*-value to obtain the observed results under the null hypothesis H_o of no association. The use of a *P*-value has been criticized as the same *P*-value can have different implications, depending on the factors that affect the statistical power of the test, such as the number of the studied SNPs, the size of the study, and the MAF of the SNP: for example, a *P*-value of 0.001 can be considered significant in a candidate gene study including only a few SNPs, whereas a *P*-value of 10^{-6} is deemed only as suggestive evidence of association in a

GWAS due to the burden of multiple testing (See 3.4.2). On the other hand, obtaining a significant association with $P < 5\times10^{-8}$ for a SNP with low power due to low MAF or small number of samples is so unlikely that the result should be regarded with caution despite the significant *P*-value. To overcome the limitations of the *P*-value based analyses, the Bayesian statistics have been suggested for the assessment of associations [Stephens and Balding 2009].

The Bayesian methods aim to define the posterior probability of an association (PPA), i.e. the probability that a SNP is truly associated with the phenotype, given the prior assumptions and the observed data. The PPA can be compared with the posterior probability of no association (PPnA) based on the model parameters θ, observed data y, and the posterior density $p(\theta|y)$ of the model parameters, as follows [Gelman 2004]:

$$\frac{PPA}{PPnA} = \frac{p(\theta_1|y)}{p(\theta_0|y)} = \frac{p(\theta_1)p(y|\theta_1)}{p(\theta_0)p(y|\theta_0)}$$

In the genetic association studies, the two sets of parameters θ_1 and θ_0 correspond to hypotheses H_o (no association) and H_1 (association). $p(\theta_1)$ is the prior probability of a SNP being associated with the phenotype, and is often interpreted as an estimate of the proportion of SNPs that are associated with a phenotype, π. Values from π = 1/1,000 to π = 1/100,000 have been suggested [Stephens and Balding 2009]. Finally, the Bayes factor (*BF*) is the ratio between the probabilities of the observed data under H_1 and H_0.

Using the above notation, the PPA can be calculated in two steps: first, posterior odds (*PO*) as in the equation above:

$$PO = \frac{PPA}{PPnA} = \frac{\pi}{1-\pi} \times BF$$

Then, the PPA [Stephens and Balding 2009],

$$PPA = \frac{PO}{1+PO}$$

In general, the ranking of SNPs is similar based on PPAs and *P*-values, except for SNPs with low MAF. The flexibility of the Bayesian models allows for simultaneous evaluation of additive (used in typical GWASs), dominant and recessive effects. In addition, the PPA can be directly interpreted as a probability, irrespective of the number of studied SNPs, statistical power, or sample size [Stephens and Balding 2009].

Multi-SNP methods for SNP discovery: The common, complex diseases are assumed to have multiple genetic risk factors. Ideally, all the SNPs should be analyzed simultaneously to better capture how they interact, or affect the disease given the existence of other risk factors: even a weak effect may become more apparent when the other risk factors are taken into account [Hoggart 2008]. Most of the current multimarker approaches rely on penalized (logistic) regression models [Hoggart 2008, Ayers and Cordell 2010, He and Lin 2011] or Bayesian analysis [Sebastiani 2008, Sambo 2012, Hartley 2012].

A machine learning method based on Naïve Bayes Classifiers, called Bag of Naïve Bayes (BoNB) [Sambo 2012], is presented in more detail in the Methods section. Naïve Bayes classifiers are a simple but efficient supervised classification method, derived from the Bayes' theorem assuming that all the factors contribute independently to the posterior probability. With this assumption, the posterior probability of a class variable C given the factors $F_1 \ldots F_n$ can be formulated as

$$p(C \mid F_1, \ldots, F_n) = \frac{1}{Z} \times p(C) \times \prod_{i=1}^{n} p(F_i \mid C) \quad ,$$

where Z is a constant depending only on the factors. For example, the probability of a subject being case or control based on three SNPs can be calculated as follows:

$$p(case|SNP_1, SNP_2, SNP_3) \propto p(case) \times p(SNP_1|case) \times p(SNP_2|case) \times p(SNP_3|case)$$

$$p(ctrl \mid SNP_1, SNP_2, SNP_3) \propto p(ctrl) \times p(SNP_1 \mid ctrl) \times p(SNP_2|ctrl) \times p(SNP_3 \mid ctrl)$$

The rationale behind the BoNB algorithm is i) to generate many slightly different datasets with Bootstrap resampling of the original patient set, ii) to create a Naïve Bayes classifier for each data set, and iii) to define the SNPs that are selected by multiple Naïve Bayes classifiers as significant ones. Bootstrapping, or repeated random sampling with replacement, divides the patients into a training set to select the SNPs and to learn the model, and into an independent test set ("out-of-bag set") that can be used to test the model performance (Figure 7). The out-of-bag sets are comparable to the use of independent replication cohorts in conventional GWAS studies. The Bootstrap sampling is repeated multiple times (typically ~100). The signals identified by such a procedure are robust to small changes in the patient set.

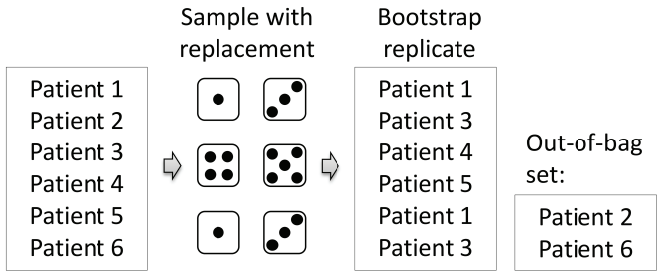


Figure 7: Illustration of sampling with replacement from a dataset of six patients using six rolls of dice. The random Bootstrap sample includes patients in the second list, whereas patients 2 and 6 form the out-of-bag set. Because of the "sample replacement" during the sampling, some patients are selected multiple times to the Bootstrap sample, see for example patients 1 and 3.

Multimarker enrichment analysis: The most commonly used multimarker analyses for GWASs are the gene set enrichment analyses. The significance of each SNP is evaluated with conventional single-SNP approaches. Based on the single-SNP results, gene set enrichment analysis can be used to infer genetic pathways or other defined gene sets that are enriched for significantly associated SNPs. This can give additional information on the affected biological processes, which can be difficult to detect when inspecting each SNP or gene separately. The gene set enrichment

analyses are performed either by selecting SNPs with a *P*-value below a pre-defined significance threshold, or by inferring the significance threshold based on the data.

4 Genetics of diabetic kidney disease

4.1 Candidate genes for diabetic kidney disease

The candidate gene studies of DN were based on functional candidates i.e. genes that were assumed to affect DN, or later on, positional candidates based on the genomic regions implied in the linkage studies. Some of the much-studied candidate genes and pathways are for example the angiotensin-converting enzyme (*ACE*) [Hadjadj 2007] and the angiotensin II receptor, type 1 gene (*AGTR1*) [Ding 2012] in the renin-angiotensin-aldosterone system, the aldose reductase (*AKR1B1*) as the first and the rate-limiting enzyme of the polyol pathway [Neamat-Allah 2001], the apolipoprotein E (*APOE*) that is part of the lipid metabolism [Araki 2000], the erythropoietin (*EPO*) involved in the angiogenesis [Tong 2008] and the nitric oxide synthase 3 (*NOS3*) affecting oxidative stress [McKnight 2010a].

To summarize all candidate gene association studies on DN, Mooyaart *et al.* reported a meta-analysis of reproduced DN associations found by a literature search [Mooyaart 2011]. Their meta-analysis included 132 publications describing 153 studies and resulted in 24 genetic variants in 17 distinct loci significantly associated with DN in T1D and/or T2D. Variants in *ACE*, *AKR1B1*, *APOC1*, *EPO*, *GREM1, HSPG2, UNC13B* and *VEGFA* were deemed significant also in patients with T1D alone (Table 2). Literature based meta-analyses, however, may suffer from publication bias since positive findings are easier to publish than negative reports, and thus, may give overly positive results. Therefore the role of these variants remains ambiguous.

4.2 Linkage studies for diabetic kidney disease

Genome-wide family based linkage studies were launched with the aim to detect new, unforeseen chromosomal regions harboring genes for DN. These studies were based on sibling pairs both with T1D, and either both affected with DN ("affected sib-pairs") or discordant for their DN status ("discordant sib-pairs") [Osterholm 2007, Rogus 2008, Wessman 2011]. All these genome-wide linkage scans suggestively support a linkage peak on chromosome 3q that was first reported for DN in T1D in a candidate gene linkage analysis of the *AGTR1* gene (genetic position 157cM on chromosome 3q; LOD score = 3.1; P=7.7×10^{-5}) [Moczulski 1998]. However, the reported linkage peak locations vary between the studies from 134 to 157cM, flanking the chromosomal region 3q21-25, and a subsequent fine-mapping

*Table 2: Candidate genes that have been associated with DN in T1D. *Significance results from a literature mining based meta-analysis by Mooyaart et al [Mooyaart 2011]. OR: odds ratio; 95% CI: 95% confidence interval*

Gene name and variant	Significance	Candidate pathway	Gene function regarding DN
ACE rs1799752	OR 1.13 (95% CI 1.04 – 1.23), 14 studies*	renin-angiotensin system	The plasma levels of angiotensin converting enzyme (ACE) are associated with DN. ACE inhibitors are a recommended treatment for DN [Lewis 1993].
AKR1B1 rs759853	OR 1.58 (95% CI 1.01-2.46), 4 studies*	polyol pathway	*AKR1B1* encodes aldose reductase which catalyses the reduction of a variety of carbonyl-containing compounds, e.g. glucose to sorbitol. Aldose reductase is a key enzyme of the polyol pathway [Neamat-Allah 2001].
APOC1 rs4420638	OR 1.54 (95% CI 1.29-1.83), 2 studies*	Lipid metabolism, cardio-vascular disease	*APOC1* encodes an apolipoprotein C1. Dyslipidemia is a risk factor for DN [McKnight 2009].
EPO rs1617640	OR 0.67 (95% CI 0.58-0.76) 2 studies*	angiogenesis	*EPO* encodes erythropoietin, which is a potent angiogenic factor expressed in both retina and kidney. Erythropoietin participates in the erythropoiesis and is used to treat anemia resulting from renal failure or chemotherapy [Tong 2008].
GREM1 rs1129456	OR 1.53 (95% CI 1.25-1.89), 2 studies*	Cell growth and differ-entiation	*GREM1* promotes the development of diabetic nephropathy in animal models [McKnight 2010b]
HSPG2 rs3767140	OR 0.64 (95% CI 0.49– 0.84)* from 2 studies	glomerular basement membrane	*HSPG2* encodes the perlecan protein, which is involved in maintenance of glomerular basement membrane electrostatic charge [Mooyaart 2011].
NCK1 rs1866813	P=7.07×10^{-6}, OR 1.33 (95% CI 1.17-1.51) from 3 studies	Region 3q21 – 3q25 implicated in linkage studies	*NCK1* encodes Nck1, involved in actin polymerization. In kidney podocytes Nck1 links nephrin, a key protein in the slit diaphragm, to the actin cytoskeleton [He 2009].
UNC13B rs13293564	OR 1.23 (95% CI 1.11-1.35), 4 studies*	apoptosis	*UNC13B* is activated and upregulated by hyperglycemia in renal cells [Tregouet 2008].
VEGFA rs833061	OR 0.48 (95% CI 0.37-0.61), 2 studies*	angiogenesis	Vasular endothelial growth factor (VEGF) is implicated in the pathogenesis of microvascular complications of diabetes [McKnight 2007]

association analysis of the affected region suggested the *NCK1* gene on chromosome 3q22 as the likely culprit behind the linkage peak (Table 2) [He 2009]. As the reported linkage peak locations vary substantially, it is also possible that these peaks represent different signals [Wessman 2011].

In addition, significant linkage has been reported for chromosomes 19q (maximum likelihood score (MLS) = 3.1) [Rogus 2008] and for 22q11 (LOD=3.58) [Wessman 2011], but no causal genetic variants affecting DN have been decisively

identified based on these linkage analyses. Linkage studies on albuminuria as a continuous trait have not been performed in patients with T1D.

4.3 GWAS on diabetic kidney disease

Prior to this thesis, only one GWAS had been published on DN in T1D, based on the Genetics of Kidneys in Diabetes US (GoKinD US) study including 1,500 T1D patients [Pezzolesi 2009a]. No locus reached genome-wide statistical significance ($P<5\times10^{-8}$) in their analysis, but they reported multiple suggestive associations ($P<10^{-5}$) on chromosomes 7p in the *CHN2* gene, on 9q near the *FRMD3* gene, on 11p in and near the *CARS* gene and on 13q on an intergenic region near the *IRS2* gene. Their subsequent analysis using imputed GWAS data resulted in four additional suggestively associated loci on chromosomes 10q at the *SORBS1* gene, on 8p near the *TRPS1* gene and between the *CDCA2* and *EBF2* genes, and on 10q near the *BUB3* and *GPR26* genes [Pezzolesi 2010]. Among these loci, the variants near the *CARS* gene and on the *FRMD3* gene were suggestively replicated in further studies with diabetic patients [Pezzolesi 2009a]. The function of these genes remain unknown, but the *FRMD3* gene is likely related to the maintenance of cellular shape and the gene is known to be expressed in kidneys as well [Pezzolesi 2009a].

A parallel GWAS on ESRD was performed in the same GoKinD US patients using pooled DNA for cases and controls. This analysis suggested associations in *ZMIZ1* and *MSC* genes, and supported the association on chromosome 13q identified by Pezzolesi *et al*. [Craig 2009].

Further GWASs on DN have been performed on patients with T2D and with diverse ethnicity, although none of the loci have reached genome-wide statistical significance. An early GWAS including 188 Japanese patients with T2D genotyped for 80,000 SNPs suggested that rs741301 and eight other SNPs in the *ELMO1* gene were associated with DN in T2D [Shimazaki 2005]. The reported variants have not been replicated in other studies, but later investigations in T1D patients of European origin and in African American patients with T2D identified other variants in the *ELMO1* gene suggestively associated with DN [Leak 2009, Pezzolesi 2009b]. A GWAS on ESRD in African American patients with T2D supported the association on *FRMD3* gene when adjusted for major ESRD risk variants for non-diabetic ESRD in African Americans [Freedman 2011]. Further GWASs on DN in T2D have suggested associations in the *ACACB* gene in Japanese patients and in or near *RPS12*, *LIMK2*, *SFI1* and other genes in African American patients [McDonough 2010].

No GWASs have been performed on the continuous variables of albuminuria or kidney function in diabetic patients. However, these traits have been studied in non-diabetic subjects. A missense mutation rs1801239 in the *CUBN* gene was identified as a risk locus for albuminuria in non-diabetic patients, and the same variant was associated with microalbuminuria in patients with diabetes. *CUBN* encodes cubilin,

which is essential for the reuptake of albumin and other low-molecular-weight proteins in the proximal tubules [Boger 2011]. Furthermore, multiple loci have been identified for reduced kidney function in non-diabetic subjects, evaluated with eGFR. These include variants in or near the *UMOD, SHROOM3, GATM-SPATA5L1, CST* and *STC1* genes [Kottgen 2009], *LASS2, GCKR, ALMS1, TFDP2, DAB2, SLC34A1, VEGFA, PRKAG2, PIP5K1B, ATXN2, DACH1, UBE2Q2* and *SLC7A9* genes [Kottgen 2010], and *MPPED2, DDX1, SLC47A1, CDK12, CASP9,* and *INO80* genes [Pattaro 2012].

5 Aims of the study

The aim of this thesis is to dissect the genetic background of diabetic kidney disease by analyzing single nucleotide polymorphisms, SNPs, across the genome.

The specific research questions are:

1. Given the numerous candidate gene studies and other studies that have reported putative associations between SNPs and diabetic kidney disease, can we validate these findings with a large study with sufficient statistical power?
2. Which are the genetic susceptibility loci for diabetic kidney disease in patients with type 1 diabetes?
3. Given the gender difference in the risk of the most severe form of diabetic kidney disease, the end-stage renal disease (ESRD), are there genetic risk factors for ESRD that only affect men or women?
4. What is the heritability of albuminuria, the main marker and risk factor for diabetic kidney disease? Are there genetic risk factors associated with albuminuria in type 1 diabetes?
5. Can we discover additional susceptibility loci for diabetic kidney disease using advanced data mining methods?

Each of these specific research questions are answered with a journal article, and each journal article provides a partial solution to the research problem. These journal articles are combined in this dissertation summary.

6 Materials and methods

6.1 Study design

Publication I was a case – control association study to investigate earlier reported risk SNPs for DN, whereas Publications II-V had a two stage study design consisting of a discovery stage and a subsequent replication or stage II analysis. All publications report SNP based search of genetic risk factors in non-related individuals.

Publication I: The aim was to replicate previously reported significant risk markers for DN. The selected SNPs were analyzed in the Finnish Diabetic Nephropathy (FinnDiane) Study and the All Ireland Warren 3, Genetics of Kidneys in Diabetes UK (UK-ROI) study. As the GoKinD US study was part of most of the original publications, results from that study were not reported for all loci. In addition, we subsequently reanalyzed the GoKinD US GWAS data after intensive quality control and compared the results with the original findings.

Publication II: The discovery stage consisted of GWASs on DN and ESRD in the FinnDiane, UK-ROI and GoKinD US studies. Based on the meta-analysis of the three GWASs, the most significant loci were selected for phase II analysis and genotyping in nine additional cohorts. Finally, we performed a combined meta-analysis of all studies.

Publication III: We explored if gender specific risk factors exist for ESRD in the FinnDiane GWAS. The significant SNP association in women was replicated in three additional studies that included a substantial number of women with ESRD.

Publication IV: We studied which SNPs are associated with elevated AER in the FinnDiane GWAS. The most significant findings were replicated in seven additional studies. In addition, we evaluated the heritability of AER using the FinnDiane genome-wide genotype data.

Publication V: We applied an advanced data mining method on the genome-wide genotype data from the FinnDiane study to detect SNPs associated with various case – control definitions of DN. The validity of the findings was tested with a permutation procedure. The validated markers were further tested for association in three additional studies.

Additional analyses: Many additional cross-sectional and longitudinal association analyses, as well as *in silico* and *in vitro* functional analyses, were performed for Publications II-V to further characterize the main findings.

6.2 Phenotype definitions

6.2.1 Definition of T1D

T1D was defined as the age at onset of diabetes no more than 35 years in Publications I-III and no more than 40 years in Publications IV and V. In addition, permanent insulin treatment had to be initiated within one year after the diagnosis of diabetes. Additional data, such as C-peptide concentrations, were used in some of the studies for the diagnosis of T1D. In case of incomplete information, the diagnosis was based on the attending physician's own classification.

6.2.2 Definition of DN

DN stages were defined according to the diagnostic albuminuria thresholds presented in Table 1 (page 6). Because of a relatively large day-to-day variability of the urinary albumin excretion, two out of three consecutive measurements were required to surpass a given threshold. In addition, patients with normal AER were required to have duration of T1D of at least 15 years in all publications. The DN phenotype in Publications I-II and V is defined as either macroalbuminuria or ESRD. Subjects with known non-diabetic kidney disease were excluded from the analyses. Divergence from these DN and T1D phenotype definitions are noted in the study specific phenotype descriptions in Section 6.4.1.

6.3 The FinnDiane Study

The discovery stage of the Publications III-V consisted entirely of Finnish patients with T1D, recruited by the FinnDiane Study. The FinnDiane study was also the largest study in Publications I and II.

FinnDiane is a nationwide multicenter study with the aim to detect genetic, environmental, clinical and biochemical risk factors for DN and diabetic complications in general. The FinnDiane study includes patients from all five Finnish University Central Hospitals, all 16 central hospitals, and 56 regional hospitals and health care centers (Figure 8). The patients are recruited to the study by their attending physician, who completes the main questionnaire and provides the latest laboratory results for a selection of biochemical variables. Blood and urine samples are sent to the central laboratory of the FinnDiane study, where more biochemical variables are measured centrally at the Helsinki University Central Hospital laboratory.

The FinnDiane Study also involves a prospective phase, in which the patients are restudied roughly 5-7 years after the baseline visit. The clinical phenotypes are continuously updated based on the patient records. Furthermore, information on the major clinical events such as the onset of ESRD can be retrieved from the Finnish Hospital Discharge Registry (HILMO).

The FinnDiane GWAS included also 554 patients recruited across Finland by the Finnish National Institute of Health and Welfare (THL). Their clinical phenotype

and additional biochemical and anthropometric data were defined based on the patients' medical records. These patients were analyzed together with the FinnDiane patients using the same phenotype definitions and inclusion criteria.

The main baseline characteristics for the patients included in the GWAS are shown in Table 3, according to the inclusion criteria in Publication V.

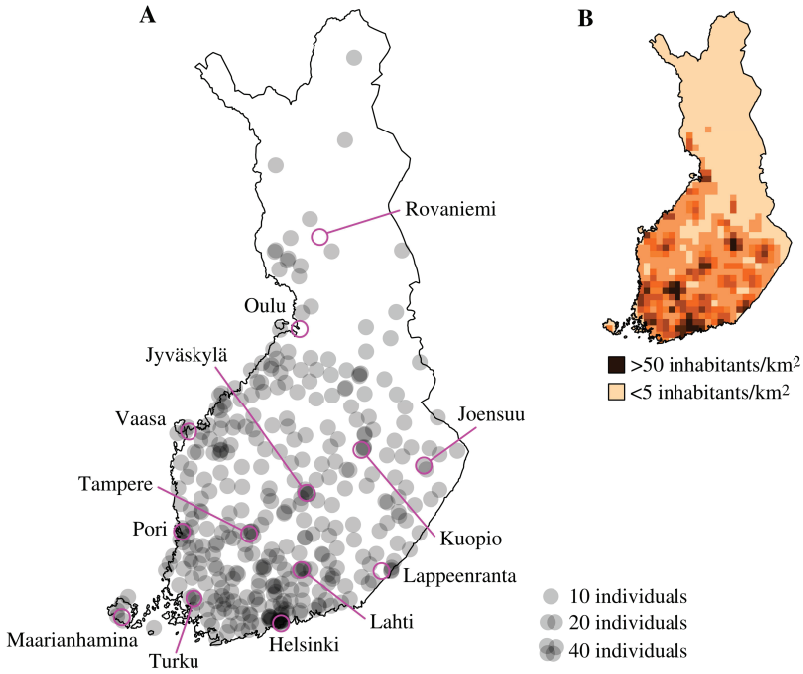


Figure 8: Geographical distribution of the FinnDiane patients. A: Current residence of the FinnDiane patients (n=4,130 addresses available). B: Regional population density in Finland (data from Statistics Finland). Figure is modified from [Mäkinen 2010].

6.4 Genome-wide genotyping and computational data preparation

In publication II, the discovery stage included three studies with GWAS data: the FinnDiane study, the GoKinD US and the UK-ROI studies. The three GWASs were genotyped on different genotyping platforms, but the post-genotyping quality control and imputation procedures were unified between the three cohorts.

6.4.1 Patients in the genome-wide association studies

The FinnDiane patients were described in the previous Section. The main clinical characteristics of the patients in the UK-ROI and GoKinD US studies are presented in Table 4.

*Table 3: Clinical characteristics of the FinnDiane patients, measured at the screen visit (apart from the DN status and the variables marked with *, which refer to the last known measurement).*

Parameters	Normal n=1,637	Micro-albuminuria n=472	Macro-albuminuria n=694	ESRD n=661
Men, n (%)	678 (41%)	270 (57%)	414 (60%)	396 (60%)
Age (years)*	43.3 ± 11.5	41.9 ± 12.8	45.4 ± 11.2	46.6 ± 9.4
Diabetes duration (years)*	27.7 ± 9.4	28.1 ± 11.2	31.9 ± 9.6	33.5 ± 8.5
Age at diabetes onset (years)	15.6 ± 8.9	13.8 ± 8.9	13.4 ± 8.6	13.1 ± 7.9
BMI (kg/m^2)	25.0 ± 3.3	25.6 ± 3.4	26.0 ± 3.8	24.7 ± 4.2
Laser treatment, n (%†)	299 (21%)	207 (48%)	339 (73%)	497 (90%)
Antihypertensive medication, n (%†)	281 (19%)	292 (67%)	412 (90%)	512 (93%)
HbA_{1C} (%)	8.2 ± 1.3	8.6 ± 1.4	8.9 ± 1.5	8.9 ± 1.7
HbA_{1C} (mmol/mol)	66 ± 14	70 ± 15	74 ± 16	74 ± 19
Systolic blood pressure (mmHg)	131 ± 16	136 ± 17	141 ± 18	151 ± 23
Diastolic blood pressure (mmHg)	78 ± 9	80 ± 10	82 ± 10	85 ± 12
Total cholesterol (mmol/L)	4.9 ± 0.8	5.0 ± 0.9	5.3 ± 0.9	5.4 ± 1.3
LDL cholesterol (mmol/L)	3.0 ± 0.8	3.1 ± 0.8	3.3 ± 0.8	3.5 ± 1.1
Triglycerides (mmol/L)	0.9 (0.7–1.2)	1.1 (0.8–1.6)	1.4 (1.0–2.1)	1.5 (1.0–2.1)
HDL cholesterol in men (mmol/L)	1.3 ± 0.3	1.2 ± 0.3	1.1 ± 0.4	1.2 ± 0.4
HDL cholesterol in women (mmol/L)	1.5 ± 0.4	1.4 ± 0.4	1.3 ± 0.4	1.3 ± 0.5
eGFR (ml/min per 1.73 m2)	91 ± 18	87 ± 21	71 ± 26	NA

†Percentages are expressed with respect to the number of measured values for each parameter. Data are given as mean ± SD or N (percentage) or median (interquartile range). Table is modified from Publication V.

UK-ROI: The study includes 1,904 white individuals with T1D, diagnosed before 31 years of age, whose parents and grandparents were born in the UK or Ireland. DN was defined as persistent proteinuria (>500 mg/24 h) developing more than 10 years after the diagnosis of diabetes, the presence of hypertension (>135/85 mmHg and/or treatment with AHT medication), and retinopathy; or ESRD. The controls had a duration of T1D ≥ 15 years, persistently normal AER, no AHT medication, and no history of treatment with ACE inhibitors [McKnight 2010b].

GoKinD US: Similar to the UK-ROI, the GoKinD US study consists of 1,792 self-reported white patients with T1D diagnosed before 31 years of age. Individuals were recruited at the George Washington University and the Joslin Diabetes Centre. DN was defined as ESRD or persistent macroalbuminuria (at least two out of three tests positive for albuminuria by dipstick ≥1+, or ACR >300 μg albumin/mg of urine

creatinine). The controls were defined using the same inclusion criteria as in UK-ROI [Pezzolesi 2009a].

Table 4: Characteristics of samples successfully analyzed in the UK-ROI and GoKinD studies in the Publication II. Cases comprise patients with macroalbuminuria or ESRD. Controls are patients with normal AER. Values are given as mean ± standard deviation. Table is modified from Publication II.

	UK-ROI, n=1,826		GoKinD US, n=1,595	
	Cases (n=823)	Controls (n=903)	Cases (n=774)	Controls (n=821)
Gender (men/women)	478/345	395/508	402/372	342/479
Duration of T1D (years)	32.9 ± 9.6	27.0 ± 8.6	31.4 ± 7.8	25.4 ± 7.7
Age at diagnosis of T1D (years)	14.5 ± 7.7	14.5 ± 7.8	11 ± 6.6	13 ± 7.3
HbA_{1C} (%)	9.0 ± 1.9	8.7 ± 1.6	7.5 ± 1.9	7.5 ± 1.2
HbA_{1C} (mmol/mol)	75 ± 21	72 ± 18	58 ± 21	58 ± 13
BMI (kg/m^2)	26.3 ± 4.7	26.2 ± 4.2	25.7 ± 5.2	26.1 ± 4.3
ESRD (%)	29.9	0	65.6	0

6.4.2 Genome-wide genotyping

In the FinnDiane Study, a total of 3,651 patients were genotyped at the Institute of Molecular Medicine Finland (FIMM, Helsinki, Finland) on the Illumina's BeadArray 610 Quad array (Illumina, San Diego, CA, USA).

DNA samples for 1,830 individuals in the UK-ROI collection were genotyped using Illumina's Omni1-Quad array at the Broad Institute. Illumina's BeadStudio clustering algorithm was used to call genotypes in both the UK-ROI and the FinnDiane.

The GoKinD US GWAS data, genotyped with the Affymetrix 500K platform (Affymetrix, Santa Clara, CA, USA), were downloaded from the dbGAP (phs000018.v2.p1, retrieved June 2010). The downloaded version 2 data was amended by updated genotype calling for a previously reported problematic plate [Pluzhnikov 2010] and additional quality control steps performed by NHLBI.

6.4.3 Quality control and population structure

An extensive genotype quality control procedure was applied for all three discovery GWAS datasets (UK-ROI, FinnDiane, GoKinD US). SNPs were filtered for those with high genotyping call rate, sufficient minor allele frequency (MAF ≥ 1%), concordance with (HWE), no difference in missingness by haplotype or by phenotype and no evidence of plate differences. Samples were included based on high individual genotyping rate and no extreme sample heterozygosity. From each pair of subjects with cryptic relatedness, defined as first-degree relatives, one was removed. In the UK-ROI and the FinnDiane, the samples were additionally excluded if there was discordance with previous genotypes. The quality control steps (detailed in Table 5) were performed using PLINK [Purcell 2007] and custom R scripts.

Table 5: Number of SNPs and samples filtered during the quality control steps for the three studies with GWAS data. Modified from Publication II.

	US GoKinD		UK-ROI		FinnDiane	
Quality control step	Subjects	SNPs	Subjects	SNPs	Subjects	SNPs
Raw GWAS Data (no QC)	1,792	364,292	1,830	975,120	3652	599,010*
Pre-quality control steps	162	734	2	–	11	–
Unsuccessful genotyping (sample call rate = 0)	–	–	–	–	35	–
Filter on SNP call rate (>90%)	–	0	–	296	–	2,252
Filter on SNP MAF (>1%)	–	25	–	179,985	–	44,617
Filter on subject call rate (>95%)	0	–	27	–	2	–
Filter on extreme heterozygosity	16	–	14	–	19	–
Filter on IBD/cryptic relatedness	4	–	22	–	39	–
Outlier detection (PCA)	15	–	39	–	–	–
Filter on HWE ($P<1\times10^{-7}$)	–	234	–	1,417	–	185
Missing by haplotype ($P<1\times10^{-7}$)	–	2,200	–	1,421	–	2,100
Missing by phenotype ($P<1\times10^{-7}$)	–	0	–	282	–	237
Test for plate effects ($P<1\times10^{-7}$)	–	200	–	32	–	89
Final GWAS data	1,595	360,899	1,726	791,687	3,546	549,530

QC: quality control. IBD: Identity by descent. PCA: Principal component analysis. Last row: final counts after all the quality control steps.
*Number of SNPs that were released by FIMM after their initial quality control.

Samples with significant evidence of admixture were identified and removed with clustering approaches such as multidimensional scaling and principal component analysis applied on the genome-wide genotype data, separately for each of the three studies. Principal component analysis was performed with the EIGENSTRAT program [Price 2006]. Genetic outliers were defined as more than six standard deviations away from the center of distribution along any of the ten first principal components (PCs) and they were iteratively removed until no outliers were detected. After this filtering procedure, the remaining samples in each study were combined with the genotype data of the three HapMap II populations [International HapMap Consortium 2007] and the PCs were recalculated and plotted to identify additional admixed individuals. Detailed results of each quality control step for each study are reported in Table 5 together with the final number of samples and SNPs passing the quality control.

After performing all the quality control steps, the final PCs were calculated for the remaining individuals. Depending on the Publication, either the two or the ten first PCs were employed to adjust the association analysis for any residual population structure.

6.4.4 Genotype imputation

After quality control, the genotype imputation was performed with the hidden Markov model (HMM) based Markov Chain Haplotyping algorith (MaCH 1.0 software) [Li 2009, Li 2010] using the HapMap II CEU samples as the reference panel for the haplotypes [International HapMap Consortium 2007]. Imputation was performed in two steps. In a randomly selected subset of ~300 patients, we first iteratively estimated (in 50 iteration rounds) the study specific model parameters linking the study population to the reference haplotypes: a "cross-over" parameter to estimate the probability to swithch from one reference haplotype to another between each SNP, and an "error rate" parameter for each SNP to allow differences from the reference panel. . The genotype imputation was then performed with the greedy algorithm and maximum likelihood method to infer the haplotypes for each subject and finally to fill in the missing SNPs based on the HapMap II reference haplotypes. SNPs with an estimated squared correlation between the imputed and the true genotypes $r^2 \leq 0.3$ were removed in the post-imputation quality control. The imputation procedure resulted in expected allele dosage data for ~2.4 million SNPs for each cohort.

The main statistical analyses were performed with the estimated allele dosages as the main explanatory variable. The estimated allele dosage is the expected count of the reference allele *a* in a genotype, ranging from 0 to 2: allele dosage $d(a) = 2 \times$ probability of the *aa* genotype + 1 × probability of the *Aa* genotype + 0 × probability of the *AA* genotype. For the purpose of some additional analyses such as the longitudinal models in Publication II, the maximal likelihood genotypes were employed instead; the genotype posterior probability of 0.9 was required for the genotype calling.

6.5 Statistical analysis of the GWAS data

6.5.1 Genome-wide association analysis

In Publications I – IV the genome-wide association tests were conducted with PLINK v1.07 [Purcell 2007] with the case – control status or AER as the dependent variable, and the estimated allele dosage data for one SNP at the time as the main explanatory variable.

Logistic regression was employed for the case – control phenotypes, and linear regression for the analysis of continuous variables. Logarithmic values rather than raw values were applied for the continuous AER measurements that followed more closely a log-normal distribution than a normal distribution (Figure 9). Models were adjusted for sex, age or age at onset of diabetes, duration of diabetes and PCs. In Publication II, the UK-ROI and GoKinD US data were additionally adjusted for the study center.

In Publication IV, subjects with or without AHT medication were analyzed separately. Results from the two groups were combined with fixed effects meta-analysis (See Section 6.5.3, Meta-analysis, for more details).

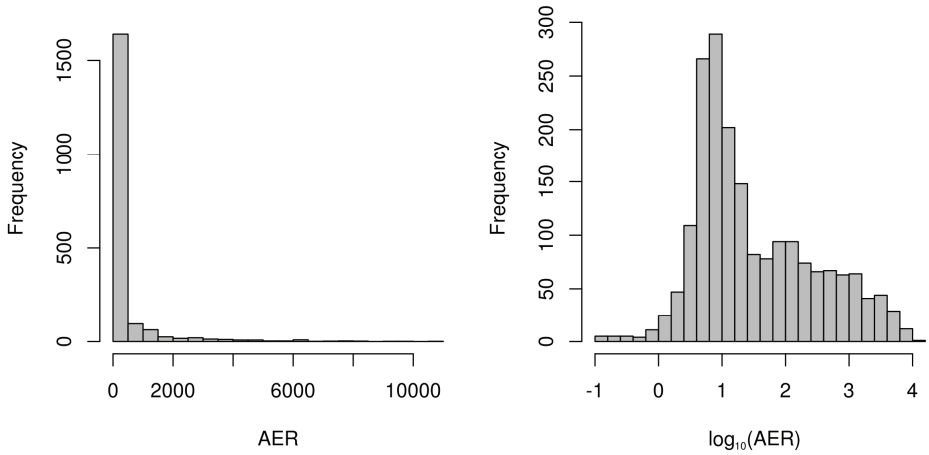


Figure 9: Histograms of 24-h AER distribution in Publication IV. A: untransformed AER value. B: log_{10} transformed AER.

Genome-wide quality control of the results: The results of the genome-wide association analyses were subject to genome-wide quality control. *P*-values are assumed to follow a uniform distribution under the null hypothesis of no association. The genomic inflation factor λ_{GC} was calculated as the median observed test statistic (i.e. χ^2 calculated from the *P*-values) divided by its expectation under the 1 degree of freedom (d.f.) χ^2 distribution (0.455). If λ_{GC} indicated inflation of the results, $\lambda_{GC} > 1$, the GWAS results were adjusted for the study specific λ_{GC} by dividing the test statistic by λ_{GC} and recalculating the *P*-value based on the obtained test statistic. Similarly, the standard errors (SE) of the effect size β coefficients were adjusted with $SE_{GC} = SE \times \sqrt{\lambda_{GC}}$ [de Bakker 2008]. Values of $\lambda_{GC} > 1.05$ were considered indicative of stratification or other issues in the statistical analysis.

Quantile-Quantile plots (QQ-plots) were plotted based on the $-\log_{10}$(*P*-values). The observed *P*-values were sorted and plotted on the y-axis, and the expected values *P'* on the x-axis (Figure 10). Expected values follow uniform distribution and were obtained with

$$P' = i/(L+1)$$

where *L* is the number of observed *P*-values, and *i* has values from 1 to *L*. It is common to use $-\log_{10}$(*P*-values) rather than *P*-values in the QQ-plots of GWAS data to help emphasize the smallest – and most interesting – *P*-values. High λ_{GC} is reflected in the QQ-plots as a significant deviation from the diagonal for the majority of the SNPs. On the other hand, significant deviation from the diagonal on the top-right corner of the QQ-plots indicates smaller-than-expected *P*-values and thus significant findings [Balding 2006].

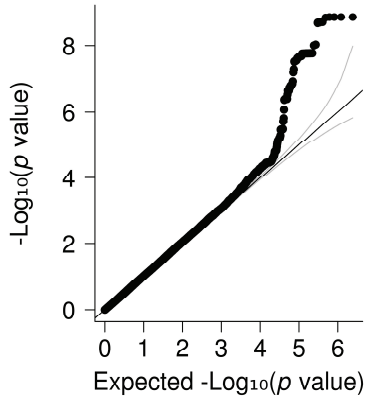


Figure 10: Quantile-quantile (QQ) plot of GWAS results. X-axis: expected P-values; Y-axis: observed P-values. Gray lines show the 95% confidence interval. No systematic inflation of results is observed, as the majority of the P-values adhere to the diagonal. Deviation from the diagonal on the top-right corner indicates better-than-expected P-values, and thus, evidence of significant associations.

Visualization of the GWAS results: Manhattan plots are commonly used to visualize the GWAS results on the genome-wide scale at a glance: the strong association signals rise from the plot like the skyscrapers in Manhattan. Manhattan plots show the chromosomal position on the x-axis and the -$\log_{10}$(*P*-values) on the y-axis (Figure 11).

LocusZoom plots were used to visualize smaller chromosomal regions around the main association signals [Pruim 2010]. Similar to the Manhattan plots, the chromosomal position is given on the x-axis, and the -$\log_{10}$(*P*-values) on the y-axis. Gene locations are superimposed in the figures, as well as the recombination rate. The SNPs are colored according to their LD (r^2) with the index SNP. Depending on

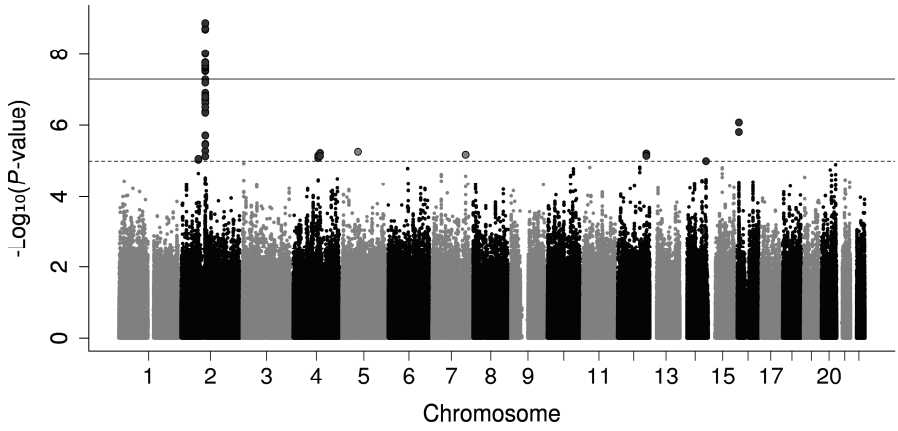


Figure 11: Manhattan plot of GWAS results indicating a strong association signal on chromosome 2. X-axis: chromosomal position of SNPs. Y-axis: significance of the SNPs (-Log10(P-value)). The dashed horizontal line shows a suggestive P-value threshold of $P < 10^{-5}$, and the solid horizontal line indicates the P-value threshold for genome-wide statistical significance ($P<5\times10^{-8}$).

the data availability, either the HapMapII CEU samples or the 1000 Genomes European samples were employed as the reference in order to calculate the LD structure.

6.5.2 Genome-wide SNP discovery with Naïve Bayes Classifiers

In Publication V, we applied a recently proposed Bag of Naïve Bayes (BoNB) algorithm [Sambo 2012] on the FinnDiane GWAS genotype data to explore genetic variants associated with different stages of DN. The BoNB algorithm is a multivariate data mining method to identify SNPs associated with a case – control phenotype in genome-wide genotype data. In brief, an ensemble of Naïve Bayes classifiers is trained with 100 bootstrap replicates and the performance of each Naïve Bayes classifier is tested with the out-of-bag sets. The marginal utility of the repeatedly selected SNPs is tested by genotype permutation (Figure 12).

The BoNB algorithm proceeds as follows:

1. 100 bootstrap replicates and out-of-bag sets are generated from the original dataset. A typical bootstrap replicate of the 3,464 FinnDiane samples included ~2,190 unique samples, with ~920 samples represented at least twice. The out-of-bag sets had typically ~1,270 samples.
2. For each bootstrap replicate, the SNPs are ranked according to their ability to classify the subjects to cases and controls. The classification is based on a simple Naïve Bayes classifier including only one SNP at the time, assuming a general 2-d.f. genotypic model. The classification performance is measured

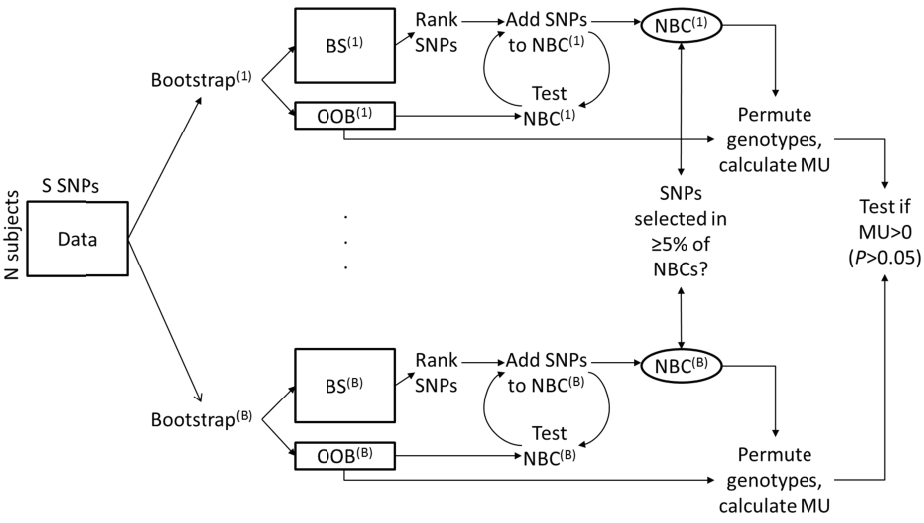


Figure 12: Schematic presentation of the BoNB algorithm. Bootsrap samples $BS^{(1)}$ … $BS^{(B)}$ are drawn with bootstrapping from the original data. Naïve Bayes classifiers (NBCs) are trained with the Bootstrap samples, and SNPs are added to NBCs until the test performance on the out-of-bag (OOB) samples starts to decrease. SNPs included in ≥5 NBCs are tested furter with permutation: the genotypes of these SNPs are permuted, and the marginal utility (MU) of the SNPs is calculated as the decrease in the classification performance of the NBC. Finally, the significant SNPs are defined as those with marginal utility significantly > 0 across the permuted NBCs. Modified from Publication V.

with the Matthews Correlation Coefficient (MCC) [Baldi 2000]. Then, a multimarker Naïve Bayes classifier is created for each bootstrap replicate, initially without any SNPs. SNPs are added iteratively to the Naïve Bayes classifier, adding always the SNPs with the highest scores. Each time SNPs are added, all the SNPs in LD with the added SNPs (defined as < 1 megabase (Mb, 1,000,000 bases) away and LD r^2>0.1) are removed from the list of remaining SNPs. After each addition of SNPs, the classification performance of the Naïve Bayes classifier is tested on the corresponding out-of-bag set with MCC. SNPs are added until the classification performance starts to decrease (the Naïve Bayes classifier starts to over-fit the model).

3. All the SNPs in the genetic regions (distance < 1 Mb) that are included in at least 5 Naïve Bayes classifiers are selected for a permutation procedure in the out-of-bag sets. The genotypes of the selected SNPs are randomly permuted, one at the time, in the out-of-bag sets of the Naïve Bayes classifiers containing the SNP. The marginal utility of the SNP (or corresponding 1 Mb region) in a Naïve Bayes classifier is computed as the relative decrease in classification performance on the out-of-bag set due to the permutation.
4. Genetic markers are defined as those with marginal utility significantly greater than zero across all the tested Naïve Bayes classifiers (Wilcoxon signed-rank test, *P*-value < 0.05).

6.5.3 Meta-analysis

Fixed effect meta-analysis was performed with the METAL software to combine the results from multiple studies [Willer 2010]. In Publications I-IV the meta-analysis was calculated based on the effect size estimate β's (i.e. natural logarithm of the odds ratio (OR) for the case – control phenotypes) and standard errors using the inverse variance method, where the β's are weighted by the standard errors to obtain an overall Z-score which can be converted to a *P*-value.

In Publication V, where the models were evaluated with genotypic association models rather than with allelic association models, we performed two different meta-analyses: first, based on the *P*-values and sample sizes without considering the effect direction, and second, with the inverse variance method taking into account the direction of effect in the best fitting bimodal mode of inheritance (additive, recessive or dominant). The *P*-value based approach converts the *P*-value observed in each study into a Z-score. The overall Z-score is the weighted sum of the Z-score in each study, weighted proportional to the square-root of the sample size of the study [Willer 2010].

6.5.4 Heritability estimates

The narrow-sense heritability of AER, defined as the phenotype variance explained by the additive effects of the genotyped SNPs, was estimated from the GWAS data of the unrelated FinnDiane patients (the first degree relatives were removed during the GWAS quality control). The GCTA software utilizes the residual relatedness

structure within the genome to assess the heritability of a trait [Yang 2011]. The method gives a lower limit for the heritability because it can only account for the heritability that is correlated with the SNPs in the employed genotyping platform.

6.5.5 Longitudinal analysis

Longitudinal data from the FinnDiane discovery cohort was used in Publication II to evaluate the association between the SNPs associated with ESRD and the duration from the onset of T1D until the diagnosis of microalbuminuria, macroalbuminuria or ESRD. These time-to-event phenotypes are illustrated in Figure 13. Additionally, we analyzed time from onset of macroalbuminuria to ESRD. The most recent kidney status data were utilized for each patient. The year of onset of the complication (microalbuminuria, macroalbuminuria, ESRD) was determined from the FinnDiane study questionnaires. The latest data for ESRD were obtained from the Finnish Hospital Discharge Registry (HILMO, as per December 31, 2009), and these data were available for all participants.

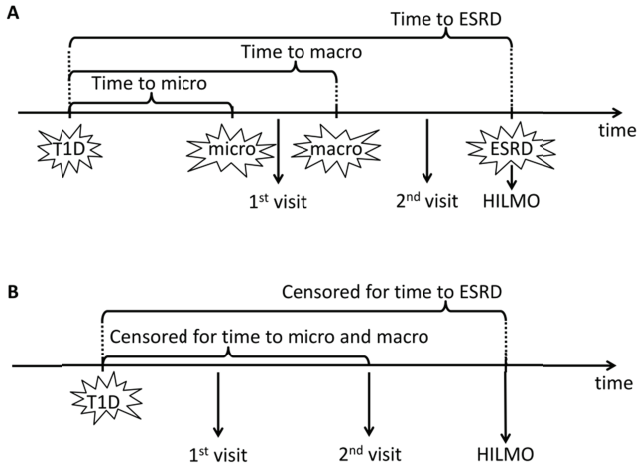


Figure 13: Illustration of the time-to-event analysis phenotypes. A: patient who develops macroalbuminuria by the time of their second FinnDiane visit, and ESRD before the HILMO registry look-up (30.12.2009). B: A patient who does not develop any kidney complications by their second FinnDiane visit, and does not develop ESRD by the time of HILMO registry look-up (December 31, 2009).

We also examined whether the SNPs were associated with mortality using data from the Finnish Death Registry (as per September 30, 2010). As patients with macroalbuminuria have a significantly increased risk of mortality, the time until death was analyzed separately for participants with and without DN. For the analysis before DN, we used time from T1D onset to death or latest record, with patients who developed DN censored out at the time of the onset of DN. For patients with DN, we analyzed time from onset of DN to death/latest record and time from onset of ESRD to death/latest record. In this retrospective study setting, selection bias can arise if the SNP is associated with the ESRD specific mortality. In order to avoid such a selection bias, we also performed the analysis of time from ESRD onset to death in the patients with incident ESRD.

The time-to-event analyses were performed using Kaplan-Meier and Cox proportional hazards regression, implemented in the 'survival' package in R software (version 2.36-10, http://cran.r-project.org/web/packages/survival) with the most likely genotypes rather than the allele dosage data.

6.5.6 Linkage disequilibrium structure

The LD (both *D'* and r^2) between SNPs was investigated to define "proxy" SNPs that can be used as surrogate markers for the reported signals (Publication I), to evaluate which SNPs represent individual signals (Publications II-V), and to discover the causal markers that could explain the observed association on the lead SNP (Publication IV). Pair-wise SNP correlations were estimated with PLINK [Purcell 2007]. Regional LD structure was explored more comprehensively in Publication IV around the *GLRA3* association region with the HaploView software [Barrett 2009a].

6.5.7 Pathway analyses

Publications II-IV include gene set enrichment analyses of the GWAS results performed with the Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA) software [Segre 2010]. MAGENTA software first maps SNPs into genes based on the chromosomal position, with gene regions spanning 110 kb upstream and 40 kb downstream of the gene's most extreme transcript boundaries. Second, the genes are given an association score based on the *P*-values of the SNPs within the gene region. The scores are adjusted for possible confounding factors such as gene length and SNP density. Finally, the pre-defined gene sets are tested for enrichment of highly ranked gene association scores compared with random gene sets with a permutation procedure. We employed the 95th percentile cutoff for the gene score rank to define enrichment of the genes as suggested in [Segre 2010]. The enrichment analyses contained a total of 2,580 gene sets, including 186 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, 1,785 gene ontology terms, 217 PANTHER biological processes, 217 PANTHER molecular functions, 94 PANTHER pathways and entries, and 81 Ingenuity pathways.

In Publication IV for the GWAS on AER, we employed additionally a more conventional pathway analysis, performed with the PANTHER database. The analysis included all the SNPs with a *P*-value < 0.001 in the GWAS, and the SNPs were assigned to genes based on the Ensembl database annotations. The overrepresentation of genes in the PANTHER pathways was estimated with the binomial statistics. The resulting *P*-values for enrichment were adjusted for multiple testing with Bonferroni correction using the number of non-overlapping ontology classes as the number of independent tests [Mi and Thomas 2009, Mi 2013].

6.5.8 *In silico* annotation of loci

Many publicly available databases exist for annotation and *in silico* analysis of SNPs. Positional annotations were obtained from the dbSNP (www.ncbi.nlm.nih.gov/SNP) or Ensembl (www.ensembl.org) databases. Functional annotation of intergenic SNPs was performed with the RegulomeDB [Boyle 2012] (Publications III-V).

RegulomeDB was designed for the functional annotation of SNPs located in non-coding and intergenic regions of the human genome. The database combines annotations from experimental data sets and *in silico* annotations showing evidence of transcription factor binding (e.g. chromatin immunoprecipitation (ChIP-seq) of transcription factors and histones, DNase I hypersensitivity assays, DNase I footprints, transcription factor binding motifs and expression quantitative trait loci (eQTL) data) in order to evaluate if a SNP overlaps with any experimentally detected regulatory regions. In Publication III, potential transcription factor binding sites and estrogen responsive elements were additionally explored with the MatInspector from the Genomatix software suite (Genomatix Software, GmbH, Munich, Germany), based on known transcription factor binding motifs. We also studied if the main SNPs in Publication III affected the gene expression level of any of the genes within a 1 Mb region up- and downstream of the signal, evaluated in the HapMap3 lymphoblastoid cell lines [Stranger 2012] using the Genevar software (www.sanger.ac.uk/resources/ software/genevar).

The disease specific differential gene expression was evaluated in kidney tissue for the genes located near the main findings. In Publications III and IV, we used data from the publicly available Nephromine database (www.nephromine.org). The two employed data sets contain renal biopsy specimens from patients with DN and healthy living kidney donors or patients with minimal-change kidney disease [Schmid 2006, Woroniecka 2011]. Publication II used data from renal biopsies in early DN (NIDDK Pima Indian Cohort) and from healthy living donor kidney transplant biopsies as described by Berthier *et al.* [Berthier 2009].

6.6 Validation of the results in follow-up studies

A total of ten studies consisting of patients with T1D were included in the follow-up studies in Publications II-V to validate the findings. In addition, patients with T2D and different ethnic background from the Family Investigation of Nephropathy and Diabetes (FIND) Study were analyzed in Publication III.

6.6.1 Patients in the follow-up studies

The patient inclusion criteria and the phenotype definitions were similar to the ones presented in Section 6.2. The follow-up studies are briefly described in this section.

DCCT/EDIC: The DCCT was a multi-center randomized clinical trial to compare the effect of intensive and conventional insulin therapy on the development and progression of early vascular and neurological complications of T1D. The follow-up study is called Epidemiology of Diabetes Interventions and Complications (EDIC) [The DCCT Research Group 1986, Molitch 1993, The DCCT Research Group 1995, Writing Team for the DCCT/EDIC Research Group 2003]. The DCCT study included 1,304 white subjects with genotype data. Renal outcomes were defined as time in years from DCCT baseline until the event. AERs were measured annually in the

DCCT and every other year in the EDIC. The patients were followed for 17.5 ± 2.6 years (mean ± SD) in the DCCT/EDIC with 12 ± 2 AER measures.

Steno: The Steno T1D study aims to study the genetic risk factors for the development of diabetic complications [Tarnow 2008]. All adult white patients with T1D attending the outpatient clinic at Steno Diabetes Center between years 1993 and 2000 were invited to participate in the study. In total, 458 Steno patients had DN defined as persistent albuminuria, the presence of retinopathy, and the absence of other kidney or urinary tract disease. Controls were defined as patients with persistent normal AER after more than 15 years of T1D in patients not treated with ACE inhibitors or angiotensin-II receptor blockers. In total, 442 subjects were included as controls. All urinary AER values were measured from 24-h urine collections.

Scania Diabetes Registry: The Scania Diabetes Registry (SDR) aims to register all individuals with diabetes in the Malmö region in Southern Sweden [Lindholm 2001]. Patients of non-Scandinavian origin were excluded from the analysis. The diagnosis of kidney disease otherwise followed the definitions in Section 6.2.2, but ESRD was defined as having dialysis or kidney transplant, or eGFR <15 ml/min. After genotype quality control in Publication II, there were 290 individuals with normoalbuminuria, 103 individuals with macroalbuminuria and 35 individuals with ESRD. A total of 494 patients with AER values were included in Publication IV.

France-Belgium, GENEDIAB & GENESIS Cohorts: The GENEDIAB [Marre 1997] and GENESIS [Hadjadj 2004] patients were recruited in France and Belgium in 1994-1995 and 1999-2001, respectively. The studies include more than 800 patients with long-standing T1D. The kidney status was classified based on the three highest AER and/or albumin concentration measurements within the last 5 years. DN cases were required to have past or present retinopathy.

Italy: The Italian cohort comprises of 356 unrelated, white, Italian patients with T1D, i.e. 188 cases with established DN and 168 control patients with normal AER. Cases with DN had concomitant diabetic retinopathy and absence of other renal or urinary tract disease or clinical or laboratory evidence of cardiac failure. Patients were recruited and studied at the Complications of Diabetes Unit of the San Raffaele Scientific Institute, Milan, Italy [Del Bo 2006].

Sweden: The Swedish cohort was collected from the Department of Endocrinology in Stockholm and the Department of Medicine in Umeå, Sweden [Mollsten 2008]. All patients with T1D were Swedish and they had diabetes diagnosed before 30 years of age. Kidney disease classification followed the definitions in 6.2.2.

RomDiane: RomDiane is a Romanian cross-sectional study of T1D, with two participating centers – Bucharest and Craiova. The study aims to identify risk factors for DN and other chronic complications in patients with T1D in the Romanian population. The data regarding diabetic complications, cardiovascular status, use of medication, and personal and family medical history were assessed by a standardized questionnaire completed by the investigators.

UK Nephropathy Family Study and Oxford Regional Prospective Study (NFS-ORPS): ORPS consists of children diagnosed with T1D before 16 years of age, recruited between 1986 and 1996 in the geographic region of the Oxford Health Authority [Amin 2008]. The NFS recruited adolescents aged 10–16 years with T1D between 2000 and 2005 throughout England [Marcovecchio 2009]. Both cohorts have been monitored with annual assessments of ACR.

FIND Study: The study was designed to investigate genetic risk factors for DN. The GWAS effort conducted in FIND consisted of 885 unrelated samples from European Americans, 1,460 from African Americans, 889 from American Indians, and 1,535 from Mexican Americans. Patients with any type of diabetes were eligible for the study, but the majority of the patients had T2D. Overt proteinuria was defined as proteinuria ≥ 500mg/24h, or AER ≥ 300mg/24h or protein to creatinine ratio >0.3 [Knowler 2005].

6.6.2 SNP selection criteria for the targeted genotyping

Table 6 summarizes the criteria applied in each Publication for the SNP selection for the follow-up studies. In Publication I the selection was based on earlier reported associations, and in Publications II-V based on the discovery GWAS or GWASs.

Table 6: SNP selection criteria for replication and phase two analysis.

Publication: phenotype	Replication selection criteria	Selected SNPs
Publication I: DN	i. All loci from the literature that reported genome-wide significance ($P<5\times10^{-8}$) ii. Suggestive signals from previous GWASs on DN in T1D (SNPs with $P<10^{-5}$) iii. Suggestive signals from previous GWASs on DN in T2D, with regional support in T1D iv. SNPs cited in a recent meta-analysis of candidate loci for DN	i. 1 ii. 8 iii. 679 iv. 30
Publication II: DN, ESRD	Loci with $P < 10^{-5}$ in the meta-analysis of the discovery cohorts; The lead SNP plus a proxy was selected when available.	42 SNPs at 24 loci
Publication III: ESRD in women/ men	i. Loci with P < 5×10-8 within men or women in the FinnDiane GWAS. ii. In silico replication attempt of loci with P < 1×10-5 within men/women in the FinnDiane GWAS	i. 1 ii. 6
Publication IV: AER	Independent SNPs with $P < 1\times10^{-4}$ in the FinnDiane GWAS; 3 additional SNPs selected for the main locus	64
Publication V: DN, ESRD, macro-albuminuria	SNPs included in at least five Naïve Bayes classifiers and significant marginal utility after genotype permutation (Wilcoxon signed-rank test, $P<0.05$).	5

6.6.3 Targeted genotyping

The targeted genotyping was performed either as small scale *de novo* genotyping, or extracted from directly genotyped or imputed GWAS data. The applied genotyping methods are summarized in Table 7.

The quality control of the genotypes included filtering for low sample genotyping rate, low SNP genotyping rate, low MAF and deviation from HWE. Additionally SNPs with low minor allele count (MAC<10) within the corresponding case and control groups were excluded in order to ensure the stability of the statistical analysis. If the genotype data were obtained from a GWAS, then the data were extracted from the quality controlled genome-wide data, according to the filters applied in each study.

Table 7: Genotyping methods used in the studies. The discovery GWASs are high-lighted with gray background. Other studies were used for targeted genotyping in replication purposes.

Cohort	Publication I: DN	Publication II: DN, ESRD	Publication III: ESRD in women/ men	Publication IV: AER	Publication V: DN, ESRD, macro-albuminuria
FinnDiane	TaqMan, GWAS	Discovery GWAS	Discovery GWAS	Discovery GWAS	Discovery GWAS
UK-ROI	iPLEX, TaqMan, GWAS	Discovery GWAS	GWAS	iPLEX, TaqMan	GWAS
GoKinD US	GWAS*	Discovery GWAS	GWAS		GWAS
Steno		iPLEX		OpenArray	TaqMan
Italy		iPLEX	iPLEX	OpenArray	
DCCT/EDIC		GWAS			
SDR		iPLEX		iPLEX, TaqMan	
UK-ROI replication		iPLEX			
France-Belgium		iPLEX			
Sweden		iPLEX		OpenArray	
RomDiane		iPLEX			
FinnDiane replication		iPLEX		GWAS	
NFS-ORPS				iPLEX, TaqMan	
FIND (T2D)			GWAS		

*GoKinD US was included only when it was not part of the original publication. TaqMan: TaqMan chemistry (Applied Biosystems, Foster City, CA, USA). iPLEX: Sequenom iPLEX (Sequenom Inc., San Diego, CA, USA). OpenArray: TaqMan chemistry OpenArray (Applied Biosystems, Life Technologies, Carlsbad, CA) in a 64-SNP format.

6.6.4 **Association analysis in the replication studies**

Association analysis of the replication cohorts was performed similarly to the genome-wide association studies using logistic or linear regression. For the *de novo* genotyped SNPs, the models were not adjusted for PCs, as their calculation requires genome-wide genotype data.

The study design of the DCCT/EDIC study differs considerably from the other studies. DCCT/EDIC was used for replication in Publication II using Cox proportional hazards analysis of discrete time-to-event outcome with additive

genotype coding. The main outcome was severe nephropathy, corresponding to the DN definition in Publication II. During the follow-up, 10% of the patients developed the severe nephropathy outcome (132 events vs. 1172 censored).

7 Results and Discussion

This section presents the main genetic results from the Publications I-V. In Publication I, we examined previously published genetic associations with diabetic nephropathy (DN) in a large set of 6,366 patients with type 1 diabetes (T1D). Using the patients from the same three studies on T1D, we then performed GWASs and meta-analyses on DN and end stage renal disease (ESRD), reported in Publication II. In Publication III, we explored genetic variants associated with ESRD in men and women separately. In Publication IV we investigated the genetic risk factors for albuminuria as a continuous variable. In Publication V we applied a novel data mining algorithm on the FinnDiane GWAS to explore if more susceptibility loci for DN could be discovered with a Bayesian approach.

7.1 Replication attempt of previous DN susceptibility loci

Many studies have been carried out to identify the genetic risk factors for DN, and these have resulted in hundreds of putative genetic susceptibility loci [Mooyaart 2011]. However, very few of the findings have been compellingly replicated. In the GENIE consortium, we performed comprehensive association testing of the previously reported variants using T1D patients from the UK-ROI and the FinnDiane studies, and combined those with the re-analyzed data from the GoKinD US Study. We examined all the genetic variants that had shown high levels of statistical significance in previous candidate gene studies, had been successfully replicated in independent studies, or had originated from genome-wide association studies on DN in T1D.

7.1.1 *EPO* promoter polymorphism

By the time of Publication I, the only genome-wide significant association with DN in T1D was reported for the erythropoietin (*EPO*) gene promoter polymorphism rs1617640 ($P = 2.8\times10^{-11}$) identified in a candidate gene association study [Tong 2008]. We studied this association by *de novo* genotyping of the SNP in the FinnDiane and the UK-ROI studies. As in the original report, the cases had both ESRD and proliferative diabetic retinopathy. No significant association was observed in either the UK-ROI (P=0.19) or the FinnDiane collections (P=0.60). Despite little evidence of association in the GENIE consortium, the association retained genome-wide statistical significance in the meta-analysis of the FinnDiane,

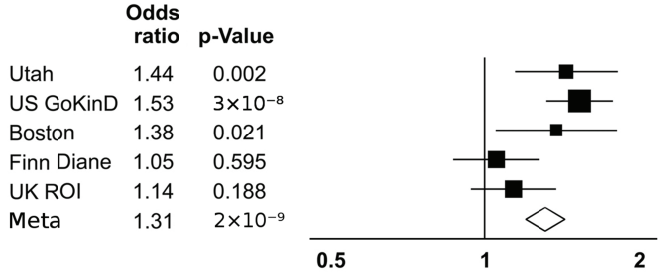


Figure 14: Meta-analysis of the previously published and new results (FinnDiane and UK-ROI) provides evidence of association between the EPO promoter SNP rs1617640 and the combined phenotype of ESRD and proliferative diabetic retinopathy in T1D.

UK-ROI and the original studies, including 3,162 cases and 3,845 control subjects in total (Figure 14).

7.1.2 Variants in *ELMO1*

A high-throughput genome-wide SNP genotyping study of DN in Japanese patients with T2D suggested variants in the *ELMO1* gene as genetic risk factors for DN, with the strongest association obtained for rs741301 [Shimazaki 2005]. Subsequently, Pezzolesi *et al.* studied the association between 118 variants within the *ELMO1* gene region and DN in T1D patients of European origin in the GoKinD US GWAS study. They found no association between rs741301 and DN, but reported eight other SNPs nominally associated (P=0.05) with DN in T1D (P=0.0017 for rs11769038) [Pezzolesi 2009b]. However, we found only low to moderate LD between rs741301 and the eight reported SNPs (r^2 ranging from 0.38 to 0.65). Thus the study by Pezzolesi *et al.* cannot be considered as a confirmation of the role of *ELMO1* in DN in patients with T1D.

We first examined the association between the original *ELMO1* risk marker rs741301 and DN in UK-ROI, FinnDiane, and the meta-analysis of the two, but found no evidence of association (P=0.40). No association was found at the SNPs reported in the GoKinD US either.

Lack of replication may be due to false positive findings in the original publication, poor statistical power in the replication study, distinct pathology of DN in T1D versus T2D [Fioretto and Mauer 2010], or different LD pattern in populations of diverse ancestries [Leak 2009]. To address the difference in the LD structure, we expanded our analysis to all available SNPs within the *ELMO1* gene and 20 kb up and downstream of the gene. However, we did not find any significant association in the UK-ROI or the FinnDiane cohorts individually, for these two cohorts in the meta-analysis, or in combination with the reported risk variants from the US GoKinD ($P<4.3\times10^{-4}$ required for significance to account for multiple testing based on 2,199 tested SNPs with 113.7 effect-independent SNPs as evaluated with the SNPSpD software [Nyholt 2004]).

7.1.3 Putative susceptibility variants from GWAS on DN in T1D

The US GoKinD GWAS reported four distinct loci with a total of 11 SNPs suggestively associated with the risk of developing DN in T1D [Pezzolesi 2009a]. Pruning out the SNPs in high LD, we selected eight independent SNPs within the four susceptibility loci. We re-evaluated the associations at the eight reported SNPs using the updated GoKinD US GWAS data with an improved genotype calling and quality control measures. The SNPs near the *FRMD3* gene and on the chromosome 12q region had similar *P*-values and effect sizes as reported in the original publication ($P=2\times10^{-7}$ – $P=9.5\times10^{-5}$). In contrast, the significance of the SNPs in the *CPVL/CHN2* and *CARS* regions was drastically reduced from $P = 6.5\times10^{-7}$ to 0.0020 and from $P = 6.4\times10^{-6}$ to 0.0022, respectively.

Association analysis of these eight SNPs in the UK-ROI and FinnDiane revealed no significant associations in either cohort ($P>0.05$). Furthermore, the extended analysis 20 kb up- and downstream of the eight SNPs showed no association in the two cohorts separately or in their meta-analysis after adjustment for multiple testing ($P>4.3\times10^{-4}$ before Bonferroni correction). In the combined meta-analysis of FinnDiane, UK-ROI and US GoKinD, two SNPs downstream of *FRMD3* remained associated with DN after adjusting for experiment-wide multiple testing (rs1888747 $P=1.5\times10^{-4}$, rs13288659 $P=9.7\times10^{-5}$).

7.1.4 Variants associated with DN in a literature based meta-analysis

Finally, we examined all the available SNPs cited by Mooyaart *et al.* in the most comprehensive literature based meta-analysis of candidate genetic variants for DN in T1D and T2D published to date. Mooyaart *et al.* identified 24 genetic variants repeatedly associated with DN [Mooyaart 2011]. Two of these SNPs were nominally associated with DN in the FinnDiane GWAS (rs13293564 at *UNC13B*: $P=0.01$; rs179975 at *ACE:* $P=0.03$) and one SNP in the UK-ROI GWAS (rs39075 at *CPVL/CHN2*: $P=0.05$). The *ACE* SNP rs179975 was nominally significant also in the meta-analysis of the two cohorts, $P=0.04$. The only signal remaining significant after Bonferroni correction was the *FRMD3* signal at rs1888747, with $P = 1.5\times10^{-4}$ when the results from the US GoKinD were included in the meta-analysis. However, the association originates from the US GoKinD GWAS, and thus, this cannot be considered as replication of the signal.

7.1.5 Summary of the association testing of previously reported susceptibility loci

The main results of the Publication I are summarized in Table 8. In our examination of two large studies of DN in T1D, FinnDiane and UK-ROI, we observed some nominally significant associations ($P<0.05$) with DN for the putative susceptibility variants previously published in candidate gene studies, but none of the associations remained significant after correction for multiple testing. Furthermore, our findings in the FinnDiane and UK-ROI studies do not support the previously reported GWAS results for DN in T1D in the US GoKinD study either. Nevertheless, the association

Table 8: Summary of the association testing results in Publication I

Original study	Reported association/ studied region	FinnDiane	UK-ROI	FinnDiane + UK-ROI	US GoKinD	FinnDiane + UK-ROI + US GoKinD
EPO: candidate gene study on DN in T1D [a]	rs1617640, $P=2.8\times10^{-11}$	NS ($P>0.05$)	NS ($P>0.05$)			FinnDiane + UK-ROI + original studies $P=2\times10^{-9}$
GWAS on DN in T2D [b]	*ELMO1* rs741301, $P=8\times10^{-6}$	NS ($P>0.05$)	NS ($P>0.05$)	NS ($P>0.05$)	NS ($P>0.05$)[c]	
ELMO1: candidate gene study [c]	20 kb up- or downstream rs741301	NS	NS	NS	8 SNPs $P=0.002–0.05$ [c]	NS
GWAS on DN in T1D (GoKinD US) [d]	8 loci with $P<10^{-5}$	NS ($P>0.05$)	NS ($P>0.05$)	NS ($P>0.05$)	2 SNPs in *FRMD3* $P<10^{-5}$ after re-analysis	rs1888747 (*FRMD3*) $P=1.5\times10^{-4}$
Meta-analysis of DN in T1D and T2D [e]	30 loci	rs13293564 (*UNC13B*) $P=0.01$; rs179975 (*ACE*) $P=0.03$	rs39075 (*CPVL/ CHN*) $P=0.05$	rs179975 (*ACE*) $P=0.04$		rs1888747 (*FRMD3*) $P=1.5\times10^{-4}$

Summary of the tested candidate gene associations. US GoKinD study was part of all the four original publications. Thus, results for GoKinD US are reported only when appropriate. NS: non-significant. Study-wise *P*-value cut-off for statistical significance after multiple testing was 4.4×10^{-4}. This cut-off is applied unless otherwise stated.
[a][Tong 2008]; [b][Shimazaki 2005]; [c][Pezzolesi 2009b]; [d][Pezzolesi 2009a]; [e][Mooyaart 2011]

between the *EPO* promoter polymorphism and the combined phenotype of ESRD and proliferative diabetic retinopathy remained genome-wide significant ($P=1\times10^{-9}$) after meta-analysis of the original and our data (Box 1).

The negative results were unexpected, since the combined set of UK-ROI and FinnDiane studies is substantially larger than most of the previous studies, and has high statistical power to detect the reported effect sizes. Thus, it is unlikely that our results are false-negative findings (type II error). These negative findings suggest that many of the previously reported associations with DN may instead be false-positive findings (type I error). We were also surprised to see the dramatic loss of significance for the *CPVL/CHN2* and *CARS* loci in the re-analyzed GoKinD US study, suggesting a spurious signal.

Box 1: What is known about the implicated genes?

EPO	*EPO* encodes erythropoietin, a plasma protein that regulates red cell production by promoting erythroid differentiation and initiating hemoglobin synthesis. Erythropoietin is produced in kidneys among other tissues [Tong 2008].

Our failure to replicate these associations underscores the need to apply stringent statistical thresholds of significance, maximize power through meta-analysis of all available data, and seek replication in independent samples, as has been previously proposed [NCI-NHGRI Working Group on Replication in Association Studies 2007, McCarthy 2008]. It remains unclear whether the genetic risk factors for DN are shared between T1D and T2D.

7.2 GWAS of diabetic nephropathy

In order to systematically search for susceptibility loci for diabetic kidney complications in T1D, in Publication II, we performed a meta-analysis of three large GWASs: the FinnDiane, UK-ROI and GoKinD US studies, with a total of 6,691 T1D patients in the analysis. To validate the main findings, we further analyzed the most significant SNPs in 5,156 additional T1D patients from nine studies of T1D. In order to gain supporting evidence and biological understanding of the observed associations, we performed additional *in silico* and *in vitro* analyses that are summarized in Figure 15.

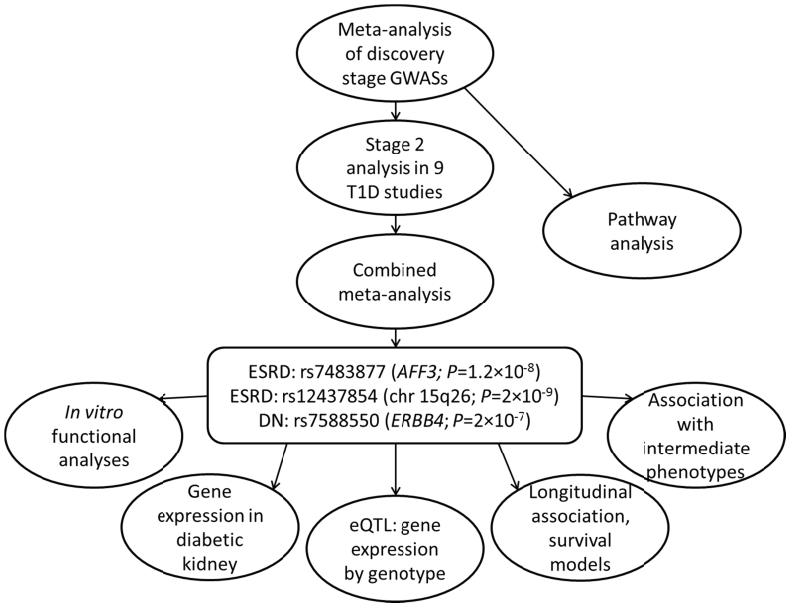


Figure 15: Flow chart of the key analyses that were performed in Publication II.

7.2.1 Discovery stage meta-analyses

Our primary phenotype of interest was DN, defined by the presence of persistent macroalbuminuria or ESRD. In addition, we analyzed a more extreme ESRD phenotype where cases with ESRD were compared to the rest of the T1D patients. Meta-analysis of the three discovery stage GWASs (FinnDiane, UK-ROI and GoKinD US) resulted in five independent signals with a suggestive $P<10^{-5}$ for the DN phenotype. In the meta-analysis of ESRD, rs7583877 on chromosome 2q11.2-q12 in

the *AFF3* gene reached genome-wide significance (P=4.8×10^{-9}). In addition, six other independent loci achieved a suggestive $P < 10^{-5}$ for association with ESRD (Figure 16).

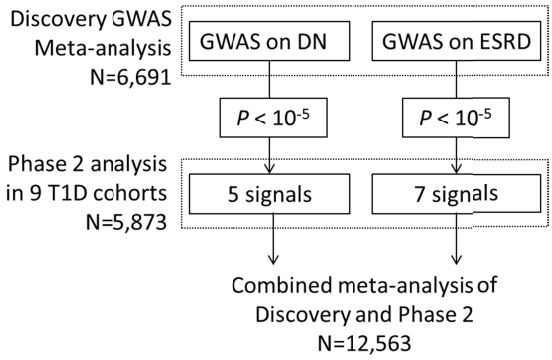


Figure 16: Two stage study design in Publication II.

7.2.2 Phase two analysis in nine additional T1D studies

All loci with $P < 10^{-5}$ were selected for the second stage analysis comprising of nine additional studies including 5,873 patients with T1D. The two-stage study design is illustrated in Figure 16. For the DN phenotype, an intronic SNP rs7588550 in the *ERBB4* gene had consistent protective effects in the stage two samples (OR 0.67, 95% CI 0.49 – 0.92, P=0.01) and the statistical significance was improved when the results from the discovery and the second stage were combined (OR 0.66, 95% CI 0.56-0.77, P=2.1×10^{-7}, Table 9).

For the analysis of ESRD, the association between the rs7583877 in *AFF3* and ESRD was not significant in the second phase studies alone, but retained genome-wide significance in the meta-analysis of the discovery studies and the second phase cohorts (OR=1.29, 95% CI 1.18-1.40, P=1.2×10^{-8}, Table 9). On chromosome 15q26 between the *RGMA* and *MCTP2* genes, rs12437854 reached genome-wide significance for association with ESRD in the combined meta-analysis (OR 1.80, 95% CI 1.48-2.17, P=2.0×10^{-9}, Table 9).

Table 9: Results from discovery, second stage and combined meta-analysis for supported markers. Modified from Publication II.

		Discovery		Stage 2		Combined	
SNP (Gene)	Fr(EA)	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value
ESRD							
rs12437854 (*RGMA-MCTP2*)	0.04	1.72 (1.36-2.18)	7.6×10^{-6}	1.95 (1.41-2.7)	5.4×10^{-5}	1.80 (1.48-2.17)	2.0×10^{-9}
rs7583877 (*AFF3*)	0.29	1.34 (1.22-1.48)	4.8×10^{-9}	1.11 (0.93-1.34)	0.25	1.29 (1.18-1.40)	1.2×10^{-8}
DN							
rs7588550 (*ERBB4*)	0.05	0.65 (0.55-0.79)	5.3×10^{-6}	0.67 (0.49-0.92)	0.01	0.66 (0.56-0.77)	2.1×10^{-7}

Fr(EA): The effect allele frequency.

7.2.3 Refined analysis of the affected phenotypes

Somewhat surprisingly, the strongest associations (in terms of the smallest *P*-values) were obtained for the ESRD phenotype, despite the smaller number of cases in the analysis compared with the DN phenotype. According to the liability model, the subjects with an extreme phenotype – ESRD in this case – are likely to carry more genetic risk factors [Gibson 2012]. In line with that idea, DN has traditionally been viewed as a continuous trait commencing with microalbuminuria, progressing to macroalbuminuria, and in an extreme case, culminating in ESRD. Recently, this paradigm has been called into question, with a suggestion that the syndrome may perhaps be composed of varying phenotypes [Kramer 2003, Perkins 2005]. In order to assess which sub-phenotype is the most affected by the identified genetic markers, we compared the effect sizes for the three associated loci using various case – control definitions for different stages of kidney disease as shown in Table 10. The signals in *AFF3* and on chromosome 15q26 showed the strongest effects for the ESRD case definition and were non-significant for macroalbuminuria, whereas the signal in *ERBB4* had similar effect size for both macroalbuminuria and ESRD.

Table 10: Association results for various case-control phenotypes for the three top signals. Results are for the meta-analysis of the three discovery cohorts. Non-significant associations are indicated with gray font. All odds ratios are given considering the minor allele as the effect allele. P-values are not directly comparable, as the number of samples and thus, the statistical power of the test varies between the comparisons.

	rs7583877 (*AFF3*)		rs12437854 (15q26)		rs7588550 (*ERBB4*)	
Analysis	OR (95%CI)	*P*	OR (95%CI)	*P*	OR (95%CI)	*P*
Normal AER vs.						
Macro	1.00 (0.90 – 1.11)	0.95	1.14 (0.86 – 1.50)	0.35	0.64 (0.51 – 0.81)	2.2×10^{-4}
ESRD	1.33 (1.19 – 1.48)	4.9×10^{-7}	1.82 (1.39 – 2.39)	1.6×10^{-5}	0.67 (0.53 – 0.84)	7.5×10^{-4}
Macro or ESRD	1.14 (1.05 – 1.25)	0.0023	1.43 (1.14 – 1.78)	0.0016	0.65 (0.55 – 0.79)	5.3×10^{-6}
Non-ESRD vs.						
ESRD	1.34 (1.22 – 1.48)	4.8×10^{-9}	1.72 (1.36 – 2.18)	7.7×10^{-6}	0.78 (0.63 – 0.97)	0.026

We then further tested the two loci associated with ESRD (rs7583877 in *AFF3*, rs12437854 on chromosome 15q26) for their association with relevant kidney endpoints using longitudinal time-to-event data for participants in the FinnDiane discovery collection. Consistent with our case-control analyses, the strongest association for rs7583877 was obtained for the time from T1D diagnosis to development of ESRD (hazard ratio (HR) 1.33, 95% CI 1.18-1.49), but also the time from T1D diagnosis to development of macroalbuminuria and the time from macroalbuminuria to ESRD reached nominal significance (Table 11). rs12437854 on chromosome 15q26 was associated with time from T1D diagnosis to development of macroalbuminuria (HR 1.31, 95% CI 1.03-1.67) and ESRD (HR 1.35, 95% CI 1.02-1.77).

An alternative explanation for an observed statistical association with ESRD might be an underlying association with survival. The mortality rates are extremely high in patients with kidney disease, with at least 25% of the patients with macroalbuminuria dying before they reach ESRD [Forsblom 2011]. Mortality in ESRD is 18-fold compared with T1D patients without albuminuria [Groop 2009]. Thus, the selection of patients with ESRD may be biased towards the patients who have stayed alive despite severe kidney disease. To assess the possibility of survival bias, we used the time until death as the final end point in the longitudinal analysis. Neither of the two studied ESRD loci was associated with mortality, suggesting that these loci are associated with ESRD per se rather than with survival (Table 11).

Table 11: Longitudinal analyses in the FinnDiane discovery cohort for rs7583877 (AFF3) and rs12437854 (15q26). P-values are not comparable between the analyses, as the sample numbers differ. The minor allele is the effect allele. HR = hazard ratio, 95% CI = 95% confidence interval. [a]Subjects that developed DN were censored out at the onset of DN. Modified from [Sandholm 2012].

	rs7583877		rs12437854	
Time-to-Event Analysis	HR (95% CI)	*P*	HR (95 % CI)	*P*
Time from T1D onset to micro	1.07 (0.97 - 1.18)	0.17	1.18 (0.92 - 1.51)	0.20
Time from T1D onset to macro	1.15 (1.04 - 1.27)	0.0065	1.31 (1.03 - 1.67)	0.030
Time from T1D onset to ESRD	1.33 (1.18 - 1.49)	1.9×10^{-6}	1.35 (1.02 - 1.77)	0.034
Time from macro to ESRD	1.16 (1.01 - 1.33)	0.040	1.16 (0.83 - 1.61)	0.38
Time from T1D onset to death[a]	0.97 (0.7 - 1.36)	0.88	0.39 (0.1 - 1.52)	0.17
Time from macro to death	1.05 (0.87 - 1.26)	0.60	1.03 (0.66 - 1.62)	0.88
Time from ESRD to death	1.09 (0.91 - 1.29)	0.35	0.86 (0.56 - 1.31)	0.48

7.2.4 **Association with intermediate phenotypes**

To explore whether the three SNPs contribute to DN/ESRD via related intermediate phenotypes, such as blood glucose, obesity, fasting lipid levels, or blood pressure, we explored the association results in publicly available GWAS datasets (Figure 17)[Willer 2008, Dupuis 2010, Heid 2010, International Consortium for Blood Pressure Genome-Wide Association Studies 2011].

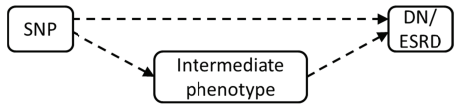


Figure 17: SNPs may affect the risk of kidney disease through intermediate phenotypes such as obesity, blood pressure, or blood lipid levels.

We found nominal, directionally consistent associations of rs12437854 with fasting glucose (*P*=0.03) [Dupuis 2010] and of rs7583877 with waist-hip ratio (*P*=0.04) [Heid 2010]. We also investigated if previously published genetic risk factors for T1D and for CKD were associated with DN or ESRD in our GWAS meta-analyses. Eight out of the 80 SNPs associated with T1D [Burren 2011] showed nominal significance

with DN or ESRD (including three at *AFF3* that are in weak LD (r^2 0.030 – 0.046 in CEU) with the SNPs described here), while no CKD SNPs were associated with DN or ESRD in our data [Kottgen 2009, Kottgen 2010, Chambers 2010]. The lack of association between the CKD SNPs and DN or ESRD in diabetes suggests that the genetic risk factors for DN differ from the genetic risk factors for CKD in the non-diabetic population.

7.2.5 Pathway analysis of GWAS results on DN and ESRD

To generate further biological hypotheses based on our GWAS results, we performed gene set enrichment analysis of Gene Ontology terms, KEGG and Ingenuity pathways and PANTHER database entries using the MAGENTA software [Segre 2010]. For the DN phenotype, genes related to "sugar binding" (P=0.0006), "double stranded DNA binding" (P=0.001) and "nucleic acid binding" (P=0.004) were enriched in the GWAS results. In the analysis of ESRD, the terms "sequence-specific DNA binding" (P=0.003), "positive regulation of transcription" (P=0.003), and "homeobox transcription factor" (P=0.004) were enriched for significant associations.

7.2.6 Exploration of the biological mechanisms

None of the three SNPs identified in the two-stage GWAS analysis are known to directly change the amino acid sequence of a known protein. Nevertheless, other SNPs in LD with these SNPs may directly affect the protein structure or regulate the gene expression of a nearby (or even distant) gene.

The *ERBB4* gene was found to be downregulated in tubulointerstitial enriched kidney biopsy tissue of patients with DN, compared with healthy kidney donors. In addition, the *SPAG16* gene near rs7588550 was upregulated in DN. *AFF3* was not differentially expressed in the kidney biopsies of DN vs. healthy subjects. Instead, near rs7583877 (intronic in *AFF3*) the expressions of the *LIPT1* and *TXNDC9* genes were upregulated, while *TSGA10* and *NPAS2* gene expressions were downregulated in tubulointerstitial and/or glomerular enriched kidney biopsies of patients with DN. No expression data were available for the two closest flanking genes for rs12437854 on chromosome 15q26, *RGMA* and *MCTP2* [Berthier 2009].

7.2.7 Discussion of the GWAS and meta-analysis on DN and ESRD

The GWAS meta-analysis performed in Publication II is the largest effort to date to define the genetic risk factors for DN in individuals with T1D. The main findings were the two genome-wide significant associations between ESRD and variants in *AFF3* and on the *RGMA – MCTP2* region, and a suggestive signal in the *ERBB4* gene with functional evidence (Box 2). Variants in and upstream *AFF3* have previously been associated with autoimmune diseases, including a suggestive association reported for T1D. Since the association between ESRD and rs7583877 in *AFF3* was not statistically significant in the follow-up cohorts, it is possible that the association is a false positive finding. However, this association was strong in both the FinnDiane and UK-ROI studies, and reached a genome-wide statistical significance

Box 2: What is known about the implicated genes?

AFF3	*AFF3* encodes a nuclear transcriptional activator that can bind to DNA and ribonucleic acid (RNA) [Melko 2011]. Variants upstream and in the 5' end of the *AFF3* gene have been suggestively associated with autoimmune diseases, including juvenile idiopathic arthritis [Hinks 2010], rheumatoid arthritis [Barton 2009], Graves' disease [Todd 2007] and T1D [Todd 2007]. Our meta-analysis suggested two association peaks for ESRD in the *AFF3* gene – the primary signal in the middle of the gene and a secondary signal in the 5' end of *AFF3*, close to the autoimmune disease signal. *In vitro* functional analyses on *AFF3* expression levels suggested that *AFF*3 may play a role in the TGF-β1-induced fibrotic responses of renal epithelial cells [Sandholm 2012].
ERBB4	*ERBB4* encodes a member of the epidermal growth factor receptor subfamily. *ERBB4* has been implicated in the development of cardiac, mammary gland and neural tissues [Gassmann 1995, Tidcombe 2003]. Mutations in *ERBB4* have been reported in cancer [Prickett 2009]. Research on kidney cell models and conditional *ERBB4* over-expression and knock-out mice suggest that *ERBB4* is important for the development of the kidneys [Zeng 2007, Veikkolainen 2012]. The gene expression studies in Publication III indicated co-expression with collagen related genes, suggesting that *ERBB4* may play a role in renal fibrosis.
RGMA	*RGMA* encodes a repulsive guidance molecule a (RGMa), an axon guidance protein on the retina [Monnier 2002]. Repulsive guidance molecules are also co-receptors in the bone morphogenetic protein signaling pathway that affects the tissue architecture across the body [Halbrooks 2007].
MCTP2	*MCTP2* gene encodes a transmembrane protein that binds Ca^{2+} and is involved in intercellular signal transduction [Shin 2005].

with $P = 4.8\times10^{-9}$. Of note, only a few studies on DN in T1D exist in the world with a substantial number of patients with ESRD, making the replication a challenging task. Among the 5,873 patients in the second phase studies, only 363 were included in the ESRD analysis case definition, resulting in a low statistical power to replicate the original finding. Despite the low number of cases, the meta-analysis of the stage two studies trended in the same direction with an OR of 1.11, and the meta-analysis including all the studies remained genome wide significant with $P = 1.2\times10^{-8}$. Longitudinal analyses in FinnDiane supported the role of rs7583877 especially in ESRD, whereas no association was observed for mortality given the stage of DN. Therefore, the signal is unlikely to originate due to survival bias.

The association on chromosome 15q26 was strongly associated with ESRD in the stage two analyses. Refined analyses on various phenotypes supported the

association with ESRD, but some evidence was also obtained for the combined DN phenotype and time from the diabetes onset until macroalbuminuria. The mechanism of this association might be through the glucose metabolism, as association was seen with fasting glucose levels [Dupuis 2010]. The rs12437854 is located on an intergenic chromosomal region, and none of the nearby genes have previously been linked to DN or related phenotypes. Therefore, the functional mechanism behind the association remains unclear.

Despite the larger number of cases in the analysis of DN, no association reached genome-wide significance ($P<5\times10^{-8}$) for DN in the discovery stage. This might be an indication of heterogeneity in patients with macroalbuminuria; defects in distinct biological processes may lead to onset of DN. In addition, the clinical definition of DN is based on a somewhat arbitrary cut-off of albuminuria. Nevertheless, rs7588550 in *ERBB4* was replicated in the second stage analyses with an OR of 0.67 highly similar to that found at the discovery stage (OR=0.65). The subsequent analyses supported the original hypothesis that the variant is associated with both macroalbuminuria and ESRD. Among the main findings, the *ERBB4* locus has the most biological support, as knock-out and over-expression mouse models have shown that the gene is important for the development of kidneys. Interestingly, ErbB4 has been suggested as a therapeutic target molecule for cancer and psychiatric and cardiovascular disorders, and ErbB4 binding ligands have already been patented for enhancing the ErbB4 signaling (Box 2) [Paatero and Elenius 2008].

7.3 GWAS on ESRD in women

In Publication III, we performed gender specific GWASs on ESRD in order to assess if gender specific genetic risk factors for ESRD exist in patients with T1D. The discovery GWAS analysis was performed in the FinnDiane study with 387 men and 258 women with ESRD.

7.3.1 GWAS identifies a susceptibility locus for ESRD in women with T1D

The GWAS on ESRD revealed two correlated SNPs ($r^2=1$) on chromosome 2q31.1 that were associated with ESRD in women with T1D with genome-wide significance and high odds ratios: rs4972593 with an OR of 2.39 (95% CI 1.75 – 3.25, $P=3.0\times10^{-8}$) and rs530673 with an OR of 2.38 (95% CI 1.75 – 3.23, $P=3.5\times10^{-8}$). Despite the 99% statistical power to detect the same association at rs4972593 in men with an $\alpha=0.05$ significance level, no association was observed ($P=0.78$, OR 0.97 (95% CI 0.78 – 1.21)), suggesting that the effect of the SNP is specific to women. No other loci reached $P < 5\times10^{-8}$ in men or women.

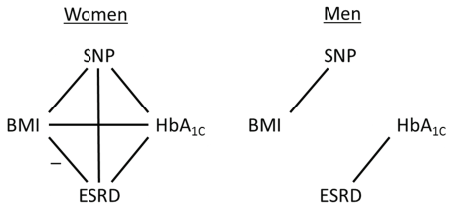


Figure 18: Observed associations in women and men between rs4972593, ESRD, and these clinical covariates that were associated with rs4972593. Inverse associations are marked with a minus sign.

7.3.2 Adjustment for covariates

We then asked if the gender specificity of the association between rs4972593 and ESRD could be due to gender differences in the intermediate phenotypes that may affect the risk of ESRD. Association testing with possible confounding factors revealed that rs4972593 was nominally associated with BMI and HbA_{1C} in women and with BMI in men. The observed associations between rs4972593, ESRD, and potential confounders are illustrated in Figure 18. Adjusting the model for BMI enhanced the association, and the observed association was only slightly attenuated when adjusted for both BMI and HbA_{1C} (Table 12). Importantly, the HbA_{1C} levels did not differ between men and women. Thus, it is unlikely that the gender specificity of the association between the SNP and ESRD would be driven by a gender related confounder.

Table 12: Association with rs4972593 and ESRD in women after adjustment for different covariates.

Adjustment model	OR	95% CI	*P*	N
Basic	2.39	1.76 - 3.23	1.8×10^{-8}	1,193
Basic + BMI	2.64	1.92 - 3.63	2.7×10^{-9}	1,123
Basic + HbA_{1C}	2.07	1.49 - 2.86	1.1×10^{-5}	1,144
Basic + BMI + HbA_{1C}	2.31	1.65 - 3.24	1.3×10^{-6}	1,110

Basic covariates: T1D duration, age, and the ten first principal components. To facilitate comparison, the results are not adjusted for the genome-wide inflation factor (λ=1.034).

7.3.3 Replication of the genetic association

To validate the finding, we tested the association at rs4972593 in three additional T1D cohorts with a substantial number of T1D women with ESRD (UK-ROI N=113; GoKinD US N=252; Italy N=68). Meta-analysis of the three replication studies resulted in a combined *P*-value of 0.02 for the replication in women (OR 1.41, 95% CI 1.05 – 1.90; Figure 19). The association effect was in the same direction in all replication studies, and the OR of 2.07 in the GoKinD US was similar to that in the FinnDiane discovery cohort. The association remained genome-wide significant after combined meta-analysis of the FinnDiane and replication cohorts ($P=3.9\times10^{-8}$, OR 1.81, 95% CI 1.47 – 2.24). No association was observed between rs4972593 and ESRD in men in the replication cohorts either ($P=0.90$) and the results remained

non-significant after combining the FinnDiane and the replication cohorts (P=0.78, OR=0.97, 95% CI 0.78 – 1.21).

7.3.4 Meta-analysis of three GWAS on ESRD

We further explored if other loci were associated with ESRD in a gender specific manner in a genome-wide meta-analysis of the FinnDiane, UK-ROI and GoKinD US GWAS data. However, the association between rs4972593 and ESRD in women remained the only signal with genome-wide significance (i.e. $P<5\times10^{-8}$).

7.3.5 Association with ESRD in women with T2D

We investigated if the association at rs4972594 could also be seen in women with T2D in the FIND study, including 570 African American, 165 European American, and 413 Mexican American diabetic women with ESRD. However, no association was observed between ESRD and rs4972593 or its proxies in any of the studied populations despite a good statistical power in European and Mexican Americans. Thus the reported association seems specific to women with T1D.

7.3.6 *In silico* analysis of the biological role of rs4972593

The associated SNPs rs4972593 and rs530673 are located in an intergenic region on chromosome 2q31.1 between the *SP3* and the *CDCA7* genes. *In silico* evaluation of the putative transcription factor binding sites at rs4972593 with the Genomatix software sute (Genomatix Software, GmbH, Munich, Germany) suggested that several transcription factor binding sites, e.g. for E-box and hypoxia inducible factors, are lost when a person carries the minor A allele of rs4972593. The rs530673 indicated potential regulatory activity in the RegulomeDB database as it was located in a DNase hypersensitivity peak and showed suggestive evidence of GATA2 and SMAD4 binding [Boyle 2012]. However, the genotypes of rs4972593, or of any other SNP in high LD with rs4972593, were not associated with gene expression levels of any gene within a 1 Mb region up- or down-stream of rs4972593 in the human HapMap3 lymphoblastoid cell line [Stranger 2012].

We further investigated if the gene-expression of the flanking genes *SP3* and *CDCA7* differs by gender and/or the DN status using publicly available gene expression databases. *CDCA7* was not significantly expressed in the renal tissue in studies of DN [Schmid 2006, Woroniecka 2011], but interestingly, *SP3* showed

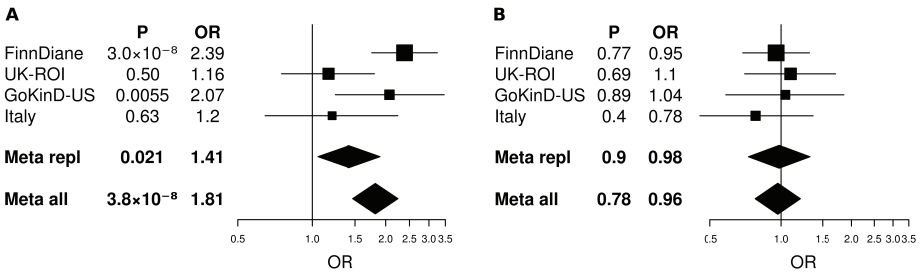


Figure 19: Forest plot of association between rs4972593 and ESRD in A) women and B) men.

higher level of gene expression in the glomeruli of women compared with men (*P*-value 0.004, fold change -1.45) [Woroniecka 2011].

7.3.7 Discussion of the findings

We identified SNPs on chromosome 2q31.1 between the *SP3* and *CDCA7* genes associated with ESRD in women with T1D, whereas no association was observed in men. *In silico* functional analyses suggested that the locus may affect transcription factor binding, but the target gene, possibly further away than the *SP3* and *CDCA7* genes, remains unknown. Based on the literature search, the *SP3* gene seems a plausible culprit as Sp3 is known to regulate expression of *CD2AP* that is important for the glomeruli, and can form a receptor complex with the estrogen receptor (Box 3).

Motivation for this sex-specific analysis arose from the notion that diabetic men are at higher risk of developing ESRD (inversely, women seem protected from ESRD compared with men), and that the main risk factors of ESRD have different effect size in men and women. On the other hand, the gender difference in the risk of ESRD is smaller in diabetic than in the non-diabetic population, and in some cases the female protection entirely disappears: the women who developed diabetes before 10 years of age, have as high risk of ESRD as the diabetic men [Harjutsalo 2011]. Regrettably, the sample sizes were too small to allow analysis stratified by sex and age at diabetes onset.

Box 3: What is known about the implicated genes?

SP3	*SP3* encodes the bi-functional Sp3 transcription factor that may either stimulate or repress the transcription of the target gene. One of the interesting Sp3 target genes is the *CD2AP* gene that encodes a protein that interacts with two important glomerular slit diaphragm proteins – nephrin and podocin – and is essential for the glomerular filtration barrier [Shih 2001, Schwarz 2001]. The gender specificity of the observed association might be explained by the finding that Sp3 directly interacts with the estrogen receptor α (ERα), forming a receptor complex for estradiol. This Sp3/ERα receptor complex binds to GC-rich promoter regions, either activating or suppressing the expression of the target gene when estradiol is bound to the receptor. The Sp3/ERα complex targets for example the vascular endothelial growth factor A (*VEGFA*) gene [Stoner 2000, Stoner 2004], which has been proposed as a common pathogenic factor behind diabetic retinopathy and nephropathy [Tremolada 2007], and associated with glomerular filtration rate in non-diabetic individuals [Kottgen 2010].
CDCA7	*CDCA7* has been suggested to affect the c-Myc mediated cell transformation. In special, the *CDCA7* overexpression enhances the transformation of lymphoblastoid cells. *CDCA7* is often overexpressed in human cancers [Osthus 2005].

This association between the *SP3* and *CDCA7* genes was not observed in the previous Publication II because the non-stratified analysis including women and men – even when adjusted for sex – only resulted in a moderate *P*-value of 7×10^{-5}. This finding highlights the importance of analyzing more homogenous groups of patients separately, despite the loss of statistical power when fewer samples are included in the analysis. In future, larger study sets may reveal even more gender-specific genetic risk factors for ESRD. This is also a step towards the idea of personalized medicine, where the treatment is based on the characteristics of the patient.

7.4 GWAS on albuminuria

Albuminuria is often the first clinically detectable manifestation of diabetic nephropathy. The strongest results in Publications II and III were found for the ESRD phenotype, but so far there are no large scale genetic association studies for the albumin excretion phenotype in T1D. Furthermore, heritability of AER has been evaluated in individuals without diabetes or in patients with T2D, but not in patients with T1D. In Publication IV we therefore performed a GWAS and estimated the heritability of AER in T1D.

7.4.1 Heritability of AER

We first estimated the narrow-sense heritability of AER, defined as the proportion of variability of AER that can be explained by the additive effects of the SNPs on the employed genotyping platform. Correlating the relatedness of the subjects based on the GWAS data, and the similarity of their AER values, we estimated that the directly genotyped GWAS SNPs explain 27% of the total AER variability. This estimate is in accordance with the earlier, family based estimates of AER heritability in non-diabetic subjects or patients with T2D. When AER was adjusted for the main known risk factors – age at diabetes onset, duration of diabetes, sex and use of AHT medication – the genetic factors explained 38% of the remaining AER variability.

7.4.2 GWAS on AER

The relation between AHT medication and AER is bidirectional and complex: AHT mediation is prescribed in response to elevated AER with the objective to reduce or at least slow down the increase of the AER. However, the effect of AHT medication depends on many factors and varies between individuals, and no general estimates exist for how much the AHT medication lowers AER. To reduce the bias due to AHT medication, we stratified the GWAS on AER by the use of AHT medication, and the results from the two strata were combined together with meta-analysis. This meta-analysis of 1,925 FinnDiane patients revealed five SNPs in the *GLRA3* gene on chromosome 4q34.1 with a genome-wide significant *P*-value ($P<1.5\times10^{-9}$ for rs10011025). A total of 62 distinct genetic loci reached a *P*-value $< 1\times10^{-4}$.

7.4.3 Replication of the putative susceptibility loci for AER

A total of 64 SNPs at 62 loci with *P*-value less than 10^{-4} were selected for replication in seven studies with 3,750 patients with T1D and data on AER on ACR. The strongest replication was seen for rs2410601 on chromosome 8p22 between the *PSD3* and *SH2D4A* genes (*P*=0.026). Meta-analysis of the FinnDiane and replication studies improved the discovery stage *P*-value from 2.5×10^{-5} to 3.9×10^{-6}.

In the *GLRA3* gene, the rs1564939 with $P = 8.4\times10^{-9}$ in the discovery stage reached a nominally significant *P*-value of 0.04 in the replication studies. However, the association was in the opposite direction compared with the discovery cohort. Interestingly, in the Finnish replication cohort the effect was in the same direction as the discovery study, with the minor C allele associated with higher AER (*P*=NS). A subsequent meta-analysis of the non-Finnish subjects had the *P*-value 0.03, with the minor C allele associated with lower AER.

7.4.4 LD structure and targeted sequencing of the *GLRA3* susceptibility locus

We first assessed if regional differences in the linkage disequilibrium (LD) structure i.e. the correlation of the SNPs could explain the opposite effect directions. Comparison of the LD structure of the common SNPs around the associated region showed no differences between the FinnDiane discovery samples and individuals of European origin (HapMap II, CEU population). We then hypothesized that population specific rare variants may constitute a synthetic association observed at rs1564939: if one or more rare variants are by chance more often inherited in the same haplotype with a common SNP, the common SNP will show evidence of association (Figure 20)[Dickson 2010]. The rare variants would not be observed on the LD plot of common variants. Moreover, synthetic associations may show inconsistent effects between populations if the causal rare variants are population specific. Importantly, the Finnish population has undergone population bottlenecks and been genetically isolated, leading to different rare variants in Finland than in the rest of Europe [Norio 2003a, Norio 2003b].

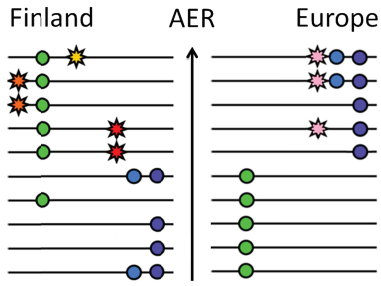


Figure 20: Illustration of a synthetic association arising from population specific rare variants. Lines represent DNA sequences that are ordered according to the AER value of the subject. Common variants are depicted with circles, whereas stars illustrate rare variants. Modified from [Anderson 2011].

In order to study the rare variants near the association signal in the *GLRA3* gene, we sequenced an 11 kb region around rs1564939 and rs10011025 in 48 FinnDiane and 48 UK subjects. Within the sequenced region, 43 SNPs were observed in the Finnish population, of which two were novel SNPs (ss647894785, ss647894811), whereas 38 SNPs were found in the UK population. However, these identified SNPs do not directly change the amino acid sequence of the GLRA3 protein structure.

7.4.5 Putative susceptibility locus in patients without AHT medication

We also performed a sub-analysis including only subjects with no AHT medication, with the aim to avoid bias caused by differences in the AHT medication. In this analysis, rs2097443 located between the *PARVG* and *LDOC1L* genes was associated with albuminuria with a combined *P*-value of 0.02 in the replication cohorts. Of note, the association was significant in two out of three replication studies with SNP data for rs2097443 (*P*-values 0.028 and 0.04 for the NFS-ORPS and UK-ROI studies, respectively) and the third replication study showed an effect in the same direction as well.

7.4.6 Comparison with results for non-diabetic subjects and patients with T2D

A previous GWAS on albuminuria identified a non-synonymous SNP rs1801239 (i.e. the SNP changes the resulting protein structure) in the *CUBN* gene associated with ACR in non-diabetic subjects [Boger 2011]. However, in our GWAS, no association was seen in patients with T1D with either AER or ACR (*P*=0.61 and *P*=0.72, respectively). We also tested if the 64 SNPs selected for replication in our GWAS showed evidence of association in the 63,153 non-diabetic individuals examined by Boger *et al.*, but none of the loci were significant after adjustment for multiple testing. Moreover, none of the 64 SNPs are located in the linkage peaks reported for ACR in T2D [Krolewski 2006]. The lack of overlap with the loci for T2D may be due to differences in the linkage and association study settings, but the lack of overlap with non-diabetic population supports the assumption that the genetic background of albuminuria is different in diabetes than in non-diabetic subjects.

7.4.7 Pathway analysis of the GWAS on AER

The gene set enrichment analysis performed with the MAGENTA software suggested enrichment of signals in genes related to natural killer cell mediated immunity ($P=8\times10^{-6}$, false discovery rate (FDR) = 0.003). The gene set overrepresentation analysis of the PANTHER pathways in the GWAS data highlighted the role of the metabotrobic glutamate receptor group 1 pathway ($P=6.7\times10^{-5}$, $P=0.012$ after correction for multiple testing). Both metabotrobic glutamate receptor (mGluR) group 1 members, mGluR1 and mGluR5, are expressed in mouse podocytes and their activation was shown to protect against albuminuria and podocyte apoptosis [Gu 2012].

7.4.8 Discussion

Publication IV reports the first GWAS on albuminuria in T1D, with variants in the *GLRA3* showing strong evidence of association. However, in the replication cohorts the association was observed in the opposite direction. The discrepancy in the effect direction may be due to false positive finding in the discovery cohort (type I error), and many of the 64 SNPs with $P < 10^{-4}$ selected for replication are likely to be false positive findings. However, the original association in *GLRA3* had strong statistical evidence in both directly genotyped and imputed SNPs (smallest *P*-value = 2×10^{-9}). The replication indicated only nominal statistical significance for association (P=0.04), not sufficient for significance after correction for multiple testing. The difficulty to reach significant *P*-values in replication may be due to the variability of the albuminuria phenotype, employment of different measures of albuminuria, or differences related to the age and disease severity between the replication studies. To reduce the methodological variability, we validated the high correlation between AER and ACR in our data and showed the robustness of the GWAS results for both phenotypes. In addition, we employed the mean of multiple albuminuria measurements when available.

A third explanation for the opposite effect direction is the synthetic association theory stating that common variants may reflect the effects of multiple rare variants that happen to be disproportionally distributed between the alleles of the common SNP [Dickson 2010]. If the rare variants are population specific, the association observed at the common SNP may be inconsistent across populations. Our sequencing effort identified more variants in the Finnish than in the UK subjects supporting the possibility of synthetic association, but none of them directly affect the protein structure. With the 96 individuals sequenced in the Finnish and the British populations, we had > 99% statistical power to detect variants with MAF ≥ 0.05. However, with 48 individuals sequenced in each group, we only had moderate statistical power of 62% to detect population specific rare variants with MAF = 0.01, suggesting that we may have missed many population-specific non-common variants. Moreover, synthetic associations may be constituted of SNPs much more distant than the sequenced 11 kb region, and consequently, Dickson *et al.* suggest that largescale sequencing efforts are required to detect the rare variants behind the synthetic associations [Dickson 2010].

Finally, this GWAS suggested other loci that are of interest for future studies: The strongest replication was obtained for rs2410601 on chromosome 8p22 between the *PSD3* and *SH2D4A* genes. Interestingly, Sh2d4a encoded by the *SH2D4A* is localized in the podocyte slit diaphragm, making it a strong biological candidate for albuminuria in diabetes [Patrakka 2007]. When only subjects without AHT medication were considered, the strongest replication was found for rs2097443 between the *PARVG* and *LDOC1L* genes. The overrepresentation analysis of the PANTHER pathways in the GWAS results highlights the mGluR group 1 proteins mGluR1 and mGluR5, which are expressed in the podocytes as well. The genes

encoding for these proteins had only moderate *P*-values in the GWAS and were thus not selected for replication. Putative association signals in the genes encoding for these mGluRs and other molecules in the pathway are of interest for further studies (Box 4).

Box 4: What is known about the implicated genes?

GLRA3	*GLRA3* encodes the α3 subunit of the glycine receptor (GlyR), best known for its function in the nervous system, but also found elsewhere in the body. In the pancreatic α-cells, activation of the GlyR by glycine results in release of glucagon from the α-cells; glucagon is a hormone with an opposite effect of insulin (Figure 21) [Li 2013]. Glycine has been shown to protect kidneys in ischemia, but it is not known if the effect is mediated through the GlyR in the kidneys [Yin 2002, den Eynden 2009] .
SH2D4A	*SH2D4A* is expressed in the glomerular podocytes in the slit diaphragms and co-localizes with nephrin, an essential protein in the podocytes [Patrakka 2007].
GRM1 and *GRM5*	Metabotrobic glutamate receptor (mGluR) group 1 includes two proteins, mGluR1 and mGluR5, encoded by *GRM1* and *GRM5*. Glutamate is a neurotransmitter, but it is an extracellular signaling mediator in the non-nervous tissue as well. Both mGluR1 and mGluR5 are expressed in the podocytes. Activation of the two proteins by a mGluR1/5 selective agonist (S)-3,5-dihydroxyphenylglycine (DHPG) attenuated proteinuria and protected from podocyte apoptosis in proteinuric mice [Gu 2012] .

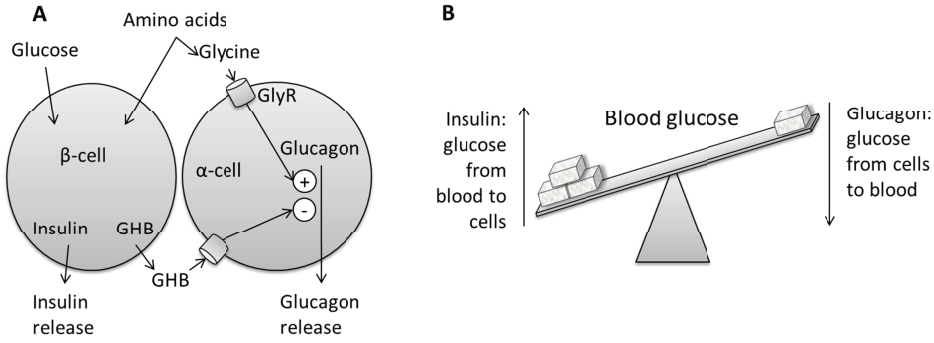


Figure 21: Glycine receptor (GlyR) is active in the pancreatic α-cells, where it regulates the glucagon release in response to circulating glycine. Modified from [Li 2013]. B) Glucagon counterbalances the effect of insulin.

7.5 Data mining of the GWAS data

In Publication V we applied a novel data mining method, BoNB [Sambo 2012], on the FinnDiane GWAS data to detect additional loci associated with various DN and ESRD phenotype definitions. The BoNB is a supervised classification algorithm, based on learning an ensemble of Naïve Bayesian Classifiers that divide subject into cases and controls according to SNPs. The GWAS data are sampled repeatedly with Bootstrap sampling to obtain multiple slightly different data sets to ensure robustness of the results.

7.5.1 Loci identified with the BoNB algorithm

The BoNB algorithm reported eight SNPs that were repeatedly selected to the Naïve Bayes classifiers and improved the classification in the independent out-of-bag test sets. In the genotype permutation of these eight SNPs, five SNPs had significant marginal utility across the tested classifiers: rs2838302 intronic in the *SIK1* gene and rs12917114 between the *SEMA6D* and *SLC24A5* genes were predictors for ESRD when compared with all the other patients; rs12137135 between the *WNT4* and *ZBTB40* genes, rs1670754 upstream the *MAPRE1P2* pseudogene, and rs17709344 between the *RGMA* and *MCTP2* genes were predictors for ESRD when compared with patients with normal AER.

The five identified loci were then evaluated with more conventional association testing. All loci had a *P*-value < 10^{-4} in the genotypic association models (i.e. allowing that the effect of the three genotypes aa, Aa, and AA are independent of each other; Table 13). We further defined for each SNP the association model with the best statistical significance, considering the binary models of recessive, dominant, or additive association. The dominant model was the most significant model for rs17709344 between the *RGMA* and *MCTP2* genes, whereas the best *P*-value was obtained using the recessive model for the four other SNPs. The ORs for these models were remarkably high and varied from 2.5 to 5.2 (Table 13).

Table 13: Statistical testing of the loci selected with BoNB

SNP	Genes	*P* perm	*P* genotypic	Best model (*P*, OR)
rs12137135 [a]	*WNT4 – ZBTB40*	0.031	5.2×10^{-5}	Rec (1.3×10^{-5}, 3.1)
rs17709344 [a]	*RGMA – MCTP2*	0.031	2.6×10^{-5}	Dom (2.4×10^{-5}, 2.5)
rs1670754 [a]	*MAPRE1P2*	0.031	3.5×10^{-5}	Rec (7.7×10^{-6}, 3.4)
rs12917114 [b]	*SEMA6D – SLC24A5*	0.024	1.4×10^{-5}	Rec (4.8×10^{-6}, 3.2)
rs2838302 [b]	*SIK1*	0.031	5.4×10^{-5}	Rec (2.0×10^{-4}, 5.2)

P perm: *P*-value for marginal utility > 0 using genotype permutation. *P* genotypic: *P* value of association using genotypic model. Best model: association model with the best *P*-value, Rec = Recessive, Dom = Dominant.
[a] ESRD vs. normal AER; [b] ESRD vs. no ESRD phenotype.

7.5.2 Replication in independent studies

The five selected loci were further tested for association in three additional studies: the Steno, UK-ROI and GoKinD US studies. Using the genotypic association model,

significant associations (P<0.05) were obtained for rs12137135 (*WNT4 – ZBTB40*) in Steno (P=0.009) and rs12917114 (*SEMA6D – SLC24A5*) in UK-ROI (P=0.04). Combining the P-values of the replication cohorts, irrespective of the effect direction, resulted in significant P-values for rs17709344 (*RGMA – MCTP2; P*=0.01) and rs12917114 (*SEMA6D – SLC24A5; P*=0.005).

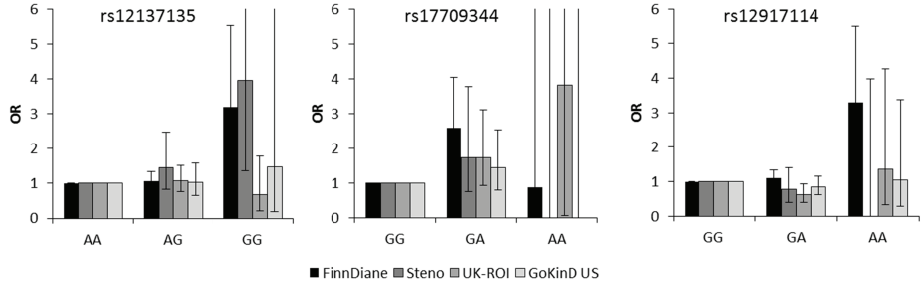


Figure 22: Odds ratios by genotype in the FinnDiane and replication studies for rs12137135 (WNT4 – ZBTB40), rs17709344 (RGMA – MCTP2) and rs12917114 (SEMA6D – SLC24A5). The major homozygous genotype is the reference genotype with OR = 1. Error bars indicate the 95% CI for the OR, and are truncated at 6.

Of note, the number of the rare genotype carriers of these three SNPs varied from 0 to 28 in cases and/ or controls. Consequently, the rare genotype effects varied greatly across the studies, and the 95% CIs were wide (Figure 22). Nevertheless, the recessive association model for rs12137135 (*WNT4 – ZBTB40),* defined as the best fitting model for the SNP in the FinnDiane study, was significantly replicated in Steno (P=0.008). Furthermore, the meta-analysis of the replication cohorts was significant for dominant association between ESRD and rs17709344 genotypes (*RGMA – MCTP2; P*=0.008).

7.5.3 Association with intermediate phenotypes

We tested in the FinnDiane and the Steno studies if the five selected SNPs were associated with intermediate phenotypes including HbA_{1C}, blood pressure, blood lipid measurements, and obesity related phenotypes. The rs17709344 (*RGMA – MCTP2*) was associated with serum LDL cholesterol in both studies, suggesting that the mechanism behind the observed association may be related to the lipid metabolism.

7.5.4 Discussion of the findings

The aim of the Publication V was to explore novel susceptibility loci for DN using advanced data mining methods to complement the findings of the conventional methods. We identified five susceptibility loci for ESRD, of which the locus between the *RGMA* and *MCTP2* genes was identified in Publication II in the meta-analysis of > 12,000 patients with T1D. This overlapping finding supports the idea of finding more signals with less samples if advanced data mining methods are used; In Publication II, the same SNP rs17709344 had a moderate P-value of 0.0006 in

FinnDiane alone, and thus, would not had been selected as a main signal without the large meta-analysis.

On the other hand, the earlier association with ESRD at *AFF3* (Publication II, $P<5\times10^{-8}$ in FinnDiane) was not observed in this study due to the imbalance between the number of analyzed cases and controls, leading to no predictive value for common SNPs of moderate effect size. Therefore, the applied method is complementary to the conventional GWAS methods, rather than replacing the existing methodology.

Similar to Publications II and III, the main findings originated from the ESRD phenotype. ESRD is a more clearly defined phenotype than macroalbuminuria, where the diagnosis is based on a somewhat arbitrary cut-off. It may also be that more or stronger genetic risk factors exist for the transition from macroalbuminuria to ESRD. A third possibility is that according to the liability model the patients with ESRD carry more risk factors as they represent the extreme end of the disease continuum, and thus, provide the most genetic findings.

Some evidence of replication was found for rs12137135 (*WNT4 – ZBTB40*), rs17709344 (*RGMA – MCTP2*) and rs12917114 (*SEMA6D – SLC24A5*) using the genotypic model of association. Replication was also attempted on the best fitting models. However, for four out of the five selected SNPs the model with the smallest *P*-value in the FinnDiane was a recessive model, and the number of rare homozygous genotypes was extremely low in cases and controls, ranging from 0 to 54. Thus, the statistical power of replication was low and the 95% confidence intervals were too wide for the interpretation of the results. However, the BoNB algorithm was designed to reduce the necessity of replication; the robustness of the findings is increased with the bootstrap sampling of the data so that signals were considered only if they were suggested at least by 5 out of 100 Naïve Bayes Classifiers. To be supported by a classifier, addition of the SNP has to improve the prediction performance of the independent out-of-bag test set, which can be considered as an internal replication cohort. To further validate the findings, the SNPs need to prove useful in the genotype permutation procedure. Therefore, all the five selected SNPs show high level of evidence and warrant future studies, even though not consistently replicated in independent cohorts. Of special interest is the *WNT4* gene 200 kb away from rs12137135, as the WNT signaling pathway has been implicated in DN (Box 5)[Zhou 2012].

Box 5: What is known about the implicated genes?

WNT4	*WNT4* is a member of the WNT gene family that encodes secreted signalling proteins. Wnt-4 is required for the development of renal tubules, and thus plays a critical role in renal morphogenesis [Stark 1994, Kispert 1998]. Wnt-4 is expressed in kidneys also during recovery after kidney injury [Surendran 2002] or acute renal failure [Terada 2003]. Wnt/beta-catenin signalling has been shown to affect survival of high glucose-stressed mesangial cells [Lin 2006], relating the kidney findings to diabetic setting as well.
SEMA6D	Semaphorins are best known for their involvement in the axon guidance, but *SEMA6D* is also involved in regulation of the late phase of T cell primary immune responses [O'Connor 2008].
RGMA	See Box 2, page 56.
MCTP2	See Box 2, page 56.

8 Conclusions and future prospects

In this thesis we have applied a wide range of computational methods on a large scale, genome-wide genotyping data on subjects with type 1 diabetes (T1D) with the aim to define the genetic risk factors behind diabetic kidney complications. As the most concrete results of this thesis, we identified four novel genetic loci affecting the risk of diabetic kidney disease in T1D: *AFF3* and *RGMA – MCTP2* associated with end-stage renal disease (ESRD), *CDCA7 – SP3* associated with ESRD in women with T1D, and *GLRA3* associated with albuminuria. These findings represent the first robust associations with genome-wide statistical significance for different stages of DN. In addition, we evaluated the previously reported suggestive susceptibility loci for diabetic nephropathy (DN), but found little evidence of association. The only exception was the *EPO* locus that remained genome-wide significantly associated with the combined ESRD and retinopathy phenotype after combination with our data. Furthermore, we identified putative susceptibility loci that warrant further evaluation: variants in the *ERBB4* gene were strongly although not genome-wide significantly associated with DN, and the *WNT4 – ZBTB40* and *SEMA6D – SLC24A5* loci were associated with ESRD using advanced data mining methods. Pathway analyses, on the other hand, suggested that the metabotrobic glutamate receptors *GRM1* and *GRM5* could be interesting candidate genes for albuminuria.

Identification of the loci is the first step of genetic discovery, but still far away from the ultimate aim of discovering new therapeutic target molecules or biomarkers for DN, let alone personalized medicine or prediction of disease risk based on the individuals' genetic profile. For convenience, the genetic loci are annotated in this thesis according to the gene/genes closest to the association signal. However, the current understanding of genetics suggests that the closest gene is not necessarily the causal one, even when the association signal is located within a gene [Smemo 2014]. The strongest support for the causality of a gene is found for the association signal intronic in the *ERBB4* gene. *ERBB4* gene expression is lower in the kidneys of patients with DN compared with healthy controls, and *ERBB4* knock-out and overexpression mice have demonstrated that the gene is important for the development of the kidneys. The continuously improving *in silico* analyses and annotation tools and biological databases can be useful in the search for the culprit gene of the other association signals. Furthermore, improvements in the genotype imputation reference panel, especially from the 1000 Genomes project, can refine the localization of the association signals, thus, helping the interpretation of the

functional mechanism behind the associations. Nevertheless, *in vivo* and/or *in vitro* studies are important for the evaluation of the mechanisms and validation of the functional hypotheses resulting from the *in silico* methods.

Future directions

For many other common diseases, the key factor for additional genetic findings has been the increase in the number of studied subjects. For example, the largest GWAS meta-analysis on T1D was based on over 7,500 cases and 9,000 controls and resulted in 41 susceptibility loci for T1D [Barrett 2009b]; for T2D, the largest GWAS meta-analysis of 8,000 cases and 40,000 controls brought the number of T2D susceptibility loci to 38 [Voight 2010], and even more loci have been identified with genome-wide-scale candidate gene approach. However, it is challenging to obtain such a number of patients for the analysis of DN, as DN is a "disease within a disease": both cases and controls should have diabetes. Therefore, the largest theoretical number of subjects for the analysis of DN in T1D will always remain smaller than for T1D. The largest GWAS meta-analysis on DN in T1D to date, presented in Publication II, included 7,300 patients in the discovery stage.

Larger GWAS meta-analyses on DN are underway, but the brute force of large numbers is not the only way forward. Publication III, identifying a susceptibility locus for ESRD in women with T1D, was a good example that novel signals can be found by dividing the patients to plausible sub-groups of more homogenous subjects, despite reducing the number of samples. Improving the phenotypic data and searching the susceptibility loci from different angles as in this thesis may help to identify additional genetic risk factors for DN and common diseases in general.

Complex diseases, such as DN, are affected by complex biological processes. It is likely that multiple genetic factors increase the propensity for such a disease and also interact with each other. Ideally, all the genetic factors should be evaluated jointly to fully capture the interaction effects. Multimarker data mining methods exist, but the vast majority of the published GWASs are restricted to the conventional single-marker analysis. Due to the limited number of available patients in the GWASs on DN, advanced data mining methods are particularly interesting for suggesting novel genetic risk factors for DN. In Publication V, we were able to detect an association on the *RGMA-MCTP2* region in the FinnDiane GWAS of 3,450 patients, whereas in Publication II, the signal became evident only after the meta-analysis of more than 10,000 patients. Apart from the *RGMA – MCTP2* locus, we suggest four other susceptibility loci for DN identified with the BoNB algorithm. Despite the lack of genome-wide statistical significance, we believe that these are of potential interest to the scientific community and propose new research avenues to evaluate and validate the role of these variants in DN.

A major challenge for the genetic discovery in DN is the availability of replication cohorts. Ideally, the replication step should contain at least the same amount of samples as the discovery step, and preferably twice the number. However, studies on

DN are often rather small, including some hundreds of subjects in total. This may, in part, explain why only a few candidate gene associations have been robustly replicated: small replication studies with low statistical power may result in false negative findings i.e. lack of replication. Especially the number of patients with the most severe form of DN, ESRD, is rarely over one hundred in a study. This is in contrast with our notion that the ESRD phenotype has provided the most susceptibility loci for DN in T1D. The study design for more advanced DN related phenotypes should take into account the limited replication possibilities and consider other alternatives such as permutation or cross-validation for the validation of the signals.

The four identified susceptibility loci for ESRD and albuminuria are unlikely to be the only genetic risk factors for DN. Despite the large number of genetic risk factors catalogued for many other common diseases such as T1D and T2D, the identified loci still explain only a small proportion of the estimated heritability. One suggested source of the "missing heritability" of the common diseases are multiple rare variants, potentially even located in the same genes that harbor common risk factors identified with the GWASs; there may be multiple ways to break a protein, but the most severe defects will remain rare as there would be negative selection against them in the evolution. Extending on this idea, the synthetic association theory suggests that some of the common variants identified in GWASs may actually reflect the combined effects of multiple rare variants that are by change inherited together with the common SNP allele. We hypothesize that the association signal for albuminuria in the *GLRA3* gene in Publication IV is an example of such synthetic association, explaining the inconsistent direction of effect between the Finnish and non-Finnish subjects.

While some of the rare variants may be indirectly detected with the GWAS approach, the GWAS platforms were originally designed to detect common variants, whereas direct sequencing provides a better study setting to detect the rare variation. Next generation sequencing approaches, and the whole exome sequencing studies in special, are now emerging to detect the rare causal variants behind the common diseases. Protein coding exon sequences constitute only approximately 1-2% of the human genome, but mutations on those regions may directly alter or truncate the protein structure. Therefore they are a plausible place to search for rare variants with large effect size.

In addition to the four novel identified susceptibility loci for ESRD and albuminuria, this thesis discusses many novel concepts and hot topics in the genetics of common disease, ranging from the gender specific risk factors – and risk factors for more homogenous patient groups in general – to the theory of rare variants contributing to the synthetic associations and to the use of advanced data mining methods for the detection of additional genetic risk factors.

9 Acknowledgements

This thesis is a result of collaboration between the Department of Biomedical Engineering and Computational Science (BECS) at the Aalto University School of Science, and the Folkhälsan Research Center. I wish to acknowledge Professors Jouko Lampinen and Ari Koskelainen, the heads of BECS, Professor Anna-Elina Lehesjoki, the head of Folkhälsan Research Center, and Professor Eero Honkanen, the head of the Department of Medicine, Division of Nephrology at Helsinki University Central Hospital for providing the research facilities and making my cross-disciplinary thesis possible.

I am grateful to my thesis supervisor Professor Kimmo Kaski at BECS for his vision and enthusiasm, as he was one of the key persons to launch the study program on bioinformation technologies, and has guided and supervised me on my scientific road from Otaniemi to Frankfurt and to Biomedicum. I also wish to give my warmest thanks to my thesis advisor Professor Per-Henrik Groop, the head of the FinnDiane Study Group at Folkhälsan Research Center, for the inspiration, advice and support, and for opening the door to the wonderful research community of diabetic nephropathy.

I followed the path from BECS to FinnDiane in the footsteps of Dr. Ville-Petteri Mäkinen, my second thesis advisor. I warmly thank Ville for being such a good role model of an engineer in the medical world. I would also like to express my deepest gratitude for teaching me so much about writing, and not sparing the red ink and suggestions when reading my early manuscript drafts.

Dr. Carol Forsblom and Dr. Valma Harjutsalo have been central figures for my research, both of them always working for the best our group, and always supporting and having time to answer my questions. I am grateful to Carol for teaching me so much about the clinical side of diabetes and its complications, and for being such an inexhaustible source of information and new study ideas. I also wish to thank Carol for his good company on the many trips on both sides of the Atlantic.

I wish to express my gratitude to the FinnDiane co-authors Aila, Aino, Anna S, Daniel, Emma, Janne, Jenny, Johan, Kustaa, Lena, Maikki, Markku L, Markku S, Milla, Nina, Outi, Nicu and Raija for their valuable contribution to this work. Warm thanks go also to Emma, Hanna, Maija, Miimu, Nadja, Nanna and Tomi with whom I have shared the office room and many interesting discussions. I am thankful for Maikki and her experience for skillfully running the lab, and I wish to thank all the present and past members of the FinnDiane group for making my thesis possible by seeing the patients and collecting the phenotypes and genotypes. I thank also the

two newer colleagues with a technical background, Erkka and Iiro, for all the help and scientific discussions. Special thanks to Anna-Reetta, Anna S, Asta, Chris, Heidi, Jaana, Jukka, Lina, Mari, Mervi, Mira, Nanne, Niklas, Rasmus and Tuula for the enjoyable coffee breaks and unforgettable conference trips, and for the inspiring atmosphere that we have.

I give my warmest thanks to Drs. Rany Salem and Amy Jayne McKnight for teaching me so much about computational genetics. We formed a determined and compact analysis team with great problem solving capabilities, and I would like to thank you for taking me part of that group. The Finnish nigthless summer nights have a totally different meaning with the two of you. Rany, now I can finally answer you that I will soon have my thesis done! I would also like to acknowledge the other members of the GENIE consortium for the scientific discussions. You were instrumental for making my thesis what it finally became.

Our collaboration with the Italians started at a SUMMIT plenary meeting where Dr. Francesco Sambo presented his visions and new data mining methods. I want to thank Francesco, as well as Dr. Alberto Malovini and Dr. Monica Stavarachi for the enjoyable weeks in Helsinki when science was done, and for persevering through the manuscript writing process. It was a delight to work with you three! I also thank Francesco for the excellent tour around Venice, and I hope you have enjoyed your stays in Helsinki as much as I enjoyed Italy. I also wish to acknowledge the other members of the SUMMIT consortium for their support.

I want to thank also all the international colleagues participating in this work, especially Drs. Emma Ahlqvist, Harshal Deshmukh, Jason Cooper, Loredana Marcovecchio, and Maria Lajer, who have been running analyses and discussing science in numerous e-mails, teleconferences, skype calls, and face-to-face meetings.

This work was supported by the Folkhälsan Research Foundation, the Wilhelm and Else Stockmann Foundation, Liv och Hälsa Foundation, Helsinki University Central Hospital Research Funds (EVO), the Sigrid Juselius Foundation, the Signe and Ane Gyllenberg Foundation, Finska Läkaresällskapet, Novo Nordisk Research Foundation, Academy of Finland, Tekes, the European Union's Seventh Framework Program (FP7/2007-2013) for the Innovative Medicine Initiative under grant agreement n° IMI/115006 (the SUMMIT consortium) and the Finnish Graduate School for Computational Sciences.

I want to thank all my friends and especially Annika, Katriina, Laura, Maiju, Teija and Terttu for the enduring friendship, as well as Aino, Katri, Jenni and the rest of the extended bio'03 for the amazing Otaniemi experience and the life thereafter. I would like to express my deepest gratitude to my entire family and especially my mom and dad Päivi and Raimo, and my little brother Juha for all their love and help. I owe my warmest thanks to Philippe for always supporting me, even when this thesis took more than the planned two years.

Helsinki, June 27, 2014
Niina Sandholm

10 Bibliography

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, *et al.* 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56-65.

American Diabetes Association 2013. Economic costs of diabetes in the U.S. in 2012. *Diabetes Care* 36: 1033-1046.

Amin R, Widmer B, Prevost AT, Schwarze P, Cooper J, Edge J, *et al.* 2008. Risk of microalbuminuria and progression to macroalbuminuria in a cohort with childhood onset type 1 diabetes: prospective observational study. *BMJ (Clinical Research Ed.)* 336: 697-701.

Andersen AR, Christiansen JS, Andersen JK, Kreiner S & Deckert T 1983. Diabetic nephropathy in Type 1 (insulin-dependent) diabetes: an epidemiological study. *Diabetologia* 25: 496-501.

Anderson CA, Soranzo N, Zeggini E & Barrett JC 2011. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biology* 9: e1000580.

Araki S, Moczulski DK, Hanna L, Scott LJ, Warram JH & Krolewski AS 2000. APOE polymorphisms and the development of diabetic nephropathy in type 1 diabetes: results of case-control and family-based studies. *Diabetes* 49: 2190.

Ayers KL & Cordell HJ 2010. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology* 34: 879-891.

Baldi P, Brunak S, Chauvin Y, Andersen CA & Nielsen H 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics (Oxford, England)* 16: 412-424.

Balding DJ 2006. A tutorial on statistical methods for population association studies. *Nature Reviews.Genetics* 7: 781-791.

Banting FG, Best CH, Collip JB, Campbell WR & Fletcher AA 1922. Pancreatic extracts in the treatment of diabetes mellitus: preliminary report. *Canadian Medical Association Journal = Journal De L'Association Medicale Canadienne* 12: 141-146.

Barrett JC 2009a. Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harbor Protocols* 2009: pdb.ip71.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, *et al.* 2009b. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics* 41: 703-707.

Barton A, Eyre S, Ke X, Hinks A, Bowes J, Flynn E, *et al.* 2009. Identification of AF4/FMR2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further pan-autoimmune susceptibility genes. *Human Molecular Genetics* 18: 2518-2522.

Beamer BA, Yen CJ, Andersen RE, Muller D, Elahi D, Cheskin LJ, *et al.* 1998. Association of the Pro12Ala variant in the peroxisome proliferator-activated receptor-gamma2 gene with obesity in two Caucasian populations. *Diabetes* 47: 1806-1808.

Bentley DR 2000. The Human Genome Project--an overview. *Medicinal Research Reviews* 20: 189-196.

Berglund J, Lins PE, Adamson U & Lins LE 1987. Microalbuminuria in long-term insulin-dependent diabetes mellitus. Prevalence and clinical characteristics in a normotensive population. *Acta Medica Scandinavica* 222: 333-338.

Berthier CC, Zhang H, Schin M, Henger A, Nelson RG, Yee B, *et al.* 2009. Enhanced expression of Janus kinase-signal transducer and activator of transcription pathway members in human diabetic nephropathy. *Diabetes* 58: 469-477.

Boger CA, Chen MH, Tin A, Olden M, Kottgen A, de Boer IH, *et al.* 2011. CUBN is a gene locus for albuminuria. *Journal of the American Society of Nephrology : JASN* 22: 555-570.

Borch-Johnsen K, Norgaard K, Hommel E, Mathiesen ER, Jensen JS, Deckert T*, et al.* 1992. Is diabetic nephropathy an inherited complication. *Kidney Int* 41: 719-722.

Borch-Johnsen K & Kreiner S 1987. Proteinuria: value as predictor of cardiovascular mortality in insulin dependent diabetes mellitus. *British Medical Journal (Clinical Research Ed.)* 294: 1651-1654.

Borch-Johnsen K, Andersen PK & Deckert T 1985. The effect of proteinuria on relative mortality in type 1 (insulin-dependent) diabetes mellitus. *Diabetologia* 28: 590-596.

Botstein D & Risch N 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics* 33 Suppl: 228-237.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M*, et al.* 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* 22: 1790-1797.

Brenner BM, Cooper ME, de Zeeuw D, Keane WF, Mitch WE, Parving HH*, et al.* 2001. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *The New England Journal of Medicine* 345: 861-869.

Burren OS, Adlem EC, Achuthan P, Christensen M, Coulson RM & Todd JA 2011. T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research. *Nucleic Acids Research* 39: D997-1001.

Carlson R 2003. The pace and proliferation of biological technologies. *Biosecurity and Bioterrorism : Biodefense Strategy, Practice, and Science* 1: 203-214.

Chambers JC, Zhang W, Lord GM, van der Harst P, Lawlor DA, Sehmi JS*, et al.* 2010. Genetic loci influencing kidney function and chronic kidney disease. *Nature Genetics* 42: 373-375.

Conway BR & Maxwell AP 2009. Genetics of diabetic nephropathy: are there clues to the understanding of common kidney diseases? *Nephron.Clinical Practice* 112: c213-21.

Coonrod BA, Ellis D, Becker DJ, Bunker CH, Kelsey SF, Lloyd CE*, et al.* 1993. Predictors of microalbuminuria in individuals with IDDM. Pittsburgh Epidemiology of Diabetes Complications Study. *Diabetes Care* 16: 1376-1383.

Craig DW, Millis MP & DiStefano JK 2009. Genome-wide SNP genotyping study using pooled DNA to identify candidate markers mediating susceptibility to end-stage renal disease attributed to Type 1 diabetes. *Diabetic Medicine : A Journal of the British Diabetic Association* 26: 1090-1098.

Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ*, et al.* 2011. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *Lancet* 378: 31-40.

de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S & Voight BF 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics* 17: R122-8.

Del Bo R, Scarlato M, Ghezzi S, Maestroni A, Sjolind L, Forsblom C*, et al.* 2006. VEGF gene variability and type 1 diabetes: evidence for a protective role. *Immunogenetics* 58: 107-112.

den Eynden JV, Ali SS, Horwood N, Carmans S, Brone B, Hellings N*, et al.* 2009. Glycine and glycine receptor signalling in non-neuronal cells. *Frontiers in Molecular Neuroscience* 2: 9.

Diabetes Epidemiology Research International Group 1988. Geographic patterns of childhood insulin-dependent diabetes mellitus. *Diabetes* 37: 1113-1119.

Dickson SP, Wang K, Krantz I, Hakonarson H & Goldstein DB 2010. Rare variants create synthetic genome-wide associations. *PLoS Biology* 8: e1000294.

Ding W, Wang F, Fang Q, Zhang M, Chen J & Gu Y 2012. Association between two genetic polymorphisms of the renin-angiotensin-aldosterone system and diabetic nephropathy: a meta-analysis. *Molecular Biology Reports* 39: 1293-1303.

Doria A, Patti ME & Kahn CR 2008. The emerging genetic architecture of type 2 diabetes. *Cell Metabolism* 8: 186-200.

Doublier S, Lupia E, Catanuto P & Elliot SJ 2011. Estrogens and progression of diabetic kidney damage. *Current Diabetes Reviews* 7: 28-34.

Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU*, et al.* 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics* 42: 105-116.

ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C*, et al.* 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74.

Fagerudd JA, Pettersson-Fernholm KJ, Gronhagen-Riska C & Groop PH 1999. The impact of a family history of Type II (non-insulin-dependent) diabetes mellitus on the risk of

diabetic nephropathy in patients with Type I (insulin-dependent) diabetes mellitus. *Diabetologia* 42: 519-526.

Finne P 2010. Suomen munuaistautirekisteri vuosiraportti 2009.

Finne P, Reunanen A, Stenman S, Groop PH & Gronhagen-Riska C 2005. Incidence of end-stage renal disease in patients with type 1 diabetes. *JAMA : The Journal of the American Medical Association* 294: 1782-1787.

Fioretto P & Mauer M 2010. Diabetic nephropathy: diabetic nephropathy-challenges in pathologic classification. *Nature Reviews.Nephrology* 6: 508-510.

Forbes JM & Cooper ME 2013. Mechanisms of diabetic complications. *Physiological Reviews* 93: 137-188.

Forsblom C, Harjutsalo V, Thorn LM, Waden J, Tolonen N, Saraheimo M*, et al.* 2011. Competing-risk analysis of ESRD and death among patients with type 1 diabetes and macroalbuminuria. *Journal of the American Society of Nephrology : JASN* 22: 537-544.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA*, et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.

Freedman BI, Langefeld CD, Lu L, Divers J, Comeau ME, Kopp JB*, et al.* 2011. Differential effects of MYH9 and APOL1 risk variants on FRMD3 Association with Diabetic ESRD in African Americans. *PLoS Genetics* 7: e1002150.

Gassmann M, Casagranda F, Orioli D, Simon H, Lai C, Klein R*, et al.* 1995. Aberrant neural and cardiac development in mice lacking the ErbB4 neuregulin receptor. *Nature* 378: 390-394.

Gelman A, Carlin JB, Stern HS & Rubin DB 2004. Bayesian Data Analysis. 2nd edition ed. Taylor & Francis.ISBN 9781420057294.

Gibson G 2012. Rare and common variants: twenty arguments. *Nature Reviews.Genetics* 13: 135-145.

Groop PH, Thomas MC, Moran JL, Wadèn J, Thorn LM, Mäkinen VP*, et al.* 2009. The presence and severity of chronic kidney disease predicts all-cause mortality in type 1 diabetes. *Diabetes* 58: 1651.

Groop LC, Kankuri M, Schalin-Jantti C, Ekstrand A, Nikula-Ijas P, Widen E*, et al.* 1993. Association between polymorphism of the glycogen synthase gene and non-insulin-dependent diabetes mellitus. *The New England Journal of Medicine* 328: 10-14.

Gu L, Liang X, Wang L, Yan Y, Ni Z, Dai H*, et al.* 2012. Functional metabotropic glutamate receptors 1 and 5 are expressed in murine podocytes. *Kidney International* 81: 458-468.

Guan Y & Stephens M 2008. Practical issues in imputation-based association mapping. *PLoS Genetics* 4: e1000279.

Hadjadj S, Tarnow L, Forsblom C, Kazeem G, Marre M, Groop PH*, et al.* 2007. Association between angiotensin-converting enzyme gene polymorphisms and diabetic nephropathy: case-control, haplotype, and family-based study in three European populations. *Journal of the American Society of Nephrology* 18: 1284.

Hadjadj S, Pean F, Gallois Y, Passa P, Aubert R, Weekers L*, et al.* 2004. Different patterns of insulin resistance in relatives of type 1 diabetic patients with retinopathy or nephropathy: the Genesis France-Belgium Study. *Diabetes Care* 27: 2661-2668.

Halbrooks PJ, Ding R, Wozney JM & Bain G 2007. Role of RGM coreceptors in bone morphogenetic protein signaling. *Journal of Molecular Signaling* 2: 4.

Halme M & Kajosaari M 2006. Kystinen fibroosi – harvinainen monielinsairaus. *Duodecim; Laaketieteellinen Aikakauskirja* 122: 1341-1346.

Harjutsalo V, Sjöberg L & Tuomilehto J 2008. Time trends in the incidence of type 1 diabetes in Finnish children: a cohort study. *The Lancet* 371: 1777-1782.

Harjutsalo V, Katoh S, Sarti C, Tajima N & Tuomilehto J 2004. Population-based assessment of familial clustering of diabetic nephropathy in type 1 diabetes. *Diabetes* 53: 2449-2454.

Harjutsalo V, Sund R, Knip M & Groop PH 2013. Incidence of type 1 diabetes in Finland. *JAMA : The Journal of the American Medical Association* 310: 427-428.

Harjutsalo V, Maric C, Forsblom C, Thorn L, Waden J, Groop PH*, et al.* 2011. Sex-related differences in the long-term risk of microvascular complications by age at onset of type 1 diabetes. *Diabetologia* 54: 1992-1999.

Hartley SW, Monti S, Liu CT, Steinberg MH & Sebastiani P 2012. Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction. *Frontiers in Genetics* 3: 176.

He B, Österholm AM, Hoverfält A, Forsblom C, Hjörleifsdóttir EE, Nilsson AS*, et al.* 2009. Association of genetic variants at 3q22 with nephropathy in patients with type 1 diabetes mellitus. *The American Journal of Human Genetics* 84: 5-13.

He Q & Lin DY 2011. A variable selection method for genome-wide association studies. *Bioinformatics (Oxford, England)* 27: 1-8.
Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E*, et al.* 2008. Investigation of the fine structure of European populations with applications to disease association studies. *European Journal of Human Genetics : EJHG* 16: 1413-1429.
Hedrick PW 1987. Gametic disequilibrium measures: proceed with caution. *Genetics* 117: 331-341.
Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V*, et al.* 2010. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics* 42: 949-960.
Hinks A, Eyre S, Ke X, Barton A, Martin P, Flynn E*, et al.* 2010. Association of the AFF3 gene and IL2/IL21 gene region with juvenile idiopathic arthritis. *Genes and Immunity* 11: 194-198.
Hoggart CJ, Whittaker JC, De Iorio M & Balding DJ 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics* 4: e1000130.
Hovind P, Tarnow L, Rossing K, Rossing P, Eising S, Larsen N*, et al.* 2003. Decreasing incidence of severe diabetic microangiopathy in type 1 diabetes. *Diabetes Care* 26: 1258-1264.
Howie BN, Donnelly P & Marchini J 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5: e1000529.
International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD*, et al.* 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478: 103-109.
International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL*, et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.
International HapMap Consortium 2005. A haplotype map of the human genome. *Nature* 437: 1299-1320.
International Human Genome Sequencing Consortium 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945.
Kangas T 2001. Diabeetikoiden ja verrokkien terveyspalvelujen käyttö ja kustannukset Helsingissä. 56: 1525-1531.
Kispert A, Vainio S & McMahon AP 1998. Wnt-4 is a mesenchymal signal for epithelial transformation of metanephric mesenchyme in the developing kidney. *Development (Cambridge, England)* 125: 4225-4234.
Knowler WC, Coresh J, Elston RC, Freedman BI, Iyengar SK, Kimmel PL*, et al.* 2005. The Family Investigation of Nephropathy and Diabetes (FIND): design and methods. *Journal of Diabetes and its Complications* 19: 1-9.
Kottgen A, Pattaro C, Boger CA, Fuchsberger C, Olden M, Glazer NL*, et al.* 2010. New loci associated with kidney function and chronic kidney disease. *Nature Genetics* 42: 376-384.
Kottgen A, Glazer NL, Dehghan A, Hwang SJ, Katz R, Li M*, et al.* 2009. Multiple loci associated with indices of renal function and chronic kidney disease. *Nature Genetics* 41: 712-717.
Kramer HJ, Nguyen QD, Curhan G & Hsu CY 2003. Renal insufficiency in the absence of albuminuria and retinopathy among adults with type 2 diabetes mellitus. *JAMA : The Journal of the American Medical Association* 289: 3273-3277.
Krolewski AS, Poznik GD, Placha G, Canani L, Dunn J, Walker W*, et al.* 2006. A genome-wide linkage scan for genes controlling variation in urinary albumin excretion in type II diabetes. *Kidney International* 69: 129-136.
Kruglyak L 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* 22: 139-144.
Leak TS, Perlegas PS, Smith SG, Keene KL, Hicks PJ, Langefeld CD*, et al.* 2009. Variants in intron 13 of the ELMO1 gene are associated with diabetic nephropathy in African Americans. *Annals of Human Genetics* 73: 152-159.
Lewis EJ, Hunsicker LG, Bain RP & Rohde RD 1993. The effect of angiotensin-converting-enzyme inhibition on diabetic nephropathy. The Collaborative Study Group. *The New England Journal of Medicine* 329: 1456-1462.
Li C, Liu C, Nissim I, Chen J, Chen P, Doliba N*, et al.* 2013. Regulation of glucagon secretion in normal and diabetic human islets by gamma-hydroxybutyrate and glycine. *The Journal of Biological Chemistry* 288: 3938-3951.

Li Y, Willer CJ, Ding J, Scheet P & Abecasis GR 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* 34: 816-834.

Li Y, Willer C, Sanna S & Abecasis G 2009. Genotype imputation. *Annual Review of Genomics and Human Genetics* 10: 387-406.

Lin CL, Wang JY, Huang YT, Kuo YH, Surendran K & Wang FS 2006. Wnt/beta-catenin signaling modulates survival of high glucose-stressed mesangial cells. *Journal of the American Society of Nephrology : JASN* 17: 2812-2820.

Lindholm E, Agardh E, Tuomi T, Groop L & Agardh CD 2001. Classifying diabetes according to the new WHO clinical stages. *European Journal of Epidemiology* 17: 983-989.

Maeda S, Osawa N, Hayashi T, Tsukada S, Kobayashi M & Kikkawa R 2007. Genetic variations associated with diabetic nephropathy and type II diabetes in a Japanese population. *Kidney International.Supplement* (106): S43-8.

Mäkinen V 2010. Computational Analysis of the Metabolic Phenotypes in Type 1 Diabetes and Their Associations with Mortality and Diabetic Complications. The doctoral dissertations of the Helsinki University of Technology.ISBN ISBN 978-952-60-3013-5.

Marchini J & Howie B 2010. Genotype imputation for genome-wide association studies. *Nature Reviews.Genetics* 11: 499-511.

Marcovecchio ML, Dalton RN, Schwarze CP, Prevost AT, Neil HA, Acerini CL*, et al.* 2009. Ambulatory blood pressure measurements are related to albumin excretion and are predictive for risk of microalbuminuria in young people with type 1 diabetes. *Diabetologia* 52: 1173-1181.

Maric C & Sullivan S 2008. Estrogens and the diabetic kidney. *Gender Medicine* 5 Suppl A: S103-13.

Marre M, Jeunemaitre X, Gallois Y, Rodier M, Chatellier G, Sert C*, et al.* 1997. Contribution of genetic polymorphism in the renin-angiotensin system to the development of renal complications in insulin-dependent diabetes: Genetique de la Nephropathie Diabetique (GENEDIAB) study group. *The Journal of Clinical Investigation* 99: 1585-1595.

Marre M, Chatellier G, Leblanc H, Guyene TT, Menard J & Passa P 1988. Prevention of diabetic nephropathy with enalapril in normotensive diabetics with microalbuminuria. *BMJ (Clinical Research Ed.)* 297: 1092-1095.

Marshall SM 2012. Diabetic nephropathy in type 1 diabetes: has the outlook improved since the 1980s? *Diabetologia* 55: 2301-2306.

Mathiesen ER, Hommel E, Giese J & Parving HH 1991. Efficacy of captopril in postponing nephropathy in normotensive insulin dependent diabetic patients with microalbuminuria. *BMJ (Clinical Research Ed.)* 303: 81-87.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP*, et al.* 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews.Genetics* 9: 356-369.

McDonough CW, Palmer ND, Hicks PJ, Roh BH, An SS, Cooke JN*, et al.* 2010. A genome-wide association study for diabetic nephropathy genes in African Americans. *Kidney International* .

McKnight A, Patterson C, Sandholm N, Kilner J, Buckham T, Parkkonen M*, et al.* 2010a. Genetic Polymorphisms in Nitric Oxide Synthase 3 Gene and Implications for Kidney Disease: A Meta-Analysis. *American Journal of Nephrology* 32: 476-481.

McKnight AJ, Patterson CC, Pettigrew KA, Savage DA, Kilner J, Murphy M*, et al.* 2010b. A GREM1 gene variant associates with diabetic nephropathy. *Journal of the American Society of Nephrology : JASN* 21: 773-781.

McKnight AJ, Maxwell AP, Fogarty DG, Sadlier D, Savage DA & Warren 3/UK GoKinD Study Group 2009. Genetic analysis of coronary artery disease single-nucleotide polymorphisms in diabetic nephropathy. *Nephrology, Dialysis, Transplantation : Official Publication of the European Dialysis and Transplant Association - European Renal Association* 24: 2473-2476.

McKnight AJ, Maxwell AP, Patterson CC, Brady HR & Savage DA 2007. Association of VEGF-1499C-->T polymorphism with diabetic nephropathy in type 1 diabetes mellitus. *Journal of Diabetes and its Complications* 21: 242-245.

Melko M, Douguet D, Bensaid M, Zongaro S, Verheggen C, Gecz J*, et al.* 2011. Functional characterization of the AFF (AF4/FMR2) family of RNA-binding proteins: Insights into the molecular pathology of FRAXE intellectual disability. *Human Molecular Genetics* 20: 1873-1885.

Mi H, Muruganujan A & Thomas PD 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research* 41: D377-86.

Mi H & Thomas P 2009. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods in Molecular Biology (Clifton, N.J.)* 563: 123-140.

Moczulski DK, Rogus JJ, Antonellis A, Warram JH & Krolewski AS 1998. Major susceptibility locus for nephropathy in type 1 diabetes on chromosome 3q: results of novel discordant sib-pair analysis. *Diabetes* 47: 1164.

Mogensen CE, Chachati A, Christensen CK, Close CF, Deckert T, Hommel E*, et al.* 1985. Microalbuminuria: an early marker of renal involvement in diabetes. *Uremia Investigation* 9: 85-95.

Molitch ME, Steffes MW, Cleary PA & Nathan DM 1993. Baseline analysis of renal function in the Diabetes Control and Complications Trial. The Diabetes Control and Complications Trial Research Group [corrected. *Kidney International* 43: 668-674.

Mollsten A, Kockum I, Svensson M, Rudberg S, Ugarph-Morawski A, Brismar K*, et al.* 2008. The effect of polymorphisms in the renin-angiotensin-aldosterone system on diabetic nephropathy risk. *Journal of Diabetes and its Complications* 22: 377-383.

Monnier PP, Sierra A, Macchi P, Deitinghoff L, Andersen JS, Mann M*, et al.* 2002. RGM is a repulsive guidance molecule for retinal axons. *Nature* 419: 392-395.

Mooyaart A, Valk EJJ, van Es L, Bruijn J, de Heer E, Freedman B*, et al.* 2011. Genetic associations in diabetic nephropathy: a meta-analysis. *Diabetologia* : 1-10.

Mueller PW, Rogus JJ, Cleary PA, Zhao Y, Smiles AM, Steffes MW*, et al.* 2006. Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes. *Journal of the American Society of Nephrology : JASN* 17: 1782-1790.

Nathan DM, Zinman B, Cleary PA, Backlund JYC, Genuth S, Miller R*, et al.* 2009. Modern-day clinical course of type 1 diabetes mellitus after 30 years' duration: the diabetes control and complications trial/epidemiology of diabetes interventions and complications and Pittsburgh epidemiology of diabetes complications experience (1983-2005). *Archives of Internal Medicine* 169: 1307.

NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ*, et al.* 2007. Replicating genotype-phenotype associations. *Nature* 447: 655-660.

Neamat-Allah M, Feeney S, Savage D, Maxwell A, Hanson R, Knowler W*, et al.* 2001. Analysis of the association between diabetic nephropathy and polymorphisms in the aldose reductase gene in Type 1 and Type 2 diabetes mellitus. *Diabetic Medicine* 18: 906-914.

Nordwall M, Bojestig M, Arnqvist HJ, Ludvigsson J & Linkoping Diabetes Complications Study 2004. Declining incidence of severe retinopathy and persisting decrease of nephropathy in an unselected population of Type 1 diabetes-the Linkoping Diabetes Complications Study. *Diabetologia* 47: 1266-1272.

Norio R 2003a. Finnish Disease Heritage I: characteristics, causes, background. *Human Genetics* 112: 441-456.

Norio R 2003b. Finnish Disease Heritage II: population prehistory and genetic roots of Finns. *Human Genetics* 112: 457-469.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A*, et al.* 2008. Genes mirror geography within Europe. *Nature* 456: 98-101.

Nyholt DR 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics* 74: 765-769.

O'Connor BP, Eun SY, Ye Z, Zozulya AL, Lich JD, Moore CB*, et al.* 2008. Semaphorin 6D regulates the late phase of CD4+ T cell primary immune responses. *Proceedings of the National Academy of Sciences of the United States of America* 105: 13015-13020.

Osterholm AM, He B, Pitkaniemi J, Albinsson L, Berg T, Sarti C*, et al.* 2007. Genome-wide scan for type 1 diabetic nephropathy in the Finnish population reveals suggestive linkage to a single locus on chromosome 3q. *Kidney International* 71: 140-145.

Osthus RC, Karim B, Prescott JE, Smith BD, McDevitt M, Huso DL*, et al.* 2005. The Myc target gene JPO1/CDCA7 is frequently overexpressed in human tumors and has limited transforming activity in vivo. *Cancer Research* 65: 5620-5627.

Paatero I & Elenius K 2008. ErbB4 and its isoforms: patentable drug targets? *Recent Patents on DNA & Gene Sequences* 2: 27-33.

Pambianco G, Costacou T, Ellis D, Becker DJ, Klein R & Orchard TJ 2006. The 30-year natural history of type 1 diabetes complications: the Pittsburgh Epidemiology of Diabetes Complications Study experience. *Diabetes* 55: 1463-1469.

Parving HH, Hommel E & Smidt UM 1988. Protection of kidney function and decrease in albuminuria by captopril in insulin dependent diabetics with nephropathy. *BMJ (Clinical Research Ed.)* 297: 1086-1091.

Parving HH & Smidt UM 1986. Hypotensive therapy reduces microvascular albumin leakage in insulin-dependent diabetic patients with nephropathy. *Diabetic Medicine : A Journal of the British Diabetic Association* 3: 312-315.

Patrakka J, Xiao Z, Nukui M, Takemoto M, He L, Oddsson A*, et al.* 2007. Expression and subcellular distribution of novel glomerulus-associated proteins dendrin, ehd3, sh2d4a, plekhh2, and 2310066E14Rik. *Journal of the American Society of Nephrology : JASN* 18: 689-697.

Pattaro C, Kottgen A, Teumer A, Garnaas M, Boger CA, Fuchsberger C*, et al.* 2012. Genome-wide association and functional follow-up reveals new loci for kidney function. *PLoS Genetics* 8: e1002584.

Perkins BA, Nelson RG, Ostrander BE, Blouch KL, Krolewski AS, Myers BD*, et al.* 2005. Detection of renal function decline in patients with diabetes and normal or elevated GFR by serial measurements of serum cystatin C concentration: results of a 4-year follow-up study. *Journal of the American Society of Nephrology : JASN* 16: 1404-1412.

Pezzolesi MG, Poznik GD, Mychaleckyj JC, Paterson AD, Barati MT, Klein JB*, et al.* 2009a. Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. *Diabetes* 58: 1403-1410.

Pezzolesi MG, Skupien J, Mychaleckyj JC, Warram JH & Krolewski AS 2010. Insights to the genetics of diabetic nephropathy through a genome-wide association study of the GoKinD collection. *Seminars in Nephrology* 30: 126-140.

Pezzolesi MG, Katavetin P, Kure M, Poznik GD, Skupien J, Mychaleckyj JC*, et al.* 2009b. Confirmation of genetic associations at ELMO1 in the GoKinD collection supports its role as a susceptibility gene in diabetic nephropathy. *Diabetes* 58: 2698-2702.

Pluzhnikov A, Below JE, Konkashbaev A, Tikhomirov A, Kistner-Griffin E, Roe CA*, et al.* 2010. Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *American Journal of Human Genetics* 87: 123-128.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA & Reich D 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38: 904-909.

Prickett TD, Agrawal NS, Wei X, Yates KE, Lin JC, Wunderlich JR*, et al.* 2009. Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in ERBB4. *Nature Genetics* 41: 1127-1132.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP*, et al.* 2010. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics (Oxford, England)* 26: 2336-2337.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D*, et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81: 559-575.

Quinn M, Angelico MC, Warram JH & Krolewski AS 1996. Familial factors determine the development of diabetic nephropathy in patients with IDDM. *Diabetologia* 39: 940-945.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ*, et al.* 2001. Linkage disequilibrium in the human genome. *Nature* 411: 199-204.

Reichard P, Nilsson BY & Rosenqvist U 1993. The effect of long-term intensified insulin treatment on the development of microvascular complications of diabetes mellitus. *New England Journal of Medicine* 329: 304-309.

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

Rogus JJ, Poznik GD, Pezzolesi MG, Smiles AM, Dunn J, Walker W*, et al.* 2008. High-Density Single Nucleotide Polymorphism Genome-Wide Linkage Scan for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes. *Diabetes* 57: 2519.

Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A*, et al.* 2008. Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PloS One* 3: e3519.

Salonia A, Lanzi R, Scavini M, Pontillo M, Gatti E, Petrella G*, et al.* 2006. Sexual function and endocrine profile in fertile women with type 1 diabetes. *Diabetes Care* 29: 312-316.

Sambo F, Trifoglio E, Di Camillo B, Toffolo GM & Cobelli C 2012. Bag of Naive Bayes: biomarker selection and classification from genome-wide SNP data. *BMC Bioinformatics* 13 Suppl 14: S2-2105-13-S14-S2. Epub 2012 Sep 7.

Sandholm N, Salem RM, McKnight AJ, Brennan EP, Forsblom C, Isakova T*, et al.* 2012. New susceptibility Loci associated with kidney disease in type 1 diabetes. *PLoS Genetics* 8: e1002921.

Sanyal A, Lajoie BR, Jain G & Dekker J 2012. The long-range interaction landscape of gene promoters. *Nature* 489: 109-113.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S & Snyder M 2012. Linking disease associations with regulatory information in the human genome. *Genome Research* 22: 1748-1759.

Scheet P & Stephens M 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics* 78: 629-644.

Schmid H, Boucherot A, Yasuda Y, Henger A, Brunner B, Eichinger F*, et al.* 2006. Modular activation of nuclear factor-kappaB transcriptional programs in human diabetic nephropathy. *Diabetes* 55: 2993-3003.

Schwarz K, Simons M, Reiser J, Saleem MA, Faul C, Kriz W*, et al.* 2001. Podocin, a raft-associated component of the glomerular slit diaphragm, interacts with CD2AP and nephrin. *The Journal of Clinical Investigation* 108: 1621-1629.

Seaquist ER, Goetz FC, Rich S & Barbosa J 1989. Familial clustering of diabetic kidney disease. *New England Journal of Medicine* 320: 1161-1165.

Sebastiani P, Wang L, Nolan VG, Melista E, Ma Q, Baldwin CT*, et al.* 2008. Fetal hemoglobin in sickle cell anemia: Bayesian modeling of genetic associations. *American Journal of Hematology* 83: 189-195.

Segre AV, DIAGRAM Consortium, MAGIC investigators, Groop L, Mootha VK, Daly MJ*, et al.* 2010. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics* 6: e1001058.

Seissler J & Scherbaum WA 2006. Autoimmune diagnostics in diabetes mellitus. *Clinical Chemistry and Laboratory Medicine : CCLM / FESCC* 44: 133-137.

Shih NY, Li J, Cotran R, Mundel P, Miner JH & Shaw AS 2001. CD2AP localizes to the slit diaphragm and binds to nephrin via a novel C-terminal domain. *The American Journal of Pathology* 159: 2303-2308.

Shimazaki A, Kawamura Y, Kanazawa A, Sekine A, Saito S, Tsunoda T*, et al.* 2005. Genetic variations in the gene encoding ELMO1 are associated with susceptibility to diabetic nephropathy. *Diabetes* 54: 1171-1178.

Shin OH, Han W, Wang Y & Sudhof TC 2005. Evolutionarily conserved multiple C2 domain proteins with two transmembrane regions (MCTPs) and unusual Ca2+ binding properties. *The Journal of Biological Chemistry* 280: 1641-1651.

Silbiger S & Neugarten J 2008. Gender and human chronic renal disease. *Gender Medicine* 5 Suppl A: S3-S10.

Sircar S 2008. Principles of medical physiology. Stuttgart ; New York: Thieme.ISBN 1588905721; 9781588905727; 9783131440617; 3131440619.

Skyler JS & Oddo C 2002. Diabetes trends in the USA. *Diabetes/Metabolism Research and Reviews* 18 Suppl 3: S21-6.

Slatkin M 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature Reviews.Genetics* 9: 477-485.

Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C*, et al.* 2014. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507: 371-375.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU*, et al.* 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* 42: 937-948.

Stark K, Vainio S, Vassileva G & McMahon AP 1994. Epithelial transformation of metanephric mesenchyme in the developing kidney regulated by Wnt-4. *Nature* 372: 679-683.

Steiner DF, Cunningham D, Spigelman L & Aten B 1967. Insulin biosynthesis: evidence for a precursor. *Science (New York, N.Y.)* 157: 697-700.

Stephens M & Balding DJ 2009. Bayesian statistical methods for genetic association studies. *Nature Reviews.Genetics* 10: 681-690.

Stoner M, Wormke M, Saville B, Samudio I, Qin C, Abdelrahim M*, et al.* 2004. Estrogen regulation of vascular endothelial growth factor gene expression in ZR-75 breast cancer cells through interaction of estrogen receptor alpha and SP proteins. *Oncogene* 23: 1052-1063.

Stoner M, Wang F, Wormke M, Nguyen T, Samudio I, Vyhlidal C*, et al.* 2000. Inhibition of vascular endothelial growth factor expression in HEC1A endometrial cancer cells through interactions of estrogen receptor alpha and Sp3 proteins. *The Journal of Biological Chemistry* 275: 22769-22779.

Strachan T & Read AP 1999. Genetic mapping of mendelian characters.

Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE*, et al.* 2012. Patterns of cis regulatory variation in diverse human populations. *PLoS Genetics* 8: e1002639.

Surendran K, McCaul SP & Simon TC 2002. A role for Wnt-4 in renal fibrosis. *American Journal of Physiology.Renal Physiology* 282: F431-41.

Tarnow L, Groop PH, Hadjadj S, Kazeem G, Cambien F, Marre M*, et al.* 2008. European rational approach for the genetics of diabetic complications--EURAGEDIC: patient populations and strategy. *Nephrology, Dialysis, Transplantation : Official Publication*

of the European Dialysis and Transplant Association - European Renal Association 23: 161-168.

Terada Y, Tanaka H, Okado T, Shimamura H, Inoshita S, Kuwahara M*, et al.* 2003. Expression and function of the developmental gene Wnt-4 during experimental acute renal failure in rats. *Journal of the American Society of Nephrology : JASN* 14: 1223-1233.

The DCCT Research Group 1997. Clustering of long-term complications in families with diabetes in the diabetes control and complications trial. *Diabetes* 46: 1829-1839.

The DCCT Research Group 1995. Effect of intensive therapy on the development and progression of diabetic nephropathy in the Diabetes Control and Complications Trial. *Kidney International* 47: 1703-1720.

The DCCT Research Group 1986. The Diabetes Control and Complications Trial (DCCT). Design and methodologic considerations for the feasibility phase. *Diabetes* 35: 530-545.

The DIAMOND Project Group 2006. Incidence and trends of childhood Type 1 diabetes worldwide 1990-1999. *Diabetic Medicine* 23.

Thorn LM, Forsblom C, Fagerudd J, Pettersson-Fernholm K, Kilpikari R, Groop PH*, et al.* 2007. Clustering of risk factors in parents of patients with type 1 diabetes and nephropathy. *Diabetes Care* 30: 1162-1167.

Tidcombe H, Jackson-Fishert A, Mathers K, Stern DF, Gassmann M & Golding JP 2003. Neural and mammary gland defects in Erbb4 knockout mice genetically rescued from embryonic lethality. *Proceedings of the National Academy of Sciences of the United States of America* 100: 8281-8286.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V*, et al.* 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics* 39: 857-864.

Tolonen N, Forsblom C, Thorn L, Waden J, Rosengard-Barlund M, Saraheimo M*, et al.* 2009. Lipid abnormalities predict progression of renal disease in patients with type 1 diabetes. *Diabetologia* 52: 2522-2530.

Tong Z, Yang Z, Patel S, Chen H, Gibbs D, Yang X*, et al.* 2008. Promoter polymorphism of the erythropoietin gene in severe diabetic eye and kidney complications. *Proceedings of the National Academy of Sciences* 105: 6998.

Tregouet DA, Groop PH, McGinn S, Forsblom C, Hadjadj S, Marre M*, et al.* 2008. G/T substitution in intron 1 of the UNC13B gene is associated with increased risk of nephropathy in patients with type 1 diabetes. *Diabetes* 57: 2843-2850.

Tremolada G, Lattanzio R, Mazzolari G & Zerbini G 2007. The therapeutic potential of VEGF inhibition in diabetic microvascular complications. *American Journal of Cardiovascular Drugs : Drugs, Devices, and Other Interventions* 7: 393-398.

Tuomilehto J, Borch-Johnsen K, Molarius A, Forsen T, Rastenyte D, Sarti C*, et al.* 1998. Incidence of cardiovascular disease in Type 1 (insulin-dependent) diabetic subjects with and without diabetic nephropathy in Finland. *Diabetologia* 41: 784-790.

U.S. Renal Data System 2011. USRDS 2011 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

Veikkolainen V, Naillat F, Railo A, Chi L, Manninen A, Hohenstein P*, et al.* 2012. ErbB4 modulates tubular cell polarity and lumen diameter during kidney development. *Journal of the American Society of Nephrology* 23: 112-122.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG*, et al.* 2001. The sequence of the human genome. *Science (New York, N.Y.)* 291: 1304-1351.

Viberti G, Mogensen C, Passa P, Bilous R & Mangili R 1994. St Vincent Declaration, 1994: Guidelines for the Prevention of Diabetic Renal Failure.

Visscher PM, Hill WG & Wray NR 2008. Heritability in the genomics era--concepts and misconceptions. *Nature Reviews.Genetics* 9: 255-266.

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP*, et al.* 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* 42: 579-589.

Watson JD & Crick FH 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171: 737-738.

Wellcome Trust Case Control Consortium 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-678.

Wessman M, Forsblom C, Kaunisto MA, Soderlund J, Ilonen J, Sallinen R*, et al.* 2011. Novel susceptibility locus at 22q11 for diabetic nephropathy in type 1 diabetes. *PloS One* 6: e24053.

Willer CJ, Li Y & Abecasis GR 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* 26: 2190-2191.

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, *et al.* 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics* 40: 161-169.

Wiseman M, Viberti G, Mackintosh D, Jarrett RJ & Keen H 1984. Glycaemia, arterial pressure and micro-albuminuria in type 1 (insulin-dependent) diabetes mellitus. *Diabetologia* 26: 401-405.

Woroniecka KI, Park AS, Mohtat D, Thomas DB, Pullman JM & Susztak K 2011. Transcriptome analysis of human diabetic kidney disease. *Diabetes* 60: 2354-2369.

Writing Team for the DCCT/EDIC Research Group 2003. Sustained effect of intensive treatment of type 1 diabetes mellitus on development and progression of diabetic nephropathy: the Epidemiology of Diabetes Interventions and Complications (EDIC) study. *JAMA : The Journal of the American Medical Association* 290: 2159-2167.

Yang J, Lee SH, Goddard ME & Visscher PM 2011. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* 88: 76-82.

Yin M, Zhong Z, Connor HD, Bunzendahl H, Finn WF, Rusyn I, *et al.* 2002. Protective effect of glycine on renal injury induced by ischemia-reperfusion in vivo. *American Journal of Physiology.Renal Physiology* 282: F417-23.

Zeng F, Zhang M-, Singh AB, Zent R & Harris RC 2007. ErbB4 isoforms selectively regulate growth factor-induced Madin-Darby canine kidney cell tubulogenesis. *Molecular Biology of the Cell* 18: 4446-4456.

Zhou T, He X, Cheng R, Zhang B, Zhang RR, Chen Y, *et al.* 2012. Implication of dysregulation of the canonical wingless-type MMTV integration site (WNT) pathway in diabetic nephropathy. *Diabetologia* 55: 255-266.

11 List of abbreviations

A	Adenine
AER	Albumin excretion rate
ACE	Angiotensin converting enzyme
ACR	Albumin-to-creatinine ratio
ARB	Angiotensin II receptor blocker
AHT	Anti-hypertensive (treatment)
BF	Bayes Factor
BMI	Body mass index
BoNB	Bag of Naïve Bayes
C	Cytosine
CEU	Caucasian individuals from Utah, USA, from the Centre d'Etude du Polymorphisme Humain collection
CHB	Han Chinese individuals in Beijing, China
CI	Confidence interval
CKD	Chronic kidney disease
D'	"D prime", a measure of linkage disequilibrium
DCCT	Diabetes Control and Complications Trial
d.f.	Degree of freedom
DN	Diabetic nephropathy
DNA	Deoxyribose nucleic acid
EDIC	The Epidemiology of Diabetes Interventions and Complications
eGFR	Estimated glomerular filtration rate
eQTL	Expression quantitative trait loci
ERα	Estrogen receptor α
ESRD	End stage renal disease
FDR	False discovery rate
FIMM	The Institute of Molecular Medicine Finland
FIND	Family Investigation of Nephropathy and Diabetes Study
FinnDiane	The Finnish Diabetic Nephropathy Study
G	Guanine
GFR	Glomerular filtration rate
GlyR	Glycine receptor
GoKinD US	Genetics of Kidneys in Diabetes US Study
GWAS	Genome-wide association study

HbA_{1C}	Proportion of glycosylated hemoglobin
HILMO	The Finnish hospital discharge registry
HMM	Hidden Markov models
HR	Hazard ratio
HWE	Hardy-Weinberg equilibrium
JPT	Individuals from Tokyo, Japan
kb	kilobase, 1,000 bases
KEGG	Kyoto Encyclopedia of Ggenes and Genomes
LADA	latent autoimmune diabetes of adults
LD	Linkage disequilibrium
LOD	Logarithm of odds
MAC	Minor allele count
MAF	Minor allele frequency
MAGENTA	Meta-Analysis Gene-set Enrichment of variaNT Associations
Mb	Megabase, 1,000,000 bases
MCC	Matthews correlation coefficient
mGluR	Metabotrobic glutamate receptor
MLS	Maximum likelihood score
mRNA	Messenger ribonucleic acid
NFS-ORPS	UK Nephropathy Family Study and Oxford Regional Prospective Study
NS	Non-significant
OOB	Out-of-bag
OR	Odds ratio
PC	Principal component
PCA	Principal component analysis
PPA	Posterior probability of association
PPnA	Posterior probability of no association
QC	Quality control
QQ-plot	Quantile-quantile plots
r^2	Squared correlation coefficient, a measure of LD
RNA	Ribonucleic acid
SDR	Scania Diabetes Registry
SE	Standard error
SNP	Single nucleotide polymorphism
T	Thymine
T1D	Type 1 diabetes
T2D	Type 2 diabetes
THL	Finnish National Institute of Health and Welfare
UK-ROI	All Ireland Warren 3, Genetics of Kidneys in Diabetes UK Study
YRI	Individuals from Yoruba in Ibadan, Nigeria