On the origin of obesity and type 2 diabetes

Clara Elbers

The research presented in this thesis was performed at the Department of Medical Genetics and the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands, and was financially supported by IOP Genomics (grant IGE05012), a program by NL Innovatie.

ISBN 978-90-393-5413-1

Printed by
Ipskamp Drukkers B.V.

Cover design by
Sietse van der Wal

On the origin of obesity and type 2 diabetes

Over het ontstaan van obesitas and type 2 diabetes
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus prof. dr. J. C. Stoof,
in gevolge het besluit van het college voor promoties
in het openbaar te verdedigen op donderdag 7 oktober 2010
des middags te 4.15 uur

door

Clara Charlotte Elbers

geboren op 27 februari 1980
te Utrecht

Promotores

Prof. dr. C. Wijmenga

Prof. dr. Y.T. van der Schouw

Copromotor

Dr. N.C. Onland-Moret

Table of contents

Chapter 1

Preface and outline of the thesis

Preface and outline of the thesis

The incidence of type 2 diabetes (T2D) is rising rapidly worldwide and there are already more than 180 million diabetic subjects. T2D risk factors include ethnic background, age, hypertension, overweight, increased abdominal fat, and lack of physical exercise. Obesity is considered to be the most important risk factor for T2D and the main one driving the current epidemic as 90% of T2D patients are obese. Worldwide obesity has also reached epidemic proportions, with 300 million adults classified as clinically obese. T2D and obesity are multifactorial disorders in which both genetic and non-genetic (environmental and lifestyle) factors play a role. The past years has witnessed substantial advances in understanding the genetic basis of obesity and T2D. To date, 17 common obesity loci and 18 common T2D loci have been identified. However, only around 10% of the genetic risk for these traits can be explained. Therefore, many more risk loci for obesity and T2D still need to be discovered.

The high prevalence of T2D in many human populations poses a further evolutionary question: Why is the disease so common, when it should disappear as those genetically susceptible to it are removed by natural selection?

In the present thesis we focus on (I) evaluating alternative methods to find candidate genes for T2D and obesity, (II) studying genetic and environmental risk factors for T2D and obesity, and (III) studying the origin of the high prevalence of T2D and obesity in modern societies.

Both obesity and T2D are complex genetic traits but they share some non-genetic risk factors. In the introduction, **chapter 2**, we describe the genes recently identified for T2D and obesity by genome-wide association studies (GWAS) and evaluate their functions in an effort to determine whether there is any support for the hypothesis that T2D and obesity share some underlying mechanism(s).

In the first, methodological, part of the thesis (part I), we use an alternative strategy to find candidate genes for obesity and T2D and explore alternative methods for the investigation of GWAS data to obtain valuable information on the biology and evolutionary origin of T2D. In **chapter 3** we combine six tools for disease gene identification to analyse the overlapping T2D and obesity susceptibility loci to pinpoint shared candidate genes for T2D and obesity. In this study, we evaluated alternative methods to study GWAS data. Instead of focusing on the single nucleotide

polymorphisms (SNPs) with the highest statistical significance, we took advantage of prior biological information and tried to detect overrepresented pathways in the GWAS data in **chapter 4**. We evaluated whether pathway classification analysis can help prioritize the biological pathways most likely to be involved in the disease etiology.

Part II of the thesis reports on studies investigating genetic and environmental risk factors for T2D and obesity. In **chapter 5** we investigated the role of variants in *NPY1R, NPY2R* and *NPY5R* genes, involved in the hypothalamic pathway, in total and nutrient-specific energy intake. In **chapter 6**, we investigate whether we can replicate the recently reported associations of the susceptibility loci with different obesity related phenotypes and explored the effect of variation in the currently implicated obesity genes affects on dietary energy and macronutrient intake. In **chapter 7**, we assessed the association between both parity and age at first full-term pregnancy with the risk of T2D in women.

The studies, described in part III, aim to investigate the evolutionary explanation of obesity and T2D. In **chapter 8** we tested a theory on the evolutionary origin of obesity and T2D, the thrifty gene hypothesis, by investigating whether recently identified T2D and obesity risk alleles have been under recent positive selection. In **chapter 9** we investigate whether sub- or infertility predicts later-in-life T2D risk. To investigate body weight from a historical perspective, we studied weight distribution in an 18th century criminal gang (**chapter 10**). **Chapter 11** provides a general discussion on the origin of obesity and T2D.

Chapter 2

Gene variants for obesity and type 2 diabetes mellitus: Shared aetiology?

Clara C. Elbers, Marcel G.M. Wolfs, Timon W. van Haeften

Eur Endocrinology 2009;5(1):27-30

Abstract

The incidence of type 2 diabetes (T2D) is rising rapidly worldwide, mainly due to the increase in the incidence of obesity. Both obesity and T2D are complex genetic traits but they share some non-genetic risk factors. Hence, it is tempting to speculate that the susceptibility to T2D and obesity may also involve shared underlying genetic factors acting on common molecular mechanisms. Recent genome-wide association (GWA) studies identified 17 common loci for obesity and 18 common loci for T2D. This review explores whether the susceptibility loci for T2D and obesity can indicate potential overlapping mechanisms in the disorders. Additionally, we touch upon the challenges regarding follow-up of confirmed GWA signals as well as alternative approaches to analysing GWA data to a fuller potential.

Introduction

The incidence of type 2 diabetes (T2D) is rising rapidly worldwide and there are already more than 180 million diabetic subjects. T2D risk factors include ethnic background, age, hypertension, overweight, increased abdominal fat, and lack of physical exercise. Obesity is considered to be the most important risk factor for T2D and the main one driving the current epidemic as 90% of T2D patients is obese. Worldwide obesity has also reached epidemic proportions, with 300 million adults classified as clinically obese (based on data from the World Health Organization (WHO)). Up to 50% of these obese individuals will develop T2D at some stage in their life, depending on the age when they became obese.

T2D and obesity are multifactorial disorders in which both genetic and non-genetic (environmental and lifestyle) factors play a role. While the life-time risk for T2D in the Western world is around 10%, first-degree relatives of patients have a 20-40% risk for the disease, and concordance rates for identical twins have been estimated to be 57% or higher (up to 90%) for T2D in male twins [1]. These observations clearly indicate there is a genetic component to the disease; however, the model seems to be more complex, involving multiple genes and environmental factors.

Common obesity and T2D share some non-genetic factors as both are influenced by diet and physical inactivity. Both conditions are characterised by insulin resistance, suggesting a shared pathology. It has been proposed that the susceptibility to develop T2D and obesity is, in part, due to shared underlying genetic factors involved in common molecular mechanisms. This review explores the genes recently identified for T2D and obesity by genome-wide association (GWA) studies and evaluates their functions in an effort to determine whether there is any support for the hypothesis that T2D and obesity share some underlying mechanism(s).

Common T2D and obesity susceptibility loci

Before the era of GWA studies, genes were prioritized as candidate disease genes because of their function and/or position, and they were then studied for association with obesity and T2D. These approaches had limited success as replication of associated genes only proved possible for variants in or near *PPARG, KCNJ11, TCF2* and *WFS1* with T2D [2] and for variants in or near *BDNF, MC4R* and *SH2B1* with

obesity. Recently, a number of GWA studies have identified 17 common loci for obesity and 18 common loci for T2D [3-14] (table 1, table 2). One interesting finding is that the results of GWA studies often point towards genes with currently unknown or poorly described functions. This is one of the reasons why the previous approaches to gene hunting had limited success. Most of the recently identified genes were not tested for association, simply because their biological functions were unknown and they were not therefore suspected of being involved in the disease.

GWA studies further indicate that associated common markers have only a minor impact on disease susceptibility. The known risk variants for T2D are all relatively frequent in the population, ranging from 0.26 for *TCF7L2* to 0.85 for *PPARG* in the European population, and have a low effect size, with odds ratios (ORs) ranging from 1.10 (confidence interval (CI): 1.07-1.14) for *TCF2* to 1.37 (CI: 1.31-1.43) for *TCF7L2*. For obesity, each of the associated variants has a very modest effect ranging from 0.06 kg/m2 for *KCTD15* to 0.33 kg/m2 for *FTO* per allele change in BMI. All these variants together can only explain a small percentage of the genetic susceptibly of T2D and obesity. However, these variants are generally not the causal variants; the ORs of the causal variants should be higher and will presumably explain a larger percentage of the genetic susceptibility to these two conditions. In addition, it is likely that there are many more genes contributing a similar or smaller effect.

Functions of T2D and obesity genes

The susceptibility loci for obesity and T2D can provide insight into the aetiology of the traits, yet it is difficult to link genetic associations to biological mechanisms. It is important to keep in mind that most of the observed associations are located in non-coding regions of the genome and that the presented T2D and obesity genes are mostly genes near the associated markers. We assume that at least some of these nearby genes are truly involved in the traits as we discuss the function of these genes to gain more insight into the disease pathology.

Table 1. Overview of genetic variants associated with T2D and their putative role in disease pathogenesis.

Marker	Chr	Closest gene(s)	Putative mechanism in T2D
rs7578597	2	*THADA*	Apoptosis of beta cells
rs10010131	4	*WFS1*	Apoptosis of beta cells
rs10923931	1	*NOTCH2*	Beta cell growth and development
rs4402960	3	*IGF2BP2*	Beta cell growth and development
rs10946398	6	*CDKAL1*	Beta cell growth and development
rs1111875	10	*HHEX-IDE*	Beta cell growth and development
rs7901695	10	*TCF7L2*	Beta cell growth and development
rs4430796	17	*TCF2*	Beta cell growth and development
			Cell cycle
rs864745	7	*JAZF1*	Cell cycle
rs10811661	9	*CDKN2A-2B*	Cell cycle
rs12779790	10	*CDC123-CAMK1D*	Cell cycle
rs13266634	8	*SLC30A8*	Insulin secretion
rs2237892	11	*KCNQ1*	Insulin secretion
rs5215	11	*KCNJ11*	Insulin secretion
rs10830963	11	*MTNR1B*	Insulin secretion
rs17036101	3	*SYNC, PPARG*	Unknown
rs4607103	3	*ADAMTS9*	Unknown
rs1153188	12	*DCD*	Unknown
rs7961581	12	*TSPAN8*	Unknown

Type 2 diabetes

The main feature of T2D is the inability of an individual to maintain proper blood glucose levels. The key player in this homeostasis is the peptide hormone insulin which is produced in the beta-cells of the pancreas. After a meal this hormone is released into the bloodstream and transported to its several target tissues, where it diminishes hepatic glucose output and triggers glucose uptake and storage as either fat or glycogen. T2D starts with the failure of several tissues, such as adipose tissue and muscle, to respond to the stimulus of insulin (this is often referred to as insulin resistance). As a result, insulin levels rise and it is presumed that, after a certain time, beta-cells are not able to keep up with the growing demands for insulin release. At this point a (second) vicious cycle of higher blood glucose levels and a higher demand for insulin is entered, which increases the strain on the beta-cells and leads to beta-cell apoptosis and ultimately a complete inability to produce insulin [15].

The T2D risk variants currently pinpointed appear to act through interference with beta-cell insulin secretion rather than through insulin sensitivity of insulin target tissues. This indicates that disturbances in beta-cell function are ultimately decisive for the actual development of T2D. The known T2D genes can be classified into subgroups for their potential role in beta-cell function, based on what is known about their molecular function [15]. It has been proposed that *KCNJ11, KCNQ1, MTNR1B* and *SLC30A8* are involved in insulin secretion; *CDKAL1, IGF2BP2, HHEX-IDE, NOTCH2, JAZF1, TCF7L2* and *TCF2* in beta-cell growth and development; *TCF2, CDKN2A-2B, CDC123* and *JAZF1* in the cell cycle; and *THADA* and *WFS1* in the apoptosis of beta-cells.

Obesity

Overweight and obesity result from a long-lasting imbalance between food intake and energy expenditure, leading to storage of excess calories as body fat. Control of energy balance involves the integration of satiety signals from the gastrointestinal tract, adipose tissue and nutrient-related signals. Adiposity signals provide feedback information from body energy stores to various hypothalamic regions and are mediated via the circulating hormones leptin and insulin and others. As a result, body weight remains remarkably stable most of the time in most people, but any defects in this system can lead to a deregulation of body weight. It is known that hypothalamic defects in either insulin or leptin signalling are associated with increased food intake and/or heavier body weight [16,17].

Many of the recently established obesity susceptibility loci are located near genes that are highly expressed in the brain and/or have been shown to have a function in neuronal development or activity. These suggest a key role for the hypothalamic pathways in regulating food intake and energy homeostasis in the architecture of obesity. The role of *BDNF, MC4R* and *SH2B1* genes in obesity pathogenesis is well known from functional studies: MC4R is the key signalling neuropeptide, inhibiting food intake and increasing metabolic rate [18]. BDNF decreases food intake in response to nutritional status and MC4R signalling [19]. Studies show that *SH2B1* regulates energy balance, body weight, peripheral insulin sensitivity, and glucose homeostasis, at least in part by enhancing hypothalamic leptin sensitivity [20]. The roles of the other recently identified loci in obesity pathogenesis are not yet clear: *FTO* is suggested to participate in the central control of energy

homeostasis, where it is regulated by feeding and fasting [21]. *NEGR1* and *TMEM18* are involved in neural development [11,13], whereas *NPC1* is involved in endosomal cholesterol trafficking in the central nervous system, liver and macrophages [22].

Because satiation signals influence how many calories are eaten during individual meals, it is interesting to further explore the effect of obesity susceptibility variants on total and nutrient-specific dietary intake. While total energy intake is a vital aspect of food intake, macronutrient composition or diet patterns may be equally important factors underlying the development of obesity. Common variants near the FTO and MC4R genes were recently found to be associated with total energy intake and a variant near MC4R was also found to be associated with dietary fat [23-25]. In addition, in a study population of 1700 Dutch females, the susceptibility loci near *NEGR1, TMEM18, BDNF, MTCH2* and *SH2B1* showed association with macronutrient intake [32]. It can be argued the genes associated with food intake play a role in satiation signalling. Further insight into the function of these genes may yield valuable clues for lifestyle intervention and therapeutics.

Table 2. Overview of genetic variants associated with obesity and their putative role in pathogenesis.

Marker	Chr	Closest gene(s)	Putative mechanism in obesity
rs4074134	11	*BDNF*	Energy homeostasis
rs7498665	16	*SH2B1*	Energy homeostasis
rs9939609	16	*FTO*	Energy homeostasis
rs17782313	18	*MC4R*	Energy homeostasis
rs1805081	18	*NPC1*	Lipid transport
rs2815752	1	*NEGR1*	Neural development
rs6548238	2	*TMEM18*	Neural development
rs10838738	11	*MTCH2*	Satiation signalling
rs10913469	1	*SEC16B*	Unknown
rs7647305	3	*ETV5*	Unknown
rs10938397	4	*GNPDA2*	Unknown
rs2844479	6	*NCR3*	Unknown
rs4712652	6	*PRL*	Unknown
rs10508503	10	*PTER*	Unknown
rs7138803	12	*BCDIN3D*	Unknown
rs1424233	16	*MAF*	Unknown
rs11084753	19	*KCTD15*	Unknown

GWA studies give insight into shared disease aetiology

So far, results from the recent GWA studies do not point towards both obesity and T2D having an increased risk from shared disease susceptibility loci. It seems that the susceptibility genes for obesity are involved at the start of the trait (energy imbalance) and those for T2D at a later stage of the disease (beta-cell defect).

The initial finding of association of the *FTO* gene with T2D was subsequently shown to be entirely due to an obesity risk. The gene was found to be highly associated with T2D in several study populations [26], but failed to replicate in studies where they matched the cases and controls on BMI or selected relatively lean cases [7,9]. The loci near *GNPDA2, BDNF* and *TMEM18* that were associated with obesity were also found to be weakly associated with T2D [11,13]. Follow-up studies should explore whether these associations with T2D mainly act through an effect on weight regulation; they need to take into account not only BMI, but also other measures of obesity. Recently it has become clear that not only the amount of body fat, but especially its distribution is important in determining disease risk; independently of BMI, a larger waist circumference (as a measure of abdominal obesity) is related to chronic disease risk, like T2D [27].

However, because there are still many more obesity and T2D genes to discover, we cannot rule out that the susceptibility for developing T2D and obesity is partly due to shared underlying genetic factors. In a previous study we compared all the published genome scans for T2D and obesity and identified five overlapping chromosomal regions for both entities [28]. However, the shared genetic effect may be smaller than we initially thought or obesity could simply be a non-genetic risk factor for T2D because it provokes insulin resistance.

Follow-up of confirmed associations and alternative gene-hunting approaches

The next challenge is to go from the statistical association of the markers to a functional link between the genomic region and T2D and obesity. The associated SNP will either be the disease-causing variant or be in strong linkage disequilibrium (LD) with the causal variant, i.e. the 'associated SNP' and the causal variant are inherited together. Re-sequencing of the susceptibility loci is needed to establish the causal genes. This will be a daunting task because the LD blocks can be extensive. In GWA

studies, allele frequencies of approximately 300,000-500,000 common SNPs across the human genome are compared one by one between patients and healthy controls. Although this approach successfully identified new T2D and obesity genes, 99.9% of the GWA data has not yet been analysed to its full potential.

It is possible that single locus methods do not reflect the correct underlying model of association. There is growing evidence that gene-gene and gene-environment interactions contribute to complex diseases rather than single genes [29]. Several models for epistasis (i.e. gene-gene interactions) have been proposed [30], including ones in which the genes alone have no effect on disease aetiology but where their interaction modifies disease risk. In addition, it is likely that genetic variation contributes to disease risk through complex biological pathways. It is unlikely that the genes involved in these pathways will be picked up using traditional single-locus analyses, and different methods will be needed to extract this information from GWA datasets [31].

Conclusion

Common obesity and T2D share some non-genetic factors as both are influenced by diet and physical inactivity. Both conditions are characterised by insulin resistance, suggesting a shared pathology. However, results from recent GWA studies do not point towards shared disease susceptibility loci with an increased risk for both obesity and T2D.

Currently it seems that the susceptibility genes for obesity are involved at the start of the trait (energy imbalance) and those for T2D at a later stage of the disease (beta-cell defect). It is suggested that the shared genetic effect may be smaller than we thought or obesity could simply be a non-genetic risk factor for T2D because it provokes insulin resistance. Discovering more obesity and T2D genes will provide a broader insight into the shared disease pathology.

Acknowledgments

This review was financially supported by SenterNovem (IOP genomics grant IGE05012). We thank Jackie Senior for critically reading the manuscript. The authors have declared that no competing interests exist.

References

1. Newman B, Selby JV, King MC, Slemenda C, Fabsitz R, et al. (1987) Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. Diabetologia 30: 763-768.
2. Frayling TM (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. Nat Rev Genet 8: 657-662.
3. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.
4. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, et al. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40: 768-775.
5. Lyssenko V, Nagorny CL, Erdos MR, Wierup N, Jonsson A, et al. (2009) Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. Nat Genet 41: 82-88.
6. Meyre D, Delplanque J, Chevre JC, Lecoeur C, Lobbens S, et al. (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nat Genet 41: 157-159.
7. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331-1336.
8. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341-1345.
9. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.
10. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, et al. (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39: 770-775.
11. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41: 18-24.
12. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, et al. (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 40: 1098-1102.
13. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, et al. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41: 25-34.
14. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40: 638-645.
15. Stumvoll M, Goldstein BJ, van Haeften TW (2008) Type 2 diabetes: pathogenesis and treatment. Lancet 371: 2153-2156.
16. Niswender KD, Schwartz MW (2003) Insulin and leptin revisited: adiposity signals with overlapping physiological and intracellular signaling capabilities. Front Neuroendocrinol 24: 1-10.
17. Bruning JC, Gautam D, Burks DJ, Gillette J, Schubert M, et al. (2000) Role of brain insulin receptor in control of body weight and reproduction. Science 289: 2122-2125.
18. Farooqi IS, Keogh JM, Yeo GS, Lank EJ, Cheetham T, et al. (2003) Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. N Engl J Med 348: 1085-1095.
19. Xu B, Goulding EH, Zang K, Cepoi D, Cone RD, et al. (2003) Brain-derived neurotrophic factor regulates energy balance downstream of melanocortin-4 receptor. Nat Neurosci 6: 736-742.
20. Ren D, Zhou Y, Morris D, Li M, Li Z, et al. (2007) Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. J Clin Invest 117: 397-406.
21. Fredriksson R, Hagglund M, Olszewski PK, Stephansson O, Jacobsson JA, et al. (2008) The obesity gene, FTO, is of ancient origin, up-regulated during food deprivation and expressed in neurons of feeding-related nuclei of the brain. Endocrinology 149: 2062-2071.
22. Ikonen E (2008) Cellular cholesterol trafficking and compartmentalization. Nat Rev Mol Cell Biol 9: 125-138.
23. Cecil JE, Tavendale R, Watt P, Hetherington MM, Palmer CN (2008) An obesity-associated FTO gene variant and increased energy intake in children. N Engl J Med 359: 2558-2566.
24. Qi L, Kraft P, Hunter DJ, Hu FB (2008) The common obesity variant near MC4R gene is associated with higher intakes of total energy and dietary fat, weight change and diabetes risk in women. Hum Mol Genet 17: 3502-3508.
25. Wardle J, Llewellyn C, Sanderson S, Plomin R (2009) The FTO gene and measured food intake in children. Int J Obes (Lond) 33: 42-45.

26. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.
27. Pischon T, Boeing H, Hoffmann K, Bergmann M, Schulze MB, et al. (2008) General and abdominal adiposity and risk of death in Europe. N Engl J Med 359: 2105-2120.
28. Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, et al. (2007) A strategy to search for common obesity and type 2 diabetes genes. Trends Endocrinol Metab 18: 19-26.
29. Kelada SN, Eaton DL, Wang SS, Rothman NR, Khoury MJ (2003) The role of genetic polymorphisms in environmental health. Environ Health Perspect 111: 1055-1064.
30. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37: 413-417.
31. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, et al. (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol. 2009 Feb 20.
32. Bauer F, Elbers CC, Adan R, Loos RJF, Onland-Moret NC, et al. (2009) GWAS identified obesity genes are associated with adiposity measures and potentially with nutrient-specific food preference. Am J Clin Nutr. 2009 Oct;90(4):951-9.

Chapter 3

A strategy to search for common obesity and type 2 diabetes genes

Clara C Elbers, N Charlotte Onland-Moret, Lude Franke,
Anne G Niehoff, Yvonne T van der Schouw, Cisca Wijmenga

Trends Endocrinol Metab. 2007 Jan-Feb;18(1):19-26

Abstract

Worldwide the incidence of type 2 diabetes (T2D) is rising rapidly, mainly owing to the increase in the incidence of obesity, which is an important risk factor for T2D. Both obesity and T2D are complex genetic traits and they share some non-genetic risk factors. Hence, it is tempting to speculate that the susceptibility to T2D and obesity may also partly be due to shared genes. By comparing all published genome scans for T2D and obesity, five overlapping chromosomal regions for both diseases (encompassing 612 candidate genes) were identified. By analyzing these five susceptibility loci for T2D and obesity using six freely available bioinformatics tools for disease gene identification, 27 functional candidate genes were pinpointed that are involved in eating behaviour, metabolism and inflammation. These genes may reveal a molecular link between the two disorders.

Obesity and type 2 diabetes

Worldwide the incidence of type 2 diabetes (T2D) is rising rapidly and there are already more than 170 million diabetics. T2D results from the body's inability to respond properly to the action of insulin produced by the pancreas; this results from impairment in both insulin sensitivity and insulin secretion [1]. T2D is a multifactorial disorder in which both genetic and non-genetic (environmental and life-style) factors play a role. The concordance rate of T2D amongst monozygotic twins is 76% compared to 40% amongst dizygotic twins, providing convincing evidence that genetic factors contribute to the development of T2D. In addition, there is a 3.5-fold increased risk for a first degree relative of a T2D patient to develop the disease [1]. Although both observations clearly imply there is a genetic component in the disease, the model seems to be more complex, involving multiple genes and environmental factors. A number of genes have been implicated that may contribute significantly to the risk of T2D, including peroxisome proliferator-activated receptor gamma (*PPARG*), potassium inwardly-rectifying channel, subfamily J, member 11 (*KCNJ11*) [2] and, more recently, transcription factor 7-like 2 (*TCF7L2*) [3]. These genes are known to explain only part of the underlying genetic component [2] so there are likely to be other, not yet identified, genes that are also important contributors to T2D susceptibility.

Besides a positive family history, T2D risk factors include ethnic background, age, hypertension, overweight, increased abdominal fat, and lack of physical exercise. Obesity is considered to be the most important risk factor for T2D and the main one driving the current epidemic as 90% of T2D patients are obese. Worldwide obesity has also reached epidemic proportions, with 300 million adults classified as clinically obese (based on data from the World Health Organization (WHO)); 20% of these obese individuals suffer from T2D. Obesity is commonly assessed by the body mass index (BMI), defined as the weight in kilograms divided by the square of the height in meters (kg/m2). A BMI value higher than 30 is defined as obesity according to the WHO.

Obesity and T2D share some non-genetic factors as both are influenced by diet and physical inactivity. Both traits are characterized by insulin resistance, suggesting a shared pathology. It has been proposed that low-grade inflammation in visceral fat may be a potential mechanism whereby obesity results in insulin

resistance [4]. It is tempting to speculate that the susceptibility to develop T2D and obesity is, in part, due to shared underlying genetic factors involved in common molecular mechanisms. This review explores whether there is any support for the hypothesis that T2D and obesity share some underlying susceptibility genes.

Susceptibility loci for T2D and obesity

Genome scans are a useful approach to define susceptibility loci for disease candidate genes [5]. Genome-wide linkage scans involve the typing of families and sibling pairs using polymorphic markers that are positioned across the whole genome, followed by calculation of the degree of linkage of the marker to a disease trait. Positional candidate genes can then be identified by examining the regions around the peaks of linkage that are obtained. Linkage-based studies have implicated many susceptibility loci for both T2D and obesity. Bell *et al.* [6] collected and evaluated genome scans performed on obesity till 2004 based on 31 papers. Since 2004 five additional genome scans on obesity have been published [7-11]. For T2D, 33 genome scans have been reported since 1996 [12-44] (Supplementary table 1).

A total of 14 susceptibility loci for obesity (blue bars, figure 1) and 18 susceptibility loci for T2D (red bars, figure 1) fulfilled the inclusion criteria according to the methods assessed in box 1. Supplementary table 2 gives a complete overview of the susceptibility loci reported for both T2D and obesity. Five of these chromosomal regions, encompassing a total of 612 genes, were found to be linked to both obesity and T2D.

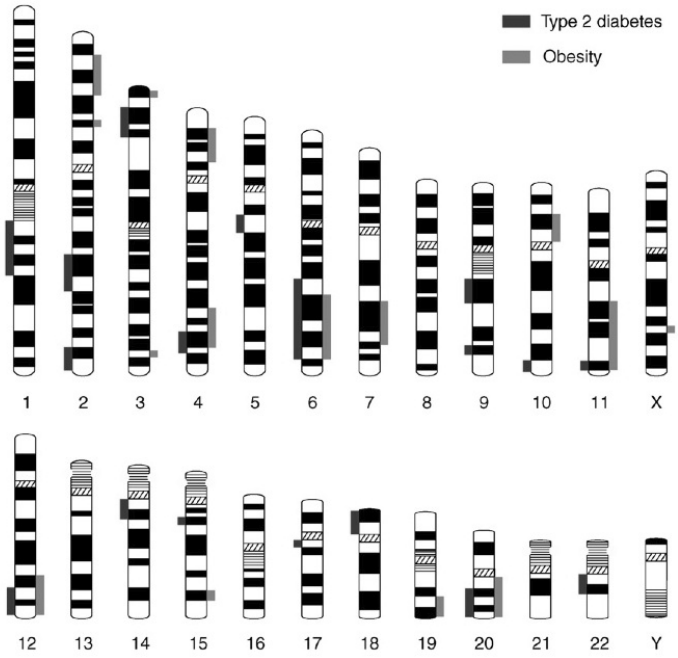


Figure 1. Genetic linkage map for obesity and T2D. The grey bars indicate susceptibility loci for T2D and the light grey bars for obesity. Five chromosomal regions (4q32, 6q22– 6q24, 11q24, 12q24 and 20q12–20q13) were found to be linked to both obesity and T2D. For detailed information on chromosomal locations, see the supplementary material online.

Finding candidate genes using disease gene identification methods

Unfortunately, the data from linkage studies do not directly indicate the gene of interest and identifying a potential gene is usually rather difficult [45, 46] as linkage intervals can contain dozens to hundreds of candidate genes. To identify the gene of interest, a dense map of single nucleotide polymorphisms (SNPs) encompassing the candidate region needs to be tested for genetic association in very large case-control studies. This strategy is based on the "common disease/common variant" hypothesis [47] and assumes that common disease susceptibility alleles are involved in complex traits. If a risk polymorphism exists, it will either be genotyped directly, or be in strong linkage disequilibrium (LD) with one of the genotyped SNPs. The recent completion of the HapMap project phase I has already resulted in a public database of more than 4 million common SNP variants across the genome [48]. This resource makes it feasible to carry out comprehensive genetic association studies needing a high density of SNPs (usually more than one SNP every 10 kb). Consequently, large

numbers of genotypes need to be generated so these studies are rather expensive, despite the recent drop in cost per genotype.

An attractive alternative strategy is to first prioritize the positional candidate genes based on the function of the individual genes. Functional candidates are genes with products that can, in some way, be related to the pathogenesis of a disorder. Evidence for involvement of a gene in the disease process can, for example, be expression in the appropriate tissue or distribution of the gene product in a cell of interest. In the case of obesity, fat, adipose, hypothalamus, pituitary and gut are relevant tissues, whereas for T2D pancreas, fat, adipose, liver, kidney, gut and muscle tissues are considered relevant. Evidence for involvement can also be drawn from a similarity to phenotypes associated with naturally occurring or engineered mutations in other species. For example, the ob/ob mouse that has a defect in the leptin gene is an excellent model for studying obesity. Strong mechanistic support can also come from a causal relationship of the phenotype with a variant nucleotide, with altered protein expression, gene expression or function.

Until recently, investigating each gene separately for its likely involvement in the disease process had to be done manually with the aid of available public databases such as OMIM, Entrez, and genome browsers, but now there are some promising bioinformatics tools (disease gene identification methods) for disease gene identification. Six of these new tools are freely available online: Prioritizer [49], Geneseeker [50], PROSPECTR&SUSPECTS (P&S) [51], Disease Gene Prediction (DGP) [52], Genes2Diseases (G2D) [53] and Endeavour [54] (table 1). These tools use information extracted from public online databases, such as sequence data, medical literature, gene ontology/function annotation (GO), and information on biology, function and gene expression. Although the different tools have the same goal, they are based on different principles: Prioritizer ranks genes based on their functional interaction with genes on different susceptibility loci, assuming that disease genes in a specific disorder are usually functionally related. Geneseeker points to genes which are expressed in disease-related tissues. PROSPECTR differentiates between those genes that are likely to be involved in diseases and those which are not using sequence-based features like gene length, protein length and the percent identity of homologs in other species. SUSPECTS scores candidate genes using PROSPECTR and also assesses how similar their annotation is to already known disease genes. DGP assigns probabilities to genes that could indicate involvement in hereditary

Table 1. Principles of the computational disease gene identification tools and detailed information on them

Tool and website	Principle	Repositories used	Input
Endeavour www.esat.kuleuven.be/endeavour/	Ranks a test gene based on its similarity with the training genes	MEDLINE[a] abstracts, LocusLink, GO, InterPro and BIND protein–protein interactions, KEGG pathways, Microarray and EST-based expression data, TFBS, *cis*-regulatory modules, sequence similarity by BLAST	Susceptibility loci Known disease genes
Prioritizer www.prioritizer.nl	Ranks genes in different loci if they interact functionally	Protein–protein interactions from Reactome, HPRD and BIND. KEGG pathways, coexpression data from GEO, Y2H interactions in various species, GO	Susceptibility loci
G2D www.bork.embl-heidelberg.de/g2d/	Scores all GO terms according to their relevance to each diseaseFil	MEDLINE, MesH-C and MesH-D terms, GO, RefSeq collection	Susceptibility loci
Geneseeker www.cmbi.kun.nl/GeneSeeker/	Filters candidate genes based on expression and phenotypic data from humans and mice	OXFORD, MIMMAP, HGMD, MGD, Zuerich, PubMed, OMIM, UniProt, MLC, TBASE, GeneCards	OMIM # Susceptibility loci Disease-related tissue
DGP cgg.ebi.ac.uk/services/dgp/	mutate based on their sequence properties	OMIM, NCBI Locuslink, Ensembl database, CoGenT, SwissPROT, BLAST protein database	Susceptibility loci
PROSPECTR and SUSPECTS www.genetics.med.ed.ac.uk/suspects/	Trained to differentiate between genes likely to be involved in a disease and those that are not, based on sequence features. (PROSPECTR) Scores similarity between the annotation and already known disease genes (SUSPECTS)	OMIM, HGMD, NCBI Homologene, InterPro protein domains, SwissPROT, Novartis, GO, Ensembl expression data	Susceptibility loci

[a] Abbreviations: BIND, biomolecular interaction network; BLAST, the basic local alignment search tool for finding regions of local similarity between sequences; cis-regulatory modules, combinations of TFBSs; CoGenT, complete genome-tracking database; Ensembl, project to produce and maintain automatic annotation on selected eukaryotic genomes; EST, expressed sequence tag; GeneCards, human gene-centric database; GEO, gene expression omnibus; HGMD, human gene mutation database; HPRD, human protein reference database; InterPro, database of protein families, domains and functional sites; KEGG, Kyoto encyclopaedia of genes and genomes; LocusLink, provides a single query interface to curated sequences and descriptive information about genetic loci; MEDLINE, medical literature analysis and retrieval system online; MGD, mouse genome database; MIMMAP, reformatted version of OMIM; MLC, mouse locus catalogue; NCBI, National Centre for Biotechnology Information; Novartis, Gene Expression Atlas of the Genomics Institute of the Novartis Research Foundation; OXFORD, human to mouse translation chromosomal locations; Reactome, curated database of biological processes in humans; RefSeq, reference sequence; SwissPROT, protein sequence database; TBASE, transgenic animals and targeted mutations; TFBS, transcription factor binding sites; UniProt; universal protein database; Y2H interactions, yeast two-hybrid interactions; Zuerich, chromosomal deletion and duplication map of malformation.

disease using the parameters conservation, phylogenetic extent, protein length and paralogy. G2D scores all the terms of GO according to their relevance to the disease. Endeavour is a software application for the computational prioritization of test genes based on a training set of genes already known to be involved in the disease of interest. The ranking of a test gene is based on its similarity with the training genes.
These six tools were combined to analyse the five overlapping T2D and obesity loci encompassing the 612 positional candidate genes (methodology box 1). This strategy resulted in 27 candidate genes (table 2), many of which qualify as "shared disease" candidate genes when looking at their function.

When the 27 genes were grouped based on GO terms, it was found that five of the identified genes toll-like receptor 2 (*TLR2*), friends leukemia virus integration 1 (*FLI1*) fibrinogen beta/gamma chain (*FGB, FGG*) and scavenger receptor class B member 1 (*SCARB1*) were involved in immunity and defense. It is known that low-grade inflammation in visceral fat of obese individuals causes insulin resistance and subsequently T2D. Although there is little evidence to date that causally links inflammation and obesity, there are very recent data showing a role for inflammation in weight control [55]. Mice deficient for interleukin 18 (IL18) show an increased food intake resulting in accumulation of fat tissue. The insulin resistance seen in these IL18-deficient mice is secondary to obesity and involves an enhanced expression of genes associated with gluconeogenesis in the liver, resulting from defective phosphorylation of signal transducer and activator of transcription 3 (STAT3). Hence, *TLR2, FLI1, FGG, FGB* and *SCARB1* might be interesting candidates to further investigate the link between satiety and inflammation and are promising "shared disease" candidate genes.

Box 1. Methodology

Identification of susceptibility loci

The degree of evidence for all reported T2D loci was quantified as follows: a locus with a LOD score of 3 or more was considered significant, a locus with a LOD score between 2.2 and 3 was considered suggestive, and a locus with a LOD score between 1 and 2.2 was considered nominal. For T2D we included only those loci that were found significant at least once, or were found suggestive in at least one study and at least nominal in two or more studies. The inclusion of the second category of loci was based on a study by Wiltshire *et al.* [69] in which it was postulated that locus counting is a useful additional tool for the evaluation of genome scan data for complex trait loci. We used the same two criteria to determine the loci from the five papers published on obesity since 2004 and combined these loci with those from Bell *et al*.. As obesity phenotypes BMI, serum leptin levels, abdominal subcutaneous and visceral fat, percentage body fat were included. All these phenotypes were used as continuous quantitative traits as well as with various cut-off levels.

Gene identification methods

Prioritizer, Endeavour, DGP, Geneseeker, G2D and P&S were combined to analyse the five overlapping T2D and obesity loci encompassing the 612 positional candidate genes. The loci 4q32, 6q22-6q24, 11q24, 12q24 and 20q12-20q13 were run in all systems as input. Additionally, Endeavour and P&S had to be trained with a set of genes. *ACDC, ADRA2A, ADRA2B, ADRB1, ADRB2, ADRB3, LEP, LEPR, NR3C1, UCP1, UCP2, UCP3 6, PPARG, KCNJ11* and *TCF7L2* [2] are already known to be involved in the diseases and were therefore used as training genes. Geneseeker required disease-related tissue as input. Fat, adipose, hypothalamus, pituitary, gut, liver, kidney and muscle were used. G2D needed OMIM # as an additional input, for T2D this is #125853. OMIM #601665 for obesity was not recognized by the program and was therefore omitted.

Identification of candidate genes

Geneseeker pinpoints to genes that show expression in disease-related tissue. Therefore we took all genes pinpointed by this method into consideration. All other tools produce rankings and therefore the top 20 genes from each method was included for comparison. A gene was considered interesting as a candidate gene if it was indicated by three or more of the tools. Because Endeavour, DGP and P&S partly use the same input information and show quite similar output, candidate genes were excluded if they were solely identified by these three methods.

Table 2. List of 27 genes selected by the six disease gene identification tools, showing how the genes were prioritized and whether they are already genetically or functionally associated with obesity or T2D

Gene	Tool						Genetic or functional association with type 2 diabetes or obesity	Refs
	Geneseeker	Prioritizer	DGB	PandS	G2B	Endeavour		
Chr4								
NPY1R			*	*	*	*	Involved in physiological regulation of energy balance	[7]
NPY2R		*	*	*		*	Lower BMI (P = 0.017) in 585T>C homozygous men. Lower allele and homozygosity frequency of 585T>C in obese and morbid obese men (P = 0.007 and P = 0.002, respectively). Obese T2D phenotype of ob/ob mice partially mediated by signalling through the NPY2R	[73] [74,75]
NPY5R		*	*			*	NPY5R mediates food intake in lean rats	[76]
CPE	*	*	*	*		*	CPE mRNA expression levels higher in visceral adipose tissue compared with subcutaneous adipose tissue in morbidly obese subjects (P<0.03). BKS mice homozygous for the CPEfat mutation are severely obese, hyperinsulinaemic and hyperglycaemic	[77,78]
FGB		*	*	*		*		
FGG			*	*	*	*		
CTSO		*		*		*	Homologue of CTSS. Obesity is characterized by high circulating levels of CTSS (P<0.0001); moreover, CTSS mRNA levels (P = 0.006) and protein levels (P<0.05) in obese subcutaneous adipose tissue are increased compared with lean subjects	[79]
GLRB		*	*			*		
TLR2	*			*	*	*	Homologue of TLR4. TLR4 mRNA is induced during adipocyte differentiation and its level is enhanced in the fat tissues of obese db/db mice. TLR4 activation in 3T3-L1 adipocytes provokes insulin resistance	[80]

Gene	Tool						Genetic or functional association with type 2 diabetes or obesity	Refs
	Geneseeker	Prioritizer	DGB	PandS	G2B	Endeavour		
Chr6								
GRM1	*		*	*		*		
OPRM1			*	*	*	*	A single SNP (rs648007) in OPRM1 associated with T2D	
							(P = 0.013) in African-Americans	
LATS1				*	*	*		
ESR1	*	*	*	*		*		
ENPP1	*		*	*	*		Upregulation of ENPP1 transcription in liver (P = 0.025) and brain (P = 0.034) of diabetic rabbits compared with controls. ENPP1 121Q allele predicts susceptibility to T2D in south Asians (P = 0.01) and Caucasians (P = 0.003). The Q allele of K121Q (OR = 1.6) and the T allele of rs997509 (OR = 4.7) were found to be associated with T2D in obese subjects from Poland. A risk haplotype was found to be associated with childhood obesity (OR = 1.69), adult morbid and moderate obesity (OR = 1.5 and OR = 1.37, respectively) and T2D (OR = 1.56)	[66-69]
Chr11								
FLI1				*	*	*		
KCNJ5		*	*	*	*	*	Homologue of KCNJ11. This gene contributes significantly to the risk of T2D	[2]
ROBO4		*		*		*		
Chr12								
AACS		*	*	*				
SCARB1			*	*		*	Key component in the reverse cholesterol transport pathway.	[81]
							Genetically associated with differences in insulin sensitivity	
							in healthy subjects	
TCF1	*	*	*	*			Responsible for MODY, an uncommon monogenetic form of early onset T2D	[2]
NCOR2		*	*	*		*	NCOR2 has an important role in the adipocyte by inhibiting	[65]
							adipocyte differentiation via repression of PPAR-g activity	

Gene	Tool						Genetic or functional association with type 2 diabetes or obesity	Refs
	Geneseeker	Prioritizer	DGB	PandS	G2B	Endeavour		
Chr20								
GNAS	*		*	*		*	Targeted disruption of the GNAS gene in mice leads to distinct phenotypes in heterozygotes, depending on whether the maternal (m /+) or paternal (+/p) allele is mutated. m /+ mice become obese, whereas +/p mice are thinner than normal. Both m /+ and +/p mice have greater sensitivity to insulin, with low to normal fasting glucose levels, low fasting insulin levels, improved glucose tolerance and exaggerated hypoglycaemic response to administered insulin	[82]
LAMA5			*	*		*		
PCK1		*	*	*		*	The OR for T2D among subjects with one or two copies of −232G compared with −232C homozygotes was found to be 1.9 in a native Canadian Oji-Cree sample and 2.8 in a Caucasian sample. This association was not replicated in a German Caucasian population. An in vitro experiment showed that the −232G construct was resistant to downregulation by insulin compared with a construct containing 232C	[62-64]
PPGB			*	*		*		
HNF4A	*		*	*	*		Responsible for MODY, an uncommon monogenetic form of early onset T2D	[2]
PTGIS			*	*	*			

Key and abbreviations: *, gene is pointed out by Geneseeker in general or is ranked in the top 20 by all other bioinformatics tools; AACS, acetoacetyl-CoA synthetase; BKS mice, a diabetes-permissive inbred strain that arose through a genetic contamination of the C57BL/6J (B6) strain, probably by DBA/2J; CTSO, cathepsin O; CTSS, cathepsin S; db/db mice, mouse model for diabetic dyslipidaemia; ESR1, oestrogen receptor 1; GLRB, glycine receptor b; GRM1, glutamate receptor metabotropic 1; LAMA5, laminin a5; LATS1, large tumour suppressor homologue; MODY, maturity onset diabetes of the young; NPY1R, neuropeptide Y receptor Y1; NPY2R, neuropeptide Y receptor Y2; NPY5R, neuropeptide Y receptor Y5; ob/ob mice, mice with a defect in the leptin gene, and an excellent model for studying obesity; OPRM1, opioid receptor m1; OR, odds ratio; PPGB, protective protein for b-galactosidase; PTGIS, prostaglandin I2 (prostacyclin) synthase; ROBO4, roundabout homolog 4, magic roundabout (Drosophila).

The thrifty gene hypothesis

The group of 27 genes also contained 10 genes involved in metabolism, sloth and gluttony (table 3). This observation may point towards a role for thrifty genes as being important in the shared molecular basis of obesity and T2D.

Table 3. Genes from the set of 27 genes pointed out by multiple gene identification systems which are associated with the following thrifty GO terms: metabolic, sloth and gluttony.

Thrifty GO-term	Genes
metabolic	
fatty acid metabolism	*AACS, PTGIS*
gluconeogenesis	*PCK1*
lipid, fatty acid and steroid metabolism	*SCARB1, PTGIS, AACS*
glucose metabolism	*NPY1R*
energy reserve metabolism	*GNAS*
insulin processing	*CPE*
sloth	
locomotory behavior	*NPY1R, NPY2R*
gluttony	
eating behavior	*NPY1R, NPY2R, NPY5R*

Human evolution has shaped the genome of modern man and one major driver of natural selection is famine [56]. During the periods of prolonged famine that plagued our early ancestors, a survival advantage would have been conferred by genes favouring the economical use and storage of energy; the so-called thrifty genes. This theory was initially proposed by Neel [57] who focused on the efficient use of glucose as a biological fuel. He suggested that evolutionary pressure to preserve glucose for use by the brain during starvation led to a genetic propensity towards insulin resistance in peripheral tissue. In the Western world, food is, in general, easily available and plentiful, so these thrifty genes are maladaptive in modern society and may now contribute to susceptibility for obesity and T2D. However, although these evolutionary theories, focussing on the potential survival advantages of thrifty genes that are now maladaptive, are of great interest, they are speculative and difficult to prove [58]. Thriftiness can take many forms: (i) metabolic, an energy-sparing super-

efficient metabolism, (ii) adipogenic, a propensity to rapid fat gain, (iii) physiological, an ability to switch off non-essential processes such as reproductive, thermogenic and immune capabilities, (iv) gluttony, a tendency to gorge when food is available, and (v) sloth, a tendency to conserve energy through inactivity [59]. Physiological thriftiness is not very likely to cause obesity and/or T2D, as the ability to switch off non-essential processes during famine will not be clearly maladaptive during normal or excessive food intake. However, the other forms of thriftiness could be plausible characteristics of genes that are maladaptive in modern society. Although the ten genes presented in table 3 would fit in with a "thrifty gene theory" based on their function (as they may influence mechanisms such as energy reserve metabolism and eating behaviour), in-depth genetic studies are needed to prove this theory. It would be interesting to compare the allele frequencies of these genes among different human populations with respect to food supply (past and present) and native climate. It would also be interesting to study the effect of long-term energy restriction on the expression of these genes [60].

Candidate T2D and obesity genes

In addition to the inflammatory and thrifty genes mentioned above, the computational disease gene identification methods indicated some interesting genes already known to be associated with T2D or obesity. These include transcription factor 1 (*TCF1*), hepatocyte nuclear factor 4, alpha (*HNF4A*), opioid receptor mu 1 (*OPRM1*), phosphoenolpyruvate carboxykinase 1 (*PCK1*), neuropeptide Y receptor 2 (*NPY2R*), ectonucleotide pyrophosphatase/phosphodiesterase 1 (*ENPP1*), guanine nucleotide binding protein alpha stimulatin complex (*GNAS*), carboxypeptidase (*CPE*), and nuclear receptor co-repressor 2 (*NCOR2*). Many of these genes show genetic association to either T2D or obesity, or are in some other way functionally associated with either one of the disorders (for details see table 2).

Of particular interest is the *PCK1* gene, a main control point for the regulation of gluconeogenesis. A promoter SNP (-232C→G) in *PCK1* is associated with T2D [61]. The odds ratio (OR) for T2D among individuals with one or two copies of -232G compared with -232C/C homozygotes was 1.9 in a Canadian Oji-Cree Indian sample, and 2.8 in a Caucasian sample. However, this association was not replicated in a German Caucasian population [62]. An in vitro experiment in three different cell

lines showed that the -232G construct was resistant to down-regulation by insulin compared to a construct that did contain 232C [61]. The common assumption is that mutations in *PCK1* lead to excessive glucose production through hepatic gluconeogenesis. However, there is an alternative explanation in which mutations at the *PCK1* locus could selectively affect *PCK1* expression in adipose tissue. This would result in changes in glyceroneogenesis that would affect the storage and releases of fatty acids. Beale *et al.* [63] therefore proposed the *PCK1* gene as a candidate gene for both T2D and obesity.

Another interesting gene is *NCOR2*. The protein encoded by this gene (*NCOR2*) interacts with PPARγ). PPAR)γ is an inflammatory factor that is also involved in the development of adipose tissue. Genetic studies have implicated the *PPARG* gene with obesity as well as T2D. NCOR2 also plays an important role in the adipocyte by inhibiting adipocyte differentiation via repression of PPARγ activity [64]. Hence, the *NCOR2* gene is another interesting candidate gene to investigate for its susceptibility to both obesity and T2D.

The *ENPP1* gene was also indicated by multiple gene identification systems. It is an inhibitor of the insulin receptor. Quantitative PCR analysis revealed a significant upregulation of *ENPP1* transcription in liver (p=0.025) and brain (p=0.034) of diabetic rabbits compared with controls [65]. The polymorphic *ENPP1* 121Q allele predicted genetic susceptibility to T2D in a South Asian sample (P=0.01) and a Caucasian sample (p=0.003). A three-allele risk haplotype also showed association with childhood obesity (OR=1.69), adult morbid and moderate obesity (OR=1.5 and OR=1.37, respectively) and T2D (OR=1.56, p=0.00002) 66. This makes *ENPP1* the first example of a common genetic link between childhood obesity, adult obesity and T2D.

Discussion

Complex traits such as obesity and T2D pose special challenges for genetic analyses because of gene-gene and gene-environment interactions, genetic heterogeneity, and low penetrance of the individual genes. The heterogeneity means it is difficult to generalize genome scan results over different populations and ethnicities. In addition, the multifactorial nature of complex traits assumes that the contribution of each of the susceptibility genes is likely to be small, and that only the joint effect of several

susceptibility genes in combination with environmental factors will lead to disease 46. It is therefore not surprising that large numbers of chromosomal regions have been implicated in disease susceptibility of both T2D and/or obesity, and hence, analysing all the individual positional candidate genes and loci will be a daunting task. Applying computational disease gene identification methods can be hugely helpful in the hunt for complex disease genes.

Recently, Tiffin *et al.* [67] analyzed 9556 positional candidate genes using multiple bioinformatics tools that could be implicated in T2D and/or obesity. Their approach was different from the approach in this study and resulted in a different list of genes for the following reasons. They included all susceptibility loci for either one of the traits which resulted in inclusion of nearly half the genome. Hence, they will have indicated genes that could be responsible for either one of the disorders, whereas the approach in the present review focused only on the overlapping T2D and obesity loci. In addition, two extra computational methods (Prioritizer and Endeavour) were used in the present review that incorporate a wider range of biological data sources than the other tools.

This review has yielded an interesting list of candidate genes by investigating the overlapping chromosomal linkage regions for T2D and obesity, using a combination of computational disease gene identification methods. Many of these identified genes are excellent candidates to study further for their role in the shared disease aetiology between obesity and T2D, and a few have already been genetically or functionally associated with both disorders (*ENPP1, NPY2R*). Although this cannot be taken as evidence that these computational methods work, it is tempting to assume that at least some of these genes may be true candidates, especially as this list includes genes belonging to the inflammatory pathway recently suggested to form an important molecular link between obesity and T2D [4]. Based on the candidate gene list presented here, it can be speculated that the molecular link between obesity and T2D extends beyond low-grade inflammation and may also contain thrifty genes. It will be interesting to see whether high-resolution SNP typing of these candidate genes in obese and/or T2D cohorts can be used to establish genetic association. In addition, these genes might be interesting candidates for identifying quantitative trait loci (QTLs) affecting obesity and T2D phenotypes in mice [68]. The list of 27 genes and the pathways identified may also help in the further interpretation of genome-wide genetic association data for T2D and obesity.

Acknowledgements

The work described in this review was made possible through financial support from SenterNovem (IOP genomics grant IGE05012). We thank Jackie Senior for critically reading the manuscript.

References

1. (2003) Report of the expert committee on the diagnosis and classification of diabetes mellitus. Diabetes Care 26 Suppl 1, S5-20
2. McCarthy, M.I., and Zeggini, E. (2006) Genetics of type 2 diabetes. Curr Diab Rep 6, 147-154
3. Grant, S.F., et al. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 38, 320-323
4. Dandona, P., et al. (2004) Inflammation: the link between insulin resistance, obesity and diabetes. Trends Immunol 25, 4-7
5. Botstein, D., and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl, 228-237
6. Bell, C.G., et al. (2005) The genetics of human obesity. Nat Rev Genet 6, 221-234
7. Chen, G., et al. (2005) A genome-wide scan for quantitative trait loci linked to obesity phenotypes among West Africans. Int J Obes (Lond) 29, 255-259
8. Cornes, B.K., et al. (2005) Sex-limited genome-wide linkage scan for body mass index in an unselected sample of 933 Australian twin families. Twin Res Hum Genet 8, 616-632
9. Lewis, C.E., et al. (2005) Sex-specific findings from a genome-wide linkage analysis of human fatness in non-Hispanic whites and African Americans: the HyperGEN study. Int J Obes (Lond) 29, 639-649
10. Norris, J.M., et al. (2005) Quantitative trait loci for abdominal fat and BMI in Hispanic-Americans and African-Americans: the IRAS Family study. Int J Obes (Lond) 29, 67-77
11. Wilson, S.G., et al. (2006) Linkage and potential association of obesity-related phenotypes with two genes on chromosome 12q24 in a female dizygous twin cohort. Eur J Hum Genet 14, 340-348
12. Lindgren, C.M., et al. (2002) Genomewide search for type 2 diabetes mellitus susceptibility loci in Finnish families: the Botnia study. Am J Hum Genet 70, 509-516
13. Aulchenko, Y.S., et al. (2003) A genome-wide search for genes involved in type 2 diabetes in a recently genetically isolated population from the Netherlands. Diabetes 52, 3001-3004
14. Avery, C.L., et al. (2004) Linkage analysis of diabetes status among hypertensive families: the Hypertension Genetic Epidemiology Network study. Diabetes 53, 3307-3312
15. Busfield, F., et al. (2002) A genomewide search for type 2 diabetes-susceptibility genes in indigenous Australians. Am J Hum Genet 70, 349-357
16. Duggirala, R., et al. (1999) Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. Am J Hum Genet 64, 1127-1140
17. Ehm, M.G., et al. (2000) Genomewide search for type 2 diabetes susceptibility genes in four American populations. Am J Hum Genet 66, 1871-1881
18. Elbein, S.C., et al. (1999) A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. Diabetes 48, 1175-1182
19. Ghosh, S., et al. (2000) The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. Am J Hum Genet 67, 1174-1185
20. Hanis, C.L., et al. (1996) A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nat Genet 13, 161-166
21. Hanson, R.L., et al. (1998) An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. Am J Hum Genet 63, 1130-1138
22. Hsueh, W.C., et al. (2003) Genome-wide and fine-mapping linkage studies of type 2 diabetes and glucose traits in the Old Order Amish: evidence for a new diabetes locus on chromosome 14q11 and confirmation of a locus on chromosome 1q21-q24. Diabetes 52, 550-557
23. Hunt, K.J., et al. (2005) Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study. Diabetes 54, 2655-2662
24. Iwasaki, N., et al. (2003) Mapping genes influencing type 2 diabetes risk and BMI in Japanese subjects. Diabetes 52, 209-213
25. Luo, T.H., et al. (2001) A genome-wide search for type II diabetes susceptibility genes in Chinese Hans. Diabetologia 44, 501-506
26. Mahtani, M.M., et al. (1996) Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families. Nat Genet 14, 90-94
27. Martin, L.J., et al. (2002) A quantitative trait locus influencing type 2 diabetes susceptibility maps to a region on 5q in an extended French family. Diabetes 51, 3568-3572
28. Mori, Y., et al. (2002) Genome-wide search for type 2 diabetes in Japanese affected sib-pairs confirms susceptibility genes on 3q, 15q, and 20q and identifies two new candidate Loci on 7p and 11p. Diabetes 51, 1247-1255

29. Nawata, H., et al. (2004) Genome-wide linkage analysis of type 2 diabetes mellitus reconfirms the susceptibility locus on 11p13-p12 in Japanese. J Hum Genet 49, 629-634
30. Ng, M.C., et al. (2004) Genome-wide scan for type 2 diabetes loci in Hong Kong Chinese and confirmation of a susceptibility locus on chromosome 1q21-q25. Diabetes 53, 1609-1613
31. Parker, A., et al. (2001) A gene conferring susceptibility to type 2 diabetes in conjunction with obesity is located on chromosome 18p11. Diabetes 50, 675-680
32. Permutt, M.A., et al. (2001) A genome scan for type 2 diabetes susceptibility loci in a genetically isolated population. Diabetes 50, 681-685
33. Pratley, R.E., et al. (1998) An autosomal genomic scan for loci linked to prediabetic phenotypes in Pima Indians. J Clin Invest 101, 1757-1764
34. Rotimi, C.N., et al. (2004) A genome-wide search for type 2 diabetes susceptibility genes in West Africans: the Africa America Diabetes Mellitus (AADM) Study. Diabetes 53, 838-841
35. Sale, M.M., et al. (2004) A genome-wide scan for type 2 diabetes in African-American families reveals evidence for a locus on chromosome 6q. Diabetes 53, 830-837
36. Silander, K., et al. (2004) A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. Diabetes 53, 821-829
37. van Tilburg, J.H., et al. (2003) A genome-wide scan in type 2 diabetes mellitus provides independent replication of a susceptibility locus on 18p11 and suggests the existence of novel Loci on 2q12 and 19q13. J Clin Endocrinol Metab 88, 2223-2230
38. Vionnet, N., et al. (2000) Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. Am J Hum Genet 67, 1470-1480
39. Watanabe, R.M., et al. (2000) The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. II. An autosomal genome scan for diabetes-related quantitative-trait loci. Am J Hum Genet 67, 1186-1200
40. Wiltshire, S., et al. (2001) A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. Am J Hum Genet 69, 553-569
41. Xiang, K., et al. (2004) Genome-wide search for type 2 diabetes/impaired glucose homeostasis susceptibility genes in the Chinese: significant linkage to chromosome 6q21-q23 and chromosome 1q21-q24. Diabetes 53, 228-234
42. Zhao, J.Y., et al. (2005) An autosomal genomic scan for loci linked to type 2 diabetes in northern Han Chinese. J Mol Med 83, 209-215
43. Reynisdottir, I., et al. (2003) Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. Am J Hum Genet 73, 323-335
44. Frayling, T.M., et al. (2003) Young-onset type 2 diabetes families are the major contributors to genetic loci in the Diabetes UK Warren 2 genome scan and identify putative novel loci on chromosomes 8q21, 21q22, and 22q11. Diabetes 52, 1857-1863
45. Carlson, C.S., et al. (2004) Mapping complex disease loci in whole-genome association studies. Nature 429, 446-452
46. Glazier, A.M., et al. (2002) Finding genes that underlie complex traits. Science 298, 2345-2349
47. Becker, K.G. (2004) The common variants/multiple disease hypothesis of common complex genetic disorders. Med Hypotheses 62, 309-317
48. (2005) A haplotype map of the human genome. Nature 437, 1299-1320
49. Franke, L., et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78, 1011-1025
50. van Driel, M.A., et al. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res 33, W758-761
51. Adie, E.A., et al. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics 22, 773-774
52. Lopez-Bigas, N., and Ouzounis, C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32, 3108-3114
53. Perez-Iratxeta, C., et al. (2002) Association of genes to genetically inherited diseases using data mining. Nat Genet 31, 316-319
54. Aerts, S., et al. (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24, 537-544
55. Netea, M.G., et al. (2006) Deficiency of interleukin-18 in mice leads to hyperphagia, obesity and insulin resistance. Nat Med 12, 650-656

56. Darwin, C. (1859) On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. John Murray, Albemarle Street
57. Neel, J.V. (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14, 353-362
58. Lazar, M.A. (2005) How obesity causes diabetes: not a tall tale. Science 307, 373-375
59. Prentice, A.M., et al. (2005) Insights from the developing world: thrifty genotypes and thrifty phenotypes. Proc Nutr Soc 64, 153-161
60. Kagawa, Y., et al. (2002) Single nucleotide polymorphisms of thrifty genes for energy metabolism: evolutionary origins and prospects for intervention to prevent obesity-related diseases. Biochem Biophys Res Commun 295, 207-222
61. Cao, H., et al. (2004) Promoter polymorphism in PCK1 (phosphoenolpyruvate carboxykinase gene) associated with type 2 diabetes mellitus. J Clin Endocrinol Metab 89, 898-903
62. Gouni-Berthold, I., et al. (2006) Association of the promoter polymorphism -232C/G of the phosphoenolpyruvate carboxykinase gene (PCK1) with Type 2 diabetes mellitus. Diabet Med 23, 419-425
63.Beale, E.G., et al. (2004) Disregulated glyceroneogenesis: PCK1 as a candidate diabetes and obesity gene. Trends Endocrinol Metab 15, 129-135
64. Cohen, R.N. (2006) Nuclear receptor corepressors and PPARgamma. Nucl Recept Signal 4, e003
65. Eller, P., et al. (2006) Hepatic ENPP1 expression is induced in diabetic rabbits. Mamm Genome 17, 886-891
66. Meyre, D., et al. (2005) Variants of ENPP1 are associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. Nat Genet 37, 863-867
67. Tiffin, N., et al. (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res 34, 3067-3081
68. Singer, J.B., et al. (2004) Genetic dissection of complex traits with chromosome substitution strains of mice. Science 304, 445-448
69. Wiltshire, S., et al. (2002) Evaluating the results of genomewide linkage scans of complex traits by locus counting. Am J Hum Genet 71, 1175-1182

Supplementary Data

Table S1. Overview of type 2 diabetes genome-based linkage studies.

Study	Population	Configuration of families	Number typed	Phenotypes	Analyses used
ulchenko *et al.* 2003 [1]	isolated population, the Netherlands	79 nuclear families	217	T2D	likelihood method Terwilliger (1995)
very *et al.* 2004 [2]	hypertension families of Caucasian and African American origin	567 diabetic participants in 437 families	1545 Caucasians, 1608 African Americans	T2D	threshold model, QTL
usfield *et al.* 2002 [3]	Australian Aboriginal	one large pedigree from isolated population	138	T2D or IGH	PLA
uggirala *et al.* 1999 [4]	Mexican American	27 extended families	440	T2D + age at onset	VC
hm *et al.* 2000 [5]	U.S. Caucasian	162 mpx sibships (247 ASPs)	835	T2D or IGH	NPLA
	Mexican American	117 mpx sibships (259 ASPs)	591	T2D or IGH	NPLA
	African American	65 mpx sibships (62 ASPs)	229	T2D or IGH	NPLA
	Japanese American	18 mpx sibships (45 ASPs)	128	T2D or IGH	NPLA
lbein *et al.* 1999 [6]	Utah Mormon	42 extended pedigrees	468-618	IGT or T2D	PLA, NPLA
rayling *et al.* 2003 [7]	U.K. Warren 2	573 families	1206	young-onset T2D	NPLA
ihosh *et al.* 1998 [8]	Finnish	478 mpx sibships (719 ASPs)	1175 affected	T2D + other measures of glycaemia	NPLA, VC
anis *et al.* 1996 [9]	Mexican American	330 sib pairs	408	T2D	NPLA, PLA
anson *et al.* 1998 [10]	Pima Indian	90 extended families (containing 264 nuclear pedigrees)	966 offspring + parents	T2D + other related traits	NPLA, VC
such *et al.* 2003 [11]	Old Amish	691 individuals from founder population	691	T2D and/or IGH, glucose levels	NPLA
unt *et al.* 2005 [12]	Mexican American	39 pedigrees	906	T2D	threshold model
vasaki *et al.* 2003 [13]	Japanese	164 families, 256 ASP	368	T2D, BMI	NPL, QTL
indgren *et al.* 2002 [14]	Finland	58 mpx extended families, including 26 from previous studies	440	T2D or IGH	NPLA
uo *et al.* 2001 [15]	Chinese Hans	102 families	478	T2D	NPLA
fahtani *et al.* 1996 [16]	Finland	26 extended pedigrees	217	T2D or IGH + subsetting on insulin levels	NPLA
fartin *et al.* 2004 [17]	Caucasian	single extended family	480	T2D	VC
fori *et al.* 2002 [18]	Japanese	159 mpx families (224 ASPs)	359	T2D	NPLA
lawata *et al.* 2004 [19]	Japanese	102 ASP		T2D	NPLA
g *et al.* 2004 [20]	Hong Kong Chinese	64 families		T2D	NPLA
arker *et al.* 2001 [21]	Scandinavian	353 mpx families (480 ASPs)	1488	T2D or IGH stratified by age	NPLA

References Table S1

1. Aulchenko, Y.S., et al. (2003) A genome-wide search for genes involved in type 2 diabetes in a recently genetically isolated population from the Netherlands. Diabetes 52, 3001-3004
2. Avery, C.L., et al. (2004) Linkage analysis of diabetes status among hypertensive families: the Hypertension Genetic Epidemiology Network study. Diabetes 53, 3307-3312
3. Busfield, F., et al. (2002) A genomewide search for type 2 diabetes-susceptibility genes in indigenous Australians. Am J Hum Genet 70, 349-357
4. Duggirala, R., et al. (1999) Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. Am J Hum Genet 64, 1127-1140
5. Ehm, M.G., et al. (2000) Genomewide search for type 2 diabetes susceptibility genes in four American populations. Am J Hum Genet 66, 1871-1881
6. Elbein, S.C., et al. (1999) A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. Diabetes 48, 1175-1182
7. Frayling, T.M., et al. (2003) A genome-wide scan in families with maturity-onset diabetes of the young: evidence for further genetic heterogeneity. Diabetes 52, 872-881
8. Ghosh, S., et al. (1999) Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. Proc Natl Acad Sci U S A 96, 2198-2203
9. Hanis, C.L., et al. (1996) A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nat Genet 13, 161-166
10. Hanson, R.L., et al. (1998) An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. Am J Hum Genet 63, 1130-1138
11. Hsueh, W.C., et al. (2003) Genome-wide and fine-mapping linkage studies of type 2 diabetes and glucose traits in the Old Order Amish: evidence for a new diabetes locus on chromosome 14q11 and confirmation of a locus on chromosome 1q21-q24. Diabetes 52, 550-557
12. Hunt, K.J., et al. (2005) Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study. Diabetes 54, 2655-2662
13. Iwasaki, N., et al. (2003) Mapping genes influencing type 2 diabetes risk and BMI in Japanese subjects. Diabetes 52, 209-213
14. Lindgren, C.M., et al. (2002) Genomewide search for type 2 diabetes mellitus susceptibility loci in Finnish families: the Botnia study. Am J Hum Genet 70, 509-516
15. Luo, T.H., et al. (2001) A genome-wide search for type II diabetes susceptibility genes in Chinese Hans. Diabetologia 44, 501-506
16. Mahtani, M.M., et al. (1996) Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families. Nat Genet 14, 90-94
17. Martin, L.J., et al. (2002) A quantitative trait locus influencing type 2 diabetes susceptibility maps to a region on 5q in an extended French family. Diabetes 51, 3568-3572
18. Mori, Y., et al. (2002) Genome-wide search for type 2 diabetes in Japanese affected sib-pairs confirms susceptibility genes on 3q, 15q, and 20q and identifies two new candidate loci on 7p and 11p. Diabetes 51, 1247-1255
19. Nawata, H., et al. (2004) Genome-wide linkage analysis of type 2 diabetes mellitus reconfirms the susceptibility locus on 11p13-p12 in Japanese. J Hum Genet 49, 629-634
20. Ng, M.C., et al. (2004) Genome-wide scan for type 2 diabetes loci in Hong Kong Chinese and confirmation of a susceptibility locus on chromosome 1q21-q25. Diabetes 53, 1609-1613
21. Parker, A., et al. (2001) A gene conferring susceptibility to type 2 diabetes in conjunction with obesity is located on chromosome 18p11. Diabetes 50, 675-680
22. Permutt, M.A., et al. (2001) A genome scan for type 2 diabetes susceptibility loci in a genetically isolated population. Diabetes 50, 681-685
23. Pratley, R.E., et al. (1998) An autosomal genomic scan for loci linked to prediabetic phenotypes in Pima Indians. J Clin Invest 101, 1757-1764
24. Reynisdottir, I., et al. (2003) Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. Am J Hum Genet 73, 323-335
25. Rotimi, C.N., et al. (2004) A genome-wide search for type 2 diabetes susceptibility genes in West Africans: the Africa America Diabetes Mellitus (AADM) Study. Diabetes 53, 838-841
26. Sale, M.M., et al. (2004) A genome-wide scan for type 2 diabetes in African-American families reveals evidence for a locus on chromosome 6q. Diabetes 53, 830-837
27. Silander, K., et al. (2004) A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. Diabetes 53, 821-829

28. van Tilburg, J.H., et al. (2003) A genome-wide scan in type 2 diabetes mellitus provides independent replication of a susceptibility locus on 18p11 and suggests the existence of novel Loci on 2q12 and 19q13. J Clin Endocrinol Metab 88, 2223-2230
29. Vionnet, N., et al. (2000) Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. Am J Hum Genet 67, 1470-1480
30. Watanabe, R.M., et al. (2000) The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. II. An autosomal genome scan for diabetes-related quantitative-trait loci. Am J Hum Genet 67, 1186-1200
31. Wiltshire, S., et al. (2004) Evidence from a large U.K. family collection that genes influencing age of onset of type 2 diabetes map to chromosome 12p and to the MODY3/NIDDM2 locus on 12q24. Diabetes 53, 855-860
32. Xiang, K., et al. (2004) Genome-wide search for type 2 diabetes/impaired glucose homeostasis susceptibility genes in the Chinese: significant linkage to chromosome 6q21-q23 and chromosome 1q21-q24. Diabetes 53, 228-234
33. Zhao, J.Y., et al. (2005) An autosomal genomic scan for loci linked to type 2 diabetes in northern Han Chinese. J Mol Med 83, 209-21539

Table S2. Review of chromosomal regions that show linkage in type 2 diabetes and obesity.

Chromosomal region	Study	Ref.	Locus marker	LOD	Disorder
1q21-1q25	Elbein et al.1999	[1]	CRP-APOA2	4.3	T2D
	Frahling et al.2003	[2]		2.4	T2D
	Hsueh et al.2003	[3]	D18S858	2.4	T2D
	Ng et al.2004	[4]	APOA2-D1S194	3.1	T2D
	Sale et al.2004	[5]	D1S1589		T2D
	Silander et al.2004	[6]	D1S1677	1.0a	T2D
	Vionnet et al.2000	[7]	APOA2-D1S484	3.0a	T2D
	Xiang et al.2004	[8]	D1S1589	3.3a	T2D
	Zhao et al.2004	[9]	D1S2851	2.2b	T2D
	Wiltshire et al.2001	[10]		2.4	T2D
2p23-2p21	Bell et al.2005	[11]			obesity
2p14	Chen et al.2005	[12]	D2S2739	3.3	obesity
2q24-2q32	Avery et al.2004	[13]		1.9	T2D
	Busfield et al.2002	[14]	D2S2345	3	T2D
	Iwasaki et al.2003	[15]	D2S1353-D2S1776	1	T2D
2q36-2q37	Hanis et al.1996	[16]	D2S125	4	T2D
	Elbein et al.1999	[1]	D2S336	2.2	T2D
	Luo et al.2001	[17]	D2S126	2.1	T2D
3p26	Norris et al..2005	[18]	D3S2387-MFD433	3.7	obesity
3p24-3p22	Lindgren et al.2002	[19]		2.2 b	T2D
	Ehm et al.2000	[20]	D3S2432	2.4	T2D
	Duggirala et al.1999	[21]	D3S2394-GAT128C02	2.5	T2D
	Iwasaki et al.2003	[15]	D3S3038	1.6 a	T2D
3q27	Bell et al.2005	[11]			obesity
4p15-4p14	Bell et al.2005	[11]			obesity

	Lewis et al.2005	[22]		1.7	obesity
4q31-4q32	Bell et al.2005	[11]			obesity
4q32-4q33	Lindgren et al.2002	[19]	D4S1595-D4S3047	2.5 b	T2D
	Ng et al.2004	[4]	D4S2349-D4S1644	2.6	T2D
	Sale et al.2004	[5]	D4S1629		T2D
	Aulchenko et al.2003	[23]	D4S431	1.3	T2D
	Rotimi et al.2004	[24]	D4S623-D4S2394	1.4	T2D
	Hsueh et al.2003	[3]	D4S1575-D4S424	1.3	T2D
	Permutt et al.2001	[25]			T2D
5q13	Ehm et al.2000	[20]	D5S1404	3.3	T2D
	Frahling et al.2003	[2]		1.3	T2D
6q21-6q25	Sale et al.2004	[5]	D6S1035	2.3	T2D
	Xiang et al.2004	[8]	D6S1040	6.2 a	T2D
	Aulchenko et al.2003	[23]	D6S1277-D6S1027	1.9	T2D
	Hanson et al.1998	[26]	D6S1009-D6S1003	1.4	T2D
	Iwasaki et al.2003	[15]	D6S1009	1.4	T2D
	Luo et al.2001	[17]	D6S264	1.4	T2D
6q22-6q25	Bell et al.2005	[11]			obesity
7q31-7q32	Bell et al.2005	[11]			obesity
9q21	Lindgren et al.2002	[19]	D9S1874-D9S153	3.9 b	T2D
	Luo et al.2001	[17]	D9S171	3.3 b	T2D
9q33	Iwasaki et al.2003	[15]	D9S282	5.3	T2D
	Aulchenko et al.2003	[23]	D9S1682	1	T2D
10p12-10p11	Bell et al.2005	[11]			obesity
10q26	Duggirala et al.1999	[21]		3.8	T2D
	Sale et al.2004	[5]	D10S217-D10S212		T2D
	Aulchenko et al.2003	[23]	D10S212	1.1	T2D
11q14-11q24	Bell et al.2005	[11]			obesity

11q24	Duggirala et al.1999	[21]	D11S4464-D11S912	2.3	T2D
	Hanson et al.1998	[26]	D11S4464-D11S912	1.7	T2D
	Elbein et al.1999	[1]	D11S925-D11S912	1.2	T2D
12q23-12q24	Bell et al.2005	[11]			obesity
	Cornes et al.2005	[27]		3	obesity
	Lewis et al.2005	[22]	D12S395-D12S2078	3.8	obesity
	Norris et al.2005	[18]	D12S1052-D12S1064	2.9	obesity
	Wilson et al.2006	[28]	D12S1612	3.2 b	obesity
12q24	Zhao et al.2004	[9]	D12S86	1.3 b	T2D
	Lindgren et al.2002	[19]	D12S2070-D12S324	2.1 b	T2D
	Mahtani et al.1996	[29]	D12S1349	3.3	T2D
	Rotimi et al.2004	[24]	D12S2070-D12S395	1.9	T2D
14q11.2-14q12	Hsueh et al.2003	[3]		3.5	T2D
	Aulchenko et al..2003	[23]	D14S283	1.2	T2D
	Wiltshire et al.2003	[10]	D14S70-D14S288	1.4	T2D
15q14	Mori et al.2002	[30]	D15S994	3.9	T2D
	Hanis et al.1996	[16]	CYP19	4	T2D
15q25.3	Lewis et al.2005	[22]	D15S655	3	obesity
17q11.2	Avery et al.2004	[13]		3	T2D
18p11	Aulchenko et al.2003	[23]	D18S63	2.3	T2D
	Ng et al.2004	[4]	D18S843	1	T2D
	Elbein et al.1999	[1]	D18S59	2.4	T2D
	Zhao et al.2004	[9]	D18S53	1.3 b	T2D
	van Tilburg et al.2003	[31]	D18S471-D18S843	2.3	T2D
	Parker et al.2001	[32]		3.8	T2D
	Reynisdottir et al.2003	[33]	D18S63	1.3	T2D
19q13.33-19q13.43	Bell et al.2005	[11]			obesity
20q12-20q13	Elbein et al.1999	[1]	D20S197	2.4	T2D
	Silander et al.2004	[6]	D20S197	2.5 a	T2D
	Zhao et al.2004	[9]	D20S178	1.0 b	T2D

	Permut et al.2001	[25]		2	T2D
	Rotimi et al.2004	[24]	D20S480-D20S171	2.6	T2D
	Mori et al.2002	[30]	D20S119	2.3	T2D
	Luo et al.2001	[17]	D20S197	1.5 b	T2D
	Iwasaki et al.2003	[15]	D20S107	2.0 a	T2D
20q11-20q13	Bell et al.2005	[11]			obesity
22q11-22q12	Iwasaki et al.2003	[15]	D22S420	2.2 a	T2D
	Frahling et al.2003	[2]	D22S420	2.5	T2D
	Zhao et al.2004	[9]	D22S274	1.4 b	T2D
	Avery et al.2004	[13]		3.4	T2D
X	Bell et al.2005	[11]			obesity

[a] multi point LOD score, [b] non parametric linkage score

References Table S2

1. Elbein, S.C., et al. (1999) A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. Diabetes 48, 1175-1182
2. Frayling, T.M., et al. (2003) Young-onset type 2 diabetes families are the major contributors to genetic loci in the Diabetes UK Warren 2 genome scan and identify putative novel loci on chromosomes 8q21, 21q22, and 22q11. Diabetes 52, 1857-1863
3. Hsueh, W.C., et al. (2003) Genome-wide and fine-mapping linkage studies of type 2 diabetes and glucose traits in the Old Order Amish: evidence for a new diabetes locus on chromosome 14q11 and confirmation of a locus on chromosome 1q21-q24. Diabetes 52, 550-557
4. Ng, M.C., et al. (2004) Genome-wide scan for type 2 diabetes loci in Hong Kong Chinese and confirmation of a susceptibility locus on chromosome 1q21-q25. Diabetes 53, 1609-1613
5. Sale, M.M., et al. (2004) A genome-wide scan for type 2 diabetes in African-American families reveals evidence for a locus on chromosome 6q. Diabetes 53, 830-837
6. Silander, K., et al. (2004) A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. Diabetes 53, 821-829
7. Vionnet, N., et al. (2000) Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. Am J Hum Genet 67, 1470-1480
8. Xiang, K., et al. (2004) Genome-wide search for type 2 diabetes/impaired glucose homeostasis susceptibility genes in the Chinese: significant linkage to chromosome 6q21-q23 and chromosome 1q21-q24. Diabetes 53, 228-234
9. Zhao, J.Y., et al. (2005) An autosomal genomic scan for loci linked to type 2 diabetes in northern Han Chinese. J Mol Med 83, 209-215
10. Wiltshire, S., et al. (2001) A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. Am J Hum Genet 69, 553-569
11. Bell, C.G., et al. (2005) The genetics of human obesity. Nat Rev Genet 6, 221-234
12. Chen, G., et al. (2005) A genome-wide scan for quantitative trait loci linked to obesity phenotypes among West Africans. Int J Obes (Lond) 29, 255-259
13. Avery, C.L., et al. (2004) Linkage analysis of diabetes status among hypertensive families: the Hypertension Genetic Epidemiology Network study. Diabetes 53, 3307-3312
14. Busfield, F., et al. (2002) A genomewide search for type 2 diabetes-susceptibility genes in indigenous Australians. Am J Hum Genet 70, 349-357
15. Iwasaki, N., et al. (2003) Mapping genes influencing type 2 diabetes risk and BMI in Japanese subjects. Diabetes 52, 209-213
16. Hanis, C.L., et al. (1996) A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nat Genet 13, 161-166
17. Luo, T.H., et al. (2001) A genome-wide search for type II diabetes susceptibility genes in Chinese Hans. Diabetologia 44, 501-506
18. Norris, J.M., et al. (2005) Quantitative trait loci for abdominal fat and BMI in Hispanic-Americans and African-Americans: the IRAS Family study. Int J Obes (Lond) 29, 67-77
19. Lindgren, C.M., et al. (2002) Genomewide search for type 2 diabetes mellitus susceptibility loci in Finnish families: the Botnia study. Am J Hum Genet 70, 509-516
20. Ehm, M.G., et al. (2000) Genomewide search for type 2 diabetes susceptibility genes in four American populations. Am J Hum Genet 66, 1871-1881
21. Duggirala, R., et al. (1999) Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. Am J Hum Genet 64, 1127-1140
22. Lewis, C.E., et al. (2005) Sex-specific findings from a genome-wide linkage analysis of human fatness in non-Hispanic whites and African Americans: the HyperGEN study. Int J Obes (Lond) 29, 639-649
23. Aulchenko, Y.S., et al. (2003) A genome-wide search for genes involved in type 2 diabetes in a recently genetically isolated population from the Netherlands. Diabetes 52, 3001-3004
24. Rotimi, C.N., et al. (2004) A genome-wide search for type 2 diabetes susceptibility genes in West Africans: the Africa America Diabetes Mellitus (AADM) Study. Diabetes 53, 838-841
25. Permutt, M.A., et al. (2001) A genome scan for type 2 diabetes susceptibility loci in a genetically isolated population. Diabetes 50, 681-685
26. Hanson, R.L., et al. (1998) An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. Am J Hum Genet 63, 1130-1138

27. Cornes, B.K., et al. (2005) Sex-limited genome-wide linkage scan for body mass index in an unselected sample of 933 Australian twin families. Twin Res Hum Genet 8, 616-632
28. Wilson, S.G., et al. (2006) Linkage and potential association of obesity-related phenotypes with two genes on chromosome 12q24 in a female dizygous twin cohort. Eur J Hum Genet 14, 340-348
29. Mahtani, M.M., et al. (1996) Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families. Nat Genet 14, 90-94
30. Mori, Y., et al. (2002) Genome-wide search for type 2 diabetes in Japanese affected sib-pairs confirms susceptibility genes on 3q, 15q, and 20q and identifies two new candidate loci on 7p and 11p. Diabetes 51, 1247-1255
31. van Tilburg, J.H., et al. (2003) A genome-wide scan in type 2 diabetes mellitus provides independent replication of a susceptibility locus on 18p11 and suggests the existence of novel loci on 2q12 and 19q13. J Clin Endocrinol Metab 88, 2223-2230
32. Parker, A., et al. (2001) A gene conferring susceptibility to type 2 diabetes in conjunction with obesity is located on chromosome 18p11. Diabetes 50, 675-680
33. Reynisdottir, I., et al. (2003) Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. Am J Hum Genet 73, 323-335

Chapter 4.1

Using genome-wide pathway analysis to unravel the etiology of complex diseases

Clara C. Elbers, Kristel R. van Eijk, Lude Franke, Flip Mulder,
Yvonne T. van der Schouw, Cisca Wijmenga, N. Charlotte Onland-Moret

Genet Epidemiol. 2009 Jul;33(5):419-31

Abstract

Several genome-wide association studies (GWAS) have been published on various complex diseases. Although, new loci are found to be associated with these diseases, still only very little of the genetic risk for these diseases can be explained. As GWAS are still underpowered to find small main effects, and gene-gene interactions are likely to play a role, the data might currently not be analyzed to its full potential. In this study, we evaluated alternative methods to study GWAS data. Instead of focusing on the single nucleotide polymorphisms (SNPs) with the highest statistical significance, we took advantage of prior biological information and tried to detect overrepresented pathways in the GWAS data. We evaluated whether pathway classification analysis can help prioritize the biological pathways most likely to be involved in the disease etiology.

In this study, we present the various benefits and limitations of pathway-classification tools in analyzing GWAS data. We show multiple differences in outcome between pathway tools analyzing the same dataset. Furthermore, analyzing randomly selected SNPs always results in significantly overrepresented pathways, large pathways have a higher chance of becoming statistically significant and the bioinformatics tools used in this study are biased towards detecting well-defined pathways.

As an example, we analyzed data from two GWAS on type 2 diabetes (T2D); the Diabetes Genetics Initiative (DGI) and the Wellcome Trust Case Control Consortium (WTCCC). Occasionally the results from the DGI and the WTCCC GWAS showed concordance in overrepresented pathways, but discordance in the corresponding genes. Thus, incorporating gene networks and pathway classification tools into the analysis can point towards significantly overrepresented molecular pathways, which cannot be picked up using traditional single-locus analyses. However, the limitations discussed in this study, need to be addressed before these methods can be widely used.

Introduction

Recently many genome-wide association studies (GWAS) have been published on several complex diseases, such as type 2 diabetes (T2D) [9,20,21,22,23,30]. Although these methods have been successful in finding new susceptibility genes for various complex diseases, not all the GWAS data is analyzed to its full potential. In most GWAS, single-locus case-control comparisons are used to identify SNPs associated with a disease. However, these methods are still underpowered to detect small risk effects. For instance, *PPARG*, a gene known to be associated with T2D, is not found to be associated in the individual GWAS on T2D. Moreover, the odds ratios (OR) of the genes found in T2D vary between 1.10 (confidence interval (CI): 1.07-1.14) for *TCF2* and 1.37 (CI: 1.31-1.43) for *TCF7L2* [8]. Even the largest GWAS on T2D, comprising 1,924 cases and 2,938 controls [1,30], had only 20% power to detect effects of this size. Therefore, it is likely that there are many more genes contributing similar or smaller effect sizes.

It is possible that single locus methods do not reflect the correct underlying model of association. There is growing evidence that gene-gene and gene-environment interactions contribute to complex diseases rather than single genes [14]. Several models for epistasis (i.e. gene-gene interactions) have been proposed [16], including models in which the genes alone have no effect on disease etiology but where their interaction modifies disease risk. The genetic contribution might even include higher-order interactions between genetic and non-genetic factors in complex biological pathways. The genes involved in these complex underlying pathways will probably not be picked up using traditional single-locus analyses, and different methods are needed to extract this information from GWAS datasets.

It can be assumed that only a limited number of biological pathways contribute to the etiology of complex traits [3]. This implies that a large proportion of the disease susceptibility genes will be functionally related and/or interact with one another in biological pathways. Several publicly accessible bioinformatics tools are now available for pathway classification analysis [27]; they sort genes into pre-defined pathways of cellular processes based on genomic and molecular information.
In this study we assessed the usefulness of different pathway classification tools (Webgestalt 'KEGG' and 'BioCarta' [31], GATHER [4] and DAVID [11], PANTHER [18]) to detect biologically overrepresented pathways in GWAS datasets.

We evaluated whether pathway classification analysis can help prioritize the biological pathways most likely to be involved in the disease etiology. As an example, we used data from two GWAS on T2D (the Diabetes Genetics Initiative (DGI) and the Wellcome Trust Case Control Consortium (WTCCC)) to investigate whether we could find overlapping overrepresented pathways between the datasets which cannot be picked up using traditional single-locus analyses. To study the robustness of our results, we analyzed the datasets with different in- and exclusion criteria. Furthermore, we analyzed 30 sets of randomly selected SNPs to test for bias of the pathway classification tools.

Methods

Study populations T2D GWAS

Both DGI and WTCCC have recently published a GWAS on T2D using a case-control design, using the Affymetrix GeneChip Human Mapping 500k Array Set [20,30]. The summaries of their results are publicly available online at http://www.broad.mit.edu/diabetes/scandinavs/type2.html and http://www.wtccc.org.uk/info/summary_stats.shtml. We used these datasets for our study.

DGI

The DGI is a collaboration of the Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Lund University, Sweden, and Novartis Institutes for Biomedical Research. The GWAS for T2D comprised 1,464 T2D cases and 1,467 controls from Finland and Sweden. The subjects were obtained from family-based studies (326 sibships discordant for T2D; 442 cases and 392 euglycemic controls) and from population-based studies (1,022 T2D cases and 1,075 euglycemic controls, matched for gender, age, BMI, and region of origin). All subjects were characterized for anthropometric measures, glucose tolerance and insulin secretion, lipids and apolipoproteins, and blood pressure. Genotyping of 500,568 SNPs was attempted in each sample and the overall call rate for passing SNPs was 99.2%. After filtering rare and monomorphic variants and applying stringent quality control filters, we had high-quality genotypes for 386,731 common SNPs for our analyses.

WTCCC

The WTCCC study comprised 1,924 T2D cases and 2,938 controls from the UK. The patients were all of British or Irish decent, and were obtained from the Diabetes UK Warren 2 study. They were recruited as part of family-based studies or as isolated cases. Diagnosis of T2D was based on currently prescribed diabetes-specific medication or on laboratory evidence of hyperglycemia. Other forms of diabetes were excluded. The controls were obtained from two different sources: the 1958 birth cohort and the UK Blood Service donors, both distributed nationwide. No relevant phenotypic data for any of these samples was available for this study. Of the 490,032 autosomal SNPs, 459,447 SNPs passed the initial quality control of the WTCCC [21].

Inclusion criteria

We obtained data on 386,731 and 459,447 SNPs from DGI and WTCCC, respectively. The DGI dataset contained fewer SNPs because of more stringent quality control filters than the WTCCC dataset [20.21]. Because of issues related to the analysis of X-chromosomal SNPs, we excluded X-chromosomal SNPs from our study. We also excluded a set of SNPs that did not meet our inclusion criteria (combined minor allele frequency (MAF) cases and controls > 0.01; call rate cases > 0.95; call rate controls > 0.95; Hardy Weinberg equilibrium (HWE) p-value controls < 0.001). Ultimately, 370,519 SNPs from the DGI and 390,025 SNPs from the WTCCC fulfilled our inclusion criteria.

Both DGI and WTCCC were designed to identify T2D susceptibility loci, but the two studies differed considerably in study design. Combining two different datasets increases the population size and therefore the power to detect associations, but does not take study design differences into account. We therefore decided to investigate pathways in the DGI and WTCCC datasets separately as well as in a combined dataset (supplementary information 1). We performed a Mantel-Haenszel procedure to obtain pooled estimates for the overlapping SNPs [19].

SNP selection

For our analyses we selected all SNPs that showed association with T2D with a p-value lower than 0.003 from each dataset. This resulted in 1,179 SNPs that were selected from DGI and 1,712 SNPs from WTCCC (figure 1). The threshold of $p < 0.003$ was chosen because this p-value was in all cases lower than the cut-off value where the observed p-value distribution deviated from expected. We also varied this threshold but, because of input-size related limitation of the bioinformatics tools the threshold of $p < 0.003$ was most suitable for our study.

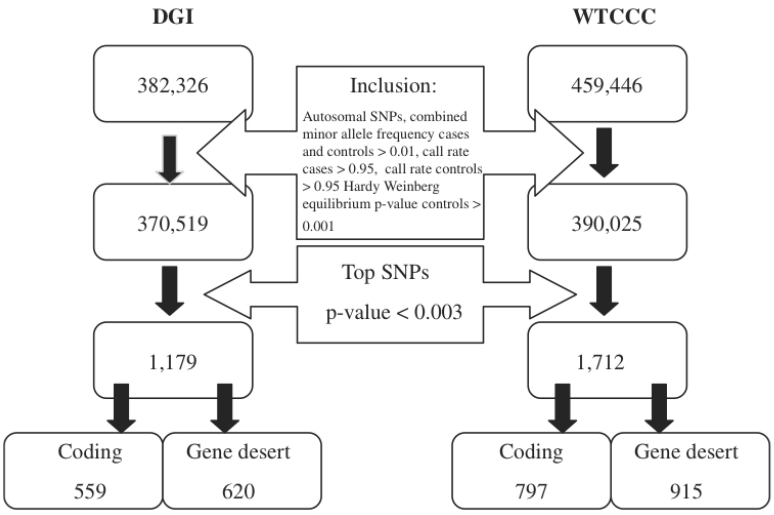


Fig. 1. Pipeline showing numbers of SNPs included at various stages of analyses.

Mapping SNPs to haplotype blocks

Because parts of the genome are inherited together, each SNP gives information about several other variants on that piece of chromosome. Because of these linkage disequilibrium (LD) patterns in the genome, a SNP can be mapped back to an LD block containing several genes. To perform pathway analyses on the GWAS data, we first allocated the genotyped SNPs to genomic regions, based on the LD structure around the SNPs. Unfortunately, we did not have access to the crude genotype data from the DGI or WTCCC. However, our lab has recently performed a GWAS in 778 UK individuals with celiac disease and 1,422 UK population controls using the Illumina HumanHap300 BeadChips [26]. The genome-wide genotype data from this study was used to characterize the LD structure throughout the genome. It is valid to use this population because these controls overlap with the 1958 birth cohort from the

WTCCC GWAS study. The patterns were based on a large dataset, making the prediction of the haplotype blocks somewhat more robust compared to estimates based on, for instance, the HapMap database.

We established the LD blocks around the top SNPs from the DGI and WTCCC data. We defined haploblocks as SNPs that were in LD with at least an r2 of 0.25 with the selected SNP. SNPs that were located in a block that contained one or more genes were called 'mappable SNPs'. SNPs which were located in gene deserts and mapped only to non protein-coding regions were excluded from further analyses ('non-mappable SNPs'). From the 1,179 SNPs in the DGI dataset and the 1,712 SNPs in the WTCCC dataset that showed association with a p-value lower than 0.003, 559 and 797 SNPs, respectively, were mappable (figure 1).

Network analyses

Most of our 'mappable SNPs' mapped back to loci that encompassed multiple genes. By including all these genes for pathway analyses, genes that were not involved in the disease could add noise to the results. To select the most likely T2D susceptibility gene(s) from each locus, we first used the gene network tool 'Prioritizer' [7]. This tool prioritizes those genes that are functionally related or that interact with genes in the other selected loci, assuming that true disease genes are mostly functionally related and will therefore be closer to each other in a gene interaction network than false-positive genes that have been randomly selected. Prioritizer uses a Bayesian approach to reconstruct a functional gene network, based upon known functional interactions from the Human Protein Reference Database (HPRD), Biomolecular Interaction Network Database (BIND) and Kyoto Encyclopedia of Genes and Genomes (KEGG). It also makes use of circumstantial evidence, derived from sharing of Gene Ontology (GO) terms, co-expression information from microarray data, deposited in the Gene Expression Omnibus (GEO), ~70,000 known protein-protein interactions in other organisms that have been mapped to orthogolous pairs of genes [15] and 3,000 predicted human protein-protein interactions [24]. Positional candidate genes, located in different loci but functionally closely related in a gene network are assigned higher interaction scores than positional candidate genes that were functionally further apart from each other. Detailed information on this scoring system is described elsewhere [7]. The lower the interaction p-value assigned to this interaction score, the more

likely it is that this specific gene is functionally closely related to another gene in the network.

Although this network tool generates only theoretical evidence, it can pinpoint probable disease-causing genes. For example, Prioritizer predicted the *NPY2R* gene as being associated with obesity and T2D [6], and a few months later Campbell et al. published a paper stating they had found association of *NPY2R* with both traits in 2,800 Caucasian individuals [2].

All the haploblocks that we defined based on the selected "mappable" SNPs were analyzed using the Prioritizer program. From each haploblock we selected those gene(s) that were, according to Prioritizer, functionally related to genes in other selected regions with an interaction p-value below 0.05 or, if none of the genes on a locus met this criterion, we selected the most significant gene per locus.

Pathway tools

After selecting the most probable candidate genes per locus, we sorted all genes into pathways using five pathway-classification tools. These tools test whether the number of genes from each pathway in our list of predicted candidate genes is higher than expected given the number of genes selected from the total number of genes. For example, from the total of 20,000 genes in the human genome, we sampled n genes in our study. Let there be y genes in a certain pathway according to the pathway-classification tool of which we sampled x genes in our study, these tools test whether x out of y is what you expect given n out of 20,000. These tools use either the Fisher exact test or a Baysian tests. This resulted in a ranking of biological pathways by p-value.

We explored different pathway tools to investigate whether different methods resulted in consistent outcomes. First, we used three pathway tools that classified genes based on KEGG pathways: Webgestalt 'KEGG' [31], GATHER [4] and DAVID [11]. Second, we investigated two tools that used other pathway classification methods: PANTHER [18] and Webgestalt 'BioCarta' [31].

Different inclusion and exclusion criteria

Because the in- and exclusion criteria that we chose might influence our results, we checked the consistency of our results by analyzing the datasets with several different in- and exclusion criteria:

1. **Threshold p-value.** We studied the influence of the threshold p-value that was chosen for the SNP selection, on the results by analyzing the datasets with different thresholds (0.001; 0.002; 0.003).

2. **Exclusion of the HLA region**. The HLA region encodes proteins of classical HLA class I and class II genes in the major histocompatibility complex (MHC) and is essential in immune recognition. This region is highly polymorphic and its LD extends across multiple HLA and non-HLA genes in the MHC [5]. The extended haploblocks with functionally related genes could bias the analysis towards interactions and pathways of immune functions, so this region could possibly have influenced our network and pathway analysis. We therefore performed our analysis without the genes in the HLA region. We defined the HLA region in this study as the region on chromosome 6, from base pair 20,000,000 till base pair 40,000,000.

3. **Network analysis without GO terms**. Prioritizer incorporates relatively unbiased genome-wide experimental datasets on molecular interactions like micro-array co-expression data (MA), human yeast two-hybrid interactions (Y2H), and high-throughput protein-protein interactions (PPI) from lower eukaryotes, but also the relatively more biased GO data source. GO-based analyses have a gene-centered view rather than focusing on physical and functional interactions between genes and are therefore more biased towards well-studied gene functions compared to PPI, Y2H and the MA. We therefore ran our network analysis twice, with and without the GO data source.

4. **Pathway analysis without network analysis**. To further evaluate whether the network analysis would bias our results towards certain well studied pathway, we then performed pathway analysis on our original gene set without pre-selection by 'Prioritizer' network analysis.

5. **Inclusion of all known T2D loci.** The ten recently discovered T2D susceptibility loci for T2D in the original GWAS studies were not all present in our selected gene list. However, because these genes are known to be involved in T2D-related pathways, we added the following ten genes, *PPARG, KCNJ11, TCF7L2, TCF2, WFS1, HHEX-IDE, SLC30A8, CDKAL1, CDKN2A-2B* and *IGF2BP2*, to both our network and pathway analyses.

Random checks

Genes vary greatly in size, LD blocks vary in the number of genes they contain, and pathways vary in the number of genes they contain. These differences influence the chances of each gene in each pathway to be selected in the gene set. In addition, larger pathways have more power to detect a certain difference in observed versus expected number of genes compared to smaller pathways. To investigate whether certain biological pathways were biased because of the problems outlined above, we repeated our analysis with 30 sets of randomly selected SNPs.

Results

Results of different pathway tools analyzing T2D GWAS as an example

To compare the outcomes of the different pathway tools, we focused on the ten strongest, overrepresented pathways in the GWAS datasets.

In the DGI dataset, the pathways 'cytokine-cytokine receptor interaction' and 'Jak-STAT signaling pathway' were overrepresented in the top-10 in all three KEGG pathway tools. The pathways 'Calcium signaling pathway', 'Huntington disease', 'ECM-receptor interaction' and 'Valine, leucine and isoleucine degradation' were present in the top-10 in two out of three pathway tools (table 1). The analyses with the pathway tools Panther and Webgestalt 'BioCarta' resulted in similar outcomes. Again the ten strongest, overrepresented pathways in the DGI dataset pointed towards chemokine- and cytokine-mediated inflammation pathways, while the Panther results also contained the 'Huntington disease' pathway (table 2).

In the WTCCC dataset, the 'cell cycle' and 'Wnt signaling' pathways were overrepresented in the top-10 in all three KEGG pathway tools. The 'adherens junction', 'neuroactive ligand-receptor interaction', 'C21-Steroid hormone metabolism', 'ECM-receptor interaction', 'apoptosis' and 'alkaloid biosynthesis II' pathways were present in the top-10 in two out of three pathway tools (table 3). Panther and Webgestalt 'BioCarta' pointed more towards pathways involved in inflammation and immunology, but the 'apoptosis' and 'Wnt signaling' pathways were also present in the top-10 most overrepresented pathways in the WTCCC dataset (table 4). The results of the combined dataset of the DGI and the WTCCC are presented in the supplementary information 1.

Overall we observed similarities as well as differences in outcome between the tools when analyzing the same dataset and discordance as well as concordance in the results from the DGI and the WTCCC GWAS.

Randomly selected SNPs

Although the SNPs and the corresponding genes were randomly selected, all random datasets resulted in significantly overrepresented pathways using the three pathway classification tools, Webgestalt, GATHER and PANTHER (supplementary figure 1). Only random analysis using DAVID did not always result in a statistically overrepresented pathway. As randomly selected genes should not result in overrepresented pathways, this suggests that the tools might be too sensitive for the detection of observed vs. expected differences. Therefore, the use of these bioinformatics tools for pathway analysis in GWAS might be limited.

In addition, the random selection of SNPs and relevant genes gives information on whether some pathways are more likely to be included in our analysis by chance than others. A high number of expected genes represented large pathways, a low number of expected genes represented small pathways. Supplementary figure 1 shows the correlation between the expected number of genes per pathway and the corresponding p-value for overrepresentation. We observed that especially in the pathway classification tools 'Panther' and Webgestalt KEGG, large pathways were favored to become significantly overrepresented in our analysis.

Analyzing T2D GWAS using different in- and exclusion criteria

To study the robustness of our results, we analyzed the datasets with different in- and exclusion criteria (table 5, results DGI). These changes did not affect our results and we found parallel pathways overrepresented. Although the ranking of the pathways did not change much with or without network analysis, the p-values for overrepresentation for each pathway was smaller after network analysis, indicating that making a pre-selection from all genes selected on knowledge of functional relatedness, increases the power to detect associated pathways. The analysis with different p-value thresholds showed that with p-values below 0.003 the number of genes that could be selected as input for the pathway analysis was too small to be able to detect significantly overrepresented pathways.

Table 1. Overrepresented KEGG pathways in the top results of the DGI's genome-wide association study on type 2 diabetes

KEGG pathway from Webgestalt	P-value	KEGG pathway from GATHER	P-value	KEGG pathway from DAVID	P-value
Cytomine-cytokine receptor interaction	8.08×10^{-6}	Huntington's disease	0.60	Jak-STAT signalling pathway	5.24×10^{-2}
Calcium signalling pathway	1.07×10^{-5}	Fatty acid biosynthesis (path 2)	0.61	Cytokine-cytokine receptor interaction	6.19×10^{-2}
Huntington's disease	1.08×10^{-5}	Benzoate degradation via hydroxylation	0.68	Fatty acid elongation in mitochondria	8.55×10^{-2}
Cell adhesion molecules (CAMs)	1.74×10^{-5}	Valine, leucine and isoleucine degradation	0.68	Valine, leucine and isoleucine degradation	0.21
Jak-STAT signalling pathway	5.87×10^{-5}	Vitamin B6 metabolism	0.68	Bile acid biosynthesis	0.23
MAPK signalling pathway	8.59×10^{-5}	d-arginine and d-ornithine metabolism	0.68	Fc epsilon ri signalling pathway	0.24
Long-term potentiation	2.16×10^{-4}	Calcium signalling pathway	0.72	Tight junction	0.25
ECM-receptor interaction	7.29×10^{-4}	Cytokine-cytokine receptor interaction	0.72	Hematopoietic cell lineage	0.28
Regulation of actin cytoskeleton	7.92×10^{-4}	Jak-STAT signalling pathway	0.72	Complement and coagulation cascade	0.32
Leukocyte transdothelial migration	9.30×10^{-4}	Toll-like receptor signalling pathway	0.72	Fatty acid metabolism	0.37

SNPs which showed association with T2D (P>0.003) were included in this study and were mapped backed to regions on the genome, and the predicted candidate genes were used for analysis. The number of observed genes for each KEGG pathway was compared to the number of expected genes per KEGG pathway. The top-10 ranking KEGG pathways per method are shown.

Table 2. Overrepresented PANTHER and BioCarta pathways in the top results of the DGI's genome-wide association study on type 2 diabetes

Panther Pathways	P-value	BioCarta pathways from Webgestalt	P-value
Ionotropic glutamate receptor pathway	3.77×10^{-4}	Dendritic cells in regulating TH1 and TH2 Development	6.21×10^{-5}
Wnt signalling pathway	7.00×10^{-4}	Cytokines and inflammatory response Th1/Th2 Differentiation	1.50×10^{-4}
Inflammation mediated by chemokine and cytokine signalling pathway	1.63×10^{-3}	Th1/Th2 Differentiation	6.22×10^{-4}
Angiogenesis	1.68×10^{-3}	The role of eosinophils in the chemokine network of allergy	6.87×10^{-4}
Oxidative stress response	4.17×10^{-3}	GATA3 participate in activating the Th2 cytokine genes expression	1.16×10^{-3}
Huntington disease	4.69×10^{-3}	Monocyte and its surface molecules	1.46×10^{-3}
Heterotrimeric G-protein signalling pathway-Gq α- and Go α-mediated pathway	5.61×10^{-3}	Selective expression of chemokine receptors during T-cell polarization	1.48×10^{-3}
Heterotrimeric G-protein signalling pathway-Gi α- and Gs α-mediated pathway	6.13×10^{-3}	Antigen dependent B cell activation	1.80×10^{-3}
Axon guidance mediated pathway by Slit/Robo	7.33×10^{-3}	FAS signalling pathway (CD95)	2.13×10^{-3}
O-antigen biosynthesis	7.63×10^{-3}	Acute myocardial infarction	2.19×10^{-3}

SNPs which showed association with T2D ($P<0.003$) were included in this study and were mapped backed to regions on the genome and the predicted candidate genes were used for analysis. The number of observed genes for each pathway was compared to the number of expected genes per pathway. The highest 10 ranking pathways per method are shown for Webgestalt ''BioCarta'' and PANTHER.

Table 3. Overrepresented KEGG pathways in the top results of the WTCCC's genome-wide association study on type 2 diabetes

KEGG pathways from Webgestalt	P-value	KEGG pathways from GATHER	P-value	KEGG pathways from DAVID	P-value
Focal adhesion	2.84×10^{-7}	Cell cycle	1.19×10^{-2}	Cell cycle	3.58×10^{-2}
Cell cycle	4.45×10^{-6}	C21-Steroid hormone metabolism	2.82×10^{-2}	Wnt signalling pathway	0.12
Inflammation mediated by chemokine and Wnt signalling pathway	1.28×10^{-5}	Ethylbenzene degradation	2.82×10^{-2}	C21-Steroid hormone metabolism	0.14
Neuroactive ligand-receptor interaction	2.31×10^{-5}	Adherens junction	5.18×10^{-2}	Apoptosis	0.19
MAPK signalling pathway	2.68×10^{-5}	Alzheimer's disease	5.34×10^{-2}	Cholera infection	0.23
Axon guidance	4.67×10^{-5}	Wnt signalling pathway	7.81×10^{-2}	Ethylbenzene degradation	0.29
Chronic myeloid leukemia	8.28×10^{-5}	Apoptosis	7.89×10^{-2}	Ephithelial cell signalling in *Helicobacter pylori* infection	0.29
Adherens junction	9.12×10^{-5}	Alkaloid biosynthesis II	8.80×10^{-2}	Complement and coagulation cascades	0.40
Regulation of actin cytoskeleton	9.49×10^{-5}	ECM-receptor interaction	9.07×10^{-2}	Alkaloid biosynthesis II	0.40
ECM-receptor interaction	2.04×10^{-4}	Huntington's disease	9.63×10^{-2}	Neuroactive ligand-receptor interaction	0.49

SNPs which showed association with T2D ($P<0.003$) were included in this study and were mapped backed to regions on the genome, and the predicted candidate genes were used for analysis. The number of observed genes was compared to the number of expected genes for each KEGG pathway. The top-10 ranking KEGG pathways per pathway classification tool are shown.

Table 4. Overrepresented PANTHER and BioCarta pathways in the top results of the WTCCC's genome-wide association study on type 2 diabetes

PANTHER pathways	P-value	BioCarta pathways Webgestalt	P-value
T cell activation	6.37×10^{-6}	Basic mechanism of action of PPARa, PPARb(d) and PPARg, and effects on gene expression	3.65×10^{-4}
Inflammation mediated by chemokine and cytokine signalling pathway	8.24×10^{-6}	Cell cycle: G1/S check point	4.20×10^{-4}
FAS signalling pathway	7.04×10^{-4}	Induction of apoptosis through DR3 and DR4/5 death receptors	5.79×10^{-4}
B cell activation	1.04×10^{-3}	Regulation of spermatogenesis by CREM	1.19×10^{-3}
Axon guidance mediated by semaphorins	2.41×10^{-3}	Activation of Csk by camp-dependent protein kinase inhibits signalling through the T cell receptor	1.32×10^{-3}
Integrin signalling pathway	2.96×10^{-3}	HIV-I Nef: negative effector of Fas and THF	1.69×10^{-3}
Wnt signalling pathway	3.29×10^{-3}	Lectin-induced complement pathway	2.11×10^{-3}
Huntington disease	4.90×10^{-3}	Rho cell motility signalling pathway	2.51×10^{-3}
Toll receptor signalling pathway	5.01×10^{-3}	Cyclins and cell cycle regulation	2.51×10^{-3}
Apoptosis signalling pathway	5.38×10^{-3}	NFkB activation by non-typeable *Hemophilus influenzae*	2.89×10^{-3}

SNPs which showed association with T2D ($P<0.003$) were included in this study and were mapped backed to regions on the genome, and the predicted candidate genes were used for analysis. The number of observed genes was compared to the number of expected genes for each KEGG pathway. The top-10 ranking KEGG pathways per pathway classification tool are shown.

Table V. Overrepresented pathways (Panther in the top results of the DGI's genome-wide association study on Type 2 diabetes

Overrepresented pathways DGI	P<0.001	P<0.002	P<0.003	P<0.003 exclusive HLA	P<0.003 without GO analysis	P<0.003 with known T2D loci	P<0.03 without network analysis
Ionotropic glutamate receptor pathway	**0.23**	**5.04×10^{-3}**	**3.77×10^{-4}**	1.26×10^{-3}	7.14×10^{-3}	**4.42×10^{-4}**	**4.26×10^{-3}**
Wnt signalling pathway	**1.19×10^{-3}**	**8.29×10^{-3}**	**7.00×10^{-4}**	**5.82×10^{-4}**	1.30×10^{-2}	**9.50×10^{-4}**	2.52×10^{-2}
Inflammation mediated by chemokine and cytokine signalling pathway	**0.05**	0.19	**1.63×10^{-3}**	**5.56×10^{-4}**	0.20	2.13×10^{-3}	1.95×10^{-2}
Angiogenesis	0.45	**3.30×10^{-3}**	1.68×10^{-3}	**2.01×10^{-4}**	**1.75×10^{-4}**	**2.12×10^{-3}**	2.24×10^{-2}
Oxidative stress response	0.38	0.09	4.17×10^{-3}	2.52×10^{-3}	**1.89×10^{-3}**	4.75×10^{-3}	2.87×10^{-2}
Huntington disease	1	0.26	4.69×10^{-3}	7.15×10^{-3}	0.17	5.63×10^{-3}	3.03×10^{-2}
Heterotrimeric G-protein signalling pathway-Gq α- and Go α-mediated pathway	1	0.57	5.61×10^{-3}	2.88×10^{-3}	2.94×10^{-3}	6.64×10^{-3}	6.25×10^{-2}
Heterotrimeric G-protein signalling pathway-Gi α- and Gs α-mediated pathway	1	1	6.13×10^{-3}	3.00×10^{-3}	0.19	7.34×10^{-3}	7.73×10^{-2}
Axon guidance mediated by Slit/Robo	1	0.75	7.33×10^{-3}	5.12×10^{-3}	**1.05×10^{-3}**	8.04×10^{-3}	3.00×10^{-2}
O-antigen biosynthesis	1	1	7.63×10^{-3}	6.25×10^{-3}	7.21×10^{-2}	8.04×10^{-3}	**1.72×10^{-2}**
VEGF signalling pathway	1	1	8.34×10^{-3}	1.09×10^{-3}	3.55×10^{-3}	9.44×10^{-3}	**1.84×10^{-2}**
Xanthine and guanine salvage pathway	1	1	1.02×10^{-2}	8.40×10^{-3}	3.58×10^{-3}	1.08×10^{-2}	2.29×10^{-2}
Intergrin signalling pathway	1	1	1.13×10^{-2}	5.47×10^{-3}	2.50×10^{-2}	1.36×10^{-2}	0.14
Cadherin signalling pathway	1	0.23	3.09×10^{-2}	1.81×10^{-2}	5.59×10^{-3}	3.53×10^{-2}	0.10
Hypoxia response via HIF activation	1	1	4.67×10^{-2}	3.62×10^{-2}	1.15×10^{-2}	4.99×10^{-2}	0.13
Alzheimer disease—presenilin pathway	1	0.47	9.12×10^{-2}	6.22×10^{-2}	9.70×10^{-3}	0.10	0.34

LD, linkage disequilibrium; T2D, type 2 diabetes; GO, gene ontology database. SNPs which showed association with T2D (Po0.003) were included in the analysis. The number of observed genes was compared to the number expected for each pathway, using Fisher's exact test. P-values from this test are shown. The top-10 ranking pathways per method are shown in gray, with the top-3 ranking pathways shown in bold. To check for consistency we analyzed the datasets with different inclusion and exclusion criteria. The results of analyses with different thresholds in P-value, without the HLA region, with the known T2D loci, without GO terms in the network analysis, and with no network analysis at all, are shown. The top-10 ranking pathways per method are shown in gray.

Discussion

The validity of the results presented in the example of T2D GWAS data depends greatly on the strengths and limitations of the methods used (box 1). The results in this study are all hypothetical until they are validated on real data. While the concept of pathway analysis is simple and attractive, it is restricted by our limited knowledge of cellular processes. The majority of genes in the genome are relatively unknown and their biological function still needs to be established. Because network and pathway tools make use of functional information from gene and protein databases, they are biased towards the well-studied genes, interactions and pathways.

Differences in pathway definitions

Differences in outcome between the tools by analyzing the same dataset could be due to different updates of gene lists and pathways, different human reference gene sets, and different statistical tests. For pathway tools not based on KEGG, differences could also be due to different definitions and classifications of pathways. For example, a consistently overrepresented pathway in the results from PANTHER in the DGI and the WTCCC data is that of inflammation mediated by chemokine and cytokine signaling. In PANTHER, 'inflammation' is a large pathway comprising 315 genes, whereas in other tools, this pathway is divided into smaller inflammation and inflammation-related pathways, like the 'Jak-STAT signaling pathway' and 'cytokine-cytokine receptor interaction'. This makes it difficult to compare results from multiple pathway tools and it shows that there needs to be more consensus on pathway classification and definition. The difficulty is that biological processes usually involve more than one pathway and pathways interact with each other. Therefore, well-defined pathways are hard to establish and making good computational predictions of cellular processes from genomic and molecular information will be a great challenge for the future.

Our results also clearly show that consensus of the tools depends on whether the tools keep their gene lists, pathways and human reference sets updated according to the latest versions of NCBI, KEGG, etc. We would emphasize that online pathway classification tools and other bioinformatics tools that are not regularly updated quickly become useless for classification purposes.

Box 1. Troubleshooting for analysing GWAS and microarray data using pathway classification tools

Problem	Description	Remark
Differences in outcome between pathway tools	***Use of different updates of gene assembly and gene builds*:** Some genes are not recognized by pathway tools and other genes do not exist or are located elsewhere in the most recent version of NCBI	Publicly available online pathway tools that are not regularly updated quickly become unusable for classification purposes
	***Use of different human reference gene sets*:** For example, Panther uses NCBI: Homo Sapiens, but other tools use privately composed reference sets or reference sets with only 4,000 genes	The source of the reference gene set should be clear and easy accessible
	***Different statistical tests*:** Fisher's exact test, hypergeometric test, binomal test, Bayes' factor	
	***Different definition and classification of pathways*:** For example, in PANTHER,''inflammation'' is a large pathway comprising 315 genes, whereas in other tools, this pathway is divided into smaller inflammation and inflammation-related pathways, like the ''Jak-STAT signaling'' and ''cytokine-cytokine receptor interaction'' pathways	There needs to be more consensus on pathway classification and definition. However, biological processes usually involve more than one pathway and pathways interact with each other. Thus, making good computational predictions of cellular processes from genomic and molecular information will be a great challenge for the future
There are always significantly overrepresented pathways	***Samples taken at random resulted in significantly overrepresented pathways*** using the pathway classification tools, whereas analyzing randomly selected gene sets should hypothetically not result in overrepresented pathways	

Some pathways have a higher chance of becoming statistically significant	***Some features make a gene more likely to be included in our analysis by chance*:** For example, large genes, genes in large LD blocks, and genes in pathways that contain more genes ***Large pathways were favored to become significantly overrepresented in our analysis.*** This could be due to the statistical attribute that the power of the test increases when the numbers compared become larger, as is the case in analyzing larger pathways ***The Affymetrix GeneChip Human Mapping 500k Array SetGeneChip covers the whole genome, but not all regions are equally well covered.*** This might have affected our analyses because some gene sets and pathways are better covered and therefore more likely to show up in the results	Although the statistical test we applied partly corrected for these biases, permutation and bootstrapping of the data would greatly improve the results of the analysis
The bioinformatics tools used in this study are biased toward detecting well-defined pathways	***Network and pathway tools make use of functional information from gene and protein databases.*** However, the majority of genes in the genome are relatively unknown and their biological function still needs to be established	

Large pathways in favor

We observed that especially in the pathway classification tools 'Panther' and Webgestalt KEGG, large pathways were favored to become significantly overrepresented in our analysis. This could be due to the statistical attribute that the power of the test increases when the numbers for comparison become larger, as is the case in analyzing larger pathways. In addition, because we created large LD blocks around each selected SNP, multiple genes per SNP will be selected and the chance of

selecting a false-positive gene is large. Genes from large pathways with many genes will then have a higher chance of being selected, although we partly corrected for this feature statistically.

Large genes, genes in large LD blocks, and genes in pathways that contain more genes, were all more likely to be included in our analysis by chance, although the statistical test used should partly correct for these features.

Gene distribution per pathway

The number of genes per pathway is not equally distributed. Figure 2 shows that the pathways in KEGG differ from 4 till 471 genes per pathway. This wide distribution of genes per pathway could again influence our results. Although pathways are not expected to have equal numbers of genes, if a smaller range is allowed, many of these problems will be eliminated. For example, Webgestalt Biocarta classifies the genes in pathways within a small range and this results in less bias towards larger pathways when analyzing the data (supplementary figure 1). Also, one can argue that a group of 4 genes can hardly be considered a pathway, but a group of 471 genes could in fact be subdivided into several smaller pathways.

Coverage Affymetrix GeneChip Human Mapping 500k Array

The DGI and WTCCC both performed a GWAS using the Affymetrix GeneChip Human Mapping 500k Array Set. Although this gene chip covers the whole genome, not all regions are equally well covered. This might have affected our analyses because some gene sets and pathways are better covered and are therefore more likely to show up in the results (figure 3 and 4). The distribution of the average SNPs/gene is not normal on the Affymetrix GeneChip. The range varies from 7.84 SNPs/gene for the ‘fatty acid biosynthesis’ pathway to 441.5 SNPs/gene for the ‘antigen processing and presentation’ pathway. In addition, the SNP density, measured by the average number of basepairs (bp)/SNPs on the Affymetrix GeneChip varies from 69.72 bp/SNP for the ‘antigen processing and presentation’ pathway to 5398.80 bp/SNP for the pathway ‘Heparan sulfate biosynthesis’. Permutation and bootstrapping might reduce these problems. Also, SNP imputation could solve the problem of bias in coverage of the Affymetrix GeneChip Human Mapping 500k Array, but wouldn’t solve the problem of bias of overall SNP coverage in the genome.

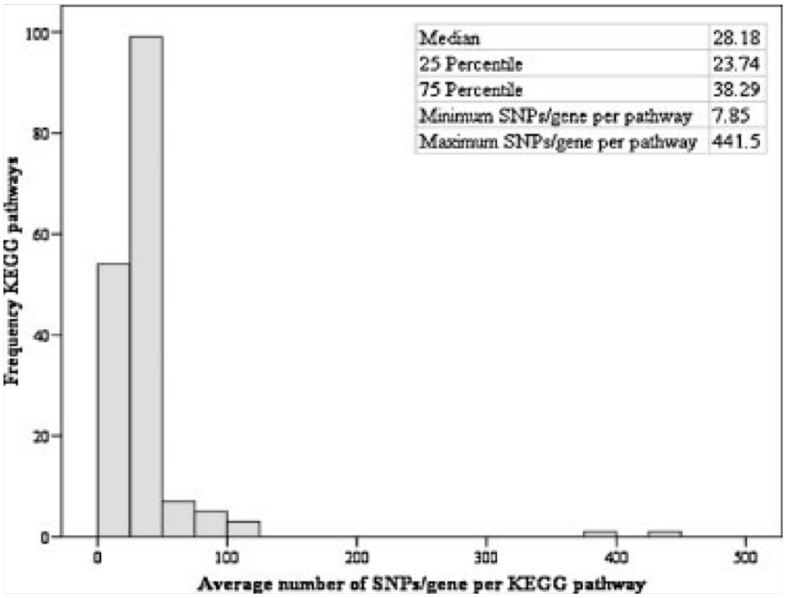


Fig. 2. Coverage KEGG pathways Affymetrix GeneChip Human Mapping 500k Array. Although this gene chip covers the whole genome, not all regions are equally well covered. The distribu- tion of the average SNPs/gene is not normally distributed on the Affymetrix GeneChip. The range per pathway varies from 7.84 to 441.5 SNPs per gene.

P-value thresholds

The analysis with different p-value thresholds showed that with p-values below 0.003 the number of genes that was selected as input for the pathway analysis was too small to be able to detect significantly overrepresented pathways. This would suggest that a minimum amount of genes is needed for these analyses. The analysis with p-value thresholds above 0.003 resulted in complications, because the bioinformatics tools only could handle a certain amount of genes as an input. This is a limitation of these methods, because we couldn't study the effect of using high p-value thresholds on our results. However, although we cannot predict what the highest appropriate p-value threshold is, the cut-off should not be too low, because it is not convenient to include the whole genome for pathway analysis.

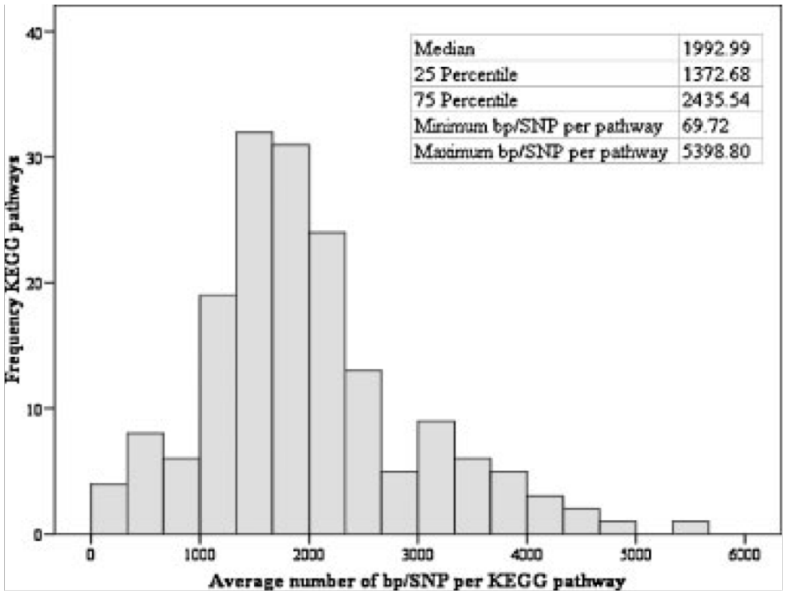


Fig. 3. Coverage of base pairs (bp) per SNP per KEGG pathway Affymetrix GeneChip Human Mapping 500k. The range per pathway varies from 69.72 to 5,398.80 bp per SNP.

Differences between GWAS T2D datasets

The results from the DGI and the WTCCC GWAS show concordance as well as discordance. This could be due to differences in study design, genetic heterogeneity between the study populations, differential biases and errors across studies, and random effects [12]. Both DGI and WTCCC were designed to identify T2D susceptibility loci but the two study designs differed considerably. In the DGI study, the cases and controls were matched for gender, age, BMI and region of origin. This makes the DGI study capable of detecting T2D susceptibility genes and pathways that confer risk independently of obesity. Cases and controls from the WTCCC were not matched and anthropometric measurements, such as BMI, in the controls groups were unknown. Identified genes and pathways from the WTCCC can be both T2D genes as well as BMI genes. Therefore, the fact that not all pathways overlap between the DGI and WTCCC might just reflect a difference in pathways that are picked-up because of this difference in study design.

Replication of GWAS results, genes versus pathways

The data from the DGI and WTCCC overlapped in overrepresented pathways, but frequently we picked up different genes in the same pathway. As an example, figure 5 shows the inflammation mediated by chemokine and cytokine signaling and the genes involved in this pathway. This figures show that although this similar pathway seems important in both studies, the genes from this pathway differ per study. These results may therefore be difficult to replicate when investigating single loci, but incorporating gene networks and pathway classification tools into your analysis can point towards significantly overrepresented molecular pathways.

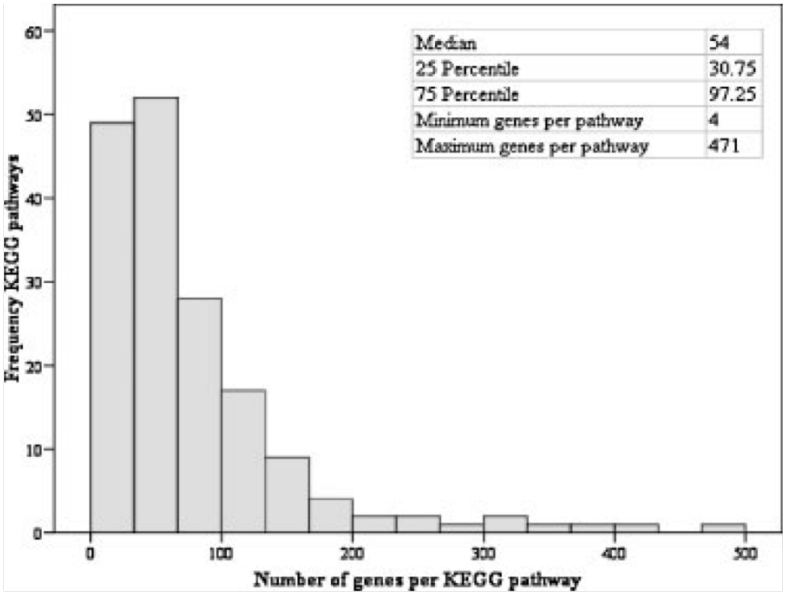


Fig. 4. Distribution of genes per KEGG pathway, showing that the pathways in KEGG differ from 4 genes till 471 genes per pathway. This wide distribution is far from normal and could have influenced our results.

New T2D genes

Recently, after this study was performed, a meta-analysis was published on three T2D GWAS, including the DGI and the WTCCC. Eleven new susceptibility loci for T2D were found, containing the genes *JAZF1, CAMK1D-CDC123, TSPAN8-LGR5, THADA, ADAMTS9, NOTCH2, DCD, SYNC2-PPARG, VEGFA, BCL11A* and

ADAM30 [29]. Furthermore, 2 Japanese studies recently found the T2D susceptibility gene *KCNQ1* [25,28].

In our initial gene set that is based on the top most associated SNPs, *JAZF1, CAMK1D-CDC123, TSPAN* and *PPARG* were present in the WTCCC data, and *NOTCH2* and *ADAM30* were present in the DGI dataset. Subsequently, *JAZF1*, *CAMK1D* and *PPARG* were predicted by Prioritizer as candidate genes for T2D in the WTCCC data and *NOTCH2* and *ADAM30* in the DGI dataset. This shows that from the five recently unknown T2D susceptibility genes (*PPARG* was already known to be involved in T2D) that were selected as input for our network analysis, Prioritizer was able to predict four.

The identified T2D genes on the eleven susceptibly loci also illustrate the limitations of pathway classification tools. The biological functions of most of these genes are unknown and only *NOTCH2* and *VEGFA* could be sorted in pathways at all (i.e. notch signaling for *NOTCH2*). The pathway in which *VEGFA* is sorted fits the predicted pathways for T2D of cytokine-cytokine receptor interaction and focal adhesion.

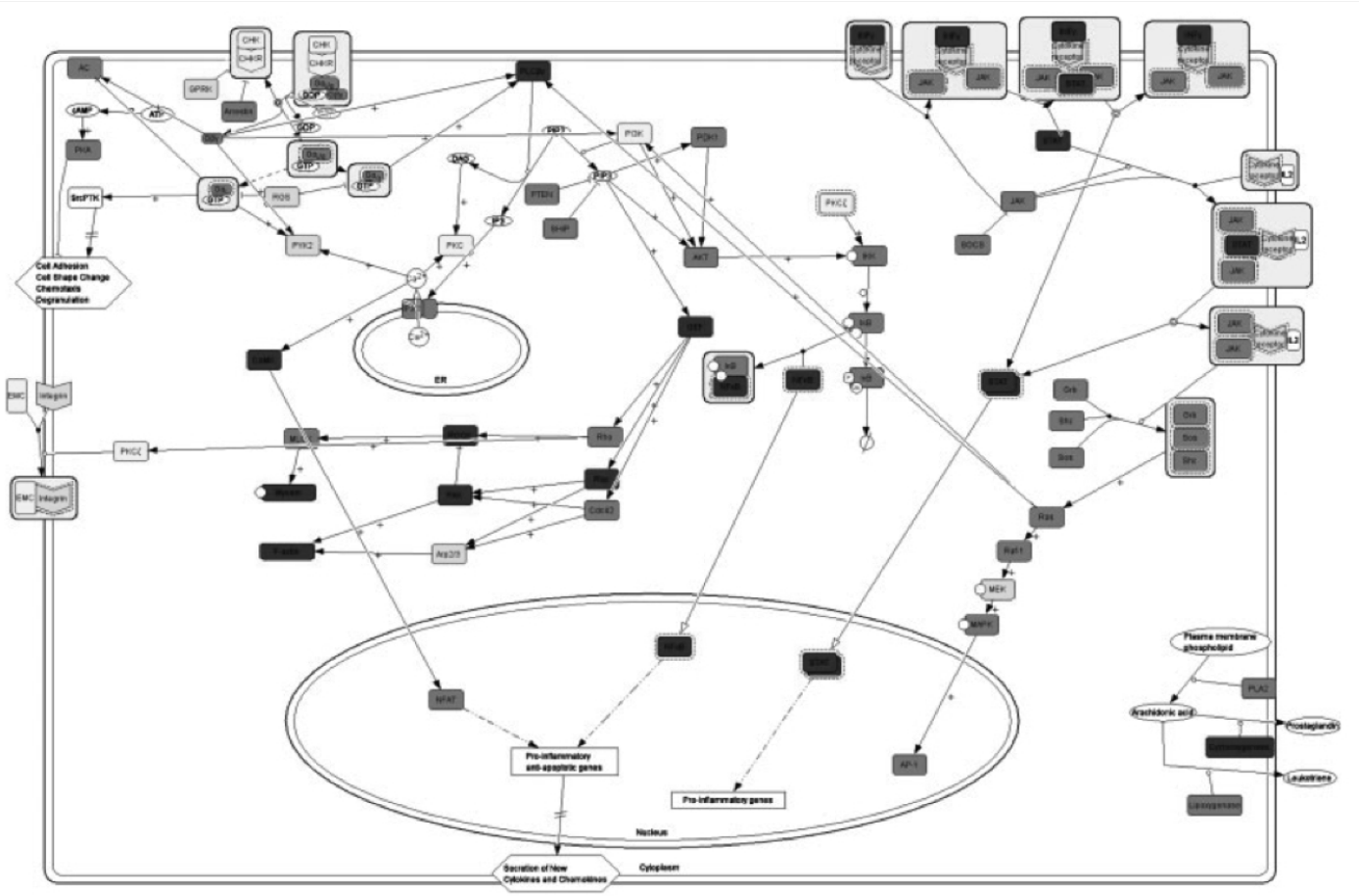


Fig. 5. Inflammation mediated by the chemokine and cytokine signaling pathway. Genes in grey are in LD with SNPs that showed association with T2D (Po0.003) in the DGI GWAS. Genes in dark grey are in linkage disequilibrium (LD) with SNPs that showed association with type 2 diabetes (T2D) (Po0.003) in the WTCCC GWAS. Genes in light grey are in LD with SNPs that showed association with T2D (Po0.003) in both datasets. The results from the DGI and the WTCCC studies showed concordance in overrepresented pathways, but discordance in the corresponding genes. This shows that the small effect sizes due to the genes make it difficult to replicate disease association, but incorporating interaction analysis into the model can reveal significantly overrepresented molecular pathways.

Biology of T2D

Because of the above mentions problems, our study does not provide major mechanistic insight in the etiology of T2D. However, our approach does show some interesting observations in the T2D GWAS data using prior biological information.

It has been suggested that the genes found in the T2D GWAS imply a role for β cell development and function in T2D etiology [8,17]. However, the results of our pathway analysis show that by taking a different view of the top SNPs it is possible to detect other biological mechanisms as well. The pathways that were most strongly overrepresented and showed the highest consistency throughout our results were all pathways involved in inflammation. It is known that low-grade inflammation in visceral fat of obese individuals causes insulin resistance and subsequently T2D [10]. However, our analysis does not rule out the involvement of β cell development and function in the underlying biology of T2D. Oxidative stress as well as various inflammatory cytokines have also been proposed to play an important role in mediating β cell destruction [13]. Unfortunately, none of the pathway classification tools contained a pathway of β cell development and function, probably because it is relatively unknown and still needs to be described in detail in the literature and in pathway databases [1].

Conclusion

Incorporating gene networks and pathway classification tools into your analysis can point towards significantly overrepresented molecular pathways, which cannot be picked up using traditional single-locus analyses. However, while the concept of pathway analysis is simple and attractive, it is restricted by our limited knowledge of cellular processes. Pathway tools still suffer from several limitations and the next challenge is how to make better computational predictions of cellular processes from genomic and molecular information.

Acknowledgments

The study was financially supported by SenterNovem (IOP genomics grant IGE05012). We thank Jackie Senior for critically reading the manuscript.

The authors have declared that no competing interests exist.

References

1. Bell GI, Polonsky KS. 2001. Diabetes mellitus and genetically programmed defects in beta-cell function. Nature 414(6865):788-91.
2. Campbell CD, Lyon HN, Nemesh J, Drake JA, Tuomi T, Gaudet D, Zhu X, Cooper RS, Ardlie KG, Groop LC and others. 2007. Association studies of BMI and type 2 diabetes in the neuropeptide Y pathway: a possible role for NPY2R as a candidate gene for type 2 diabetes in men. Diabetes 56(5):1460-7.
3. Carlborg O, Haley CS. 2004. Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5(8):618-25.
4. Chang JT, Nevins JR. 2006. GATHER: a systems approach to interpreting genomic signatures. Bioinformatics 22(23):2926-33.
5. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M and others. 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 38(10):1166-72.
6. Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, Wijmenga C. 2007. A strategy to search for common obesity and type 2 diabetes genes. Trends Endocrinol Metab 18(1):19-26.
7. Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. 2006. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78(6):1011-25.
8. Frayling TM. 2007. Genome-wide association studies provide new insights into type 2 diabetes etiology. Nat Rev Genet 8(9):657-62.
9. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM and others. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-61.
10. Hotamisligil GS. 2006. Inflammation and metabolic disorders. Nature 444(7121):860-7.
11. Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8(9):R183.
12. Ioannidis JP, Patsopoulos NA, Evangelou E. 2007. Heterogeneity in meta-analyses of genome-wide association investigations. PLoS ONE 2(9):e841.
13. Kaneto H, Matsuoka TA, Katakami N, Kawamori D, Miyatsuka T, Yoshiuchi K, Yasuda T, Sakamoto K, Yamasaki Y, Matsuhisa M. 2007. Oxidative Stress and the JNK Pathway are Involved in the Development of Type 1 and Type 2 Diabetes. Curr Mol Med 7(7):674-86.
14. Kelada SN, Eaton DL, Wang SS, Rothman NR, Khoury MJ. 2003. The role of genetic polymorphisms in environmental health. Environ Health Perspect 111(8):1055-64.
15. Lehner B, Fraser AG. 2004. A first-draft human protein-interaction map. Genome Biol 5(9):R63.
16. Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37(4):413-7.
17. McCarthy MI, Zeggini E. 2007. Genome-wide association scans for Type 2 diabetes: new insights into biology and therapy. Trends Pharmacol Sci.
18. Mi H, Guo N, Kejariwal A, Thomas PD. 2007. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res 35(Database issue):D247-52.
19. Robins J, Greenland S, Breslow NE. 1986. A general estimator for the variance of the Mantel-Haenszel odds ratio. Am J Epidemiol 124(5):719-23.
20. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ and others. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316(5829):1331-6.
21. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU and others. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316(5829):1341-5.
22. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S and others. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445(7130):881-5.
23. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S and others. 2007. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39(6):770-5.

24. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S and others. 2005. A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6):957-68.
25. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jorgensen T and others. 2008. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet.
26. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MC, Bethel G, Holmes GK and others. 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 39(7):827-9.
27. Werner T. 2008. Bioinformatics applications for pathway analysis of microarray data. Curr Opin Biotechnol.
28. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y and others. 2008. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet.
29. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G and others. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40(5):638-45.
30. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM and others. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316(5829):1336-41.
31. Zhang B, Kirov S, Snoddy J. 2005. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33(Web Server issue):W741-8.

Supplementary information

The combined dataset of the DGI and WTCCC

Methods

After filtering the data according to our inclusion criteria, 353,584 overlapping SNPs remained in the GWAS datasets. Both DGI and WTCCC were designed to identify T2D susceptibility loci, but the two studies differed considerably in study design. Combining two different datasets increases the population size and therefore the power to detect associations, but does not take study design differences into account. We therefore decided to investigate pathways in the DGI and WTCCC datasets separately as well as in a combined dataset. We performed a Mantel-Haenszel procedure to obtain pooled estimates for the overlapping SNPs and found 1,340 SNPs that showed association with a p-value lower than 0.003.

To perform pathway analyses on the GWAS data, we first allocated the genotyped SNPs to genomic regions, based on the LD structure around the SNPs. From the pooled estimates of the results, 483 SNPs from the 1,340 SNPs were mappable. We included these SNPs in the analysis for this study.

Results

Several pathways were pinpointed as the strongest overrepresented pathways in the DGI and the WTCCC combined dataset by two out of the three KEGG pathway classification tools (table 1S): 'Wnt signaling pathway', 'focal adhesion', 'T and B cell receptor signaling pathway', 'regulation of actin cytoskeleton', 'MAPK signaling pathway', 'epithelial cell signaling in Helicobacter pylori infection', 'apoptosis' and 'toll-like receptor signaling pathway'. The other pathway classification tools both pointed to 'toll-like receptor signaling', and 'Wnt signaling' as an overrepresented BioCarta pathway (table 2S).

Table 1S. Overrepresented KEGG pathways in the top results of the pooled data of the GWAS from DGI and WTCCC on type 2 diabetes.

KEGG pathways from Webgestalt	p-value	KEGG pathways from GATHER	p-value	KEGG pathways from DAVID	p-value
Wnt signaling pathway	2.06X 10-6	Wnt signaling pathway	1.96×10^{-2}	T cell receptor signaling pathway	2.74×10^{-2}
Focal adhesion	1.51×10^{-5}	Riboflavin metabolism	4.64×10^{-2}	B cell receptor signaling pathway	0.12
T cell receptor signaling pathway	1.58×10^{-5}	Apoptosis	7.14×10^{-2}	Epithelial cell signaling in Helicobacter pylori infection	0.20
Regulation of actin cytoskeleton	1.45×10^{-4}	Cell cycle	7.65×10^{-2}	MAPK signaling pathway	0.22
MAPK signaling pathway	1.49×10^{-4}	Toll-like receptor signaling pathway	7.65×10^{-2}	Apoptosis	0.49
Purine metabolism	2.05×10^{-4}	Focal adhesion	9.16×10^{-2}	Toll-like receptor signaling pathway	0.51
B cell receptor signaling pathway	4.64×10^{-4}	pantothenate and CoA biosynthesis	9.16×10^{-2}		
Epithelial cell signaling in Helicobacter pylori infection	6.15×10^{-4}	Alzheimer's disease	9.93×10^{-2}		
Adipocytokine signaling pathway	7.20×10^{-4}	Insulin signaling pathway	0.12		
Pancreatic cancer	7.98×10^{-4}	regulation of actin cytoskeleton	0.12		

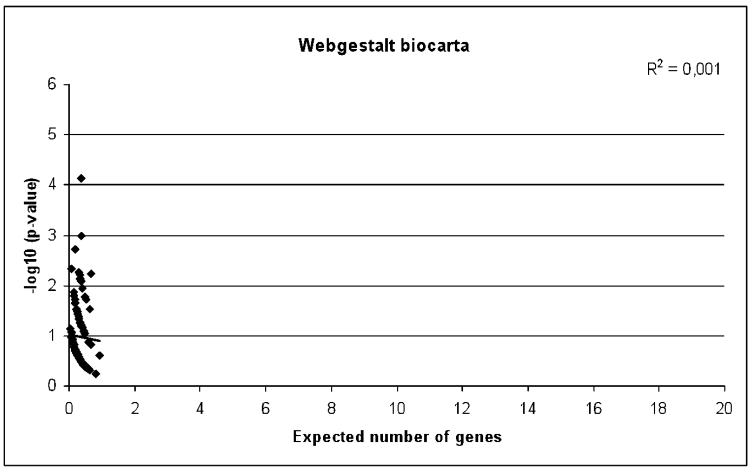
SNPs which showed association with T2D ($p < 0.003$) were included in this study and were mapped backed to regions on the genome, and the predicted candidate genes were used for analysis. The number of observed genes was compared to the number of expected genes for each KEGG pathway. The top-10 ranking KEGG pathways per pathway classification tool are shown.

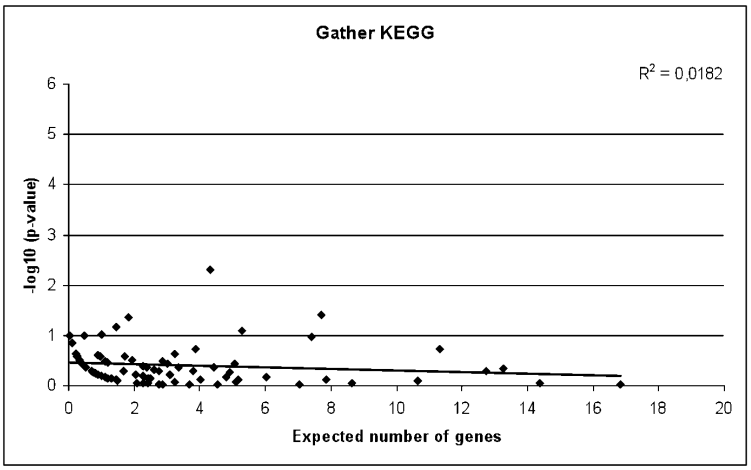
Table 2S. Overrepresented PANTHER and Biocarta pathways in the top results of the combined GWAS data of the DGI and the WTCCC on type 2 diabetes.

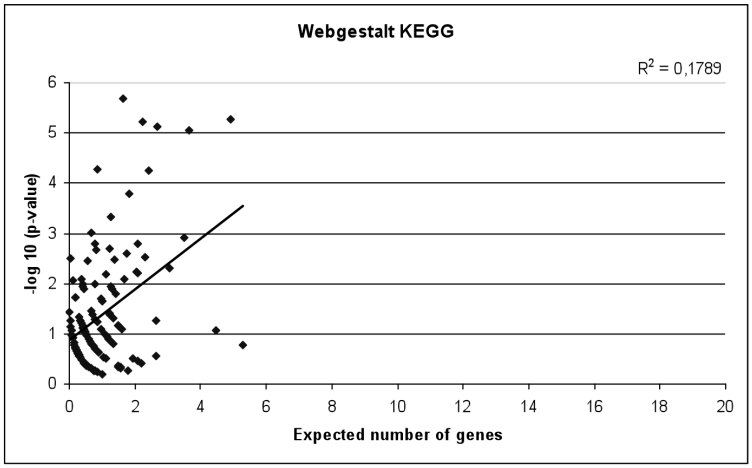
BioCarta pathways Webgestalt	p-value	Panther pathways	p-value
AKT signaling pathway	2.34×10^{-4}	T cell activation	6.32×10^{-7}
NFkB activation by nontypeable hemophilus influenzae	3.89×10^{-4}	B cell activation	1.10×10^{-4}
WNT signaling pathway	4.36×10^{-4}	Inflammation mediated by chemokine and cytokine signaling pathway	4.33×10^{-4}
Basic mechanism of action of PPARa. PPARb(d) and PPARg and effects on gene expression	4.46×10^{-4}	Angiogenesis	3.81×10^{-3}
Toll-like receptor pathway	1.20×10^{-3}	Axon guidance mediated by netrin	5.09×10^{-3}
Double stranded rna induced gene expression	1.61×10^{-3}	Cytoskeletal regulation by Rho GTPase	1.57×10^{-2}
?-arrestins in GPCR desensitization	2.26×10^{-3}	Huntington disease	1.67×10^{-2}
Acetylation and deacetylation of RelA in the nucleus	3.02×10^{-3}	Toll receptor signaling pathway	2.35×10^{-2}
CD40L signaling pathway	3.44×10^{-3}	Ascorbate degradation	3.36×10^{-2}
TACI and BCMA stimulation of B cell immune responses.	3.44×10^{-3}	Axon guidance mediated by Slit/Robo	3.65×10^{-2}

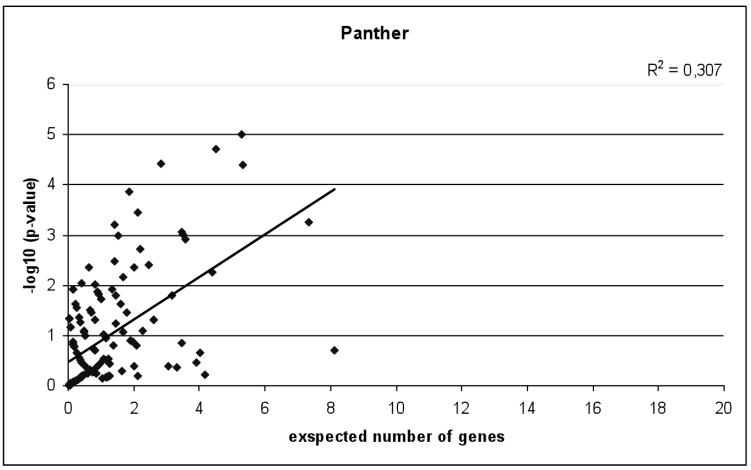
SNPs which showed association with T2D ($p < 0.003$) were included in this study and were mapped backed to regions on the genome, and the predicted candidate genes were used for analysis. The number of observed genes was compared to the number of expected genes for each pathway. The top-10 ranking KEGG pathways per method are shown for Webgestalt 'Biocarta' and PANTHER.

Supplementary figure 1. Correlation between the expected number of genes per pathways and corresponding p-value of overrepresentation per pathway.









This figure shows the correlation of a sample taken at random. We compared our observed number of genes per analysis to the number of genes expected by chance. A high number of expected genes represent large pathways, a low number of expected genes represent small pathways. We observed that some pathway classification tools favor large pathways leading to significant overrepresentation in our analysis.

Chapter 4.2

Interrogating Type 2 Diabetes Genome-Wide Association Data Using a Biological Pathway-Based Approach

Clara C. Elbers, Yvonne T. van der Schouw, Cisca Wijmenga,
N. Charlotte Onland-Moret

Diabetes. 2009 Sep;58(9):e9

Interrogating Type 2 Diabetes Genome-Wide Association Data Using a Biological Pathway-Based Approach

Perry *et al.* [1] performed a pathway-based approach aiming to identify biological pathways associated with type 2 diabetes. They used genome-wide association (GWA) data from the type 2 diabetes study in the U.K. Wellcome Trust Case Control Consortium (WTCCC) for the initial analysis and validated the findings with data from the Diabetes Genetics Initiative (DGI) and Finland–United States Investigation of NIDDM Genetics (FUSION) studies. The Wnt signaling pathway was the most strongly associated, and they therefore postulated this was the most interesting candidate pathway. However, after correcting for multiple testing, none of the top-ranking pathways reached statistical significance. Perry *et al.* concluded that type 2 diabetes genes are likely to reside in multiple pathways.

We recently performed comparable genome-wide pathway analysis in two of the three GWA datasets used by Perry *et al.* (the WTCCC and DGI) and found overlapping but also different results to theirs [2]. However, we encountered several problems using these pathway methods. Our main conclusion is therefore that pathway-based approaches have many limitations that need to be addressed before these methods can be used to provide accurate results and conclusions can be drawn.

First, in classification systems like Kyoto Encyclopedia of Genes and Genomes (KEGG) or BioCarta, the majority of human genes are currently not sorted on any pathway. Of the 18 type 2 diabetes susceptibility loci recently identified, only 5 (*CDKN2A-2B, PPARG, NOTCH2, VEGFA*, and *TCF7L2*) could be assigned to known biological pathways. In addition, β-cell function, one of the mechanisms suggested to underlie type 2 diabetes, has not been specifically described as a pathway in either KEGG or BioCarta. Thus, although type 2 diabetes genes may well play a role in multiple pathways, we feel that this conclusion cannot be drawn based on the results from pathway-based analyses.

Second, as Perry *et al.* discuss, larger pathways are favored to become significantly overrepresented in pathway analysis. This is due to the statistical attribute that the power of tests increases as the numbers for comparison become larger, which is the case in analyzing lager pathways. One of the top associated pathways in both our study and that of Perry et al. is the Wnt signaling pathway, which comprises many genes. It is therefore highly likely to become statistically

overrepresented in pathway analyses. We analyzed 30 randomly selected sets of genes, encompassing around 1,500 genes per set, and in 16 of the 30 sets the Wnt signaling pathway was in the list of the top 10 ranked pathways, and in 5 of the 30 sets it was even ranked in the top 3.

We would like to emphasize that the limitations of pathway-based analyses in GWA data should be kept in mind when drawing conclusions based on overrepresented pathways.

References

1. Perry JR, McCarthy MI, Hattersley AT, Zeggini E, the Wellcome Trust Case Control Consortium, Weedon MN, Frayling TM: Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes 2009; 58: 1463– 1467
2. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC: Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 2009; 33: 419– 431

Chapter 5

Variants in Neuropeptide Y Receptor 1 and 5 Are Associated with Nutrient-Specific Food Intake and Are Under Recent Selection in Europeans

Clara C. Elbers, Carolien G. F. de Kovel, Yvonne T. van der Schouw, Juliaan R. Meijboom, Florianne Bauer, Diederick E. Grobbee, Gosia Trynka, Jana V. van Vliet-Ostaptchouk, Cisca Wijmenga, N. Charlotte Onland-Moret

PLoS ONE. 2009 Sep 17;4(9):e7070

Abstract

There is a large variation in caloric intake and macronutrient preference between individuals and between ethnic groups, and these food intake patterns show a strong heritability. The transition to new food sources during the agriculture revolution around 11,000 years ago probably created selective pressure and shaped the genome of modern humans. One major player in energy homeostasis is the appetite-stimulating hormone neuropeptide Y, in which the stimulatory capacity may be mediated by the neuropeptide Y receptors 1, 2 and 5 (NPY1R, NPY2R and NPY5R).

We assess association between variants in the *NPY1R, NPY2R* and *NPY5R* genes and nutrient intake in a cross-sectional, single-center study of 400 men aged 40 to 80 years, and we examine whether genomic regions containing these genes show signatures of recent selection in 270 HapMap individuals (90 Africans, 90 Asians, and 90 Caucasians) and in 846 Dutch bloodbank controls.

Our results show that derived alleles in NPY1R and NPY5R are associated with lower carbohydrate intake, mainly because of a lower consumption of mono- and disaccharides. We also show that carriers of these derived alleles, on average, consume meals with a lower glycemic index and glycemic load and have higher alcohol consumption. One of these variants shows the hallmark of recent selection in Europe.

Our data suggest that lower carbohydrate intake, consuming meals with a low glycemic index and glycemic load, and/or higher alcohol consumption, gave a survival advantage in Europeans since the agricultural revolution. This advantage could lie in overall health benefits, because lower carbohydrate intake, consuming meals with a low GI and GL, and/or higher alcohol consumption, are known to be associated with a lower risk of chronic diseases.

Introduction

One major player in energy homeostasis is the appetite-stimulating hormone neuropeptide Y (NPY)[1]. In rodents, NPY evokes eating behavior, inducing particularly carbohydrate intake. Injection of NPY in the brain elicits a strong feeding response even in satiated animals, eventually leading to obesity [2]. The effect of NPY is mediated by the neuropeptide Y receptors (NPYRs) [3]. Especially the Y1, Y2, and Y5 receptors (NPY1R, NPY2R, NPY5R) appear to be candidates for mediating the appetite stimulatory capacity of NPY [4,5] through binding of NPY. These are receptors in the arcuate and paraventricular nuclei of the hypothalamus. Variants in genes coding for these receptors may therefore influence energy intake, which could influence an individual's susceptibility to becoming obese and developing T2D. We have previously pinpointed *NPY1R, NPY2R* and *NPY5R* as positional candidate genes for both obesity and T2D [6].

Large variations in caloric intake and macronutrient preference between individuals have been reported and these food intake patterns show a strong heritability [7]. There are also large differences in food intake and percentage of nutrient-specific energy intake among different ethnic groups [8,9]. These ethnic differences in total and nutrient-specific energy intake might be caused by the natural selection of mutations providing an advantage for a particular environment or type of agriculture. The transition to different food sources during the agricultural revolution, which started around 11,000 years ago, was an important selective pressure and the changes in food intake helped shape the genome of modern humans [10]. Genome-wide sequence and SNP data of living humans can be used to study the recent natural selection over the past 30,000 years [11,12]. Under neutral selection, the linkage disequilibrium (LD) around variants in the genome will decay over time due to recombination, so that older (common) alleles typically have short-range LD and younger (rare) alleles have long-range LD. However, when an allele is under positive selection, its frequency rises rapidly in the population over a short time span and the haplotype carrying the advantageous allele therefore breaks down more slowly than an allele with the same frequency under neutral selection.

In this study we investigated the role of single nucleotide polymorphisms (SNPs) in *NPY1R, NPY2R* and *NPY5R* genes in the total and nutrient-specific energy intake in a Dutch study population of 400 healthy older men. To see whether

changing environments in the past may have caused adaptation or maladaption to our current life style, we examined whether these loci showed a signature of recent selection, using genome-wide SNP data from the HapMap populations and a genome-wide SNP dataset of 846 Dutch bloodbank controls.

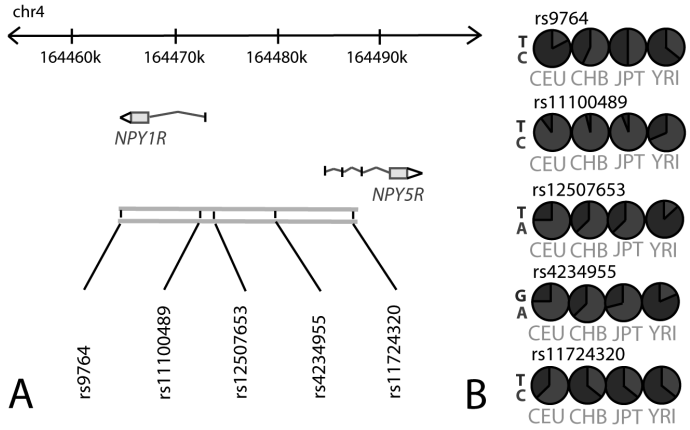


Figure 1. Characteristics of the NPY1R/NPY5R region.
Figure 1A. tSNPs that optimally cover the genetic variation in the haplotype containing the *NPY1R* and *NPY5R* genes so that all SNPs with a minor allele frequency of ≥0.10 were captured with r2≥0.8. Figure 1B. The global allele frequency distributions per SNP are shown. *NPY1R* neuropeptide Y receptor 1; *NPY5R* neuropeptide Y receptor 5; CEU Utah residents with Northern and Western European ancestry from the CEPH collection; CHB Han Chinese in Beijing, China; JPT Japanese in Tokyo, Japan; YRI Yoruba in Ibadan, Nigeria.

Results

NPY1R, NPY2R and NPY5R Variation and Macronutrient Intake in Healthy Older Men

Five tSNPs in the *NPY2R* gene and another five in the *NPY1R* and *NPY5R* genes were genotyped in the Hamlet population. The genotype success rates for all ten tSNPs were above 95%. There were no discordances in the genotypes of any of the CEPH sample and all genotypes were in agreement with Hardy-Weinberg equilibrium ($p > 0.01$). Age, body mass index (BMI) and macronutrient intake of the participants are shown in Table 1. As neither age nor BMI were associated with any of the *NPY2R* and *NPY1R/NPY5R* SNPs, they do not confound the relation in this study. Therefore they were not included as covariates in the model.

Table 1. Characteristics of the Hamlet study population.

Characteristic	N	Mean (SD)
Age (years)	382	60.40 (11.22)
BMI (kg/m2)	382	26.27 (3.44)
Energy intake (kcal)	380	2255.30 (517.48)
Protein intake (% of total energy intake)	380	15.02 (2.01)
Fat intake (% of total energy intake)	380	35.59 (5.09)
Carbohydrate intake (% of total energy intake)	380	42.98 (6.64)
Alcohol intake (% of total energy intake)	380	6.41 (6.46)

We did not find an association between any of the SNPs and total energy intake. However, by studying macronutrient-specific energy intake, we observed associations between SNPs in the *NPY1R/NPY5R* genes and carbohydrate intake, and with alcohol intake.

For rs17724320 in the *NPY1R/NPY5R* genes, we found a dose-response relationship of the derived T allele with carbohydrate intake ($p < 0.01$ for trend), meaning that carbohydrate intake was lowest in men carrying two ancestral C alleles and that it increased with each extra derived allele (Figure 2a). The haplotype analysis showed that carriers of the TTTGT haplotype consumed, on average, 6.2% more total carbohydrates than carriers of the reference haplotype TCAAC ($p = 0.003$) (Figure 2b).

There are many types of carbohydrates and the physiological responses to these vary substantially. We therefore also studied the association between SNPs in the *NPY1R/NPY5R* genes and relative mono- and disaccharide intake, relative polysaccharide intake, and GI and GL. The association appeared to be mainly restricted to mono- and disaccharides. For rs11100489, rs12507653, rs4234955 and rs17724320 in the *NPY1R/NPY5R* genes, we found the same dose-response relationship with mono- and disaccharide intake as for total carbohydrates for the derived allele (all showed a $p < 0.05$ trend) (Figure 3a). Men carrying two derived C alleles of rs11100489 ate 1.9% less mono- and disaccharides compared to men carrying one or two ancestral T alleles ($p = 0.02$). For rs12507653, men carrying one or two derived A alleles consumed 2.3% and 3.0% less mono- and disaccharides, respectively, than men homozygote for the ancestral T allele ($p = 0.04$ and $p = 0.008$, respectively). For rs11100489, the same genotype that was associated with a decrease in mono- and disaccharide intake was also associated with an increase in polysaccharide intake of 1.1% ($p = 0.05$) (Figure 4a).

The haplotype analysis showed that carriers of the TTTGT haplotype consumed 6.2% more mono- and disaccharides than carriers of the reference TCAAC haplotype ($p = 0.002$) (Figure 3b). We found no difference in polysaccharide intake between the different haplotypes (Figure 4b).

There were no associations between single SNP genotypes and GI and GL (Figures 5a and 6a). However, the GI of the daily food intake of individuals carrying the TCTGT haplotype was significantly higher than of individuals carrying the reference haplotype TCAAC (0.533 versus 0.509; $p = 0.01$) (Figure 5b). The daily food intake of individuals who carry the ancestral TCTGT, TCAGC or TTTGT haplotypes had a significantly higher GL than individuals carrying the reference haplotype TCAAC (147, 160 and 130 versus 118, respectively; $p = 0.001$, $p = 0.03$ and $p < 0.0001$, respectively)) (Figure 6b).

For alcohol intake there was an association with rs11724320 in the *NPY1R/NPY5R* genes (Figure 7a). Men homozygote for the derived allele consumed 2.4% more alcohol than men homozygote for the ancestral allele ($p = 0.04$). Carriers of the TTTGT and CCAAC haplotypes showed a difference of 4.0% and 2.0% in relative consumption of alcohol compared to the reference haplotype TCAAC (both p-values: 0.03) (Figure 7b).

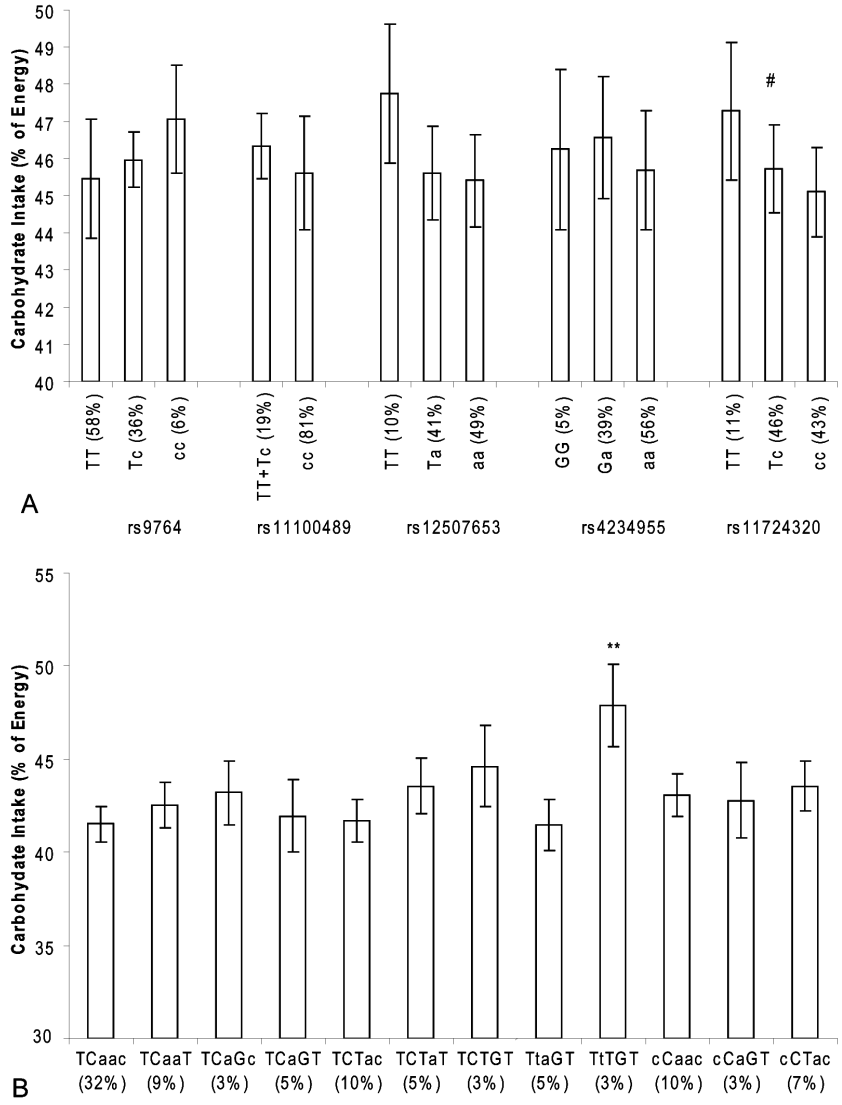


Figure 2. NPY1R/NPY5R variants and carbohydrate intake in the Hamlet population. Figure 2a shows the association of SNPs in the NPY1R/NPY5R region with carbohydrate intake as percentage of total energy intake. The ancestral alleles are indicated as capital letters. # $p<0.01$ for trend. Figure 2b shows the association of NPY1R/NPY5R haplotypes with carbohydrate intake as percentage of total energy intake. The haploblocks consist of the SNPs rs9764, rs11100489, rs12507653, rs4234955 and rs11724320 and the ancestral alleles are indicated as capital letters. **$p<0.01$ (compared with linear regression model).

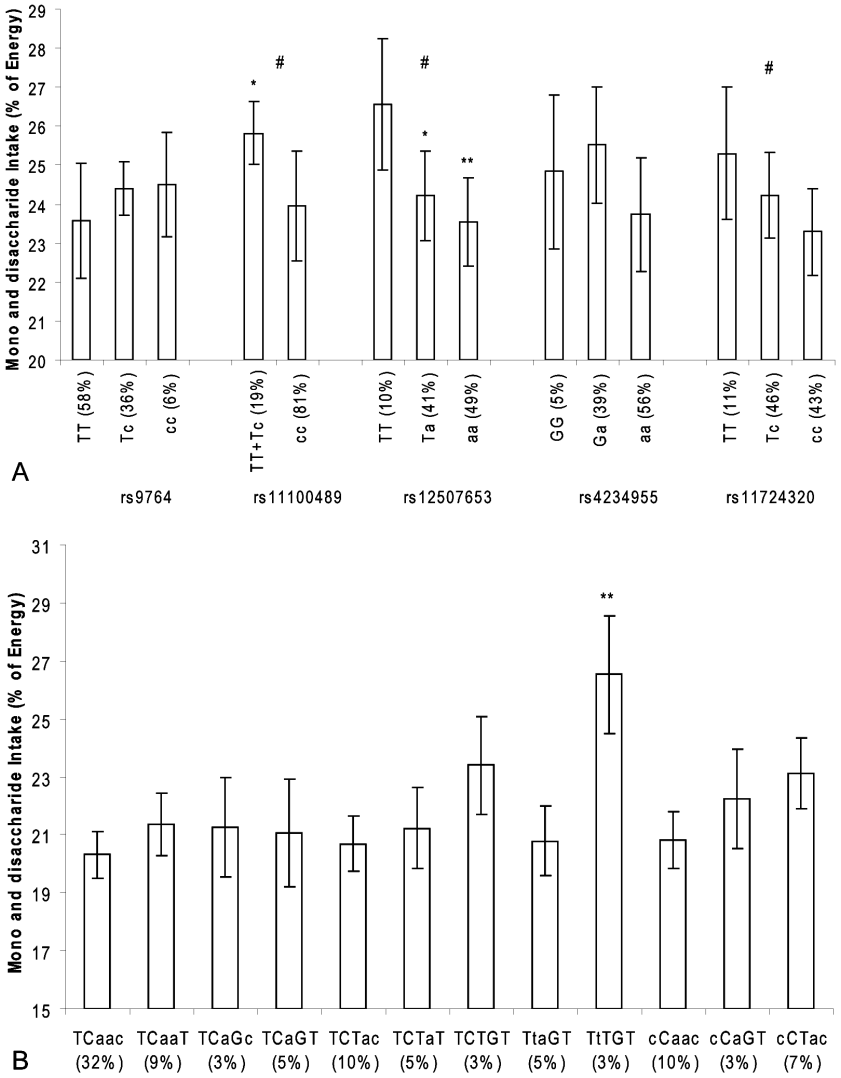


Figure 3. NPY1R/NPY5R variants and mono- and disaccharide intake in the Hamlet population.

Figure 3a shows the association of SNPs in the NPY1R/NPY5R region with mono- and disaccharide intake as percentage of total energy intake. The ancestral alleles are indicated as capital letters.*p<0.05, **p<0.01(compared with linear regression model), # P<0.01 for trend. Figure 3b shows the association of NPY1R/NPY5R haplotypes with mono- and disaccharide intake as percentage of total energy intake. The haploblocks consist of the SNPs rs9764, rs11100489, rs12507653, rs4234955 and rs11724320 and the ancestral alleles are indicated as capital letters. *p<0.05 (compared with linear regression model).

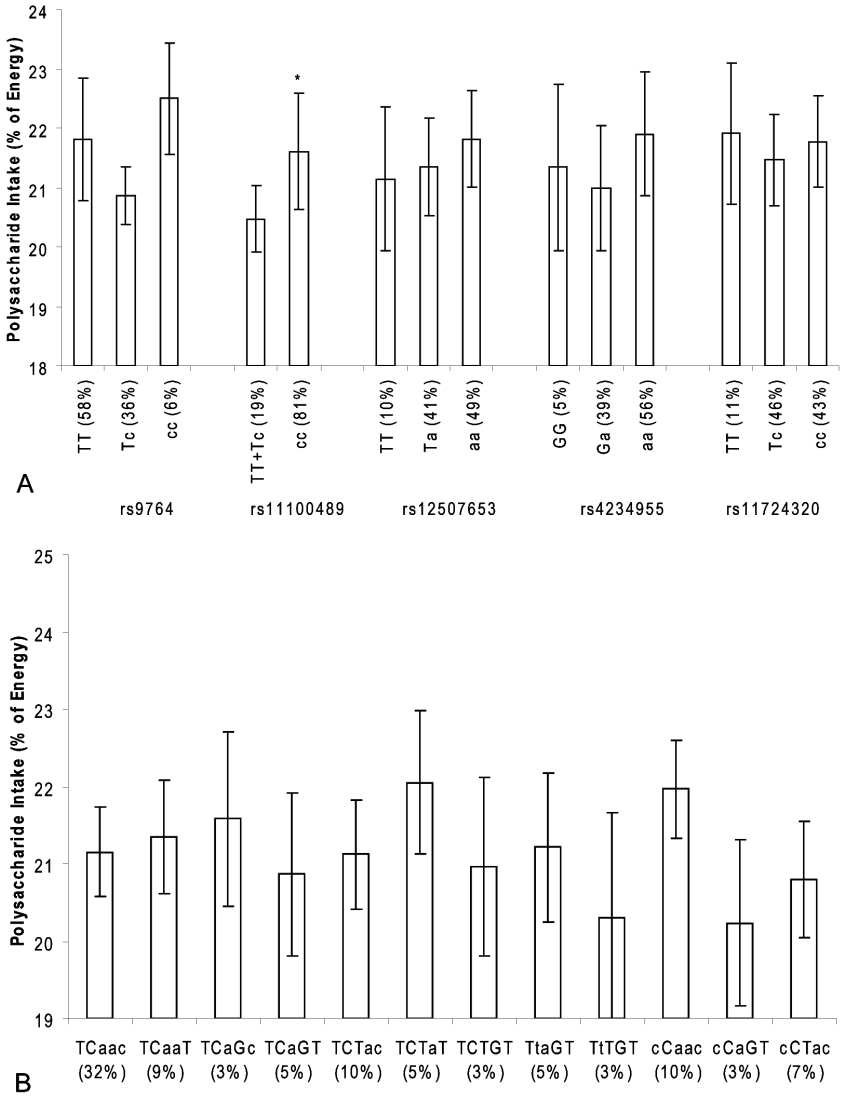


Figure 4. NPY1R/NPY5R variants and polysaccharide intake in the Hamlet population.
Figure 4a shows the association of SNPs in the NPY1R/NPY5R region with polysaccharide intake as percentage of total energy intake. The ancestral alleles are indicated as capital letters.*p<0.05(compared with linear regression model) Figure 4b shows the association of NPY1R/NPY5R haplotypes with polysaccharide intake as percentage of total energy intake. The haploblocks consist of the SNPs rs9764, rs11100489, rs12507653, rs4234955 and rs11724320 and the ancestral alleles are indicated as capital letters. *p<0.05 (compared with linear regression model).

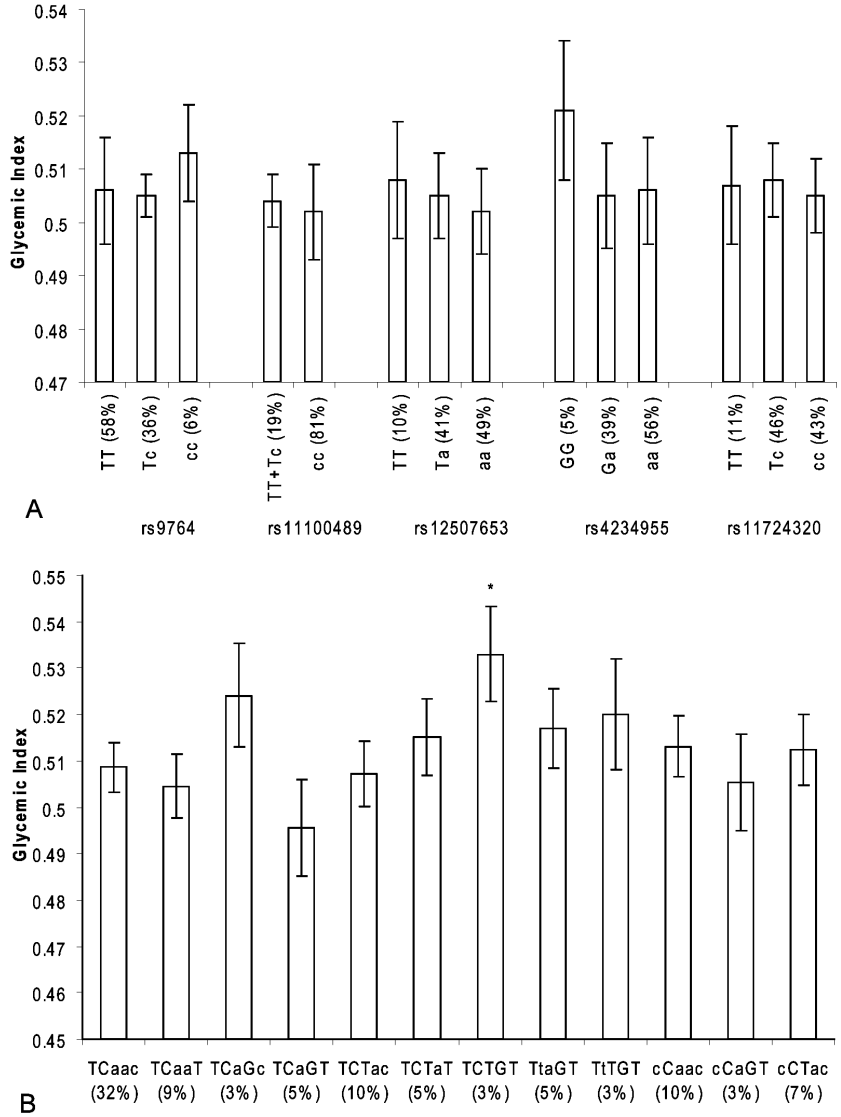


Figure 5. NPY1R/NPY5R variants and dietary glycemic index in the Hamlet population.

Figure 5a shows the association of SNPs in the NPY1R/NPY5R region with dietary glycemic index. The ancestral alleles are indicated as capital letters. Figure 5b shows the association of NPY1R/NPY5R haplotypes with dietary glycemic index. The haploblocks consist of the SNPs rs9764, rs11100489, rs12507653, rs4234955 and rs11724320 and the ancestral alleles are indicated as capital letters. *$p<0.05$ (compared with linear regression model).

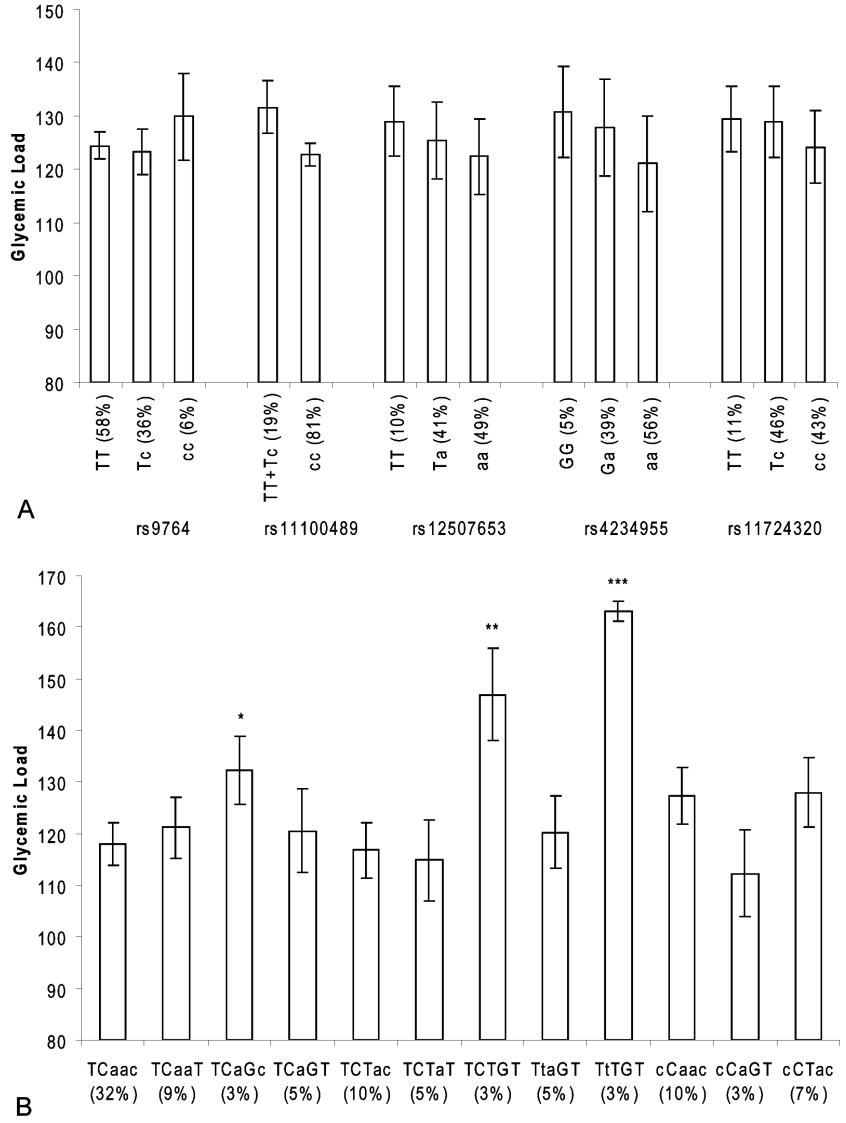


Figure 6. NPY1R/NPY5R variants and dietary glycemic load in the Hamlet population.

Figure 6a shows the association of SNPs in the NPY1R/NPY5R region with dietary glycemic load. The ancestral alleles are indicated as capital letters. Figure 6b shows the association of NPY1R/NPY5R haplotypes with dietary glycemic index. The haploblocks consist of the SNPs rs9764, rs11100489, rs12507653, rs4234955 and rs11724320 and the ancestral alleles are indicated as capital letters. *$p<0.05$, **$p<0.001$, ***$p<0.0001$ (compared with linear regression model).

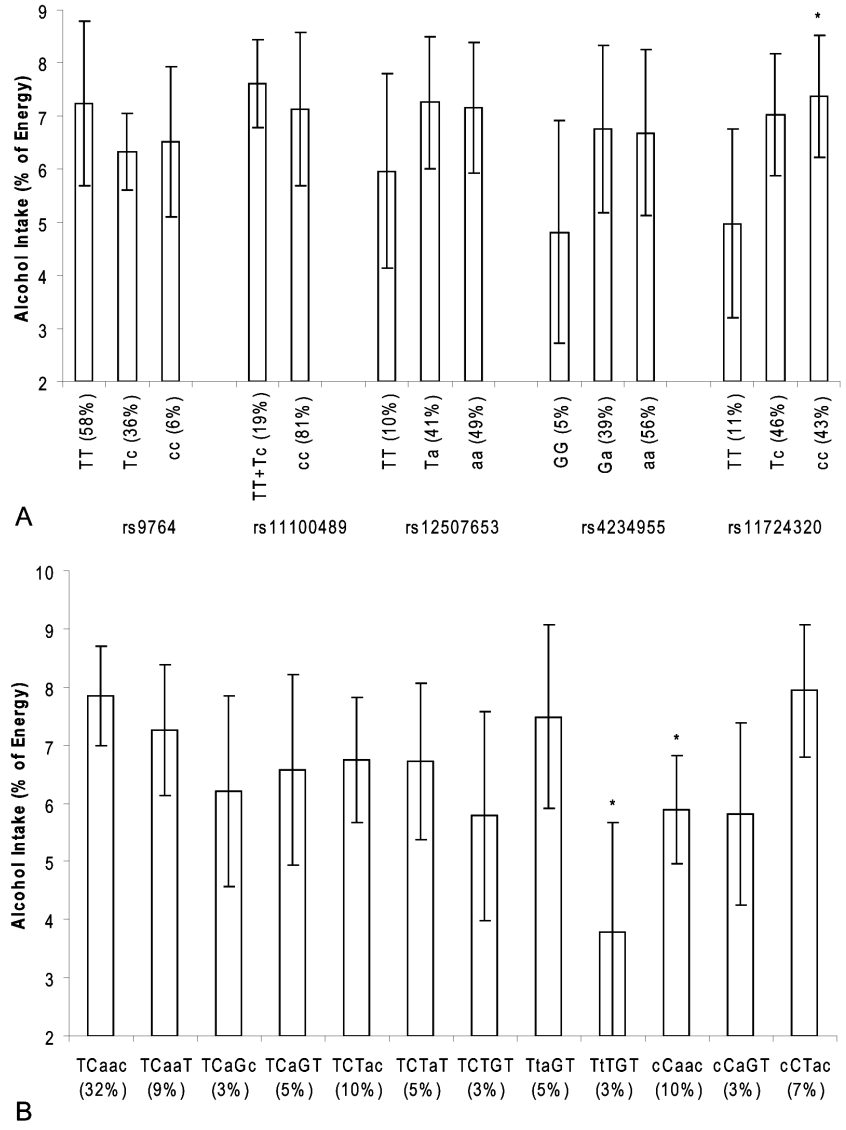


Figure 7. NPY1R/NPY5R variants and alcohol intake in the Hamlet population.
Figure 3a shows the association of SNPs in the NPY1R/NPY5R region with alcohol intake as percentage of total energy intake. The ancestral alleles are indicated as capital letters. *p<0.05, **p<0.01(compared with linear regression model), # P<0.01 for trend. Figure 3b shows the association of NPY1R/NPY5R haplotypes with alcohol intake as percentage of total energy intake. The haploblocks consist of the SNPs rs9764, rs11100489, rs12507653, rs4234955 and rs11724320 and the ancestral alleles are indicated as capital letters. *p<0.05 (compared with linear regression model).

Signatures of Recent Selection in the *NPY1R/NPY5R* Gene

As evident from Figure 8a, the derived C allele of rs11724320, located in the NPY1R/NPY5R region, is positioned on an unusually long haplotype compared to the ancestral T allele in the European HapMap individuals. In the African and Asian HapMap individuals, the haplotype lengths around rs11724320 are much shorter and there is no difference in haplotype lengths between the derived locus and the ancestral locus (Figure 8b: Africans).

The standardized iHS score is –2.160 inHapMap Caucasians, indicating that the haplotypes on the derived allele background are significantly longer than the haplotypes associated with the ancestral allele (empirical p-value p=0.03). This indicates that the locus is under recent positive selection in the European HapMap population.

To replicate these results, we calculated the haplotype decay of the derived and the ancestral alleles around the same SNP in 846 Caucasians, using EHH and iHS analysis. In this population, the derived C allele frequency was 63% and that for the ancestral T allele 37%. Although the derived C allele is very common in Europeans, it has long-range linkage disequilibrium (Figure 9). The standardized iHS score is –2.12 in for the NPY1R/NPY5R locus in the Dutch dataset and this correlates with a empirical p-value of 0.05. This implies that the allele frequency rose rapidly in the population over a short period and it confirms our previous findings that the C allele of rs11724320 is under positive selection in Europeans.

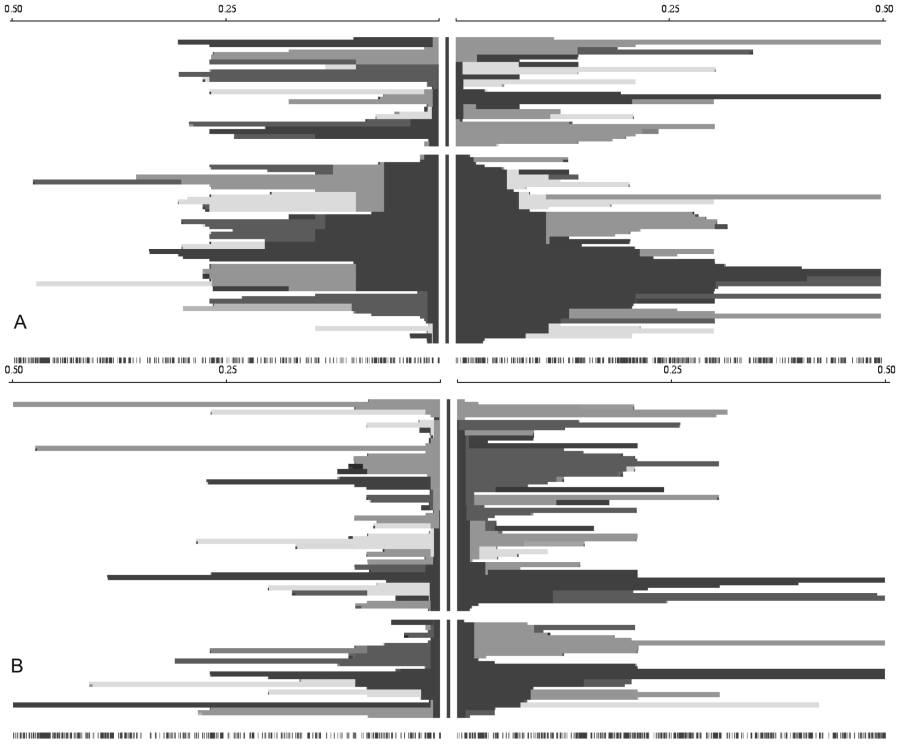


Figure 8. Haplotype decay around rs11724320 in Europeans and Africans.
Figure 8a shows the haplotype decay in genomic region of 1Mb around rs11724320, located in the NPY1R/NPY5R region, in 90 Utah residents with Northern and Western European ancestry from the CEPH collection (European HapMap individuals). Each horizontal line represents a haplotype and the center column represents the core SNP rs11724320 with the derived C-allele below and the ancestral T-allele above. The derived C allele of rs11724320is positioned on an unusually long haplotype compared to the ancestral T allele in the European HapMap individuals. Figure 8b shows the haplotype decay in genomic region of 1Mb around rs11724320, located in the NPY1R/NPY5R region, in 90 YRI Yoruba in Ibadan from Nigeria (African HapMap individuals). The haplotype lengths around the derived C allele and the ancestral T allele of rs11724320 do not significantly differ in the African population.

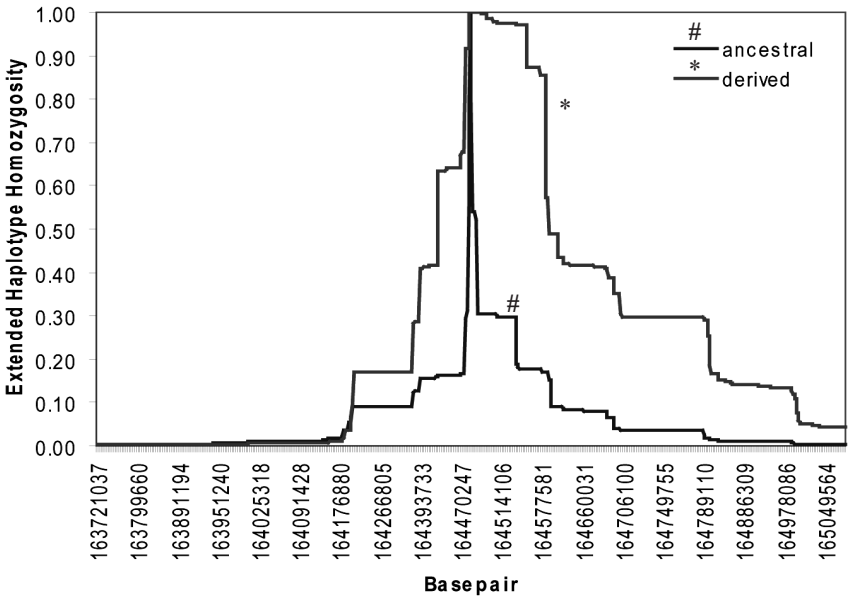


Figure 9. Haplotype decay around rs11724320 in 864 Caucasians from the Netherlands.
Extended haplotype homozygosity and distance from core region rs11724320 in GWAS data with the derived allele shown (*) and the ancestral allele (#). For this analysis we used the derived and ancestral haplotypes of 864 Caucasians from the Netherlands. Although the derived C allele is very common in this population (allele frequency of 63%), it has long-range LD. This implies that the allele frequency increased rapidly in the population over a short time span.

Age Estimation of Locus under Selection

We used the haplotype of the 846 Dutch Caucasians to obtain a crude estimate of the time of the selective sweep. At EHH = 0.25, the haplotype length around the derived allele of rs11724320 is 436,713 base pairs long. The number of generations g therefore equals (ln 0.25 / –2*0.436713) * 100 ≈ 160. Taking the generation time to be 25 years, the ancestor time becomes t = 25g and the selective sweep therefore started 25*160 ≈ 4000 years ago. The support interval, calculated at EHH = 0.15 and EHH = 0.35 is ~3800 to ~4300 years ago.

Discussion

The results of this study show that derived alleles in *NPY1R* and *NPY5R* are associated with lower relative carbohydrate intake, mainly because of a lower

consumption of mono- and disaccharides. We also show that carriers of these derived alleles on average consume meals with a lower GI and GL. However, the same alleles are associated with increased alcohol consumption. The derived allele of rs11724320 appears to be under recent selection in the European population, and probably originates from around 4,000 years ago.

A predicted selective sweep of around 4,000 years ago fits the theory of adaptation to novel food sources during the agriculture revolution, which started in Europe around 6,000 years ago and was gradually further developed from that point on. Our data suggest that a lower carbohydrate intake, consumption of meals with a low GI and GL, and/or higher alcohol consumption gave a survival advantage in Europeans during the agricultural revolution.

Adaptation to New Food Sources

The consumption of new food sources or the transition to novel dietary habits can lead to selective pressure when certain gene variants are better adapted to a particular dietary habit than others, resulting in a survival advantage for humans carrying the better adapted variants of the gene. One famous example is the selective advantage of variants in the lactase gene (LCT) which preserves the ability to digest lactose, the major sugar in milk, after weaning and throughout adult life [13]. This lactase persistence is considered an adaptation to dietary change brought about by the development of agriculture and animal domestication and husbandry.

The agricultural revolution started around 11,000 years ago in the area of the Black Sea and was accompanied by major changes in diet for many human populations [14]. In Western Europe, agriculture was started about 6,000 years ago and was gradually further developed from that point on. Therefore the selection of the derived allele of rs11724320, which originated about 4,000–5,000 years ago, might well have been driven by the transition to novel food habits.

The Mediterranean agriculture that developed in Europe at that time comprised livestock husbandry, which supplied much more protein and fat than the agricultures that developed in warmer parts of the world [14]. It can be argued that the allele under selection in the *NPY1R/NPY5R* genes, which is associated with reduced carbohydrate intake, is in fact adapted to this increased amount of protein and fat in the diet after the agriculture revolution. A higher percentage of dietary protein and fat necessarily results in a lower percentage of other macronutrients like carbohydrates.

However, we did not find an association between SNP variants in the *NPY1R/NPY5R* genes and fat and protein intake as a percentage of total energy intake, even when we took the percentage of fat and protein together in the linear regression model. This could be due to power-related issues, because one limitation of our study is the small size of the Hamlet population. Since, however, a reduced relative carbohydrate uptake must be compensated for by increased fat, protein and alcohol intake, we cannot exclude the possibility that the selective pressure was in fact due to an increased use of fat and proteins rather than reduced (simple) sugar consumption.

Thrifty genes, favoring the economical use and storage of energy, confer a survival advantage in times of food scarcity [15]. Currently, these thrifty genes are maladapted to our 'Westernized' diet and lifestyle and may nowadays be contributing to the occurrence of obesity and T2D worldwide. However, it is possible that Europeans had already started adapting genetically to a 'Westernized' diet with high fat and protein intake after the rise of Mediterranean agriculture. The lower frequency of T2D in Europeans compared to other ethnic groups that are now adopting a 'Westernized' diet and lifestyle supports this hypothesis [16].

An alternative hypothesis is that new food sources may have helped Europeans to adapt to colder climates with less sunlight [17]. Ultraviolet radiation (UVR) can damage the bare human skin, but it is also important for the synthesis of vitamin D. This vitamin plays an essential role in the mineralization and normal growth of bone during infancy and childhood. Apart from the lighter skin pigmentation, a demand for adequate vitamin D synthesis in the less sunny northern European climate may therefore have favored adaptation to vitamin D deficiency with more consumption of high fat and high protein products like liver, fish, oils, eggs and milk products (these products contain vitamin D) [17]. Caucasians living in Western Europe may have required efficient thermogenesis to cope with cold climates. Lipids from fat, but not glucose, contribute to thermogenesis during exposure to cold [18]. However, a recent study that investigated the Y chromosome and mitochondrial DNA (both parts of the DNA which do not show recombination) in 2000 Dutch men showed that 80% originated from hunter-gatherers that already populated Western Europe 25,000 year ago (http://www.nrcnext.nl/nieuws/wetenschap/article2030713.ece, article in Dutch). The ancestors of modern residents of the Netherlands had already lived in Europe long

before the agriculture revolution began and had probably already adapted to colder climates with less sunlight.

Another possibility is that instead of a survival advantage for high fat and high protein intake, the selective pressure may have been due to a survival advantage for low-carbohydrate and/or low GI/GL diets. Many studies have assessed the effectiveness and safety of different weight-loss diets, including the low-carbohydrate diet without calorie restriction.[19-21] After 12 to 48 months, the participants who were on the low-carbohydrate diet not only showed significantly more weight-loss compared to participants in other diet groups, but also experienced positive changes in overall metabolic effects and lipid profiles. They also experienced a decrease in C-reactive protein levels and in blood pressure. Thus, reducing carbohydrate intake as a percentage of total energy intake results in overall health benefits, even though the total energy intake is not restricted.

All dietary carbohydrates can be digested or converted into glucose. However, there are several types of carbohydrates (like monosaccharides, disaccharides, oligosaccharides, and starch and non-starch polysaccharides) and people's physiological glycemic and insulinemic responses to these different carbohydrates vary substantially [22] A high GI meal is followed by rapid absorption of glucose and rapid stimulation of insulin secretion and other hormones. Within the first 2 hours of consuming a high GI meal, plasma glucose levels can become twice as high as after consuming a low GI meal containing identical nutrition and energy. This rapid response after ingestion of a high GI meal challenges the mechanism of energy homeostasis; acute metabolic effects follow a high GI meal [22].

A meta-analysis of observational studies on the effects of dietary GI and GL on the risk of chronic diseases showed an association of high-GI and/or high-GL with an increased risk of chronic diseases, such as type 2 diabetes, coronary heart disease, gall bladder disease, and breast cancer [23].

Finally, as we also found an increased alcohol intake for the derived allele carriers, we cannot exclude the possibility that increased alcohol consumption had a survival advantage. Rodent studies indicate that ethanol consumption and resistance are inversely related to NPY signaling. Both the NPY- and NPY1R-deficient mice showed increased ethanol consumption and reduced sensitivity to ethanol-induced sedation [24]. Multiple studies in humans show that moderate alcohol consumption has a protective effect for T2D possibly due to increased insulin sensitivity [25-26].

Also, anti-inflammatory effects of moderate alcohol consumption may be involved in this risk reduction. [27]

NPY1R/NPY5R - The Hypothalamus Pathway and Nutrient-Specific Food Intake

To our knowledge this is the first report of variants in the *NPY1R/NPY5R* genes being associated with nutrient-specific food intake. Our findings correspond with rodent studies in which NPY evoked feeding behavior, inducing particularly carbohydrate intake.

Two other genes from the hypothalamus pathway also have been found to be associated with nutrient-specific food intake, but not with total energy intake. The Ala67Thr SNP in the agouti-related protein (AGRP) gene was associated with lower fat intake and higher carbohydrate intake [28] and the rs2272382 SNP in the TUB gene was shown to be associated with an increased energy intake from carbohydrates, mainly because of consuming more mono- and disaccharides. The same SNP was shown to be associated with a higher daily GL food intake [29]. This implies that the hypothalamus pathway plays an important role in controlling nutrient-specific food intake.

Correcting for multiple testing

In this study the p-values of the results of the association analysis are presented without correction for multiple testing. We justify this firstly, because we do not test hypothesis-free. Secondly, both the macronutrient intake (in percentage of total energy intake) and the tagging SNPs are not independent measurements.

In the first stage of the analyses we tested 5 SNPs in *NPY2R* and 5 SNPs in *NPY1R/NPY5R* region for association with macro-nutrient intake (as percentage of total energy intake). In the second stage of the analyses we continued with the most interesting findings and therefore we tested 5 SNPs and 12 haplotypes in the *NPY1R/NPY5R* region for association with subgroups of carbohydrate intake (mono- and disaccharides, polysaccharides, GI and GL).

In the first stage, after controlling for testing 10 SNPs for association with macronutrient intake by the False Discovery Rate (FRD) procedure none of the associations remain statistically significant at a threshold (q) of 0.10. However in the second stage, after controlling for testing 5 SNPs together with 12 haplotypes in the

NPY1R/NPY5R region for association with subgroups of carbohydrate intake the association of haplotype TTTGT with carbohydrate intake, the associations of haplotype TTTGT and rs12507653 with mono- and disaccharides and the associations of haplotypes TTTGT, TCTGT and TCAGA and GL remain statistically significant at a threshold (q) of 0.10.

Further studies should be done to confirm these associations in other populations.

Conclusion

We show that derived alleles in NPY1R and NPY5R are associated with lower carbohydrate intake, mainly because of a lower consumption of mono- and disaccharides. We also show that carriers of these derived alleles, on average, consume meals with a lower glycemic index and glycemic load and have higher alcohol consumption. One of these variants shows the hallmark of recent selection in Europe.

Our data suggest that lower carbohydrate intake, consuming meals with a low glycemic index and glycemic load, and/or higher alcohol consumption, gave a survival advantage in Europeans since the agricultural revolution. This advantage could lie in overall health benefits, because lower carbohydrate intake, consuming meals with a low GI and GL, and/or higher alcohol consumption, are known to be associated with a lower risk of chronic diseases.

Methods

Hamlet study

The Hamlet study is a cross-sectional, single-center study in 400 men aged 40 to 80 years living independently. The recruitment of the participants has been described elsewhere.[30] In brief, participants visited the study center twice for physical examinations, including drawing of blood, and filled in a validated food frequency questionnaire (FFQ) on their dietary intake, which is designed to estimate regular intake of 178 food items in the year before enrolment.[31,32] We calculated and assigned the values (grams/day) for total energy, fat, carbohydrates, protein and

alcohol for each food item in the FFQ (described in detail by de Kleijn et al. [33]). Energy-adjusted intake was calculated using the nutrient-density method.[34]

We calculated glycemic load (GL) by multiplying the glycemic index (GI) of a food item with its carbohydrate content, then multiplied this value with its frequency of consumption and summed the values over all food items.[35,36] Glycemic load thus represented both quality and quantity of carbohydrates, and interaction between the two. Each unit of dietary glycemic load represented the equivalent of 1 g carbohydrate from glucose. The overall glycemic index of a man's diet was calculated by dividing the dietary glycemic load by the total amount of carbohydrate consumed. Such expression of dietary glycemic index per gram of carbohydrate thus reflects the overall quality of the daily carbohydrate intake.

The Pearson correlation coefficient between the FFQ and twelve monthly recall questionnaires (each for a 24-hour period) ranged from 0.61 to 0.85 for the macronutrients, energy intake and alcohol intake.

All participants gave written informed consent before enrolment and the study was approved by the institutional review board of the University Medical Center Utrecht. Data collection took place between March 2001 and April 2002.

HapMap

A total of 270 people are included in the HapMap database (Phase II) [36]: 30 trios of US residents with Northern and Western European ancestry (CEU), 30 trios of Yoruba people from Ibadan, Nigeria (YRI), 45 unrelated Japanese individuals from the Tokyo area (ASN), and 45 unrelated Chinese individuals from Beijing (ASN).

GWAS data

We used a genome-wide dataset of 846 Dutch blood bank controls. More details on this study are described elsewhere [37]. All individuals gave their informed consent. This study was approved by the Medical Ethical Committee of the University Medical Center Utrecht.

Genotyping in Hamlet

Information about SNPs in the *NPY1R, NPY2R, NPY5R* genes was obtained from the HapMap project (www.hapmap.org, HapMap data Rel#21/phase II Jul 06). Tagging SNPs (tSNPs) were selected using Haploview version 3.2, which is based on Tagger

software (www.broad.mit.edu/mpg/tagger/) [38]. so that all SNPs with a minor allele frequency (MAF) of $\geq$ 0.10 were captured with r2 $\geq$ 0.8. We selected five tSNPs (rs6849115, rs1021868, rs12507396, rs1047214, rs9990860) for the *NPY2R* gene and five more (rs9764, rs11100489, rs12507653, rs4234955 and rs17724320) for the *NPY1R* and *NPY5R* genes, as these two genes are located together in the human genome (Figure 1).

These SNPs were genotyped in the Hamlet study using Taqman assays-on-demand (Applied Biosystems, Nieuwerkerk a/d IJssel, the Netherlands), performed according to the manufacturer's specifications. The sequence information for all primers and probes is available upon request. The genotypes were analyzed using a TaqMan 7900HT (Applied Biosystems, Nieuwerkerk a/d IJssel, the Netherlands). The DNA samples were processed in 384-well plates. Each plate contained 8 negative controls and 16 genotyping controls, which consisted of four duplicates of four different samples obtained from the Centre d'Etude du Polymorphisme Humain (CEPH).

Data Analysis in Hamlet

The genotype frequencies were tested for Hardy–Weinberg equilibrium by χ 2 analysis. Association between genotypes (as the independent variable) and macronutrient intake (as dependent variables) was determined using linear regression analysis. We studied single SNP associations with total energy intake and macronutrient-specific energy intake using the ancestral allele as reference in the linear regression model, although it was not always the most frequent allele. The ancestral allele was based on alignment to the chimpanzee sequence. As we wished to study the effect on macronutrient intake independent of total energy intake, this total intake was included in the models as an explanatory variable. We also performed trend analyses to test a dose-response effect for the derived alleles.

The False Discovery Rate (FDR) method from Benjamini and Hochberg was used to control for multiple testing [39].

All statistical analyses were performed using SPSS, version 15.0 for Windows (SPSS, Chicago, IL, USA). Haplotype analyses in Hamlet were performed using the haplo.stats package of R (version 2.7.1). The ancestral haplotype had an allele frequency of 0.03 in the Hamlet population and we therefore included all allele frequencies > 0.02 for analysis. The TCAAC haplotype was the most common, with

an allele frequency of 0.33, and this haplotype was used as a reference in the linear regression model. We did not use the ancestral haplotype as a reference because of its low frequency in the Hamlet population.

Integrated Haplotype Score (iHS) Analysis

We used the web-based tool haplotter to calculate extended haploblocks around our SNPs in HapMap and the online available software to calculate extended haploblocks around SNPs in our genome wide dataset, using the iHS method.[12] iHS is a statistic that was developed to detect evidence of recent positive selection (< 30,000 years ago) at a locus, and is based on the differential levels of linkage disequilibrium surrounding a positively selected allele compared to the background allele at the same position. An extremely positive iHS score (> 2) means that haplotypes on the ancestral allele background are longer than the derived allele background, while an extremely negative iHS score (< –2) means that the haplotypes on the derived allele background are longer than the haplotypes associated with the ancestral allele.

Extended Haplotype Homozygosity Analysis in the GWAS Data

A region of 1 Mb around *NPY1R/NPY5R* was extracted from the imputed GWAS dataset. We used the Beagle software program to infer haplotypes from genotypes of the Dutch subjects [40]. Then we calculated haplotype decay around the SNPs in the NPY1R/NPY5R region by performing extended haplotype homozygosity (EHH) analysis, using R (version 2.7.1). An EHH value stands for the probability that all haplotypes are homozygote at a recombination distance r from the selected site. We started by choosing the ancestral allele of the core SNP; at this point EHH = 1 for that allele. Next we compared the ancestral allele of the core SNP with the first proximate SNP upstream and looked for the most frequent haplotype between the ancestral allele of the core SNP and each of the alleles of the first proximate SNP. This meant that all individuals with the most frequent haplotype remained for analysis; all individuals with the other haplotype were excluded forever. If, for example, 80% of the individuals showed the most frequent haplotype between the ancestral allele and the first proximate SNP upstream, then EHH = 0.80 at that point. Subsequently, the first proximate allele upstream was compared with the second one and the same comparison and inclusion was done. We repeated this analysis with all proximate alleles on the upstream side of the ancestral allele of the core SNP until all subjects

were excluded (EHH = 0) and we performed the same procedure with all alleles located downstream from the core SNP. We then repeated this whole procedure for the derived allele of the core SNP.

Age Estimation

We used the data from the EHH analysis of the 864 Dutch individuals to obtain a crude estimate of the age of expansion of the derived variant as a result of recent selection, the so-called selective sweep. For this analysis we assumed a star phylogeny of the haplotypes. The recombination distance r is the distance in cM/Mbp between EHH = x to the left of the core SNP and EHH = x to the right of the core SNP. For a chosen x, r can be obtained from the data. As both x and r are then known, the generation time g can be calculated as: $g = (\ln x / -2r) * 100$. Assuming an average generation length of 25 years, the age of the selective sweep equals 25g. For this study, we calculated r at EHH = 0.25 (support interval EHH = 0.15 – EHH = 0.35).

Acknowledgements

This study was funded by an IOP Genomics grant IGE05012 from SenterNovem. We thank Jackie Senior for critically reading the manuscript, Mohammad Hadi Zafarmand for statistical advice and Sacha Zhernakova for data imputation.

References

1. Bell CG, Walley AJ, Froguel P (2005) The genetics of human obesity. Nat Rev Genet 6: 221-234.
2. Stanley BG, Daniel DR, Chin AS, Leibowitz SF (1985) Paraventricular nucleus injections of peptide YY and neuropeptide Y preferentially enhance carbohydrate ingestion. Peptides 6: 1205-1211.
3. Kalra SP, Kalra PS (2004) NPY and cohorts in regulating appetite, obesity and metabolic syndrome: beneficial effects of gene therapy. Neuropeptides 38: 201-211.
4. Inui A (1999) Neuropeptide Y feeding receptors: are multiple subtypes involved? Trends Pharmacol Sci 20: 43-46.
5. Sainsbury A, Schwarzer C, Couzens M, Fetissov S, Furtinger S, et al. (2002) Important role of hypothalamic Y2 receptors in body weight regulation revealed in conditional knockout mice. Proc Natl Acad Sci U S A 99: 8938-8943.
6. Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, et al. (2007) A strategy to search for common obesity and type 2 diabetes genes. Trends Endocrinol Metab 18: 19-26.
7. Rankinen T, Bouchard C (2006) Genetics of food intake and eating behavior phenotypes in humans. Annu Rev Nutr 26: 413-434.
8. Kagawa Y, Yanagisawa Y, Hasegawa K, Suzuki H, Yasuda K, et al. (2002) Single nucleotide polymorphisms of thrifty genes for energy metabolism: evolutionary origins and prospects for intervention to prevent obesity-related diseases. Biochem Biophys Res Commun 295: 207-222.
9. Stephen AM, Sieber GM, Gerster YA, Morgan DR (1995) Intake of carbohydrate and its components--international comparisons, trends over time, and effects of changing to low-fat diets. Am J Clin Nutr 62: 851S-867S.
10. Diamond J (2002) Evolution, consequences and future of plant and animal domestication. Nature 418: 700-707.
11. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.
12. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
13. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31-40.
14. Kiple KF, Ornelas KC (2000) The Cambridge world history of food. Cambridge ; New York: Cambridge University Press.
15. Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14: 353-362.
16. Dowse GK, Zimmet PZ, Finch CF, Collins VR (1991) Decline in incidence of epidemic glucose intolerance in Nauruans: implications for the "thrifty genotype". Am J Epidemiol 133: 1093-1104.
17. Jobling M, Hurles M, Tyler-Smith C (2004) Human evolutionary genetics. New York: Garland Science. xx, 523 p. p.
18. McDonald RB, Murtagh CM, Horwitz BA (1994) Age and gender effects on glucose utilization in skeletal muscle and brown adipose tissue of cold-exposed rats. Proc Soc Exp Biol Med 207: 102-109.
19. Shai I, Schwarzfuchs D, Henkin Y, Shahar DR, Witkow S, et al. (2008) Weight loss with a low-carbohydrate, Mediterranean, or low-fat diet. N Engl J Med 359: 229-241.
20. Gardner CD, Kiazand A, Alhassan S, Kim S, Stafford RS, et al. (2007) Comparison of the Atkins, Zone, Ornish, and LEARN diets for change in weight and related risk factors among overweight premenopausal women: the A TO Z Weight Loss Study: a randomized trial. Jama 297: 969-977.
21. Nordmann AJ, Nordmann A, Briel M, Keller U, Yancy WS, Jr., et al. (2006) Effects of low-carbohydrate vs low-fat diets on weight loss and cardiovascular risk factors: a meta-analysis of randomized controlled trials. Arch Intern Med 166: 285-293.
22. Ludwig DS (2002) The glycemic index: physiological mechanisms relating to obesity, diabetes, and cardiovascular disease. Jama 287: 2414-2423.
23. Barclay AW, Petocz P, McMillan-Price J, Flood VM, Prvan T, et al. (2008) Glycemic index, glycemic load, and chronic disease risk--a meta-analysis of observational studies. Am J Clin Nutr 87: 627-637.
24. Thiele TE, Badia-Elder NE (2003) A role for neuropeptide Y in alcohol intake control: evidence from human and animal research. Physiol Behav 79: 95-101.
25. Beulens JW, Stolk RP, van der Schouw YT, Grobbee DE, Hendriks HF, et al. (2005) Alcohol consumption and risk of type 2 diabetes among older women. Diabetes Care 28: 2933-2938.
26. Davies MJ, Baer DJ, Judd JT, Brown ED, Campbell WS, et al. (2002) Effects of moderate alcohol intake on fasting insulin and glucose concentrations and insulin sensitivity in postmenopausal women: a randomized controlled trial. Jama 287: 2559-2562.

27. Sierksma A, van der Gaag MS, Kluft C, Hendriks HF (2002) Moderate alcohol consumption reduces plasma C-reactive protein and fibrinogen levels; a randomized, diet-controlled intervention study. Eur J Clin Nutr 56: 1130-1136.
28. Loos RJ, Rankinen T, Rice T, Rao DC, Leon AS, et al. (2005) Two ethnic-specific polymorphisms in the human Agouti-related protein gene are associated with macronutrient intake. Am J Clin Nutr 82: 1097-1101.
29. van Vliet-Ostaptchouk JV, Onland-Moret NC, Shiri-Sverdlov R, van Gorp PJ, Custers A, et al. (2008) Polymorphisms of the TUB gene are associated with body composition and eating behavior in middle-aged women. PLoS ONE 3: e1405.
30. Muller M, den Tonkelaar I, Thijssen JH, Grobbee DE, van der Schouw YT (2003) Endogenous sex hormones in men aged 40-80 years. Eur J Endocrinol 149: 583-589.
31. Ocke MC, Bueno-de-Mesquita HB, Goddijn HE, Jansen A, Pols MA, et al. (1997) The Dutch EPIC food frequency questionnaire. I. Description of the questionnaire, and relative validity and reproducibility for food groups. Int J Epidemiol 26 Suppl 1: S37-48.
32. Ocke MC, Bueno-de-Mesquita HB, Pols MA, Smit HA, van Staveren WA, et al. (1997) The Dutch EPIC food frequency questionnaire. II. Relative validity and reproducibility for nutrients. Int J Epidemiol 26 Suppl 1: S49-58.
33. de Kleijn MJ, van der Schouw YT, Wilson PW, Adlercreutz H, Mazur W, et al. (2001) Intake of dietary phytoestrogens is low in postmenopausal women in the United States: the Framingham study(1-4). J Nutr 131: 1826-1832.
34. Willett W, Stampfer MJ (1986) Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 124: 17-27.
35. Liu S, Willett WC, Stampfer MJ, Hu FB, Franz M, et al. (2000) A prospective study of dietary glycemic load, carbohydrate intake, and risk of coronary heart disease in US women. Am J Clin Nutr 71: 1455-1461.
36. (2005) A haplotype map of the human genome. Nature 437: 1299-1320.
37. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 39: 827-829.
38. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, et al. (2005) Efficiency and power in genetic association studies. Nat Genet 37: 1217-1223.
39. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met 57: 289-300.
40. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81: 1084-1097.

Chapter 6

Obesity genes identified in genome-wide association studies are associated with adiposity measures and potentially with nutrient-specific food preference

Florianne Bauer, Clara C Elbers, Roger AH Adan, Ruth JF Loos,
N Charlotte Onland-Moret, Diederick E Grobbee,
Jana V van Vliet-Ostaptchouk, Cisca Wijmenga,
Yvonne T van der Schouw

Am J Clin Nutr. 2009 Oct;90(4):951-9

Abstract

Background: New genetic loci, most of which are expressed in the brain, have recently been reported to contribute to the development of obesity. The brain, especially the hypothalamus, is strongly involved in regulating weight and food intake.

Objectives: We investigated whether the recently reported obesity loci are associated with measures of abdominal adiposity and whether these variants affect dietary energy or macronutrient intake.

Design: We studied 1700 female Dutch participants in the European Prospective Investigation into Cancer and Nutrition (EPIC). Their anthropometric measurements and intake of macronutrients were available. Genotyping was performed by using KASPar chemistry. A linear regression model, with an assumption of an additive effect, was used to analyze the association between genotypes of 12 single nucleotide polymorphisms (SNPs) and adiposity measures and dietary intake.

Results: Seven SNPs were associated ($P < 0.05$) with weight, body mass index (BMI), and waist circumference (unadjusted for BMI). They were in or near to 6 loci: FTO, MC4R, KCTD15, MTCH2, NEGR1, and BDNF. Five SNPs were associated with dietary intake ($P < 0.05$) and were in or near 5 loci: SH2B1 (particularly with increased fat), KCTD15 (particularly with carbohydrate intake), MTCH2, NEGR1, and BDNF.

Conclusions: We confirmed some of the findings for the newly identified obesity loci that are associated with general adiposity in a healthy Dutch female population. Our results suggest that these loci are not specifically associated with abdominal adiposity but more generally with obesity. We also found that some of the SNPs were associated with macronutrient-specific food intake.

Introduction

Obesity is a risk factor for developing several diseases, including type 2 diabetes and certain types of cancer [1]. Obesity, defined as a body mass index (BMI; in kg/m2) of 30, results from an imbalance in energy intake and energy expenditure [2]. However, the underlying mechanisms are largely unknown and are being intensively studied. Recently, it has become clear that not only the amount of body fat but also its distribution is important in determining disease risk: an increasing waist circumference as a measure of abdominal obesity is related to increased chronic disease risk and mortality, independent of BMI as a measure of general obesity [3].

Although environmental factors play an important role in the development of obesity, multiple twin and family studies have indicated that genetic factors make a significant contribution to its aetiology [4]. Many genetic loci have been identified as being associated with obesity; however, these loci only explain a small part of the genetic variance underlying the development of obesity [5, 6]. Recently, genome-wide association studies (GWAS) have expanded the number of genetic susceptibility loci for obesity by identifying several new single nucleotide polymorphisms (SNPs) consistently associated with both BMI and weight, and thus, contributing to obesity risk [7, 8]. The loci identified are located in or near the genes *FTO, MC4R, TMEM18, GNPDA2, SH2B1, KCTD15, MTCH2, NEGR1, BDNF, and ETV5* [7, 8].

These loci are likely to be involved in many biological pathways because they are expressed in numerous tissues. Notably, some of the new obesity genes (*FTO, MC4R, TMEM18, GNPDA2, SH2B1, KCTD15, and BDNF*) are expressed particularly in the hypothalamus, a crucial centre for energy balance and regulation of food intake [2, 9–13]. Whereas total energy intake is a vital aspect of food intake, the macronutrient composition of food or dietary patterns may be equally important as factors underlying the development of obesity. However, the long-term risks and benefits of high-fat, low-carbohydrate diets or high-protein diets are a matter of lively scientific debate [14–16].

We set out first to investigate whether the recently reported obesity loci are more specifically associated with abdominal obesity—an important contributor to increased morbidity and mortality, independent of the total amount of body fat. Second, we explored the effect of variation in the loci implicated with obesity on

dietary energy and macronutrient intakes in 1700 healthy Dutch women to investigate whether food intake is involved in the development of obesity.

Methods

Study population

The study population consisted of Dutch female participants in the European Prospective Investigation into Cancer and Nutrition (EPIC), conducted in Utrecht, Netherlands (Prospect-EPIC) [17]. Between 1993 and 1997, 17,357 women aged 49–70 y and residing in or near Utrecht were recruited through a regional, population-based, breast cancer screening program. All of the women gave written informed consent, and the study was approved by the University Medical Center Utrecht Review Board. At recruitment, each participant filled in a general questionnaire on lifestyle factors, gynaecologic and obstetric history, and past and current morbidity as well as a validated, semi quantitative, food-frequency questionnaire (FFQ) with the aim of capturing the habitual diet during the year preceding enrolment. In addition, pulse rate, blood pressure, and anthropometric measurements were recorded, a blood sample was taken, and serum, plasma, erythrocytes, and buffy coat samples were stored at –196°C. A random sample of 1736 (10%) women was selected for biochemical analyses. Buffy coat samples were missing for 36 women; therefore, our study population comprised 1700 women. For the analyses of energy and macronutrient intakes, we excluded women who did not fill in the dietary questionnaire (n = 11). In addition, we excluded women with an implausibly low total energy intake of <800 kcal/d (n = 9).

Adiposity measures

Body height was measured to the nearest 0.5 cm with a wall-mounted stadiometer (Lameris, Utrecht, Netherlands). Body weight was measured with the subjects wearing light indoor clothing and no shoes to the nearest 0.5 kg with a floor scale (Seca, Atlanta, GA). Body mass index was calculated as weight divided by height squared (kg/m2). Waist circumference was measured to the nearest 0.5 cm with a standard household tape measure. We considered BMI to be a measure of general adiposity and waist circumference of abdominal adiposity.

Food-frequency questionnaire

The FFQ contained questions on the usual frequency of consumption of 77 main food items during the year preceding enrolment. Further information was sought on consumption frequency for different sub items, preparation methods, and additions. Colour photographs were used to estimate portion size for 28 food items. Overall, the questionnaire allows the estimation of the average daily consumption of 178 foods by asking about sub items for several foods, such as fruit and vegetables, in additional questions. Food consumption data were converted into macro- and micronutrients by using an updated version of the computerized Dutch food composition table 1996 [18]. The FFQ was validated in pilot studies before the start of our study [19]. Energy-adjusted intake was calculated by using the nutrient-density method [20].

DNA extraction and genotyping

Genomic DNA was extracted from buffy coats with the use of the QIAamp Blood Kit (Qiagen Inc, Valencia, CA). The following 13 SNPs were selected for genotyping in the random sample of the Prospect-EPIC study from the SNPs previously reported to be associated with BMI [7, 8] or from SNPs reported to be in linkage disequilibrium (LD) with the associated SNPs (see Supplementary Table 1 under "Supplemental data" in the online issue): rs1121980 (*FTO*) (proxy for rs9939609; r2 = 0.84), rs17700633 (*MC4R*), rs17782313 (*MC4R*), rs6548238 (*TMEM18*), rs10938397 (*GNDAP2*), rs7498665 (*SH2B1*), rs368794 (*KCTD15*) (proxy for rs11084753; r2 = 1), rs10838738 (*MTCH2*), rs2568958 (*NEGR1*) (proxy for rs2815752; r2 = 1), rs1488830 (*BDNF*), rs925946 (*BDNF*), rs7647305 (*ETV5*), and r2844479 (locus at 16p21). Genotyping of the 13 SNPs was performed by Kbioscience (Hoddesdon Herts, United Kingdom) by using KASPar chemistry, which is a competitive allele-specific PCR SNP genotyping system that uses FRET quencher cassette oligos (http://www.kbioscience.co.uk/genotyping/genotyping_chemistry.html, last accessed 9 February 2009). Blind duplicates, plate-identifying blank wells, and Hardy-Weinberg equilibrium tests were used as quality-control tests. Typing of the 13 SNPs resulted in genotype success rates >95%, except for rs368794 (93.5%) and rs2844479 (88.4%). We included 12 SNPs with a genotype success rate 93.5% for data analysis. None of the genotype distributions of the SNPs deviated significantly from Hardy-Weinberg equilibrium ($P > 0.01$). The SNPs located in the *MC4R* and *BDNF* loci

were not in LD in our study population: r2 = 0.13 (rs17700633, rs17782131) and r2 = 0.12 (rs1488830, rs925946), respectively.

Data analysis

Population characteristics are expressed as means ± SDs for continuous, normally distributed traits, and frequencies are expressed as percentages for categorical variables. The genotype frequencies were tested for Hardy-Weinberg equilibrium by using a chi-square test with 1 df. A linear regression model was used to analyze the association between the 12 SNP genotypes and adiposity measures and dietary energy and macronutrient intakes (energy-adjusted). We assessed this association under an additive genetic model, which assumes that there is a linear gradient in risk of each additional risk allele. Individuals homozygous for the risk allele, previously defined as the risk rising effect allele associated with BMI [7, 8], served as a reference group.

We adjusted our analyses of waist circumference for BMI. The estimates of dietary energy and macronutrient intakes did not differ after BMI was adjusted for. The false discovery rate (FDR) method from Benjamini and Hochberg was used to control for multiple testing [21]. All statistical analyses were performed by using SPSS, version 14.0 for Windows (SPSS Inc, Chicago, IL).

Table 1. Anthropometric and food intake characteristics of 1700 healthy Dutch women from the Prospect–European Prospective Investigation into Cancer and Nutrition (EPIC) study

Characteristic	Mean ± SD
Age (y)	57.22 ± 6.06
Height (cm)	164.20 ± 6.01
Weight (kg)	69.84 ± 11.48
BMI (kg/m^2)	25.90 ± 4.02
Waist (cm)	83.32 ± 10.05
Total energy intake (kcal/d)	1797.71 ± 430.65
Protein intake (g/d)[1]	70.53 ± 9.08
Fat intake (g/d)[1]	68.56 ± 10.27
Carbohydrate intake (g/d)[1]	195.31 ± 27.98
Alcohol (g/d)[1]	9.08 ± 12.73

[1] Energy-adjusted

Results

Population and genotype characteristics of the random sample of the Prospect-EPIC study are shown in Table 1 and Table 2. The 80% power we had in Prospect-EPIC to detect effect sizes with the lowest minor allele frequency of 0.22 of all typed SNPs is shown in Table 3.

Table 2. Genotypic information for 1700 healthy Dutch women from the Prospect–European Prospective Investigation into Cancer and Nutrition (EPIC) study[1]

SNP (locus)	Tested allele	MAF	HWE	LD
rs1121980 (*FTO*)	*A*	0.42	0.04	
rs17700633 (*MC4R*)	*A*	0.3	0.98	
rs17782313 (*MC4R*)	*C*	0.26	0.94	$r^2 = 0.13$
rs6548238 (*TMEM18*)	*C*	0.84	0.49	
rs10938397 (*GNPDA2*)	*G*	0.42	0.76	
rs7498665 (*SH2B1*)	*G*	0.42	0.38	
rs368794 (*KCTD15*)	*T*	0.67	0.95	
rs10838738 (*MTCH2*)	*G*	0.33	0.22	
rs2568958 (*NEGR1*)	*A*	0.58	0.11	
rs1488830 (*BDNF*)	*T*	0.78	0.33	
rs925946 (*BDNF*)	*T*	0.29	0.47	$r^2 = 0.12$
rs7647305 (*ETV5*)	*C*	0.8	0.69	

[1] SNP, single nucleotide polymorphism; HWE, Hardy-Weinberg equilibrium; LD, linkage disequilibrium; MAF, minor allele frequency.

Table 3. Power (80%) to detect effect sizes with the lowest minor allele frequency of 0.22 of all typed single nucleotide polymorphisms in 1700 healthy Dutch women from the Prospect–European Prospective Investigation into Cancer and Nutrition (EPIC) study

Genetic model	β (Additive)
Height (cm)	0.7
Weight (kg)	1.4
Waist (cm)	1.2
Hip (cm)	1
Waist-to-hip ratio	0.007
BMI (kg/m^2)	0.5
Protein intake (g/d)[1]	1.1
Fat intake (g/d)[1]	1.2
Carbohydrate intake (g/d)[1]	3.3
Alcohol (g/d)[1]	1.5

[1] Energy-adjusted

Analysis of association between SNPs and adiposity measures

Linear regression analyses of the adiposity measures with the 12 SNPs are shown in Table 4 and Table 5. We found statistically significant associations ($P < 0.05$) with different adiposity measures and 7 of the 12 analyzed SNPs located in or near *FTO, MC4R, KCTD15, MTCH2, NEGR1*, and *BDNF*. An increase in weight, ranging from 0.90 to 1.36 kg per allele, was shown in the analyzed SNPs in or near *FTO, MC4R*, and *BDNF*. The SNPs in or near the *FTO, MC4R, MTCH2,* and *BDNF* loci were associated with an increase in BMI ranging from 0.30 to 0.56 per allele. Women carrying the effect allele in only the SNPs in or near the *FTO* and *MC4R* loci had an increase in waist circumference of 1.23 cm (95% CI: 0.55, 1.91) and 1.38 cm (95% CI: 0.63, 2.13) per allele, respectively. However, after BMI was adjusted for, waist circumference was no longer associated with the SNPs in or near *FTO* and *MC4R*. Other SNPs did not show statistically significant associations with adiposity measures but did show some trends in the same directions.

Table 4. Relation between new obesity loci and adiposity measures in 1700 healthy Dutch women from the Prospect–European Prospective Investigation into Cancer and Nutrition (EPIC) study by single nucleotide polymorphism[1]

	Per *A* allele change			
	rs1121980 (*FTO*)		rs17700633 (*MC4R*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Height (cm)	–0.31 (–0.72, 0.10)	0.14	0.11 (–0.34, 0.56)	0.63
Weight (kg)	1.28 (0.50, 2.06)	0.001	1.36 (0.51, 2.21)	0.002
BMI (kg/m^2)	0.56 (0.29, 0.83)	0.0001	0.47 (0.17, 0.77)	0.002
Waist (cm)	1.23 (0.55, 1.91)	0.0004	1.38 (0.63, 2.13)	0.0003
Waist adjusted for BMI (cm)	0.001 (–0.33, 0.33)	0.99	0.34 (–0.03, 0.70)	0.07
	Per *C* allele change			
	rs17782313 (*MC4R*)		rs6548238 (*TMEM18*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Height (cm)	0.52 (0.05, 0.99)	0.03	–0.06 (–0.63, 0.51)	0.85
Weight (kg)	0.98 (0.08, 1.87)	0.03	0.68 (–0.40, 1.76)	0.22
BMI (kg/m^2)	0.22 (–0.10, 0.53)	0.18	0.26 (–0.12, 0.64)	0.17
Waist (cm)	0.57 (–0.22, 1.35)	0.16	0.62 (–0.32, 1.57)	0.2
Waist adjusted for BMI (cm)	0.11 (–0.27, 0.49)	0.58	0.07 (–0.39, 0.52)	0.78

[1] Data were derived from a linear regression model.

Table 4. (continued)

	Per *G* allele change			
	rs10938397 (*GNPDA2*)		rs7498665 (*SH2B1*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Height (cm)	0.01 (–0.41, 0.43)	0.97	0.01 (–0.41, 0.44)	0.95
Weight (kg)	–0.05 (–0.84, 0.74)	0.9	0.61 (–0.19, 1.42)	0.14
BMI (kg/m^2)	–0.01 (–0.29, 0.27)	0.93	0.20 (–0.09, 0.48)	0.17
Waist (cm)	–0.16 (–0.86, 0.54)	0.66	0.45 (–0.26, 1.15)	0.21
Waist adjusted for BMI (cm)	–0.13 (–0.46, 0.21)	0.45	0.02 (–0.32, 0.36)	0.91

[1] Data were derived from a linear regression model.

Table 5. Relation between new obesity loci and adiposity measures in 1700 healthy Dutch women from the Prospect–European Prospective Investigation into Cancer and Nutrition (EPIC) study by single nucleotide polymorphism[1]

	Per *T* allele change		Per *G* allele change	
	rs368794 (*KCTD15*)		rs10838738 (*MTCH2*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Height (cm)	0.53 (0.08, 0.98)	0.02	–0.06 (–0.49, 0.38)	0.8
Weight (kg)	0.72 (–0.12, 1.57)	0.09	0.78 (–0.04, 1.61)	0.06
BMI (kg/m^2)	0.09 (–0.21, 0.39)	0.55	0.30 (0.01, 0.59)	0.04
Waist (cm)	0.07 (–0.68, 0.82)	0.85	0.61 (–0.11, 1.33)	0.1
Waist adjusted for BMI (cm)	–0.10 (–0.47, 0.26)	0.58	–0.07 (–0.42, 0.27)	0.68
	Per *A* allele change		Per *C* allele change	
	rs2568958 (*NEGR1*)		rs7647305 (*ETV5*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Height (cm)	–0.44 (–0.85, –0.03)	0.04	–0.15 (–0.68, 0.37)	0.57
Weight (kg)	–0.34 (–1.12, 0.44)	0.39	0.72 (–0.27, 1.70)	0.15
BMI (kg/m^2)	0.01 (–0.27, 0.28)	0.97	0.32 (–0.03, 0.66)	0.07
Waist (cm)	–0.42 (–1.10, 0.27)	0.23	0.44 (–0.43, 1.30)	0.33
Waist adjusted for BMI (cm)	–0.43 (–0.76, –0.10)	0.01	–0.25 (–0.67, 0.17)	0.24
	Per *T* allele change			
	rs1488830 (*BDNF*)		rs925946 (*BDNF*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Height (cm)	–0.14 (–0.64, 0.37)	0.59	–0.03 (–0.49, 0.44)	0.91
Weight (kg)	–0.04 (–1.00, 0.92)	0.94	0.90 (0.02, 1.77)	0.05
BMI (kg/m^2)	0.04 (–0.30, 0.38)	0.81	0.32 (0.02, 0.63)	0.04
Waist (cm)	–0.11 (–0.95, 0.73)	0.8	0.70 (–0.07,1.46)	0.08
Waist adjusted for BMI (cm)	–0.19 (–0.59, 0.22)	0.37	0.03 (–0.34, 0.40)	0.87

[1] Data were derived from a linear regression model.

Analysis of association between SNPs and dietary energy and macronutrient intakes

Linear regression analyses of dietary energy and macronutrient intakes with the 12 SNPs are shown in Table 6 and Table 7. We found statistically significant associations ($P < 0.05$) between 5 of the 12 SNPs and macronutrient intake located in or near *SH2B1, KCTD15, MTCH2, NEGR1,* and *BDNF*. The risk allele at rs7498665 (SH2B1) was associated with increased total fat (per allele effect: 1.08 g/d energy-adjusted; 95% CI: 0.36, 1.81), saturated fat (per allele effect: 0.60 g/d energy-adjusted; 95% CI: 0.22, 0.97), and monounsaturated fat intake (per allele effect: 0.37 g/d energy-adjusted; 95% CI: 0.04, 0.69). A decrease in monounsaturated fat intake was shown for the risk alleles of the SNPs in or near *KCTD15* and *NEGR1*, whereas carriers of the risk allele for *NEGR1* also had lower saturated fat intakes. Carriers of the risk allele in or near *KCTD15* consumed less monounsaturated fat (per allele effect: –0.37; 95% CI: –0.72, –0.02), and for *NEGR1* they consumed less saturated fat (per allele effect: –0.40 g/d energy-adjusted; 95% CI: –0.77, –0.04) and monounsaturated fat (per allele effect: –0.34 g/d energy-adjusted; 95% CI: –0.65, –0.03).

In addition to the association of carriers with the risk allele of the SNP in or near *KCTD15* consuming less fat, carriers of this risk allele ate more total carbohydrate (per allele effect: 2.50 g/d energy-adjusted; 95% CI: 0.39, 4.60) and mono- and disaccharides (per allele effect: 2.62 g/d energy-adjusted; 95% CI: 0.69, 4.55).

Carriers with the risk allele of the SNP rs10838738 (*MTCH2*) consumed less polysaccharides (per allele effect: –1.33 g/d energy-adjusted; 95% CI: –2.61, –0.05). Women with the risk allele at rs925946 (*BDNF*) consumed less alcohol (per allele effect: –1.15 g/d energy-adjusted; 95% CI: –2.14, –0.17).

After multiple testing using FDR was controlled for, the associations between rs1121980 (*FTO*) and BMI and waist and the association between rs17700633 (MC4R) and waist remained statistically significant at an FDR threshold of 0.05, whereas the associations between rs1121980 (*FTO*) and weight, rs17700633 (*MC4R*) and weight and BMI and rs7498665 (*SH2B1*) and fat and saturated fat intakes remained statistically significant at an FDR threshold of 0.10.

Table 6. Relation between new obesity loci and dietary energy and macronutrient intakes in 1700 healthy Dutch women from the Prospect–European Prospective Investigation into Cancer and Nutrition (EPIC) study by single nucleotide polymorphism[1]

	Per *A* allele change			
	rs1121980 *(FTO)*		rs17700633 *(MC4R)*	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Energy (kcal)	–19.0 (–48.1, 10.1)	0.2	3.4 (–28.5, 35.6)	0.83
Fat (g/d)[2]	0.55 (–0.15, 1.26)	0.12	–0.09 (–0.86, 0.68)	0.83
Saturated fat (g/d)[2]	0.25 (–0.12, 0.61)	0.19	0.12 (–0.28, 0.52)	0.56
Monounsaturated fat (g/d)[2]	0.22 (–0.10, 0.53)	0.17	–0.10 (–0.44, 0.24)	0.57
Polyunsaturated fat (g/d)[2]	0.08 (–0.16, 0.32)	0.51	–0.11 (–0.37, 0.15)	0.42
Carbohydrate (g/d)[2]	–1.74 (–3.65, 0.18)	0.08	–0.08 (–2.17, 2.02)	0.94
Mono- and disaccharides (g/d)[2]	–1.38 (–3.14, 0.38)	0.12	0.59 (–1.34, 2.52)	0.55
Polysaccharide (g/d)[2]	–0.31 (–1.51, 0.90)	0.62	0.68 (–2.10, 0.66)	0.32
Protein (g/d)[2]	0.25 (–0.45, 0.95)	0.49	0.36 (–0.41, 1.12)	0.37
Vegetable protein (g/d)[2]	0.01 (–0.27, 0.29)	0.97	–0.19 (–0.50, 0.12)	0.22
Animal protein (g/d)[2]	0.25 (–0.51, 1.02)	0.52	0.54 (–0.30, 1.38)	0.21
Alcohol (g/d)[2]	0.33 (–0.55, 1.21)	0.47	–0.16 (–1.12, 0.81)	0.75
	Per *C* allele change			
	rs17782313 *(MC4R)*		rs6548238 *(TMEM18)*	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Energy (kcal)	12.0 (–21.4, 45.5)	0.48	8.9 (–31.4, 49.1)	0.67
Fat (g/d)[2]	–0.12 (–0.93, 0.69)	0.78	0.50 (–0.47, 1.48)	0.31
Saturated fat (g/d)[2]	–0.14 (–0.56, 0.28)	0.5	0.44 (–0.07, 0.94)	0.09
Monounsaturated fat (g/d)[2]	0.00 (–0.36, 0.36)	1	0.02 (–0.41, 0.45)	0.94
Polyunsaturated fat (g/d)[2]	0.03 (–0.25, 0.30)	0.84	0.05 (–0.28,0.38)	0.77
Carbohydrate (g/d)[2]	–0.98 (–3.18, 1.23)	0.39	0.47 (–2.17, 3.10)	0.73
Mono- and disaccharides (g/d)[2]	–0.57 (–2.60, 1.46)	0.58	0.00 (–2.42, 2.42)	1
Polysaccharide (g/d)[2]	–0.42 (–1.80, 0.96)	0.55	0.45 (–1.21, 2.11)	0.59
Protein (g/d)[2]	0.04 (–0.77, 0.84)	0.93	–0.11 (–1.07, 0.86)	0.83
Vegetable protein (g/d)[2]	–0.16 (–0.48, 0.16)	0.33	0.21 (–0.17, 0.60)	0.28
Animal protein (g/d)[2]	0.18 (–0.70, 1.06)	0.69	–0.32 (–1.38, 0.74)	0.55
Alcohol (g/d)[2]	0.76 (–0.26, 1.77)	0.14	–1.01 (–2.22, 0.19)	0.1

[1] Data were derived from a linear regression model.
[2] Energy-adjusted.

Table 6. (continued)

	Per *G* allele change			
	rs10938397 *(GNPDA2)*		rs7498665 *(SH2B1)*	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Energy (kcal)	–13.5 (–43.3, 16.3)	0.37	–9.2 (–39.2, 20.9)	0.55
Fat (g/d)[2]	0.18 (–0.54, 0.90)	0.63	1.08 (0.36, 1.81)	0.003
Saturated fat (g/d)[2]	–0.08 (–0.45, 0.30)	0.68	0.60 (0.22, 0.97)	0.002
Monounsaturated fat (g/d)[2]	0.11 (–0.21, 0.43)	0.49	0.37 (0.04, 0.69)	0.03
Polyunsaturated fat (g/d)[2]	0.14 (–0.10, 0.39)	0.25	0.12 (–0.13, 0.37)	0.34
Carbohydrate (g/d)[2]	–0.26 (–2.22, 1.69)	0.79	–0.96 (–2.94, 1.02)	0.34
Mono- and disaccharides (g/d)[2]	–0.35 (–2.14, 1.45)	0.71	–1.03 (–2.85, 0.78)	0.26
Polysaccharide (g/d)[2]	0.13 (–1.11, 1.36)	0.84	0.11 (–1.14, 1.36)	0.86
Protein (g/d)[2]	–0.47 (–1.18, 0.24)	0.19	–0.33 (–1.05, 0.40)	0.38
Vegetable protein (g/d)[2]	0.13 (–0.16, 0.41)	0.39	–0.07 (–0.36, 0.22)	0.66
Animal protein (g/d)[2]	–0.59 (–1.37, 0.19)	0.14	–0.25 (–1.03, 0.54)	0.54
Alcohol (g/d)[2]	0.17 (–0.72, 1.06)	0.7	–0.70 (–1.60, 0.21)	0.13

[1] Data were derived from a linear regression model.
[2] Energy-adjusted.

Table 7. Relation between new obesity loci and dietary energy and macronutrient intake in 1700 healthy Dutch women from the Prospect–European Prospective Investigation into Cancer and Nutrition (EPIC) study by single nucleotide polymorphism[1]

	Per *T* allele change rs368794 (*KCTD15*)		Per *G* allele change rs10838738 (*MTCH2*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Energy (kcal)	–7.60 (–39.6, 24.3)	0.64	2.21 (–28.3, 32.8)	0.89
Fat (g/d)[2]	–0.64 (–1.42, 0.14)	0.11	–0.34 (–1.08, 0.41)	0.38
Saturated fat (g/d)[2]	–0.32 (–0.72, 0.09)	0.12	–0.12 (–0.51, 0.27)	0.54
Monounsaturated fat (g/d)[2]	–0.37 (–0.72; –0.02)	0.04	–0.08 (–0.41, 0.26)	0.66
Polyunsaturated fat (g/d)[2]	0.06 (–0.20, 0.33)	0.63	–0.15 (–0.41, 0.10)	0.23
Carbohydrate (g/d)[2]	2.50 (0.39, 4.60)	0.02	0.35 (–1.68, 2.37)	0.74
Mono- and disaccharide (g/d)[2]	2.62 (0.69, 4.55)	0.008	1.66 (–0.19, 3.52)	0.08
Polysaccharide (g/d)[2]	–0.11 (–1.43, 1.22)	0.87	–1.33 (–2.61, –0.05)	0.04
Protein (g/d)[2]	–0.42 (–1.19, 0.35)	0.28	0.57 (–0.17, 1.31)	0.13
Vegetable protein (g/d)[2]	0.01 (–0.30, 0.32)	0.97	–0.05 (–0.34, 0.25)	0.76
Animal protein (g/d)[2]	–0.42 (–1.26, 0.42)	0.32	0.61 (–0.20, 1.41)	0.14
Alcohol (g/d)[2]	–0.42 (–1.39, 0.55)	0.39	–0.19 (–1.12, 0.74)	0.69

[1] Data were derived from a linear regression model.
[2] Energy-adjusted.

Table 7. (continued)

	Per *A* allele change rs2568958 (*NEGR1*)		Per *T* allele change rs1488830 (*BDNF*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Energy (kcal)	–1.80 (–30.8 27.2)	0.9	–10.1 (–45.8, 25.6)	0.58
Fat (g/d)[2]	–0.56 (–1.26, 0.14)	0.12	0.25 (–0.62, 1.11)	0.58
Saturated fat (g/d)[2]	–0.40 (–0.77, –0.04)	0.03	0.11 (–0.34, 0.56)	0.64
Monounsaturated fat (g/d)[2]	–0.34 (–0.65, –0.03)	0.03	–0.03 (–0.41, 0.36)	0.89
Polyunsaturated fat (g/d)[2]	0.19 (–0.05, 0.43)	0.11	0.15 (–0.14, 0.45)	0.31
Carbohydrate (g/d)[2]	0.67 (–1.25, 2.58)	0.49	–0.14 (–2.49, 2.21)	0.91
Mono- and disaccharides (g/d)[2]	0.52 (–1.24, 2.27)	0.57	0.56 (–1.59, 2.72)	0.61
Polysaccharide (g/d)[2]	0.14 (–1.07, 1.35)	0.82	–0.70 (–2.19, 0.79)	0.35
Protein (g/d)[2]	–0.19 (–0.89, 0.51)	0.59	0.35 (–0.51, 1.21)	0.42
Vegetable protein (g/d)[2]	0.28 (0.00, 0.56)	0.05	–0.22 (–0.57, 0.12)	0.21
Animal protein (g/d)[2]	–0.47 (–1.23, 0.30)	0.23	0.57 (–0.36, 1.51)	0.23
Alcohol (g/d)[2]	0.45 (–0.42, 1.33)	0.31	0.35 (–1.42, 0.73)	0.53
	Per *T* allele change rs925946 (*BDNF*)		Per *C* allele change rs7647305 (*ETV5*)	
	β (95% CI)	*P* value	*β* (95% CI)	*P* value
Energy (kcal)	1.26 (–31.2, 33.74)	0.94	–4.18 (–41.0, 32.7)	0.82
Fat (g/d)[2]	0.24 (–0.55, 1.03)	0.55	0.02 (–0.88, 0.91)	0.97
Saturated fat (g/d)[2]	0.16 (–0.26, 0.57)	0.46	0.15 (–0.32, 0.61)	0.54
Monounsaturated fat (g/d)[2]	0.10 (–0.25, 0.45)	0.57	–0.09 (–0.49, 0.31)	0.65
Polyunsaturated fat (g/d)[2]	–0.04 (–0.30, 0.23)	0.79	–0.04 (–0.34, 0.27)	0.81
Carbohydrate (g/d)[2]	0.59 (–1.55, 2.74)	0.59	0.80 (–1.64, 3.24)	0.52
Mono- and disaccharides (g/d)[2]	0.54 (–1.43, 2.50)	0.59	0.34 (–1.91, 2.58)	0.78
Polysaccharide (g/d)[2]	0.04 (–1.32, 1.40)	0.95	0.48 (–1.06, 2.02)	0.54
Protein (g/d)[2]	0.64 (–0.14, 1.42)	0.11	–0.11 (–1.00, 0.78)	0.81
Vegetable protein (g/d)[2]	–0.01 (–0.33, 0.30)	0.94	0.17 (–0.19, 0.52)	0.36
Animal protein (g/d)[2]	0.65 (–0.21, 1.50)	0.14	–0.28 (–1.25, 0.69)	0.57
Alcohol (g/d)[2]	–1.15 (–2.14, –0.17)	0.02	–0.30 (–1.42, 0.82)	0.6

[1] Data were derived from a linear regression model.
[2] Energy-adjusted.

Discussion

Two large-scale GWA studies recently identified new genetic loci associated with measures of obesity [7, 8]. In this study we evaluated 12 common variants from these loci and confirmed the effect of the *FTO, MC4R, MTCH2*, and *BDNF* genes on weight and on BMI in a healthy Dutch female population. In addition to finding an association with general adiposity, we also found evidence of an association between

these new loci and macronutrient-specific food intake. In our population, only 4 SNPs in or near *FTO* (rs1121980), *MC4R* (rs17700633), *MTCH2* (rs10838738), and *BDNF* (rs925946) were statistically significantly associated with BMI. Although the size of our study limited our power (6–67%) [22] to identify the previously reported effect sizes of BMI, the trends we observed for the associations of the SNPs with BMI were all in the same direction, as previously reported [7, 8]. Also, for the *MC4R* SNP rs17782313 that was previously associated with dietary fat intake [23], we had 96% power to detect the same effect size.

Notably, almost all the new loci have an effect on BMI. Because visceral fat accumulation in particular is related to health risk, we chose to use waist circumference as a measure specific for the amount of intra-abdominal (visceral) fat, because this might be a more specific measure of obesity than a general adiposity measure, such as BMI [24, 25]. In this study, we had detailed information on both general and abdominal adiposity measures. We observed that the associations and trends between the new analyzed loci and BMI agreed with the associations and trends with other adiposity measures, such as weight and waist circumference. However, when we adjusted waist circumference for BMI, it was no longer associated. These results suggest that the identified loci are not specifically associated with abdominal adiposity, but merely represent loci associated with general obesity. This suggests that these loci are important in determining fat gain in general, but not in the distribution of fat in the body.

The development of obesity is due to various possible mechanisms, in which food intake also plays a role. Our results suggest that the new obesity loci might also play a role in the choice and preference of specific macronutrient intake. For the *SH2B1, KCTD15, MTCH2,* and *NEGR1* loci, the obesity-risk alleles were associated with dietary intake of saturated fat, carbohydrates, mono- and disaccharides, and polysaccharides. These results agree with previous associations found between intake of fat and carbohydrates and adiposity measures [26–29]. In this study of 1700 females, we had detailed dietary information obtained through a validated instrument, the FFQ. The Pearson correlation coefficient between the FFQ and 12 monthly 24-h recall questionnaires ranged from 0.61 to 0.85 for intake of dietary energy and macronutrients, including alcohol intake [19]. Despite observations that people who are overweight tend to underestimate their food intake [30–32], we did find associations between some SNPs and fat, carbohydrate, and alcohol intakes. With a

minor allele frequency of 0.22, we had 80% power to detect differences in energy-adjusted nutrient intakes ranging from 1.10 to 3.30 g/d at a significance level of 0.05, assuming an additive model of inheritance. We cannot exclude the possibility that we may have missed even smaller effects of dietary energy and macronutrient intakes.

Many of the new genes are highly expressed in the brain, and several are particularly evident in the hypothalamus, which is consistent with central neural system processes playing an important role in regulating body weight. The hypothalamus plays a key role in regulating energy homeostasis and food intake. Disturbances in the hypothalamic region can lead to deregulation of body weight because of changes in eating behaviour [11]. Interestingly, a few candidate genes for obesity in the hypothalamic pathway, such as *AGRP* and *TUB*, were reported to be associated with macronutrient intake. *AGRP* polymorphisms were associated with total energy, fat, and carbohydrate intakes [33], whereas variants in the *TUB* gene, associated with body weight and BMI, were also shown to be associated with eating behaviour: carriers of the risk alleles for obesity had a diet high in carbohydrates and low in fats [34]. Recently, common variants in *FTO* and *MC4R*, also related to the hypothalamic pathway, were associated with energy intake [23, 35–37], where rs17782313 near *MC4R* was also associated with dietary fat intake [23]. However, we could not confirm these findings. This may have been due to differences in the SNPs studied, the type of participants (e.g., mainly children), a smaller number of subjects, and differences in dietary intake measurement not comparable with our FFQ. However, the association with dietary fat intake and *MC4R* might also be a chance finding, because we had 96% power to find at least this same effect.

We found associations with dietary macronutrient intake and the new *SH2B1, KCTD15, MTCH2, NEGR1*, and *BDNF* loci. SNPs in or near *SH2B1, KCTD15*, and *NEGR1* were associated with total fat, saturated, and monounsaturated fat intakes. SNPs in or near *KCTD15* and *MTCH2* were associated with total carbohydrate, mono- and disaccharide, and polysaccharide intakes. To understand whether these associations can be implicated in the energy balance and food intake via a role in the hypothalamus, it is necessary to know the underlying function of the new loci. Unfortunately, little is known about these loci as yet. For *SH2B1*, there is a possible role in regulating body weight via its role in enhancing leptin signalling [38, 39]. The function of *KCTD15* is unknown. *MTCH2* may function in cellular apoptosis [40, 41], and *NEGR1* may affect neural outgrowth [42, 43]. *BDNF* is a neurotrophic factor that

promotes the differentiation and survival of developing neurons and their maintenance in the adult nervous system [44, 45]. Thus, the precise functions need to be determined to reveal the possible mechanisms in food intake of these new genes.

Because obesity is a result of an imbalance between food intake and energy expenditure (e.g., because of limited physical activity), we examined whether the new obesity genes also have an effect on physical activity. In our population we found no evidence of a relation between the novel loci and physical activity, as measured with a questionnaire validated in elderly people [46] (data not shown). We had no data on basal metabolic rate or thermogenesis, so we cannot exclude the possibility that the reported loci have an effect on obesity through energy expenditure. It is also important to note that not only genes play a role in weight regulation. Genes might interact with each other or with environmental factors such as food nutrients to play a role in the development of adiposity, but further research is necessary to investigate these mechanisms.

In conclusion, our study showed that the new loci are associated with obesity phenotypes through general adiposity. Our results further suggest that these loci play a role in nutrient-specific choice and dietary preference. These results need to be confirmed.

Acknowledgement

We thank SenterNovem (IOP Genomics IGE05012) for financial support and Jackie Senior for critically reading the manuscript.

References

1. World Health Organization. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. World Health Organ Tech Rep Ser 2000;894:1–253.
2. Schwartz, MW & Porte, D, Jr. Diabetes, obesity, and the brain. Science 2005;307:375–9.
Pischon, T, Boeing, H, Hoffmann, K, et al.. General and abdominal adiposity and risk of death in Europe. N Engl J Med 2008;359:2105–20.
3. Loos, RJ & Bouchard, C. Obesity—is it a genetic disorder? J Intern Med 2003;254:401–25.
4. Mutch, DM & Clement, K. Unraveling the genetics of human obesity. PLoS Genet 2006;2:e188.
5. Rankinen, T, Zuberi, A, Chagnon, YC, et al.. The human obesity gene map: the 2005 update. Obesity (Silver Spring) 2006;14:529–644.
6. Thorleifsson, G, Walters, GB, Gudbjartsson, DF, et al.. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 2009;41:18–24.
7. Willer, CJ, Speliotes, EK, Loos, RJ, et al.. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 2009;41:25–34.
8. Adan, RA & Vanderschuren, LJ, la Fleur SE. Anti-obesity drugs and neural circuits of feeding. Trends Pharmacol Sci 2008;29:208–17.
9. Bell, CG, Walley, AJ & Froguel, P. The genetics of human obesity. Nat Rev Genet 2005;6:221–34.
10. Berthoud, HR & Morrison, C. The brain, appetite, and obesity. Annu Rev Psychol 2008;59:55–92.
11. Berthoud, HR, Sutton, GM, Townsend, RL, Patterson, LM & Zheng, H. Brainstem mechanisms integrating gut-derived satiety signals and descending forebrain information in the control of meal size. Physiol Behav 2006;89:517–24.
12. Hofker, M & Wijmenga, C. A supersized list of obesity genes. Nat Genet 2009;41:139–40.
13. Astrup, A, Meinert Larsen, T & Harper, A. Atkins and other low-carbohydrate diets: hoax or an effective tool for weight loss? Lancet 2004;364:897–9.
14. Brehm, BJ & D'Alessio, DA. Weight loss and metabolic benefits with diets of varying fat and carbohydrate content: separating the wheat from the chaff. Nat Clin Pract Endocrinol Metab 2008;4:140–6.
15. Krieger, JW, Sitren, HS, Daniels, MJ & Langkamp-Henken, B. Effects of variation in protein and carbohydrate intake on body mass and composition during energy restriction: a meta-regression 1. Am J Clin Nutr 2006;83:260–74.
16. Boker, LK, van Noord, PA, van der Schouw, YT, et al.. Prospect-EPIC Utrecht: study design and characteristics of the cohort population. European Prospective Investigation into Cancer and Nutrition. Eur J Epidemiol 2001;17:1047–53.
17. NEVO-Table. Dutch food composition table 1996. The Hague, Netherlands: Voorlichtingsbureau voor de Voeding, 1996.
18. Ocke, MC, Bueno-de-Mesquita, HB, Pols, MA, Smit, HA, van Staveren, WA & Kromhout, D. The Dutch EPIC food frequency questionnaire. II. Relative validity and reproducibility for nutrients. Int J Epidemiol 1997;26(suppl_1):S49–58.
19. Willett, W & Stampfer, MJ. Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 1986;124:17–27.
20. Benjamini, Y & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Met 1995;57:289–300..
21. Gauderman, W & Morrison, J. QUANTO 1.1: a computer program for power and sample size calculations for genetic-epidemiology studies. 2006. Available from: http://hydra.usc.edu/gxe
22. Qi, L, Kraft, P, Hunter, DJ & Hu, FB. The common obesity variant near MC4R gene is associated with higher intakes of total energy and dietary fat, weight change and diabetes risk in women. Hum Mol Genet 2008;17:3502–8.
23. Despres, JP & Lemieux, I. Abdominal obesity and metabolic syndrome. Nature 2006;444:881–7.
24. Kuk, JL, Janiszewski, PM & Ross, R. Body mass index and hip and thigh circumferences are negatively associated with visceral adipose tissue after control for waist circumference. Am J Clin Nutr 2007;85:1540–4.
25. Field, AE, Willett, WC, Lissner, L & Colditz, GA. Dietary fat and weight gain among women in the Nurses' Health Study. Obesity (Silver Spring) 2007;15:967–76.
26. Gardner, CD, Kiazand, A, Alhassan, S, et al.. Comparison of the Atkins, Zone, Ornish, and LEARN diets for change in weight and related risk factors among overweight premenopausal women: the A TO Z Weight Loss Study: a randomized trial. JAMA 2007;297:969–77.
27. Nordmann, AJ, Nordmann, A, Briel, M, et al.. Effects of low-carbohydrate vs low-fat diets on weight loss and cardiovascular risk factors: a meta-analysis of randomized controlled trials. Arch Intern Med 2006;166:285–93.

28. Shai, I, Schwarzfuchs, D, Henkin, Y, et al.. Weight loss with a low-carbohydrate, Mediterranean, or low-fat diet. N Engl J Med 2008;359:229–41.
29. Kretsch, MJ, Fong, AK & Green, MW. Behavioral and body size correlates of energy intake underreporting by obese and normal-weight women. J Am Diet Assoc 1999;99:300–6.(quiz 307-8).
30. Lichtman, SW, Pisarska, K, Berman, ER, et al.. Discrepancy between self-reported and actual caloric intake and exercise in obese subjects. N Engl J Med 1992;327:1893–8.
31. Prentice, AM, Black, AE, Coward, WA, et al.. High levels of energy expenditure in obese women. Br Med J (Clin Res Ed) 1986;292:983–7.
32. Loos, RJ, Rankinen, T, Rice, T, et al.. Two ethnic-specific polymorphisms in the human Agouti-related protein gene are associated with macronutrient intake. Am J Clin Nutr 2005;82:1097–101.
33. van Vliet-Ostaptchouk, JV, Onland-Moret, NC, Shiri-Sverdlov, R, et al.. Polymorphisms of the TUB Gene are associated with body composition and eating behavior in middle-aged women. PLoS One 2008;3:e1405.
34. Cecil, JE, Tavendale, R, Watt, P, Hetherington, MM & Palmer, CN. An obesity-associated FTO gene variant and increased energy intake in children. N Engl J Med 2008;359:2558–66.
35. Haupt, A, Thamer, C, Staiger, H, et al.. Variation in the FTO gene influences food intake but not energy expenditure. Exp Clin Endocrinol Diabetes 2009;117:194–7.
36. Wardle, J, Llewellyn, C, Sanderson, S & Plomin, R. The FTO gene and measured food intake in children. Int J Obes (Lond) 2009;33:42–5.
37. Li, Z, Zhou, Y, Carter-Su, C, Myers, MG, Jr & Rui, L. SH2B1 enhances leptin signaling by both Janus kinase 2 Tyr813 phosphorylation-dependent and -independent mechanisms. Mol Endocrinol 2007;21:2270–81.
38. Ren, D, Zhou, Y, Morris, D, Li, M, Li, Z & Rui, L. Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. J Clin Invest 2007;117:397–406.
39. Grinberg, M, Schwarz, M, Zaltsman, Y, et al.. Mitochondrial carrier homolog 2 is a target of tBID in cells signaled to die by tumor necrosis factor alpha. Mol Cell Biol 2005;25:4579–90.
40. Pagliarini, DJ, Calvo, SE, Chang, B, et al.. A mitochondrial protein compendium elucidates complex I disease biology. Cell 2008;134:112–23.
41. Marg, A, Sirim, P, Spaltmann, F, et al.. Neurotractin, a novel neurite outgrowth-promoting Ig-like protein that interacts with CEPU-1 and LAMP. J Cell Biol 1999;145:865–76.
42. Schafer, M, Brauer, AU, Savaskan, NE, Rathjen, FG & Brummendorf, T. Neurotractin/kilon promotes neurite outgrowth and is expressed on reactive astrocytes after entorhinal cortex lesion. Mol Cell Neurosci 2005;29:580–90.
43. Huang, EJ & Reichardt, LF. Neurotrophins: roles in neuronal development and function. Annu Rev Neurosci 2001;24:677–736.
44. Jones, KR & Reichardt, LF. Molecular cloning of a human gene that is a member of the nerve growth factor family. Proc Natl Acad Sci USA 1990;87:8060–4.
45. Voorrips, LE, Ravelli, AC, Dongelmans, PC, Deurenberg, P & Van Staveren, WA. A physical activity questionnaire for the elderly. Med Sci Sports Exerc 1991;23:974–9.

Chapter 7

Young age at first full-term pregnancy is associated with increased risk for type 2 diabetes in women

Clara C. Elbers, N. Charlotte Onland-Moret, Timon W. van Haeften,
René Eijkemans, Cisca Wijmenga, Petra H.M. Peeters,
Yvonne T. van der Schouw

Abstract

Background Multiparty is found to be associated with increased type 2 diabetes (T2D) risk in previous studies. However, it remains unclear whether body mass mediate the observed association. Also it is unknown whether variation in age at first full-term pregnancy influences T2D risk.

Methods We assessed the association between parity and age at first full-term pregnancy in a cohort of 17,357 Dutch women, aged 49-70 at baseline using Cox proportional hazards models. Analyses were adjusted for multiple confounders. To investigate whether BMI, waist and gestational diabetes mediate the observed associations, analyses were additionally adjusted for these variables.

Results At baseline, 332 women had T2D. During a mean follow-up of 9.1 ± 3.6 years, 535 T2D cases occurred. Compared to women with no children, women with 4 or more children had an increased risk for T2D with a multivariate adjusted HR of 1.27 (95% CI 0.99-1.63). This association was mediated by adiposity, as adjustment for BMI and waist attenuated the association to null (HR for women with ≥ 4 children=1.04; 95% CI: 0.81–1.33).

An older age at first delivery reduced the risk of T2D, with a multivariate adjusted HR of 0.60 (95% CI 0.44–0.83) for women ≥31 years compared to those ≤ 20 years at first childbirth (p for linear trend = 2.93x10-4). Adjustment for BMI and waist attenuated this association as well (HR ≥31 years versus ≤ 20 years 0.70 (0.50-0.99)), although not completely.

Conclusion The association of parity and T2D risk was found to be mediated by increased body mass. Furthermore, age at first full-term pregnancy was inversely associated with the subsequent development of T2D. In this case body mass attenuated the association, but could not fully account for it.

Introduction

Previous studies showed a link between multiparity and future T2D risk [1,2,3], but other studies could not confirm this [4,5,6]. It has been suggested that the observed associations are in fact a result of the increase in body mass associated with child bearing [5,6]. Parity has been associated with larger waist circumference many years after childbearing [7,8] and prospective studies have observed substantial increases in waist circumference with multiple pregnancies [9]. Pregnancy promotes abdominal obesity, which is an increased risk factor for developing insulin resistance [10] that can subsequently lead to T2D [11]. However, the degree to which potential intermediates such as waist and BMI affect the association between parity and T2D remains partly unclear.

Apart from the association through body mass with T2D, another possible biological mechanism for the association between parity and T2D risk in women could be through reduced estrogen exposure. It has been hypothesized that pregnancy permanently resets ovarian function, leading to a reduced lifetime exposure to estrogen [12]. Previously, it was shown that estrogen levels were significantly lower in premenopausal parous women compared to nulliparous women [12,13]. As it has been suggested that high levels of endogenous estrogen protect against T2D in premenopausal women [14], this reduced exposure to estrogen in parous women could affect future T2D risk. Also, it can be argued that the duration of this exposure is important. We therefore hypothesize that the longer the reduced estrogen exposure, due to early childbirth, the higher the risk for T2D may be. As parity is the start of this change in level of exposure to estrogens, age at first full-term pregnancy (AFFTP) is a good marker for the duration of this reduced exposure. Therefore, we assessed the association between both parity and age at first full-term pregnancy with the risk of T2D in the large Prospect-EPIC cohort comprising of 17,357 Dutch women.

Methods

Subjects

The Prospect-EPIC cohort is one of the two Dutch contributions to the European Prospective Investigation into Cancer and Nutrition (EPIC) [15]. It is a prospective

cohort study in 17,357 women aged 49–70 who lived in Utrecht and vicinity and who participated in the breast cancer screening program between 1993 and 1997 [16]. All participants gave their written informed consent and the study was approved by the Institutional Review Board. The design, sampling strategies, and examination techniques of the cohort have been described previously [16].

Data collection

Baseline measurements

At baseline, all participants filled out detailed questionnaires on usual diet, reproductive history, presence of chronic diseases and related potential risk factors. They underwent a brief medical examination and a blood sample was drawn.

Parity and age at first childbirth

Parity was assessed from the question: "how many live born children do you have and what were their birthdays?" The age at first full-term pregnancy was calculated as the interval between the birthday of the mother and the birthday of her first live born child. For analyses, we categorized subjects into five categories for parity (no children, 1 child, 2 children, 3 children and ≥ 4 children) and four categories for age at first full-term pregnancy (≤ 20, 21-25, 26-30, ≥31).

Covariates

Body weight was measured in light indoor clothing without shoes to the nearest 0.5 kg with a floor scale (Seca, Atlanta, GA, USA). Additionally, height, waist and hip circumference were measured. Body mass index (BMI) was calculated as weight divided by height squared (kg/m2).

Alcohol consumption was assessed by a validated food frequency questionnaire (FFQ) [17]. Baseline alcohol intake was determined by multiplying the consumption of each beverage by its ethanol content and was calculated to grams per week (g/week). Subsequently, we categorized subject into four alcohol consumption categories: <0.05 g/week, 0.05-5.5 g/week, 5.5-10.5 g/week, >10.5 g/week.

Duration and types of physical activity during the year preceding study recruitment were assessed by a set of questions that was used in all EPIC cohorts. By combining occupational physical activity with time spent on cycling and sporting in

summer and winter, the validated Cambridge Physical Activity Index (CPAI) was calculated [18]. Based on this index participants were divided in four physical activity categories: inactive, moderately inactive, moderately active and active.
Smoking behaviour was categorized as never, former or current smoking.

To define the socio-economical status, the highest attained level of education of the participants was used and classified into three categories: low (primary education up to completing intermediate vocational education), middle (up to higher secondary education) and high (those with higher vocational education and university).

The number of years of oral contraception use was self reported, and participants were divided into four groups: never, 1-4 years, 4-10 years, >10 years. Self reported gestational diabetes status during pregnancy was indicated as yes or no.

Missing value analyses

Missing values for BMI, waist, alcohol intake, CPAI, smoking, gestational diabetes, socio-economical status, number of years of oral contraceptives use and number of life born children were imputed using multiple imputation [19], which we repeated 5 times to account for uncertainties in imputed data. None of the variables had > 5% missing values; the percentage of missing values ranged from 0.1% for BMI to 2.9% for years of oral contraceptives use.

Morbidity and mortality follow-up

Occurrence of T2D during follow-up was obtained via self-report in two follow-up questionnaires sent to the participants with intervals of three to five years, linkage to the Dutch register of hospital discharge diagnoses (HDD) and a mailed urinary glucose strip test (part of the cohort) (I. Sluijs, Neth J Med, under revision). Potential cases of incident T2D were verified against information from the participants' general practitioner or pharmacist through mailed questionnaires. T2D was defined present when the general practitioner or pharmacist confirmed the diagnosis. Information on vital status was obtained through linkage with the municipal administration registries [20]. Causes of death were obtained from the Dutch Central Bureau of Statistics, coded according to the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10). For our analyses, T2D was the endpoint of interest

and follow up ended at the date of diagnosis or at the date of death. All others were censored on January 1st 2006.

Data analysis

Means and standard deviations (for normally distributed variables) and numbers and frequencies (for categorical variables) were presented.

The person-time for each woman was calculated from birth to the month of diagnosis of the endpoint (T2D), the month of death from other causes, or the end of follow-up (January 1, 2006). Cox proportional hazards regression models were used to estimate the Hazards ratio's (HRs) for the T2D event with 95% confidence intervals (CIs). Age adjusted analysis (model 1) was performed to identify the relation between parity, AFFTP and T2D. Additionally, four multivariate models were used: model 2, including the potential confounders age, alcohol intake, physical activity, smoking, socio-economic status and oral contraceptives use; model 3, including all confounders from model 2 and BMI; model 4, including all confounders from model 2 and waist; model 5, including all confounders from model 2 and gestational diabetes. As BMI, waist and gestational diabetes could be potential intermediates, we studied these variables in separate models. As maternal age at first childbirth could influence parity and therefore possibly affect T2D risk, we additionally adjusted the models containing BMI or waist for parity.

We also performed trend analyses with categorical variables as continuous in the model to test a dose-response effect for parity and AFFTP and the risk for T2D. Results were considered statistically significant at 2-sided $P \leq 0.05$. All statistical analyses were performed using SPSS (PASW Statistics 18).

Results

Table 1 shows the baseline characteristics of the population included in this study. The study had a mean follow-up of 9.1 ± 3.6 years and comprised 157,964 person-years. The mean age of the study group at baseline was 57.1 ± 6.0 years. In total, the study contained 867 verified T2D patients; 332 prevalent cases and 535 incident cases that occurred during the period of follow-up. When calculating follow-up time from birth, this resulted in a mean follow-up of 66.9 ± 6.7 years with a corresponding 1,160,428 person-years.

Table 1, Baseline characteristics of 17,357 women from the PROSPECT cohort

	Mean ± SD
Follow-up time (yr)	9.1 ± 3.5
Follow-up time from birth (yr)	66.9 ± 6.7
Age at intake (yr)	57.1 ± 6.0
BMI (kg/m2)	26.0 ± 4.1
Waist (cm)	83.8 ± 10.2
Alcohol intake (g/week)	9.1 ± 12.6
Pill use (yr)	5.4 ± 6.8
Live born children	2.4 ± 1.5
Age at first full-term pregnancy (yr)	25.1 ± 4.0
	N (%)
Smoking	
Current smoker	3783 (20.0)
Former smoker	5982 (34.7)
Never smoker	7466 (43.3)
Physical activity	
Inactive	1301 (7.5)
Moderate inactive	4612 (26.6)
Moderate active	4437 (25.6)
Active	7007 (40.4)
Educational level	
Low	13311 (76.7)
Medium	1249 (7.2)
High	2270 (13.1)
Gestational diabetes	
No	16714 (96.3)
Yes	643 (3.7)

Table 1, Hazard ratio's for Type 2 Diabetes risk by parity

	HR	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	p trend
Parity	0	1	2	3	≥ 4	
Subjects (n)	2131	1487	6623	4059	2741	
T2D cases (%)	97 (4.4)	60 (4.0)	312 (4.7)	179 (4.4)	219 (7.4%)	
Model 1[a]	1	0.99 (0.72-1.37)	1.23 (0.98-1.54)	1.02 (0.80-1.30)	1.46 (1.15-1.85)	0.01
Model 2[b]	1	0.86 (0.62-1.20)	1.14 (0.90-1.45)	0.97 (0.75-1.25)	1.27 (0.99-1.63)	0.05
Model 3[c]	1	0.84 (0.60-1.17)	1.08 (0.85-1.37)	0.90 (0.70-1.16)	1.04 (0.81-1.33)	0.84
Model 4[d]	1	0.84 (0.61-1.17)	1.10 (0.87-1.39)	0.93 (0.72-1.20)	1.04 (0.81-1.33)	0.83
Model 5[e]	1	0.81 90.58-1.13)	1.07 (0.84-1.36)	0.91 (0.71-1.18)	1.16 (0.91-1.49)	0.19

[a] Model 1 = Adjusted for age at baseline (continuous)

[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), alcohol intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)

[c] Model 3 = Model 2 plus BMI (continuous)

[d] Model 4 = Model 2 plus waist (continuous)

[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

On average, women had 2.4 (± 1.5) live born children (table 1). Compared to women with no children, women with 4 or more children had an increased risk for T2D (HR=1.27; 95% CI 0.99-1.63; ptrend=0.05; table 2, model 2). Adding BMI or waist to the model (table 2, model 3 and model 4) attenuated the association between parity and T2D risk (HR=1.04; 95% CI: 0.81-1.33). Adding gestational diabetes to the multivariate model also diminished the risk for T2D and resulted in a HR of 1.16 [95% CI 0.91-1.49] for women with ≥ 4 children compared to women without children (table 2, model 5).

The mean maternal age at first delivery was 25.1 (± 4.0) years (table 1). In the multivariate adjusted analysis, increasing maternal age was significantly associated with a reduced T2D risk (HR AFFTP ≥31 vs. ≤ 20 years= 0.60; 95% CI 0.44–0.83; ptrend=2.93x10-4, table 3). After additional adjustment for BMI or waist (table 3, model 3 and model 4 respectively) the association attenuated but remained statistically significant. The BMI adjusted HR of maternal age ≥31 years versus ≤ 20 years was 0.74 [95% CI 0.54-1.03; ptrend=0.04] and the waist adjusted HR of maternal age ≥31 years versus ≤ 20 years was 0.72 [95% CI 0.52-0.99; ptrend=0.03]. Adding gestational diabetes did not change these results (HR=0.56 [95% CI 0.40-0.78; table 3, model 5).

As maternal age at first childbirth could influence parity and therefore possibly affect T2D risk, we additionally adjusted the analyses for AFFTP for parity (table 3, model 6 and model 7). This did not substantially influence the results.

Table 2, Hazard ratio's for Type 2 Diabetes risk by maternal age at first childbirth in 13,396 women

	HR	HR (95% CI)	HR (95% CI)	HR (95% CI)	p trend
Maternal age at first childbirth (years)	≤ 20	21- 25	26- 30	≥ 31	
Subjects (n)	1,531	7196	4893	1311	
T2D cases (%)	116 (7.6)	366 (5.1)	220 (4.5)	59 (4.5)	
Model 1[a]	1	0.63 (0.51-0.77)	0.50 (0.40-0.63)	0.48 (0.35-0.65)	2.78×10^{-8}
Model 2[b]	1	0.74 (0.60-0.91)	0.65 (0.51-0.81)	0.60 (0.44-0.83)	2.93×10^{-4}
Model 3[c]	1	0.81 (0.66-1.00)	0.77 (0.61-0.97)	0.74 (0.54-1.03)	0.04
Model 4[d]	1	0.81 (0.65-1.00)	0.77 (0.61-0.97)	0.72 (0.52-0.99)	0.03
Model 5[e]	1	0.73 (0.59-0.90)	0.62 (0.49-0.78)	0.56 (0.40-0.78)	3.77×10^{-5}
Model 6[f]	1	0.81 (0.66-1.01)	0.77 (0.61-0.98)	0.74 (0.53-1.04)	0.05
Model 7[g]	1	0.80 (0.65-0.99)	0.76 (0.60-0.97)	0.70 (0.50-0.99)	0.03

[a] Model 1 = Adjusted for age at baseline (continuous)

[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), alcohol intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)

[c] Model 3 = Model 2 plus BMI (continuous)

[d] Model 4 = Model 2 plus waist (continuous)

[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

[f] Model 6 = Model 2 plus BMI (continuous) and parity (no children, 1 child, 2 children, 3 children and ≥ 4 children)

[g] Model 7 = Model 2 plus waist (continuous) and parity (no children, 1 child, 2 children, 3 children and ≥ 4 children)

Discussion

In this large cohort of 17,357 Dutch women, multivariate adjusted parity and age at first full-term pregnancy were statistical significantly associated with the risk of T2D. The increased risk of T2D associated with higher parity was found to be mediated by increased body fat, as adjusting for BMI as well as for waist completely attenuated the association. Age at first full-term pregnancy was inversely associated with the subsequent development of T2D. In this case body fat did attenuate the association, but could not fully account for it. To our knowledge, this is the first study to report that having your first child at a later age decreases the risk for T2D.

Before interpreting the data, some strengths and limitations need to be discussed. The main advantages of this study are its prospective nature, the long follow-up time and the large sample size. Furthermore, potential cases of T2D were verified by the participants' general practitioner or pharmacist and T2D was only defined present when one of them confirmed the diagnosis. In this study, the person-time for each woman was calculated from birth to the month of T2D diagnosis, to the month of death from other causes, or the end of follow-up. In studies with prospective cohorts, the person-time is usually calculated from baseline, therefore only including incident T2D cases. As our outcome variables 'parity' and 'age at first childbirth' were almost always established before T2D disease onset, we choose to calculate person-time from birth. This allowed us to include an extra 332 prevalent T2D cases for analyses, which would otherwise have been excluded. We additionally studied the association of parity and age at first childbirth and T2D risk with person-time calculated from baseline, to explore the effect of using a person-time from baseline versus from birth. The results were comparable (data not shown) with similar HR, but wider 95% CI.

The increased risk for T2D due to the number of children appears to be mediated by increased body fat associated with past child bearing, a finding which is corroborated by other studies [5,6]. Most of the previous studies that did find an association between multiparity and T2D risk did not adjust for potential confounding by body fat [5]. However, the Atherosclerosis Risk in Communities study reported that grandmultiparity (having $\geq$ 5 children) increases the risk for T2D, even after adjustment for BMI and waist [3].

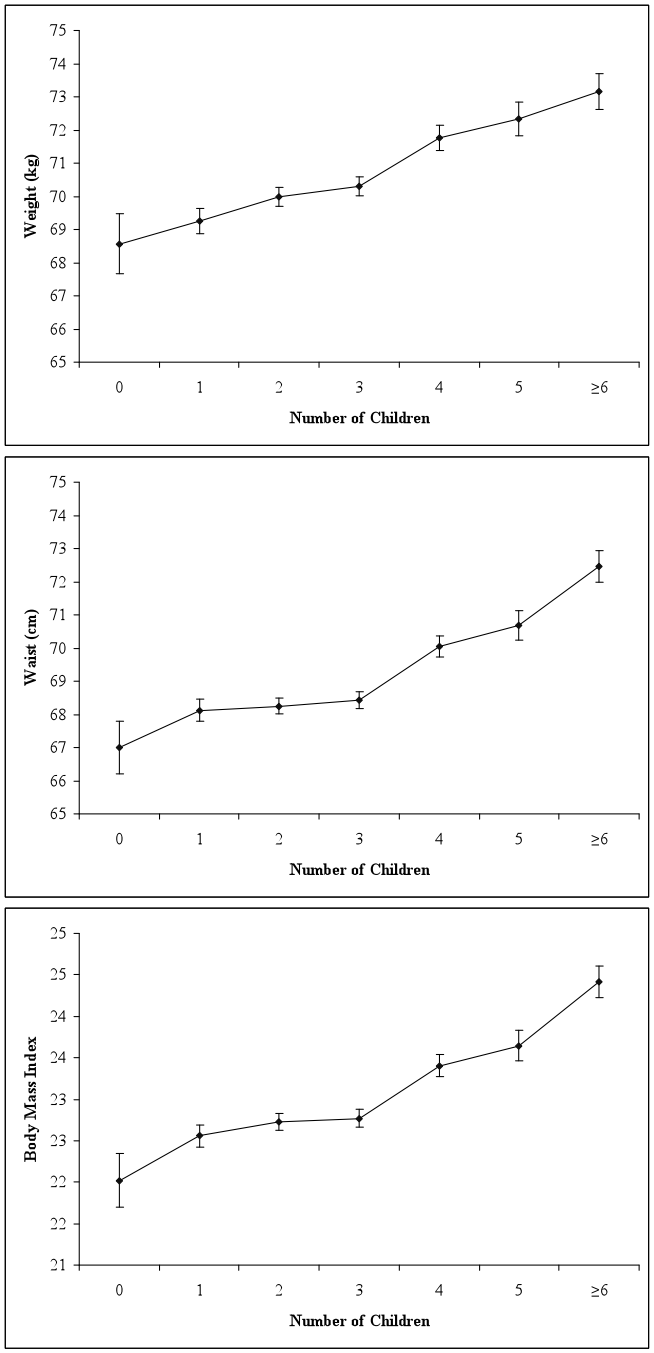


Figure 1. Body mass in 17,357 Dutch women aged 49-70 by number of live born children; age-adjusted weight (1A), age-adjusted waist (1B) and age-adjusted BMI (1C)

In our study we found an increased weight, BMI and waist circumference with each additional child, as is shown in figure 1. Pregnancy is a time in most women's lifes where substantial weight is gained. The incidence of major weight gain in adult life is highest in persons aged 25 to 34 years and is twice as high in women (men, 3.9%; women, 8.4%) [21]. Several long-term prospective studies show that gestational weight gain is associated with increased maternal weight decades after pregnancy [22,23]. In another perspective study it was found that childbearing women showed a threefold greater increase in visceral fat deposition from preconception to postpartum compared to women not bearing children [24]. This extra adipose tissue deposition during pregnancy serves as a nutritional reserve to ensure an adequate energy supply to the newborn. This is supported by findings that breastfeeding reduces postpartum weight retention [25]. Anthropological studies on weaning age of infants suggest a mean weaning period of 2.8 years in hunter-gatherer population [26,27]. Among 14,929 Prospect women, the average period for breastfeeding their first child was 10.66 (± 11.51) weeks. This is a 13-fold decrease compared to hunter gather women. Possibly the body mass increase per child is in part caused by the difference between adipose tissue that is prepared for by the human body for long-term breastfeeding, and the adipose tissue that is actually used for breastfeeding. This extra abdominal fat is nowadays an increased risk factor for developing insulin resistance and subsequently T2D [10,11].

This is the first study to show that having the first child at a later age decreases the risk for T2D. In 1992, Manson et al. studied the association of AFFTP and T2D risk in the Nurses Health Study (NHS) [5]. The authors did not find an age- and BMI-adjusted association between AFFTP and T2D risk. These results are in discordance with our study. This discrepancy could be due to age at baseline differences as women from the NHS were between the age of 30 and 55 years at baseline compared to the women of the Prospect-EPIC cohort who were aged 49 to 70 at baseline. It is known that older women have the highest prevalence of T2D and are more likely to be post-menopausal and therefore less protected from T2D by ovarian estrogens [28].

It has been suggested that high levels of endogenous estrogens protect against T2D in humans, although evidence is thus far conflicting. The overall prevalence of T2D is lower in premenopausal women compared to men, a trend that is reversed after menopause [14]. Additionally, results from previous oral estrogen therapy trials showed a lower risk for T2D among post-menopausal women who used estrogen

treatment [29,30,31]. On the contrary, studies on endogenous postmenopausal estrogen levels have found that higher levels are associated with increased T2D risk [32,33]. However, postmenopausal estrogen levels depend on peripheral aromatization of androgens, and do likely not reflect premenopausal ovarian production.

One of the possible biological mechanisms for the association between young age at first childbirth and T2D in women could be that pregnancy permanently resets ovarian function, leading to a reduced lifetime exposure to estrogens [12]. As a late age at first full-term pregnancy results in a shorter period of the reduced exposure to estrogens, this could then relatively protect against the development of T2D. Data from animal studies suggest that the ovarian estrogen 17 β-estradiol maintains β-cell function, by protecting β-cells from apoptosis, as (i) females from T2D mouse models are protected from β-cell death and hyperglycemia [34] and (ii) it was shown that estrogens protect pancreatic β-cells from apoptosis and thus preventing insulin-deficient diabetes in mice [35].

Although it has often been assumed that β-cell deterioration occurs only late in the course of development of T2D with insulin resistance occurring long before the onset of beta cell function, recent studies conclude otherwise. In studies in which hyperglycaemic clamps were performed in nondiabetic subjects it was found that β-cell function is lower in first-degree relatives of diabetes patients, even before the onset of impairments in glucose tolerance [36,37,38]. However, whether long-term reduced oestrogen exposure in women may relate to alterations in β-cell function of women with a young AFFTP before menopause is unknown and should be studied further.

The hypothesis that young age at first full-term pregnancy leads to long-term reduced estrogens exposure, assumes that pregnancy leads to permanent changes in the hormonal profile of parous women. However, although we know that pregnancy exposes the body to an altered hormone profile, it is unclear whether this indeed leads to permanent changes [39]. Estradiol levels were decreased in parous versus non-parous women in two studies [12,13] but unchanged in another small study [40]. The problem of assessing different hormones at different stages of the reproductive cycle makes it difficult to define a specific altered hormone profile for parous women.

Conclusion

In women, parity and a younger AFFTP increased the risk of T2D. The association of parity and T2D risk was found to be mediated by increased body fat. We show an increased weight, BMI and waist with each additional child. Adjustment for body mass attenuated the inverse association of AFFTP and T2D but could not fully account for it. A possible underlying mechanism could be that young age at first full-term pregnancy leads to long-term reduced estrogen exposure subsequently leading to reduced β-cell function. However, this mechanism should be further studied in future research.

References

1. Beral V (1985) Long term effects of childbearing on health. J Epidemiol Community Health 39: 343-346.
2. Gunderson EP, Jacobs DR, Jr., Chiang V, Lewis CE, Tsai A, et al. (2009) Childbearing is associated with higher incidence of the metabolic syndrome among women of reproductive age controlling for measurements before pregnancy: the CARDIA study. Am J Obstet Gynecol 201: 177 e171-179.
3. Nicholson WK, Asao K, Brancati F, Coresh J, Pankow JS, et al. (2006) Parity and risk of type 2 diabetes: the Atherosclerosis Risk in Communities Study. Diabetes Care 29: 2349-2354.
4. Gunderson EP, Lewis CE, Tsai AL, Chiang V, Carnethon M, et al. (2007) A 20-year prospective study of childbearing and incidence of diabetes in young women, controlling for glycemia before conception: the Coronary Artery Risk Development in Young Adults (CARDIA) Study. Diabetes 56: 2990-2996.
5. Manson JE, Rimm EB, Colditz GA, Stampfer MJ, Willett WC, et al. (1992) Parity and incidence of non-insulin-dependent diabetes mellitus. Am J Med 93: 13-18.
6. Fowler-Brown AG, de Boer IH, Catov JM, Carnethon MR, Kamineni A, et al. (2010) Parity and the Association with Diabetes in Older Women. Diabetes Care.
7. Bjorkelund C, Lissner L, Andersson S, Lapidus L, Bengtsson C (1996) Reproductive history in relation to relative weight and fat distribution. Int J Obes Relat Metab Disord 20: 213-219.
8. Troisi RJ, Wolf AM, Mason JE, Klingler KM, Colditz GA (1995) Relation of body fat distribution to reproductive factors in pre- and postmenopausal women. Obes Res 3: 143-151.
9. Gunderson EP, Murtaugh MA, Lewis CE, Quesenberry CP, West DS, et al. (2004) Excess gains in weight and waist circumference associated with childbearing: The Coronary Artery Risk Development in Young Adults Study (CARDIA). Int J Obes Relat Metab Disord 28: 525-535.
10. Pischon T, Boeing H, Hoffmann K, Bergmann M, Schulze MB, et al. (2008) General and abdominal adiposity and risk of death in Europe. N Engl J Med 359: 2105-2120.
11. Kahn SE, Hull RL, Utzschneider KM (2006) Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature 444: 840-846.
12. Bernstein L, Pike MC, Ross RK, Judd HL, Brown JB, et al. (1985) Estrogen and sex hormone-binding globulin levels in nulliparous and parous women. J Natl Cancer Inst 74: 741-745.
13. Dorgan JF, Reichman ME, Judd JT, Brown C, Longcope C, et al. (1995) Relationships of age and reproductive characteristics with plasma estrogens and androgens in premenopausal women. Cancer Epidemiol Biomarkers Prev 4: 381-386.
14. Wild S, Roglic G, Green A, Sicree R, King H (2004) Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care 27: 1047-1053.
15. Riboli E (1992) Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 3: 783-791.
16. Boker LK, van Noord PA, van der Schouw YT, Koot NV, Bueno de Mesquita HB, et al. (2001) Prospect-EPIC Utrecht: study design and characteristics of the cohort population. European Prospective Investigation into Cancer and Nutrition. Eur J Epidemiol 17: 1047-1053.
17. Ocke MC, Bueno-de-Mesquita HB, Pols MA, Smit HA, van Staveren WA, et al. (1997) The Dutch EPIC food frequency questionnaire. II. Relative validity and reproducibility for nutrients. Int J Epidemiol 26 Suppl 1: S49-58.
18. Wareham NJ, Jakes RW, Rennie KL, Schuit J, Mitchell J, et al. (2003) Validity and repeatability of a simple index derived from the short physical activity questionnaire used in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Public Health Nutr 6: 407-413.
19. van der Heijden GJ, Donders AR, Stijnen T, Moons KG (2006) Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 59: 1102-1109.
20. Herings RM, Bakker A, Stricker BH, Nap G (1992) Pharmaco-morbidity linkage: a feasibility study comparing morbidity in two pharmacy based exposure cohorts. J Epidemiol Community Health 46: 136-140.
21. Williamson DF, Kahn HS, Remington PL, Anda RF (1990) The 10-year incidence of overweight and major weight gain in US adults. Arch Intern Med 150: 665-672.
22. Rooney BL, Schauberger CW (2002) Excess pregnancy weight gain and long-term obesity: one decade later. Obstet Gynecol 100: 245-252.
23. Mamun AA, Kinarivala M, O'Callaghan MJ, Williams GM, Najman JM, et al. (2010) Associations of excess weight gain during pregnancy with long-term maternal overweight and obesity: evidence from 21 y postpartum follow-up. Am J Clin Nutr 91: 1336-1341.

24. Gunderson EP, Sternfeld B, Wellons MF, Whitmer RA, Chiang V, et al. (2008) Childbearing may increase visceral adipose tissue independent of overall increase in body fat. Obesity (Silver Spring) 16: 1078-1084.
25. Baker JL, Gamborg M, Heitmann BL, Lissner L, Sorensen TI, et al. (2008) Breastfeeding reduces postpartum weight retention. Am J Clin Nutr 88: 1543-1551.
26. Alvarez HP (2000) Grandmother hypothesis and primate life histories. Am J Phys Anthropol 113: 435-450.
27. Sellen DW (2007) Evolution of infant and young child feeding: implications for contemporary public health. Annu Rev Nutr 27: 123-148.
28. Szmuilowicz ED, Stuenkel CA, Seely EW (2009) Influence of menopause on diabetes and diabetes risk. Nat Rev Endocrinol 5: 553-558.
29. Kanaya AM, Herrington D, Vittinghoff E, Lin F, Grady D, et al. (2003) Glycemic effects of postmenopausal hormone therapy: the Heart and Estrogen/progestin Replacement Study. A randomized, double-blind, placebo-controlled trial. Ann Intern Med 138: 1-9.
30. Bonds DE, Lasser N, Qi L, Brzyski R, Caan B, et al. (2006) The effect of conjugated equine oestrogen on diabetes incidence: the Women's Health Initiative randomised trial. Diabetologia 49: 459-468.
31. Margolis KL, Bonds DE, Rodabough RJ, Tinker L, Phillips LS, et al. (2004) Effect of oestrogen plus progestin on the incidence of diabetes in postmenopausal women: results from the Women's Health Initiative Hormone Trial. Diabetologia 47: 1175-1187.
32. Ding EL, Song Y, Malik VS, Liu S (2006) Sex differences of endogenous sex hormones and risk of type 2 diabetes: a systematic review and meta-analysis. JAMA 295: 1288-1299.
33. Ding EL, Song Y, Manson JE, Rifai N, Buring JE, et al. (2007) Plasma sex steroid hormones and risk of developing type 2 diabetes in women: a prospective study. Diabetologia 50: 2076-2084.
34. Louet JF, LeMay C, Mauvais-Jarvis F (2004) Antidiabetic actions of estrogen: insight from human and genetic mouse models. Curr Atheroscler Rep 6: 180-185.
35. Le May C, Chu K, Hu M, Ortega CS, Simpson ER, et al. (2006) Estrogens protect pancreatic beta-cells from apoptosis and prevent insulin-deficient diabetes mellitus in mice. Proc Natl Acad Sci U S A 103: 9232-9237.
36. van Haeften TW, Pimenta W, Mitrakou A, Korytkowski M, Jenssen T, et al. (2002) Disturbances in beta-cell function in impaired fasting glycemia. Diabetes 51 Suppl 1: S265-270.
37. Bonadonna RC, Stumvoll M, Fritsche A, Muggeo M, Haring H, et al. (2003) Altered homeostatic adaptation of first- and second-phase beta-cell secretion in the offspring of patients with type 2 diabetes: studies with a minimal model to assess beta-cell function. Diabetes 52: 470-480.
38. van Haeften TW, Pimenta W, Mitrakou A, Korytkowski M, Jenssen T, et al. (2000) Relative conributions of beta-cell function and tissue insulin sensitivity to fasting and postglucose-load glycemia. Metabolism 49: 1318-1325.
39. Britt K, Ashworth A, Smalley M (2007) Pregnancy and the risk of breast cancer. Endocr Relat Cancer 14: 907-933.
40. Musey VC, Collins DC, Brogan DR, Santos VR, Musey PI, et al. (1987) Long term effects of a first pregnancy on the hormonal environment: estrogens and androgens. J Clin Endocrinol Metab 64: 111-118.

Chapter 8

The thrifty genes hypothesis is not supported by evolutionary analysis in Europeans

Clara C. Elbers, N. Charlotte Onland-Moret, Alexandra Zhernakova, Renske A. Kuiper, Yvonne T. van der Schouw, Cisca Wijmenga, Carolien G. F. de Kovel

Abstract

Type 2 diabetes (T2D) and obesity have a strong genetic basis and they are quite common despite their negative effects on human health. Like all species, Homo sapiens has been shaped by evolutionary processes and the worldwide incidence of T2D and obesity suggests that genes underlying these traits may have been favored by the process of natural selection. In trying to explain this observation, James Neel proposed the 'thrifty genotype theory', according to which our early ancestors frequently suffered periods of prolonged famine, during which a survival advantage would have been conferred by genes favouring the economical use and storage of energy, the so-called thrifty genes.

Genome-wide association studies have identified 19 T2D and 17 obesity susceptibility alleles. In combination with genome-wide single nucleotide polymorphism data, this allows us to test the thrifty genes hypothesis. We examined whether these susceptibility variants have been under positive selection in 3,657 Dutch and UK individuals and 1,301 HapMap III samples.

Our results do not support the thrifty gene hypothesis as we found no signs of positive selection around the known T2D and obesity risk alleles. However, some protective variants for T2D and obesity do show suggestive signs of positive selection in our European data and it can be argued that Europeans have already adapted genetically to a Western diet. The lower frequency of T2D in Europeans compared to other ethnic groups which are now adopting a 'Westernized' diet and lifestyle supports this hypothesis.

Introduction

Although type 2 diabetes (T2D) is a severe medical condition, it is quite common, with a prevalence of 2.8% worldwide and a prevalence of 8% in the USA (according to the American Diabetes Association). The concordance rate of T2D in monozygotic twins is 76% compared to 40% in dizygotic twins, providing convincing evidence that genetic factors contribute to the development of T2D. One interesting question is why the phenotypes of T2D and obesity, the main risk factor for T2D, are so common despite their negative effects on human health. Like all species, *Homo sapiens* has been shaped by evolutionary processes and the fact that so many people are susceptible to developing T2D and obesity suggests that genes underlying these traits may have been favored by the process of natural selection. In trying to explain this observation, James Neel proposed the 'thrifty genotype theory' in 1962 [1], according to which our early ancestors frequently suffered periods of prolonged famine, during which a survival and/or reproductive advantage would have been conferred by genes favouring the economical use and storage of energy, the so-called thrifty genes. The theory focuses on the efficient use of glucose as a biological fuel and suggests that evolutionary pressure to preserve glucose for use by the brain during starvation led to a genetic propensity towards insulin resistance in peripheral tissue. In the Western world, food is, in general, easily available and plentiful, so these thrifty genes are maladaptive in modern society and may now contribute to the widespread susceptibility for T2D and obesity [1].

Although the thrifty genes hypothesis is popular and frequently cited, it is also controversial and has been discussed for decades in many scientific papers [2]. The most powerful argument against the theory came from anthropological studies that suggested that during the past 2.5 million years of human history, famines were not sufficiently frequent and severe to cause evolutionary pressure [3-4]. There are several alternative theories explaining the genetic basis of T2D and obesity [2]. Till now, all these hypotheses have been speculative and there was no consensus in the field.

However, the 19 T2D susceptibility alleles [5-11] and 17 obesity susceptibility alleles [12-16], in combination with genome-wide single nucleotide polymorphism (SNP) data, now allows us to investigate whether genetic variants underlying the traits have been favored by positive natural selection. When a genetic variant is under

positive selection, it increases in frequency in a population and this leaves a 'signature' or pattern in the human genome. These signatures can be identified by comparison with the background distribution of genetic variation in humans, which is generally argued to have evolved largely under neutrality [17-18].

We tested the thrifty gene hypothesis by investigating whether recently identified T2D and obesity risk alleles have been under recent positive selection in four large study populations and 11 HapMap III populations. We did this by studying haplotype lengths and allele frequency differences between populations in multiple genome-wide SNP data sets, encompassing 3,657 individuals for the UK and the Netherlands, and 1,301HapMap III individuals.

Results

We studied all the T2D and obesity susceptibility loci known up to March 2009 for signs of recent selection. The 19 T2D and 17 obesity variants are SNPs that were found to be associated with genome-wide significance and were replicated in other independent populations (within the original study). For the analysis we used 3,657 individuals from UK and Dutch populations. We investigated signatures of selection by studying haplotype lengths and allele frequency differences between populations. The results were highly reproducible across all the populations tested for both alleles with unusually long haplotypes (table 3) as well as other haplotypes (data not shown).

iHS analysis

In the first stage of the study we analysed a region of 1 Mb around the susceptibility loci in the 1958BC and in the second stage we replicated the most promising results ($p<0.1$) in multiple data sets encompassing 2,215 individuals from Europe. The p-values presented are empirical (2-sided) p-values. A locus with a p-value below 0.05 means that the locus is a 5% outlier compared to a set of 8,500 randomly chosen background SNPs. We excluded a susceptibility SNP near to the *NOTCH2* gene from the analysis because the locus was located too close to the centromere to calculate a reliable iHS value. Another SNP, near the *PTER* gene, was excluded from analysis, because its major allele frequency was above 95% and the iHs method has low power to detect selective sweeps that have reached a frequency above 95%.

Of the 19 studied T2D susceptibility loci, the non-risk alleles near *THADA, PPARG, JAZF1, SLC30A8,* and the risk allele near *CDC123-CAMK1D*, were positioned on an unusually long haploblock compared to the background allele at the same position ($p < 0.1$) in the 1958BC data set (table 1). This suggests that the protective alleles of rs7578597 near *THADA*, rs17036101 near *PPARG* and rs864745 near *JAZF1*, and the risk allele near *CDC123-CAMK1D*, show signatures of positive selection.

Table 1. Recent selection for T2D susceptibility loci in 1,442 individuals from the 1958 Birth Cohort

SNP ID	Chr	Closest gene(s)	Risk allele	Derived allele freq	iHScorr 1958BC	*p* value	Allele under (suggestive) selection
rs10923931	1	*NOTCH2*	T	0.88	NA [a]	NA [a]	
rs7578597	2	*THADA*	T*	0.10	-1.46	**0.10**	non-risk allele
rs17036101	3	*SYNC-PPARG*	G	0.94	3.54	**0.002**	non-risk allele
rs4402960	3	*IGF2BP2*	T	0.33	1.60	0.14	
rs4607103	3	*ADAMTS9*	C*	0.22	-0.16	0.89	
rs10010131	4	*WFS1*	G*	0.41	1.05	0.35	
rs10946398	6	*CDKAL1*	C*	0.46	-0.18	0.87	
rs864745	7	*JAZF1*	T*	0.47	-1.60	**0.07**	non-risk allele
rs13266634	8	*SLC30A8*	C*	0.31	-1.85	**0.04**	non-risk allele
rs10811661	9	*CDKN2A-2B*	T	0.84	-0.22	0.85	
rs1111875	10	*HHEX-IDE*	T*	0.59	0.74	0.49	
rs12779790	10	*CDC123-CAMK1D*	G	0.17	-1.62	**0.06**	risk allele
rs7901695	10	*TCF7L2*	C*	0.67	0.06	0.98	
rs10830963	11	*MTNR1B*	G	0.29	-0.04	0.98	
rs2237892	11	*KCNQ1*	C*	0.08	-0.50	0.65	
rs5215	11	*KCNJ11*	C	0.36	-0.86	0.38	
rs1153188	12	*DCD*	A*	0.75	-0.14	0.90	
rs7961581	12	*TSPAN8*	C	0.28	-0.85	0.40	
rs4430796	17	*TCF2*	A	0.57	0.83	0.45	

iHScorr, corrected iHs value; 1958BC, 1958 Birth Cohort; * ancestral allele.
[a] SNP rs10923931 near to the NOTCH2 gene was located too close to the centromere to calculate a reliable iHS value.

Of the 17 studied obesity susceptibility loci, three non-risk alleles near the genes *SEC16B, TMEM18* and *BDNF* were surrounded by an extended haplotype ($p<0.1$, table 2), implying that the protective alleles showed signs of selection.
Finding 5 out of the 19 SNPs in the $p<0.10$ region (instead of the expected 2) has a probability of 0.008, while finding 3 out of 17 in this region has a probability of 0.08. Combined (8 out of 36) has a probability of 0.008. This indicates that we can reasonably assume a selection pressure for at least some of these alleles.

In the replication stage, we observed similar iHS values across three different European study populations (table 3), indicating that the findings in the 1958BC were unlikely to be due to sampling oddities.

Table 2. Recent selection for obesity susceptibility loci in 1,442 individuals from the 1958 Birth Cohort.

SNP ID	Chr	Closest gene(s)	Risk allele	Derived allele frq	iHScorr 1958BC	*p* value	Allele under (suggestive) selection
rs10913469	1	*SEC16B*	A	0.79	1.99	**0.06**	non-risk allele
rs2815752	1	*NEGR1*	A	0.60	-0.23	0.54	
rs6548238	2	*TMEM18*	C*	0.09	-1.46	**0.09**	non-risk allele
rs7647305	3	*ETV5*	C	0.80	0.19	0.84	
rs10938397	4	*GNPDA2*	G	0.43	-0.87	0.38	
rs2844479	6	*NCR3*	T*	0.48	-0.40	0.72	
rs4712652	6	*PRL*	A*	0.43	0.73	0.50	
rs10508503	10	*PTER*	C*	<0.05	NA [a]	NA [a]	
rs10838738	11	*MTCH2*	G	0.31	0.20	0.84	
rs6265	11	*BDNF*	G*	0.91	-1.95	**0.03**	non-risk allele
rs7138803	12	*BCDIN3D*	A	0.36	-0.35	0.76	
rs1424233	16	*MAF*	A*	0.54	-0.61	0.56	
rs7498665	16	*SH2B1*	G*	0.60	0.90	0.39	
rs9939609	16	*FTO*	A*	0.60	1.64	0.13	
rs17782313	18	*MC4R*	C	0.23	-0.29	0.80	
rs1805081	18	*NPC1*	A*	0.41	-0.47	0.68	
rs11084753	19	*KCTD15*	G*	0.31	-0.39	0.72	

iHScorr, corrected iHs value; 1958BC, 1958 Birth Cohort; * ancestral allele.
[a]The derived allele frequency of SNP rs10508503 near to the PTER gene was too low (<0.05) to calculate a reliable iHS value.

Table 3. Follow-up of the iHS analysis in three European populations.

SNP	chr	Closest gene(s)	Risk allele	Allele under selection	778 CD cases UK		508 CD cases NL		929 control NL	
					iHScorr	*p* value	1HScorr	*p* value	1HScorr	*p* value
rs7578597	2	*THADA*	T	T	-1.61	0.07	-1.31	0.13	-1.30	0.13
rs10736101	3	*SYNC-PPARG*	G	G	3.13	0.01	3.23	0.01	3.31	0.004
rs864745	7	*JAZF1*	T	T	-1.57	0.07	-1.54	0.08	-1.52	0.08
rs13266634	8	*SLC30A8*	C	C	-1.87	0.04	-2.o2	0.02	-1.94	0.03
rs12779790	10	*CDC123-CAMK1D*	G	A	-1.61	0.07	-1.54	0.08	-1.56	0.07
rs10913469	1	*SEC16B*	A	A	1.71	0.11	2.45	0.03	2.45	0.03
rs6548238	2	*TMEM18*	C	C	-1.60	0.07	-1.47	0.09	-1.46	0.09
rs6265	11	*BDNF*	G	G	-1.80	0.04	-`.46	0.09	-1.70	0.05

iHScorr, corrected iHS value; CD celiac disease; UK, United Kingdom; NL, the Netherlands

Fst analysis

We studied whether the Fst values of the susceptibility alleles between the HapMap III population were outliers in contrast to an empirical genome-wide distribution of Fst (figures 1 and 2) [19]. An Fst value of 0 means that different populations are identical in allele frequency, whereas an Fst value of 1 means that different populations are fixed for different alleles.

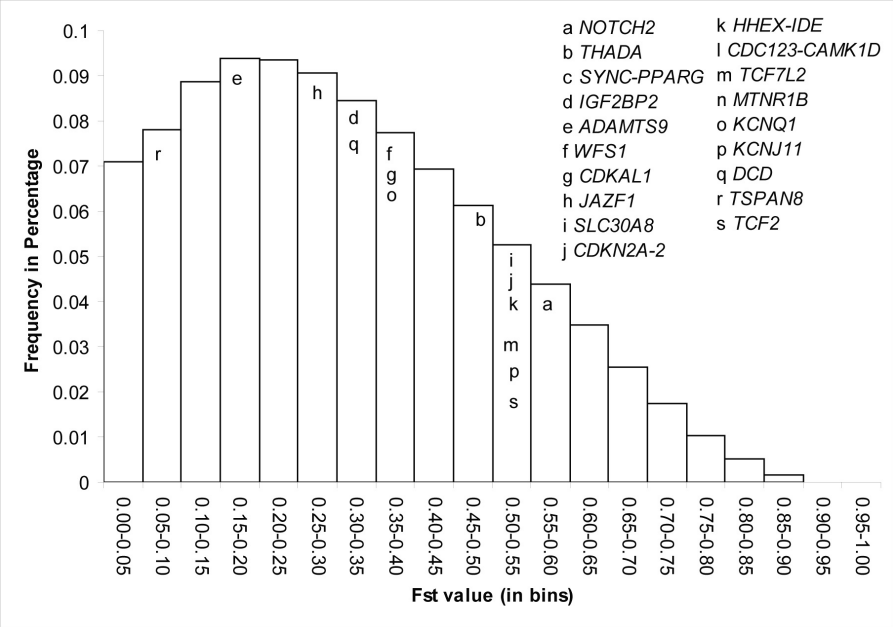


Figure 1. Genome-wide distribution of F_{st} values between HapMap III populations. F_{st} values for T2D susceptibility alleles are indicated in the relevant bars. For *SYNC-PPARG, CDC123-CAMK1D* and *MTNR1B,* the susceptibility alleles were not available for all HapMap III populations and therefore the F_{st} value could not be calculated globally.

None of the studied SNPs had an Fst value that was a significant outlier compared to a genome-wide distribution. Only rs6265 nearby the *BDNF* gene had an Fst value of 0.645 that was located in the 10% outlier region of the genome-wide Fst distribution. The population of African origin all had ancestral risk-allele frequencies of rs6265 above 95%, the Native American populations both had an ancestral allele frequency of 83%, European populations showed allele frequencies between 75% and 80% and the Eastern Asia populations all had allele frequencies of the ancestral allele below 65% (figure 3). Furthermore, we observed no significant allele frequency

differences across different populations for the associated tagSNPs that would have indicated any signs of differential selection.

For *TMEM18, GNPDA2, PRL, FTO, SYNC-PPARG, CDC123-CAMK1D, MTNR1B*, the susceptibility alleles were not available for all HapMap III populations. Therefore, the global Fst could not be calculated for these SNPs. However, Fst values between separate population groups could be calculated for these SNPs and showed no significant differential allele frequency (data not shown).

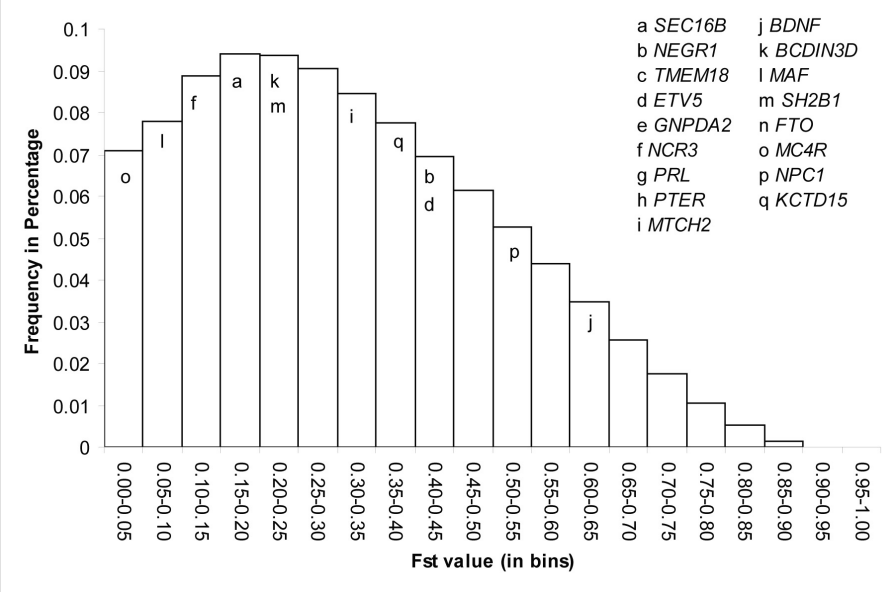


Figure 2. Genome-wide distribution of global F_{st} values between HapMap III populations.
F_{st} values for obesity susceptibility alleles are indicated in the relevant bars. For *TMEM18, GNPDA2, PRL, FTO* the susceptibility alleles were not available for all HapMap III populations and therefore the F_{st} value could not be calculated globally.

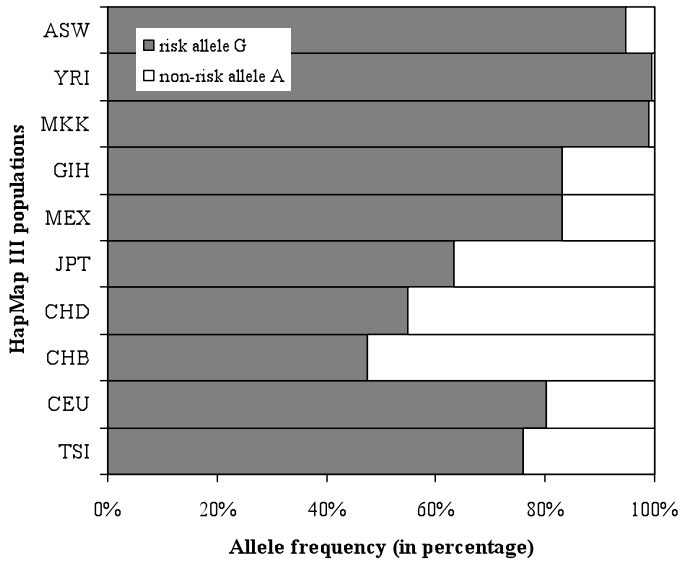


Figure 3. Allele frequency differences for rs6265 nearby the BDNF gene of the ancestral risk allele G and the derived non-risk allele A in 10 HapMap III populations. ASW, African ancestry from Southwest USA; YRI, Yoruba from Ibadan, Nigeria; MKK, Maasai in Kinyawa, Kenya; GIH, Gujarati Indians from Houston; MEX, Mexican ancestry from Los Angeles; JPT, Japanese from the Tokyo area; CHD, Chinese from Denver; CHB, Han Chinese from Beijing; CEU, Northern and Western European ancestry; TSI, Toscans from Italy. Frequency data of rs6265 was not available for Luhya in Webuye, Kenya (LWK).

Discussion

This paper describes a comprehensive analysis of signatures for recent selection on T2D and obesity susceptibility loci in four large study populations. The majority of the variants do not show any signature of selection, although some protective variants for T2D and obesity show signs of positive selection in our European data. Our results therefore do not support the thrifty gene hypothesis, as we found no signs of positive selection around the known T2D and obesity risk alleles.

However, for the T2D susceptibility loci in or near *THADA, PPARG, JAZF1, SLC30A8, and* the obesity susceptibility loci in or near *SEC16B, TMEM18* and *BDNF*, we did find (suggestive) signatures of selection, not for the risk allele but for the protective allele. This suggests that the risk allele was, in fact, the subject of recent negative selection in the European populations. In addition, we show that the T2D and obesity risk alleles are no more differentiated in the HapMap phase III population than random SNPs in the genome. All these findings argue against the

thrifty genes hypothesis [20-21]. Only one risk allele for T2D, nearby *CDC123-CAMK1D*, was found to be suggestive for being under positive selection.

Although the evolutionary analyses performed in this study argue against the thrifty genes hypothesis, our data is not complete enough to reject the theory entirely. Firstly, although the GWA studies have improved our understanding of the genetic basis of T2D and obesity, we can still only explain around 10% of the genetic risk for these traits [22]. Thus, the majority of T2D and obesity loci are still unknown and cannot be tested for signatures of selection. It is worth remembering that GWAS studies do not capture information about rare SNPs or other genetic variants, like copy number variations and inversion/deletion variants.

Secondly, 'thrifty genes' that cause susceptibility for T2D and obesity could have reached fixation in the population (i.e. all individuals of the population carry the same risk allele), and they therefore cannot be picked up by genome-wide association studies using case and control data. Older selection pressure could also have acted on these variants so that their signature is no longer visible in the genome.

Thirdly, if an associated tag SNPs does not shown a signature of selection, it does not necessarily mean that that the causal variant will not show a signature of selection. Although it is likely that the tag SNP and the causal SNP are located on the same haplotype (and therefore show similar iHs values), the fact that the allele frequencies of the associated tag SNPs are not differentially distributed among different populations does not mean that the causal variants are not significantly differentially distributed. The recent GWA studies were mainly focussed on populations of European ancestry and there are large differences in correlation between SNPs among populations. Although the causal variant is likely to be in LD with the associated SNP in the European populations, the LD of causal variants with the associated SNP in other populations might be much weaker. Therefore, a difference in allele frequency distribution of tag SNPs does not necessarily reflect the allele frequency distribution of the causal SNP. To deal with this problem, instead of studying the Fst of associated SNPs, Pickrell et al. calculated the maximum Fst in a 100 kb window surrounding the susceptibility loci for T2D in the Human Genome Diversity CEPH Panel (HGPD) [23]. They observed that regions encompassing T2D susceptibility alleles can significantly differentiate Europeans and East Asians from Africans, with the regions surrounding *TCF7L2, TSPAN8, JAZF1* and *ADAMTS9* as strong outliers. These findings suggest that these regions have experienced recent

positive selection, but how this relates to the thrifty genes hypothesis is unclear. Therefore, in our study, we chose to analyse only tag SNPs instead of a broader genomic region for population allele frequency differentiation. In order to test the thrifty genes hypothesis, we were especially interested in whether the risk allele or the protective allele showed signatures of selection. The direction of the selection is very important in this case and when the causal variant is unknown and the tag SNP is not an outlier in allele frequency distribution, it is impossible to interpret the data with respect to the thrifty genes hypothesis.

The known T2D and obesity risk alleles do not show signs of recent positive selection in this study and therefore cannot be considered as 'thrifty genes'. Although the thrifty gene hypothesis is well known and attractive, it might be that the theory is a bit too simple. Therefore we here discuss the effect of malnutrition and famine on mortality to see whether these could have been major drivers of selection. Periods of food insecurity are relatively common and historically have occurred about once in every ten years [24]. It is likely that privileged groups, like the aristocrats or other elites, did not have problems of food shortage. If we add that the elite probably enjoyed better-than-average environmental conditions, a significantly lower-than-average mortality should be expected. Past experiences do not support this. Hollingsworth [25] studied the British elite in cohorts born in each quarter century between 1550 and 1750. They exhibited a life expectancy at birth (e0) of between 30–38.8 for males and 33.7–38.3 for females. The estimates of e0 for the same historical interval and the normal population are between 33.1–38.7 [26]. They almost exactly match the elite population and it can therefore be concluded that differential mortality in the past – with nutrition as the discriminating variable – seems to have been modest. The majority of the episodes of extraordinary and catastrophic mortality in the past were caused by infectious diseases.

In this study we find signatures of recent selection (<30,000 years ago) for protective alleles for T2D and obesity in several European study populations. This could have had something to do with diet in Europe. The agricultural revolution started around 11,000 years ago in the Middle East and gradually spread towards north-west Europe around 6,000 years ago; it was accompanied by major changes in diet for many human populations [27]. The Mediterranean-type agriculture that developed in Europe comprised livestock that supplied much more protein and fat than the agricultures that developed in warmer parts of the world [27]. Data on diet in

these times are scarce, but historical sources from medieval times from the Netherlands show that, in addition to grain products of wheat and rye, middle-class people ate large amounts of meat and, at least in the Netherlands, the poorer classes generally consumed cheese and milk as their main food source [28]. On the contrary, the average citizen in pre-revolutionary France lived on a poor quality diet containing less than 1400 calories per day [29]. What additionally sets Europeans apart from other ethnicities is that the 'escape from hunger', a dramatic change of event in humans, occurred 200 years in European populations before it did anywhere else in the world [30].

According to the thrifty genes theory, T2D and obesity susceptibility genes conferred a survival advantage in times of food scarcity, but are nowadays maladaptive to a 'Westernized' diet and lifestyle [1]. However, it can be suggested that from the Mediterranean agriculture onwards, Europeans already started to adapt genetically to a 'Westernized' diet with high fat and protein intake. Therefore, natural positive selection possibly already reduced the European frequencies of those thrifty genotypes in previous centuries. The lower frequency of T2D in Europeans compared to other ethnic groups which are now adopting a 'Westernized' diet and lifestyle supports this hypothesis. In the USA, the T2D frequencies are 6.6% in European-Americans, 7.5% in Asian-Americans, 10.4% in Hispanics, and 11.8% in African Americans. Although T2D is a late-onset disease in Europe, it is an early-onset disease in some parts of the world, with the extreme example of the Nauru Islanders and the Pima Indians [31]. Humans with early-onset T2D have a reduced fitness and if we argue that, in Europe, T2D used to be a disease that occurred much earlier in life than nowadays, there might have been a selective pressure against T2D and obesity risk alleles. Some studies also suggest that late-onset T2D patients might have had reduced fertility earlier in life, which represents further reduced fitness [32]. This is also proposed in the 'fertility first' hypothesis by Corbett *et al* [33].

It is possible that studying unusually long haploblocks is much less suitable for susceptibility variants of complex traits than monogenetic traits. The selective pressure is divided among multiple loci, or eventually there may be a balance in fitness [34]. An alternative explanation for the high frequency of individuals susceptible to T2D and obesity could be that obesity and T2D risk alleles rose in allele frequency in the population due to mutation and random drift. Speakman proposed the 'predator release theory' [35] where he postulates a stabilizing selection

for body fatness. Carrying around large fat reserves may enhance the probability of surviving a period of food shortage, but could in the meantime increase the probability of being killed by a predator. During the early period of human evolution (6-2 million years ago), humans were preyed on by large predatory animals. Absence of predation in more recent times led to a change in the population distribution of body fatness due to random mutation and drift.

Conclusion

Our results do not support the thrifty gene hypothesis, because we found no signs of positive selection around T2D and obesity risk alleles. However, some protective variants for T2D and obesity do show (suggestive) signs of positive selection in our European data and it can be argued that Europeans are already adapting genetically to a Western diet by purging genetic variants leading to type 2 diabetes and obesity.

Methods

T2D and obesity susceptibility alleles

We studied all 19 T2D and 17 obesity susceptibility loci indentified up to March 2009 for signs of natural selection. All the variants are single nucleotide polymorphisms (SNPs) that were found to be associated with genome-wide significance and were replicated in other independent populations (within the original study). The T2D and obesity susceptibility loci are described in tables 1 and 2, giving SNPs, the nearest gene, the risk and non-risk alleles, and the ancestral state per allele.

Study populations

In the first stage of the analysis to study haplotype length we used 1,442 individuals from the 1958 UK Birth Cohort (1958BC), genotyped on Illumina HumanHap300 BeadChips for the celiac disease GWA project [36]. In the second stage we tried to replicate our most promising findings in three other genome-wide SNP data sets comprising 2,215 individuals in total: a population of 929 healthy Dutch blood bank controls, a UK study population of 778 celiac disease cases and a Dutch population of 508 cases. We also studied the distribution of the T2D and obesity risk alleles in HapMap phase III. These study populations are described in more detail below.

Dutch subjects

The 929 Dutch controls were unrelated individuals who were blood bank donors selected at random. The 508 cases were unrelated Dutch children and adults with celiac disease. All the cases and controls were from the Netherlands and of European descent, with at least three grandparents born in the Netherlands. Use of the data was approved by the Medical Ethics Committee of the University Medical Center Utrecht. These populations are described in more detail elsewhere [37].

UK subjects

We recruited 778 adult celiac disease patients from outpatient clinics at seven UK hospital sites. They were all of Northern European origin. Ethics committee approval (Oxfordshire REC B) and local approval were obtained for all these study populations [36].

HapMap phase III subjects

Because large allele frequency differences between populations may be the result of differential selection pressures, we studied the allele frequency distribution of the T2D and obesity risk alleles in HapMap phase III data. A total of 1,301 samples from 11 populations are included in the HapMap III database [38]: 90 individuals of African ancestry from Southwest USA (ASW), 180 US residents with Northern and Western European ancestry (CEU), 90 Han Chinese individuals from Beijing (CHB), 100 Gujarati Indians from Houston, Texas (GIH), 100 Chinese from Denver, Colorado (CHD), 91 Japanese individuals from the Tokyo area (JPT), 100 Luhya in Webuye, Kenya (LWK), 90 individuals with Mexican ancestry from Los Angeles (MEX), 180 Maasai in Kinyawa, Kenya (MKK), 100 Toscans from Italy (TSI) and 180 Yoruba people from Ibadan, Nigeria (YRI).

According to their continental origin, samples were divided into four geographical groups: CHB, CHD and JPT were grouped as Eastern Asian ancestry, YRI, ASW, LWK and MKK as African ancestry, GIH and MEX as Native American ancestry and CEU and TSI as European ancestry.

Data analysis

Because not all the SNPs were genotyped directly in the genome-wide data sets, we imputed some genotypes using PLINK (v1.04) with the phase II HapMap CEU data as a reference panel [39]. Only genotypes that could be accurately imputed with a call

rate of above 80% per SNP were used for analyses. To check the effect of imputation on our results, we analysed several regions with and without imputation and found the results to be largely comparable.

Regions of 1 Mb around the T2D or obesity susceptibility alleles were extracted from the imputed data sets and we used the Beagle software program to estimate phased haplotypes from genotypes [40]. Based on chimpanzee alignment, we assigned the ancestral state of all the SNPs in the data files. All the T2D and obesity susceptibility alleles had known ancestral states.

Integrated Haplotype Score (iHS)

We used the freely available, online iHS software to calculate extended haploblocks around the T2D and obesity susceptibility loci in the genome-wide SNP data sets. Briefly, the iHS is a statistic developed to detect evidence of recent positive selection (< 30,000 years ago) at a locus. It is based on the differential levels of linkage disequilibrium (LD) surrounding a positively selected allele compared to the background allele at the same position. Under neutral selection, the LD around variants in the genome will decay over time due to recombination. Hence, older (common) alleles typically have short-range LD and younger (rare) alleles have long-range LD. Without positive selection, new alleles need considerable time to become common. Thus, alleles with a high frequency are typically old, and are therefore expected to have short-range LD: they sit on short haplotypes. However, when an allele is under positive selection, its frequency rises rapidly in the population over a short time span and the haplotype carrying the advantageous allele will be longer relative to haplotypes around equally frequent alleles that have become common purely by random genetic drift. The iHS software has been described in detail elsewhere [18].

An extremely positive iHS score means that haplotypes on the ancestral allele background are longer than the derived allele background (thus suggesting a recent selection on the ancestral allele), while an extremely negative iHS score means that the haplotypes on the derived allele background are longer than the haplotypes associated with the ancestral allele.

We standardized the iHS values using derived frequency bins in a set of 8,500 randomly chosen SNPs surrounding the T2D and obesity susceptibility regions. After

standardization, the iHS was normally distributed. We calculated the p-value with a 2-sided test based on the empirical distribution of the iHS values.

Fixation index (Fst)

When a genetic variation is under positive selection, it increases in prevalence in a population. Because diet, climate and pathogen load vary across the world, there are population differences in selective pressure resulting in global allele frequency variations. Therefore, allele frequency differences between populations could indicate that the alleles were under selection in a certain population (although it could also point towards a population bottleneck). The Fst is a measure of population differentiation based on data of genetic variation and the statistic compares the genetic variability within and between populations [41].

Under neutral selection, Fst is determined by genetic drift and will therefore affect all loci in the genome in a similar way. We study the global Fst values of the T2D and obesity susceptibility loci in the HapMap III populations and compared these values with an empirical genome-wide distribution. We used computed Fst values from the SNP@Evolution database [19].

Acknowledgements

This work was supported by SenterNovem (IOP genomics grant IGE05012), by the Netherlands Genomics Initiative and partly funded by the Dutch Government (grant BSIK03009 to CW) and the Netherlands Organization for Scientific Research (NWO-VICI grant 918.66.620 to CW and grant number 917.66.315 for CGFdK). We thank Flip Mulder for bioinformatics support and Jackie Senior for critically reading the manuscript.

References

1 Neel, J.V. (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet, 14, 353-362.

2 Walley, A.J., Asher, J.E. and Froguel, P. (2009) The genetic contribution to non-syndromic human obesity. Nat Rev Genet, 10, 431-442.

3 Cohen, M.N. (1989) Health and the rise of civilization, New Haven, CT Yale University Press, 1989.

4 Baschetti, R. (1998) Diabetes epidemic in newly westernized populations: is it due to thrifty genes or to genetically unknown foods? J R Soc Med, 91, 622-625.

5 Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S. et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature, 445, 881-885.

6 Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U. et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science, 316, 1341-1345.

7 Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J. et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science, 316, 1331-1336.

8 Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D.P., Holmkvist, J., Borch-Johnsen, K., Jorgensen, T. et al. (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet, 40, 1098-1102.

9 Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S. et al. (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet, 39, 770-775.

10 Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G. et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet, 40, 638-645.

11 Lyssenko, V., Nagorny, C.L., Erdos, M.R., Wierup, N., Jonsson, A., Spegel, P., Bugliani, M., Saxena, R., Fex, M., Pulizzi, N. et al. (2009) Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. Nat Genet, 41, 82-88.

12 Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W. et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science, 316, 889-894.

13 Loos, R.J., Lindgren, C.M., Li, S., Wheeler, E., Zhao, J.H., Prokopenko, I., Inouye, M., Freathy, R.M., Attwood, A.P., Beckmann, J.S. et al. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet, 40, 768-775.

14 Thorleifsson, G., Walters, G.B., Gudbjartsson, D.F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I. et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet, 41, 18-24.

15 Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U., Lamina, C. et al. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet, 41, 25-34.

16 Hofker, M. and Wijmenga, C. (2009) A supersized list of obesity genes. Nat Genet, 41, 139-140.

17 Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature, 419, 832-837.

18 Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome. PLoS Biol, 4, e72.

19 Cheng, F., Chen, W., Richards, E., Deng, L. and Zeng, C. (2009) SNP@Evolution: a hierarchical database of positive selection on the human genome. BMC Evol Biol, 9, 221.

20 Lohmueller, K.E., Mauney, M.M., Reich, D. and Braverman, J.M. (2006) Variants associated with common disease are not unusually differentiated in frequency across populations. Am J Hum Genet, 78, 130-136.

21 Myles, S., Davison, D., Barrett, J., Stoneking, M. and Timpson, N. (2008) Worldwide population differentiation at disease-associated SNPs. BMC Med Genomics, 1, 22.

22 Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J. (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet, 10, 241-251.
23 Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W. et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res.
24 Minnesota. University. Laboratory of Physiological Hygiene. and Keys, A.B. (1950) The biology of human starvation. University of Minnesota Press, Minneapolis,.
25 Hollingsworth, T.H. (1964) The demography of the British peerage. Population Investigation Committee, London,.
26 Rotberg, R.I., Rabb, T.K. and Boserup, E. (1985) Hunger and history : the impact of changing food production and consumption patterns on society. Cambridge University Press, Cambridge [Cambridgeshire] ; New York.
27 Kiple, K.F. and Ornelas, K.C. (2000) The Cambridge world history of food. Cambridge University Press, Cambridge ; New York.
28 Jobse-van Putten, J. (1996) Simple But Nutritious: Cultural History of Daily Meals in the Netherlands. SUN Memoria / P. J. Meertens Instituut, Amsterdam 1995, Amsterdam.
29 Toutain, J.-C. (1971) La consommation alimentaire en France de 1789 a 1964. INED, Paris.
30 Diamond, J. (2003) The double puzzle of diabetes. Nature, 423, 599-602.
31 Dowse, G.K. (1996) Incidence of NIDDM and the natural history of IGT in Pacific and Indian Ocean populations. Diabetes Res Clin Pract, 34 Suppl, S45-50.
32 Sam, S. and Dunaif, A. (2003) Polycystic ovary syndrome: syndrome XX? Trends Endocrinol Metab, 14, 365-370.
33 Corbett, S.J., McMichael, A.J. and Prentice, A.M. (2009) Type 2 diabetes, cardiovascular disease, and the evolutionary paradox of the polycystic ovary syndrome: a fertility first hypothesis. Am J Hum Biol, 21, 587-598.
34 Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W. and Pritchard, J.K. (2009) The role of geography in human adaptation. PLoS Genet, 5, e1000500.
35 Speakman, J.R. (2007) A nonadaptive scenario explaining the genetic predisposition to obesity: the "predation release" hypothesis. Cell Metab, 6, 5-12.
36 van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C., Bethel, G., Holmes, G.K. et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet, 39, 827-829.
37 Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D. et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet, 40, 395-402.
38 Li, C.Y., Chiu, C.H., Huang, H.S., Lin, C.H. and Wu, T.S. (2006) High-performance liquid chromatographic method for simultaneous quantification of eight major biologically active ingredients in 'Da-Chai-Hu-Tang' preparation. Biomed Chromatogr, 20, 305-308.
39 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 81, 559-575.
40 Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet, 81, 1084-1097.
41 Lewontin, R.C. and Krakauer, J. (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics, 74, 175-195.

Chapter 9

Type 2 diabetes and earlier-in-life sub- and infertility

Clara C. Elbers, N. Charlotte Onland-Moret, René Eijkemans,

Cisca Wijmenga, Diederick E. Grobbee, Yvonne T. van der Schouw

Abstract

Background Fertility problems are frequently followed by early menopause and early menopause has been associated with diabetes. Thus far, it is unknown whether sub- or infertility is independently associated with future type 2 diabetes risk.

Methods We assessed the association between measures of sub- and infertility in a cohort of 17,357 Dutch women, aged 49-70 at baseline using Cox proportional hazards models. Analyses were adjusted for various confounders. To investigate whether BMI, waist and gestational diabetes mediate the observed associations, analyses were additionally adjusted for these variables.

Results At baseline, 332 women had T2D. During a mean follow-up of 9.1 ± 3.6 years, 535 T2D cases occurred. Compared to women with a regular cycle length of 27-29 days, women with irregular menstrual cycles had an increased, albeit non-significant risk for T2D, with a multivariate adjusted HR of 1.20 (95% CI 0.97-1.49). The association strengthened after adjusting for BMI (HR for irregular cycle=1.25; 95% CI 1.00-1.54) and waist (HR for irregular cycle=1.22; 95% CI 0.98-1.51). None of the other measures of sub- or infertility were associated with increased risk for T2D.

Conclusion Generally, measures of sub- and infertility did not independently predict subsequent development of T2D. However, most T2D patients were diagnosed after menopause. Therefore, future studies should further investigate the association between reduced fertility and premenopausal T2D, especially as the developing epidemic of obesity has shown a substantial reduction in the age of onset of T2D and is starting to emerge in women of childbearing age.

Introduction

Compared to all other populations with a modern lifestyle, the age-adjusted type 2 diabetes (T2D) prevalence in populations of European ancestry is relatively low [1]; T2D prevalence in U.S. Europeans, U.S. Africans, U.S. Hispanics and U.S. Pima Indians is respectively 7.6%, 13%, 17% and 50% [1]. It has been proposed that these differences in T2D susceptibility between European and non-European populations are the genetic and evolutionary consequences of geographical differences in food history [2,3].

The 'thrifty genes theory' hypothesized that the T2D phenotype gives a survival advantage during periods of famine, but is maladaptive in societies with high food abundance [4]. Historical data show that starting from about 1600, European societies became capable to efficiently intervene famine, by redistributing over-abundance grain to areas of food scarcity [5]. Diamond [2] suggested that as a result, Europeans should have undergone an epidemic in T2D starting several centuries before present as a result of the new reliability of sufficient food supplies, and eliminated the most T2D-prone genotypes by processes of natural selection [2].

Natural selection works through differential reproductive success rather than simple differential survival. Because fertility is a driving force behind evolution, infertility could be one of the underlying causes that decreases the T2D genotype frequencies in Europeans, especially since T2D is a late-onset disease and therefore not directly acting on survival. However, it is unknown whether T2D is associated with earlier in life reproductive problems, although there is some indirect evidence suggesting a link. Fertility problems are frequently followed by early menopause [6], and early menopause has been associated with type 1 diabetes and premenopausally diagnosed T2D [7]. An earlier decline in the ovarian follicle pool has been suggested as a cause of early menopause in women with type 1 diabetes [7,8]. Also obesity, the most important risk factor for T2D, is associated with reduced fertility. Previously, a U-shaped association between BMI and relative risk of ovarian infertility was observed in the Nurses' Health Study II, with increased risk for ovarian infertility for women with a BMI below 20 and above 24 kg/m2 [9].

Thus far, it is unknown whether sub- or infertility is independently associated with future risk of developing T2D. Therefore, we assessed the association between

measures of sub- and infertility and T2D risk in the Prospect cohort comprising 17,357 Dutch women.

Methods

Subjects

The Prospect-EPIC cohort is one of the two Dutch contributions to the European Prospective Investigation into Cancer and Nutrition (EPIC) [10]. It is a prospective cohort study among 17,357 women aged 49–70 who lived in Utrecht and vicinity and who participated in the breast cancer screening program between 1993 and 1997 [11]. All participants gave their written informed consent and the study was approved by the Institutional Review Board. The design, sampling strategies, and examination techniques of the cohort have been described previously [11].

Data collection

Baseline measurements

At baseline, all participants filled out detailed questionnaires on usual diet, reproductive history, presence of chronic diseases and related potential risk factors. They underwent a brief medical examination and a blood sample was drawn.

Measures of reduced fertility

To define reduced, sub- or infertility, we used the following variables: (I) having had an irregular menstrual cycle pattern between age 30-40 years, (II) having consulted a physician for fertility problems, (III) nulliparity, (IV) uniparity (V) having had a miscarriage (VI), a long time interval between the birth of the first and the second child (VII) a short reproductive time between the age of menarche and the age of menopause.

For each analysis, an appropriate subpopulation was defined: (I) menstrual cycle irregularity, in all women reporting on menstrual cycle pattern. The information on menstrual cycle pattern concerned the period between age 30 and 40 years and irregularity of the menstrual cycle pattern was self-defined; (II) subfertility, by studying ever consulting a medical doctor for fertility problems in all women who

reported that they have tried to achieve pregnancy; (III) nulliparity , in women who reported that they have tried to achieve pregnancy; (IV) having only one child, in all parous women; (V) at least one miscarriage, in all women who were ever pregnant; (VI) time interval >5 years between birth of first and second child, in women with at least two live born children; (VII) reproductive time, in all women who reached a natural menopause at baseline.

Potential covariates

Because of the potential for confounding, we adjusted our analyses for BMI, waist, gestational diabetes, age, alcohol intake, physical activity, smoking, socio-economic status and oral contraceptives use.

Body weight was measured in light indoor clothing without shoes to the nearest 0.5 kg with a floor scale (Seca, Atlanta, GA, USA). Additionally, height, waist and hip circumference were measured. Body mass index (BMI) was calculated as weight divided by height squared (kg/m2).

Alcohol consumption was assessed by a validated food frequency questionnaire (FFQ). Baseline alcohol intake was determined by multiplying the consumption of each beverage by its ethanol content and was calculated to grams per week (g/week). Subsequently, we categorized subject into four alcohol consumption categories: <0.05 g/week, 0.05-5.5 g/week, 5.5-10.5 g/week, >10.5 g/week.

Duration and types of physical activity during the year preceding study recruitment were assessed by a set of questions that was used in all EPIC cohorts. By combining occupational physical activity with time spent on cycling and sporting in summer and winter, the validated Cambridge Physical Activity Index (CPAI) [12] was calculated. Based on this Index participants were divided in four physical activity categories: inactive, moderately inactive, moderately active and active.

Smoking behaviour was categorized as no, former or current smokers.

To define the socio-economical status, the highest attained level of education of the participants was used and classified into three categories: low (primary education up to completing intermediate vocational education), middle (up to higher secondary education) and high (those with higher vocational education and university).

The number of years of oral contraception use was self reported, and participants were divided into four groups: never, 1-4 years, 4-10 years, >10 years.

Self reported gestational diabetes status during pregnancy was indicated as yes or no.

Missing value analyses

Missing values for BMI, waist, alcohol intake, physical activity, CPAI, smoking, gestational diabetes, socio-economical status, years of oral contraceptives use, number of miscarriages and age of menarche were imputed using multiple imputation [13], repeated 5 times to account for uncertainties in imputed data. None of the variables had > 5% missing values; the percentage of missing values ranged from 0.1% for BMI to 2.9% for years of oral contraceptives use.

Morbidity and mortality follow-up

Occurrence of T2D during follow-up was obtained via self-report in two follow-up questionnaires sent to the participants within intervals of three to five years, linkage to the Dutch register of hospital discharge diagnoses (HDD) and a mailed urinary glucose strip test (part of the cohort) (I Sluijs Neth J Med, under revision). Potential cases of T2D were verified against information from the participants' general practitioner or pharmacist through mailed questionnaires. T2D was defined present when the general practitioner or pharmacist confirmed the diagnosis. Information on vital status was obtained through linkage with the municipal administration registries [14]. Causes of death were obtained from the Dutch Central Bureau of Statistics, coded according to the International Classification of Diseases, Tenth Revision, clinical Modification (ICD-10). For our analyses, T2D was the endpoints of interest and follow up ended at the date of diagnosis or at the date of death. All others were censored on January 1st 2006.

Data analysis

Population characteristics are described using means and standard deviations (for normally distributed variables) and numbers and frequencies (for categorical variables).

The person-time for each woman was calculated from birth to the month of diagnosis of the endpoint (T2D), the month of death from other causes, or the end of follow-up (January 1, 2006). Hazard ratio's (HRs) and 95% confidence intervals (CIs) for risk on T2D were estimated using Cox regression analysis. We used a stepwise approach to adjust for potential confounders and study the role of the potential

intermediate factors BMI, waist and gestational diabetes, using five multivariate models: model 1, including age at baseline; model 2, including potential confounders age, alcohol intake, physical activity, smoking, socio-economic status and oral contraceptives use; model 3, including all confounders from model 2 and BMI; model 4, including all confounders from model 2 and waist; model 5, including all confounders from model 2 and gestational diabetes.

We also performed trend analyses with categorical variables entered as continuous variables in the model to test a dose-response effect for measures of sub- and infertility and the risk for T2D.

Results were considered statistically significant at 2-sided $P \leq 0.05$. All statistical analyses were performed using SPSS.

Results

Table 1 shows the baseline characteristics of the population included in this study. The study had a mean follow-up of 9.1 ± 3.6 years and comprised 157,964 person-years. During follow-up, 535 new T2D cases occurred. When calculating follow-up time from birth, this resulted in a mean follow-up of 66.9 ± 6.7 years with a corresponding 1,160,428 person-years. The mean age of the study group at baseline was 57.1 ± 6.0 years. In total, the study contained 867 verified T2D patients; 332 prevalent and 535 incident cases.

Table 1. Baseline characteristics of 17,357 women from the PROSPECT cohort

	Mean ± SD
Follow-up time (yr)	66.86 ± 6.66
Age at intake (yr)	57.14 ± 6.03
BMI (kg/m2)	26.03 ± 4.09
Waist (cm)	83.75 ± 10.17
Alcohol intake (g/week)	9.11 ± 12.61
Pill use (yr)	5.38 ± 6.75
Live born children	2.37 ± 1.51
	N (%)
Smoking	
Current smoker	3783 (20.0)
Former smoker	5982 (34.7)
Non-smoker	7466 (43.3)
Physical activity	
	N (%)
Inactive	1301 (7.5)
Moderate inactive	4612 (26.6)
Moderate active	4437 (25.6)
Active	7007 (40.4)
Educational level	
Low	13311 (76.7)
Medium	1249 (7.2)
High	2270 (13.1)
Gestational diabetes	
No	16714 (96.3)
Yes	643 (3.7)

A total of 13,991 women reported on regularity of natural menstruation between the ages of 30 to 40 years, of whom 14% reported to have an irregular menstrual cycle. Compared to women with a regular cycle length of 27-29 days, women with irregular menstrual cycles had an increased, but non-significant, risk for T2D (model 2, Table 2), with a multivariate adjusted HR of 1.20 [95% CI 0.97-1.49]. The association strengthened after adjusting for BMI or waist, with a BMI adjusted HR of 1.25 [95% CI 1.00-1.54] and a waist adjusted HR of 1.22 [95% CI 0.98-1.51]. Adjustment for gestational diabetes did not alter the results.

Table 2. Hazard ratio's for Type 2 Diabetes risk by menstrual period

	HR	HR (95% CI)	HR (95% CI)	HR (95% CI)
Menstrual period	≤ 26	27-29	≤ 30	irregular
Subjects (n)	3,737	6,779	1,528	1,947
T2D cases (%)	177 (4.7)	312 (4.6)	83 (5.4)	117 (6.0)
Model 1[a]	1.06 (0.88-1.27)	1	1.19 (0.93-1.51)	1.36 (1.10-1.68)
Model 2[b]	1.01 (0.84-1.22)	1	1.19 (0.93-1.52)	1.20 (0.97-1.49)
Model 3[c]	1.05 (0.87-1.26)	1	1.20 (0.94-1.53)	1.25 (1.00-1.54)
Model 4[d]	1.09 (0.90-1.31)	1	1.13 (0.95-1.54)	1.22 (0.98-1.51)
Model 5[e]	1.01 (0.84-1.21)	1	1.17 (0.92-1.50)	1.19 (0.96-1.47)

[a] Model 1 = Adjusted for age at baseline (continuous)
[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), alcohol intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)
[c] Model 3 = Model 2 plus BMI (continuous)
[d] Model 4 = Model 2 plus waist (continuous)
[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

Out of the 15,708 Prospect women who wanted to get pregnant, 11.8% consulted a physician for fertility problems. Consulting a physician for fertility problems was not associated with risk of T2D; multivariate adjusted HR for women who consulted a physician for fertility problem versus women who did not was 0.95 [95% CI 0.76-1.20], a pattern that did not change after BMI, waist or gestational diabetes adjustment (Table 3).

Of the 15,708 women who tried to get pregnant, 700 (4.5%) remained childless. No relationship was found between nulliparity and future T2D risk (model 2, table 4); multivariate adjusted HR for nulliparous women compared to parous women was 0.98 [95% CI 0.68-1.41]. This did not change after adjustment for BMI, waist and gestational diabetes (Table 4).

Table 3. Hazard ratio's for Type 2 Diabetes risk by fertility consult in 15,708 women who tried to get pregnant

	HR	HR (95% CI)
fertility consult	no	yes
Subjects (n)	14,726	1,849
T2D cases (%)	766 (5.1)	83 (4.3)
Model 1[a]: age	1	0.90 (0.72-1.13)
Model 2 [b]: multiple confounders	1	0.95 (0.76-1.20)
Model 3 [c]: model 2 + BMI	1	1.04 (0.83-1.30)
Model 4 [d]: model 2 + waist	1	1.02 (0.81-1.28)
Model 5 [e]: model 2 + gestational diabetes	1	0.93 (0.74-1.17)

[a] Model 1 = Adjusted for age at baseline (continuous)
[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)
[c] Model 3 = Model 2 plus BMI (continuous)
[d] Model 4 = Model 2 plus waist (continuous)
[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

Table 4. Hazard ratio's for Type 2 Diabetes risk by nulliparty in 15,708 women who tried to get pregnant

	IR (95% CI)	HR
Number of children	≥ 1	0
Subjects (n)	15,008	700
T2D cases (%)	762 (5.1)	31 (4.4)
Model 1[a]: age	1	0.88 (0.62-1.26)
Model 2 [b]: multiple confounders	1	0.98 (0.68-1.41)
Model 3 [c]: model 2 + BMI	1	1.04 (0.72-1.49)
Model 4 [d]: model 2 + waist	1	1.05 (0.73-1.51)
Model 5 [e]: model 2 + gestational diabetes	1	1.04 (0.73-1.50)

[a] Model 1 = Adjusted for age at baseline (continuous)
[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)
[c] Model 3 = Model 2 plus BMI (continuous)
[d] Model 4 = Model 2 plus waist (continuous)
[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

Of the 15,129 Prospect women who had children, 1,487 (9.8%) women were uniparous. Compared to women with two or more children, women with only one child had a decreased risk for T2D (model 2, Table 5), with a multivariate adjusted HR of 0.77 [95% CI 0.59-1.00]. However, after adjustment for BMI or waist (table 5, model 3 and 4 respectively), the association between uniparity and reduced T2D risk was no longer statistically significant; BMI and waist adjusted HRs for women with one child compared to women with two or more children were 0.82 [95% CI 0.63–1.08] and 0.81 [95% CI 0.62-1.06], respectively. Adjustment for gestational diabetes again did not alter the results.

Table 5. Hazard ratio's for Type 2 Diabetes risk by uniparity in 15,129 women with children

	HR (95% CI)	HR
Number of children	≥ 2	1
Subjects (n)	13,642	1,487
T2D cases (%)	710 (5.2)	60 (4.0)
Model 1[a]: age	1	0.81 (0.62-1.06)
Model 2 [b]: multiple confounders	1	0.77 (0.59-1.00)
Model 3 [c]: model 2 + BMI	1	0.82 (0.63-1.08)
Model 4 [d]: model 2 + waist	1	0.81 (0.62-1.06)
Model 5 [e]: model 2 + gestational diabetes	1	0.77 (0.58-1.01)

[a] Model 1 = Adjusted for age at baseline (continuous)
[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)
[c] Model 3 = Model 2 plus BMI (continuous)
[d] Model 4 = Model 2 plus waist (continuous)
[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

Out of the 15,708 Prospect women who wanted to get pregnant, 3,979 (25.3%) had one or miscarriages, with an average of 1.4 (± 0.9) miscarriages and a maximum of 10 miscarriages. Women who had one or more miscarriages showed no differential risk for T2D compared to women who did not have a miscarriage (table 6), with a multivariate adjusted HR of 1.12 [95% CI 0.96-1.31], a BMI adjusted HR of 1.07 [95% CI 0.91-1.25], a waist adjusted HR of 1.02 [95% CI 0.87-1.19] and a gestational diabetes adjusted HR of 1.08 [95% CI 0.93-1.27].

Table 6. Hazard ratio's for Type 2 Diabetes risk by number of miscarriages in 15,708 women who tried to get pregnant

	HR	HR (95% CI)
Number of miscarriages	none	≥ 1
Subjects (n)	11,729	3979
T2D cases (%)	569 (4.9)	223 (5.6)
Model 1[a]: age	1	1.12 (0.96-1.31)
Model 2 [b]: multiple confounders	1	1.12 (0.96-1.31)
Model 3 [c]: model 2 + BMI	1	1.07 (0.91-1.25)
Model 4 [d]: model 2 + waist	1	1.02 (0.87-1.19)
Model 5 [e]: model 2 + gestational diabetes	1	1.08 (0.93-1.27)

[a] Model 1 = Adjusted for age at baseline (continuous)
[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)
[c] Model 3 = Model 2 plus BMI (continuous)
[d] Model 4 = Model 2 plus waist (continuous)
[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

The average time interval between the first and second child of Prospect women was 32.7 (± 20.4) months. Time interval between children was not associated with future T2D risk (multivariate adjusted p-value for trend 0.40). Again, adjustment for BMI, waist and gestational diabetes did not alter the results (p-values for trend were 0.36, 0.59 and 0.41 respectively) (table 7).

In Prospect, 6,292 women reported to have a natural menopause. The average time interval between menarche and natural menopause, i.e. the reproductive time, was 36.9 (± 4.5) years. Reproductive time was not associated with T2D risk with a multivariate adjusted HR of 1.22 [95% CI 0.93-1.61] for women with a reproductive time of ≥ 40 years compared to women with a reproductive time ≤ 35 (multivariate adjusted p-value for trend= 0.06; model 2, table 8). After adjustment for both BMI and waist separately (model 3 and 4 respectively, table 8) the association between reproductive time and reduced T2D risk attenuated. The HR for women with a reproductive time of ≥ 40 years versus women with a reproductive time of ≤ 35 years was 1.09 [95% CI 0.83-1.44] for BMI adjusted analysis and 1.15 [95% CI 0.87-1.51] for waist adjusted analysis. Again adjustment for gestational diabetes did not change the results.

Table 7. Hazard ratio's for Type 2 Diabetes risk by number of month between the first and second child in 13,396 women with two or more children with known birth years

	HR	HR (95% CI)	HR (95% CI)	HR (95% CI)	P trend
Month between first and second child	≤ 18	19-27	28-40	≥ 41	
Subjects (n)	3,093	3,459	3,658	3,186	
T2D cases (%)	178 (5.8)	170 (4.9)	167 (4.6)	181 (5.7)	
Model 1[a]	1	0.91 (0.74-1.12)	0.87 (0.70-1.08)	1.02 (0.83-1.25)	0.97
Model 2[b]	1	0.93 (0.76-1.15)	0.87 (0.70-1.08)	0.93 (0.76-1.14)	0.40
Model 3[c]	1	0.96 (0.77-1.18)	0.87 (0.70-1.07)	0.93 (0.76-1.15)	0.36
Model 4[d]	1	0.95 (0.77-1.17)	0.88 (0.71-1.09)	0.96 (0.78-1.19)	0.59
Model 5[e]	1	0.93 (0.76-1.15)	0.87 (0.70-1.07)	0.93 (0.76-1.15)	0.41

[a] Model 1 = Adjusted for age at baseline (continuous)
[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), alcohol intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)
[c] Model 3 = Model 2 plus BMI (continuous)
[d] Model 4 = Model 2 plus waist (continuous)
[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

Table 8. Hazard ratio's for Type 2 Diabetes risk by number of reproduction years in 6,411 women with natural menopause

	HR	HR (95% CI)	HR (95% CI)	HR (95% CI)	P trend
Reproduction years	≤ 35	36-37	38-39	≥ 40	
Subjects (n)	2,021	1,265	1,397	1,728	
T2D cases (%)	100 (4.9)	53 (4.2)	88 (6.3)	108 (6.3)	
Model 1[a]	1	0.83 (0.59-1.15)	1.21 (0.91-1.61)	1.12 (0.86-1.48)	0.17
Model 2[b]	1	0.91 (0.65-1.27)	1.30 (0.97-1.74)	1.22 (0.93-1.61)	0.06
Model 3[c]	1	0.89 (0.64-1.25)	1.29 (0.97-1.73)	1.09 (0.83-1.44)	0.26
Model 4[d]	1	0.89 (0.64-1.24)	1.35 (1.01-1.81)	1.15 (0.87-1.51)	0.12
Model 5[e]	1	0.89 (0.64-1.25)	1.26 (0.95-1.69)	1.22 (0.92-1.60)	0.07

[a] Model 1 = Adjusted for age at baseline (continuous)
[b] Model 2 = Adjusted for age at baseline (continuous), smoking (never, past, current), alcohol intake (<0.05 g/w, 0.05-5.5 g/w, 5.5-10.5 g/w, >10.5 g/w), socio-economical status (low, middle, high), pill years (never, 1-4 years, 4-10 years, >10 years), and physical activity (inactive, moderate inactive, moderate active, active)
[c] Model 3 = Model 2 plus BMI (continuous)
[d] Model 4 = Model 2 plus waist (continuous)
[e] Model 5 = Model 2 plus gestational diabetes (yes/no)

Discussion

In this large cohort of 17,357 women, measures of sub- and infertility did not independently predict subsequent development of T2D. To our knowledge, this is the first study to investigate the association between various measures of sub- and infertility and future T2D risk in a prospective cohort.

Before interpreting the data, some strengths and limitations need to be discussed. The main advantages of this study are its prospective nature, the long follow-up time and the large sample size. Furthermore, the women were extensively questioned on their reproductive history. Also, potential cases of T2D were verified by the participants' general practitioner or pharmacist and T2D was only defined present when one of them confirmed the diagnosis. In this study, the person-time for each woman was calculated from birth to the month of T2D diagnosis, to the month of death from other causes, or to the end of follow-up. In prospective cohort studies, the person-time is usually calculated from baseline, therefore only including incident T2D cases. As our variables for measures of sub- or infertility were established long before T2D onset in cases, we choose to calculate person-time from birth. This allowed us to include an extra 332 prevalent T2D cases for analyses, which were otherwise excluded. We additionally studied the association of parity and age at first childbirth and T2D risk with person-time calculated from baseline, to explore the effect of using a person-time from baseline versus from birth. The results were comparable (data not shown) with similar HR, but wider 95% CI.

In this study we used the variables 'having had an irregular menstrual cycle', 'having consulted a physician for fertility problems', 'nulliparity', 'uniparity', 'having had a miscarriage', 'a long time interval between the birth of the first and the second child' and 'a short reproductive time' as measures of sub- and infertility. It can be discussed whether these variables truly represent sub- or infertility in women. 'Having consulted a physician for fertility problems', 'nulliparity', 'uniparity' and 'a long time interval between the birth of the first and the second child' could also have been caused by male infertility. Of the 1,849 Prospect women who, together with their partners, consulted a physician for fertility problems, eventually 845 (46%) couples got a diagnosis of sub- or infertility; in 353 couples (42%) the female was diagnosed with sub- or infertility, in 362 couples (43%) the male was diagnosed with sub- or infertility and in 130 couples (15%) both female and male were diagnosed

with sub- or infertility. However, the associations between measures of sub- and infertility and T2D risk did not change, when we excluded women with sub- or infertile partners (data not shown). Although some misclassification cannot be excluded, this is likely to be non-differential, since misclassification of sub- or infertile women occurred independently of T2D.

The variable 'time interval between birth of first and second child' is a substitute for time to pregnancy, which is widely used to estimate the degree of subfertility [15]. However, we were unable to directly determine time to pregnancy in our cohort. Even though the interval between first and second child comprises for a major part unintentional waiting time, most likely subfertility in this analysis is of relatively minor magnitude, because all women in this studied subpopulation were able to conceive at least twice.

Long or highly irregular menstrual cycles have been associated with insulin resistance, higher glucose levels and increased risk of T2D in previous studies [16,17,18]. We previously reported that compared to women with a regular cycle length of 27-29 days, women with irregular menstrual cycles had a non-significant increased risk for T2D, and a significantly increased risk of coronary heart disease [18]. Here we showed that the association with T2D slightly strengthened after adjusting for both BMI and waist separately. However, the link between irregular menstrual cycles and T2D remains unknown. Both the association with T2D and the association with coronary heart disease could not be explained by metabolic risk factors or altered hormone levels [18].

In the Prospect cohort, 91.2% of the T2D patients were diagnosed after menopause. As the developing epidemic of obesity currently show a substantial reduction in the age of onset of T2D and is emerging in women of childbearing age, it is important to further investigate the association between reduced infertility and premenopausal T2D. Previous studies provide some evidence for the connection between infertility and premenopausal T2D. First of all, one common cause of sub- and infertility, the polycystic ovary syndrome (PCOS), is already known to be associated with impaired glucose tolerance and T2D in adolescent girls and premenopausal women [19,20]. The syndrome is a heritable form of ovarian infertility that clinically affects 5-10% of reproductive women and is characterized by a long history of chronic anovulation in association with insulin resistance and androgen excess [21,22]. Secondly, reproductive abnormalities are often part of the metabolic

syndrome when it occurs in premenopausal women [23]. The metabolic syndrome is recognized as a major risk factor for T2D [24]. Thirdly, pregnancy losses, predominantly through stillbirth, are high in women with type 1 and type 2 diabetes [25]. However, it is unknown whether T2D associated phenotypes cause sub- or infertility or whether sub- and infertility are markers for unknown factors increasing T2D risk in menopausal women. As our data show that sub- and infertility do not predict subsequent development of T2D, it is tempting to speculate that premenopausal T2D is causal for reduced fertility rather then the other way around. Unfortunately, we were not able to study the association between infertility and premenopausal T2D risk in Prospect, due to the low number of premenopausal T2D cases.

Our data show that general measures of sub- and infertility are not associated with T2D later in life. However, most T2D patients were diagnosed after menopause. Future studies should investigate the association between reduced infertility and premenopausal T2D, especially as the developing epidemic of obesity currently show a substantial reduction in the age of onset of T2D and its emergence in women of childbearing age.

References

1. King H, Aubert RE, Herman WH (1998) Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. Diabetes Care 21: 1414-1431.
2. Diamond J (2003) The double puzzle of diabetes. Nature 423: 599-602.
3. Corbett SJ, McMichael AJ, Prentice AM (2009) Type 2 diabetes, cardiovascular disease, and the evolutionary paradox of the polycystic ovary syndrome: a fertility first hypothesis. Am J Hum Biol 21: 587-598.
4. Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14: 353-362.
5. Rotberg RI, Rabb TK, Boserup E (1985) Hunger and history : the impact of changing food production and consumption patterns on society. Cambridge: Cambridge University Press. 336 p. p.
6. Kok HS, van Asselt KM, van der Schouw YT, Grobbee DE, te Velde ER, et al. (2003) Subfertility reflects accelerated ovarian ageing. Hum Reprod 18: 644-648.
7. Dorman JS, Steenkiste AR, Foley TP, Strotmeyer ES, Burke JP, et al. (2001) Menopause in type 1 diabetic women: is it premature? Diabetes 50: 1857-1862.
8. Soto N, Iniguez G, Lopez P, Larenas G, Mujica V, et al. (2009) Anti-Mullerian hormone and inhibin B levels as markers of premature ovarian aging and transition to menopause in type 1 diabetes mellitus. Hum Reprod 24: 2838-2844.
9. Rich-Edwards JW, Spiegelman D, Garland M, Hertzmark E, Hunter DJ, et al. (2002) Physical activity, body mass index, and ovulatory disorder infertility. Epidemiology 13: 184-190.
10. Riboli E (1992) Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 3: 783-791.
11. Boker LK, van Noord PA, van der Schouw YT, Koot NV, Bueno de Mesquita HB, et al. (2001) Prospect-EPIC Utrecht: study design and characteristics of the cohort population. European Prospective Investigation into Cancer and Nutrition. Eur J Epidemiol 17: 1047-1053.
12. Wareham NJ, Jakes RW, Rennie KL, Schuit J, Mitchell J, et al. (2003) Validity and repeatability of a simple index derived from the short physical activity questionnaire used in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Public Health Nutr 6: 407-413.
13. van der Heijden GJ, Donders AR, Stijnen T, Moons KG (2006) Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 59: 1102-1109.
14. Herings RM, Bakker A, Stricker BH, Nap G (1992) Pharmaco-morbidity linkage: a feasibility study comparing morbidity in two pharmacy based exposure cohorts. J Epidemiol Community Health 46: 136-140.
15. Greenhall E, Vessey M (1990) The prevalence of subfertility: a review of the current confusion and a report of two new studies. Fertil Steril 54: 978-983.
16. Roumain J, Charles MA, de Courten MP, Hanson RL, Brodie TD, et al. (1998) The relationship of menstrual irregularity to type 2 diabetes in Pima Indian women. Diabetes Care 21: 346-349.
17. Weiss DJ, Charles MA, Dunaif A, Prior DE, Lillioja S, et al. (1994) Hyperinsulinemia is associated with menstrual irregularity and altered serum androgens in Pima Indian women. Metabolism 43: 803-807.
18. Gast GC, Grobbee DE, Smit HA, Bueno-de-Mesquita HB, Samsioe GN, et al. (2010) Menstrual cycle characteristics and risk of coronary heart disease and type 2 diabetes. Fertil Steril.
19. Moran LJ, Misso ML, Wild RA, Norman RJ (2010) Impaired glucose tolerance, type 2 diabetes and metabolic syndrome in polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod Update.
20. Legro RS, Kunselman AR, Dodson WC, Dunaif A (1999) Prevalence and predictors of risk for type 2 diabetes mellitus and impaired glucose tolerance in polycystic ovary syndrome: a prospective, controlled study in 254 affected women. J Clin Endocrinol Metab 84: 165-169.
21. (2004) Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). Hum Reprod 19: 41-47.
22. Azziz R, Marin C, Hoq L, Badamgarav E, Song P (2005) Health care-related economic burden of the polycystic ovary syndrome during the reproductive life span. J Clin Endocrinol Metab 90: 4650-4658.
23. Sam S, Dunaif A (2003) Polycystic ovary syndrome: syndrome XX? Trends Endocrinol Metab 14: 365-370.
24. Reusch JE (2002) Current concepts in insulin resistance, type 2 diabetes mellitus, and the metabolic syndrome. Am J Cardiol 90: 19G-26G.

25. Macintosh MC, Fleming KM, Bailey JA, Doyle P, Modder J, et al. (2006) Perinatal mortality and congenital anomalies in babies of women with type 1 or type 2 diabetes in England, Wales, and Northern Ireland: population based study. BMJ 333: 177.

Chapter 10

Historical perspective on body characteristics: physical appearance of members of the 18th century Dutch gang of Calotte

Clara C. Elbers, N. Charlotte Onland-Moret, Yvonne T. van der Schouw, Cisca Wijmenga, Lotte C. van de Pol

Abstract

During the last couple of decades, Europeans increased both in length and width.

Before the 19th century there is very little concrete information on which to base conclusions on heights and especially weight. In the Municipal Archives of 's-Hertogenbosch, in the Netherlands, we found a list dated 1766 with 319 names and very detailed physical descriptions of a group of people who were all members of a criminal gang. Based on these descriptions we were able to obtain information on the different body characteristics during this part in time.

Individuals on the list were all between 16 and 65 years of age, with a mean age of 32.8 (±10.3). Of the group, 55.8% was male, and the origin of the gang members was Northern-European (90%), Jewish (8.8%) and Roma (1.4%). The height of males was normally distributed with a mean of 1.58 (±0.05) meters (m) for males. Men from Jewish origin were significantly shorter compared to men from Northern-European origin ($p<0.0001$), with an average height of 1.54 (±0.05) m for Jewish versus 1.59 (±0.05) m for Northern-European men. For body weight we found that the majority of the people was described as average of weight, however, more individuals were fatter than average compared to thinner than average. Of the study population, 24% was described as very fat and additionally, body characteristics concerning fat were mentioned, such as 'a huge belly', 'fat legs' or 'a round and plump face'.

The short average height of 1.58 m for males suggests a poor biological well-being for the individuals of the gang of Calotte. For body weight, a substantial part of the study population was described as 'fat' (17.2%) or 'very fat' (28.8%). This might suggest that a part of the Dutch population was overweight long before the current obesity epidemic, even in less prosperous time and even among individuals from the lowest economical classes.

Introduction

During the last couple of decades, Europeans increased both in length and width [1]. In the Netherlands, male height grew from 1.69 in 1900 (based on measurements of army recruits) up to 1.81 in 2009 (www.cbs.nl). For body weight, the percentage of Dutch adults being overweight increased from 34.9% in 1990 to 46.9% in 2008 (www.cbs.nl). It is often suggested that the current overweight and obesity epidemic in Europe is a trend of the last couple of decades; however this is only an assumption, based on little scientific evidence. In the discussion on anthropometric and physical characteristics of human population in relation to food and disease, historical data is therefore important.

Before the 19th century there is very little information on which to base conclusions on heights and especially weight and other human features. We know what the upper-class Europeans looked like, ate and died from, because they left us paintings, diaries and letters. Much less though is known about the general population. Data on height among the lower classes started to be available from the 18th century, but is mainly based on measurements of army recruits, and so concerns only young and reasonably healthy males [1,2]. As armies usually had a minimum standard for height, the data is likely to be biased towards a longer average in soldiers. Weight was not measured and other physical characteristics were not usually given.

In the Municipal Archives of ‘s-Hertogenbosch, in the Netherlands, we found a list dated the year 1766 with 319 names and very detailed physical descriptions of a group of people who were all members of a criminal gang. Based on these descriptions we were able to study relations between different body characteristics in this population. We investigated the average height and body weight distribution of the gang members to get a historical perspective on body characteristics.

Box 1. 18th century bandits and criminal gangs in the Netherlands.

In the 18th century there was a scare all over Europe over bandits and their gangs: Cartouche (France), Dick Turpin (England) and Schinderhannes (Germany) were known all over Europe [16,17]. In the Netherlands from the late seventeenth century on, there were several gangs of robbers active, some of which like the infamous Bokkenrijders (Riders of billy goats) or Zwartmakers (Black makers) would enter popular folklore for generations [18]. These gangs of robbers did not usually operate in the rich province of densely populated and urbanised Holland, but in the poor southern provinces of North Brabant and Limburg, where farmers were less protected and where they could and did easily skip the borders to other jurisdictions. Members of these criminal groups, however, where not strangers to the communities they robbed, but usually came from the margins of these societies, including members of despised professions like skinners, itinerant workers, showmen at fairs and tinkers, former or deserted soldiers, Jews, and gypsies. Women were also members of these criminal groups. They specialised in picking pockets at markets and fairs, and fencing of stolen goods. As wives, concubines and mothers of the robbers, too, they were an integral part of these criminal gangs.

The authorities took these gangs very seriously. Many dozens of men and women ended their lives at the gallows; because of the active prosecution the judicial records contain rich material on which the characteristics, lives and networks of these people can and have been be reconstructed.

Methods

Study population

In 1766 in the city of s-Hertogenbosch, the main city of North Brabant in the Netherlands, seven members of a gang known as ‘the gang of Calotte’ were arrested, including the leader Calotte, the Flemish born 32-years old Jozef de Vriese, who was also known as Prince Charles, the Seedy Student, and Captain-of- a-Hundred-Rogues. For months, the gang members were questioned, and were pressed to give detailed information about the other members of their group. On the 13th of December 1766 five of the group were broken on the wheel, the other two hung. On the basis of these interrogations, a list was put together with the names and descriptions of 319 people

that belonged to their gang or were closely connected with it. This ‘List of Ruffians, Rogues, Highwaymen and other suspect Persons’ was printed and copies, marked ‘secret’ were sent to other courts of justice in the country, to help identify vagrant undesirables.

Human characteristics

The list with personal descriptions of gang members was not the only list of its kind, but it was unique in the completeness of the physical characteristics provided for each person, and the fact that it concerns a group consisting of men and women, and was of a large age range. The list includes not only age and the colour of hair and eyes, but also height, a typing of body posture and fat distribution, facial condition (smooth-pockmarked) and a detailed description of body posture and physical deformities by which the persons could be identified. For women, the number and age of their children was written down. It must be stressed, though, that height and weight are not measured but taken from the estimates given by the seven robbers who were snitching on accomplishes.

The age of individuals was often given as ‘around 50’ or ‘between 28 and 30’. The informants may not have remembered the precise age of all members; moreover, people of the lower classes themselves did not always know the precise date of their birth. For analyses, we used the around age of a person or took the average of two ages of which a person was described to be in between of.

The height of males was given in feet and thumbs. We conversed this into meters, using the ‘s-Hertogenbosse foot of 0.287 m and the Hondbosche thumb of 2.87 cm that were used in the area at the time [3]. Height of females was never defined in feet and thumbs, but was described in words, like ‘very tall’ or ‘average of height’. We categorized the height of women into 3 groups: short, average and tall.
Bodyweight in both men and women was circumscribed in words as slender, fragile, thin, average weight, reasonably fat, fat and very fat. We categorized individuals into 5 groups: very thin, thin, average, fat and very fat.

A person was pockmarked as its face was described as ‘pockmarked’, ‘pockpitted’ or ‘pocked’. Eye problems included being squint-eyed, having running eyes, having a spot on the eye, being blind and having only one eye. Posture and limb deformities included stooping, bent, walking with knees/feet inwards, old fractures and lame leg.

Data analyses

Population characteristics are described using means and standard deviations (for age and height) and numbers and frequencies (for categorical variables). To study differences in height between different ethnicities we used the Student's t-test. Results were considered statistically significant at 2-sided $P \leq 0.05$. All statistical analyses were performed using SPSS (version PASW Statistics 18).

Results

Study population

The 319 people on the 'list of Ruffians, Rogues, Highwaymen and other suspect Persons' belonging to or closely associated with the 'gang of Calotte' were all between 16 and 65 years of age, with a mean age of 32.8 (±10.3). Of the group, 55.8% was male and age for males and females is presented in table 1 in three age-groups. The average age was 31.5 (±9.2) for males and 34.7 (±11.2) for females. Of the total study population 90% was of Northern-European origin, 8.8% was of Jewish origin and 1.3% was Roma.

Table 1. Age and ethnicity in 319 members from the gang of Calotte in 1766.

	Total (%)	Men (%)	Women (%)
N	319	178	141
Age in years			
16-25	88 (27.6)	56 (31.5)	32 (22.7)
25-35	130 (40.8)	75 (42.1)	55 (39.0)
35-65	101 (31.7)	47 (26.4)	54 (38.3)
Ethnicity			
Northern-European	287 (90.0)	156 (87.6)	131 (92.9)
Jewish	28 (8.8)	21 (11.8)	7 (5.0)
Roma	4 (1.3)	1 (0.6)	3 (2.1)

Height

Height in men was normally distributed, with a mean of 1.58 (±0.05) meters (figure 1). Height in women appeared to be normally distributed, but was never defined in feet and thumbs, so an average height could not be calculated.

Men from Jewish origin were significantly shorter compared to men from Northern-European origin ($p<0.0001$), with an average height of 1.54 (±0.05) meters for Jewish men versus 1.59 (±0.05) for Northern-European men.

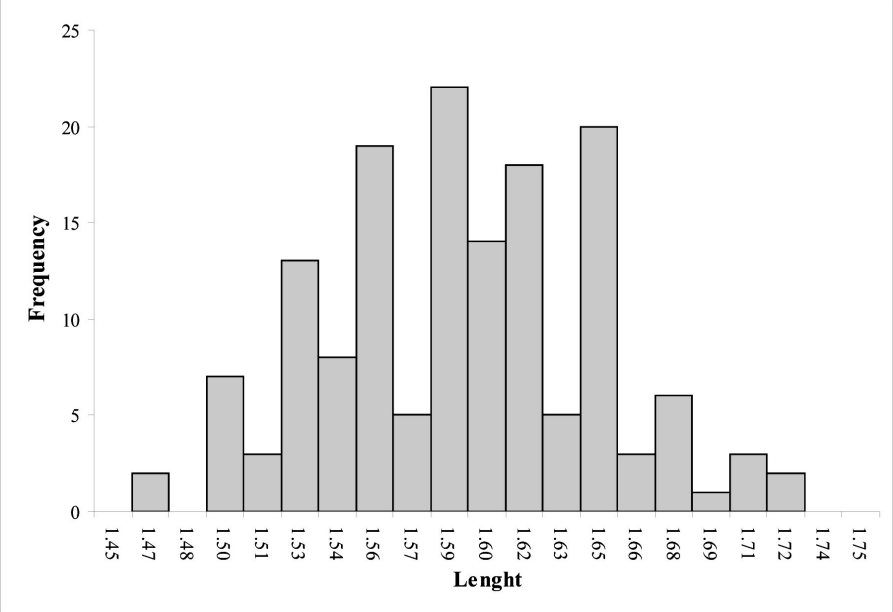


Figure 1. Height distribution 152 males of the gang of Calotte in 1766

Body weight

Among the members of the gang of Calotte, body weight was not normally distributed; 21 individuals were described as very thin (6.6%), 38 as thin (11.9%), 105 as average (32.9%), 55 as fat (17.2%), 92 as very fat (28.8%) and for 8 individuals (2.5%) weight was not described (figure 2).

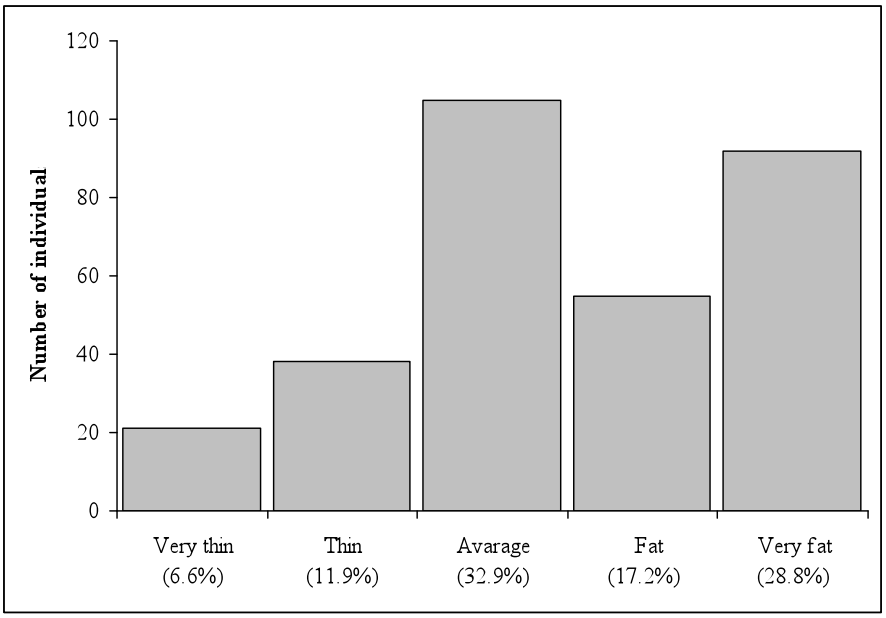


Figure 2. Categories of posture definitions as used in the 1766 dataset of members of the gang of Calotte

Discussion

We studied various body characteristics in a group of 319 individuals of the lowest social classes, living in the18th century. The mean age of the population was 32.8 (±10.3). Height was normally distributed with an average height of 1.58 (±0.05) meters. Body weight was not normally distributed and 32.9% was described as average of weight.

Physical stature is a useful measure of biological well-being; malnutrition and infectious diseases during childhood are associated with a short height in adulthood and a large part of the worldwide variation in height can therefore be attributed to socioeconomic status. The average height of the gang members was 1.58 (±0.05) meter and is therefore respectively 11 cm and 23 cm shorter than the average height of army recruits in 1900 and present-living Dutch males. Therefore, this average height suggests a poor biological well-being for the individuals of the study population. The people on the 1766 list were born between 1720 and 1750, and they

grew up and lived their lives in the middle of the 18th century. In the Netherlands, this was a period of economic decline, and also one of a bad food situation for the lower classes, with recurring epidemics of cattle plague and high grain prices. Additionally, in the mid-18th century smallpox was a major epidemic everywhere in the world and the disease was a leading cause of death in Europe, killing an estimated 400,000 Europeans each year [4]. It is very likely that a substantial part of the gang members had been infected by smallpox earlier in life. Smallpox survivors were most often badly scarred for the rest of their lives, especially on their faces [5]. One in five persons on the list (19.5%), were said to have a pockmarked or poxy face. Also, other permanent complications which are associated with smallpox, like eyes problems and limb deformities are commonly noted to describe specific features for individuals. The smallpox epidemic could have had an effect of the average height in this population. The little information that is available about 18th century heights, suggest a low point in average height in the middle of the 18th century [2]. However, the average height of 1.64 meters in Saxon, German and Scotch military in the same time period is still higher than the average height 1.58 of our study population.

The differences in height between various historical sources could also be due to differences in conversion from foot and thumb into meters. Before the introduction of the metric system, at the end of the 18th century, weights and measures were not standardised; there were numerous differences all over the Netherlands, and also differences in its uses. There was a widely accepted standard for feet and thumbs used in most official papers, which was the 'Rijnlandse' foot of 31.40 cm and a 'Rijnlandse' thumb was 2.61 cm [3]. However, this was used for land surveying, and is not likely to have the measure Calotte and his companions had in mind; for one thing, it would result in an average height of males of 1.71 (±0.05) m, which surely is implausible. Therefore, we think that the average height of 1.58 meters is a valid estimation of the true average.

Men from Jewish origin were significantly shorter compared to men from Northern-European origin ($p<0.0001$), with an average height of 1.54 (±0.05) meters for Jewish men versus 1.59 (±0.05) for Northern-European men. The majority of Jews were from the lowest socioeconomic classes, but Jews were, more than the other robbers, from crowded slums in the cities, mainly from Amsterdam. During childhood, they could therefore have been more exposed to infectious diseases and malnutrition compared to Northern-Europeans, especially if their choice of food was

additionally restricted because of a kosher diet. However, height is also a heritable trait with a heritability of ~0.8, meaning that within populations, about 80% of the variation in height among individuals is due to genetic factors [6]. Therefore, it is reasonable to think that the difference in height between the two ethnicities has in part a genetic base.

Figure 3. Adriaen van Ostade, Dancing Couple. Rijksmuseum Amsterdam

Most people were average of weight, however, many more people (48%) were described as being fatter then average compared to thinner then average (figure 2). Of course, these body weight definitions cannot be compared to BMI distributions of the present-day Dutch population. The BMI of someone being described as ‘fat’ or ‘very fat’ in the 18th century is probably not the same as an obese or morbid obese person in the 21st century. However, it can be argued that a substantial part of the gang of Calotte that was described as ‘fat’ or ‘very fat’, was genuinely overweight. Additional evidence comes from the fact that for these individuals, also other body characteristics concerning fat were mentioned, such as ’a huge belly’, ‘fat legs’ or ‘a round and plump face’. Also, other historical data suggest that the Dutch were not the skinniest people.

In Dutch art, peasants are often depicted as fat. For example, in the many peasants scenes painted by Adriaen van Ostade (1610-1685), the people from the countryside are depicted as of sturdy built and are often eating and drinking (figure 3). This was an artistic tradition which cannot be directly taken as a depiction of reality. In the same period, however, the common people of Holland, especially the farmers, are called fat by foreigners. This is connected to their immoderate consumption of butter and cheese. Also, they are said to drink heavily. The Englishman Fynes Morrison, for example, wrote about the diet of the common people: 'Butter is the first and last dish at the Table ... and thereupon by strangers they are merrily called Butter-mouths. They are much delighted with white meats, and the Bawers [=farmers] drinke milk instead of beer, and [...] passing in boates from city to city for trade, carry with them cheese and boxes of butter for their foode, whereupon in like sort strangers call them Butter-boxes' [7]. 'A Dutch man', sneers a anti-Dutch pamphlet of 1665, written during the times of the Anglo-Dutch naval wars, 'is a lusty, fat, two legged cheese-worm: A creature, that is so addicted to eating butter, drinking fat drink and sliding [=skating], that all the world knows him for a slippery fellow' [7].

It may be questioned whether all common people in the Dutch Republic were well fed and fat, but it is certain that in the diet of the Dutch butter, cheese and milk were very important; as was fish, especially the (fat) herring [8,9]. From the fourteenth century onwards, there was an extensive dairy production in the countryside of Holland; around 1500, more than half of the households of the countryside and even a third of the rural households were involved in dairy production [10]. Most of the butter and cheese was produced for the market and sold all over Europe, but the consumption of butter, cheese and milk within the Netherlands was also substantial [8,9].

It should be noted that the citations on the Dutch being fat were from the 17th century (the golden age) in the rich and dairy producing province of Holland. The members of the gang of Calotte grew up and lived their lives in the middle of the 18th century, a period of economic decline, in the poor south of the Republic. However, it is still remarkably that a large part of the study population was described as 'fat' or 'very fat'. It suggest that a part of the Dutch population was overweight, even in less prosperous time and even among individuals from the lowest economical classes.

Overweight and obesity is a major risk factor for developing type 2 diabetes (T2D). Compared to all other populations with a modern lifestyle, the age-adjusted T2D prevalence in populations of European ancestry is relatively low [11]; T2D prevalence in U.S. Europeans, U.S. Africans, U.S. Hispanics and U.S. Pima Indians is respectively 7.6%, 13%, 17% and 50% [11]. It has been proposed that these differences in T2D susceptibility between European and non-European populations are the genetic and evolutionary consequences of geographical differences in food history [12,13].

The 'thrifty genes theory' hypothesized that the T2D phenotype gives a survival advantage during periods of famine, but is maladaptive in societies with high food abundance [14]. Historical data show that starting from about 1600, European societies became capable to efficiently intervene famine, by redistributing over-abundance grain to areas of food scarcity [15]. This was especially true for the Netherlands, as Amsterdam was the centre and staple market of the international grain trade. Diamond [12] suggested that as a result, Europeans should have undergone an epidemic in T2D starting several centuries before present as a result of the new reliability of sufficient food supplies, and eliminated the most T2D-prone genotypes by processes of natural selection [12]. Our data show that such a scenario could have been possible, as in our study population a substantial part of the population showed characteristics of being overweight and therefore, those individuals were higher at risk for developing T2D.

Conclusion

The average height of 1.58 meters for males suggests a poor biological well-being for the individuals of the gang of Calotte. For body weight, a substantial part of the study population was described as 'fat' (17.2%) or 'very fat' (28.8%). Also other body characteristics concerning fat were mentioned, such as 'a huge belly', 'fat legs' or 'a round and plump face'. This might suggest that a part of the Dutch population was overweight long before the current obesity epidemic, even in less prosperous time and even among individuals from the lowest economical classes.

.

References

1. de Beer H (2001) Voeding, gezondheid en arbeid in Nederland tijdens de negentiende eeuw. Een bijdrage tot de antropometrische geschiedschrijving. Amsterdam: Aksant.
2. Komlos J, Cinnirella F (2005) European heights in the early 18th century. Econ Hum Biol.
3. Verhoeff JM (1983) De oude Nederlandse maten en gewichten. Amsterdam: P.J.Meertens-Instituut.
4. Rutten W (1997) 'De vreselijkste aller Harpijen'. Pokkenepidemieën en pokkenbestrijding in Nederland in de achttiende en negentiende eeuw: een sociaal-historische en historisch-demografische studie. Wageningen: A.A.G. Bijdragen 36.
5. Hays JN Epidemics and pandemics: their impact of human history.
6. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9: 255-266.
7. Meijer-Drees M (1997) Andere landen, andere mensen. De beeldvorming van Holland versus Spanje en Engeland omstreeks 1650 Den Haag.
8. Burema L (1953) Voeding in Nederland van de Middeleeuwen tot de twintigste eeuw. Assen: Van Gorcum.
9. Jobse-van Putten J (1996) Eenvoudig maar voedzaam: cultuurgeschiedenis van de dagelijkse maaltijd in Nederland. Nijmegen: SUN.
10. van Bavel B, Gelderblom O (2009) The economic origins of cleanliness in the Dutch Golden Age. Past and Present 205 41-69.
11. King H, Aubert RE, Herman WH (1998) Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. Diabetes Care 21: 1414-1431.
12. Diamond J (2003) The double puzzle of diabetes. Nature 423: 599-602.
13. Corbett SJ, McMichael AJ, Prentice AM (2009) Type 2 diabetes, cardiovascular disease, and the evolutionary paradox of the polycystic ovary syndrome: a fertility first hypothesis. Am J Hum Biol 21: 587-598.
14. Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14: 353-362.
15. Rotberg RI, Rabb TK, Boserup E (1985) Hunger and history : the impact of changing food production and consumption patterns on society. Cambridge: Cambridge University Press. 336 p. p.
16. Egmond F (1986) Banditisme in de Franse tijd. Profile van de Grote Nederlandse Bende 1790-1799. Zwolle: De Bataafsche Leeuw.
17. Egmond F (1993) Underworlds. Organized Crime in the Netherlands 1650-1800. Cambridge: Polity Press.
18. Blok A (1991) De Bokkenrijders. Roversbenden en geheime genootschappen in de landen van Overmaas (1730-1774). Amsterdam: Prometheus.

Chapter 11

General discussion

General discussion

The studies described in this thesis aim not only to study *how* humans are susceptible to obesity and T2D but also *why*. Although *why* and *how* may sound similar in the perspective of disease risk, they are actually very distinct. By studying the question *how*, the aim is to find the proximate cause of the disease, i.e. factors that are closest to or immediately responsible for causing the trait. For obesity and T2D, proximate causes could be genetic and lifestyle risk factors or an underlying molecular mechanisms of insulin resistance. In contrast, by studying the question *why*, the aim is to find the higher-level ultimate cause of the disease, i.e. the evolutionary explanation of the trait.

A truly full explanation of a complex disease needs both a proximate explanation of how things work and a complementary evolutionary explanation of why it got that way. However, only very few researchers systematically apply evolutionary biology into medical science [1,2]. Scientists in the field of evolutionary medicine, like Randolph Nesse and George Williams, advocate the application of the evolutionary theory to understand health and disease, which they name Darwinian medicine [3]. Darwinian medicine provides a complementary approach to the present mechanistic explanations that dominate medical science and education [4].

In this chapter I will discuss human obesity and T2D from an evolutionary perspective.

Natural selection

Natural selection is the key mechanism of evolution. The theory of evolution, as proposed by Charles Darwin in 1859 [5], has some basic principles:

1. There is variation in traits among individuals within a species;
2. Traits are heritable;
3. There is differential reproduction; i.e. those individuals who best fit their environment are likely to survive, reproduce, and pass the traits to the next generation.

The process of natural selections makes heritable traits become more common in a population over successive generations when these traits make it more likely for an organism to survive and successfully reproduce.

Darwin never found out exactly how organisms pass traits to their offspring. He died in 1882 and only in the mid-twentieth century, after the discovery of chromosomes by Theodor Boveri, genes by Gregor Mendel and the DNA double helix by Watson and Crick (using x-ray diffusion data collected by Rosalind Franklin) [6], the importance of genetic variation to evolution was revealed.

Natural selection acts on phenotypes, but this selection is recorded in genotypes and genomic regions. It is the main driver resulting in the diversity of species and their genomes. Because of genetic variation, individuals have varying capacities to survive and reproduce in different environments. Genetic variation that reduces the fitness of the carrier is subject to negative selection, whereas genetic variation that increases fitness undergoes positive selection.

Clash between our genes and modern life

Evolution is the foundation for biology and biology is the basis for medicine. Seeing human organisms as machines that are optimally designed and engineered is misleading. Instead, human traits are full of compromises, which are shaped by natural selection to maximize reproduction and not to maximize health. There are many unavoidable tradeoffs and constraints in human biology [3].

Therefore, the idea that common complex traits are caused by a few defective genes is incorrect. The recent findings from genome-wide association studies (GWAS) have made it clear that complex diseases are much more heterogeneous and polygenic than previously believed [7-11]. An evolutionary view suggests that many genetic variants interact with environmental factors and other genes during development to influence disease phenotypes. Human genetic variation that increase disease resistance can have costs and some variants that increase disease vulnerability can have benefits at some stages in life. This view can in general help to explain why complex diseases are so prevalent and difficult to prevent.

Because biological evolution is much slower than cultural changes, many diseases arise from a clash between our genes and modern life. For instance, the increase in human obesity and T2D is largely due to a mismatch between adaptive biological characteristics of our species and the modern environment, which has changed dramatically over a relatively short period of time from the environment under which we evolved. This concept is very important in understanding obesity and

T2D [12]. We carry our human past with us in our genes and this affects how the body reacts to the environment.

Theories on the genetic basis of obesity and type 2 diabetes

Although T2D is a severe medical condition, it is quite common, with a prevalence of 2.8% worldwide and a prevalence of 7.8% in the USA in 2007 (http://www.diabetes.org) The concordance rate of T2D in monozygotic twins is 76% compared to 40% in dizygotic twins, providing convincing evidence that genetic factors contribute to the development of T2D. One interesting question is why the phenotypes of T2D and obesity, the main risk factor for T2D, are so common despite their negative effects on human health. Like all species, *Homo sapiens* has been shaped by evolutionary processes and the fact that so many people are susceptible to developing T2D and obesity suggests that genes underlying these traits may have been favored by the process of natural selection. In trying to explain this observation, James Neel proposed the 'thrifty genotype theory' in 1962 [13], according to which our early ancestors frequently suffered periods of prolonged famine, during which a survival and/or reproductive advantage would have been conferred by genes favouring the economical use and storage of energy, the so-called thrifty genes. The theory focuses on the efficient use of glucose as a biological fuel and suggests that evolutionary pressure to preserve glucose for use by the brain during starvation led to a genetic propensity towards insulin resistance in peripheral tissue. In the Western world, food is, in general, easily available and plentiful, so these thrifty genes are maladaptive in modern society and may now contribute to the widespread susceptibility for T2D and obesity [14].

Although the thrifty genes hypothesis is popular, it is also controversial and has been discussed for decades in many scientific papers [15-25]. The most powerful argument against the theory came from anthropological studies that suggested that during the past 2.5 million years of human history, famines were not sufficiently frequent and severe to cause evolutionary pressure [26]. Also paleoanthropologists questioned whether ancient foraging people were truly subject to cycles of 'feast and famine' [19]. There is now abundant evidence that primitive foraging populations had access to a wider range of foods and were taller, had better dentition and less infectious disease than the agricultural populations, which started to develop around

10,000 years ago [27]. Neel himself revised his hypothesis after the distinction between type 1 and type 2 diabetes became clear; it was tissue-specific resistance to the action of insulin, particularly by skeletal muscle, rather than a rapid insulin response, which was the hallmark of his 'thrifty genotype' [28]. Neel then hypothesized that when calorie intake was low, muscle insulin resistance would enable the economic use of metabolic fuel by reducing insulin-stimulated uptake of glucose in skeletal muscle, disturbing the suppression of hepatic glucose synthesis from lipid precursors. By 1998 Neel admitted that his hypothesis, even as revised, entailed an oversimplification of the impacts of modern nutrition on the "fine old genes" involved in energy homeostasis [14]. Despite the doubts of Neel himself regarding the theory, the thrifty gene hypothesis remained very popular and is cited frequently, even today (Figure 1).

Another popular explanation for the high frequency of individuals susceptible to T2D and obesity is the ‘predator release theory’, proposed by Speakman [22]. According to his hypothesis, obesity and T2D risk alleles rose in allele frequency in the population due to mutation and random drift. Carrying around large fat reserves may enhance the probability of surviving a period of food shortage, but could in the meantime increase the probability of being killed by a predator. Therefore, Speakman postulates a stabilizing selection for body fatness. During the early period of human evolution (6-2 million years ago), humans were preyed on by large predatory animals. Absence of predation in more recent times led to a change in the population distribution of body fatness due to random mutation and drift. This hypothesis fits in the idea that common variants may only explain a fraction of the genetic risk for complex traits like obesity and T2D, while the remainder of the risk alleles are rare.
There are several alternative theories that try to explain the genetic basis of T2D and obesity [15-25]. Till now, all these hypotheses have been speculative and there was no consensus in the field. Box 1 summarizes various theories on obesity and T2D.

Box 1. Hypothesis on the origin and the current epidemic of obesity and T2D

The thrifty gene hypothesis
James Neel proposed the 'thrifty genotype theory' in 1962 [13], according to which our early ancestors frequently suffered periods of prolonged famine, during which a survival and/or reproductive advantage would have been conferred by genes favouring the economical use and storage of energy, the so-called thrifty genes. The theory focuses on the efficient use of glucose as a biological fuel and suggests that evolutionary pressure to preserve glucose for use by the brain during starvation led to a genetic propensity towards insulin resistance in peripheral tissue. In the Western world, food is, in general, easily available and plentiful, so these thrifty genes are maladaptive in modern society and may now contribute to the widespread susceptibility for T2D and obesity.
The drifty gene hypothesis
Obesity and T2D risk alleles rose in allele frequency in the population due to mutation and random drift. Speakman proposed the 'predator release theory' where he postulates a stabilizing selection for body fatness [22]. Carrying around large fat reserves may enhance the probability of surviving a period of food shortage, but could in the meantime increase the probability of being killed by a predator. During the early period of human evolution (6-2 million years ago), humans were preyed on by large predatory animals. Absence of predation in more recent times led to a change in the population distribution of body fatness due to random mutation and drift.
The thrifty phenotype
Malnutrition in the environment within the womb during development induces thrifty mechanisms, because it predicts a future of starvation. This hypothesis is based on the finding that low birth weight is associated with an increased risk of T2D.
The behavioral switch hypothesis
Insulin resistance evolved as a socio-ecological and socio-nutritional adaptation rather than thriftiness. Insulin resistance is an adaptation to (i) a transition in reproductive strategy from 'a large number of offspring with little investment per individual' to 'a smaller number of offspring with more investment per individual' and (ii) a transition from a muscle dependent lifestyle to a brain dependent lifestyle [23].
The thrifty epigenotype hypothesis
The capacity for efficient storage and use of energy is an ancient and complex trait

and has become robust against genetic mutations. Genetic variants play a minor role in the etiology of obesity and T2D. Instead, disease susceptibility is predominantly determined by epigenetic variation and corresponding epigenotypes have the potential to be inherited across generations [21].

The 'genetically unknown foods' hypothesis

Theoretically, it can be expected that thrifty genes are more common in Europeans, as Europe has low abundant vegetation due to long harsh winters and has almost continuously been devastated by wars, characterized by famine and starvation. In real, the prevalence of T2D in Europeans is low. According to the 'genetically unknown foods' hypothesis, proposed by Baschetti (1998) [15,26], humans are genetically still unequipped for some foods that were unavailable to our hunter-gatherer ancestors. These foods, that are currently commonly available in western diets, may be responsible for common complex traits, like obesity and T2D. The low T2D prevalence in Europeans reflects their moderate adaptation to those foods, which has been achieved through natural selection in the last millennium.

Cryoprotective evolutionary adaptation

Previous animal research showed that high concentrations of glucose, glycerol and other sugar derivatives depress the freezing point of body fluids and prevent the formation of ice crystals in cells through cooling, thus acting as a cryoprotectant for vital organs as well as muscle tissue. Moalem et al. hypothesize that factors predisposing to elevated levels of glucose may have been selected for as adaptive measures in exceedingly cold climates [49]. However, critics of this theory argue that humans are unable to allow their body temperatures to cool much below 30oC before they experience heart failure and death. Hence, the importance of elevated blood glucose levels for cryoprotection as the climate cools seems at best marginal.

The 'fertility first' hypothesis

Polycystic ovary syndrome (PCOS) is a heritable form of ovarian infertility and is characterized by a long history of chronic anovulation in association with insulin resistance and androgen excess. Women with PCOS have a four times increased prevalence of T2D. Corbett et al. introduced the 'fertility first' hypothesis, proposing that the PCOS, T2D and the metabolic syndrome are modern phenotypic expressions of a metabolic genotype attuned to the dietary and energetic conditions of our early ancestors [18]. This metabolic 'fertility first' genotype, rather then the 'thrifty'

genotype persisted at high prevalence because it conferred a fertility advantage in periods of food shortage. The hypothesis predicts that the increasing rate of T2D will be tempered by natural selection against the underlying genes, driven by sub- or infertility.

The unbalanced 'autonomic nervous system' hypothesis

During the last century, life has changed dramatically in industrialized countries. Food has become abundant and the necessity for physical effort became considerably reduced. Additionally, physical activity does not need to coincide with a day and night rhythm. As a result, the environment, sensed by the brain, has become metabolically flattened and arrhythmic. Kreier et al. hypothesize that in such conditions the human brain loses its feeling for internal and external rhythms and propose an unbalanced and arrhythmic autonomic nervous system as a major cause of the metabolic syndrome [50].

Although the theories on the origin of obesity and T2D differ, most theories assume that the high prevalence of the disorders must be sustained by some compensating advantage that outweighs the morbidity and mortality. In this thesis (Chapter 8) we investigated whether the known genetic variants underlying obesity and T2D [8-11,29] have indeed been favoured by positive natural selection, as is suggested in the 'thrifty gene hypothesis' and other theories. When a genetic variant is under positive selection, it increases in frequency in a population and this leaves a 'signature' or pattern in the human genome [30]. These signatures can be identified by comparison with the background distribution of genetic variation in humans, which is generally argued to have evolved largely under neutrality [30-32]. In genome-wide genetic data from Europeans, we did not find signs of positive selection around the currently known T2D and obesity risk alleles and our findings therefore do not support the theory that these alleles had a survival advantage in the recent (<30,000 years ago) past. However, as we discussed in chapter 8, our data was not complete enough to reject the theory either. First of all, 'thrifty genes' that cause susceptibility for T2D and obesity could have reached fixation in the population (i.e. all individuals of the population carry the same risk allele), and they therefore cannot be picked up by GWAS using case and control data. Older selection pressure could also have acted on these variants so that their signature is no longer visible in the

genome. Secondly, although the GWAS have improved our understanding of the genetic basis of T2D and obesity, we can still only explain around 10% of the genetic risk for these traits [33]. Thus, the majority of T2D and obesity loci are still unknown and cannot be tested for signatures of selection.

However, what we did find was that some risk variants for T2D and obesity show suggestive signs of negative selection in our European data, indicating that the risk allele for the traits had a survival disadvantage in Europeans.

Worldwide T2D prevalence

King et al. estimated the age-adjusted prevalence of T2D in all countries in the world [34]. Table 1 shows the standardized T2D prevalence in several human populations, taking differences in age distributions of the various populations into account. [34]. For developing countries, T2D prevalence was calculated separately for urban, rural and/or traditional living populations, as these populations have distinct lifestyle and dietary habits.

Table 1. T2D prevalence in different human populations

Population	Region	% T2D prevalence
Europeans	The Netherlands	2
	U.K.	2.1
	U.S.	7.6
	Australia	8
Native Americans	Chile Mapuche	1
	U.S. Hispanics	17
	U.S. Pima Indians	50
Aboriginals Australia	Traditional	0
	Westernized	23
Africans	Rural Tanzania	1
	Urban South Africa	8
	U.S. Africans	13
Asia	Rural China	0
	Urban Singapore	9
	Rural India	0
	Urban India	12

Prevalence of diabetes in the world was estimated to be 4% in 1995 and to rise to 5.4% by the year 2025. However, age-adjusted T2D prevalence differs enormously among human populations. The explosion in T2D prevalence is occurring especially in developing countries, at about 50% per decade. The epidemic is just beginning in the world's two most populated countries, India and China, and therefore by the year 2025 more than half of the world's diabetics will be Asians. Furthermore, table 1 shows that traditionally living human populations have a much lower T2D prevalence compared to populations with the same ethnic background that live in urban environments.

Compared to all other populations with a modern lifestyle, the T2D prevalence in European populations is relatively low, even though Europeans are the richest and best-fed humans in the world. Indeed, Europeans (living in Europe and through the rest of the world) are the 'inventors' of the Western lifestyle [26]. Although the number of European individuals with T2D is rising, as it is in all population, the prevalence of the disease is still lower in Europeans than in any other non-European population [34]. In chapter 8 we show that some protective variants for T2D and obesity do show suggestive signs of positive selection in European genetic data, suggesting an advantage for the obesity and T2D for the protective allele. It can be argued that Europeans are already adapting genetically to a Western diet by purging genetic variants leading to type 2 diabetes and obesity. The lower frequency of T2D in Europeans compared to other ethnic groups which are now adopting a 'Westernized' diet and lifestyle supports this hypothesis.

Jared Diamond proposed that the genetic and evolutionary consequences of geographical differences in food history may provide the answer [35]. One interesting theory is that starting from about 1600, European societies became capable to efficiently intervene famine, by redistributing over-abundance grain to areas of food scarcity. This was possible because of well-organized state polities and the increasingly efficient food transport by land and by sea. Therefore prolonged famines gradually disappeared in Europe, starting from about 1650 in the Netherlands and in Great Britain and proceeding in the late 1800s in southern France and Italy. Furthermore, increasingly diversified agriculture broadened the base of European agriculture, thereby reducing the risk of starvation from failure of a single crop. As a result, the view of Jared Diamond on European food history is that several centuries before present, Europeans should have undergone an epidemic in T2D that resulted

from the new reliability of sufficient food supplies and eliminated most diabetes-prone bearers of the thrifty genotype [35]. Therefore, Europe's food abundance would have increased gradually over the course of several centuries and the result, between the 1400s and 1700s, would have been a slow rise in T2D prevalence.

The increase in food availability in previous centuries is also shown by Wansink and Wansink [36]. They recently demonstrated that the portion size in painted meals of Jesus Christ's last supper, generally increased with time. Over the last millennium, the relative sizes of the main dish, bread and plates have linearly increased. This supports that the production, availability, abundance and affordability of food in Europe has been increased. In chapter 10 from this thesis, we showed weight distribution in an 18th century cohort population. Many of the people were estimated to have a normal weight, however significantly more individuals are described heavier than average compared to lighter than average. People were often described to have a huge stomach or fat legs. Also, weight was increased with age, just as it is in modern populations.

Therefore, it can be suggested that adaptation to dietary habits resulted in purging genetic variants leading to T2D and obesity in European populations.

Adaptation to diet

A major selective force during human evolution has been diet. Modern humans originated in Africa within the past ~200,000 years and then spread across the rest of the earth within the past 100,000 years [37,38,39]. Our early ancestors were hunter-gatherers and only in the relatively recent past (< 10,000 years ago) have humans developed plant and animal domestication (e.g. agriculture and pastoralism); this transition was accompanied by major changes in diet for most human populations [40]. Analysis of the diets of modern hunter-gatherer populations indicates they obtain 19-35% of their energy from protein, 22-40% from carbohydrate and 28-58% from fat [41]. In comparison, adults in the western world obtain some 16% of their energy from protein, nearly 50% from carbohydrates, approx. 34% from fat and about 3% from alcohol.

Large variations in caloric and macronutrient intake and preference between individuals have been reported and these food intake patterns show a strong heritability, indicating a genetic basis [42]. There are also large differences in food

intake and percentage of nutrient-specific energy intake among different ethnic groups [35]. These ethnic differences in total and nutrient-specific energy intake might be caused by natural selection of mutations providing an advantage for a particular environment or type of diet. Diet as a driver of selection can be a condition beyond human control, like availability of a particular food source in the environment, but also culture is increasingly considered to be a strong driver of selection in humans [43]. One famous example how diet shaped the human genome is the co-evolution of diary farming and adult lactose tolerance. The ability to digest lactose disappears after early childhood in most humans; however in some populations the lactase enzyme that is essential to break down lactose stays active in adulthood. Different regulatory variants nearby the lactase gene are responsible for the lactose tolerance in northern Europe and African pastoralist populations, demonstrating convergent adaptation to drinking milk. Dairying created the selection pressure that drove alleles for lactose tolerance to high frequency in these populations [44]. In chapter 5 we show that derived alleles in *NPY1R* and *NPY5R* are associated with lower carbohydrate intake. One of these variants shows the hallmark of recent selection in Europe [45]. Our data suggest that lower carbohydrate intake gave a survival advantage in Europeans since the agricultural revolution. This advantage could lie in overall health benefits, because lower carbohydrate intake, consuming meals with a low GI and GL, and/or moderate alcohol consumption, are known to be associated with a lower risk of chronic diseases.

It is likely that there are many more dietary habits that caused allele frequency shifts in populations, yet to be identified. Recent genome-wide scans for signatures of selection in several human populations pinpointed towards multiple genes that have been favored by recent natural selection [31,32]. The great challenge is now to find the connection between genotype, phenotype and drivers of selection like diet and other mechanisms underlying obesity and T2D [43].

Polygenic adaptation

As I stated earlier, obesity and T2D are traits that are affected by a large number of loci [33]. Most methods to find signatures of selection focus on models of selection at one locus. This is in contrast to classical models of natural and quantitative genetics, where it is assumed that most traits are influenced by variation at many loci [46]. If

the environment shifts so that there is a new phenotypic optimum, then the population will adapt by allele frequency shifts at many loci. Once the phenotype in the population matches the new optimum, selection will weaken. This means that it may be very common for selection to push alleles upwards in frequency but generally not to fixation [47]. In principle, this type of process could allow very rapid adaptation, yet be difficult to detect using most current population genetic methods. If there was a sudden onset of strong selection for or against obesity and/or T2D, a rapid shift in average BMI or T2D prevalence could be expected. However, the response to selection would be generated by modest allele frequency shifts at many loci. Even with strong selection and a strong phenotypic response, standard methods for detected selective sweeps would have limited power. This mechanism of polygenic adaptation would allow rapid phenotypic adaptation, without necessarily generating any large differences in allele frequencies between populations.

The challenge for the future is to develop methods for studying polygenic adaptation, to be able to study whether complex diseases as obesity and T2D truly were subjects of selection in the human history.

Future perspective

How should we study complex diseases, like obesity and T2D from an evolutionary point of view? Which research questions should we make and how could we test this. For obesity and T2D, I propose the following research questions and approaches:

1. Could obesity and T2D characteristics influence reproductive success? In chapter 9, we studied whether T2D patients have earlier in life reproductive success. Generally, measures of sub- and infertility did not independently predict subsequent development of T2D. However, most T2D patients were diagnosed after menopause. Therefore, future studies should further investigate the association between reduced fertility and premenopausal T2D.
2. Do T2D and obesity risk alleles have yet-to-be-identified advantageous that outweigh their costs?
3. Do historical legacies account for disease susceptibility in humans? Or do novel environmental factors mainly contribute to the traits?
4. Did different human societies adapt to different diets? And if different human populations are genetically adapted to different dietary patterns; can the food

guide pyramid and the food balance wheel, which suggests optimal daily nutrition guidelines for each food category, be applied to all human populations? The African continent contains the highest amount of genetic, phenotypic, cultural, and linguistic diversity in the world [37,48]. African populations have distinct diets and lifestyle, including hunter-gatherers, pastoralists, agriculturalists, and agro-pastoralists. They likely have experienced local adaptation and have population or region-specific genetic variation. Therefore, African populations are very suitable for studying how the human genome has been shaped by evolutionary processes, like diet. More research on diet and human evolution should be done in Africans population.

References

1. Nesse RM, Bergstrom CT, Ellison PT, Flier JS, Gluckman P, et al. (2010) Evolution in health and medicine Sackler colloquium: Making evolutionary biology a basic science for medicine. Proc Natl Acad Sci U S A 107 Suppl 1: 1800-1807.
2. Nesse RM, Stearns SC, Omenn GS (2006) Medicine needs evolution. Science 311: 1071.
3. Nesse RM, Williams GC (1994) Why we get sick : the new science of Darwinian medicine. New York: Times Books. xi, 291 p. p.
4. Stearns SC, Nesse RM, Govindaraju DR, Ellison PT (2010) Evolution in health and medicine Sackler colloquium: Evolutionary perspectives on health and medicine. Proc Natl Acad Sci U S A 107 Suppl 1: 1691-1695.
5. Darwin C, Matthews JW (1859) The origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life. London ; New York: Unit Library. viii, 487 p. p.
6. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737-738.
7. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41: 18-24.
8. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40: 638-645.
9. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, et al. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316: 1336-1341.
10. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331-1336.
11. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341-1345.
12. Power ML, Schulkin J (2009) The Evolution of Obesity. Baltimore: The John Hopkins University Press.
13. Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14: 353-362.
14. Neel JV (1999) The "thrifty genotype" in 1998. Nutr Rev 57: S2-9.
15. Baschetti R (1999) Genetically unknown foods or thrifty genes? Am J Clin Nutr 70: 420-421.
16. Belsare PV, Watve MG, Ghaskadbi SS, Bhat DS, Yajnik CS, et al. (2010) Metabolic syndrome: aggression control mechanisms gone out of control. Med Hypotheses 74: 578-589.
17. Colagiuri S, Allen JS, Brand Miller JC, Cheer SM (1997) The thrifty genotype hypothesis. Diabet Med 14: 504.
18. Corbett SJ, McMichael AJ, Prentice AM (2009) Type 2 diabetes, cardiovascular disease, and the evolutionary paradox of the polycystic ovary syndrome: a fertility first hypothesis. Am J Hum Biol 21: 587-598.
19. Cordain L, Miller J, Mann N (1999) Scant evidence of periodic starvation among hunter-gatherers. Diabetologia 42: 383-384.
20. Reaven GM (1998) Hypothesis: muscle insulin resistance is the ("not-so") thrifty genotype. Diabetologia 41: 482-484.
21. Stoger R (2008) The thrifty epigenotype: an acquired and heritable predisposition for obesity and diabetes? Bioessays 30: 156-166.
22. Speakman JR (2007) A nonadaptive scenario explaining the genetic predisposition to obesity: the "predation release" hypothesis. Cell Metab 6: 5-12.
23. Watve MG, Yajnik CS (2007) Evolutionary origins of insulin resistance: a behavioral switch hypothesis. BMC Evol Biol 7: 61.
24. Prentice AM, Rayco-Solon P, Moore SE (2005) Insights from the developing world: thrifty genotypes and thrifty phenotypes. Proc Nutr Soc 64: 153-161.
25. Miller JC, Colagiuri S (1994) The carnivore connection: dietary carbohydrate in the evolution of NIDDM. Diabetologia 37: 1280-1286.
26. Baschetti R (1998) Diabetes epidemic in newly westernized populations: is it due to thrifty genes or to genetically unknown foods? J R Soc Med 91: 622-625.
27. Armelagos GJ, Cohen MN, State University of New York College at Plattsburgh., Wenner-Gren Foundation for Anthropological Research. (1984) Paleopathology at the origins of agriculture. New York: Academic Press. xx, 615 p p.

28. Koberling J TR, editors (1982) The Genetics of Diabetes Mellitus. Amsterdam: Academic Press.
29. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, et al. (2010) Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet 42: 142-148.
30. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.
31. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
32. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826-837.
33. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-251.
34. King H, Aubert RE, Herman WH (1998) Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. Diabetes Care 21: 1414-1431.
35. Diamond J (2003) The double puzzle of diabetes. Nature 423: 599-602.
36. Wansink B, Wansink CS (2010) The largest Last Supper: depictions of food portions and plate size increased over the millennium. Int J Obes (Lond).
37. Reed FA, Tishkoff SA (2006) African human diversity, origins and migrations. Curr Opin Genet Dev 16: 597-605.
38. Campbell MC, Tishkoff SA (2010) The evolution of human genetic and phenotypic variation in Africa. Curr Biol 20: R166-173.
39. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, et al. (2009) The genetic structure and history of Africans and African Americans. Science 324: 1035-1044.
40. Diamond J (2002) Evolution, consequences and future of plant and animal domestication. Nature 418: 700-707.
41. Eaton SB, Konner M (1985) Paleolithic nutrition. A consideration of its nature and current implications. N Engl J Med 312: 283-289.
42. Rankinen T, Bouchard C (2006) Genetics of food intake and eating behavior phenotypes in humans. Annu Rev Nutr 26: 413-434.
43. Laland KN, Odling-Smee J, Myles S (2010) How culture shaped the human genome: bringing genetics and the human sciences together. Nat Rev Genet 11: 137-148.
44. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31-40.
45. Elbers CC, de Kovel CG, van der Schouw YT, Meijboom JR, Bauer F, et al. (2009) Variants in neuropeptide Y receptor 1 and 5 are associated with nutrient-specific food intake and are under recent selection in Europeans. PLoS One 4: e7070.
46. Daly AK, Day CP (2009) Genetic association studies in drug-induced liver injury. Semin Liver Dis 29: 400-411.
47. Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20: R208-215.
48. Tishkoff SA, Williams SM (2002) Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet 3: 611-621.
49. Moalem S, Storey KB, Percy ME, Peros MC, Perl DP (2005) The sweet thing about Type 1 diabetes: a cryoprotective evolutionary adaptation. Med Hypotheses 65: 8-16.
50. Kreier F, Yilmaz A, Kalsbeek A, Romijn JA, Sauerwein HP, et al. (2003) Hypothesis: shifting the equilibrium from activity to food leads to autonomic unbalance and the metabolic syndrome. Diabetes 52: 2652-2656.

Chapter 12

Summary

The incidence of type 2 diabetes (T2D) is rising rapidly worldwide and there are already more than 180 million diabetic subjects. T2D risk factors include ethnic background, age, hypertension, overweight, increased abdominal fat, and lack of physical exercise. Obesity is considered to be the most important risk factor for T2D and the main one driving the current epidemic as 90% of T2D patients are obese. Worldwide obesity has also reached epidemic proportions, with 300 million adults classified as clinically obese. T2D and obesity are multifactorial disorders in which both genetic and non-genetic (environmental and lifestyle) factors play a role. In the present thesis we focused on (I) evaluating alternative methods to find candidate genes for T2D and obesity, (II) studying genetic and environmental risk factors for T2D and obesity, and (III) studying the origin of the high prevalence of T2D and obesity in modern societies.

Chapter 2 is a review that explores the genes recently identified for T2D and obesity by genome-wide association (GWA) studies and evaluates their functions in an effort to determine whether there is any support for the hypothesis that T2D and obesity share some underlying mechanisms. By evaluating the function of currently known risk alleles it seems that the susceptibility genes for obesity are involved at the start of the trait (energy imbalance) and those for T2D at a later stage of the disease (beta-cell defect). It is suggested that the shared genetic effect may be smaller than we thought or obesity could simply be a non-genetic risk factor for T2D because it provokes insulin resistance. Discovering more obesity and T2D genes will provide a broader insight into the shared disease pathology.

Part I of the thesis discusses alternative gene-hunting strategies for T2D and obesity. Data from linkage studies do not directly indicate the gene of interest and identifying a potential gene is usually rather difficult as linkage intervals can contain dozens to hundreds of candidate genes. To identify the gene of interest, a dense map of single nucleotide polymorphisms (SNPs) encompassing the candidate region needs to be tested for genetic association in very large case-control studies. An attractive alternative strategy is to first prioritize the positional candidate genes based on the function of the individual genes using bioinformatics tools. In **chapter 3**, all published genome scans for T2D and obesity (till 2006) were compared and five overlapping chromosomal regions for both diseases (encompassing 612 candidate genes) were identified. By analyzing these five susceptibility loci for T2D and obesity using six freely available bioinformatics tools for disease gene identification, 27

functional candidate genes were pinpointed that are involved in eating behaviour, metabolism and inflammation. These genes may reveal a molecular link between the two disorders. Although these bioinformatics tools for disease gene prioritisation are attractive, they still suffer from several limitations. As network and pathway tools make use of functional information from gene and protein databases, they are biased towards the well-studied genes, interactions and pathways. This is clearly shown in **chapter 4.1**, where we evaluated whether pathway classification analysis can help prioritize the biological pathways most likely to be involved in the disease etiology. Instead of focusing on SNPs with the highest statistical significance, we took advantage of prior biological information and tried to detect overrepresented pathways in genome-wide association (GWA) data. We show multiple differences in outcome between pathway tools analyzing the same dataset. Furthermore, analyzing randomly selected SNPs always results in significantly overrepresented pathways, large pathways have a higher chance of becoming statistically significant and the bioinformatics tools used in this study are biased towards detecting well-defined pathways. In **chapter 4.2**, we additionally describe several problems that we encountered using these pathway methods. We would like to emphasize that the limitations of pathway-based analyses in GWA data should be kept in mind when drawing conclusions based on overrepresented pathways.

Part II of the thesis investigates genetic and lifestyle risk factors for obesity and T2D. Obesity is the result of an imbalance between energy intake and energy expenditure. There is a large variation in caloric intake and macronutrient preference between individuals and between ethnic groups, and these food intake patterns show a strong heritability. One major player in energy homeostasis is the appetite-stimulating hormone neuropeptide Y, in which the stimulatory capacity may be mediated by the neuropeptide Y receptors 1, 2 and 5 (NPY1R, NPY2R and NPY5R). In **chapter 5** we assessed the association between variants in the *NPY1R, NPY2R* and *NPY5R* genes and nutrient intake in a cross-sectional, single-centre study of 400 older men. Our data show that derived alleles in *NPY1R* and *NPY5R* are associated with lower carbohydrate intake, mainly because of a lower consumption of mono- and disaccharides. We also show that carriers of these derived alleles, on average, consume meals with a lower glycaemia index and glycaemia load and have higher alcohol consumption. One of these variants shows the hallmark of recent selection in Europe. Our data suggest that lower carbohydrate intake, consuming meals with a low

glycaemia index and glycaemia load, and/or higher alcohol consumption, gave a survival advantage in Europeans since the agricultural revolution. This advantage could lie in overall health benefits, because lower carbohydrate intake, consuming meals with a low glycaemia index and glycaemia load, and/or higher alcohol consumption, are known to be associated with a lower risk of chronic diseases.

Recently, GWA studies have identified several common loci for obesity. In **chapter 6** we investigated whether the recently reported obesity loci are more specifically associated with abdominal obesity—an important contributor to increased morbidity and mortality, independent of the total amount of body fat. Additionally, we explored the effect of variation in the obesity susceptibility loci on dietary energy and macronutrient intakes in 1700 healthy Dutch women. Our data show that the obesity susceptibility loci are not specifically associated with abdominal adiposity, but merely represent loci associated with general obesity. We did find an association with dietary macronutrient intake and several obesity susceptibility loci. SNPs in or near *SH2B1*, *KCTD15,* and *NEGR1* were associated with total fat, saturated, and monounsaturated fat intakes. SNPs in or near *KCTD15* and *MTCH2* were associated with total carbohydrate, mono- and disaccharide, and polysaccharide intakes. These results suggest that the new obesity loci might play a role in the choice and preference of specific macronutrients.

Another lifestyle factor that is linked to T2D risk is parity. Having 4 or more children is found to be associated with increased T2D risk in women. It has been suggested that the observed associations are mediated by body mass as child bearing is associated with the increase in body mass. However, the degree to which waist and BMI affect the association between parity and T2D remains unclear. Apart from the association with T2D through body mass, another possible biological mechanism for the association between parity and T2D risk in women could be through reduced oestrogen exposure. It has been hypothesized that pregnancy permanently resets ovarian function, leading to a reduced lifetime exposure to oestrogen. As the first pregnancy is the start of this change in level of exposure to estrogens, age at first full-term pregnancy is a good marker for the duration of this reduced exposure. Therefore, in **chapter 7**, we assessed the association between both parity and age at first full-term pregnancy with the risk of T2D in the large Prospect-EPIC cohort comprising of 17,357 Dutch women. Our results show that the association of parity and T2D risk was found to be mediated by increased body mass. We show an increased weight,

BMI and waist with each additional child. Furthermore, age at first full-term pregnancy was inversely associated with the subsequent development of T2D. Body mass attenuate the association, but could not fully account for it. We argue that a possible underlying mechanism could be that young age at first full-term pregnancy leads to long-term reduced estrogens exposure subsequently leading to reduced β-cell function.

Part III of the thesis studies the origin of the high prevalence of T2D and obesity in modern societies. Although T2D is a severe medical condition, it is quite common. One interesting question is why the phenotypes of T2D and obesity, the main risk factor for T2D, are so common despite their negative effects on human health. Like all species, Homo sapiens has been shaped by evolutionary processes and the fact that so many people are susceptible to developing T2D and obesity suggests that genes underlying these traits may have been favoured by the process of natural selection. There are several theories that try to explain the genetic basis of T2D and obesity of which 'the thrifty gene hypothesis' is most known. In brief, in this theory it is hypothesized that the T2D phenotype gives a survival advantage during periods of famine, but is maladaptive in societies with high food abundance. In **chapter 8** we investigated whether the known genetic variants underlying obesity and T2D have indeed been favoured by positive natural selection using genome-wide SNP data from several European populations, as is suggested in the 'thrifty gene hypothesis' and other theories. When a genetic variant is under positive selection, it increases in frequency in a population and this leaves a 'signature' or pattern in the human genome. These signatures can be identified by comparison with the background distribution of genetic variation in humans, which is generally argued to have evolved largely under neutrality. In genome-wide genetic data from Europeans, we did not find signs of positive selection around the currently known T2D and obesity risk alleles and our findings therefore do not support the theory that these alleles had a survival advantage in the recent (<30,000 years ago) past. However, our data was not complete enough to reject the theory either. First of all, 'thrifty genes' that cause susceptibility for T2D and obesity could have reached fixation in the population (i.e. all individuals of the population carry the same risk allele), and they therefore cannot be picked up by GWAS using case and control data. Older selection pressure could also have acted on these variants so that their signature is no longer visible in the genome. Secondly, although the GWAS have improved our understanding of the

genetic basis of T2D and obesity, we can still only explain around 10% of the genetic risk for these traits. Thus, the majority of T2D and obesity loci are still unknown and cannot be tested for signatures of selection. However, what we did find was that some risk variants for T2D and obesity show suggestive signs of negative selection in our European data, indicating that the risk allele for the traits had a survival disadvantage in Europeans.

Compared to all other populations with a modern lifestyle, the age-adjusted T2D prevalence in populations of European ancestry is relatively low. It has been proposed that these differences in T2D susceptibility between European and non-European populations are the genetic and evolutionary consequences of geographical differences in food history. Historical data show that, starting from about 1600, European societies became capable to efficiently intervene famine, by redistributing over-abundance grain to areas of food scarcity. Jared Diamond suggested that as a result, Europeans should have undergone an epidemic in T2D starting several centuries before present as a result of the new reliability of sufficient food supplies, and eliminated the most T2D-prone genotypes by processes of natural selection. Natural selection works through differential reproductive success rather than simple differential survival. Because fertility is a driving force behind evolution, reduced fertility or infertility could be one of the underlying causes that decreased the T2D genotype frequencies in Europeans, especially because T2D is a late-onset disease and therefore not directly acting on survival. Thus far, it is unknown whether sub- or infertility is associated with future T2D risk. Therefore, in **chapter 9**, we assessed the association between measures of sub- and infertility and T2D risk in the Prospect cohort comprising 17,357 Dutch women. Our data show that general measures of sub- and infertility are not associated with T2D later in life. However, most T2D patients were diagnosed after menopause. Future studies should further investigate the association between reduced infertility and premenopausal T2D, especially as the developing epidemic of obesity has seen a substantial reduction in the age of onset of T2D and its emergence in women of childbearing age.

In the discussion on anthropometric and physical characteristics of human population in relation to food and disease, historical data is important. However, before the 19th century there is very little concrete information on which to base conclusions on heights and especially weight and other human features. In **chapter 10** we investigated a list, that was put together in 1766 in ‘s-Hertogenbosch, with names

and descriptions of 319 men and women that belonged to the gang of Calotte. This list is unique in respect of the physical characteristics provided for each person on the list, which included height, descriptions of body posture and fat distribution, facial condition (smooth-pockmarked) and a detailed description of physical deformities by which the persons could be identified. Based on descriptions of these gang members we could do multiple interesting observations and study relations between different body characteristics. Among the members of the gang of Calotte, body weight was not normally distributed (just as it is nowadays). Although, the majority of the people was described as average of weight, more individuals were fatter then average compared to thinner then average. It is remarkably that a part of the study population was described as 'fat' (17.2%) or 'very fat' (28.8%). Also other body characteristics concerning fat were mentioned, such as 'a huge belly', 'fat legs' or 'a round and plump face'. It suggest that a part of the Dutch population was overweight long before the current obesity epidemic, even in less prosperous time and even among individuals from the lowest economical classes.

Chapter 11 discusses human obesity and T2D from an evolutionary perspective as a truly full explanation of a complex disease needs both a proximate explanation of how things work and a complementary evolutionary explanation of why it got that way.

Nederlandse samenvatting

In de hele wereld stijgt het aantal mensen met type 2 diabetes (T2D) snel en op dit moment zijn er al meer dan 180 miljoen mensen met dit type suikerziekte. Risicofactoren voor T2D zijn onder meer een bepaalde etnische achtergrond, leeftijd (vandaar de traditionele naam 'ouderdomssuiker'), hoge bloeddruk, overgewicht, abdominaal vet (buikvet), en te weinig lichaamsbeweging. Van deze risicofactoren is obesitas de belangrijkste: 90% van de T2D patiënten is zwaarlijvig. Ook obesitas heeft wereldwijd epidemische proporties aangenomen. Op dit moment zijn er 300 miljoen volwassenen die als klinisch zwaarlijvig kunnen worden beschouwd. T2D en obesitas zijn multifactoriële aandoeningen waarbij zowel genetische en niet-genetische (milieu-en leefstijl) factoren een rol spelen. In dit proefschrift richten we ons op de evaluatie van alternatieve methoden om kandidaatgenen te vinden voor T2D en obesitas (deel I, hoofdstuk 3-4), onderzoek naar genetische en niet-genetische risicofactoren voor T2D en obesitas (deel II, hoofdstuk 5-7), en de vraag naar de oorsprong van de hoge prevalentie van T2D en obesitas in de moderne maatschappij (deel III, hoofdstuk 8-11).

Hoofdstuk 2 is een overzichtsartikel waarin we alle genen beschrijven die zijn geassocieerd met T2D en obesitas in genoomwijde associatie studies. We evalueren de biologische functie van deze genen om te kijken of er onderliggende biologische mechanismen zijn die zowel bij het ontstaan van obesitas als van T2D een rol spelen. Bij de evaluatie van de functie van de nu bekende risico-allelen lijkt het erop dat de reeds bekende obesitasgenen betrokken zijn bij het begin van de aandoening (verstoring van de energiebalans) en T2D genen in een later stadium (β cel defect). Dit doet vermoeden dat het gedeelde genetische effect kleiner is dan oorspronkelijk werd gedacht. Het kan ook zijn dat obesitas een niet-genetische risicofactor voor T2D is omdat het insulineresistentie veroorzaakt. De identificatie van een groter aantal obesitas en T2D genen kan dus een beter inzicht geven in de gedeelde ziekte pathologie.

In **deel** I van dit proefschrift (hoofdstuk 3 en 4) onderzoeken en evalueren we alternatieve strategieën om genen voor obesitas en type 2 diabetes te identificeren. Data van linkage studies wijzen niet direct naar kandidaatgenen en de identificatie van een mogelijk gen is meestal erg moeilijk omdat linkage intervallen wel honderden kandidaatgenen kunnen bevatten. Om een kandidaatgen te identificeren moeten er veel SNPs (genetische variaties met een basepaar verschil) worden getest voor associatie met een ziekte in een groot aantal patiënten en controles. Een interessante

alternatieve strategie is om kandidaatgenen te identificeren op basis van biologische functie met behulp van diverse bioinformatica programma's.

In **hoofdstuk 3** hebben we alle tot en met 2006 gepubliceerde genoomscans voor obesitas en T2D met elkaar vergeleken en konden zo vijf regio's identificeren die voor beide ziekten overlappen. Deze vijf regio's bevatten samen 612 kandidaatgenen. Door deze regio's nader te bestuderen door middel van programma's om kandidaatgenen te vinden, konden we 27 genen in deze regio's prioriteren die betrokken zijn bij eetgedrag, metabolisme en inflammatie. Deze genen kunnen ons meer vertellen over de moleculaire link tussen de twee aandoeningen.
Het gebruik van netwerk- en pathway programma's om kandidaatgenen te vinden is een goede, maar helaas ook beperkte methode, omdat deze programma's gebruik maken van functionele informatie over genen en eiwitdatabases en daarom vooral betrekking hebben op reeds uitvoerig bestudeerde genen, interacties en netwerken.

Dit wordt verder uitgewerkt in **hoofdstuk 4.1**, waarin we onderzoeken of en zo ja welke netwerkclassificatie-analyses kunnen helpen bij het vinden van biologische interacties die mogelijk betrokken zijn bij de etiologie van een ziekte. In plaats van ons te concentreren op de SNPs met de meest significante associaties, maakten we gebruik van al bekende biologische informatie en probeerden we de biologische netwerken te vinden die relatief het meest voorkomen in genoomwijde associatie-data. We laten zien dat het gebruik van verschillende netwerkclassificatie programma's verschillende uitkomsten geven. Bovendien blijkt dat random geselecteerde SNPs altijd in een overgerepresenteerd netwerk resulteren, en dat grotere netwerken een grotere kans hebben om significant overgerepresenteerd te zijn. De bioinformaticaprogramma's die we gebruikten in ons onderzoek bleken de neiging te hebben om goed gedefinieerde netwerken te detecteren. Verdere problemen die we tegenkwamen toen we deze netwerkprogramma's gebruikten worden beschreven in **hoofdstuk 4.2**. We willen benadrukken dat bij het trekken van conclusies op basis van deze netwerkprogramma's rekening gehouden moet worden met de beperkingen ervan.

In **deel II** van dit proefschrift onderzoeken we de genetische en de niet-genetische risicofactoren voor obesitas en T2D. Obesitas wordt veroorzaakt door een verstoorde balans tussen energie-inname en energieverbruik. Er is een grote variatie in de calorie-inname en macronutriënten-inname tussen individuen en tussen verschillende etnische groepen. Deze patronen van voedingsinname hebben een grote

erfelijke component. Het hormoon neuropeptide Y speelt een belangrijke rol in de energiebalans omdat dit hormoon het hongergevoel stimuleert. Dit stimulerende effect gaat waarschijnlijk via de Neuropeptide receptor 1, 2 en 5 (NPY1R, NPY2R en NPY5R).

In **hoofdstuk 5** onderzoeken we de associatie tussen variaties in deze NPYR genen en nutriënt inname in een populatie van 400 oudere mannen. Onze data laat zien dat recentelijk ontstane allelen in NPY1R en NPY5R geassocieerd zijn met een lagere kolydrateninname, met name met een lagere mono- en disaccharide-inname. We laten ook zien dat dragers van deze genvariaties over het algemeen maaltijden consumeren met een lagere glycemische index. Ze hebben wel een hogere alcoholconsumptie. Een van deze variaties bevat een patroon van natuurlijke selectie in Europa. Onze data suggereren dat voor Europeanen vanaf de ontwikkeling van de landbouw het eten van minder koolhydraten en het drinken van meer alcohol een overlevingsvoordeel kan hebben opgeleverd, omdat dit consumptiepatroon is geassocieerd is met een lager risico op bepaalde chronische ziekten, zoals uit eerdere studies blijkt.

In de afgelopen tijd hebben genoomwijde associatiestudies onlangs meerdere veelvoorkomende genetische variaties gevonden die geassocieerd zijn met obesitas. In **hoofdstuk 6** onderzoeken we of deze variaties ook specifiek geassocieerd zijn met abdominaal vet (buikvet). Onafhankelijk van iemands totale lichaamsvet, geeft buikvet een verhoogd risico op ziekte en voortijdige sterfte. Ook bestuderen we het effect van deze genetische variaties op de totale voedingsinname en op de macronutriënteninname in 1700 gezonde Nederlandse vrouwen. Onze data laten zien dat de obesitasgenen niet specifiek geassocieerd zijn met abdominaal vet, maar wel met obesitas in het algemeen. We vonden een verband tussen macronutriënteninname en genetische variaties in obesitasgenen. Variaties in de genen SH2B1, KCTD15 en NEGR1 waren geassocieerd met respectievelijk de inname van vet algemeen, van verzadigd vet en van enkelvoudig verzadigd vet. Genetische variaties in KCTD15 en MTCH2 waren geassocieerd met respectievelijk de inname van koolhydraten, van enkelvoudige suikers en van meervoudige suikers. Dit resultaat geeft aan dat de recent gevonden obesitasgenen een rol kunnen spelen bij de keuze en voorkeur van specifieke macronutriënten.

Een van de andere leefstijlfactoren die worden gelinkt aan T2D is pariteit, dat wil zeggen het aantal kinderen dat iemand krijgt. Vrouwen met vier of meer kinderen

hebben een verhoogde kans op T2D. Omdat het baren van kinderen is geassocieerd met gewichtstoename, is er geopperd dat hier de oorzaak ligt, maar in hoeverre BMI en tailleomtrek de associatie tussen pariteit en T2D beïnvloeden is tot nu toe onduidelijk. Een andere theorie die deze associatie mogelijk verklaart is dat bij de eerste zwangerschap de ovarium functie permanent wordt 'gereset', hetgeen kan leiden tot een levenslang verlaagde blootstelling aan oestrogeen. De leeftijd van een vrouw bij de geboorte van haar eerste kind is dan een goede marker van de duur van deze verlaagde blootstelling.

Uitgaande van deze hypothese onderzoeken we in **hoofdstuk 7** de associatie met zowel pariteit als de leeftijd van de eerste voldragen zwangerschap met het risico op T2D in 17.357 Nederlandse vrouwen van het Prospect-EPIC cohort. Onze resultaten laten zien dat de associatie van pariteit met T2D gemedieerd wordt door verhoogd lichaamsgewicht. Elk extra kind zorgt voor een toename in gewicht, BMI en tailleomtrek. Ook vonden we dat de leeftijd van de eerste zwangerschap omgekeerd geassocieerd was met het risico op T2D. Deze associatie was deels gemedieerd door lichaamsgewicht, maar het effect kon hier niet volledig door verklaard worden. Een mogelijk mechanisme kan zijn dat een jonge leeftijd van de eerste zwangerschap leidt tot een langdurig verminderde blootstelling aan oestrogeen en dat dit weer een verlaagde β-cel functie veroorzaakt.

In **deel III** van dit proefschrift bestuderen we de oorsprong van de hoge prevalentie van obesitas en T2D in de moderne maatschappij. Waarom komen deze aandoeningen zo veel voor, terwijl ze een negatieve effect hebben op de algemene gezondheid? De mens is immers gevormd via evolutionaire processen en het feit dat zo veel mensen een natuurlijke aanleg hebben om obesitas en T2D te krijgen lijkt daarom op te wijzen dat de genen die daarbij betrokken zijn begunstigd zijn door processen van natuurlijke selectie. Er zijn verschillende theorieën die de sterke genetische basis van obesitas en T2D proberen te verklaren en de bekendste daarvan is de 'zuinige genen theorie' (in het Engels: 'thrifty genes hypothesis'). De theorie houdt in het kort in dat het T2D fenotype in tijden van ontberingen en hongersnood een overlevingsvoordeel geeft, maar dat het fenotype nadelig is in een samenleving met een overvloed aan voedsel.

In **hoofdstuk 8** kijken we of bekende genen voor obesitas en T2D zijn bevoordeeld door positieve natuurlijk selectie, zoals met name gesuggereerd in the 'zuinige genen theorie'. We onderzochten dit in genoomwijde SNP data sets van

verschillende Europese populaties. Als een genetische variatie onder positieve selectie staat, dan stijgt de frequentie van deze variant in de populatie en dit laat een specifiek patroon achter in het genoom van de mens. Deze patronen kunnen worden geïdentificeerd door ze te vergelijken met de distributie van genetische variatie in de mens, omdat over het algemeen wordt aangenomen dat dit voornamelijk is geëvolueerd onder een neutraal model.

In genoomwijde genetische data van Europeanen konden we geen bewijs vinden van positieve natuurlijke selectie op obesitas en T2D risicogenen. Dit laat zien dat in de afgelopen 30.000 jaar deze allelen geen substantieel overlevingsvoordeel hebben gehad. Onze data waren echter niet compleet genoeg om de theorieën die uitgaan van evolutionair voordeel te verwerpen. Het is mogelijk dat de 'zuinige genen' al gefixeerd zijn in de populatie (dit betekent dat alle individuen uit een populatie hetzelfde risico-allel hebben), waardoor ze niet kunnen worden gevonden door het vergelijken van patiënten met een controle groep van niet-patiënten. Daar komt bij dat de nu bekende obesitas en T2D genen slechts 10% van het genetische risico op deze aandoeningen verklaren. Het merendeel van de obesitas en T2D genen is dus onbekend en kan niet getest worden voor signalen van selectie. Wat we wel zagen is dat een aantal risico genen een (zwak) signaal van negatieve selectie laten zien in Europeanen. Dit kan betekenen dat risicovarianten voor obesitas en T2D wel een negatief effect hebben gehad op de overleving van Europeanen.

In vergelijking met andere populaties met een moderne levensstijl, is de leeftijdsgecorrigeerde prevalentie van T2D in Europeanen relatief laag. Er wordt geopperd dat dit veroorzaakt wordt door de genetische en evolutionaire gevolgen van historische verschillen in de voedselvoorziening tussen Europeanen en andere populaties. Historische gegevens laten zien dat in de Europese samenlevingen vanaf ongeveer 1600 langdurige hongersnoden zeldzaam waren. Jared Diamond heeft de hypothese geformuleerd dat daardoor de epidemie van T2D in Europa al enkele eeuwen geleden is begonnen, en dat hierdoor de sterkste T2D genen in Europa door natuurlijke selectie zijn geëlimineerd.

Natuurlijke selectie werkt via verschillen in overleving, maar nog meer via verschillen in voortplantingssucces. Het zou dus mogelijk kunnen zijn dat verminderde vruchtbaarheid of onvruchtbaarheid de onderliggende oorzaak is voor de vermindering van T2D fenotypes in Europeanen, vooral omdat T2D een ziekte is pas later in het leven opspeelt en daarom niet direct een effect heeft op iemands

voortplantingssucces. Het is nog onbekend of verminderde vruchtbaarheid of onvruchtbaarheid geassocieerd is met het risico op T2D op latere leeftijd.

Daarom onderzoeken we in **hoofdstuk 9** of we een associatie konden vinden tussen vruchtbaarheid en een later risico op T2D in 17.357 Nederlandse vrouwen. Onze data laten een dergelijke associatie niet zien, maar daarbij moet aangetekend worden dat meeste vrouwen met T2D in ons onderzoeksbestand de diagnose T2D pas kregen na de menopauze. In vervolgstudies zou beter gekeken moeten worden naar de associatie van verminderde vruchtbaarheid met voor de menopauze gediagnosticeerde T2D.

Voor de discussie over de ontwikkeling van antropometrische en fysieke kenmerken van menselijke populatie in relatie tot voeding en ziektes zijn historische gegevens belangrijk. Vóór de negentiende eeuw is er echter weinig concrete informatie voorhanden die geschikt is om conclusies over lengte, gewicht en andere menselijke uiterlijke kenmerken op te baseren. Voor eerdere tijden moeten we genoegen nemen met toevallige vondsten.

In **hoofdstuk 10** bestuderen we een dergelijke vondst, een lijst van 317 mannen en vrouwen, die behoorden tot de zogenaamde 'bende van Calotte', een criminele groepering afkomstig uit de armste bevolking, waarvan de leiders in 1766 berecht werden in 's-Hertogenbosch, De lijst is bijzonder omdat het voor elke persoon behalve naam en leeftijd ook de lengte en een typering van het postuur gegeven wordt, aangevuld met andere uiterlijke kenmerken. Aan de hand daarvan konden we verschillende interessante observaties doen en relaties tussen verschillende lichamelijke kenmerken binnen deze groep vaststellen.

Opvallend was dat de statistische verdeling van lichaamsgewicht niet normaal was, evenmin als dat het geval is in de moderne tijd. Het grootste deel van de personen werd omschreven als 'gemiddeld van gewicht', maar van de overigen waren er veel meer dik dan dun. Het is opmerkelijk dat 17,2 % van deze mensen werd getypeerd als 'dik' en 28,8% zelfs als 'heel dik', en dat er kenmerken gegeven werden als 'een enorme buik', 'dikke benen' of 'een rond en pafferig gezicht'. Dit kan erop duiden dat een deel van de Nederlandse populatie al kampte met overgewicht lang voor de huidige obesitasepidemie, zelfs onder mensen van de laagste economische klasse en in een economisch moeilijke tijd.

In **hoofdstuk 11**, tenslotte, worden obesitas en T2D besproken vanuit een evolutionair perspectief. Een volledige verklaring van complexe ziektes vergt immers

zowel een uitleg van de directe oorzaak, die de vraag beantwoordt hoe iets in zijn werk gaat, als een aanvullende evolutionaire verklaring, die ingaat op de vraag waarom het zo in zijn werk in kan zijn gegaan.

Dankwoord

Als kind van ouders die altijd maar bezig waren met hun proefschrift schrijven, begreep ik dat promoveren wel iets heel erg verschrikkelijks moest zijn. Ik nam me destijds dus voor om maar zo vroeg mogelijk met mijn proefschrift te beginnen: dan was ik er maar alvast vanaf. Later kwam ik erachter dat promoveren een keuze is, dus dat was een grote opluchting. Helaas kon ik na mijn afstuderen de verleiding toch niet weerstaan om verder te gaan in het onderzoek en ook een proefschrift te schrijven. Hier is-ie dan. En uiteindelijk viel het toch allemaal best wel mee. Ik ben heel blij met het resultaat! Dit boekje was er natuurlijk niet gekomen zonder de hulp van collega's en steun (en afleiding) van vrienden en familie.

Cisca en Yvonne, ik ben heel blij dat ik mijn promotieonderzoek onder jullie vleugels heb kunnen doen. Jullie zijn (allebei op jullie eigen manier) een voorbeeld voor me hoe je succesvol kan zijn in de wetenschap. Bedankt, ik heb heel veel van jullie geleerd!

Beste Cisca, dankjewel voor al je commentaar en goede ideeën voor mijn onderzoek. Je nuchtere houding maakte dat problemen waar ik tegenaan liep na een gesprek met jou opeens veel minder erg leken. Dat kon ik altijd goed gebruiken. Ook de fijne gesprekken over het vervolg van mijn carrière en je goede raad hierover zijn heel belangrijk voor me geweest. Bedankt dat ik een tijdje bij jou en Marten op zolder mochten 'wonen' in jullie prachtige huis. Ik vond het heel gezellig, vooral de leuke en lekkere etentjes en alle gesprekken 's ochtends aan de keukentafel!

Beste Yvonne, bedankt voor je enthousiasme en het meedenken met alle studies in dit proefschrift. Te gek dat je me op het laatste moment toch nog hebt weten te overtuigen dat epidemiologie leuk is! Heel erg bedankt ook voor de laatste weken waarin je er altijd voor me was en je dag en nacht naast je computer zat te wachten op nieuwe hoofdstukken van mijn proefschrift (althans, zo leek het tenminste;)). Zonder jou was dit proefschrift vast een stuk dunner geworden! Tenslotte wil ik je graag bedanken voor de gesprekken en je interesse toen het allemaal wat minder ging. Dat heeft me erg geholpen.

Beste Charlotte de wekelijkse werkbesprekingen waren altijd heel leerzaam en gezellig! Je hebt me geleerd om altijd kritisch na te blijven denken over methodologie

en dat is erg nuttig gebleken voor mijn onderzoek. Bedankt voor alle leuke brainstormsessies over genetica en methodes, je hulp bij het oplossen van de vele problemen waar we tijdens dit project tegenaan liepen en de goede gesprekken over van alles en nog wat. Je hebt een belangrijke bijdrage geleverd aan de meeste hoofdstukken van dit proefschrift. Bedankt, ik heb er heel veel aan gehad!

Het verzamelen, genotyperen en analyseren van de EPIC-NL samples nam een groot deel van mijn AIOproject in beslag. Ook al valt er in dit proefschrift helaas niet veel te over lezen (maar wordt vervolgd..), toch wil ik iedereen graag heel hartelijk bedanken die heeft bijgedragen aan deze enorme kluif. Allereerst Florianne, bedankt voor je hulp bij al het monnikenwerk van EPIC-NL! Veel succes nog met de laatste loodjes van je eigen proefschrift. Samples lichten en verzamelen is nog niet zo gemakkelijk als het lijkt en ik ben daarom erg blij dat ik hulp heb gehad van Joline, Ivonne, Anita, Heleen, Jens, José, Marlies, Robert-Jan, Jan, Martijn, Sandra, Bernard en Suzanne bij het achterna bellen van huisartsen voor het valideren van de T2D patiënten, het lichten van de samples in de krochten van het RIVM en databeheer en dataopslag van de samples. Iedereen bedankt!
Beste Pieter, Elvira en Mathieu, jullie verdienen echt een ereplaats in dit dankwoord! Jullie zijn fantastische collega's en ik ben heel blij dat ik met jullie heb mogen werken. Toen we met honderden verprutste DNA samples in Groningen aankwamen, durfde ik er bijna niet meer in te geloven dat het nog goed zou komen met de experimenten. Gelukkig bleek al gauw dat we bij jullie in goede handen waren: jullie hadden heel veel kennis van zaken, dachten mee en hadden voor elk probleem wel een oplossing. Bedankt voor jullie waardevolle bijdrage aan het EPIC-NL project! Lieve Elvira, je enthousiasme over de Illumina, de Beagle, Hawaï en heel veel andere dingen waren aanstekelijk en je stond altijd klaar om ergens mee te helpen. Wat ontzettend erg dat je er niet meer bent.
Beste Soesma en Elinda, ook jullie wil ik graag bedanken voor alle hulp bij de Illumina experimenten. Dear Asia and Gosia, thank you for showing me how to handle the large Illumina dataset. Beste Paul, dankjewel voor je hulp en expertise bij het opschonen, analyseren en interpreteren van de EPIC-NL data. Heel fijn dat je ik mede dankzij jou nu ook betrokken ben bij de T2D meta-analyse. Daar ben ik echt heel erg blij mee! Beste Jessica, geweldig dat ik er de laatste maanden zo'n fijne

collega bij kreeg om me te helpen bij de EPIC QC en alle analyses. Bas, bedankt voor de fijne samenwerking, het was gezellig.

Het onderzoek in dit proefschrift maakte deel uit van het IOP Genomics project en ik wil graag iedereen bedanken die hierbij betrokken is geweest. Beste Prof. Gert-Jan van Ommen en leden van de begeleidingscommissie, dankjewel voor jullie ideeën en suggesties voor het onderzoek. Beste Tonnie en Annelies, bedankt voor alle praktische hulp bij het project. Beste Marten, Jana, Marcel, Timon, Sander, Yanti, Wim, Anouk, Marcella, Ellen en Veerle, bedankt voor de fijne samenwerking. Beste Marten, ook jou wil ik graag bedanken voor de gastvrijheid toen ik een tijdje bij jou en Cisca op zolder 'woonde'. Ik vond het heel gezellig. Beste Timon, zonder jou had ik waarschijnlijk 4 jaar lang promotieonderzoek gedaan over obesitas en type 2 diabetes zonder iets over het ziekte-aspect te weten. Je bent een wandelende encyclopedie en ik ben blij dat ik altijd langs mocht komen en dat je dan ook echt de tijd voor me nam om projecten en ideeën te bespreken. Lieve Jana, wat fijn om in Groningen een echte vriendin te hebben waarbij ik altijd welkom was. De gezellige etentjes bij jou thuis en alle fijne gesprekken over leuke (en minder leuke dingen) van onderzoek en het AIO-zijn en van alles en nog wat heb ik erg gewaardeerd. Ik hoop dat we in de toekomst nog vaak over genetica, onderzoek en nog meer kunnen praten.

Lieve flexkamergenootjes: Flip, Simone, Sasha, Jelena, Ester, Ruben, Eric, Kristel, Linda, Jacobien, Karen, Anette en Jessica. Het was niet altijd gemakkelijk in Utrecht, als left-over-AIO van Cisca, maar dat lag zeker niet aan jullie. Als ik het even niet meer zag zitten dan hadden jullie altijd een luisterend oor of kwamen jullie aan zetten met lieve kadootjes als een ijsbeer met gedicht of de Sabeti-boekenlegger. Ik zal ook altijd met veel plezier terug blijven denken aan de wii-, cocktail-, spontane, semispontane, stiekeme, the Hoff-, brainstorm-, koek&zopie-borrels en alle (kansloze) experimenten met wodka-brain-jelly, aquasaurussen en het droogijszwembad. Bedankt! Mede door jullie ging ik elke dag met plezier naar mijn werk. Simone, 'my family', snel weer gezellig een bustour doen met naamstickers? Hier in Amerika zijn ze er dol op! Flip, dankjewel voor al je fijne programmaatjes. Je maakte altijd precies het programma wat ik nodig had. Dat zal ik zeker gaan missen en ik moet er hier maar aan geloven om zelf hardcore te gaan programmeren. Sasha, heel leuk om samen aan het evolutie project te werken! Je bent vol van positieve

energie en dat maakt jou zo'n leuke collega. Ik hoop dat we in de toekomst nog veel leuke discussie kunnen voeren over genetica en alles wat daarbij komt kijken.

Beste overige collega's van de biomedische genetica, Harry, Jackie, Carolien, Carla, Berooz, Alienke, Albertine, Rian, Erica, Bart, Magdalena, Wigard, Ivo, Bobby, Marianna, Edwin, Roel, Dalila, Martin, Marc, Stephan, Judith, Richard, Monique, Jacqueline, Noelline Lude, Frederieke, bedankt voor de fijne samenwerking. Beste Carolien, jij bent de spil van de afdeling. Fijn dat je me een beetje onder je hoede hebt genomen na het vertrek van Cisca. Bedankt voor alle leuke discussies over populatiegenetica en voor je waardevolle bijdrage aan hoofdstuk 5 en 8. Beste Jackie, dankjewel voor het nakijken van alle manuscripten en voor al je goede adviezen over van alles en nog wat in het leven. Beste Lude, bedankt voor alle leuke discussies over genetica en je belangrijke bijdrage aan hoofdstuk 3 en 4.

Beste studenten Iris, Dirkjan, Renske, Kristel, Sietske, leuk om jullie te mogen begeleiden. Ik vond het erg gezellig en heb ook veel van jullie geleerd!

Werken (en 'wonen') in Groningen zou niet half zo leuk geweest zijn zonder alle gezellige Groningers. Bedankt aan iedereen die me tijdens mijn verblijf in het hoge Noorden uitnodigde voor het eten, de bioscoop, een biertje in de kroeg en/of om te blijven logeren: Jana & Peter & Marsha, Noortje & Mathieu, Kike & Reinier, Jihane, Agatha, Huub, Eva en iedereen die ik nog ben vergeten. Mede dankzij jullie voelde ik me helemaal thuis in Groningen!

Lieve Linde, Anne, Kike en Sanne, wat bijzonder om samen op te groeien ☺ Lieve Linde, we zijn al 21 jaar hartsvriendinnen en van jongs af aan bespreek ik al mijn problemen en geheimen als eerste met jou. Urenlang (dagenlang?) hebben we met elkaar aan de telefoon gehangen en ik vindt het fijn dat je altijd een luisterend oor hebt en gewoon je eigen mening geeft. Ook bedankt voor alle leuke concerten, onverwachtse activiteiten en gezellige avondjes. Lieve Anne, in groep 3 zaten we al bij elkaar in de klas en de afgelopen jaren hebben we allebei in het Stratenum gewerkt. Fijn om met je over onderzoek, promotiestress, mannen en van alles en nog wat te kunnen praten. Linde en Anne, fijn dat jullie naast me staan op 7 oktober:

mocht ik een black-out krijgen dat vertrouw ik erop dat jullie de promotiecommissie ondertussen zullen afleiden met een spetterende serenade.

Lieve Michèle, dit proefschrift is er ook een beetje dankzij jou, want zonder jou had ik het Stedelijk echt niet overleefd (figuurlijk dan;)). Bedankt voor de goede oude vriendschap. Friends 4-ever!

Lieve biologievrienden; Martine, Deborah, Emmalie, Maaike, Simone, Kevin en Walter, biologie studeren was dubbel zo leuk omdat jullie mijn studiegenootjes waren. Ik zal de fantastische/kansloze/hilarische/dramatische (biologie)feesten PUT-avonden en weekenden nooit vergeten! Lieve Martine, wat hebben we de afgelopen jaren een avonturen beleefd: karnaval in Rio, safari door Kakadou National Parc, langlauf-huttentocht in Oslo, snowboardvakanties en niet te vergeten alle stap- en hangavondjes in Utrecht. Op en buiten mijn werk ging niet altijd alles over rozen, maar als mijn beste vriendin heb je me door veel moeilijke tijden heen gesleept! Ik ben blij dat je je droomman/baan gevonden hebt in Olso, maar ik mis je wel. Dear Knut, you are the very nice and interesting Norseman-friend I always wanted to meet. Take good care of Martine and little ♥! Lieve Deborah, Emmalie, Maaike en Simone, bedankt voor het organiseren van mijn gezellige en zonnige afscheidsfeestje! Ik heb ervan genoten en ik ben ook erg blij met mijn boek met foto's om af en toe bij heimwee met een doos tissues doorheen te bladeren. Ook al woon ik nu even iets verder weg, toch hoop ik dat er in de toekomst nog veel gezellige ladies-nights gaan aankomen. Lieve Kevin, bedankt voor alle gezellige hangavondjes toen we nog 'buren' waren en te gek dat jij en Marije onze getuigen wilden zijn op onze Elvis-wedding in Las Vegas!

Lieve Allegra-vriendinnen; Heleen, Ilse, Sonja, Ruth & Lies, wat was het fijn om jarenlang met jullie in het USKO te zingen. Ik heb onvergetelijke herinneringen aan USKO- kampen en concertreizen en de Johannes, de Matthäus en de Hohe Messe zijn voor mij onlosmakelijk met jullie verbonden. Bedankt voor alle gezellige Allegra-avondjes en brunches bij 'Le Journal' waar we met elkaar de kleine en grote problemen van het leven bespraken. Momenteel zien we elkaar minder vaak (en is er wel heel veel testosteron bijgekomen in Allegra), maar jullie zijn nog steeds dierbare

vriendinnen. Lieve Heleen, heel grappig dat we in de loop der jaren toch nog zulke goede vriendinnen zijn geworden.

Lieve Diva's (Roos, Marijn, Tiffany, Marijke, Meta, Mees, Sophie, Janneke, Merel, Mechel, Judith, Miriam, Melynda, Liesbeth, Jitske, Feike, Marjon, Petra, Sanne, Lydia) en Pepijn, wat ga ik jullie ontzettend missen! Ik vond het fantastisch om met jullie op het podium te staan en naast jullie voelde ik me een echte Diva! Promoveren was niet altijd gemakkelijk, maar na een maandagavondje zingen met Divina kon ik er altijd weer helemaal tegenaan. Bedankt voor de mooie muziek en de gezelligheid.

Verder nog bedankt aan alle vrienden en familie die me de afgelopen jaren hebben gesteund en gezorgd hebben voor afleiding: Oma, de Elbersen (met name Riet en Jos) de van de Pollen (in het bijzonder mijn tantes Alice en Batja), Bouke, Martijn, Floor, Michiel, Maud, Karlijn, Sicco, Rob, Tim, Sietse, Stefan, Roelinka, de Kochjes, de Schinkels, oud-Uskieten, UBV-ers, de PUTcie, oud-huisgenootjes en iedereen die ik nog ben vergeten. Lieve Oma, wat fijn dat ik zo goed met je kan praten over alle dingen van het leven, ondanks ons leeftijdsverschil. Lieve Steef en Hannie, dankjewel voor jullie hartelijkheid! Er valt altijd iets te bleven met jullie. Heel fijn dat ik mijn promotiefeestje bij jullie in het Schillertheater mag vieren. Lieve Sietse, dankjewel voor het ontwerpen van de voorkant van mijn boekje. Ik ben er ontzettend blij mee! Ook bedankt voor al je praktische hulp bij het vervoeren van spullen en de poes tijdens onze verhuizingen. Stefan, ook jij bedankt voor alle goede zorgen voor Molly.

Lieve Ferwerda's, toen ik verliefd werd op Bart kreeg ik jullie er gratis bij. Met zijn allen kunnen jullie enorm druk zijn, maar jullie zijn vooral heel erg hartelijk, betrokken en lief! Ik voelde me vanaf de eerste dag welkom in jullie gezin en ik waardeer het zeer dat jullie altijd met me mee hebben geleefd, in goede en minder goede tijden. Ik ben heel blij dat ik bij zo'n grote gezellige familie mag horen!

Lieve Ed en Lotte, bedankt (in volgorde van belangrijkheid;)) voor de genen en de omgevingsfactoren! Jullie zijn er altijd voor me en hebben altijd in me geloofd en ik kan me geen betere en lievere ouders voorstellen dan jullie. Ik vind het ook heel leuk dat we onze passie en interesse voor de wetenschap delen. De ontelbare keren dat jullie me hebben geholpen met het oplossen van praktische problemen, dat jullie

meedachten met dilemma's en adviezen gaven over van alles en nog wat in het leven zijn heel belangrijk voor me geweest. Van jongs af aan hebben jullie me geleerd om situaties en problemen van meerdere kanten te bekijken. Dit is zeer waardevol gebleken in mijn werk en in het leven. Bedankt, ik hou van jullie. Lieve Lotte, te gek dat onze ideeën en goede samenwerking hebben geleid tot hoofdstuk 10 in dit proefschrift!

Lieve Lies, bedankt dat je altijd voor me klaar staat! Als we samen zijn is het altijd gezellig en ik kan nog net zo met je lachen als vroeger, toen we nog twee kleine zusjes waren. Je sarcastische 'van de Pol' humor is er alleen maar op vooruit gegaan (ook al blijft 'geen jas aan?' natuurlijk een klassieker;)). Jij bent de avontuurlijkste van ons tweeën en het avontuur heeft je dit keer naar Lyon gebracht. Ik bewonder je dat je zo stoer bent er altijd helemaal voor gaat om er weer iets moois van te maken. I hope that you, Fabien and little ♥ will be very happy! Thanks for already spreading my genes while I am still busy working on my career.

Allerliefste Bart, wat ben ik blij dat ik je heb gevonden! Promoveren (en veel andere dingen in het leven) zijn sindsdien alleen maar leuker geworden. Je passie voor biologie en je enthousiasme over de wetenschap zijn aanstekelijk en de leukste discussies over mijn onderzoek heb ik met jou gevoerd. Ook heel erg bedankt voor je steun, voor al je praktische hulp, voor de lol die we samen hebben en voor je lieve en rustige karakter als ik weer eens ergens gestrest over was. Ik heb veel zin in alle avonturen die we nog samen gaan beleven. Ik hou van je!

List of publications

Vossen CY, **Elbers CC**, Hasstedt SJ, Koeleman BPC, Rosendaal FR, Bovill EG. Computational Candidate Gene Prioritization Programs for Identifying Genetic Risk Factors for Venous Thrombosis. *J Thromb Haemost. 2010 Aug ;8(8):1869-1871.*

Elbers CC§, Zhernakova A§, Ferwerda B, Romanos J, Trynka G, Dubois PC, de Kovel CGF, Oosting M, Barisani D, Bardella MT, Finnish Celiac Disease Study Group, Joosten LAB, Saavalainen P, van Heel DA, Catassi C, Netea MG, Wijmenga C. Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a selective force against bacterial infection. *Am J Hum Genet. 2010 Jun 11;86(6):970-7.* §equal contribution

Bauer F, Onland-Moret NC, Niehoff AG, **Elbers CC**, Grobbee DE, Wijmenga C, van der Schouw YT. PTPN1 polymorphisms are associated with total and low-density lipoprotein cholesterol. *Eur J Cardiovasc Prev Rehabil. 2010 Feb;17(1):28-34.*

Ferwerda B, Ferwerda G, Plantinga TS, Willment JA, van Spriel AB, Venselaar H, **Elbers CC**, Johnson MD, Cambi A, Huysamen C, Jacobs L, Jansen T, Verheijen K, Masthoff L, Morré SA, Vriend G, Williams DL, Perfect JR, Joosten LA, Wijmenga C, van der Meer JW, Adema GJ, Kullberg BJ, Brown GD, Netea MG. Human dectin-1 deficiency and mucocutaneous fungal infections. *N Engl J Med. 2009 Oct 29;361(18):1760-7.*

Bauer F, **Elbers CC**, Adan R, Loos RJF, Onland-Moret NC, Grobbee DE, van Vliet-Ostaptchouk JV, Wijmenga C, van der Schouw YT. GWAS identified obesity genes are associated with adiposity measures and potentially with nutrient-specific food preference. *Am J Clin Nutr. 2009 Oct;90(4):951-9.*

Elbers CC, de Kovel CGF, van der Schouw YT, Meijboom JR, Bauer F, Grobbee DE, van Vliet-Ostaptchouk JV, Wijmenga C, Onland-Moret NC. Variants in neuropeptide Y receptor 1 and 5 are associated with nutrient-specific food intake and are under recent selection in Europeans. *PLoS ONE. 2009 Sep 17;4(9):e7070.*

Elbers CC, van der Schouw YT, Wijmenga C, Onland-Moret NC. Comment on: Perry et al. (2009) Interrogating Type 2 Diabetes Genome-Wide Association Data Using a Biological Pathway-Based Approach. *Diabetes. 2009 Sep;58(9):e9*

Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC. Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol. 2009 Jul;33(5):419-31.*

Elbers CC, Wolfs MGM, van Haeften TW. Gene variants for obesity and type 2 diabetes mellitus: Shared aetiology? *European Endocrinology 2009;5(1):27-30*

Helder S, Jansen-Vreeken C, **Elbers CC**, Adan R. Genetisch onderzoek naar obesitas wijst op rol hersenen. *Ned Tijdschr voor Voeding & Dietiek. 2009;64(2)*

Bauer F, Onland-Moret NC, Niehoff AG, **Elbers CC**, Grobbee DE, Wijmenga C, van der Schouw YT. No Association of PTPN1 Polymorphisms with Macronutrient Intake and Measures of Adiposity. *Obesity (Silver Spring). 2008 Dec;16(5):2767-71*

Van Vliet-Ostaptchouk JV, Onland-Moret NC, van Haeften TW, Franke L, **Elbers CC**, Shiri-Sverdlov R, van der Schouw YT, Hofker MH, Wijmenga C. HHEX gene polymorphisms are associated with type 2 diabetes in the Dutch Breda cohort. *Eur J Hum Genet. 2008 May;16(5):652-6.*

Niehoff AG, van Haeften TW, Onland-Moret NC, **Elbers CC**, Wijmenga C, van der Schouw YT. C-reactive protein is independently associated with glucose but not with insulin resistance in healthy men. *Diabetes Care. 2007 Jun;30(6):1627-9.*

Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, Wijmenga C. A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol Metab. 2007 Jan-Feb;18(1):19-26.*

Curriculum vitae

Clara Elbers werd op 27 februari 1980 geboren te Utrecht. Ze haalde haar middelbare school diploma aan het Stedelijk Gymnasium te Utrecht en in 1999 begon ze met haar studie Biologie aan de Universiteit Utrecht. Haar stages liep ze bij de afdeling Psychiatrie van het Erasmus Medisch Centrum te Rotterdam en het Institute for Risk Assessment Sciences (IRAS) te Utrecht.

Tijdens haar studie speelde ze cello bij strijkorkest Zoroaster en zong ze bij het Utrecht Studenten Koor en Orkest (USKO). Met het USKO toerde ze verschillende malen door Europa om de Johannes Passion, de Matthäus Passion en de Hohe Messe van J.S. Bach op te voeren.

In 2006 begon ze met haar promotie traject aan het Universitair Medisch Centrum Utrecht waar ze werkte bij de afdeling Biomedische Genetica en het Julius Centrum onder de supervisie van Prof. Cisca Wijmenga en Prof. Yvonne van der Schouw. Het onderzoek wat ze daar deed resulteerde in dit proefschrift.

Naast het promoveren zong ze de afgelopen jaren in close harmony koor Divina. Samen met Divina stond ze op het podium met de succesvolle theater-show ‘nieuwe collectie’ en won ze diverse publieksprijzen bij nationale korenfestivals.

Momenteel werkt ze op de afdeling Genetica in het lab van Prof. Sarah Tishkoff aan de University of Pennsylvania. In 2010 kreeg Clara Elbers de NWO Rubicon-subsidie toegekend om te onderzoeken hoe verschillende voedingspatronen het menselijk genoom in Afrika hebben beïnvloed.

Clara Elbers woont in Philadelphia samen met Bart Ferwerda.