MATTHIJS J.H.M. VAN DER LOOS

Molecular Genetics and Hormones

New Frontiers in Entrepreneurship Research

Molecular Genetics and Hormones: New Frontiers in Entrepreneurship Research

Molecular Genetics and Hormones: New Frontiers in Entrepreneurship Research

Moleculaire genetica en hormonen:
Nieuwe, onontgonnen gebieden in het ondernemerschapsonderzoek

THESIS

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.G. Schmidt

and in accordance with the decision of the Doctorate Board.

The public defense shall be held on
Thursday, June 20, 2013, at 9:30 hours

by

Mattheus Josef Henricus Maria van der Loos
born in Heerlen.

Doctoral Committee

Promotors:	Prof.dr. A.R. Thurik
	Prof.dr. P.J.F. Groenen
	Prof.dr. A. Hofman
Other members:	Prof.dr. H. Bleichrodt
	Prof.dr. M. Johannesson
	Prof.dr. H.W. Tiemeier
Copromotor:	Dr. P.D. Koellinger

Erasmus Research Institute of Management – ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

ERIM PhD Series in Research in Management, 287
ERIM reference number: EPS-2013-287-S&E
ISBN 978-90-5892-330-1
© 2013, Matthijs J.H.M. van der Loos

Design: B&T Ontwerp en advies www.b-en-t.nl

This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: http://www.haveka.nl/greening

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the author.

Contents

Preface (Voorwoord) **vii**

1 Introduction and Conclusion **1**

1.1 Motivation and Contribution 3

1.2 Thesis Outline, Research Questions, and Main Results 7

1.3 Discussion 9

1.4 Publication Status of Chapters 10

2 Candidate Gene Studies and the Quest for the Entrepreneurial Gene **13**

2.1 Introduction 15

2.2 Replication Study 15

2.3 Results 16

2.4 Discussion 19

2.5 Conclusion 21

3 Genome-Wide Association Studies in Economics and Entrepreneurship Research: Promises and Limitations **23**

3.1 Introduction 25

3.2 Basic Genetic Concepts and GWAS 28

3.3 Identification of True Positives 31

3.4 How To Interpret GWAS Results 40

3.5 Conclusion: Genetics in Economics Research? 44

4 The Molecular Genetic Architecture of Self-Employment **47**

4.1 Introduction 49

4.2 Materials and Methods 51

4.3 Results 56

4.4 Discussion 65

5 The Molecular Genetics of Serial Self-Employment 71

5.1 Introduction 73

5.2 Basic Genetic Concepts 77

5.3 Data 78

5.4 Evidence for Heritability 79

5.5 Genome-Wide Association Study (GWAS) 84

5.6 Genetic Risk Prediction 90

5.7 Discussion 92

5.8 Conclusion 95

6 Measures of Bioactive Serum Testosterone Are Not Associated with Entrepreneurial Behavior in Two Independent Observational Studies 97

6.1 Introduction 99

6.2 Materials and Methods 100

6.3 Results 103

6.4 Discussion 104

Appendix A: Supplementary Tables to Chapter 4 109

Appendix B: Supplementary Tables to Chapter 5 129

Appendix C: Details on the GREML Procedure 133

Appendix D: Supplementary GWAS Results to Chapter 5 137

Summary 143

Nederlandse Samenvatting (Summary in Dutch) 145

References 147

About the Author 169

Preface (Voorwoord)

Eind november 2007 bevond ik mij ergens onderweg van Cairns naar Melbourne tijdens een welverdiende vakantie, toen ik het bevrijdende bericht ontving dat de aanvraag voor mijn promotieonderzoek was goedgekeurd door de Vaste Commissie voor de Wetenschapsbeoefening van de Erasmus School of Economics (ESE). Ik zou mee gaan werken aan de zoektocht naar het "ondernemerschapsgen" als onderdeel van een samenwerkingsverband tussen de ESE en het Erasmus MC dat een paar maanden eerder was opgezet door mijn promotoren Roy Thurik, Patrick Groenen en Albert Hofman, en mijn copromotor Philipp Koellinger. Binnen een half jaar zou ik aan de slag kunnen gaan!

Een week later kreeg ik het bericht dat het samenwerkingsverband reeds zijn vruchten had afgeworpen en er zeer hoopvolle bevindingen waren die lieten zien dat er inderdaad ondernemerschapsgenen bestonden. "Genoeg werk aan de winkel dus", aldus Patrick Groenen in een e-mail. Men wilde dat ik zo snel mogelijk zou beginnen en zo startte ik na een overbruggingsperiode van twee maanden bij onderzoeksbureau EIM halverwege april 2008 met mijn promotieonderzoek.

Inmiddels is het vijf jaar later wanneer ik dit voorwoord schrijf en achteraf gezien bleek alles toch nét iets moeilijker te liggen dan die eerste, veelbelovende bevindingen suggereerden (die overigens later het resultaat bleken te zijn van een meetfout). Vooral de eerste jaren van mijn promotietraject vergden zo nu en dan het nodige van mijn doorzettingsvermogen, maar uiteindelijk resulteerden alle investeringen in een aantal mooie publicaties.

Hoewel de vondst van specifieke ondernemerschapsgenen vooralsnog is uitgebleven, zijn er tijdens mijn promotietraject wel veel inzichten opgedaan over de genetica van ondernemerschap en de methoden waarmee deze kennis kan worden vergaard. Bovendien zijn deze inzichten en methoden ook op andere onderzoeksgebieden in de economie toepasbaar. Mijns inziens is het belangrijkste resultaat van dit proefschrift dan ook dat het de basis vormt voor een nieuw onderzoeksgebied dat moleculaire genetica in de economie integreert. Dit gegeven geeft mij veel voldoening en compenseert alle door mij gedane investeringen ruimschoots!

Zonder de hulp van anderen was het echter niet mogelijk geweest om dit proefschrift te voltooien. Deze mensen wil ik daarvoor graag bedanken. Allereerst mijn begeleiders die samen met mij de sprong in het diepe waagden. Ik bedank Roy Thurik voor zijn motiverende begeleiding en nimmer aflatende enthousiasme en vertrouwen in dit onderzoek. Ook wil ik graag Patrick Groenen en Albert Hofman bedanken voor hun rol als (pro)motor. I am much indebted to my daily supervisor and copromotor, Philipp Koellinger, for his continued support and belief in this research. Without these, this thesis would never have materi-

alized. Hoewel Fernando Rivadeneira, Frank van Rooij en André Uitterlinden officieel geen begeleiders zijn, hebben zij wel een cruciale rol gespeeld in dit onderzoek, waarvoor ik hen graag wil bedanken.

In de loop der jaren zijn er drie promovendi in mijn voetsporen getreden en ik wil hen graag bedanken voor de aangename en stimulerende samenwerking. Ten eerste bedank ik Niels Rietveld, met name voor de vele discussies over al dan niet aan ons onderzoek gerelateerde materie. Dit waren altijd welkome onderbrekingen van de werkdag. Ik wens je veel succes met het afronden van je promotie de komende tijd. Eén ding is zeker: jouw proefschrift zal wel significante bevindingen bevatten! Secondly, I would like to thank Aysu Okbay and Ronald de Vlaming, who joined the ESE past year, for the pleasant collaboration. Although our employment at the ESE had only a short overlap, I enjoyed our time together. I wish you all the best in completing your PhDs.

Collaboration is a sine qua non for research in molecular genetics. Accordingly, the papers that compose this thesis have been co-authored with a substantial number of people. In particular, I would like to thank Dan Benjamin, David Cesarini, Magnus Johannesson, Chris Dawes, Jonathan Beauchamp, and Chris Chabris for the fruitful collaboration that resulted in several major publications. I truly enjoyed every part of it! I am also grateful to all my other co-authors for their efforts, but please excuse me for not listing all of you here because that would probably require a spreadsheet.

I would like to thank Han Bleichrodt, Ingmar Franken, Magnus Johannesson, Henning Tiemeier, Mirjam van Praag, and André Uitterlinden for serving on either the inner or plenary committees.

Naast dat mijn tijd bij de ESE werd veraangenaamd door mijn directe collega's, hebben ook de overige (ex-)leden van de ondernemerschapsgroep hieraan bijgedragen. Ik wil in het bijzonder Peter van der Zwan bedanken voor de vele gesprekken over de ins en outs van het promoveren, tips en trucs voor Word, en andere aangelegenheden. Ook zou ik graag Jolanda Hessels, Brigitte Hoogendoorn en Ingrid Verheul willen bedanken voor de vele lunches, diners en congresbezoeken. Tevens bedank ik Anka, Gerda, Kim, Nita en Ramona voor de secretariële ondersteuning.

We zijn nu aangekomen bij de laatste groepen personen die van invloed zijn geweest op de totstandkoming van dit proefschrift, namelijk familie en vrienden. Allereerst wil ik Marianne van der Loos en Robert Zuurbier bedanken voor hun rol als paranimf. Ik ben er erg blij mee dat jullie mij op deze belangrijke dag willen bijstaan!

Voor de nodige afleiding naast het werk werd door vrienden en vriendinnen gezorgd tijdens vele (nachtelijke) excursies, etentjes en wat dies meer zij. Ik zal niet iedereen bij naam noemen, want dan vergeet ik geheid iemand, maar ik wil graag iedereen die dit leest en zich aangesproken voelt hiervoor bedanken!

Ten slotte wil ik mijn ouders, broer en zus bedanken voor hun interesse en steun op allerlei gebieden. Frieda (noem ik je zo wel eens?), je weet hoeveel beter jij bent in dit soort dingen en daarom bedank ik je maar gewoon voor “alles”, maar vooral voor je altijd positieve instelling en voor de bij tijd en wijle broodnodige “schop onder mijn kont”. Je begrijpt wat ik bedoel.

Rotterdam, april 2013 Matthijs van der Loos

CHAPTER 1

Introduction and Conclusion

Partly based on Van der Loos, Koellinger, Groenen, and Thurik (2010).

Abstract

The research presented in this thesis involves an investigation of the molecular genetics of entrepreneurship and of a hormonal correlate through which genes may influence entrepreneurial behavior. The initial two chapters discuss the challenges and pitfalls of using methods that enable identification of genes associated with entrepreneurial behavior using actual molecular genetic data. The next two chapters use these and related methods to examine the molecular genetics of entrepreneurship. The final chapter focuses on the relationship between the hormone testosterone and entrepreneurial behavior. The remainder of this introductory chapter is structured as follows. Section 1.1 sets the background of the research by discussing the motivation and contribution. The thesis outline, research questions, and main results are presented in Section 1.2. The results are briefly discussed in Section 1.3 as well as some implications. Section 1.4 concludes with an overview of the publication status of each chapter. All chapters can be read independently of the others.

1.1 Motivation and Contribution

For nearly a century, the Standard Social Science Model (SSSM) has provided the conceptual foundations of the social sciences, including entrepreneurship research (Thurik, 2012; Tooby & Cosmides, 1992). Briefly put, the SSSM postulates that the human mind is initially a blank slate that is only programmed by culture and socialization. This view implies that individual preferences and choices, such as occupational choice, are the result of nurture, i.e., the environment, and not of nature, i.e., genes, or the interplay between these two. Accordingly, entrepreneurship research has sought to answer its quintessential question of what makes an entrepreneur by studying environmental factors and how these shape occupational choice. Although the best explanatory factor of who becomes an entrepreneur is the occupation of one's parents, it was assumed, based on the SSSM, that becoming an entrepreneur is only explained by "learned individual differences or situational factors" (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008, p. 167). As a result, researchers have offered incomplete and uncertain answers of who becomes an entrepreneur (Gartner, 1988; Shane & Venkataraman, 2000).

In recognizing this limitation, Nicolaou, Shane, Cherkas, Hunkin, and Spector (2008) suggested that genetic factors may have an influence on entrepreneurial propensity. By comparing identical (monozygotic) and fraternal (dizygotic) twins, who share half of their genetic material, on average, they estimated that approximately 50% of the variance in entrepreneurial tendency can be attributed to genetic factors. This estimate for heritability of entrepreneurship was corroborated in several later studies (Nicolaou & Shane, 2010; Nicolaou, Shane, Cherkas, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2009; Shane, Nicolaou, Cherkas, & Spector, 2010; Zhang et al., 2009). The relevance of genes in entrepreneurial behavior raises various new research questions, including which interactions of genes and environmental conditions tend to result in particular outcomes; how people with particular genes fit with given environments or self-select into them; and how the interplay of individuals and their environment results in prosperity and satisfaction of people or a lack thereof. However, while studies of twins facilitate developing a more complete understanding of the determinants of entrepreneurship by showing that genes apparently do matter, they lack the ability to pinpoint exactly the genes or biological pathways involved, which is a first step towards addressing these questions. Identification of genes associated with entrepreneurial behavior is enabled by genetic association studies that incorporate data on genetic variation on the molecular level. These studies and several related methods are the topic of this thesis and, in particular, the challenges and pitfalls of using them in entrepreneurship research.

From an economics research perspective, understanding the motivations and circumstances under which individuals engage in entrepreneurial activity is important because entrepreneurship is a vital element of well-functioning economies. It is sometimes denoted

as their "scarcest input factor." Entrepreneurs introduce innovations into the economic system and may contribute towards higher productivity levels and hence economic growth (Audretsch & Keilbach, 2004; Carree & Thurik, 2006; Van Stel, Carree, & Thurik, 2005). In addition, market entry by entrepreneurial activity is vital in adjusting markets towards competitive levels (Kirzner, 1973), and even purely imitative entrepreneurial activity can have growth-enhancing effects by stimulating efficiency and promoting the diffusion of technologies (Schmitz, 1989). Furthermore, entrepreneurs create jobs for others (Roessler & Koellinger, 2012) and empirical evidence indicates that entrepreneurship is an early indicator of recovery from recessions (Koellinger & Thurik, 2012).

The research presented in this thesis may ultimately also inform epidemiology and public health policy given that occupations are associated with mortality and morbidity (Adler et al., 1994; Adler & Ostrove, 1999; Marmot, Kogevinas, & Elston, 1987; Winkleby, Jatulis, Frank, & Fortmann, 1992). Moreover, mortality, morbidity, and occupational choice have all been shown to be heritable to a certain extent (Herskind et al., 1996; Manolio et al., 2009; McGue, Vaupel, Holm, & Harvald, 1993; Mitchell et al., 2001; Nicolaou & Shane, 2010; Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; V.B. Hjelmborg et al., 2006; Zhang et al., 2009), suggesting two possible causal pathways through which genes may influence mortality and morbidity: either the genes that influence occupational choice also influence mortality and morbidity, or the effect of genes on mortality and morbidity is mediated by occupational choice and the environment. For example, a mismatch between genetic predisposition and occupational choice may result in decreased happiness or stress which in turn influences mortality and morbidity. In similar vein, a lack of desired social status seems to be associated with earlier death (Rablen & Oswald, 2008).

At first, the search for entrepreneurial genes using genetic association studies may seem more than challenging because entrepreneurship is an ill-defined concept (Shane & Venkataraman, 2000) and the causal pathways from genes to entrepreneurial behavior are long and broad. On the other hand, results from twin studies show that not only variation in entrepreneurial tendency but also in related individual characteristics such as preferences and personality traits can be attributed to genetic factors. For example, twin studies have shown that risk seeking (Cesarini, Dawes, Johannesson, Lichtenstein, & Wallace, 2009), novelty seeking (Ebstein et al., 1996; Kluger, Siegfried, & Ebstein, 2002), general cognitive ability and intelligence (Deary, Spinath, & Bates, 2006; Plomin, 1999; Plomin & Kosslyn, 2001; Plomin & Spinath, 2004), educational attainment (Miller, Mulvey, & Martin, 2001), and overconfidence (Cesarini, Lichtenstein, Johannesson, & Wallace, 2009) are heritable. In addition, entrepreneurship used to be a more obvious (or even natural) occupation, because paid employment is a relatively recent development (Cipolla, 1993; Finley, 1973).

Early genetic association studies were so-called candidate gene studies that examined the association between a *phenotype* (an observable characteristic of an individual that is the result of genes and the environment) and a small number of genetic variants that are selected based on an a priori hypothesis that is derived from information about their biological functioning (Beauchamp et al., 2011). However, these studies often failed to replicate because of low statistical power (due to optimistic expectations about effect sizes) and/or publication bias. Examples of non-replication of candidate gene studies include general intelligence (Chabris et al., 2012), personality (De Moor et al., 2012; Ebstein et al., 1996; Lesch et al., 1996; Paterson, Sunoharo, & Kennedy, 1999; Terracciano et al., 2009; Verweij et al., 2010), and trust (Apicella et al., 2010; Israel et al., 2009). Chapter 2 further discusses shortcomings of candidate gene studies, specifically within the setting of entrepreneurship research by providing an example of a failed replication of an association between a genetic variant in the *DRD3* gene and entrepreneurship.

Fuelled by recent technological developments and the completion of the Human Genome (Venter et al., 2001; Lander et al., 2001) and HapMap projects (The International HapMap Consortium, 2005), genotyping[1] hundreds of thousands of genetic variants in large samples has become financially feasible. This launched an unprecedented era of genetic discoveries by the application of the genome-wide association design where genes are not selected a priori but a very large number of genetic variants spread around the genome are tested individually for association with a phenotype. Genome-wide association studies (GWASs) have been successful in identifying genetic variants associated with numerous complex quantitative traits and diseases (Hindorff et al., 2009; Visscher, Brown, McCarthy, & Yang, 2012).

GWASs focus on genetic variants known as single nucleotide polymorphisms (SNPs, pronounced "snips") covering a high proportion of the common genetic variation in the genome. The first GWAS used only 10,000 genotyped SNPs in 100 individuals (Hu et al., 2005), but the field has evolved enormously. Decreasing genotyping costs and improved statistical techniques have made it possible to analyze millions of SNPs. However, with the increase in the number of SNPs and consequently the number of statistical tests it can be expected on the basis of pure chance that a large number of SNPs will show significant associations. For example, assume that none of the one million SNPs in a GWAS are associated with a phenotype, i.e., that the statistical null hypothesis is correct. If we adopt a 5% significance level for hypothesis testing, performing one million tests will yield 50,000 expected incorrect rejections of the null hypothesis. Hence, to keep the false positive rate at an acceptable level, very stringent significance levels are required in GWAS to adjust for

[1] Genotyping refers to directly measuring the molecular genetic make-up of an individual by examining the individual's DNA sequence using biological assays.

multiple testing. The commonly used Bonferroni correction, for example, suggests a *p*-value lower than 5×10^{-8} if the significance level for the whole family of one million tests is supposed to be 5%. Given this very strict significance level, very large sample sizes are needed to discover associations with weak effects (McCarthy et al., 2008). As a consequence, collaborative research consortia have been assembled to share GWAS data usually analyzed in the form of meta-analysis. The large sample sizes and replication of associations therein most likely reflect that genome-wide significant findings are true positives. Chapter 3 discusses in detail the promises and limitations of GWAS within the setting of economics and entrepreneurship research.

To perform a well-powered GWAS of entrepreneurship we have set up an international research consortium that we have termed the "Gentrepreneur Consortium" (Van der Loos et al., 2010). To the best of our knowledge, this is the first attempt to apply GWAS to an economic outcome of a relatively general nature, entrepreneurship, and will reveal potentials and limitations of this approach for economics research. The first challenge was the need for an accurate phenotype definition. As entrepreneurship is a phenomenon that can materialize in many different forms, different definitions and operationalizations coexist (Shane & Venkataraman, 2000). Moreover, in GWASs there is always a trade-off between phenotype heterogeneity and sample size, and opting for either one has consequences for statistical power. In Chapter 4 we perform a GWAS of entrepreneurship where we have chosen for sample size and investigate the molecular genetics of being self-employed at least once (at least once self-employment), the most widely available proxy for entrepreneurship. Conversely, in Chapter 5 we have opted to minimize phenotype heterogeneity and study the genetics of serial self-employment, a much stricter and scarcely available measure of entrepreneurship. In these chapters, we also perform twin studies to estimate the heritability of the tendency to engage in at least once and serial self-employment. In addition, we use a novel method from molecular genetics to estimate heritability based on molecular genetic data (Yang et al., 2010). This SNP-based heritability can be interpreted as the proportion of variance that is explained by genotyped SNPs in the tendency to engage in at least once and serial self-employment. We also examine whether at least once and serial self-employment can be predicted solely using molecular genetic data.

Finally, Chapter 6 studies the effect of testosterone on entrepreneurial behavior. This is a potential causal pathway from genes to entrepreneurial behavior because testosterone and entrepreneurship are both heritable, and there is some research to suggest that testosterone and entrepreneurial behavior are associated with each other (White, Thornhill, & Hampson, 2006). However, this evidence is based on a small sample size study that has not been replicated. The aim of the study described in Chapter 6 is to verify this previous report of an association using two large, independent, population-based samples of Dutch and German males.

1.2 Thesis Outline, Research Questions, and Main Results

The remainder of this thesis consists of five chapters that answer six research questions. These questions are described in detail below including the main results.

Research question 1: Is a genetic variant in the *DRD3* gene associated with entrepreneurship? (Chapter 2)

A recent small sample size candidate gene study suggests that a genetic variant in the *DRD3* gene is associated with entrepreneurial tendency (Nicolaou, Shane, Adi, Mangino, & Harris, 2011). However, it is now widely accepted that results from such studies generally fail to replicate (Ioannidis, Tarone, & McLaughlin, 2011; Siontis, Patsopoulos, & Ioannidis, 2010). Chapter 2 attempts to replicate this previously reported association in three much larger, independent samples of Dutch males. Using self-employment as a proxy for entrepreneurship we find no evidence to support the hypothesis that the genetic variant in the *DRD3* gene is associated with entrepreneurship. We provide several explanations for this non-replication and discuss the candidate gene approach specifically within the setting of entrepreneurship research.

Research question 2: What are the promises and limitations of GWASs in economics and entrepreneurship research? (Chapter 3)

Within medical genetics, the GWAS approach has been very successful in discovering genetic variants associated with numerous quantitative traits and diseases (Hindorff et al., 2012; Visscher, Brown, et al., 2012). In economics and entrepreneurship research most genetic association studies have been candidate gene studies but the fields have started to adopt the GWAS approach (Beauchamp et al., 2011; Martin et al., 2011; Rietveld et al., 2013; Van der Loos et al., 2010). Chapter 3 discusses basic genetic concepts, the GWAS design, and how GWAS can be used in economics and entrepreneurship research in a way accessible to economists. Moreover, we perform a simulation study to estimate that the required sample size for a well-powered GWAS of entrepreneurship is at least 30,000 participants.

Research question 3: Is entrepreneurship heritable and if so how much is accounted for by SNPs? (Chapters 4 and 5)

Several papers have recently shown that entrepreneurial behavior is heritable (Nicolaou & Shane, 2010; Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Nicolaou et al., 2009; Shane et al., 2010; Zhang et al., 2009). In Chapters 4 and 5 we estimate the heritability of entrepreneurship using twin studies and entrepreneurship operationalized as at least once and serial self-employment to compare the results with estimates from a novel method from molecular genetics to estimate herita-

bility based on molecular genetic data (Yang et al., 2010). For at least once self-employment, we estimate the twin-based heritability to be approximately 50%, in agreement with earlier studies, and show that about half is accounted for by all SNPs considered jointly. The results for serial self-employment also indicate significant twin-based heritability and the SNP-based heritability estimates suggest that a large share is accounted for by SNPs.

Research question 4: Which SNPs / genes underlie the heritability of entrepreneurship? (Chapters 4 and 5)

Given that entrepreneurship is heritable it should, in principle, be possible to discover SNPs that are associated with entrepreneurship in a GWAS. Chapters 4 and 5 perform a GWAS on at least once and serial self-employment, respectively, to identify SNPs that are robustly associated with self-employment. In addition, these chapters perform gene-based tests to examine if SNPs considered jointly within genes are associated with at least once and serial self-employment. For at least once self-employment, we do not identify associated SNPs or genes in a meta-analysis of GWASs across sixteen studies comprising approximately 50,000 participants. For serial self-employment, we find one associated SNP in a GWAS meta-analysis of approximately 8,700 individuals. However, Bayesian reasoning prevents us from interpreting this finding as a true positive. Gene-based tests reveal no associations between any known genes and serial self-employment.

Research question 5: Can entrepreneurship be predicted solely from molecular genetic data? (Chapters 4 and 5)

It has been suggested that molecular genetic data may be used by companies, including banks or venture capitalists, to maximize profits and guide individuals' vocational choices (Nicolaou et al., 2011; Shane, 2010). How likely it is that these promises will be borne out in the near future depends on how well entrepreneurship can be predicted solely from molecular genetic data. Chapters 4 and 5 examine to what extent at least once and serial self-employment can be predicted from currently available molecular genetic data. The results indicate that at least once and serial self-employment cannot be predicted solely using molecular genetic data.

Research question 6: Is testosterone associated with entrepreneurship? (Chapter 6)

Because it is unlikely that genes influence entrepreneurial behavior directly, indirect pathways from genes to entrepreneurship should exist. The effect of testosterone on entrepreneurial behavior is a potential causal pathway because testosterone and entrepreneurship are both heritable, and there is some research to suggest that testosterone and entrepreneurial behavior are associated with each other (White et al., 2006). However, this evidence is

based on a small sample size study that has not been replicated. The aim of the study described in Chapter 6 is to verify this previous report of an association using two large, independent, population-based samples of Dutch and German males. In these samples we find no evidence to support the hypothesis that testosterone and entrepreneurship, operationalized as self-employment, are associated. We discuss several causes for this non-replication and conclude that most likely the previous report of a significant association was due to limited statistical power and/or publication bias.

1.3 Discussion

Until recently, the SSSM took the view that individuals are not born with the necessary skills and personality traits to become entrepreneurs, but that entrepreneurial behavior is solely shaped by environmental factors. For example, Drucker (1985) famously wrote: "Most of what you hear about entrepreneurship is all wrong. It's not magic; it's not mysterious; and it has nothing to do with genes. It's a discipline and, like any discipline, it can be learned." This view stands in stark contrast to recent evidence from twin studies showing that entrepreneurship, at least to a certain extent, is influenced by genetic factors (Nicolaou & Shane, 2010; Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Nicolaou et al., 2009; Shane et al., 2010; Zhang et al., 2009). While these studies suggest that genes can predispose people to entrepreneurship, they lack the ability to identify the specific genes involved. This thesis presents the results of novel research that uses unique molecular genetic data to examine the molecular genetics of entrepreneurship. The results have important consequences for research on the molecular genetics of entrepreneurship and other economic outcomes and behaviors as well.

First, using twin studies, we report significant heritability estimates for entrepreneurship, defined as at least once and serial self-employment, which are in line with previous twin studies (Chapters 4 and 5). Moreover, using a recently-developed method from molecular genetics, we present novel evidence based on actual molecular genetic data that around half of the heritability of at least once and serial self-employment is accounted for by SNPs.

Second, we show in Chapters 4 and 5 that the heritable variation in at least once and serial self-employment is not explained by genes that have been suggested in the literature as candidate entrepreneurship genes (Nicolaou & Shane, 2009; Nicolaou et al., 2011; Shane, 2010). In addition, we find that a specific association between a genetic variant in the *DRD3* gene and the tendency to engage in entrepreneurship fails to replicate in three independent, much larger samples of Dutch males (Chapter 2). When examining the entire genome for associations between SNPs and at least once and serial self-employment in a GWAS (Chapter 3), we fail to find robust evidence for associations (Chapters 4 and 5). It

follows that it is currently impossible to predict entrepreneurial behavior solely from molecular genetic data.

Third, we show that the hormone testosterone is not associated with entrepreneurial behavior in two large, independent, population-based samples of Dutch and German males although this has been hypothesized as a potential causal pathway through which genes may act on entrepreneurial propensity (White et al., 2006).

Taken as a whole, the results presented in this thesis suggest that the heritable variation in entrepreneurship is not the result of one or several well-defined genes, but is likely to be influenced by hundreds or even thousands of genes with a very small effect size each. This is known as a *polygenic* molecular genetic architecture. The results have several implications for future research on the molecular genetics of entrepreneurship. First, the large part of the twin-based heritability of entrepreneurship that is captured by SNPs suggests that research on the molecular genetics of entrepreneurship is, in principle, feasible. We should be able to discover SNPs that are associated with entrepreneurship. Second, the results suggest that the effects of individual SNPs are likely to be very small. It follows that very large sample sizes will be needed in future research, even larger than in the current research, to establish robust associations.

Because it is likely that other economic variables will have similar polygenic molecular genetic architectures (see, for example, Benjamin, Cesarini, Van der Loos, et al., 2012), the results presented in thesis suggest that, although very large sample sizes will be needed, studying the molecular genetics of these other variables should also be feasible. In this respect, this thesis may serve as a practical guide by presenting and discussing relevant methods to study the molecular genetics of economic variables. For example, the GWAS of educational attainment conducted by Rietveld et al. (2013) already builds on and extends the research presented in this thesis. In conclusion, this thesis helps to build the foundations for a novel research field that integrates molecular genetics into economics.

1.4 Publication Status of Chapters

Table 1.1 indicates for each chapter of this thesis the publication status and the research question(s) it addresses. Four chapters have been published in international peer-reviewed journals, one is currently under review, and one is in preparation for submission. The table also lists five other papers to which I contributed and that are either published or under review. The paper by Van der Loos, Groenen, et al. (2011) is a version of Chapter 3 that was published in a Dutch journal. The remaining papers examine the molecular genetics of different traits, but related to entrepreneurship, namely, educational attainment (Beauchamp et al., 2011; Benjamin, Cesarini, Van der Loos, et al., 2012; Rietveld et al., 2013), general intelligence (Chabris et al., 2012), and economic and political preferences (Benjamin, Cesarini, Van der Loos, et al., 2012).

Table 1.1. Publication status of chapters and five other papers.

Chapter	Title	Research question(s)	Publication status	Reference
1	Genome-Wide Association Studies and the Genetics of Entrepreneurship		Published in *European Journal of Epidemiology*	Van der Loos et al. (2010)
2	Candidate Gene Studies and the Quest for the Entrepreneurial Gene	1	Published in *Small Business Economics*	Van der Loos, Koellinger, et al. (2011)
3	Genome-Wide Association Studies in Economics and Entrepreneurship Research: Promises and Limitations	2	Published in *Small Business Economics*	Koellinger et al. (2010)
4	The Molecular Genetic Architecture of Self-Employment	3, 4, and 5	Published in *PLOS ONE*	Van der Loos, Rietveld, et al. (2013)
5	The Molecular Genetics of Serial Self-Employment	3, 4, and 5	Manuscript in preparation	Koellinger et al. (2012)
6	Measures of Bioactive Serum Testosterone Are Not Associated with Entrepreneurial Behavior in Two Independent Observational Studies	6	Manuscript submitted for publication	Van der Loos, Haring, et al. (2013)
	Other papers			
	De Genetica van Ondernemerschap		Published in *ESB Dossier*	Van der Loos, Groenen, et al. (2011)
	Molecular Genetics and Economics		Published in *Journal of Economic Perspectives*	Beauchamp et al. (2011)
	The Molecular Genetic Architecture of Economic and Political Preferences		Published in *Proceedings of the National Academy of Sciences of the United States of America*	Benjamin, Cesarini, Van der Loos, et al. (2012)
	Most Reported Genetic Associations with General Intelligence Are Probably False Positives		Published in *Psychological Science*	Chabris et al. (2012)
	GWAS of 126,559 individuals identifies common genetic variants associated with educational attainment		Manuscript submitted for publication	Rietveld et al. (2013)

CHAPTER 2

Candidate Gene Studies and the Quest for the Entrepreneurial Gene

Based on Van der Loos, Koellinger, et al. (2011).

Abstract

Candidate gene studies of human behavior are gaining interest in economics and entrepreneurship research. Performing and interpreting these studies is not straightforward because the selection of candidates influences the interpretation of the results. As an example, Nicolaou et al. (2011) report a significant association between a common genetic variant in the *DRD3* gene and the tendency to be an entrepreneur. We fail to replicate this finding using a much larger, independent data set. In addition, we discuss the candidate gene approach and give suggestions to avoid the publication of false positives.

2.1 Introduction

In a recent paper, Nicolaou et al. (2011) report a significant association between a common genetic variant (a single nucleotide polymorphism, or SNP) in the dopamine receptor D3 (*DRD3*) gene and the tendency to be an entrepreneur, in a group of 1,335 British subjects. In this candidate gene study, polymorphisms in a set of nine genes were tested for an association with the tendency to be an entrepreneur, resulting in a single significant association. The set of candidate genes consisted of five dopamine receptor genes associated with novelty or sensation seeking and four genes associated with attention deficit hyperactivity disorder (ADHD). These specific genes were selected based upon the notions that ADHD and sensation seeking are more common among entrepreneurs. The authors claim that this is the first evidence of an association between variants in a specific gene and entrepreneurship.

We tried to replicate their findings by performing an association analysis of the 18 SNPs reported in Nicolaou et al. (2011), including the significant association between a SNP in the *DRD3* gene and entrepreneurship, in three much larger, independent groups of Dutch subjects from the Rotterdam Study (Hofman, Grobbee, de Jong, & Van den Ouweland, 1991; Hofman et al., 2009). However, we failed to replicate their finding, and, therefore, we postulate that the reported association is a false positive, probably arising from several shortcomings in the study by Nicolaou et al. We discuss these shortcomings and provide suggestions for future research.

2.2 Replication Study

2.2.1 Data

Our replication study uses data from The Rotterdam Study (Hofman et al., 1991, 2009), a large population-based prospective cohort study of elderly Caucasians ongoing since 1990 in the city of Rotterdam in the Netherlands. The study started with a pilot phase in the second half of 1989. From January 1990 to September 1993, 7,983 participants were successfully recruited in the well-defined Ommoord district in Rotterdam. This formed the initial cohort called Rotterdam Study I (RS-I). The participants were all 55 years of age or over when entering the study. From February 2000 to December 2001, an additional 3,011 participants older than 55 were gathered within a second cohort and interviewed: Rotterdam Study II (RS-II). From February 2006 to December 2008, a third cohort was gathered, Rotterdam Study III (RS-III), consisting of 3,932 individuals of 45 years and older.

In RS-I, 5,974 participants have been successfully genotyped, 2,129 in RS-II and 2,030 in RS-III. Genotyping is performed using the Illumina 550 and 610 K arrays. As the type of array differs between the candidate gene study and our replication study, not all 18 reported SNPs were readily available in the Rotterdam Study cohorts. Therefore, we im-

puted these SNPs from the available genotype data using MACH (Li, Willer, Ding, Scheet, & Abecasis, 2006; Li, Willer, Sanna, & Abecasis, 2009).

We construct a binary variable indicating whether a subject had (1) never been self-employed or (2) been self-employed at least once during his/her complete working life (RS-I) or in his/her current or last occupation (RS-II and RS-III). For RS-I, individuals with an incomplete working life history and individuals who had never had a job are excluded from our study, except those who are classified as self-employed at least once. The rationale for this is that incomplete working life histories could "contaminate" the control group with people who were self-employed at least once. Complete SNP and self-employment data are available for 5,374 subjects (531 cases, 4,843 controls) in RS-I, 2,066 subjects (197 cases, 1,869 controls) in RS-II, and 1,925 subjects (209 cases, 1,716 controls) in RS-III. In this way, our measure of entrepreneurship is equivalent to the definition used by Nicolaou et al. (2011), i.e., "have you ever started a business in your working life". This equivalence is confirmed by a correlation coefficient of 0.87 between the two constructs of self-employment and starting a new business (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008).

2.2.2 Methods

Association analysis is performed for each SNP by logistic regression using the program mach2dat (Li et al., 2006, 2009), which is accessed through a web-based interface called GRIMP (Estrada et al., 2009). For each SNP, two models are estimated: model 1 including the SNP as an independent variable, and model 2 controlling for sex and possible population stratification by including the first four principal components of the genotypic covariance–variance matrix. For RS-III, a dummy for age (≥ 50) is included in the latter model.

To adjust for multiple testing, a Bonferroni correction[2] is applied resulting in a significance level of 0.0028 (0.05 / 18 tests), which corresponds to a significance level of 0.05 for all tests. However, we will argue below that this significance level is arbitrary. Several other choices of significance levels could also be justified, although this does not change our conclusions.

2.3 Results

Tables 2.1, 2.2, and 2.3 show the association results for RS-I, RS-II, and RS-III, respectively, between the 18 reported SNPs and at least once self-employment. In RS-II and RS-

[2] Testing multiple hypotheses will inflate the false positive rate for the entire family of tests. For example, accepting a significance level of 5% and performing 100 tests will yield 5 (100×0.05) expected incorrect rejections of the null hypothesis. One possible solution to keep the number of false positives at an acceptable level is the Bonferroni correction. Applying this often-used adjustment consists of dividing the desired family-wise significance level by the number of independent tests performed to obtain a test-wise significance level.

Table 2.1. Association results using two logit models of at least once self-employment for RS-I.

				Model 1		Model 2	
SNP	**Allele**	**Chr.**	**Freq.**	**Beta**	***p*-value**	**Beta**	***p*-value**
rs1486011	C	3	0.063	0.352	0.0056	0.348	0.0068
rs393795	T	5	0.195	0.064	0.4330	0.046	0.5781
rs409588	T	5	0.193	0.068	0.4021	0.051	0.5402
rs456082	G	5	0.193	0.067	0.4082	0.050	0.5478
rs458860	A	5	0.192	0.068	0.4005	0.051	0.5384
rs460000	T	5	0.191	0.070	0.3880	0.053	0.5229
rs460700	C	5	0.195	0.064	0.4314	0.046	0.5761
rs463379	C	5	0.192	0.069	0.3955	0.051	0.5321
rs464528	T	5	0.192	0.069	0.3972	0.051	0.5342
rs250682	C	5	0.196	0.063	0.4424	0.045	0.5893
rs456774	C	5	0.207	0.104	0.1918	0.090	0.2688
rs1486008	T	3	0.056	0.374	0.0025	0.387	0.0020
rs16822416	A	3	0.056	0.374	0.0025	0.388	0.0020
rs1486009	G	3	0.056	0.374	0.0025	0.388	0.0020
rs464061	A	5	0.211	0.043	0.6117	0.027	0.7542
rs3732783	C	3	0.046	0.365	0.0090	0.384	0.0067
rs4436578	T	11	0.886	0.032	0.7584	0.012	0.9115
rs2975292	G	5	0.640	−0.023	0.7326	−0.002	0.9772

Table 2.2. Association results using two logit models of at least once self-employment for RS-II.

				Model 1		Model 2	
SNP	**Allele**	**Chr.**	**Freq.**	**Beta**	***p*-value**	**Beta**	***p*-value**
rs1486011	C	3	0.057	0.020	0.9330	0.017	0.9420
rs393795	T	5	0.203	−0.038	0.7811	−0.037	0.7860
rs409588	T	5	0.200	−0.038	0.7792	−0.037	0.7852
rs456082	G	5	0.200	−0.038	0.7789	−0.037	0.7848
rs458860	A	5	0.200	−0.038	0.7793	−0.037	0.7854
rs460000	T	5	0.199	−0.038	0.7792	−0.037	0.7855
rs460700	C	5	0.203	−0.038	0.7810	−0.037	0.7859
rs463379	C	5	0.200	−0.038	0.7791	−0.037	0.7853
rs464528	T	5	0.200	−0.038	0.7792	−0.037	0.7853
rs250682	C	5	0.203	−0.037	0.7814	−0.037	0.7861
rs456774	C	5	0.214	−0.011	0.9314	−0.013	0.9241

Table 2.2. (continued)

rs1486008	T	3	0.050	−0.001	0.9969	−0.009	0.9711
rs16822416	A	3	0.050	−0.001	0.9965	−0.009	0.9708
rs1486009	G	3	0.050	−0.001	0.9966	−0.009	0.9709
rs464061	A	5	0.219	−0.071	0.6122	−0.072	0.6074
rs3732783	C	3	0.041	0.063	0.8110	0.052	0.8459
rs4436578	T	11	0.891	0.087	0.6143	0.068	0.6964
rs2975292	G	5	0.648	0.056	0.6234	0.052	0.6495

Table 2.3. Association results using two logit models of at least once self-employment for RS-III.

				Model 1		Model 2	
SNP	**Allele**	**Chr.**	**Freq.**	**Beta**	***p*-value**	**Beta**	***p*-value**
rs1486011	C	3	0.067	−0.068	0.7674	−0.040	0.8652
rs393795	T	5	0.194	0.139	0.2745	0.157	0.2250
rs409588	T	5	0.194	0.139	0.2747	0.156	0.2254
rs456082	G	5	0.194	0.139	0.2746	0.157	0.2252
rs458860	A	5	0.194	0.139	0.2748	0.156	0.2254
rs460000	T	5	0.194	0.139	0.2751	0.156	0.2259
rs460700	C	5	0.194	0.139	0.2744	0.157	0.2249
rs463379	C	5	0.194	0.139	0.2749	0.156	0.2256
rs464528	T	5	0.194	0.139	0.2748	0.156	0.2255
rs250682	C	5	0.194	0.139	0.2750	0.157	0.2253
rs456774	C	5	0.208	0.125	0.3266	0.145	0.2593
rs1486008	T	3	0.059	−0.151	0.5283	−0.104	0.6690
rs16822416	A	3	0.059	−0.151	0.5284	−0.104	0.6690
rs1486009	G	3	0.059	−0.151	0.5283	−0.104	0.6690
rs464061	A	5	0.214	0.151	0.2509	0.175	0.1896
rs3732783	C	3	0.050	−0.013	0.9583	0.033	0.8952
rs4436578	T	11	0.894	−0.086	0.6003	−0.075	0.6519
rs2975292	G	5	0.644	−0.021	0.8467	−0.053	0.6324

III, none of the SNPs are even remotely significant in both models, while the estimation results for RS-I require more explanation.

Nicolaou et al. (2011) report a significant association between SNP rs1486011 and the tendency to be an entrepreneur. This SNP is not significantly associated in RS-I at the chosen level of significance of 0.0028. Moreover, the positive coefficient suggests the opposite; carrying the C allele seems not to decrease the probability of being self-employed at least once, as reported by Nicolaou et al., but to increase the odds.

Further inspection of the results indicates that three SNPs within the *DRD3* gene, rs1486008, rs16822416, and rs1486009, survive our Bonferroni-corrected significance level of 0.0028. However, the direction of the effects is opposite to the associations reported in the candidate gene study. Although we cannot reject the hypothesis that the *DRD3* gene is associated with entrepreneurship based on these results, they do not support the effect of the G allele of SNP rs1486011 reported by Nicolaou et al. (2011).

2.4 Discussion

We performed an association analysis of 18 SNPs in the *DRD2*, *DRD3*, and *SLC6A3* genes in three independent groups of Dutch subjects. The set of analyzed SNPs includes a SNP previously reported to be significantly associated with entrepreneurship by Nicolaou et al. (2011). Our study fails to replicate this association and, in fact, finds several other significant associations with opposite effects to those reported by Nicolaou et al.

There are several shortcomings with the candidate gene study that lead us to suspect that the reported association is a false positive and that our results should also be interpreted with care. These shortcomings are lessons learned from the era of candidate gene studies, usually pursued with ill-defined markers across genes, small samples, and/or lacking replication.

Indeed, there are numerous examples of small-scale candidate gene studies that report significant associations with behavioral traits that could not be replicated. For instance, Israel et al. (2009) report an association between a variant of the *OXTR* gene and the dictator game. Apicella et al. (2010) fail to replicate this association. Other studies report an association between a genetic variant in the serotonin transporter gene and anxiety-related traits such as harm avoidance (Lesch et al., 1996; Vormfelde et al., 2006) that others fail to replicate (Becker, El-Faddagh, Schmidt, & Laucht, 2007; Lang et al., 2004). Hence, the decisive proof of a true association is replication in an independent study, a feature that the study of Nicolaou et al. (2011) lacks. Lastly, Ioannidis (2005) shows that the pre-study probability of a genetic association being true is generally extremely low, and consequently, the post-study probability is also low.

With regard to the candidate gene study, first, we believe that the selection of candidates by Nicolaou et al. (2011), although seemingly sound, is largely arbitrary. The set comprises genes previously thought to be associated with novelty or sensation seeking and ADHD, characteristics that are hypothesized to be more common among entrepreneurs. Following this line of thought, there are many other candidate genes, such as the serotonin 2A and 1B transporters (*HTR2A* and *HTR2B*), dopamine and serotonin transporters (*SLC6A4*), dopamine beta-hydroxylase (*DBH*), monoamine oxidase B (*MAOB*), and genes associated with testosterone level. Furthermore, probably more than half of all genes are related to brain function or to the expression of proteins in the brain (Sandberg et al., 2000)

and could therefore be candidates. This leads to hundreds of thousands of potential candidate loci and makes the candidate gene approach infeasible for the study of complex behaviors such as entrepreneurship.

Second, the selection criteria of SNPs within the chosen candidate genes are confined to the coding regions. A complete overview of the selected SNPs is lacking, although Nicolaou et al. (2011) report that the SNPs from the coding regions of the nine candidate genes were selected. SNPs in regulatory non-coding regions are not considered, although these could have substantial effects on a given phenotype (for an overview, see Hindorff et al., 2012).

Third, the hypothesis that dopamine receptor genes are associated with novelty or sensation seeking is itself based on mixed evidence from small-scale studies that could not always be replicated. For example, Ebstein et al. (1996) report a significant association between a variant of the *DRD4* gene and novelty seeking, which could not be replicated by Malhotra et al. (1996). A recent meta-analysis by Munafo, Yalcin, Willis-Owen, and Flint (2008) concludes that the *DRD4* gene may be associated with measures of novelty seeking and impulsivity, but significant evidence of publication bias was found. Finally, Verweij et al. (2010) report that the *DRD4* gene is not significantly associated with the novelty seeking dimension of Cloninger's temperament scales, although the study had 91.5% power to detect SNPs that explain 1% of the variance.

Obviously, the choice of candidate genes is limited by knowledge of the biological function of genes and their possible relationship with entrepreneurship. Recent technological advancements have enabled so-called genome-wide association studies (GWASs), which are considered hypothesis-free as no prior knowledge about gene function is needed. Instead of hypothesizing relationships between genes and a trait a priori, a GWAS systematically interrogates the entire genome for associations between genetic variants (SNPs) and a trait. In current GWASs, millions of SNPs are statistically tested for association, leading to a severe multiple testing problem. Therefore, it is conventional wisdom to apply a very stringent significance level of $p < 5 \times 10^{-8}$ (McCarthy et al., 2008) to each tested SNP to control the false positive rate. Despite this, GWASs have been remarkably successful in uncovering associations between common genetic variation and human traits and diseases (Hindorff et al., 2009) and are gaining interest in the social sciences (Koellinger et al., 2010; Van der Loos et al., 2010).

Given that GWASs are currently the way forward in genetics research and that genome-wide data are available in the data set of Nicolaou et al. (2011; see also http://boss.blogs.nytimes.com/2009/09/21/literally-born-entrepreneurs/), a comprehensive, hypothesis-free GWAS of entrepreneurship is an attractive alternative to the hypothesis-based candidate gene study. Obviously, the reported association would not have reached the accepted genome-wide significance level of $p < 5 \times 10^{-8}$. Associations are often re-

ported to be false positives if a set of candidate genes is selected, while not all relevant genes and SNPs are considered (e.g., Apicella et al., 2010; Becker et al., 2007; Israel et al., 2009; Lang et al., 2004; Lesch et al., 1996; Vormfelde et al., 2006).

2.5 Conclusion

We tried to replicate the significant association between a variant in the DRD3 gene and entrepreneurship reported by Nicolaou et al. (2011), using three much larger, independent groups of Dutch subjects from the Rotterdam Study, and fail to do so. In fact, we find that the reported association has an opposite, insignificant effect in our study. Moreover, we find several other associations with opposite effects among the SNPs reported by Nicolaou et al. As explained above, it is difficult to choose a level of significance. All associations would be rendered insignificant using the level of significance commonly used in the GWAS approach ($p < 5 \times 10^{-8}$), which is the superior method, in our view.

As another extreme, we can argue that not all 18 SNPs in our analysis are independent, but are correlated, i.e., they are in *linkage disequilibrium*. Consequently, the number of independent statistical tests would be less than 18, and a higher significance level could have been used. Assuming that, for simplicity, SNPs within a gene are highly correlated, we could effectively perform three independent statistical tests (with the *DRD2*, *DRD3*, and *SLC6A3* genes), resulting in a Bonferroni-adjusted significance level of 0.0167 (0.05 / 3). Adopting this significance level, SNPs rs1486011 and rs3732783 would become significantly associated with entrepreneurship next to the three other SNPs reported above, but again with opposite effects to those reported by Nicolaou et al. (2011). Thus, relaxing or tightening the significance level does not change our conclusion; we fail to replicate the results of the candidate gene study, and we emphasize that a hypothesis-free GWAS in an adequately powered setting is the preferred approach.

CHAPTER 3

Genome-Wide Association Studies in Economics and Entrepreneurship Research: Promises and Limitations

Based on Koellinger et al. (2010).
A version in Dutch has been published as Van der Loos, Groenen, et al. (2011).

Abstract

The recently developed genome-wide association study (GWAS) design enables the identification of genes specifically associated with economic outcomes such as occupational and other choices. This is a promising new approach for economics research which we aim to apply to the choice for entrepreneurship. However, due to multiple testing issues, very large sample sizes are needed to differentiate between true and false positives. For a GWAS on entrepreneurship, we expect that a sample size of at least 30,000 observations is required.

3.1 Introduction

There are two popular views on what makes an entrepreneur. The first is that anyone can learn the necessary skills provided (s)he puts in enough time and effort. The second is that people are either born with the right personality and skills or they are not, and there is not much that one can do about it. Obviously, which of these two stories is true has far-reaching implications for individual behavior and economic policies. As we discuss below, there is increasing evidence that inherited qualities play a role in occupational choice with recent scientific advances showing different pathways through which genes can influence entrepreneurial behavior. However, in contrast to popular views, a genetic influence does not imply any kind of determinism, irrelevance of the environment, or of free will, as we discuss later.

The recent scientific breakthroughs that make it possible to discover the genetic basis of human behavior and traits are linked to the results of the Human Genome (Collins, Morgan, & Patrinos, 2003) and HapMap projects (The International HapMap Consortium, 2005). These projects decoded the human genome and identified those genetic regions where humans frequently exhibit differences, which is only a very small part of the entire genome. This resulted in new technological developments that allow the genotyping[3] of hundreds of thousands of markers in large samples at reasonable costs. In particular, a new generation of studies of variations across the entire human genome, called genome-wide association studies (GWASs), have launched an unprecedented era of genetic discoveries, already resulting in more than 500 published studies, identifying common variants associated with numerous complex quantitative traits and diseases (Hindorff et al., 2009). GWASs focus on so-called single nucleotide polymorphisms (SNPs, pronounced "snips"), base pairs that differ between members of a species, which cover a high proportion of the common genetic variation within the genome. This study design provides insights into biological processes and improves our understanding of the biological origins of differences among human beings. This is an important step toward putting the old debate about whether entrepreneurs are born or made through a rigorous scientific test focusing on the genes. Furthermore, this study design is also applicable to various other outcomes of economic relevance, such as educational attainment, risk preferences, and income.

From an economics perspective, the idea that genes influence behavior seems far-fetched at first glance. Typically economists focus on understanding the role of the environment in shaping human behavior, the interactions of people, and the consequences of these interactions. Economists sometimes find it convenient to study the behavior of *repre-*

[3] Genotyping refers to determining the genotype of an individual by the use of biological assays which are also called DNA microarrays. These microarrays integrate several laboratory functions on a single chip that is suitable for high-throughput screening methods.

sentative agents (Hartley, 1996; Kirman, 1992). However, there is ample scope for individuality in economic models, which is typically formalized in the form of preferences[4] or productivity values[5] that depend on personal characteristics. Such individual differences are likely to have important economic implications. For example, we know that occupational choice depends on risk and uncertainty preferences (Iyigun & Owen, 1998; Kihlstrom & Laffont, 1979; Knight, 1921), as well as on non-monetary preference for independence (Benz & Frey, 2008; Block & Koellinger, 2009), educational attainment (Evans & Leighton, 1989), skills (Jovanovic, 1994; Laussel & Le Breton, 1995; Lazear, 2004, 2005; Roessler & Koellinger, 2012), gender (Grilo & Thurik, 2008) and a tendency to be overconfident and overly optimistic (Camerer & Lovallo, 1999; Koellinger, Minniti, & Schade, 2007). All these individual attributes are candidates for an indirect genetic influence on occupational choice.[6]

In fact, there is growing empirical evidence from studies of twins that individual characteristics, which can affect the tendency to become an entrepreneur, are indeed at least partially due to genetic differences. Examples include preferences for risk seeking (Cesarini, Dawes, et al., 2009), altruism in dictator games (Israel et al., 2008; Knafo et al., 2008), job satisfaction (Arvey, Bouchard, Segal, & Abraham, 1989), vocational interests (Betsworth et al., 1994), work values (Keller, Bouchard, Arvey, Segal, & Dawes, 1992), novelty seeking (Ebstein et al., 1996; Kluger et al., 2002), gambling (Comings et al., 1996; Pérez de Castro, Ibáñez, Torres, Sáiz-Ruiz, & Fernández-Piqueras, 1997), general cognitive ability and intelligence (Deary et al., 2006; Plomin, 1999; Plomin & Kosslyn, 2001; Plomin & Spinath, 2004), educational attainment (Miller et al., 2001), and overconfidence (Cesarini, Lichtenstein, et al., 2009).

In addition, empirical evidence suggests that entrepreneurship tends to run in families. Lentz and Laband (1990) observe that around half of all US self-employed proprietors are second-generation business owners. Evans and Leighton (1989) find that the likelihood of self-employment increases if the father is a manager, and decreases if the father is unskilled. Furthermore, Dunn and Holtz-Eakin (2000) find that parental self-employment both increases the fraction of time that offspring spend in self-employment and reduces the age at which they enter. Colombier and Masclet (2008) find intergenerational correlation for self-employment in France. Andersson and Hammarstedt (2010) show that having both a self-employed father and a self-employed grandfather positively affects self-employment

[4] In economics, the term preference typically refers to theoretical assumptions about the rank order between different choices according to the degree of desirability to an individual.

[5] For example, the labor productivity of a person measures output per labor-hour, given a particular production technology and capital input. Differences in labor productivity are often attributed to personal characteristics such as education or experience.

[6] Other attributes of an environmental nature such as (the threat of) unemployment (Thurik, Carree, Van Stel, & Audretsch, 2008) and the institutional environment (Freytag & Thurik, 2007) may play moderating roles.

propensities for third-generation male immigrants in Sweden. Finally, Van der Zwan, Thurik, and Grilo (2010) show that people with self-employed parents climb the *entrepreneurial ladder* more quickly than those without such parents. It seems likely that self-employed parents transfer relevant skills and familiarity with entrepreneurial behavior to children. But it could also be that inherited characteristics explain the observed intergenerational effects. Indeed, several comparative twin studies suggest a potential genetic influence on the propensity to become self-employed (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Zhang et al., 2009).

In late 2007, these thoughts and findings encouraged us to start investigating the human genome to identify genetic causes of entrepreneurial behavior using GWAS. We assembled a multidisciplinary research group of economists and genetic epidemiologists, establishing the Gentrepreneur Consortium (Van der Loos et al., 2010). To the best of our knowledge, this is the first attempt to apply GWAS to an economic outcome of a relatively general, and hence complex, nature. We are aware that the entrepreneurial choice is possibly a very complex one to explain because entrepreneurship is a multidimensional phenomenon about which there is no general agreement. Not only have psychology, economics, anthropology, and business studies widely different views but also the more popular view of what entrepreneurship is, seems to vary with time and space (Bygrave & Hofer, 1991; Verheul, Uhlaner, & Thurik, 2005; Wennekers & Thurik, 1999). In the present study we measure entrepreneurial activity with self-employment which is an established and widely used measure.

Our GWAS uses data from the Rotterdam Study (Hofman et al., 2009). The Rotterdam Study is a prospective cohort study, hosted at the Erasmus University Rotterdam Medical Center. The majority of the genotyped individuals in the Rotterdam Study provided data on their occupational status, allowing us to study entrepreneurial behavior by looking at self-employment. The present chapter is inspired by our ongoing work with this data and employs simulations to illustrate important identification issues in GWAS in general. Our simulations mimic several characteristics of the Rotterdam Study, such as sample size ($n \approx 10{,}000$), the prevalence of self-employment (~10% of the sample) and the number of SNPs (~550,000).

We presented preliminary results using RS-I at the Behavior Genetics Association June 2008 conference in Louisville, Kentucky (Groenen et al., 2008). Since then our work has focused on replicating results using independent samples and we have now embedded our effort to assemble a working group (Gentrepreneur) within the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium (Psaty et al., 2009). As we discuss below, replication is crucial for this type of research and our efforts to replicate the findings from our discovery cohort are ongoing.

In this chapter, we describe the GWAS design and how it can be applied to study economic outcomes. We illustrate the statistical multiple testing problem that arises in this context when using simulation studies that closely mimic a GWAS setting on entrepreneurship. Following current best practice in genetics research, we discuss how strict confidence levels in combination with large sample sizes are required to identify genes that are truly associated with entrepreneurship or other economic outcomes. Furthermore, the interpretation of findings from GWAS on economic outcomes is not straightforward and this chapter provides several guidelines in this regard.

We begin by describing some basic genetic concepts and the principles underlying genome-wide association studies (GWASs) in Section 3.2. From the set-up of GWASs, the multiple testing problem arises, which we describe in detail and illustrate with a series of simulations in Section 3.3. The interpretation of results from GWASs is discussed in Section 3.4. Section 3.5 concludes and outlines some possible future potential of GWASs for economics and entrepreneurship research.

3.2 Basic Genetic Concepts and GWAS

The human genome comprises all genetic information in human cells and consists of 23 chromosomal pairs (46 in total); half is inherited from the mother and half from the father. These chromosomes "package" DNA molecules that encode the genetic information in a linear sequence of chemical bases along two DNA strands. A DNA strand is a polymer of nucleotides. Each nucleotide is a building block consisting out of a phosphate, a sugar, and a base. The base in a nucleotide can be Adenine (A), Cytosine (C), Guanine (G), or Thymine (T); thus there are four distinct nucleotides. DNA is structured as a double helix in which two DNA strands are held together by weak hydrogen bonds to form a DNA duplex. Hydrogen bonding occurs between the bases of opposing nucleotides along the two strands: Adenine always binds to Thymine and Cytosine always binds to Guanine. Consequently, two DNA strands of a DNA duplex are said to have complementary sequences and the sequence of one DNA strand can easily be inferred if the DNA sequence of its complementary strand is already known. It is usual, therefore, to describe a DNA sequence by writing the sequence of bases for only one strand. For example, one individual may have inherited the AA nucleotides for one particular position on a pair of chromosomes (i.e., a *genotype*). This would imply the individual inherited an A base from the paternal chromosome and an A base from the maternal chromosome. Another individual may have inherited AG nucleotides at the same position, a different base from each of the two parents, while a third may have inherited both GG nucleotides from each parent. Alternative bases in a nucleotide at the same physical locus are called *alleles*. A DNA sequence on one position of the genome that exhibits at least 1% variation between members of a species is called a single nucleotide polymorphism (SNP). The minor allele frequency (MAF) refers

to the frequency of the less common allele of a SNP in a population. People having two copies of the same allele are said to be homozygous for this allele. On the other hand, individuals having two different alleles are called heterozygous.

Almost all human DNA—99.9 percent of the three billion nucleotides that make up the human genome—is identical from person to person. The remaining 0.1% of the genome varies by SNPs (and other types of genomic variation), which is what makes humans different from each other. The total number and locations of SNP markers that need to be genotyped in order to detect an association between common genetic variants and an outcome of interest (also known as the phenotype of an individual) was identified by the HapMap project (The International HapMap Consortium, 2005). Facilitated by the results of the HapMap project, high throughput array-based technologies for whole-genome SNP analysis were recently developed.

GWAS is facilitated by a phenomenon called linkage disequilibrium (LD). LD refers to the non-random way SNPs are inherited together, i.e., many SNPs on the human genome are systematically correlated. SNPs in perfect linkage disequilibrium are inherited together, while SNPs in perfect linkage equilibrium are inherited randomly. LD makes it possible to discover which SNP is causing an outcome even if the SNP is not genotyped. In this case SNPs that are genotyped and in LD with the causal SNP are associated with the outcome. Thus, when a significant association is found between a SNP and an outcome, the association is not necessarily causal. However, the known systematic correlations of SNPs may still enable researchers to identify the causal gene by looking up SNPs that are in LD to the candidate loci.[7] LD patterns in the human genome have been charted by the HapMap project and are used to reduce the number of SNPs that need to be included in an assay to cover a broad spectrum of the genome.

Typically, genotyping is currently done with 550,000 SNP arrays that, after data cleaning, tend to deliver information about the specific alleles for around 500,000 SNPs that are available for statistical analysis. Although this already gives a high resolution image of the human genome, next generation microarrays will allow researchers to assay 2–12 million markers per sample, including comprehensive coverage of both common and rare variants.

The basic GWAS design is to associate an outcome of interest, such as the presence of a disease, an IQ score, or the employment status of an individual (called phenotype), with SNPs on the chip, usually by carrying out a bivariate statistical test for each SNP For a binary outcome like entrepreneurship (with $y = 1$ meaning the individual is an entrepreneur, and $y = 0$ otherwise), the bivariate test performed for each SNP typically consists of a Pearson's χ^2 test for independence within a two-by-two contingency table. The columns in

[7] In practice, the identification of the true causal gene is limited by the fact that the biological function of most SNPs is still not well understood.

Table 3.1. Genotype counts for a particular SNP and binary outcome.

Genotype	$y = 1$	$y = 0$
AA	a	b
AG	c	d
GG	e	f

Table 3.2. Allele counts for a particular SNP and binary outcome.

Allele	$y = 1$	$y = 0$
A	$2a + c$	$2b + d$
G	$2e + c$	$2f + d$

this table indicate the status of the outcome, while the rows indicate one of two alleles for a certain SNP. The table is constructed by collapsing a three-by-two table into the two-by-two table. For example, Table 3.1 classifies individuals according to their genotype for a certain SNP, which is AA, AG, or GG, and according to their status of the outcome. The table shows there are a subjects with $y = 1$ with genotype AA for this SNP, b subjects with $y = 0$ with genotype AA, and so on. This table can be collapsed into a two-by-two table by counting the number of alleles for each allele of this SNP (A and G). This results in Table 3.2, where the letters refer to the ones used in Table 3.1. This is done to increase the power of the test, as the test within the two-by-two table is a 1 degree of freedom test in contrast to the 2 degrees of freedom test within the three-by-two table. A χ^2 test for independence is carried out for each SNP in the study based on tables like Table 3.2.

Alternatively, a logistic regression on the outcome of interest can be carried out for each SNP, and each regression equation may include additional control variables. If 500,000 SNPs are available for statistical analyses, this implies that 500,000 χ^2 tests or 500,000 logistic regressions must be conducted. Typically, these analyses are carried out using specialized software such as PLINK (Purcell et al., 2007).

However, collapsing the two-by-three table into a two-by-two table can only be done under the assumption that the so-called Hardy-Weinberg Equilibrium (HWE) (Hardy, 1908; Weinberg, 1908) holds within the complete sample (Guedj, Nuel, & Prum, 2008; Sasieni, 1997). HWE is a population genetics law stating that genotype and allele proportions are constant in a population from generation to generation, given that the population is large, mating is random, there are no mutations, and there is no selection or migration. Deviations from HWE may indicate one of the above-mentioned phenomena, but may also imply genotyping errors or population stratification. Therefore, it is imperative to test SNPs for HWE before running a GWAS. Understanding HWE and how to test for it requires some knowledge of the mathematics of HWE. It is quite straightforward and as follows. Assume that the proportions of the alleles A and G in a population are given by p and q, respectively. Furthermore, assume that p and q are identical for females and males in a population, mating is random, and that the population is large. Under these assumptions, a so-called Punnett square can be constructed (see Table 3.3) to derive the possible genotypes in the next generation and their proportions. Table 3.3 shows that offspring in

Table 3.3. Punnett square for the alleles A and G with proportions p and q. Proportions are given in parentheses.

	Mother allele	
Father allele	**A (p)**	**G (q)**
A (p)	AA (p^2)	AG (pq)
G (q)	AG (pq)	GG (q^2)

the next generation randomly receive either the A or the G allele from their mother and father. This results in the three possible genotypes: AA, AG, and GG. Furthermore, under the independence assumption the expected proportions of the genotypes in the next generation are p^2 for A homozygotes, $2pq$ for heterozygotes, and q^2 for G homozygotes. Finally, based on the above, and given that the allele and genotype proportions must sum to one, we can derive two equations: $p + q = 1$ and $p^2 + 2pq + q^2 = 1$.

There are three steps to perform the test of whether a specific SNP fulfills HWE: First, the proportion of the observed alleles in the population is calculated. Second, using the latter equation and the computed allele proportions, the expected genotype proportions can be obtained. Finally, the expected genotype proportions can be compared to the genotype proportions observed in the population using a simple one degree of freedom χ^2 test (Crisp, Beaumont, Flowerdew, & Vardy, 1978).

3.3 Identification of True Positives

The very large numbers of independent statistical tests that must be carried out in this research design lead to a severe multiple testing problem. In other words, it is expected that just on the basis of pure chance a large number of SNPs will show highly significant associations even if there is no actual relationship between a SNPs and the studied outcome. For example, assume that none of the analyzed 500,000 SNPs are truly associated with the outcome, i.e., the statistical null hypothesis of no association between SNP and outcome is correct. If we adopt a 1% significance level for hypothesis testing, performing 500,000 tests will yield 5,000 expected incorrect rejections of the null hypothesis (i.e., false positives). Even an apparently stringent significance level of 0.00001 (equivalent to a p-value of 10^{-5}) still leads to 5 false positives on average. Not surprisingly, many GWAS often yield SNPs with p-values in this range, even studies with relatively small samples. As a result, many early GWAS studies reported findings that could not be later replicated (Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002). Hence, to keep the false positive rate at an acceptably low level, stringent significance levels are now used in GWAS to compensate for multiple testing. The often-used Bonferroni correction, for example, sug-

gests a significance level of 10^{-7} for individual tests in order to obtain a 5% significance level for the whole family of 500,000 tests. On the other hand, due to linkage disequilibrium one is essentially conducting more tests than the number of genotyped SNPs. The generally accepted opinion is to account for at least 1 million independent tests in a European descent GWAS (Hoggart, Clark, De Iorio, Whittaker, & Balding, 2008; McCarthy et al., 2008; The International HapMap Consortium, 2005). Based on this, the Bonferroni correction proposes a significance level of 5×10^{-8} to obtain a family-wise significance level of 5%. This level is often referred to as *genome-wide significance* and only SNPs that pass this threshold are typically considered to be true positives. However, to reach such high levels of significance, very large sample sizes are needed to be able to discover associations with weak effects (McCarthy et al., 2008).

To demonstrate the need for large sample sizes in order to find small effects, we performed several simulation studies that mimic the situation of a GWAS on entrepreneurship. We simulated datasets of three different sizes (n = 1,000, n = 10,000, and n = 30,000) with 550,000 SNPs for each observation. The SNPs are unlinked and in perfect linkage equilibrium for different sample sizes. Subsequently, a GWAS was performed on the simulated data sets. Simulation and association was performed using PLINK software (Purcell et al., 2007). For the simulation of SNPs a trait prevalence of 10% in the population was assumed, which is roughly comparable to the prevalence of entrepreneurship in both the Netherlands and in our discovery cohort, the Rotterdam Study. Therefore, to mimic the true setting as closely as possible, the ratio between non-entrepreneurs and entrepreneurs is also 9 to 1 in the simulated data sets. The allele frequencies range from 0 to 1 and the effect allele is assumed to act multiplicatively, i.e., the odds ratio for people having two copies of the effective allele is the square of the odds ratio associated with having just one copy of the effective allele. Note that this amounts to an additive effect on the log scale.

Before the association analysis, SNPs that failed a test of Hardy-Weinberg equilibrium (HWE) at the 10^{-6} level in subjects with $y = 0$ were dropped. In the data sets for n = 10,000 and n = 30,000 this resulted in 1 and 18 SNPs, respectively, being dropped.

No SNPs were dropped due to HWE testing in the other data set. After that, alleles with a minor allele frequency (MAF) smaller than 5% were also filtered out. For all three data sets approximately 55,000 SNPs failed the MAF filter and were dropped from the analysis. As said, testing for HWE in subjects with $y = 0$ is necessary for the χ^2 test within a two-by-two table to be valid. SNPs in the simulation study can be out of HWE because they are generated randomly not taking HWE into account, whereas in practice, in absence of true association, deviation from HWE proportions will very likely reflect genotyping errors. Of the 550,000 SNPs, five sets of each thirty SNPs were simulated with a known association with the trait with odds ratios of 1.2, 1.5, 1.7, 2 and 3. The remaining SNPs (549,850 in total) were simulated with an odds ratio of exactly one and, consequently, are

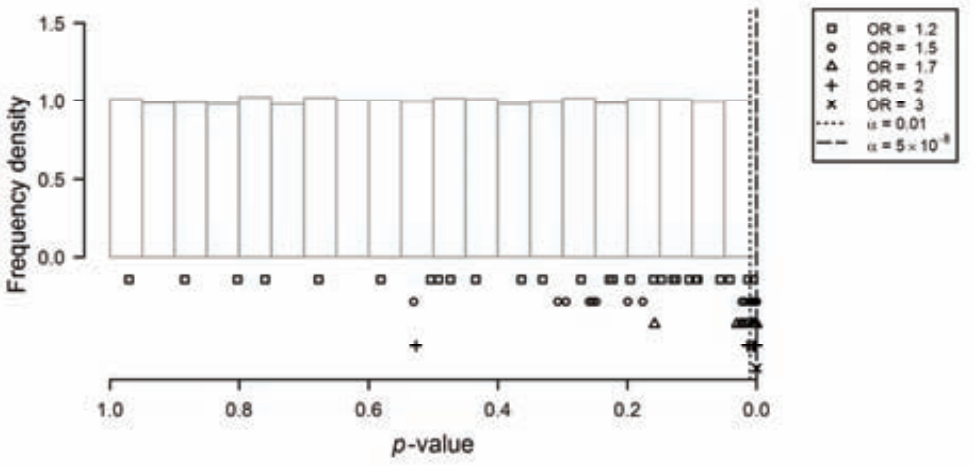

Figure 3.1. Histogram of simulated *p*-values with $n_{y=1} = 100$ and $n_{y=0} = 900$. In total 550,000 SNPs are simulated, including 5 sets of 30 SNPs with odds ratios greater than 1, i.e., 150 SNPs have a known association with $y = 1$. Frequency density is the relative frequency divided by the bin width so that the area of all bins sums to one.

not associated with the trait. Finally, the analysis was performed using χ^2 tests for independence in a two-by-two table for each SNP, as described above.

The results of the simulation studies are plotted as density histograms of the *p*-values (Figures 3.1 and 3.2). The two figures are only different in the sample size used for the analysis, with 1,000 observations for Figure 3.1 and 10,000 observations for Figure 3.2. In these histograms the *y*-axis is the frequency density, which is the relative frequency divided by the bin width, which is 0.05 given that 20 bins of equal width are used on a scale from 0 to 1 (Sturges, 1926). Note that the total area of all bins sums to one and the bin size multiplied by its density is the relative frequency of the observations falling in that bin. Furthermore, below the histograms the *p*-values of the associated SNPs are plotted using different symbols to indicate different odds ratios. The dotted line indicates the conventional significance level of 0.01. Finally, the dashed line is the genome-wide significance level of 5×10^{-8}. Figures 3.1 and 3.2A show the entire range of *p*-values generated by the GWAS. As one can expect, the randomly generated SNPs result in an almost uniform distribution of *p*-values. Importantly, the true positives with an odds ratio of greater than one cannot be differentiated from the other SNPs when small samples, with $n = 1{,}000$, are used, as shown in Figure 3.1. Most true positives with odds ratios of smaller than 2 do not even reach the conventional significance level of 0.01. Apparently, this study is severely underpowered to detect true positives with low odds ratios.

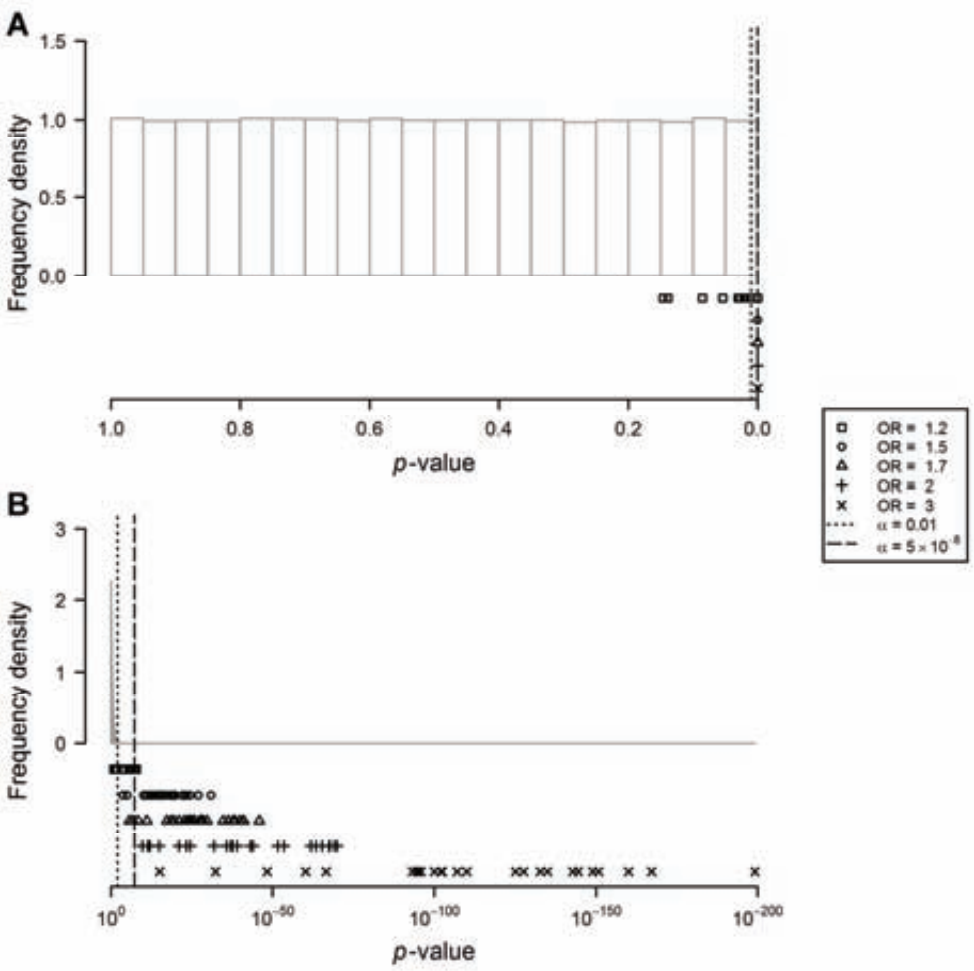

Figure 3.2. Histograms of simulated *p*-values with $n_{y=1}$ = 1,000 and $n_{y=0}$ = 9,000. In total 550,000 SNPs are simulated, including 5 sets of 30 SNPs with odds ratios greater than 1, i.e., 150 SNPs have a known association with $y = 1$. Frequency density is the relative frequency divided by the bin width so that the area of all bins sums to one. Panels A and B show the same data, except that in panel B the x-axis is transformed to the $-\log_{10}$ scale.

In contrast, when using the larger sample size of n = 10,000, clustering of true positives with an odds ratio of 1.5 and higher is exhibited around the dashed line at the right of panel A of Figure 3.2, which indicates genome-wide significance. However, from this figure it is not clear how well the study differentiates between true and false positives at *p*-values smaller than 0.01. One must zoom into this area to learn more about the power of

GWAS at larger samples sizes. One way to zoom into the relevant area of low *p*-values in panel A of Figure 3.2 is by transforming the *x*-axis to the $-\log_{10}(p\text{-value})$. This is done in panel B of Figure 3.2.

Another way to better visualize false and true positives is to plot the different odds ratios against the *p*-values using a $-\log_{10}$ scale on the *x*-axis (Figures 3.3A–C). These plots make it possible to see how associated SNPs compare to un-associated SNPs at different sample sizes and odds ratios. Again, the conventional confidence level of 0.01 is indicated by a dotted line and genome-wide significance at 5×10^{-8} with a dashed line.

Figure 3.3A plots the results of the GWAS with a sample size of $n = 1{,}000$, similar to Figure 3.1. Apparently it is impossible to differentiate between false positives and true positives with an odds ratio less than 3 with this sample size, while true positives with an odds ratio greater than or equal to 3 will probably be detected. In addition, most true positives with an odds ratio smaller than 1.7 do not even reach the conventional significance level of 0.01 and will remain undetected. A GWAS on entrepreneurship with a sample size of $n = 1{,}000$ is severely underpowered to detect true positives with low odds ratios.

A tenfold increase in sample to $n = 10{,}000$ resolves these problems for most SNPs with an odds ratio of 1.5 or higher, as shown in Figure 3.3B. However, the low odds ratios of 1.2 still cannot be differentiated from false positives because they are unlikely to reach the threshold level of genome-wide significance. Furthermore, we see that the genome-wide significance threshold is rather conservative: No false positives cross this threshold, but a few true positives with odds ratios of 1.4 and higher fall slightly below the cut-off significance level. Hence, these SNPs will be reported as false negatives although they have very low *p*-values between 10^{-6} and 5×10^{-8}.

Figure 3.3C repeats the same exercise with a sample size of $n = 30{,}000$. This time, all SNPs with an odds ratio of 1.5 or higher are correctly identified. Also, a majority of SNPs with an odds ratio of 1.2 are detected and can be differentiated from ineffective SNPs.

In addition to the effect size (odds ratio) of the effective allele, other factors also influence the power of genetic association studies, such as the chosen type 1 error, the minor allele frequency (MAF), the linkage disequilibrium of the marker, and the true-associated variant. There are also confounding factors such as population structure and geography, misclassification errors, and selection biases (Wang, Barratt, Clayton, & Todd, 2005). Based on the genetic power calculator by Purcell, Cherny, and Sham (2003), Figure 3.4 illustrates the joint influence of MAF and odds ratios on the required sample size that is needed to detect true positives in a sample with a 10% share of individuals who exhibit $y = 1$, again closely matching the typical set-up of a GWAS on entrepreneurship. For example, with a MAF of 20% and an odds ratio of 1.3, the figure shows that a sample of approximately $n = 15{,}000$ is needed to have an 80% probability of detecting a true association.

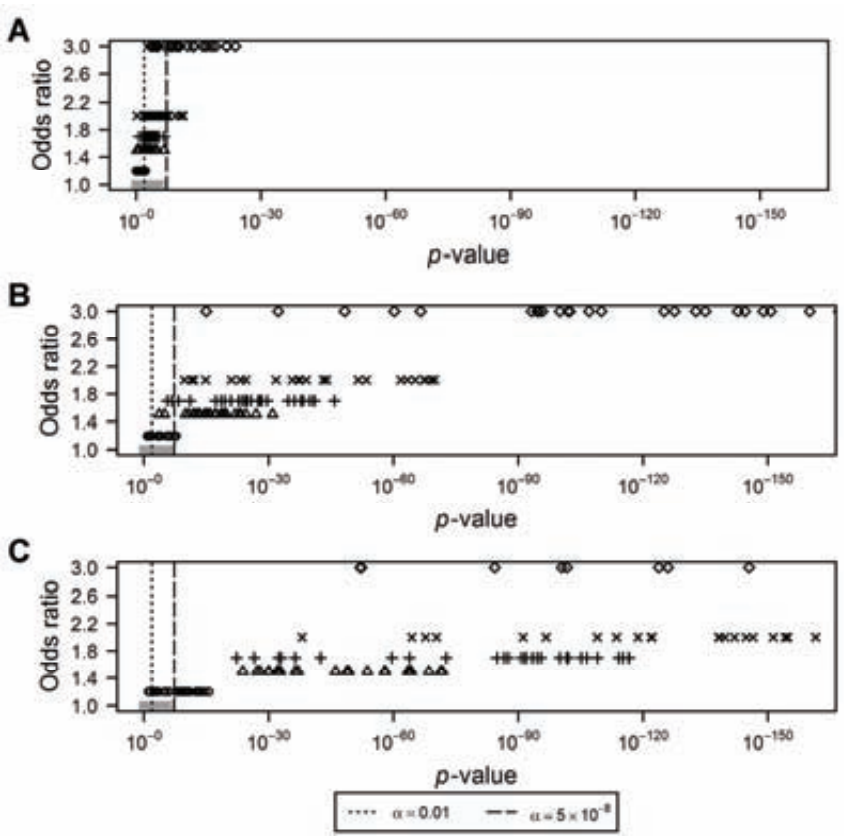

Figure 3.3. *p*-values versus odds ratios for three different sample sizes. Panel A $n_{y=1} = 100$ and $n_{y=0} = 900$, panel B $n_{y=1} = 1{,}000$ and $n_{y=0} = 9{,}000$, and panel C $n_{y=1} = 3{,}000$ and $n_{y=0} = 27{,}000$. For each panel 550,000 SNPs are simulated, including 30 SNPs with odds ratios greater than 1, i.e., 150 SNPs have a known association with $y = 1$. Some SNPs with high odds ratios achieved *p*-values smaller than 10^{-160} and are therefore not shown.

It is obvious that the sample size required to detect true positives can easily become enormous if the effective minor allele has a frequency of less than 0.2 and if the odds ratio is smaller than 1.3. Unfortunately, there is no way of ruling out that most or even all true positives lie in this range of parameters, ex ante. How likely is it that the genetic factors associated with economic behavior such as entrepreneurship will have small odds ratios? Medicine has already discovered many genetic disorders that are complex, multifactorial, or polygenic; disorders likely to be associated with multiple genes in combination with lifestyle and environmental factors. Some examples of such genetically complex diseases identified under GWAS on are listed in Table 3.4. Frequently, weak effects of single loci

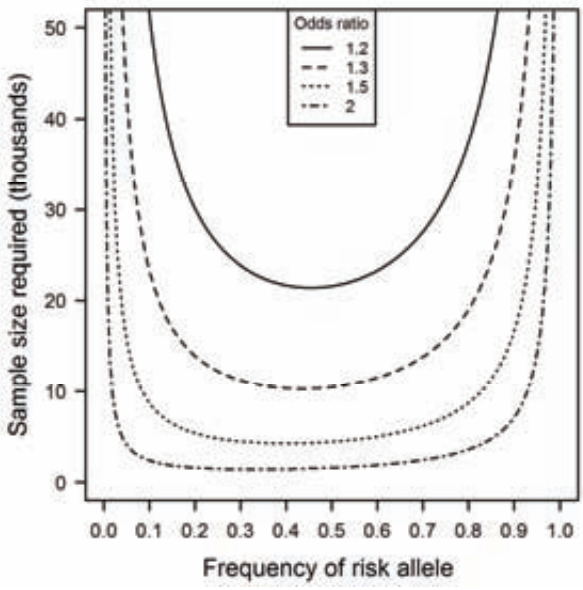

Figure 3.4. Effects of allele frequency and allelic odds ratio on sample size requirements. Numbers shown are for a statistical power of 80% for a χ^2 test within a two-by-two table at a significance level of p-value $< 5 \times 10^{-8}$ for a sample with 10% of $y = 1$.

are found with odds ratios that are in the range of, or smaller than, 1.2 (e.g., Alzheimer's disease, bipolar disorder, breast cancer, lung cancer, multiple sclerosis, and type 2 diabetes). If these genetically complex diseases are any guideline, we should expect that the SNPs associated with entrepreneurship and other complex behaviors will also have weak effects. This demonstrates that very large sample sizes are needed to find small effects. Not only are such large genotyped samples very costly to obtain, but most datasets of genotyped cohorts that are currently available are not nearly large enough for this purpose. At this point in time, the best available solution lies in the meta-analysis of several independent cohorts. In this study design, a consortium of different cohorts is formed that includes genotyped individuals and sufficient information on the outcome of interest such as the presence of a disease or an economic outcome like self-employment or educational attainment. Within the consortium, independent GWAS are performed on each sample, following harmonized standards for the phenotype definition, SNP filtering, and model specification. The results of each GWAS are then meta-analyzed using software such as METAL (Willer, Li, & Abecasis, 2010).

Table 3.4. Sample of results from GWASs on genetically complex traits.

Trait/ Disease	$y = 1$	$y = 0$	Sample Size	Lowest OR[a]	p-value	Highest OR[a]	p-value	Ref.
Alzheimer's disease	5,964	10,188	16,152	1.16	1×10^{-9}	2.53	2×10^{-157}	Harold et al. (2009)
Bipolar disorder	1,868	2,938	4,806	1.03	7×10^{-6}	2.08	6×10^{-8}	Burton et al. (2007)
Breast cancer	27,036	25,253	52,289	1.04	9×10^{-6}	1.26	2×10^{-76}	Easton et al. (2007)
Lung cancer	9,531	9,674	19,205	1.15	8×10^{-9}	1.24	5×10^{-10}	Wang et al. (2008)
Multiple sclerosis	4,839	9,336	14,175	1.10	2×10^{-7}	2.75	4×10^{-225}	De Jager et al. (2009)
Type 2 diabetes	3,836	12,562	16,398	1.15	3×10^{-6}	1.20	8×10^{-9}	Steinthorsdottir et al. (2007)

[a] Refers to the lowest and highest overall odds ratio achieved by combining the discovery and replication samples.

Given the typical sample size of genotyped cohorts used in medical research, often more than five large independent cohorts must be included for a meta-analysis to reach sufficient statistical power. This implies that setting up and managing a consortium requires substantial time investments and a long-term commitment to the research project before publishable findings become available. An additional challenge is that any two GWASs will typically use overlapping but non-identical SNP maps due to the variety of genotyping technologies available. Thus, not every SNP is genotyped in every study, but one still wants to obtain a measure of statistical significance for each individual SNP, taking into account all evidence ("direct" and "indirect") from all studies. The typical strategy in such cases is to impute genotypes for all "missing" SNPs in all cohorts, and carrying out the analysis as if the imputed data were observed. This is possible because the HapMap project provides independent samples of haplotypes[8] that can be used as reference to impute missing alleles in a study using software such as MACH (Li & Abecasis, 2006) or IMPUTE (Marchini, Howie, Myers, McVean, & Donnelly, 2007). The resulting imputed samples often have more than 2 million SNPs, which decreases the power of the analysis even further and in principal requires the use of even stricter confidence levels to avoid false positive according to the Bonferroni correction (4×10^{-9} if 2 million tests are carried out and a significance level of 1% is desired for the entire family of tests). Nevertheless, there are several degrees of correlation (linkage disequilibrium) between markers resulting

[8] A haplotype is the specific combination of alleles at several loci on a single chromosome that are inherited together.

in the use of a significance threshold of 5×10^{-8} which takes into account the number of independent common variants (tests) in the genome.

Hence, the replication and meta-analysis of several samples in one study is often necessary to identify small genetic effects. In addition to addressing the multiple testing problem, the meta-analysis study design has a secondary effect that may be either desirable or undesirable from the point of view of an economist interested in genetic causes of behavior: meta-analysis has a bias for identifying loci that have a similar association with economic behavior in different environments. This is because the cohorts included in a meta-analysis are unlikely to be collected from identical geographic, economic and cultural settings. For example, to conduct a meta-analysis on entrepreneurship it is necessary to include cohorts from various regions since no single homogenous region is likely to have a sufficient number of genotyped individuals available. In the Gentrepreneur Consortium we are currently running (Van der Loos et al., 2010) cohorts from the Netherlands, the United States, Great Britain, Germany, and Iceland are included. This study design tends to suppress alleles that are effective in only one country, but ineffective in any of the other countries. This is desirable if the research objective is to identify genetic factors that are characteristic of entrepreneurs across different economic and cultural environments. However, if the objective is to identify and to compare different genetic determinants of entrepreneurship in different environments, very large samples in each country will be needed that allow the identification of true positives, with or without the application of meta-analysis. At this point of time, this will only be possible for very few countries, if at all.

One final factor aggravating the need for large samples is that one cannot assume that the economic behavior of men and women are triggered by the same SNPs. For example, in an empirical study on the correlates of nascent entrepreneurship Wagner (2007) demonstrates that men and women cannot be pooled in one sample because the estimated coefficients of the unpooled samples differ significantly from each other, in particular with respect to the effect of fear of failure. Grilo, Thurik, Verheul, and Van der Zwan (2007) present similar results using the concept of the entrepreneurial ladder. Arguably, men and women face different demands and constraints when making decisions, whether about education, the formal job market, the amount of time to invest in money earning activities, and whether to choose self-employment or a wage job (Cowling & Taylor, 2001; Verheul, Carree, & Thurik, 2009). If men and women face different environmental circumstances that require different skills and attitudes to cope with to achieve a particular economic outcome, the same genes would not necessarily be linked to this economic outcome. An example is entrepreneurship because both genders may face different motivations and environmental constraints in their occupational choice, which may lead to different genes being associated with entrepreneurship for men and women. As a result, separate GWAS

should be conducted on entrepreneurship for both genders. However, this requires even larger samples to identify SNPs with relatively small effects.

In summary, the required sample size for a GWAS on an economic outcome of interest can easily get very large (> 50,000 observations) if the effective alleles have odds ratios of 1.2 or lower; if the frequency of the effective minor allele is low (< 20%), if the outcome of interest is rare (< 20% of the population), and if there is an interaction between genes and the environment that leads to country-, time-, or gender-specific associations. On the other hand, strong genetic effects with odds ratios of 3 or greater can already be detected with sample sizes of around 1,000 observations. However, given previous findings on genetically complex behaviors, it is unlikely that many economic outcomes will be found that can be clearly linked to a limited number of genes with strong effects. Based on our discussion, we expect sample sizes of at least 30,000 observations will be required to identify SNPs associated with economic outcomes such as entrepreneurship.

3.4 How To Interpret GWAS Results

Given the discussion above, it is obvious that GWAS results below the threshold of genome-wide significance are likely to be false positives. But what does it mean if a particular study does not find SNPs that reach genome-wide significance? Does this mean that genes are not important? Although this is one of the possible causes for not finding genome-wide significance, it is certainly not the only one. Our simulations demonstrate that a plausible alternative explanation is that the study is underpowered: it does not have enough observations to find SNPs with weak effects. In addition, it could be that different genes are associated with an economic outcome in different environments, which will make it difficult to detect the SNPs with meta-analysis using data from different regions or time periods. For example, let us assume that one way genes influence the propensity to become an entrepreneur is via risk preferences. Generally, greater willingness to tolerate risk should increase the probability that an individual will choose self-employment over wage work (Cramer, Hartog, Jonker, & Van Praag, 2002; Kihlstrom & Laffont, 1979). However, the risk of entrepreneurship for the individual may depend on the cultural and economic context. To illustrate, assume that entrepreneurship is less risky in the USA than in Japan because failing with an entrepreneurial business may be more severely punished in Japan than in the USA. While failure may result in severely restricted job offers and lower wages in Japan, it may actually be regarded as positive in the USA, which could lead to better job offers following the entrepreneurial episode of an individual. Consequently, genes influencing risk preferences may be more strongly associated with entrepreneurial behavior in Japan than in the USA. In fact, the relationship may be completely absent or even reversed in the USA, if the above assumptions are true. Hence, a meta-analysis pooling observations

from both countries would tend to "overlook" genes associated with risk preferences although they may be highly relevant in the Japanese context.

Furthermore, there are a number of limitations in current SNP arrays that could be responsible for not finding the true genetic determinants of behavior. For example, rare SNPs (i.e., base pairs that vary in only a very small part of a population) are not covered by current SNP arrays and the most effective SNPs may be found there.

In addition, there are exceptions to the rule that people carry exactly two copies of each SNP, one inherited from each parent (Redon et al., 2006). Instead, there are regions in the genome where people have copy number variants, ranging from zero to 14 copies of a gene. Such copy number variants are not yet recorded in most SNP arrays although they occur frequently throughout the genome and cover hundreds of genes, disease loci, functional elements, and segmental duplications. These copy number variants could not only influence the susceptibility to diseases (Estivill & Armengol, 2007; Freeman et al., 2006), but also the tendency to become an entrepreneur or other economic outcomes.

Another way how inherited changes could influence economic behavior is via epigenetics. Epigenetics refers to the fact that changes in phenotype (appearance) or gene expression can be caused by mechanisms other than changes in the underlying DNA sequence. These changes may persist through cell divisions for the remainder of the cell's life and may also be passed on for numerous generations without any change in the underlying DNA sequence. An example of epigenetic effects is the process of cell differentiation: a single fertilized egg cell changes into the many cell types including bones, muscles, blood vessels, organs, etc. as it continues to divide. It does so by activating some genes while inhibiting others (Reik, 2007). Epigenetic effects occur via several mechanisms, including changes in nutrition. For example, an experiment on agouti mice, which are yellow, fat, and susceptible to life-shortening diseases like cancer and diabetes, found that a change in diet of mother mice could have far-reaching consequences for their offspring (Waterland & Jirtle, 2003). In the experiment, a test group of mother mice were fed a diet rich in methyl donors, small chemical clusters that can attach to, and turn off, a gene. Although these mother mice passed on the agouti gene to their offspring, their children were slender and mousy brown instead of yellow and fat. Furthermore, in addition to living longer, the offspring did not display the susceptibility to cancer and diabetes found in the parents.

Epigenetic effects can also be induced after birth and can be long lasting, passing through several generations without changing the underlying DNA sequence. Medical scientists are still coming to understand the many ways that epigenetic changes unfold. However, there is increasing evidence that genes are not necessarily fate and genetic functions can be switched on and off through nutrition or exposure to environmental risk fac-

tors. Of course, epigenetic effects that cause heritable differences in behavior are not revealed by the GWAS design.

Rare SNPs, copy number variants, and epigenetic effects are just examples illustrating that the GWAS design does not cover all the potential pathways in which traits and economic outcomes can be heritable. In addition, the vast sample sizes required to identify SNPs with small effective could be prohibitively large. Consequently, not finding any genome-wide significant hits in a GWAS does not allow for the conclusion that genes are not relevant for the outcome under investigation.

This is an important insight because it helps reconcile findings from GWAS and from twin studies that appear conflicting at first glance. For example, numerous studies of twins find that intelligence and general cognitive ability are heritable to a significant extent (Deary et al., 2006; Plomin, 1999; Plomin & Kosslyn, 2001; Plomin & Spinath, 2004). Yet, until now GWAS have not been successful at identifying loci associated with IQ scores; even though this is one of the most intensively researched traits in behavioral genetics.

As discussed in Section 3.3, one reason for this discrepancy is that it is very difficult to identify effective SNPs with low odds ratios in GWAS since very large samples are required. However, if intelligence, entrepreneurship and many other human traits and behaviors are genetically complex traits, then it is reasonable to expect that many SNPs with low odds ratios will be found once sufficiently large sample sizes have been gathered. Furthermore, numerous small genetic effects can easily add up in total importance, especially if the effective SNPs interact with each other. This could help to explain the discrepancy between relatively weak effects of SNPs discovered in GWAS and the strong estimated importance of genes often reported in studies of twins. However, it could also be that particular SNPs are only effective in narrowly defined environments. In this case, it is unlikely that even extremely large scale GWAS meta-analyses will discover genome-wide significant SNPs although different genes may be important for the behavior of individuals in their particular environment.

Another reason for the discrepancy between the results in twin studies and GWAS are shortcomings in the design of twin studies that may lead to an overestimation of the importance of genes. First, twin studies require the assumption of a shared, identical environment for twins. A violation of this assumption can lead to an overestimation of genetic effects (Rutter, 2006). A second potential shortcoming of twin studies is the assumption that MZ twins are genetically identical. It was recently found that this is not true as even MZ twin pairs often exhibit different copy number variation profiles (Bruder et al., 2008). Not much is known about how these different profiles can influence behavior and if these differences are more pronounced among MZ or DZ twin pairs. Not accounting for such genetic differences among twins introduces a potential bias into the twin study design. Finally, the structural equation models (SEM) used in twin studies only indicate one possi-

ble explanation of the underlying correlation matrix. They do not allow one to conclude that the model is true or unique because other SEM models may also fit the same data.

Given the methodological difficulties of GWAS and twin studies, it is possible that twin studies tend to give an "upper bound" for the relevance of genes in explaining an outcome of interest, while GWAS give a "lower bound," potentially overlooking many important heritable factors. We emphasize here that GWAS is designed to identify common factors, i.e., those with a population frequency of at least 5%. For the identification of rarer genetic risk factors different technology and study designs are necessary. Furthermore, most, if not all, current GWAS analyses are focused on identifying genetic risk factors with an additive effect (rather than dominant or recessive effects), and on identifying individual genetic risk factors (rather than gene-gene and gene-environment interactions). In other words, effects departing from these assumptions are easily missed and require particular attention and sufficiently powered samples.

The essential question is what does it mean if a GWAS finds genome-wide significance for one or several loci. Does this imply genetic determinism together with environmental irrelevance and lack of free will? Luckily, such interpretation is usually not warranted. Firstly, ongoing epigenetic research has identified various mechanisms that affect how genes, the environment, and behavior can interact, thus leading to long-lasting differences in cell functions. Secondly, so far most research on behavioral genetics and on the genetics of diseases does not find a strict determinism between a particular gene and a specific outcome. In most cases, genes are neither a necessary nor sufficient condition for an outcome to occur. Rather, genes influence the probability that an event occurs, often conditional on non-genetic factors such as the exposition to an environmental risk factor or choice (Rutter, 2006). For example, there are genes associated with smoking (Caporaso et al., 2009) and with lung cancer (Wang et al., 2008). Not everyone with these genes smokes or gets lung cancer, and many people who smoke and get lung cancer do not have these genes. Yet, having these genes significantly increases the probability of smoking and lung cancer. In particular, a genetic predisposition to lung cancer is likely to be amplified by smoking, while an absence of the exposure to smoke is likely to reduce the risk of lung cancer even for those with the genetic predisposition. Similar interactions between genes, the environment, and free will can be expected for economic outcomes such as entrepreneurship. Hence, if GWAS finds effective alleles for entrepreneurship, this will most probably indicate a genetic predisposition to an outcome that will only materialize in the presence of appropriate environmental conditions and conscious choice. It is then necessary to understand the causal pathway of the genetic effect to draw economically valuable conclusions. Is the genetic effect present in different environments? Is it linked to other well-known characteristics of entrepreneurs such as a low degree of risk aversion or overconfidence? Are there differences in the way genes influence the entrepreneurial propensity of

men and women? And how do the biological, psychological, and economic mechanisms work that lead to these associations?

From our point of view, the absence of genome-wide significant results in GWAS on entrepreneurship does not necessarily contradict results from twin studies that suggest a high importance of heritability. Rather, it most likely means that the available sample sizes are not large enough or that the most important factors cannot be found among common SNPs. Increasing sample size and looking at other places such as rare SNPs, copy number variants, and epigenetic effects would be the next step. While the discovery of specific SNPs associated with entrepreneurship would be an important finding, it would not be the end of the quest. The next step would be understanding how the causal pathway from genes to behavior works and how robust these findings are in different environments. Only then could one start drawing conclusions regarding optimal individual behavior and economic policy.

3.5 Conclusion: Genetics in Economics Research?

For economists who believe that a better understanding of economic behavior is an end in itself, the virtues of GWAS and genetics in economic research are evident. Genetics can help us understand the root of individual differences, for example with respect to preferences and productivity values. Also, genetics can help discover new dimensions of individuality that influence economic behavior; those not yet part of established theory. We find this prospect exciting and promising enough to justify this time-consuming and risky endeavor using this approach.

But beyond curiosity as a motivation, are there clear, tangible results economists could expect to get out of this approach? Arguably, the history of science teaches us that the social relevance of many discoveries is not readily and rapidly apparent. Also, most discoveries have little if any social relevance and it is difficult to determine ex ante which research agenda is the most promising. Yet, we dare to speculate about some potential benefits of economists looking at genes here.

First, genetic differences across populations may be discovered that will help to explain aggregate economic outcomes. For example, Global Entrepreneurship Monitor data shows that immigrant countries have a higher share of nascent entrepreneurs than other countries (Ali et al., 2008; Levie, 2007). One potential explanation, from a genetic point of view, are founder effects (a special case of genetic drift): if a small group from a population splinters off and founds a new population in a geographically distant area, the new population is likely to exhibit different shares of alleles at specific loci in the DNA. In this case, the genetic predisposition of the founders will have very strong and long-lasting

effects on the genetic make-up of the new population far into the future that could influence their behavior.[9] If the spin-off population is very small, it will not be possible to represent all genetic variants found in the original population. Furthermore, the spin-off decision could have genetic determinants. For example, if there are genetic predispositions to low risk aversion and novelty seeking, these genes will tend to be overrepresented in immigrant nations that were initially populated by mavericks and explorers. Consequently, there could be a higher average genetic predisposition to entrepreneurial behavior in such countries, independent from the institutional framework conditions or push effects resulting from social marginalization or isolation. This higher level of (genetically induced) entrepreneurial behavior could then have further repercussions in productivity figures, available job offers, wages, and the ability of the economy to adapt quickly to structural changes.

Second, detailed insights into the genetics of economic behavior and its causal pathways may improve our understanding of the scope and potential boundaries for economic policies. For example, a poor fit between genetic predisposition and occupational choice may result in lower monetary income. In addition, not attaining desired social status can affect life expectancy (Rablen & Oswald, 2008) and potentially other non-monetary determinants of utility such as general health. Hence, people may have a genetic predisposition for a particular occupation and there may be a price tag on not finding one's "occupational destiny." Insights along these lines may enable more targeted, maybe even personalized support for people during their educational and work life.

Our conclusion is that GWAS is a promising approach to investigate the genetic causes of economic outcomes. However, as with other genetically complex traits, we expect that very large sample sizes will be needed: in the magnitude of several ten thousand observations, which will lead to a high cost for each true positive discovered. Doubtless the financial and administrative resources necessary to gather the necessary data are beyond the means of most economics departments and research institutes. This, and the rapid progress in the fields of genetics, implies that close cooperation between economics and medical departments will be imperative for finding genetic determinants of economic outcomes.

Finally, it is worth emphasizing that genetics is still a young and rapidly developing research field. GWAS are a significant improvement to earlier approaches in genetics; approaches that have already delivered a wealth of invaluable new insights. However, it is unlikely to be the final word. Rapid scientific and technological progress will enable even better and cheaper insights in the human genome in the future. This will lead to more and

[9] A well-known example for this effect is the high prevalence of people suffering the Ellis-van Creveld syndrome in the North American Amish population, which can be traced back to two members of the new colony started in Pennsylvania in 1744 (Cavalli-Sforza et al., 1996).

better data availability and methodological improvements that can also be used for studying economic outcomes of interest, with entrepreneurship being just one prominent example. Hence our belief that economists can and will learn something useful by looking at genes.

CHAPTER 4

The Molecular Genetic Architecture of Self-Employment

Based on Van der Loos, Rietveld, et al. (2013).

Abstract

Economic variables such as income, education, and occupation are known to affect mortality and morbidity, such as cardiovascular disease, and have also been shown to be partly heritable. However, very little is known about which genes influence economic variables, although these genes may have both a direct and an indirect effect on health. We report results from the first large-scale collaboration that studies the molecular genetic architecture of an economic variable—entrepreneurship—that was operationalized using self-employment, a widely-available proxy. Our results suggest that common SNPs when considered jointly explain about half of the narrow-sense heritability of self-employment estimated in twin data ($h^2 = 54\%$). However, a meta-analysis of genome-wide association studies across sixteen studies comprising 50,627 participants did not identify genome-wide significant SNPs. 58 SNPs with $p < 10^{-5}$ were tested in a replication sample ($n = 3{,}271$), but none replicated. Furthermore, a gene-based test shows that none of the genes that were previously suggested in the literature to influence entrepreneurship reveal significant associations. Finally, a SNP-based genetic score did not significantly predict self-employment out-of-sample. Our results are consistent with a highly polygenic molecular genetic architecture of self-employment, with many genetic variants of small effect. Although self-employment is a multi-faceted, heavily environmentally influenced, and biologically distal trait, our results are similar to those for other genetically complex and biologically more proximate outcomes, such as height, intelligence, personality, and several diseases.

4.1 Introduction

Economic variables such as income, education, and occupation are well-known to be related to health outcomes and longevity (Adler et al., 1994; Adler & Ostrove, 1999; Dowd et al., 2011; Ettner, 1996; Lager & Torssander, 2012; Marmot et al., 1987; Matthews, Kelsey, Meilahn, Kuller, & Wing, 1989; Steenland, Henley, & Thun, 2002; Van Kippersluis, O'Donnell, & Van Doorslaer, 2011; Winkleby et al., 1992). Specifically, there is a consistent inverse relation between indicators of socioeconomic status and cardiovascular disease (Kaplan & Keil, 1993). For example, occupational choice is associated with the incidence of coronary heart disease among women (Haynes & Feinleib, 1980). Intriguingly, health outcomes, longevity, income, educational attainment, and occupational choice have all been shown to be partly heritable (see Manolio et al., 2009, for complex diseases, Herskind et al., 1996; McGue et al., 1993; Mitchell et al., 2001; V.B. Hjelmborg et al., 2006, for longevity, Behrman & Taubman, 1976; Benjamin, Cesarini, Van der Loos, et al., 2012; Lichtenstein, Pedersen, & McClearn, 1992; Miller et al., 2001; Scarr & Weinberg, 1994, for education, Björklund, Jäntti, & Solon, 2007; Sacerdote, 2007; Taubman, 1976, for income, and Nicolaou & Shane, 2010; Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Zhang et al., 2009, for occupational choice). This suggests that the same genetic factors could be linked to socioeconomic status and health outcomes, or that indirect causal pathways from genetic variants to health outcomes exist that are mediated by individual behavior and the environment. For example, a potential mismatch between personal disposition and occupational choice may result in stress and decreased happiness, which have been shown to negatively affect (cardiovascular) disease incidence and longevity (Argyle, 1997; Cooper & Marshall, 1976; Cooper & Smith, 1985; Schnall, Landsbergis, & Baker, 1994). Therefore, knowledge about the specific molecular genetic architecture of socioeconomic variables and about the effects of mismatches between genetic predispositions and realized choices could yield important insights for epidemiology and public health policy. Unfortunately, most efforts to investigate the influence of genes on economic variables were until now limited to candidate gene studies that often failed to replicate later (Beauchamp et al., 2011; Benjamin, Cesarini, Chabris, et al., 2012).

This study reports results from the first large-scale collaboration that studies the molecular genetic architecture of a specific economic behavior—entrepreneurship—using data from high-density SNP arrays. Entrepreneurship has been associated with poor health (Lewin-Epstein & Yuchtman-Yaar, 1991; Rees & Shah, 1986), relatively low average incomes (Hamilton, 2000), and also with greater job and life satisfaction (Benz & Frey, 2008; Blanchflower & Oswald, 1998; Block & Koellinger, 2009). The analysis of entrepreneurship is complicated by the fact that it is a multi-faceted phenomenon (Shane & Venkataraman, 2000). Individuals may engage in entrepreneurial activity for a variety of reasons. For example, certain individuals may be motivated to pursue a business oppor-

tunity or to gain independence, whereas others may do so because of unemployment and a lack of viable alternatives in paid employment. Despite this complexity, empirical evidence suggests that entrepreneurship tends to run in families (Andersson & Hammarstedt, 2010; Colombier & Masclet, 2008; Dunn & Holtz-Eakin, 2000; Evans & Leighton, 1989; Lentz & Laband, 1990; Van der Zwan et al., 2010), and recent twin studies consistently estimate the heritability of this behavior to be on the order of 50% (Nicolaou & Shane, 2010; Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Zhang et al., 2009). As these results suggest that entrepreneurship is partly influenced by genetic variation, specific markers that are associated with entrepreneurship should, in principle, exist. Research that is aimed at discovering these specific markers has thus far been limited to one candidate gene study. Nicolaou et al. (2011) found evidence for an association between a specific genetic variant in the *DRD3* gene and entrepreneurship in a sample of $n = 1{,}335$. However, a more recent study by Van der Loos, Koellinger, et al. (2011) failed to replicate this association in three larger samples of $n = 5{,}374$, $n = 2{,}066$, and $n = 1{,}925$.

The molecular genetic architecture of entrepreneurship therefore remains largely unknown. A variety of alternative architectures could account for heritable variation. For example, there may be a small number of rare variants with strong effects, multiple common variants with small or modest effects, or some combination of these possibilities (Verweij et al., 2012; Visscher, Goddard, Derks, & Wray, 2012). Therefore, we aimed to identify the molecular genetic architecture of entrepreneurship to facilitate a more sophisticated understanding of the nature of the associated heritable variation.

We use self-employment as a proxy for entrepreneurship in this study, which is the most widely available proxy for entrepreneurship. Self-employment is defined as having started, owned, and managed a business. Initially, we used a classical twin design to estimate the heritability of the tendency to engage in self-employment. We performed this analysis to determine the comparability of our results with (1) estimates of previous twin studies, and (2) estimates from a novel method from molecular genetics. This recently described method (Yang et al., 2010) is used here to quantify the proportion of variance that is explained by common SNPs (and unknown causal variants that are in linkage disequilibrium with these SNPs) in the tendency to engage in self-employment.

Furthermore, we performed a meta-analysis of genome-wide association studies (GWASs) of self-employment from sixteen studies to identify genetic variants that are robustly associated with self-employment. Together, these studies comprised 50,627 participants of European ancestry who are part of the Gentrepreneur Consortium (Koellinger et al., 2010; Van der Loos et al., 2010). This study is the first large-scale effort to identify common genetic variants that are associated with an economic variable. We also tested whether self-employment could be predicted out-of-sample solely using genotype data and the results of our meta-analysis.

Theoretical and empirical evidence from entrepreneurship research suggests that there may be differences between males and females with respect to the type of businesses they start. These differences also extend to individuals' motivations, goals, and resources (Bird & Brush, 2002; Du Rietz & Henrekson, 2000; Georgellis & Wall, 2005; Koellinger, Minniti, & Schade, 2013; Verheul, Thurik, Grilo, & Van der Zwan, 2012) and exist because women face different—and typically more—barriers to entrepreneurship than men (Bates, 2002; Riding & Swift, 1990; Verheul & Thurik, 2001). Therefore, we performed both pooled and sex-stratified analyses for all of our investigations.

4.2 Materials and Methods

4.2.1 Participating Studies and Self-Employment Measures

The analyses were performed within the Gentrepreneur Consortium (Koellinger et al., 2010; Van der Loos et al., 2010), which included two out of the five studies that participate in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (Psaty et al., 2009) and fourteen additional studies. The discovery studies included the Age, Gene/Environment Susceptibility–Reykjavik Study (AGES), the Austrian Stroke Prevention Study (ASPS), the Erasmus Rucphen Family study (ERF), the Gutenberg Health Study (GHS), Health 2000 (H2000), the Helsinki Birth Cohort Study (HBCS), the Health and Retirement Study (HRS), the Cooperative Health Research in the Region of Augsburg (KORA S4), the Northern Finland Birth Cohort 1966 (NFBC1966), the Netherlands Twin Register Cohort 1 (NTR1), the Netherlands Twin Register Cohort 2 (NTR2), the Rotterdam Study Baseline (RS-I), the Rotterdam Study Extension of Baseline (RS-II), the Rotterdam Study Young (RS-III), the SardiNIA Study of Aging (SardiNIA), the Study of Health in Pomerania (SHIP), The Hellenic study of Interactions between SNPs & Eating in Atherosclerosis Susceptibility (THISEAS), the UK Adult Twin Registry (TwinsUK), and the Cardiovascular Risk in Young Finns Study (YFS). The Swedish Twin Registry (STR) served as an in silico replication study, as genome-wide data were only available following the completion of the discovery stage.

The studies collected data regarding occupational status using questionnaires or interviews, from which self-employment status was distilled. Self-employment measures were defined in collaboration with the consortium leaders to minimize heterogeneity across participating studies. The cases were defined as individuals who were self-employed at least once, and the controls were defined as individuals who were never self-employed during their working life. However, for a number of studies, reliable data regarding work-life history were unavailable, possibly resulting in the inclusion of previously self-employed individuals in the control group. The details regarding the background and self-employment measures of each of the discovery studies and of the replication study are given in Table A1.

4.2.2 Ethics Statement

All participating studies were approved by the relevant institutional review boards or the local research ethics committees, including the Icelandic National Bioethics Committee (VSN: 00-063), the Icelandic Data Protection Authority, and the Institutional Review Board for the National Institute on Aging (AGES); the Ethics Committee of the Medical Faculty of the University of Graz (ASPS); the Medical Ethics Committee at Erasmus University which approved the protocols for the ascertainment and examination of human subjects (ERF); the local ethics committee and data safety commissioner, the sampling design was approved by the federal data safety commissioner (GHS); the Ethics Committee for Epidemiology and Public Health in the Hospital District of Helsinki and Uusimaa in Finland, in accordance with the ethical standards of the Declaration of Helsinki (H2000); the Ethics Committee of Epidemiology and Public Health of the Hospital District of Helsinki and Uusimaa (HBCS); the Health Sciences Institutional Review Board at the University of Michigan (HRS); the Ethics Committee of the Bavarian Medical Association (KORA S4); the Ethics Committee of the University Hospital of Oulu (NFBC1966); the VU University Medical Ethical Committee (NTR); the Medical Ethics Committee of the Erasmus Medical Center (RS); the local Ethics Committee for the Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche and the MedStar Research Institute, responsible for intramural research at the National Institute of Aging (SardiNIA); the Ethics Committee of the University of Greifswald (SHIP), the Ethical Review Board in Stockholm (STR); the Bioethics Committee of the Harokopio University of Athens (THISEAS); the NRES Committee London-Westminster (TwinsUK); the local Ethics Committees of the participating universities (YFS). Written informed consent was provided by all of the participants.

4.2.3 Genotyping, Imputation, and Quality Control

The seventeen participating studies used a variety of commercially available SNP genotyping platforms to genotype their participants. Each study performed quality control of their genotypic data and imputed the genotypes of each participant to a common set of approximately 2.5 million SNPs from the HapMap CEU population. The exceptions to this were THISEAS, which only supplied results for directly genotyped SNPs, and HRS, which imputed to the 1,000 Genomes Project Phase I v3 panel. Prior to the meta-analysis, we performed parallel quality control of the association results for each study. SNPs were excluded on the basis of minor allele frequency (MAF < 0.01 or MAF < 0.05 if deemed necessary) and if the imputation quality (a measure of the observed variance divided by the expected variance of the imputed allele dosage from the imputation software output) was less than 0.4. Following these exclusions, approximately 2.4 million SNPs remained.

Study-specific details regarding the genotyping, imputation, and quality control are given in Table A2.

4.2.4 Statistical Analysis

Tetrachoric correlations were used to calculate self-employment correlations for monozygotic (MZ) and dizygotic (DZ) twin pairs. This analysis assumes a latent normally distributed tendency to engage in self-employment. We estimated the heritability of the tendency to engage in self-employment in the replication study using standard twin study methods, which were implemented in the program Mx (Neale, Boker, Xie, & Maes, 2003). Only complete twin pairs with data regarding self-employment status were included in the analysis, and opposite-sex DZ twin pairs were excluded. Specifically, for pooled males and females, males only, and females only, we fitted the three following nested models using the maximum likelihood approach on the raw data: (1) a model including an additive genetic effect, a shared common environment effect, and an individual-specific environment effect (the *ACE* model); (2) a model that included only an additive genetic and an individual-specific environment effect (the *AE* model); and (3) a model including only a common environment effect and an individual-specific environment effect (the *CE* model). For all of the samples, we controlled for a *z*-score of age by estimating age-specific thresholds. For the pooled sample, we additionally controlled for sex in a similar way.

We used the method that was recently developed by Yang et al. (2010) to estimate the proportion of variance in the tendency to engage in self-employment that is explained by all of the common genotyped SNPs. The method is implemented in the GCTA software (Yang, Lee, Goddard, & Visscher, 2011) and hinges on the assumption that in a sample of unrelated individuals, environmental factors segregate independently in the pedigree from the degree of genetic relatedness. In contrast to the twin study design, genetic relatedness is not inferred from the pedigree but is estimated directly from genome-wide SNP data. Under the assumption of no confounding by environmental variables, we can then estimate the accounted-for variance by relating the estimated genetic relatedness between pairs of individuals to their phenotypic correlation. The resulting estimate is actually a lower bound of the heritability that is estimated from classic twin and family studies. The reason for this is that twin and family studies capture the variation that is due to all of the additive causal variants, whereas the more recently developed method only captures the variants that are either directly genotyped or in linkage disequilibrium.

We used a combined sample of individuals from one of the discovery studies (RS-I) and the replication study (STR) to estimate the accounted-for variance. We restricted the sample from each study to individuals for whom data regarding self-employment were available. Additionally, we included only one randomly selected individual from each family in the STR sample. A second round of quality control of the genotypic data was

then performed for both studies. In the RS-I sample, we excluded 3,748 SNPs because they failed a test of Hardy-Weinberg equilibrium at $p < 1 \times 10^{-6}$. We removed 24,993 SNPs with minor allele frequencies that were lower than 0.01 and another 6,665 due to data missingness greater than 5%. In total, 5,374 individuals and 561,466 autosomal SNPs were included in the analysis. In the STR sample, we removed two SNPs because they failed a test of Hardy-Weinberg equilibrium at $p < 1 \times 10^{-6}$. Another 628 SNPs with a minor allele frequency lower than 0.01 were removed, as were two SNPs with data missingness greater than 5%. Therefore, 643,924 autosomal SNPs and 2,589 individuals were included in the analysis.

We then estimated the genetic relationships among 7,963 individuals in the combined sample from the 301,115 common autosomal SNPs. We dropped one of any pair of individuals with an estimated genetic relationship that was > 0.025 while maximizing the remaining sample size to exclude the possibility of ascribing shared environmental effects to genetic effects and/or including the effects of causal variants not correlated with the genotyped SNPs but captured by the pedigree. The maximum relatedness in the remaining sample therefore approximately corresponds to cousins two to three times removed (Yang et al., 2010).

Next, the linear mixed model $y = \mu + g + e$ was fitted, where y is the binary phenotype, g the total additive genetic effect of the SNPs, and e is a residual effect. The restricted maximum likelihood (REML) was used to estimate the variance of the total additive genetic effect σ_g^2 of the SNPs by fitting the genetic relationships as the covariance structure. Because the analyzed phenotype is binary, σ_g^2 is the variance of the total additive genetic effects on the observed 0–1 scale. A latent normally distributed tendency to engage in self-employment was assumed when transforming the explained variance from the observed 0–1 scale to the latent scale using the transformation that is derived in the appendix of Dempster and Lerner (1950). For all of the analyses, we controlled for a z-score of age, study, and the first ten principal components of the genetic relationships of the combined sample. In the pooled sample, we also controlled for sex.

The genome-wide association analysis of self-employment was independently performed by each study according to a predefined analysis plan. The analyses were performed for pooled males and females, males only, and females only using an additive genetic model, controlling for age (≤ 29 [reference]; 30–39; 40–49; ≥ 50) and sex in the pooled sample. To control for population stratification, the first four principal components of the genotypic data were also included if available. We provide details regarding the statistical analysis within each study in Table A2.

Following the association analyses, the genomic inflation factor λ was calculated for each sample to quantify any remaining population stratification or cryptic relatedness. The lowest inflation factor was 0.989, and the highest was 1.156, although this latter value was

for a study that did not include the first four principal components of the genotypic data in the analysis (Table A3). Genomic control (Devlin & Roeder, 1999) was applied in samples with inflation factors that were greater than one by adjusting the test statistics.

We next performed fixed-effect meta-analyses of the association results from the discovery studies for pooled males and females, males only, and females only using METAL software (Willer et al., 2010). As the phenotype definitions differed across studies, the effect sizes could not be readily compared. Therefore, we combined the association results using weighted *z*-scores that were based on the *p*-values and the direction of the effects. This method first computes a per-study signed *z*-score for each SNP based on its *p*-value and the effect direction. The *z*-scores are then summed with weights that are proportional to the square root of the sample size of each study. Following the meta-analyses, only autosomal SNPs that were present in the Hapmap Phase II CEU panel (release 22, NCBI build 36) and in at least half of the contributing samples in each meta-analysis were retained prior to both reporting *p*-values and the creation of the Q–Q and Manhattan plots. We a priori set the genome-wide significance threshold to $p < 5 \times 10^{-8}$. SNPs with $p < 1 \times 10^{-5}$ were considered suggestive and also carried forward to the replication stage. The heterogeneity of the test statistics between the studies was assessed using the I^2 metric (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003) and Cochran's *Q* statistic (Cochran, 1954).

Replication was attempted for significant and suggestive SNPs from each meta-analysis using an in silico replication study. The association results for these SNPs were looked up in the replication study and meta-analyzed together with the discovery samples for pooled males and females, males only, and females only.

We used the discovery meta-analyses results to calculate gene-based *p*-values using the VEGAS program (Liu et al., 2010). The positions of the UCSC Genome Browser hg18 assembly were employed to assign SNPs to genes, which included regions that were ±50 kb from the 5' and 3' UTRs.

For the prediction analyses, we followed the approach that was pioneered by The International Schizophrenia Consortium (Purcell et al., 2009) and used the association results from the discovery meta-analyses to predict self-employment in the STR. Specifically, twelve overlapping sets of SNPs that were nominally associated in the discovery meta-analyses were created for different significance thresholds ($p_T < 0.01$, $p_T < 0.05$, $p_T < 0.1$, $p_T < 0.2$, $p_T < 0.3$, $p_T < 0.4$, $p_T < 0.5$, $p_T < 0.6$, $p_T < 0.7$, $p_T < 0.8$, $p_T < 0.9$, and $p_T \leq 1$). These sets were used as inputs for score calculation in the STR.

Table 4.1. Tetrachoric correlations in the tendency to engage in self-employment for MZ and DZ twin pairs in STR for pooled males and females, males only, and females only.

	Pooled		Males		Females	
	MZ	DZ	MZ	DZ	MZ	DZ
n	1,062	1,170	419	469	643	701
Tetrachoric ρ	0.560	0.363	0.677	0.332	0.401	0.230
Standard error	0.042	0.052	0.053	0.072	0.078	0.090

n refers to the number of twin pairs.

Prior to calculating the scores for each individual in the STR, we followed Purcell et al. (2009) and selected all of the autosomal SNPs, pruning those in strong linkage disequilibrium with other SNPs. This process was performed using a pairwise r^2 threshold of 0.25 in a window of 200 SNPs that slides in five SNP increments. Following this exclusion process, 135,856 SNPs remained. The PLINK (Purcell et al., 2007) *score* function was then used to calculate the total score for each individual in the STR. The score is defined as the sum of the number of score alleles, weighted by the estimated coefficients from the discovery meta-analyses, divided by the number of non-missing genotypes. If an individual was missing a genotype, it was imputed as the mean genotype based on the score allele frequency in the STR. On average, the score was calculated from approximately 120,000 SNPs given that (1) the coefficients were only estimated for SNPs in the HapMap CEU population in the discovery meta-analyses, and (2) the overlap with the genotyped SNPs was not perfect. Lastly, we regressed self-employment onto the score using a logistic regression model. The variance that was explained by the score was estimated using the Nagelkerke pseudo-R^2 of the fitted model. We also calculated the area under the receiver operating characteristic curve (AUC) to evaluate the prediction accuracy.

4.3 Results

4.3.1 Heritability of Self-Employment and the Degree of Variance That Is Accounted for by Common SNPs

We used data from the Swedish Twin Registry (STR) and the classical twin design to estimate the heritability of the tendency to engage in self-employment. We computed the tetrachoric correlations between the tendencies to engage in self-employment within MZ and DZ twin pairs. Table 4.1 indicates that the correlations within the MZ twin pairs were consistently higher than within the DZ twin pairs for males only, for females only, and for

Table 4.2. Results of fitting *ACE*, *AE*, and *CE* models to the tendency to engage in self-employment in STR for pooled males and females, males only, and females only.

Model	*A*	(95% CI)	*C*	(95% CI)	*E*	(95% CI)
			Pooled			
ACE	0.54	(0.25, 0.63)	0.01	(0.00, 0.25)	0.45	(0.37, 0.55)
AE	0.55	(0.46, 0.63)	—	—	0.45	(0.37, 0.54)
CE	—	—	0.42	(0.35, 0.49)	0.58	(0.51, 0.65)
			Males			
ACE	0.67	(0.33, 0.76)	0.00	(0.00, 0.28)	0.33	(0.24, 0.44)
AE	0.67	(0.56, 0.76)	—	—	0.33	(0.24, 0.44)
CE	—	—	0.50	(0.41, 0.59)	0.50	(0.41, 0.59)
			Females			
ACE	0.38	(0.00, 0.53)	0.02	(0.00, 0.38)	0.60	(0.47, 0.76)
AE	0.40	(0.26, 0.53)	—	—	0.60	(0.47, 0.75)
CE	—	—	0.31	(0.19, 0.42)	0.69	(0.58, 0.81)

For pooled males and females the analyses are based on 2,232 twin pairs (1,062 MZ and 1,170 DZ), for males only on 888 twin pairs (419 MZ and 469 DZ), and for females only on 1,344 twin pairs (643 MZ and 701 DZ). The share of self-employed was 21% for the pooled, 32% for the male, and 13% for the female sample. In all samples we controlled for age and in the pooled sample for sex; *A*: additive genetic component; *C*: shared common environment component; *E*: individual-specific environment component; 95% CI: 95% confidence interval.

pooled males and females.[10] Applying Falconer's (1960) formula to these correlations yields h^2 estimates of 0.39 for pooled males and females, 0.69 for males only, and 0.34 for females only.

A maximum likelihood approach was employed to estimate the relative contributions of the additive genetic (*A*), shared common environment (*C*), and individual-specific environment (*E*) components. This approach was performed using an *ACE* model and two nested submodels for pooled males and females, males only, and females only. Table 4.2 gives the estimates of the *A* component as 0.54 for pooled males and females, 0.67 for males only, and 0.38 for females only. The estimates of the *C* component were 0.01 for pooled males and females, 0.00 for males only, and 0.02 for females only. The *A* component was significant at the 95% confidence level for pooled males and females, and for males only, although the confidence intervals were very wide. This component was not significant for the females only analysis. When we removed the *C* component from the model, the estimate for the *A* component for females only did not change markedly but was

[10] We note that the correlation within DZ twin pairs in the pooled sample is higher than for the DZ correlations in males and females when the two sexes are considered separately. This effect most likely results from imprecise estimation of the tetrachoric correlations due to the small number of cases. When we computed Pearson correlations, the pooled DZ twin pairs correlation was in between the male and female DZ twin pairs correlations.

Table 4.3. Variance in the tendency to engage in self-employment explained by all autosomal SNPs in a combined sample of RS-I and STR for pooled males and females, males only, and females only.

Sample	σ_g^2/σ_P^2	s.e.	*p*-value	*n*
Pooled	0.25	0.14	0.032	6,223
Males	0.25	0.24	0.152	2,986
Females	0.00	0.28	0.499	3,835

The genetic relationships were estimated from 301,115 directly genotyped autosomal SNPs that were available in both studies. All analyses controlled for age, study, and the first 10 principal components of the genetic similarity matrix of the combined sample of RS-I and STR. In the pooled sample we also controlled for sex. The results did not change markedly when 4 or 20 principal components were included. The share of self-employed in this combined RS-I and STR sample was 14.5% overall, 20.7% for males, and 9.2% for females; σ_g^2/σ_P^2: proportion of phenotypic variance explained by the variance of the total additive genetic effects of the 301,115 autosomal SNPs; s.e.: standard error; *p*-value: *p*-value from a likelihood ratio (LR) test assuming that the LR is distributed as a 50:50 mixture of zero and χ_1^2.

significant at the 95% confidence level. In this submodel, the estimates of the *A* component for pooled males and females, and males only were 0.55 and 0.67, respectively; these results were significant.

The recently developed method by Yang et al. (2010) was employed to estimate the degree of variance in the tendency to engage in self-employment that is explained by all of the genotyped autosomal SNPs in the GWAS datasets. The proportion of the explained variance was estimated for pooled males and females, males only, and females only. To maximize the power of the analysis, we used a combined sample of one of the discovery studies (Rotterdam Study Baseline [RS-I]) and the STR. We estimated that 25% ($p = 0.032$) of the variance in the tendency to engage in self-employment could be explained by the common genotyped autosomal SNPs for pooled males and females (Table 4.3). The variance that could be explained for males only and for females only was 25% ($p = 0.152$) and 0% ($p = 0.499$), respectively.[11] Overall, the results for pooled males and females and for males indicated that the degree of variance in the tendency to engage in self-employment that is explained by all of the common autosomal SNPs simultaneously is only approximately half of the narrow-sense heritability that is estimated using the STR and the classical twin design.[12]

[11] The estimates for males and females separately are not significantly different from one other. The fact that the variance that is explained is zero for females is most likely due to the very low number of female cases ($n = 353$) compared to the number of controls ($n = 3,482$). The estimation of the explained variance is therefore very imprecise.

[12] We also estimated the variance that was explained for pooled males and females, males only, and females only in the RS-I and the STR separately. The estimates were not significant because the standard errors of these esti-

Table 4.4. Descriptive statistics of the sixteen discovery studies and the replication study.

	Pooled		Males		Females		Demographics	
Study	Cases	Controls	Cases	Controls	Cases	Controls	Mean Age	SD Age
AGES	529	2,690	439	913	90	1,777	51.2	6.5
ASPS	46	788	26	336	20	452	65.2	8.1
ERF	214	857	113	366	101	491	47.2	13.4
GHS	424	2,706	282	1,332	142	1,374	55.9	10.9
H2000	228	1,895	145	890	83	1,005	50.7	11.1
HBCS	265	1,459	141	595	124	864	61.5	2.9
HRS	1947	4273	1048	1780	899	2493	63.6	7.9
KORA S4	177	1,588	121	760	56	828	53.8	8.8
NFBC1966	462	3,772	322	1,718	140	2,054	31.0	0.0
NTR1	201	1,354	94	494	107	860	46.4	13.3
NTR2	166	818	77	355	89	463	51.0	13.8
RS-I	531	4,843	319	1,994	212	2,849	68.8	8.8
RS-II	197	1,869	113	848	84	1,021	64.8	8.0
RS-III	209	1,716	138	746	71	970	56.1	5.8
SardiNIA	740	3,402	515	1,207	225	2,195	46.3	17.1
SHIP	157	3,906	107	1,891	50	2,015	49.7	16.3
THISEAS	204	481	176	243	28	238	51.1	11.2
TwinsUK[a]	822	2,333	—	—	730	2,165	54.5	12.4
YFS	215	2,143	89	1,194	126	949	37.6	5.0
Total discovery	7,734	42,893	4,265	17,662	3,377	25,063	53.4	9.4
STR	737	2,534	484	925	253	1,609	60.6	4.3
Total combined	8,471	45,427	4,749	18,587	3,630	26,672	53.8	9.1

AGES: Age, Gene/Environment Susceptibility–Reykjavik Study; ASPS: Austrian Stroke Prevention Study; ERF: Erasmus Rucphen Family study; GHS: Gutenberg Health Study; H2000: Health 2000; HBCS: Helsinki Birth Cohort Study; HRS: Health and Retirement Study; KORA S4: Cooperative Health Research in the Region of Augsburg; NFBC1966: Northern Finland Birth Cohort 1966; NTR1: Netherlands Twin Register Cohort 1; NTR2: Netherlands Twin Register Cohort 2; RS-I: Rotterdam Study Baseline; RS-II: Rotterdam Study Extension of Baseline; RS-III: Rotterdam Study Young; SardiNIA: SardiNIA Study of Aging; SHIP: Study of Health in Pomerania; THISEAS: The Hellenic study of Interactions between SNPs & Eating in Atherosclerosis Susceptibility; TwinsUK: the UK Adult Twin Registry; YFS: the Cardiovascular Risk in Young Finns Study; STR: Swedish Twin Registry; Cases: number of participants that were at least once self-employed; Controls: number of participants that were not, and ideally never, self-employed; SD: standard deviation.

[a] The number of male participants was insufficient for a male stratified analysis.

mates depend heavily on the sample size. However, considered in their entirety, the results were consistent with the estimates that we present for the combined RS-I and STR samples.

4.3.2 Meta-analyses of Genome-Wide Association Studies

We performed genome-wide association analyses of self-employment using the data from sixteen discovery studies. These studies comprised 7,734 participants who had been self-employed at least once and 42,893 participants who did not report being self-employed. Table 4.4 includes the descriptive statistics for the studies. The mean ages in the pooled samples of males and females ranged from 31 to 68.8 years, and the average age across all of the studies was 53.4 years. Following independent association analyses for each study, we performed a fixed-effect meta-analysis of the study-level results for approximately 2.4 million SNPs using a pooled *z*-score approach.

The discovery meta-analysis Q–Q plot (Figure 4.1A) did not indicate a strong deviation for the lowest *p*-values. However, no confounding issues related to population stratification, cryptic relatedness, or genotyping errors were detected, as no systematic deviation from the expectation under the null hypothesis of no association was observed (Pearson & Manolio, 2008). As illustrated in the Manhattan plot (Figure 4.2A), we observed twenty SNPs with $4.1 \times 10^{-6} \leq p < 1 \times 10^{-5}$ (Tables 4.5 and A4). The SNP with the lowest *p*-value, rs6906622 ($p = 4.10 \times 10^{-6}$), was located near the *RNF144B* gene, with most studies indicating that the minor allele increased the probability of being self-employed (Table 4.5).

We next attempted to replicate in silico the twenty suggestive SNPs in the STR. Two of the twenty SNPs associated with self-employment were statistically significant at the 5% level in the replication study. However, the SNP effects were not in the same direction as in the majority of the discovery studies (Table A4), indicating that these SNPs were potential false positives. We then performed a combined meta-analysis of the discovery and replication studies. For all SNPs, the *p*-values were larger in the combined sample than in the discovery sample and did not reach genome-wide significance (Table A4).

The Q–Q plot for the male only meta-analysis (Figure 4.1B) gave a certain degree of suggestive evidence of association; however, no evidence of population stratification, cryptic relatedness, or genotyping errors was observed, as only certain SNPs—those with particularly low *p*-values—deviated from their expectation under the null hypothesis of no association. The female only meta-analysis Q–Q plot (Figure 4.1C) did not indicate a strong deviation for the lowest *p*-values and no evidence of population stratification, cryptic relatedness, or genotyping errors was observed. No SNPs reached genome-wide significance in the sex-stratified meta-analyses (Table 4.5), as can be observed in the Manhattan plots (Figures 4.2B and C). The male meta-analysis resulted in 22 suggestive SNPs with $p < 1 \times 10^{-5}$, and the female meta-analysis resulted in sixteen suggestive SNPs (Tables 4.5, A5, and A6). The top SNP in males, rs6738407 ($p = 1.52 \times 10^{-7}$), was located in the *HECW2* gene, and most studies reported that carrying the minor allele decreased the probability of being self-employed. The top SNP in females, rs2331548 ($p = 1.93 \times 10^{-6}$), was

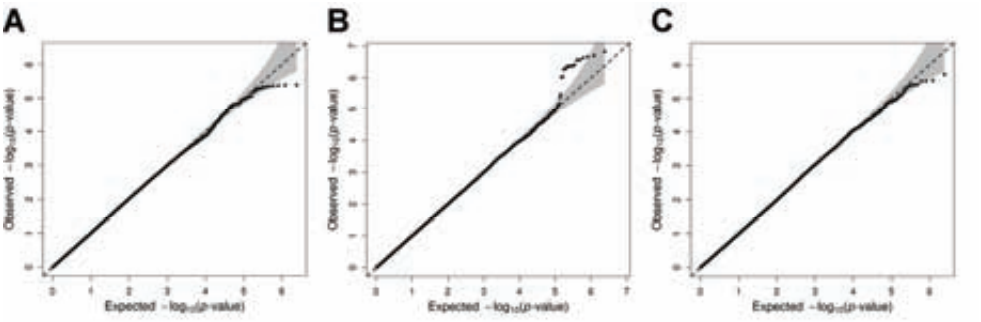

Figure 4.1. Q–Q plots of the self-employment discovery meta-analyses. Q–Q plot of the self-employment discovery meta-analysis for (A) pooled males and females, (B) males only, and (C) females only. The grey shaded areas in the Q–Q plots represent the 95% confidence bands around the *p*-values.

located near the *CBR4* gene, and most studies estimated that carrying the minor allele decreased the probability of being self-employed.

The replication strategy for the 38 suggestive SNPs from the sex-stratified meta-analysis that were carried forward into the replication stage was similar to that used for the meta-analysis replication of the pooled data. We performed an in silico replication study using the data from the STR. None of the SNPs reached nominal significance ($p < 0.05$) in the replication study for males only (Table A5) and females only (Table A6). In addition, for the majority of the suggestive SNPs, the direction of the effect was not consistently in the same direction as was reported in the majority of the discovery studies, again indicating that these SNPs were potential false positives. We meta-analyzed the results from the sex-stratified discovery meta-analysis and the replication study in a combined meta-analysis. For males, five SNPs had lower *p*-values compared to the male discovery meta-analysis, although none reached genome-wide significance (Table A5). In the combined meta-analysis for females, we observed that one SNP, rs562487, had a smaller *p*-value in this combined meta-analysis; however, this SNP did not reach genome-wide significance ($p = 4.01 \times 10^{-6}$; Table A6).

4.3.3 Gene-Based Association Analyses

The findings from the discovery meta-analyses were used to perform gene-based association tests for seventeen genes that have been previously suggested to be candidate genes for entrepreneurship (Nicolaou et al., 2011; Shane, 2010), including *ADORA2A*, *ADRA2A*, *COMT*, *DDC*, *DRD1*, *DRD2*, *DRD3*, *DRD4*, *DRD5*, *DYX1C1*, *HTR1B*, *HTR1E*, *HTR2A*, *KIAA0319* (*DYX2*), *ROBO1*, *SLC6A3* (*DAT1*), and *SNAP25*. Genes with $p < 0.003$ (0.05 / 17 genes) were considered significant, but none of the candidate genes reached this

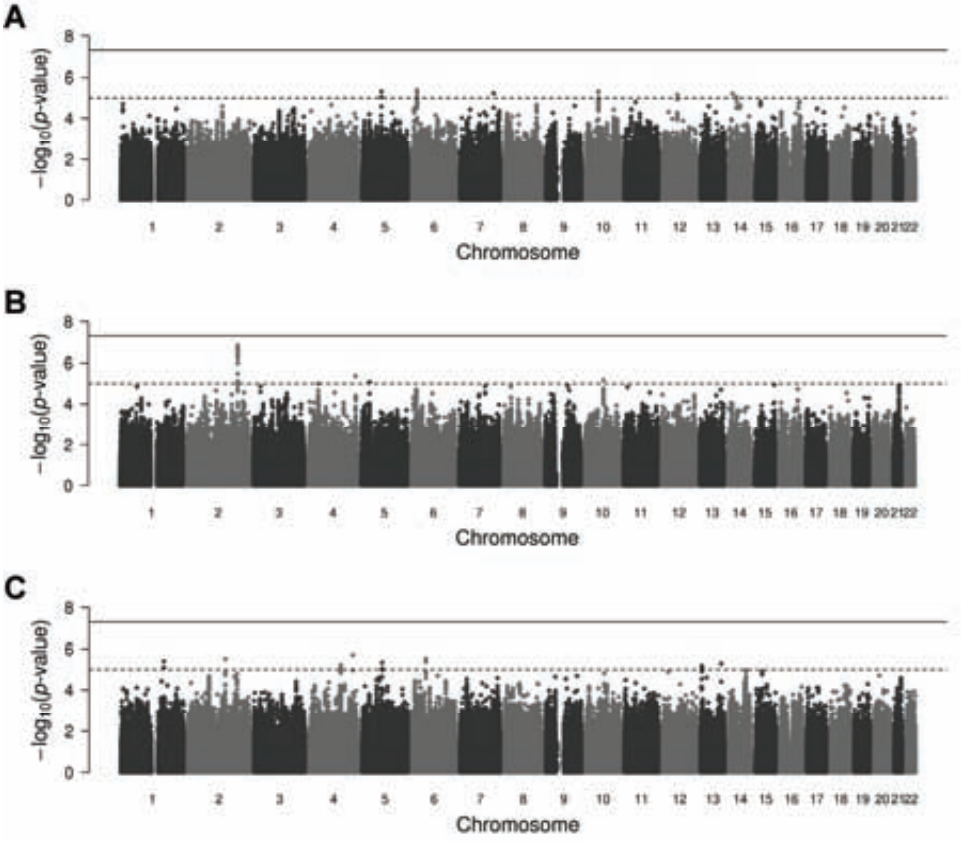

Figure 4.2. Manhattan plots of the self-employment discovery meta-analyses. Manhattan plot of the self-employment discovery meta-analysis for (A) pooled males and females, (B) males only, and (C) females only. SNPs are plotted on the *x*-axis according to their position on each chromosome against association with self-employment on the *y*-axis shown as $-\log_{10}(p\text{-value})$. The solid line indicates the threshold for genome-wide significance ($p < 5 \times 10^{-8}$) and the dashed line the threshold for suggestive SNPs ($p < 1 \times 10^{-5}$).

level (Table A7).

To identify novel genes that may be associated with self-employment, we tested 17,697 genes for pooled males and females, 17,698 genes for males only, and 17,699 genes for females only, implying a significance level of $p < 2.8 \times 10^{-6}$. None of the analyzed genes reached this predetermined significance level (Tables A8, A9, and A10). The gene with the lowest *p*-value was *SLC15A3* for the pooled male and female analysis ($p = 1.63 \times 10^{-4}$). For males only, the lowest *p*-value was for *TMEM156* ($p = 1.61 \times 10^{-4}$), and for females only, the lowest *p*-value was for *PCP4* ($p = 4.70 \times 10^{-5}$).

Table 4.5. Top SNPs ($p < 1 \times 10^{-5}$) from the self-employment discovery meta-analyses for pooled males and females, males only, and females only.

SNP	Chr.	Pos.	Effect / non-effect allele	EAF	*p*-value	Direction	Nearest gene	Number of SNPs in region
					Pooled			
rs6906622	6	18,596,287	T/C	0.21	4.10×10^{-6}	++-++++++++-++++?++	*RNF144B*	12
rs2358531	5	75,515,542	A/G	0.71	4.79×10^{-6}	---?------+--+--?--	*SV2C*	2
rs10776614	10	49,433,172	T/C	0.16	4.79×10^{-6}	-+-----+--------?--	*ARHGAP22*	2
rs17166082	7	131,363,900	A/G	0.06	5.82×10^{-6}	-?-?--?--+------?--	*PLXNA4*	1
rs994208	14	33,531,622	C/G	0.66	6.11×10^{-6}	-+--------------?--	*EGLN3*	1
rs3847697	12	57,282,257	T/C	0.44	6.79×10^{-6}	--+-----?-+-+-?-?--	*LRIG3*	1
rs3742467	14	49,709,284	T/C	0.88	9.11×10^{-6}	+++++-?+-+++-+++?++	*SOS2*	1
					Males			
rs6738407	2	196,851,876	A/G	0.20	1.52×10^{-7}	---------+------?-	*HECW2*	18
rs6825440	4	183,636,063	A/T	0.24	4.25×10^{-6}	-+--------+---+-?-	*ODZ3*	1
rs7904494	10	72,056,694	A/T	0.78	6.74×10^{-6}	+-+---?--++-----?-	*PRF1*	1
rs4867424	5	32,331,331	T/C	0.49	8.39×10^{-6}	--+-------------?-	*MTMR12*	1
rs2712008	4	38,752,396	T/G	0.14	9.94×10^{-6}	+-++++?+++++-+++?+	*KLHL5*	1
					Females			
rs2331548	4	170,199,179	A/G	0.96	1.93×10^{-6}	??+?++++++++++++?++	*CBR4*	1
rs521326	6	52,927,336	A/G	0.61	2.92×10^{-6}	------------+---?--	*GSTA4*	5
rs1022335	2	145,813,253	A/T	0.37	3.02×10^{-6}	------?-----+---?--	*ZEB2*	1
rs10753804	1	168,583,032	T/C	0.49	3.92×10^{-6}	------?----+----?--	*SCYL1BP1*	2
rs562487	5	78,442,190	A/G	0.48	4.49×10^{-6}	+++++-++-+-++-++?++	*BHMT*	2
rs9557259	13	99,031,403	T/C	0.06	5.16×10^{-6}	??-?++?++++++?????+	*TM9SF2*	1
rs1383043	4	123,562,066	A/G	0.38	6.05×10^{-6}	--+----+--------??+	*ADAD1*	2
rs9578700	13	23,775,308	A/G	0.67	6.53×10^{-6}	-+++-----------?-+	*SPATA13*	2

Chr.: chromosome; Pos.: position; EAF: average effect allele frequency; In the column "direction", the studies are in the following order: 1. AGES, 2. ASPS, 3. ERF, 4. GHS, 5. H2000, 6. HBCS, 7. HRS, 8. KORA, 9. NFBC1966, 10. NTR1, 11. NTR2, 12. RS-I, 13. RS-II, 14. RS-III, 15. SardINIA, 16. SHIP, 17. THISEAS, 18. TwinsUK (pooled and female sample) / YFS (male sample), 19. YFS (pooled and female sample); A question mark indicates that the SNP was not tested in that specific study; For SNPs that were located close together in the same region, only the most significant SNP is included in the table. The last column shows the number of neighboring SNPs that exceed the threshold for suggestive SNPs.

We also sought to replicate the association that was reported by Nicolaou et al. (2011) to exist between a common variant, rs1486011, which is located in the *DRD3* gene, and the tendency to be an entrepreneur. The SNP was nominally significant in the discovery meta-analysis ($p = 0.011$; Table A11); however, most studies reported a positive effect of the C allele—opposite to that reported by Nicolaou et al., corroborating the results from an earlier replication study (Van der Loos, Koellinger, et al. (2011). We also sought to replicate this SNP in the sex-stratified discovery meta-analyses. In this analysis, we observed a certain degree of evidence for a positive effect of the C allele in males ($p = 0.046$; Table A11) but not in females ($p = 0.112$; Table A11).

4.3.4 Predicting Self-Employment from Genotype Data

We examined whether the results from the discovery meta-analyses could be used to predict self-employment in the replication study (Purcell et al., 2009). We pruned the set of SNPs to a subset of approximately 120,000 SNPs that are in approximate linkage equilibrium. In an initial prediction analysis, we included only the subset of these 120,000 SNPs that reached a 1% significance level. We calculated a predictive score for each individual in the replication study by determining, for each SNP, the product of the individual's number of effect alleles and the estimated regression coefficient from the discovery meta-analysis. This product was then summed across the included SNPs and divided by the number of included SNPs. We evaluated the predictive power of the SNPs by calculating the degree of variance in the tendency to engage in self-employment that was explained by the score and the area under the receiver operating characteristic curve (AUC). We repeated this prediction analysis eleven additional times, each time with a less stringent significance threshold required for a SNP to be included in the score. Hence, each time this analysis was performed, a larger subset of the 120,000 SNPs was analyzed.

For the pooled analysis of males and females, for males only, and for females only, the results indicated that the score was never nominally associated with self-employment in the replication study, the AUC was under 0.54 for all of the SNP sets, and the variance that was explained by the score was always lower than 0.32% (Table A12). Furthermore, we did not observe a positive relationship between the variance in the tendency to engage in self-employment that was explained by the score and the significance threshold p_T (Figure 4.3).

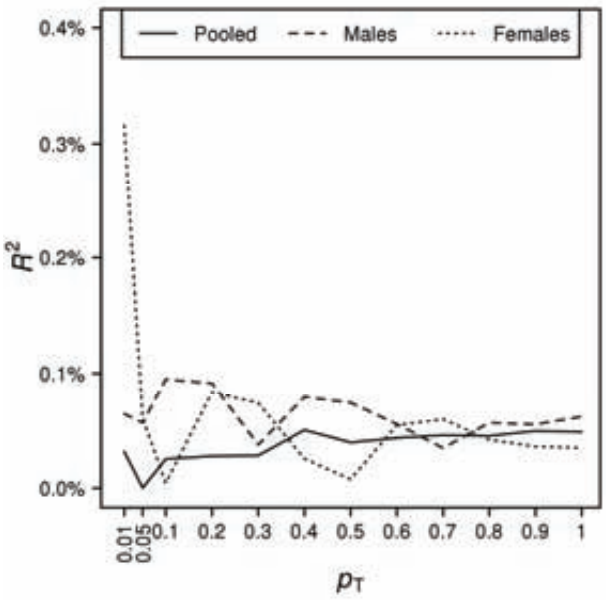

Figure 4.3. Prediction results. Variance explained (Nagelkerke pseudo-R^2 from logistic regression) vs. *p*-value threshold p_T for including SNPs in the score calculation.

4.4 Discussion

We present results from four methods of analysis, three of which are based on genome-wide molecular genetic data, to investigate the molecular genetic architecture of self-employment.

First, using a classical twin design, we report that 54% of the variance in the tendency to engage in self-employment is due to additive genetic effects, with higher heritability for males (67%) than for females (38%). Our estimates are in agreement with those of previous twin studies. These earlier studies suggested heritabilities of 48% in a sample of primarily female British twins (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008) and of 38% in a sample of US twins (Nicolaou & Shane, 2010). In addition, Zhang et al. (2009) estimated the heritability of current business ownership and self-employment in a sample of Swedish twins and observed evidence of a significant additive genetic effect for females but not for males. Our results actually suggest the opposite, i.e., significant heritability among males but not among females; however, the confidence intervals of the estimates are very wide for both our study and for that of Zhang et al. At least a portion of the differ-

ences between these two studies may be explained by imprecision and/or by the different samples and definitions of entrepreneurship that were used.

Second, by applying a method that was recently developed by Yang et al. (2010) to entrepreneurship, we estimate that approximately 25% of the variance in the tendency to engage in self-employment (about half of the h^2 estimated in twin studies) could in principle be explained by the additive effects of common SNPs that are in linkage disequilibrium with the unknown causal variants. These results are in line with previous studies, which have estimated that common SNPs account for one-quarter to half of the narrow-sense heritability for height (Yang et al., 2010), intelligence (Chabris et al., 2012; Davies et al., 2011), personality (Verweij et al., 2010; Vinkhuyzen et al., 2012), several common diseases (Lee, Wray, Goddard, & Visscher, 2011), schizophrenia (Lee et al., 2012), and recently for several economic and political preferences (Benjamin, Cesarini, Van der Loos, et al., 2012).

Several explanations may explain why the heritability estimate for self-employment using common SNPs is approximately half of the estimate that was obtained using the classical twin design. First, the causal variants may be in regions of the genome that are currently not covered by the available SNP arrays. Second, it is possible that the genotyped SNPs and the causal variants are not in complete linkage disequilibrium because, for example, the true causal variants have on average lower minor allele frequencies than the genotyped SNPs.[13] Both of these explanations imply that the estimates that we obtained for self-employment using the more novel method are at the lower bounds of the heritability that is commonly estimated in twin and family studies. A third, alternative, explanation for the different results that were obtained using these techniques is that the twin-based heritability estimates are biased upwards because of, for example, genetic interactions (Zuk, Hechter, Sunyaev, & Lander, 2012) or a violation of the identical common environment assumption in twin studies (Charney, 2008).

Third, we perform the first meta-analysis of GWASs of an economic behavior (i.e., self-employment) using data from sixteen studies that together comprise approximately 50,000 participants. The discovery stage had 80% power to detect a variant at genome-wide significance with a minor allele frequency of 0.25 and odds ratios of approximately 1.11 for pooled males and females, 1.15 for males only, and 1.17 for females only (Purcell et al., 2003), assuming we had a non-noisy, harmonized measure of self-employment

[13] Yang et al. (2010) provide evidence for this in the case of human height. They estimated that 45% of the variance in height is accounted for by common SNPs, while the heritability of height is consistently estimated to be approximately 80%. The authors then developed a method that estimated the variance that was accounted for by common SNPs, assuming imperfect linkage disequilibrium between the genotyped SNPs and the unobserved causal variants. This method revealed that 84% of the variance in height, the complete heritability, could be explained by the causal variants. Twin and family studies do not suffer from this issue, as genetic relatedness is inferred from the expected relationships within the pedigree and include all of the additive genetic variation.

across studies. Yet, we do not identify genome-wide significant associations. This result suggests that there are no common SNPs for self-employment with moderate to large effect sizes, thus placing an upper bound on the effect sizes of common SNPs that we can expect to exist. Gene-based tests for approximately 17,700 genes, including several candidate genes for entrepreneurship that have been previously suggested in the literature (Nicolaou et al., 2011; Shane, 2010), do not reveal significant associations. In addition, we are unable to replicate a previously reported correlation, namely, rs1486011, a SNP that is located in the *DRD3* gene. This common variant was identified by Nicolaou et al. (2011), who reported its association with the tendency to be an entrepreneur. The non-replication of associations is common in candidate gene studies of human traits and behaviors. This failure to identify replicable associations is likely due to a combination of underpowered sample sizes (due to optimistic assumptions regarding plausible effect sizes) and publication bias (Ioannidis, 2005). Examples of non-replication of candidate genes studies on complex human traits include general intelligence (Chabris et al., 2012), personality (De Moor et al., 2012; Ebstein et al., 1996; Lesch et al., 1996; Paterson et al., 1999; Terracciano et al., 2009; Verweij et al., 2010), and trust (Apicella et al., 2010; Israel et al., 2009). We therefore stress that caution is warranted when interpreting claims from candidate gene studies of SNPs or genes with strong effects on complex behavioral traits like self-employment.

Finally, we report that a genetic score that was estimated in our meta-analysis sample has no significant predictive power in our replication study. The variance that was explained by the score was always lower than 0.32%. However, this result does not contradict our finding that approximately half of the narrow-sense heritability can be explained by common SNPs. This latter heritability analysis uses the measured SNPs to estimate realized relatedness between individuals, and given the large number of SNPs in a dense SNP array, realized relatedness can be estimated fairly accurately. In contrast, estimating a strongly predictive score from a sample requires good estimates of the effects of individual SNPs. If our discovery sample was infinitely large, it would have been possible to precisely estimate all of the SNP effects and to obtain a score with the theoretically highest possible predictive power, as estimated using the Yang et al. (2010) method. The smaller the discovery sample, the noisier the estimates of the individual SNP effects; therefore, the predictive power of the score will be lower (Goddard, Wray, Verbyla, & Visscher, 2009; Visscher, Yang, & Goddard, 2010). Our estimates of the effects of the individual SNPs are still too imprecise to allow out-of-sample prediction with SNP data that would have practical utility.

Together, our results demonstrate that common SNPs jointly account for a substantial share of the variance in the tendency to engage in self-employment ($\sigma_g^2/\sigma_P^2 = 25\%$). However, because we do not find specific SNPs in our large-scale meta-analyses of GWASs that examined self-employment, this heritability is not due to SNPs with moderate to large

effects. A plausible interpretation of these results therefore appears to be that the molecular genetic architecture of self-employment is highly polygenic, implying that there are hundreds or thousands of variants that individually have a small effect and which together explain a substantial proportion of the heritability. We cannot rule out the possibility that rare genetic variants, or other, currently unmeasured, variants that are insufficiently correlated with the SNPs on the genotyping platforms, have large effects on an individual's tendency to be self-employed. However, if these genetic variants are rare, they would still not contribute a great deal to the population-based variance in self-employment, and large samples would still be required to identify these variants (Lee et al., 2011; Verweij et al., 2012; Wray, Purcell, & Visscher, 2011).

Our results are similar to those that have been reported for biologically more proximate human traits (Chabris et al., 2012; Davies et al., 2011; Verweij et al., 2012; Vinkhuyzen et al., 2012; Yang et al., 2010) and diseases (Lee et al., 2011, 2012; Purcell et al., 2009) for which a polygenic molecular genetic architecture has also been suggested. One implication of this similarity is that, with sufficiently large sample sizes, SNPs that are associated with self-employment—and possibly also other economic variables—can in principle be discovered, as has been the case for, e.g., height (Lango Allen et al., 2010) and BMI (Speliotes et al., 2010). However, a discovery sample of approximately 50,000 individuals is apparently still too small for a meta-analysis of GWASs on a biologically distal, complex, and relatively rare human behavior such as self-employment.

Given the need for very large samples in meta-analyses of GWASs on complex traits, an important challenge of the present study was to identify a measure of entrepreneurship that is available in a sufficiently large sample. We opted to maximize the available sample size in this study and operationalized entrepreneurship as self-employment, which is also the most frequently used measure of entrepreneurship in the economics literature (Parker, 2009).

We included every study we were aware of in the analysis that included a measure of self-employment and which was willing to contribute data, although this approach necessitated that data from diverse populations (e.g., Eastern German self-employed individuals and US business owners) were pooled. The available measures of self-employment varied across studies, including different single- and multiple-item measures, data from stand-alone surveys, and data from repeated measures or retrospective employment histories of the participants. For a number of studies, this approach resulted in a lack of detailed and reliable data regarding work-life history. Substantial measurement error, especially with respect to the definition of the control group, was therefore unavoidable. Ideally, the control group would encompass only participants who had never been self-employed and who will never be self-employed. Such an analysis would have required data regarding the complete work-life history of participants and participants who had reached an appropriate

age. However, only data regarding current employment status were available in the majority of the contributing studies. It is therefore possible that there was a certain degree of misclassification in the studies that included only single-item, single-response measures of self-employment, thereby adding noise to the phenotype definition and potentially reducing the statistical power with respect to association detection.

Statistical power may have also been reduced by heterogeneity within the case group, as this group comprised individuals who became self-employed for very different reasons. For example, certain individuals may have chosen self-employment because they had no viable alternatives in paid employment, whereas others may have done so because of their desire to pursue a business opportunity. The motivations, goals, and resources of these two groups of individuals are obviously very different, and the genetics underlying these various characteristics may likewise differ greatly. Unfortunately, more detailed information regarding the motivations, activities, and success of entrepreneurs was unavailable for most of the genotyped samples.

In general, GWASs face a practical trade-off between phenotype quality and sample size. Surprisingly, statistical power calculations suggest that studying a more noisy phenotype in a larger sample is often more likely to be successful than studying a perfect phenotype in a small sample. For example, assume that a common SNP exists with a minor allele frequency of 0.5 that increases the odds for all types of entrepreneurship by a factor of 1.13 on average (assuming 15% of the population are entrepreneurs and the data are population samples). The required sample size to detect this SNP with 80% power for a perfectly-measured outcome is approximately 30,000. Measuring entrepreneurship perfectly would require a lengthier survey that is administered more than once. Such a large genotyped sample with perfect measures of entrepreneurship does not currently exist. Smaller samples with perfect measures would be underpowered to detect the SNP. In contrast, if the available measures for entrepreneurship are noisy and have a test-retest reliability of only 0.6—which is typical for behavioral traits measured by brief surveys (Ansolabehere, Rodden, & Snyder, 2008; Loomis, 1989; Weertman, Arntz, Dreessen, Van Velzen, & Vertommen, 2003)—80% power to detect this SNP requires a discovery sample of approximately 50,000 individuals. Thus, our study was well-powered to detect effects of this magnitude even if there was substantial measurement error and noise in the data.

The results of our study have three implications for this future research agenda. First, the high share of variance in self-employment that can be attributed towards interpersonal differences in common SNPs suggests that this research agenda is in principle feasible. Second, to investigate if and how genes that are related to economic variables influence medical outcomes, it will be necessary in the future to identify either the specific genetic variants that are underlying the heritability of economic variables (i.e., to investigate causal pathways from genes to medical outcomes), or to calculate genetic scores that have at

least moderate out-of-sample predictive power (i.e., to investigate the medical consequences of a mismatch between genetic predisposition and economic outcomes). Even larger samples than what we had available in our present study will be needed to identify genome-wide significant SNPs and to estimate more accurate genetic scores for economic variables. Third, our results suggest that the effects of single SNPs on self-employment are likely to be very small. Given these effect sizes, statistical power calculations suggests that a research strategy that aims to maximize sample size by pooling data with slightly inaccurate measures of self-employment is more likely to be successful than a research strategy that aims to collect perfect phenotype measures in a much smaller sample. If successful, this research could shed new light on the complex interaction of genes, environment, and personal choices on health and longevity.

CHAPTER 5

The Molecular Genetics of Serial Self-Employment

Based on Koellinger et al. (2012).

Abstract

We study the genetic architecture of entrepreneurship, as represented by serial self-employment. Using a classical twin study design, we estimate that a significant proportion of variance is attributable to genetic factors. We corroborate positive heritability in males using a recently developed method of estimating heritability from genome-wide data. In an attempt to identify specific genetic variants underlying the heritable variation, we conduct genome-wide association studies in two samples. Pooling the results, we test over two million genetic markers for association and attempt to replicate the findings in a third sample. Furthermore, we test to which extent a genetic risk score calculated from the results of the two discovery samples can predict serial self-employment in the replication sample. We find that none of the genes that have been previously suggested as candidates for entrepreneurship are significantly associated with serial self-employment. We identify a novel genetic variant that replicates, but Bayesian reasoning suggests that a cautious interpretation is warranted. Furthermore, the out-of-sample predictive accuracy of a genetic risk score is virtually zero. The overall pattern of results suggests that the heritable variation in entrepreneurship in our samples is accounted for by a large number of genes with very small effects. We discuss the implications of these findings for research and practice.

5.1 Introduction

A large and growing body of research is focused on estimating how much of the behavioral trait variation across individuals can be statistically accounted for by genetic factors, including several studies that investigate entrepreneurial behavior (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Zhang et al., 2009). Most of these analyses are twin studies, which estimate the heritability of a trait by comparing the correlation of the trait in monozygotic (MZ) twin pairs to the correlation in dizygotic (DZ) twin pairs. Such studies suggest that entrepreneurship and a wide range of other important economic behaviors and outcomes—including income and education (Bowles & Gintis, 2002; Taubman, 1976), investing behavior (Cesarini, Johannesson, Lichtenstein, Sandewall, & Wallace, 2010), overconfidence (Cesarini, Lichtenstein, et al., 2009; Cesarini, Johannesson, Magnusson, & Wallace, 2012), risk taking (Cesarini, Dawes, et al., 2009), and leadership (Chaturvedi, Zyphur, Arvey, Avolio, & Larsson, 2012)—are moderately heritable.

Although twin and family studies can establish that genetic factors account for some of the variation in a trait, they do not identify specific genes or the biological pathways through which genes function. Information about the genetic pathways would be valuable for several reasons. First, such knowledge has the potential to improve our understanding of the causes and consequences of individual differences. Second, genetic variants known to be associated with behavioral traits could be used in (otherwise non-genetic) empirical work as control variables or as measures of otherwise-unobserved traits of interest. For example, empirical studies that attempt to identify factors that make some managers more successful than others may benefit from controlling for the otherwise unobserved entrepreneurial tendencies of managers, which may be predictable from genetic data. Management scholars have also argued that the prediction of entrepreneurial propensity using genetic data could have practical applications in business and for individual decision making (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou & Shane, 2010; Shane, 2010).

Virtually all of the work to date in the social sciences aimed at finding molecular genetic markers uses the candidate gene approach (for reviews, see Ebstein, Israel, Chew, Zhong, & Knafo, 2010, and Benjamin, Cesarini, Van der Loos, et al., 2012). The candidate gene approach consists of selecting a set of markers based on what is known or believed about their biological function. In principle, this approach of formulating and testing ex ante hypotheses is quite reasonable. As an empirical matter, however, the findings from candidate gene studies of complex traits have rarely replicated (Ioannidis, 2005), especially in the social sciences (Beauchamp et al., 2011; Benjamin, Cesarini, Chabris, et al., 2012; Benjamin, Cesarini, Van der Loos, et al., 2012). As a result, very few robust insights into the genetics of complex human behavior are available. Existing work on the genetics of

entrepreneurship is a case in point. Nicolaou et al. (2011) report an association between variation in the *DRD3* gene and entrepreneurial behavior in a sample of 1,335 female twins. A particular genetic variant was reported to be associated with an 80% increase in the odds of being an entrepreneur. Van der Loos, Koellinger, et al. (2011) fail to replicate this association in a sample seven times larger.

Several factors seem to account for the replication problems of candidate gene studies in social science (Beauchamp et al., 2011; Benjamin, Cesarini, Chabris, et al., 2012). First, many candidate gene studies cannot effectively control for a confound known as *population stratification* (i.e., the problem that genetic variation is often correlated with environmental confounders).[14] If the relationship between the confound and the genetic variant is responsible for the association in the original sample and is absent in the replication sample, the association will not replicate. Second, the typical dataset in genetic association work has many behavioral measures and many genetic markers. Hence, false positives arise due to multiple hypothesis testing (Shaffer, 1995) because the *p*-values are usually not adjusted to reflect the model selection and pretesting. In principle, having an ex ante theory to guide the research should reduce the number of hypotheses being tested. Much of the discipline provided by focusing on biologically plausible hypotheses is illusory, however, because there is a huge pool of plausible hypotheses linking genetic variation to complex behaviors. Indeed, 70% of all genes (over 14,000 in total) are expressed in the brain (Ramsköld, Wang, Burge, & Sandberg, 2009), and for many of these, a biologically plausible link to behavior—including entrepreneurship—can be constructed.

Replication failures of candidate gene studies are not only frequent in the social sciences but also in the study of complex traits in the medical sciences (Hirschhorn et al., 2002; Ioannidis, 2005).[15] As a result, candidate gene studies have been largely replaced by genome-wide association studies (GWASs) in medical genetics research. In a GWAS, hundreds of thousands of genetic markers (also called *variants*) spread across the genome are individually tested for association with the outcome of interest without any prior hypothesis. GWASs have led to many scientific and biological discoveries in medical research that consistently replicate in independent samples (Visscher, Brown, et al., 2012). This recent development has been made possible by the dramatic decline in the cost of genotyping.

[14] The problem of population stratification in genetic association studies is nicely illustrated in a thought experiment by Lander and Schork (1994). Consider a genetic discovery study on chopstick use among a population of individuals of European and Asian descent. Any genetic variant that differs in frequency between the two groups will be associated with chopstick use. However, these associations are of course not causal. Stratification can also be a problem in ethnically homogenous samples because the frequency of genetic variants can vary across subgroups within an ethnicity.

[15] An illustrative example is Obeidat et al. (2011), who analyzed 130 genes that had previously been reported to have been associated with lung function and found that only one replicated consistently.

The GWAS design helps to overcome some of the problems of the candidate gene approach. First, it is possible and standard practice in a GWAS to use the whole genome data to identify population structure using principal component analysis of the genetic data (Price et al., 2006). Controlling for the principal components has been shown to quite effectively address concerns about population stratification. Second, the hypothesis-free study design of GWAS makes the need to correct for multiple testing transparent. As a result, stringent *p*-value thresholds have emerged for GWASs in the medical literature (McCarthy et al., 2008). Furthermore, the large number of published GWASs from large samples (Hindorff et al., 2012) have convincingly shown that most traits are influenced by a large number of genes (i.e., they are genetically complex) and that individual genetic markers with an R^2 greater than 0.3% are very rare (Benjamin, Cesarini, Chabris, et al., 2012). These data have helped researchers to adjust their priors about the plausible effect sizes of individual genetic variants. Because of these insights, medical researchers have increasingly recognized the importance of very large datasets that have sufficient power to detect even very small genetic effects at conservative *p*-values. The GWAS approach has enabled an unprecedented surge in genetic discoveries that replicate consistently (Hindorff et al., 2012; Visscher, Brown, et al., 2012).

The availability of whole-genome data has also spurred the development of methods that use such data to answer new research questions. First, to what extent can an outcome be predicted by jointly considering the effects of all of the variants estimated in a GWAS? Researchers typically approach this question by forming a predictor, called the genetic risk score, from a linear combination of genetic effects, estimated in a discovery sample (Purcell et al., 2007). The predictive power of the genetic risk score is then assessed by calculating the correlation between the score and the outcome in a replication sample. The predictive power is increasing in the size of the discovery sample because the coefficients that enter the risk score are estimated with greater precision. Second, what would be the predictive power of a genetic risk score if an infinitely large discovery sample would be available that would allow to estimate the effect of every single genetic variant precisely? Methods have recently been developed for answering this second question, as well (Yang et al., 2010).

In this chapter, we use data from the Gentrepreneur Consortium (Koellinger et al., 2010; Van der Loos et al., 2010) to introduce GWAS as well as these new methods from molecular genetics research to the management literature. The Gentrepreneur Consortium is the earliest large-scale interdisciplinary research collaboration aimed at using molecular genetics data to shed light on an important economic behavior: entrepreneurship. Our results allow us to address several important questions in entrepreneurship research. First, we test if the propensity to engage in entrepreneurship has a genetic component using data from a large sample of comprehensively genotyped individuals. Thus, rather than only

inferring the heritability of entrepreneurship from assumptions about the genetic similarity of people in the same family, we can also estimate heritability from actually observed genetic (dis)similarities among unrelated people. Second, we test if a predisposition to entrepreneurial behavior can be traced back to a small number of well-understood genes with strong effects, as the previous literature has suggested (Nicolaou & Shane, 2009; Nicolaou et al., 2011; Shane, 2010). This will help to clarify the meaning of the term *heritability* in the context of entrepreneurship research. Third, we attempt to identify novel genetic variants that are robustly associated with entrepreneurship. Generally, genes that are robustly associated with entrepreneurship could lead to new insights into the important question what differentiates entrepreneurs and non-entrepreneurs (Shane & Venkataraman, 2000). Fourth, we investigate if it is possible to predict entrepreneurial propensity from genetic data. Previous literature has suggested that companies, including banks or venture capitalists, may (mis)use such information to maximize profits and that genetic information may also guide the vocational choices of individuals (Nicolaou et al., 2011; Shane, 2010). Accurate genetic predictions may also be used to improve empirical research (Benjamin, Cesarini, Chabris, et al., 2012). Thus, answers to this fourth question have practical implications as well as consequences for future research.

Because of the large number of hypotheses tested in a GWAS, conservative *p*-value thresholds must be applied. To attain adequate statistical power given plausible effect sizes for behavioral traits (Benjamin, Cesarini, Chabris, et al., 2012), sample sizes larger than those presently available in any individual dataset are required (Koellinger et al., 2010). To obtain a sufficiently large sample to study entrepreneurship, it is therefore necessary to pool results from several samples. For this reason, it is important to use a proxy for entrepreneurship that is available in multiple genotyped samples.

A general challenge for empirical research on entrepreneurship is that the literature does not agree on a single definition of entrepreneurship (Davidsson, 2005; Koppl, 2007; Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008). This is partly because scholars who are interested in entrepreneurship study different research questions that necessitate different definitions and operationalizations (Parker, 2009). Hence, different empirical proxies for entrepreneurship co-exist, including self-employment, business ownership, starting a business, or commercializing new ideas. Earlier twin studies that investigated the heritability of entrepreneurship used self-employment as a proxy (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Zhang et al., 2009). However, self-employment captures little innovative, growth-oriented, and opportunity-recognizing activities (Koellinger, 2008; Parker, 2009). The only available proxy for entrepreneurship that is measured in several samples with genetic data and that captures some of the innovative, growth-oriented dimension of entrepreneurship is serial self-employment (Hyytinen & Ilmakunnas, 2007; Westhead, Ucbasaran, Wright, & Binks, 2005). An indi-

vidual is said to be serially self-employed if she has experienced at least two episodes of self-employment. Scholars have argued that serial entrepreneurs display a unique mindset (McGrath & MacMillan, 2000). They are more likely to enjoy the excitement of starting a business from scratch, realizing an idea and taking it to the market, and they are more likely to run innovative businesses than people who are self-employed only once (Hyytinen & Ilmakunnas, 2007; Westhead et al., 2005).[16] To the extent that management scholars are interested in the innovative dimension of entrepreneurship, serial self-employment is therefore a better proxy than those previously used to study the heritability of entrepreneurship, i.e., once self-employment or business ownership (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Zhang et al., 2009). For these reasons, we focus on serial self-employment in this chapter.

5.2 Basic Genetic Concepts

The human genome consists of 23 pairs of chromosomes; one of each pair is inherited from the mother, and the other is inherited from the father. Each chromosome is composed of two intertwined strands of DNA, each made up of a sequence of nucleotide molecules. There are four distinct nucleotide molecules, called bases: A (adenine), C (cytosine), T (thymine), and G (guanine). The base A on one strand is always paired with the base T on the other strand, and the base C is always paired with the base G. Because the bases are strictly paired in this way, DNA is conventionally described by writing the sequence of bases for only one strand. For example, at a particular locus (i.e., position on the genome), suppose an individual has inherited the AT base pair from the mother and the GC base pair from the father. At this locus, this person's genotype would be written as AG.

While there are a number of ways that individuals differ from each other genetically, the most common form of genetic variation are single nucleotide polymorphisms (SNPs); SNPs are very frequent and account for approximately 90% of the variation in the human genome (Ziegler & König, 2010). A SNP occurs when individuals differ in which base pair they have at a particular locus. Each of the two possible base pairs is called an allele for that SNP. The allele that is less common in the population is called the minor allele. Individuals who inherited the same allele from each parent are called homozygous for that SNP, while individuals who inherited different alleles are called heterozygous. For each measured SNP, an individual's genotype is coded as a 0, 1 or 2, depending on the number of minor alleles (where 0 and 2 identify the two possible types of homozygotes, and 1 identifies a heterozygote). The most commonly used nomenclature for SNPs is based on

[16] In one of the datasets participating in our study, STR (discussed below), respondents who indicated on a survey that they are self-employed were asked a follow-up question about the type of business they are running. 30% of the serially self-employed answered they have growth and innovation ambitions, compared to only 16% among respondents who were self-employed only once ($n = 2{,}568$, $p_{\chi 2} < 0.001$).

rs-numbers, which provide a unique name for every SNP as assigned by the National Center for Biotechnology Information (USA).

Measuring, or genotyping, SNPs is performed using specialized, array-based technology that allows fast genotyping of hundreds of thousands of SNPs per individual. Current arrays contain approximately 500,000 SNPs, but versions with over 4 million SNPs are already available and this number is expected to continue to increase.

It was recently estimated that the total number of SNPs in humans is approximately 50 million, out of a total of ~3.2 billion base pairs in the human genome (National Center for Biotechnology Information, 2012a). Because SNPs located close to each other are highly correlated with each other, commercial SNP arrays covering only a fraction of all existing SNPs nevertheless "tag" a large part of the genetic variation in a population. The correlation structure of SNPs is now well understood largely due to the availability of reference populations whose entire genomes have been sequenced as part of projects such as HapMap phase II (The International HapMap Consortium, 2005) or 1000 Genomes (The 1000 Genomes Project Consortium, 2010). This information can be used to impute unobserved SNPs with high accuracy. For example, imputed SNPs using the HapMap CEU reference panel typically have an average accuracy of $R^2 \approx 0.95$ in a European population (Huang et al., 2009). Using imputed SNP data facilitates pooling association results from several cohorts that have been genotyped with different SNP arrays. This practice increases the statistical power to detect genetic associations in the combined analysis of several datasets (Halperin & Stephan, 2009; Marchini et al., 2007).

5.3 Data

The three cohorts from the Gentrepreneur Consortium (Koellinger et al., 2010; Van der Loos et al., 2010) that have detailed data on serial self-employment contributed to this study: The Rotterdam Study (RS, see Hofman et al., 1991, 2009), the Erasmus Rucphen Family Study (ERF, see Henneman et al., 2008), and the Swedish Twin Registry (STR, see Lichtenstein et al., 2006). We use the first two studies as our *discovery cohorts*, meaning that they were used for the initial analysis. These analyses were conducted in 2010. The Swedish Twin sample, which was genotyped in 2011, comprises our *replication sample*, meaning that it was used to try to replicate the initial findings.

Each study used a different type of genotyping array, but in every case the directly-genotyped SNPs have been imputed to the HapMap II CEU panel to allow comparison and pooling. After imputation, each sample contains over two million SNPs. Our study uses these imputed data.

All three cohorts consist almost entirely of white European Caucasians. As is standard in GWAS meta-analyses (McCarthy et al., 2008), individuals whose genetic data revealed that they had a different genetic ancestry were removed from the analyses.

We refer to an observation that fits the definition of serial self-employment (i.e., more than one spell of self-employment) as a *case*. We compare the cases to *controls*, defined as individuals without any recorded self-employment spells. Individuals with only one known spell of self-employment are excluded from the analysis.[17] In Table B1, we describe the three cohorts and the available measures in greater detail. Detailed information about genotyping, imputation, and SNP quality control can be found in Table B2.

5.4 Evidence for Heritability

5.4.1 Twin-Based Estimates

As a preliminary, we report traditional twin-based estimates of the heritability of serial self-employment using data from the Swedish sample. In twin studies, heritability is estimated by comparing the resemblance of monozygotic (MZ) twins to dizygotic (DZ) twins. The outcome of interest, or phenotype, P, is modeled as follows:

$$(1) \qquad P = aA + cC + eE,$$

where C is the common environment, A is the sum of additive genetic affects, and E represents unique environmental effects. Under some strong functional form and independence assumptions (Kempthorne, 1997; Plomin, DeFries, McClearn, & McGuffin, 2008), the variation can be partitioned into three components: a^2, c^2, and e^2, where $a^2 + c^2 + e^2 = 1$. Here, a^2 is called heritability and represents the share of variance explained by genetic differences—in other words, a^2 is what the R^2 would be in a regression of the phenotype P on the genetic endowment A, if A were observed (rather than being a latent variable) so that such a regression could be run. The assumptions underlying twin studies are the subject of much research; for a discussion, see Benjamin, Cesarini, Chabris, et al. (2012).

One difficulty that arises when applying the model is that observed serial self-employment is a binary variable. In such cases the convention in the literature is to assume that a continuous latent variable underlies the observed binary variable. The binary variable tells us whether the individual in question is on the right-hand side of some threshold on the underlying liability scale. This threshold is estimated as part of the model (Neale, 2003). The latent variable is assumed to have a standard normal distribution. The heritability estimate is then interpreted as the proportion of variance in the latent variable that is due to genetic differences.

[17] Leaving individuals that were self-employed only once in the control group would make it difficult to interpret the results because the estimated coefficients would not only reflect the difference between being serially self-employed and not being self-employed but would also reflect the difference between being serially self-employed and being self-employed only once.

Table 5.1. Sibling tetrachoric correlations for serial self-employment in STR.

	Pooled		Males		Females	
	MZ	**DZ**	**MZ**	**DZ**	**MZ**	**DZ**
n	818	884	299	313	519	571
Tetrachoric ρ	0.775	0.523	0.792	0.491	0.620	0.203
Standard error	0.049	0.081	0.061	0.104	0.125	0.233

n refers to the number of twin pairs.

Nicolaou, Shane, Cherkas, Hunkin, and Spector (2008) and Nicolaou, Shane, Cherkas, and Spector (2008) used this standard method to estimate the heritability of self-employment with data from the TwinsUK registry, which consists almost entirely of women. They estimated that heritability is approximately 40% and found some evidence that the genetic factors underlying entrepreneurship co-vary with the genetic factors underlying sensation seeking. Zhang et al. (2009) applied the same method to data from the Swedish Twin Registry (STR) to study the heritability of current business ownership and self-employment separately in men and women. Their point estimates suggest that heritability is higher in females than males, but the estimates in these studies are imprecise. We extend this line of research and report heritability estimates on serial self-employment in STR, using new data from the recently administered SALTY questionnaire (Cesarini et al., 2010), which covers a different subset of twins than those studied by Zhang et al. (2009).

All individuals in STR with valid data on serial self-employment from the SALTY questionnaire were included in the analysis. The analyses were run in Mx (Neale et al., 2003). In total 1,636 MZ twins and 1,768 same-sex DZ twins were included.[18] We report the raw MZ and DZ correlations in Table 5.1 and the results from the variance decomposition in Table 5.2. The tetrachoric correlations in Table 5.1 show that serial self-employment is more strongly correlated among MZ twins than among DZ twins, which is consistent with positive heritability. In the *ACE* model in Table 5.2, which includes components for additive genetic effects and common and unique environmental effects, the estimated heritability is 0.60 in males and 0.59 in females. All of these analyses control for an age effect (*z*-score) on the threshold. The pooled estimate, which also corrects for the different prevalence of serial self-employment among males and females, is 0.61. The estimate of heritability is significantly different from zero in the pooled model and in the model for males but not in the model for females. Although our sample is large, our estimates are quite imprecise due to the small number of cases. In fact, the confidence intervals are so large that there are no statistically distinguishable differences between the heritability estimates reported in Table 5.2 and in earlier studies (Nicolaou, Shane, Cherkas,

[18] Singletons and opposite-sex twins were excluded.

Table 5.2. Heritability of serial self-employment using twin study estimates in STR.

Model	a^2	(95% CI)	c^2	(95% CI)	e^2	(95% CI)
			Males			
ACE	0.60	(0.13, 0.87)	0.17	(0.00, 0.57)	0.23	(0.12, 0.37)
AE	0.79	(0.65, 0.88)	—	—	0.21	(0.12, 0.35)
CE	—	—	0.64	(0.51, 0.75)	0.36	(0.25, 0.49)
			Females			
ACE	0.59	(0.00, 0.79)	0	(0.00, 0.61)	0.41	(0.21, 0.69)
AE	0.59	(0.32, 0.79)	—	—	0.41	(0.21, 0.69)
CE	—	—	0.48	(0.23, 0.67)	0.52	(0.33, 0.77)
			Pooled			
ACE	0.61	(0.18, 0.82)	0.11	(0.00, 0.48)	0.28	(0.18, 0.41)
AE	0.73	(0.61, 0.83)	—	—	0.27	(0.17, 0.39)
CE	—	—	0.60	(0.48 – 0.69)	0.40	(0.31, 0.52)

Analyses are based on 612 male twin pairs (299 MZ and 313 DZ) and 1,090 female twin pairs (519 MZ and 571 DZ) where both twins have valid information on serial self-employment. The share of serial self-employed is 14.5% for males, 3.0% for females, and 7.1% pooled in this sample. All three analyses control for age (*z*-score) and the pooled analysis controls in addition for sex.

Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Zhang et al., 2009). What all of these estimates have in common, however, is that they suggest some positive degree of heritability of entrepreneurship, even though measured differently and across different samples.

5.4.2 Evidence from Molecular Genetic Data

We provide further evidence on the heritability of entrepreneurship using a recently-developed method that uses whole-genome SNP array data to estimate a lower bound of the heritability of a trait (Yang et al., 2010). This technique—known as Genomic-Relatedness-Matrix Restricted Maximum Likelihood (GREML)—has recently been applied to study height (Yang et al., 2010), intelligence (Davies et al., 2011), schizophrenia (Lee et al., 2012) and various economic and political preferences (Benjamin, Cesarini, Van der Loos, et al., 2012).

We conduct a GREML analysis of entrepreneurship for two reasons. The first is that twin studies are based on strong assumptions. If these assumptions are violated, the resulting estimates usually will be biased (Kempthorne, 1997). GREML provides estimates of heritability that are based on a different set of assumptions than traditional twin studies, providing complementary evidence. Unlike traditional twin and family studies, GREML uses unrelated individuals and estimates the genetic similarity between two individuals directly from genome-wide SNP data. The key assumption in GREML is that among such

unrelated individuals, environmental factors are independent from the degree of genetic relatedness between people. Under that assumption, an estimate of heritability can be obtained by examining how the correlation in phenotype between pairs of individuals relates to the realized genetic distance between those individuals (see Appendix C for technical details). The resulting estimate is a lower bound for heritability for two reasons: (1) the method assumes that genetic effects are additive (it attributes any interaction effects to the environment), and (2) the estimated relationship between phenotype and genetic relatedness is attenuated because relatedness is measured imperfectly; the common SNPs typed on the genotyping chip capture much but not all of the variation in genetic variation across individuals (Yang et al., 2010).

The second reason we conduct GREML is to estimate how much of the heritable variation is captured by the presently-available molecular-genetic data. This is of interest because much of the excitement regarding the heritability of entrepreneurship derives from the possibility that genetic data may eventually be used to predict it. As explained in Section 5.6 below, a genetic risk score for predicting entrepreneurship can be constructed by summing the additive effects of individual SNPs measured on a genotyping chip. Since GREML is a consistent estimator for the cumulative effects of these SNPs, it generates an estimate of the ultimate predictive power (in terms of R^2) that could be obtained from a genetic risk score if the effects of the SNPs were known (or estimated from a sufficiently large sample).

For our analyses, we pooled the genotyped data from the two largest available samples, RS-I and STR. Pooling the samples increases the statistical power of the analysis considerably, which is especially valuable because serial self-employment is a rare event in the data.[19] We randomly selected one individual per family in STR for inclusion in the estimation. We used the SNP data to estimate the relatedness of every pair of individuals in our sample. As is conventional in the literature, we restricted the analyses to pairs of individuals whose relatedness did not exceed 0.025. The reason for this restriction is that at very low levels of expected relatedness, a larger fraction of the variation in relatedness is random (Hill & Weir, 2011). We estimated the total fraction of variance accounted for by the genotyped SNPs, h^2_{SNPs}. To examine whether the results are driven by population structure, we estimated the first 20 principal components (PCs) of the combined sample (Price et al., 2006). We controlled for varying numbers of PCs and found that the results stabilized when five or more were included as controls. In all models, we included as controls cohort dummies and a z-score for age.

[19] None of the GREML results on heritability is statistically significant in RS-I or STR alone. Further pooling of samples would increase the precision of the heritability estimates. However, the data from ERF is not well suited for this purpose since GREML requires a sample of unrelated individuals, whereas ERF is, due to the study design, a cohort with a very high degree of relatedness among individuals.

Table 5.3. Heritability of serial self-employment using molecular genetic data from RS-I and STR.

Sample	h^2_{SNPs}	s.e.	*p*-value	$n_{RS\text{-}I}$	n_{STR}	n_{total}
Males	0.75	0.44	0.040	1,780	822	2,602
Females	0.00	0.70	0.500	2,360	1232	3,592
Pooled	0.17	0.28	0.272	3,684	1,988	5,672

The genetic relationships in the combined RS-I and STR sample are estimated from 301,115 directly genotyped SNPs that were available in both cohorts, controlling for cohort, sex, age (*z*-score), and the first twenty principal components of the genetic similarity matrix of the pooled sample of RS-I and STR. The results do not change qualitatively if four or ten principal components are included. In STR, one individual from each family was randomly chosen. The share of serial self-employed is 8.15% for males, 2.53% for females, and 4.90% overall in this combined sample; s.e.: standard error; *p*-value: *p*-value from a likelihood ratio (LR) test assuming that the LR is distributed as a 50:50 mixture of zero and χ_1^2.

We estimated separate models by sex, as well as a pooled model in which we controlled for sex. The estimates (Table 5.3) show a high degree of heritability of serial self-employment for males ($h^2_{SNPs} = 75\%$, $p = 0.04$) and zero heritability for females. The pooled sample of men and women together has a non-significant heritability estimate of 17% ($p = 0.27$).

Although imprecise, we interpret the evidence as a whole as supportive of positive heritability for entrepreneurship, especially among males, for whom the twin and SNP-based heritability estimates are both statistically distinguishable from zero with overlapping confidence intervals. Among males, the GREML estimates further suggest that a large share of the observed genetic influence is accounted for by the measured SNPs.

Previous applications of GREML to the study of body height (Yang et al., 2010), intelligence (Davies et al., 2011), schizophrenia (Lee et al., 2012) and preferences (Benjamin, Cesarini, Van der loos, et al., 2012) have shown that ⅓ to ½ of the heritability estimates in traditional twin studies can be explained by the SNPs measured on existing platforms. This gap may imply that the additive effects of the measured SNPs only account for ⅓ to ½ of the genetic variation, that the twin-based estimates of heritability are biased upward, that non-additive effects of SNPs exist, or that combinations of these reasons are true. Earlier studies either had larger sample sizes and a larger fraction of cases ($n_{schizophrenia} = 21{,}258$; $n_{cases,\ schizophrenia} = 9{,}087$) than the present study, or focused on outcomes that are measured on a continuous scale (height, intelligence, preferences), which generally increases the observed variance in the outcomes of interest. These sample characteristics greatly improved the precision of the GREML estimates in these studies (e.g., $h^2_{SNPs,\ schizophrenia}$ is estimated with s.e. = 1%). In light of the previous evidence, it is likely that the null result for females in our study is due to a lack of statistical power because of the rare occurrence of serial self-employment among females ($n_{cases,\ serial\ self\text{-}employed,\ female} = 144$). Furthermore, it

is likely that the very high point estimate for males in our study would be smaller in a larger sample.

5.5 Genome-Wide Association Study (GWAS)

To investigate which specific SNPs are associated with serial self-employment, we carried out a genome-wide association study (GWAS; see Koellinger et al., 2010 and Beauchamp et al., 2011) using the two discovery cohorts (ERF and RS-I) and then meta-analyzed the cohort-specific results. After this discovery stage was complete in early 2011, we sought to replicate the most promising associations in the Swedish Twin Registry (STR) sample, which only became available in July 2011. Because our examination of existing proposed candidate genes (Nicholaou et al., 2011; Shane, 2010) is based on the GWAS results, we discuss the results from the GWAS first.

Any GWAS must confront the three major challenges that we outline below. In Appendix D, we provide additional details on the data and methods we used to address these challenges.

5.5.1 Multiple Hypothesis Testing

The very large number of statistical tests that are carried out in a GWAS leads to a severe multiple-hypothesis-testing problem. In our sample, there are over two million imputed SNPs. Because SNPs are locally correlated (in so-called linkage disequilibrium), testing each individual variant for association with some outcome has been shown to be approximately equivalent to testing one million independent hypotheses (Hoggart et al., 2008; McCarthy et al., 2008; The International HapMap Consortium, 2005). Stringent significance levels are used in GWASs to maintain the false positive rate at an acceptably low level. A Bonferroni correction for one million independent tests suggests that a significance level of 5×10^{-8} is necessary to obtain a family-wide significance level of 5%. This level is often referred to as genome-wide significance (McCarthy et al., 2008). We use this significance threshold in the present study. In addition, and in line with standard practice in medical genetics, SNPs with a p-value of $10^{-5} > p > 5 \times 10^{-8}$ are categorized as suggestive hits that enter the replication stage along with the genome-wide significant hits. In the replication stage, it is customary to apply a Bonferroni correction based on the number of independent hits that were tested for replication. For example, if ten independent suggestive hits enter the replication stage, the corrected significance for a family-wide significance level of 5% would be $p = 0.005$. However, for SNPs that reached genome-wide significance in the discovery stage, a nominal significance at the 5% level is typically considered to be sufficient for reporting a positive replication, under the condition that the overall p-value of the meta-analysis improves when the replication cohort is included.

5.5.2 Population Stratification

The standard approach in GWAS for dealing with the problem of population stratification is threefold. First, restrict attention to individuals with a relatively homogenous ethnic background. Our samples satisfy this requirement (see Section 5.3). Second, control for any remaining population substructure by including principal components (PCs) of the genome-wide data as controls in the regressions of the phenotype on the individual SNPs (Price et al., 2006). In the RS-I and the STR data, we follow the standard practice in medical genetics of controlling for the first four PCs. In the ERF data, controlling for PCs is not appropriate because the sample consists essentially of only one large family and population stratification is therefore not an issue. Third, apply genomic control (Devlin & Roeder, 1999) to the results to correct for remaining population stratification in the cohorts. This is a simple, conservative, linear adjustment of the estimated *p*-values. The appropriate adjustment factor, called λ, is estimated for each cohort and analysis separately.[20]

5.5.3 Quality Control for Genetic Data

To avoid spurious findings it is standard in molecular genetics research to apply strict quality controls to the genotypic data. We closely followed these conventions. Markers that did not meet the standard quality criteria were not used for imputation. We also omitted SNPs that did not satisfy at least one of the following criteria. First, imputed SNPs had to be known to have reasonable accuracy. Second, the SNP had to have a minor allele frequency above 0.01. The reason is that genotyping errors are more common in SNPs with lower minor allele frequencies. Finally, to be included in the meta-analysis, the SNP had to be available in both RS-I and ERF.

5.5.4 Model Specification and Meta-analysis

In a GWAS, the trait of interest is tested for association with one genetic marker at a time. The two discovery cohorts ran logistic regressions of serial self-employment on each individual SNP available in the imputed data after quality control. The estimated model is as follows:

$$(2) \qquad P(y_i = 1 \mid \mathbf{x}_i) = \exp(\boldsymbol{\beta}'\mathbf{x}_i) / (1 + \exp(\boldsymbol{\beta}'\mathbf{x}_i)),$$

[20] We report the λ for each cohort and analyses in the notes to Tables 5.5, D1, and D2. A correction factor of $\lambda \leq 1$ is typically interpreted as evidence against population stratification in the sample, and the estimated *p*-values of the coefficients are not adjusted in this case. Otherwise, the *p*-values are linearly adjusted upward (Devlin & Roeder, 1999).

Table 5.4. Sample sizes for genome-wide association studies.

	Males		Females		Pooled	
	n	Cases	*n*	Cases	*n*	Cases
ERF	400	8.5%	526	6.7%	926	7.5%
RS-I	2,082	4.2%	2,922	2.5%	5,004	3.2%
STR	1,111	16.7%	1,660	3.1%	2,771	8.6%
Total	3,593	8.5%	5,108	3.1%	8,701	5.4%

where y_i is a dummy for serial self-employment for individual i, $\mathbf{x}_i$ is a vector of regressors, and $\boldsymbol{\beta}$ is the vector of coefficients. The primary regressor of interest, $x_{i,1}$, is the number of minor alleles at this particular locus. If the genotype was not imputed, then this value is always an integer equal to 0, 1, or 2. If the genotype was imputed at this locus, then the variable is equal to the imputed (expected) number of minor alleles. We also control for a set of age dummies ($x_{i,2,\ldots,5}$ = age in four categories) and four principal components of the genotypic data in RS-I and STR ($x_{i,6,\ldots,9}$). In the pooled analysis, we also control for sex ($x_{i,10}$). Details about the statistical analysis within each study can be found in Table B2.

We performed meta-analysis of the cohort-specific results for equation 2 using the METAL software (Willer et al., 2010). The software first selects a reference allele for each marker and calculates a weighted z-score characterizing the evidence for association across studies, with weights proportional to the square-root of the sample size of each study. The overall p-value of each allele is given by $p = 2\Phi(|-z|)$. Extreme negative z-scores yield a small p-value and indicate an allele associated with lower entrepreneurial propensity, whereas extreme positive z-scores yield a small p-value and indicate a positive association with entrepreneurial propensity.

5.5.5 GWAS Results

Table 5.4 provides descriptive statistics for our sample. Overall, 5,930 observations are available in the discovery stage ($n = 3{,}448$ for females and $n = 2{,}482$ for males). In the replication stage, 2,771 additional observations from STR are available ($n = 1{,}660$ for females and $n = 1{,}111$ for males). Overall, the cases comprise 5.4% of the sample. None of the results from the gender-stratified GWAS analyses reached genome-wide significance (see Appendix D). Table 5.5 displays the top SNPs from the pooled analysis. The listed SNPs belong to different loci (i.e., are distant from each other on the genome) and are approximately uncorrelated. For each set of correlated SNPs, we include only the one with the lowest p-value in the discovery stage.

Table 5.5. Top SNPs, females and males pooled.

Discovery meta-analysis					STR		Combined meta-analysis		
SNP	Chr.	Avg. Freq.	*p*-value	Nearest gene	*p*-value	Freq.	*p*-value	Direc-tion	Im-prove ment
rs3774790	3	0.86	3.56×10^{-7}	*ABHD5*	6.56×10^{-5}	0.84	1.08×10^{-10}	−−−	yes
rs4748739	10	0.32	9.33×10^{-6}	*NEBL*	0.11	0.29	5.22×10^{-6}	+++	yes
rs17775594	14	0.88	6.50×10^{-6}	*C14orf101*	0.22	0.89	9.91×10^{-6}	−−−	no
rs4479388	2	0.95	1.75×10^{-6}	*KCNJ3*	0.51	0.93	3.52×10^{-4}	−−+	no
rs9514109	13	0.23	6.13×10^{-6}	*SLC10A2*	0.71	0.24	4.26×10^{-4}	−−+	no
rs4776126	15	0.28	9.08×10^{-7}	*WDR72*	0.09	0.28	2.00×10^{-3}	++−	no

Chr.: chromosome; Single genomic control used (Devlin & Roeder, 1999); The applied imputation accuracy thresholds for including SNPs are info > 0.40 (IMPUTE) and Rsq > 0.40 (MACH); In the column "direction", the studies are in the following order: 1. ERF ($\lambda = 1.108$), 2. RS-I ($\lambda = 1.001$), 3. STR ($\lambda = 0.988$), where λ is the genomic control parameter; A "?" indicates that this SNP was not available in the respective study; Only the SNP with the lowest *p*-value of each identified locus is listed here; SNPs are ordered by *p*-value in the combined meta-analysis.

Six suggestive loci with $10^{-5} > p > 5 \times 10^{-8}$ entered the replication stage in the pooled analysis. One that is located on the *ABHD5* gene (rs3774790) replicates in STR ($p = 3.56 \times 10^{-7}$ in the discovery cohorts and $p = 6.56 \times 10^{-5}$ in STR), yielding a genome-wide significant combined *p*-value of 1.08×10^{-10}. Our top hit rs3774790 has an odds ratio of 1.89 in the discovery meta-analysis and an odds ratio of 1.042 in the replication cohort. *ABHD5* codes a protein that is involved in the storage of fats in the body (National Center for Biotechnology Information, 2012b), and defects in *ABHD5* have been linked to a rare medical condition called the Chanarin-Dorfman syndrome (Emre et al., 2010). Nothing about this gene's known function would suggest a straightforward link to entrepreneurial behavior.

We interpret this finding very cautiously despite the fact that it reaches genome-wide significance because statistically significant findings obtained from small samples often fail to replicate (Ioannidis, 2005). A heuristic Bayesian calculation along the lines of Benjamin, Cesarini, Chabris, et al. (2012) helps illustrate why our caution is warranted. Suppose it is known that there is either no association (an odds ratio of 1) or there is an association with an odds ratio of 1.042. This alternative hypothesis of 1.042 is the effect size in the replication sample, which is an unbiased estimate of the true effect size. (In contrast, the effect size in the discovery sample is inflated due to the well-known *winner's curse* problem; see Garner, 2007.) Using Bayes' Rule, the probability that a genome-wide significant SNP in our sample with an odds ratio of 1.042 is truly associated with entrepreneurship is Pr(*true* | *significant*) = Pr(*significant* | *true*) Pr(*true*) / [Pr(*significant* | *true*) Pr(*true*) + Pr(*significant* | *false*) Pr(1 − *true*)]. For Pr(*significant* | *false*), we can use the threshold *p*-value of genome-wide significance, 5×10^{-8}. For Pr(*significant* | *true*), we can use the

power to detect a statistically significant effect at the genome-wide significance level of 5×10^{-8} in a combined sample size of 8,701 given a true odds ratio of 1.042, which is equal to 1.07×10^{-7} (using the genetic power calculator by Purcell et al., 2003). The posterior probability that the finding is true, Pr(*true* | *significant*), only exceeds 50% if the prior probability Pr(*true*) exceeds 32%. Given that there was no reason to expect that this particular SNP is associated with entrepreneurship, such a high prior probability is difficult to justify.

A much more realistic prior is based on the assumption that entrepreneurship is a genetically highly complex trait, with 10,000 loci that are independently associated with it. Because there are approximately 1 million independent loci (also called *haplotypes*) in the human genome (The International HapMap Consortium, 2005), the chance of randomly selecting one of them is 1%. Using this prior, the posterior probability that our finding is true is only 2.1%. This reasoning suggests we should infer little from this evidence of a statistically significant association.

5.5.6 Test of Candidate Genes

Nicolaou et al. (2011) reported an association between a variant in the *DRD3* gene (rs1486011) and entrepreneurship. Van der Loos, Koellinger, et al. (2011) subsequently failed to replicate the association using self-employment as a proxy for entrepreneurship in the RS-I data, one of our discovery cohorts. Here, we examine the association for serial self-employment in all three cohorts and find that the association also fails to replicate in the RS-I, ERF, and STR samples, and in the pooled results. The point estimate of the effect is insignificant and has the opposite sign to that reported by Nicolaou et al.[21]

Next, we sought to systematically investigate a number of candidate genes that have been proposed in the literature (Nicolaou et al., 2011; Shane, 2010), but for which there are no published significant findings. No SNP in or near any of these genes reached our threshold for suggestive significance in our GWAS. We had 80% statistical power to detect common SNPs with a minor allele frequency of > 0.2 at $p < 5 \times 10^{-8}$ if the SNP had an odds ratio > 1.55 (Purcell et al., 2003); the odds ratio for the SNP association reported by Nicolaou et al. (2011) was 1.8. Thus, if the result of Nicolaou et al. (2011) would be true or if SNPs with similarly large effects would exist, we were well-powered to detect it in our data. Our results show that this is not the case.

[21] The *p*-value of rs1486011 in the combined sample of RS-I, ERF, and STR is 0.353 for males, 0.005 for females, and 0.413 in the pooled sample.

Table 5.6. Gene-based *p*-values for 17 candidate entrepreneurship genes for males and females pooled, males only, and females only.

Gene	Pooled	Males	Females
ADORA2A	0.730	0.514	0.773
ADRA2A	0.871	0.478	0.113
COMT	0.095	0.697	0.063
DDC	0.345	0.372	0.806
DRD1	0.908	0.952	0.483
DRD2	0.034	0.390	0.144
DRD3	0.171	0.155	0.160
DRD4	0.668	0.256	0.739
DRD5	0.606	0.627	0.164
DYX1C1	0.663	0.531	0.756
HTR1B	0.812	0.946	0.246
HTR1E	0.273	0.041	0.397
HTR2A	0.745	0.362	0.491
KIAA0319 (*DYX2*)	0.980	0.506	0.073
ROBO1	0.885	0.592	0.762
SLC6A3 (*DAT1*)	0.792	0.664	0.801
SNAP25	0.648	0.944	0.611

We additionally tested the 17 candidate genes proposed by Shane (2010) for association with entrepreneurship using a gene-based test of association implemented by the VEGAS software (Liu et al., 2010). This gene-based test uses information from all the measured SNPs in a gene and may reject the null hypothesis of no association even if no individual SNP reaches the significance threshold. We conduct this test for males and females separately and in the pooled sample, for a total of 51 tests. Table 5.6 reports the results of the gene-based test. Under the assumption that the 51 tests are independent, we should have expected 5.1 significant associations at the 10% level and 2.55 significant associations at the 5% level. Overall, we observe five associations in Table 5.6 at a nominal significance level of 10%, and two of these are also significant at a nominal significance level of 5%. This is even less than what one would expect to observe under the null hypothesis of no association.

For completeness, we also conducted the gene-based test on the nearly 18,000 genes in our sample. Again, we conducted this analysis in the pooled sample, as well as the female and male samples separately. No gene reaches statistical significance at the Bonferroni adjusted significance level of $p < 2.81 \times 10^{-6}$.

5.6 Genetic Risk Prediction

Our GREML estimates suggest that the measured SNPs taken together have substantial predictive power, at least in men. Here, we test to what extent a genetic risk score constructed using a linear combination of the SNP effects estimated by the GWAS can be used to predict entrepreneurship out of sample. In the limit of an infinite sample, the out-of-sample R^2 from a regression of serial self-employment on such a genetic risk score should be the same as the GREML estimate of the proportion of variance captured by measured SNPs. However, the smaller the sample used to construct the genetic risk score, the less precisely estimated is the effect of each individual SNP. This will introduce measurement error in the genetic risk score that will attenuate the predictive power of the score. In this section, we construct a genetic risk score using the results of the discovery stage of the GWAS meta-analysis. We then ask how predictive this genetic risk score is of serial self-employment in the Swedish sample.

The standard method for constructing the genetic risk score was developed by Purcell et al. (2009). To avoid double counting SNPs, we use the software package PLINK (Purcell et al., 2007) to select a pruned set of SNPs that are approximately uncorrelated. An individual's genetic risk score at the *p*-value threshold τ is then defined by,

$$(3) \qquad g_{i,\tau} = \Sigma_j \, x_{ij} \, \hat{\beta}_j \, 1(\tau),$$

where j is a pruned SNP, $j = 1,\ldots, J$, $\hat{\beta}_j$ is the regression coefficient of SNP j estimated from the GWAS, x_{ij} is the number of reference alleles for individual i at SNP j, and $1(\tau)$ is an indicator variable taking the value 1 if the *p*-value of $\hat{\beta}_j$ is less than τ and 0 otherwise. We construct the scores using the thresholds 0.01, 0.05, 0.1, 0.2,…, 0.9, 1.0.

This method has been used to generate genetic risk scores with non-negligible out-of-sample predictive power in applications to schizophrenia (Lee et al., 2012) and height (Lango Allen et al., 2010), with sample sizes of 54,171 and 183,727, respectively. For example, a regression of height on a genetic risk score using an independent sample yields an R^2 of a little over 10% (Lango Allen et al., 2010).

In the context of entrepreneurship, we evaluate the predictive power of the SNPs by running univariate logit regressions of serial self-employment on the genetic risk score in the STR. We run the regressions separately for men and women and for the pooled sample. A random selection of one person per family was used in the prediction sample (804 males and 1212 females, with 138 and 41 cases, respectively).[22]

[22] We only include one family member in STR because we cannot control for family structure in the prediction equation.

Table 5.7. Results of the prediction analyses in STR for males and females pooled, males only, and females only.

	Pooled					Males					Females				
Threshold	***n* SNPs**	**Coeff.**	***p*-value**	**AUC**	**R^2 (%)**	***n* SNPs**	**Coeff.**	***p*-value**	**AUC**	**R^2 (%)**	***n* SNPs**	**Coeff.**	***p*-value**	**AUC**	**R^2 (%)**
$p_T < 0.01$	994	0.207	0.177	0.520	0.001	997	−0.156	0.839	0.531	0.002	1,023	−0.318	0.929	0.565	0.007
$p_T < 0.05$	5,167	0.046	0.224	0.519	0.001	5,298	0.064	0.302	0.493	0.001	5,512	−0.097	0.670	0.533	0.001
$p_T < 0.1$	10,480	0.058	0.262	0.511	0.000	11,046	−0.108	0.811	0.520	0.002	11,294	0.256	0.185	0.531	0.003
$p_T < 0.2$	20,980	0.111	0.216	0.518	0.001	22,601	−0.208	0.856	0.531	0.002	23,016	0.235	0.182	0.584	0.003
$p_T < 0.3$	31,551	0.207	0.137	0.517	0.001	34,309	−0.211	0.821	0.527	0.002	34,743	0.068	0.411	0.491	0.000
$p_T < 0.4$	42,243	0.209	0.188	0.517	0.001	46,016	−0.316	0.878	0.533	0.003	46,445	0.143	0.349	0.531	0.000
$p_T < 0.5$	52,786	0.250	0.188	0.516	0.001	57,671	−0.401	0.930	0.550	0.005	58,158	−0.017	0.516	0.500	0.000
$p_T < 0.6$	63,222	0.367	0.132	0.524	0.001	69,495	−0.331	0.869	0.540	0.003	69,675	0.098	0.417	0.490	0.000
$p_T < 0.7$	73,810	0.397	0.146	0.522	0.001	81,133	−0.338	0.852	0.529	0.002	81,346	−0.104	0.582	0.533	0.000
$p_T < 0.8$	84,459	0.393	0.178	0.516	0.001	92,855	−0.327	0.824	0.531	0.002	93,031	−0.101	0.572	0.537	0.000
$p_T < 0.9$	94,991	0.420	0.189	0.520	0.001	104,610	−0.401	0.867	0.527	0.003	104,685	−0.172	0.615	0.539	0.000
$p_T \leq 1.0$	105,432	0.436	0.204	0.512	0.001	116,221	−0.289	0.773	0.522	0.001	116,214	−0.429	0.752	0.532	0.001

Prediction results are based on a logistic regression of serial self-employment on the genetic risk score. The first column specifies that threshold *p*-value of SNPs in the discovery meta-analysis that was used to exclude SNPs from the prediction score; the last row did not apply a filter. The column *n* SNPs indicates the number of directly genotyped SNPs that were used to calculate the score in STR for 12 overlapping significance thresholds. The *p*-value indicates the significance of the regression coefficient on the genetic risk score (one-sided Wald test). Column AUC gives the *C*-statistic for the Area Under the Curve. The R^2 is the Nagelkerke pseudo-R^2 from the logistic regression.

In all specifications, the Nagelkerke R^2 of all models is below 0.01% (Table 5.7), and the coefficient is never statistically significant. Furthermore, in all instances the area under the curve (AUC, see Zhou, Obuchowski, & McClish, 2002) is very close to 0.5, indicating that the prediction accuracy is very close to a random draw. Evidently, a discovery sample of 5,930 is too small to generate any statistically detectable predictive power.

5.7 Discussion

Our findings from a twin study and from molecular genetic data provide additional evidence for the heritability of entrepreneurship. These results add to a growing literature suggesting that variation in many economic behaviors and outcomes can be accounted for by genetic factors (Benjamin, Cesarini, Van der Loos, et al., 2012; Cesarini, Dawes, et al., 2009; Cesarini et al., 2010, 2012; Cesarini, Lichtenstein, et al., 2009; Chaturvedi et al., 2012; Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Taubman, 1976; Zhang et al., 2009).

The evidence that entrepreneurship is heritable has prompted scholars to speculate about the use of molecular genetic data for understanding and predicting entrepreneurship (Shane, 2010). It has been suggested that molecular genetic data could allow researchers to tackle interesting new questions such as whether managers and entrepreneurs have similar genetic endowments (Nicolaou & Shane, 2010). It has also been proposed that it may become possible to develop genetic tests that score an individual's propensity to become an entrepreneur (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008). However, whether genes could eventually be used to predict outcomes such as entrepreneurship is not only a function of heritability, but also depends on the *molecular genetic architecture* of the trait, i.e., the joint distribution of effect sizes and allele frequencies of the causal genetic variants (Lander, 2011).

A high heritability implies that genetic factors can in principle explain a large share of the observed differences across people and that genes could be in principle be used to make relatively accurate predictions of outcomes and behaviors. However, the molecular genetic architecture of a trait has implications for if and when these potentials may eventually be realized. If a few genes with relatively large effects on entrepreneurship exist, it would be possible to detect them in relatively small samples. If, however, the heritability is accounted for by a large number of genetic variants, each with very small individual effects, identifying these variants and using them for practical purposes would require much larger sample sizes. Given the sample sizes used in previous work, scholars seeking to identify genes for entrepreneurship were implicitly operating under the assumption that a few genes with large effects exist (Nicolaou & Shane, 2009; Nicolaou et al., 2011; Shane, 2010).

Our results indicate that this assumption is likely to be wrong. Although our combined sample is more than six times larger than the most recent candidate gene study on entrepreneurship, none of the candidate genes that were proposed to have large effects in the previous literature are significantly associated with serial self-employment in our study. Moreover, our evidence in its entirety suggests that there are no common SNPs in our data with strong effects.

Our top hit has an odds ratio of 1.042 in the replication sample. This effect is much smaller than what is reported in the candidate gene studies that have not survived replication (Nicolaou et al., 2011; Van der Loos, Koellinger, et al., 2011). The relatively small effect size is not surprising given the recent evidence from large scale GWA studies on complex traits such as body height (Lango Allen et al., 2010; Yang et al., 2010), cognitive ability (Davies et al., 2011), personality (De Moor et al., 2012; Terracciano et al., 2010; Verweij et al., 2010), economic preferences (Benjamin, Cesarini, Van der Loos, et al., 2012), and even clinical diagnoses such as schizophrenia (Lee et al., 2012), type 2 diabetes (Steinthorsdottir et al., 2007), or depression (Shyn et al., 2011), none of which have identified individual genes with large effects.[23]

In light of the fact that most true associations probably have small effects (Benjamin, Cesarini, Chrabris, et al., 2012), discovering the genetic variants linked to entrepreneurship is likely to require much larger datasets than are currently available. To calibrate expectations about an appropriate sample size for discovering individual SNPs, we calculated the sample size that would be required for 80% statistical power to replicate the top hit from this study (rs3774790) at a nominal significance level of $p = 0.05$. Assuming a replication sample with the same rate of serial self-employment, an odds ratio of 1.042, and the same minor allele frequency that we observe in our data, more than 1.2 million observations would be needed (Purcell et al., 2003). Even larger samples are required to identify SNPs with smaller effect sizes. Such enormous sample sizes will most likely not be available in the next few years. Therefore, many of the suggested practical uses of genetic data (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Shane, 2010) do not appear to be feasible in the immediate future.

Nevertheless, our GREML results demonstrate that a substantial part of the observed heritability among males can be accounted for by the common SNPs that are measured in our data. Thus, the combined evidence from the GWAS and GREML analyses suggest that many common genetic variants, each with very small effect sizes, contribute to the herita-

[23] In principle, it is possible that there exist rare genetic variants, or other currently unmeasured variants not sufficiently correlated with the SNPs in our genotyping platform, with large effects on social science outcomes (Freeman et al., 2006). It is also possible that genetic interaction contributes significantly towards the heritability of entrepreneurship. If this is true, large samples would still be required to detect these effects in the future (Lee et al., 2012; Verweij et al., 2012; Wray et al., 2011).

bility of entrepreneurship. If, as our results suggest, the genetic architecture of entrepreneurship is diffuse, accurate prediction will require very large sample sizes because the effects of each SNP need to be estimated with sufficient accuracy (Visscher et al., 2010). This reasoning is confirmed by our finding that all (> 2 million) SNPs in our discovery samples together jointly predict less than 0.01% of the variance in serial self-employment in the replication sample.

Our findings are largely consistent with a study by Van der Loos, Rietveld, et al. (2013) on being self-employed at least once. Arguably, being self-employed at least once is a noisier proxy for entrepreneurship than serial self-employment, but using this measure allowed these authors to investigate eighteen instead of only three samples, yielding a total sample size of 62,820 individuals. That study estimated that being self-employed at least once is moderately heritable ($h^2 \approx 50\%$) and that approximately half of this heritability is due to individual variation in common SNPs. However, the prediction accuracy of these common SNPs is still very limited ($R^2 < 0.32\%$). Their GWAS analysis of single self-employment did also not find any genome-wide significant results, although they had 80% power to detect common SNPs with a minor allele frequency in excess of 0.25 and an odds-ratio of 1.11 at $p = 5 \times 10^{-8}$.

Given these conclusions regarding the molecular genetic architecture of entrepreneurship, we propose two avenues for future research on the genetics of entrepreneurship. A first way forward is to shift the focus to biologically proximate variables that mediate the relationship between genes and entrepreneurship. Examples of such variables that can be measured in large samples may include preferences toward risk and uncertainty, confidence, and optimism. One advantage of this approach is that genetic effects on more biologically proximate outcomes are likely to be stronger and hence easier to detect, for a given sample size, than the genetic effects on biologically distal outcomes, such as entrepreneurship. In addition, it seems likely that the genetic factors that predict these biologically proximate outcomes vary less across the different environments that must be pooled when conducting a GWAS meta-analysis. In contrast, genetic markers associated with entrepreneurship may be relatively difficult to detect in part because the relative importance of the determinants of entrepreneurship differs across environments.

A second and complementary way forward is to use new statistical approaches that make more efficient use of the genetic data in the aggregate to shed light on the biology underlying a trait. For example, a recently developed extension of GREML asks whether relatedness estimated from a set of SNPs that are believed to be functionally related explain a disproportionate share of variation. Using this method, one paper recently found that SNPs involved in the central nervous system explain a disproportionate amount of variance in the liability to schizophrenia (Lee et al., 2012). Although this method does not

directly identify individual SNPs, it could implicate specific biological systems and thereby help focus the search for specific variants that are associated with entrepreneurship.

If these future research directions eventually enable robust discoveries of specific genetic variants, they may lead to a successful revival of candidate gene approaches. With empirically well-supported hypotheses to guide research, it will then be possible to systematically investigate gene-gene and gene-environment interactions, which likely play an important role in complex behavioral traits such as entrepreneurship.

The conclusion that the heritable variation in serial entrepreneurship is likely to be accounted for by the sum of a large number of tiny genetic effects echoes similar findings for other social science traits (Benjamin, Cesarini, Van der Loos, et al., 2012). While the focus of this chapter has been on serial entrepreneurship, we conjecture that these lessons also generalize to most other behavioral traits in the management sciences and the social sciences.

5.8 Conclusion

The "quest for the entrepreneurial gene" (Van der Loos, Koellinger, et al., 2011) is largely motivated by the struggle of scholars to understand entrepreneurs better, what motivates them, and what makes them different from other people. Various research approaches, including tools and theories from economics, psychology, and sociology have been proposed and applied to these questions, yet the answers to "what makes an entrepreneur" remain uncertain and incomplete (Shane & Venkataraman, 2000). Evidence that genes may be part of the answer (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Zhang et al., 2009) has been received with both great hopes and enthusiasm, as well as with skepticism and critique among scholars and in the media. Here, we contribute to this debate by investigating the genetic architecture of entrepreneurship, measured as serial self-employment, using three large samples of comprehensively genotyped individuals from the Netherlands and Sweden and state-of-the-art methods from genetic epidemiology. These methods are likely to gain importance in the social sciences as well (Beauchamp et al., 2011; Benjamin, Cesarini, Chabris, et al., 2012; Benjamin, Cesarini, Van der Loos et al., 2012).

Our results are consistent with earlier findings from twin studies (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Zhang et al., 2009) suggesting that the propensity to engage in entrepreneurship is to some extent genetically influenced. This implies that individual differences should remain an important topic for future research on entrepreneurship (Shane & Venkatamaran, 2000). However, the partial heritability of entrepreneurship that we report does not suggest any kind of determinism or absence of free will. Rather, the estimated heritability parameters simply reflect how much of the variance in entrepreneurial behavior can be attributed towards genetic

differences among people in the observed data. Thus, heritability estimates are population parameters that can fluctuate across populations and depending on environmental conditions, rather than universal truths.

Furthermore, our results help to put the insight that entrepreneurship is partially heritable into perspective. Specifically, our empirical evidence is inconsistent with the view that a small number of well-understood genes with strong effects are responsible for the observed heritability of entrepreneurship. Rather, our findings suggest that entrepreneurship is a genetically very complex trait, with possibly thousands of genes that each exercises a small influence on entrepreneurial propensity. We demonstrate an important practical implication of this insight: It is currently not possible yet to predict the entrepreneurial propensity of an individual from genetic data with a practically relevant degree of accuracy. Thus, hopes and fears related to possible applications of genetic predictions of entrepreneurial propensity (Nicolaou, Shane, Cherkas, Hunkin, & Spector, 2008; Nicolaou, Shane, Cherkas, & Spector, 2008; Shane, 2010) are currently not warranted yet. The extent to which genetic prediction will become possible in the future depends on two factors. First, prediction accuracy increases with the sample size of comprehensively genotyped individuals that can be used to study entrepreneurial behavior. Our sensitivity analyses suggest that population-based samples of more than one million individuals are likely to be needed to make substantial progress using the currently available methods. However, such data do not exist yet. Second, prediction accuracy depends on how stable the influence of specific genes on entrepreneurial propensity is in different environments. Given the currently available evidence, it is too early to conclude that specific genes exist that have a universal influence on entrepreneurship that is independent from environmental conditions.

However, the available empirical evidence also suggests that future research could identify such genes that influence entrepreneurial behavior, even though the effect of every single gene that will be discovered is likely to be small. Future research in this direction using the available research methods will need to employ much larger samples than currently possible. Alternatively, novel research strategies and statistical approaches will be developed and applied to make progress. Thus, we expect that research on genetic and biological influences on entrepreneurship and human behavior in general will remain an active and potentially rewarding field of research.

CHAPTER 6

Measures of Bioactive Serum Testosterone Are Not Associated with Entrepreneurial Behavior in Two Independent Observational Studies

Based on Van der Loos, Haring, et al. (2013).

Abstract

Previous research has suggested a positive association between testosterone (T) and entrepreneurial behavior in males. However, this evidence was found in a study with a small sample size and has not been replicated. In the present study, we aimed to verify this association using two large, independent, population-based samples of males. We tested the association of T with entrepreneurial behavior, operationalized as self-employment, using data from the Rotterdam Study ($n = 587$) and the Study of Health in Pomerania ($n = 1{,}697$). Total testosterone (TT) and sex hormone-binding globulin (SHBG) were measured in the serum. Free testosterone (FT), non-SHBG-bound T (non-SHBG-T), and the TT / SHBG ratio were calculated and used as measures of bioactive serum T, in addition to TT adjusted for SHBG. Using logistic regression models, we found no significant associations between any of the serum T measures and self-employment in either of the samples. To our knowledge, this is the first large-scale study on the relationship between serum T and entrepreneurial behavior. Most likely, the absent association between T and entrepreneurship in our study suggests that the previous report of a positive association was the result of publication bias and/or low statistical power.

6.1 Introduction

Entrepreneurial behavior is an important element and is the driving force of dynamic changes in modern economies (Schumpeter, 1934). Empirical evidence suggests that important economic stimuli ensue from entrepreneurship (Kirzner, 1973; Koellinger & Thurik, 2012; Roessler & Koellinger, 2012). Thus, understanding the motivations underlying entrepreneurial behavior is highly relevant. Individual socio-demographic characteristics such as age, sex, and educational attainment have traditionally been a major research focus (Shane & Venkataraman, 2000), but recently, increased testosterone (T) levels have been suggested to be a biological predisposing factor for entrepreneurial behavior.

Specifically, White et al. (2006) observed that many of the features that characterize entrepreneurs correlate with T. For example, risk-taking behavior is a much-debated feature of entrepreneurship (Stewart & Roth, 2001, 2004) and has been shown to be associated with T (Apicella et al., 2008; Brañas-Garza & Rustichini, 2011; Coates & Hebert, 2008; Goudriaan et al., 2010; Sapienza, Zingales, & Maestripieri, 2009; Stanton, Liening, & Schultheiss, 2011). Based on such relationships, White et al. argued that higher T levels may induce entrepreneurial behavior, and they developed a theoretical basis for this relationship by drawing upon evolutionary psychology theory. These authors hypothesized that individuals with higher T levels are more likely to engage in new venture creation, the author's measure of entrepreneurship, and that this relationship is partially mediated by an individual's risk propensity. White et al. found evidence for their hypothesis using a salivary T measure in a sample of 110 male North American MBA students. However, this study was limited by a small sample size and has not been replicated. Although, there is some evidence to suggest that the 2D:4D digit ratio, a hypothesized proxy for prenatal and adult T levels (Manning, Scutt, Wilson, & Lewis-Jones, 1998; Tan, 2008), and entrepreneurial behavior are associated (Guiso & Rustichini, 2011; Trahms, Coombs, & Barrick, 2010; Unger, Rauch, Narayanan, Weis, & Frese, 2009), there is contradictory evidence about the validity of this digit ratio as a proxy for T (see Folland, Mc Cauley, Phypers, Hanson, and Mastana, 2012, for the most recent discussion). The precise role of T in entrepreneurial behavior therefore remains unknown.

Thus, the aim of our study was to evaluate the relationship between entrepreneurship, operationalized as self-employment, in a much larger sample of males than previously used. In particular, we utilized two large, independent, population-based samples of males and measured their serum T levels, in contrast to the salivary T measure used by White et al. (2006). In the serum, T is mainly bound to sex hormone-binding globulin (SHBG) and albumin, leaving only a small fraction of T unbound or free. By binding to T, SHBG prohibits T from diffusing from the bloodstream into target tissue cells and perform its biological function. Hence, free testosterone (FT) is generally regarded as bioactive. It is unclear if albumin-bound T is also bioactive (Manni et al., 1985; Mendel, 1989). In our analyses,

we used FT and non-SHBG-bound T (albumin-bound and free) as measures of bioactive T, in addition to the total T (TT) / SHBG ratio and TT adjusted for SHBG. Our measures are similar to the salivary T measure used by White et al. because salivary T reflects the part of serum T that is free (Vining & McGinley, 2006). For completeness, we also tested for an association between TT and self-employment. Based on previous findings, we hypothesized that there is a positive association between the serum T measures and self-employment.

6.2 Materials and Methods

6.2.1 Ethics Statement

All of the participants provided written informed consent, and the study was approved by the medical ethics committee of the Erasmus Medical Center and the local Ethics Committee of the University of Greifswald.

6.2.2 Participants

We used cross-sectional data from two population-based cohorts: the Rotterdam Study (RS) and the Study of Health in Pomerania (SHIP). The RS is a large, population-based cohort study of the elderly that has been ongoing since 1990 in the city of Rotterdam in the Netherlands (Hofman et al., 1991, 2011). From 1990 to 1993, 10,215 inhabitants aged 55 and over from the Ommoord district were invited to participate, and 7,983 (response 78%) took part in the baseline examination, including 3,105 males. In addition to the original cohort used here (RS-I), there are two other cohorts included in the Rotterdam Study (RS-II and RS-III), but T was not measured in these cohorts. We excluded those participants who used sexual hormones (ATC code G03), testosterone 5a-reductase inhibitors (ATC code G04CB), sexual hormone antagonists (ATC code L02B), or anabolic steroids (ATC code A14A) because of the effects of these drugs on the serum T level. Males with missing data on hormone levels, self-employment, or covariates were also excluded, leaving 589 males from the RS in our sample.

The SHIP is a population-based cohort study ongoing in West Pomerania, a region in northeastern Germany (Völzke et al., 2011). A two-stage cluster sampling method was adopted from the WHO MONICA Project (Augsburg, Germany) to select a sample of 7,008 individuals from the entire population of 212,157 people living in the area using the population registration offices, where all German inhabitants are registered. Individuals without German citizenship and those who did not reside in the study area were excluded. The final sample comprised 4,308 participants at baseline (response 69%), including 2,116 males. We excluded users of sexual hormones (ATC code G03), testosterone 5a-reductase inhibitors (ATC code G04CB), sexual hormone antagonists (ATC code L02B), or anabolic

steroids (ATC code A14A). After excluding males with missing data on self-employment, hormone levels, or covariates, 1,697 males from the SHIP were available for the analyses.

6.2.3 Hormone Measurements

The serum TT and sex hormone-binding globulin (SHBG) levels were measured using coated tube (T) or double antibody (SHBG) radioimmunoassays (Diagnostic Systems Laboratories, Inc., Webster, TX) in the RS and using competitive chemiluminescent enzyme immunoassays (Siemens Immulite 2500 Total Testosterone, ref. L5KTW, lot 110; Immulite 2550 SHBG ref. L5KSH, lot 119; Siemens Healthcare Medical Diagnostics, Bad Nauheim, Germany) in the SHIP. Further details have been described previously (Friedrich et al., 2008; Hak et al., 2002; Koch et al., 2011).

As measures of bioactive T, we used FT, non-SHBG-bound T (non-SHBG-T), the TT / SHBG ratio, and TT adjusted for SHBG. The serum FT and non-SHBG-T levels were calculated according to the method of Södergard, Bäckström, Shanbhag, and Carstensen (1982) using the equations described in De Ronde et al. (2005) and a fixed albumin level of 40 g/l.

6.2.4 Self-Employment and Covariate Measures

At baseline, participants from the RS were interviewed at home and asked for their complete work-life histories. The participants' occupations and employment status (employed, self-employed, or a collaborating family member) for each occupation were recorded. Based on this information, we were able to identify individuals who were self-employed at some point during their working careers and individuals who had not been self-employed. Individuals who had never had a job and individuals with an incomplete work-life history except those who were classified as self-employed at least once were excluded from our study. The rationale for excluding these individuals is that individuals with incomplete work-life histories could have been self-employed at least once in the past, which would make it impossible to interpret the coefficient for self-employment.

In the SHIP, participants were asked about their current or last occupational status using questionnaires. We coded individuals as self-employed if they reported that they were farmers with more than 10 hectares of property (2.5% of the self-employed), university graduates with a liberal profession, e.g., physician, lawyer, or tax accountant (8.6% of the self-employed), or self-employed in business, craft, or the tertiary sector (88.9% of the self-employed). The self-employment rate was lower in the SHIP than in the RS, in agreement with the fact that the SHIP is located in the former German Democratic Republic, where self-employment was systematically discouraged (Fritsch, 2004).

Body mass index (BMI) was calculated as weight in kilograms divided by the square of height in meters. In the RS, weight and height were measured during the research facili-

ty visit while participants were wearing indoor clothing and no shoes. SHIP participants were wearing lightweight clothing and no shoes during height and weight measurements. Current smoking status was assessed using a computerized questionnaire during the home interview in the RS and using computer-assisted personal interviews in the SHIP. To harmonize educational attainment measures across the RS and the SHIP, we first transformed the study-specific measures to an internationally comparable measure of educational attainment according to UNESCO's International Standard Classification of Education (ISCED) scale (United Nations Educational, Scientific and Cultural Organization, 2006). The ISCED levels were then converted to US years of schooling equivalents.

6.2.5 Statistical Analysis

Categorical data are reported as percentages, and continuous data are represented as the mean together with the standard deviation. Differences between groups were tested using Pearson's χ^2 tests for categorical data and t-tests for continuous data.

We used logistic regression models to investigate the association between self-employment and serum T measures. These models were adjusted for age, age^2, and educational attainment because age has been shown to exhibit an inverted U-shaped relationship with entrepreneurship (Parker, 2009) and because of the positive effect of education on entrepreneurship (Block, Hoogerheide, & Thurik, 2013). We also controlled for BMI, current smoking, and time of blood sampling, as these are well-known confounders of serum T (Dai, Gutai, Kuller, & Cauley, 1988; Diver, Imtiaz, Ahmad, Vora, & Fraser, 2003; Feldman et al. (2002); Field, Colditz, Willett, Longcope, & McKinlay, 1994). All regression analyses were performed separately for the RS and SHIP samples, after a Chow test discouraged data pooling ($p < 0.001$ for all models).

We performed the following sensitivity analyses. First, we used the multivariable fractional polynomial (MFP) algorithm to test for non-linear effects of serum T measures (Sauerbrei & Royston, 1999) and to avoid categorization (Royston, Altman, & Sauerbrei, 2006; Schmidt, Ittermann, Schulz, Grabe, & Baumeister, 2013). Second, we performed all regression analyses in a subset of 743 males from the SHIP who were younger than 60 years and were part- or full-time (non-)self-employed at the time of the hormone measurements to ensure that T could not have been influenced by other factors in the period between the self-employment period and when the hormone measurements were taken. *p*-values smaller than 0.05 were considered significant. All statistical analyses were performed using Stata version 12.1 (Stata Corporation, College Station, TX, USA).

Table 6.1. Characteristics of the participants from the Rotterdam study (RS) and the Study of Health in Pomerania (SHIP).

	RS (n = 589)		SHIP (n = 1,697)		p for difference
Self-employed (%)	12.4		4.8		< 0.001
Age (years)	68.29	(7.63)	50.68	(16.56)	< 0.001
Educational attainment (US years of schooling)	11.47	(3.82)	13.56	(3.79)	< 0.001
Body mass index (kg/m2)	25.81	(2.94)	27.64	(4.09)	< 0.001
Current smoker (%)	28.9		34.2		0.017
TT (nmol/L)	11.36	(3.93)	16.76	(6.01)	< 0.001
SHBG (nmol/L)	35.9	(14.42)	51.6	(25.60)	< 0.001
FT (nmol/L)	0.27	(0.11)	0.34	(0.11)	< 0.001
Non-SHBG-T (nmol/L)	6.66	(2.60)	8.36	(2.81)	< 0.001
TT / SHBG ratio	0.37	(0.24)	0.37	(0.16)	0.984

TT: total testosterone; SHBG: sex hormone-binding globulin; FT: free testosterone; non-SHBG-T: non-SHBG-bound testosterone. Categorical data are presented as percentages, and continuous data are presented as the mean (standard deviation). p-values were calculated using Pearson's χ^2 tests for categorical data and t-tests for continuous data.

6.3 Results

The descriptive statistics for the participating males from the RS and the SHIP are reported in Table 6.1. The self-employment rate in the RS was higher than in the SHIP, and males from the RS were on average older than males from the SHIP. Educational attainment was higher in the SHIP than in the RS. BMI was lower in the RS than in the SHIP, and a larger percentage of males in the SHIP were smokers than in the RS. The serum TT, SHBG, FT, and non-SHBG-T levels were all lower in the RS than in the SHIP, although the TT / SHBG ratio was similar between the two. The differences between the RS and SHIP were all significant except for the difference in the TT / SHBG ratio. The mean serum TT, FT, and non-SHBG-T levels did not significantly differ between self-employed and non-self-employed males in the RS or the SHIP (Figure 6.1; $p > 0.05$ for all comparisons). The logistic regression results indicated that none of the serum T measures were associated with self-employment in either of the samples (Table 6.2). Sensitivity analyses using MFP models did not provide evidence for a non-linear association between serum T measures and self-employment. Finally, analyses in a subset of males from the SHIP who were (non)-self-employed at the time of hormone measurement did not reveal an association between any of the serum T measures and self-employment ($p > 0.05$ for all serum T measures).

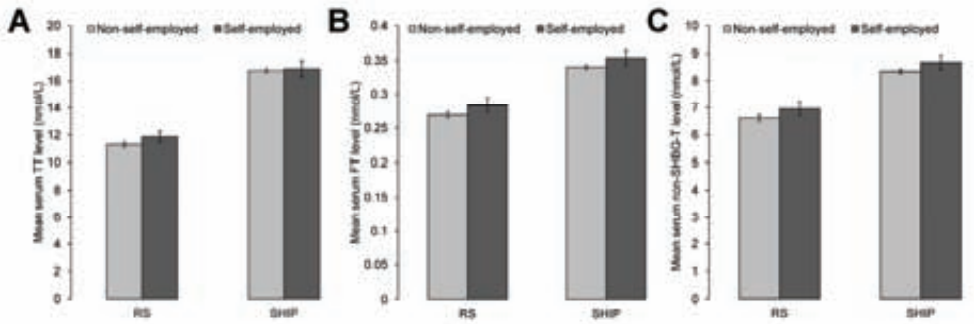

Figure 6.1. Mean serum testosterone measures by self-employment status for participants from the Rotterdam study (RS) and the Study of Health in Pomerania (SHIP). The figure shows the mean serum levels of total testosterone (TT; panel A), free testosterone (FT; panel B), and non-SHBG-bound testosterone (non-SHBG-T; panel C) by self-employment status for participants from the RS and the SHIP. The errors bars indicate the standard error of the mean. The mean serum TT, FT, and non-SHBG-T levels did not significantly differ between self-employed and non-self-employed participants from the RS or the SHIP ($p > 0.05$ for all comparisons).

6.4 Discussion

To our knowledge, this is the first large-scale investigation of the suggested association between T and entrepreneurial behavior based on serum T measures. We observed no association between any of the serum T measures and self-employment in two large, independent, population-based samples of males. Several sensitivity analyses were conducted to confirm the robustness of our results. First, we verified that T did not have a non-linear effect on self-employment. Second, in a subset of males from the SHIP who were younger than 60 years and part- or full-time (non-)self-employed at the time of the hormone measurements, no significant associations were found. These findings confirm that the association was not masked by the fact that a substantial proportion of the participants in our study were elderly males in which we tried to associate T—which could have been influenced by many other causes over time—to occupational choices that happened much earlier in their lives. Our findings contradict earlier evidence of a positive and significant relationship between salivary T levels and entrepreneurial behavior (White et al., 2006). There are at least four reasons to explain these divergent results. First, White et al. (2006) operationalized entrepreneurship as being involved full-time in new venture creation, whereas we defined entrepreneurship as being self-employed at least once during the working career, the most frequently used measure of entrepreneurial behavior in the economics literature (Parker, 2009). The different definition used by White et al. (2006) may imply that T

Table 6.2. Association between serum testosterone measures and self-employment in the Rotterdam study (RS) and the Study of Health in Pomerania (SHIP).

Serum T measure	β	(95% CI)	p-value
	RS		
TT	0.045	(−0.018, 0.109)	0.163
FT	1.479	(−0.905, 3.863)	0.224
Non-SHBG-T	0.060	(−0.037, 0.157)	0.224
TT / SHBG ratio	0.227	(−0.749, 1.204)	0.648
TT adjusted for SHBG	0.048	(−0.017, 0.113)	0.146
	SHIP		
TT	0.011	(−0.030, 0.053)	0.602
FT	0.762	(−1.583, 3.108)	0.524
Non-SHBG-T	0.031	(−0.065, 0.127)	0.524
TT / SHBG ratio	0.447	(−1.269, 2.162)	0.610
TT adjusted for SHBG	0.019	(−0.034, 0.072)	0.486

TT: total testosterone; SHBG: sex hormone-binding globulin; FT: free testosterone; non-SHBG-T: non-SHBG-bound testosterone. All models were adjusted for age, age^2, educational attainment, time of blood sampling, BMI, and current smoking status.

merely plays a role in the initial phases of starting a new business and is irrelevant to being self-employed. However, new venture creation and self-employment are strongly correlated ($\rho > 0.7$, see Nicolaou, Shane, Cherkas, Hunkin, and Spector, 2008), and we would thus expect, at least to a certain extent, an association between T and self-employment.

Second, it is known that measurements of T in saliva, as used by White et al. (2006), can be influenced by sample handling, leakage of blood (plasma) into the saliva, and the storage conditions of archived samples (Granger, Shirtcliff, Booth, Kivlighan, & Schwartz, 2004), challenging the validity of these measurements.

Third, White et al.'s (2006) strongly selected study sample, comprising North American MBA students, is very different from the Dutch and German population-based samples used in the present study. The non-replication may be an effect of this heterogeneity, meaning that the effects of T may differ depending on the environment.

Fourth, it is well known that non-replication is especially pronounced in studies with small sample sizes because such studies are underpowered (Ioannidis, 2005). Our calculations indicate that White et al.'s (2006) sample was seriously underpowered to find a true effect of T on entrepreneurial behavior. For example, if we adopt a 5% significance level and assume that, similar to White et al.'s sample, 28% of the individuals in a population are involved full-time in new venture creation, then the smallest detectable odds ratio with 80% power in a sample of $n = 110$ is approximately 1.9. However, White et al. (2006, Table 4, column 5) estimated that the odds ratio of T (in pg/ml) for being engaged full-time

in new venture creation is approximately 1.03 ($e^{0.03}$). If we assume that their odds ratio is an unbiased estimate of the population odds ratio, then they would have needed approximately 45,000 participants to achieve 80% power to detect such an effect. Therefore, the original association is more likely to be due to chance than to a true difference. The replication of the original association in an independently obtained sample would have increased the probability that the initial findings were true (Moonesinghe, Khoury, & Janssens, 2007).

Our findings are in agreement with the results of a randomized clinical trial of T supplementation that did not reveal any significant effects of serum T on a number of economic behaviors (Zethraeus et al., 2009). Because this clinical trial was performed in females, who are generally considered more risk averse and less competitive than males, the potential effect of T supplementation, if any, was expected to be even more pronounced than in males.

Our results do not necessarily rule out an influence of T on entrepreneurial behavior. Our largest sample had 80% power to detect odds ratios larger than 1.4, which is equivalent to an odds ratio of approximately 1.9 for being engaged in new venture creation given the correlation with self-employment. Therefore, we cannot exclude the existence of an effect of T on entrepreneurial behavior with a smaller effect size. Larger samples will be needed in future studies to draw definitive conclusions regarding the potential association between T and entrepreneurial behavior. However, we believe that the practical utility of such an effort will be very small. If we assume that our estimates of the regression coefficients are the true population parameters, then the odds ratio of FT (in pg/ml) for being at least once self-employed is almost one.

The cross-sectional nature of our study does not allow causal inferences to be drawn. Additionally, there may have been omitted variables, selection bias, reverse causality, or measurement error, which could have led to endogeneity for T and subsequently the inconsistent estimation of the logistic regression model parameters. In such cases, Mendelian Randomization is an attractive approach to tackle the endogeneity problem and allows the inference of causal relationships with cross-sectional data (Davey Smith & Ebrahim, 2003; Von Hinke Kessler Scholder, Davey Smith, Lawlor, Propper, & Windmeijer, 2011). Statistically, Mendelian Randomization is the use of instrumental variables (IV) regression using genes as instruments (Wehby, Ohsfeldt, & Murray, 2008). We considered performing a Mendelian Randomization analysis but found that the only candidate instrument for our serum T measures that is currently available, single nucleotide polymorphism (SNP) rs5934505 on chromosome X (Ohlsson et al., 2011), was correlated with self-employment and thus violated the crucial *exclusion restriction*. The other two SNPs identified in the genome-wide association study of T were only associated with TT (Ohlsson et al., 2011) and therefore could not be used as instruments for measures of bioactive T.

Another limitation of our study is the different age ranges in the RS and SHIP. Whereas the RS participants were aged over 55 years at baseline and had data on their complete working-life history, the age range of the SHIP participants was 20 to 80 years, implying a right-censored self-employment status. Furthermore, our data included only male participants, and a role of T in females cannot be ruled out a priori. We restricted our analyses to males because the original association was based on results from a male-only sample. Furthermore, T levels were unavailable for females in the SHIP and were available only for postmenopausal females in the RS.

In conclusion, using two large, independent, population-based samples of males, we did not find evidence to support the hypothesis that T is associated with entrepreneurial behavior and thereby failed to replicate findings from a previous study. The previously reported association was most likely a false-positive finding due to a combination of low statistical power, which implies a high false discovery rate, and publication bias. Hence, findings from studies with small sample sizes that report associations between biological characteristics and (economic) behaviors should be interpreted with care and, ideally, replicated in adequately powered, independent samples to avoid the publication of false-positive results.

APPENDIX A

Supplementary Tables to Chapter 4

Table A1. Study design, sample size, sample quality control, and self-employment measure within each study.

Study			Sample QC			
Abbreviation	Full name	Study design	Total sample size (*n*)	Call rate	Other exclusions	Sample size after QC (*n*)	Self-employment measure
AGES	Age, Gene/Environment Susceptibility–Reykjavik Study	Population-based	3,219	≥ 98%	(1) Mismatch previous genotypes; (2) gender discrepancy; (3) missing self-employment status.	3,219	Questionnaire: Are you an employer or manager?
ASPS	Austrian Stroke Prevention Study	Population-based	2,008	≥ 98%	(1) Excess autosomal heterozygosity; (2) mismatch between called and phenotypic gender; (3) outliers in the IBD analysis; (4) missing self-employment status.	834	Participants were systematically asked about their life-long professional activities and this information is recorded in the subjects' charts. In addition information was gathered as to whether the subjects conducted their professional activities as employees or self-employed.
ERF	Erasmus Rucphen Family study	Family-based	3,500	≥ 98%	(1) Heterozygosity: FDR < 1%; (2) ethnic outliers; (3) duplicates; (4) gender mismatch; (5) excess IBS incompatible with pedigree; (6) missing self-employment status.	1,071	Individuals were asked if they were ever self-employed (no; yes—once; yes—more than once; always; not applicable/no response) and coded as self-employed if they were self-employed at least than once. The control group consists of people who were never self-employed.
GHS	Gutenberg Health Study	Population-based	3,500	≥ 97%	(1) Deviations in expected heterozygosity; (2) doubtful IBS patterns; (3) missing self-employment status.	3,130	Categorization of the actual job or last job (for participants who retired); questions on each job phase of the professional career were collected in a computer-assisted personal interview (CAPI).

Table A1. (continued).

H2000	Health 2000	Popula-tion-based	8,028 ≥ 95%	(1) Excess heterozygosity; (2) relatedness and/or failed gender check; (3) missing self-employment status.	2,123	Participants were asked "Are you/were you 1) Salary earner, 2) Agricultural entrepreneur, 3) Other entrepreneur, 4) Self-employed person or freelancer, 5) Working in a family member's farm without a salary, 6) Working in a family member's business without a salary, 7) Other, 8) Never been working in a full time job, 9) don't know" Salary earners were coded as controls and participants that answered 2,3 or 4 were coded as cases. Other participants were excluded.
HBCS	Helsinki Birth Cohort Study	Birth cohort study	8,760 ≥ 95%	(1) Excess heterozygosity; (2) relatedness and/or failed gender check; (3) missing self-employment status.	1,724	In the HBCS, data on classification of the socio-economic status and self-employment were derived from the Statistics Finland. These data were available with five-year-interval from 1970 to 2000. Participants were classified as self-employed if at any of the time points they indicated self-employment. Participants were excluded if data in more than two (out of seven) time points were missing ($n = 4$). In addition, those controls with indication of self-employment were excluded.

Table A1. (continued).

HRS	Health and Retirement Study	Population-based	12,507 ≥ 98%	(1) Relatedness check; (2) ethnic outliers; (3) missing self-employment status.	6,220	From the HRS RAND v.L dataset the binary variables r*slfemp were used that indicate if an individual was self-employed or working for someone else in wave *. Individuals are coded as self-employed if they responded in at least one data wave to be self-employed. The control group consists of people who were never self-employed but indicated at least once to work for someone else.
KORA S4	Cooperative Health Research in the Region of Augsburg	Population-based	4,261 > 93%	(1) Gender mismatch; (2) missing self-employment status.	1,724	Questionnaire based. "Which position do you/did you have in your job?"
NFBC1966	Northern Finland Birth Cohort 1966	Population-based	12,231 ≥ 95%	(1) Gender discrepancy with genetic data from X-linked markers; (2) withdrawn consent; (3) duplicates and first and second degree relatives; (4) contaminated samples; (5) missing self-employment status.	4,234	Based on questionnaire questions Q1-Q3, the subjects were classified into 9 groups according to the instructions from Statistics Finland (Reference: Tilastokeskus Käsikirjoja 17: Sosioekonomisen aseman luokitus, 1989): 1) Farm businessmen 2) Other entrepreneurs 3) Upper white-collar workers 4) Lower white-collar workers 5) Blue-collar workers 6) Students 7) Pensioners 8) Unemployed (incl. long-term unemployed and unclassified) 9) Socio-economic status unknown. For the present analysis groups 1) and 2) are considered as cases and others as controls, excluding group 7) pensioners and from group 8) long-term unemployed and unclassified.

Table A1. (continued).

NTR1	Netherlands Twin Register Cohort 1	Twin study	29,852 > 90%	(1) Presence of genetic data; (2) gender discrepancy with genetic data; (3) unexpected IBS sharing; (4) contaminated samples; (5) duplicates and first and second degree relatives; (6) missing self-employment status.	1,555	Data came from eight surveys. Participants were asked to indicate whether they were self-employed (1991, 1993, 1995, 2004, 2009) or to indicate the type of organization that they worked in, with being self-employed as one of the answer categories (1997, 2000, 2002). If they indicated to be self-employed in any one of the surveys, they were classified as self-employed.
NTR2	Netherlands Twin Register Cohort 2	Twin study	29,852 > 90%	(1) Presence of genetic data; (2) gender discrepancy with genetic data; (3) unexpected IBS sharing; (4) contaminated samples; (5) duplicates and first and second degree relatives; (6) missing self-employment status.	984	Data came from eight surveys. Participants were asked to indicate whether they were self-employed (1991, 1993, 1995, 2004, 2009) or to indicate the type of organization that they worked in, with being self-employed as one of the answer categories (1997, 2000, 2002). If they indicated to be self-employed in any one of the surveys, they were classified as self-employed.

Table A1. (continued).

RS-I	Rotterdam Study Baseline	Population-based	7,983	≥ 97.5%	(1) Gender mismatch with typed X-linked markers; (2) excess autosomal heterozygosity > 0.336 ~ FDR > 0.1%; (3) duplicates and/or 1st or 2nd degree relatives using IBS probabilities >97% from PLINK; (4) ethnic outliers using IBS distances > 3SD from PLINK; (5) missing self-employment status.	5,374 Detailed background information on the entire work-life history of all participants was available, including the number and duration of self-employment spells. Participants who had at least one spell in self-employment were coded as cases, those with zero spells as controls. Participants with incomplete work-life histories were excluded.
RS-II	Rotterdam Study Extension of Baseline	Population-based	3,011	≥ 97.5%	(1) Gender mismatch with typed Xlinkedmarkers; (2) excess autosomal heterozygosity ($F < -0.055$); (3) duplicates and/or 1st degree relatives using IBD PiHAT >40% from PLINK; (4) ethnic outliers IBS distances > 4SD mean HapMap CEU cluster from PLINK; (5) missing self-employment status.	2,066 Information on current employment status was available and self-employed participants were coded as cases. Participants that indicated employment or that were collaborating family members were coded as controls. Other participants were excluded.

Table A1. (continued).

RS-III	Rotterdam Study Young	Population-based	3,932 ≥ 97.5%	(1) Gender mismatch with typed Xlinked markers; (2) excess autosomal heterozygosity ($F < -0.055$); (3) duplicates and/or 1st degree relatives using IBD PiHAT >40% from PLINK; (4) ethnic outliers IBS distances > 4SD mean HapMap CEU cluster from PLINK; (5) missing self-employment status.	1,925	Information on current or last employment status (if retired) was available and self-employed participants were coded as cases. Participants that indicated employment or that were collaborating family members were coded as controls. Other participants were excluded.
SardINIA	SardiNIA Study of Aging	Population-based	6,148 ≥ 95%	(1) missing genotype; (2) missing self-employment status.	4,142	Information on current employment status was available and self-employed participants were coded as cases.
SHIP	Study of Health in Pomerania	Population-based	4,308 ≥ 92%	(1) Duplicate samples (by IBS); (2) reported/genotyped gender mismatch; (3) missing self-employment status.	4,063	"Which occupational position do you currently have?" Participants were coded as self-employed if they answered that they were farmers with more than 10 hectare property, university graduates with a liberal profession (physician, lawyer, tax accountant, etc.) or self-employed (in business, craft, or the tertiary sector).

Table A1. (continued).

STR	Swedish Twin Registry	Twin study	10,946	≥ 97%	(1) Sex-check (heterozygosity of X-chromosomes); (2) deviations in heterozygosity of more than 5 SD from the population mean; (3) cryptically relatedness check; (4) missing self-employment status.	3,271	Participants were asked "(1) have you ever run a business?" and "(2) how many businesses have you in total been part of starting?" Participants were coded as cases if they answered "Yes" to the first and "≥1" to the second question. Controls were those that answered "No" to the first and "0" to the second question. Other participants were excluded.
THISEAS	The Hellenic study of Interactions between SNPs & Eating in Atherosclerosis Susceptibility	CAD case-control	1,877	> 95%	(1) Missing self-employment status; (2) heterozygosity; (3) gender mismatch; (4) ethnic outliers.	685	Information on current employment status was available. Participants were asked whether they were: 1) civil servant, 2) private employee, 3) self-employed, 4) part-time employee, 5) retired, or 6) house holding. Self-employed participants were coded as cases. Participants who were retired were excluded.
TwinsUK	The UK Adult Twin Registry	Twin study	4,427	≥ 95%	(1) Heterozygosity across all SNPs ≥2 s.d. from the sample mean; (2) evidence of non-European ancestry as assessed by PCA comparison with HapMap3 populations; (3) observed pairwise IBD probabilities suggestive of sample identity errors; (4) missing self-employment status.	3,155	Information on the duration of self-employment was available. Participants that indicated to have spent some time in self-employment were coded as cases, those that indicated to have never been self-employed were coded as controls. Other participants were excluded.

Table A1. (continued).

YFS	The Cardiovascular Risk in Young Finns Study	Population-based cohort	3,596 ≥ 95%	(1) Excess heterozygosity; (2) relatedness and/or failed gender check.	2,358	Study subjects were asked in years 1986, 1989, 1992 and 2001 if they were 1) Salary earners 2) Farm businessmen 3) Unpaid workers at their family farm 4) Entrepreneurs 5) Unpaid workers at their family business or 6) Others. Persons who indicated to be farm businessmen or entrepreneurs in any of the surveys were classified as cases. Those who hadn't in any time point reported to be an entrepreneur or farm businessman were selected as controls.

Table A2. Genotyping, imputation, SNP quality control, and statistical analysis within each study.

	Genotyping				Imputation and quality control before meta-analysis				Association analysis	
	SNP inclusion criteria				SNP inclusion criteria					
Study	MAF	Call rate	HWE	Genotyped SNPs after QC	Imputation software	MAF	Imputation quality	Imputed SNPs after QC	Software	Covariates
AGES	≥ 1%	≥ 98%	$\geq 10^{-6}$	317,344	MACH	≥ 1%	Rsq ≥ 0.4	2,176,303	ProbABEL	(1) Sex in pooled sample; (2) first four PCs in all samples; (3) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
ASPS	≥ 1%	≥ 98%	$\geq 10^{-6}$	550,635	MACH	≥ 5%	Rsq ≥ 0.4	2,164,654	GenABEL	(1) Sex in pooled sample; (2) first four PCs.
ERF	≥ 1%	≥ 98%	$\geq 10^{-6}$	650,197	MACH	≥ 1%	Rsq ≥ 0.4	2,353,164	ProbABEL	(1) Sex in pooled sample; (2) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
GHS	≥ 1%	≥ 98%	$\geq 10^{-4}$	649,182	IMPUTE	≥ 5%	info ≥ 0.4	2,220,912	SNPTEST	(1) Sex in pooled sample; (2) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50
Health 2000	≥ 1%	≥ 95%	$\geq 10^{-6}$	555,418	MACH	≥ 1%	Rsq ≥ 0.4	2,463,699	ProbABEL	(1) Sex in pooled sample; (2) first four PCs; (3) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
HBCS	≥ 1%	≥ 95%	$\geq 10^{-6}$	533,491	MACH	≥ 1%	Rsq ≥ 0.4	2,416,556	ProbABEL	(1) Sex in pooled sample; (2) first four PCs in all samples.
HRS	≥ 1%	≥ 98%	$\geq 10^{-4}$	2,195,306	MACH	≥ 1%	Rsq ≥ 0.4	2,227,690	PLINK	(1) Sex in pooled sample; (2) first four PCs in all samples; (3) age dummies for the categories 30–39 (reference), 40–49, ≥ 50.
KORA S4	—	—	—	909,622	IMPUTE	≥ 1%	info ≥ 0.4	2,521,850	SNPTEST	(1) Sex in pooled sample; (2) age categories (≤ 39; 40–49; ≥ 50).

Table A2. (continued).

NFBC1966	≥ 5%	≥ 95%	$\geq 10^{-4}$	328,007	IMPUTE	≥ 1%	info ≥ 0.4	2,405,775	SNPTEST	(1) Sex in pooled sample; (2) first four PCs.
NTR1	≥ 1%	≥ 95%	$> 10^{-5}$	427,049	IMPUTE	≥ 1%	info ≥ 0.4	2,420,149	SNPTEST	(1) Sex in pooled sample; (2) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
NTR2	> 1%	> 95%	$> 10^{-5}$	528,072	IMPUTE	≥ 1%	info ≥ 0.4	2,532,400	SNPTEST	(1) Sex in pooled sample; (2) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
RS-I	≥ 1%	≥ 98%	$\geq 10^{-6}$	512,349	MACH	≥ 1%	Rsq ≥ 0.4	2,433,150	MACH2DAT	(1) Sex in pooled sample; (2) first four PCs.
RS-II	≥ 1%	≥ 97.5%	$\geq 10^{-6}$	466,389	MACH	≥ 1%	Rsq ≥ 0.4	2,432,613	MACH2DAT	(1) Sex in pooled sample; (2) first four PCs.
RS-III	≥ 1%	≥ 97.5%	$\geq 10^{-6}$	514,073	MACH	≥ 1%	Rsq ≥ 0.4	2,436,797	MACH2DAT	(1) Sex in pooled sample; (2) first four PCs; (3) dummy for age ≥ 50.
SardINIA	≥ 5%	≥ 95%	$\geq 10^{-6}$	356,359	MACH	≥ 5%	Rsq ≥ 0.4	1,972,533	Merlin	(1) Sex in pooled sample.
SHIP	—	—	—	869,224	IMPUTE	≥ 1%	info ≥ 0.4	2,514,047	QUICKTEST	(1) Sex in pooled sample; (2) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
STR	≥ 1%	≥ 97%	$\geq 10^{-7}$	644,556	IMPUTE	≥ 1%	info ≥ 0.4	2,481,423	Merlin	(1) Sex in pooled sample; (2) first four PCs.
THISEAS	≥ 5%	≥ 95%	$\geq 10^{-6}$	96,015	—	≥ 1%	—	95,510[a]	PLINK	(1) Age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
TwinsUK	≥ 5%	≥ 95%	$\geq 10^{-5}$	536,559	IMPUTE	≥ 1%	info ≥ 0.4	2,262,054	GenABEL	(1) Sex in pooled sample; (2) first four PCs; (3) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
YFS	≥ 1%	≥ 95%	$\geq 10^{-6}$	546,677	MACH	≥ 1%	Rsq ≥ 0.4	2,409,746	ProbABEL	(1) Sex in pooled sample; (2) first four PCs in all samples; (3) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.

[a] Number of genotyped SNPs after filtering on minor allele frequency ≥ 1%.

Table A3. Genomic inflation factors.

Study	Pooled	Males	Females
AGES	1.000	1.019	0.992
ASPS	1.117	0.998	0.998
ERF	1.105	1.099	1.042
GHS	0.996	1.006	0.999
H2000	0.990	0.994	0.995
HBCS	1.005	0.995	1.007
HRS	1.008	1.006	0.998
KORA S4	1.000	1.008	1.001
NFBC1966	1.003	1.007	1.004
NTR1	1.006	1.015	1.002
NTR2	1.004	1.021	1.022
RS-I	1.022	1.008	1.003
RS-II	1.008	1.015	0.998
RS-III	1.000	1.004	1.013
SardINIA	1.074	1.038	1.156
SHIP	1.010	1.021	0.989
STR	0.996	0.997	0.990
THISEAS[a]	—	—	—
TwinsUK[b]	1.015	—	1.011
YFS	1.010	1.015	1.007

[a] THISEAS did not provide results using genome-wide SNP data and genomic control lambdas were therefore not estimated.

[b] The number of male subjects was insufficient for a male stratified analysis.

Table A4. Replication results of the twenty suggestive SNPs ($p < 1 \times 10^{-5}$) from the self-employment discovery meta-analyses for pooled males and females.

Discovery meta-analysis									Swedish Twin Registry		Combined meta-analysis		
SNP	Chr.	Effect / non-effect allele	Nearest gene	I^2	Cochran's *Q* test *p*-value	*n*	*p*-value	Overall freq.	*p*-value	Freq.	*p*-value	Direction	Improvement ?
rs6906622	6	T/C	*RNF144B*	0.0	0.644	49,942	4.10×10^{-6}	0.21	0.30	0.19	2.54×10^{-5}	++-++++++++-++++?++-	no
rs477588	6	T/C	*RNF144B*	0.0	0.646	49,942	4.15×10^{-6}	0.21	0.32	0.19	2.46×10^{-5}	++-++++++++-++++?++-	no
rs9350075	6	A/C	*RNF144B*	0.0	0.640	49,942	4.26×10^{-6}	0.79	0.31	0.81	2.63×10^{-5}	--+++++++++----?--+	no
rs3920593	6	T/C	*RNF144B*	0.0	0.641	49,942	4.35×10^{-6}	0.79	0.30	0.81	2.68×10^{-5}	--+++++++++----?--+	no
rs6928004	6	A/G	*RNF144B*	0.0	0.640	49,942	4.47×10^{-6}	0.79	0.31	0.81	2.71×10^{-5}	--+++++++++----?--+	no
rs4716268	6	A/C	*RNF144B*	0.0	0.639	49,942	4.60×10^{-6}	0.21	0.31	0.19	2.76×10^{-5}	++-++++++++-++++?++-	no
rs10776614	10	T/C	*ARHGAP22*	0.0	0.980	49,942	4.79×10^{-6}	0.16	0.78	0.17	6.78×10^{-6}	-+-----+--------?---	no
rs2358531	5	A/G	*SV2C*	0.0	0.660	46,812	4.79×10^{-6}	0.71	0.01	0.72	1.42×10^{-4}	---?------+--+--?--+	no
rs747533	10	A/G	*ARHGAP22*	0.0	0.973	49,942	5.26×10^{-6}	0.26	0.92	0.26	1.14×10^{-5}	-+-----+--------?--+	no
rs510367	6	A/G	*RNF144B*	0.0	0.677	49,942	5.33×10^{-6}	0.21	0.31	0.19	3.20×10^{-5}	++-++++++++-++++?++-	no
rs6899634	6	A/C	*RNF144B*	0.0	0.619	49,942	5.54×10^{-6}	0.21	0.32	0.19	3.19×10^{-5}	++-++++++++-++++?++-	no
rs17166082	7	A/G	*PLXNA4*	0.0	0.821	39,758	5.82×10^{-6}	0.06	0.88	0.05	9.37×10^{-6}	-?-?--?--+------?---	no
rs994208	14	C/G	*EGLN3*	0.0	0.741	49,942	6.11×10^{-6}	0.66	0.62	0.68	2.05×10^{-5}	-+---------------?--+	no
rs3847697	12	T/C	*LRIG3*	0.0	0.762	41,566	6.79×10^{-6}	0.44	0.64	0.43	2.19×10^{-5}	--+-----?-+-+-?-?--+	no
rs9371065	6	A/T	*RNF144B*	0.0	0.682	49,942	8.84×10^{-6}	0.22	0.32	0.19	4.84×10^{-5}	++-++++++++-++++?++-	no
rs2057556	6	T/C	*RNF144B*	0.0	0.693	49,942	8.98×10^{-6}	0.78	0.33	0.81	4.88×10^{-5}	--+++++++++----?--+	no
rs10514062	5	A/T	*SV2C*	0.0	0.636	49,942	8.99×10^{-6}	0.72	0.01	0.72	2.44×10^{-4}	----------+--+--?--+	no
rs3742467	14	T/C	*SOS2*	3.5	0.413	43,722	9.11×10^{-6}	0.88	0.44	0.89	3.97×10^{-5}	+++++-?+-+++-+++?++-	no
rs1324564	6	T/C	*RNF144B*	0.0	0.698	49,942	9.21×10^{-6}	0.78	0.33	0.81	4.97×10^{-5}	--+++++++++----?--+	no
rs1359339	6	T/G	*RNF144B*	0.0	0.775	49,942	9.77×10^{-6}	0.22	0.33	0.19	5.24×10^{-5}	++-++++++++-++++?++-	no

Chr.: chromosome; Pos.: position; Overall freq.: average effect allele frequency; In the column "direction", the studies are in the following order: 1. AGES, 2. ASPS, 3. ERF, 4. GHS, 5. H2000, 6. HBCS, 7. HRS, 8. KORA S4, 9. NFBC1966, 10. NTR1, 11. NTR2, 12. RS-I, 13. RS-II, 14. RS-III, 15. SardINIA, 16. SHIP, 17. THISEAS, 18. TwinsUK, 19. YFS, 20. STR; A question mark indicates that the SNP was not tested in that specific study.

Table A5. Replication results of the 22 suggestive SNPs ($p < 1 \times 10^{-5}$) from the self-employment discovery meta-analyses for males only.

Discovery meta-analysis									Swedish Twin Registry		Combined meta-analysis		
SNP	Chr.	Effect / non-effect allele	Nearest gene	I^2	Cochran's *Q* test *p*-value	*n*	*p*-value	Overall freq.	*p*-value	Freq.	*p*-value	Direction	Improvement?
rs6738407	2	A/G	*HECW2*	0.0	0.819	21,508	1.52×10^{-7}	0.20	0.26	0.20	7.93×10^{-8}	---------+------?--	yes
rs7558876	2	T/C	*HECW2*	0.0	0.676	21,508	2.02×10^{-7}	0.25	0.98	0.26	4.85×10^{-7}	---------+-----?-+	no
rs4850390	2	A/T	*HECW2*	0.0	0.807	21,508	2.18×10^{-7}	0.20	0.26	0.20	1.14×10^{-7}	---------+------?--	yes
rs12619045	2	A/G	*HECW2*	0.0	0.707	21,508	2.41×10^{-7}	0.75	0.97	0.74	5.81×10^{-7}	+++++++++-+++++?+-	no
rs17829296	2	A/G	*HECW2*	0.0	0.643	21,508	2.75×10^{-7}	0.80	0.77	0.81	9.04×10^{-7}	+++++++-++++++++?+-	no
rs1869794	2	T/G	*HECW2*	0.0	0.650	21,508	2.82×10^{-7}	0.20	0.81	0.19	8.72×10^{-7}	-------+--------?-+	no
rs4145361	2	T/C	*HECW2*	0.0	0.767	21,508	3.94×10^{-7}	0.80	0.27	0.80	2.12×10^{-7}	+++++++++-++++++?++	yes
rs11889112	2	A/G	*HECW2*	0.0	0.624	21,508	4.23×10^{-7}	0.20	0.81	0.19	1.27×10^{-6}	-------+--------?-+	no
rs11901261	2	A/C	*HECW2*	0.0	0.623	21,508	4.28×10^{-7}	0.80	0.81	0.81	1.28×10^{-6}	+++++++-++++++++?+-	no
rs11902488	2	A/G	*HECW2*	0.0	0.623	21,508	4.31×10^{-7}	0.80	0.81	0.81	1.29×10^{-6}	+++++++-++++++++?+-	no
rs10190368	2	A/G	*HECW2*	0.0	0.618	21,508	4.50×10^{-7}	0.80	0.87	0.81	1.22×10^{-6}	+++++++-++++++++?+-	no
rs6757097	2	T/C	*HECW2*	0.0	0.705	21,508	4.87×10^{-7}	0.25	0.95	0.26	1.16×10^{-6}	---------+-----?-+	no
rs2889155	2	T/G	*HECW2*	0.0	0.704	18,680	5.27×10^{-7}	0.82	0.61	0.82	6.10×10^{-7}	++++++?-++++++++?++	no
rs2124384	2	T/C	*HECW2*	0.0	0.646	19,786	5.49×10^{-7}	0.80	0.83	0.81	1.58×10^{-6}	+++++++-++++++?+?+-	no
rs12473229	2	T/G	*HECW2*	0.0	0.681	18,680	9.25×10^{-7}	0.75	0.94	0.74	2.16×10^{-6}	++++++?++-+++++?+-	no
rs2305567	2	A/G	*HECW2*	0.0	0.690	21,508	9.84×10^{-7}	0.75	0.94	0.74	2.30×10^{-6}	+++++++++-+++++?+-	no
rs6730618	2	T/C	*HECW2*	0.0	0.720	21,508	3.34×10^{-6}	0.78	0.92	0.79	5.89×10^{-6}	+-+++++++++-++++?++	no
rs6825440	4	A/T	*ODZ3*	24.1	0.175	21,508	4.25×10^{-6}	0.24	0.07	0.21	9.68×10^{-7}	-+--------+---+-?--	yes
rs7904494	10	A/T	*PRF1*	15.2	0.280	18,680	6.74×10^{-6}	0.78	0.32	0.80	3.79×10^{-5}	+-+---?--++-----?-+	no
rs1378626	2	T/C	*HECW2*	0.0	0.667	21,508	7.62×10^{-6}	0.21	0.54	0.20	7.10×10^{-6}	-+---------+----?--	yes
rs4867424	5	T/C	*MTMR12*	0.0	0.919	21,508	8.39×10^{-6}	0.49	0.83	0.50	2.00×10^{-5}	--+-------------?-+	no
rs2712008	4	T/G	*KLHL5*	0.0	0.815	18,680	9.94×10^{-6}	0.14	0.44	0.11	4.26×10^{-5}	+-++++?+++++-+++?+-	no

Chr.: chromosome; Pos.: position; Overall freq.: average effect allele frequency; In the column "direction", the studies are in the following order: 1. AGES, 2. ASPS, 3. ERF, 4. GHS, 5. H2000, 6. HBCS, 7. HRS, 8. KORA S4, 9. NFBC1966, 10. NTR1, 11. NTR2, 12. RS-I, 13. RS-II, 14. RS-III, 15. SardINIA, 16. SHIP, 17. THISEAS, 18. YFS, 19. STR; A question mark indicates that the SNP was not tested in that specific study.

Table A6. Replication results of the sixteen suggestive SNPs ($p < 1 \times 10^{-5}$) from the self-employment discovery meta-analyses for females only.

Discovery meta-analysis									Swedish Twin Registry		Combined meta-analysis		
SNP	Chr.	Effect / non-effect allele	Nearest gene	I^2	Cochran's *Q* test *p*-value	*n*	*p*-value	Overall freq.	*p*-value	Freq.	*p*-value	Direction	Improvement?
rs2331548	4	A/G	*CBR4*	0.0	0.913	24,319	1.93×10^{-6}	0.96	0.95	0.96	4.30×10^{-6}	??+?++++++++++?++-	no
rs521326	6	A/G	*GSTA4*	0.0	0.558	28,174	2.92×10^{-6}	0.61	0.48	0.63	1.33×10^{-5}	----------+---?--+	no
rs1022335	2	A/T	*ZEB2*	0.0	0.762	24,782	3.02×10^{-6}	0.37	0.48	0.36	1.37×10^{-5}	------?-----+---?--+	no
rs487642	6	A/G	*GSTA4*	0.0	0.557	28,174	3.05×10^{-6}	0.39	0.47	0.38	1.42×10^{-5}	++++++++++-+++?++-	no
rs575861	6	A/C	*GSTA4*	0.0	0.599	28,174	3.83×10^{-6}	0.61	0.50	0.63	1.63×10^{-5}	----------+---?--+	no
rs10753804	1	T/C	*SCYL1BP1*	0.0	0.848	24,782	3.92×10^{-6}	0.49	0.98	0.42	7.97×10^{-6}	------?----+----?--+	no
rs521664	6	T/C	*GSTA4*	0.0	0.593	28,174	3.95×10^{-6}	0.61	0.53	0.63	1.60×10^{-5}	----------+---?--+	no
rs594614	6	A/C	*GSTA4*	0.0	0.528	28,440	4.15×10^{-6}	0.39	0.48	0.38	1.83×10^{-5}	++++++++++-+++-++-	no
rs562487	5	A/G	*BHMT*	0.0	0.630	28,174	4.49×10^{-6}	0.48	0.50	0.50	4.01×10^{-6}	+++++-++-+-++-++?+++	yes
rs9557259	13	T/C	*TM9SF2*	0.0	0.592	12,506	5.16×10^{-6}	0.06	0.34	0.03	2.92×10^{-5}	??-?++?++++++?????+-	no
rs1383043	4	A/G	*ADAD1*	24.9	0.167	25,279	6.05×10^{-6}	0.38	0.71	0.35	1.78×10^{-5}	--+----+--------??++	no
rs9578700	13	A/G	*SPATA13*	38.9	0.047	28,174	6.53×10^{-6}	0.67	0.34	0.70	3.65×10^{-5}	-+++----------?-++	no
rs4481907	1	T/C	*SCYL1BP1*	0.0	0.835	28,174	7.88×10^{-6}	0.51	0.99	0.58	1.48×10^{-5}	++++++++++-++++?+++	no
rs947230	13	A/C	*SPATA13*	40.1	0.041	28,174	8.02×10^{-6}	0.67	0.33	0.70	4.39×10^{-5}	-+++----------?-++	no
rs7684834	4	A/G	*KIAA1109*	31.7	0.097	28,174	8.69×10^{-6}	0.62	0.82	0.65	1.27×10^{-5}	++-++++-++++++++?--+	no
rs557302	5	A/G	*BHMT2*	0.0	0.716	28,174	8.77×10^{-6}	0.49	0.69	0.49	1.06×10^{-5}	-----+--+-+-----?---	no

Chr.: chromosome; Pos.: position; Overall freq.: average effect allele frequency; In the column "direction", the studies are in the following order: 1. AGES, 2. ASPS, 3. ERF, 4. GHS, 5. H2000, 6. HBCS, 7. HRS. 8. KORA S4, 9. NFBC1966, 10. NTR1, 11. NTR2, 12. RS-I, 13. RS-II, 14. RS-III, 15. SardINIA, 16. SHIP, 17. THISEAS, 18. TwinsUK, 19. YFS, 20. STR; A question mark indicates that the SNP was not tested in that specific study.

Table A7. Gene-based *p*-values for the candidate entrepreneurship genes for pooled males and females, males only, and females only.

Gene	Pooled	Males	Females
ADORA2A	0.228	0.464	0.293
ADRA2A	0.007	0.011	0.183
COMT	0.528	0.999	0.073
DDC	0.334	0.758	0.604
DRD1	0.666	0.331	0.366
DRD2	0.749	0.843	0.786
DRD3	0.012	0.010	0.603
DRD4	0.483	0.366	0.221
DRD5	0.689	0.803	0.417
DYX1C1	0.384	0.164	0.347
HTR1B	0.892	0.975	0.511
HTR1E	0.953	0.518	0.597
HTR2A	0.079	0.030	0.685
KIAA0319 (DYX2)	0.324	0.477	0.419
ROBO1	0.554	0.692	0.435
SLC6A3 (DAT1)	0.679	0.645	0.627
SNAP25	0.118	0.209	0.888

Table A8. Gene-based *p*-values for the top 25 genes associated with self-employment in the discovery meta-analysis for pooled males and females.

Chr.	Gene	Number of SNPs	Start position	Stop position	*p*-value
11	*SLC15A3*	78	60,461,135	60,475,833	1.63×10^{-4}
11	*TMEM132A*	67	60,448,488	60,461,207	2.46×10^{-4}
11	*PRPF19*	78	60,414,777	60,430,632	2.61×10^{-4}
11	*TMEM109*	68	60,438,252	60,447,491	2.73×10^{-4}
11	*CD6*	109	60,495,690	60,544,424	3.05×10^{-4}
9	*FANCG*	65	35,063,834	35,070,013	3.15×10^{-4}
6	*FBXL4*	132	99,428,321	99,502,570	3.31×10^{-4}
9	*PIGO*	68	35,078,687	35,086,579	3.51×10^{-4}
7	*SLC26A5*	112	102,780,412	102,873,834	3.53×10^{-4}
11	*ZP1*	73	60,391,590	60,399,740	3.84×10^{-4}
9	*DNAJB5*	62	34,979,784	34,988,428	4.02×10^{-4}
9	*VCP*	68	35,046,064	35,062,739	4.03×10^{-4}
9	*C9orf131*	59	35,031,101	35,035,988	5.02×10^{-4}
9	*KIAA1539*	68	35,094,117	35,105,893	5.05×10^{-4}
9	*STOML2*	67	35,089,888	35,093,154	5.55×10^{-4}
14	*C14orf138*	90	49,645,099	49,653,047	5.63×10^{-4}
9	*KIAA1045*	76	34,948,191	34,972,541	6.03×10^{-4}
10	*SHOC2*	95	112,713,902	112,763,413	6.81×10^{-4}
5	*PRLR*	293	35,099,984	35,266,334	7.03×10^{-4}
6	*IHPK3*	205	33,797,420	33,822,660	7.24×10^{-4}
14	*LOC196913*	102	49,620,118	49,629,111	1.00×10^{-3}
14	*SOS2*	159	49,653,595	49,767,849	1.02×10^{-3}
6	*C6orf125*	181	33,773,323	33,787,482	1.05×10^{-3}
19	*TMEM190*	41	60,580,015	60,581,424	1.22×10^{-3}
15	*TMOD2*	138	49,831,101	49,889,635	1.26×10^{-3}

Table A9. Gene-based *p*-values for the top 25 genes associated with self-employment in the discovery meta-analysis for males only.

Chr.	Gene	Number of SNPs	Start position	Stop position	*p*-value
4	*TMEM156*	171	38,644,835	38,710,436	1.61×10^{-4}
4	*KLHL5*	146	38,723,053	38,800,224	3.21×10^{-4}
11	*SLCO2B1*	139	74,539,810	74,594,947	3.57×10^{-4}
5	*STARD4*	96	110,861,920	110,876,056	4.19×10^{-4}
8	*TMEM67*	50	94,836,268	94,899,523	4.48×10^{-4}
17	*TNFSF12-TNFSF13*	63	7,393,139	7,405,649	4.67×10^{-4}
1	*OR2M2*	89	246,409,910	246,410,954	4.71×10^{-4}
4	*ING2*	74	184,663,213	184,669,243	4.73×10^{-4}
17	*SAT2*	47	7,470,280	7,471,889	4.98×10^{-4}
17	*TNFSF12*	61	7,393,098	7,401,931	5.16×10^{-4}
17	*TNFSF13*	49	7,402,339	7,405,642	5.16×10^{-4}
17	*EIF4A1*	44	7,416,780	7,423,048	5.18×10^{-4}
17	*SENP3*	46	7,406,042	7,416,011	5.29×10^{-4}
1	*OR2M5*	75	246,375,072	246,376,011	6.39×10^{-4}
2	*HECW2*	496	196,772,221	197,165,580	7.41×10^{-4}
17	*CD68*	44	7,423,528	7,426,153	7.41×10^{-4}
17	*SHBG*	48	7,474,215	7,477,395	8.32×10^{-4}
1	*OR2M3*	98	246,432,992	246,433,931	8.35×10^{-4}
17	*SOX15*	46	7,432,221	7,434,212	9.16×10^{-4}
17	*MPDU1*	48	7,427,853	7,432,247	1.00×10^{-3}
2	*PAX3*	205	222,772,850	222,871,944	1.02×10^{-3}
8	*RBM12B*	35	94,812,903	94,822,400	1.07×10^{-3}
15	*TMOD2*	139	49,831,101	49,889,635	1.10×10^{-3}
17	*FXR2*	58	7,435,271	7,458,796	1.22×10^{-3}
19	*TMEM190*	42	60,580,015	60,581,424	1.28×10^{-3}

Table A10. Gene-based *p*-values for the top 25 genes associated with self-employment in the discovery meta-analysis for females only.

Chr.	Gene	Number of SNPs	Start position	Stop position	*p*-value
21	*PCP4*	227	40,161,216	40,223,192	4.70×10^{-5}
9	*MELK*	94	36,562,904	36,667,679	2.02×10^{-4}
2	*FLJ20160*	155	190,981,325	191,075,286	2.48×10^{-4}
5	*BHMT2*	107	78,401,338	78,421,031	2.66×10^{-4}
4	*ADAD1*	62	123,519,617	123,570,389	2.98×10^{-4}
4	*KIAA1109*	93	123,311,207	123,503,357	3.76×10^{-4}
4	*IL2*	47	123,592,075	123,597,100	3.81×10^{-4}
5	*BHMT*	74	78,443,359	78,463,869	4.26×10^{-4}
15	*CSPG4*	54	73,753,717	73,792,244	5.15×10^{-4}
4	*IL21*	106	123,753,232	123,761,661	5.46×10^{-4}
5	*ACTBL2*	128	56,811,599	56,814,393	6.98×10^{-4}
5	*ACTBL2*	128	56,811,599	56,814,393	7.33×10^{-4}
8	*PCMTD1*	194	52,892,692	52,974,288	7.53×10^{-4}
15	*SNX33*	36	73,728,402	73,738,023	8.43×10^{-4}
15	*ODF3L1*	37	73,803,373	73,807,082	9.80×10^{-4}
15	*ODF3L1*	37	73,803,373	73,807,082	9.90×10^{-4}
2	*HIBCH*	203	190,777,604	190,892,804	1.09×10^{-3}
2	*GKN2*	121	69,025,867	69,033,606	1.15×10^{-3}
22	*C22orf30*	74	30,402,241	30,438,731	1.16×10^{-3}
5	*C5orf35*	92	56,240,859	56,248,770	1.17×10^{-3}
6	*STX11*	120	144,513,346	144,554,769	1.17×10^{-3}
6	*GPX6*	59	28,579,051	28,591,549	1.32×10^{-3}
2	*INPP1*	110	190,916,440	190,944,636	1.34×10^{-3}
7	*SKAP2*	333	26,673,212	26,870,866	1.44×10^{-3}
2	*MGC13057*	140	190,710,730	190,776,455	1.58×10^{-3}

Table A11. Meta-analysis association results for SNP rs1486011 for pooled males and females, males only, and females only.

Sample	Effect / non-effect allele	EAF	*n*	*p*-value	Direction
Pooled	C/G	0.074	45,800	0.011	+++++++−+−−++−?+?−+
Males	C/G	0.077	19,786	0.046	+−++++−−+−−+−+?+?+
Females	C/G	0.072	25,754	0.112	+−−++++−−++++−?+?−+

EAF: average allele frequency; In the column "direction", the studies are in the following order: 1. AGES, 2. ASPS, 3. ERF, 4. GHS, 5. H2000, 6. HBCS, 7. HRS, 8. KORA S4, 9. NFBC1966, 10. NTR1, 11. NTR2, 12. RS-I, 13. RS-II, 14. RS-III, 15. SardINIA, 16. SHIP, 17. THISEAS, 18. TwinsUK (pooled and female sample) / YFS (male sample), 19. YFS (pooled and female sample); A question mark indicates that the SNP was not tested in that specific study.

Table A12. Results of the prediction analyses in STR for pooled males and females, males only, and females only.

	Pooled				Males				Females			
Threshold	**Coeff.**	***p*-value**	**R^2 (%)**	**AUC**	**Coeff.**	***p*-value**	**R^2 (%)**	**AUC**	**Coeff.**	***p*-value**	**R^2 (%)**	**AUC**
$p_T < 0.01$	−0.168	0.760	0.032%	0.510	−0.146	0.756	0.065%	0.516	−0.466	0.938	0.316%	0.531
$p_T < 0.05$	−0.004	0.525	0.000%	0.499	0.035	0.258	0.057%	0.504	−0.051	0.751	0.061%	0.520
$p_T < 0.1$	0.061	0.264	0.025%	0.513	0.071	0.200	0.095%	0.513	−0.021	0.573	0.004%	0.498
$p_T < 0.2$	0.100	0.254	0.028%	0.510	0.110	0.205	0.091%	0.515	−0.146	0.786	0.083%	0.516
$p_T < 0.3$	0.134	0.252	0.028%	0.512	0.093	0.298	0.038%	0.509	−0.185	0.773	0.074%	0.511
$p_T < 0.4$	0.219	0.187	0.050%	0.513	0.168	0.221	0.079%	0.515	−0.134	0.669	0.025%	0.507
$p_T < 0.5$	0.234	0.215	0.040%	0.512	0.195	0.228	0.074%	0.515	−0.087	0.595	0.008%	0.495
$p_T < 0.6$	0.029	0.204	0.044%	0.513	0.198	0.260	0.056%	0.510	−0.274	0.740	0.055%	0.507
$p_T < 0.7$	0.034	0.198	0.046%	0.514	0.180	0.305	0.035%	0.508	−0.331	0.749	0.060%	0.503
$p_T < 0.8$	0.038	0.199	0.045%	0.516	0.261	0.258	0.057%	0.510	−0.032	0.713	0.042%	0.501
$p_T < 0.9$	0.045	0.188	0.050%	0.516	0.289	0.260	0.055%	0.514	−0.033	0.699	0.036%	0.502
$p_T \leq 1.0$	0.049	0.190	0.049%	0.515	0.034	0.248	0.062%	0.514	−0.036	0.696	0.035%	0.501

Prediction results are based on a logistic regression of self-employment on the score controlling for the number of non-missing genotypes. Approximately 120,000 directly genotyped SNPs were used to calculate the score in STR for 12 overlapping significance thresholds. The R^2 is the Nagelkerke pseudo-R^2 from the logistic regression. The *p*-value indicates the significance of the score coefficient (one-sided Wald test, since the alternative hypothesis is that the score is positively correlated with self-employment). The AUC is the area under the receiver operating characteristic curve.

APPENDIX B

Supplementary Tables to Chapter 5

Table B1. Study design, sample size, sample quality control, and self-employment measure within each study.

Study				Sample QC			
Abbreviation	**Full name**	**Study design**	**Total sample size (*n*)**	**Call rate**	**Other exclusions**	**Sample size after QC (*n*)**	**Self-employment measure**
ERF	Erasmus Rucphen Family study	Family-based	3,500	≥98%	(1) Heterozygosity: FDR < 1%; (2) ethnic outliers; (3) duplicates; (4) gender mismatch; (5) excess IBS incompatible with pedigree; (6) missing serial self-employment status.	926	Individuals were asked if they were ever self-employed (no; yes—once; yes— more than once; always; not applicable/no response) and coded as serial self-employed if they were self-employed more than once. The control group consists of people who were never self-employed.
RS-I	Rotterdam Study Baseline	Population-based	7,983	≥97.5%	(1) Gender mismatch with typed X-linked markers; (2) excess autosomal heterozygosity > 0.336 ~ FDR > 0.1%; (3) duplicates and/or 1st or 2nd degree relatives using IBS probabilities > 97% from PLINK; (4) ethnic outliers using IBS distances > 3SD from PLINK; (5) missing serial self-employment status.	5,004	Detailed background information on the entire work-life history of all participants was available, including the number and duration of self-employment spells. Participants who had at least two spells in self-employment were coded as cases, those with zero spells as controls. Participants with one spell in self-employment or incomplete work-life histories were excluded.
STR	Swedish Twin Registry	Twin study	10,946	≥97%	(1) Sex-check (heterozygosity of X-chomosomes); (2) deviations in heterozygosity of more then 5 SD from the population mean; (3) cryptically relatedness check; (4) missing serial self-employment status.	2,771	Participants were asked "(1) have you ever run a business?" and "(2) how many businesses have you in total been part of starting?" Participants were coded as cases if they answered "Yes" to the first and ">1" to the second question. Controls were those that answered "No" to the first and "0" to the second question. Other participants were excluded.

Table B2. Genotyping, imputation, SNP quality control, and statistical analysis within each study.

	Genotyping				Imputation and quality control before meta-analysis				Association analysis	
	SNP inclusion criteria					SNP inclusion criteria				
Study	MAF	Call rate	HWE	Genotyped SNPs after QC	Imputation software	MAF	Imputation quality	Imputed SNPs after QC	Software	Covariates
ERF	≥ 1%	≥98%	$\geq 10^{-6}$	650,197	MACH	≥1%	Rsq≥0.4	2,174,485	ProbABEL	(1) Sex in pooled sample; (2) age dummies for the categories ≤ 29 (reference), 30–39, 40–49, ≥ 50.
RS-I	≥ 1%	≥98%	$\geq 10^{-6}$	512,349	MACH	≥1%	Rsq≥0.4	2,433,150	MACH2DAT	(1) Sex in pooled sample; (2) first four PCs.
STR	≥ 1%	≥97%	$\geq 10^{-7}$	644,556	IMPUTE	≥1%	info≥0.4	2,481,423	Merlin	(1) Sex in pooled sample; (2) first four PCs.

APPENDIX C

Details on the GREML Procedure

GREML partitions the observed variance in phenotypes into unobserved genetic effects and unobserved environmental effects (Visscher, Hill, & Wray, 2008) in a way that is analogous to the traditional *AE* model in twin studies, with the *C* component constrained to equal zero and assuming no correlation between *A* and *E*. But where twin studies assume genetic relatedness (1 for MZs and 0.5 DZs), GREML calculates a genetic relationship matrix (GRM) **A** based on SNP data. First, the elements of the *person* × *SNP* data matrix **X**[24] are scaled to:

(C1) $w_{ij} = (x_{ij} - 2p_j) / \sqrt{[2p_j(1-p_j)]}$,

for each individual $i = 1,\ldots, N$ and SNP $j = 1,\ldots, m$, with p_j denoting the prevalence of the reference allele, to obtain the scaled *person* × *SNP* data matrix **W**. The GRM **A** is obtained by the matrix multiplication $\mathbf{WW}'/m$, and contains the pairwise relationships between individuals at the SNPs. Estimates of relationships are always relative to a base population in which the average relationship is zero. Here, the base population is the set of all the individuals in the sample and due to the scaling of **W**, the average relationship between all pairs of individuals is 0 and the average relationship of an individual with oneself is 1.[25]

The GRM **A** is used in a variance-component analysis to quantify the amount of variance of a phenotype that can be explained by the SNPs. This variance-analysis is based on the linear model:

(C2) $\mathbf{y} = \mu\mathbf{1} + \mathbf{g} + \mathbf{e}$,

where the phenotype vector **y** is explained by the mean term μ (with **1** a $N \times 1$ vector of ones), additive genetic effect vector **g** and the residual vector $\mathbf{e} \sim N(\mathbf{0},\sigma_e^2\mathbf{1})$. The **e** vector can be interpreted as the environmental effect, which shows the equivalence with the AE twin models (i.e., there are no common environment effect in the model). We assume that the SNPs have random effects $\mathbf{u} \sim N(\mathbf{0},\sigma_u^2\mathbf{I})$, with **I** being a $m \times 1$ vector of ones. Then the $N \times 1$ total additive genetic effect vector $\mathbf{g} \sim N(\mathbf{0},\sigma_g^2\mathbf{1} = m\sigma_u^2\mathbf{1})$. This vector **g** can be calculated for all individuals by the matrix multiplication **Wu**. Now we take the variance of both sides of equation C2 assuming $\text{cov}(\mathbf{g},\mathbf{e}) = 0$ and see that the variance–covariance matrix of the phenotype **y** can be expressed as:

---

[24] Matrix **X** contains the SNP values of all persons in the sample. For each *person* × *SNP* combination it takes the value 0, 1, or 2 depending on the number of reference allele copies the person has for the specific SNP.

[25] This number can differ from 1 for an individual due to inbreeding. Inbreeding occurs when genetically related persons have children and results in increased homozygosity. On a population level these children have more homozygous SNPs than expected based on the population allele frequency.

(C3) $\text{var}(\mathbf{y}) = \text{var}(\mathbf{g}) + \text{var}(\mathbf{e}) = \text{var}(\mathbf{Wu}) + \text{var}(\mathbf{e}) = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}_{N \times N}$,

with $\mathbf{I}_{N \times N}$ the $N \times N$ identity matrix. The variance components can be estimated with restricted maximum likelihood (Yang et al., 2011). The proportion of variance in the phenotype that can be explained by SNPs (h^2_{SNPs}) is equal to σ_g^2 divided by var($\mathbf{y}$).

For binary traits such as serial entrepreneurship, this estimate has to be transformed to the underlying liability scale to make it independent of prevalence, to correct for ascertainment bias, and to facilitate comparability with the twin study estimates given above. The derivations of Dempster and Lerner (1950) and Lee et al. (2011) show that:

(C4) $h^2_{\text{SNPs}} = h^2_{\text{SNPs,binary_trait}}\,[K(1-K) / z^2] \times [K(1-K) / P(1-P)]$,

with K and P being the population and sample prevalence of the trait, respectively, and z being the height of the standard normal probability density function at the truncation threshold. People that exceed this threshold are considered as cases, otherwise as controls.

The significance of the σ_g^2 term in the model is tested with a log-likelihood ratio test. First, equation C3 is being estimated with σ_g^2 restricted to 0. The p-values are calculated assuming that the difference in the two log-likelihoods is asymptotically distributed as a 50:50 mixture of zero and a χ_1^2 under the null hypothesis (Lee et al., 2011, 2012).

APPENDIX D

Supplementary GWAS Results to Chapter 5

Table D1. Top SNPs females.

Discovery meta-analysis					STR		Combined meta-analysis		
SNP	Chr.	Avg. Freq.	*p*-value	Nearest gene	*p*-value	Freq.	*p*-value	Direc-tion	Im-provem ent?
rs3868899	3	0.75	4.98×10^{-6}	*ST6GAL1*	0.122	0.76	3.62×10^{-6}	−−−	yes
rs2542702	12	0.30	8.03×10^{-6}	*MED13L*	0.201	0.32	1.10×10^{-5}	+++	no
rs187232	20	0.02	8.12×10^{-7}	*SPTLC3*	0.627	0.02	1.50×10^{-5}	+++	no
rs441770	6	0.95	5.06×10^{-6}	*EGFL11*	0.642	0.96	5.99×10^{-5}	+−−	no
rs252170	12	0.17	5.71×10^{-6}	*NEDD1*	0.626	0.17	6.18×10^{-5}	+++	no
rs6416659	16	0.01	9.03×10^{-6}	*ERCC4*	0.636	0.02	8.95×10^{-5}	+++	no
rs12666194	7	0.98	6.31×10^{-6}	*NFE2L3*	0.741	0.98	9.69×10^{-5}	−−?	no
rs13134145	4	0.03	4.03×10^{-6}	*CENPC1*	0.863	0.03	1.02×10^{-4}	+++	no
rs12028458	1	0.74	5.48×10^{-6}	*EPHB2*	0.943	0.72	1.60×10^{-4}	−−−	no
rs1015283	8	0.70	8.27×10^{-6}	*CSMD1*	0.791	0.73	4.45×10^{-4}	−−+	no
rs3774790	3	0.86	9.22×10^{-6}	*ABHD5*	0.699	0.84	6.19×10^{-4}	−−+	no
rs8073791	17	0.09	4.58×10^{-6}	*HRNBP3*	0.059	0.07	7.15×10^{-3}	++−	no

Chr.: chromosome; Single genomic control used (Devlin & Roeder, 1999); The applied imputation accuracy thresholds for including SNPs are info > 0.40 (IMPUTE) and Rsq > 0.40 (MACH); In the column "direction", the studies are in the following order: 1. ERF ($\lambda = 1.025$), 2. RS-I ($\lambda = 0.999$), 3. STR ($\lambda = 1.007$), where λ is the genomic control parameter; A "?" indicates that this SNP was not available in the respective study; Only the SNP with the lowest *p*-value of each identified locus is listed here; SNPs are ordered by *p*-value in the combined meta-analysis.

Tables D1 and D2 show the results from the gender stratified GWAS meta-analyses. None of the suggestive loci in the gender-stratified analyses (12 for females, 10 for males) reaches $p < 0.05$ in the replication sample. Typically, the *p*-value in the combined meta-analysis is weaker than in the discovery meta-analysis, suggesting that most of the loci are false positives. Only the top hit in the analysis on females formally improves its *p*-value in the combined meta-analysis, but also fails to reach nominal significance in STR ($p = 0.122$).

In addition, we present additional graphical analyses based on the combined meta-analysis of the GWAS results from all three cohorts (ERF, RS-I, and STR). These graphical analyses have emerged as standard tools in the medical literature (McCarthy et al., 2008; Pearson & Manolio, 2008). They help to indicate whether the study has generated more significant results than expected by chance and to put such findings in context. Figures D1A–C display so-called quantile–quantile plots for males, females, and the pooled samples, respectively. The figures plot the ordered $-\log_{10}(p\text{-values})$ of SNPs of the fitted theoretical uniform distribution[26] on the *x*-axis to the observed data on the *y*-axis (Thode, 2002, p. 21). In the absence of significant findings, the points in the Q–Q plot will approximately

[26] Under the null hypothesis of no association it is expected that the *p*-values are uniformly distributed.

Table D2. Top SNPs males.

Discovery meta-analysis					STR		Combined meta-analysis		
SNP	Chr.	Avg. Freq.	*p*-value	Nearest gene	*p*-value	Freq.	*p*-value	Direc-tion	Im-provem ent?
rs11959890	5	0.80	1.74×10^{-7}	*DKFZp686 D0972*	0.092	0.80	6.59×10^{-4}	−−+	no
rs2734166	7	0.04	7.73×10^{-6}	*TRY1*	—	—	7.73×10^{-6}	++?	no
rs6795266	3	0.69	1.05×10^{-6}	*SIAH2*	0.897	0.70	3.64×10^{-5}	−−−	no
rs2116691	3	0.69	1.05×10^{-6}	*SIAH2*	0.900	0.70	3.67×10^{-5}	−−−	no
rs4479388	2	0.95	5.95×10^{-6}	*KCNJ3*	0.882	0.93	2.33×10^{-4}	−−+	no
rs10959372	9	0.04	7.57×10^{-6}	*PTPRD*	0.887	0.05	2.71×10^{-4}	++−	no
rs8001253	13	0.98	8.79×10^{-6}	*KL*	0.600	0.98	6.67×10^{-4}	−−+	no
rs2226332	21	0.95	1.74×10^{-6}	*APP*	0.881	0.95	1.00×10^{-4}	−−+	no
rs11150412	16	0.09	7.21×10^{-6}	*LOC283902*	0.377	0.10	1.20×10^{-3}	++−	no
rs2853585	1	0.13	5.16×10^{-6}	*HEATR1*	0.330	0.12	1.17×10^{-3}	++−	no

Chr.: chromosome; Single genomic control used (Devlin & Roeder, 1999); The applied imputation accuracy thresholds for including SNPs are info > 0.40 (IMPUTE) and Rsq > 0.40 (MACH); In the column "direction", the studies are in the following order: 1. ERF ($\lambda = 1.055$), 2. RS-I ($\lambda = 1.003$), 3. STR ($\lambda = 0.997$), where λ is the genomic control parameter; A "?" indicates that this SNP was not available in the respective study; Only the SNP with the lowest *p*-value of each identified locus is listed here; SNPs are ordered by *p*-value in the combined meta-analysis.

lie on the 45 degree line. Because the strongest associations have the smallest *p*-values (e.g., 10^{-6}), their $-\log_{10}$ values will be the greatest (e.g., 6). Upward divergence from the 45 degree line for higher *p*-values would indicate the presence of more statistically significant SNPs than would be expected by chance alone. Genome-wide significant findings have an observed $y > 7.3 = -\log_{10}(p\text{-value})$ (with $p = 5 \times 10^{-8}$), which would lie clearly above the 45 degree line. Downward divergence from the 45 degree line would indicate that the theoretical distribution of *p*-values is more dispersed than the observed distribution in the data, which could indicate misspecification of the estimated model or other data problems. The shaded area in the figure indicates the 95% confidence band.

The Q–Q plots show no downward deviation from a random distribution of *p*-values. Additionally, there is no evidence that the mean or median *p*-value over all SNPs would be inflated, which could result from artificial differences in allele frequencies due to population stratification, uncontrolled genetic relatedness among individuals, or genotyping errors (Yang et al., 2011). In summary, the Q–Q plots give no evidence of obvious data problems or model misspecifications.

The plot for the pooled analysis (Figure D1A) shows a slight upward divergence at $p \approx 10^{-5}$ and four isolated SNPs that reach genome-wide significance (including rs3774790, see Table 5.5). The plot for the analysis of females (Figure D1B) resembles a

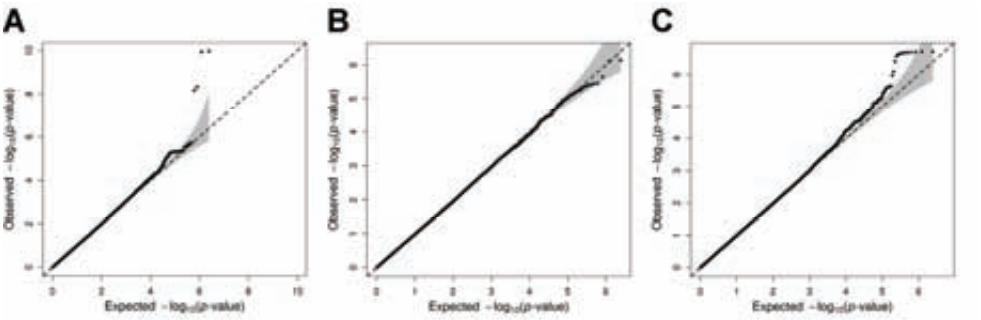

Figure D1. Q–Q plots of the serial self-employment combined meta-analyses. Q–Q plot of the serial self-employment combined meta-analysis using single genomic control (Devlin & Roeder, 1999) for (A) pooled males and females, (B) females only, and (C) males only. The grey shaded areas in the Q–Q plots represent the 95% confidence bands around the *p*-values.

purely random distribution of *p*-values. Either there is no association between SNPs and serial self-employment among females, or, more likely, the current sample size ($n_{\text{cases, female}} = 158$) is too small to detect it. The plot for males (Figure D1C) shows some upwards divergence starting at $p \approx 10^{-4}$, possibly indicating the presence of a number of SNPs with very small effect sizes that could be identified in a larger sample.

Figures D2A through D2C are so-called Manhattan plots that show the *p*-values, again on a $-\log_{10}$ scale, but with the genomic coordinates on the *x*-axis. Several regions in the Manhattan plots exhibit so-called "chimneys," i.e., genetic regions of a number of correlated SNPs that are jointly associated with serial self-employment and, therefore, stand out.

One genome-wide significant locus on chromosome 3 is shown in Figure D2A, which displays the meta-analysis results for females and males together. This association result comes from rs3774790, the top hit in Table 5.5, and its most closely associated neighboring SNPs, which are all located on the *ABHD5* gene. Noticeably, no other signal in this plot even remotely reaches genome-wide significance, although several loci reach $-\log_{10}(p\text{-value}) > 5$. Figure D2B (females) and Figure D2C (males) show several suggestive regions with $-\log_{10}(p\text{-value}) > 5$ on different chromosomes, often from single SNPs that have a much lower *p*-values than other SNPs in the same area. In Figure D2C (males), two regions on chromosome 3 and 4 reach *p*-values of 2.3×10^{-7} and 1.98×10^{-7}, respectively. The signal on chromosome 3 is the same one as in the pooled analysis and comes from the *ABHD5* gene, with rs3774790 again as the top hit. The top hit from the signal on chromosome 4 is rs13145299. This SNP is located in a non-coding region between the *TRCP3* and the *KIAA1109* gene, neither of which shows a direct association with serial self-employment. Very little is currently known about the biological function of this region.

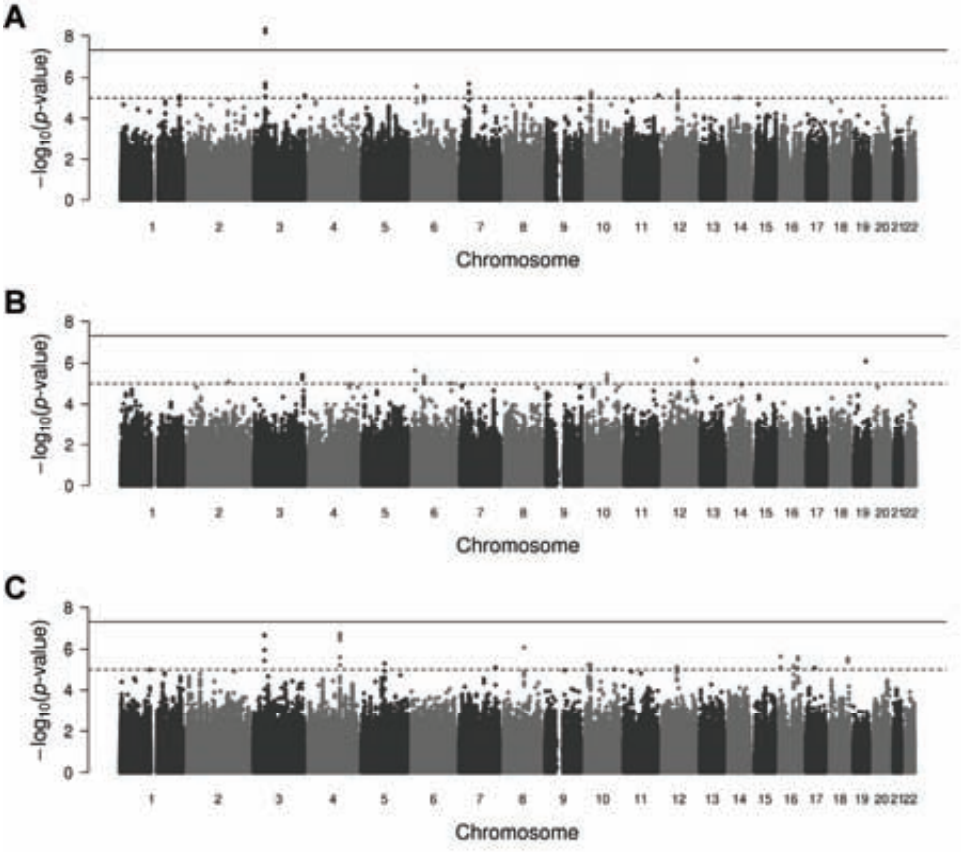

Figure D2. Manhattan plots of the serial self-employment combined meta-analyses. Manhattan plot of the serial self-employment combined meta-analysis using single genomic control (Devlin & Roeder, 1999) for (A) pooled males and females, (B) females only, and (C) males only. SNPs are plotted on the *x*-axis according to their position on each chromosome against association with self-employment on the *y*-axis shown as $-\log_{10}(p\text{-value})$. The solid line indicates the threshold for genome-wide significance ($p < 5 \times 10^{-8}$) and the dashed line the threshold for suggestive SNPs ($p < 1 \times 10^{-5}$).

Summary

Recent studies suggest that entrepreneurship is partly heritable, but are unable to pinpoint the specific genes involved. This thesis presents results from novel research aiming to identify genes associated with entrepreneurship using genetic data on the molecular level. In addition, the relationship between testosterone and entrepreneurship is examined based on the hypothesis that genes may exert their influence through this hormone.

The thesis begins with a discussion of candidate gene studies where genes are tested for association that are selected based upon an a priori hypothesis that is derived from information about their biological functioning. However, these studies often fail to replicate because of low statistical power (due to optimistic expectations about effect sizes) and/or publication bias. As an example, we show that a previously published association between the *DRD3* gene and entrepreneurship fails to replicate.

Next, the promises and limitations of the recently developed genome-wide association study (GWAS) design are discussed. This design enables a genome-wide, hypothesis-free search for associated genes. However, due to multiple testing, large sample sizes are needed in GWASs to differentiate between true and false positives. A simulation study shows that at least 30,000 participants are needed for a well-powered GWAS on entrepreneurship.

The following part first reports evidence that entrepreneurship, defined both as being self-employed at least once and serial self-employment, is partly heritable. Second, using a novel method from molecular genetics, evidence is presented showing that around half of the heritability is accounted for by actual molecular genetic data. Third, we perform GWASs on (serial) self-employment, but these fail to identify robustly associated genes. Fourth, prediction exercises show that it is currently impossible to predict (serial) self-employment solely from molecular genetic data.

In the final part, we show that, in contrast to earlier findings, testosterone is not associated with entrepreneurship.

Taken as a whole, the results suggest that entrepreneurship is likely to be influenced by hundreds if not thousands of genes with a very small effect size each, which has several implications for future research on the molecular genetics of entrepreneurship. First, the large part of the heritability of entrepreneurship that is captured by molecular genetic data suggests that this research is, in principle, feasible. Second, the results suggest that the effects of individual genes are likely to be very small, implying that very large sample sizes will be needed in future research to establish robust associations.

Most importantly, this thesis may serve as a practical guide for studying the molecular genetics of other economic variables. In conclusion, this thesis helps to build the foundations for a novel research field that integrates molecular genetics into economics.

Nederlandse Samenvatting (Summary in Dutch)

Recent onderzoek toont aan dat ondernemerschap deels erfelijk is, maar niet welke genen hierbij betrokken zijn. Dit proefschrift gaat op zoek naar deze genen door gebruik te maken van genetische data op moleculair niveau. Daarnaast wordt ingegaan op de relatie tussen testosteron en ondernemerschap op basis van de hypothese dat genen via dit hormoon invloed kunnen uitoefenen op ondernemerschap.

Het proefschrift begint met een beschrijving van kandidaatgenstudies waarin mogelijk gerelateerde genen worden bestudeerd die vooraf zijn geselecteerd op basis van kennis over hun biologische functie. Het blijkt echter dat dit soort studies vaak niet kan worden herhaald vanwege laag onderscheidend vermogen (wegens optimistische verwachtingen over effectgroottes) en/of het selectief publiceren van onderzoeksresultaten. Als voorbeeld laten we zien dat in onze data geen bewijs kan worden gevonden voor een eerder gepubliceerd verband tussen het *DRD3* gen en ondernemerschap.

Vervolgens worden de kansen en beperkingen van genoombreed associatieonderzoek besproken. Deze methode maakt, in tegenstelling tot kandidaatgenstudies, een hypothesevrije speurtocht door het hele menselijke genoom naar gerelateerde genen mogelijk. Door het grote aantal statistische toetsen zijn echter enorme steekproefgroottes nodig om statistisch significante verbanden te kunnen vinden. Een simulatiestudie toont aan dat voor genoombreed associatieonderzoek naar ondernemerschap een minimale steekproefgrootte van dertigduizend personen nodig is.

Het volgende deel laat ten eerste zien dat ondernemerschap, gedefinieerd zowel als zelfstandig als serieel ondernemerschap, deels erfelijk is. Ten tweede blijkt, door gebruik te maken van een nieuwe methode uit de moleculaire genetica, dat ongeveer de helft van de erfelijkheid kan worden verklaard door de momenteel beschikbare moleculair genetische data. Ten derde vinden we in genoombreed associatieonderzoek naar (serieel) ondernemerschap geen statistisch significante verbanden met genen. Ten vierde tonen we aan dat voorspellen van (serieel) ondernemerschap op basis van moleculair genetische data op dit moment niet mogelijk is.

Het laatste deel laat zien dat, in tegenstelling tot een eerdere studie, er geen statistisch significant verband is tussen ondernemerschap en testosteron.

Alle bevindingen samen suggereren dat ondernemerschap zeer waarschijnlijk het resultaat is van honderden zo niet duizenden genen met ieder een zeer klein effect. Deze conclusie heeft een aantal gevolgen voor toekomstig onderzoek naar de moleculaire genetica van ondernemerschap. Ten eerste suggereert het grote deel van erfelijkheid dat door de

momenteel beschikbare moleculair genetische data wordt verklaard dat het vinden van genen voor ondernemerschap in principe mogelijk is. Ten tweede laten de bevindingen zien dat de effecten van individuele genen zeer klein zijn wat tot gevolg heeft dat enorme steekproefgroottes nodig zijn in toekomstig onderzoek om aan te kunnen tonen dat verbanden statistisch significant zijn.

De voornaamste bijdrage van dit proefschrift is echter dat het als leidraad dient voor onderzoek naar de moleculaire genetica van andere economische variabelen. Concluderend kunnen we stellen dat dit proefschrift de fundamenten legt voor een nieuw onderzoeksgebied dat moleculaire genetica in de economie integreert.

References

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, S. L. (1994). Socioeconomic status and health: The challenge of the gradient. *American Psychologist*, *49*, 15–24.

Adler, N. E., & Ostrove, J. M. (1999). Socioeconomic status and health: What we know and what we don't. *Annals of the New York Academy of Sciences*, *896*, 3–15.

Ali, A., Allen, I. E., Brush, C., Bygrave, W. D., De Castro, J., Lange, J., … Suhu, A. (2008). *What entrepreneurs are up to: 2008 National entrepreneurial assessment for the United States of America*. Wellesley, MA, USA: Babson College.

Andersson, L., & Hammarstedt, M. (2010). Intergenerational transmissions in immigrant self-employment: Evidence from three generations. *Small Business Economics*, *34*, 261–276.

Ansolabehere, S., Rodden, J., & Snyder, J. M. (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, *102*, 215–232.

Apicella, C. L., Cesarini, D., Johannesson, M., Dawes, C. T., Lichtenstein, P., Wallace, B., … Westberg, L. (2010). No association between oxytocin receptor (OXTR) gene polymorphisms and experimentally elicited social preferences. *PLOS ONE*, *5*(6), e11153.

Apicella, C. L., Dreber, A., Campbell, B., Gray, P. B., Hoffman, M., & Little, A. C. (2008). Testosterone and financial risk preferences. *Evolution and Human Behavior*, *29*, 384–390.

Argyle, M. (1997). Is happiness a cause of health? *Psychology & Health*, *12*, 769–781.

Arvey, R. D., Bouchard, T. J., Jr., Segal, N. L., & Abraham, L. M. (1989). Job satisfaction: Environmental and genetic components. *Journal of Applied Psychology*, *74*, 187–192.

Audretsch, D., & Keilbach, M. (2004). Entrepreneurship capital and economic performance. *Regional Studies*, *38*, 949–959.

Bates, T. (2002). Restricted access to markets characterizes women-owned businesses. *Journal of Business Venturing*, *17*, 313–324.

Beauchamp, J. P., Cesarini, D., Johannesson, M., Van der Loos, M. J. H. M., Koellinger, P. D., Groenen, P. J. F., … Christakis, N. A. (2011). Molecular genetics and economics. *Journal of Economic Perspectives*, *25*, 57–82.

Becker, K., El-Faddagh, M., Schmidt, M. H., & Laucht, M. (2007). Is the serotonin transporter polymorphism (5-HTTLPR) associated with harm avoidance and internalising problems in childhood and adolescence? *Journal of Neural Transmission*, *114*, 395–402.

Behrman, J., & Taubman, P. (1976). Intergenerational transmission of income and wealth. *American Economic Review*, *66*, 436–440.

Benjamin, D. J., Cesarini, D., Chabris, C. F., Glaeser, E. L., Laibson, D. I., Guðnason, V., … Lichtenstein, P. (2012). The promises and pitfalls of genoeconomics. *Annual Review of Economics*, *4*, 627–662.

Benjamin, D. J., Cesarini, D., Van der Loos, M. J. H. M., Dawes, C. T., Koellinger, P. D., Magnusson, P. K. E., … Visscher, P. M. (2012). The molecular genetic architecture of economic and political preferences. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 8026–8031.

Benz, M., & Frey, B. S. (2008). Being independent is a great thing: Subjective evaluations of self-employment and hierarchy. *Economica*, *75*, 362–383.

Betsworth, D. G., Bouchard, T. J, Jr., Cooper, C. R., Grotevant, H. D., Hansen, J.-I. C. Scarr, S., & Weinberg, R. A. (1994). Genetic and environmental influences on vocational interests assessed using adoptive and biological families and twins reared together and apart. *Journal of Vocational Behavior*, *44*, 263–278.

Bird, B., & Brush, C. (2002). A gendered perspective on organizational creation. *Entrepreneurship Theory and Practice*, *26*, 41–65.

Björklund, A., Jäntti, M., & Solon, G. (2007). Nature and nurture in the intergenerational transmission of socioeconomic status: Evidence from Swedish children and their biological and rearing parents. *B.E. Journal of Economic Analysis & Policy*, *7*(2), article 4.

Blanchflower, D. G., & Oswald, A. J. (1998). What makes an entrepreneur? *Journal of Labor Economics*, *16*, 26–60.

Block, J., & Koellinger, P. (2009). I can't get no satisfaction—Necessity entrepreneurship and procedural utility. *Kyklos*, *62*, 191–209.

Block, J. H., Hoogerheide, L., & Thurik, A. R. (2013). Education and entrepreneurial choice: An instrumental variables analysis. *International Small Business Journal*, *31*, 23–33.

Bowles, S., & Gintis, H. (2002). The inheritance of inequality. *Journal of Economic Perspectives*, *16*, 3–30.

Brañas-Garza, P., & Rustichini, A. (2011). Organizing effects of testosterone and economic behavior: Not just risk taking. *PLOS ONE*, *6*(12), e29842.

Bruder, C. E. G., Piotrowski, A., Gijsbers, A. A. C. J., Andersson, R., Erickson, S., Díaz de Ståhl, T., … Dumanski, J. P. (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *American Journal of Human Genetics*, *82*, 763–771.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., … Worthington, J. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*, 661–678.

Bygrave, W. D., & Hofer, C. W. (1991). Theorizing about entrepreneurship. *Entrepreneurship Theory and Practice*, *16*, 13–22.

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*, *89*, 306–318.

Caporaso, N., Gu, F., Chatterjee, N., Sheng-Chih, J., Yu, K., Yeager, M., … Bergen, A. W. (2009). Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLOS ONE*, *4*(2), e4653.

Carree, M., & Thurik, A. R. (2006). *Entrepreneurship and economic growth*. Cheltenham, UK: Edward Elgar.

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1996). *The history and geography of human genes*. Princeton, NJ, USA: Princeton University Press.

Cesarini, D., Dawes, C. T., Johannesson, M., Lichtenstein, P., & Wallace, B. (2009). Genetic variation in preferences for giving and risk taking. *Quarterly Journal of Economic*s, *124*, 809–842.

Cesarini, D., Johannesson, M., Lichtenstein, P., Sandewall, Ö., & Wallace, B. (2010). Genetic variation in financial decision making. *Journal of Finance*, *65*, 1725–1754.

Cesarini, D., Johannesson, M., Magnusson, P. K. E., & Wallace, B. (2012). The behavioral genetics of behavioral anomalies. *Management Science*, *58*, 21–34.

Cesarini, D., Lichtenstein, P., Johannesson, M., & Wallace, B. (2009). Heritability of overconfidence. *Journal of the European Economic Association*, *7*, 617–627.

Chabris, C. F., Hebert, B. M., Benjamin, D. J. Beauchamp, J. P., Cesarini, D., Van der Loos, M. J. H. M., … Laibson, D. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, *23*, 1314–1323.

Charney, E. (2008). Genes and ideologies. *Perspectives on Politics*, *6*, 299–319.

Chaturvedi, S., Zyphur, M. J., Arvey, R. D., Avolio, B. J., & Larsson, G. (2012). The heritability of emergent leadership: Age and gender as moderating factors. *Leadership Quarterly*, *23*, 219–232.

Cipolla, C. M. (1993). *Before the industrial revolution: European society and economy. 1000–1700* (3rd ed.). London, UK: Routledge.

Coates, J. M., & Herbert, J. (2008). Endogenous steroids and financial risk taking on a London trading floor. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6167–6172.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from large-scale biology. *Science*, *300*, 286–290.

Colombier, N., & Masclet, D. (2008). Intergenerational correlation in self employment: Some further evidence from French ECHP data. *Small Business Economics*, *30*, 423–437.

Comings, D. E., Rosenthal, R. J., Lesieur, H. R., Rugle, L. J., Muhleman, D., Chiu, C., … Gade, R. (1996). A study of the dopamine D2 receptor gene in pathological gambling. *Pharmacogenetics*, *6*, 223–234.

Cooper, C. L., & Marshall, J. (1976). Occupational sources of stress: A review of the literature relating to coronary heart disease and mental ill health. *Journal of Occupational Psychology*, *49*, 11–28.

Cooper, C. L., & Smith, M. (1985). *Job Stress and Blue Collar Work*. Chichester, UK: Wiley.

Cowling, M., & Taylor, M. (2001). Entrepreneurial women and men: Two different species? *Small Business Economics*, *16*, 167–175.

Cramer, J. S., Hartog, J., Jonker, N., & Van Praag, C. M. (2002). Low risk aversion encourages the choice for entrepreneurship: An empirical test of a truism. *Journal of Economic Behavior and Organization*, *48*, 29–36.

Crisp, D. J., Beaumont, A. R., Flowerdew, M. W., & Vardy, A. (1978). The Hardy-Weinberg test—A correction. *Marine Biology*, *46*, 181–183.

Dai, W. S., Gutai, J. P., Kuller, L. H., & Cauley, J. A. (1988). Cigarette smoking and serum sex hormones in men. *American Journal of Epidemiology*, *128*, 796–805.

Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian Randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, *32*, 1–22.

Davidsson, P. (2005). *Researching Entrepreneurship*. New York, NY, USA: Springer.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., … Deary, I. J. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*, *16*, 996–1005.

Deary, I. J., Spinath, F. M., & Bates, T. C. (2006). Genetics of intelligence. *European Journal of Human Genetics*, *14*, 690–700.

De Jager, P. L., Jia, X., Wang, J., De Bakker, P. I. W., Ottoboni, L., Aggarwal, N. T., … Oksenberg, J. R. (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature Genetics*, *41*, 776–782.

De Moor, M. H. M., Costa, P. T., Terracciano, A., Krueger, R. F., De Geus, E. J. C., Toshiko, T., … Boomsma, D. I. (2012). Meta-analysis of genome-wide association studies for personality. *Molecular Psychiatry*, *17*, 337–349.

Dempster, E. R., & Lerner, I. M. (1950). Heritability of threshold characters. *Genetics*, *35*, 212–236.

De Ronde, W., Van der Schouw, Y. T., Muller, M., Grobbee, D. E., Gooren, L. J., Pols, H. A. P., & De Jong, F. H. (2005). Associations of sex-hormone-binding globulin (SHBG) with non-SHBG-bound levels of testosterone and estradiol in independently living men. *Journal of Clinical Endocrinology & Metabolism*, *90*, 157–62.

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*, 997–1004.

Diver, M. J., Imtiaz, K. E., Ahmad, A. M., Vora, J. P., & Fraser, W. D. (2003). Diurnal rhythms of serum total, free and bioavailable testosterone and of SHBG in middle-aged men compared with those in young men. *Clinical Endocrinology*, *58*, 710–717.

Dowd, J. B., Albright, J., Raghunathan, T. E., Schoeni, R. F., LeClere, F., & Kaplan, G. A. (2011). Deeper and wider: Income and mortality in the USA over three decades. *International Journal of Epidemiology*, *40*, 183–188.

Drucker, P. F. (1985). *Innovation and entrepreneurship*. New York, NY, USA: Harper & Row.

Dunn, T., & Holtz-Eakin, D. (2000). Financial capital, human capital, and the transition to self-employment: Evidence from intergenerational links. *Journal of Labor Economics*, *18*, 282–305.

Du Rietz, A., & Henrekson, M. (2000). Testing the female underperformance hypothesis. *Small Business Economics*, *14*, 1–10.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., … Ponder, B. A. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, *447*, 1087–1093.

Ebstein, R. P., Israel, S., Chew, S. H., Zhong, S., & Knafo, A. (2010). Genetics of human social behavior. *Neuron*, *65*, 831–844.

Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., … Belmaker, R. H. (1996). Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of novelty seeking. *Nature Genetics*, *12*, 78–80.

Emre, S., Ünver, N., Ersoy-Evans, S., Yüzbaşıoğlu, A., Gürakan, F., Gümrük, F., & Karaduman, A. (2010). Molecular analysis of Chanarin-Dorfman syndrome (CDS) patients: Identification of novel mutations in the ABHD5 gene. *European Journal of Medical Genetics*, *53*, 141–144.

Estivill, X., & Armengol, L. (2007). Copy number variants and common disorders: Filling the gaps and exploring complexity in genome-wide association studies. *PLOS Genetics*, *3*(10), e190.

Estrada, K., Abuseiris, A., Grosveld, F. G., Uitterlinden, A. G., Knoch, T. A., & Rivadeneira, F. (2009). GRIMP: A web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data. *Bioinformatics*, *25*, 2750–2752.

Ettner, S. L. (1996). New evidence on the relationship between income and health. *Journal of Health Economics*, *15*, 67–85.

Evans, D. S., & Leighton, L. S. (1989). Some empirical aspects of entrepreneurship. *American Economic Review*, *79*, 519–535.

Falconer, D. S. (1960). *Introduction to quantitative genetics*. New York, NY, USA: Ronald Press.

Feldman, H. A., Longcope, C., Derby, C. A., Johannes, C. B., Araujo, A. B., Coviello, A. D., … McKinlay, J. B. (2002). Age trends in the level of serum testosterone and other hormones in middle-aged men: Longitudinal results from the Massachusetts Male Aging Study. *Journal of Clinical Endocrinology & Metabolism*, *87*, 589–598.

Field, A. E., Colditz, G. A., Willett, W. C., Longcope, C., & McKinlay, J. B. (1994). The relation of smoking, age, relative weight and dietary intake to serum adrenal steroids, sex hormones and sex hormone-binding globulin in middle-aged men. *Journal of Clinical Endocrinology & Metabolism*, *79*, 1310–1316.

Finley, M. I. (1973). *The ancient economy*. London, UK: Chatto & Windus.

Folland, J. P., Mc Cauley, T. M., Phypers, C., Hanson, B., & Mastana, S. S. (2012). Relationship of 2D:4D finger ratio with muscle strength, testosterone, and androgen receptor CAG repeat genotype. *American Journal of Physical Anthropology*, *148*, 81–87.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., … Lee, C. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, *16*, 949–961.

Freytag, A. & Thurik, A. R. (2007). Entrepreneurship and its determinants in a cross-country setting, *Journal of Evolutionary Economics*, *17*, 117–131.

Friedrich, N., Völzke, H., Rosskopf, D., Steveling, A., Krebs, A., Nauck, M., & Wallaschofski, H. (2008). Reference ranges for serum dehydroepiandrosterone sulfate and testosterone in adult men. *Journal of Andrology*, *29*, 610–617.

Fritsch, M. (2004). Entrepreneurship, entry and performance of new business compared in two growth regimes: East and West Germany. *Journal of Evolutionary Economics*, *14*, 525–542.

Garner, C. (2007). Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology*, *31*, 288–295.

Gartner, W. B. (1988). "Who is an entrepreneur?" Is the wrong question. *American Journal of Small Business*, *12*, 11–32.

Georgellis, Y., & Wall, H. J. (2005). Gender differences in self-employment. *International Review of Applied Economics*, *19*, 321–342.

Goddard, M. E., Wray, N. R., Verbyla, K., & Visscher, P. M. (2009). Estimating effects and making predictions from genome-wide marker data. *Statistical Science*, *24*, 517–529.

Goudriaan, A. E., Lapauw, B., Ruige, J., Feyen, E., Kaufman, J.-M., Brand, M., & Vingerhoets, G. (2010). The influence of high-normal testosterone levels on risk-taking in healthy males in a 1-week letrozole administration study. *Psychoneuroendocrinology*, *35*, 1416–1421.

Granger, D. A., Shirtcliff, E. A., Booth, A., Kivlighan, K. T., & Schwartz, E. B. (2004). The "trouble" with salivary testosterone. *Psychoneuroendocrinology*, *29*, 1229–1240.

Grilo, I., & Thurik, R. (2008). Determinants of entrepreneurial engagement levels in Europe and the US. *Industrial and Corporate Change*, *17*, 1113–1145.

Grilo, I., Thurik, R., Verheul, I., & Van der Zwan, P. (2007). *Climbing the entrepreneurial ladder: The role of gender* (ERIM Report Series Research in Management ERS-2007-098-ORG). Rotterdam, The Netherlands: Erasmus Research Institute of Management.

Groenen, P. J. F., Hofman, A., Koellinger, P. D., Van der Loos, M. J. H. M., Rivadeneira, F., Van Rooij, F. J. A., … Uitterlinden, A. (2008). Genome-wide association for loci influencing entrepreneurial behavior: The Rotterdam Study. *Behavior Genetics*, *38*, 628–629.

Guedj, M., Nuel, G., & Prum, B. (2008). A note on allelic tests in case-control association studies. *Annals of Human Genetics*, *72*, 407–409.

Guiso, L., & Rustichini, A. (2011). *Understanding the size and profitability of firms: The role of a biological factor* (EUI Working Paper ECO 2011/01). Retrieved from European University Institute website: http://hdl.handle.net/1814/15642

Hak, A. E., Witteman, J. C. M., De Jong, F. H., Geerlings, M. I., Hofman, A., & Pols, H. A. P. (2002). Low levels of endogenous androgens increase the risk of atherosclerosis in elderly men: The Rotterdam Study. *Journal of Clinical Endocrinology & Metabolism*, *87*, 3632–3639.

Halperin, E., & Stephan, D. A. (2009). SNP imputation in association studies. *Nature Biotechnology*, *27*, 349–351.

Hamilton, B. H. (2000). Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political Economy*, *108*, 604–631.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, *28*, 49–50.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., … Williams, J. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics*, *41*, 1088–1093.

Hartley, J. E. (1996). Retrospectives: The origins of the representative agent. *Journal of Economic Perspectives*, *10,* 169–177.

Haynes, S. G., & Feinleib, M. (1980). Women, work and coronary heart disease: Prospective findings from the Framingham Heart Study. *American Journal of Public Health*, *70*, 133–141.

Henneman, P., Aulchenko, Y. S., Frants, R. R., Van Dijk, K. W., Oostra, B. A, & Van Duijn, C. M. (2008). Prevalence and heritability of the metabolic syndrome and its individual components in a Dutch isolate: The Erasmus Rucphen Family study. *Journal of Medical Genetics*, *45*, 572–577.

Herskind, A. M., McGue, M., Holm, N. V., Sørensen, T. I., Harvald, B., & Vaupel, J. W. (1996). The heritability of human longevity: A population-based study of 2872 Danish twin pairs born 1870–1900. *Human Genetics*, *97*, 319–323.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*, 557–560.

Hill, W. G., & Weir, B. S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, *93*, 47–64.

Hindorff, L. A., MacArthur, J., Morales, J., Junkins, H. A., Hall, P. N., Klemm, A. K., & Manolio, T. A. (2012, November 6). A catalog of published genome-wide association studies. Retrieved from www.genome.gov/gwastudies

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 9362–9367.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, *4*, 45–61.

Hofman, A., Breteler, M. M. B., Van Duijn, C. M., Janssen, H. L. A., Krestin, G. P., Kuipers, E. J., … Witteman, J. C. M. (2009). The Rotterdam Study: 2010 Objectives and design update. *European Journal of Epidemiology*, *24*, 553–572.

Hofman, A., Grobbee, D. E., De Jong, P. T. V. M., & Van den Ouweland, F. A. (1991). Determinants of disease and disability in the elderly: The Rotterdam Elderly Study. *European Journal of Epidemiology*, *7*, 403–422.

Hofman, A., Van Duijn, C. M., Franco, O. H., Ikram, M. A., Janssen, H. L. A., Klaver, C. C. W., … Witteman, J. C. M. (2011). The Rotterdam Study: 2012 Objectives and design update. *European Journal of Epidemiology*, *26*, 657–686.

Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C., & Balding, D. J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, *32*, 179–185.

Hu, N., Wang, C., Hu, Y., Yang, H. H., Giffen, C., Tang, Z. Z., … Lee, M. P. (2005). Genome-wide association study in esophageal cancer using GeneChip mapping 10K array. *Cancer Research*, *65*, 2542–2546.

Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., & Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *American Journal of Human Genetics*, *84*, 235–250.

Hyytinen, A., & Ilmakunnas, P. (2007). What distinguishes a serial entrepreneur? *Industrial and Corporate Change*, *16*, 793–821.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124.

Ioannidis, J. P. A., Tarone, R., & McLaughlin, J. K. (2011). The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*, *22*, 450–456.

Israel, S., Lerer, E., Shalev, I., Uzefovsky, F. Reibold, M., Bachner-Melman, R., … Ebstein, R. P. (2008). Molecular genetic studies of the arginine vasopressin 1a receptor (AVPR1a) and the oxytocin receptor (OXTR) in human behaviour: From autism to altruism with some notes in between. *Progress in Brain Research*, *170*, 435–449.

Israel, S., Lerer, E., Shalev, I., Uzefovsky, F. Reibold, M., Laiba, E., … Ebstein, R. P. (2009). The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task. *PLOS ONE*, *4*(5), e5535.

Iyigun, M. F., & Owen, A. L. (1998). Risk, entrepreneurship, and human capital accumulation. *American Economic Review*, *88*, 454–457.

Jovanovic, B. (1994). Firm formation with heterogeneous management and labor skills. *Small Business Economics*, *6*, 185–191.

Kaplan, G. A., & Keil, J. E. (1993). Socioeconomic factors and cardiovascular disease: A review of the literature. *Circulation*, *88*, 1973–1998.

Keller, L. M., Bouchard, T. J., Jr., Arvey, R. D., Segal, N. L., & Dawes, R. V. (1992). Work values: Genetic and environmental influences. *Journal of Applied Psychology*, *77*, 79–88.

Kempthorne, O. (1997). Heritability: Uses and abuses. *Genetica*, *99*, 109–112.

Kihlstrom, R. E., & Laffont, J.-J. (1979). A general equilibrium entrepreneurial theory of the firm based on risk aversion. *Journal of Political Economy*, *87*, 719–748.

Kirman, A. P. (1992). Whom or what does the representative individual represent? *Journal of Economic Perspectives*, *6*, 117–136.

Kirzner, I. M. (1973). *Competition and entrepreneurship*. Chicago, IL, USA: The University of Chicago Press.

Kluger, A. N., Siegfried, Z., & Ebstein, R. P. (2002). A meta-analysis of the association between DRD4 polymorphism and novelty seeking. *Molecular Psychiatry*, *7*, 712–717.

Knafo, A., Israel, S., Darvasi, A., Bachner-Melman, R., Uzefovsky, F., Cohen, L., … Ebstein, R. P. (2008). Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. *Genes, Brain and Behavior*, *7*, 266–275.

Knight, F. H. (1921). *Risk, uncertainty, and profit*. Boston, MA, USA: Houghton Mifflin Company.

Koch, B., Glaser, S., Schaper, C., Krebs, A., Nauck, M., Dorr, M., … Friedrich, N. (2011). Association between serum testosterone and sex hormone-binding globulin and exercise capacity in men: Results of the Study of Health in Pomerania (SHIP). *Journal of Andrology*, *32*, 135–143.

Koellinger, P. (2008). Why are some entrepreneurship more innovative than others? *Small Business Economics*, *31*, 21–37.

Koellinger, P., Minniti, M., & Schade, C. (2007). "I think I can, I think, I can": Overconfidence and entrepreneurial behaviour. *Journal of Economic Psychology*, *28*, 502–527.

Koellinger, P., Minniti, M., & Schade, C. (2013). Gender differences in entrepreneurial propensity. *Oxford Bulletin of Economics and Statistics*, *75*, 213–234.

Koellinger, P. D., & Thurik, A. R. (2012). Entrepreneurship and the business cycle. *Review of Economics and Statistics*, *94*, 1143–1156.

Koellinger, P. D., Van der Loos, M. J. H. M., Groenen, P. J. F., Thurik, A. R., Rivadeneira, F., Van Rooij, F. J. A., … Hofman, A. (2010). Genome-wide association studies in economics and entrepreneurship research: Promises and limitations. *Small Business Economics*, *35*, 1–18.

Koellinger, P. D., Van der Loos, M. J. H. M., Rietveld, C. A., Benjamin, D. J., Cesarini, D., Eklund, N., … Thurik, A. R. (2012). *The molecular genetics of serial self-employment*. Manuscript in preparation.

Koppl, R. (2007). Entrepreneurial behavior as a human universal. In M. Minniti (Ed.), *Entrepreneurship: The engine of growth* (Vol.1, pp. 1–20). Westport, CT, USA: Praeger.

Lager, A. C. J., & Torssander, J. (2012). Causal effect of education on mortality in a quasi-experiment on 1.2 million Swedes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 8461–8466.

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, *470* 187–197.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, *265*, 2037–2048.

Lang, U. E., Bajbouj, M., Wernicke, C., Rommelspacher, H., Danker-Hopfe, H., & Gallinat, J. (2004). No association of a functional polymorphism in the serotonin transporter gene promoter and anxiety-related personality traits. *Neuropsychobiology*, *49*, 182–184.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., … Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*, 832–838.

Laussel, D., & Le Breton, M. (1995). A general equilibrium theory of firm formation based on individual unobservable skills. *European Economic Review*, *39*, 1303–1319.

Lazear, E. P. (2004). Balanced skills and entrepreneurship. *American Economic Review*, *94*, 208–211.

Lazear, E. P. (2005). Entrepreneurship. *Journal of Labor Economics*, *23*, 649–680.

Lee, S. H., Decandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., … Wray, N. R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, *44*, 247–250.

Lee, S. H., Wray, N. R., Goddard, M. E., & Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics*, *88*, 294–305.

Lentz, B. F., & Laband, D. N. (1990). Entrepreneurial success and occupational inheritance among proprietors. *Canadian Journal of Economics*, *23*, 563–579.

Lesch, K.-P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B. D., Petri, S., … Murphy, D. L. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science*, *274*, 1527–1531.

Levie, J. (2007). Immigration, in-migration, ethnicity and entrepreneurship in the United Kingdom. *Small Business Economics*, *28*, 143–169.

Lewin-Epstein, N., & Yuchtman-Yaar, E. (1991). Health risks of self-employment. *Work and Occupations*, *18*, 291–312.

Li, Y., & Abecasis, G. R. (2006). Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics*, *S79*, 2290.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2006). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*, 816–834.

Li, Y., Willer, C. J., Sanna, S., & Abecasis, G. R. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, *10*, 387–406.

Lichtenstein, P., Pedersen, N. L., & McClearn, G. (1992). The origins of individual differences in occupational status and educational level. *Acta Sociologica*, *35*, 13–31.

Lichtenstein, P. Sullivan, P. F., Cnattingius, S., Gatz, M., Johansson, S., Carlström, E., … Pedersen, N. L. (2006). The Swedish Twin Registry in the third millennium: An update. *Twin Research and Human Genetics*, *9*, 875–882.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., … Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics*, *87*, 139–145.

Loomis, J. B. (1989). Test-retest reliability of the contingent valuation method: A comparison of general population and visitor responses. *American Journal of Agricultural Economics*, *71*, 76–84.

Malhotra, A. K., Virkkunen, M., Rooney, W., Eggert, M., Linnoila, M., & Goldman, D. (1996). The association between the dopamine D4 receptor (D4DR) 16 amino acid repeat polymorphism and novelty seeking. *Molecular Psychiatry*, *1*, 388–391.

Manni, A., Pardridge, W. M., Cefalu, W., Nisula, B. C., Bardin, C. W., Santner, S. J., & Santen, R. J. (1985). Bioavailability of albumin-bound testosterone. *Journal of Clinical Endocrinology & Metabolism*, *61*, 705–710.

Manning, J. T., Scutt, D., Wilson, J., & Lewis-Jones, D. I. (1998). The ratio of 2nd to 4th digit length: A predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. *Human Reproduction*, *13*, 3000–3004.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*, 906–913.

Marmot, M. G., Kogevinas, M., & Elston, M. A. (1987). Social/economic status and disease. *Annual Review of Public Health*, *8*, 111–135.

Martin, N. W., Medland, S. E., Verweij, K. J. H., Lee, S. H., Nyholt, D. R., Madden, P. A., … Martin, N. G. (2011). Educational attainment: A genome wide association study in 9538 Australians. *PLOS ONE*, *6*(6), e20128.

Matthews, K. A., Kelsey, S. F., Meilahn, E. N., Kuller, L. H., & Wing, R. R. (1989). Educational attainment and behavioral and biologic risk factors for coronary heart disease in middle-aged women. *American Journal of Epidemiology*, *129*, 1132–1144.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*, 356–369.

McGrath, R. G., & MacMillan, I. C. (2000). *The entrepreneurial mindset*. Boston, MA, USA: Harvard Business School Press.

McGue, M., Vaupel, J. W., Holm, N., & Harvald, B. (1993). Longevity is moderately heritable in a sample of Danish twins born 1870–1880. *Journal of Gerontology*, *48*, B237–B244.

Mendel, C. M. (1989). The free hormone hypothesis: A physiologically based mathematical model. *Endocrine Reviews*, *10*, 232–274.

Miller, P., Mulvey, C., & Martin, N. (2001). Genetic and environmental contributions to educational attainment in Australia. *Economics of Education Review*, *20*, 211–224.

Mitchell, B. D., Hsueh, W. C., King, T. M., Pollin, T. I., Sorkin, J., Agarwala, R., … Shuldiner, A. R. (2001). Heritability of life span in the Old Order Amish. *American Journal of Medical Genetics 102*, 346–352.

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most published research findings are false—But a little replication goes a long way. *PLOS Medicine*, *4*(2), e28.

Munafo, M. R., Yalcin, B., Willis-Owen, S. A., & Flint, J. (2008). Association of the dopamine D4 receptor (DRD4) gene and approach-related personality traits: Meta-analysis and new data. *Biological Psychiatry*, *63*, 197–206.

National Center for Biotechnology Information. (2012a). *dbSNP*. Retrieved from http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi

National Center for Biotechnology Information. (2012b). *ABHD5 abhydrolase domain containing 5*. Retrieved from http://www.ncbi.nlm.nih.gov/gene/51099

Neale, M. C. (2003). Twin studies: Software and algorithms. In D. Cooper (Ed.), *Nature encyclopedia of the human genome* (pp. 679–683). London, UK: Macmillan, Nature Publishing Group.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling* (6th ed.). Retrieved from http://www.vipbg.vcu.edu/~vipbg/software/mxmanual.pdf

Nicolaou, N., & Shane, S. (2009). Can genetic factors influence the likelihood of engaging in entrepreneurial activity? *Journal of Business Venturing*, *24*, 1–22.

Nicolaou, N., & Shane, S. (2010). Entrepreneurship and occupational choice: Genetic and environmental influences. *Journal of Economic Behavior & Organization*, *76*, 3–14.

Nicolaou, N., Shane, S., Adi, G., Mangino, M., & Harris, J. (2011). A polymorphism associated with entrepreneurship: Evidence from dopamine receptor candidate genes. *Small Business Economics*, *36*, 151–155.

Nicolaou, N., Shane, S., Cherkas, L., Hunkin, J., & Spector, T. D. (2008). Is the tendency to engage in entrepreneurship genetic? *Management Science*, *54*, 167–179.

Nicolaou, N., Shane, S., Cherkas, L., & Spector, T. D. (2008). The influence of sensation seeking in the heritability of entrepreneurship. *Strategic Entrepreneurship Journal*, *2*, 7–21.

Nicolaou, N., Shane, S., Cherkas, L., & Spector, T. D. (2009). Opportunity recognition and the tendency to be an entrepreneur: A bivariate genetics perspective. *Organizational Behavior and Human Decision Processes*, *110*, 108–117.

Obeidat, M., Wain, L. V., Shrine, N., Kalsheker, N., Soler-Artigas, M., Repapi, E., … Hall, I. P. (2011). A comprehensive evaluation of potential lung function associated genes in the SpiroMeta general population sample. *PLOS ONE*, *6*(5), e19382.

Ohlsson, C., Wallaschofski, H., Lunetta, K. L., Stolk, L., Perry, J. R. B., Koster, A., … Haring, R. (2011). Genetic determinants of serum testosterone concentrations in men. *PLOS Genetics*, *7*(10), e1002313.

Parker, S. C. (2009). *The economics of entrepreneurship*. Cambridge, UK: Cambridge University Press.

Paterson, A. D., Sunohara, G. A., Kennedy, J. L. (1999). Dopamine D4 receptor gene: Novelty or nonsense? *Neuropsychopharmacology*, *21*, 3–16.

Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA*, *299*, 1335–1344.

Pérez de Castro, I., Ibáñez, A., Torres, P., Sáiz-Ruiz, J., & Fernández-Piqueras, J. (1997). Genetic association study between pathological gambling and a functional DNA polymorphism at the D4 receptor gene. *Pharmacogenetics*, *7*, 345–348.

Plomin, R. (1999). Genetics and general cognitive ability. *Nature*, *402*(Supp), C25–C29.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral Genetics* (5th ed.). New York, NY, USA: Worth.

Plomin, R., & Kosslyn, S. M. (2001). Genes, brain and cognition. *Nature Neuroscience*, *4*, 1153–1155.

Plomin, R., & Spinath, F. M. (2004). Intelligence: Genetics, genes, and genomics. *Journal of Personality and Social Psychology*, *86*, 112–129.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909.

Psaty, B. M., O'Donnell, C. J., Gudnason, V., Lunetta, K. L., Folsom, A. R., Rotter, J. I., … Boerwinkle, E. (2009). Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from five cohorts. *Circulation: Cardiovascular Genetics*, *2*, 73–80.

Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, *19*, 149–150.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, *81*, 559–575.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., … Scolnick, E. M. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*, 748–752.

Rablen, M. D., & Oswald, A. J. (2008). Mortality and immortality: The Nobel Prize as an experiment into the effect of status upon longevity. *Journal of Health Economics*, *27*, 1462–1471.

Ramsköld, D., Wang, E. T., Burge, C. B., & Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLOS Computational Biology*, *5*(12), e1000598.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*, 444–454.

Rees, H., & Shah, A. (1986). An empirical analysis of self-employment in the U.K. *Journal of Applied Economics*, *1*, 95–108.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, *447*, 425–432.

Riding, A. L., & Swift, C. S. (1990). Women business owners and terms of credit: Some empirical findings of the Canadian experience. *Journal of Business Venturing*, *5*, 327–340.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., … Koellinger, P. D. (2013). *GWAS of 126,559 individuals identifies common genetic variants associated with educational attainment*. Manuscript submitted for publication.

Roessler, C., & Koellinger, P. (2012). Entrepreneurship and organization design. *European Economic Review*, *56*, 888–902.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*, 127–141.

Rutter, M. (2006). *Genes and behavior: Nature-nurture interplay explained*. Oxford, UK: Blackwell.

Sacerdote, B. (2007). How large are the effects from changes in family environment? A study of Korean American adoptees. *Quarterly Journal of Economics*, *122*, 119–157.

Sandberg, R., Yasuda, R., Pankratz, D. G., Carter, T. A., Del Rio, J. A., Wodicka, L., … Barlow, C. (2000). Regional and strain-specific gene expression mapping in the adult mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11038–11043.

Sapienza, P., Zingales, L., & Maestripieri, D. (2009). Gender differences in financial risk aversion and career choices are affected by testosterone. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 15268–15273.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics*, *53*, 1253–1261.

Sauerbrei, W., & Royston, P. (1999). Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*, 71–94.

Scarr, S., & Weinberg, R. A. (1994). Educational and occupational achievements of brothers and sisters in adoptive and biologically related families. *Behavior Genetics*, *24*, 301–325.

Schmidt, C. O., Ittermann, T., Schulz, A., Grabe, H. J., & Baumeister, S. E. (2013). Linear, nonlinear or categorical: How to treat complex associations in regression analyses? Polynomial transformations and fractional polynomials. *International Journal of Public Health*, *58*, 157–160.

Schmitz, J. A., Jr. (1989). Imitation, entrepreneurship, and long-run growth. *Journal of Political Economy*, *97*, 721–739.

Schnall, P. L., Landsbergis, P. A., & Baker, D. (1994). Job strain and cardiovascular disease. *Annual Review of Public Health*, *15*, 381–411.

Schumpeter, J. A. (1934). *The theory of economic development*. Cambridge, MA, USA: Harvard University Press.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584.

Shane, S. (2010). *Born entrepreneurs, born leaders: How your genes affect your work life.* New York, NY, USA: Oxford University Press.

Shane, S., Nicolaou, N., Cherkas, L., & Spector, T. D. (2010). Genetics, the Big Five, and the tendency to be self-employed. *Journal of Applied Psychology*, *95*, 1154–1162.

Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of Management Review*, *25*, 217–226.

Shyn, S. I., Shi, J., Kraft, J. B., Potash, J. B., Knowles, J. A., Weissman, M. M., … Hamilton, S. P. (2011). Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Molecular Psychiatry*, *16*, 202–215.

Siontis, K. C. M., Patsopoulos, N. A., & Ioannidis, J. P. A. (2010). Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *European Journal of Human Genetics*, *18*, 832–837.

Södergard, R., Bäckström, T., Shanbhag, V., & Carstensen, H. (1982). Calculation of free and bound fractions of testosterone and estradiol-17β to human plasma proteins at body temperature. *Journal of Steroid Biochemistry*, *16*, 801–810.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., … Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*, 937–948.

Stanton, S. J., Liening, S. H., & Schultheiss, O. C. (2011). Testosterone is positively associated with risk taking in the Iowa Gambling Task. *Hormones and Behavior*, *59*, 252–256.

Steenland, K., Henley, J., & Thun, M. (2002). All-cause and cause-specific death rates by educational status for two million people in two American cancer society cohorts, 1959–1996. *American Journal of Epidemiology*, *156*, 11–21.

Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G. B., … Stefansson, K. (2007). A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature Genetics*, *39*, 770–775.

Stewart, W. H., & Roth, P. L. (2001). Risk propensity differences between entrepreneurs and managers: A meta-analytic review. *Journal of Applied Psychology*, *86*, 145–153.

Stewart, W. H., & Roth, P. L. (2004). Data quality affects meta-analytic conclusions: A response to Miner and Raju (2004) concerning entrepreneurial risk propensity. *Journal of Applied Psychology*, *89*, 14–21.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, *21*, 65–66.

Taubman, P. (1976). The determinants of earnings: Genetics, family, and other environments: A study of white male twins. *American Economic Review*, *66*, 858–870.

Tan, U. (2008). Ratio of fourth to second fingertip extensions in relation to serum estradiol and testosterone levels in men and women. *Perceptual and Motor Skills*, *107*, 3–13.

Terracciano, A., Balaci, L., Thayer, J., Scally, M., Kokinos, S., Ferrucci, L., … Costa, P. T., Jr. (2009). Variants of the serotonin transporter gene and NEO-PI-R Neuroticism: No association in the BLSA and SardiNIA samples. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *150B*, 1070–1077.

Terracciano, A., Sanna, S., Uda, M., Deiana, B., Usala, G., Busonero, F., … Costa, P. T., Jr. (2010). Genome-wide association scan for five major dimensions of personality. *Molecular Psychiatry*, *15*, 647–656.

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*, 1061–1073.

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*, 1299–1320.

Thode, H. C., Jr. (2002). *Testing for normality*. New York, NY, USA: Marcel Dekker.

Thurik, A. R. (2012, August). Entrepreneurship and modern biology. In A. Hofman (Chair), *Erasmus Summer Programme Lectures*. Symposium conducted at the Erasmus Summer Programme 2012, Rotterdam, The Netherlands.

Thurik, A. R., Carree, M. A., Van Stel, A., & Audretsch, D. B. (2008). Does self-employment reduce unemployment? *Journal of Business Venturing*, *23*, 673–686.

Tooby, J. & Cosmides, L. (1992). The Psychological Foundations of Culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York, NY, USA: Oxford University Press.

Trahms, C. A., Coombs, J. E., & Barrick, M. (2010). Does biology matter? How prenatal testosterone, entrepreneur risk propensity, and entrepreneur risk perceptions influence venture performance. *Frontiers of Entrepreneurship Research*, *30*(5), article 4.

Unger, J. M., Rauch, A., Narayanan, J., Weis, S., & Frese, M. (2009). Does prenatal testosterone predict entrepreneurial success? Relationships of 2D:4D and business success. *Frontiers of Entrepreneurship Research*, *29*(5), article 15.

United Nations Educational, Scientific and Cultural Organization. (2006). *International Standard Classification of Education 1997: ISCED 1997. May 2006 Re-edition*. Retrieved from http://www.uis.unesco.org/Library/Documents/isced97-en.pdf

Van der Loos, M. J. H. M., Groenen, P. J. F., Hofman, A., Koellinger, P. D., Rivadeneira, F., Van Rooij, F. J. A., … Uitterlinden, A. G. (2011). De genetica van ondernemerschap. *ESB Dossier*, *96*(4609S), 30–36.

Van der Loos, M. J. H. M., Haring, R., Rietveld, C. A., Baumeister, S. E., Groenen, P. J. F., Hofman, A., … Thurik, A. R. (2013). *Measures of bioactive serum testosterone are not associated with entrepreneurial behavior in two independent observational studies*. Manuscript submitted for publication.

Van der Loos, M. J. H. M., Koellinger, P. D., Groenen, P. J. F., Rietveld, C. A., Rivadeneira, F., Van Rooij, F. J. A., … Thurik, A. R. (2011). Candidate gene studies and the quest for the entrepreneurial gene. *Small Business Economics*, *37*, 269–275.

Van der Loos, M. J. H. M., Koellinger, P. D., Groenen, P. J. F., & Thurik, A. R. (2010). Genome-wide association studies and the genetics of entrepreneurship. *European Journal of Epidemiology*, *25*, 1–3.

Van der Loos, M. J. H. M., Rietveld, C. A., Eklund, N., Koellinger, P. D., Rivadeneira, F., Abecasis, G. R., … Thurik, A. R. (2013). The molecular genetic architecture of self-employment, *PLOS ONE*, *8*(4), e60542.

Van der Zwan, P., Thurik, R., & Grilo, I. (2010), The entrepreneurial ladder and its determinants, *Applied Economics*, *42*, 2183–2191.

Van Kippersluis, J. L. W., O'Donnell, O. A., & Van Doorslaer, E. K. A. (2011). Long run returns to education: Does education lead to an extended old age? *Journal of Human Resources*, *94*, 695–721.

Van Stel, A., Carree, M., & Thurik, A. R. (2005). The effect of entrepreneurial activity on national economic growth. *Small Business Economics*, *24*, 311–321.

V.B. Hjelmborg, J., Iachine, I., Skytthe, A., Vaupel, J. W., McGue, M., Koskenvuo, M., … Christensen, K. (2006). Genetic influence on human lifespan and longevity. *Human Genetics*, *119*, 312–321.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*, 1304–1351.

Verheul, I., Carree, M., & Thurik, R. (2009). Allocation and productivity of time in new ventures of female and male entrepreneurs. *Small Business Economics*, *33*, 273–291.

Verheul, I., & Thurik, A. R. (2001). Start-up capital: "Does gender matter?". *Small Business Economics*, *16*, 329–346.

Verheul, I., Thurik, A., Grilo, I., Van der Zwan, P. (2012). Explaining preferences and actual involvement in self-employment: Gender and the entrepreneurial personality. *Journal of Economic Psychology*, *33*, 325–341.

Verheul, I., Uhlaner, L. & Thurik, R. (2005). Business accomplishments, gender and entrepreneurial self-image. *Journal of Business Venturing*, *20*, 483–518.

Verweij, K. J. H., Yang, J., Lahti, J., Veijola, J., Hintsanen, M., Pulkki-Råback, L., … Zietsch, B. P. (2012). Maintenance of genetic variation in human personality: Testing evolutionary models by estimating heritability due to common causal variants and investigating the effect of distant inbreeding. *Evolution*, *66*, 3238–3251.

Verweij, K. J. H., Zietsch, B. P., Medland, S. E., Gordon, S. D., Benyamin, B., Nyholt, D. R., … Wray, N. R. (2010). Genome-wide association study of Cloninger's temperament scales: Implications for the evolutionary genetics of personality. *Biological Psychology*, *85*, 306–317.

Vining, R. F., & McGinley, R. A. (2006). The measurement of hormones in saliva: Possibilities and pitfalls. *Journal of Steroid Biochemistry*, *27*, 81–94.

Vinkhuyzen, A. A. E., Pedersen, N. L., Yang, J., Lee, S. H., Magnusson, P. K. E., Iacono, W. G., … Wray, N. R. (2012). Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Translational Psychiatry*, *2*(4), e102.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*, 7–24.

Visscher, P. M., Goddard, M. E., Derks, E. M., & Wray, N. R. (2012). Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Molecular Psychiatry*, *17*, 474–485.

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era—Concepts and misconceptions. *Nature Reviews Genetics*, *9*, 255–266.

Visscher, P. M., Yang, J., & Goddard, M. E. (2010). A commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010) *Twin Research and Human Genetics*, *13*, 517–524.

Völzke, H., Alte, D., Schmidt, C. O., Radke, D., Lorbeer, R., Friedrich, N., … Hoffmann, W. (2011). Cohort profile: The Study of Health in Pomerania. *International Journal of Epidemiology*, *40*, 294–307.

Von Hinke Kessler Scholder, S., Davey Smith, G., Lawlor, D. A., Propper, C., & Windmeijer, F. (2011). Mendelian Randomization: The use of genes in instrumental variable analyses. *Health Economics*, *20*, 893–896.

Vormfelde, S. V., Hoell, I., Tzvetkov, M., Jamrozinski, K., Sehrt, D., Brockmöller, J., & Leibing, E. (2006). Anxiety- and novelty seeking-related personality traits and serotonin transporter gene polymorphisms. *Journal of Psychiatric Research*, *40*, 568–576.

Wagner, J. (2007). What a difference a Y makes—Female and male nascent entrepreneurs in Germany. *Small Business Economics*, *28*, 1–21.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics*, *6*, 109–118.

Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., … Houlston, R. S. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nature Genetics*, *40*, 1407–1409.

Waterland, R. A., & Jirtle, R. L. (2003). Transposable elements: Targets for early nutritional effects on epigenetic gene regulation. *Molecular and Cellular Biology*, *23*, 5293–5300.

Weertman, A., Arntz, A., Dreessen, L., Van Velzen, C., & Vertommen, S. (2003). Short-interval test-retest interrater reliability of the Dutch version of the Structured Clinical Interview for DSM-IV personality disorders (SCID-II). *Journal of Personality Disorders*, *17*, 562–567.

Wehby, G. L., Ohsfeldt, R. L., & Murray, J. C. (2008). 'Mendelian Randomization' equals instrumental variable analysis with genetic instruments. *Statistics in Medicine*, *27*, 2745–2749.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, *64*, 368–382.

Wennekers, S., & Thurik, R. (1999). Linking entrepreneurship and economic growth. *Small Business Economics*, *13*, 27–55.

Westhead, P., Ucbasaran, D., Wright, M., & Binks, M. (2005). Novice, serial and portfolio entrepreneur behavior and contributions. *Small Business Economics*, *25*, 109–132.

White, R. E., Thornhill, S., & Hampson, E. (2006). Entrepreneurs and evolutionary biology: The relationship between testosterone and new venture creation. *Organizational Behavior and Human Decision Processes*, *100*, 21–34.

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*, 2190–2191.

Winkleby, M. A., Jatulis, D. E., Frank, E., & Fortmann, S. P. (1992). Socioeconomic status and health: How education, income, and occupation contribute to risk factors for cardiovascular disease. *American Journal of Public Health*, *82*, 816–820.

Wray, N. R., Purcell, S. M., & Visscher, P. M. (2011). Synthetic associations created by rare variants do not explain most GWAS results. *PLOS Biology*, *9*(1), e1000579.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*, 565–569.

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*, 76–82.

Zethraeus, N., Kocoska-Maras, L., Ellingsen, T., Von Schoultz, B., Hirschberg, A. L., & Johannesson. M. (2009). A randomized trial of the effect of estrogen and testosterone on economic behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 6535–6538.

Zhang, Z., Zyphur, M. J., Narayanan, J., Arvey, R. D., Chaturvedi, S., Avolio, B. J., … Larssong, G. (2009). The genetic basis of entrepreneurship: Effects of gender and personality. *Organizational Behavior and Human Decision Processes*, *110*, 93–107.

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York, NY, USA: John Wiley & Sons.

Ziegler, A., & König, I. R. (2010). *A statistical approach to genetic epidemiology* (2nd ed.). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA.

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 1193–1198.

About the Author

Matthijs J.H.M. van der Loos (1984) completed his grammar school education in Rotterdam in 2002. He attended the Erasmus University Rotterdam and graduated cum laude in 2007 with a Master of Science degree in Economics and Informatics. After spending a semester at the University of Sydney, he embarked on his PhD in economics under the supervision of Professors Roy Thurik, Patrick Groenen, and Albert Hofman, and Associate Professor Philipp Koellinger. His research focused on methods that enable identification of genes associated with entrepreneurship as well as a hormonal correlate through which these genes may influence entrepreneurial behavior. In addition, he coauthored several papers regarding the molecular genetics of various other economic outcomes and behaviors. His work has been published in international peer-reviewed journals such as *Small Business Economics*, the *European Journal of Epidemiology*, and *PLOS ONE*. He has presented his work at various international conferences including the *Babson College Entrepreneurship Research Conference*, the *Behavior Genetics Association Annual Meeting*, and the *European Association for Research in Industrial Economics Annual Conference*. Matthijs is currently on the job market in the private sector.

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ERIM PH.D. SERIES RESEARCH IN MANAGEMENT

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://hdl.handle.net/1765/1 ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

DISSERTATIONS LAST FIVE YEARS

Acciaro, M., *Bundling Strategies in Global Supply Chains*, Promoter(s): Prof.dr. H.E. Haralambides, EPS-2010-197-LIS, http://hdl.handle.net/1765/19742

Agatz, N.A.H., *Demand Management in E-Fulfillment*, Promoter(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2009-163-LIS, http://hdl.handle.net/1765/15425

Alexiev, A., *Exploratory Innovation: The Role of Organizational and Top Management Team Social Capital*, Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-208-STR, http://hdl.handle.net/1765/20632

Asperen, E. van, *Essays on Port, Container, and Bulk Chemical Logistics Optimization*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2009-181-LIS, http://hdl.handle.net/1765/17626

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-Frequency Data*, Promoter(s): Prof.dr. D.J.C. van Dijk, EPS-2013-273-F&A, http://hdl.handle.net/1765/38240

Benning, T.M., *A Consumer Perspective on Flexibility in Health Care: Priority Access Pricing and Customized Care*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2011-241-MKT, http://hdl.handle.net/1765/23670

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2013-278-S&E, http://hdl.handle.net/1765/39128

Betancourt, N.E., *Typical Atypicality: Formal and Informal Institutional Conformity, Deviance, and Dynamics*, Promoter(s): Prof.dr. B. Krug, EPS-2012-262-ORG, http://hdl.handle.net/1765/32345

Bezemer, P.J., *Diffusion of Corporate Governance Beliefs: Board Independence and the Emergence of a Shareholder Value Orientation in the Netherlands*, Promoter(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-192-STR, http://hdl.handle.net/1765/18458

Binken, J.L.G., *System Markets: Indirect Network Effects in Action, or Inaction*, Promoter(s): Prof.dr. S. Stremersch, EPS-2010-213-MKT, http://hdl.handle.net/1765/21186

Blitz, D.C., *Benchmarking Benchmarks*, Promoter(s): Prof.dr. A.G.Z. Kemna & Prof.dr. W.F.C. Verschoor, EPS-2011-225-F&A, http://hdl.handle.net/1765/22624

Borst, W.A.M., *Understanding Crowdsourcing: Effects of Motivation and Rewards on Participation and Performance in Voluntary Online Activities*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2010-221-LIS, http://hdl.handle.net/1765/21914

Budiono, D.P., *The Analysis of Mutual Fund Performance: Evidence from U.S. Equity Mutual Funds*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-185-F&A, http://hdl.handle.net/1765/18126

Burger, M.J., *Structure and Cooptition in Urban Networks*, Promoter(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.R. Commandeur, EPS-2011-243-ORG, http://hdl.handle.net/1765/26178

Camacho, N.M., *Health and Marketing: Essays on Physician and Patient Decision-making*, Promoter(s): Prof.dr. S. Stremersch, EPS-2011-237-MKT, http://hdl.handle.net/1765/23604

Carvalho, L., *Knowledge Locations in Cities: Emergence and Development Dynamics*, Promoter(s): Prof.dr. L. van den Berg, EPS-2013-274-S&E, http://hdl.handle.net/1765/38449

Carvalho de Mesquita Ferreira, L., *Attention Mosaics: Studies of Organizational Attention*, Promoter(s): Prof.dr. P.M.A.R. Heugens & Prof.dr. J. van Oosterhout, EPS-2010-205-ORG, http://hdl.handle.net/1765/19882

Chen, C.-M., *Evaluation and Design of Supply Chain Operations Using DEA*, Promoter(s): Prof.dr. J.A.E.E. van Nunen, EPS-2009-172-LIS, http://hdl.handle.net/1765/16181

Defilippi Angeldonis, E.F., *Access Regulation for Naturally Monopolistic Port Terminals: Lessons from Regulated Network Industries*, Promoter(s): Prof.dr. H.E. Haralambides, EPS-2010-204-LIS, http://hdl.handle.net/1765/19881

Deichmann, D., *Idea Management: Perspectives from Leadership, Learning, and Network Theory*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2012-255-ORG, http://hdl.handle.net/1765/31174

Desmet, P.T.M., *In Money we Trust? Trust Repair and the Psychology of Financial Compensations*, Promoter(s): Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2011-232-ORG, http://hdl.handle.net/1765/23268

Diepen, M. van, *Dynamics and Competition in Charitable Giving*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2009-159-MKT, http://hdl.handle.net/1765/14526

Dietvorst, R.C., *Neural Mechanisms Underlying Social Intelligence and Their Relationship with the Performance of Sales Managers*, Promoter(s): Prof.dr. W.J.M.I. Verbeke, EPS-2010-215-MKT, http://hdl.handle.net/1765/21188

Dietz, H.M.S., *Managing (Sales)People towards Performance: HR Strategy, Leadership & Teamwork*, Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2009-168-ORG, http://hdl.handle.net/1765/16081

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promoter(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-272-LIS, http://hdl.handle.net/1765/38241

Doorn, S. van, *Managing Entrepreneurial Orientation*, Promoter(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-258-STR, http://hdl.handle.net/1765/32166

Douwens-Zonneveld, M.G., *Animal Spirits and Extreme Confidence: No Guts, No Glory*, Promoter(s): Prof.dr. W.F.C. Verschoor, EPS-2012-257-F&A, http://hdl.handle.net/1765/31914

Duca, E., *The Impact of Investor Demand on Security Offerings*, Promoter(s): Prof.dr. A. de Jong, EPS-2011-240-F&A, http://hdl.handle.net/1765/26041

Duursema, H., *Strategic Leadership: Moving Beyond the Leader-follower Dyad*, Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2013-279-ORG, http://hdl.handle.net/1765/39129

Eck, N.J. van, *Methodological Advances in Bibliometric Mapping of Science*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2011-247-LIS, http://hdl.handle.net/1765/26509

Eijk, A.R. van der, *Behind Networks: Knowledge Transfer, Favor Exchange and Performance*, Promoter(s): Prof.dr. S.L. van de Velde & Prof.dr.drs. W.A. Dolfsma, EPS-2009-161-LIS, http://hdl.handle.net/1765/14613

Essen, M. van, *An Institution-Based View of Ownership*, Promoter(s): Prof.dr. J. van Oosterhout & Prof.dr. G.M.H. Mertens, EPS-2011-226-ORG, http://hdl.handle.net/1765/22643

Feng, L., *Motivation, Coordination and Cognition in Cooperatives*, Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2010-220-ORG, http://hdl.handle.net/1765/21680

Gertsen, H.F.M., *Riding a Tiger without Being Eaten: How Companies and Analysts Tame Financial Restatements and Influence Corporate Reputation*, Promoter(s): Prof.dr. C.B.M. van Riel, EPS-2009-171-ORG, http://hdl.handle.net/1765/16098

Gharehgozli, A.H., *Developing New Methods for Efficient Container Stacking Operations*, Promoter(s): Prof.dr.ir. M.B.M. de Koster, EPS-2012-269-LIS, http://hdl.handle.net/1765/37779

Gijsbers, G.W., *Agricultural Innovation in Asia: Drivers, Paradigms and Performance*, Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2009-156-ORG, http://hdl.handle.net/1765/14524

Gils, S. van, *Morality in Interactions: On the Display of Moral Behavior by Leaders and Employees*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2012-270-ORG, http://hdl.handle.net/1765/38028

Ginkel-Bieshaar, M.N.G. van, *The Impact of Abstract versus Concrete Product Communications on Consumer Decision-making Processes*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-256-MKT, http://hdl.handle.net/1765/31913

Gkougkousi, X., *Empirical Studies in Financial Accounting*, Promoter(s): Prof.dr. G.M.H. Mertens & Prof.dr. E. Peek, EPS-2012-264-F&A, http://hdl.handle.net/1765/37170

Gong, Y., *Stochastic Modelling and Analysis of Warehouse Operations*, Promoter(s): Prof.dr. M.B.M. de Koster & Prof.dr. S.L. van de Velde, EPS-2009-180-LIS, http://hdl.handle.net/1765/16724

Greeven, M.J., *Innovation in an Uncertain Institutional Environment: Private Software Entrepreneurs in Hangzhou, China*, Promoter(s): Prof.dr. B. Krug, EPS-2009-164-ORG, http://hdl.handle.net/1765/15426

Hakimi, N.A., *Leader Empowering Behaviour: The Leader's Perspective: Understanding the Motivation behind Leader Empowering Behaviour*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2010-184-ORG, http://hdl.handle.net/1765/17701

Hensmans, M., *A Republican Settlement Theory of the Firm: Applied to Retail Banks in England and the Netherlands (1830-2007)*, Promoter(s): Prof.dr. A. Jolink & Prof.dr. S.J. Magala, EPS-2010-193-ORG, http://hdl.handle.net/1765/19494

Hernandez Mireles, C., *Marketing Modeling for New Products*, Promoter(s): Prof.dr. P.H. Franses, EPS-2010-202-MKT, http://hdl.handle.net/1765/19878

Heyden, M.L.M., *Essays on Upper Echelons & Strategic Renewal: A Multilevel Contingency Approach*, Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-259-STR, http://hdl.handle.net/1765/32167

Hoever, I.J., *Diversity and Creativity: In Search of Synergy*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2012-267-ORG, http://hdl.handle.net/1765/37392

Hoogendoorn, B., *Social Entrepreneurship in the Modern Economy: Warm Glow, Cold Feet*, Promoter(s): Prof.dr. H.P.G. Pennings & Prof.dr. A.R. Thurik, EPS-2011-246-STR, http://hdl.handle.net/1765/26447

Hoogervorst, N., *On The Psychology of Displaying Ethical Leadership: A Behavioral Ethics Approach*, Promoter(s): Prof.dr. D. De Cremer & Dr. M. van Dijke, EPS-2011-244-ORG, http://hdl.handle.net/1765/26228

Huang, X., *An Analysis of Occupational Pension Provision: From Evaluation to Redesign*, Promoter(s): Prof.dr. M.J.C.M. Verbeek & Prof.dr. R.J. Mahieu, EPS-2010-196-F&A, http://hdl.handle.net/1765/19674

Hytönen, K.A., *Context Effects in Valuation, Judgment and Choice*, Promoter(s): Prof.dr.ir. A. Smidts, EPS-2011-252-MKT, http://hdl.handle.net/1765/30668

Jalil, M.N., *Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics*, Promoter(s): Prof.dr. L.G. Kroon, EPS-2011-222-LIS, http://hdl.handle.net/1765/22156

Jaspers, F.P.H., *Organizing Systemic Innovation*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2009-160-ORG, http://hdl.handle.net/1765/14974

Jiang, T., *Capital Structure Determinants and Governance Structure Variety in Franchising*, Promoter(s): Prof.dr. G. Hendrikse & Prof.dr. A. de Jong, EPS-2009-158-F&A, http://hdl.handle.net/1765/14975

Jiao, T., *Essays in Financial Accounting*, Promoter(s): Prof.dr. G.M.H. Mertens, EPS-2009-176-F&A, http://hdl.handle.net/1765/16097

Kaa, G. van, *Standard Battles for Complex Systems: Empirical Research on the Home Network*, Promoter(s): Prof.dr.ir. J. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-166-ORG, http://hdl.handle.net/1765/16011

Kagie, M., *Advances in Online Shopping Interfaces: Product Catalog Maps and Recommender Systems*, Promoter(s): Prof.dr. P.J.F. Groenen, EPS-2010-195-MKT, http://hdl.handle.net/1765/19532

Kappe, E.R., *The Effectiveness of Pharmaceutical Marketing*, Promoter(s): Prof.dr. S. Stremersch, EPS-2011-239-MKT, http://hdl.handle.net/1765/23610

Karreman, B., *Financial Services and Emerging Markets*, Promoter(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.P.G. Pennings, EPS-2011-223-ORG, http://hdl.handle.net/1765/22280

Kwee, Z., *Investigating Three Key Principles of Sustained Strategic Renewal: A Longitudinal Study of Long-Lived Firms*, Promoter(s): Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-174-STR, http://hdl.handle.net/1765/16207

Lam, K.Y., *Reliability and Rankings*, Promoter(s): Prof.dr. P.H.B.F. Franses, EPS-2011-230-MKT, http://hdl.handle.net/1765/22977

Lander, M.W., *Profits or Professionalism? On Designing Professional Service Firms*, Promoter(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2012-253-ORG, http://hdl.handle.net/1765/30682

Langhe, B. de, *Contingencies: Learning Numerical and Emotional Associations in an Uncertain World*, Promoter(s): Prof.dr.ir. B. Wierenga & Prof.dr. S.M.J. van Osselaer, EPS-2011-236-MKT, http://hdl.handle.net/1765/23504

Larco Martinelli, J.A., *Incorporating Worker-Specific Factors in Operations Management Models*, Promoter(s): Prof.dr.ir. J. Dul & Prof.dr. M.B.M. de Koster, EPS-2010-217-LIS, http://hdl.handle.net/1765/21527

Li, T., *Informedness and Customer-Centric Revenue Management*, Promoter(s): Prof.dr. P.H.M. Vervest & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-146-LIS, http://hdl.handle.net/1765/14525

Liang, Q., *Governance, CEO Identity, and Quality Provision of Farmer Cooperatives,* Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2013-281-ORG, http://hdl.handle.net/1765/39253

Lovric, M., *Behavioral Finance and Agent-Based Artificial Markets*, Promoter(s): Prof.dr. J. Spronk & Prof.dr.ir. U. Kaymak, EPS-2011-229-F&A, http://hdl.handle.net/1765/22814

Maas, K.E.G., *Corporate Social Performance: From Output Measurement to Impact Measurement*, Promoter(s): Prof.dr. H.R. Commandeur, EPS-2009-182-STR, http://hdl.handle.net/1765/17627

Markwat, T.D., *Extreme Dependence in Asset Markets Around the Globe*, Promoter(s): Prof.dr. D.J.C. van Dijk, EPS-2011-227-F&A, http://hdl.handle.net/1765/22744

Mees, H., *Changing Fortunes: How China's Boom Caused the Financial Crisis*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2012-266-MKT, http://hdl.handle.net/1765/34930

Meuer, J., *Configurations of Inter-Firm Relations in Management Innovation: A Study in China's Biopharmaceutical Industry*, Promoter(s): Prof.dr. B. Krug, EPS-2011-228-ORG, http://hdl.handle.net/1765/22745

Mihalache, O.R., *Stimulating Firm Innovativeness: Probing the Interrelations between Managerial and Organizational Determinants*, Promoter(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-260-S&E, http://hdl.handle.net/1765/32343

Milea, V., *New Analytics for Financial Decision Support*, Promoter(s): Prof.dr.ir. U. Kaymak, EPS-2013-275-LIS, http://hdl.handle.net/1765/38673

Moonen, J.M., *Multi-Agent Systems for Transportation Planning and Coordination*, Promoter(s): Prof.dr. J. van Hillegersberg & Prof.dr. S.L. van de Velde, EPS-2009-177-LIS, http://hdl.handle.net/1765/16208

Nederveen Pieterse, A., *Goal Orientation in Teams: The Role of Diversity*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-162-ORG, http://hdl.handle.net/1765/15240

Nielsen, L.K., *Rolling Stock Rescheduling in Passenger Railways: Applications in Short-term Planning and in Disruption Management*, Promoter(s): Prof.dr. L.G. Kroon, EPS-2011-224-LIS, http://hdl.handle.net/1765/22444

Niesten, E.M.M.I., *Regulation, Governance and Adaptation: Governance Transformations in the Dutch and French Liberalizing Electricity Industries*, Promoter(s): Prof.dr. A. Jolink & Prof.dr. J.P.M. Groenewegen, EPS-2009-170-ORG, http://hdl.handle.net/1765/16096

Nijdam, M.H., *Leader Firms: The Value of Companies for the Competitiveness of the Rotterdam Seaport Cluster*, Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2010-216-ORG, http://hdl.handle.net/1765/21405

Noordegraaf-Eelens, L.H.J., *Contested Communication: A Critical Analysis of Central Bank Speech*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2010-209-MKT, http://hdl.handle.net/1765/21061

Nuijten, A.L.P., *Deaf Effect for Risk Warnings: A Causal Examination applied to Information Systems Projects*, Promoter(s): Prof.dr. G. van der Pijl & Prof.dr. H. Commandeur & Prof.dr. M. Keil, EPS-2012-263-S&E, http://hdl.handle.net/1765/34928

Nuijten, I., *Servant Leadership: Paradox or Diamond in the Rough? A Multidimensional Measure and Empirical Evidence*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-183-ORG, http://hdl.handle.net/1765/21405

Oosterhout, M. van, *Business Agility and Information Technology in Service Organizations*, Promoter(s): Prof.dr.ir. H.W.G.M. van Heck, EPS-2010-198-LIS, http://hdl.handle.net/1765/19805

Oostrum, J.M. van, *Applying Mathematical Models to Surgical Patient Planning*, Promoter(s): Prof.dr. A.P.M. Wagelmans, EPS-2009-179-LIS, http://hdl.handle.net/1765/16728

Osadchiy, S.E., *The Dynamics of Formal Organization: Essays on Bureaucracy and Formal Rules*, Promoter(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2011-231-ORG, http://hdl.handle.net/1765/23250

Otgaar, A.H.J., *Industrial Tourism: Where the Public Meets the Private*, Promoter(s): Prof.dr. L. van den Berg, EPS-2010-219-ORG, http://hdl.handle.net/1765/21585

Ozdemir, M.N., *Project-level Governance, Monetary Incentives and Performance in Strategic R&D Alliances*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2011-235-LIS, http://hdl.handle.net/1765/23550

Peers, Y., *Econometric Advances in Diffusion Models*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-251-MKT, http://hdl.handle.net/1765/30586

Pinçe, Ç., *Advances in Inventory Management: Dynamic Models*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2010-199-LIS, http://hdl.handle.net/1765/19867

Porras Prado, M., *The Long and Short Side of Real Estate, Real Estate Stocks, and Equity*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2012-254-F&A, http://hdl.handle.net/1765/30848

Poruthiyil, P.V., *Steering Through: How Organizations Negotiate Permanent Uncertainty and Unresolvable Choices*, Promoter(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. S. Magala, EPS-2011-245-ORG, http://hdl.handle.net/1765/26392

Potthoff, D., *Railway Crew Rescheduling: Novel Approaches and Extensions*, Promoter(s): Prof.dr. A.P.M. Wagelmans & Prof.dr. L.G. Kroon, EPS-2010-210-LIS, http://hdl.handle.net/1765/21084

Pourakbar, M., *End-of-Life Inventory Decisions of Service Parts*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2011-249-LIS, http://hdl.handle.net/1765/30584

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promoter(s): Prof.dr. H.R. Commandeur & Prof.dr. H.J.H.M. Claassen, EPS-2013-282-S&E, http://hdl.handle.net/1765/39654

Rijsenbilt, J.A., *CEO Narcissism: Measurement and Impact*, Promoter(s): Prof.dr. A.G.Z. Kemna & Prof.dr. H.R. Commandeur, EPS-2011-238-STR, http://hdl.handle.net/1765/23554

Roelofsen, E.M., *The Role of Analyst Conference Calls in Capital Markets*, Promoter(s): Prof.dr. G.M.H. Mertens & Prof.dr. L.G. van der Tas RA, EPS-2010-190-F&A, http://hdl.handle.net/1765/18013

Rosmalen, J. van, *Segmentation and Dimension Reduction: Exploratory and Model-Based Approaches*, Promoter(s): Prof.dr. P.J.F. Groenen, EPS-2009-165-MKT, http://hdl.handle.net/1765/15536

Roza, M.W., *The Relationship between Offshoring Strategies and Firm Performance: Impact of Innovation, Absorptive Capacity and Firm Size*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2011-214-STR, http://hdl.handle.net/1765/22155

Rus, D., *The Dark Side of Leadership: Exploring the Psychology of Leader Self-serving Behavior*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-178-ORG, http://hdl.handle.net/1765/16726

Schellekens, G.A.C., *Language Abstraction in Word of Mouth*, Promoter(s): Prof.dr.ir. A. Smidts, EPS-2010-218-MKT, http://hdl.handle.net/1765/21580

Sotgiu, F., *Not All Promotions are Made Equal: From the Effects of a Price War to Cross-chain Cannibalization*, Promoter(s): Prof.dr. M.G. Dekimpe & Prof.dr.ir. B. Wierenga, EPS-2010-203-MKT, http://hdl.handle.net/1765/19714

Srour, F.J., *Dissecting Drayage: An Examination of Structure, Information, and Control in Drayage Operations*, Promoter(s): Prof.dr. S.L. van de Velde, EPS-2010-186-LIS, http://hdl.handle.net/1765/18231

Sweldens, S.T.L.R., *Evaluative Conditioning 2.0: Direct versus Associative Transfer of Affect to Brands*, Promoter(s): Prof.dr. S.M.J. van Osselaer, EPS-2009-167-MKT, http://hdl.handle.net/1765/16012

Tarakci, M., *Behavioral Strategy: Strategic Consensus, Power and Networks*, Promoter(s): Prof.dr. P.J.F. Groenen & Prof.dr. D.L. van Knippenberg, EPS-2013-280-ORG, http://hdl.handle.net/1765/39130

Teixeira de Vasconcelos, M., *Agency Costs, Firm Value, and Corporate Investment*, Promoter(s): Prof.dr. P.G.J. Roosenboom, EPS-2012-265-F&A, http://hdl.handle.net/1765/37265

Tempelaar, M.P., *Organizing for Ambidexterity: Studies on the Pursuit of Exploration and Exploitation through Differentiation, Integration, Contextual and Individual Attributes*, Promoter(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-191-STR, http://hdl.handle.net/1765/18457

Tiwari, V., *Transition Process and Performance in IT Outsourcing: Evidence from a Field Study and Laboratory Experiments*, Promoter(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.dr. P.H.M. Vervest, EPS-2010-201-LIS, http://hdl.handle.net/1765/19868

Tröster, C., *Nationality Heterogeneity and Interpersonal Relationships at Work*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2011-233-ORG, http://hdl.handle.net/1765/23298

Tsekouras, D., *No Pain No Gain: The Beneficial Role of Consumer Effort in Decision Making*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-268-MKT, http://hdl.handle.net/1765/37542

Tzioti, S., *Let Me Give You a Piece of Advice: Empirical Papers about Advice Taking in Marketing*, Promoter(s): Prof.dr. S.M.J. van Osselaer & Prof.dr.ir. B. Wierenga, EPS-2010-211-MKT, hdl.handle.net/1765/21149

Vaccaro, I.G., *Management Innovation: Studies on the Role of Internal Change Agents*, Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-212-STR, hdl.handle.net/1765/21150

Verheijen, H.J.J., *Vendor-Buyer Coordination in Supply Chains*, Promoter(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2010-194-LIS, http://hdl.handle.net/1765/19594

Verwijmeren, P., *Empirical Essays on Debt, Equity, and Convertible Securities*, Promoter(s): Prof.dr. A. de Jong & Prof.dr. M.J.C.M. Verbeek, EPS-2009-154-F&A, http://hdl.handle.net/1765/14312

Vlam, A.J., Customer First? *The Relationship between Advisors and Consumers of Financial Products*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-250-MKT, http://hdl.handle.net/1765/30585

Waard, E.J. de, *Engaging Environmental Turbulence: Organizational Determinants for Repetitive Quick and Adequate Responses*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr. J. Soeters, EPS-2010-189-STR, http://hdl.handle.net/1765/18012

Wall, R.S., *Netscape: Cities and Global Corporate Networks*, Promoter(s): Prof.dr. G.A. van der Knaap, EPS-2009-169-ORG, http://hdl.handle.net/1765/16013

Waltman, L., *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*, Promoter(s): Prof.dr.ir. R. Dekker & Prof.dr.ir. U. Kaymak, EPS-2011-248-LIS, http://hdl.handle.net/1765/26564

Wang, Y., *Information Content of Mutual Fund Portfolio Disclosure*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2011-242-F&A, http://hdl.handle.net/1765/26066

Wang, Y., *Corporate Reputation Management: Reaching Out to Find Stakeholders*, Promoter(s): Prof.dr. C.B.M. van Riel, EPS-2013-271-ORG, http://hdl.handle.net/1765/38675

Weerdt, N.P. van der, *Organizational Flexibility for Hypercompetitive Markets: Empirical Evidence of the Composition and Context Specificity of Dynamic Capabilities and Organization Design Parameters*, Promoter(s): Prof.dr. H.W. Volberda, EPS-2009-173-STR, http://hdl.handle.net/1765/16182

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value of Dutch Firms*, Promoter(s): Prof.dr. A. de Jong, EPS-2013-277-F&A, http://hdl.handle.net/1765/39127

Wubben, M.J.J., *Social Functions of Emotions in Social Dilemmas*, Promoter(s): Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2009-187-ORG, http://hdl.handle.net/1765/18228

Xu, Y., *Empirical Essays on the Stock Returns, Risk Management, and Liquidity Creation of Banks*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-188-F&A, http://hdl.handle.net/1765/18125

Yang, J., *Towards the Restructuring and Co-ordination Mechanisms for the Architecture of Chinese Transport Logistics*, Promoter(s): Prof.dr. H.E. Harlambides, EPS-2009-157-LIS, http://hdl.handle.net/1765/14527

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promoter(s): Prof.dr. M.B.M. de Koster, EPS-2013-276-LIS, http://hdl.handle.net/1765/38766

Zhang, D., *Essays in Executive Compensation*, Promoter(s): Prof.dr. I. Dittmann, EPS-2012-261-F&A, http://hdl.handle.net/1765/32344

Zhang, X., *Scheduling with Time Lags*, Promoter(s): Prof.dr. S.L. van de Velde, EPS-2010-206-LIS, http://hdl.handle.net/1765/19928

Zhou, H., Knowledge, *Entrepreneurship and Performance: Evidence from Country-level and Firm-level Studies*, Promoter(s): Prof.dr. A.R. Thurik & Prof.dr. L.M. Uhlaner, EPS-2010-207-ORG, http://hdl.handle.net/1765/20634

Zwan, P.W. van der, *The Entrepreneurial Process: An International Analysis of Entry and Exit*, Promoter(s): Prof.dr. A.R. Thurik & Prof.dr. P.J.F. Groenen, EPS-2011-234-ORG, http://hdl.handle.net/1765/23422

Erasmus Research Institute of Management - ERiM

MOLECULAR GENETICS AND HORMONES
NEW FRONTIERS IN ENTREPRENEURSHIP RESEARCH

Recent studies suggest that entrepreneurship is partly heritable, but are unable to pinpoint the specific genes involved. This thesis presents results from novel research aiming to identify genes associated with entrepreneurship using genetic data on the molecular level. In addition, the relationship between testosterone and entrepreneurship is examined since genes may exert their influence through this hormone.

The thesis starts by reviewing candidate gene studies that test a pre-specified set of genes for association, but which often fail to replicate. An example within the setting of entrepreneurship research is provided to illustrate this last point. Next, the genome-wide association study (GWAS) design is presented that scans the entire genome for associations. However, due to multiple testing, GWAS requires very large sample sizes to establish robust associations and we perform a simulation study to estimate the minimum sample size needed for a GWAS on entrepreneurship. The following part reports evidence that entrepreneurship is partly heritable and around half of the heritability is accounted for by actual molecular genetic data. However, a GWAS on entrepreneurship does not identify robustly associated genes and prediction exercises show that it is currently impossible to predict entrepreneurship solely from molecular genetic data. In the final part, we show that, in contrast to earlier findings, testosterone is not associated with entrepreneurship.

Taken as a whole, the results suggest that entrepreneurship is likely to be influenced by hundreds if not thousands of genes with a very small effect size each, implying that very large sample sizes will be needed in future research to discover associated genes. Most importantly, this thesis may serve as a practical guide for studying the molecular genetics of other economic variables. In conclusion, this thesis helps to build the foundations for a novel research field that integrates molecular genetics into economics.

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERIM PhD Series
Research in Management

Erasmus Research Institute of Management - ERiM
Rotterdam School of Management (RSM)
Erasmus School of Economics (ESE)
Erasmus University Rotterdam (EUR)
P.O. Box 1738, 3000 DR Rotterdam,
The Netherlands

Tel. +31 10 408 11 82
Fax +31 10 408 96 40
E-mail info@erim.eur.nl
Internet www.erim.eur.nl

Design & layout: B&T Ontwerp en advies (www.b-en-t.nl) Print: Haveka (www.haveka.nl)