
BEHAVIOR
GENETICS
ABDEL ABDELLAOUI

“To kill an error is as good a service as,
and sometimes even better than,
the establishing of a new truth or fact.
- *Charles Darwin*

Reading committee:
Prof. dr. G.J. van Ommen
Prof. dr. P. de Bakker
Prof. dr. P. van Lier
Prof. dr. M.C. Cornel
Prof. dr. P.D. Koellinger
Prof. dr. G.E. Davies
Dr. M. Distel

Paranymphs:
Farid Benmbarek
Mohamed Elhour

Acknowledgements:
I gratefully acknowledge support from the Center for Medical Systems Biology (CMSB, NWO Genomics), as well as from the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI–NL), the VU University's Institute for Health and Care Research (EMGO+) and the Neuroscience Campus Amsterdam (NCA). I had the opportunity to work with data collected by NTR and NESDA, which were funded through multiple grants, including funding from NWO, the European Community's Seventh Framework Program, and the European Science Council (ERC Advanced, 230374). Most of the genotyping was done at the Avera Institute for Human Genetics, Sioux Falls, South Dakota and Rutgers University Cell and DNA Repository, New Jersey (USA). Genotyping was funded in part by the Genetic Association Information Network (GAIN) of the Foundation for the US National Institutes of Health (NIMH, MH081802) and by the Grand Opportunity grants 1RC2MH089951-01 and 1RC2 MH089995-01 from the NIMH.

Printed by: Ipskamp Drukkers
Coverart: Jay Sunsmith
Layout: Abdel Abdellaoui

ISBN 978-94-6259-319-0

VRIJE UNIVERSITEIT

BEHAVIOR ↔ GENETICS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Psychologie en Pedagogiek
op woensdag 1 oktober 2014 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Abderrahman Abdellaoui

geboren te Amsterdam

promotoren:	Prof. dr. D.I. Boomsma
	Prof. dr. B.W.J.H. Penninx
copromotoren:	Prof. dr. E.C.J. de Geus
	Dr. J.J. Hottenga

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

My choice for getting involved in scientific research was initially motivated by an interest in the human brain and the key role it plays in our behavior, emotions, and cognition. My fascination began to grow when I realized that the "software" coding for the biological pieces of this complex puzzle resided in a single DNA molecule that was copied throughout our bodies. One of the more effective approaches scientists have used to unravel human brain functioning was by studying circumstances in which these functions fail,[1] which can lead to psychiatric illness, which contributes to ~13% of the global disease burden.[2] This motivated me to take on a PhD project focused on learning about biological processes underlying psychiatric illness that was made possible by the Netherlands Twin Registry (NTR) funding from Centre for Medical Systems Biology (CMSB: www.cmsb.nl). This thesis is about genetic variation in relation to several complex behavioral and psychiatric (endo)phenotypes, namely: educational attainment, the location where people choose to live, religion, thought problems, attention problems, and major depressive disorder.

Behavior ← Genetics

I started my PhD with twin studies, which have demonstrated that much of our behaviors, emotions, cognitions, and psychiatric disorders have a strong genetic component. Twin studies test the hypothesis that monozygotic (MZ) twins show a larger phenotypic resemblance than dizygotic (DZ) twins because of their larger genetic relatedness.[3; 4] **Part I** of this thesis describes two twin studies conducted on data collected by the Netherlands Twin Registry (NTR)[5; 6] to explore the methodology and assumptions leading to heritability estimates. Both chapters explored issues regarding the measurement and heritability of Thought Problems, an endophenotype for psychiatric traits. It was investigated whether the difference between paternal and maternal reports about their children's problem behavior can be incorporated into the heritability estimates (*chapter 2*), and whether self reports measure the same construct across age and gender, with the aim of giving reliable longitudinal heritability estimates for Thought Problems (*chapter 3*).

The recent technological advancements for large scale assessment of DNA variants in humans made it possible to measure several hundred thousand to more than a million variants at a genome-wide scale. While I was working on twin studies, NTR began to enrich the phenotype datasets with genetic variants measured on microarrays that include probes for single nucleotide polymorphisms (SNPs) and for copy number variants (CNVs). SNPs are single nucleotides that vary between members of a species and are the most abundant form of DNA sequence variation. CNVs are segments of DNA that vary in copy number (CN) between individuals, where 2 is the "normal" amount of copies (one from each parent). I was eager to

search those data for directly measured genetic variation influencing these traits. Since 2005, the year of the first GWAS,[7] genome-wide microarray data have been used predominantly for genome-wide association studies (GWASs),[8] and are often even referred to as "GWAS data" in colloquial speech. In a GWAS, each available SNP (usually at least a few hundred thousand) is tested individually in order to find variants associated with the phenotype of interest. My first experience with a GWAS was one conducted on Thought Problems during an internship. This study was done on a small sample of 549 subjects, which unsurprisingly did not yield any genome-wide significant results. Even if there were real signals among all the tested variants, they would have drowned among the hundreds of thousands of variants tested through multiple testing correction. Performing that many individual tests requires extremely small significance levels that are even hard to reach with sample sizes more than a hundred times greater than the GWAS I conducted.

Uniquely, NTR has collected micro-array data in both members of MZ twin pairs. Although monozygotic (MZ) twins are assumed to be genetically identical, they can occasionally show surprisingly large phenotypic discordances, even when considering very heritable traits.[9] **Part II** addresses the question whether rare post-twinning mutations that lead to genetic difference between MZ twins can be detected using microarray data, and whether these genetic differences can be linked to phenotypic differences. CNVs are more likely to differ between MZ twins due to their greater mutation rate, and perhaps more likely to cause phenotypic discordance, as they can affect more causal regions when they are larger.[10;11] Even though they are less numerous than SNPs, together they encompass a larger proportion of the genome.[12;13] To answer the question, CNVs were called in a group of 50 MZ twin pairs that were selected for concordance and discordance on Attention Problems, a heritable trait strongly related to ADHD, and searched for differences in CNVs between the MZ twins (*chapter 4*). In an attempt to describe at the population level what the level of discordance in MZ twins might be, we repeated this exercise, but in a larger set of ~1,100 unselected MZ pairs (*chapter 5*). An important question that is addressed in these two chapters is whether MZ differences that are found based on array data can be validated by quantitative PCR (qPCR). The possibility that genome-wide microarray technology lead to the discovery of false positive CNV mutations needed to be considered.

Behavior → Genetics?

In GWAS projects, it is difficult to find causal genetic variation by testing the association between a phenotype and each individual SNP. Even though hundreds of SNPs have been associated with many phenotypes, the vast majority of the expected additive genetic influences predicted by twin studies have not

yet been linked to individual SNPs analyzed in GWASs, especially for behavioral and psychiatric traits.[14; 15] It is estimated that there are between 6,300 and 10,200 SNPs that contribute to the risk for schizophrenia,[16] one of the more heritable psychiatric traits (~80%).[17] Recently, 128 consistent signals have been detected, which required a sample size of 35,000 cases and 47,000 controls,[18] making it the most successful psychiatric trait to date with respect to the number of discovered causal genetic variants. In contrast, after many GWAS efforts there have been no robust and replicable findings yet for major depressive disorder (MDD), with the latest large-scale effort being a mega-analysis of 16,000 cases and 60,000 controls.[19] Part of the problem, as noted above, is the level of significance required to declare a finding. Other problems may include the difficulty of capturing rare variants and structural variants with microarray chips, which are also likely to contribute additive genetic influences. This lack of explained heritability, also referred to as the "missing heritability", was an issue that needed to be addressed.[20; 21] Some doubt whether those additive genetic signals exist among the genetic variants at all and reject the heritability estimates of twin studies.[22] A large part of this question was likely answered recently by assessing the collective influences of common and rare SNPs on height, which has long served as a classic model for complex traits.[23-25] Yang et al[26] showed that the majority of the additive genetic signals were indeed very likely to be captured by the measured common SNPs, but that the individual SNP effects were just too small to be picked up with the current methodology, unless much larger sample sizes were available.

Instead of looking at individual SNPs, which requires much larger sample sizes than I had direct access to, I diverted from my initial goal of finding individual genetic variants that influence behavioral and/or psychiatric traits, and instead investigate in **parts III and VI** how genetic variants behave collectively. This half of the thesis is about the larger patterns of genetic variation in the Netherlands in relation to geographic location, complex (behavioral) phenotypes, and historical context. This had not yet been done in the Dutch population, but may potentially result in useful information for GWASs. GWASs have to account for genome-wide patterns of variation in their dataset, since systematic differences between phenotypic groups that do not necessarily reflect causal genetic variants may confound real signals.[27] Principal Component Analysis (PCA) was a promising statistical tool to start with, since it can mathematically summarize the largest patterns of variation for many (genetic) measurements points, and has been used to infer population structure from genetic data for several decades.[28] It is up to the researcher to find out what those PCs mean by interpreting what binds the measurement points that cluster together (by comparing the clustering patterns with phenotypic information or microarray quality metrics for example). Previous studies on genome-wide patterns of variation between more global populations found out that those patterns usually

reflect ancestry differences and correlate with geography. If migration levels (and thus gene flow) are relatively low, genetic similarity should decrease as distance increases.[29] These genome-wide signals for ancestry differences on a continental level are likely to contain diversifying evolutionary selection pressures.[30] Would similar ancestry signals also be detectable within a more homogeneous population in a small geographic area such as the Netherlands? **Part III** of the thesis reports the results of PCAs and a variety of follow-up analyses conducted to examine and explain the largest patterns of genetic variation in the Netherlands. *Chapter 6* summarizes genome-wide SNPs from microarray data on a large sample from a collaboration between the NTR and the Netherlands Study of Depression and Anxiety (NESDA).[31] *Chapter 7* explores genome-wide variation beyond SNPs by mapping genome-wide patterns of short insertions and deletions (<20 bp, also referred to as indels) and larger deletions (20 – 10,000 bp) from Next Generation Sequence (NGS) data from a collaborative project between four Dutch biobanks called *The Genome of the Netherlands* (GoNL) in the framework of the Dutch biobank infrastructure BBMRI-NL (www.bbmri.nl).[32; 33]

Another metric for genome-wide variation is the F_{roh} measure. F_{roh} is the proportion of the genome consisting of runs of homozygosity (ROHs: multiple contiguous homozygous SNPs). Offspring carry longer ROHs when their parents are more closely related. Long ROHs increase the risk of diseases influenced by recessive alleles,[34] which may also include psychiatric disorders.[35] F_{roh} is generally used to study the often harmful effects of inbreeding in humans and other animals,[36] but **part IV** addresses whether F_{roh} can vary in a relatively outbred human population due to social behavior and organization that can lead to differences in how closely parents are related (for example by causing differences in how much they share of the genome-wide patterns described in part III). *Chapter 8* and *chapter 9* focus on how collective behaviors can influence collective SNP variation on a population level by examining the relationships between F_{roh}, the major patterns of genome-wide variation detected in part III, migration, assortative mating, religion, major depressive disorder, and educational attainment.

New discoveries are more likely to be made when new data is available. These studies would not have been possible without the large and rich datasets I had access to.The sample size for the phenotypic data on Thought Problems in **part I** was ~18,000 and ~20,000 twin individuals respectively for both chapters, which is larger than the individual population sizes of ~71.8% of Dutch municipalities. The genotyping of MZ twins for **part II** on Affymetrix 6.0 microarray was done at Avera Institute for Human Genetics (AIHG; South Dakota, USA) and Rutgers University Cell and DNA Repository (RUCDR; New Jersey, USA) and probably represents the largest dataset to date of MZ pairs in which both members were

genotyped. For **parts III and IV** I had access to a large group of Dutch subjects from NTR and NESDA that were also genotyped at AIHG and RUCDR and had a sufficiently random geographical distribution across the Netherlands. In addition, **part III** also included NGS data provided by four Dutch biobanks: the LifeLines Cohort Study from Groningen, the Leiden Longevity Study, the Netherlands Twin Registry, and the Rotterdam Studies from the Erasmus University.

At the end of this thesis a summary of the main results and conclusions will be presented, and I will provide my perspective on the future of the field.

References

1. Kandel, E.R., Schwartz, J.H., and Jessell, T.M. (2000). Principles of neural science (McGraw-Hill New York).
2. Collins, P.Y., Patel, V., Joestl, S.S., March, D., Insel, T.R., Daar, A.S., Bordin, I.A., Costello, E.J., Durkin, M., et al. (2011). Grand challenges in global mental health. *Nature* 475, 27-30.
3. Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics* 3, 872-882.
4. Martin, N., Boomsma, D., and Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics* 17, 387-392.
5. van Beijsterveldt, C.E., Groen-Blokhuis, M., Hottenga, J.J., Franić, S., Hudziak, J.J., Lamb, D., Huppertz, C., de Zeeuw, E., Nivard, M., et al. (2013). The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin Research & Human Genetics* 16.
6. Willemsen, G., Vink, J.M., Abdellaoui, A., den Braber, A., van Beek, J.H., Draisma, H.H., van Dongen, J., van't Ent, D., Geels, L.M., et al. (2013). The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Research and Human Genetics* 16, 271-281.
7. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385-389.
8. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics* 90, 7-24.
9. Zwijnenburg, P.J., Meijers-Heijboer, H., and Boomsma, D.I. (2010). Identical but not the same: the value of discordant monozygotic twins in genetic research. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 153, 1134-1149.
10. van Ommen, G.-J.B. (2005). Frequency of new copy number variation in humans. *Nature Genetics* 37, 333-334.
11. Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. *Nature genetics* 39, S43-S47.
12. Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* 453, 56-64.
13. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318, 420-426.
14. Burmeister, M., McInnis, M.G., and Zöllner, S. (2008). Psychiatric genetics: progress amid controversy. *Nature Reviews Genetics* 9, 527-540.
15. Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics* 13, 537-551.
16. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kahler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., et al. (2013). Genome-wide association analysis identifies 13

new risk loci for schizophrenia. *Nature Genetics* 45, 1150-1159.
17. Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry* 60, 1187-1192.
18. Ripke, S. (2013). Psychiatric Genomics Consortium quadruples schizophrenia GWAS sample-size to 35,000 cases and 47,000 controls. *XXIst World Congress of Psychiatric Genetics: Redefining mental illness through genetics* (Boston, Massachusetts).
19. Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne, E.M., Blackwood, D.H., Boomsma, D.I., et al. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular psychiatry* 18, 497-511.
20. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747-753.
21. Maher, B. (2008). The case of the missing heritability. *Nature* 456, 18-21.
22. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* 109, 1193-1198.
23. Fisher, R.A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399-433.
24. Galton, F. (1886). Hereditary stature. *Nature* 33, 317.
25. Hemani, G., Yang, J., Vinkhuyzen, A., Powell, J.E., Willemsen, G., Hottenga, J.-J., Abdellaoui, A., Mangino, M., Valdes, A.M., et al. (2013). Inference of the Genetic Architecture Underlying BMI and Height with the Use of 20,240 Sibling Pairs. *The American Journal of Human Genetics* 93, 865-875.
26. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42, 565-569.
27. Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* 11, 459-463.
28. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The history and geography of human genes (Princeton university press).
29. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., et al. (2008). Genes mirror geography within Europe. *Nature* 456, 98-101.
30. McEvoy, B.P., Montgomery, G.W., McRae, A.F., Ripatti, S., Perola, M., Spector, T.D., Cherkas, L., Ahmadi, K.R., Boomsma, D., et al. (2009). Geographical structure and differential natural selection among North European populations. *Genome Research* 19, 804-814.
31. Penninx, B.W., Beekman, A.T., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W., et al. (2008). The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research* 17, 121-140.
32. Boomsma, D.I., Wijmenga, C., Slagboom, E.P., Swertz, M.A., Karssen, L.C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., et al. (2014). The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics* 22, 221-227.
33. The Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*.
34. Charlesworth, D., and Willis, J.H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics* 10, 783-796.
35. Keller, M.C., Simonson, M.A., Ripke, S., Neale, B.M., Gejman, P.V., Howrigan, D.P., Lee, S.H., Lencz, T., Levinson, D.F., et al. (2012). Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genetics* 8, e1002656.
36. Keller, M.C., Visscher, P.M., and Goddard, M.E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189, 237-249.

PART I

TWIN STUDIES: RATER EFFECTS, MEASUREMENT INVARIANCE, AND LONGITUDINAL HERITABILITY OF THOUGHT PROBLEMS

CHAPTER 2

GENETIC INFLUENCES ON THOUGHT PROBLEMS IN 7-YEAR-OLDS: A TWIN-STUDY OF GENETIC, ENVIRONMENTAL, AND RATER EFFECTS

This chapter is based on:

Abdel Abdellaoui, Meike Bartels, James J Hudziak, Patrizia Rizzu, Toos CEM van Beijsterveldt, Dorret I Boomsma: Genetic Influences on Thought Problems in 7-Year-Olds: a Twin-Study of Genetic, Environmental and Rater Effects. *Twin Research and Human Genetics 2009*; 11(6):571-8.

Abstract

The Thought Problems scale (TP) of the CBCL assesses symptoms such as hallucinations, OCD-symptoms and strange thoughts/behaviors and has also been associated with other behavioral disorders. This study uses parental reports to examine the etiology of variation in TP, about which relatively little is known, in 7-year old twins.

Parental ratings on TP were collected in 8,962 seven-year-old twin pairs. Because the distribution of TP scores was highly skewed scores were categorized into 3 classes. The data were analyzed under a threshold liability model with genetic structural equation modeling. Ratings from both parents were simultaneously analyzed to determine the rater agreement phenotype (or common phenotype [TPc]) and the rater specific phenotype [TPs] which represents rater disagreement caused by rater bias, measurement error and/or a unique view of the parents on the child's behavior.

Scores on the TP-scale varied as a function of rater (fathers rated fewer problems), sex (boys scored higher) and zygosity (DZ twins scored higher). The TPc explained 67% of the total variance in the parental ratings. Variation in TPc was influenced mainly by the children's genotype (76%). Variance in TPs also showed a contribution of genetic factors (maternal reports: 61%, paternal reports: 65%), indicating that TPs does not only represent rater bias. Shared environmental influences were only found in the TPs. No sex-differences in genetic architecture were observed.

These results indicate an important contribution of genetic factors to thought problems in children as young as 7 years.

Introduction

The Thought Problems scale (TP) of the Child Behavior Checklist (CBCL)[1] is an empirically derived set of items that cover symptoms such as hallucinations, obsessive-compulsive symptoms and strange thoughts and behaviors. The strong statistical relation between the TP-items was seen in the factor analysis through which the scale was derived.[1] The internal consistency of the TP-scale was shown to be sufficient in the CBCL.[2;3] Compared to the other empirically derived CBCL scales, the TP-scale received relatively little attention. In a longitudinal study of Ferdinand et al (2001),[4] Thought Problems in childhood were associated with substance abuse (alcohol and tobacco) in young adulthood. Other studies showed TP in children to be associated with disorders,such as Obsessive-Compulsive Disorder (OCD),[5] Multiple Complex Developmental Disorder (MCDD)[6] and Fragile X syndrome.[7] The TP-scale also predicted DSM-III-R diagnoses of simple phobia, social phobia, separation anxiety disorder, mood disorders and psychotic disorders.[8] There have been a few studies on the heritability of the TP-scale in young twins that found evidence for significant heritability.[2;3;9-11] The aim of this study is to obtain more insight into the genetic and environmental contributions to variation in TP in a large group of 7-year-old twins. We obtained ratings from both parents and modeled the extent to which both parents agreed on the presence or absence of TP.

Parental ratings, such as those collected with the CBCL, provide meaningful information about a child's behavioral and emotional problems. Parents observe their children for long periods and in natural situations, which makes them ideal informants regarding their children's behavior. Studies that collect information from both parents tend to show that parents agree for a substantial part in their assessment of behavioral problems in their children. An analysis by Achenbach et al (1987)[12] showed a correlation of 0.60 between the paternal and maternal ratings of the same child. Although correlations are high, they are not perfect. To explore the processes underlying disagreements in parental ratings, Hewitt et al (1992)[13] developed a series of models, in which the disagreement of the parents' reports reflect rater bias and a unique accurately assessed part of the child's behavior.

Rater bias may be caused by the parents' own characteristics (a projection bias), or by parents' response biases (e.g. stereotyping, employing different normative standards or having certain response styles, i.e. judging certain types of behavior more or less severely). The unique views on the phenotype may arise firstly because parents observe their child in distinct situations and environments. For instance, the mother could have the task of bringing the child to school, while the father is the one who accompanies the child to soccer games. Secondly, the way the mother interacts with the child may differ from the way the father interacts with

the child.[12] Another cause for different views in parent reports could be that men are more sensitive to different kinds of input from their children than women.[14-16]

A number of quantitative genetic studies have examined problem behavior in children with the CBCL, with the main focus on internalizing and externalizing behavior problems,[9;17-26] and a number of these studies have modeled the agreement and disagreements among parental reports. For example, in a study by Rowe and Kandel (1997),[27] parents rated the internalizing and externalizing behavior of their two oldest children (between 9 and 17 years old). "Individual view" and "shared view" models were used. The parents assessed similar aspects of the child's behavior, and in addition the mother and father ratings contained a significant individual view component.

Twin studies provide the opportunity to analyze whether the variance of the unique view of the parental assessments can be explained by genetic factors. If genetic factors are found, this implies that the unique view of the parents partly represents real behavior of the child. This rules out the possibility of the rater disagreements containing only rater bias. Hewitt et al. (1992)[13] studied internalizing behavior in pre-pubertal (8-11 years) and pubertal twins (12-16 years), and found evidence for such genetic effects. Dutch twin studies on internalizing and externalizing behaviors also found that the unique viewpoint of parents does not solely reflect rater bias.[17;18;28-30]

When data are derived from questionnaires developed to measure the degree of dysfunctional behavior such as the CBCL, a large degree of skewness is often observed. These non-normal distributions can be explained by the fact that in symptom data a majority of the subjects displays few or no symptoms.[31] This is especially the case for the TP subscale, which is one of the subscales with the lowest mean scores in general population samples.[1] Logarithmic and square root transformations are often not enough to correct for this non-normality. Categorizing the observations and analyzing the data with a threshold model has shown to be a successful way to decrease bias in parameter estimates.[32] A disadvantage of categorizing the data may be that it reduces the statistical power of the analysis, and therefore large sample sizes are required.

This study estimates the genetic and environmental influences on variation in thought problems in a large group of 7-year-old Dutch twin pairs, while taking the agreement and disagreements between the parent reports into account. To overcome biased estimates due to skewness of the phenotype, the continuous TP scale was categorized into three classes (low, middle, high) and analyzed with threshold models.

Methods

Subjects

The participants were all registered with the Netherlands Twin Registry (NTR), which was established by the Department of Biological Psychology at the Vrije Universiteit in Amsterdam.[17;33] Parents of young twins receive questionnaires when their twins are 1, 2, 3, 5, 7, 10 and 12 years old. For this study, data of 7-year-old twin pairs from birth cohorts 1986-1997 were used. The questionnaires were mailed to the parents within three months of the twins' 7th birthday. Reminders were sent after two to three months. The response rate was 62 % (N = 8,962). Reasons for families not participating at this wave of data collection vary; some families request to no longer take part in the research due to various reasons, while other families move to new addresses without notifying the registry staff. Bartels et al (2007)[17] showed that the drop-out was largely random according to the definition of Little and Rubin (1987),[34] making generalizations of the results more valid.

Zygosity was determined for 1492 same-sex twin-pairs by blood group (n = 389) or DNA polymorphisms (n = 1,103). The zygosity of the other same-sex twins was determined using a discriminant analysis of questionnaire items answered by the parents. The questionnaire led to correct classification of the zygosity in about 93 % of the cases.[35]

The sample contained 1,466 monozygotic male (MZM), 1,516 dizygotic male (DZM), 1,675 monozygotic female (MZF), 1,445 dizygotic female (DZF), and 2,860 dizygotic opposite-sex (DOS) twin pairs. Response rates for mothers were higher than for fathers and the group could be further divided into twin pairs for which both mother and father had replied (1,053 MZM, 1,105 DZM, 1,226 MZF, 999 DZF, 2,002 DOS), pairs with only mother-reports (395 MZM, 392 DZM, 423 MZF, 424 DZF, 805 DOS) and pairs with only father-reports (n = 113).

Measures

The Child Behavior Checklist (CBCL 4-18)[1] was used to assess emotional and behavioral problems including Thought Problems. The questionnaire consists of 113 items, and measures behavior during the preceding 6 months. There are 7 TP items, such as "shows odd behavior", "hears sounds that are not there" or "sees things that are not there". The reliability of the CBCL has been confirmed by in Dutch epidemiological samples. The 2-week test-retest correlation for TP was 0.74.[36] The TP scores were transformed from continuous to categorical data with two thresholds to limit the number of categories to three (low, middle and high levels of TP). The thresholds were chosen in such a way that there were no empty cells in the contingency tables of twin 1 versus twin 2 scores.

Genetic analyses

Data from monozygotic (MZ) and dizygotic (DZ) twins were used to decompose the variation in the liability of Thought Problems into a contribution of additive genetic, shared environmental and non-shared environmental components. The categorical trait Thought Problems was modeled to have an unobserved, underlying continuous distribution with 2 thresholds that divide the distribution into 3 categories (low, middle, high). Such underlying distributions have been termed the liability or vulnerability.[37] The continuous variation in liability may be genetic or environmental in origin. The mean and variance of the liability distribution were standardized with mean zero and unit variance.

The additive genetic influence (A) on the variation in liability represents the sum of allelic influences at each locus in the genome contributing to the phenotype. The environmental influences can be shared and non-shared. Shared environment is common (C) to both twins growing up in the same family. Non-shared environmental influences (E) are unique to each twin and do not lead to twin resemblance.

Model Fitting

With multiple raters (2 parents in this study), the variance of the liability distribution can be distinguished into 2 parts that represent agreement (commonly rated phenotype by both parents) and disagreement (unique parental views). Both parts of the variance can be influenced by genetic (A), shared (C) and non-shared (E) factors and not necessarily to the same degree. We use the psychometric model[13] as presented in Figure 1. All variables enclosed in circles are latent (unobserved) that influence observed traits (enclosed in rectangles). The influence of latent variables on other latent variables or on observed traits is given by factor loadings 'a', 'c', and 'e'. These factor loadings come with 'm' or 'f ' if the latent variable influences a unique view of the father or the mother.

Mx[38] was used to obtain parameter estimates for thresholds and factor loadings. Genetic models were fitted to the raw data with maximum likelihood estimation procedures. In a saturated model thresholds and polychoric twin correlations were estimated separately for MZM, DZM, MZF, DZF, DOSMF and DOSFM groups. This model was also used to test whether the thresholds could be constrained to be equal for mother and father ratings, for MZ and DZ-twins, for boys and girls and for the youngest and the oldest twin. Next, the fit of the psychometric model was compared with the fit of the saturated model.

The psychometric model was used to test for sex differences in factor loadings. Next, the significant contribution of the common and unique C variance components was tested. The significance of the unique 'A' component was tested, to investigate whether the rater specific parts represented not only measurement

error and rater bias, but also reflected meaningful variation.

Significance of the estimates was established by comparing the full model with a simplified model. The more parsimonious nested model is chosen over the full model when the analysis shows a low non-significant χ^2 test statistic ($p > .05$). In addition to the χ^2 test statistic, Akaike's Information Criterion (AIC = $\chi^2 - 2$ x degrees of freedom) was computed. The lower the AIC, the better the fit of the model to the observed data.

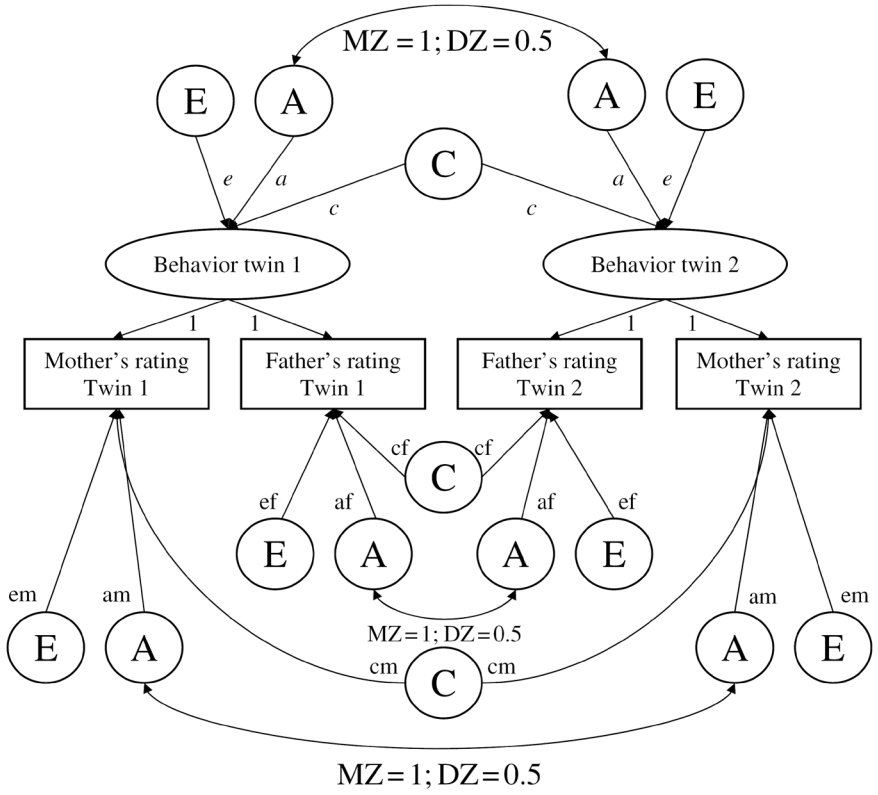


Figure 1: The psychometric-model

Results

Description of the data and threshold differences

For descriptive purposes, Table 1 summarizes the means and variances for the untransformed measures of Thought Problems by sex for mother and father reports.

Table 2 shows the thresholds from the saturated model in Mx. The thresholds are higher for all father ratings, indicating a tendency for mothers to give higher ratings (i.e. lower thresholds) for both boys and girls (χ^2 (24) = 84.19, $p < .01$, see Table 4). The thresholds also indicate more TP (i.e. lower thresholds) for the DZ-twins than for the MZ-twins (χ^2 (16) = 29.75, $p = .02$). Finally, Table 2 shows higher thresholds for the liability for TP in girls than in boys (i.e. fewer thought problems for girls) (χ^2 (24) = 42.13, $p = .01$). The overall differences between the thresholds of the youngest and oldest twin were not significant (χ^2 (24) = 32.35, $p = .11$).

Table 1: Means and standard deviations (SD) for untransformed maternal and paternal ratings of Child Behavior Checklist Thought Problems (age 7) for first- and second born twins.

	Mother ratings		Father ratings	
	Boys Mean (SD)	Girls Mean (SD)	Boys Mean (SD)	Girls Mean (SD)
Oldest twin	.39 (.97)	.33 (.86)	.31 (.80)	.24 (.70)
Youngest twin	.40 (.95)	.31 (.80)	.34 (.82)	.25 (.74)

Correlations

From the saturated model polychoric correlations between twins, between raters and cross-twin cross-rater were obtained. Table 3 shows correlations between twins rated by the same parent in the first and second columns. The last four columns show the cross-correlations between twins each rated by a different parent. The inter-parent correlations were comparable for both first- and second-born twins. On average the inter-parent correlation is .66, which is in the same range as the parental agreement found in previous studies of about .60.[1]

Model Fitting Results for the Psychometric Model

The model fitting results of the simultaneous analysis of maternal and paternal ratings are summarized in Table 4. Sex differences in factor loadings were not detected (χ^2 (7) = 7.35, $p = .60$). The common environmental effects on the common view of parents were not significant (χ^2 (1) = .972, $p = .32$), but the contribution of

Table 2: Maximum likelihood estimates of thresholds for the three levels of liability for thought problems (low, middle, high), with the scores ranging from 0 to 2.

		Mother		Father	
		Oldest twin	Youngest twin	Oldest twin	Youngest twin
MZM	Threshold 1	.89	.81	.94	.90
	Threshold 2	1.48	1.40	1.55	1.48
DZM	Threshold 1	.72	.67	.81	.70
	Threshold 2	1.31	1.27	1.46	1.37
MZF	Threshold 1	.90	.86	1.04	1.00
	Threshold 2	1.42	1.52	1.67	1.78
DZF	Threshold 1	.77	.82	.92	.90
	Threshold 2	1.34	1.35	1.52	1.61
DOSMF	Threshold 1	.70	.94	.77	.99
	Threshold 2	1.24	1.56	1.41	1.62
DOSFM	Threshold 1	.92	.71	1.04	.82
	Threshold 2	1.49	1.34	1.61	1.44

Note: MZM/DZM = monozygotic/dizygotic males, MZF/DZF = monozygotic/dizygotic females. DOS = dizygotic opposite sex twins

C to the variance of the unique views of father and mother was significant (χ^2 (2) = 19.54, $p < .01$). Dropping the additive genetic component on the unique views also gave a significant deterioration of fit (χ^2 (2) = 18.09, $p < .01$), indicating a significant contribution of A to the variance of the rater specific parts of the phenotype. Table 5 summarizes the estimates of genetic and environmental influences. The largest part of the variance in liability for thought problems could be explained by genetic factors.

The commonly assessed part of the phenotype explains 67% (51% [=A] + 16% [=E], see Table 5) of the total variance in maternal and paternal ratings (total variance = sum of all maternal/paternal parameters) and was influenced mainly by the children's genotype. Heritability of the commonly assessed TP phenotype was 76% (51% [=A] / 67% [=A+E]). The remaining variance of the commonly assessed phenotype was explained by non-shared environmental factors (16% [=E] / 67% [=A+E] = 24%).

The rater specific genetic influences seem somewhat higher for the father than the mother ratings (14 % vs. 10 %). These significant genetic influences on the unique views indicate that fathers and mothers asses reliable and rater specific information regarding TP in their children. The rater specific shared environmental influences explain about 13 % of the variance both parent-ratings. This is the part of the variance which could represent rater bias. Higher maternal specific

Table 3: Correlations (ratings by the same parent) and cross-correlations (ratings given by different parents) between the twins and between the parents

	Same Rater		Different Raters			
	Twins		Twins		Interparent	
	M/M	F/F	M/F	F/M	O	Y
MZM	.72	.78	.49	.46	.60	.65
DZM	.47	.49	.32	.29	.73	.67
MZF	.74	.76	.46	.51	.65	.67
DZF	.46	.42	.22	.23	.67	.59
DOSMF	.41	.44	.26	.23	.72	.60
DOSFM	.44	.53	.35	.33	.71	.69

Note: MZM/DZM = monozygotic/ dizygotic males, MZF/DZF = monozygotic/ dizygotic females. DOS = dizygotic opposite sex twins. Same rater twins = correlation between the oldest and the youngest twin, rated by M/M = mothers or F/F = fathers. Different raters twins = cross-correlation: either oldest twin rated by mothers and youngest by fathers (M/F) or the other way around (F/M). Different raters interparent: O = correlation between mother and father ratings for the oldest child;Y = idem for the youngest child.

non-shared environmental influences are observed (10 % vs. 6 %), which could reflect reliable information about the non-shared environmental influences on the variance of the phenotype, but may also include measurement error.

Discussion

This study examined the influence of genetic and environmental factors on thought problems in 7-year-old twins, based on parental reports. The analyses modeled genetic and environmental influences on both the commonly agreed upon phenotype and on unique views of the parents. Parents agreed to a large extent on the occurrence of thought problems (TP) in their twin offspring. The commonly assessed phenotype explained 67% (51% [=A] + 16% [=E]) of the total variance in maternal and paternal ratings and is influenced mainly by genetic factors. Heritability of the commonly assessed TP phenotype was 76% (51% [=A] / 67% [=A+E], see Table 5). There was no evidence that shared family environment contributed to variance of the common phenotype. In contrast, shared environmental influences contributed to the unique views of parents. The rater specific shared environmental influences may be due to rater bias, but explain only a small part of the total variance (13% [=C] / 33% [=A+C+E], see Table 5). The heritability estimates of this study are slightly higher than previously found heritabilities for the TP-scale,[3; 9; 11] which used mostly maternal reports. The large

Table 4: Summary of model fitting results of simultaneous analysis of paternal and maternal ratings of Thought Problems

	-2 LL	Nr. of parameters	df	$\Delta\chi^2$	Δdf	*P*	AIC	Compared to:
Saturated Model:								
Saturated	33966.377	84	30470				-26973.623	
Equal thresholds for:								
Mother and Father	34050.563	60	30494	84.186	24	<.01	-26937.437	Saturated
MZ and DZ	33996.124	68	30486	29.746	16	.02	-26975.876	Saturated
Boys and Girls	34008.503	60	30494	42.125	24	.01	-26979.497	Saturated
Oldest/Youngest	33998.724	60	30494	32.347	24	.11	-26989.276	Saturated
Psychometric Model:								
Full psychometric	33985.752	66	30492	19.375	22	.62	-26998.248	Saturated
Simplification Psychometric model:								
Sex differences								
No sex diff.	33993.106	57	30499	7.354	7	.60	-27008.894	Full Psychometric
Factor Estimates								
No sex diff./No common C	33994.078	56	30500	.972	1	.32	-27009.992	No sex differences
No sex diff./No unique C	34015.649	55	30501	19.543	2	<.01	-26990.351	No sex differences
No sex diff./No unique A	34011.198	55	30501	18.092	2	<.01	-26994.802	No sex differences

Table 5: Variation explained by genetic, shared environmental and non-shared environmental factors

	Mother	Father
Genetic Factors		
Common	51 %	51 %
Unique	10 %	14 %
Shared Environmental Factors		
Common	-	-
Unique	13 %	13 %
Non-shared Environmental Factors		
Common	16 %	16 %
Unique	10 %	6 %

heritability estimates found for boys by Kuo et al (2003)[2] - using parental reports which were also mostly maternal - were more in line with our findings. The gender differences in genetic and environmental influences found by Kuo et al (2003)[2] were not replicated in our study. Kuo et al (2003)[2] also observed significant small non-additive genetic influences for boys. Shared environmental influences were not significant in any of the previous studies.[9; 11] In the study by Polderman et al (2006),[10] TP was rated by teachers and familial influences were detected, but to explain the familial influences, it was not possible to distinguish between common environment and genetic factors.

The genetic influences on the variance of the rater-specific parts are higher for the father reports in the analyses. Previous studies of internalizing and externalizing disorders in young twins usually found these estimates to be larger in maternal ratings.[18; 28-30; 39] This may support the notion that that there are sex differences in the perception of parents with respect to this phenotype in their children and that fathers may add important extra information regarding the Thought Problems phenotype. Sex differences in human perception have been suggested from a very early age onwards.[15]

We observed sex differences in TP with boys showing more TP than girls. A similar sex difference is found in OCD, which is associated with TP.[40-42] We also observed differences between mono- and dizygotic twins, with MZ twins obtaining lower TP scores than DZ twins. This suggests the presence of negative social interactions in twin pairs, since under interaction the prevalence rates for a categorical variable between MZ en DZ twins are expected to differ if that trait has an underlying continuous distribution.[43]

The distribution of the TP scale was highly skewed and we analyzed the data using a threshold model. The means of the untransformed (continuous) data

shown in Table 1 are also indeed lower than most other CBCL-subscales,[1] which indicates that a majority of the subjects display few or no symptoms, which is expected in such a skewed distribution.[31] A simulation study by Derks, Dolan and Boomsma (2004)[32] showed that analysis of L-shaped distributed data results in an underestimation of additive genetic, and shared environmental influences and an overestimation of non-shared environmental effects. After conducting the Mx-analyses on the untransformed data (results not shown), these differences were also found when comparing those estimates with the estimates made through analysis of the categorical data. Categorical analysis also has disadvantages. First, the statistical power is reduced, and requires large samples. This study used a large sample of 8,962 twin pairs. Another disadvantage of the categorical data analyses is that they are computationally more demanding.

The most important finding from these data is the substantial heritability that is observed in young children. This strengthens the notion that the TP-scale measures a true syndrome. Other ways to support this finding may include linkage and genome wide association analyses. Further study is needed to investigate the differences between the parental reports. Further research is also needed to obtain insight into estimates of heritability in other age groups and to obtain information on the relation between Thought Problems and other disorders such as autism, OCD and psychotic disorders, to examine whether the Thought Problems phenotype could also have a predictive value for these disorders. It is important that such studies in children include both maternal and paternal reports.

References

1. Achenbach, T.M. (1991). Manual for the Child Behavior Checklist/4-18 and 1991 profile (Burlington, VT: University of Vermont, Department of Psychiatry).
2. Kuo, P.-H., Lin, C.C., Yang, H.-J., Soong, W.-T., and Chen, W.J. (2004). A twin study of competence and behavioral/emotional problems among adolescents in Taiwan. *Behavior Genetics* 34, 63-74.
3. Lin, C.C., Kuo, P.-H., Su, C.-H., and Chen, W.J. (2006). The Taipei Adolescent Twin/sibling Family Study I: behavioral problems, personality features, and neuropsychological performance. *Twin Research and Human Genetics* 9, 890-894.
4. Ferdinand, R.F., Blüm, M., and Verhulst, F.C. (2001). Psychopathology in adolescence predicts substance use in young adulthood. *Addiction* 96, 861-870.
5. Geller, D. (2004). Re-examining comorbidity of obsessive compulsive and attention-deficit hyperactivity disorder using an empirically derived taxonomy. *European Child & Adolescent Psychiatry* 13, 83-91.
6. de Bruin, E.I., de Nijs, P.F., Verheij, F., Hartman, C.A., and Ferdinand, R.F. (2007). Multiple complex developmental disorder delineated from PDD-NOS. *Journal of autism and developmental disorders* 37, 1181-1191.
7. Hessl, D., Dyer-Friedman, J., Glaser, B., Wisbeck, J., Barajas, R.G., Taylor, A., and Reiss, A.L. (2001). The influence of environmental and genetic factors on behavior problems and autistic symptoms in boys and girls with fragile X syndrome. *Pediatrics* 108, e88-e88.
8. Kasius, M.C., Ferdinand, R.F., Berg, H., and Verhulst, F.C. (1997). Associations between different diagnostic approaches for child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry* 38, 625-632.
9. Edelbrock, C., Rende, R., Plomin, R., and Thompson, L.A. (1995). A twin study of competence and problem behavior in childhood and early adolescence. *Journal of Child Psychology and Psychiatry* 36, 775-785.
10. Polderman, T.J., Posthuma, D., De Sonneville, L.M., Verhulst, F.C., and Boomsma, D.I. (2006). Genetic analyses of teacher ratings of problem behavior in 5-year-old twins. *Twin Research and Human Genetics* 9, 122-130.
11. Schmitz, S., Fulker, D.W., and Mrazek, D.A. (1995). Problem behavior in early and middle childhood: An initial behavior genetic analysis. *Journal of Child Psychology and Psychiatry* 36, 1443-1458.
12. Achenbach, T.M., McConaughy, S.H., and Howell, C.T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin* 101, 213.
13. Hewitt, J.K., Silberg, J., Neale, M., Eaves, L., and Erickson, M. (1992). The analysis of parental ratings of children's behavior using LISREL. *Behavior Genetics* 22, 293-317.
14. Baron-Cohen, S., and Hammer, J. (1997). Is autism an extreme form of the" male brain"? *Advances in Infancy research* 11, 193-218.
15. Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki, A., and Ahluwalia, J. (2000). Sex differences in human neonatal social perception. *Infant Behavior and Development* 23, 113-118.
16. Rosenthal, R., Hall, J.A., DiMatteo, M.R., Rogers, P.L., and Archer, D. (1979). Sensitivity to nonverbal communication: The PONS test (Johns Hopkins University Press Baltimore).
17. Bartels, M., Beijsterveldt, C., Derks, E.M., Stroet, T.M., Polderman, T.J., Hudziak, J.J., and Boomsma, D.I. (2007). Young Netherlands Twin Register (Y-NTR): a longitudinal multiple informant study of problem behavior. *Twin Research and Human Genetics* 10, 3-11.
18. Bartels, M., Boomsma, D.I., Hudziak, J.J., Rietveld, M.J., van Beijsterveldt, T.C., and van den Oord, E.J. (2004). Disentangling genetic, environmental, and rater effects on internalizing and externalizing problem behavior in 10-year-old twins. *Twin Research* 7, 162-175.
19. Gjone, H., and Stevenson, J. (1997). The association between internalizing and externalizing behavior in childhood and early adolescence: Genetic or environmental common

influences? *Journal of Abnormal Child Psychology* 25, 277-286.
20. Hudziak, J.J., Rudiger, L.P., Neale, M.C., Heath, A.C., and Todd, R.D. (2000). A twin study of inattentive, aggressive, and anxious/depressed behaviors. *Journal of the American Academy of Child & Adolescent Psychiatry* 39, 469-476.
21. Leve, L.D., Winebarger, A.A., Fagot, B.I., Reid, J.B., and Goldsmith, H.H. (1998). Environmental and genetic variance in children's observed and reported maladaptive behavior. *Child Development* 69, 1286-1298.
22. Oord, E.J., Koot, H.M., Boomsma, D.I., Verhulst, F.C., and Orlebeke, J. (1995). A Twin-Singleton Comparison of Problem Behaviour in 2-3-Year-Olds. *Journal of Child Psychology and Psychiatry* 36, 449-458.
23. Schmitz, S., and Mrazek, D.A. (2001). Genetic and environmental influences on the associations between attention problems and other problem behaviors. *Twin Research* 4, 453-458.
24. Silberg, J.L., Erickson, M.T., Meyer, J.M., Eaves, L.J., Rutter, M.L., and Hewitt, J.K. (1994). The application of structural equation modeling to maternal ratings of twins' behavioral and emotional problems. *Journal of Consulting and Clinical Psychology* 62, 510.
25. van der Valk, J.C., Verhulst, F.C., Stroet, T.M., and Boomsma, D.I. (1998). Quantitative genetic analysis of internalising and externalising problems in a large sample of 3-year-old twins. *Twin Research* 1, 25-33.
26. Zahn-Waxler, C., Schmitz, S., Fulker, D., Robinson, J., and Emde, R. (1996). Behavior problems in 5-year-old monozygotic and dizygotic twins: Genetic and environmental influences, patterns of regulation, and internalization of control. *Development and Psychopathology* 8, 103-122.
27. Rowe, D.C., and Kandel, D. (1997). In the eye of the beholder? Parental ratings of externalizing and internalizing symptoms. *Journal of Abnormal Child Psychology* 25, 265-275.
28. Bartels, M., Hudziak, J.J., Boomsma, D.I., Rietveld, M.J., and Van Beijsterveldt, T.C. (2003). A study of parent ratings of internalizing and externalizing problem behavior in 12-year-old twins. *Journal of the American Academy of Child & Adolescent Psychiatry* 42, 1351-1359.
29. Van der Valk, J., Van den Oord, E., Verhulst, F., and Boomsma, D. (2001). Using Parental Ratings to Study the Etiology of 3-year-old Twins' Problem Behaviors: Different Views or Rater Bias? *Journal of Child Psychology and Psychiatry* 42, 921-931.
30. Van der Valk, J., Van den Oord, E., Verhulst, F., and Boomsma, D. (2003). Using shared and unique parental views to study the etiology of 7-year-old twins' internalizing and externalizing problems. *Behavior Genetics* 33, 409-420.
31. Oord, E.J., Pickles, A., and Waldman, I.D. (2003). Normal variation and abnormality: an empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry* 44, 180-192.
32. Derks, E.M., Dolan, C.V., and Boomsma, D.I. (2004). Effects of censoring on parameter estimates and power in genetic modeling. *Twin Research* 7, 659-669.
33. Boomsma, D.I., De Geus, E.J., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.-J., Posthuma, D., Van Beijsterveldt, T.C., Hudziak, J.J., et al. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics* 9, 849-857.
34. Little, R.J., and Rubin, D.B. (1987). Statistical Analysis With Missing Data (New York: Wiley).
35. Rietveld, M., van Der Valk, J., Bongers, I., Stroet, T., Slagboom, P., and Boomsma, D. (2000). Zygosity diagnosis in young twins by parental report. *Twin Research* 3, 134-141.
36. Verhulst, F.C., van der Ende, J., and Koot, J.M. (1996). Handleiding voor de CBCL/4-18 (Afdeling Kinder-en Jeugdpsychiatrie, Sophia Kinderziekenhuis/Academisch Ziekenhuis Rotterdam/Erasmus Universiteit Rotterdam).
37. Falconer, D. (1981). Introduction to Quantitative Genetics. *Hong Kong: Longman*.
38. Neale, M.C., Boker, S.M., Xie, G., and Maes, H.M. (1999). Mx: Statistical modeling. *Richmond, Virginia: Department of Psychiatry*.
39. Boomsma, D., Van Beijsterveldt, C., and Hudziak, J. (2005). Genetic and environmental influences on anxious/depression during childhood: a study from the Netherlands Twin Register. *Genes, Brain and Behavior* 4, 466-481.
40. Geller, D.A., Biederman, J., Jones, J., Shapiro, S., Schwartz, S., and Park, K.S. (1998). Obsessive-

compulsive disorder in children and adolescents: a review. *Harvard Review of Psychiatry* 5, 260-273.
41. Hanna, G.L. (1995). Demographic and clinical features of obsessive-compulsive disorder in children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry* 34, 19-27.
42. Zohar, A.H. (1999). The epidemiology of obsessive-compulsive disorder in children and adolescents. *Child and Adolescent Psychiatric Clinics of North America*.
43. Carey, G. (1992). Twin imitation for antisocial behavior: Implications for genetic and family environment research. *Journal of Abnormal Psychology* 101, 18.

CHAPTER 3

THOUGHT PROBLEMS FROM ADOLESCENCE TO ADULTHOOD: MEASUREMENT INVARIANCE AND LONGITUDINAL HERITABILITY

This chapter is based on:

Abdel Abdellaoui, Marleen H M de Moor, Lot M Geels, Jenny H D A van Beek, Gonneke Willemsen, Dorret I Boomsma: Thought Problems from Adolescence to Adulthood: Measurement Invariance and Longitudinal Heritability. *Behavior Genetics* 2011; 42(1):19-29.

Abstract

This study investigates the longitudinal heritability in Thought Problems (TP) as measured with ten items from the Adult Self Report (ASR). There were ~9,000 twins, ~2,000 siblings and ~3,000 additional family members who participated in the study and who are registered at the Netherlands Twin Register. First an exploratory factor analysis was conducted to examine the underlying factor structure of the TP-scale. Then the TP-scale was tested for measurement invariance (MI) across age and sex. Next, genetic and environmental influences were modeled on the longitudinal development of TP across three age groups (12-18, 19-27 & 28-59 year olds) based on the twin and sibling relationships in the data.

An exploratory factor analysis yielded a one-factor solution, and MI analyses indicated that the same TP-construct is assessed across age and sex. Two additive genetic components influenced TP across age: the first influencing TP throughout all age groups, while the second arises during young adulthood and stays significant throughout adulthood. The additive genetic components explained 37% of the variation across all age groups. The remaining variance (63%) was explained by unique environmental influences. The longitudinal phenotypic correlation between these age groups was entirely explained by the additive genetic components.

We conclude that the TP-scale measures a single underlying construct across sex and different ages. These symptoms are significantly influenced by additive genetic factors from adolescence to late adulthood.

Introduction

The Thought Problems (TP) scale is one of the empirically defined syndrome scales from the Achenbach System of Empirically Based Assessment (ASEBA), a widely used series of instruments for the assessment of mental health[1] across different ages and raters. The TP-scale measures symptoms common in several mental disorders: hallucinations, OCD-symptoms, strange thoughts and behaviors, self-harm and suicide attempts. TP has been associated with psychiatric disorders such as OCD,[2; 3] pediatric bipolar disorder,[4] mania,[5] 22q11 deletion syndrome,[6] and several psychotic features.[7] When considered together with the Rule Breaking syndrome scale from the ASEBA, TP is predictive for schizophrenia.[8] Together with the Somatic Complaints scale, the TP-scale can be predictive for mania or hypomania.[8]

The TP-scale has received less attention than the other subscales of the ASEBA. It is mainly comprised of low-prevalence items and is the subscale with the lowest internal consistency (Cronbach's alpha = .51).[1] TP also has a relatively low long-term stability (.36 for a mean interval of ~3.5 years).[1] These features make the TP scale difficult to analyze, unless large sample sizes are available.

The heritability of TP has been estimated in children (4 to 16 years old) and ranged from .32 to .75, while shared environmental influences ranged from 0 (not detectable) to .21. Dominant (non-additive) genetic influences have not been reported for this age group.[9-14] These estimates were based on parental or teacher ratings of children's behavior. The study with the largest sample size (~9000 7-year old twin pairs) estimated the heritability at 61% and 65% for ratings from the twins' mothers and fathers, respectively.[9] This study also concluded that the rater agreement on TP between the parents was 67%, while the remaining 33% consisted of a unique view on the phenotype and/or measurement error.

The current study analyzes TP-data from self-reports in 12 to 59 year old subjects. It could be argued that, given the content of some of the items of the TP-scale, self-ratings might assess the phenotype differently. Since TP-scores seem to change more with age than scores of the other ASEBA problem scales, the influence of genes and environment may also differ from estimates obtained in children.

This study is conducted in a large sample of adolescent and adult twins and their family members, who between 1991 and 2010 took part in longitudinal survey studies. We first investigate whether the TP-scale assesses a single or multiple constructs through an exploratory factor analysis. Based on the outcome of this analysis, we test whether the TP-scale measures the same construct(s) across different ages and sex in measurement invariance (MI) analyses.[15-18] This is important, because in order to consider genotype by sex and genotype by age interaction, it needs to be established that different patterns in familial resemblance in these groups are

not caused by differences in measurement.[19] If the TP-scale is indeed measurement invariant, genetic and environmental influences on the longitudinal development of TP can be examined with data from monozygotic (MZ) and dizygotic (DZ) twins and their siblings.

Methods

Participants

Data came from the longitudinal survey study of the Netherlands Twin Registry (NTR), in which Dutch twins and their family members are assessed every 2-4 years since 1991.[20] Details about sample selection and response rates are described elsewhere.[20; 21] We analyzed data from twins, siblings, offspring, parents and spouses collected in 1991, 1995, 1997 and 2009/2010. Data from twins were available at all time points, while for the other family members data were available for the surveys collected in 1997 and 2009/2010.

For the EFA and MI analyses , the sample was divided into three age groups (12-18, 19-27 and 28-59 year olds) and two sex groups, which resulted in six groups (3 age groups × 2 sex groups). For each subject one random measurement was chosen from the longitudinal dataset. Additional MI analyses (within age groups) were carried out analyzing data from twins. Here, one random measurement was chosen per age group (which could lead to twins being included in multiple age groups).

For the longitudinal genetic modeling, data from twin pairs and two additional siblings (brother and sister) were analyzed. The ages of subjects within each survey varied greatly, therefore the data were reorganized so that the longitudinal design was based on age intervals instead of survey intervals.[22] The sample was divided into three age groups (12-18, 19-27 and 28-59 year olds). Multiple measurements for each subject were included, but only one measurement per age group (chosen at random).

For the EFA and MI analysis 15,320 subjects were included (twins and family members). Data from 9,067 twins were analyzed for the additional MI analysis (MI within age groups; 4080 measurements in the first, 5814 in the second, and 3307 in the third age group). For the longitudinal genetic analyses, data from 11,107 subjects were included (8,446 subjects with one, 2,126 with two and 535 with three measurements). A breakdown by age group, sex and zygosity of all samples is given in Appendix Tables 1 to 3.

DNA or blood group polymorphisms were used to determine zygosity for 38 % of the same-sex twin pairs. For the other 62 % zygosity was determined from surveys completed by parents and twins. The surveys asked questions regarding the resemblance of the twins and whether they were mistaken for each other as

children by family members and strangers. When there was inconsistency across time or persons, the majority of the judgments determined the outcome. If there were inconsistencies between survey questions and DNA, the DNA zygosity was used. Correspondence between zygosity determined by survey questions and DNA was 98 % if there were no (longitudinal or rater) inconsistencies in the parental and twin questionnaire reports, otherwise it was 97 %.[23]

Measures

Behavioral and emotional problems were assessed with the Adult Self Report (ASR),[1] which is part of the Achenbach System of Empirically Based Assessment (ASEBA). The ASR consists of 126 items. The TP-scale consists of 10 items (shown in Table 1). The items have three response categories: (0) not true; (1) somewhat or sometimes true; (2) very true or often true.

For the factor analyses (EFA & MI) item scores were analyzed. The Cronbach's alpha was .57 in the complete sample of 15,320 individuals, which is slightly higher than .51 as reported in the ASEBA manual.[1] Missing items were handled with the weighted least square estimation (WLSMV) with missing data in Mplus (for the EFA and MI analyses), and the raw data maximum likelihood approach in Mx (for the additional MI analyses), allowing the use of all available data.[24; 25]

For the genetic modeling the log-transformed sum scores were analyzed only in subjects who had at most two missing items. If one or two items were missing, these were given the average value of the available items for an individual. Of the 14303 measurements, there were 505 with 1 item missing (166 from age group 1, 179 from age group 2, and 160 from age group 3) and 146 with 2 items missing (22 from age group 1, 49 from age group 2, and 75 from age group 3). Including the individuals with (a) missing item(s) did not lead to a decreased variance.

Exploratory Factor Analysis (EFA)

The software package Mplus Version 5.21[24] was used to explore the factor structure of the TP-items in an exploratory factor analysis (EFA) for ordinal data with the WLSMV estimator. An underlying normal distribution was assumed for each item, where the three response categories are divided by two thresholds estimated from the data. Dependency among observations of family members was corrected for with the 'complex' option, which has shown to be effective in the context of family data.[26] Mplus gives several descriptive model fit statistics to help determine how many common factors to include in the model to adequately account for the correlation among the item scores. In this study, model fit was evaluated with the root mean square error of approximation (RMSEA), because it

performs well in factor models with categorical data and is robust to large sample sizes and model complexity.[27; 28] An RMSEA value smaller than .05 is considered a good fit, between .05 and .08 an adequate fit, between .08 and .10 a mediocre fit, and values > .10 are not considered acceptable.[27; 28] The decision for the factor model was based on parsimony, the eigenvalues and whether the fit was acceptable (good or adequate, i.e. the cutoff value of the RMSEA was .08).

Measurement Invariance

An essential step in examining age and sex differences is testing for measurement invariance (MI).[15; 16] MI was tested for the six age × sex groups with a multi-group confirmatory factor analysis (MGCFA) for ordinal data, assuming an underlying continuously distributed liability subject to a series of thresholds that categorize the phenotype. For each item, two thresholds are estimated because there are 3 response categories (visualized in Figure 1), meaning that the factor model is only indirectly connected to the measured variables. Flora and Curran (2004)[29] showed that especially with large sample sizes, confirmatory factor analyses perform well with ordinal data.

Four models reflecting four levels of MI are tested that form a nested hierarchy and are represented by increasing levels of cross-group equality constraints. The first level of measurement invariance is *configural invariance*, which implies that the same factor structure holds for all six groups, but parameter estimates may vary across groups. Configural invariance is tested by fitting the hypothesized factor model in each of the age × sex groups separately and in a multigroup analysis of the total sample. If the model fits well, the next level of MI, *metric invariance*, is tested. Metric invariance means that the latent factor scores predict the item responses equally well across groups, i.e. that the common factors have the same meaning across groups. This is tested by constraining the factor loadings to be equal across the six groups. The third level of MI is *strong factorial invariance*, which implies that comparisons of group means are meaningful, i.e. that differences in latent response means reflect differences in factor means. Strong factorial invariance holds if factor loadings and thresholds can both be constrained to be equal across groups. The fourth and most stringent step is testing for *strict factorial invariance*. This is tested by constraining factor loadings, thresholds and the residual variances of the latent responses to be equal across groups. If strict factorial invariance holds, comparisons of latent response means and observed variances across groups are meaningful (i.e. they reflect true differences in the latent factor mean and variance, hence the factor represents the same construct across groups). Figure 1 shows a visual representation of the constraints for each level of MI. See Flora and Curran (2004)[29] and Millsap and Yun-Tein (2004)[30] for a more detailed description on ordered-categorical measures in this context.

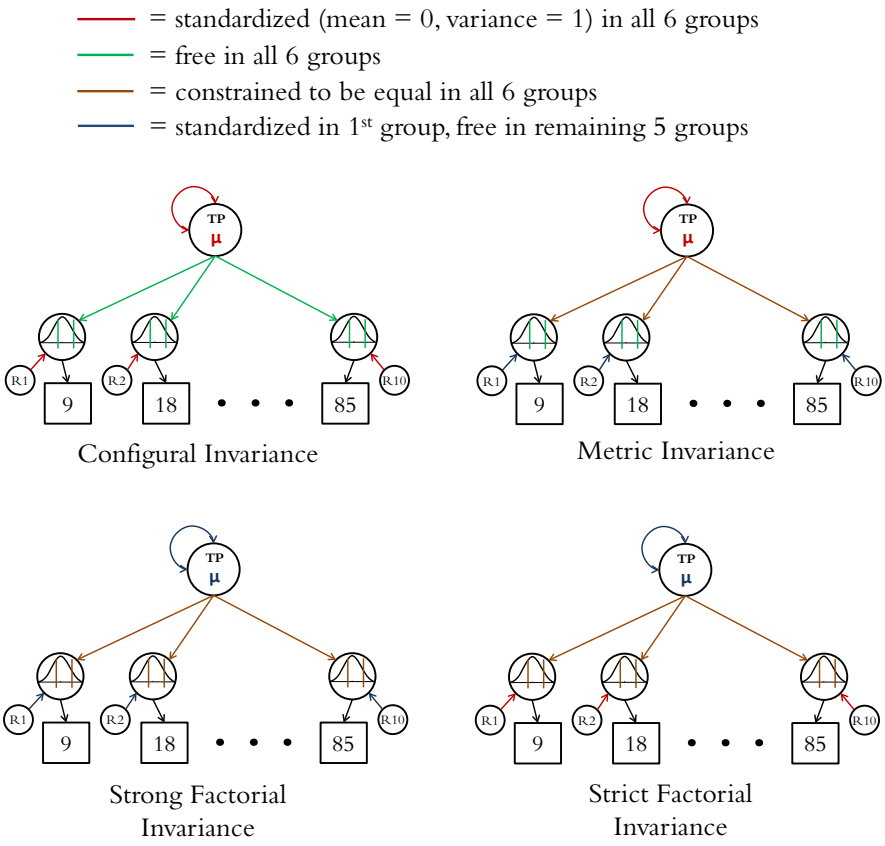


Figure 1: (note: only the first, second and last of the ten items are displayed) Configural invariance model: The 10 residual item variances are fixed at 1 in all groups, in order to estimate the 10 factor loadings freely in all groups. The unmeasured latent TP-factor is also standardized (mean = 0, variance = 1) in all groups in order to estimate the thresholds. This model was fitted for each of the six age × sex groups separately, as well as in a multigroup analysis of the total sample. Metric invariance model: The factor loadings are constrained to be equal in all six groups (i.e. estimated once). The 10 residual item variances are fixed at 1 in the first group only, in order to estimate the 10 factor loadings. The unmeasured latent TP-factor is standardized (mean = 0, variance = 1) in all groups in order to estimate the thresholds. Strong factorial invariance model: The factor loadings and the thresholds of the latent responses are constrained to be equal in all six groups (i.e. estimated once). The 10 residual item variances are fixed at 1 in the first group only, in order to estimate the 10 factor loadings. The unmeasured latent TP-factor is standardized (mean = 0, variance = 1) for the first group only in order to estimate the thresholds. Strict factorial invariance model: The factor loadings and the thresholds of the latent responses are constrained to be equal in all six groups (i.e. estimated once). The 10 residual item variances are fixed at 1 in all six groups. The unmeasured latent TP-factor is standardized (mean = 0, variance = 1) for the first group only in order to estimate the thresholds.

Mplus Version 5.21[24] was used to test for MI, using the THETA parameterization. As for the EFA, the WLSMV estimator was used, the 'complex' option was used to correct for dependency among observations of family members, and the RMSEA was used as a model fit index.

By testing for MI between the three age groups, it is assumed that MI also holds within the age groups. This assumption is tested by investigating MI as a continuous function of age in Mx.[25; 31; 32] With this approach, due to practical limitations, we chose to test MI in twins and with respect to factor loadings and thresholds only, similar to the metric invariance and strong factorial invariance tests respectively in the between group MI tests. For a more detailed description of these tests, see the Appendix and Appendix Figure 2.

Genetic Modeling

The contribution of genetic and environmental influences on TP can be inferred from the resemblance between MZ twins, DZ twins and siblings. This design is based on the assumption that DZ twins and siblings share on average ~50% of their segregating genes and MZ twins share ~100% of their genome. Therefore, genetic effects are assumed to be present if MZ twin correlations are larger than DZ twin correlations. A more detailed description of how additive genetic (A), non-additive or dominant genetic (D), shared environmental (C) and unique environmental influences (E, also includes measurement error) are inferred from twin and sibling correlations can be found elsewhere.[33; 34]

The genetic analyses were done in Mx.[25] All models were fitted to the raw data with maximum likelihood estimation procedures. First, correlations, means and variances of TP sum scores were computed for sibs and twins of all zygosity groups (MZM, DZM, MZF, DZF, DOS) in a fully saturated model. The difference between DZ and sibling correlations was tested by constraining them to be equal and comparing the fit to the fit of the fully saturated model. Sex differences between twin/sibling correlations were tested in the same way. Homogeneity of means was tested by constraining the means to be equal across zygosity (twins and siblings), sex and age groups. To test whether the large range of ages within the age groups needs to be corrected for, it was tested whether including age as a covariate (linear and quadratic) on the means in the saturated model led to a better fit. For the linear age covariate age was standardized (to z-scores) and for the quadratic age covariate age was standardized and then squared, to reduce the correlation between the two covariates. Based on the twin correlations, it was determined whether to estimate the A, C and E or the A, D and E parameters, since a model that includes A, C, D and E would not be identified. If MZ twin correlations are more than twice the DZ correlation, an ADE model would be more sensible, otherwise the ACE model is fitted.[33; 34]

Next, a Cholesky decomposition,[35] with constraints/covariates based on their significance in the saturated model, was fitted to the TP-data. This model is described in the path diagram in Figure 2 for an opposite-sex twin pair with a male and a female sibling. The addition of siblings to this classical twin design has been shown to increase the power to detect dominant genetic and shared environmental influences.[36] The measured phenotypes are represented in rectangles, and the unmeasured latent sources of variance are in circles. The genetic (A & D) and environmental (C & E) sources of variance are each represented by three factors: the first influencing the variances and covariances of TP for all three age groups, the second explaining the variances and covariances of only the second and third age group, and the third explaining the variances and covariances of the third age-group only. This model allows for the investigation of longitudinal changes in the genetic/environmental factors (in the form of new genetic factors arising, like A2 or A3 in Figure 2, for additive genetic influences) and longitudinal stability of the genetic/enviromental influences (in the form of longitudinal correlations, derived from a21, a31 and a32 in Figure 2, for additive genetic influences).

Significance of the estimated parameters and differences between groups (sex, age groups, zygosities) in the saturated and Cholesky models were obtained by comparing the full models with the constrained models. In Mx, the fit of different models can be compared by means of likelihood ratio tests.[35] The χ^2 value is obtained by subtracting the -2 log likelihood (-2LL) of the more restricted model from the -2LL of the less restricted model. The Δdf is the difference between the degrees of freedom of the two models. According to the standard approach, if the χ^2 test results in a non-significant *p*-value ($p \geq .05$), the constrained model is preferred. The χ^2 value however is inflated when using large sample sizes and complex models, causing small discrepancies in large samples to seem significant. Given the large sample sizes and the complexity of the Cholesky model with three age groups and two siblings, we chose an alternative fit index: the Bayesian Information Criterion (BIC),[37] which performs well with large sample sizes and complex models.[38] Models with a lower BIC value were chosen as a better fit over the model with a higher BIC.

Results

EFA and MI

The endorsement frequencies of the items for subjects in the EFA and MI analyses are shown in Table 1. The endorsement of the positive answer categories was almost identical in these datasets and was relatively low. The frequencies of the positive answer categories were highest for items 9 (category 1: .33, category 2: .08) and 63 (category 1: .29, category 2: .05), and lowest for the item on suicide attempts (item 18: category 1: .01 for the total sample, category 2: .003) and the hallucination

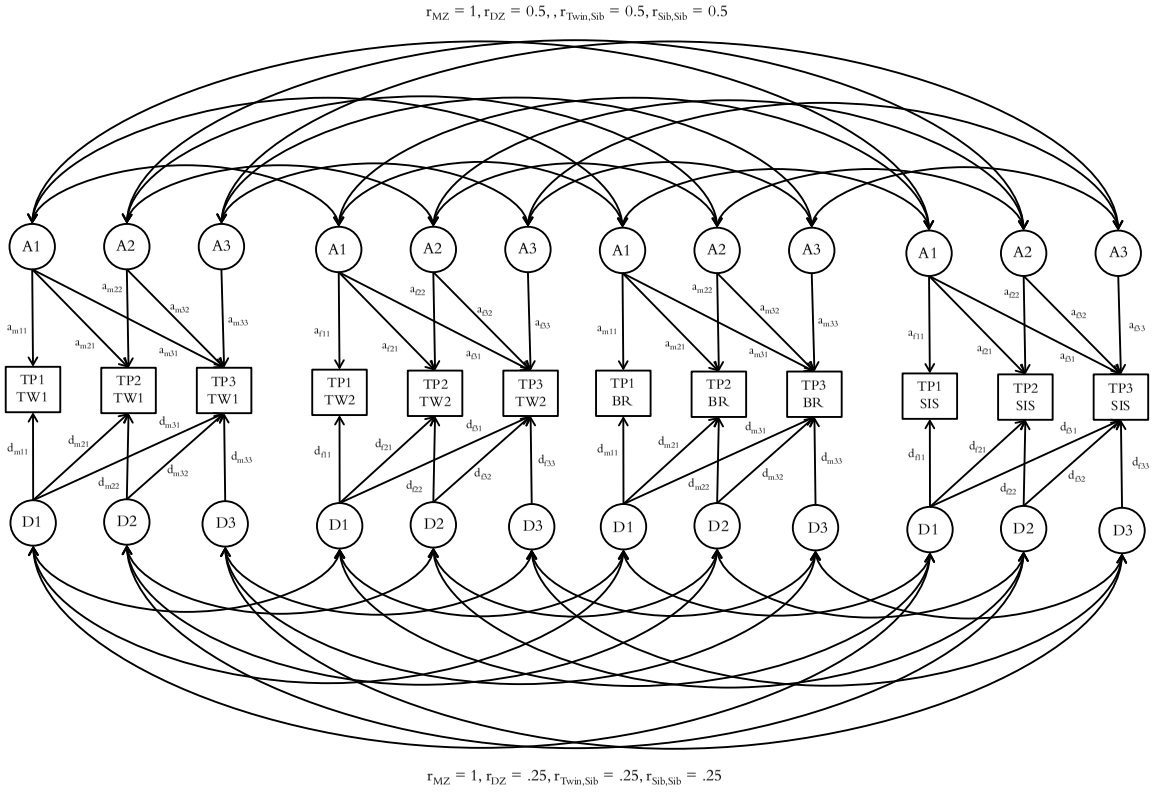


Figure 2: Path diagram for longitudinal ADE model on Thought Problems (TP) for three age groups. The figure shows data from an opposite sex twin pair (TW1 = male, TW2 = female) and their two siblings (BR = brother, SIS = sister). The rectangles represent the log-transformed TP sum-scores (TP1 = TP measured at ages 12-18, TP2 = 19-27, TP3 = 28-59). The circles the latent unmeasured factors (A = additive genetic effects, D = dominant genetic effects, E = non-shared environmental effects, and is omitted in the figure for simplicity, but is modeled in a similar way). In parameter subscripts, m stands for male and f stands for female.

Table 1: Frequencies of the item responses in samples from the EFA and MI analyses and the factor loadings as estimated in the EFA

	Frequencies of item responses						
	EFA + MI between age groups (i.e., all available subjects)			MI within age groups (i.e., twins only)			Factor loadings (EFA)
	0	1	2	0	1	2	
9: I can't get my mind off certain thoughts	.59	.33	.08	.59	.32	.09	.52
18: I deliberately try to hurt or kill myself	.99	.01	.003	.98	.01	.004	.62
36: I accidentally get hurt a lot	.87	.12	.02	.84	.14	.02	.39
40: I hear sounds or voices that other people think aren't there	.98	.02	.004	.97	.02	.005	.70
46: Parts of my body twitch/make nervous movements	.91	.07	.02	.90	.08	.02	.51
63: I would rather be with older people than people my own age	.66	.29	.05	.62	.32	.06	.38
66: I repeat certain acts over and over	.94	.05	.01	.94	.05	.01	.56
70: I see things that other people think aren't there	.98	.02	.01	.97	.02	.01	.71
84: I do things that other people think are strange	.88	.11	.02	.87	.11	.02	.73
85: I have thoughts that other people would think are strange	.88	.10	.02	.87	.11	.02	.84

Response categories: (0) not true; (1) somewhat or sometimes true; (2) very true or often true.

symptoms (items 40: category 1: .02, category 2: .004; item 70: category 1: .02, category 2: .01).

The EFA yielded a one-factor solution as a good fit for the ten items with an RMSEA of .038. The eigenvalues also strongly support the one-factor solution (eigenvalues 1 to 10: 4.20, 0.93, 0.90, 0.85, 0.79, 0.67, 0.61, 0.52, 0.28, 0.27).

Table 1 shows the factor loadings from the EFA. Item 85 (*I have thoughts that other people would think are strange*, factor loading = .84) has the highest factor loading. Items 36 and 63 have the lowest factor loadings (.39 & .38 respectively). Removing these two items lead to a worse fit (RMSEA = .048) and a lower first eigenvalue (eigenvalues 1 to 10: 3.85, 0.90, 0.86, 0.68, 0.61, 0.55, 0.28, 0.27). Therefore all items were retained, also allowing for comparisons with previous

studies using this scale.

The fit of the configural invariance models was good in all groups (RMSEA < .05), except in the adult males, where it could be considered adequate (RMSEA = .065). In the multigroup analysis, the configural invariance model also had a good fit (RMSEA = .044), indicating that the one-factor model holds in all age × sex groups. Of the remaining MI tests, the metric invariance model showed a good fit (RMSEA = .047), while the strong factorial and strict factorial invariance had an adequate fit (RMSEA = .053 & .060 respectively; see Table 2). Testing for MI within the age groups yielded similar results. MI with respect to both factor loadings and thresholds across age held within in all three age groups. For more details on the MI tests within age groups, see the Appendix, Appendix Figure 2 and Appendix Table 4.

Table 2: Model fitting results for measurement invariance tested across sex and age

	N	Free parameters	RMSEA
Exploratory Factor Analysis – one-factor solution	15320	10	.038
Configural invariance – Males – 12-18 years old	1255	30	.041
Configural invariance – Females – 12-18 years old	1488	30	.032
Configural invariance – Males – 19-27 years old	2129	30	.044
Configural invariance – Females – 19-27 years old	3284	30	.035
Configural invariance – Males – 28-59 years old	2497	30	.065
Configural invariance – Females – 28-59 years old	4667	30	.037
Configural invariance – *Total sample*	15320	180	.044
Metric Invariance	15320	180	.047
Strong Factorial Invariance	15320	90	.053
Strict Factorial Invariance	15320	40	.060

Longitudinal Genetic Analysis

There were no significant mean or variance differences for the TP-score between the different zygosities, sibs or sex based on the BIC values (values not shown). The mean TP-scores were equal for adolescents and young adults (non-transformed mean TP-score = 1.34), but dropped significantly in later adulthood (non-transformed mean TP-score = .91). The variance did not differ significantly between the age groups. BIC values also indicated that the age covariate effects were not significant in the saturated model, and were therefore not included in the ACE/ADE Cholesky model (see Table 4).

The within-person longitudinal correlations were .37 between adolescence

and young adulthood, .37 between adolescence and adulthood, and .26 between young adulthood and adulthood. Table 3 shows the cross-twin-within-time and the cross-twin-cross-time correlations. The DZ correlations did not differ significantly from the sibling-correlations as indicated by BIC values (see Table 4). The MZ-correlations are consistently higher than the DZ correlations in all three age groups, indicating genetic influences on the TP-scores. The twin correlations within age also suggest dominant genetic influences in young adults and adults, indicated by MZ correlations larger than twice the DZ correlations. The cross-twin-cross-time correlations show that past TP-scores of one twin are more predictive of future TP-scores for the co-twin in MZ pairs than in DZ/sibling pairs. This suggests that the longitudinal stability of TP-scores may be explained by genetic factors.

The significance of each A and D parameter was tested, as well as longitudinal correlations of unique environmental effects (= e21, e31 and e32 only). The first and second additive genetic factor in the longitudinal model significantly influenced TP in the three age groups (see Table 4). The genetic correlation between TP in adolescence and young adulthood was .92, between young adulthood and adulthood .87, and .62 between adolescence and young adulthood. The longitudinal correlations among the unique environmental influences were not significant. The proportions of variance explained by genetic and unique environmental influences did not differ between the three age groups. The variance explained by additive

Table 3: Cross-Twin-Within-Time and Cross-Twin-Cross-Time correlations as estimated in the saturated model (with and without sex differences)

	Cross-Twin – Within-Time			Cross-Twin – Cross-Time		
	12-18	19-27	28-59	12-18 – 19-27	12-18 – 28-59	19-27 – 28-59
MZM	.29	.35	.24	.20	.28	.29
DZM	.17	.11	.16	.11	.12	.04
MZF	.39	.43	.31	.28	.14	.26
DZF	.30	.21	.07	.23	.07	.12
DOS (mf / fm)	.24	.15	.08	.10/.17	.12/.01	.08/.17
MZ	.34	.40	.30	.24	.17	.27
DZ	.27	.19	.10	.19	.10	.10

MZM = male monozygotic twin pairs. DZM = male dizygotic twin/sibling pairs. MZF = female monozygotic twin pairs. DZF = female dizygotic twin/sibling pairs. DOS = opposite sex dizygotic twin/sibling pairs, MZ = all monozygotic twin pairs, DZ = all dizygotic twin/sibling pairs, mf = male – female correlation, fm = female – male correlation.

Table 4: Summary of the Model Fitting Results of the longitudinal genetic analyses

	-2 LL	#par	df	χ^2	Δdf	*P*	BIC
Saturated Model:							
1. Fully saturated	1961.155	120	13834				-58715.507
2. rDZ = rSib (versus 1)	1980.646	100	13854	19.491	20	0.490	-58792.065
3. No sex differences for twin/sibling correlations (versus 1)	2005.494	81	13873	44.339	39	0.257	-58861.629
4. Covariate standardized age dropped (versus 1)	2004.964	114	13840	43.809	6	<.001	-58719.494
5. Covariate squared standardized age dropped (versus 1)	1983.825	114	13840	22.670	6	0.001	-58730.063
ADE model (models include restrictions from models 2 to 5):							
6. ADE-model	2211.950	29	13929				-59000.051
7. AE-model (versus 6)	2222.021	23	13935	10.071	6	0.122	-59020.906
8. AE-model – A3 dropped (versus 7)	2228.579	22	13936	6.558	1	0.010	-59021.943
9. AE-model – A3 & a22 dropped (versus 8)	2239.785	21	13937	11.206	1	0.001	-59020.655
10. AE-model – A3 & a32 dropped (versus 8)	2237.624	21	13937	9.045	1	0.003	-59021.735
11. AE-model – A3, e21, e31 & e32 dropped (versus 8)	2251.165	19	13939	22.586	3	<.001	-59023.595
12. AE-model – A3, e21, e31 & e32 dropped + proportion variance explained by A equal for all age groups (versus 11)	2263.957	15	13943	12.792	4	0.012	-59034.460

genetic influences was 37% in all age groups, and the remaining 63% was explained by unique environmental influences or measurement error. The unstandardized genetic components also barely change over time. The unstandardized genetic components for A are: .029 for young adolescents, .027 for young adults and .025 for adults. The unstandardized genetic covariance components are: .026 between adolescents and young adults, .023 between young adults and adults, and .016 between adolescents and adults. The unstandardized components for E are: .049 for young adolescents, .047 for young adults and .042 for adults.

Discussion

This study investigated the strength and the structure of the relations between the TP-items with an exploratory factor analysis (EFA), whether the TP-scale is measurement invariant across age and sex, and examined the longitudinal genetic and environmental influences on the TP-scale using the genetic relatedness of the twin subjects and their siblings.

The EFA yielded a one-factor structure. Further examination of the one-factor structure in a multigroup confirmatory factor analysis led to the conclusion that the TP-scale is measurement invariant between adolescent, young adult and adult males and females. Testing for MI within age groups confirmed MI with respect to both factor loadings and thresholds. This means that between and within the age groups, differences between observed thresholds and observed variances across age and sex appear to be due to common factor variation and real differences in the TP-construct/factor mean.

The longitudinal genetic analyses detected additive genetic influences on TP. TP was influenced by the same additive genetic component from adolescence to adulthood, but an additional genetic component arises during young adulthood, and keeps influencing the trait throughout adulthood. The additive genetic factor explained 37% of the variance across all age groups. The genetic correlation between adolescents and young adults was very high (.92). The genetic correlation between young adults and adults was .87, and .62 between adolescents and young adults. This indicates that the largest part of the young adult variation was explained by the same genetic component as in adolescents, and that the genetic component that arose during young adulthood explained the largest part of the adult variation. Dominant genetic and shared environmental influences were not detectable. The remaining variance was explained by unique environmental influences and may also partly reflect measurement error. There were no significant longitudinal correlations between the unique environmental factors, i.e. unique environmental factors in one age group did not influence TP in another age group. The mean scores were about equal in the first two age groups, and decreased significantly in

the adult group.

The results of the EFA, MI and the longitudinal heritability analyses imply that (1) there is a single construct underlying the ten TP-items, (2) longitudinal changes in the TP-scores can be explained as true changes in the underlying TP-construct, and (3) there are two genetic components that accompany the longitudinal development of TP: the first influencing TP throughout all age groups, while the second arises during young adulthood and stays influent throughout adulthood. The longitudinal stability is reported to be lower for this scale than for other ASR scales. The ASR-manual reports a longitudinal stability of .36 for a mean interval of ~3.5 years.[1] The longitudinal correlations are in the same range in this study (.37 between adolescence and young adulthood, .37 between adolescence and adulthood, and .26 between young adulthood and adulthood). The results of this study imply that the longitudinal correlation is not due to environmental factors, but can be explained entirely by genetic factors.

The one-factor structure for the ten TP-items and the fact that the total TP-scores share additive genetic influences across age suggest that the Thought Problems scale may be measuring an underlying liability for multiple symptoms. When taking a closer look at the items, they seem to point towards schizo-obsessive symptoms. There is growing evidence that comorbidity of schizophrenic and obsessive-compulsive symptoms may possibly result from a pathophysiological linkage between the two disorders. Schizophrenia and OCD occur together more often than expected, based on their separate lifetime prevalence rates, and seem to share common functional circuits and dysfunctions of neurotransmitter systems.[39; 40] See Tibbo & Warneke (1999),[40] Stein (2000),[39] Reznik et al (2001),[41] Bottas et al (2005),[42] and Poyurovski et al (2006)[43] for reviews and discussions about the schizo-obsessive disorder as a new diagnostic entity.

The TP-scale includes items that cover classical OCD-symptoms and are also included in the Obsessive Compulsive Scale of the Achenbach questionnaire (items 9, 66, 84 and 85).[44] TP also includes items that cover symptoms that could be interpreted as OCD-symptoms as well as psychotic symptoms (items 84, 85, 40, 70). Besides being a classical schizophrenic symptom, hallucinations - covered by items 40 (= auditory hallucinations) and 70 (= visual hallucinations) - are not uncommon in OCD-patients.[45; 46] Studies have linked intrusive cognitions - such as hallucinations and obsessions - with inhibitory dysregulation in the brain, which both schizophrenic and OCD patients suffer from.[47-49] Studies of schizophrenic patients, with and without OCD, showed that subjects with OCD showed more suicide attempts (item 18) and motor symptoms (item 46) than patients without OCD.[50-53] Effective treatment strategies also differed between the two groups for the motor symptoms. Items 36 and 63 have considerable lower factor loadings (see Table 1) and are more difficult to relate to schizo-obsessive disorders. Item 36

could perhaps be linked to the motor symptoms. Item 63 however not only has the lowest factor loading of all ten items in the EFA (see Table 1), but is also hardest to fit theoretically into the construct the TP-scale seems to measure. The significant genetic influences on the variation of this scale support previous findings about the heritability of TP and are in line with the findings that relatives of OCD-schizophrenia patients had significantly higher risks for OCD-schizophrenia than relatives of schizophrenia patients without OCD.[54]

There are certain limitations in this study that should be considered when interpreting these results. Because of the highly varying ages in each of the four surveys used in this study (1991, 1995, 1997 & 2009/2010), relatively large age intervals had to be defined for the age subgroups in the genetic modeling analyses, resulting in a somewhat low temporal resolution of the longitudinal results. Also, since we only included one measurement per age group and data from siblings were collected only in 1997 and 2009/2010, the majority of the subjects only had one measurement in the longitudinal analyses. Another limitation is the overall low score of the TP-scale in this sample, which makes it more difficult to draw conclusions at a clinical level.

It appears that the ten TP-items measure a single TP-construct, that measurement invariance holds for the TP-scale and that there are significant additive genetic influences on its variation in different age groups that correlate high over time. When considering the symptoms the TP-items cover, the most plausible known corresponding clinical entity is the schizo-obsessive disorder. Further investigation is needed on the relationship between the TP-scale and schizo-obsessive disorder. Future studies also have to determine the effectiveness of this scale in clinical settings. Since the TP scale measures the same construct influenced by the same genes in younger and older subjects and in males and females, pooling their data together in linkage-analyses and (genome-wide) association studies may increase power in candidate gene studies.

Appendix – Measurement Invariance within age groups

Methods

By testing for MI between the three wide age groups, it is assumed that MI also holds within the age groups. This assumption is tested in three (adolescents [12-18], young adults [19-27] and adults [28-59]) single group item-factor analyses.[31; 32] For a detailed description of how this model is applied to ordinal data, see Kubarych et al (2010)[31] and Wirth & Edwards (2007).[55] The path model with one TP factor underlying the 10 items is shown in Appendix Figure 1. Boxes represent the ten observed TP-items; solid line circles represent factors; broken line circles represent special nodes used to estimate the covariate moderation effects; diamonds represent the covariate effects (age, transformed to a z-score); triangles represent unit constants for estimating means and threshold covariate effects; single-headed arrows represent linear regression effects; and double headed arrows represent variances and covariances. The covariate effects on the factor mean and variance are represented by *B* and *D* respectively through the special nodes *DF*. The factor loadings are denoted $L_{\#}$, and the covariate effects on the factor loadings are represented by $J_{\#}$ through the special nodes *DL*. The moderation effects of the item thresholds ($m_{\#}$) are estimated by parameters $K_{\#}$. For each item, two thresholds are estimated because there are 3 response categories. Separate MZ and DZ twin correlations are only allowed between the TP factors (*TP TW1* and *TP TW2*) and between the item residuals ($R1_{\#}$ and $R2_{\#}$).

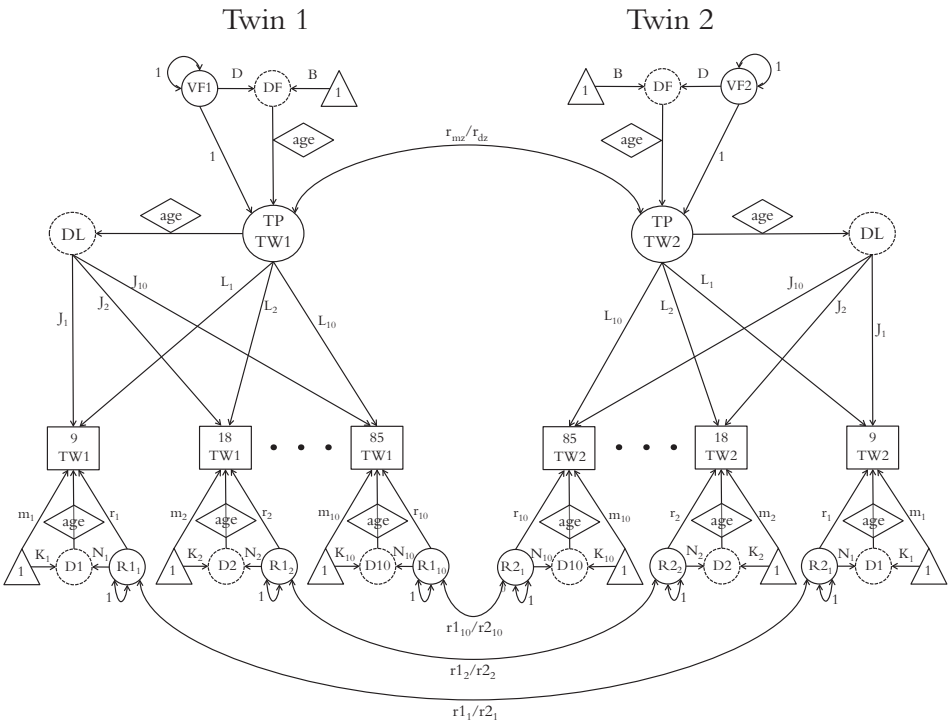
Given this model, MI can be evaluated at two levels. 1) If the factor loadings change as a function of age, this may bias the factor mean and variance. This can be tested by comparing the fit of a model with moderated factor variance (*D* free) with the fit of a model where the moderation on the factor loadings is allowed ($J_{\#}$ free). If the latter fits better, the TP scale may not be measurement invariant. 2) If the item thresholds change as a function of age due to causes other than the factor, the factor mean may be biased. Analogous to the first test, this can be tested by comparing the fit of a model where only the factor mean is allowed to vary as a function of age (by freeing *B*) with the fit of a model where the item threshold locations are allowed to vary across age (by freeing all $K_{\#}$). If the model with moderated item thresholds fits better than the model with the moderated factor mean, the TP scale would not be considered measurement invariant. Hence, we distinguish between the genuine effects, reflected by changes in variance and factor mean, and changes in the functioning of the measurement instrument, which may be reflected by changes in the factor loadings and items thresholds.

Models were tested in Mx,[32] which compares the fit of different models by likelihood ratio tests.[35] The χ^2 value is obtained by subtracting the -2 log likelihood

(-2LL) of the more restricted model from the -2LL of the less restricted model. The Δdf is the difference between the degrees of freedom of the two models. According to the standard approach, if the χ^2 test results in a non-significant *p*-value ($p \geq .05$), the constrained model is preferred. The χ^2 value however is inflated when using large sample sizes, causing small discrepancies in large samples to seem significant. Given the large sample sizes and the complexity of model, we chose an alternative fit index: the Bayesian Information Criterion (BIC),[37] which has been shown to perform well with large sample sizes and complex models.[38] Models with a lower BIC value were chosen as a better fit over the model with a higher BIC.

Results

The results of the MI tests are shown in Appendix Table 4. In all three age groups, first a full MI baseline model (model 1) was fitted, where the covariate effects were constrained to zero. Freeing the covariate effects of age on the latent factor variance in model 2 did not result in a better fit than model 1 in any of the age groups, indicating that the variance does not change over time within the age groups. Freeing the covariate effects of age on the factor loadings in model 3 did not result in a better fit than model 2 in any of the age groups, indicating that the TP-scale is measurement invariant on this level. In model 4 the covariate effect on the factor mean of age was freely estimated. Based on the BIC, comparisons with model 1 suggested a better fit for freely estimated age parameters in adolescents only, indicating factor mean changes across age in that age group. Model 5 (with freely estimated age effects on item thresholds) did not show a better fit than model 4 in any of the three age groups, indicating that allowing the thresholds to vary across age does not results in a better fit than allowing only the factor mean to vary across age. This suggests that differences in thresholds across age are due to differences in the factor mean in all three age groups, i.e. the TP-scale is measurement invariant on this level as well.



Appendix Figure 1: Path diagram for the MI analyses for twins. Boxes: observed TP-items (only the first, second and last of the ten items are displayed); solid line circles: unobserved variables (factors); broken line circles: special nodes used to estimate the covariate moderation effects; diamonds: the covariate effects (age); triangles: unit constants for estimating means and threshold covariate effects; single-headed arrows: linear regression effects; double headed arrows: variances and covariances. TP TW1: TP-factor for twin 1; TP TW2: TP factor for twin 2; VF: factor variance; $m_{\#}$: item thresholds; $r_{\#}$: item variances; $L_{\#}$: factor loadings; B: covariate effects on the factor mean; D: covariate effects on factor variance; $K_{\#}$: covariate effects on item thresholds; $J_{\#}$: covariate effects on the factor loadings. rMZ/rDZ = estimated factor twin correlations for monozygotic (MZ) and dizygotic (DZ) twins, $r1_{\#}/r2_{\#}$ = twin correlations between item residuals.

Appendix Table 1: Configuration and sample size for the four analyses

	EFA + MI between age groups	MI within age groups – 12-18	MI within age groups – 19-27	MI within age groups –28-59	Longitudinal genetic analysis
Males from MZ twin-pairs	1366	653	890	432	1355
Males from same sex DZ twin-pairs	1044	537	717	259	1040
Females from MZ twin-pairs	2664	1051	1580	1275	2650
Females from same sex DZ twin-pairs	1654	674	1053	664	1650
Males from opposite-sex twin-pairs	1039	557	694	267	1028
Females from opposite-sex twin-pairs	1300	608	880	410	1292
Twins with unknown zygosity	379	-	-	-	-
Male siblings	922	-	-	-	809
Female siblings	1563	-	-	-	1283
Mothers	1626	-	-	-	-
Fathers	1006	-	-	-	-
Spouses of twins	526	-	-	-	-
Spouses of siblings	10	-	-	-	-
Offspring of twins	164	-	-	-	-
Offspring of siblings	57	-	-	-	-
Total	15320	4080	5814	3307	11107

Appendix Table 2: Number of twins and siblings per age group with data for the longitudinal genetic analyses

	Twins			Brother			Sister		
	12-18	19-27	28-59	12-18	19-27	28-59	12-18	19-27	28-59
MZM	594	793	420	17	60	61	23	69	72
DZM	471	643	249	15	45	48	4	51	73
MZF	874	1419	1248	26	93	104	19	119	171
DZF	565	965	651	16	63	71	9	82	101
DOS	1015	1408	657	25	96	98	21	125	152
Separate Sibs	-	-	-	1	20	44	2	61	274
Total Twins	3519	5226	3225	-	-	-	-	-	-
Total Sibs	-	-	-	100	377	426	78	507	843

MZM/DZM = monozygotic/ dizygotic males. MZF/DZF = monozygotic/ dizygotic females. DOS = dizygotic opposite sex twins.

Appendix Table 3: Number of twins and siblings with more than one datapoint in the longitudinal genetic analyses

	Twins				Brother				Sister			
	12-18 & 19-27	12-18 & 28-59	19-27& 28-59	12-18 - 28-59	12-18 & 19-27	12-18 & 28-59	19-27& 28-59	12-18 - 28-59	12-18 & 19-27	12-18 & 28-59	19-27& 28-59	12-18 - 28-59
MZM	187	51	68	73	0	0	14	0	2	3	14	0
DZM	146	30	65	41	0	1	11	0	0	1	11	0
MZF	194	120	159	209	2	1	23	0	3	6	36	0
DZF	145	70	120	98	2	1	17	0	3	2	16	0
DOS	302	87	143	114	0	1	20	0	4	5	36	0
Separate Sibs	-	-	-	-	0	0	0	0	0	0	3	0
Total Twins	974	358	555	535	-	-	-	-	-	-	-	-
Total Sibs	-	-	-	-	2	5	85	0	12	17	116	0

Appendix Table 4: Summary of the Model Fitting Results of the MI analyses for ages 12 - 18

	-2ln*L*	#par	df	$\Delta\chi^2$	Δ df	*p*	BIC	Effect Size
Adolescents (ages 12 – 18):								
1. Full measurement invariance: no covariate effects	33320.109	52	40331	-	-	-	-138205.21	-
2. Age effects on latent variance (versus 1)	33320.057	53	40330	0.052	1	0.820	-138201.40	-.003
3. Age effects on factor loadings (versus 2)	33307.202	62	40321	12.855	9	0.169	-138173.27	-
4. Age effects on latent mean (versus 1)	33301.679	53	40330	18.43	1	< .001	-138210.59	.089
5. Age effects on thresholds (versus 4)	33261.634	62	40321	40.045	9	< .001	-138196.05	-
Young adults (ages 19 – 27):								
1. Full measurement invariance: no covariate effects	46373.794	52	57505	-	-	-	-210051.31	-
2. Age effects on latent variance (versus 1)	46373.703	53	57504	0.091	1	0.763	-210047.30	.030
3. Age effects on factor loadings (versus 2)	46364.929	62	57495	8.774	9	0.458	-210015.19	-
4. Age effects on latent mean (versus 1)	46374.521	53	57504	-0.727	1	NA	-210046.90	-.049
5. Age effects on thresholds (versus 4)	46353.477	62	57495	21.044	9	0.012	-210020.91	-
Adults (ages 28 – 59):								
1. Full measurement invariance: no covariate effects	20215.571	52	32714	-	-	-	-116676.06	-
2. Age effects on latent variance (versus 1)	20215.824	53	32713	-0.253	1	NA	-116672.06	-.011
3. Age effects on factor loadings (versus 2)	20207.303	62	32704	8.521	9	0.483	-116641.44	-
4. Age effects on latent mean (versus 1)	20211.715	53	32713	3.856	1	0.050	-116674.11	-.042
5. Age effects on thresholds (versus 4)	20184.145	62	32704	27.57	9	0.001	-116653.02	-

References

1. Achenbach, T., and Rescorla, L. (2003). Manual for the ASEBA adult forms & profiles. *University of Vermont, Research Center for Children, Youth, & Families, Burlington, VT.*
2. Geller, D. (2004). Re-examining comorbidity of obsessive compulsive and attention-deficit hyperactivity disorder using an empirically derived taxonomy. *European Child & Adolescent Psychiatry* 13, 83-91.
3. Ivarsson, T., Melin, K., and Wallin, L. (2008). Categorical and dimensional aspects of co-morbidity in obsessive-compulsive disorder (OCD). *European Child & Adolescent Psychiatry* 17, 20-31.
4. Diler, R.S., Birmaher, B., Axelson, D., Goldstein, B., Gill, M., Strober, M., Kolko, D.J., Goldstein, T.R., Hunt, J., et al. (2009). The Child Behavior Checklist (CBCL) and the CBCL-bipolar phenotype are not useful in diagnosing pediatric bipolar disorder. *Journal of Child and Adolescent Psychopharmacology* 19, 23-30.
5. Diler, R.S., Uguz, S., Seydaoglu, G., and Avci, A. (2008). Mania profile in a community sample of prepubertal children in Turkey. *Bipolar Disorders* 10, 546-553.
6. Sobin, C., Kiley-Brabeck, K., Monk, S.H., Khuri, J., and Karayiorgou, M. (2009). Sex differences in the behavior of children with the 22q11 deletion syndrome. *Psychiatry Research* 166, 24-34.
7. Kasius, M.C., Ferdinand, R.F., Berg, H., and Verhulst, F.C. (1997). Associations between different diagnostic approaches for child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry* 38, 625-632.
8. Morgan, C.J., and Cauce, A. (1999). Predicting DSM-III-R Disorders From the Youth Self-Report: Analysis of Data From a Field Study. *Journal of the American Academy of Child & Adolescent Psychiatry* 38, 1237-1245.
9. Abdellaoui, A., Bartels, M., Hudziak, J.J., Rizzu, P., Van Beijsterveldt, T.C., and Boomsma, D.I. (2008). Genetic influences on thought problems in 7-year-olds: A twin-study of genetic, environmental and rater effects. *Twin Research and Human Genetics* 11, 571-578.
10. Edelbrock, C., Rende, R., Plomin, R., and Thompson, L.A. (1995). A twin study of competence and problem behavior in childhood and early adolescence. *Journal of Child Psychology and Psychiatry* 36, 775-785.
11. Kuo, P.-H., Lin, C.C., Yang, H.-J., Soong, W.-T., and Chen, W.J. (2004). A twin study of competence and behavioral/emotional problems among adolescents in Taiwan. *Behavior Genetics* 34, 63-74.
12. Lin, C.C., Kuo, P.-H., Su, C.-H., and Chen, W.J. (2006). The Taipei Adolescent Twin/sibling Family Study I: behavioral problems, personality features, and neuropsychological performance. *Twin Research and Human Genetics* 9, 890-894.
13. Polderman, T.J., Posthuma, D., De Sonneville, L.M., Verhulst, F.C., and Boomsma, D.I. (2006). Genetic analyses of teacher ratings of problem behavior in 5-year-old twins. *Twin Research and Human Genetics* 9, 122-130.
14. Schmitz, S., Fulker, D.W., and Mrazek, D.A. (1995). Problem behavior in early and middle childhood: An initial behavior genetic analysis. *Journal of Child Psychology and Psychiatry* 36, 1443-1458.
15. Horn, J.L., and McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research* 18, 117-144.
16. Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525-543.
17. Vandenberg, R.J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods* 5, 139-158.
18. Vandenberg, R.J., and Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 3, 4-70.
19. Lubke, G.H., Dolan, C.V., and Neale, M.C. (2004). Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. *Twin*

Research 7, 292-298.
20. Boomsma, D.I., Vink, J.M., Van Beijsterveldt, T.C., de Geus, E.J., Beem, A.L., Mulder, E.J., Derks, E.M., Riese, H., Willemsen, G.A., et al. (2002). Netherlands Twin Register: a focus on longitudinal research. *Twin Research* 5, 401-406.
21. Boomsma, D.I., De Geus, E.J., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.-J., Posthuma, D., Van Beijsterveldt, T.C., Hudziak, J.J., et al. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics* 9, 849-857.
22. Mehta, P.D., and West, S.G. (2000). Putting the individual back into individual growth curves. *Psychological Methods* 5, 23.
23. Willemsen, G., Posthuma, D., and Boomsma, D.I. (2005). Environmental factors determine where the Dutch live: results from the Netherlands twin register. *Twin Research and Human Genetics* 8, 312-317.
24. Muthén, L., and Muthén, B. (2006). Version 5 Mplus user's guide. *Muthén & Muthén, Los Angeles*.
25. Neale, M.C., Boker, S.M., Xie, G., and Maes, H.M. (2006). Mx: Statistical modeling (6th edition). *Richmond, Virginia: Department of Psychiatry*.
26. Rebollo, I., De Moor, M.H., Dolan, C.V., and Boomsma, D.I. (2006). Phenotypic factor analysis of family data: correction of the bias due to dependency. *Twin Research and Human Genetics* 9, 367-376.
27. Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research* 8, 23-74.
28. Yu, C.-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. University of California Los Angeles.
29. Flora, D.B., and Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods* 9, 466.
30. Millsap, R.E., and Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research* 39, 479-515.
31. Kubarych, T.S., Aggen, S.H., Kendler, K.S., Torgersen, S., Reichborn-Kjennerud, T., and Neale, M.C. (2010). Measurement non-invariance of DSM-IV narcissistic personality disorder criteria across age and sex in a population-based sample of Norwegian twins. *International Journal of Methods in Psychiatric Research* 19, 156-166.
32. Neale, M.C., Aggen, S.H., Maes, H.H., Kubarych, T.S., and Schmitt, J.E. (2006). Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors* 31, 1010-1034.
33. Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics* 3, 872-882.
34. Plomin, R. (2008). Behavioral genetics (Macmillan).
35. Neale, M., and Cardon, L. (1992). Methodology for genetic studies of twins and families (Springer).
36. Posthuma, D., and Boomsma, D.I. (2000). A note on the statistical power in extended twin designs. *Behavior Genetics* 30, 147-158.
37. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6, 461-464.
38. Markon, K.E., and Krueger, R.F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics* 34, 593-610.
39. Stein, D.J. (2000). Neurobiology of the obsessive–compulsive spectrum disorders. *Biological Psychiatry* 47, 296-304.
40. Tibbo, P., and Warneke, L. (1999). Obsessive-compulsive disorder in schizophrenia: epidemiologic and biologic overlap. *Journal of Psychiatry and Neuroscience* 24, 15.
41. Reznik, I., Mester, R., Kotler, M., and Weizman, A. (2001). Obsessive-compulsive schizophrenia: a new diagnostic entity? *The Journal of Neuropsychiatry & Clinical Neurosciences* 13, 115-116.
42. Bottas, A., Cooke, R.G., and Richter, M.A. (2005). Comorbidity and pathophysiology of obsessive–compulsive disorder in schizophrenia: Is there evidence for a schizo-obsessive subtype of schizophrenia? *Journal of Psychiatry and Neuroscience* 30, 187.
43. Poyurovski, M., Weizman, A., and Weizman, R. (2006). Obsessive-Compulsive Disorder and Comorbidity; Chapter 4: Schizo-Obsessive Disorder (Nova Science Publishers).

44. Hudziak, J.J., Althoff, R.R., Stanger, C., Beijsterveldt, C.v., Nelson, E.C., Hanna, G.L., Boomsma, D.I., and Todd, R.D. (2006). The Obsessive Compulsive Scale of the Child Behavior Checklist predicts obsessive-compulsive disorder: a receiver operating characteristic curve analysis. *Journal of Child Psychology and Psychiatry* 47, 160-166.
45. Fontenelle, L.F., Lopes, A.P., Borges, M.C., Pacheco, P.G., Nascimento, A.L., and Versiani, M. (2008). Auditory, visual, tactile, olfactory, and bodily hallucinations in patients with obsessive-compulsive disorder. *CNS Spectrums: The International Journal of Neuropsychiatric Medicine* 13.
46. Hermesh, H., Konas, S., Shiloh, R., Reuven, D., Marom, S., Weizman, A., and Gross-Isseroff, R. (2004). Musical hallucinations: prevalence in psychotic and nonpsychotic outpatients. *Journal of Clinical Psychiatry*.
47. Badcock, J.C., Waters, F.A., and Maybery, M. (2007). On keeping (intrusive) thoughts to one's self: Testing a cognitive model of auditory hallucinations. *Cognitive Neuropsychiatry* 12, 78-89.
48. Michie, P.T., Badcock, J.C., Waters, F.A., and Maybery, M.T. (2005). Auditory hallucinations: Failure to inhibit irrelevant memories. *Cognitive Neuropsychiatry* 10, 125-136.
49. Waters, F.A., Badcock, J.C., Maybery, M.T., and Michie, P.T. (2003). Inhibition in schizophrenia: association with auditory hallucinations. *Schizophrenia Research* 62, 275-280.
50. Krüger, S., Bräunig, P., Höffler, J., Shugar, G., Börner, I., and Langkrär, J. (2000). Prevalence of obsessive-compulsive disorder in schizophrenia and significance of motor symptoms. *The Journal of Neuropsychiatry and Clinical Neurosciences* 12, 16-24.
51. Patel, D., Laws, K., Padhi, A., Farrow, J., Mukhopadhaya, K., Krishnaiah, R., and Fineberg, N. (2010). The neuropsychology of the schizo-obsessive subtype of schizophrenia: a new analysis. *Psychological Medicine* 40, 921.
52. Sevincok, L., Akoglu, A., and Kokcu, F. (2007). Suicidality in schizophrenic patients with and without obsessive-compulsive disorder. *Schizophrenia Research* 90, 198-202.
53. Tibbo, P., Kroetsch, M., Chue, P., and Warneke, L. (2000). Obsessive–compulsive disorder in schizophrenia. *Journal of Psychiatric Research* 34, 139-146.
54. Poyurovsky, M., Kriss, V., Weisman, G., Faragian, S., Schneidman, M., Fuchs, C., Weizman, A., and Weizman, R. (2005). Familial aggregation of schizophrenia-spectrum disorders and obsessive-compulsive associated disorders in schizophrenia probands with and without OCD. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 133, 31-36.
55. Wirth, R., and Edwards, M.C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods* 12, 58.

PART II

COPY NUMBER VARIANTS: POST-TWINNING MUTATIONS AND DISCORDANCE BETWEEN MONOZYGOTIC TWINS

CHAPTER 4

DE NOVO AND INHERITED CNVS IN MZ TWIN PAIRS SELECTED FOR DISCORDANCE AND CONCORDANCE ON ATTENTION PROBLEMS

This chapter is based on:

Erik A Ehli*, Abdel Abdellaoui*, Yueshan Hu, Jouke Jan Hottenga, Mathijs Kattenberg, Toos van Beijsterveldt, Meike Bartels, Robert R Althoff, Xiangjun Xiao, Paul Scheet, Eco J de Geus, James J Hudziak, Dorret I Boomsma*, Gareth E Davies*: De Novo and Inherited CNVs in MZ Twin Pairs Selected for Discordance and Concordance on Attention Problem. *European Journal of Human Genetics* 2012; 20(10):1037-43. * *these authors contributed equally*

Abstract

Copy number variations (CNVs) have been reported to be causal suspects in a variety of psychopathologic traits. We investigate whether *de novo* and/or inherited CNVs contribute to the risk for Attention Problems (AP) in children. Based on longitudinal phenotyping, 50 concordant and discordant monozygotic (MZ) twin pairs were selected from a sample of ~3200 MZ pairs. Two types of *de novo* CNVs were investigated: (1) CNVs shared by both MZ twins, but not inherited (pre-twinning *de novo* CNVs), which were detected by comparing copy number (CN) calls between parents and twins; (2) CNVs not shared by co-twins (post-twinning *de novo* CNVs), which were investigated by comparing the CN calls within MZ pairs. The association between the overall CNV burden and AP was also investigated for CNVs genome-wide, CNVs within genes and CNVs outside of genes. Two *de novo* CNVs were identified and validated via qPCR: a pre-twinning *de novo* duplication in a concordant unaffected twin pair, and a post-twinning deletion in the higher scoring twin from a concordant affected pair. For the overall CNV burden analyses, affected individuals had significantly larger CNVs that overlapped with genes than unaffected individuals (p = .008). This study suggests that the presence of larger CNVs may increase the risk for AP, since they are more likely to affect genes, and confirms that MZ twins are not always genetically identical.

Introduction

Copy number variants (CNVs) are polymorphisms in the number of copies of chromosomal segments (duplications and deletions) ranging from 1 kb to several Mb, and have been recognized as a major contributor to human genetic variability. CNVs collectively encompass a larger part of the genome than single-nucleotide polymorphisms (SNPs).[1-3] Mutation rates for CNVs are two to four times higher than those of point mutations and affect larger segments of the genome.[4;5] CNVs have been shown to correlate with changes in gene expression levels.[6-9] Changes in copy number can also lead to the generation of new combinations of exons between different genes, causing protein changes in structure and modified protein activities.[10;11] Therefore CNVs are likely to be involved in phenotypic variation, including disease susceptibility, especially when they are large and affect multiple genes. CNVs can be either inherited or *de novo*, with the assumption that *de novo* CNVs are more likely to have deleterious effects.[12] CNVs have been linked to several neuropsychiatric disorders including schizophrenia, autism, and attention deficit hyperactivity disorder (ADHD).[13-16]

We investigated whether there is an association between CNVs (*de novo* and inherited) and Attention Problems (AP) in a selected sample of concordant and discordant monozygotic (MZ) twin pairs. The AP scale has been shown to be predictive for ADHD. Children who score low on the AP scale of the Child Behavior Checklist (CBCL) have a non-ADHD diagnosis in 96% of the cases, and children with a high AP score have a positive diagnosis for ADHD in 36% (girls) and 59% (boys) of cases.[17] In addition, the sensitivity and specificity of the measure is increased if longitudinal scores on AP are considered. Heritability estimates for AP and ADHD in children are about 70% and 75%, respectively[18;19] and ~75% of the covariance between the AP scale and ADHD has been estimated to be explained by genetic influences.[20] Previous work that included part of the current MZ sample showed structural[21] and functional[22] brain differences in addition to significant behavior differences among the discordant twin pairs[23].

In this study, MZ twins discordant and concordant for AP are examined for the presence of two types of *de novo* CNVs (1) pre-twinning *de novo* CNVs: CNVs that emerged during parental meiosis, and are therefore shared by the MZ twins, but not by the parents (parental genotypes were available for more than half of the subjects); and (2) post-twinning *de novo* CNVs: CNVs that undergo a copy number (CN) change in mitosis during the development of one of the twins, causing a discordance between the MZ twins. Post-twinning *de novo* mutations could result in a genetic discordance in all tissues (due to a premorula mutation, most likely at the two-cell stage) or in somatic mosaicism for genetic discordance (due to mutation at the four-cell stage or later).[24] *De novo* CNVs have been

demonstrated in monozygotic (MZ) twins[25] and are one mechanism by which phenotypic discordance in MZ twins may be explained. Validation of *de novo* CNVs identified through a genome-wide scan is important due to the tendency to discover false positive mutations when using SNP microarray technology.[26] In this study, we employ the use of quantitative PCR (qPCR) to confirm the *de novo* CNVs indentified from the genome-wide scan for CNVs. In addition, the association between the genome-wide CNV burden and AP is investigated for CNVs genome-wide, CNVs overlapping with genes and CNVs outside of genes (for the *de novo* and inherited CNVs pooled together).

Methods

Subjects

Fifty monozygotic (MZ) twin pairs were selected from the Netherlands Twin Register (NTR).[27] Selection was based on longitudinal maternal reports from the Attention Problems (AP) scale of the CBCL.[28] The AP scale has been used to identify children at risk for clinical ADHD and consists of 11 items (e.g. "can't sit still, restless, or hyperactive", "can't concentrate, can't pay attention for long", "impulsive or acts without thinking", etc.). Normative scores are provided for the AP scale, which allows for determining whether a child is at risk for ADHD based on gender and age specific T-scores.[23] The AP scale was collected at ages 7, 10 and 12 years and eligible twin pairs were selected from a total sample of 3228 MZ twin pairs. A total of 1966 MZ twin pairs (birth cohorts 1986 – 1994) had measures from at least two time points and an additional 1256 pairs had longitudinal ratings from all three time points. Children were identified as affected if they had a T-score greater than 60 at all available time-points and a T-score of at least 65 at one or more time points. Children were classified as unaffected if they had a T-score of lower than 55 at all time points. A T-score of 65 represents the clinical cut-off for ADHD.[17] The criterion of longitudinal discordance in MZ twins represents a severe selection measure, as only 18 of the 1966 pairs with longitudinal data available meet this criterion. There were 52 concordant high-scoring twin pairs (both twins affected), 962 concordant low-scoring twin pairs (both twins unaffected) and 18 discordant twin pairs (one affected and one unaffected). The twins were only selected on AP, and not on the presence or absence of any other disorders. AP was not measured for the parents. DNA samples were available for 50 MZ pairs: 17 concordant high (6 male & 11 female pairs), 22 concordant low (8 male & 14 female pairs), and 11 discordant (4 male & 7 female pairs) twin pairs and 36 parent pairs. The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, and an Institutional

Review Board certified by the US Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180).

Genotyping

Twins and their parents provided buccal swabs for DNA extraction. Methods for buccal swab collection, genomic DNA extraction, and zygosity testing have been described previously.[29] Genotyping was performed on the Affymetrix Human Genome-Wide SNP 6.0 Array according to the manufacturer's protocol. This array contains 906,600 SNP and 940,000 copy number probes. Of the copy number probes, 800,000 are evenly spaced across the genome and the rest across 3700 known CNV regions. A total of 172 individuals were genotyped (50 MZ twin pairs and 36 parent pairs). Twins were randomly distributed across plates with respect to AP scores and twins from the same twin pair were genotyped on separate plates. Parents were genotyped together, but not on the same plate as their offspring. Quality control (QC) was done according to protocol and resulted in a total sample size of 153 individuals comprising 45 complete twin pairs (21 concordant low, 10 discordant, and 14 concordant high). Of these 45 complete twin pairs, 25 sets had DNA from both parents pass QC, four complete twin pairs had DNA from one parent pass QC, the unpaired twins had DNA from one parent pass QC, and one unpaired twin had DNA from both parents that passed QC. CNVs were called with the Birdsuite[30] and PennCNV[31] algorithms. CN segments were only included in further analyses if the following conditions were met: 1) the CN calls agreed between both algorithms, 2) the overlapping part of the segments from both algorithms was > 100 kb, and 3) the segment was not in a centromere. Since calling algorithms can produce artificially split CNV calls, adjacent CNV calls were merged after visual inspection of LRR and BAF plots, if the gap in between was ≤ 50% of the entire length of the newly merged CNV (see Supplementary Figure 1 in the online Supplementary Materials of the published version of this article for LRR and BAF plots of all these CNVs for LRR and BAF plots of all these CNVs). The CNV calling and QC procedures are described in more detail in the Appendix.

Pre-twinning de novo CNV detection

CN calls from the 25 MZ twin pairs who had both parents pass QC were examined to detect possible pre-twinning *de novo* CNV events. These segments were identified with a script written in Perl (scripts are available at the online Supplementary Materials of the published version of this article, where segments with the same start and end positions between both twins and both parents as well as overlapping segments were compared. If overlapping segments showed the same

CN between twins and a discrepancy with the parental CN calls, and the overlap was > 100 kb, the overlapping part was included as a *de novo* CNV segment. In order to judge whether a CNV is inherited or *de novo*, allele specific CN information is needed from the parents. Since allele specific CN calls were not available, the allele specific CNs was assumed to be as follows: if CN = 2, each allele is assumed to have a CN of 1 (1-1), if CN = 3, 1-2 is assumed, if CN = 4, 2-2 is assumed, if CN = 1, 1-0 is assumed, if CN = 0, 0-0 is assumed. If possible *de novo* CNVs were detected, these were tested for confirmation using qPCR (see Appendix for more details on the qPCR replication).

Post-twinning de novo CNV detection

The CN calls, passing the above per sample and per CNV quality control thresholds, of the 45 complete MZ twin pairs were analyzed to detect possible post-twinning *de novo* CNV events. These segments were identified with a program written in Perl (scripts are available at the online Supplementary Materials of the published version of this article), where segments with the same start and end positions between twins as well as overlapping segments were compared. If two overlapping segments showed a different CN between twins and a size > 100 kb the overlapping part was identified as a *de novo* CNV segment. Putative *de novo* CNVs were tested for confirmation using qPCR (see Appendix for more details on the qPCR replication).

Statistical analysis for genome-wide CNV burden and AP

Genome-wide CNV burden linked to AP was analyzed with permutation tests in Plink[32] in the 45 complete twin pairs and four unpaired twins. Phenotypes were not permuted between males, females or related individuals, thereby correcting for sex and twin relations. The amount of CNV events as well as the average size was tested for association with AP status. This was done for three groups of CNV events with any deviation from the expected CN (CN = 0, 1, 3 or 4): CNVs genome-wide, CNVs that overlap with genes and CNVs that do not. Significant results were followed by post-hoc tests, by testing gains (CN = 3 or 4), losses (CN = 1 or 0), losses of one copy (CN = 1), losses of two copies (CN = 0), one copy gains (CN = 3), and gains of two copies (CN = 4). Inherited as well as *de novo* CNVs were included in the analysis (*de novo* CNVs that were not validated by qPCR were removed from the analysis). For the male participants, the CNs of the X and Y chromosomes were transformed by adding one copy to the observed CN, in order to include the sex chromosomes with the autosomes in the permutation analysis (i.e., the expected CN of 1 was turned into a CN of 2, like in the autosomes). This transformation was not applied to the pseudoautosomal regions (PARs), since these already have an expected CN of 2.

Results

De novo CNVs

A total of 26 *de novo* CNV events were identified from the microarray data: eight pre-twinning and 18 post-twinning CNVs. CNV qPCR targets for 18 regions in the human genome were identified which would validate all 26 *de novo* CNVs. The primer and probe binding sites for qPCR were selectively chosen in regions within the CNV for which (1) there is no polymorphic SNP, (2) there is no homology to other regions in the genome, and (3) in regions for which there are no common repetitive elements. Based on these criteria, primers and probes could only be selected for 11 of the 18 CNV targets, allowing for testing the validity of 17 of the 26 *de novo* CNVs using the qPCR method (three pre-twinning and 14 post-twinning CNVs).

Of the three possible pre-twinning *de novo* CNVs that could be included in the qPCR replication study, one was validated on chromosome 15q11.2 in a male concordant unaffected twin pair (see Appendix Table 1, Figure 1, and Appendix Figure 1a). In this pedigree, both the microarray and qPCR data show that both parents have a CN of 2 in this region, and that both twins have a CN of 3. Of the 16 putative post-twinning *de novo* CNVs that were included in the qPCR replication study, qPCR experiments validated one *de novo* CNV event: a 1.3 Mb deletion in a male concordant high twin pair, in the higher scoring co-twin (see Appendix Table 2, Figure 1b, and Appendix Figure 1b). In addition, a 116 kb duplication, was not validated nor rejected by the qPCR experiments in the affected twin of a male discordant pair (see Appendix Table 2, Figure 1c, and Appendix Figure 1c).

The 1.3 Mb deletion was initially called as two separate CNVs of 848 kb and 334 kb by Birdsuite and PennCNV. The qPCR targets were designed for both these regions, and the gap in between. All three qPCR experiments resulted in a deletion for the oldest twin, and a CN of 2 for the youngest, confirming that this is indeed one large deletion that was artificially split by the calling algorithms.

Interestingly, qPCR was not able to reject or validate the microarray supported hypothesis of a 116 kb *de novo* CNV duplication in the affected twin of a discordant pair on 17p13.2. Despite both calling algorithms supporting a duplication in this region, the LRR and BAF plots (see Figure 1c) were visually ambiguous, so it was decided to add a second qPCR target to this region 30 kb downstream. The qPCR experiments did not unequivocally validate or refute the presence of a duplication in this region (see Appendix Figure 1c). The experiment was repeated three different times for each target assay, with four sample replicates in each experiment. In each instance the calculated CN for the affected twin was greater than that of unaffected twin (2.34 vs. 1.91 and 2.40 vs. 1.97 for the chr17:5921845 and chr17:5951803 targets respectively).

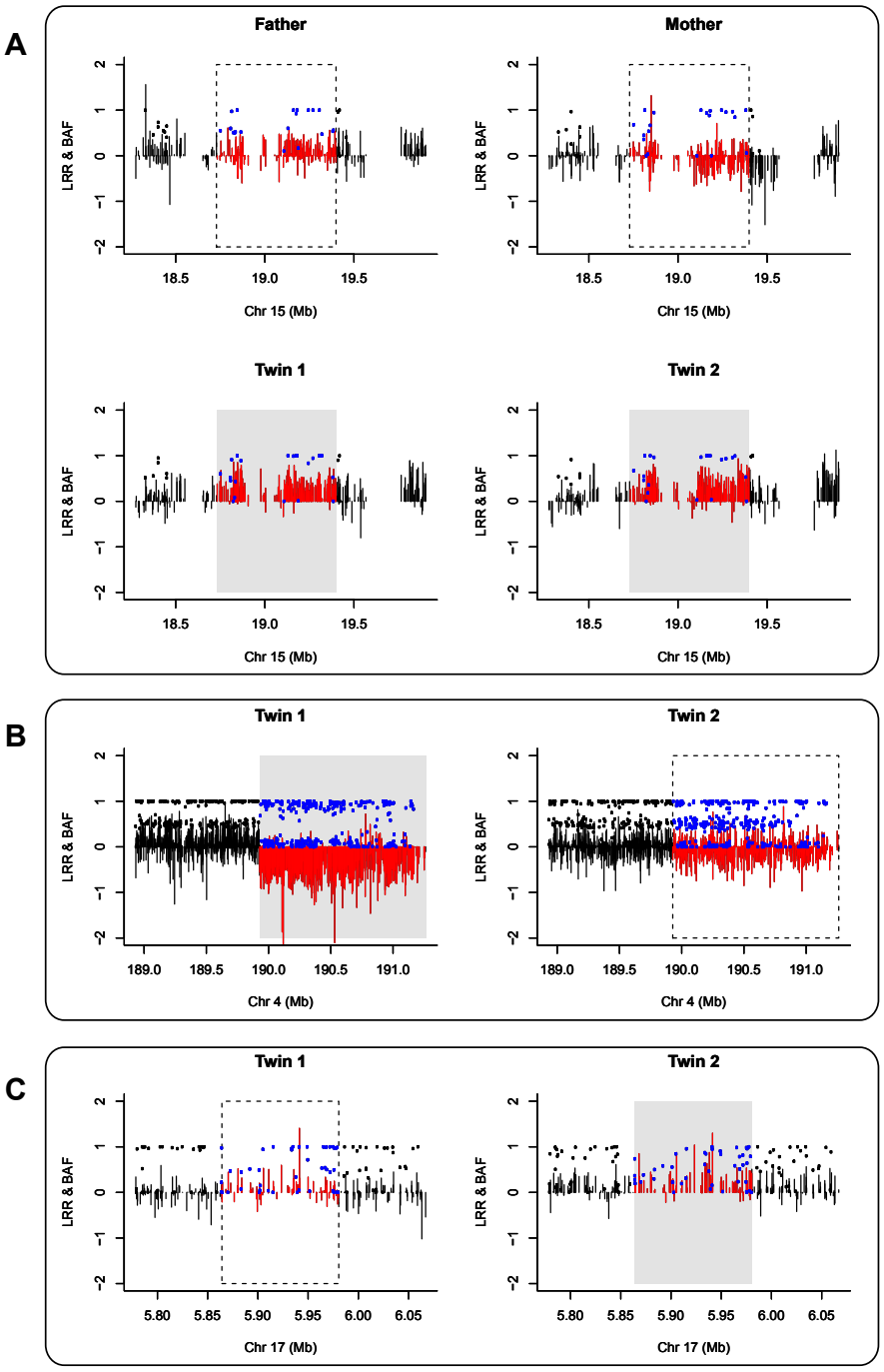

A
Father
Mother
Twin 1
Twin 2
LRR & BAF
Chr 15 (Mb)
B
Twin 1
Twin 2
LRR & BAF
Chr 4 (Mb)
C
Twin 1
Twin 2
LRR & BAF
Chr 17 (Mb)

Figure 1 (previous page). The pre- and post-twinning *de novo* CNVs. Each plot shows LogR Ratio (LRR; vertical bars) and B-allele frequency (BAF; solid points). The LRR and BAF are shown in color in the region of the CNV (red and blue respectively), and in black in the flanking regions. The actual deletion/duplication is highlighted by a gray rectangle, while a CN call of 2 is highlighted by a dashed rectangle. ***A*** depicts the region of the pre-twinning *de novo* duplication in family 34 for both parents and both twins (both unaffected for AP). The duplication is mainly characterized by an increase in LRR in the twins compared to the parents. The clustering of BAF does not show striking differences between the twins and the parents, most likely because there are relatively few SNP probes in this region (CN probes do not have BAF values). ***B*** shows the region of the post-twinning deletion in family 5 for both twins (both affected with AP). The deletion is characterized by a decrease in LRR and an altered clustering of BAF, only seen in twin 1 (the oldest twin). ***C*** shows the region of the possible post-twinning duplication in family 33 for both twins (discordant), where twin 2 is affected with AP. Although both calling algorithms called a *de novo* duplication, the LRR and BAF values do not show striking differences when inspected visually, which is why extra qPCR experiments were conducted for this region.

Table 1: Genes within each confirmed *de novo* CNV region. Genes that are in the RefSeq database (http://www.ncbi.nlm.nih.gov/gene) as well as in the Ensembl database (http://www.ensembl.org/) are reported.

Gene ID	Description	Tissue Expressed (transcripts per million)[a]	Gene Ontology (functioning of gene products)[b]
chr15:18,728,578-19,399,146:			
HERC2P3	Hect domain and RLD 2 pseudogene 3, non-coding RNA	Brain (69); Thyroid (42); Uterus (38); Thymus (36); Testis (24); Pharynx (24); Mammary Gland (13); Stomach (10); Placenta (10); Eye (9); Blood (8); Connective Tissue (6); Prostate (5); Embryonic Tissue (4); Intestine (4); Kidney (4); Liver (4); Skin (4)	*Molecular functions:* metal ion binding; ubiquitin-protein ligase activity.
chr4:189,928,060-191,261,904:			
HSP90AA4P	Heat shock protein 90kDa alpha (cytosolic), class A member 4, pseudogene , non-coding RNA	Ascites (24); Skin (4)	*Cellular components:* cytoplasm. *Molecular functions:* ATP binding; nucleotide binding; unfolded protein binding. *Biological processes:* protein folding; response to stress.
G1	FSHD region gene 1, mRNA.	Bone Marrow (143); Pharynx (72); Pituitary Gland (60); Blood (56); Lymph Node (54); Salivary Gland (49); Ovary (48); Parathyroid (48); Eye (47); Muscle (46); Liver (33); Mammary Gland (32); Stomach (31); Uterus (30); Adrenal Gland (30); Embryonic Tissue (27); Lung (26); Prostate (21); Testis (21); Cervix (20); Connective Tissue (20); Pancreas (18); Kidney (14); Skin (14); Bone (13); Heart (11); Intestine (8); Brain (4); Placenta (3)	*Cellular components:* cajal body; catalytic step 2 spliceosome; nuclear speck; nucleolus; nucleus. *Biological processes:* nuclear mRNA splicing, via spliceosome; RNA splicing; rRNA processing.
TUBB4Q	Tubulin, beta polypeptide 4, member Q, pseudogene, mRNA	Skin (4)	*Cellular components*: cytoplasm; cytoskeleton; microtubule. *Molecular functions*: GTP binding; GTPase activity; nucleotide binding; structural molecule activity *Biological processes:* 'de novo' posttranslational protein folding; cellular protein metabolic process; microtubule-based movement; protein folding; protein polymerization.
FRG2	FSHD region gene 2, mRNA	NA	*Cellular components*: nucleus.
DUX4 Family	Double homeobox 4, mRNA	NA	*Cellular components*: nucleus. *Molecular functions*: sequence-specific DNA binding; sequence-specific DNA binding transcription factor activity.
chr17:5,864,185-5,980,521:			
WSCD1	WSC domain containing 1, mRNA	Muscle (120); Umbilical Cord (73); Pituitary Gland (60); Eye (56); Testis (48); Brain (45); Kidney (37); Ovary (29); Thyroid (21); Bone Marrow (20); Vascular (19); Embryonic Tissue (18); Spleen (18); Placenta (17); Lung (14); Mouth (14); Bone (13); Connective Tissue (13); Heart (11); Pancreas (9); Blood (8)	*Cellular components:* integral to membrane (i.e., penetrating at least one phospholipid bilayer of a membrane); membrane (i.e., double layer of lipid molecules that encloses all cells, and many organelles; may be a single or double lipid bilayer; also includes associated proteins) *Molecular functions:* sulphotransferase activity.

[a]Data from NCBI UniGene (http://www.ncbi.nlm.nih.gov/unigene)

[b]Data from the Gene Ontology Project (http://www.geneontology.org)

NA – Tissue specific gene expression data not available.

Genes located within each of the *de novo* CNV regions are summarized in Table 1. Figure 1 shows the LRR and BAF plots and Appendix Figure 1 displays the qPCR replication data of the *de novo* CNV regions. In addition, Appendix Figure 2 places each of these *de novo* CNVs in a more global context by showing all of the catalogued structural variations from the Database of Genomic Variants (DGV).

Genome-wide CNV burden and AP

There was a nominally significant association with AP and the average size of CNVs within genes, where the affected individuals had larger CNV events than the unaffected group (>120 kb more on average, p = .00830, cf. a level of .00833 (= .05/6) maintains a family-wise type-I error of .05, see Table 2). The post-hoc tests showed that each type of CNV showed the same trend (a larger average CNV size in the affected group, see Table 2), except for the CNVs with deletions of 2 copies (CN = 0), which was the least common type; occurring only seven times (five events in affected individuals and two events in unaffected individuals). None of these types showed a significant signal, suggesting that the significant effect of burden is due to the combined effect of both losses and gains. The average size of the CNVs did not differ significantly between affected and unaffected individuals for the regions outside of genes. The number of CNVs also did not show significant differences, both within and outside of genes.

Table 2: Results for permutation tests for the number of CNVs genome-wide and their size versus AP

	Mean number of CNVs			Average size of CNVs (kb)		
CNV events	Unaffected	Affected	nr of CNVs vs. AP	Unaffected	Affected	CNV size vs. AP
CNVs genome-wide (CN=0, 1, 3, 4)	4.528	3.805	p = .961	242.2	280.0	p = .058
CNVs overlapping genes (CN=0, 1, 3, 4)	2.566	1.854	p = .989	266.6	388.7	p = .008
CNVs outside of genes (CN=0, 1, 3, 4)	2.094	2.000	p = .638	210.7	197.2	p = .738

Discussion

This study investigated the importance of the number and size of CNVs for Attention Problems in "identical" twins. The presence of *de novo* CNV mutations and effects of genome-wide CNV burden were examined.

The pre-twinning *de novo* CNVs were examined for a subset of the sample (25 twin pairs) that had genomic DNA from both parents available and passed QC.

One pre-twinning *de novo* CNV mutation was detected that resulted in both MZ twins having a duplication (CN=3) on chromosome 15q11.2. This region contains the gene *HERC2P3*, which is expressed in the human brain (see Table 1). However, both individuals in this twin set scored in the normal range for AP. We assume this to be a *de novo* pre-twinning CNV event, but we recognize the possibility of a rare condition that one of the parents carries two copies for one allele and zero copies on the other allele, in which case this would not be a *de novo* CNV event.

A post-twinning *de novo* deletion of ~1.3 Mb on 4q35.2 was confirmed with three qPCR experiments in a concordant affected twin pair. The twin with the deletion had a higher AP score, 20% lower birth weight than the co-twin, scored in the clinical range for the DSM oriented CBCL scale for Conduct Problems and performed worse at school according to longitudinal parental and teacher reports. The 4q35.2 subtelomeric deletions found in this twin have been suggested to contribute to co-morbid psychiatric illness and mental retardation.[33] The deletion contains the *FRG1* gene which is expressed in the human brain. In addition, chromosome 4q35 contains a polymorphic D4Z4 macrosatellite repeat, consisting of 10-100 tandem 3.3-kb D4Z4 repeats. An identical copy of the *DUX4* gene (double homeobox) is located in each of the 3.3-kb repeat elements. Contractions in this polymorphic region have been implicated in Facioscapulohumeral Muscular Dystropy (FSHD).[34] The *DUX4* protein has been shown to function as a transcriptional activator of the paired –like homeodomain transcription factor 1 (PITX1),[35] which is expressed in the pituitary gland and brain. *DUX4* is a nuclear protein also capable of acting as a pro-apoptotic protein, inducing cell death through caspase 3/7 activity when overexpressed.[36] Although *FRG1* and *DUX4* have been highly implicated in the pathophysiology of FSHD, our findings and the molecular mechanisms of these proteins make them possible targets for follow-up study on how they may impact the developing brain.

The microarray supported hypothesis of a 116 kb duplication on 17p13.2 in the affected twin of a discordant pair could not be validated or rejected using qPCR (see Appendix Figure 1c). The algorithm for predicting copy number is based on the delta C_T of the reference target (in this case RNaseP) to the CNV target of interest. Although experimental variation can affect the calculated CN of the genomic DNA in a qPCR experiment (e.g. technical reproducibility, genomic DNA quality, etc.),[37;38] in all instances (12 replicates for two assay targets) the affected twin had a larger calculated copy number for this region of the genome. Considering that the Genomic DNA was normalized and the fact that these samples are monozygotic twins makes interpretation of the data difficult. We hypothesize that the duplication in 17p13.2 is a somatic mutation resulting in mosaicism of the affected twin. Somatic mosaicism is generally defined as the presence of genetically distinct populations of cells for a given tissue in the same

organism. It has been suggested that somatic mosaicism in pathogenic genes may be relatively common.[25] We cannot conclusively determine this hypothesis, but it was only possible to detect/suspect this by examining MZ twin pairs. Regions in 17p13.2 have been associated with autism spectrum disorder.[39-41] The *WSCD1* gene from the duplication in 17p13.2 in the affected twin of the discordant pair is expressed in the brain and is involved in the phospholipid bilayer of the membrane (see Table 1), which has been suggested to play a major role in the high degree of comorbidity between ADHD, dyspraxia and autism spectrum disorders[42], which have all been reported by the parents and teachers of the carrier of the putative *de novo* duplication. The unaffected co-twin had an above average IQ and had no health or other problems reported.

Each of the *de novo* CNVs identified in this study has been compared to the catalog of structural variants from the Database of Genomic Variation (Appendix Figure 2). There have been several duplications and deletions reported for the pre-twinning *de novo* CNV on 15q11.2 and the post-twinning deletion on 4q35.2. Interestingly, a slightly larger deletion of 4q35.2, was identified from the Vrije University Hospital clinical database in a child with autism, ADHD, and developmental delay without dysmorphism (personal communication Petra Zwijnenburg). There have not been any duplications reported in the Database for Genomic Variation for the putative *de novo* CNV in the affected twin of a discordant pair on 17p13.2.

The CNVs that were not identified as *de novo* were assumed to be inherited and were included with the *de novo* CNVs in the genome-wide CNV burden association analysis. The association analysis of genome-wide CNV burden and AP showed that CNVs that overlap with genes were larger in size in affected than in unaffected subjects (p = .008). Deletions and duplications showed the same trend, but no significant signals, indicating that both contributed to the main effect. The CNVs that were larger in subjects with high AP scores were scattered across the genome. This suggests that AP might be influenced by many CNVs with small effects, which has been recently revealed to be the case for SNP effects on complex traits as well.[43] Since the majority of human genes are expressed in the cortex,[44] randomly located CNVs affecting genes are likely to have an effect on highly heritable cognitive traits, such as AP. An alternative hypothesis is that neuropsychiatric disorders are caused by rare and highly penetrant CNVs, which often disrupt the balance of dosage sensitive genes.[13;45;46] Studying the genes affected by this disruption may provide important insights into the susceptibility of disease.

Rare events, such as *de novo* CNVs, are hard to detect when the tools used to measure them are relatively noisy, as is the case with CNV signals from microarray chips that are currently available. In this study, this could be especially problematic when trying to detect post-twinning *de novo* CNVs by comparing

twin pairs that were genotyped on separate plates. Stringent QC procedures might not be enough to distinguish real signal from noise, which made replication with qPCR a necessary step to validate the presence of these apparent mutations. In order to accurately detect *de novo* CNVs, it is important to confirm the mutation using a molecular assay more sensitive to copy number alterations than the microarrays used to initially screen for them. qPCR has been shown to be highly effective in the validation of CNVs from microarray data.[26; 47; 48] The outcome of this study shows that even when only considering large CNVs (>100 kb), there can still be a substantial amount of false positives among the few CN differences between the MZ twins, reflecting the difficulty in measuring CNVs accurately. We excluded the source DNA (buccal-derived) as a major factor. In a different sample of twin families in which blood and buccal derived DNAs were collected, we have shown that the CNV calls between blood and buccal sources did not show a greater discordance than those from the same source (e.g. both samples from blood), indicating that buccal-derived DNA is suitable for the microarray chip used in the present study (PS et al, unpublished data). The validated *de novo* CNV however confirms that MZ twins are not always 100% genetically identical and that these differences are detectable. An important question remains: how common are these post-twinning *de novo* mutations? To answer this question in more detail, high-throughput CNV calling methods are needed with higher resolutions and accuracy than the microarray chips currently available. Most heritability studies rely on the assumption that MZ twins are 100% identical.[49; 50] Our study largely supports this assumption, but also suggests that the rare post-twinning *de novo* events may lead to phenotypic discrepancies. As a result, the classical twin design may slightly underestimate the genetic effects of a trait. If CNV discordance between MZ twins contributes to phenotypic discordance, the CNV effect on the phenotype would be inadvertently attributed to unique environmental effects in a classical twin study design.

In conclusion, this study found that CNVs that overlap with genes tend to be larger in individuals that consistently score high on AP and who may also have associated elevations in other behavioral problems. Also, two *de novo* CNVs were detected: a pre-twinning duplication, and a post-twinning deletion that resulted in a discordance in CN between the MZ twins. Replication studies with larger sample sizes are needed to validate the effect of the size of CNVs on AP, and to investigate the effects of the regions where the *de novo* CNVs were found.

Appendix

Quality Control and CNV calling

Affymetrix Genotyping Console (GTC) 4.0 was used to calculate Contrast Quality Control (CQC) and Median of the Absolute values of all Pairwise Differences (MAPD) metrics for a measure of per sample quality control (QC). CQC is a metric for how well the allele intensities separate into clusters, with lower CQC values indicating a higher difficulty for the algorithm in distinguishing homozygotic genotypes from heterozygotic genotypes. MAPD is an estimate of variability or standard deviation, where increased variability decreases the quality of CN calls. Samples with a CQC < 0.4 or an MAPD > 0.35 were excluded. The dataset is considered problematic if more than 10 % of the samples do not pass the CQC cutoff of 0.4 or when the mean CQC is smaller than 1.70.[51] Only eight of the 172 samples had a CQC less than 0.4 and the mean CQC was > 1.70. Of the 164 remaining samples that passed the CQC cutoff, 11 samples had an MAPD > 0.35, leaving a total sample size of 153 individuals comprising 45 complete twin pairs (21 concordant low, 10 discordant, and 14 concordant high). Of these 45 complete twin pairs, 25 sets had DNA from both parents passing QC, four complete twin pairs had DNA from one parent pass QC, the unpaired twins had DNA from one parent passing QC, and one unpaired twin had DNA from both parents that passed QC. Zygosity was confirmed for all these twins by their very high SNP concordance (the lowest was 97.72%, the highest 99.95%, with a mean SNP concordance of 99.46%). The .CEL files for the 153 samples were imported to Birdsuite 1.5.5 and PennCNV to make the CNV calls.

For Birdsuite, the Affymetrix Powertool (APT-1.10.2, plug-in to Birdsuite 1.5.5) was used for plate-wise normalization. All samples passing the initial QC were included in Birdsuite (including the separate parents that were not included in further downstream analyses to aid in the estimation of the probe-specific means and variances used in the Birdseye algorithm).[30] The Birdseye algorithm from Birdsuite 1.5.5 was one of the two algorithms used to make the CN calls. This algorithm searches for consistent evidence for CNVs across multiple neighboring probes. Information from neighboring probes is integrated into a copy number (CN) call (0, 1, 2, 3 or 4) for the segment covered by the probes using a hidden Markov model (HMM) based algorithm.[30] Birdsuite only calls CNs up to 4, because the Affymetrix platform is not designed for detecting CNs above this level. Higher-order CNs will likely be called as 4 due to saturation of probe intensities. A logarithm of the odds ratio (LOD) score was generated for each CNV segment, indicating the likelihood of a CNV relative to no CNV in the region. CNV segments were only included if they had a LOD-score > 10. An additional level of CNV quality control was generated by also calling CNVs with a second algorithm

(PennCNV).

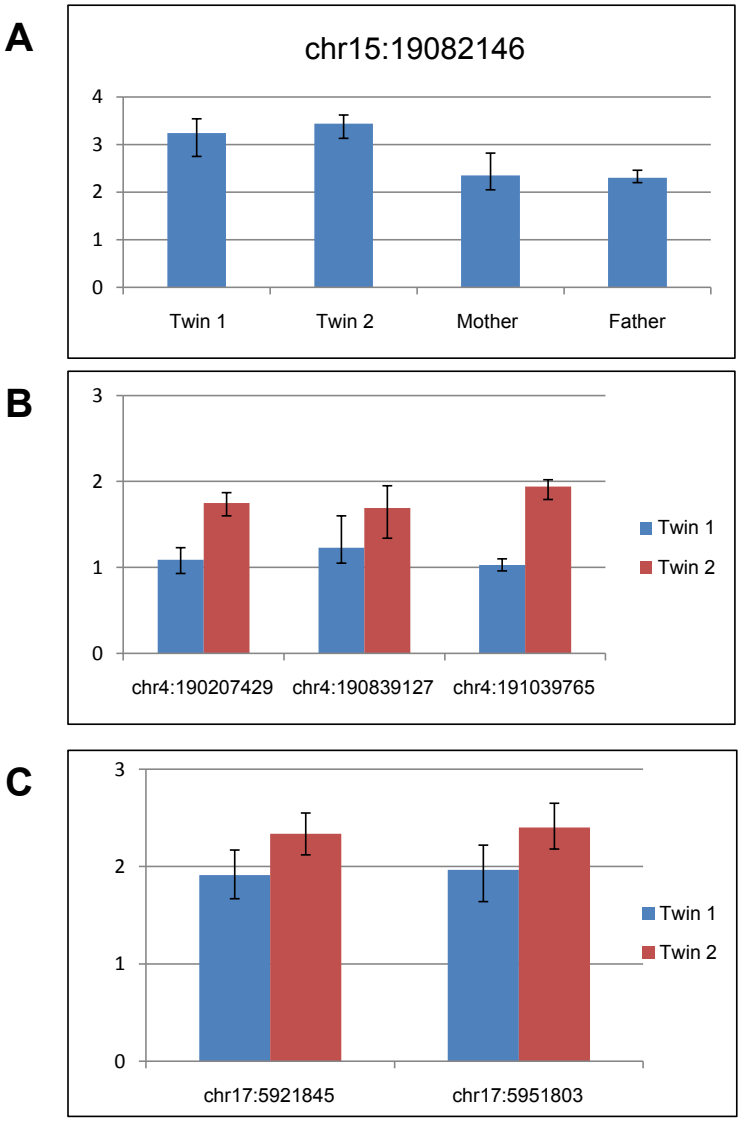
PennCNV (Aug. 2010 version), with a workflow described elsewhere,[31] was used to call genotypes, extract allele-specific signal intensities, cluster canonical genotypes and finally generate a standard input file including log-R ratio (LRR) values and the "B allele" frequency (BAF) for each marker in each individual. PennCNV uses a HMM based approach for kilobase-resolution detection of CNVs. Copy number (CN) calls (0, 1, 2, 3, and 4) for fragments on chromosomes were generated with at least 2 markers. A "genomic waves" effect in calling CNVs was determined by checking whether waviness factor is less than -0.04 or higher than 0.04 and this effect was minimized through an improved version of wave adjustment procedure in PennCNV. CNVs on chromosome X and Y were called by following a specific protocol.

The CN calls of Birdsuite and PennCNV were compared with a script written in Perl (scripts are available at the online Supplementary Materials of the published version of this article). CN segments were only included in further analyses if the following conditions were met: 1) the CN calls agreed between both algorithms, 2) the overlapping part of the segments from both algorithms was > 100 kb, and 3) the segment was not in a centromere. Calls were also included if the CN call in Birdsuite was equal to the expected CN (CN = 2 for autosomes, CN = 1 & 1 and CN = 2 & 0 for X and Y in males and females respectively) and the segment was not present in the PennCNV output, since PennCNV only gives the CN state when the CN deviates from the expected CN, and Birdsuite gives CN states for all segments. Since calling algorithms can produce artificially split CNV calls, adjacent CNV calls were merged after manual inspection of LRR and BAF plots, if the gap in between was ≤ 50% of the entire length of the newly merged CNV (see Supplementary Figure 1 in the online Supplementary Materials of the published version of this article for LRR and BAF plots of all these CNVs). After QC, the highest observed per-sample LRR SD of the probes in the remaining CNVs was 0.154.

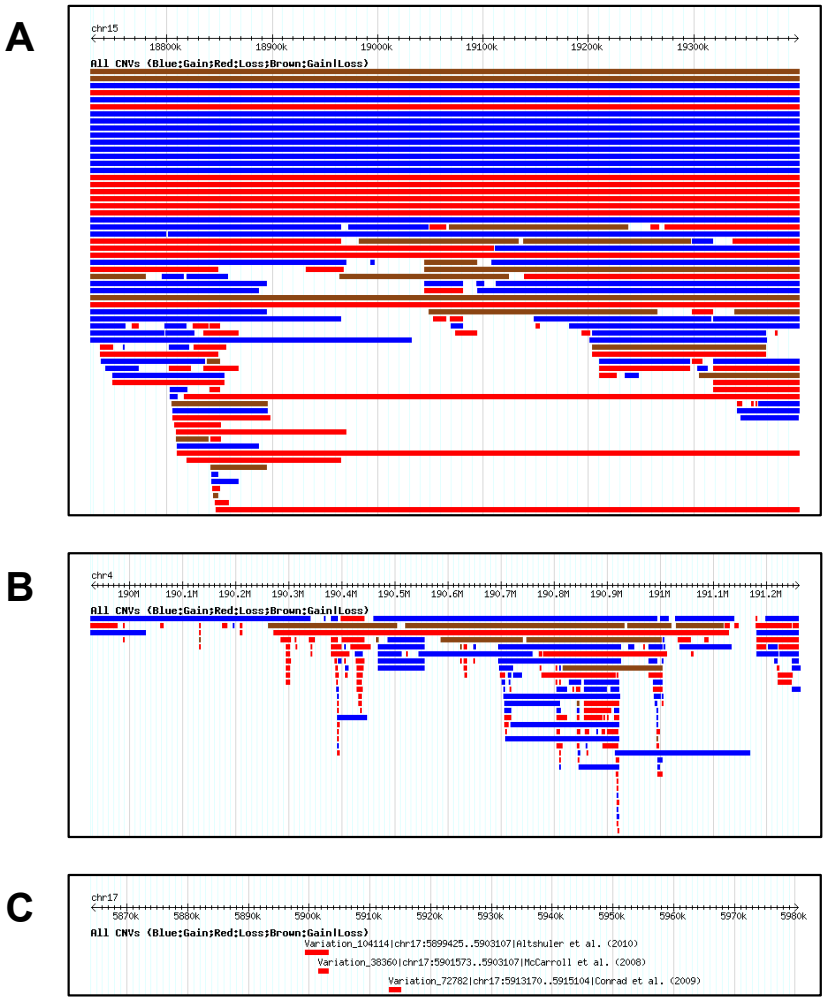
De novo CNV validation with qPCR

The CN of regions with possible *de novo* CNVs (pre- and post-twinning) was re-examined with quantitative Real-Time Polymerase Chain Reaction (qPCR) utilizing TaqMan fluorescently labeled oligonucleotide DNA probes; using RNaseP as an internal reference and a differentially labeled fluorescent probe for the target of interest within the putative CNV. A PCR reaction was performed with the genomic DNA and both probe/primer sets. Briefly, 10ng of genomic DNA was mixed with 1X TaqMan Genotyping MasterMix, 1X VIC labeled RNaseP assay mix (internal reference), and 1X FAM labeled target assay mix using DH_2O for a final reaction volume of 10µl (Applied Biosystems, Foster City, CA, USA). Each sample

was replicated at least four times for accuracy. Cycling conditions consisted of an initial denaturation at 95°C for 10 minutes, 40 cycles of denaturation at 95°C for 15 sec, and annealing and elongation at 60°C for 60 sec on an Applied Biosystems 7900HT Real-time PCR machine. Raw data (C_T) was exported to CopyCaller Software V1.0 (Applied Biosystems, Foster City, CA, USA). Copy Number calls were determined using the software algorithm (copy number assignment without a calibrator sample) when compared to the reference signal from RNaseP, which is assumed to be present at 2 copies in a diploid organism. The CopyCaller Software provides a CN calculated value and a CN predicted value from the raw data. Although the integer for CN calculated could be a whole number with a fractional part, the predicted CN is a whole number (0,1,2,3, etc.) derived from the calculated CN.



Appendix Figure 1. Each bar graph depicts the qPCR data for each of the *de novo* CNV regions. The bar represents the mean calculated copy number with the error line denoting the mean maximum and mean minimum copy number from each of the experimental replicates. ***A*** depicts the qPCR results from the pre-twinning *de novo* CNV on 15q11.2 where both twins have a CN=3 and both parents have a CN =2. ***B*** confirms the deletion on 4q35.2 of twin 1 using three different targets spaced across the CNV and ***C*** summarizes the qPCR data for the putative duplication on 17p13.2 in the affected twin (twin 2) for two different copy number target assays spaced approximately 30 kb apart. Coordinates are from hg18.



Appendix Figure 2. Graphical representation of all the copy number variations from the Database of Genomic Variants (http://projects.tcag.ca/variation/) for each of the de novo CNVs identified in the study. Each CNV in the catalog is denoted by a different color bar (Blue=Gain; Red=Loss; Brown= Gain/Loss). ***A*** shows the catalogued structural variations in the region representing the pre-twinning *de novo* CNV on 15q11.2 (chr15:18728578-19399146) where both twins have a CN=3 and both parents have a CN =2. ***B*** depicts the structural variations in the database for the region comprising the deletion on 4q35.2 (chr4:189928060-191261904) of twin 1 in a concordant affected twin pair. ***C*** summarizes a total of three small deletions in the database from the region of the putative duplication on 17p13.2 (chr17:5864185-5980521) in the affected twin (twin 2) of a discordant twin pair. All coordinates are from hg18.

Appendix Table 1: Pre-twinning de novo CNV events (rows in bold are confirmed by qPCR).

Family ID	CL/D/ CH	Chr	Start	End	Size	Affy6 CN father	Affy6 CN mother	Affy6 CN twins	qPCR CN father	qPCR CN mother	qPCR CN twins
30	CL	9	68312776	68476231	163455	2	2	1	NA	NA	NA
30	CL	9	69100250	69205261	105011	2	2	1	NA	NA	NA
34	**CL**	**15**	**18728578**	**19399146**	**670568**	**2**	**2**	**3**	**2**	**2**	**3**
38	CL	15	18846002	19566875	720873	2	2	1	2	3	2
39	CL	15	19393208	19548935	155727	2	2	3	NA	NA	NA
43	CL	17	31552238	31653809	101571	2	2	3	NA	NA	NA
44	CL	17	41987366	42107479	120113	2	2	1	NA	NA	NA
6	CH	15	18846002	19566875	720873	2	2	3	2	3	3

CL = concordant low, D = discordant, CH = concordant high. Genome coordinates are based on NCBI36/hg18.

NA - Not available, qPCR probes could not be designed.

Appendix Table 2: List of post-twinning de novo CNVs (rows in bold are confirmed by qPCR).

Family ID	CL/D/ CH	Chr	Start	End	Size	Affy 6 CN twin 1	Affy 6 CN twin 2	qPCR CN twin 1	qPCR CN twin 2	AP twin 1	AP twin 2	Mean AP twin 1	Mean AP twin 2
28	CL	9	43523459	43720905	197447	2	3	3	3	UA	UA	47.95	43.70
29	CL	15	18522250	18655543	133294	2	1	1	1	UA	UA	44.57	43.70
33	CL	1	16758722	17076084	317362	3	2	2	2	UA	UA	44.16	50.43
33	CL	2	87481276	87833445	352170	3	2	3	3	UA	UA	44.16	50.43
33	CL	8	12284675	12487426	202752	3	2	NA	NA	UA	UA	44.16	50.43
38	CL	1	16741950	16843043	101094	1	2	1	1	UA	UA	49.70	43.37
45	CL	1	16741950	16859438	117489	3	2	2	2	UA	UA	41.38	40.32
49	CL	15	19488342	19794591	306250	1	2	NA	NA	UA	UA	44.60	50.23
18	D	15	18866712	19195198	328487	3	2	3	3	UA	A	47.63	94.06
18	**D**	**17**	**5864185**	**5980521**	**116337**	**2**	**3**	**2**	**3**	**UA**	**A**	**47.63**	**94.06**
23	D	10	46094186	46366738	272553	2	3	NA	NA	UA	A	49.47	65.09
26	D	8	7011977	7213846	201870	2	3	2	2	UA	A	44.01	67.06
26	D	15	18522250	19080173	557924	2	3	3	3	UA	A	44.01	67.06
27	D	17	41784437	42107479	323043	3	2	NA	NA	UA	A	50.37	66.81
5	**CH**	**4**	**189928060**	**191261904**	**1333844**	**1**	**2**	**1**	**2**	**A**	**A**	**81.22**	**64.62**
8	CH	1	147540169	147703466	163298	3	2	2	2	A	A	68.28	74.34
15	CH	15	18652835	18841578	188744	2	1	1	1	A	A	76.50	79.89
16	CH	15	18544080	19503764	959684	3	4	3	3	A	A	80.19	93.65

CL = concordant low, D = discordant, CH = concordant high. A = affected, UA = unaffected, Twin 1 = oldest twin, Twin 2 = youngest twin.
Genome coordinates are based on NCBI36/hg18

References

1. Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* 453, 56-64.
2. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318, 420.
3. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444-454.
4. Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. *Nature Genetics* 39, S43-S47.
5. van Ommen, G.J.B. (2005). Frequency of new copy number variation in humans. *Nature Genetics* 37, 333-334.
6. Aldred, P.M.R., Hollox, E.J., and Armour, J.A.L. (2005). Copy number polymorphism and expression level variation of the human -defensin genes DEFA1 and DEFA3. *Human Molecular Genetics* 14, 2045.
7. Hollox, E.J., Armour, J.A.L., and Barber, J.C.K. (2003). Extensive normal copy number variation of a [beta]-defensin antimicrobial-gene cluster. *American Journal of Human Genetics* 73, 591-600.
8. Linzmeier, R.M., and Ganz, T. (2005). Human defensin gene copy number polymorphisms: Comprehensive analysis of independent variation in [alpha]-and [beta]-defensin regions at 8p22-p23. *Genomics* 86, 423-430.
9. McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., et al. (2005). Common deletion polymorphisms in the human genome. *Nature Genetics* 38, 86-92.
10. Rotger, M., Saumoy, M., Zhang, K., Flepp, M., Sahli, R., Decosterd, L., and Telenti, A. (2007). Partial deletion of CYP2B6 owing to unequal crossover with CYP2B7. *Pharmacogenetics and genomics* 17, 885.
11. Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D., and Lupski, J.R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics* 41, 849-853.
12. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., De Bakker, P., Maller, J.B., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* 40, 1166-1174.
13. Cook Jr, E.H., and Scherer, S.W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature* 455, 919-923.
14. Merikangas, A.K., Corvin, A.P., and Gallagher, L. (2009). Copy-number variants in neurodevelopmental disorders: promises and challenges. *Trends in Genetics* 25, 536-544.
15. Moreno-De-Luca, D., and Cubells, J.F. (2011). Copy Number Variants: A New Molecular Frontier in Clinical Psychiatry. *Current psychiatry reports*, 1-9.
16. Williams, N.M., Zaharieva, I., Martin, A., Langley, K., Mantripragada, K., Fossdal, R., Stefansson, H., Stefansson, K., Magnusson, P., et al. (2010). Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *The Lancet*.
17. Derks, E.M., Hudziak, J.J., Dolan, C.V., Ferdinand, R.F., and Boomsma, D.I. (2006). The relations between DISC-IV DSM diagnoses of ADHD and multi-informant CBCL-AP syndrome scores. *Comprehensive Psychiatry* 47, 116-122.
18. Faraone, S.V., Perlis, R.H., Doyle, A.E., Smoller, J.W., Goralnick, J.J., Holmgren, M.A., and Sklar, P. (2005). Molecular genetics of attention-deficit/hyperactivity disorder. *Biological Psychiatry* 57, 1313-1323.
19. Derks, E.M., Hudziak, J.J., and Boomsma, D.I. (2009). Genetics of ADHD, Hyperactivity, and Attention Problems. In Handbook of behavior genetics (New York: Springer Verlag).

20. Derks, E.M., Hudziak, J.J., Dolan, C.V., van Beijsterveldt, T.C.E.M., Verhulst, F.C., and Boomsma, D.I. (2008). Genetic and environmental influences on the relation between attention problems and attention deficit hyperactivity disorder. *Behavior Genetics* 38, 11-23.
21. van 't Ent, D., Lehn, H., Derks, E.M., Hudziak, J.J., Van Strien, N.M., Veltman, D.J., De Geus, E.J., Todd, R.D., and Boomsma, D.I. (2007). A structural MRI study in monozygotic twins concordant or discordant for attention/hyperactivity problems: evidence for genetic and environmental heterogeneity in the developing brain. *Neuroimage* 35, 1004-1020.
22. van 't Ent, D., van Beijsterveldt, C.E., Derks, E.M., Hudziak, J.J., Veltman, D.J., Todd, R.D., Boomsma, D.I., and De Geus, E.J. (2009). Neuroimaging of response interference in twins concordant or discordant for inattention and hyperactivity symptoms. *Neuroscience* 164, 16-29.
23. Lehn, H., Derks, E.M., Hudziak, J.J., Heutink, P., van Beijsterveldt, T., and Boomsma, D.I. (2007). Attention problems and attention-deficit/hyperactivity disorder in discordant and concordant monozygotic twins: evidence of environmental mediators. *Journal of the American Academy of Child & Adolescent Psychiatry* 46, 83-91.
24. Vadlamudi, L., Dibbens, L.M., Lawrence, K.M., Iona, X., McMahon, J.M., Murrell, W., Mackay-Sim, A., Scheffer, I.E., and Berkovic, S.F. (2010). Timing of de novo mutagenesis--a twin study of sodium-channel mutations. *New England Journal of Medicine* 363, 1335.
25. Bruder, C.E.G., Piotrowski, A., Gijsbers, A.A.C.J., Andersson, R., Erickson, S., Diaz de Ståhl, T., Menzel, U., Sandgren, J., von Tell, D., et al. (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *American Journal of Human Genetics* 82, 763-771.
26. Ono, S., Imamura, A., Tasaki, S., Kurotaki, N., Ozawa, H., Yoshiura, K., and Okazaki, Y. (2010). Failure to Confirm CNVs as of Aetiological Significance in Twin Pairs Discordant for Schizophrenia. *Twin Research and Human Genetics* 13, 455-460.
27. Boomsma, D.I., de Geus, E.J.C., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.J., Posthuma, D., van Beijsterveldt, T.C.E.M., Hudziak, J.J., et al. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics* 9, 849-857.
28. Achenbach, T. (1991). Manual for the CBCL/4-18 and 1991 Profile. *Burlington: University of Vermont, Department of Psychiatry[Links]*.
29. Willemsen, G., de Geus, E.J.C., Bartels, M., van Beijsterveldt, C.E.M.T., Brooks, A.I., Estourgie-van Burk, G.F., Fugman, D.A., Hoekstra, C., Hottenga, J.J., et al. (2010). The Netherlands Twin Register Biobank: a resource for genetic epidemiological studies. *Twin Research and Human Genetics* 13, 231-245.
30. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics* 40, 1253.
31. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* 17, 1665.
32. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81, 559-575.
33. Pickard, B.S., Hollox, E.J., Malloy, M.P., Porteous, D.J., Blackwood, D.H., Armour, J.A., and Muir, W.J. (2004). A 4q35. 2 subtelomeric deletion identified in a screen of patients with co-morbid psychiatric illness and mental retardation. *BMC medical genetics* 5, 21.
34. Deutekom, J.C.T.V., Wljmenga, C., Tlenhoven, E.A.E.V., Gruter, A.M., Hewitt, J.E., Padberg, G.W., Ommen, G.J.B., Hofker, M.H., and Fronts, R.R. (1993). FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Human Molecular Genetics* 2, 2037.
35. Dixit, M., Ansseau, E., Tassin, A., Winokur, S., Shi, R., Qian, H., Sauvage, S., Mattéotti, C., Van Acker, A.M., et al. (2007). DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. *Proceedings of the National Academy*

of Sciences of the United States of America 104, 18157.
36. Kowaljow, V., Marcowycz, A., Ansseau, E., Conde, C.B., Sauvage, S., Mattéotti, C., Arias, C., Corona, E.D., Nuñez, N.G., et al. (2007). The DUX4 gene at the FSHD1A locus encodes a pro-apoptotic protein. *Neuromuscular Disorders* 17, 611-623.
37. Cukier, H.N., Pericak-Vance, M.A., Gilbert, J.R., and Hedges, D.J. (2009). Sample degradation leads to false-positive copy number variation calls in multiplex real-time polymerase chain reaction assays. *Analytical biochemistry* 386, 288-290.
38. Fernandez-Jimenez, N., Castellanos-Rubio, A., Plaza-Izurieta, L., Gutierrez, G., Irastorza, I., Castaño, L., Vitoria, J.C., and Bilbao, J.R. (2011). Accuracy in Copy Number Calling by qPCR and PRT: A Matter of DNA. *PloS one* 6, e28910.
39. Bremer, A., Giacobini, M.B., Eriksson, M., Gustavsson, P., Nordin, V., Fernell, E., Gillberg, C., Nordgren, A., Uppströmer, Å., et al. (2010). Copy number variation characteristics in subpopulations of patients with autism spectrum disorders. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics*.
40. Cannon, D.S., Miller, J.S., Robison, R.J., Villalobos, M.E., Wahmhoff, N.K., Allen-Brady, K., McMahon, W.M., and Coon, H. (2010). Genome-wide linkage analyses of two repetitive behavior phenotypes in Utah pedigrees with autism spectrum disorders. *Molecular Autism* 1, 1-13.
41. Joober, R., and El-Husseini, A. (2006). Synaptic Abnormalities and Candidate Genes in Autism. In Molecular mechanisms of synaptogenesis. (New York: Springer Verlag).
42. Richardson, A.J., and Ross, M. (2000). Fatty acid metabolism in neurodevelopmental disorder: a new perspective on associations between attention-deficit/hyperactivity disorder, dyslexia, dyspraxia and the autistic spectrum. *Prostaglandins Leukotrienes and Essential Fatty Acids* 63, 1-9.
43. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* 42, 565-569.
44. Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., et al. (2007). A survey of genetic human cortical gene expression. *Nature Genetics* 39, 1494-1499.
45. Inoue, K., and Lupski, J.R. (2003). Genetics and genomics of behavioral and psychiatric disorders. *Current Opinion in Genetics & Development* 13, 303-309.
46. Lee, J.A., and Lupski, J.R. (2006). Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron* 52, 103-121.
47. Weaver, S., Dube, S., Mir, A., Qin, J., Sun, G., Ramakrishnan, R., Jones, R.C., and Livak, K.J. (2010). Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods* 50, 271-276.
48. Zhang, D., Qian, Y., Akula, N., Alliey-Rodriguez, N., Tang, J., Gershon, E.S., and Liu, C. (2011). Accuracy of CNV Detection from GWAS Data. *PloS one* 6, e14511.
49. Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics* 3, 872-882.
50. Plomin, R., DeFries, J., McClearn, G., and McGuffin, P. (2008). Behavioral genetics (Vol. 5). *New York: Worth*.
51. Affymetrix. (2008). Affymetrix White Paper: Quality Control Assessment in Genotyping Console. http://media.affymetrix.com/support/technical/whitepapers/genotyping_console_cqc_whitepaper.pdf.

CHAPTER 5

CNV CONCORDANCE IN 1,097 MZ TWIN PAIRS

Abdel Abdellaoui, Erik A. Ehli, Jouke-Jan Hottenga, Zachary Weber, Hamdi Mbarek, Gonneke Willemsen, Toos van Beijsterveldt, Andrew Brooks, Jim J. Hudziak, Patrick F. Sullivan, Eco J. de Geus, Gareth E. Davies, Dorret I. Boomsma

Abstract

Monozygotic (MZ) twins are genetically identical at conception, making them informative subjects for studies on somatic mutations. Copy number variants (CNV) are responsible for a substantial part of genetic variation, have relatively high mutation rates, and are likely to be involved in phenotypic variation. We conducted a genome-wide survey for post-twinning *de novo* CNVs in 1,097 MZ twin pairs. Comparisons between MZ twins were made by CNVs measured in DNA from blood (mostly adults) or buccal epithelium (mostly children) with the Affymetrix 6.0 microarray and two calling algorithms.

We found a total of 153 putative post-twinning *de novo* CNVs >100 kb, of which the majority resided in 15q11.2. Based on how well the raw intensity signals visually agreed with CNV calls made by the two algorithms, a selection was made of eleven *de novo* CNVs from 15q11.2 for a first series of qPCR validation experiments. Two out of eleven post-twinning *de novo* CNVs were validated with qPCR in the same twin pair. The thirteen year old MZ twin pair that showed two discordances in CN in 15q11.2 in their buccal DNA did not show large phenotypic differences. The remaining putative *de novo* CNVs from 15q11.2 were found significantly more often in DNA from blood (mostly adults), and significantly more in older twins within the dataset with blood samples.

We validated post-twinning CNVs that did not lead to great phenotypic discordance and resided in an unstable genomic region (15q11.2) that was overrepresented among the rest of the putative *de novo* CNVs. More qPCR validation experiments are planned for the remaining putative *de novo* CNVs in order to further investigate the possible role of tissue and/or age in the post-twinning mutation rate, whether these mutations result in phenotypic discordance, or whether relatively large CNV mutations in regions associated with psychiatric disease can be phenotypically tolerated. Older twin pairs were more likely to carry a putative somatic mutation in the 15q11.2 region, suggesting that we are capturing real signals.

Introduction

Monozygotic (MZ) twins have been long assumed to be genetically identical, which is an important assumption in twin studies, where phenotypic correlations between MZ twins and dizygotic (DZ) twins are compared in order to estimate the relative contribution of genes and environment in human traits.[1] MZ twins are in fact genetically identical at conception, but can accumulate mutations after the zygote splits up; making MZ twins informative for the study of somatic mutations. Post-twinning point mutations have been reported on,[2-6] but are expected to be scarcer than post-twinning *de novo* copy number variants (CNVs). CNVs, the most studied type of structural variant (SV), are segments of DNA ranging from 1kb to several Mb that differ in copy number (CN) across different members of the species. CNVs have a higher mutation rate than single nucleotide polymorphisms (SNPs) and affect larger segments of the genome.[7-9] Even though post-twinning *de novo* CNVs are expected to be rare, they can potentially aid in finding causal variants for genomic disorders. After Bruder et al (2008)[10] demonstrated the existence of CNV discordance in MZ twins, many studies followed that tried to find CNV discordances that might be explanatory for phenotypic MZ discordances.

Table 1 shows an overview of studies attempting to detect CNV differences between MZ twins since 2008. Forsberg et al (2012)[11] conducted the largest study of this kind to date examining 159 MZ pairs, and validated five post-twinning mutations >1 Mb and five < 1 Mb, all found in the older twin pairs of their sample (> 60 years old). An estimate of the post-twinning mutation rate for CNVs is difficult to make with this design, since it is likely to depend on the age of the twins, with older individuals having an increased chance for somatic mutations,[3; 11] and likely also depends on tissue.[12] The majority of studies looking for CNV discordances in MZ twins did not detect reproducible post-twinning CNV mutations, indicating that relatively large CNV discordances between MZ twins are a considerably rare phenomenon, or are at least hard to detect, even among phenotypically discordant twins.

Most studies on post-twinning *de novo* CNVs first scan the entire genome using genome-wide microarray technology, making them only sensitive for relatively large CNVs (> 10-100 kb), and then validate suggestive signals with additional and more sensitive molecular assays, such as qPCR. In practice, CNVs have been considered relatively noisy when using currently available genome-wide microarray technologies, and qPCR has shown to be effective in validating CNV signals from microarray data.[13; 14]

We conduct a genome-wide scan for post-twinning *de novo* CNVs (>100kb) in 1,097 unselected MZ twin pairs with a wide age range (0-79 years old). DNA was extracted from blood for about half of the samples, which

Table 1: List of studies searching for post-twinning de novo CNVs.

Study	N (pairs)	Phenotype	Age	Platform	Nr of post-twinning de novo CNVs replicated with qPCR
Bruder et al., 2008[10]	19	Unselected (N=10); discordant for neurodegenerative disease (N=9)	Unselected: 37-60 Discordant: >60	32k BAC array & Illumina HumanHap 300 Duo beadchip	3 with qPCR replication + many supported only by 32k BAC array & Illumina beadchip
Baranzini et al., 2010[15]	3	Discordant for multiple sclerosis	19, 39, 56	Affymetrix 6.0, Illumina GAIIx	0
Lasa et al., 2010[16]	1	Discordant for breast cancer	61	Agilent Human Genome CGH 244k microarray	0
Ono et al., 2010[17]	3	Discordant for schizophrenia	$\mu = 53$	Affymetrix 6.0	0
Jakobsen et al., 2011[18]	1	Discordant for cleft lip	Adult	CGH & Affymetrix 5.0	0
Maiti et al., 2011[19]	2	Discordant for schizophrenia	43, 53	Affymetrix 6.0	0 (*but several without qPCR replication*)
Pamphlett and Morahan, 2011[20]	5	Healthy (N=2), discordant for amyotrophic lateral sclerosis (N=3)	35-74	Affymetrix 6.0	2 in blood in 1 discordant pair, 7 in hair in 1 discordant pair (*no qPCR replication performed*)
Sasaki et al., 2011[21]	1	Discordant for multiple system atrophy	67	Agilent SurePrint G3 Human CNV 400K Microarray; Illumina 57k whole-genome CNV beadchip	1 replicated by non-qPCR method
Breckpot et al., 2012[22]	6	Discordant for congenital heart defects	0 (N=5), 4 (N=1)	OGT CytoSure Syndrome Plus array	3 (all in 1 affected twin)
Ehli et al., 2012[23]	50	Attention Problems: concordant affected (N=17), concordant unaffected (N=22), discordant (N=11)	<13	Affymetrix 6.0	2 (1 in concordant unaffected, 1 in discordant pair)
Forsberg et al., 2012[11]	159	Unselected	<60 (N=81), >60 (N=78)	Illumina 1M-Duo, Illumina Omni-Express, Nimblegen 135K, Nimblegen 720K	5 > 1Mb, 5<1Mb

Table 1 (continued): List of studies searching for post-twinning de novo CNVs.

Study	N (pairs)	Phenotype	Age	Platform	Nr of post-twinning de novo CNVs replicated with qPCR
Halder et al., 2012[24]	1	Discordant for DiGeorge Syndrome with major cardiac malformation and cardiac failure	0	Fluorescence in situ hybridization (FISH) & Illumina HumanCytoSNP-12 BeadChip	1
Solomon et al., 2012[25]	2	Discordant for VACTERL association-type congenital malformations	Not reported	Illumina Omni1-Quad SNP array; Agilent SureSelect Human All Exon	0
Veenma et al., 2012[26]	11	Discordant for esophageal atresia (N=7), discordant for congenital diaphragmatic hernia (N=4)	Not reported	Illumina HumanCytoSNP-12	0
Baudisch et al., 2013[27]	4	Discordant for anorectal malformations (N=3) or for bladder exstrophy-epispadias complex (N=1)	0, 5, 9, 13	HumanOmni1-Quad v1	0
Bloom et al., 2013[28]	5	Discordant for schizoaffective disorder (N=2), bipolar disorder (N=2) or schizophrenia (N=1)	>60	Roche Nimblegen 2.1 Mprobe CGH array	0
Furukawa et al., 2013[29]	1	Discordant for systemic lupus erythematosus	38	Illumina Human 610-Quad BeadChips	0
Miyake et al., 2013[30]	1	Discordant for Rett syndrome	13	Illumina HumanOmni2.5-Quad, Illumina HiSeq2000	0
Laplana et al., 2014[31]	1	Discordant for autism spectrum disorder	24	Agilent 400K CGH array	0

included the majority of adult subjects >18 years old, and the other half of the samples (mainly children) had their DNA extracted from buccal swabs (see Figure 1). CNVs were measured with the Affymetrix 6.0 microarray, and after stringent QC a first series of CNV candidates were selected for qPCR replication. Phenotypic data based on extensive longitudinal questionnaires[32] are available to be examined for twin pairs with validated post-twinning mutations.

Methods

Participants

The 1,097 MZ twin pairs included in this study were registered with the Netherlands Twin Registry (NTR),[33;34] and were not selected based on phenotypic information. SNPs from the Affymetrix 6.0 microarray confirmed that all twins were indeed monozygotic. The mean age of the twins was 25.04 (SD = 15.86), and ranged from 0 to 79 years old (see Figure 1). DNA was extracted from blood for 1,163 twins (mean age = 35.53, SD = 13.24), and from buccal epithelium for 1,031 twins (mean age = 13.11, SD = 8.39). Methods for buccal and blood collection and genomic DNA extraction have been described previously.[35]

CNV calling

Data from 1,097 MZ twin pairs were extracted from a dataset containing a total of 13,188 samples that were genotyped on the Affymetrix Human Genome-Wide SNP 6.0 Array according to the manufacturer's protocol. This array contains 906,600 SNP and 940,000 copy number probes. Of the copy number probes, 800,000 are evenly spaced across the genome and the rest across 3700 known CNV regions. SNPs were called using Affymetrix Powertool, and were used during the QC stage and to confirm the zygosity of MZ twins. CNVs were called with the Birdsuite[36] and PennCNV[37] algorithms.

For Birdsuite 1.5.5 , the Affymetrix Powertool (APT-1.10.2, plug-in to Birdsuite 1.5.5) was used for plate-wise normalization. This algorithm searches for consistent evidence for CNVs across multiple neighboring probes. Information from neighboring probes is integrated into a copy number (CN) call (0, 1, 2, 3 or 4) for the segment covered by the probes using a hidden Markov model (HMM) based algorithm. A logarithm of the odds ratio (LOD) score was generated for each CNV segment, indicating the likelihood of a CNV relative to no CNV in the region. CNV segments were only included if they had a LOD-score > 10. We followed the recommendation from the manual in creating batches (http://www.broadinstitute.org/science/programs/medical-and-population-genetics/birdsuite/birdsuite-faq), and processed a maximum of 96 samples per batch. If the plate of

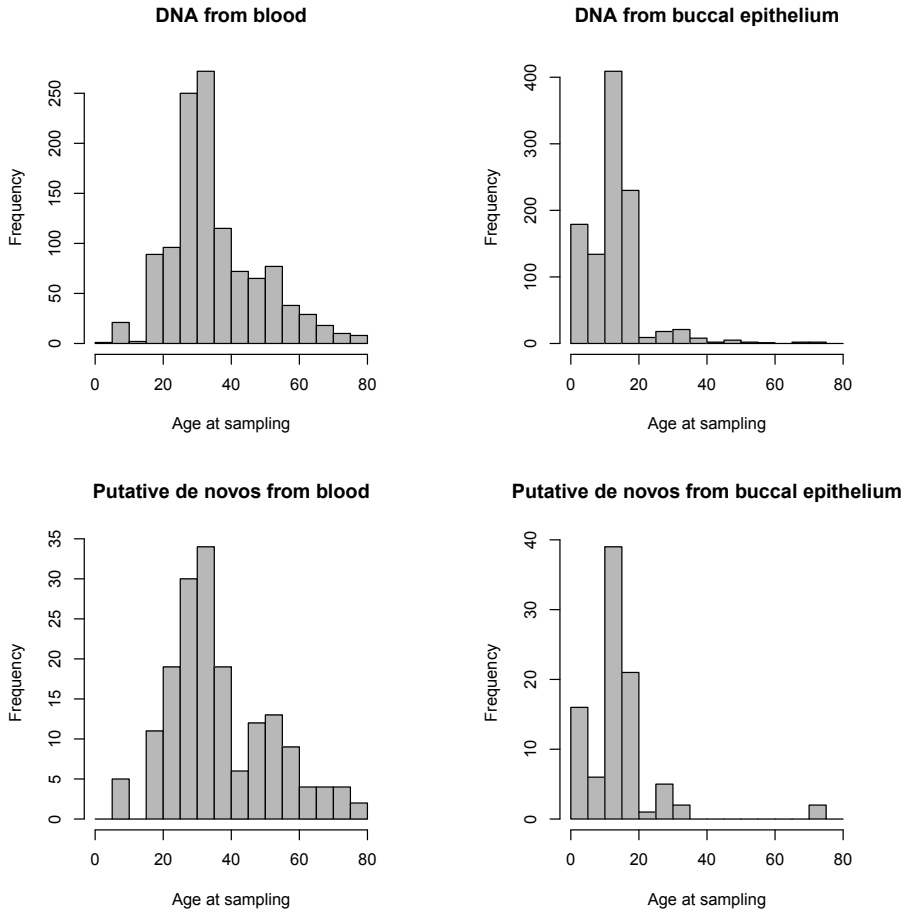


Figure 1: Age distribution of twins per tissue type for all samples (above) and the 152 putative de novo CNVs (below).

origin was known, samples from the same plate were included in the same batch, resulting in 178 batches. Samples where the plate of origin was not known (~3%) were randomly distributed across five batches.

PennCNV was used to call genotypes, extract allele-specific signal intensities, cluster canonical genotypes and finally generate a standard input file including log-R ratio (LRR) values and the "B allele" frequency (BAF) for each marker in each individual. PennCNV uses a HMM based approach for kilobase-resolution detection of CNVs. We followed the recommendation from the manual in creating batches (http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.html), and processed as many samples per calling batch as possible, resulting in four batches (one batch including all twins and duplicates with N=4,182, and three batches with N=3,002 per batch).

The CN calls of Birdsuite and PennCNV were compared with a script written in Perl. CN segments were only included in further analyses if the following conditions were met: 1) the CN calls agreed between both algorithms, 2) the overlapping part of the segments from both algorithms was > 100 kb, and 3) the segment was not in a centromere. Calls were also included if the CN call in Birdsuite was equal to the expected CN (CN=2) and the segment was not present in the PennCNV output, since PennCNV only gives the CN state when the CN deviates from the expected CN, and Birdsuite gives CN states for all segments. Since calling algorithms can produce artificially split CNV calls, adjacent CNV calls were merged after manual inspection of LRR and BAF plots, if the gap in between was ≤ 50% of the entire length of the newly merged CNV.
Individuals were excluded from CNV calling if they had: 1) CQC < .4; 2) SNP missingness > 10%; 3) F based on SNPs > .10 or < -.10; 4) if they had > 50 CNVs with CN≠2. After QC, 12,559 samples remain with a mean CQC of 2.17 (datasets are considered problematic if the mean CQC is smaller than 1.70).

CN calls of complete MZ twin pairs passing QC (N=1,097, mean CQC=2.25) were analyzed to detect possible post-twinning de novo CNV events. Segments with CN differences between MZ twins were extracted with a purpose written perl script, which compares segments with the same start and end positions between twins, as well as overlapping segments.
As an additional quality control, LRR & BAF plots were created for the putative *de novo* CNV segments and were visually inspected by AA and EE. CNVs with LRR & BAF plots that showed a discordance in the same direction as the Birdsuite and PennCNV calls were chosen as a first series of qPCR validation candidates. There were also plots in which the raw intensity signals showed a clear discordance, but were not entirely in line with the Birdsuite and PennCNV calls; these were not chosen for qPCR validation, but will be further investigated in the near future. CNVs that passed all quality control steps were validated using qPCR.

qPCR validation

Calibrator Sample Selection: We selected a sample with CN=2 in the regions included in the qPCR experiments as a calibrator sample, which was used to calibrate the qPCR assay to what a signal from CN=2 should look like. Calibrator samples were selected using Affymetrix 6.0 and next generation sequence data from the partially overlapping NTR-GoNL[38] database (total overlap between the NTR-Affymetrix 6 and GoNL dataset = 81 samples). For these 81 individuals, we first selected samples that showed CN=2 in Birdsuite and no call from PennCNV for the 15q11.2 region. From this set, we then selected samples that showed no CN calls in the GoNL sequence data for two CNV calling algorithms, CNVnator[39] and DWAC-seq (http://tools.genomes.nl/dwac-seq.html), since these algorithms, like

PennCNV, only make calls when CN≠2. After visual inspection of the LRR&BAF plots for the remaining samples (N=10), we then selected one calibrator sample with CN=2 for the qPCR experiments.

CNV Confirmation by qPCR: Samples identified as possible copy number variation (CNV) candidates (N=11 MZ pairs) were removed from -20˚C storage at the Avera Institute for Human Genetics, quantitated using Qubit 2.0 Broad Range Assay (Life Technologies, Carlsbad, CA), and normalized to 5ng/ul. Proposed copy number variations were validated using quantitative real-time polymerase chain reaction (qPCR). Four TaqMan Copy Number Assays (See Table 3) were run on a Viia7 real-time PCR machine (Life Technologies, Carlsbad, CA). TaqMan Copy Number Reference Assay RNase P (Life Technologies, Carlsbad, CA) was used as an internal reference because it is known to exist in two copies in a diploid genome. The copy number assay reporter was FAM and the RNase P reference assay reporter was VIC. All four assays were performed on genomic DNA and run in 384 well PCR plates, with individual reaction volumes of 10ul. Each sample was run with four replicates for accuracy. The four assay plates each contained the respective CNV candidates along with one non-template control sample and one calibrator sample (CN=2). Using ViiA7 Software v1.2, the Ct threshold was set to manual with a value of 0.2 and Auto-baseline was selected to "ON." PCR conditions included an initial hold at 95˚C for 10 minutes, and then 95˚C for 15 seconds followed by 60˚C for 1 minute, together repeated for 40 cycles.

Data generated from the four CNV assays were analyzed with CopyCaller Software v2.0 (Life Technologies, Carlsbad, CA). Ct values from both the copy number assay and the reference assay were exported as (.txt) files to CopyCaller. Analysis settings incorporated a calibrator sample with CN=2. Comparative Ct (ΔΔCt) relative quantitation analysis was performed and sample copy numbers were called using the software algorithm. The ΔΔCt analysis method first determines the difference in Ct value (ΔCt) between the target regions and the reference assay, then it determines the difference between those ΔCt values and the calibrator sample (ΔΔCt). With this information, the CopyCaller Software generates both a calculated and a predicted CN value.

Statistical analyses

Pearson correlations of LOD-scores between co-twins and Pearson correlations of the number of probes between co-twins were computed in IBM SPSS Statistics 21. A chi-squared test was conducted in IBM SPSS Statistics 21 in order to test whether the putative *de novo* CNVs were more likely to be obtained from blood or buccal samples. The difference in age between samples that showed a putative *de novo* CNV in 15q11.2 and the rest of the samples was tested with a t-test for blood and buccal separately in IBM SPSS Statistics 21.

Results

There were 556 CNV segments that showed a CN discordance between MZ twins >100kb. The LOD-scores from the Birdsuite calls showed a significant negative correlation within twin pairs (r = -.247, $p = 3.5 \times 10^{-9}$), as did the number of probes encompassing the CNV (r = -.248, $p = 3.1 \times 10^{-9}$). More than 70% of these calls (N=400) showed an overlap of <10% between twins from the same twin pair (note that the overlap had to be >100 kb). This indicates that many CN discordances may be caused by inaccurate CNV breakpoint estimates and/or a quality difference in CN calls. After only including CNVs with an overlap between twins of >10%, 153 putative *de novo* CNVs remained, of which the correlations of the LOD-scores and number of probes between co-twins were no longer significant (r = -.029, p = .724, and r = -.036, p = .654 respectively). Of these 153 CNVs, more than half (N=90; 58.8%) were from chromosome 15q11.2 ranging from bp positions 18,466,953 to 20,776,822 (build 36). Of the 90 CNVs in 15q11.2, 65 (72.2%) were found in blood derived samples, which mainly came from older twin pairs. LRR & BAF plots were generated for both twins for all 153 CNVs. These LRR & BAF plots were inspected manually in order to select putative *de novo* CNVs suited for qPCR replication. Thirteen CNVs showed LRR&BAF plots that visually agreed with the CNV calls (inspected by by AA and EE), of which eleven were in chromosome 15q11.2 and were followed up with qPCR validation experiments.

Two CNVs in the same twin pair showed a CNV discordance in the qPCR experiments for two CNVs in 15q11.2 (~350kb in 18,491,920-18,841,578, and ~280 kb in 19,090,388-19,369,260; see Table 2 and Figure 2). The twin pair was thirteen years old at the time of sampling, and their DNA was extracted from a buccal epithelium sample. They do not show large phenotypic differences with respect to overall health, behavior, (birth) length, (birth) weight, or other physical appearance in longitudinal parental and self-report questionnaires from age 1 to 21. The twin with CN=3 for both CNVs (twin 2 in Table 2 and Figure 2) did perform better in school and finished high school two levels higher than the twin with CN=1 and CN=2, consistent with their CITO (http://www.cito.nl/) score difference (10 points higher for twin 2).

The remaining 142 putative *de novo* CNVs were found significantly more often in DNA from blood than in DNA from buccal epithelium (65.6% of twin pairs had blood-derived DNA; χ^2 (1) = 17.40, $p = 3 \times 10^{-5}$). When excluding CNVs from the 15q11.2 region (which were also more prevalent in blood samples when excluding the eleven regions included in qPCR: 74.4%), the association between carrying a somatic mutation and source of DNA was not significant anymore (54.9% of twin pairs had blood-derived DNA; χ^2 (1) = .490, p = .484). As nearly all young twins were done on buccal epithelium,

Table 2: CNV calls from Affymetrix 6.0 and qPCR experiments for eleven putative de novo CNVs from 15q11.2 (bold = CNV discordance in qPCR experiments)

								qPCR results			
ID	Source	Age	chr	start	end	size	CN call Affy 6	CN Predicted	CN Calculated	Min CN	Max CN
1-1	BLOOD	18	15	18,527,876	18,700,552	172,676	4	3	3.14	2.80	3.46
1-2	BLOOD	18	15	18,527,876	18,700,552	172,676	2	3	3.38	3.23	3.53
2-1	BUCCAL	13	15	18,846,002	19,379,050	533,048	3	4	3.77	3.37	4.11
2-2	BUCCAL	13	15	18,846,002	19,379,050	533,048	2	4	3.72	3.47	3.92
3-1[a]	**BUCCAL**	**13**	**15**	**18,491,920**	**18,841,578**	**349,658**	**1**	**1**	**1.07**	**0.96**	**1.12**
3-2[a]	**BUCCAL**	**13**	**15**	**18,491,920**	**18,841,578**	**349,658**	**2**	**3**	**2.54**	**2.42**	**2.67**
4-1[b]	**BUCCAL**	**13**	**15**	**19,090,388**	**19,369,260**	**278,872**	**1**	**2**	**1.54**	**1.47**	**1.64**
4-2[b]	**BUCCAL**	**13**	**15**	**19,090,388**	**19,369,260**	**278,872**	**3**	**3**	**3.29**	**2.77**	**3.60**
5-1	BUCCAL	12	15	18,700,552	19,090,596	390,044	3	3	3.46	3.13	3.65
5-2	BUCCAL	12	15	18,700,552	19,090,596	390,044	2	4	3.76	3.50	4.00
6-1	BUCCAL	13	15	18,816,444	19,090,596	274,152	3	3	2.54	2.39	2.62
6-2	BUCCAL	13	15	18,816,444	19,090,596	274,152	2	2	2.43	2.39	2.46
7-1	BLOOD	34	15	18,544,080	18,841,578	297,498	2	1	1.20	1.08	1.27
7-2	BLOOD	34	15	18,544,080	18,841,578	297,498	1	1	1.20	1.12	1.31
8-1	BLOOD	29	15	18,522,250	18,810,700	288,450	1	1	1.11	1.08	1.15
8-2	BLOOD	29	15	18,522,250	18,810,700	288,450	2	1	1.08	1.01	1.13
9-1	BLOOD	50	15	19,103,576	19,438,545	334,969	2	4	3.64	3.60	3.72
9-2	BLOOD	50	15	19,103,576	19,438,545	334,969	3	4	3.66	3.56	3.89
10-1	BLOOD	25	15	19,107,947	19,466,195	358,248	2	3	3.48	3.38	3.75
10-2	BLOOD	25	15	19,107,947	19,466,195	358,248	3	3	3.47	3.34	3.60
11-1	BLOOD	40	15	18,866,712	19,488,342	621,630	1	2	1.76	1.60	1.96
11-2	BLOOD	38	15	18,866,712	19,488,342	621,630	2	2	2.04	1.98	2.09

[a,b] See Figure 2a and 2b respectively for LRR&BAF plots for these two CNVs.

Table 3: TaqMan Copy Number Assay Names and Chromosome Locations

Copy Number Assay Name	Chromosome Location (hg 19)
Hs04452190_cn	Chr15:20602985
Hs03908328_cn	Chr15:20394238
Hs03900971_cn	Chr15:20549958
Hs04450511_cn	Chr15:20863835

and adult samples in blood, we checked whether the age difference might have contributed to the overrepresentation of blood-derived samples among the 15q11.2 mutations. For both blood and buccal epithelium samples, samples with a putative *de novo* CNV showed a higher average age than the rest of the samples from the same source without a putative *de novo* CNV (39.08 vs. 35.18 for blood; 15.16 vs. 13.03 for buccal epithelium samples), but these differences were only significant for the blood samples (p = .013 and p = .413 respectively).

Discussion

We searched for post-twinning *de novo* CNV mutations >100 kb in ~1,100 unselected MZ twin pairs using the Affymetrix 6.0 microarray. CNVs were called using two algorithms, which resulted in 153 significant signals, of which the majority came from the 15q11.2 region. Eleven candidates from 15q11.2 were selected for qPCR replication based on visual inspection of 153 LRR & BAF plots. Two out of eleven were validated, suggesting the remaining 142 putative *de novo* mutations may also contain a substantial proportion of false positives. The majority of the remaining putative somatic mutations resided in 15q11.2, among which blood-derived samples were significantly overrepresented. It was recently shown in a partially overlapping sample of young subjects (genotyped on the same microarray with the same two calling algorithms) that DNA from blood is not more or less likely to show MZ discordance for CNVs than DNA derived from buccal epithelium, and that the quality of the CN calls was stable across longitudinally sampled individuals in a time interval of 12 years.[40] The significant overrepresentation of blood-derived samples among the putative somatic mutations thus may be explained by true mutations that increase with age, as 1) blood-derived samples were predominantly adult as opposed to buccal-derived samples, 2) carriers of putative 15q11.2 mutations from both blood and buccal epithelium showed a higher average age than the rest of the samples from the same tissue (significantly higher for blood), and 3) previous studies have shown that *de novo* mutations increase with age.[3; 11; 41]

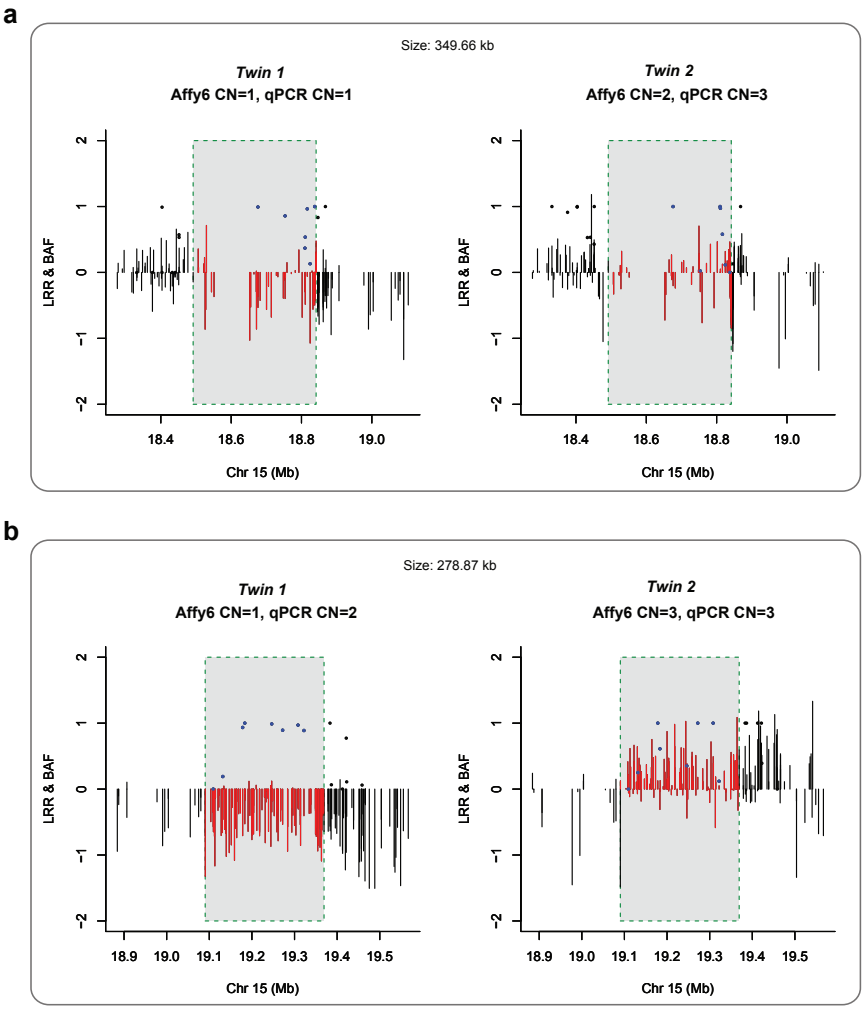


Figure 2: Log R Ration (LRR) for CNV probes & B allele frequency (BAF) for SNP probes of the two validated post-twinning mutations in the same 13 year old twin pair (see Table 1 for bp positions and more details on the qPCR results for a and b respectively). LRR is shown in vertical bars and BAF in solid points. The LRR&BAF values are shown in color in the region of the post-twinning de novo CNV (red and blue respectively), and in black in the flanking regions.

Two post-twinning CNVs in 15q11.2 were replicated in a young MZ twin pair that showed no large phenotypic differences. CNVs in 15q11.2 have been associated with Prader-Willi and Angelman syndromes,[42] schizophrenia,[43] behavioural disturbances,[44] developmental and language delay,[45] epilepsy,[46] and more recently with decreased fecundity, dyslexia, dyscalculia, and brain structure

changes that are associated with schizophrenia and dyslexia.[47] The 15q11.2 region is one of the genomic regions rich in segmental duplications,[48] which makes CNVs in these regions harder to detect, but also means this region is enriched for CNVs and more prone to *de novo* CNV mutations through non-allelic homologous recombination.[49]

None of the 153 LRR & BAF plots were as convincing as the post-twinning CNV found on chromosome 4 in a previous study we conducted on 50 MZ pairs that were selected out of ~3200 MZ pairs based on concordance or discordance for attention problems, where we used a similar methodology.[23] We did have a higher qPCR confirmation rate in this study than in Ehli et al (2012) however, due to a more stringent QC (only CNVs with an overlap between twins of >10% were allowed), but still had a relatively large amount of non-replicated *de novo* events. We plan to conduct more qPCR experiments for the remaining putative *de novo* CNVs in the near future to answer a number of important questions. To what extent do post-twinning CNVs coincide with phenotypic differences between MZ twins? Can humans phenotypically tolerate relatively large CNV mutations, even in regions associated with a wide range of pathology? Is the overrepresentation of 15q11.2 among the microarray signals due to the difficulty of accurately measuring CNVs in regions enriched for segmental duplications, or is this region truly enriched for large post-twinning *de novo* CNVs? If our signals are a result of noise due to segmental duplications, then why are signals from 15q11.2 more abundant than other regions rich in segmental duplications?

References

1. Boomsma, D.I., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. *Nat Rev Genet* 3, 872-882.
2. Vadlamudi, L., Dibbens, L.M., Lawrence, K.M., Iona, X., McMahon, J.M., Murrell, W., Mackay-Sim, A., Scheffer, I.E., and Berkovic, S.F. (2010). Timing of de novo mutagenesis--a twin study of sodium-channel mutations. *N Engl J Med* 363, 1335-1340.
3. Ye, K., Beekman, M., Lameijer, E.W., Zhang, Y., Moed, M.H., van den Akker, E.B., Deelen, J., Houwing-Duistermaat, J.J., Kremer, D., et al. (2013). Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Res Hum Genet* 16, 1026-1032.
4. Kondo, S., Schutte, B.C., Richardson, R.J., Bjork, B.C., Knight, A.S., Watanabe, Y., Howard, E., de Lima, R.L., Daack-Hirsch, S., et al. (2002). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* 32, 285-289.
5. Sakuntabhai, A., Ruiz-Perez, V., Carter, S., Jacobsen, N., Burge, S., Monk, S., Smith, M., Munro, C.S., O'Donovan, M., et al. (1999). Mutations in ATP2A2, encoding a Ca2+ pump, cause Darier disease. *Nat Genet* 21, 271-277.
6. Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., et al. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat Biotechnol* 30, 61-68.
7. Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. *Nature genetics* 39, S43-S47.
8. van Ommen, G.-J.B. (2005). Frequency of new copy number variation in humans. *Nature genetics* 37, 333-334.
9. Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J., and Eichler, E.E. (2010). De novo rates and selection of large copy number variation. *Genome Res* 20, 1469-1481.
10. Bruder, C.E., Piotrowski, A., Gijsbers, A.A., Andersson, R., Erickson, S., Diaz de Ståhl, T., Menzel, U., Sandgren, J., von Tell, D., et al. (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *The American Journal of Human Genetics* 82, 763-771.
11. Forsberg, L.A., Rasi, C., Razzaghian, H.R., Pakalapati, G., Waite, L., Thilbeault, K.S., Ronowicz, A., Wineinger, N.E., Tiwari, H.K., et al. (2012). Age-related somatic structural changes in the nuclear genome of human blood cells. *Am J Hum Genet* 90, 217-228.
12. Piotrowski, A., Bruder, C.E., Andersson, R., Diaz de Stahl, T., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., et al. (2008). Somatic mosaicism for copy number variation in differentiated human tissues. *Hum Mutat* 29, 1118-1124.
13. Weaver, S., Dube, S., Mir, A., Qin, J., Sun, G., Ramakrishnan, R., Jones, R.C., and Livak, K.J. (2010). Taking qPCR to a higher level: analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods* 50, 271-276.
14. Zhang, D., Qian, Y., Akula, N., Alliey-Rodriguez, N., Tang, J., Gershon, E.S., and Liu, C. (2011). Accuracy of CNV detection from GWAS data. *PLoS One* 6, e14511.
15. Baranzini, S.E., Mudge, J., van Velkinburgh, J.C., Khankhanian, P., Khrebtukova, I., Miller, N.A., Zhang, L., Farmer, A.D., Bell, C.J., et al. (2010). Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* 464, 1351-1356.
16. Lasa, A., y Cajal, T.R., Llort, G., Suela, J., Cigudosa, J., Cornet, M., Alonso, C., Barnadas, A., and Baiget, M. (2010). Copy number variations are not modifiers of phenotypic expression in a pair of identical twins carrying a BRCA1 mutation. *Breast cancer research and treatment* 123, 901-905.
17. Ono, S., Imamura, A., Tasaki, S., Kurotaki, N., Ozawa, H., Yoshiura, K.-i., and Okazaki, Y. (2010). Failure to confirm CNVs as of aetiological significance in twin pairs discordant for schizophrenia. *Twin Research and Human Genetics* 13, 455-460.
18. Jakobsen, L.P., Bugge, M., Ullmann, R., Schjerling, C.K., Borup, R., Hansen, L., Eiberg, H., and Tommerup, N. (2011). 500K SNP array analyses in blood and saliva showed no differences in a pair of monozygotic twins discordant for cleft lip. *American Journal of Medical Genetics*

Part A 155, 652-655.
19. Maiti, S., Kumar, K.H.B.G., Castellani, C.A., O'Reilly, R., and Singh, S.M. (2011). Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. *PLoS One* 6, e17125.
20. Pamphlett, R., and Morahan, J.M. (2011). Copy number imbalances in blood and hair in monozygotic twins discordant for amyotrophic lateral sclerosis. *Journal of Clinical Neuroscience* 18, 1231-1234.
21. Sasaki, H., Emi, M., Iijima, H., Ito, N., Sato, H., Yabe, I., Kato, T., Utsumi, J., and Matsubara, K. (2011). Copy number loss of (src homology 2 domain containing)-transforming protein 2 (SHC2) gene: discordant loss in monozygotic twins and frequent loss in patients with multiple system atrophy. *Mol Brain* 4, 24.
22. Breckpot, J., Thienpont, B., Gewillig, M., Allegaert, K., Vermeesch, J., and Devriendt, K. (2012). Differences in copy number variation between discordant monozygotic twins as a model for exploring chromosomal mosaicism in congenital heart defects. *Molecular syndromology* 2, 81-87.
23. Ehli, E.A., Abdellaoui, A., Hu, Y., Hottenga, J.J., Kattenberg, M., van Beijsterveldt, T., Bartels, M., Althoff, R.R., Xiao, X., et al. (2012). De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on Attention Problems. *European Journal of Human Genetics* 20, 1037-1043.
24. Halder, A., Jain, M., Chaudhary, I., and Varma, B. (2012). Chromosome 22q11. 2 microdeletion in monozygotic twins with discordant phenotype and deletion size. *Mol Cytogenet* 5, 13.
25. Solomon, B., Pineda-Alvarez, D., Hadley, D., Hansen, N., Kamat, A., Donovan, F., Chandrasekharappa, S., Hong, S.-K., Roessler, E., et al. (2012). Exome Sequencing and High-Density Microarray Testing in Monozygotic Twin Pairs Discordant for Features of VACTERL Association. *Molecular syndromology* 4, 27-31.
26. Veenma, D., Brosens, E., de Jong, E., van de Ven, C., Meeussen, C., Cohen-Overbeek, T., Boter, M., Eussen, H., Douben, H., et al. (2012). Copy number detection in discordant monozygotic twins of Congenital Diaphragmatic Hernia (CDH) and Esophageal Atresia (EA) cohorts. *European Journal of Human Genetics* 20, 298-304.
27. Baudisch, F., Draaken, M., Bartels, E., Schmiedeke, E., Bagci, S., Bartmann, P., Nöthen, M.M., Ludwig, M., and Reutter, H. (2013). CNV Analysis in Monozygotic Twin Pairs Discordant for Urorectal Malformations. *Twin Research and Human Genetics* 16, 802-807.
28. Bloom, R.J., Kähler, A.K., Collins, A.L., Chen, G., Cannon, T.D., Hultman, C., and Sullivan, P.F. (2013). Comprehensive analysis of copy number variation in monozygotic twins discordant for bipolar disorder or schizophrenia. *Schizophrenia research* 146, 289-290.
29. Furukawa, H., Oka, S., Matsui, T., Hashimoto, A., Arinuma, Y., Komiya, A., Fukui, N., Tsuchiya, N., and Tohma, S. (2013). Genome, epigenome and transcriptome analyses of a pair of monozygotic twins discordant for systemic lupus erythematosus. *Human immunology* 74, 170-175.
30. Miyake, K., Yang, C., Minakuchi, Y., Ohori, K., Soutome, M., Hirasawa, T., Kazuki, Y., Adachi, N., Suzuki, S., et al. (2013). Comparison of genomic and epigenomic expression in monozygotic twins discordant for Rett syndrome. *PLoS One* 8, e66729.
31. Laplana, M., Royo, J.L., Aluja, A., López, R., Heine-Sunyer, D., and Fibla, J. (2014). Absence of Substantial Copy Number Differences in a Pair of Monozygotic Twins Discordant for Features of Autism Spectrum Disorder. *Case Reports in Genetics* 2014.
32. Boomsma, D.I., De Geus, E.J., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.-J., Posthuma, D., Van Beijsterveldt, T.C., Hudziak, J.J., et al. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics* 9, 849-857.
33. Willemsen, G., Vink, J.M., Abdellaoui, A., den Braber, A., van Beek, J.H., Draisma, H.H., van Dongen, J., van't Ent, D., Geels, L.M., et al. (2013). The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Research and Human Genetics* 16, 271-281.
34. van Beijsterveldt, C.E., Groen-Blokhuis, M., Hottenga, J.J., Franić, S., Hudziak, J.J., Lamb, D., Huppertz, C., de Zeeuw, E., Nivard, M., et al. (2013). The Young Netherlands Twin

Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin Research & Human Genetics* 16.
35. Willemsen, G., De Geus, E.J., Bartels, M., Van Beijsterveldt, C., Brooks, A.I., Estourgie-van Burk, G.F., Fugman, D.A., Hoekstra, C., Hottenga, J.-J., et al. (2010). The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Research and Human Genetics* 13, 231-245.
36. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics* 40, 1253-1260.
37. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research* 17, 1665-1674.
38. Boomsma, D.I., Wijmenga, C., Slagboom, E.P., Swertz, M.A., Karssen, L.C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., et al. (2014). The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics* 22, 221-227.
39. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research* 21, 974-984.
40. Scheet, P., Ehli, E.A., Xiao, X., van Beijsterveldt, C.E., Abdellaoui, A., Althoff, R.R., Hottenga, J.J., Willemsen, G., Nelson, K.A., et al. (2012). Twins, tissue, and time: an assessment of SNPs and CNVs. *Twin Research & Human Genetics* 15.
41. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature* 488, 471-475.
42. Donlon, T. (1988). Similar molecular deletions on chromosome 15q11. 2 are encountered in both the Prader-Willi and Angelman syndromes. *Human genetics* 80, 322-328.
43. Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., et al. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature* 455, 232-236.
44. Doornbos, M., Sikkema-Raddatz, B., Ruijvenkamp, C.A., Dijkhuizen, T., Bijlsma, E.K., Gijsbers, A.C., Hilhorst-Hofstee, Y., Hordijk, R., Verbruggen, K.T., et al. (2009). Nine patients with a microdeletion 15q11. 2 between breakpoints 1 and 2 of the Prader–Willi critical region, possibly associated with behavioural disturbances. *European journal of medical genetics* 52, 108-115.
45. Burnside, R.D., Pasion, R., Mikhail, F.M., Carroll, A.J., Robin, N.H., Youngs, E.L., Gadi, I.K., Keitges, E., Jaswaney, V.L., et al. (2011). Microdeletion/microduplication of proximal 15q11. 2 between BP1 and BP2: a susceptibility region for neurological dysfunction including developmental and language delay. *Human genetics* 130, 517-528.
46. de Kovel, C.G., Trucks, H., Helbig, I., Mefford, H.C., Baker, C., Leu, C., Kluck, C., Muhle, H., Von Spiczak, S., et al. (2010). Recurrent microdeletions at 15q11. 2 and 16p13. 11 predispose to idiopathic generalized epilepsies. *Brain* 133, 23-32.
47. Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G.B., Jonsdottir, G.A., et al. (2014). CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* 505, 361-366.
48. Zody, M.C., Garber, M., Sharpe, T., Young, S.K., Rowen, L., O'Neill, K., Whittaker, C.A., Kamal, M., Chang, J.L., et al. (2006). Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature* 440, 671-675.
49. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444-454.

PART III

POPULATION GENETICS: GENOMIC STRUCTURE OF THE NETHERLANDS

CHAPTER 6

POPULATION STRUCTURE, MIGRATION, AND DIVERSIFYING SELECTION IN THE NETHERLANDS

This chapter is based on:

Abdel Abdellaoui, Jouke-Jan Hottenga, Peter de Knijff, Michel G Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, Erik A Ehli, Yueshan Hu, Gareth E Davies, James J Hudziak, Patrick F Sullivan, Toos van Beijsterveldt, Gonneke Willemsen, Eco J de Geus, Brenda W J H Penninx, Dorret I Boomsma: Population Structure, Migration, and Diversifying Selection in the Netherlands. *European Journal of Human Genetics* 2013; 21:1277-1285.

Abstract

Genetic variation in a population can be summarized through principal component analysis (PCA) on genome-wide data. PCs derived from such analyses are valuable for genetic association studies, where they can correct for population stratification. We investigated how to capture the genetic population structure in a well-characterized sample from the Netherlands and in a world-wide dataset and examined whether 1) removing long-range linkage disequilibrium (LD) regions and LD-based SNP pruning significantly improves correlations between PCs and geography and 2) whether genetic differentiation may have been influenced by migration and/or selection.

In the Netherlands, three PCs showed significant correlations with geography, distinguishing between: 1) North and South; 2) East and West; and 3) the middle-band and the rest of the country. The third PC only emerged with minimized LD, which also significantly increased correlations with geography for the other two PCs.

In addition to geography, the Dutch North-South PC showed correlations with genome-wide homozygosity (r=.245), which is in line with a serial founder effect due to northwards migration, and also with height (♂: r=.142, ♀: r=.153).

The divergence between subpopulations identified by PCs is partly driven by selection pressures. The first three PCs showed significant signals for diversifying selection (545 SNPs - the majority within 184 genes). The strongest signal was observed between North and South for the functional SNP in HERC2 that determines human blue/brown eye color.

Thus, this study demonstrates how to increase ancestry signals in a relatively homogeneous population and how those signals can reveal evolutionary history.

Introduction

Population genetic studies are of great value for detecting population substructure and making inferences about human history regarding migrations, expansions, and human evolution.[1] The genetic variation in a population can be summarized by uncorrelated principal components (PCs) through principal component analysis (PCA) on genome-wide data, usually with the explained variance monotonically decreasing with each PC. The PCs explaining most variation often show striking correlations with geography,[2-4] a consequence of the decreasing genetic similarity as geographic distance increases. Such PCs are also of value in genetic association studies, where they are used to correct for allele frequency differences due to systematic ancestry differences, i.e., population stratification.[5]

When analyzing genome-wide genetic variants, one has to consider that some regions of the genome may be overrepresented in the PCs due to elevated levels of linkage disequilibrium (LD), diluting the genome-wide patterns that reflect ancestry differences. Very strong and/or long-range LD at a particular locus can even result in PCs that only reflect genetic variation in that specific region.[6;7] Price et al therefore recommended to exclude long-range LD regions for PCAs,[7] but advised against pruning for LD, since that did not significantly affect PCs in HapMap populations.[6] We hypothesize that these LD artifacts may have larger confounding effects when carrying out a PCA in a single relatively small population, where ancestry differences are relatively small, than in a PCA that is run on a pooled dataset of multiple populations with greater between population differences. To test this hypothesis, we ran PCAs on different SNP sets with varying levels of LD on a large sample of Dutch individuals, and separately in a pooled dataset consisting of the populations from the 1000 Genomes Project[8] covering five different continents (Europe, Africa, Asia, North-, and South-America). Correlations between PCs and geography should be a good proxy for how well the PCs reflect ancestry differences. Correction for stratified phenotypes in association studies, such as height/stature, should be more effective in reducing false positives when using PCs that are a better reflection of one's ancestry.

The PCs showing the strongest ancestry signals are then used to further study the population substructure and genetic history of the Netherlands. The demographic history of this population is complex and still not completely understood. This is partly due to the highly variable Dutch geographic landscape. Large parts of the Netherlands were (and still are) well below sea level and our current landscape (apart from the urbanized areas) resembles the one of approximately 1500 AD. Before that time, large parts of The Netherlands were still covered by sea (either permanently or under strong tidal influence) and uninhabitable, hence

population sizes were probably very low. The arrival of the first Romans (56 BC) also marked the beginning of waves of immigrants from various parts of Europe. Among the first were Batavians (from Germany), during later centuries followed by large groups of economic immigrants and religious refugees from throughout Europe (mainly Iberian, French, Belgian, German, British, and Scandinavian).[9] It is estimated that the ancestors of ~75% of what we currently call the "native" Dutch population (*autochtonen* in Dutch) have immigrated into the Netherlands during the past 20 centuries.[9] Further genetic differentiation within Dutch subpopulations may have been induced by isolation due to geographic and/or social factors. Studies on marriage records of the 19th and early 20th century showed greater isolation within Southern provinces, and the North-West province of Friesland, while the urbanized West showed lower rates of homogamy.[10-12] Religion has also played a considerable role in maintaining Dutch (sub)populations during almost the entire second half of the last millennium, separating the Catholic South from the mostly Protestant North, but also maintaining substructures among the highly segregated Protestant groups. Strong religious assortment was detectable until well into the 20th century.[13] With increasing secularization during the 1960s and 1970s however, religious assortment started to decline.[14]

The first goal of this study is to explore the ability of a PCA to capture population differentiation in a relatively homogeneous population using different SNP sets varying in LD. PCs that represent ancestry differences then are to be employed to aid in investigating patterns of past human migration and the impact of selection on genetic variation in the geographically relatively small area of the Netherlands (41,543 km^2; 16,039 sq mi). Migration patterns were previously detected through correlations between distance from Addis Abbaba, Ethiopia, and genome-wide heterozygosity and LD.[15-18] To investigate the influence of adaptive selection pressures on the genetic differentiation within the Netherlands, the distribution of alleles will be compared between subpopulations identified by the PCs.

Methods

Participants

Subjects were registered at the Netherlands Twin Register (NTR, N=5,509)[19] or the Netherlands Study of Depression and Anxiety (NESDA, N=2,038).[20] Genotyping was performed on the Affymetrix Human Genome-Wide SNP 6.0 Array according to the manufacturer's protocol.

Individuals with possible non-Dutch or non-European ancestry (N=258) were identified by projecting PCs from the 1000 Genomes individuals on the Dutch individuals, and with additional help of the birth country of their parents

(see Appendix).

Only unrelated individuals were analyzed. Unrelated individuals were chosen using GCTA,[21] by excluding one of each pair of individuals with an estimated genetic relationship of >0.025 (i.e., more related than third or fourth cousin), reducing the sample from 7 547 to 4 441 subjects.

The current living address was available for 4,103 unrelated subjects (of which 1,841 also had place of birth available). Adult height was available for 3,714 unrelated subjects, self-reported eye color for 1,581 unrelated subjects (coded as blue, intermediate or brown), and self-reported hair color for 1,583 unrelated subjects (coded as blond, red, light brown, dark brown, or black).

PCA on three SNP sets

Three different SNP sets were created to run the PCAs on, varying in the amount of LD allowed: Panel 1: all SNPs that passed QC (499,849 SNPs); Panel 2: excluding 24 long-range LD regions identified by Price et al[7] (487,672 SNPs); and Panel 3: an LD-pruned SNP set without long range LD regions, where SNPs were pruned recursively in a sliding window (window size = 50, number of SNPs to shift after each step = 5) based on a variance inflation factor (VIF) of 2 (130,248 SNPs). See Appendix for details on QC. PCAs were run with the EIGENSOFT package[6] to compute 10 PCs for each of the three LD varying SNP sets using its default parameters.

Effects of LD on PCA

To determine which SNPs underlie the variation reflected by the PCs, δ (absolute allele frequency difference) was calculated for all SNPs between individuals with the highest and individuals with the lowest PC values (top and bottom 1000 for the Dutch dataset; top and bottom 250 for 1000 Genomes). To investigate the amount of LD that influenced a PC, an LD matrix of its top 500 SNPs (determined by δ) was calculated in Plink, after which all LD values (r^2) were averaged (Table 1).

To test whether LD influences correlations of PCs with geography, we compared the correlations of PCs from the three panels with the latitude and longitude coordinates with the R package *psych*, which allows testing the difference between two dependent correlations sharing one variable (the geographic location in this case).[22;23]

Traces of migration: F, haplotype block size, and F_{st}

F (genome-wide homozygosity) was calculated in Plink.[24] Haplotype blocks were calculated per chromosome in Plink for different groups of individuals. This was done with pair-wise LD calculations for SNPs within 4000 kb (the size of the largest long-range LD region: the chromosome 8p23.1 inversion), using the

largest SNP set (499,849 SNPs). The sizes of all autosomal haplotype blocks were then averaged. F_{st} was calculated as a measure for genetic differentiation between populations according to Weir and Cockerham[25] by calculating it for every SNP and then averaging all F_{st} values to obtain a genome-wide point estimate of the genetic distance.

Selection pressures as a source of genetic differentiation

Selection pressures were identified in Bayescan 2.1.[26] A comparison of several algorithms designed to achieve this goal through F_{st} outlier tests concluded that this software package had the lowest false negative and false positive rates.[27] After computing F_{st} values for all 499,849 SNPs between the top 1000 and bottom 1000 individuals for 3 PCs reflecting ancestry, F_{st} coefficients are decomposed into a population-specific component (β), shared by all loci, and a locus-specific component (α), shared by both populations. If α differs significantly from 0, it is assumed that the locus was under diversifying ($\alpha > 0$) or balancing ($\alpha < 0$) selection. Significance is based on FDR corrected q-values (< .05).

For a more detailed description of the methods, see Appendix.

Results

Increasing genome-wide ancestry signals by reducing LD

PCAs were run on three SNP sets that differed in the amount of LD: Panel 1 (499,849 SNPs), Panel 2 (excluding 24 known long-range LD regions: 487,672 SNPs), and Panel 3 (24 long-range LD regions excluded, and LD pruning: 130,248 SNPs). PCAs were run the 1000 Genomes dataset (N=1,014; no Dutch included), and for a dataset consisting of Dutch individuals only (N=4,441).

For the 1000 Genomes dataset, the three panels had almost identical components (see correlations in Appendix Tables 2 and 3). The only PC reflecting a long-range LD region was PC10 from Panel 1 (the top 449 SNPs based on δ fall within the inversion on chromosome 8p23.1). PCs extracted from the Dutch dataset showed large differences between the three Panels (see Appendix Tables 4 and 5). In Panel 1, all top 10 PCs represent variation in long-range LD regions. For the first 9 PCs, the majority or all of the top 500 SNPs fall in the 24 long-range LD regions, and for PC10 45% of the top 500 SNPs come from one of the 24 long-range LD regions. For Panel 2, the LD levels between the top 500 SNPs of the top ten PCs are slightly lower, but still somewhat in the same range as for Panel 1 (except for PC1, the North-South PC; see Table 1), while in Panel 3 the LD levels are about tenfold lower (Table 1), suggesting that these PCs are likely to represent more genome-wide patterns.

Table 1. The 95% confidence intervals (CI) of the mean r^2 values of the top 500 SNPs (determined by δ) for each PC for each dataset. The LD changes very little across panels for the 1000 Genomes dataset, as opposed to the Dutch datasets.

	1000 Genomes			The Netherlands		
	Panel 1 = All SNPs that passed QC	Panel 2 = Panel 1 without the 24 long-range LD regions	Panel 3 = Panel 2 with genome-wide LD based SNP pruning	Panel 1 = All SNPs that passed QC	Panel 2 = Panel 1 without the 24 long-range LD regions	Panel 3 = Panel 2 with genome-wide LD based SNP pruning
PC1	0.1550-0.1556	0.1534-0.1540	0.1536-0.1542	0.0669-0.0676	0.0044-0.0049 (↕)	0.0037-0.0042 (↕)
PC2	0.1008-0.1012	0.1003-0.1008	0.1004-0.1008	0.0889-0.0900 (↕)	0.0570-0.0586	0.0061-0.0066 (↔)
PC3	0.1639-0.1648	0.1680-0.1688	0.1759-0.1767	0.0930-0.0941	0.0465-0.0479 (↔)	0.0102-0.0109 (↗)
PC4	0.1721-0.1729	0.1776-0.1784	0.1888-0.1896	0.0865-0.0880	0.0516-0.0529	0.0056-0.0061
PC5	0.1699-0.1704	0.1661-0.1666	0.1496-0.1501	0.0995-0.1014	0.0662-0.0678	0.0037-0.0042
PC6	0.2077-0.2086	0.1946-0.1955	0.1764-0.1772	0.0908-0.0926	0.0678-0.0694	0.0036-0.0040
PC7	0.0315-0.0319	0.0310-0.0315	0.0460-0.0465	0.0691-0.0708	0.0547-0.0562	0.0049-0.0054
PC8	0.0343-0.0349	0.0449-0.0454	0.0342-0.0348	0.0521-0.0536 (↔)	0.0574-0.0590	0.0038-0.0042
PC9	0.1011-0.1018	0.0824-0.0830	0.0852-0.0858	0.0626-0.0641	0.0558-0.0573	0.0045-0.0050
PC10	0.0799-0.0806	0.0179-0.0186	0.0141-0.0145	0.0618-0.0633	0.0526-0.0541	0.0040-0.0044

↕: The PC with the highest correlation with the North-South gradient.
↔: The PC with the highest correlation with the East-West gradient.
↗: The PC that also showed a correlation with the East-West gradient, and separates individuals from the middle of the Netherlands from individuals from the rest of the country (illustrated in Figure 1d).

As long as current and past migration rates are not too high, correlations with geography should be a good proxy for how well the PCs reflect ancestry. For a subset of the current sample with place of birth as well as current living address available (N=1 841), the mean distance between birthplace and current living address is 33.33 km (20.71 mi; see Appendix Figure 1). To test whether the degree of LD influences the correlations of the PCs with geography significantly, correlations between PCs from the three SNP panels and the North-South/East-West gradient were compared. The results of these tests are shown in Table 2 (per SNP panel, only the correlation of the PC with the highest geographic correlation is shown and tested). The correlations with geography are significantly improved for Panel 2, and again after additional LD based SNP pruning in Panel 3. Panel 3 is the only Panel where the North-South and East-West PCs show up as the first

two PCs respectively. Panel 3 also shows an additional PC (PC3) with a significant correlation of .162 ($p < .001$) with the East-West gradient (see Figure 1d).

Table 2. Comparison of correlations with geography and λ's in GWASs for height (N=3 714) between PCs from the three SNP panels varying in LD.

Panel used for PCA	Nr. of SNPs for PCA	Correlations between PCs and North-South gradient (N = 4,103)		Correlations between PCs and East-West gradient (N = 4,103)		λ for GWAS on height including the North-South PC as a covariate
		Pearson Correlation	Difference test	Pearson Correlation	Difference test	
Panel 1 = All SNPs that passed QC	499 849	$r_{PC2,\updownarrow}$= .441	-	$r_{PC8,\leftrightarrow}$= .219	-	1.03937
Panel 2 = Panel 1 without the 24 long-range LD regions	487 672	$r_{PC1,\updownarrow}$= .589	$p = 2.0\times10^{-59}$ (versus Panel 1)	$r_{PC3,\leftrightarrow}$= .270	$p = 7.4\times10^{-11}$ (versus Panel 1)	1.03092
Panel 3 = Panel 2 with genome-wide LD based SNP pruning	130 248	$r_{PC1,\updownarrow}$=.603	$p = 2.8\times10^{-5}$ (versus Panel 2)	$r_{PC2,\leftrightarrow}$=.378	$p = 6.6\times10^{-26}$ (versus Panel 2)	1.02961

Population stratification of height

Height has been known as a stratifying variable across the world, even within relatively small areas, such as the Netherlands. Northern Dutch are taller on average than the Dutch from the Southern parts of the Netherlands.[28] Also within Europe, height correlates with its North-South axis, with Northern Europeans being taller than Southern Europeans.[29;30] In our sample however, height does not correlate very high with the North-South gradient of the current living address (males: $r = .036$, $p = .232$; females: $r = .050$, $p = .020$). The North-South PC, however, shows a higher and more significant correlation with height in both sexes (for Panel 3, the correlations are .142 for males and .153 for females, p's $< .001$). The fact that this PC does a better job of capturing the height differences between the subpopulations than their current living address, confirms that the PC is a better measure for ancestral origin than the geographical location, and that these height differences are indeed genetic.

As an illustration of the ability of the PCs to reflect the stratifying effects in the population, and the role of LD thereon, we conducted genome-wide association analyses for height in 3 714 unrelated individuals, using the North-South PC from different SNP Panels to correct for population stratification within the Netherlands. When using only sex as a covariate in the GWAS on height, the lambda (λ) is 1.0543. Introducing the North-South PC from Panel 1, λ decreases to 1.0394, and continues to decline to 1.0296 as a result of excluding long-range LD regions and LD-based SNP pruning (see Table 2).

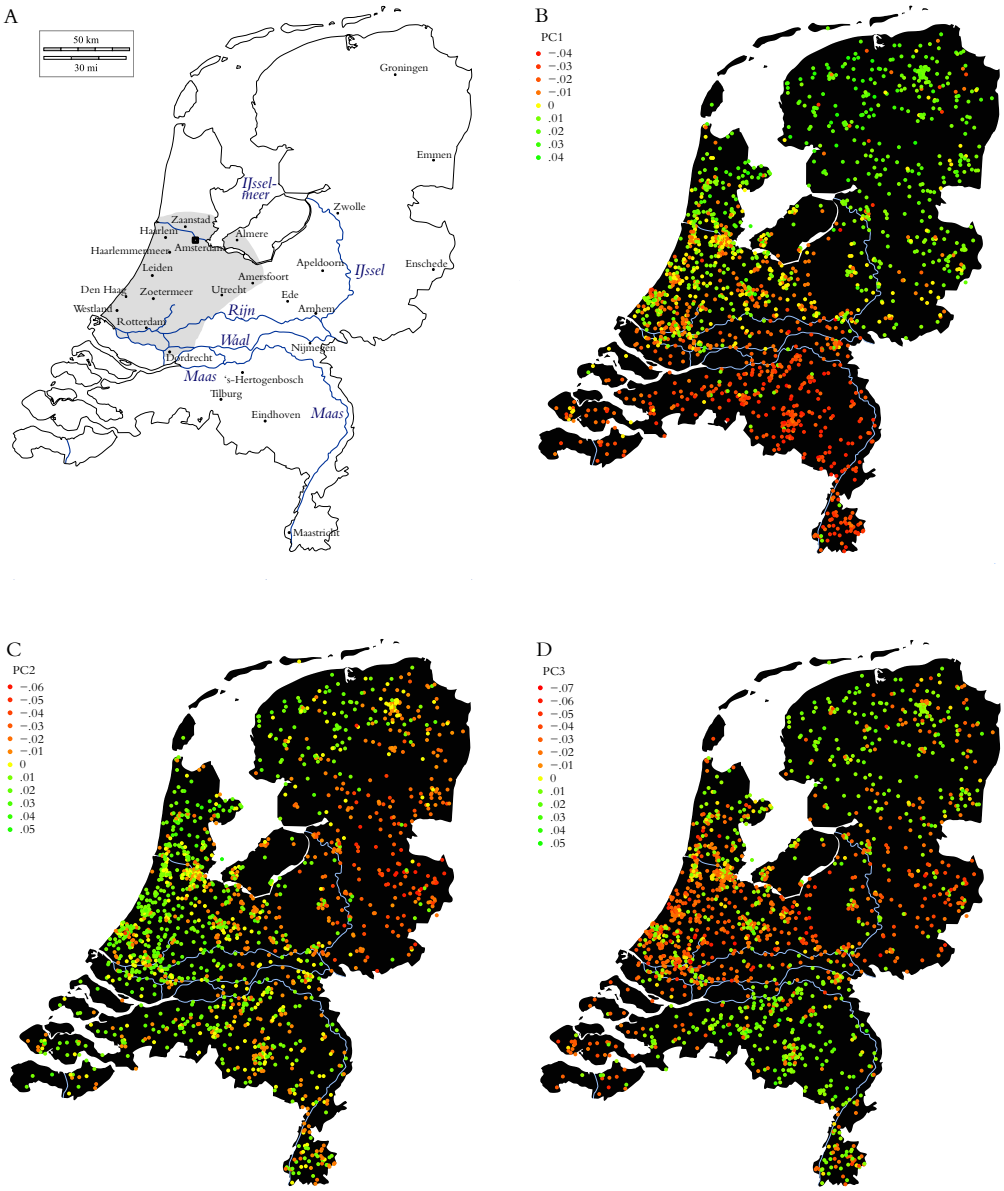


Figure 1. The three PCs showing significant correlations with geography. ***A***: Map of the Netherlands, its major rivers, and its 26 largest municipalities (population size >100k as of April 2012). The grey area represents the highly urbanized Randstad area. ***B***, ***C***, & ***D***: The colors of the points indicate the mean value per postal code of PC1, PC2 and PC3 respectively from the LD pruned SNP set without long-range LD regions (Panel 3). The plot is based on 4,130 unrelated Dutch individuals spread out across 1,635 postal codes of their current living address.

The first three PCs from Panel 3

The first three PCs of Panel 3 have the highest correlations with geography (the only ones with $p < .001$), and the eigenvalues remain relatively constant in subsequent PCs (see Appendix Figure 3). The following analyses will focus on these three PCs, which are plotted in Figure 1 and Appendix Figures 4 and 5. The first PC, which we shall refer to as the North-South PC (Figure 1b), roughly differentiates the Southern provinces (Zeeland, Noord-Brabant, and Limburg) below the three major rivers (the Maas, the Waal, and the Rhine) from the Northern provinces, with the more urbanized West falling in between. The second PC, which we shall call the East-West PC (Figure 1c), mainly differentiates the North-Eastern part of the Netherlands from the rest. The third PC, which we will call the middle-band PC (Figure 1d), separates the Northern and Southern provinces from the middle band area of the Netherlands. There were 157 complete spouse pairs in the sample, for which we calculated the spouse correlations for each of the three PCs. The North-South PC has the highest and most significant spouse correlation ($r = .555$, $p < .001$). The East-West PC and the middle-band PC show nominally significant spouse correlations ($r = .164$, $p = .040$, and $r = .179$, $p = .025$ respectively).

Traces of migration in the Netherlands

The patterns of the first three PCs from Panel 3 resemble the expected patterns of the first three PCs of Novembre and Stephens[31] (see Figure 1). Novembre and Stephens caution against drawing conclusions on migration events based on these patterns, because they resemble mathematical artifacts that may arise when PCA is conducted on spatial data where (genetic) similarity decreases with distance. The North-South PC, however, showed a moderate but significant correlation with F (inbreeding coefficient, a measure for genome-wide homozygosity) of .245 ($p < .001$), indicating that the southern people are more heterozygous than the northern individuals (PC2 and PC3 did not show significant correlations with F). It was previously observed across populations that heterozygosity is negatively correlated with the distance from Addis Ababa, Ethiopia.[16; 17] This phenomenon is consistent with a serial founder effect, where populations expanded through successive migrations of smaller subsets of the populations out of the previous location, starting from a single origin in sub-Saharan Africa. This serial founder effect also results in increased LD with increasing distance from Africa.[15]

Especially for PC1, the highly urbanized Randstad area shows an excess of intermediate PC values (see Figure 1a and 1b), which could be due to the admixture of Dutch subpopulations caused by high migration rates between rural areas and the urbanized West, as well as between the major cities in the West.[32; 33] To investigate whether this could have influenced the correlation between PC1 and F, correlations with the North-South gradient and with F were calculated for PC1 for

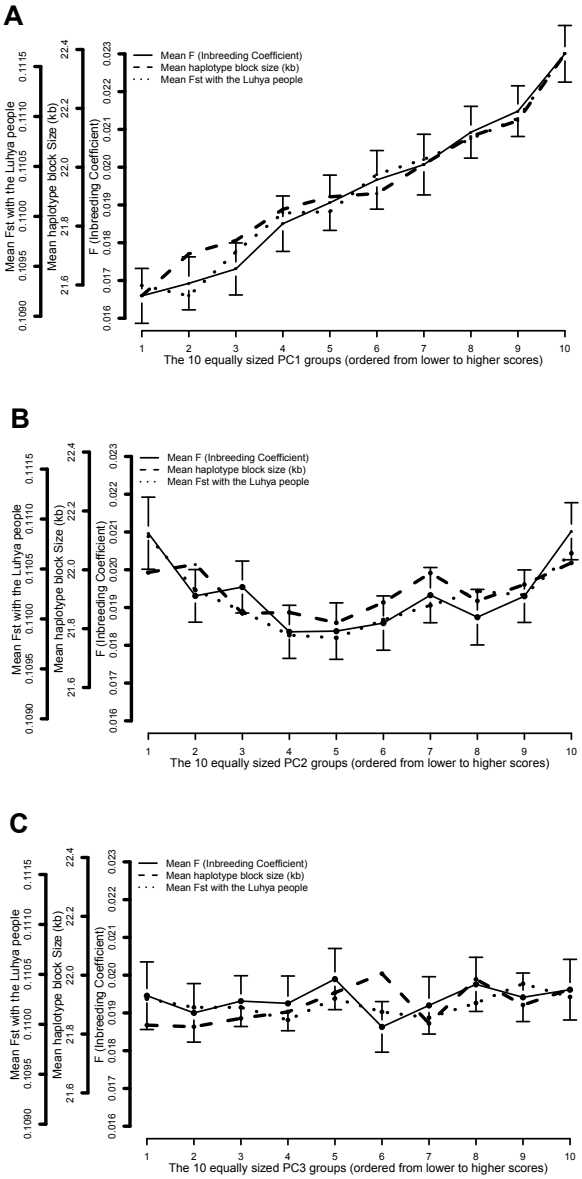


Figure 2. F, haplotype block size and F_{st} with the Luhya people in relation to the PCs. For the first three PCs (Figures ***A***, ***B*** & ***C*** respectively), the Dutch subjects were ordered in an ascending order according to their PC value (= PCs from Panel 3, the LD pruned dataset without long-range LD regions) and divided into ten equally sized groups (the first 9 groups with N=444, and group 10 with N=445). For each of the ten groups the mean F (Inbreeding Coefficient) is calculated and plotted with its 95% confidence interval. As an illustration two related measures, the genome-wide average haplotype block size and the mean Fst with the Luhya people from 1000 Genomes, are plotted as well.

individuals from the 13 largest municipalities of Randstad area (i.e., with population >100k), and individuals from the rest of the Netherlands separately (see paragraph *The Randstad* in the Appendix, and Appendix Tables 7 and 8). For individuals from the Randstad, the correlation with the North-South gradient is not significant and drops to around zero, while it increases for the rest of the country to .669. The correlation with F however is lower (r = .170), but still very significant within the Randstad as well as for the rest of the country, where the correlation increases (r = .259). This indicates that the correlation between PC1 and homozygosity observed in the entire sample is not due to local admixture or inbreeding, making the serial-founder effect hypothesis a more plausible explanation.

To further illustrate the relationship between the PCs and F, the Dutch subjects were ordered in an ascending order according to their PC value and divided into ten equally sized groups (9 groups with N=444, and 1 group with N=445). For each group the mean F was calculated, and plotted in Figure 2 with its 95% confidence interval. As an illustration, we calculated and plotted two related measures in the same Figure: genome-wide average haplotype block size, and the mean F_{st} with the Luhya people from 1000 Genomes (the 1000 Genomes population closest to Ethiopia). Figure 2 shows that all three measures show a similar linear increase as the North-South PC score increases, suggesting northwards migration. For PC2, the East-West PC, the homozygosity increases as one moves towards more positive as well as more negative values. This may be due to migration in multiple directions, but alternative explanations for this observation are also possible, such as local admixture and/or inbreeding. PC3 did not show significant differences between its 10 groups.

Selection pressures as a source of genetic differentiation

To investigate the extent of adaptive effects on the genetic differentiation within the Netherlands, F_{st} values were computed with Bayescan 2.1[26] for all 499 849 SNPs that passed QC. F_{st} values were computed between the top and bottom 1000 individuals for each PC depicted in Figure 1 (genome-wide mean F_{st}'s: PC1=.00059, PC2=.00026, PC3=.00021). Bayescan 2.1 then detected outliers with respect to F_{st} values using a Bayesian approach, allowing a distinction between divergence due to random drift and divergence that is more likely to be driven by selection pressures. After FDR correction, 273 SNPs reached significance for PC1, 172 SNPs were significant for PC2, and 100 SNPs for PC3 ($q < .05$). All significant signals were in the direction of diversifying selection, which may be partly explained by the weak power to detect balancing selection in F_{st} outlier approaches.[26; 27; 34] 58.6% of the significant SNPs for PC1 fell within 88 genes, for PC2 62.2% fell within 55 genes, and for PC3 75% fell within 41 genes (as opposed to 51.4% of all 499 849 SNPs). These elevated proportions of genic SNPs among the outliers suggest that selection pressures on functional genetic variants

played a role in the genetic differentiation between these Dutch subpopulations. Some of these genes have also been observed as highly differentiated within Europe, such as *LCT* (PC1), *HERC2* (PC1), *CADPS* (PC1), *IRF1* (PC1), *SLC44A5* (PC1), *R3HDM1* (PC1), *ACOXL* (PC3), and *BTBD9* (PC3).[35-39]

The SNP with the highest F_{st} was observed in the North-South PC (PC1), falls within the *HERC2* gene (rs8039195, $F_{st} = .0061$, $q = 0$), and has been strongly associated with hair- and eye color.[40-42] In the current Dutch dataset this SNP was also highly predictive for both eye color and hair color when analyzed in a linear regression (eye color: $p = 3.59 \times 10^{-133}$; hair color: $p = 1.65 \times 10^{-22}$). Since eye color and hair color are associated, we conducted an additional linear regression for rs8039195 on eye color with hair color as a covariate, and for hair color with eye color as a covariate. With covariates, the association was still highly significant for eye color ($p = 7.8 \times 10^{-112}$), but not for hair color ($p = .218$). The genotype frequencies of this SNP are also highly differentiated between 1000 Genomes populations, with the TT genotype having lower frequencies in populations with predominantly brown eyes, while in Northern European populations the genotype frequency can be as high as 93.5% in the Finnish, where blue eyes are much more prevalent (see Appendix Table 6). To get a higher resolution of the F_{st} values within and around the *HERC2* gene, F_{st} values were calculated for 3 495 SNPs (chr15: 28,300,000 bp - 28,600,000 bp) between the available 1000 Genomes Northern European populations (the British and Finnish) and Southern European populations (the Iberian and Toscan). Of the SNPs that were genotyped in the Dutch sample, rs8039195 had the highest F_{st} between these European populations. The highest F_{st} value of all 3 495 available 1000 Genomes SNPs was observed for rs12913832, identified recently as the functional SNP for determining human blue-brown eye color.[43; 44] This SNP is in high LD with rs8039195 ($r^2 = .394$, $D' = .993$), consistent with rs12913832 being responsible for the significant signal for diversifying selection in this population.

In addition, significant divergence was observed in all three PCs for a number of genes that play major roles in brain function, such as *GRM7* (PC1; encodes a metabotropic glutamate receptor), *GRIN2A* (PC1; encodes a subunit for the NMDA receptor), *BDNF* (PC2; encodes the brain-derived neurotrophic factor), SLC6A4 (PC3; encodes the serotonin transporter), *NRXN3* (PC3; encodes neurexin-3-alpha), *AUTS2* (PC3; autism susceptibility candidate 2). When including genes from all three PCs that showed signals of selection pressures in the clustering algorithm of Ingenuity Pathway Analysis (Ingenuity Systems, IPA spring release 2012), the top 11 biological functions are brain related ($p \leq 1.26 \times 10^{-4}$; with a large degree of overlap in molecules between the functions), with the most significant being *neurotransmission of nervous tissue* with 11 molecules and $p = 2.2 \times 10^{-6}$. To ensure this result is not due to a sampling bias (part of the sample consists of major depressive disorder [MDD] cases and controls),

the Bayescan analysis was repeated, this time comparing MDD cases (N=966) with MDD controls (N=1 522). This analysis showed no significant signals.

Other notable genes showing significant signals include *FTO* (PC1) and *HCP5* (HLA Complex P5 gene; PC1 & PC2). A full list of significant SNPs, Bayescan statistics, and the genes they fall in can be found in the supplementary excel file that is available at the online Supplementary Materials in the published version of this article.

Discussion

In an effort to elucidate the genetic substructure in a well characterized population that contributes to multiple GWAS efforts, PCAs were conducted followed by a variety of follow-up analyses. The main aims of this study were: (1) determine which of the SNP sets (varying in the amount of LD) led to the best PCs in terms of reflecting ancestral origin, (2) using these PCs to investigate patterns of past human migration, and (3) identifying genomic regions under selection pressures.

We first examined the effect of reducing LD on the ability of the PCs to capture the genome-wide patterns reflecting ancestry differences. In SNP panel 1, the PCA on the 1000 Genomes populations resulted in only 1 of the top 10 PCs (PC10) reflecting a long-range LD region, while in the Dutch dataset, all top 10 PCs reflect these regions. Price et al[6] showed that genome-wide LD-based SNP pruning did not lead to improved PCs in an analysis of HapMap data. This was confirmed in our analysis of the 1000 Genomes dataset. In the Dutch dataset however, LD-based SNP pruning did lead to improved PCs, as shown by: (1) a large decrease in LD in the top 500 SNPs of the top 10 PCs (Table 1); (2) significantly improved correlations of the PCs with geography (Table 2); (3) the emergence of a new PC among the top ten PCs (PC3, the middle-band PC) that correlates significantly with geography (Figure 1d). We thus conclude that both excluding long-range LD regions and LD pruning are necessary when studying a relatively confined population, which may consist of overlapping subpopulations. Large GWAS efforts usually consist of meta-analyses of multiple cohorts consisting of relatively homogeneous populations, which often use PCs to account for population stratification. PCs extracted from SNP sets with less LD are better suited for this goal, as we show using height as an example.

The Dutch North-South component had the highest correlation with geography, and showed the strongest levels of differentiation based on genome-wide F_{st} values. The high spouse correlation for this PC (.555) suggests that the North-South differentiation is at least to some extent still ongoing. In Europe, there

also is a consistent and reproducible distinction between Northern and Southern populations.[45; 46] When projected onto the Dutch individuals, the 1000 Genomes PC that differentiates between Northern and Southern European populations (1000 Genomes PC4 in Appendix Figure 2) shows a high and significant correlation with the Dutch North-South PC ($r = .656$, $p < .001$, in unrelated Dutch individuals). The correlation with height is also in the same direction as in Europe (i.e., Northerners are taller than Southerners on average), and blue/brown eye color as well. The Dutch North-South PC also shows a decrease in heterozygosity and an increase in mean haplotype block size in Northern as compared to Southern Dutch individuals (Figure 2), which has been observed between Northern and Southern European populations as well,[45] and is best explained by a serial founder effect. This effect is in line with the European South-North expansions expected to have occurred at least during Paleolithic, Mesolithic and Neolithic times.[47-50] This effect does not necessarily have to reflect an upward migration that took place within the Netherlands; it may also be that, more recently, Southern Europeans migrated more to the South of the Netherlands, while Northern Europeans migrated more to the Northern parts of the country, maintaining the North-South distribution within the country.

It seems that the genetic differentiation between the Dutch subpopulations led to some phenotypic differences as well, as can be seen for example in the significant correlation of height with the North-South PC. The divergence between these subpopulations is at least in part driven by diversifying selection pressures. The majority of SNPs with significant signals of selection pressures are within genes, of which several have been found to strongly differentiate within Europe as well.[35-39]

The highest F_{st} is observed for rs8039195 from the *HERC2* gene. This signal is very likely coming from the neighboring SNP rs12913832 (the strongest blue-brown eye color determinant in humans).[43; 44] It is not entirely clear yet why eye color was under such strong selection pressures. It has been proposed that European eye color may have been under frequency-dependant sexual selection,[51] which is known to favor color polymorphisms and increase their diversity in many species. The strong signal this particular SNP shows is probably due to the large effect this SNP has on the trait under selection, increasing the selective pressure on this single polymorphism.

Genes involved in brain function are significantly overrepresented among the rest of the signals. Selection pressures on brain related genes in modern humans have been reported previously.[52; 53] More research is needed on the exact variants under selection and their functional impact in order to hypothesize which of the wide range of brain functions may have been under selection and why.

Other notable genes include *FTO* (PC1) and *HCP5* (PC1 & PC2). *FTO* plays a role in metabolism, having a large enough effect on obesity to be consistently

associated with it,[54; 55] suggesting dietary influenced selection pressures, such as those expected from the transition from hunter-gatherer to agricultural societies. Selection pressures on *FTO* and other genes involved in obesity have been observed before in other populations.[56; 57] The lactase gene (LCT) is also a well-established target of dietary influenced selection pressures, and also showed significant signals in PC1. *HCP5* (HLA Complex P5 gene) from the MHC region (a long-range LD region that was excluded in the PCA) is one of two genes that appear in multiple PCs (PC1 & PC2), and plays a role in the immune system. Strong divergence of several genes from the HLA complex has been observed within Europe,[36-38] most likely due to high evolution rates in the highly polymorphic MHC region in order to maintain resistance to rapidly evolving pathogens.[58; 59] Other immunity-related genes that showed significant signals of selection in this study as well as previous studies are: *IRF1* (PC1), *ACE* (PC1), *LRRC4C* (PC2), *PLCL1* (PC3), and *HSPD1* (PC3).[60]

In interpreting these findings, one should consider the possibility of ascertainment bias in SNP selection for the microarray, which may have caused signals to be missed (especially for analyses on selection pressures). SNP selection of about half of the SNPs on this Affymetrix array however is random in order to provide sufficient genome-wide coverage, which may have decreased this bias.[61]

This is a unique population genetics study in terms of resolution, because of the large sample from a relatively small geographical area with detailed phenotypic information available for the majority of the subjects. Increasing signals for ancestry in this dataset allowed for the investigation of traces left by migration and adaptation in the genome of a region where the subpopulations have relatively subtle genetic differences. Further research is needed to identify the functional variants in the genomic regions showing significant signals for diversifying selection pressures, as these are likely to influence traits that increased fitness and/or reproductive success. Our results also confirm the importance of considering stratification in association studies of complex traits designed to detect very small effects, even when analyzing smaller supposedly homogeneous populations. In computing PCs to correct for these subtle ancestry differences, the level of LD should be minimized.

Appendix

Participants

Subjects were registered at the Netherlands Twin Register (NTR, N=5,509; 2,226 males and 3,283 females)[19] or the Netherlands Study of Depression and Anxiety (NESDA, N=2,038; 684 males, and 1,354 females).[20] The NTR sample consisted of 830 unrelated individuals, 1 431 families with two members, 372 with 3 members, 111 with four, 49 with five, and 2 families with six members (parents, twins, siblings, spouses of twins). The NESDA sample consisted of unrelated individuals only.

Genotyping was performed on the Affymetrix Human Genome-Wide SNP 6.0 Array at two sites (Avera Institute for Human Genetics [AIHG], South Dakota, USA and the Rutgers University Cell and DNA Repository [RUCDR], New Jersey, USA) according to the manufacturer's protocol. Methods for blood and buccal swab collection, genomic DNA extraction, and genotyping have been described previously.[62; 63]

The birth country of the parents was available for the majority of the subjects (N=4,485) as well as their current living addresses ($N_{relateds}$=7,092, $N_{unrelateds}$=4,103). For the current living addresses, the postal codes were translated into geographic coordinates (longitude and latitude) for each participant using the open source 6PP database,[64] in order to compute correlations between the PCs and North-South/East-West gradients. These coordinates were also used to plot the subjects on the map of the Netherlands in Figure 1. Place of birth (city or municipality of birth) was available for 1,841 subjects who also had current living address available. These were also translated into geographic coordinates using the open source 6PP database (these coordinates are less accurate however than those obtained from postal codes). Adult height (stature; age ≥ 18 years old) was available for the majority of the subjects ($N_{relateds}$=5,914, $N_{unrelateds}$=3,714). Self-reported eye color was available for 3,375 subjects (1,581 unrelated, coded as blue, intermediate or brown). Self-reported hair color was available for 3,380 subjects (1,583 unrelated, coded as blond, red, light brown, dark brown, or black). The numbers reported here (and in the rest of the manuscript) are excluding 659 individuals that were removed due to a batch effect (see Appendix: *Removing a Batch Effect*).

The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, and an Institutional Review Board certified by the US Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180). All subjects provided written informed consent.

Quality Control

Autosomal SNPs were analyzed. Quality control (QC) was conducted in Plink,[24] by removing all SNPs with a minor allele frequency (MAF) smaller than 5%, missing rate greater than 5%, a Hardy-Weinberg equilibrium (HWE) deviation with a *p*-value smaller than 0.001. SNPs were also removed if there were alleles that were incongruent between datasets, which could also represent an allele flip (this filter was applied twice: when merging the dataset from AIHG with the dataset from RUCDR, and when merging the merged Dutch dataset with a total of 1 094 subjects from the 1000 Genomes dataset). Individuals were removed if they had a missing rate greater than 5%, or excess genome-wide heterozygosity / inbreeding levels (F, as calculated in Plink on an LD-pruned set, must be greater than -0.10 and smaller than 0.10). Only SNPs that passed QC in the genotyped Dutch dataset were analyzed for 1000 Genomes samples (June 2011 release).[65]

Identifying individuals with non-European/non-Dutch ancestry

The 1000 Genomes dataset was used as a reference to aid in identifying and excluding individuals with a non-Dutch ancestry (see Appendix Figure 2). The 1000 Genomes PCs that were not pruned for LD and did not contain long-range LD regions were used for this goal (SNP Panel 2, in line with the suggestions from Price et al).[6; 7] Eight of the top ten 1000 Genomes PCs (all but PC4 and PC7) cluster the European populations together, making them useful for detecting individuals with a non-European ancestry. A Dutch individual was labeled as a potential outlier with a non-European ancestry if one of the 1000 Genomes PCs of that individual was lower than the minimum or higher than the maximum score of that particular PC of the European 1000 Genomes individuals (CEPH, Finnish, British, Iberian, and Toscan). This yielded 151 outliers from PCs 1, 2, 3, and 5. A good (albeit imperfect) indicator of one's ancestry is the country of birth of the parents, which was available for a subset of the Dutch dataset. When comparing the birthplace of the parents between the outliers and the rest of the Dutch sample, the majority of the outliers (57%) had at least one parent born outside of the Netherlands, as opposed to 4.5% of the rest of the sample. This suggests that (the majority of) these individuals are indeed likely to have a non-European ancestry; hence they were excluded. PCs 4 and 7 are the only two PCs that differentiate between European populations. For PC4, the two populations with the lowest F_{st} when compared to the Dutch, the British (F_{st} = .0005) and CEPH (F_{st} =.0002; see Appendix Table 1), are the only European populations that cluster with the Dutch. Of the Dutch individuals that fall outside the British and CEPH cluster (N=129), the majority (53.3%) have at least one parent that is born outside of the Netherlands, suggesting these individuals are also likely to have a non-Dutch ancestry component. These individuals were also excluded. The 1000 Genomes

PC7 shows three clusters that overlap with each other. PC7 separates British and CEPH individuals from Finnish, Toscan, and Puerto Rican individuals, with the rest of the populations falling in between these two clusters. The Dutch individuals also mainly fell in between the two lateral clusters, but showed a large overlap with all three clusters, making it difficult to interpret who are outliers, therefore no individuals were excluded based on this PC. Eventually a total of 258 of the 7,547 individuals were excluded from the PCA on the Dutch sample. Parental birth place information was available for 132 of these individuals, of which 73 (55.3%) had at least one parent born outside of the Netherlands (as opposed to 4% of the rest of the individuals).

Three SNP sets, varying in LD, for PCAs

Three different SNP sets were created to run the PCAs on, varying in the amount of LD allowed: a SNP set including all SNPs that passed QC (499,849 SNPs; Panel 1); a SNP set excluding 24 long-range LD regions identified by Price et al[7] (487,672 SNPs; Panel 2), and an LD-pruned SNP set without long range LD regions, where SNPs were pruned recursively in a sliding window (window size = 50, number of SNPs to shift after each step = 5) based on a variance inflation factor (VIF) of 2, resulting in a set with 130,248 SNPs (Panel 3). VIF = $1/[1-R^2]$, where R^2 is the multiple correlation coefficient for a SNP regressed on all other SNPs within the window simultaneously.[24]

Principal Component Analysis (PCA)

PCAs were run with the EIGENSOFT package[6] to compute 10 PCs for each of the three LD varying SNP sets using its default parameters. The PCA was run on unrelated individuals only, and projected onto the other subjects. Unrelated individuals were chosen using GCTA,[21] by excluding one of each pair of individuals with an estimated genetic relationship of >0.025 (i.e., more related than third or fourth cousin). The genetic relationship matrix was calculated for each population separately. First, PCs extracted from the 1000 Genomes individuals (1014 unrelated individuals, from the SNP set that was not pruned for LD and without long-range LD regions, i.e., in line with the suggestions from Price et al)[6;7] were used to detect individuals with possible non-Dutch or non-European ancestry. After excluding these individuals (N=258), 4,441 unrelated Dutch individuals were extracted with GCTA. PCA was run on these unrelated individuals for each of the three LD varying SNP sets and projected on the rest.

Delta, F_{st}, mean LD and mean haplotype block size

Delta (δ) and F_{st} were calculated using scripts written in Perl. δ is defined as the absolute allele frequency difference between two groups or populations. In order to determine which SNPs underlie the variation reflected by the PCs, δ was

calculated for all SNPs between the individuals with the highest PC values versus the individuals with the lowest PC values (top and bottom 1000 for the Dutch dataset; top and bottom 250 for the 1000 Genomes dataset).

F_{st} was calculated as a measure for genetic differentiation between populations according to Weir and Cockerham[25] by calculating it for every SNP and then averaging all F_{st} values to obtain a genome-wide point estimate of the genetic distance. F_{st} values normally range between 0 and 1. Note that F_{st} according to Weir and Cockerham (Figure 2, Appendix Table 1, and analyses on HERC2 in Northern vs. Southern European 1000 Genomes populations) gives slightly different outcomes than the F_{st}'s calculated by Bayescan 2.1[26] and should not be directly compared. For computational reasons, the latter was used only for the analyses on selection pressures, (i.e., in comparing the top 1000 versus bottom 1000 individuals for 3 PCs, described in the paragraph below).

The mean LD and average haplotype block size were calculated in Plink[24] and additional purpose-written perl scripts. To investigate the amount of LD that influenced a PC, Plink was used to calculate an LD matrix of the top 500 SNPs of the PC (determined by δ), after which all LD values (r^2) were averaged (Table 1). To examine the presence of serial founder effects, haplotype blocks were calculated per chromosome in Plink for different groups of individuals. This was done with pair-wise LD calculations for SNPs within 4000 kb (the size of the largest long-range LD region: the chromosome 8p23.1 inversion between 8 and 12 Mb), using the largest SNP set (499,849 SNPs). The sizes of all autosomal haplotype blocks were then averaged.

Identifying variants under selection

Candidate loci that may have been under selection pressures were identified in Bayescan 2.1.[26] A comparison of several algorithms designed to achieve this goal through F_{st} outlier tests concluded that this software package had the lowest false negative and false positive rates.[27] After computing F_{st} values for all 499,849 SNPs between the top 1000 and bottom 1000 individuals for 3 PCs reflecting ancestry, the F_{st} coefficients are decomposed into a population-specific component (β), shared by all loci, and a locus-specific component (α), shared by both populations. If α differs significantly from 0, it is assumed that the locus was under diversifying ($\alpha > 0$) or balancing/purifying ($\alpha < 0$) selection, although power is usually weak for detecting balancing selection.[26; 27; 34] Significance is based on FDR corrected q-values (< .05). Higher false positive rates may be observed when isolated populations are included that underwent a strong bottleneck. Since the subpopulations in our sample are not geographically isolated, we have no strong reasons to assume strong isolation and/or strong bottlenecks within the Netherlands.

Additional Analyses, Software and Bioinformatics

SPSS and additional perl scripts were used for data management. Graphics were created with R. Plink was used for computing F (inbreeding coefficient/ genome-wide homozygosity, on an LD-pruned SNP set), and for GWASs (linear regressions on unrelated individuals) on adult height (N=3,714), eye color (N=1,581) and hair color (N=1,583). All reported correlations are Pearson correlations computed with SPSS. All base pair positions are in build 37. SNP annotations and genic information about SNPs were extracted from the Ensembl database (Ensembl Genes 67, GRCh37.7).

Ingenuity Pathway Analysis (Ingenuity Systems, IPA spring release 2012), was used to examine whether particular biological functions were overrepresented among the genes showing significant signals for selection pressures. The Ingenuity database contains a large amount of information about structure, biological function, and subcellular localization of the proteins. Only biological relationships that were experimentally observed were considered in the analysis.

The Randstad

The Randstad is a metropolitan region in the Western part of the Netherlands containing >40% of the Dutch population (~7.1 million out of ~16.8 million). This region includes the four largest cities of the Netherlands (Amsterdam, Rotterdam, Den Haag, Utrecht) and surrounding areas (see Figure 1a). The term Randstad did not exist until the second half of the twentieth century, but migration records since 1800 already show considerable migration flows between rural areas and the urbanized West, as well as between the major cities in the West.[32; 33] This may have lead to more admixture between Dutch subpopulations in this region.

From the 4,155 unrelated Dutch individuals with a known current living address, three selections were made: inhabitants of the four largest municipalities with a population size of >300k (N = 624), inhabitants of the thirteen Randstad municipalities with a population size of >100k (N = 1,086), and of the 26 municipalities from the entire country with a population size >100k (N = 1,630) (see Figure 1a for the locations of the 26 largest municipalities and Appendix Table 7 for an overview of their population size in April 2012 according to the Central Bureau of Statistics).[66]

Especially for PC1, the Randstad region seems to show most of the intermediate values at face value in Figure 1a. When only including individuals living in the major municipalities in this region, the correlation between PC1 and the North-South axis is not significant (r = -.010, p = .808 for the four major Randstad municipalities with population size >300k; r = .055, p = .074 for the thirteen major Randstad municipalities with population size >100k; see Appendix Table 8). When excluding these individuals from the entire sample of unrelated

Dutch individuals, the correlation with geography increases considerably for PC1 (r = .648, p < .001 excluding the four major Randstad municipalities; r = .669, p < .001 excluding the 13 major Randstad municipalities; see Appendix Table 8). The correlation between PC1 and genome-wide homozygosity also increases slightly without the major municipalities. As opposed to the correlation with the North-South axis, the correlation with genome-wide homozygosity remains significant for the individuals from the major Randstad municipalities (r = .201, p < .001 excluding the four major Randstad municipalities; .170, p < .001 excluding the 13 major Randstad municipalities). This indicates that the correlation between PC1 and homozygosity observed in the entire sample is not due to local admixture or inbreeding, making the serial-founder effect hypothesis more plausible. When excluding all Dutch municipalities with a population size > 100k, the correlation between PC1 and the North-South axis increases further (r = .678, p < .001), but the correlation for the individuals from these municipalities is also still very significant (r = .451, p < .001).

The correlation of PC2 with the East-West gradient also increases from .378 to .405 as the major municipalities of the Randstad are excluded, and increases further to .439 when all 26 municipalities with a population size > 100k are excluded (p's < .001, see Appendix Table 8). PC2 still shows a significant correlation with the East-West gradient when only considering the 13 municipalities from the Randstad (r = .145, p < .001), and the 26 municipalities from the entire country (r = .281, p < .001). The correlation of PC3 with the East-West gradient shows little change when excluding these municipalities.

Removing a Batch Effect

Genotyping was done in 8,207 individuals, of which 320 individuals were excluded from the initial PCA because of a non-European/non-Dutch ancestry (identified as described under *Identifying individuals with non-European/non-Dutch ancestry*). PCA was run on 125,303 LD based pruned SNPs (parameters as described in Methods) in 4,666 unrelated individuals and projected onto the remainder of the sample. PC1 from this analysis showed a strong correlation with F (r = .755) and the Contrast QC (CQC, a quality metric from Affymetrix representing how well allele intensities separate into clusters; r = .596). The strong correlation was caused by a subset of individuals. We calculated the distance between the mean value of PC1 (-.0022) and the highest observed PC1 value (.0116), and subtracted this value from the mean. All individuals who scored below this value were considered outliers and excluded from subsequent analyses (N=659; see Appendix Figure 6). All numbers reported in the main manuscript and Appendix are excluding these 659 individuals.

Supplementary file

The supplementary excel file (available at the online Supplementary Materials in the published version of this article) contains all SNPs that showed significant signals for diversifying selection ($q < .05$) for all three PCs (note: each PC is on a separate sheet of the file). The following columns are included:

- **chr**: chromosome.
- **bp**: base pair position (according to build 37).
- **SNP_ID**: rs ID of the SNP.
- **fst**: the F_{st} coefficient averaged over populations. In each population F_{st} is calculated as the posterior mean using model averaging.
- **alpha**: the estimated alpha coefficient indicating the strength and direction of selection. A positive value of alpha suggests diversifying selection, whereas negative values suggest balancing or purifying selection.
- **prob**: the posterior probability for the model including selection.
- **log10PO**: the logarithm of Posterior Odds to base 10 for the model including selection. Note that this value is arbitrarily fixed to 1000 when the posterior probability is 1 (should be infinity).
- **qval**: the q-value for the model including selection.
- **Gene_ID**: The name of the gene the SNP falls in.

Appendix Table 1: Population pairwise F_{st} values for the Dutch and 1000 Genomes populations. NLD = Dutch individuals from the NTR and NESDA; ASW = HapMap African ancestry individuals from SW US; CEU = CEPH individuals; CHB = Han Chinese in Beijing; CHS = Han Chinese South; CLM = Colombian in Medellin, Colombia; FIN = HapMap Finnish individuals from Finland; GBR = British individuals from England and Scotland; IBS = Iberian populations in Spain; JPT = Japanese individuals; LWK = Luhya individuals; MXL = HapMap Mexican individuals from LA California; PEL = Peruvian in Lima, Peru; TSI = Toscan individuals; YRI = Yoruba individuals.

	NLD	ASW	CEU	CHB	CHS	CLM	FIN	GBR	IBS	JPT	LWK	MXL	PEL	TSI	YRI
NLD	0														
ASW	.08363	0													
CEU	.00020	.07524	0												
CHB	.08551	.11905	.08450	0											
CHS	.08650	.11916	.08547	.00094	0										
CLM	.01604	.06074	.01370	.06904	.07007	0									
FIN	.00615	.07925	.00589	.07820	.07937	.01691	0								
GBR	.00050	.07552	.00019	.08455	.08554	.01393	.00639	0							
IBS	.00122	.06898	.00199	.10833	.10999	.01062	.01007	.00215	0						
JPT	.08659	.11974	.08574	.00625	.00796	.06978	.07931	.08574	.11083	0					
LWK	.11488	.00951	.10414	.13597	.13658	.09144	.10707	.10443	.11163	.13715	0				
MXL	.03089	.07247	.02820	.06013	.06158	.00705	.02834	.02829	.02618	.06085	.10242	0			
PEL	.01176	.05149	.00985	.07656	.07739	.00548	.01498	.01006	.00618	.07741	.08168	.01531	0		
TSI	.00373	.07336	.00326	.08428	.08527	.01403	.01109	.00350	.00154	.08553	.10106	.02956	.00902	0	
YRI	.12253	.01051	.11282	.14222	.14321	.10133	.11524	.11311	.12653	.14375	.00678	.11224	.09134	.10960	0

Appendix Table 2: The correlations between the SNP set including all SNPs that passed QC (Panel 1), and the SNP set excluding the 24 long-range LD regions (Panel 2), for the 1000 Genomes dataset.

Pearson Correlations		1000 Genomes PCs: Panel 1 (all SNPs)									
		PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
1000 Genomes PCs: Panel 2 (all SNPs excluding 24 long-range LD regions)	**PC1**	**1.000**	.020	.031	.006	.017	.059	.001	.001	-.006	.000
	PC2	.018	**1.000**	.003	-.006	-.014	.026	-.002	.002	.002	-.003
	PC3	.032	.002	**1.000**	.059	.014	.001	-.025	.023	-.049	.001
	PC4	.007	-.006	.056	**1.000**	.001	.003	-.009	.002	-.017	.006
	PC5	.018	-.014	.007	.003	**.999**	.012	.001	.003	.000	.004
	PC6	.059	.026	.001	.001	-.005	**1.000**	.003	-.002	.000	-.006
	PC7	.001	-.002	-.021	-.010	-.005	.001	**.997**	-.011	.010	.024
	PC8	.002	.002	.021	.003	.000	-.001	.009	**.997**	.039	.013
	PC9	-.006	.002	-.048	-.015	-.003	.002	.003	-.045	**.991**	-.070
	PC10	-.001	.002	.002	.000	.000	.001	-.001	.003	.011	.071

Appendix Table 3: The correlations between the SNP set excluding the 24 long-range LD regions (Panel 2), and the LD pruned SNP set excluding the 24 long-range LD regions (Panel 3) for the 1000 Genomes dataset.

Pearson Correlations		1000 Genomes PCs: Panel 2 (all SNPs, no long-range LD)									
		PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
1000 Genomes PCs: Panel 3 (LD pruned SNPs, no long-range LD)	**PC1**	**.998**	-.043	.034	.007	.019	.057	.000	.001	-.005	-.001
	PC2	.084	**.998**	.009	-.003	-.011	.030	-.005	.001	.002	.002
	PC3	.029	-.001	**.999**	.050	.013	-.001	-.023	.021	-.047	.002
	PC4	-.008	.008	-.066	**-.997**	-.002	-.010	.004	-.004	.023	.005
	PC5	-.020	.013	-.009	-.002	**-.993**	-.060	.008	.004	.002	-.010
	PC6	.061	.026	.003	-.005	-.049	**.993**	-.007	-.006	.003	-.001
	PC7	.002	.002	-.022	-.013	.004	.012	**.983**	.017	.011	.006
	PC8	-.003	-.003	-.022	-.002	-.007	-.005	.017	**-.970**	-.019	-.011
	PC9	-.006	.003	-.046	-.009	-.003	-.001	.004	-.023	**.957**	.021
	PC10	.001	.001	.003	.005	-.011	.001	-.005	.005	-.039	**.704**

Appendix Table 4: The correlations between the SNP set including all SNPs that passed QC (Panel 1), and the SNP set excluding the 24 long-range LD regions (Panel 2), for the Dutch dataset.

Pearson Correlations		Dutch PCs: Panel 1 (all SNPs)									
		PC1	PC2 ↕	PC3	PC4	PC5	PC6	PC7	PC8 ↔	PC9	PC10
Dutch PCs: Panel 2 (all SNPs, no long-range LD)	PC1 ↕	-.111	**-.750**	**.634**	-.069	.054	-.043	.009	.002	.034	.019
	PC2	-.031	.001	.002	-.091	**-.726**	**-.638**	-.187	.052	.024	-.023
	PC3 ↔	-.013	-.019	-.030	.093	.081	-.102	-.262	**-.865**	.157	.114
	PC4	-.051	-.022	-.044	-.061	-.049	.066	.073	-.066	.241	.026
	PC5	-.001	-.020	-.027	-.086	.036	.000	-.089	-.024	.102	**-.758**
	PC6	-.005	-.005	.000	.031	.018	-.026	-.048	.053	-.166	.037
	PC7	.020	-.008	-.004	-.050	.005	-.011	-.041	-.037	-.260	-.010
	PC8	-.016	.013	.009	.077	.035	-.052	.076	.072	.034	.089
	PC9	.005	.035	.043	.024	.015	-.006	-.007	-.001	.248	-.061
	PC10	.004	.030	.042	.025	.010	-.019	-.045	-.001	-.060	.060

↕: The PC with the highest correlation with the North-South gradient.
↔: The PC with the highest correlation with the East-West gradient.

Appendix Table 5: The correlations between the SNP set excluding the 24 long-range LD regions (Panel 2), and the LD pruned SNP set excluding the 24 long-range LD regions (Panel 3), for the Dutch dataset.

Pearson Correlations		Dutch PCs: Panel 2 (no long-range LD)									
		PC1 ↕	PC2	PC3 ↔	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Dutch PCs: Panel 3 (LD pruned, no long-range LD)	PC1 ↕	**-.964**	-.024	-.026	-.024	.014	.003	-.002	.002	.002	-.004
	PC2 ↔	.025	.010	**-.752**	.011	-.073	-.098	.183	-.063	.009	-.062
	PC3 ↗	.009	-.035	-.112	.007	.061	.193	-.216	-.106	.139	-.118
	PC4	.017	-.016	.050	-.097	.031	-.112	.061	.006	.027	-.013
	PC5	.012	.018	.003	-.080	-.005	-.031	-.026	-.063	.008	-.163
	PC6	-.007	.013	.006	.007	.001	-.064	.016	-.025	.074	-.008
	PC7	-.005	.001	.032	.020	-.063	-.064	.080	-.151	.026	.091
	PC8	.002	-.019	-.011	-.046	-.083	.001	-.006	-.083	-.007	.050
	PC9	.000	.051	-.012	-.030	.004	.019	.031	.121	.027	.052
	PC10	.000	.009	.021	-.011	.033	.046	.034	-.106	.078	.034

↕: The PC with the highest correlation with the North-South gradient.
↔: The PC with the highest correlation with the East-West gradient.
↗: The PC that also showed a correlation with the East-West gradient, and separates individuals from the middle of the Netherlands from individuals from the rest of the country (illustrated in Figure 1d).

Appendix Table 6: Genotype frequencies (%) of rs8039195 and rs12913832 (from the HERC2 gene) for the Dutch population and the 1000 Genomes populations.

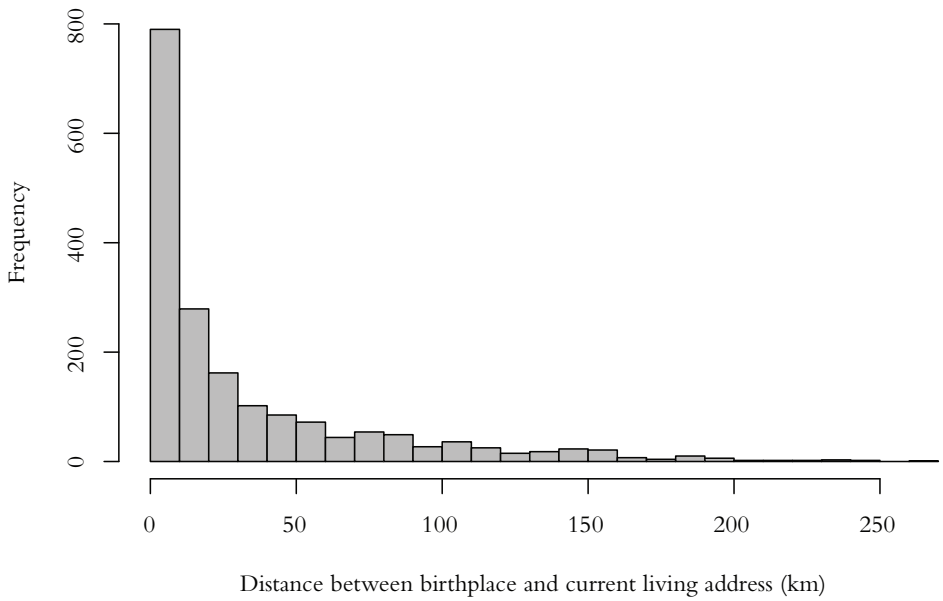
Population	rs8039195 (HERC2)			rs12913832 (HERC2)		
	CC	CT	TT	AA	AG	GG
HapMap Finnish individuals from Finland	.0	6.5	93.5	.0	19.4	80.6
Northern Dutch individuals (top 1000 PC1)	.4	13.1	86.5	-	-	-
British individuals from England and Scotland	1.2	21.4	77.4	3.4	28.1	68.5
Southern Dutch individuals (bottom 1000 PC1)	2.3	23.9	73.7	-	-	-
CEPH individuals	1.2	29.4	69.4	3.4	39.1	57.5
Iberian populations in Spain	.0	50.0	50.0	42.9	50.0	7.1
Colombian in Medellin, Colombia	9.4	43.4	47.2	51.7	41.7	6.7
Toscan individuals	16.8	42.1	41.1	30.6	53.1	16.3
HapMap Mexican individuals from LA California	16.3	46.9	36.7	75.8	15.2	9.1
Puerto Rican in Puerto Rico	9.1	56.4	34.5	65.5	29.1	5.5
Han Chinese South	39.3	48.3	12.4	99.0	1.0	.0
HapMap African ancestry individuals from SW US	43.8	45.8	10.4	70.5	27.9	1.6
Japanese individuals	68.5	23.6	7.9	100.0	.0	.0
Yoruba individuals	77.3	18.2	4.5	100.0	.0	.0
Han Chinese in Beijing	49.5	46.4	4.1	100.0	.0	.0
Luhya individuals	69.3	26.7	4.0	100.0	.0	.0

Appendix Table 7: The 26 Dutch municipalities with a population size > 100k in April 2012 according to the Central Bureau of Statistics[66].

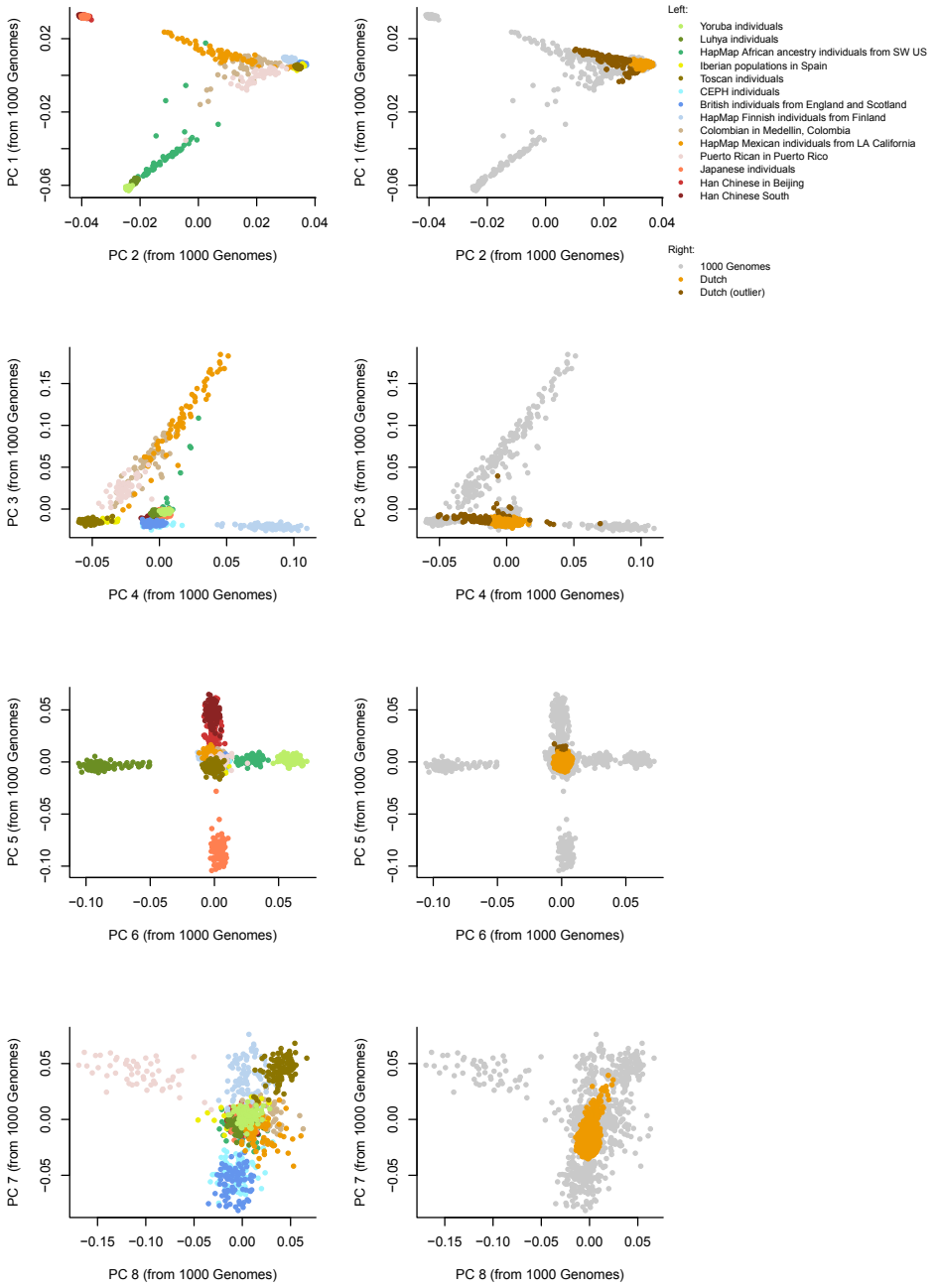
Municipality	Population Size	Randstad
Amsterdam	790,654	✓
Rotterdam	616,525	✓
Den Haag	502,683	✓
Utrecht	317,540	✓
Eindhoven	217,235	
Tilburg	207,398	
Almere	193,615	✓
Groningen	192,871	
Breda	176,835	
Nijmegen	165,262	
Enschede	158,020	
Apeldoorn	157,132	
Haarlem	152,260	✓
Arnhem	149,361	
Amersfoort	148,595	✓
Zaanstad	148,542	✓
Haarlemmermeer	143,885	✓
‘s-Hertogenbosch	141,981	
Zoetermeer	122,334	✓
Zwolle	121,733	
Maastricht	121,008	
Leiden	119,028	✓
Dordrecht	118,723	✓
Ede	108,802	
Emmen	108,779	
Westland	101,670	✓

Appendix Table 8: The correlations of the first three Dutch PCs with the North-South gradient (PC1), the East-West gradient (PC2 & PC3), and F (genome-wide homozygosity) with and without individuals from the major municipalities in the Netherlands.

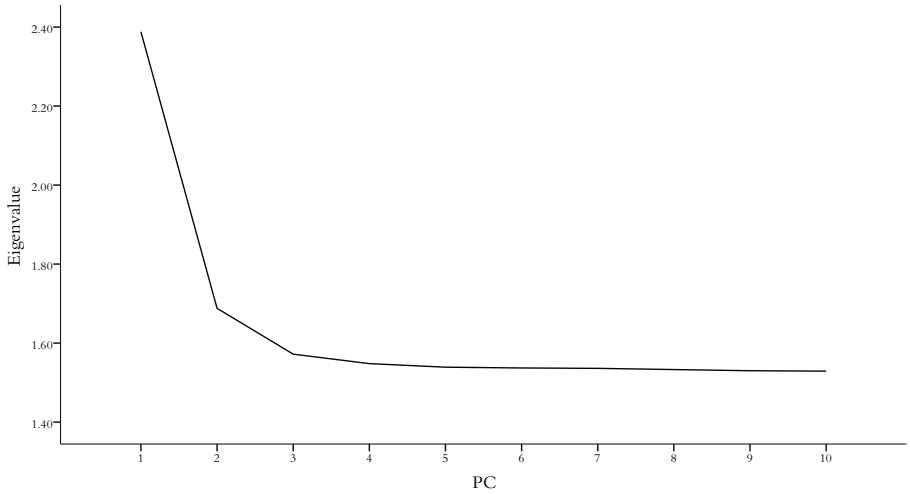
	All unrelated individuals (N = 4,441)	Excluding inhabitants of 4 Randstad municipalities with population size >300k (N = 3,531)	Inhabitants of 4 Randstad municipalities with population size >300k (N = 624)	Excluding inhabitants of 13 Randstad municipalities with population size >100k (N = 3,069)	Inhabitants of 13 Randstad municipalities with population size >100k (N = 1,086)	Excluding inhabitants of 26 municipalities with population size >100k (N = 2,525)	Inhabitants of 26 municipalities with population size >100k (N = 1,630)
$r_{PC1,\updownarrow}$	.603	.648	-.010	.669	.055	.678	.451
$r_{PC1,F}$	.245	.247	.201	.259	.170	.246	.233
$r_{PC2,\leftrightarrow}$	.378	.404	.067	.405	.145	.439	.281
$r_{PC2,F}$	.017	.033	.066	.031	.033	.019	.028
$r_{PC3,\leftrightarrow}$	.162	.171	.092	.166	.085	.164	.162
$r_{PC3,F}$	.009	.008	.063	.011	.018	.004	.036



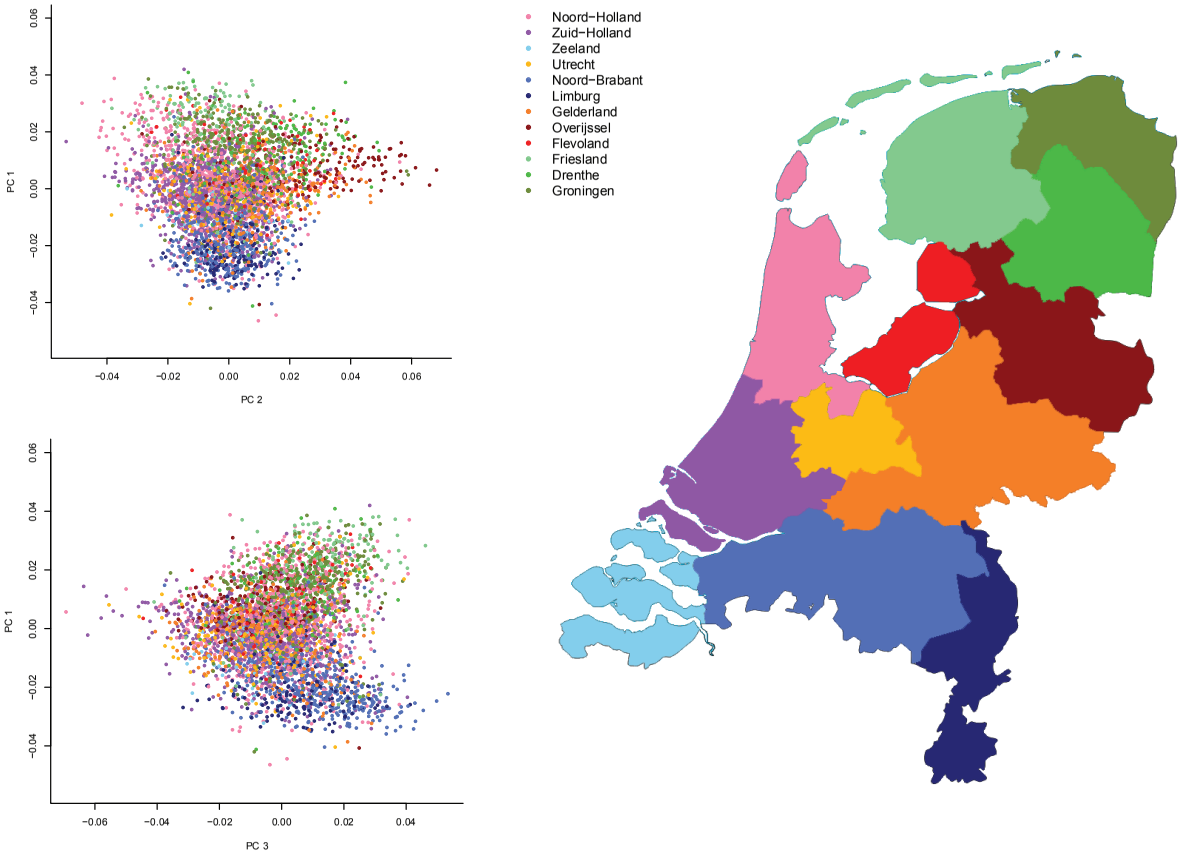
Appendix Figure 1: The distance between birthplace and current living address for 1,841 Dutch individuals. The mean distance is 33.33 km (SD=43.60) or 20.71 mi (SD=27.09).



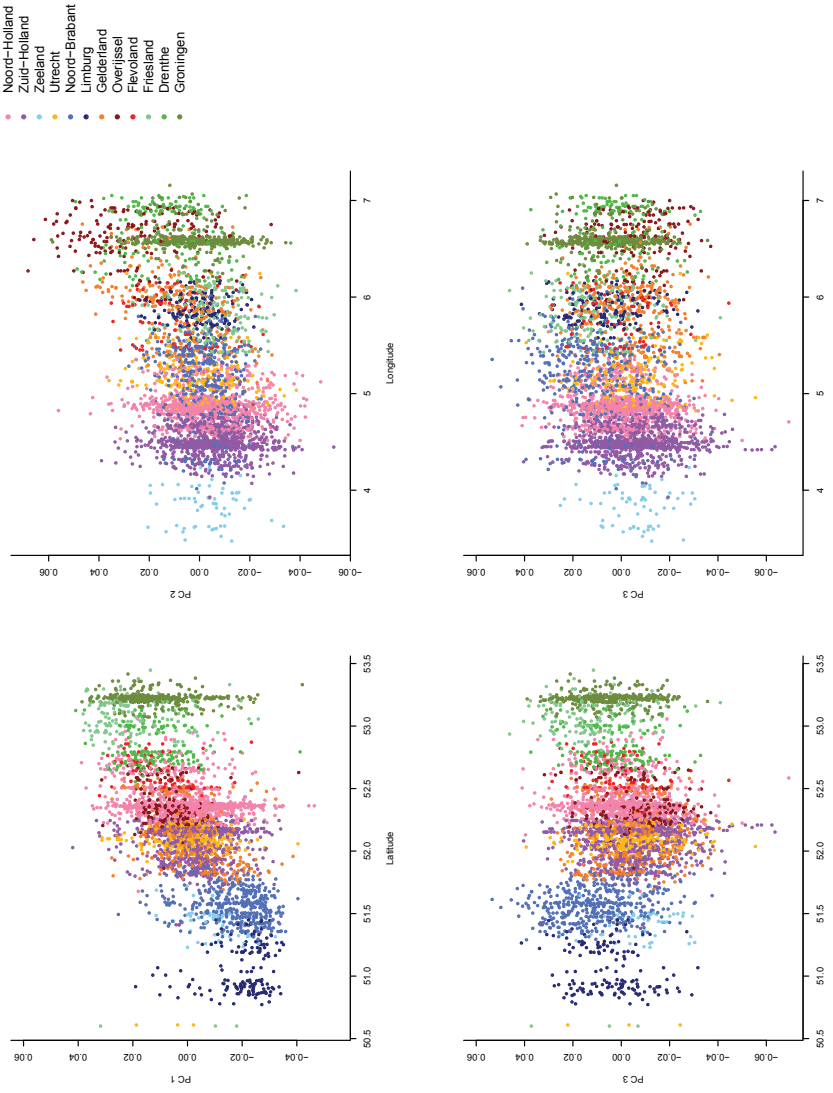
Appendix Figure 2: Identifying individuals with a non-European/non-Dutch ancestry with the projection of the 1000 Genomes PCs on the Dutch PCs.



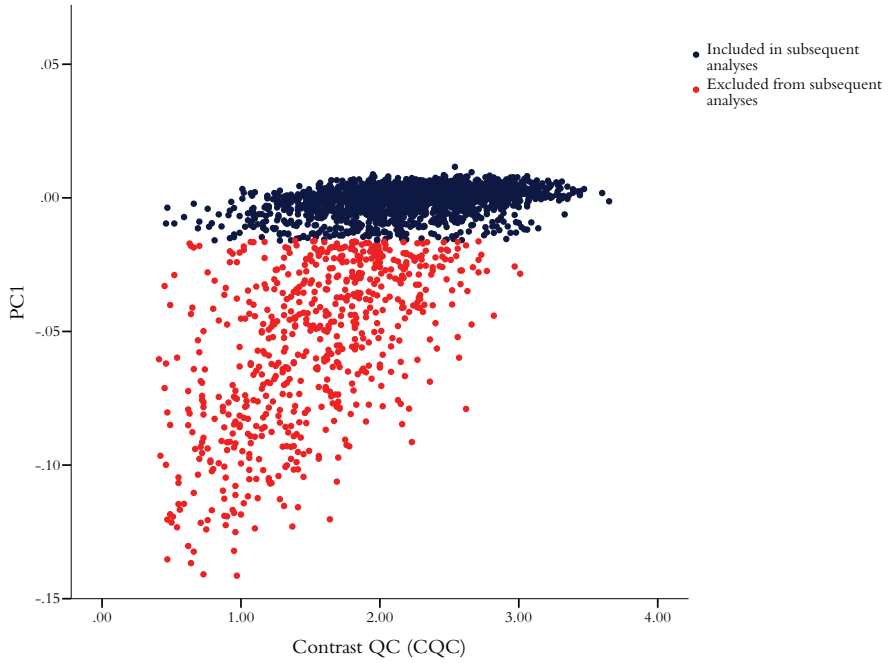
Appendix Figure 3: The eigenvalues of the Dutch PCs from the LD pruned dataset without long-range LD regions or ethnic outliers.



Appendix Figure 4: The X-Y plots for the first three PCs, with each individual colored by province of the current living address (note that PC2 has been inversed [PC2×-1], to illustrate the geographic correlation with East-West).



Appendix Figure 5: PC1 and PC2 plotted against Latitude and Longitude respectively, and PC3 plotted against both Latitude and Longitude, with each individual colored by province of the current living address (note that PC2 has been inversed here as well [PC2×−1]).



Appendix Figure 6: Scatterplot of PC1 from the PCA on 8,207 individuals and CQC.

References

1. Manni, F. (2010). Interview with Luigi Luca Cavalli-Sforza: past research and directions for future investigations in human population genetics. *Human Biology* 82, 245-266.
2. Chen, J., Zheng, H., Bei, J.X., Sun, L., Jia, W., Li, T., Zhang, F., Seielstad, M., and Zeng, Y.X. (2009). Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *American Journal of Human Genetics* 85, 775-785.
3. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., et al. (2008). Genes mirror geography within Europe. *Nature* 456, 98-101.
4. Wang, C., Zöllner, S., and Rosenberg, N.A. (2012). A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS genetics* 8, e1002886.
5. Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* 11, 459-463.
6. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38, 904-909.
7. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., et al. (2008). Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* 83, 132.
8. Durbin, R.M., Altshuler, D.L., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., Francisco, M., Donnelly, P., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-1073.
9. Schalekamp, J.C. (2009). Bataven en buitenlanders: 20 eeuwen immigratie in Nederland (Wind Publishers).
10. Knippenberg, H., and Pater, B. (1997). De eenwording van Nederland: schaalvergroting en integratie sinds 1800 (SUN).
11. Kok, J. (1998). Vrijt daar je zijt": huwelijk en partnerkeuze in Zeeland tussen 1830 en 1950. *K Mandemakers, O Hoogerhuis en A de Klerk (red), Over Zeeuwse mensen Demografische en sociale ontwikkelingen in Zeeland in de negentiende en begin twintigste eeuw Themanummer Zeeland* 7, 131-143.
12. van Poppel, F. (1994). Verbreding van de horizon? Veranderingen in de geografische herkomst van huwelijkspartners. *Acta Geograph Lovaniensia* 34, 79-88.
13. Polman, A. (1951). Geografische en confessionele invloeden bij de huwelijkskeuze in Nederland (Stenfert Kroese).
14. Hendrickx, J., Lammers, J., and Ultee, W. (1991). Religious assortative marriage in the Netherlands, 1938-1983. *Review of Religious Research*, 123-145.
15. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451, 998-1003.
16. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100.
17. Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., and Cavalli-Sforza, L.L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* 102, 15942.
18. Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, J.Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics* 91, 275-292.
19. Boomsma, D.I., De Geus, E.J.C., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.J., Posthuma, D., Van Beijsterveldt, T.C.E.M., Hudziak, J.J., et al. (2006). Netherlands Twin Register:

from twins to twin families. *Twin Research and Human Genetics* 9, 849-857.
20. Penninx, B.W.J.H., Beekman, A.T.F., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W.J., et al. (2008). The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research* 17, 121-140.
21. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2010). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*.
22. Olkin, I., and Finn, J.D. (1995). Correlations redux. *Psychological Bulletin* 118, 155.
23. Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin* 87, 245.
24. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81, 559-575.
25. Weir, B.S. (1996). Genetic data analysis II (Sunderland, MA: Sinauer).
26. Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180, 977-993.
27. Narum, S.R., and Hess, J.E. (2011). Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources* 11, 184-194.
28. CBS. (2012). Centraal Bureau voor de Statistiek, Gezondheidskenmerken naar regio, 1995-1999.
29. Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832-838.
30. Turchin, M.C., Chiang, C.W.K., Palmer, C.D., Sankararaman, S., Reich, D., and Hirschhorn, J.N. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature genetics*.
31. Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* 40, 646-649.
32. Lesger, C. (2003). Noord-Hollanders in beweging: economische ontwikkeling en binnenlandse migratie, ca. 1800-1930 (CGM).
33. Suurenbroek, F. (2001). Binnenlandse migratie naar en uit Amsterdam (1870-1890) (Centrum voor de Geschiedenis van Migranten).
34. Beaumont, M.A., and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* 13, 969-980.
35. Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research* 20, 393-402.
36. Heath, S.C., Gut, I.G., Brennan, P., McKay, J.D., Bencko, V., Fabianova, E., Foretova, L., Georges, M., Janout, V., et al. (2008). Investigation of the fine structure of European populations with applications to disease association studies. *European Journal of Human Genetics* 16, 1413-1429.
37. McEvoy, B.P., Montgomery, G.W., McRae, A.F., Ripatti, S., Perola, M., Spector, T.D., Cherkas, L., Ahmadi, K.R., Boomsma, D., et al. (2009). Geographical structure and differential natural selection among North European populations. *Genome Research* 19, 804-814.
38. Moskvina, V., Smith, M., Ivanov, D., Blackwood, D., StClair, D., Hultman, C., Toncheva, D., Gill, M., Corvin, A., et al. (2010). Genetic differences between five European populations. *Human Heredity* 70, 141-149.
39. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. *PLoS Biology* 4, e72.
40. Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, L.S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I., et al. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genetics* 6, e1000993.
41. Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., Hankinson, S.E., Hu, F.B., Duffy, D.L., et al. (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genetics* 4, e1000074.

42. Kayser, M., Liu, F., Janssens, A., Rivadeneira, F., Lao, O., van Duijn, K., Vermeulen, M., Arp, P., Jhamai, M.M., et al. (2008). Three Genome-wide Association Studies and a Linkage Analysis Identify HERC2 as a Human Iris Color Gene. *American Journal of Human Genetics* 82, 411-423.
43. Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P.N., Stark, M.S., Hayward, N.K., Martin, N.G., and Montgomery, G.W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American Journal of Human Genetics* 82, 424-431.
44. Visser, M., Kayser, M., and Palstra, R.J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Research* 22, 446-455.
45. Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., et al. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology* 18, 1241-1248.
46. Seldin, M.F., Shigeta, R., Villoslada, P., Selmi, C., Tuomilehto, J., Silva, G., Belmont, J.W., Klareskog, L., and Gregersen, P.K. (2006). European population substructure: clustering of northern and southern populations. *PLoS Genetics* 2, e143.
47. Belle, E.M.S., Landry, P.A., and Barbujani, G. (2006). Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. *Proceedings of the Royal Society of London Series B: Biological Sciences* 273, 1595-1602.
48. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The history and geography of human genes (Princeton Univ Pr).
49. Chikhi, L., Nichols, R.A., Barbujani, G., and Beaumont, M.A. (2002). Y genetic data support the Neolithic demic diffusion model. *Proceedings of the National Academy of Sciences of the United States of America* 99, 11008.
50. Torroni, A., Bandelt, H.J., Macaulay, V., Richards, M., Cruciani, F., Rengo, C., Martinez-Cabrera, V., Villems, R., Kivisild, T., et al. (2001). A signal, from human mtDNA, of postglacial recolonization in Europe. *American journal of human genetics* 69, 844.
51. Frost, P. (2006). European hair and eye color: A case of frequency-dependent sexual selection? *Evolution and Human Behavior* 27, 85-103.
52. Mekel-Bobrov, N., Gilbert, S.L., Evans, P.D., Vallender, E.J., Anderson, J.R., Hudson, R.R., Tishkoff, S.A., and Lahn, B.T. (2005). Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. *Science* 309, 1720-1722.
53. Evans, P.D., Gilbert, S.L., Mekel-Bobrov, N., Vallender, E.J., Anderson, J.R., Vaez-Azizi, L.M., Tishkoff, S.A., Hudson, R.R., and Lahn, B.T. (2005). Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science* 309, 1717-1720.
54. Fawcett, K.A., and Barroso, I. (2010). The genetics of obesity: FTO leads the way. *Trends in Genetics* 26, 266-274.
55. Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R.B., Elliott, K.S., Lango, H., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316, 889-894.
56. Klimentidis, Y.C., Abrams, M., Wang, J., Fernandez, J.R., and Allison, D.B. (2011). Natural selection at genomic regions associated with obesity and type-2 diabetes: East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. *Human Genetics* 129, 407-418.
57. Chen, R., Corona, E., Sikora, M., Dudley, J.T., Morgan, A.A., Moreno-Estrada, A., Nilsen, G.B., Ruau, D., Lincoln, S.E., et al. (2012). Type 2 Diabetes Risk Alleles Demonstrate Extreme Directional Differentiation among Human Populations, Compared to Other Diseases. *PLoS Genetics* 8, e1002621.
58. Shiina, T., Ota, M., Shimizu, S., Katsuyama, Y., Hashimoto, N., Takasu, M., Anzai, T., Kulski, J.K., Kikkawa, E., et al. (2006). Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* 173, 1555-1570.

59. Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., et al. (2004). Gene map of the extended human MHC. *Nature Reviews Genetics* 5, 889-899.
60. Barreiro, L.B., and Quintana-Murci, L. (2009). From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics* 11, 17-30.
61. Perkel, J. (2008). SNP genotyping: six technologies that keyed a revolution. *Nature Methods* 5, 447-454.
62. Boomsma, D.I., Willemsen, G., Sullivan, P.F., Heutink, P., Meijer, P., Sondervan, D., Kluft, C., Smit, G., Nolen, W.A., et al. (2008). Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *European Journal of Human Genetics* 16, 335-342.
63. Willemsen, G., de Geus, E.J.C., Bartels, M., van Beijsterveldt, C.E.M.T., Brooks, A.I., Estourgie-van Burk, G.F., Fugman, D.A., Hoekstra, C., Hottenga, J.J., et al. (2010). The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Research and Human Genetics* 13, 231-245.
64. Broek, K.v.d. (2012). 6PP database (http://www.d-centralize.nl/projects/6pp/downloads/).
65. Oxford. (2011). Mathematical Genetics and Bioinformatics Groups; 1000 Genomes June 2011 Readme file: http://mathgen.stats.ox.ac.uk/impute/README_1000G_phase1interim_jun2011.txt.
66. CBS. (2012). Centraal Bureau voor de Statistiek, Bevolkingsontwikkeling; regio per maand, April 2012.

CHAPTER 7

THE STRUCTURAL LANDSCAPE OF DUTCH GENOMES: GENOME-WIDE PATTERNS OF INDELS AND LARGER DELETIONS REFLECTING ANCESTRY

Abdel Abdellaoui, Victor Guryev, Laurent Francioli, Tobias Marschall, Alexander Schoenhuth, Jayne Y. Hehir-Kwa, Wigard Kloosterman, Eric-Wubbo Lameijer, Slavik Koval, Joep de Ligt, Najaf Amin, Freerk van Dijk, Lennart Karssen, Hailiang Mei, Jouke-Jan Hottenga, Eco J. de Geus, Gert-Jan van Ommen, Paul de Bakker, Cisca Wijmenga, Cornelia M. van Duijn, P. Eline Slagboom, Dorret I. Boomsma, Kai Ye on behalf of the GoNL Consortium

Abstract

We analyzed genome-wide patterns of indels and larger deletions from Next Generation Sequencing (NGS) data of 490 unrelated Dutch individuals who are participants in the Genome of the Netherlands (GoNL) project. Genetic variants were analyzed by Principal Component Analysis (PCA) in order to explore genome-wide patterns of variation reflecting ancestry. Geographical location and three ancestry-informative SNP PCs from a previous study were used to determine which indel PCs and larger deletion PCs were likely to reflect ancestry.

Indels showed three, and larger deletions showed five PCs with significant correlations with geography and ancestry-informative SNP PCs. The three indel PCs showed similar geographic North-South and East-West distributions as PCs based on microarray SNP data, which is likely partly explained by LD between the different forms of genetic variation. Two (uncorrelated) indel PCs showed a slightly different East-West distribution than the East-West SNP PC however, which also showed a greater genetic distance between East and West than observed with the East-West SNP PC.

F_{st} values showed that the genetic distance/differentiation between subpopulations identified by the indel PCs was similar across SNPs, indels, and larger deletions. Subpopulations identified by larger deletion PCs however show much higher and unusually large F_{st} values for larger deletions than for indels or SNPs, indicating that the variation captured by larger deletion PCs mostly reflects genome-wide patterns that are specific to larger deletions.

The GoNL project selected participants from trios (2 parents and at least one offspring) in which parents were born in the same province of the Netherlands, and thus were more likely to share recent ancestry. In contrast to SNP PCs and indel PCs, ancestry-informative larger deletion PCs did not show significant spouse correlations. This suggests that larger deletions may capture older ancestry signals than SNPs or indels, or weaker ancestry signals because of the limited amount of measured larger deletions.

Introduction

Population genetics helps us understand the larger patterns of human genetic variation and their origins, which can also be useful in the search for relationships between genetic and phenotypic variation. While microarray data have contributed much to population genetics, the higher resolution of whole-genome sequence data is expected to yield new insights into population stratification, population history, the prevalence of selection pressures, and the identification of functional variants under selection. The variation of structural variants on a population level is still relatively uncharted territory. In contrast to single nucleotide variants (SNPs), whose variation has been characterized extensively in many human populations, short insertions and short deletions (i.e., indels; the second most abundant form of genetic variation)[1] and larger structural variants (SVs; deletions, insertions, inversions, translocations, etc.) remain the more difficult types of genomic variants to discover and genotype. Next-generation sequencing (NGS) technologies offer the promise of a single experiment to detect all genomic variant types without the SNP ascertainment biases that microarrays may show. Recently the 1000 Genomes Project structural variant group applied the current optimal SV detection methods and illustrated the power of combined SV detection approaches in NGS data.[2] However, in large sequencing projects like 1000 Genomes that aim to capture human genetic variation across all populations, the number of individuals for a single population is very modest. We showed previously that considerable genetic heterogeneity may also exist within such single populations, even in populations occupying a small geographic area like the Netherlands.[3]

The Genome of the Netherlands project (GoNL),[4] based on medium coverage whole-genome sequence and a trio design, is suited for indel/SV discovery whilst enabling the study of transmission of variants across generations. The datasets, contributed by four Dutch biobanks, consists of 229 trios with one offspring and 19 trios with an additional co-twin sequenced, with both parents in the trios born in the same province. The dataset represents 11 out of 12 provinces in the Netherlands, with an approximately equal representation of the 11 provinces with respect to sample size. On each of the 763 samples, whole-genome sequencing was performed with an average coverage of 13.3x. In this study we employ these data to analyze the main patterns of common structural variation in Dutch genomes.

In order to investigate the main patterns of variation, principal component analysis (PCA) was run on common variants (MAF > .01) from genotyped indels (<20 bp) and genotyped larger deletions (20-10,000 bp). We expect good quality genotype calls to result in principal components (PCs) reflecting ancestry as they do with SNP data.[5] Ancestry-informative PCs can reveal consequences of population history. PCs based on microarray SNP data have been shown to

correlate strongly with geography in the Netherlands,[3] show genetic traces of a serial founder effect, and are significantly associated with several complex phenotypes (like height, eye color, and hair color), which is likely at least partly the cause of past diversifying selection pressures.[3;6;7] These PCs also showed significant correlations with continental variation on a genotypic and phenotypic level. Here, we explore the extent to which common short indels and common larger deletions capture the same patterns of variation as common SNPs by comparing ancestry-informative PCs derived from indels and larger deletions with ancestry-informative PCs previously extracted with microarray SNP data. We furthermore explore their relationship with geography, and with phenotypes known to reflect the North-South cline. Many indels that vary between populations have shown to be in linkage disequilibrium (LD) with the more abundant SNPs,[8] suggesting that common indels may capture common SNP variation as well.

Results

Identifying ancestry-informative PCs

A total of 1,733,833 indels (<20 bp) and 46,633 larger deletions (20 – 10,000 bp) were discovered and genotyped using a combination of multiple SV calling approaches.[9] A PCA was conducted on common (MAF > .01) and LD pruned indels and larger deletions separately in 490 unrelated subjects in order to capture genome-wide patterns of variation reflecting ancestry. The top ten PCs were compared with two so-called "indicators of ancestry" in order to select putative ancestry-informative PCs:

1. The first three PCs based on Affymetrix 6.0 SNPs in 5,166 unrelated Dutch individuals from the Netherlands Twin Registry (NTR) and Netherlands Study of Anxiety and Depression (NESDA), which have been shown to correlate strongly with geography in the Netherlands.[3] The Affymetrix 6.0 SNPs were extracted from the GoNL dataset, and formed the input data to project the three ancestry-informative PCs from the 5,116 unrelated NTR and NESDA subjects onto the 769 GoNL subjects (490 unrelated). This resulted in the same geographic distributions in the GoNL dataset as in the NTR/NESDA dataset (see Appendix Figure 1). These three PCs will be referred to as *common SNP PCs*, or *North-South SNP PC* (= PC1), *East-West SNP PC* (= PC2), and *Middle-Band SNP PC* (= PC3).
2. Geographic location (latitude and longitude). Geographic proximity is associated with shared ancestry, visible by decreasing genetic similarity with increasing distance. Geographic location was available for 480 of the 490 unrelated GoNL subjects (based on birthplace for 373 subjects and current living address for 107 subjects).

The total number of correlations used to identify ancestry-informative PCs is 100 (20 × 5): correlations between 20 SV PCs (top 10 indel PCs and top 10 larger deletions PCs) and five indicators of ancestry (three ancestry-informative *common SNP PCs*, longitude, and latitude coordinates of birthplace/current living address). This makes the significance threshold 5×10^{-4} after Bonferroni correction. To investigate the extent of sequence quality differences captured by the PCs, the correlation between PCs and the average coverage per individual was also inspected (higher coverage = more input reads = better variant discovery and genotyping).

Indels: After quality control (QC) on indels similar to QC previously performed on Affymetrix 6.0 SNPs (see reference[3] and methods), several indel PCs showed significant correlations with geography (latitude and/or longitude) and the three *common SNP PCs*. Besides a significant correlation with latitude and the *North-South SNP PC* ($r = .286$, $p = 2 \times 10^{-10}$, and $r = .220$, $p = 8.6 \times 10^{-7}$ respectively), the first PC (i.e., explaining most variation) also showed a much higher and more significant correlation with the average coverage ($r = .65$, $p = 9.8 \times 10^{-61}$). The fact that the correlations with latitude and the *North-South SNP PC* were also significant within multiple biobanks, suggests that this indel PC picked up a mixture of ancestry signals and sequence quality differences We repeated the PCA several times while decreasing the number of indels by filtering on the quality metric provided by GATK (QUAL, i.e., the Phred scaled probability that a REF/ALT polymorphism exists at the site)[10] in order to decrease the dilution of ancestry signals. Figure 1 illustrates how increasing the quality threshold increases correlations with indicators of ancestry signals for the first PC (with correlations with the *North-South SNP PC* exceeding 0.8), while the North-South cline gets decreasingly and coverage/quality differences increasingly more captured by the subsequent PCs. We decided to keep increasing the quality threshold, until the indel PCs explaining most variation captured more ancestry than quality differences (i.e., until ancestry-informative indel PCs showed higher correlations with geography and the *common SNP PCs* than with the average coverage). Figure 1 shows the selected quality threshold for this goal to be at 2500. At this threshold, the first three indel PCs show higher correlations with geography and the ancestry-informative *common SNP PCs* than with the average coverage (Table 1, Figures 1-2, Appendix Table 1). 760,276 indels out of 1,733,833 remained at this threshold (~44%). All PCAs were conducted on LD pruned sets of indels, resulting in a less severe loss of variants after additional QC and LD-pruning, as indels that do not pass the quality threshold can be replaced by higher quality indels in LD during LD-pruning (see methods for QC, and reference[3] for the reasons behind LD pruning). With a quality threshold of 2500, eventually 241,172 indels remain for PCA instead of 318,492 at a quality threshold of 0 (i.e., ~76%), indicating that the large majority of genome-wide indel variation is maintained.

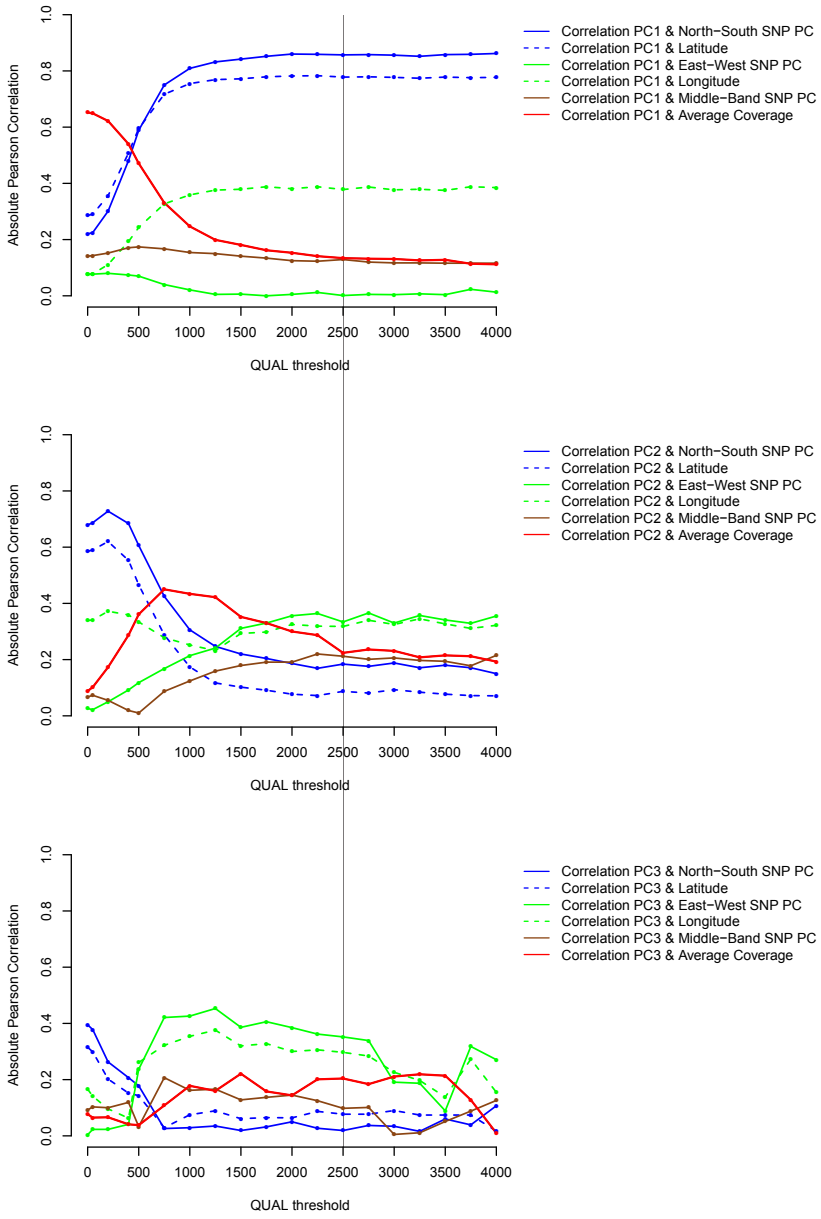


Figure 1: Pearson correlations between the first three indel PCs and several indicators of ancestry: common SNP PCs, latitude, longitude, and average coverage, for 19 PCAs (plotted in points) with quality thresholds between 0 and 4,000. The dotted black line shows the quality threshold used in further analyses, where more ancestry than quality differences are captured by the first three PCs. PC1 and PC2 are relatively constant after this point, while PC3 still shows fluctuations as the number of indels decreases.

Table 1: The three indel PCs that showed significant correlations with geography (N=480) and/or ancestry-informative common SNP PCs (N=489). The Table shows Pearson correlations and *p*-values between brackets.

	PC1	PC2	PC3
Latitude(↕)	**.78 (< 10^{-17})**	**.09 (.03)**	-.08 (.05)
Longitude (↔)	**.38 (< 10^{-17})**	**.32 (5.2 × 10^{-13})**	**-.3 (1.7 × 10^{-11})**
North-South SNP PC	**.86 (1.6 × 10^{-12})**	**.18 (3.5 × 10^{-5})**	.02 (.34)
East-West SNP PC	0 (.49)	**.36 (1.6 × 10^{-12})**	**-.37 (1.6 × 10^{-12})**
Middle-Band SNP PC	**-.14 (.001)**	**.18 (2.4 × 10^{-5})**	**-.11 (.006)**

Bold: $p < .05$; Red: $p < 5 \times 10^{-4}$

Table 2: The five PCs from larger deletions that showed significant correlations with geography (N=480) and/or ancestry-informative SNP PCs (N=489). The Table shows Pearson correlations and p-values between brackets.

	PC1	PC2	PC3	PC9	PC10
Latitude(↕)	**.26 (5 × 10^{-9})**	**.13 (.002)**	**-.19 (1.6 × 10^{-5})**	.03 (.25)	**.19 (2.3 × 10^{-5})**
Longitude (↔)	**.18 (2.9 × 10^{-5})**	.06 (.08)	-.06 (.09)	-.01 (.42)	**.08 (.04)**
North-South SNP PC	**.30 (6.7 × 10^{-12})**	**.16 (2.2 × 10^{-4})**	**-.21 (2 × 10^{-6})**	.01 (.38)	**.17 (8.4 × 10^{-5})**
East-West SNP PC	.02 (.32)	-.04 (.19)	.02 (.36)	**-.16 (2.3 × 10^{-4})**	**.12 (.004)**
Middle-Band SNP PC	-.03 (.25)	.05 (.13)	.01 (.42)	**-.17 (5.3 × 10^{-5})**	.06 (.11)

Bold: $p < .05$; Red: $p < 5 \times 10^{-4}$

Larger deletions: After QC and minimizing LD, 15,517 out of 46,633 larger deletions remained as input data for PCA. These larger deletions showed five putative ancestry-informative PCs with significant correlations with geography and/or the ancestry-informative *common SNP PCs* (Table 2 and Figure 3). None of the top ten larger deletion PCs showed a significant correlation with average coverage (Appendix Table 1).

Geographic distributions of the ancestry-informative SV PCs

Since biobanks were somewhat geographically biased in their sampling (i.e., certain biobanks are more overrepresented in certain regions, as illustrated in Figures 2 and 3 and Appendix Figure 1), correlations with geography and *common SNP PCs* were also computed within biobanks for the putative ancestry-informative PCs, in order to ensure that the geographic correlations are not due an ascertainment bias. The SV PCs showed (nominally) significant correlations within biobanks as well (see Appendix Tables 2-3), suggesting the significant geographical correlations reported in Tables 1 and 2 are not due to a sampling bias.

The indel PCs showed higher and more significant correlations with geography than the larger deletion PCs (see Figures 2 and 3 respectively). Indel PC1 mostly captured the same ancestry differences as the *North-South SNP PC*, namely the Dutch North-South cline. Indel PC2 and Indel PC3 both showed a geographic resemblance with the *East-West SNP PC* (with which they showed a correlation of .36 and .37 respectively), but with a more spread out eastern component. The geographic distribution of these two PCs is very similar, while the PCs themselves in fact are uncorrelated with each other ($r = 0$, $p = 1$; PCs are by definition uncorrelated with each other). While larger deletion PCs showed significant correlations with geography, their geographic distributions were less clear when plotted (Figure 3), although some local geographic clustering is visible (the North-Eastern region of large deletion PC1 for example). More samples and/or larger deletions may increase the resolution of the geographic distribution of the genome-wide patterns of larger deletions.

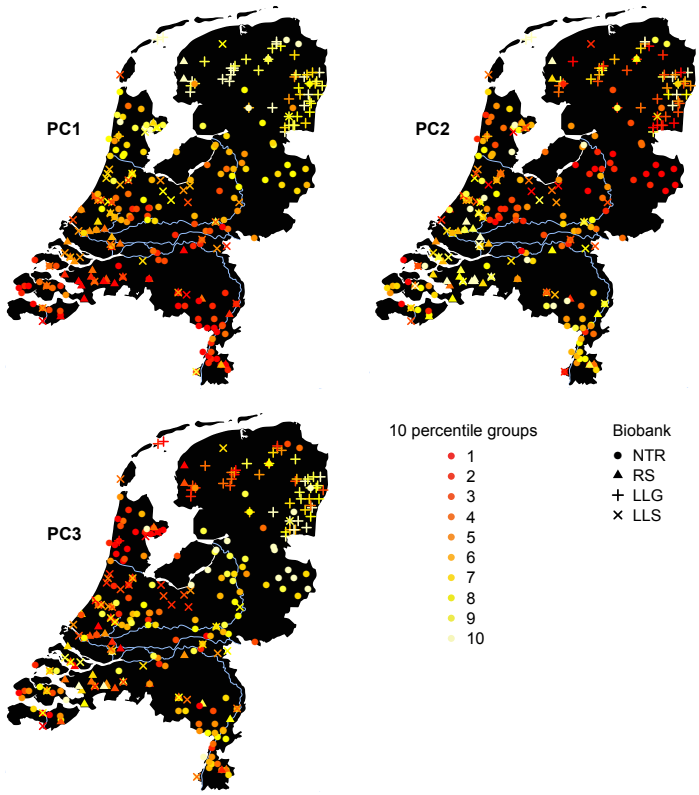


Figure 2: Ancestry-informative indel PCs showing significant correlations with geography and SNP ancestry PCs. The mean value per city was calculated, divided into ten percentile groups, and plotted. NTR = Netherlands Twin Registry; RS = Rotterdam Studies; LLG = LifeLines Groningen; LLS = Leiden Longevity Study.

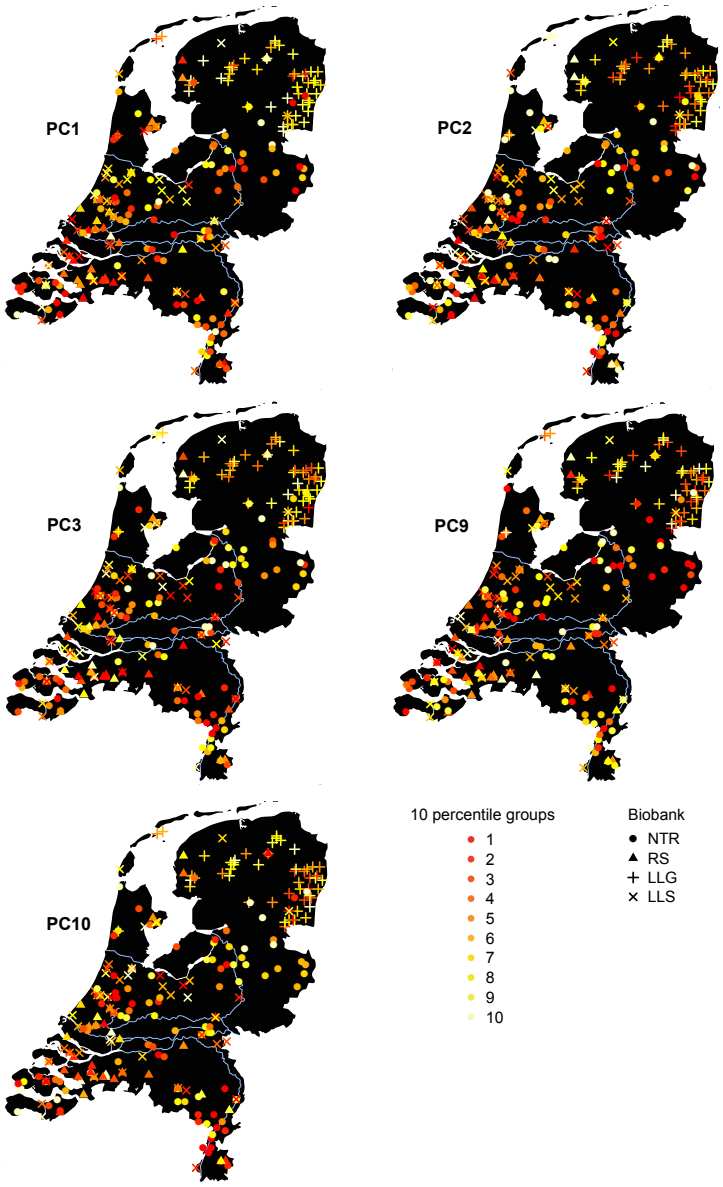


Figure 3: Ancestry-informative PCs from larger deletions showing significant correlations with geography and SNP ancestry PCs. The mean value per city was calculated, divided into ten percentile groups, and plotted. NTR = Netherlands Twin Registry; RS = Rotterdam Studies; LLG = LifeLines Groningen; LLS = Leiden Longevity Study.

Table 3: Absolute Pearson correlations with the European North-South PC from 1000 Genomes, genome-wide homozygosity (F), and the partial correlation with height (corrected for gender), and Spearman's ρ with eye and hair color.

	European North-South PC (N=490)	Genome-wide homozygosity (F) (N=490)	Height (N=423)	Eye color (N=136; *NTR subjects only)	Hair color (N=137; *NTR subjects only)
North-South SNP PC	**.78 (6.5×10^{-100})**	**.36 (1.1×10^{-16})**	**.12 (.01)**	.16 (.06)	**.30 (3.2×10^{-4})**
Indel PCs with significant correlations with the Dutch North-South cline (i.e., latitude and/or North-South SNP PC):					
PC1	**.68 (1×10^{-67})**	**.34 (1.5×10^{-14})**	.09 (.07)	.07 (.44)	**.23 (.008)**
PC2	**.20 (1.1×10^{-5})**	.07 (.13)	**.11 (.03)**	.12 (.18)	.09 (.32)
Larger deletion PCs with significant correlations with the Dutch North-South cline (i.e., latitude and/or North-South SNP PC):					
PC1	**.26 (4.6×10^{-9})**	**.10 (.03)**	.02 (.72)	.06 (.49)	**.18 (.03)**
PC2	.09 (.05)	.02 (.74)	.09 (.06)	.07 (.41)	.13 (.14)
PC3	**.15 (.001)**	.03 (.47)	.04 (.40)	.08 (.34)	.05 (.60)
PC10	**.15 (.001)**	.05 (.24)	.05 (.30)	.04 (.63)	.02 (.80)
North-South SNP PC from Affy 6.0 dataset	N=5,166 **.664 ($<10^{-100}$)**	N=5,166 **.245 (1.4×10^{-71})**	N=4,292 **.17 (3.5×10^{-29})**	N=1,538 **.13 (4.9×10^{-7})**	N=1,540 **.12 (3×10^{-6})**

Bold: $p < .05$; Red: $p < 5 \times 10^{-4}$

*Note that for the GoNL dataset, eye color and hair color are only available for NTR subjects, so the correlations might be slightly geographically biased, although the NTR part of GoNL has the most geographic coverage (Figures 2-3).

The European North-South cline: Inbreeding coefficient (F), height, eye color, hair color, and the European North-South PC from 1000 Genomes

The *North-South SNP PC* has shown several similarities with the European North-South cline in the NTR/NESDA Affymetrix 6.0 dataset (see reference[3] and bottom row of Table 3), such as a significant correlation with genome-wide homozygosity (F), height (North = taller),[6;11] eye color (North = more blue eyes),[7] hair color (North = more blond hair),[7] and the PC from 1000 Genomes that separates Southern European from Northern European populations (will be referred to as *European North-South PC*). In the GoNL dataset, the *North-South SNP PC* correlates .78 with the *European North-South PC*, and is also significantly associated with F, height, and hair color. There were multiple indel and larger deletion PCs that showed significant correlations with the Dutch North-South cline (both with latitude, and the *North-South SNP PC*), but varied in their correlations with the *European North-South PC*, F, height, eye color, and hair color (Table 3).

Indel PC1 shows the highest correlation with the *European North-South PC*

of all indel and larger deletion PCs (r = .68), likely because it mostly captures the same variation as the *North-South SNP PC*, and also shows the highest correlations with F (r = .34) and hair color (r = .23). Only indel PC2 showed a nominally significant correlation with height however, while we would have expected indel PC1 to show the association with height, as indel PC1 mostly captures the same European North-South cline as the *North-South SNP PC.* Indel PC2 did not show a significant correlation with F, eye color, or hair color.

Three out of four larger deletion PCs with significant correlations with the *North-South SNP PC* showed a (nominally) significant correlation with the *European North-South PC*. Larger deletion PC1 showed the strongest correlation with the *European North-South PC*, and also was the only larger deletion PC to show a nominally significant correlation with F and hair color. None of the larger deletion PCs showed a significant correlation with height or eye color.

Resemblances between parents from trio's

Assortment on ancestry is large when considering more differentiated populations, but can also be quite high within relatively homogeneous regions like the Netherlands. In a previous study, the three *common SNP PCs* showed significant spouse correlations in the NTR/NESDA Affymetrix 6.0 dataset in 157 spouse pairs,[3] with an exceptionally high spouse correlation for the *North-South SNP PC* (*North-South SNP PC*: r = .555, p < .001; *East-West SNP PC*: r = .164, p = .04; *Middle-Band SNP PC*: r = .179, p = .025). The *North-South SNP PC* shows the same geographic distribution as the distribution of Catholics and Protestants in the last couple of centuries. Religion has shown strong assortment for centuries and also shows a high spouse correlation of .73 in the current generation, which likely explains the fact that the spouse correlation of the *North-South SNP PC* is so much higher than that of the *East-West SNP PC* and the *Middle-Band SNP PC*.[12] When projecting the ancestry-informative *common SNP PCs* on GoNL samples, the 248 spouse pairs show a much higher and more significant spouse correlation than

Table 4: Spouse correlations of ancestry-informative PCs (N=248 spouse pairs with offspring)

Common SNP PCs	Spouse correlation	Indel PCs	Spouse correlation	Larger deletion PCs	Spouse correlation
North-South SNP PC	**.784 ($p = 8.6 \times 10^{-53}$)**	PC1	**.730 ($p = 1.4 \times 10^{-42}$)**	PC1	.008 (p = .902)
East-West SNP PC	**.737 ($p = 8.6 \times 10^{-44}$)**	PC2	**.550 ($p = 5.4 \times 10^{-21}$)**	PC2	-.048 (p = .454)
Middle-Band SNP PC	**.550 ($p = 4.8 \times 10^{-21}$)**	PC3	**.374 ($p = 1.2 \times 10^{-9}$)**	PC3	.083 (p = .193)
				PC9	041 (p = .521)
				PC10	-.019 (p = .768)

Bold: $p < 5 \times 10^{-4}$

the 157 spouse pairs from the NTR Affymetrix 6.0 dataset for all three *common SNP PCs* (Table 4). This is the result of the sampling scheme of GoNL, where both members of a spouse pair and their ancestors up to two generations (as far as ancestral birthplaces were known) had to be born in the same province.

All ancestry-informative indel PCs showed considerably high spouse correlations, ranging from .37 to .73 (Table 4). In contrast, the larger deletion PCs all show non-significant spouse correlations around zero. This is an intriguing observation, which could mean that the larger deletion PCs either capture much older ancestry signals, or very weak ancestry signals.

F_{st} values between subpopulations identified by PCs

F_{st} values in the context of comparing two populations can be used to quantify the genetic distance between populations, i.e., the extent of population differentiation due to genetic drift and selection pressures.[13] In order to examine the genetic differentiation that is captured by the ancestry-informative PCs across different forms of genetic variation, Hudson's F_{st} (an F_{st} estimator that was designed to deal with the large number of markers from NGS data)[14] was computed between the top and bottom 150 individuals of each ancestry-informative indel PC, larger deletion PC, and *common SNP PC*. For each PC, we computed the genome-wide F_{st} values for each set of variants separately: 8,685,291 common SNPs, 838,979 common indels (<20 bp), 17,340 common larger deletions (20-10,000 bp). That makes 33 F_{st} values in total (11 ancestry-informative PCs × three sets of variants), which are all shown in Table 5.

For the indel PCs and *common SNP PCs*, we can draw three main conclusions from Table 5. The first is that the three groups of genetic variants show similar F_{st} values for the same PCs, indicating that these PCs truly capture genome-wide variation of different forms of genetic variants. The second is that PCs capturing the Dutch North-South cline (*North-South SNP PC* and Indel PC1) consistently show F_{st} values that are about twice as high as the other PCs across all three types of genetic variants, consistent with the relatively strong historical segregation of the regions below and above the major rivers, with the most recent source of differentiation being the social segregation due to religious differences during the last four centuries.[12; 15] The third is that, even though indel PC2 and indel PC3 show a similar East-West geographic distribution as the *East-West SNP PC* (see Figure 1 and Appendix Figure 1), they consistently show higher F_{st} values than the *East-West SNP PC*, indicating that these indel PCs capture a stronger East-West cline.

The three sets of variants show the largest discrepancy with respect to F_{st} values when computed between subpopulations identified by the larger deletion PCs. F_{st} values for the larger deletion PCs are unusually high

when computed with larger deletions compared to when computed with SNPs or indels. This suggests that ancestry-informative larger deletion PCs capture genome-wide patterns that are more specific to larger deletions.

Table 5: Hudson's F_{st} between top and bottom 150 individuals for each ancestry-informative PC separately for SNPs, indels, and larger deletions

	SNPs	Indels	Larger deletions
Common SNPs PCs			
North-South SNP PC	.000656 (SD: .000026)	.000581 (SD: .000024)	.000583 (SD: .000063)
East-West SNP PC	.000245 (SD: .000025)	.000203 (SD: .000025)	.000069 (SD: .000042)
Middle-Band SNP PC	.000142 (SD: .000024)	.000111 (SD: .000026)	.000126 (SD: .000068)
Indel PCs			
PC1	.000656 (SD: .000022)	.000613 (SD: .000023)	.000504 (SD: .000061)
PC2	.000306 (SD: .000025)	.000317 (SD: .000028)	.000192 (SD: .000048)
PC3	.000315 (SD: .000030)	.000315 (SD: .000032)	.000254 (SD: .000054)
Larger deletion PCs			
PC1	.000344 (SD: .000020)	.000302 (SD: .000020)	.001010 (SD: .000060)
PC2	.000274 (SD: .000019)	.000218 (SD: .000018)	.000989 (SD: .000065)
PC3	.000260 (SD: .000022)	.000237 (SD: .000023)	.001020 (SD: .000053)
PC9	.000264 (SD: .000026)	.000240 (SD: .000027)	.000866 (SD: .000058)
PC10	.000269 (SD: .000024)	.000238 (SD: .000024)	.000767 (SD: .000057)

Discussion

We used NGS data to analyze genome-wide patterns of variation reflecting ancestry in indels and larger deletions within a population from a small geographic area. Indels and larger deletions from NGS partly captured ancestry differences previously observed with micro-array SNP data.

The indel PC explaining most variation (PC1) mostly captured the North-South cline previously captured with common SNPs from the Affymetrix 6.0 chip, likely due to LD with SNPs.[8] Besides correlating .86 with the *North-South SNP PC*, it showed a very similar geographic distribution, and had a similar relationship with indicators of European North-South differences as evidenced from: 1) the correlation with the *European North-South PC*, 2) the correlation with genome-wide homozygosity, likely due to a serial founder effect,[3] 3) the association with hair color, and 4) similar F_{st} values. Part of the common Affymetrix 6.0 SNPs were chosen because they were common in European populations. The indels included in this study did not have that bias, indicating that the European North-South cline captured with microarray SNPs did not explain most genetic variation in common

SNPs because of an ascertainment bias in SNPs.

Even though they were uncorrelated, indel PC2 and indel PC3 both showed a similar geographic distribution and seem to mainly separate the Eastern part of the country from the Western part, similar to the *East-West SNP PC*, with which they both showed moderate but significant correlations. Indel PC2 was the only indel PC to show a nominally significant association with height, an association we would have expected for indel PC1, since that PC mainly captures the (European) North-South cline captured by the *North-South SNP PC*, which also showed a significant association with height. Indel PC1 did show a nearly significant association with height ($p = .07$; Table 4), so we cannot exclude the lack of power due to a modest sample size as a possible explanation. Nevertheless, the association between indel PC2 and height suggests that the phenotypic associations with the North-South cline may be more complicated than previously suggested by the phenotypic associations with the *North-South SNP PC*.

Indel PC2 and PC3 show higher F_{st} values than the *East-West SNP PC* (see Table 5), suggesting the divergence between subpopulations captured by these PCs is stronger than those captured by the *East-West SNP PC*. Whether this difference is due to the stronger selection pressures in indels,[1] or due to indels capturing older variation (which would give genetic drift more time to increase F_{st} values)[13] needs to be investigated. One way to proceed would be to combine the F_{st} values with the effective population size (N_e) to compute the divergence time (T) between subpopulations identified by the ancestry-informative PCs with the equation: $2N_eF_{st}$.[16] An assumption that is likely violated with this equation is the absence of gene flow subsequent to divergence, which would drive estimations of T downwards, but may at least allow us to put the ancestry-informative PCs in chronological order, which could also be helpful in interpreting them in a historical context.

The five ancestry-informative larger deletion PCs behave in an unusual way compared to *common SNP PCs* and indel PCs. F_{st} values between subpopulations identified by larger deletion PCs suggest that they capture variation that is more specific to larger deletions, since the F_{st} values for larger deletions are higher than F_{st} values computed for SNPs or indels (see bottom five rows of Table 5). F_{st} values between subpopulations identified by indel PCs and *common SNP PCs* do not show such a discrepancy, since SNPs, indels, and larger deletions show more similar F_{st} values for these PCs, indicating that they capture more general genome-wide variation (see top six rows of Table 5). The larger deletion PCs are also the only ancestry-informative PCs that do not show significant spouse correlations, despite the spouses being chosen for sharing (recent) ancestry. The larger deletion PCs did show clear indications of capturing ancestry (significant correlations with geography, *common SNP PCs*, the *European North-South PC*, hair color, and no

correlation with coverage). The lack of spousal resemblance, the unclear geographic distributions in Figure 3 may indicate 1) that the larger deletion PCs capture older ancestry differences, which would be supported by the unusually large F_{st} values larger deletions show for these PCs, or 2) that the ancestry signals captured by larger deletions are much weaker than those captured by SNPs and indels, which would be supported by the large difference in the number of measured larger deletions and the much more abundant indels and SNPs. It was previously estimated that ~100,000 SNPs would be required to achieve sufficient power to effectively capture population differences with an F_{st} of ~.001,[5] although the larger deletions may be better at separating populations as SVs account for a greater proportion of the diversity between individuals than SNPs.[17]

Including SVs in population genetics studies can give a more complete picture of genetic variation at a population level, and may be important for structural variation studies on genotype-phenotype relations, which need to account for artifacts due to population stratification. Ancestry-informative signals for NGS data may also help identify variants under selection pressures, of which many have been observed in this population.[3] The higher resolution of NGS data may result in identifying functional genetic variants that are important for traits that increased fitness and/or reproductive success. This study aims to map genome-wide SV variation in a relatively homogeneous population and while capturing signals of genetic ancestry differences, stumbles into some interesting questions regarding unique qualities of SVs that we intend to further investigate in the near future.

Methods

Participants

The Genome of the Netherlands[4] dataset consisted of 248 unselected trios (8 trios with an additional dizygotic twin sequenced, and 11 trios with an additional monozygotic twin) with Dutch ancestry and were selected from 11 out of 12 provinces of the Netherlands. Four different biobanks contributed to the dataset: 1) The LifeLines cohort study (N=165), a three-generation population-based cohort representing ~10% of the northern part of the Netherlands;[18] 2) The Leiden Longevity Study (LLS; N=72), including subjects from families that have particularly high longevity, where at least two of the long-living siblings must still be alive to be included in the cohort;[19] 3) The Netherlands Twin Registry (NTR; N=349): a collection of families from 2 or 3 consecutive generations from the entire country based on the presence of twins, triplets, quadruplets, etc.;[20;21] 4) The Rotterdam studies (N=183), consisting of a population-based long term follow-up study from Rotterdam and its surrounding area, and the Genetic Research

in Isolated Populations program (GRIP) targeting the southwestern part of the province Noord-Brabant.[22; 23] Two parents from two trios were excluded due to traces of contamination, reducing 250 complete trios to 248 complete trios. There were a total of 643 individuals from all four biobanks with height measured (426 unrelated), 216 NTR subjects with eye color available (136 unrelated), and 216 NTR subjects with hair color measured (137 unrelated). Self-reported eye color was coded as blue, intermediate or brown, and self-reported hair color was coded as blond, red, light brown, dark brown, or black.

The 490 unrelated subjects were extracted from the 769 participants using GCTA[24] on 2,369,591 SNPs, by excluding one of each pair of individuals with an estimated genetic relationship of <0.025 (i.e., more related than third or fourth cousin). The 2,369,591 SNPs excluded 24 long-range LD regions and were pruned for LD in Plink.[25] SNPs were pruned recursively in a sliding window (window size = 50, number, of SNPs to shift after each step = 5) based on a variance inflation factor (VIF) of 2. SNP calling procedures have been described previously.[9]

SV genotyping

DNA was extracted from whole blood, and paired-end sequencing of genomic DNA was done on the Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA, USA) at the Beijing Genomics Institute (BGI) with medium coverage (13.3x); see reference[4] for more details. SVs were called using a combination of ten different tools based on five different SV detection approaches: gap alignment, split-read mapping, discordant read pair, *de novo* genome assembly and read density. Since the approaches varied in their performance dependent on size range, we divided events into three categories: (i) 1 - 20 bp, (ii) 20 - 100 bp, and (iii) > 100 bp. For each size category, a consensus set was created based on at least two methods supporting (or having discovered) the SV, and SVs had to be present in at least three families and transmitted to at least one offspring. The ten SV calling approaches are described in detail elsewhere.[9]

The combination of different SV calling approaches led to the discovery of 46,633 larger deletions of 20-10,000 bp, which were then genotyped using MATE-CLEVER.[26] This process has been described in more detail elsewhere.[9] Indels (<20 bp) were genotyped using the GATK Unified Genotyper v1.4,[10] GATK HaplotypeCaller, CLEVER,[26] and SOAP denovo.[27] Only indels detected by at least two methods with a perfect allele match were kept. This resulted in a total of 1,733,833 bi-allelic indels that were used for further downstream analyses. The indel genotyping process is described in more detail elsewhere.[9]

Principal Component Analyses (PCAs)

PCAs were run separately on common larger deletions and indels in the

490 unrelated individuals, and PCs were projected onto the rest of the individuals, using EIGENSTRAT.[5] Both indels and larger deletions were excluded from the PCA if they deviated from HWE with a *p*-value < .0001, were missing in more than 5% of individuals, had a MAF < .01, or resided in the inaccessible genome. Larger deletions and indels included in PCAs also excluded 24 long-range LD regions[28] and were pruned for LD recursively in a sliding window (window size = 50, number, of SNPs to shift after each step = 5) based on a variance inflation factor (VIF) of 2. After QC and minimizing LD, a larger deletion set with 15,517 common larger deletions remained, and indel set with 318,492 common indels remained when not filtering for quality, and 241,172 indels with a quality threshold of 2500.

Affymetrix 6.0 SNPs (LD pruned and without long-range LD regions: 129,184 SNPs)[3] were extracted from the GoNL dataset, and used to project the three ancestry-informative *common SNP PCs* from 5,116 unrelated NTR and NESDA subjects onto the 769 GoNL subjects. The *European North-South PC* was PC4 from a PCA on fourteen 1000 Genomes populations computed with the Affymetrix 6.0 SNPs, and was projected from 1000 Genomes individuals (N=1,014; no Dutch included) onto the GoNL individuals. This PC separated Southern from Northern Europeans.[3]

Hudson's F_{st}'s

Hudson's F_{st} values were computed using the EIGENSOFT package[14] between the top 150 and bottom 150 individuals for each ancestry-informative PC. The top and bottom 150 individuals were chosen from the 490 unrelated subjects. F_{st} estimates were computed for all available ancestry informative PCs (three *common SNP PCs*, three indel PCs, and five larger deletion PCs) for three sets of variants separately: 8,685,291 common SNPs, 838,979 common indels (<20 bp), 17,340 common larger deletions (20-10,000 bp). These are the number of variants that remain after initial QC of 20,467,247 SNPs, 1,733,833 indels, and 46,633 larger deletions. The initial QC consisted of excluding variants that deviated from HWE with a *p*-value < .0001, were missing in more than 5% of individuals, had a MAF < .01, or resided in the inaccessible genome.

Correlations

All reported Pearson correlations and Spearman's ρ correlations were calculated using IBM SPSS Statistics 21 in unrelated individuals only.

Appendix Table 1: Correlations between ancestry-informative PCs and average coverage (N=490 unrelated individuals)

Common SNP PCs	Correlation with average coverage	Indel PCs	Correlation with average coverage	Larger deletion PCs	Correlation with average coverage
North-South SNP PC	-.03 (p = .51)	PC1	**-.134 (p = .003)**	PC1	-.08 (p = .08)
East-West SNP PC	-.08 (p = .08)	PC2	**.224 ($p = 1 \times 10^{-6}$)**	PC2	-.024 (p = .602)
Middle-Band SNP PC	**-.116 (p = .01)**	PC3	**.204 ($p = 5 \times 10^{-6}$)**	PC3	-.046 (p = .31)
				PC9	-.045 (p = .32)
				PC10	.011 (p = .81)

Bold: $p < .05$

Appendix Table 2: The three indel PCs that showed significant correlations with geography and/or ancestry-informative SNP PCs. The Table shows Pearson correlations within biobanks and p-values between brackets.

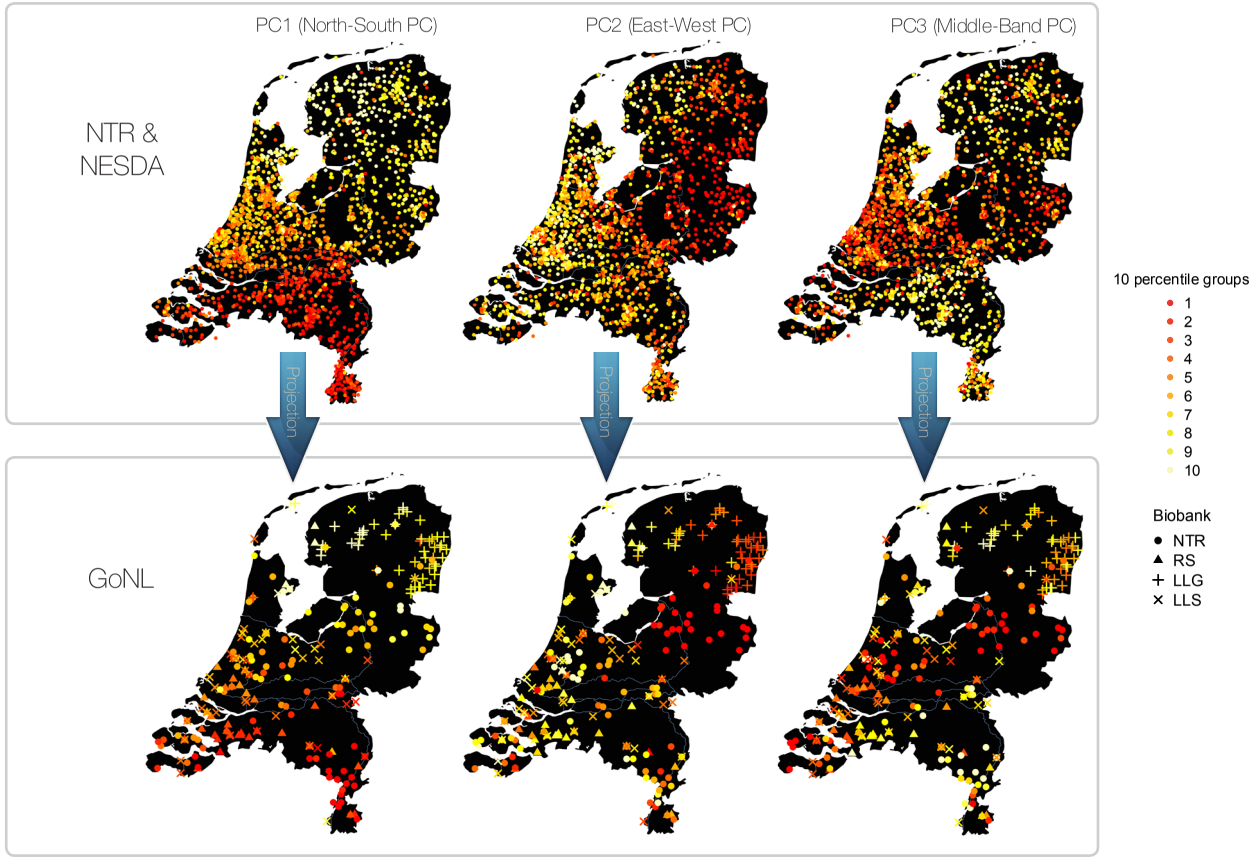
	PC1	PC2	PC3
Netherlands Twin Registry (N=215; 215 with known geographic location)			
Latitude(↕)	**.75 (5.6 × 10^{-17})**	0 (.48)	-.06 (.17)
Longitude (↔)	**.15 (.02)**	**.33 (4.3 × 10^{-7})**	**-.3 (3.7 × 10^{-6})**
North-South SNP PC	**.83 (1.6 × 10^{-12})**	.09 (.09)	-.01 (.42)
East-West SNP PC	.04 (.30)	**.42 (9.1 × 10^{-11})**	**-.4 (5.5 × 10^{-10})**
Middle-Band SNP PC	.09 (.10)	**.26 (6.4 × 10^{-5})**	**-.13 (.03)**
Rotterdam Studies (N=119; 112 with known geographic location)			
Latitude(↕)	**.64 (1.1 × 10^{-14})**	.05 (.28)	.12 (.11)
Longitude (↔)	.02 (.42)	.15 (.06)	-.06 (.27)
North-South SNP PC	**.8 (1.6 × 10^{-12})**	**.23 (.007)**	**.17 (.03)**
East-West SNP PC	-.07 (.22)	**.19 (.02)**	-.08 (.19)
Middle-Band SNP PC	.07 (.23)	**.17 (.03)**	.09 (.15)
LifeLines Groningen (N=107; 104 with known geographic location)			
Latitude(↕)	**.4 (1.1 × 10^{-5})**	**-.51 (2.1 × 10^{-8})**	**.38 (3.6 × 10^{-5})**
Longitude (↔)	**-.51 (2.2 × 10^{-8})**	**.3 (8.5 × 10^{-4})**	**-.5 (2.5 × 10^{-8})**
North-South SNP PC	**.7 (1.6 × 10^{-12})**	-.04 (.33)	**.3 (9.9 × 10^{-4})**
East-West SNP PC	**-.44 (1.4 × 10^{-6})**	**.43 (2 × 10^{-6})**	**-.57 (1.1 × 10^{-10})**
Middle-Band SNP PC	**-.55 (4 × 10^{-10})**	**.23 (.009)**	**-.38 (2.2 × 10^{-5})**
Leiden Longevity Study (N=48; 46 with known geographic location)			
Latitude(↕)	**.4 (.003)**	-.02 (.45)	.2 (.09)
Longitude (↔)	.13 (.20)	-.07 (.33)	-.01 (.46)
North-South SNP PC	**.84 (2.6 × 10^{-14})**	**.26 (.04)**	**.42 (.001)**
East-West SNP PC	-.19 (.10)	.14 (.18)	-.06 (.34)
Middle-Band SNP PC	-.06 (.34)	-.19 (.10)	-.18 (.11)

Bold: $p < .05$; Red: $p < 5 \times 10^{-4}$

Appendix Table 3: The five PCs from larger deletions that showed significant correlations with geography and/or ancestry-informative SNP PCs. The Table shows Pearson correlations within biobanks and p-values between brackets.

	PC1	PC2	PC3	PC9	PC10
Netherlands Twin Registry (N=215; 215with known geographic location)					
Latitude(↕)	**.13 (.03)**	.07 (.17)	**-.24 (1.6×10^{-4})**	-.02 (.36)	**.19 (.003)**
Longitude (↔)	-.01 (.41)	-.02 (.39)	-.06 (.21)	-.1 (.08)	.04 (.26)
North-South SNP PC	**.18 (.004)**	.1 (.08)	**-.22 (6.5×10^{-4})**	-.04 (.27)	**.14 (.02)**
East-West SNP PC	-.05 (.24)	-.02 (.37)	.02 (.37)	**-.2 (.002)**	**.19 (.003)**
Middle-Band SNP PC	.01 (.42)	.11 (.05)	.01 (.41)	**-.16 (.01)**	**.14 (.02)**
Rotterdam Studies (N=119; 112 with known geographic location)					
Latitude(↕)	**.27 (.002)**	**.18 (.03)**	**-.21 (.01)**	-.01 (.45)	-.03 (.39)
Longitude (↔)	.06 (.27)	**.21 (.02)**	.08 (.21)	.08 (.21)	.01 (.47)
North-South SNP PC	**.26 (.002)**	**.29 (7×10^{-4})**	**-.29 (5.9×10^{-4})**	-.06 (.28)	-.02 (.43)
East-West SNP PC	-.02 (.39)	-.09 (.17)	.02 (.43)	-.14 (.06)	.07 (.22)
Middle-Band SNP PC	**.21 (.01)**	**.15 (.05)**	-.09 (.16)	-.15 (.06)	-.05 (.30)
LifeLines Groningen (N=107; 104 with known geographic location)					
Latitude(↕)	**-.21 (.01)**	**.25 (.005)**	.08 (.21)	**.22 (.01)**	.08 (.21)
Longitude (↔)	.04 (.35)	**-.19 (.03)**	.03 (.39)	**-.2 (.02)**	-.06 (.26)
North-South SNP PC	**.26 (.003)**	.04 (.33)	-.01 (.46)	.09 (.17)	.13 (.10)
East-West SNP PC	.1 (.16)	**-.19 (.02)**	.07 (.22)	**-.22 (.01)**	-.02 (.42)
Middle-Band SNP PC	**-.18 (.03)**	-.13 (.08)	-.04 (.35)	-.28 (.002)	-.07 (.23)
Leiden Longevity Study (N=48; 46 with known geographic location)					
Latitude(↕)	.03 (.41)	.21 (.08)	-.14 (.18)	-.14 (.18)	.25 (.05)
Longitude (↔)	.04 (.40)	.09 (.27)	.04 (.39)	.14 (.17)	-.09 (.27)
North-South SNP PC	.04 (.40)	**.31 (.01)**	-.2 (.09)	.04 (.40)	**.28 (.03)**
East-West SNP PC	.11 (.23)	-.07 (.32)	.1 (.25)	.15 (.15)	-.12 (.21)
Middle-Band SNP PC	-.02 (.43)	-.08 (.30)	.1 (.26)	-.05 (.37)	.35 (.007)

Bold: $p < .05$; Red: $p < 5 \times 10^{-4}$



Appendix Figure 1: Three ancestry-informative PCs from Affymetrix 6 data (~500k SNPs on 5,166 NTR and NESDA subjects) projected onto the GoNL individuals. The mean value per postalcode (NTR) or city (GoNL) was calculated, divided into ten percentile groups, and plotted. NTR = Netherlands Twin Registry; RS = Rotterdam Studies; LLG = LifeLines Groningen; LLS = Leiden Longevity Study.

References

1. Montgomery, S.B., Goode, D.L., Kvikstad, E., Albers, C.A., Zhang, Z.D., Mu, X.J., Ananda, G., Howie, B., Karczewski, K.J., et al. (2013). The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Research* 23, 749-761.
2. Consortium, G.P. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-1073.
3. Abdellaoui, A., Hottenga, J.-J., de Knijff, P., Nivard, M.G., Xiao, X., Scheet, P., Brooks, A., Ehli, E.A., Hu, Y., et al. (2013). Population structure, migration, and diversifying selection in the Netherlands. *European Journal of Human Genetics*.
4. Boomsma, D.I., Wijmenga, C., Slagboom, E.P., Swertz, M.A., Karssen, L.C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., et al. (2014). The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics* 22, 221-227.
5. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38, 904-909.
6. Turchin, M.C., Chiang, C.W., Palmer, C.D., Sankararaman, S., Reich, D., Hirschhorn, J.N., and Consortium, G.I.o.A.T. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics* 44, 1015-1019.
7. Frost, P. (2006). European hair and eye color: A case of frequency-dependent sexual selection? *Evolution and Human Behavior* 27, 85-103.
8. Mills, R.E., Pittard, W.S., Mullaney, J.M., Farooq, U., Creasy, T.H., Mahurkar, A.A., Kemeza, D.M., Strassler, D.S., Ponting, C.P., et al. (2011). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Research* 21, 830-839.
9. Francioli, L.C., Menelaou, A., Pulit, S.L., Dijk, F.v., Palamara, P.F., Elbers, C.C., Neerincx, P.B.T., Ye, K., Guryev, V., et al. (2014). Whole-genome sequence variation, population structure and demographic history of the Netherlands. *Nature Genetics* (revision under review).
10. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297-1303.
11. Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832-838.
12. Abdellaoui, A., Hottenga, J.-J., Xiao, X., Scheet, P., Ehli, E.A., Davies, G.E., Hudziak, J.J., Smit, D.J., Bartels, M., et al. (2013). Association between autozygosity and major depression: Stratification due to religious assortment. *Behavior Genetics* 43, 455-467.
13. Holsinger, K.E., and Weir, B.S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics* 10, 639-650.
14. Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research* 23, 1514-1521.
15. Knippenberg, H. (1992). De religieuze kaart van Nederland: omvang en geografische spreiding van de godsdienstige gezindten vanaf de Reformatie tot heden (Uitgeverij Van Gorcum).
16. McEvoy, B.P., Powell, J.E., Goddard, M.E., and Visscher, P.M. (2011). Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Research* 21, 821-829.
17. Li, Y., Zheng, H., Luo, R., Wu, H., Zhu, H., Li, R., Cao, H., Wu, B., Huang, S., et al. (2011). Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nature Biotechnology* 29, 723-730.
18. Stolk, R.P., Rosmalen, J.G., Postma, D.S., de Boer, R.A., Navis, G., Slaets, J.P., Ormel, J., and Wolffenbuttel, B.H. (2008). Universal risk factors for multifactorial diseases. LifeLines: a three-generation population-based study. *European Journal of Epidemiology* 23, 67-74.

19. Schoenmaker, M., de Craen, A.J., de Meijer, P.H., Beekman, M., Blauw, G.J., Slagboom, P.E., and Westendorp, R.G. (2005). Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European Journal of Human Genetics* 14, 79-84.
20. Willemsen, G., Vink, J.M., Abdellaoui, A., den Braber, A., van Beek, J.H., Draisma, H.H., van Dongen, J., van't Ent, D., Geels, L.M., et al. (2013). The Adult Netherlands Twin Register: Twenty-Five Years of Survey and Biological Data Collection. *Twin Research and Human Genetics* 16, 271-281.
21. van Beijsterveldt, C.E., Groen-Blokhuis, M., Hottenga, J.J., Franic, S., Hudziak, J.J., Lamb, D., Huppertz, C., de Zeeuw, E., Nivard, M., et al. (2013). The Young Netherlands Twin Register (YNTR): Longitudinal twin and family studies in over 70,000 children. *Twin Research and Human Genetics* 16, 252.
22. Hofman, A., van Duijn, C.M., Franco, O.H., Ikram, M.A., Janssen, H.L., Klaver, C.C., Kuipers, E.J., Nijsten, T.E., Stricker, B.H.C., et al. (2011). The Rotterdam Study: 2012 objectives and design update. *European Journal of Epidemiology* 26, 657-686.
23. Hofman, A., Murad, S.D., van Duijn, C.M., Franco, O.H., Goedegebure, A., Ikram, M.A., Klaver, C.C., Nijsten, T.E., Peeters, R.P., et al. (2013). The Rotterdam Study: 2014 objectives and design update. *European Journal of Epidemiology* 28, 889-926.
24. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* 88, 76-82.
25. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81, 559-575.
26. Marschall, T., Hajirasouliha, I., and Schönhuth, A. (2013). MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics* 29, 3143-3150.
27. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20, 265-272.
28. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., et al. (2008). Long-range LD can confound genome scans in admixed populations. *American Journal of Human Genetics* 83, 132.

PART IV

RUNS OF HOMOZYGOSITY: HOW ANCESTRAL BEHAVIORS CAN INFLUENCE CURRENT GENETIC VARIATION

CHAPTER 8

ASSOCIATION BETWEEN AUTOZYGOSITY AND MAJOR DEPRESSION: STRATIFICATION DUE TO RELIGIOUS ASSORTMENT

This chapter is based on:

Abdel Abdellaoui, Jouke-Jan Hottenga, Xiangjun Xiao, Paul Scheet, Erik A Ehli, Gareth E Davies, James J Hudziak, Dirk J A Smit, Meike Bartels, Gonneke Willemsen, Andrew Brooks, Patrick F Sullivan, Johannes H Smit, Eco J de Geus, Brenda W J H Penninx, Dorret I Boomsma: Association Between Autozygosity and Major Depression: Stratification Due to Religious Assortment. *Behavior Genetics* 2013; 43:455-467.

Abstract

The effects of inbreeding on the health of offspring can be studied by measuring genome-wide autozygosity as the proportion of the genome in runs of homozygosity (F_{roh}) and relate F_{roh} to outcomes such as psychiatric phenotypes. To successfully conduct these studies, the main patterns of variation for genome-wide autozygosity between and within populations should be well understood and accounted for. Within population variation was investigated in the Dutch population by comparing autozygosity between religious and non-religious groups. The Netherlands have a history of societal segregation and assortment based on religious affiliation, which may have increased parental relatedness within religious groups. Religion has been associated with several psychiatric phenotypes, such as major depressive disorder (MDD). We investigated whether there is an association between autozygosity and MDD, and the extent to which this association can be explained by religious affiliation. All F_{roh} analyses included adjustment for ancestry-informative principal components (PCs) and geographic factors.

Religious affiliation was significantly associated with autozygosity, showing that F_{roh} has the ability to capture within population differences that are not captured by ancestry-informative PCs or geographic factors. The non-religious group had significantly lower F_{roh} values and significantly more MDD cases, leading to a nominally significant negative association between autozygosity and depression. After accounting for religious affiliation, MDD was not associated with F_{roh}, indicating that the relation between MDD and inbreeding was due to stratification.

This study shows how past religious assortment and recent secularization can have genetic consequences in a relatively small country. This warrants accounting for the historical social context and its effects on genetic variation in association studies on psychiatric and other related traits.

Introduction

There is an increasing interest in the association between runs of homozygosity (ROHs) and human disease. Alleles in long ROHs are likely to be identical-by-descent (i.e., autozygous).[1] ROHs can have harmful effects through deleterious recessive alleles that combine when related individuals mate and have offspring. Deleterious recessive alleles are usually rare because selection is unable to completely purge such mutations, since they are not damaging to heterozygote carriers; a process known as mutation-selection balance.[2] The closer or more recent inbreeding is, the longer the ROHs will be, increasing the chances of combining deleterious recessive alleles in offspring.[3] Studies on the association between autozygosity and psychiatric traits may help provide insights into both the evolutionary history and the genetic etiology of complex psychiatric disorders. Strong selection against deleterious variants should results in a bias toward both rarity and recessivity of causal variants, which in turn should increase damaging effects of inbreeding.[4] Recent reports on autozygosity as a schizophrenia risk factor suggest that purifying (negative) selection caused dominant schizophrenia risk alleles to disappear at a faster rate over evolutionary time than recessive risk alleles.[5; 6] This raises the question to which extent there may be an association between autozygosity and other psychiatric disorders. One of the goals of this study was to evaluate the relation between autozygosity and major depressive disorder (MDD), the most prevalent psychiatric disorder in adults.[7; 8] In contrast to schizophrenia, MDD is less heritable (31-42%),[9; 10] and has had limited success in identifying reliably associated genetic variants.[11; 12]

Before studies on autozygosity can be successfully conducted, the main patterns of variation for autozygosity between and within populations should be well understood and accounted for. Inbreeding is a matter of degree,[1] and the total length of ROHs reflects relatively recent inbreeding patterns and varies between worldwide populations.[13; 14] This variation is mainly due to a combination of geographic factors leading to isolation and/or bottlenecks and differences in the prevalence of consanguineous matings.Variation within populations may be driven by assortment on cultural factors and attitudes, such as religious beliefs or political preferences.[15; 16] Such a positive assortative mating strategy has been hypothesized to improve inclusive fitness by increasing genetic relatedness within groups, which can facilitate communication and altruism.[17] Religious affiliation has been reported to facilitate genetic stratification that is detectable by principal component analysis (PCA) on genome-wide single nucleotide polymorphisms (SNPs).[18] We investigate whether religious assortment may have led to increased parental relatedness by testing for systematic autozygosity differences within the Netherlands, which is a relatively small country with a high population density. If autozygosity differences

exist between religious subgroups, this may affect the outcomes of autozygosity studies on traits associated with religion.

This study focuses on variation in autozygosity within a country that had a long and relatively strict segregation in society based on religious affiliation, with religious groups for example having their own political parties, unions, schools, and universities,[19] inducing isolation, decreasing mating options, and potentially increasing parental relatedness. Studies on Dutch marriage records of the 19th and early 20th century show considerable religious assortment.[20-22] Religion is also associated with geography. The Netherlands can be roughly divided into three regions: (1) below the major rivers, where the population is mainly Catholic; (2) the middle-band, which is largely Orthodox-Protestant; (3) the Northern part, which consists largely of more liberal Protestants.[19;23] These three regions also show subtle genetic differences previously detected by PCA on genome-wide SNPs.[24] The geographic distribution of religious groups has been relatively stable for the last four centuries,[19] and has only started to change in the second half of the last century, due to increasing secularization and the influx of immigrants from other parts of the world.[19]

Religion shows significant associations with several traits, such as personality and psychiatric disorders.[25-29] Religion and psychiatry display complex associations, with studies showing both stress-buffering as well as depression-evoking effects of religious involvement.[26; 27; 29-32] However, the majority of the many studies on the relationship between religiosity and psychiatric disease (in different settings, ethnic backgrounds, age groups, and locations) reports a protective influence of religion on psychiatric disorders.[27]

Here, we investigated associations between autozygosity as quantified by ROHs based on genome-wide measured SNPs, major depressive disorder as assessed by DSM4 diagnoses, and self-reported religion to answer whether differences between religious and non-religious groups can lead to a false positive association between ROHs and MDD. Ancestry-informative PCs and living in an urban versus rural environment were accounted for, since ancestry-informative PCs can correlate with homozygosity,[24] and larger cities show a higher prevalence for psychiatric disorders[33-35] and more intermediate values of ancestry-informative PCs due to incoming migration flows.[24]

Methods

Participants

Genotyped subjects took part in the Netherlands Twin Register (NTR) biobank project[36; 37] and in the Netherlands Study of Depression and Anxiety (NESDA).[38] The NTR subjects (N=6,685; 2,678 males and 4,007 females) were randomly sampled from twin families across the Netherlands. The NESDA subjects (N=2,380; 806 males, and 1,574 females) were recruited from the general population, primary care and mental healthcare organizations. Analyses were done on unrelated individuals only. Unrelated individuals were chosen using GCTA,[39] by excluding one of each pair of individuals with an estimated genetic relationship of >0.025 (i.e., more related than third or fourth cousin). There were 4,022 genotyped unrelated individuals with religious affiliation and current living address, and 2,916 genotyped unrelated individuals with religious affiliation, current living address, and MDD case/hyper-control status. Ancestry-informative PCs were computed using 5,166 unrelated subjects and were projected onto the rest of the subjects. In addition to the genotyped subjects, data on religious affiliation were available for 25,450 subjects from the total group of NTR participants. These data were used to display the geographic distribution of religious affiliations and included 3,042 spouse pairs used to test for non-random assortment between spouses on religious affiliation.

Only individuals with Dutch ancestry were included. Individuals with a non-Dutch ancestry were identified by projecting PCs from 1000 Genomes populations on the dataset, and with additional help of the birth country of the parents. This procedure is described in more detail elsewhere.[24]

NTR and NESDA studies were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the US Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180, NESDA 03-183). All subjects provided written informed consent.

Phenotypes

Subjects were included as MDD cases when they had a lifetime diagnosis of DSM-IV MDD as determined by the Composite International Diagnostic Interview (CIDI, version 2.1).[40] The control group consisted of screened hypercontrols who never scored high on a general factor score for anxious depression and never reported a history of MDD in any survey or at the blood sampling visit (either as a complaint for which treatment was sought, reported medication use, or via the CIDI). Further details on the collection and classification of MDD cases and

controls are described elsewhere.[41; 42] There was a total of 1834 MDD cases and 2131 controls.

Religious affiliation was measured in a longitudinal questionnaire studies in the NTR and NESDA with the question "What is your religion?". Three answer categories that overlapped between NTR and NESDA were constructed: (1) none (N=2,114), (2) Roman Catholic (N=1,069), and (3) Protestant (N=839).

For the current living addresses, the postal codes were translated into geographic coordinates (longitude and latitude) for each participant using the open source 6PP database[43] and used to create Figure 2. Population sizes of the cities were obtained from the Central Bureau of Statistics[44] and recoded into a dichotomous variable reflecting whether a subject lives in a city with >100k inhabitants.

Genotyping, quality control (QC), and principal component analysis (PCA)

Methods for blood and buccal swab collection, genomic DNA extraction, and genotyping have been described previously.[24; 45] Genotyping was performed on the Affymetrix Human Genome-Wide SNP 6.0 Array (containing ~906,600 SNP probes) according to the manufacturer's protocol.

Only autosomal SNPs were analyzed. SNPs were removed if they: (1) had probes that mapped badly against NCBI Build 37/UCSC hg19 (i.e., to a "random" region, to >1 region, or to 0 regions); (2) showed a minor allele frequency (MAF) smaller than 5%; (3) had a missing rate greater than 5%; (4) deviated from Hardy-Weinberg equilibrium (HWE) with a *p*-value smaller than 0.001. After QC, 498,592 SNPs remained.

Individuals were removed if they: (1) showed a Contrast QC < .4 (CQC, a quality metric from Affymetrix representing how well allele intensities separate into clusters); (2) fell outside of the main cluster of a PC reflecting a batch effect;[24] (3) had a missing rate greater than 5%; (4) had excess genome-wide heterozygosity / inbreeding levels (F, as calculated in Plink on an LD-pruned set, must be greater than -0.10 and smaller than 0.10); (5) had non-European/non-Dutch ancestry;[24] (6) had genotypes with inconsistencies regarding reported sex or reported relatedness within families.

PCA was conducted using Eigenstrat.[46] The first three PCs correlated significantly with geography: PC1=North-South PC, PC2=East-West PC, PC3=middle-band PC. The procedure for the PCA and the three ancestry-informative PCs are described in detail elsewhere.[24]

ROH calling

ROHs were called in Plink,[47] which was found to optimally predict autozygous stretches in a recent study comparing several software packages designed for this goal.[48] Howrigan et al (2011)[48] used simulated data based on the Affymetrix

6.0 chip, making their density of SNPs in linkage disequilibrium (LD) close to the data analyzed in our study. We followed their recommendations in calling ROHs: (1) SNPs were pruned for LD (window size = 50, number of SNPs to shift after each step = 5, based on a variance inflation factor [VIF] of 2), resulting in 131,325 SNPs; (2) an ROH was defined as ≥65 consecutive homozygous SNPs with no heterozygote calls allowed. To calculate the proportion of the autosome in ROHs, the total length of ROHs were summed for each individual, and then divided by the total SNP-mappable autosomal distance (2.77×10^9 bases), resulting in the F_{roh} measure.

Statistical analyses

Chi-squared tests were performed in IBM SPSS Statistics 19 to test the association between religion and city size, and to test for non-random assortment between spouses on religious affiliation. All other statistical analyses were done in purpose written perl scripts. Regression analyses were computed with the help of the PDL::Stats::GLM perl module (see http://search.cpan.org/~maggiexyz/PDL-Stats-0.6.2/GLM/glm.pp), which allows for the computation of the R^2 change and its significance. The R^2 change in this study is the difference in explained variance of F_{roh} between regressions with and without the predictor of interest (religious affiliation or MDD). All regressions on F_{roh} also included as predictors: three PCs reflecting ancestry, city size (dichotomous), and Contrast QC (CQC, a quality metric from Affymetrix representing how well allele intensities separate into clusters, and known to correlate with heterozygosity).

Religion and autozygosity: R^2 change was computed between multiple regression analyses on F_{roh} with and without religious affiliation as a predictor of F_{roh}. To account for the non-normal distribution of F_{roh} (see Figure 3), F_{roh} was permuted 100,000 times.

Permutation analyses were repeated using the residual of F_{roh} as the dependent variable, and the residual of religious affiliation as the independent variable (i.e., residuals from regressions on the predictors used as control variables: PCs, city size, and CQC). We performed these analyses in order to check whether the association with the control variables caused a bias in the permutation analyses, since only F_{roh} was permuted. Results (data not shown) were similar and significant in the same directions.

Religion and MDD: A logistic regression analysis was run with MDD case/control status as the dependent variable, and as a predictors religion (religious versus non-religious), and city size (small versus large).

MDD and autozygosity: R^2 change was computed between multiple regressions on F_{roh} with and without MDD as a predictor. Analyses were permuted 100,000 times and were repeated adding religion as an additional predictor.

These permutations were also repeated using the residual of F_{roh} and the residual of MDD. The results (data not shown) were similar and significant in the same directions.

ROH mapping analysis: The *--homozygous-group* command in Plink[47] was used to obtain ROH segments that overlap between individuals. This was done separately for analyses on religious groups (resulting in 2,506 ROH segments, in which 2,879 allelically distinct ROHs were observed in ≥2 individuals) and MDD (resulting in 1,858 ROH segments, in which 1,773 allelically distinct ROHs were observed in ≥2 individuals). Table 3 (column "*All subjects"*) and Table 4 (row "*All subjects"*) show the sample sizes for these analyses. Two genome-wide association analyses were conducted: (1) logistic regressions for each ROH segment, where a dichotomous independent variable indicated the presence of an ROH for each individual, which may or may not allelically match; (2) logistic regressions for each allelically distinct ROH, where a dichotomous independent variable indicated the presence of an ROH for each individual. These analyses were done for four dichotomous phenotypes: (1) Catholics versus others, (2) Protestants versus others, (3) non-religious versus others, and (4) MDD cases versus controls. The dependent variable was permuted 1000 times. All regressions were corrected for PCs, city size, and CQC. The analyses for MDD also included religion (dichotomous) as a covariate.

Results

Geographic distribution and assortment on religion

Figure 2 shows that the North-South distribution of Protestants and Catholics in the Netherlands seen in previous generations is maintained today, but with an overall increase (especially in the larger cities displayed in Figure 1) of non-religious groups. When comparing the 26 cities in the Netherlands with a population size >100k with the rest of the Netherlands, the more densely populated cities show significantly fewer religious individuals ($\chi^2(1) = 289.05$, $p < .001$, Cramer's V = .261). The fact that the North-South distribution of Protestants and Catholics is still visible today may have been influenced by strong religious assortment. Spouse pairs (3,042 pairs who had a measure for religious affiliation) showed highly significant religious assortment ($\chi^2(4) = 3226.67$, $p < .001$, Cramer's V = .728; see Table 1). Of the 3,042 spouse pairs, 2,486 (81.72%) reported the same religious affiliation (Protestant, Catholic, or no affiliation).

These results confirm the stability of the religious distribution of Catholics and Protestants. The PC explaining most variation, thus showing the greatest ancestry differences in the Dutch population, shows a high correlation with the North-South

gradient based on current living address (r=.603), and mainly separates the Catholic and Protestant regions of the Netherlands (see Figure 2d). The association between geographic proximity and shared ancestry within the Catholic and Protestant regions is in line with the stable geographic distribution of religions during the last centuries, and indicates that current religious affiliation (Figures 2b and 2c) is likely to correspond to those of a person's ancestors (Figure 2a) for Catholics and Protestants.

Table 1: Distribution of religious affiliation in 3,042 spouse pairs, including χ^2 test for non-random assortment between spouses and the spouse correlation (Cramer's V).

χ^2 (4) = 3226.67, p < .001, Cramer's V = .728		*Wife*		
		Protestant	Catholic	Not Religious
Husband	Protestant	**646 (202.5)**	52 (324.8)	69 (239.8)
	Catholic	42 (312)	**1049 (500.5)**	91 (369.5)
	Not Religious	115 (288.5)	187 (462.8)	**791 (341.7)**

The numbers between brackets is the expected number of spouse pairs in that cell under the null hypothesis of no religious assortment. Observed values higher than the expected values are in bold.

Figure 1: Map of the Netherlands, its major rivers, and its 26 largest municipalities (population size >100k as of April 2012).[44]

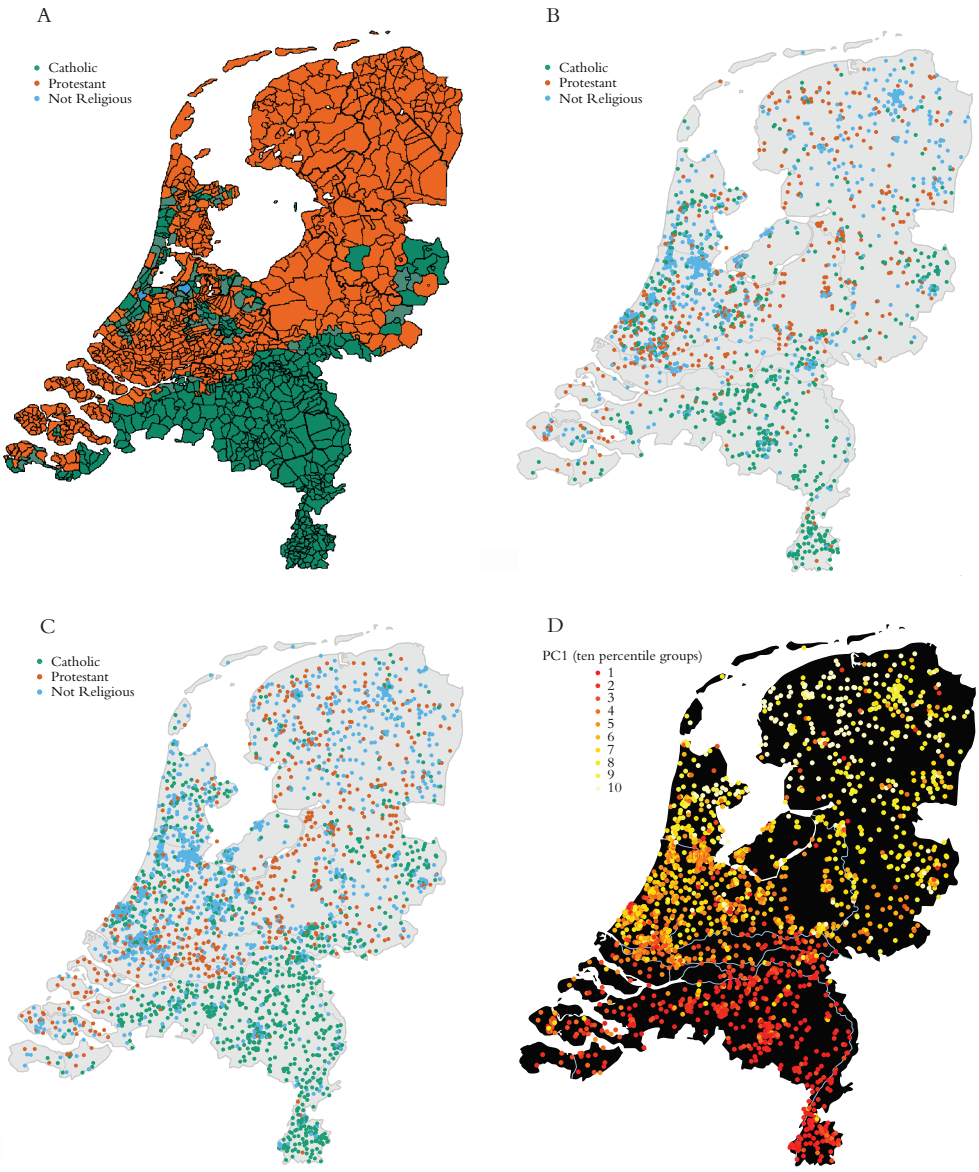


Figure 2: Geographic distribution of religious groups in the Netherlands. ***A*** shows the geographical distribution of Catholic, Protestant and not religious groups in the Netherlands in 1849.[19;49] ***B*** and ***C*** show the distribution of the current genotyped dataset (including related individuals: N=6,464) and the full dataset with a measurement for religious affiliation respectively (N=25,450). Each postal code was given the color of its most prevalent religious group. ***D*** shows the geographic distribution of the North-South PC, where the mean PC value per postal code (current living address) was computed, divided into 10 percentiles, and plotted.

Religion and autozygosity

Data on religious affiliation were available for 4,022 unrelated individuals with Dutch ancestry in three categories: Catholic (N=1,069; mean F_{roh} = .0016), Protestant (N=839; mean F_{roh} = .0019), and non-religious (N=2,114; mean F_{roh} = .0013). Linear regressions with F_{roh} as the dependent variable were performed to test whether the F_{roh} differences between these three groups were significant. Regressions included: ancestry-informative PCs (PC1=North-South PC, PC2=East-West PC, PC3=middle-band PC), city size, and CQC. To account for the non-normal distribution of F_{roh} (see Figure 3), F_{roh} was permuted 100,000 times.

Without religion as a predictor, only the North-South PC, East-West PC, and city size contributed significantly to F_{roh} variation (see Appendix 1), with the most significant contribution coming from the North-South PC, where the southern part has fewer/shorter ROHs. Including religion in the regression analysis led to a significant increase in explained variance of F_{roh}. Post- hoc analyses showed that this was due to the significant difference between the non-religious group and the two religious groups, since there was no significant difference between the two religious groups (see Table 2, Appendix 1, and Figure 3). To further investigate whether this effect is due to closer inbreeding, we removed subjects with more extreme F_{roh} values and repeated the analyses. When removing the 2.5% tail ($Z_{Froh} > 1.96$, N=103, of which 66% is religious, compared to 47% of the remaining individuals), the effect remains the same, and significant. Removing outliers with an approximate equivalent of half-cousin inbreeding ($F_{roh} > .03125$) or even distant inbreeding ($F_{roh} > .005$) also results in a significant effect (Table 2 and Appendix 1). This suggests that these effects are not due to recent or close inbreeding. The highest observed F_{roh} value was .0583, so there were no individuals with values higher than or equivalent to cousin-cousin inbreeding ($F_{roh} > .0625$).

MDD and autozygosity

Being religious was protective for MDD, while living in a larger city increased the chance for MDD: a logistic regression with MDD case/control status as the dependent variable, religion as a dichotomous (religious versus non-religious) predictor ($\beta = -.84$, $p < 10^{-16}$), and city size as a dichotomous predictor ($\beta = .62$, $p = 5.71 \times 10^{-15}$) showed a significant negative relation between MDD and religion and a significant positive relation between MDD and city size.

MDD controls showed higher F_{roh} values than MDD cases for each F_{roh} cutoff (see Table 4). Linear regressions with F_{roh} as the dependent variable were performed that included the three ancestry PCs, city size, and CQC as predictors. We did the analysis for the entire dataset (2,916 unrelated individuals with MDD and religion measured), and repeated it for the datasets excluding positive outliers (i.e., excluding

closer inbreeding). MDD showed a nominally significant association with F_{roh} for the dataset excluding the 2,5% F_{roh} tail (significantly more variance is explained with than without MDD as an independent variable: empirical $p = .042$). When repeating this analysis with religion added as a predictor, the effect disappears (empirical $p = .236$), while religion remains highly significant ($p = 2.09 \times 10^{-5}$; see Table 5 and Appendix 2).

Table 2: Results of 100k Permutations of multiple linear regression on $F_{roh,}$ including the three Dutch PCs, whether the subject lives in a city with >100k residents and CQC as predictors.

	Included in model	F (df)	R^2 change	*p*-value	Empirical *p*
All available individuals (N=4,022)					
Main test	Two religious groups and non religious group (as two dummy variables)	9.78 (2 4014)	.005	**5.80×10^{-5}**	**7×10^{-5}**
Post-hoc tests	Not religious vs. Protestant	10.27 (1 2946)	.003	**1.36×10^{-3}**	**1×10^{-3}**
	Not religious vs. Catholic	17.08 (1 3176)	.005	**3.67×10^{-5}**	**8×10^{-5}**
	Catholic vs. Protestant	.088 (1 1901)	4.57×10^{-5}	.766	.770
Individuals with $Z_{Froh} < 1.96$ (i.e., excluding the 2.5% tail; N=3,923)					
Main test	Two religious groups and non religious group (as two dummy variables)	9.99 (2 3915)	.005	**4.70×10^{-5}**	**3×10^{-5}**
Post-hoc tests	Not religious vs. Protestant	8.62 (1 2880)	.003	**3.35×10^{-3}**	**3.18×10^{-3}**
	Not religious vs. Catholic	14.38 (1 3110)	.004	**1.52×10^{-4}**	**1.8×10^{-4}**
	Catholic vs. Protestant	.001 (1 1835)	5.75×10^{-7}	.973	.973
Individuals with $F_{roh} < .03125$ (i.e., excluding half-cousin inbreeding; N=4,017)					
Main test	Two religious groups and non religious group (as two dummy variables)	13.08 (2 4009)	.006	**2.17×10^{-6}**	**$<10^{-5}$**
Post-hoc tests	Not religious vs. Protestant	14.49 (1 2943)	.005	**1.44×10^{-4}**	**1.1×10^{-4}**
	Not religious vs. Catholic	20.23 (1 3172)	.006	**7.12×10^{-6}**	**$<10^{-5}$**
	Catholic vs. Protestant	.065 (1 1898)	3.38×10^{-5}	.798	.797
Individuals with $F_{roh} < .005$ (i.e., excluding elevated levels of distant inbreeding; N=3,855)					
Main test	Two religious groups and non religious group (as two dummy variables)	12.04 (2 3847)	.006	**6.11×10^{-6}**	**$<10^{-5}$**
Post-hoc tests	Not religious vs. Protestant	13.59 (1 2835)	.005	**2.31×10^{-4}**	**2.4×10^{-4}**
	Not religious vs. Catholic	14.94 (1 3056)	.005	**1.13×10^{-4}**	**1.3×10^{-4}**
	Catholic vs. Protestant	.580 (1 1798)	3.08×10^{-4}	.446	.445

Significance threshold for post-hoc tests = .05/3 = 0.0167. Bold = significant.

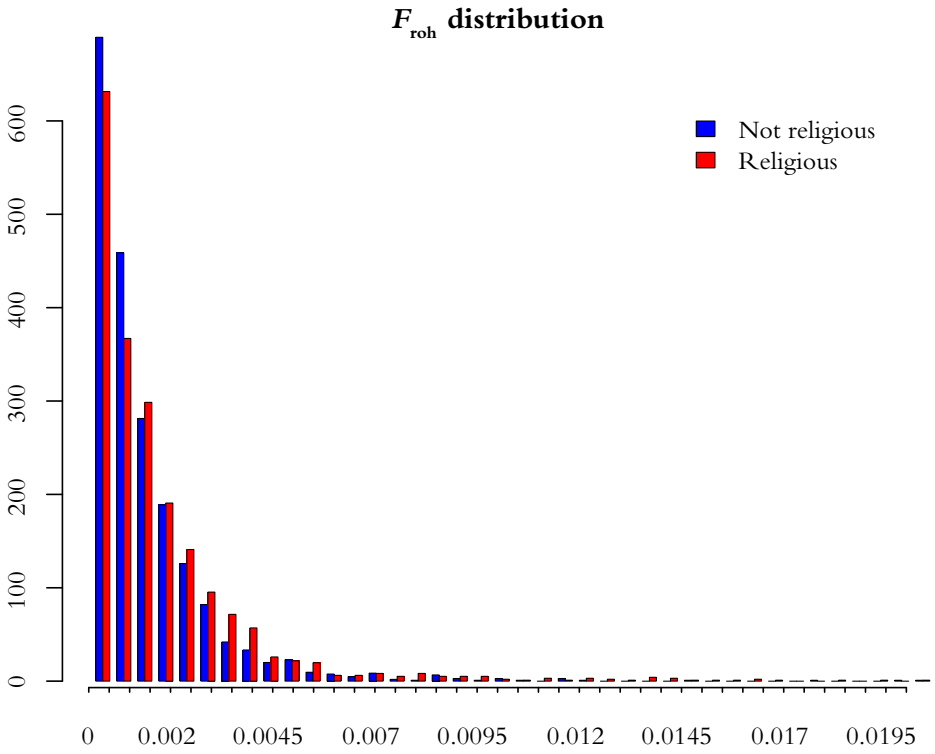


Figure 3: Density plot of the F_{roh} distribution for religious and non-religious individuals. The bars have been scaled to have a total density of 1 for both groups in order to allow a comparison between the groups.

Table 3: Sample size and mean F_{roh} for each religious group per additional F_{roh} cutoff

	All subjects		$Z_{Froh} < 1.96$		$F_{roh} < .03125$		$F_{roh} < .005$	
Religious group	N	Mean F_{roh} (SD)	N	Mean F_{roh} (SD)	N	*Mean* F_{roh} (SD)	N	*Mean* F_{roh} (SD)
Catholic	1,069	.00159 (.0031)	1,036	.00119 (.0013)	1,067	.00151 (.0024)	1,013	.00109 (.0011)
Protestant	839	.00186 (.0031)	806	.00138 (.0013)	838	.00180 (.0026)	792	.00131 (.0012)
Not Religious	2,114	.00133 (.0024)	2,081	.00112 (.0012)	2,112	.00128 (.0020)	2,050	.00105 (.0011)

Table 4: Sample size and mean F_{roh} for MDD cases and controls per additional F_{roh} cutoff

	Cases		Controls	
	N	Mean F_{roh} (SD)	N	Mean F_{roh} (SD)
All subjects (N = 2,916)	1,608	.00143 (.0025)	1,308	.00156 (.0024)
Z_{Froh} < 1.96 (N = 2,835)	1,570	.00116 (.0012)	1,265	.00124 (.0013)
Froh < .03125 (N = 2,914)	1,607	.00140 (.0022)	1,307	.00154 (.0023)
Froh < .005 (N = 2,787)	1,549	.00110 (.0011)	1,238	.00114 (.0011)

ROH mapping analysis

In order to test whether specific regions of the genome were overrepresented in one of the religious groups, or in MDD cases/controls, genome-wide analyses were conducted on ROHs. Analyses were done on each ROH segment that showed overlap between individuals. Each segment was tested in genome-wide analyses with the following dichotomous dependent variables: (1) Catholics versus others, (2) Protestants versus others, (3) non-religious versus others, and (4) MDD cases versus controls (see Tables 3-4 for sample sizes). Association analyses were run on (1) the overall burden of ROHs at each location throughout the genome, and (2) on each allelically distinct ROH. All regressions included the three ancestry PCs, city size, and CQC as covariates, and the analyses for MDD also included religion as a covariate. There were no significant results (all empirical p's > .05), indicating no association between specific genomic regions and religion or MDD.

Table 5: R^2 change, and its significance, between multiple regressions on F_{roh} with and without MDD as independent variable (once with, and once without religion as a covariate). F_{roh} was permuted 100,000 ×.

		F (df)	R^2 change	*p*	Empirical *p*
All individuals	No religion included	1.65 (1, 2909)	.0006	.199	.200
N = 2,916	Religion included	.07 (1 2908)	2.4×10^{-5}	.787	.788
Z_{Froh} < 1.96	No religion included	4.14 (1, 2828)	.0014	**.042**	**.042**
N = 2,835	Religion included	1.41 (1 2827)	.0005	.235	.236
F_{roh} < .03125	No religion included	2.84 (1, 2907)	.0010	.092	.093
N = 2,914	Religion included	.46 (1 2906)	.0002	.497	.496
F_{roh} < .005	No religion included	2.51 (1, 2780)	.0009	.113	.113
N = 2,787	Religion included	.71 (1 2779)	.0002	.400	.399

Discussion

We observed significant autozygosity differences between religious and non-religious groups, due to by greater levels of background parental relatedness in the religious groups and a more outbred non-religious group. Individuals with the same religious affiliation are more likely to share ancestry due to a combination of geographic proximity and the strong assortment and segregation of religious groups throughout the past centuries. The non-religious group showed a significantly lower ROH burden, which could not be explained by the fact that these groups are more prevalent in urban areas that attract people with a greater variation in genetic background. The decrease of ROHs in non-religious groups is likely caused by increasing variation in the gene pool of possible mates, induced by the relatively recent absence of denominational restrictions on mate selection.

In the non-religious groups there were significantly more MDD cases. Religion has often been reported to be an important coping mechanism with a positive influence on mental health.[27] Numerous studies have been conducted on the relation between religion and depressive disorder, of which the majority reported religious people to have significantly lower rates of depressive disorder or fewer depressive symptoms.[27] The association between religion and MDD led to a nominally significant negative association between autozygosity and MDD. Without knowledge of the relationship between religion and autozygosity, and between religion and MDD, the significant association may have led to the hypothesis that inbreeding protects against MDD (i.e., that recessive alleles unmasked by ROH are protective and an important part of the genetic architecture of MDD). This could be interpreted as a result of selection against the inability to develop MDD. This would be in line with the analytical rumination hypothesis which poses that MDD is an evolved response to complex problems, which functions by focusing the limited cognitive processing recourses on the analysis of the problems in the individuals life.[50] However, after including religion as a covariate, MDD did not show a significant association with ROHs, while religion still contributed significantly to F_{roh} variation. This suggests that the relation between autozygosity and MDD is a consequence of a stratification artifact.

The North-South PC contributed most significantly to F_{roh} variation in all regressions. This was no surprise, as the Dutch North-South cline significantly correlates with genome-wide homozygosity (F), making the stratification captured by the North-South PC also well detectable by F_{roh}. This correlation is likely due to a serial founder effect that is also visible in the European North-South cline that correlates highly (r = .66) with the Dutch North-South cline.[24] Unlike F_{roh}, F does not require a minimum amount of consecutive homozygous SNPs, and thus is able to capture much shorter (hence older) ROHs. Given that the serial founder effect

captured by the North-South PC is expected to have occurred in more ancient times (during the European South-North expansions), we would expect a higher correlation between PC1 and F than between PC1 and F_{roh}, which is indeed what we observe ($r_{PC1,F} = .245, p < .001$; $r_{PC1,Froh} = .082, p < .001$).

Genome-wide ROH mapping analyses did not reveal any genomic regions that were significantly more homozygous in any religious group, or in MDD cases or controls. This was not unexpected, as there was no reason to assume that past religious assortment was based on specific genomic regions, and there was no association between F_{roh} and MDD.

A larger meta-analysis also found no significant relationship between autozygosity (F_{roh}/ROHs) and MDD (9,238 MDD cases and 9,521 controls, including a subsample of the current study genotyped on a different microarray), and observed inconsistent directions of effect between the datasets.[51] The study included nine datasets from 5 countries (1 UK, 2 Australian, 3 German, 2 US, and 1 Dutch dataset), and the direction of the effects were consistent across countries (higher F_{roh} protective for MDD: Australia, Netherlands, and US; lower F_{roh} protective for MDD: UK and Germany). This suggests that there may be similar demographic/social factors associated with both autozygosity and MDD in other populations as well. In-depth analyses similar to those in the present study are needed to further explore this hypothesis in other populations.

Differences in autozygosity between social groups can unveil additional dimensions of stratification within populations. It shows in the Netherlands how recent secularization can have genetic consequences in a relatively small country. This can confound studies on traits associated with the stratifying factor, even when correcting for PCs reflecting ancestry and geographic factors. To detect effect sizes seen from known inbreeding studies in outbred populations, larger sample sizes than that of the current dataset are needed (~12,000-65,000).[1;5] The effects of within population differences as observed in the current study however are well detectable with the current sample size. It is crucial to the investigation of the effects of inbreeding on the well-being of offspring to take social factors into account that are predictive for both parental relatedness and the trait under investigation. Considering the sensitivity of F_{roh} for social stratification not detected by PCAs, we also encourage to consider accounting for these effects in GWASs.

Appendix 1: Regression coefficients (p-values between brackets) for each of the predictors included in the linear regressions.

	Religious Affiliation	Religious Affiliation	PC1 (North-South)	PC2 (East-West)	PC3 (Middle-Band)	City Size	CQC
All available individuals (N=4,022)							
No religion included	NA	NA	.0211 (3.96×10^{-11})	.0067(.037)	.0060(.060)	-.0003 (1.77×10^{-4})	.0001 (.102)
Two dummy variables	.0004 (.002)	.0004 (1.09×10^{-4})	.0230 (1.09×10^{-11})	.0071 (.025)	.0063 (.052)	-.0002 (.014)	.0001 (.093)
Not religious vs. Protestant	NA	.0004 (.001)	.0261(8.56×10^{-11})	-.0010 (.790)	.0107 (.005)	-.0002(.106)	8.15×10^{-5} (.372)
Not religious vs. Catholic	NA	.0004 (3.68×10^{-5})	.0235 (3.79×10^{-11})	.0114 (.001)	.0016 (.657)	-.0002 (.055)	.0002 (.019)
Catholic vs. Protestant	NA	-4.76×10^{-5} (.766)	.0207 (1.57×10^{-4})	.0120 (.014)	.0081 (.109)	-.0004 (.011)	.0001 (.366)
Individuals with $Z_{Froh} < 1.96$ (i.e., excluding the 2.5% tail; N=3,923)							
No religion included	NA	NA	.0178 ($< 10^{-15}$)	.0007 (.644)	-.0030 (.039)	-.0002 (7.65×10^{-5})	.0001 (.001)
Two dummy variables	.0002 (.002)	.0002 (7.63×10^{-5})	.0187 ($< 10^{-15}$)	.0009 (.523)	-.0029 (.052)	-.0001 (.008)	.0001 (.0009)
Not religious vs. Protestant	NA	.0002 (.003)	.0228 ($< 10^{-15}$)	-.0005 (.756)	.0009 (.604)	-4.44×10^{-5} (.341)	7.93×10^{-5} (.067)
Not religious vs. Catholic	NA	.0002 (1.52×10^{-4})	.0175 ($< 10^{-15}$)	.0023 (.175)	-.0033 (.046)	-.0001 (.010)	.0002 (7.39×10^{-5})
Catholic vs. Protestant	NA	2.25×10^{-6} (.974)	.0168 (7×10^{-13})	.0010 (.646)	-.0059 (.0059)	-.0002 (.0017)	.0001 (.019)
Individuals with $F_{roh} < .03125$ (i.e., excluding half-cousin inbreeding; N=4,017)							
No religion included	NA	NA	.0205 (2.89×10^{-15})	.0018 (.480)	.0039 (.134)	-.0003 (2.50×10^{-4})	.0001 (.065)
Two dummy variables	.0004 (7.49×10^{-5})	.0004 (2.28×10^{-5})	.0219 (1.78×10^{-15})	.0023 (.382)	.0040 (.127)	-.0002 (.029)	.0001 (.056)
Not religious vs. Protestant	NA	.0004 (1.44×10^{-4})	.0261 (6.00×10^{-15})	.0005 (.875)	.0090 (.004)	-.0001 (.123)	8.95×10^{-5} (.238)
Not religious vs. Catholic	NA	.0004 (7.12×10^{-6})	.0218 (2.75×10^{-14})	.0046 (.110)	-9.74×10^{-5} (.973)	-.0001 (.109)	.0002 (.011)
Catholic vs. Protestant	NA	3.30×10^{-5} (.798)	.01855 (2.54×10^{-5})	.0014 (.722)	.0048 (.240)	-.0003 (.026)	9.85×10^{-5} (.365)
Individuals with $F_{roh} < .005$ (i.e., excluding elevated levels of distant inbreeding; N=3,855)							
No religion included	NA	NA	.0161 ($< 10^{-15}$)	.0018 (.182)	-.0024 (.069)	-8.03×10^{-5} (.022)	9.22×10^{-5} (.006)
Two dummy variables	.0002 (7.81×10^{-5})	.0002 (9.61×10^{-5})	.0167 ($< 10^{-15}$)	.0020 (.136)	-.0024 (.075)	-3.32×10^{-5} (.361)	9.32×10^{-5} (.005)
Not religious vs. Protestant	NA	.0002 (2.31×10^{-4})	.0202 ($< 10^{-15}$)	.0011 (.492)	.0006 (.708)	5.85×10^{-6} (.889)	5.75×10^{-5} (.137)
Not religious vs. Catholic	NA	.0002 (1.13×10^{-4})	.0156 ($< 10^{-15}$)	.0029 (.051)	-.0021 (.159)	-2.41×10^{-5} (.542)	.0001 (.002)
Catholic vs. Protestant	NA	4.60×10^{-5} (.446)	.0149 (9.54×10^{-13})	.0018 (.353)	-.0056 (.004)	-.0001 (.048)	.0001 (.025)

Appendix 2: Regression coefficients (p-values between brackets) for each of the predictors included in the linear regressions of the MDD analyses. These results are from analyses done on individuals with $Z_{Froh} < 1.96$.

	Religious Affiliation (dichotomous)	MDD	PC1 (North-South)	PC2 (East-West)	PC3 (Middle-Band)	City Size	CQC
No Religion & No MDD	NA	NA	.0156 ($<10^{-15}$)	.0038 (.030)	-.0053 (.002)	-.0001 (.003)	.0001 (.006)
No Religion & MDD	NA	-.0001 (.042)	.0159 ($<10^{-15}$)	.0038 (.031)	-.0052 (.003)	-.0001 (.013)	.0001 (.004)
Religion & No MDD	.0002 (4.99×10^{-6})	NA	.0162 ($<10^{-15}$)	.0042 (.017)	-.0051 (.003)	-7.81×10^{-5} (.097)	.0001 (.005)
Religion & MDD	.0002 (2.09×10^{-5})	-5.66×10^{-5} (.235)	.0163 ($<10^{-15}$)	.0041 (.018)	-.0050 (.003)	-7.00×10^{-5} (.141)	.0001 (.004)

References

1. Keller, M.C., Visscher, P.M., and Goddard, M.E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189, 237-249.
2. Charlesworth, D., and Willis, J.H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics* 10, 783-796.
3. Szpiech, Z.A., Xu, J., Pemberton, T.J., Peng, W., Zöllner, S., Rosenberg, N.A., and Li, J.Z. (2013). Long Runs of Homozygosity Are Enriched for Deleterious Variation. *The American Journal of Human Genetics*.
4. DeRose, M.A., and Roff, D.A. (1999). A comparison of inbreeding depression in life-history and morphological traits in animals. *Evolution*, 1288-1292.
5. Keller, M.C., Simonson, M.A., Ripke, S., Neale, B.M., Gejman, P.V., Howrigan, D.P., Lee, S.H., Lencz, T., Levinson, D.F., et al. (2012). Runs of Homozygosity Implicate Autozygosity as a Schizophrenia Risk Factor. *PLoS Genetics* 8, e1002656.
6. Lencz, T., Lambert, C., DeRosse, P., Burdick, K.E., Morgan, T.V., Kane, J.M., Kucherlapati, R., and Malhotra, A.K. (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 104, 19942-19947.
7. Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., and Walters, E.E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 62, 593.
8. Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.U., and Kendler, K.S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. *Archives of general psychiatry* 51, 8.
9. Sullivan, P.F., Neale, M.C., and Kendler, K.S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry* 157, 1552-1562.
10. Lubke, G.H., Hottenga, J.J., Walters, R., Laurin, C., de Geus, E.J., Willemsen, G., Smit, J.H., Middeldorp, C.M., Penninx, B.W., et al. (2012). Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biological Psychiatry* 72, 707-709.
11. Ripke, S., Lewis, C.M., Lin, D.Y.U., and Wray, N. (2012). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry* 18, 497-511.
12. Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics* 13, 537-551.
13. McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., et al. (2008). Runs of homozygosity in European populations. *American Journal of Human Genetics* 83, 359-372.
14. Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, J.Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics* 91, 275-292.
15. Alford, J.R., Hatemi, P.K., Hibbing, J.R., Martin, N.G., and Eaves, L.J. (2011). The politics of mate choice. *The Journal of Politics* 73, 362-379.
16. Watson, D., Klohnen, E.C., Casillas, A., Nus Simms, E., Haig, J., and Berry, D.S. (2004). Match makers and deal breakers: Analyses of assortative mating in newlywed couples. *Journal of Personality* 72, 1029-1068.
17. Thiessen, D., and Gregg, B. (1980). Human assortative mating and genetic equilibrium: An evolutionary perspective. *Ethology and Sociobiology* 1, 111-140.
18. Haber, M., Gauguier, D., Youhanna, S., Patterson, N., Moorjani, P., Botigué, L.R., Platt, D.E., Matisoo-Smith, E., Soria-Hernanz, D.F., et al. (2013). Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture. *PloS Genetics* 9, e1003316.
19. Knippenberg, H. (1992). De religieuze kaart van Nederland: omvang en geografische spreiding

van de godsdienstige gezindten vanaf de Reformatie tot heden (Uitgeverij Van Gorcum).
20. Beekink, E., Liefbroer, A.C., and van Poppel, F. (1998). Changes in choice of spouse as an indicator of a society in a state of transition: Woerden, 1830-1930. *Historical Social Research/Historische Sozialforschung*, 231-253.
21. Kok, J. (1998). Vrijt daar je zijt": huwelijk en partnerkeuze in Zeeland tussen 1830 en 1950. *K Mandemakers, O Hoogerhuis en A de Klerk (red), Over Zeeuwse mensen Demografische en sociale ontwikkelingen in Zeeland in de negentiende en begin twintigste eeuw Themanummer Zeeland* 7, 131-143.
22. Polman, A. (1951). Geografische en confessionele invloeden bij de huwelijkskeuze in Nederland (Stenfert Kroese).
23. Donk, W.B.H.J., Jonkers, A., Kronjee, G., and Plum, R. (2006). Geloven in het publieke domein: verkenningen van een dubbele transformatie (Amsterdam University Press).
24. Abdellaoui, A., Hottenga, J.J., de Knijff, P., Nivard, M.G., Xiao, X., Scheet, P., Brooks, A., Ehli, E.A., Hu, Y., et al. (2013). Population Structure, Migration, and Diversifying Selection in the Netherlands. *European Journal of Human Genetics* 21, 1277-1285.
25. Boomsma, D.I., de Geus, E.J.C., van Baal, G.C.M., and Koopmans, J.R. (1999). A religious upbringing reduces the influence of genetic factors on disinhibition: Evidence for interaction between genotype and environment on personality. *Twin Research and Human Genetics* 2, 115-125.
26. King, M., Marston, L., McManus, S., Brugha, T., Meltzer, H., and Bebbington, P. (2012). Religion, spirituality and mental health: results from a national study of English households. *The British Journal of Psychiatry*.
27. Koenig, H.G. (2009). Research on religion, spirituality, and mental health: A review. *Canadian Journal of Psychiatry* 54, 283-291.
28. Willemsen, G., and Boomsma, D.I. (2007). Religious upbringing and neuroticism in Dutch twin families. *Twin Research and Human Genetics* 10, 327-333.
29. Miller, L., Wickramaratne, P., Gameroff, M.J., Sage, M., Tenke, C.E., and Weissman, M.M. (2012). Religiosity and major depression in adults at high risk: a ten-year prospective study. *American Journal of Psychiatry* 169, 89-94.
30. Braam, A.W., Hein, E., Deeg, D.J.H., Twisk, J.W.R., Beekman, A.T.F., and van Tilburg, W. (2004). Religious involvement and 6-year course of depressive symptoms in older Dutch citizens: results from the Longitudinal Aging Study Amsterdam. *Journal of Aging and Health* 16, 467-489.
31. Dein, S., Cook, C.C.H., Powell, A., and Eagger, S. (2010). Religion, spirituality and mental health. *The Psychiatrist* 34, 63-64.
32. Seybold, K.S., and Hill, P.C. (2001). The role of religion and spirituality in mental and physical health. *Current Directions in Psychological Science* 10, 21-24.
33. Marcelis, M., Navarro-Mateu, F., Murray, R., Selten, J.-P., and Van Os, J. (1998). Urbanization and psychosis: a study of 1942–1978 birth cohorts in The Netherlands. *Psychological medicine* 28, 871-879.
34. Sundquist, K., Frank, G., and Sundquist, J. (2004). Urbanisation and incidence of psychosis and depression Follow-up study of 4.4 million women and men in Sweden. *The British Journal of Psychiatry* 184, 293-298.
35. Van Os, J. (2004). Does the urban environment cause psychosis? *The British Journal of Psychiatry* 184, 287-288.
36. Boomsma, D.I., de Geus, E.J.C., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.J., Posthuma, D., van Beijsterveldt, T.C.E.M., Hudziak, J.J., et al. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics* 9, 849-857.
37. Willemsen, G., de Geus, E.J.C., Bartels, M., van Beijsterveldt, C.E.M.T., Brooks, A.I., Estourgie-van Burk, G.F., Fugman, D.A., Hoekstra, C., Hottenga, J.J., et al. (2010). The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Research and Human Genetics* 13, 231.
38. Penninx, B.W.J.H., Beekman, A.T.F., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W.J., et al. (2008). The Netherlands Study of Depression

and Anxiety (NESDA): rationale, objectives and methods. *International journal of methods in psychiatric research* 17, 121-140.
39. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* 88, 76.
40. WHO. (1997). World Health Organisation: Composite International Diagnostic Interview Version 2.1 (Geneva: World Health Organization).
41. Boomsma, D.I., Willemsen, G., Sullivan, P.F., Heutink, P., Meijer, P., Sondervan, D., Kluft, C., Smit, G., Nolen, W.A., et al. (2008). Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *European Journal of Human Genetics* 16, 335-342.
42. Wray, N., Pergadia, M., Blackwood, D., Penninx, B., Gordon, S., Nyholt, D., Ripke, S., Macintyre, D., McGhee, K., et al. (2010). Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Molecular Psychiatry* 17, 36-48.
43. Broek, K.v.d. (2012). 6PP database: http://www.d-centralize.nl/projects/6pp/downloads/.
44. CBS. (2012). Centraal Bureau voor de Statistiek, Bevolkingsontwikkeling; regio per maand, April 2012.
45. Scheet, P., Ehli, E.A., Xiao, X., van Beijsterveldt, C.E., Abdellaoui, A., Althoff, R.R., Hottenga, J.J., Willemsen, G., Nelson, K.A., et al. (2012). Twins, Tissue, and Time: An Assessment of SNPs and CNVs. *Twin Research and Human Genetics* 15, 737.
46. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38, 904-909.
47. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81, 559-575.
48. Howrigan, D.P., Simonson, M.A., and Keller, M.C. (2011). Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics* 12.
49. Dimitri. (2007). Kaart van Nederland met godsdienstverhoudingen per gemeente bij de volkstelling van 1849. (date retrieved: January 15, 2013) (Wikipedia, The Free Encyclopedia).
50. Andrews, P.W., and Thomson Jr, J.A. (2009). The bright side of being blue: depression as an adaptation for analyzing complex problems. *Psychological Review* 116, 620.
51. Power, R.A., Keller, M.C., Ripke, S., Abdellaoui, A., Wray, N.R., Sullivan, P.F., Group, M.P.W., and Breen, G. (2013). A Recessive Genetic Model and Runs of Homozygosity in Major Depressive Disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 165, 157-166.

CHAPTER 9

EDUCATIONAL ATTAINMENT INFLUENCES GENETIC VARIATION THROUGH MIGRATION AND ASSORTMENT

This chapter is based on:

Abdel Abdellaoui, Jouke-Jan Hottenga, Gonneke Willemsen, Meike Bartels, Toos van Beijsterveldt, Erik A. Ehli, Gareth E. Davies, Andrew Brooks, Patrick F. Sullivan, Brenda W.J.H. Penninx, Eco J. de Geus, Dorret I. Boomsma: Educational Attainment Influences Genetic Variation through Migration and Assortative Mating. *PLOS ONE* (submitted).

Abstract

Higher educated individuals are more likely to migrate, increasing the chance of meeting a spouse with a different ancestral background. In this context, the presence of strong educational assortment can result in greater ancestry differences within higher educated spouse pairs, while lower educated individuals are more likely to mate with someone with whom they share more ancestry.

We examined the association between educational attainment and F_{roh} (= the proportion of the genome consisting of runs of homozygosity [ROHs]) in ~2,000 subjects of Dutch ancestry. The subjects' own educational attainment showed a nominally significant negative association with F_{roh} (p = .045), while the contribution of parental education to offspring F_{roh} was highly significant (father: $p < 10^{-5}$; mother: $p = 9\times10^{-5}$), with higher educated parents having offspring with fewer ROHs. This association disappears after correcting for the distance between parental birthplaces, which itself was also significantly associated with F_{roh} ($p = 9 \times 10^{-5}$). Ancestry-informative principal components from the offspring showed a significantly decreasing association with geography as parental education increased, consistent with the significantly higher migration rates among higher educated parents. Parental education also showed a high spouse correlation (Spearman's ρ= .66, $p = 3 \times 10^{-262}$).

We show that less educated parents are less likely to mate with the more mobile higher educated parents, creating systematic differences in homozygosity. Understanding how behaviors influence the genomic structure of a population is highly valuable for studies on the genetic etiology of behavioral, cognitive, and social traits.

Introduction

Non-random mating can create systematic differences in parental relatedness, which can have a direct and detectable impact on genome-wide homozygosity in subsequent generations. Non-random mating in human populations can be driven by heritable behavioral traits. It is important to understand how behavior has influenced our genetic variation in order to successfully conduct and interpret studies that aim to understand the reverse, namely how genetic variation influences behavior. In the Netherlands for example, the consequences of continuous religious assortment during the last ~400 years and the relatively recent secularization are detectable through homozygosity differences between religious and non-religious groups. Such systematic differences can cause spurious associations between homozygosity and traits related to religiosity.[1]

Educational attainment (EA) is another complex trait that may induce systematic differences in parental relatedness. Education shows considerable levels of assortment.[2-5] Higher educated individuals are also more likely to have moved away from their birthplace, making the physical distance between them and their family members two to three times greater than for individuals with a lower education.[6] When ancestry shows high correlations with geography, like in the Netherlands,[7; 8] these behaviors may increase the chance for higher educated individuals to mate with someone with a different ancestral background, making their offspring more outbred, while lower educated spouse pairs are more likely to share more ancestry.

EA and its etiology have been widely studied. EA is heritable in populations in which it has been studied, with estimates ranging from ~20% to ~80% and increasing over time.[9; 10] EA is associated with many other traits, such as psychiatric disorders,[11; 12] personality,[13] life expectancy,[14] overall health,[15] and is especially deeply related to IQ.[16; 17] IQ is predictive for EA, and is a heritable complex trait[18; 19] of which the underlying genetic etiology is largely unknown. This makes EA itself an appealing trait for genetic association studies since it is more feasible to measure on a large scale than IQ.[20; 21] Higher cognitive function has recently been associated with increased homozygosity levels in a representative UK sample,[22] which is in the opposite direction of what one would expect assuming that higher educated individuals are more likely to mate with someone with different ancestry. Assortative mating on cognitive function was posed as a potential explanation for this finding, where assortment among individuals with higher cognitive ability may have induced increased homozygosity for loci that contribute to higher cognitive ability.

The current study examines how migration, ancestral background, and the proportion of the offspring genome consisting of runs of homozygosity (ROHs: multiple contiguous homozygous single nucleotide polymorphisms [SNPs]), vary

systematically between different levels of own and parental EA. The proportion of the genome consisting of ROHs is quantified by F_{roh}, which has demonstrated to be a powerful measure for shared ancestry of genetic haplotypes, and is generally used to study the deleterious effects of inbreeding in humans and other animals.[23] If parents with higher education have higher migration rates and tend to select mates with different ancestral backgrounds through assortative mating, we expect their offspring to show lower F_{roh} levels as well as weaker associations between ancestry-informative PCs and geography.

Data from a population cohort of ~2,000 unrelated subjects of Dutch ancestry included the EA of the participants and their parents, genome-wide SNPs, ancestry informative principal components (PCs), current living address, birthplace, parental birthplace, and religious affiliation.

Results

Table 1: Mean distance in km between birthplaces and p-values of t-tests testing the difference in birthplace distance between parental education levels

Educational level	Mean distance (km)	*p*-value difference test
Mean distance between paternal and own birthplace (km) per paternal education level:		
1. Primary	19.2 *(SD=32.4;N=172)*	-
2. Secondary	16.3 *(SD=30.8;N=512)*	.29 (vs. 1)
3. Higher secondary	28.7 *(SD=46.0;N=291)*	4.5×10^{-5} (vs. 2)
4. Tertiary	49.8 *(SD=55.3;N=375)*	1.1×10^{-7} (vs. 3)
Mean distance between maternal and own birthplace (km) per maternal education level:		
1. Primary	19.2 *(SD=34.7;N=245)*	-
2. Secondary	24.1 *(SD=37.8;N=722)*	.07 (vs. 1)
3. Higher secondary	34.6 *(SD=44.2;N=293)*	3.6×10^{-4} (vs. 2)
4. Tertiary	56.7 *(SD=52.6;N=223)*	6.9×10^{-7} (vs. 3)
Mean distance between paternal and maternal birthplace (km) per paternal education level:		
1. Primary	22.8 *(SD=34.9;N=144)*	-
2. Secondary	23.5 *(SD=39.7;N=483)*	.85 (vs. 1)
3. Higher secondary	34.0 *(SD=49.8;N=284)*	2.7×10^{-3} (vs. 2)
4. Tertiary	46.4 *(SD=50.5;N=383)*	1.6×10^{-3} (vs. 3)
Mean distance between paternal and maternal birthplace (km) per maternal education level:		
1. Primary	25.5 *(SD=47.8;N=145)*	-
2. Secondary	24.8 *(SD=39.9;N=641)*	.86 (vs. 1)
3. Higher secondary	36.2 *(SD=46.6;N=283)*	4×10^{-4} (vs. 2)
4. Tertiary	54.7 *(SD=52.7;N=222)*	4.7×10^{-5} (vs. 3)

Migration distance and EA

The distance between parental birthplace and own birthplace was significantly associated with EA (distance between own and paternal birthplace: $p = 8.9 \times 10^{-30}$, N=1,349; distance between own and maternal birthplace: $p = 1.2 \times 10^{-26}$, N=1,483). Post-hoc tests showed that this association is mainly driven by a significantly increasing migration distance as the educational level exceeds the Secondary Education (see Table 1 and Figure 1), with parents with a Tertiary Educational level having moved more than twice the distance than parents with Primary or Secondary Educational levels. The same effect was observed for the distance between paternal and maternal birthplace for both paternal ($p = 2.8 \times 10^{-13}$; N=1,294) and maternal ($p = 1.7 \times 10^{-16}$; N=1,291) educational levels (see Table 1), showing that higher educated individuals are more likely to mate with a partner from a different geographic region.

Educational assortment

Parental educational levels showed a high spouse correlation (Spearman's $\rho = .66$, $p = 3 \times 10^{-262}$, N=2,058; see Table 2). The majority of the parents (58.5%) shared the same educational level. The only other spouse pair combinations showing higher observed frequencies than expected are fathers with a Higher Secondary Education and mothers with a Secondary Education, or fathers with a Tertiary Education and mothers with a Higher Secondary Education (Table 2), a gender-asymmetrical pattern known as hypergamy.[24]

Table 2: Crosstab of 2,058 spouse pairs and their educational attainment: χ^2 (9) = 1496.89, p < .001, Spearman's ρ = .664.

Mother →	Primary	Secondary	Higher secondary	Tertiary
Father ↓				
Primary	**273 (78.2)**	79 (165.3)	11 (70.8)	8 (56.6)
Secondary	110 (154.4)	**528 (326.2)**	70 (139.8)	24 (111.7)
Higher secondary	38 (86.5)	**190 (182.7)**	**151 (78.3)**	31 (62.6)
Tertiary	13 (114.9)	120 (242.8)	**161 (104.1)**	**251 (83.2)**

The numbers between brackets is the expected number of spouse pairs in that cell under the null hypothesis of no assortment. Observed values higher than the expected values are in bold.

EA and F_{roh}

The subjects' own EA showed a nominally significant negative association with F_{roh} ($p = .045$, N=2,007). The association between offspring F_{roh} and parental EA was highly significant (father: $p < 10^{-5}$, N=1,989; mother: $p = 9 \times 10^{-5}$, N=1,995), with higher educated parents having offspring with lower F_{roh} levels (see Tables 3

and 4). Multiple confounders were accounted for in all regressions: we included ancestry-informative PCs, city size [i.e., living in a city with population size >100k], and religion (see reference[1] for more details on the relationship between these variables and F_{roh}). Religion significantly contributed to F_{roh} variation after including parental EA to the regression, albeit slightly less than without parental EA, indicating a (partly) independent effect of religion and EA on F_{roh} (see Table 4).

To evaluate whether the age difference between the genotyped subjects and their parents contributed to the difference between the effects of the subjects' own education and the parental education, the analyses were repeated only including individuals that were at an age where they were more likely to have completed their education. The analyses were run once only including subjects with age > 25, and once including only ages > 30. Both these analyses gave a non- significant result for the subjects' own EA (age>25: $p = .065$, N=1,641; age>30: $p = .075$, N=1,401), while parental EA remained significant (*father*: age>25: $p = 9.9 \times 10^{-4}$, N=1,610; age>30: $p = 1.6 \times 10^{-3}$, N=1,371; *mother*: age>25: $p = 3.8 \times 10^{-3}$, N=1,616; age>30: $p = .046$, N=1,376). Accounting for year of birth in order to evaluate the presence of a cohort effect also still results in a non-significant association between own EA and F_{roh} with own year of birth added as a predictor ($p = .181$, N=1,984), and a significant association between parental EA and F_{roh} with the parental year of birth added as an additional predictor (father: $p = 3.5 \times 10^{-3}$, N=1,401; mother: $p = 7.2 \times 10^{-3}$, N=1,534).

Table 3: Mean F_{roh} of the offspring, standard deviation, and sample sizes for each educational group

Educational level	Offspring education	Paternal education	Maternal education
Primary	.00192 *(SD=.003;N=74)*	.00200 *(SD=.003;N=372)*	.00184 *(SD=.003;N=439)*
Secondary	.00180 *(SD=.003;N=368)*	.00177 *(SD=.004;N=734)*	.00177 *(SD=.004;N=925)*
Higher secondary	.00170 *(SD=.003;N=659)*	.00149 *(SD=.003;N=413)*	.00127 *(SD=.002;N=397)*
Tertiary	.00141 *(SD=.003;N=988)*	.00108 *(SD=.001;N=548)*	.00100 *(SD=.001;N=314)*

Migration distance and F_{roh}

The distance between the paternal and maternal birthplaces was significantly associated with F_{roh} ($p = 9 \times 10^{-5}$, N=1,263), as was the association with the distance between parental and own birthplace (father: $p = 5.3 \times 10^{-3}$, N=1,317; mother: $p = 9.5 \times 10^{-3}$, N=1,445). After including the distance between the paternal and maternal birthplaces as a predictor, parental EA was no longer significantly associated with F_{roh} (paternal EA: $p = .077$, N=1,246; maternal EA: $p = .134$, N=1,242), while the birthplace distance still contributed significantly to F_{roh} variation (in regression including paternal EA: $p = 4.6 \times 10^{-5}$, in regression including

Table 4: Betas (and *p*-values between brackets) for each of the predictors included in the linear regressions on F_{roh}, as well as the R^2 change (= increase in explained variance after adding education as a predictor) and its empirical *p*-value from 100k permutations.

	Offspring education (N=2,007): R^2 change = .002, *Empirical p* = .045		Paternal education (N=1,989): R^2 change = .009, *Empirical p* < 10^{-5}		Maternal education (N=1,995): R^2 change = .008, *Empirical p* = 9×10^{-5}	
	No education included	*Education included*	*No education included*	*Education included*	*No education included*	*Education included*
Education	NA	-.0001 (.046)	NA	-.0003 (2.0×10^{-5})	NA	-.0003 (8.8×10^{-5})
PC1 (North-South)	.0164 (1.8×10^{-4})	.0162 (2.2×10^{-4})	.0168 (1.4×10^{-4})	.0158 (3.3×10^{-4})	.0149 (6.3×10^{-4})	.0145 (8.4×10^{-4})
PC2 (East-West)	.0118 (6.1×10^{-3})	.0117 (6.2×10^{-3})	.0122 (4.4×10^{-3})	.0128 (2.9×10^{-3})	.0131 (2.3×10^{-3})	.0141 (8.9×10^{-4})
PC3 (Middle-Band)	.0051 (.252)	.0049 (.269)	.0040 (.378)	.0037 (.407)	.0057 (.203)	.0053 (.233)
Religion (yes/no)	.0004 (4.0×10^{-3})	.0004 (4.7×10^{-3})	.0004 (4.4×10^{-3})	.0003 (.011)	.0004 (4.6×10^{-3})	.0003 (.016)
City Variable	-.0002 (.141)	-.0002 (.230)	-.0002 (.152)	-.0001 (.449)	-.0002 (.150)	-.0001 (.397)

Note that the education of the parents is regressed on F_{roh} of the subject.

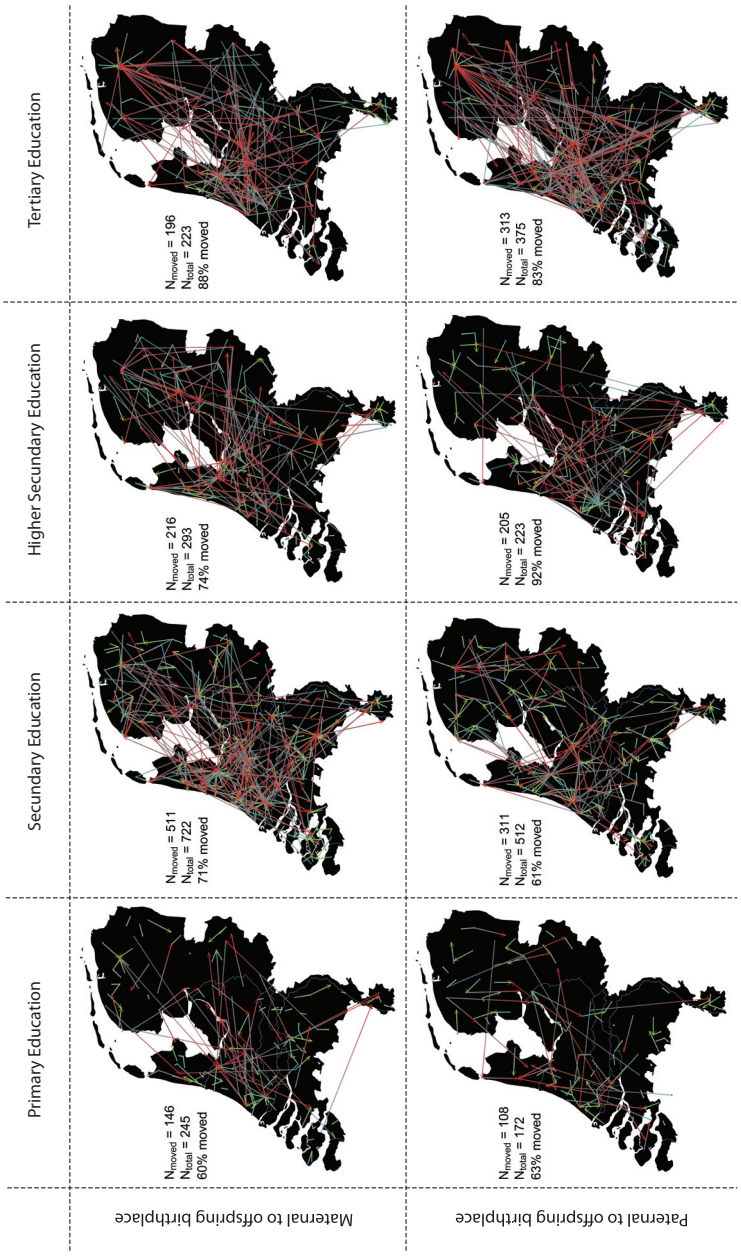


Figure 1. Plots of migrations from the parental birthplace to the offspring birthplace. The average distance the colors are based on are: father: 28.47 km (SD=44.45); mother: 30.16 km (SD=44.45). The difference between the moving distance of fathers with a Secondary Education and fathers with a Tertiary education is best suited to visualize the effect because of the almost equal sample sizes with respect to individuals plotted (i.e., moved) and the significant increase of moving distance (see Table 1); also note that fathers with Secondary Education have >25% measurements in total, which is another indicator of the difference in migration levels.

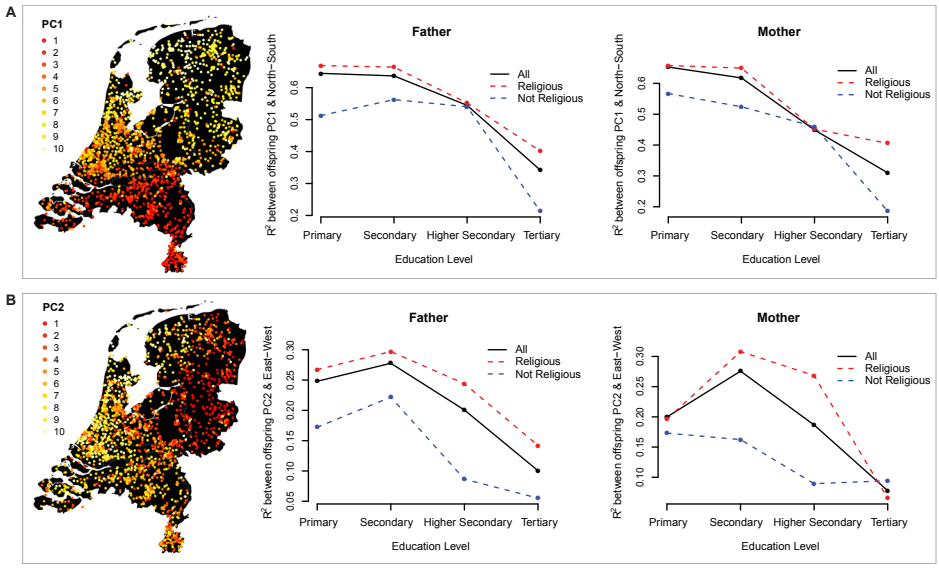


Figure 2. (***A***) Left: geographic distribution of PC1 (N = ~5,000 unrelated Dutch subjects), where the mean PC1 value per postal code (current living address) was computed, divided into 10 percentiles, and plotted. Right: the explained variance (R^2) of the offspring's PC1 by the North-South gradient based on the offspring's birthplace, per parental educational group. (***B***) Left: geographic distribution of PC2. Right: R^2 between offspring PC2 and the East-West gradient based on offspring's birth place.

maternal EA: $p = 3.1 \times 10^{-5}$). These results show that the association between F_{roh} and parental EA is explained by higher educated parents tending to have more different ancestries than lower educated parents because of higher migration levels.

Association between geography and ancestry per parental educational level

PCs from genome-wide single-nucleotide polymorphisms (SNPs) capture ancestral background, and show high correlations with geography within the Netherlands and other countries.[7; 25-27] In the current dataset, the first PC correlates .74 with the North-South gradient based on birthplace, and the second PC correlates .46 with the East-West gradient (N=1,892). The correlations between PCs and geographic location significantly differed between educational groups (PC1 for paternal EA: $p = 2.2 \times 10^{-12}$; PC1 for maternal EA: $p = 8.2 \times 10^{-12}$; PC2 for paternal EA: $p = 2.5 \times 10^{-4}$; PC2 for maternal EA: $p = 3.5 \times 10^{-4}$). Figure 2 shows a decreasing association between the PCs and geography as the parental education increases. We approximated this decrease with a linear trend ($\rho = \rho_0 + EA \times \rho_1$, where ρ is the correlation between PC and latitude/longitude, and EA is coded by 0, 1, 2, and 3), which gave us significant negative parameter estimates for ρ_1: PC1 for paternal EA: $\rho_1 = -.06$, $p = 1.7 \times 10^{-11}$; PC1 for maternal EA: $\rho_1 = -.07$, p

$= 8.9 \times 10^{-12}$; PC2 for paternal EA: $\rho_1 = -.06$, $p = 1.7 \times 10^{-4}$; PC2 for maternal EA: $\rho_1 = -.05$, $p = 6.5 \times 10^{-3}$. This would be expected if parents of subjects with higher EA tended to either live in a different geographic area than their ancestors, or that their partners live in a different geographic area than their ancestors. The effect is still visible after splitting up the sample into a religious and non-religious group (Figure 2). The non-religious group shows an overall weaker association between geography and the PCs, consistent with previously observed lower F_{roh} levels in the non-religious group in the Netherlands,[1] and suggesting migration may have also played a role in the homozygosity differences between religious and secular groups.

Discussion

The proportion of the autosomal genome in ROHs (F_{roh}) shows a nominally significant negative association with EA. In the absence of data on parental EA, geographic mobility, and ancestry, this observation could have been interpreted as the result of deleterious effects of inbreeding on cognitive ability, which would fit the existing hypotheses.[28-31] The effect was considerably more significant however when associating F_{roh} with paternal or maternal EA. We investigated whether this could be explained by a combination of migration and educational assortment. Ancestry correlates highly with geography in the Netherlands due to relatively low levels of within-country migrations in recent history.[7] Individuals with higher EA were significantly more likely to have migrated away from their birthplace and to mate with a partner from a different geographic region. In this context, educational assortment increases the chance for higher educated individuals to mate with genetically more dissimilar partners, lowering the number of homozygous alleles transmitted to their offspring, while lower educated individuals would have been more likely to mate closer to their ancestry. The association between F_{roh} and parental EA disappears after correcting for the distance between the paternal and maternal birthplaces, which itself was also significantly associated with F_{roh}. This is in line with the declining correlation between ancestry-informative PCs and the geographic location of the birthplace in subjects with higher educated parents (Figure 2). The same trend is visible after splitting up the sample in a religious and non-religious group, with the non-religious group showing consistently lower correlations between PCs and geography, suggesting migration may have also played a role in the F_{roh} differences between the religious and secular groups previously observed in this sample.[1]

A study in a UK sample found a nominally significant association between cognitive ability, which is predictive for EA, and F_{roh} in the opposite direction, with increased F_{roh} levels in individuals with higher cognitive ability.[22] Considering the

high correlation between IQ and EA,[16;17] the significant association between genetic variation and geography in the UK,[32;33] and higher migration rates for higher educated individuals within the UK,[34;35] we would have expected an association between cognitive ability and F_{roh} in the same direction as EA shows in the Dutch population. The authors hypothesized that the ROHs causing this association harbor causal variants that have become homozygous through assortative mating on cognitive function. This difference in results is reminiscent of the difference in direction between populations for the association between F_{roh} and major depressive disorder (MDD) for which the UK and the Netherlands also showed an opposite direction of effect.[1;36] The MDD-F_{roh} association disappeared in the Dutch sample after correcting for systematic differences in parental relatedness between religious and non-religious groups. These phenomena illustrate the importance of the impact of complex social, demographic, and historical processes on the genomic structure of populations. The fact that the offspring education was much less significantly associated with offspring F_{roh} than the parents' education and that the association disappeared after correcting for the distance between parental birthplaces strongly suggests that the effects we observed in the Dutch population do not reflect systematic differences in the frequency of causal genetic variants. Further analyses in a more deeply phenotyped and representative UK sample (preferably with own and parental EA & birthplace measured) are necessary in order to investigate the discrepancy in direction of effects between the UK and Dutch population, and the role of causal variants herein.

Non-random mating in human populations can be driven by heritable social traits like religion and EA. The impact of these mating behaviors on the genomic structure of a population is not always directly captured by traditional measures for population stratification, such as ancestry-informative PCs. These findings are relevant for genetic association studies, since these behaviors can be associated with additional traits of interest, like psychiatric disorders with religiosity,[1;37] or IQ with EA.[21] Deleterious effects of inbreeding studied by associating F_{roh} and the trait of interest usually require much larger sample sizes for detection than that of the current dataset (~12,000-65,000).[23] We suspect that ancestral behavior may have influenced genetic variation more systematically than genetic variation influenced the current measurable behavior in our dataset. This additional confounding and non-causal "noise" may have contributed to the difficulty of finding consistent genetic association signals for many behavioral traits, especially if the nature, effect size, and/or direction of such confounding effects would differ per population. We recommend that cohorts contributing to meta-analyses of genetic association studies on behavioral, cognitive, and social traits search for patterns of variation caused by the social/historical context of their population, so these can be accounted for accordingly in their analyses and interpretations.

Methods

Participants

Genotyped subjects were registered at the Netherlands Twin Register (NTR,[38] N=6,685; 2,678 males and 4,007 females). The NTR subjects were randomly sampled from twin families across the Netherlands. Analyses were done on unrelated individuals only. Unrelated individuals were chosen using GCTA,[39] by excluding one of each pair of individuals with an estimated genetic relationship of >0.025 (i.e., more related than third or fourth cousin). Only individuals with Dutch ancestry were included. Individuals with a non-Dutch ancestry were identified by projecting PCs from 1000 Genomes populations on the dataset, and with additional help of the birth country of the parents. This procedure is described in more detail elsewhere.[7]

This study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the US Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180). All subjects provided written informed consent.

Phenotypes

EA was measured longitudinally with the question "*What is the highest educational level that you have finished?*", "*What is the highest educational level that your father has finished?*", and "*What is the highest educational level that your mother has finished?*". The answer categories varied per survey, but could all be recoded into the following four categories: 1) Primary Education; 2) Secondary Education (VMBO, LBO, MAVO, lower secondary); 3) Higher Secondary Education (MBO, HAVO/VWO, higher secondary); 4) Tertiary Education (HBO, university, PhD). EA was available for 2,089 unrelated genotyped Dutch subjects, paternal EA was measured for 2,067 unrelated genotyped Dutch subjects, and maternal EA was available for 2,075 unrelated Dutch subjects (largely in the same subjects: 2,026 individuals had their own, their paternal, and maternal EA available).

Information on birthplace was available from survey and from city council register data for 1,892 unrelated genotyped Dutch subjects, paternal birthplace for 1,465 subjects, and maternal birthplace for 1,618 unrelated Dutch subjects; 1,371 individuals had both their own and paternal birthplace available, 1,513 individuals had their own and maternal birthplace available, 1,312 had both parental birthplaces available, and 1,227 individuals had their own, their paternal, and maternal birthplace available. Distance between birthplaces was computed with a purpose written perl script using the algorithm available on http://www.geodatasource.com/developers/perl.

The assessment of religion and city size are described in detail elsewhere.[1] Sample sizes of individuals that had EA, religion, and city size available (i.e., were included in the statistical analyses) are given in the results section and the Tables for each analysis.

Genotyping, QC, and ancestry-informative PCs

Genotyping was performed on the Affymetrix Human Genome-Wide SNP 6.0 Array according to the manufacturer's protocol. Methods for blood and buccal cell collection, genomic DNA extraction, genotyping, and QC have been described previously.[1;40;41] Only autosomal SNPs were analyzed. After QC, 498,592 SNPs remained.

Ancestry-informative PCs were computed with EIGENSTRAT[42] on 5,166 unrelated subjects with Dutch ancestry, which also included subjects from the Netherlands Study of Depression and Anxiety (NESDA).[43] The ancestry-informative PCs and their computation are described in detail elsewhere.[7]

ROHs and F_{roh}

ROHs were called using Plink.[44] A recent study comparing several software packages designed for this goal concluded that Plink predicts autozygous stretches optimally,[45] using simulated data based on the Affymetrix 6.0 chip, making their density of SNPs in linkage disequilibrium (LD) close to ours. We followed the recommendations from this study in calling ROHs: (1) SNPs were pruned for LD (window size = 50, number of SNPs to shift after each step = 5, based on a variance inflation factor [VIF] of 2), resulting in 131,325 SNPs; (2) an ROH was defined as ≥65 consecutive homozygous SNPs with no heterozygote calls allowed. F_{roh} is an overall measure of the proportion of the autosome in ROHs, which is calculated as the total length of ROHs summed for each individual, and then divided by the total SNP-mappable autosomal distance (2.77×10^9 bases).

Statistical analyses

Migration distance and EA: The relation between birthplace distances and EA was investigated firstly with a one-way ANOVA in IBM SPSS Statistics 20, with birthplace distance as the dependent variable and EA as the independent variable. Four tests were performed: 1) distance between own and paternal birthplace as the dependent variable and paternal EA as independent variable; 2) distance between own and maternal birthplace as the dependent variable and maternal EA as independent variable; 3) distance between parental birthplaces as the dependent variable and paternal EA as independent variable; 4) distance between parental birthplaces as the dependent variable and maternal EA as independent variable. Post-hoc tests were then conducted with t-tests comparing birthplace distances

between two educational levels (see Table 1), computed in IBM SPSS Statistics 20.

Educational assortment: The chi-squared test and Spearman's ρ was computed in IBM SPSS Statistics 20 to test for the assortment on EA (Table 2).

EA and F_{roh}: The R^2 change (= difference in explained variation of F_{roh}) was computed between multiple regressions on F_{roh} with and without EA as a predictor (i.e., own or parental EA [Table 4]). The regressions included as predictors: the three PCs reflecting ancestry (correlated significantly with geography: PC1=North-South PC, PC2=East-West PC, PC3=middle-band PC),[7] city size (dichotomous, i.e., living in a city with population size >100k), and religion. To evaluate the presence of a birth cohort effect, the analyses were repeated including year of birth as an additional predictor. To correct for the non-normal distribution of F_{roh}, F_{roh} was permuted 100,000 times. These analyses were done in a purpose written perl script, using the PDL::Stats::GLM perl module (see http://search.cpan.org/~maggiexyz/PDL-Stats-0.6.2/GLM/glm.pp).

Migration distance and F_{roh}: The R^2 change (= difference in explained variation of F_{roh}) was computed between multiple regressions on F_{roh} with and without the distance between birthplaces as a predictor. The regressions included as predictors: the three PCs reflecting ancestry, city size, religion, and EA. F_{roh} was permuted 100,000 times. These analyses were done in a purpose written perl script, using the PDL::Stats::GLM perl module.

Association between geography and ancestry per parental educational level: The influence of parental educational level on correlations between ancestry-informative PCs of the offspring and geographic location was tested using full information maximum likelihood estimation in OpenMx,[46] separately for maternal and paternal educational levels and the two ancestry-informative PCs. We approximated the effect of parental EA on the correlations between PC-values and geographic location with the following linear model: $\rho = \rho_0 + \text{education} \times \rho_1$, where ρ is the correlation between PC and latitude/longitude. The null hypothesis $\rho_1 = 0$ was tested by mean of the likelihood ratio test.

The R^2 from Figure 2 was computed by squaring the Pearson correlation computed in IBM SPSS Statistics 20.

References

1. Abdellaoui, A., Hottenga, J.-J., Xiao, X., Scheet, P., Ehli, E.A., Brooks, A., Davies, G.E., Hudziak, J.J., and Smit, D.J., et al. (2013). Association between Autozygosity and Major Depression: Stratification due to Religious Assortment *Behavior Genetics* 43, 455-467.
2. Blossfeld, H.-P. (2009). Educational assortative marriage in comparative perspective. *Annual Review of Sociology* 35, 513-530.
3. Mare, R.D. (1991). Five decades of educational assortative mating. *American Sociological Review*, 15-32.
4. Schwartz, C.R. (2013). Trends and Variation in Assortative Mating: Causes and Consequences. *Annual Review of Sociology* 39, 451-470.
5. Vandenberg, S.G. (1972). Assortative mating, or who marries whom? *Behavior Genetics* 2, 127-157.
6. Mulder, C., and Kalmijn, M. (2004). Even bij oma langs: NKPS laat zien hoe ver familieden van elkaar wonen. *Demos* 20, 78-80.
7. Abdellaoui, A., Hottenga, J.-J., de Knijff, P., Nivard, M.G., Xiao, X., Scheet, P., Brooks, A., Ehli, E.A., Hu, Y., et al. (2013). Population structure, migration, and diversifying selection in the Netherlands. *European Journal of Human Genetics* 21, 1277-1285.
8. Boomsma, D.I., Wijmenga, C., Slagboom, E.P., Swertz, M.A., Karssen, L.C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., et al. (2014). The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics* 22, 221-227.
9. Branigan, A.R., McCallum, K.J., and Freese, J. (2013). Variation in the heritability of educational attainment: An international meta-analysis. *Northwestern University Institute for Policy Research Working Paper* 13.
10. Heath, A.C., Berg, K., Eaves, L.J., Solaas, M.H., Corey, L.A., Sundet, J., Magnus, P., and Nance, W.E. (1985). Education policy and the heritability of educational attainment. *Nature* 314, 734-736.
11. Breslau, J., Lane, M., Sampson, N., and Kessler, R.C. (2008). Mental disorders and subsequent educational attainment in a US national sample. *Journal of Psychiatric Research* 42, 708-716.
12. Kessler, R.C., Foster, C.L., Saunders, W.B., and Stang, P.E. (1995). Social consequences of psychiatric disorders, I: Educational attainment. *American journal of psychiatry* 152, 1026-1032.
13. De Raad, B., and Schouwenburg, H.C. (1996). Personality in learning and education: A review. *European Journal of Personality* 10, 303-336.
14. Meara, E.R., Richards, S., and Cutler, D.M. (2008). The gap gets bigger: changes in mortality and life expectancy, by education, 1981–2000. *Health Affairs* 27, 350-360.
15. Mackenbach, J.P., Stirbu, I., Roskam, A.-J.R., Schaap, M.M., Menvielle, G., Leinsalu, M., and Kunst, A.E. (2008). Socioeconomic inequalities in health in 22 European countries. *New England Journal of Medicine* 358, 2468-2481.
16. Deary, I.J., Strand, S., Smith, P., and Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence* 35, 13-21.
17. Kaufman, A.S., Kaufman, J.C., Liu, X., and Johnson, C.K. (2009). How do educational attainment and gender relate to fluid intelligence, crystallized intelligence, and academic skills at ages 22–90 years? *Archives of Clinical Neuropsychology* 24, 153-163.
18. Bouchard, T.J., Lykken, D.T., McGue, M., Segal, N.L., and Tellegen, A. (1990). Sources of human psychological differences: The Minnesota study of twins reared apart. *Science* 250, 223-228.
19. Haworth, C., Wright, M., Luciano, M., Martin, N., De Geus, E., Van Beijsterveldt, C., Bartels, M., Posthuma, D., Boomsma, D., et al. (2009). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry* 15, 1112-1120.
20. Martin, N.W., Medland, S.E., Verweij, K.J., Lee, S.H., Nyholt, D.R., Madden, P.A., Heath, A.C., Montgomery, G.W., Wright, M.J., et al. (2011). Educational attainment: a genome wide association study in 9538 Australians. *PloS One* 6, e20128.
21. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., et al. (2013). GWAS of 126,559 Individuals Identifies

Genetic Variants Associated with Educational Attainment. *Science* 340, 1467-1471.
22. Power, R.A., Nagoshi, C., DeFries, J.C., Donnelly, P., Barroso, I., Blackwell, J.M., Bramon, E., Brown, M.A., Casas, J.P., et al. (2013). Genome-wide estimates of inbreeding in unrelated individuals and their association with cognitive ability. *European Journal of Human Genetics*.
23. Keller, M.C., Visscher, P.M., and Goddard, M.E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189, 237-249.
24. Esteve, A., García-Román, J., and Permanyer, I. (2012). The Gender-Gap Reversal in Education and Its Effect on Union Formation: The End of Hypergamy? *Population and Development Review* 38, 535-546.
25. Chen, J., Zheng, H., Bei, J.-X., Sun, L., Jia, W.-h., Li, T., Zhang, F., Seielstad, M., Zeng, Y.-X., et al. (2009). Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *American Journal of Human Genetics* 85, 775-785.
26. McEvoy, B.P., Montgomery, G.W., McRae, A.F., Ripatti, S., Perola, M., Spector, T.D., Cherkas, L., Ahmadi, K.R., Boomsma, D., et al. (2009). Geographical structure and differential natural selection among North European populations. *Genome Research* 19, 804-814.
27. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., et al. (2008). Genes mirror geography within Europe. *Nature* 456, 98-101.
28. Jensen, A.R. (1983). Effects of inbreeding on mental-ability factors. *Personality and Individual Differences* 4, 71-87.
29. Morton, N. (1978). Effect of inbreeding on IQ and mental retardation. *Proceedings of the National Academy of Sciences* 75, 3906-3908.
30. Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S.S., Chen, W., Hosseini, M., Behjati, F., Haas, S., et al. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* 478, 57-63.
31. Woodley, M.A. (2009). Inbreeding depression and IQ in a study of 72 countries. *Intelligence* 37, 268-276.
32. Tyler-Smith, C., and Xue, Y. (2011). A British approach to sampling. *European Journal of Human Genetics* 20, 129-130.
33. Winney, B., Boumertit, A., Day, T., Davison, D., Echeta, C., Evseeva, I., Hutnik, K., Leslie, S., Nicodemus, K., et al. (2011). People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *European Journal of Human Genetics* 20, 203-210.
34. Faggian, A., McCann, P., and Sheppard, S. (2006). An analysis of ethnic differences in UK graduate migration behaviour. *The Annals of Regional Science* 40, 461-471.
35. Faggian, A., McCann, P., and Sheppard, S. (2007). Human capital, higher education and graduate migration: an analysis of Scottish and Welsh students. *Urban Studies* 44, 2511-2528.
36. Power, R.A., Keller, M.C., Ripke, S., Abdellaoui, A., Wray, N.R., Sullivan, P.F., Group, M.P.W., and Breen, G. (2013). A Recessive Genetic Model and Runs of Homozygosity in Major Depressive Disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*.
37. Koenig, H.G. (2009). Research on religion, spirituality, and mental health: A review. *Canadian Journal of Psychiatry* 54, 283-291.
38. Boomsma, D.I., de Geus, E.J., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.-J., Posthuma, D., van Beijsterveldt, T.C., Hudziak, J.J., et al. (2006). Netherlands Twin Register: from twins to twin families. *Twin Research and Human Genetics* 9, 849-857.
39. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* 88, 76.
40. Boomsma, D.I., Willemsen, G., Sullivan, P.F., Heutink, P., Meijer, P., Sondervan, D., Kluft, C., Smit, G., Nolen, W.A., et al. (2008). Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *European Journal of Human Genetics* 16, 335-342.
41. Willemsen, G., de Geus, E.J.C., Bartels, M., van Beijsterveldt, C.T., Brooks, A.I., Estourgie-van Burk, G.F., Fugman, D.A., Hoekstra, C., Hottenga, J.-J., et al. (2010). The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Research and*

Human Genetics 13, 231.
42. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38, 904-909.
43. Penninx, B.W., Beekman, A.T., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W., et al. (2008). The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research* 17, 121-140.
44. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81, 559-575.
45. Howrigan, D., Simonson, M., and Keller, M. (2011). Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics* 12, 460.
46. Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., et al. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 76, 306-317.

CHAPTER 10

SUMMARY

This thesis consists of eight studies that can be divided into four parts. The first part consists of two twin studies on Thought Problems (TP) where the longitudinal heritability is estimated, and rater bias of parental reports of TP in children and measurement invariance across age and sex of self-reports of TP in adolescents and adults is examined. In the second part of the thesis, the search for post-twinning *de novo* CNVs in monozygotic (MZ) twins is reported. The extent of CNV discordance between MZ twins is investigated in 50 MZ pairs selected for concordance and discordance on Attention Problems (AP), and in a group of ~1,100 MZ pairs, unselected for a particular phenotype. In the third part, genome-wide patterns reflecting Dutch ancestry are examined using single-nucleotide polymorphisms (SNPs) as typed on microarrays, and indels (short insertions or deletions) and larger deletions from Next Generation Sequence (NGS) data. In the final part, two studies are performed on ancestral non-random mating through migration and assortment and its effect on genome-wide variation through runs of homozygosity (ROHs).

Part I: Twin Studies: Rater Effects, Measurement Invariance, and Longitudinal Heritability of Thought Problems

The Thought Problems (TP) scale is an empirically derived set of items that measures symptoms common in several mental disorders: hallucinations, obsessive-compulsive symptoms, strange thoughts and behaviors, self-harm, and suicide attempts.

Chapter 2, shows that more than half of the variation of TP is explained by additive genetic factors in ~9,000 7-year-old twin pairs. The heritability estimated was based on parental ratings. The analyses modeled genetic and environmental influences on the commonly agreed upon part of the phenotype and on the unique views of the parents separately. The part of TP that parents agreed on explained ~67% of the variance, of which 76% was due to additive genetic influences. The unique part of the reported TP was also mostly heritable (maternal part: 61%, paternal part: 65%), indicating that the unique views of the parents very likely reflect real behavior of the child instead of rater bias. The unique views of the parents were, unlike the part they agreed on, also influenced by shared environmental factors (maternal part: 13%, paternal part: 13%). The remaining variance (24%) was explained by unique environmental influences and may also partly reflect measurement error.

In **chapter 3**, the strength and the structure of the relations between self-reported TP-items is investigated with an exploratory factor analysis (EFA), measurement invariance (MI) of the TP-scale is investigated across age and sex, and the extent of genetic and environmental influences is estimated longitudinally using the genetic relatedness of ~9,000 twin pairs and ~2,000 siblings. The EFA yielded a one-factor structure. The one-factor structure was then used in a

multigroup confirmatory factor analysis that led to the conclusion that the TP-scale is measurement invariant between adolescent (12-18 years old), young adult (19-27 years) and adult (28-59 years) males and females, in the sense that differences between these groups reflect real differences in the mean and/or variation of the construct measured by TP. About 37% of the variation in TP throughout adolescence, young adulthood, and adulthood is estimated to be due to additive genetic influences. TP was influenced by the same additive genetic component throughout the three age groups, with an additional genetic component arising during young adulthood, which keeps influencing the trait throughout adulthood as well. The remaining variance (63%) was explained by unique environmental influences and may also partly reflect measurement error.

Part II: Copy Number Variants: Post-Twinning Mutations and Discordance between Monozygotic Twins

In **chapter 4**, 50 MZ twin pairs were selected out of ~3200 MZ twin pairs based on their data on the Attention Problems (AP) scale: 17 concordant affected, 22 concordant unaffected, and 11 discordant pairs. The AP scale has been shown to be predictive for ADHD, and both AP and ADHD are highly heritable. CNVs were measured for this group of twins and in a subset (25 pairs) also for the parents. The presence of pre- and post-twinning CNVs was investigated, and an association analysis was conducted to test whether CNVs were associated with AP. Out of 26 *de novo* CNVs suggested by the Affymetrix 6.0 chip, three were replicated using qPCR: 1) a pre-twinning mutation (duplication) in a healthy twin pair, 2) a post-twinning mutation (deletion) in a concordant affected twin pair, and 3) a post-twinning mutation (duplication) in the affected twin of a discordant pair. The post-twinning mutations overlapped with genes that were associated in previous studies with the co-morbidity of psychiatric disorders that were also diagnosed in twins carrying the mutations. Besides more behavioral problems, the carriers of the post-twinning mutations also had a lower birth weight than their co-twin. In the entire sample of 50 twin pairs, association analyses of genome-wide CNV burden and AP showed that CNVs overlapping with genes were significantly larger in affected than in unaffected subjects. CNVs in non-genic regions did not show this association. For CNVs overlapping with genes, both deletions and duplications showed the same trend (i.e., larger in affected than unaffected individuals), but no significant differences, indicating that both contributed to the effect.

In **chapter 5** a genome-wide search was conducted for post-twinning CNVs in ~1,100 unselected MZ twin pairs, of which about half had their DNA extracted from buccal (mainly children), and the other half from blood (mainly adults). A total of 153 putative post-twinning *de novo* CNVs were found. The majority of these 153 CNVs resided in 15q11.2, of which the majority was

significantly overrepresented in blood-derived DNA, and observed in significantly older twins within the dataset of blood-derived DNA. Based on visual inspection of raw intensity signals done by multiple raters independently, eleven *de novo* CNVs were deemed suitable for a first series of qPCR follow-up experiments. All eleven resided in 15q11.2, of which two mutations within the same twin pair with buccal-derived DNA were validated by the qPCR experiments (~350kb and ~280 kb). The twins were thirteen years old at the time of sampling, and did not show large phenotypic differences based on parental and self-report questionnaires from ages 1 to 21.

Part III - Population Genetics: Genomic Structure of the Netherlands

In **chapter 6**, principal component analyses (PCAs) were conducted on ~500k genome-wide SNPs in 4,441 unrelated Dutch subjects and 1,014 unrelated subjects from 14 different world-wide populations from the 1000 Genomes dataset. Removing long-range linkage disequilibrium (LD) regions and LD-based SNP pruning (resulting in 130,248 SNPs) significantly changed the composition of the top ten PCs in the Dutch dataset, but barely influenced the top ten PCs in the 1000 Genomes dataset. Minimizing LD resulted in PCs with significantly higher correlations with Dutch geography, and also resulted in three ancestry-informative PCs: 1) the North–South PC, which differentiates the Southern provinces below the three major rivers from the Northern provinces, with the more urbanized West falling in between, 2) the East–West PC, which mainly differentiates the Northeastern part of the Netherlands from the rest, and 3) the Middle-Band PC, which separates the middle-band area of the Netherlands from the rest. The North-South PC showed several similarities with the European North-South cline: 1) a correlation of .66 with the 1000 Genomes PC that separates Northern from Southern European populations, 2) a significant correlation with genome-wide homozygosity (F; North = more homozygous), 3) a significant correlation with height (North = taller), 4) a signal of significant selection pressure with a SNP associated with blue/brown eye-color (North = more blue eyes). The extent of adaptive effects on the genetic differentiation within the Netherlands was investigated by comparing the distribution of alleles for ~500k SNPs between subpopulations identified by the three PCs. Besides the signal in HERC2 (i.e., eye color, which was the strongest signal), there were significant signals of diversifying selection pressures in 544 other SNPs, of which the majority resided in 184 genes, among which genes involved in brain function were significantly overrepresented.

In **chapter 7**, it was investigated whether genome-wide patterns reflecting ancestry were also detectable in structural variants (SVs) from a Next Generation Sequencing (NGS) dataset from a collaborative effort of four Dutch biobanks called *Genome of the Netherlands* (GoNL). PCA was conducted on 490

unrelated individuals from across the Netherlands on common indels (<20 bp) and common larger deletions (20 – 10,000 bp) separately (where common means: MAF > 1%). Geographical location and the three ancestry-informative PCs from chapter 6 (projected onto GoNL individuals) were used to determine which indel PCs and larger deletion PCs were likely to reflect ancestry. Indels showed three ancestry-informative PCs, and larger deletions showed five. The indel PC explaining most variation showed a correlation of .86 with the North-South SNP PC and showed the same geographic distribution. The other two indel PCs showed a geographic distribution that was similar to the East-West SNP PC from chapter 6, but correlated only ~.36 with the East-West SNP PC, and showed a greater genetic distance between the Eastern and Western subpopulations. The five larger deletion PCs showed significant but lower correlations with the SNP PCs from chapter 6 (with the significant correlations ranging from .16 to .30) and showed no clear geographic distributions when plotted on the Dutch map (although there were significant correlations with geography ranging from .18 to .26). Ancestry-informative SNP PCs and indel PCs show similar degrees of differentiation across SNPs, indels, and larger deletions. Larger deletion PCs however seem to capture genome-wide patterns that mostly reflect larger deletions. Larger deletion PCs were also the only ancestry-informative PCs to not show a significant spouse correlation. It is not yet clear whether the discrepancy between ancestry-signals from larger deletions and from the smaller SNPs and indels are due to larger deletions capturing older variation, or due to ancestry signals being weaker due to a limited amount of measured larger deletions.

Part IV - Runs of Homozygosity: How Ancestral Behaviors Influence Current Genetic Variation

In **chapter 8**, religious affiliation shows a significant association with autozygosity, which was measured by F_{roh}: the proportion of the genome that consists of runs of homozygosity (ROHs). It also became apparent that the geographic distribution of the North-South PC that was discovered in chapter 6 matches that of the major religious denominations in the Netherlands, which has had a stable geographic distribution for about four centuries. The distribution of religions in the current dataset showed that this distribution was maintained, but with an overall increase of non-religious individuals, which have become more numerous in more recent times. The spouse correlations for religious affiliation are quite high (.73), in line with the high spouse correlation of the North-South PC and historical documentation of strong religious assortment among the Dutch. Post-hoc association analyses revealed that the significant religion-F_{roh} association was due to a difference between the religious and non-religious group, with the non-religious group showing significantly fewer/shorter ROHs, which is likely

explained by the relatively recent absence of denominational restrictions on mate selection for non-religious people, which increased variation in the gene pool of possible mates. The non-religious group also showed significantly more cases of major depressive disorder (MDD), which caused a nominally significant association between autozygosity and MDD. This spurious association disappeared after correcting for religious affiliation. Ancestry-informative PCs were not able to correct for this, since they also did not capture the autozygosity differences between religious and non-religious groups.

In **chapter 9**, a significant association between educational attainment and F_{roh} is observed. Like in chapter 8, additional evidence shows that this association is not due to causal ROHs decreasing the chances for achieving a higher educational attainment, but due to ancestral mating behavior. Parental educational attainment was much more significantly associated with offspring F_{roh}, with higher educated parents having offspring with a smaller genomic proportion of ROHs. Parents with higher educational attainment showed significantly higher distances between their own birthplace and the birthplace of their offspring and a higher distance to the birthplace of their spouse. The distance between paternal and maternal birthplaces was also significantly associated with F_{roh}, and when included in the F_{roh}-education regression, the association between F_{roh} and parental educational attainment disappeared, while the association between F_{roh} and parental birthplaces was still significant. This indicates that the association between F_{roh} and (parental) educational attainment was due to higher educated parents showing higher migration rates, increasing the chances of choosing another higher and more mobile educated spouse with a different ancestral background (as spouse correlations were also very significant). This effect was also visible through significantly decreasing correlations between ancestry-informative PCs and geography in offspring of parents with a higher education.

CHAPTER 11

GENERAL DISCUSSION

Behavior genetics is the discipline that studies the role of genetics in human and other animal behavior. In the first half of this thesis, classical behavior genetics topics are explored, such as the role of phenotypic measurement and the twin design in estimating genetic and environmental influences on human traits, and molecular genetic data are incorporated to explore the extent of genetic similarity in monozygotic twins. In the second half of the thesis, molecular genetic, geographic, and phenotypic data are incorporated to examine genetic variation on a population level within a relatively small but densely populated country, and the role of behavior therein.

Behavior ← Genetics: Twin Studies

The twin study design, developed around a century ago,[1; 2] allows for deriving heritability estimates by comparing the phenotypic resemblance between monozygotic (MZ) twins with that of dizygotic (DZ) twins.[3] Family, twin, and adoption studies demonstrated in the early 20th century that our genome plays a strong and significant role in the development of psychological and psychiatric traits.[4] In the mid-20th century however, the prevailing view was that environmental influences were largely responsible for the development of these traits, until the 1960s and 1970s, when a more balanced view on the nature-nurture debate gained ground again in the scientific community, in part through twin and adoption studies.[5] Many have utilized this design by gathering a large number of phenotypic measurements of twin pairs, often using questionnaire data, mostly self-reports when it comes to adults, and parental and/or teacher reports on children. It is estimated that over 1.5 million twins and their families participate in twin studies worldwide.[6]

Part I - Twin Studies: Rater Effects, Measurement Invariance, and Longitudinal Heritability of Thought Problems

In the first part of the thesis, the impact of genetic and environmental influences are estimated for the empirically derived combination of questions referred to as "Thought Problems" (TP), which is one of the more difficult problem behavior scales to analyze in a population-based sample, unless large sample sizes are available, due to positive item answers having a very low prevalence. This scale has often been used in the context of how predictive it is for several psychiatric traits, such as OCD, schizophrenia, bipolar disorder, substance abuse, social phobia, anxiety and mood disorders, and is likely to measure a single underlying construct. The results from chapter 3 show that the TP scale measures one factor, and shares additive genetic influences across age. As detailed in the discussion of chapter 3, the

symptoms measured by the scale (hallucinations, OCD-symptoms, strange thoughts and behaviors, self-harm and suicide attempts) suggest that TP may be measuring a liability for a so-called "schizo-obsessive" disorder. Schizophrenia and OCD have been shown to occur together more often than expected by chance, and share common functional circuits and dysfunctions of neurotransmitter systems.[7]

The heritability of TP in young children estimated in chapter 2 (63%) is considerably larger than the heritability of the multiple older age groups in chapter 3 (37%). Based on these results, we cannot yet conclude that this means that TP is more heritable in children than in adolescents and adults. The estimates in chapter 2 were obtained with the so-called rater bias model, which decomposes the variance in scores of parental reports into a part of the phenotype that both parents agree on (referred to as the common part), and parts that reflect parents' unique views on the trait (the rater specific parts), and gives estimates on genetic and environmental influences on all these components separately. In chapter 3 the heritability estimates were based on self-reports from twins and their siblings. The sudden drop in heritability from children to adolescents has been observed previously for Attention Problems (AP) as well,[8] and is likely explained by the change from parental reports to self-reports as soon as children reached the age of 12.

This drop in heritability may be the result of rater effects similar to what we modeled in chapter two and was recently addressed by Kan et al (2014).[9] Kan et al (2014)[9] showed analytically that if rater specific factors are genetically influenced, heritability estimates depend on whether both twins are rated by the same individual (which is the case in chapter 2, where the children are both rated by the father and/or by the mother), or each by a different individual (like in chapter 3, where both twins are rated by themselves, i.e., two different individuals). When they are each rated by a different individual (i.e., themselves), only the heritability of the component of the trait that they both agree on (the common part) is estimated, while the rater specific component is subscribed to environmental influences. When they are both rated by the same person, their mother for example, the effects that only the mother observes (rater specific) genetic factors are added to the heritability estimates. It seems that we assessed the following four components of the TP trait in these two studies: 1) the common part of parental reports of 7-year-olds, 2) the rater specific part of maternal reports in 7-year-olds, 3) the rater specific part of paternal reports in 7-year-olds, and 4) the common part of self-reports in young adolescents, young adults, and adults (12-59 years old). The first three components are likely to overlap with the fourth, but we would need additional data to formally test this by including parental reports on 12-59 year old twins, and self-reports for the 7-year old twins. These additional data are difficult to obtain, especially in large numbers, so for now we should be careful in interpreting different heritability estimates derived from different raters and different studies. We

cannot say much with certainty yet about the longitudinal heritability changes that may or may not occur between children and young adolescents. If we follow the reasoning of Kan et al (2014),[9] the heritability estimates of the TP in adolescents and adults (currently estimated at 37%) are likely to increase if twins would rate themselves as well as their co-twins. What makes different raters report different components of the trait? Whatever the cause, these different components do seem to be "real", as they seem to be influenced by genes. Item level analyses on data from multiple raters may give us a better idea on what the common and rater specific components of these traits represent phenotypically. These different perspectives on the same underlying trait could give us a more complete picture and important information concerning the measurement of psychiatric (endo)phenotypes. Using more informants on better defined and more accurately modeled traits may provide more insight in the different heritable parts of a trait and may thus be helpful in the quest for finding causal genetic variants.

Part II - Copy Number Variants: Post-Twinning Mutations and Discordance between Monozygotic Twins

In chapters 4 and 5 the assumption of the twin study design that MZ twins are 100% genetically identical was investigated. Genetic differences that lead to phenotypic differences between MZ twins would confound heritability estimates from twin studies (they would lead to an underestimation of the genetic contribution), but could potentially be used to identify causal genetic variants, as was demonstrated for example in the identification of the causal variant underlying Van der Woude Syndrome.[10] *De novo* copy number variants (CNVs) have been identified as possible major risk factors for several psychiatric disorders.[11-13] The majority of current molecular genetic studies focus on single nucleotide polymorphisms (SNPs), while structural variants (SVs), such as CNVs, may have a two to four times higher mutation rate and cover a larger part of the genome,[14; 15] but are unfortunately also harder to reliably measure by microarray technology. Microarrays are better at measuring whether a nucleotide is an A or a G (i.e., a SNP) than at detecting how many times a certain DNA segment occurs.[16] This makes the chance for a false positive finding much higher than finding an actual genetic difference between MZ twins. Studies using microarrays for this goal without attempting to validate their *de novo* CNVs with qPCR usually report many more CNVs differences between MZ twins than studies that do (see reference[17] for example, which reports 21 post-twinning mutations in 2 MZ pairs, in comparison to the many studies in Table 1 from chapter 5, where the majority of putative mutations did not get validated).

Thus, searching for post-twinning de novo mutations in CNVs was an endeavor with relatively low chances of succeeding. The chance of success was

likely increased by heavily selecting discordant MZ pairs from a large sample of ~3200 MZ pairs as detailed in chapter 4. Some interesting results were found when we screened 50 MZ pairs selected for AP (11 discordant), and the raw intensity signals of the 1.3 Mb post-twinning deletion on chromosome 4 (Figure 1b of chapter 4) were especially convincing, and confirmed. Another promising feature of the chapter 4 results is that the somatic mutations overlapped with genes where CNVs were previously implicated with the co-morbidity of psychiatric diseases that were also observed in the carriers of the mutations. Among the 50 MZ pairs from chapter 4, we also validated a pre-twinning *de novo* CNV (i.e., present in both twins, but not the parents) in 15q11.2, a region that was grossly overrepresented among the putative post-twinning mutations in the ~1,100 MZ pairs from chapter 5 (90 out of 152). This is either an unusually unstable genomic region, or our approach is for some reason more sensitive for detecting CNVs in this region. Eleven of the putative post-twinning 15q11.2 mutations for qPCR replication, among which two (from the same twin pair) were validated; a higher validation rate than the qPCR replication experiments in the 50 selected MZ pairs of chapter 4, where we validated 2 out of 18 putative post-twinning mutations. The validation rate may still not seem high, but the rest of the 79 putative mutations in 15q11.2 are significantly more often obtained from blood samples, of which the majority are adults, as opposed to the buccal samples, that are mainly from children. In addition, blood samples with a putative 15q11.2 mutation are also from significantly older twins than blood samples in which we did not detect a putative discordance. This is in line with recent findings of *de novo* mutation rates increasing with age.[18;19] Given this association with age, it was not expected that the one twin pair in which the discordance was validated with qPCR was a 13 year old twin pair with a buccal sample, which did not show any striking phenotypic discordance. We are currently preparing for another round of qPCR replications among the rest of the putative somatic mutations, which should give us a better idea on somatic mutation rates for relatively large (>100 kb) CNVs, and on whether large *de novo* CNV mutations in the 15q11.2 region, which have been associated with a wide range of psychiatric and cognitive disorders, can be phenotypically tolerated.

Behavior → Genetics: Ancestral Influences on Genetic Variation

The molecular structure of DNA, and thereby the molecular mechanisms underlying the coding of genetic information and inheritance, was discovered in 1953.[20-22] This discovery and the major technological advances that followed were largely responsible for commencing the so-called Genomic Era, which is still ongoing, and paved the way for mapping genetic variation through genome-wide genotyping. These advances led to a gold rush for causal genetic variants, resulting in large international collaborations analyzing vast amounts of genomes through genome-wide association studies (GWASs), where we are reaching the maximum of practically feasible sample sizes.[23] Some psychiatric traits, such as depressive or anxiety disorders, turn out to be among the most difficult complex traits with respect to finding consistent genetic association signals with the current available sample sizes.[24; 25] In the understandable hurry to disentangle the genetic etiology of complex traits and disease, understanding the larger patterns of genetic variation (and where they come from) within populations contributing to these meta- and mega-analyses may not get the urgency it deserves.

Part III - Population Genetics: Genomic Structure of the Netherlands

Because of the small geographic area, the Netherlands is among the regions sampled from when a genetically homogeneous sample is desired.[26-28] However, even in such small areas considerable genetic heterogeneity may exist. We used a principal component analysis (PCA) on genome-wide single nucleotide polymorphisms (SNPs) to summarize the largest patterns of genetic variation in the Netherlands. Principal components (PCs) reflecting ancestry differences have been shown to effectively correct for population stratification in genetic association studies, i.e., allele frequency differences due to systematic ancestry differences, which can confound signals in GWASs if they correlate with the studied phenotype.[29] Extracting reliable signals for ancestry differences from a relatively homogeneous dataset required a different and more stringent approach than in a dataset consisting of multiple more differentiated populations (e.g., the 1000 Genomes dataset). This allowed for three Dutch ancestry-informative PCs capturing relatively small ancestry differences to be extracted from microarray SNP data, which showed clear and significant correlations with geography (a North-South, an East-West, and a Middle-Band distribution), and are not independent of complex traits. SNP data extracted from the *Genome of the Netherlands* (GoNL) dataset, a Next Generation Sequence (NGS) dataset representing participants from all regions from the Netherlands from multiple Biobanks, show the same geographic distributions

when projecting the PCs derived from the Affymetrix 6 SNPs. When conducting a PCA on common indels from the GoNL dataset, the North-South and East-West cline are also clearly visible. The absence of the "Middle-Band cline" among the indel PCs might be explained by a lack of power, since that only emerged in 4,441 unrelated subjects when LD was minimized beyond the recommended levels, and when outliers due to a non-Dutch ethnic/ancestral background and non-genetic artifacts were removed. The GoNL dataset was considerably smaller (490 unrelated individuals), and non-genetic artifacts related to quality differences were still present in PCs derived from indels, which is not entirely unexpected given that they are harder to reliably measure than SNPs. Results from the GoNL dataset do confirm however that common indels are largely in LD with common SNPs included in microarrays, which means that their variation is also largely captured by common SNPs, which supports the reliability of indels imputed from microarray SNPs for GWASs.

Height is significantly correlated with the PC explaining most variation, which captures the Dutch North-South cline. Height is a classical example of a complex trait and has long served as a model for the investigation of the genetic etiology underlying complex traits[30-33]: it is influenced by many genetic variants, relatively easy to measure reliably on a large scale, and therefore turned out as one of the more successful complex traits when it comes detecting causal genetic variants.[34] We show that improving measures for ancestry by decreasing LD results in PCs that are more effective in correcting for inflated statistics in GWASs on height that are caused by systematic North-South ancestry differences, which represent more than just causal SNPs. It is very likely that the causal variants underlying the height differences between North and South differ between these regions because of selection pressures. This is the case for European North-South height differences as well.[35] The European North-South cline correlates highly with the Dutch North-South cline and shows several other similarities, such as a significant correlation with genome-wide homozygosity due to the serial founder effect that was initiated with the ancient successive out-of-Africa migrations, and selection pressures on many of the same genes. This does not necessarily mean that these events (north-ward migration and diversifying selection) took place within the borders of the Netherlands; it could also be that Southern Europeans have migrated more to the South of the Netherlands, and/or Northern Europeans more to the Northern parts.

By comparing the distribution of alleles between subpopulations identified by the ancestry-informative PCs, we were able to detect a relatively large number of SNPs in genes under diversifying selection pressures. These variants likely had relatively large effects on phenotypes that were important for survival and/or reproductive success in Dutch ancestors. The variant with the strongest signal is

the key determinant of human blue/brown eye color, which also shows significant selection pressures in Europe,[36; 37] where (like in the Netherlands) blue eyes are more prevalent in Northern regions.[38] Phenotypes under diversifying selection likely included brain-related traits, as several important and well-known brain genes showed significant signals, such as SLC6A4 (a.k.a. SERT; encodes the serotonin transporter), BDNF (encodes the brain-derived neurotrophic factor), NRXN3 (encodes neurexin-3-alpha), GRIN2A (encodes a subunit for the NMDA receptor), GRM7 (encodes a metabotropic glutamate receptor), and AUTS2 (autism susceptibility candidate 2). In addition, genes involved in neurotransmission of nervous tissue were significantly overrepresented among selection pressure signals overall. SLC6A4, one of the most studied genes in candidate gene studies in the context of psychiatric traits (especially regarding depression), showed recent selection pressures in other populations as well.[39-41] It remains to be elucidated why we are able to pick up these signals, while GWASs on behavioral and psychiatric traits rarely show significant associations in these classical candidate genes, especially when using sample sizes close to ours (<5,000 subjects). A possible explanation could be that nature "measures" these traits across many more generations than scientists do, or perhaps scientists and nature have a different opinion about what they call a "disease" or an important trait. One thing seems clear however: the ancestry differences for these genes are larger than the causal effects they have on current measurable behavioral and psychiatric traits.

Part IV - Runs of Homozygosity: How Ancestral Behaviors Influence Current Genetic Variation

Genetic variants in the Netherlands also show non-causal relationships with behavioral traits. The geographic distributions captured by the ancestry-informative PCs are visible because of relatively low levels of migration rates (and thus low levels of gene flow) in the relatively recent Dutch history. Non-random mating through assortment and migration however has created non-random genetic differences between certain phenotypic groups. The Dutch North-South PC shows the same geographic distribution of Protestants and Catholics in the last couple of centuries. When populations are separated geographically and socially for longer periods of time, they will eventually diverge from each other genetically as well (largely due to genetic drift). Protestants and Catholics have been strongly segregated in the Netherlands for centuries, which is also visible in the high levels of assortment during these times and in our dataset (spouse correlation = .73). We are likely picking up the genetic consequences of that with the North-South PC, which is also the PC that shows the strongest assortment (spouse correlation = .56, which is >3 times larger than the spouse correlation of the other two ancestry-informative PCs). Non-religious individuals however have rapidly increased in numbers during

the last 50 years. They are less restricted to mates with similar ancestries, as they are more likely to migrate and are free of denominational restrictions in their partner choice. This makes non-religious individuals less related to their mates than religious individuals, leading to offspring with significantly smaller genomic proportions of runs of homozygosity (ROHs, i.e., consecutive homozygous SNPs). We see a similar effect for education: higher educated individuals are more likely to migrate and pick a higher educated partner who is more likely to have come from a different geographic region. This makes higher educated spouse pairs genetically less similar to each other than lower educated spouse pairs, leading to offspring carrying significantly fewer/shorter ROHs. This non-random mating behavior caused the ancestry-informative PCs (which represent the largest patterns of genome-wide variation when excluding patterns due to LD) to be more mixed in the offspring, which caused these PCs to show lower correlations with geography in non-religious and higher educated individuals. These systematic differences affect many non-causal variants, which can lead to spurious associations with traits related to religion and education, and this can lead to wrong conclusions, especially when the spurious association happens to coincide with existing hypotheses. The ancestry-informative PCs did not capture systematic homozygosity differences of this subtle nature sufficiently to account for them in association analyses. These effects would probably be more absent in the GoNL dataset, where spouse pairs were not selected randomly with respect to shared ancestry (spouse pairs were chosen to be born in the same province, which was detectable through inflated spouse correlations for ancestry-informative PCs). This likely decreased the chance for sequencing higher educated or non-religious trios, since their parents are more likely to share different ancestral backgrounds.

The expected consequences of these confounding effects are not much different than those following from the classical population stratification issues, like with height, where the assumption that GWASs are conducted in genetically homogeneous samples is violated. This can especially be a problem when these systematic genetic differences within the sample are related to the trait under investigation. Religion and educational attainment are associated with many other psychiatric, behavioral, and cognitive traits of interest. We detect the confounding effects using a genome-wide measure of homozygosity and PCs summarizing genome-wide variation (both computed using ~500k directly measured SNPs), but have yet to explore the impact this might have on single-SNP associations in GWASs on these traits. Much larger sample sizes are needed to investigate this further, and it is important that the presence of these effects is also investigated in other populations contributing to meta-analyses. Educational attainment needed >125k subjects from 57 cohorts to find three borderline significant associations with p-values ranging from 2.1×10^{-9} to 4.2×10^{-9}.[42] Admittedly, educational

attainment is a difficult heterogeneous phenotype to analyze, for example because of cross-cultural and cross-generation differences in availability and access to education. It is one of the few practical proxies for IQ however that is able to reach sample sizes required to detect signals from the many alleles with weak effects among the millions of likely non-contributing variants. It is worth investigating whether these confounding effects vary between populations with different social, historical, and demographic backgrounds, and whether statistical power can be improved in meta-analyses by accounting for this non-random variation. A recent study on the relationship between MDD and F_{roh} that included nine datasets from five countries found MDD-F_{roh} associations with opposite directions of effects between datasets that were consistent across countries.[43] The association in the Dutch dataset from chapter 8, of which a considerable part was included in that study, disappeared when accounting for religious affiliation, which was significantly associated with MDD in our dataset. When dealing with behavioral, cognitive, or psychiatric traits that may correlate with social, historical, and/or demographic factors, alternative explanations for genetic associations should be considered before concluding the finding of risk increasing genetic variants. In order to do this, we must have a good understanding of what drives genetic variation on a population level.

Behavior ↔ Genetics: Main Conclusions and Future Perspective on the Field

Part I: Twin studies aiming to estimate genetic and environmental influences on complex behavioral and psychiatric traits should be aware of the impact of rater effects in their estimates. Heritability estimates may show considerable variation depending on who provides the report and on how many individuals provide the report. Including different raters when collecting phenotypic measurements may provide different perspectives on certain traits that may all turn out to contain useful information.

Part II: Searching for genetic differences between MZ twins is a promising endeavor that may lead to the discovery of novel causal genetic variants. Microarray data is not yet optimal for this goal, as it is only suited for detecting relatively large CNVs and contains a considerable amount of false positive signals. Sequencing phenotypically discordant twins may be a more suited approach, since that makes it possible to scan the entire genome without ascertainment biases that microarrays may have in their probe selection, and are better suited for finding smaller genetic differences as well.

Part III: I would recommend consortia contributing to large GWAS meta- and mega-analyses to explore the main patterns of variation in their population, and also explore how social, historical, and demographic factors shaped their population structure. Especially in more homogeneous populations, one should make sure that PCAs conducted for this goal are carried out with care, by making sure that patterns of variation such as LD patterns and non-genetic artifacts are accounted for, since those can be larger than the relatively small ancestry differences within their population. These ancestry differences may be associated with phenotypic measures of interest. Even though these ancestry differences are relatively small, they may be greater than the very small effects the many individual SNPs have on complex traits.

Part IV: The presence of more recent ancestral influences that may have led to non-random mating should also be explored, because they may create systematic differences in genome-wide homozygosity due to systematic differences in parental relatedness. Heritable behavioral traits contributing to these differences may be associated with additional traits of interest. This is especially important for genetic studies on the effects of inbreeding, but may also result in less confounded GWAS analyses.

References

1. Galton, F. (1876). The history of twins, as a criterion of the relative powers of nature and nurture. *The Journal of the Anthropological Institute of Great Britain and Ireland* 5, 391-406.
2. Rende, R.D., Plomin, R., and Vandenberg, S.G. (1990). Who discovered the twin method? *Behavior Genetics* 20, 277-285.
3. Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics* 3, 872-882.
4. Slater, E. (1936). The Inheritance of Manic-depressive Insanity (Section of Psychiatry). *Proceedings of the Royal Society of Medicine* 29, 981-990.
5. Plomin, R., Owen, M.J., and McGuffin, P. (1994). The genetic basis of complex human behaviors. *Science* 264, 1733-1739.
6. Hur, Y.-M., and Craig, J.M. (2013). Twin registries worldwide: an important resource for scientific research. *Twin Research and Human Genetics* 16, 1-12.
7. Bottas, A., Cooke, R.G., and Richter, M.A. (2005). Comorbidity and pathophysiology of obsessive–compulsive disorder in schizophrenia: Is there evidence for a schizo-obsessive subtype of schizophrenia? *Journal of Psychiatry and Neuroscience* 30, 187.
8. Kan, K.J., Dolan, C.V., Nivard, M.G., Middeldorp, C.M., van Beijsterveldt, C.E., Willemsen, G., and Boomsma, D.I. (2013). Genetic and environmental stability in attention problems across the lifespan: evidence from the Netherlands twin register. *Journal of the American Academy of Child and Adolescent Psychiatry* 52, 12-25.
9. Kan, K.-J., van Beijsterveldt, C.E.M., Bartels, M., and Boomsma, D.I. (2014). Assessing genetic influences on behavior: Informant and context dependency as illustrated by the analysis of Attention Problems. *Submitted*.
10. Kondo, S., Schutte, B.C., Richardson, R.J., Bjork, B.C., Knight, A.S., Watanabe, Y., Howard, E., de Lima, R.L.F., Daack-Hirsch, S., et al. (2002). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nature genetics* 32, 285-289.
11. Cook Jr, E.H., and Scherer, S.W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature* 455, 919-923.
12. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., et al. (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445-449.
13. Xu, B., Roos, J.L., Levy, S., Van Rensburg, E., Gogos, J.A., and Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature Genetics* 40, 880-885.
14. Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. *Nature Genetics* 39, S43-S47.
15. van Ommen, G.-J.B. (2005). Frequency of new copy number variation in humans. *Nature Genetics* 37, 333-334.
16. Baker, M. (2010). Genomics: the search for association. *Nature* 467, 1135-1138.
17. Maiti, S., Kumar, K.H.B.G., Castellani, C.A., O'Reilly, R., and Singh, S.M. (2011). Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. *PLoS One* 6, e17125.
18. Forsberg, L.A., Rasi, C., Razzaghian, H.R., Pakalapati, G., Waite, L., Thilbeault, K.S., Ronowicz, A., Wineinger, N.E., Tiwari, H.K., et al. (2012). Age-related somatic structural changes in the nuclear genome of human blood cells. *Am J Hum Genet* 90, 217-228.
19. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature* 488, 471-475.
20. Franklin, R.E., and Gosling, R.G. (1953). Molecular configuration in sodium thymonucleate. *Nature* 171, 740-741.
21. Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids. *Nature* 171, 737-738.
22. Wilkins, M.H.F., Stokes, A.R., and Wilson, H.R. (1953). Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. *Nature* 171, 738-740.
23. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery.

The American Journal of Human Genetics 90, 7-24.
24. Burmeister, M., McInnis, M.G., and Zöllner, S. (2008). Psychiatric genetics: progress amid controversy. *Nature Reviews Genetics* 9, 527-540.
25. Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics* 13, 537-551.
26. Rietschel, M., Mattheisen, M., Degenhardt, F., Kahn, R.S., Linszen, D.H., van Os, J., Wiersma, D., Bruggeman, R., Cahn, W., et al. (2011). Association between genetic variation in a region on chromosome 11 and schizophrenia in large samples from Europe. *Molecular Psychiatry* 17, 906-917.
27. Zeegers, M.P., Khan, H.S., Schouten, L.J., van Dijk, B.A., Goldbohm, R.A., Schalken, J., Shajahan, S., Pearlman, A., Oddoux, C., et al. (2010). Genetic marker polymorphisms on chromosome 8q24 and prostate cancer in the Dutch population: DG8S737 may not be the causative variant. *European Journal of Human Genetics* 19, 118-120.
28. Buizer-Voskamp, J.E., Muntjewerff, J.-W., Strengman, E., Sabatti, C., Stefansson, H., Vorstman, J.A., and Ophoff, R.A. (2011). Genome-wide analysis shows increased frequency of copy number variation deletions in Dutch schizophrenia patients. *Biological Psychiatry* 70, 655-662.
29. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38, 904-909.
30. Fisher, R.A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399-433.
31. Galton, F. (1886). Hereditary stature. *Nature* 33, 317.
32. Hemani, G., Yang, J., Vinkhuyzen, A., Powell, J.E., Willemsen, G., Hottenga, J.-J., Abdellaoui, A., Mangino, M., Valdes, A.M., et al. (2013). Inference of the Genetic Architecture Underlying BMI and Height with the Use of 20,240 Sibling Pairs. *The American Journal of Human Genetics* 93, 865-875.
33. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42, 565-569.
34. Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832-838.
35. Turchin, M.C., Chiang, C.W., Palmer, C.D., Sankararaman, S., Reich, D., and Hirschhorn, J.N. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics* 44, 1015-1019.
36. Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research* 20, 393-402.
37. McEvoy, B.P., Montgomery, G.W., McRae, A.F., Ripatti, S., Perola, M., Spector, T.D., Cherkas, L., Ahmadi, K.R., Boomsma, D., et al. (2009). Geographical structure and differential natural selection among North European populations. *Genome Research* 19, 804-814.
38. Frost, P. (2006). European hair and eye color: A case of frequency-dependent sexual selection? *Evolution and Human Behavior* 27, 85-103.
39. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. *PLoS Biology* 4, e72.
40. Crespi, B., Summers, K., and Dorus, S. (2007). Adaptive evolution of genes underlying schizophrenia. *Proceedings of the Royal Society B: Biological Sciences* 274, 2801-2810.
41. Gelernter, J. (2014). SLC6A4 polymorphism, population genetics, and psychiatric traits. *Human Genetics*, 1-3.
42. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., et al. (2013). GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science* 340, 1467-1471.
43. Power, R.A., Keller, M.C., Ripke, S., Abdellaoui, A., Wray, N.R., Sullivan, P.F., and Breen, G. (2014). A recessive genetic model and runs of homozygosity in major depressive disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*.

SAMENVATTING

Dit proefschrift bestaat uit vier delen die elk het verslag van twee onderzoeken bevatten. Het eerste deel gaat over de erfelijkheid van een verzameling psychiatrische symptomen die betrekking hebben op verminderde organisatie van je gedachten (“Thought Problems” in het Engels). De erfelijkheid, dat wil zeggen de mate waarin verschillen tussen mensen worden beinvloed door genetische verschillen, werd geschat met een tweelingendesign. Hierin wordt de gelijkenis tussen een- en twee-eiige tweelingen voor de eigenschap (het ‘fenotype’) gebruikt om het belang van genetische en omgevingsinvloeden te schatten. Dit design berust op de aanname dat eeneiige tweelingen genetisch identiek zijn, terwijl twee-eiige tweelingen ongeveer 50% delen van alle genen die in de populatie verschillen. In het tweede deel van het proefschrift wordt gezocht naar zeldzame mutaties die genetische verschillen tussen eeneiige tweelingen veroorzaken, en wordt bekeken of dergelijke mutaties ook tot fenotypische verschillen kunnen leiden. In het derde deel worden patronen van genetische variatie in Nederland in kaart gebracht en gebruikt om gevolgen van de evolutionaire geschiedenis bloot te leggen. In deel vier tenslotte wordt beschreven hoe het gedrag van (voor)ouders (met name via partnerkeuze en migratie) genetische variatie van de huidige bevolking beïnvloedt.

Deel I - Tweelingstudies: Beoordelaar effecten, meetinvariantie, en longitudinale erfelijkheid van “Thought Problems”

“Thought Problems” (TP) wordt gemeten door een verzameling vragen die statistisch met elkaar samenhangen en die symptomen meten uit verscheidene psychiatrische aandoeningen: hallucinaties, obsessief-compulsieve symptomen, vreemde gedachtes en gedragingen, en neigingen tot zelfbeschadiging.

In hoofdstuk 2 zien we dat ongeveer de helft van de individuele verschillen in TP in een groep van ~9000 7-jarige tweelingparen verklaard kunnen worden door genetische factoren. De erfelijkheid werd geschat door ouderbeoordelingen van de symptomen (zowel van vaders als van moeders) voor eeneiige tweelingen te vergelijken met die voor twee-eiige tweelingen. Gegeven de aanname dat eeneiige tweelingen genetisch identiek zijn, moeten eeneiige tweelingen voor erfelijke fenotypes meer gelijkenissen vertonen dan twee-eiige tweelingen. TP werd opgedeeld in twee componenten waarvoor genetische en omgevingsinvloeden werden geschat: het gedeelte van het fenotype waarover beide ouders het eens zijn (wat 67% van de variantie verklaarde, waarvan 76% erfelijk), en het gedeelte dat uniek door beide ouders wordt gerapporteerd (33% van de variantie: voor de moeder is 61% hiervan erfelijk, voor de vader 65%). Het feit dat het deel van de symptomen waar de ouders het niet over eens waren ook erfelijk is, geeft aan dat het unieke perspectief van beide ouders mede gebaseerd is op de symptomen van het kind, en niet op rater bias of op meetfout. Het unieke perspectief van de ouders wordt ook beïnvloed door “gedeelde omgevingsinvloeden”, die wel kunnen

samenhangen met rater bias (moeder: 13%; vader: 13%). De resterende individuele verschillen konden verklaard worden door unieke omgevingsinvloeden, die ook meetfouten kunnen bevatten.

TP wordt in hoofdstuk 3 verder geanalyseerd in tweelingen en familieleden van 12 – 59 jaar oud. Een exploratieve factor analyse bevestigt dat er één onderliggend construct wordt gemeten door TP. De vragenlijst is meetinvariant over leeftijd en sekse, wat betekent dat hetzelfde construct wordt gemeten in mannen en vrouwen van verschillende leeftijden. Met het tweelingmodel, uitgebreid met informatie van broers en zussen, werd de longitudinale erfelijkheid geschat. Individuele verschillen in TP worden in adolescenten (12-18 jaar), jong volwassenen (19-27 jaar) en oudere volwassenen (28-59 jaar) voor ongeveer 37% beïnvloed door genetische invloeden (lager dan in kinderen). Vanaf de adolescentie beïnvloeden dezelfde genen TP in alle drie de leeftijdsgroepen, met uitzondering van een additionele genetische component, die pas bij volwassenen een rol begint te spelen. De resterende individuele verschillen kunnen verklaard worden door unieke omgevingsfactoren (die deels ook meetfouten kunnen bevatten).

Deel II - Copy Number Varianten: Mutaties na de splitsing van eeneiige tweelingen

In hoofdstuk 4 wordt de DNA sequentie van 50 eeneiige tweelingparen bestudeerd die geselecteerd zijn op aandachtsproblemen uit een groep van ~3200 eeneiige tweelingparen: 17 concordante paren met aandachtsproblemen, 22 concordante paren zonder aandachtsproblemen en 11 discordante paren. Aandachtsproblemen zijn op meerdere leeftijden gemeten met een schaal die een voorspellende waarde heeft voor ADHD en net als ADHD een hoge erfelijkheid heeft. In deze groep tweelingen werden *Copy Number Varianten* (CNVs) gemeten (bij 25 paren ook in de ouders). CNVs zijn DNA segmenten die een variabel aantal keren aanwezig zijn (“normaal” aantal kopieen is twee: één van allebei de ouders, maar bij CNVs kan het aantal kopieën ook nul, één, drie of meer zijn). Er is gezocht naar CNV mutaties die voor of na de splitsing van de bevruchte eicel plaats vonden, en er is een analyse uitgevoerd om te testen of CNVs geassocieerd zijn met aandachtsproblemen. Er is één mutatie gevonden die voor de splitsing van de eicel plaats vond (aanwezig bij beide tweelingen, maar niet bij de ouders) in een tweelingpaar zonder aandachtsproblemen. Er zijn twee mutaties gevonden die na de splitsing hebben plaatsgevonden (aanwezig in een van de tweelingen) in een concordant tweelingpaar met aandachtsproblemen (een deletie op chromosoom 4) en in de persoon met aandachtsproblemen in een discordant tweelingpaar (een duplicatie op chromosoom 17). Deze CNV mutaties liggen op plekken die overlappen met genen die eerder in verband zijn gebracht met psychiatrische aandoeningen. Behalve meer gedragsproblemen hadden de dragers van de mutatie

ook een lager geboortegewicht dan hun tweelingbroer. Een analyse bij alle 50 tweelingparen liet zien dat kinderen met meer aandachtsproblemen, verspreid over het hele genoom gemiddeld grotere CNVs hadden, vooral bij CNVs die met genen overlapten.

In hoofdstuk 5 is het hele genoom gescand voor CNV verschillen binnen eeneiige tweelingen bij ~1100 niet geselecteerde eeneiige tweelingparen. Van ongeveer de helft van de tweelingparen kwam het DNA uit wangcellen (voornamelijk kinderen), en van de andere helft uit bloed (voornamelijk volwassenen). Er zijn 153 mogelijke CNV mutaties gevonden, waarvan de meerderheid uit dezelfde instabiele regio: 15q11.2. De meerderheid hiervan werd geobserveerd in DNA uit bloed (dus bij de volwassenen). De 15q11.2 mutaties die in bloed zijn waargenomen werden significant vaker in oudere tweelingparen gemeten. Een eerste selectie van 11 CNV mutaties (bij kinderen en volwassenen) zijn met qPCR ter validatie opnieuw gemeten, waarvan er uiteindelijk 2 uit wangcellen door qPCR zijn geconfirmeerd binnen hetzelfde gezonde 13-jarige tweelingpaar. Er zijn geen grote fenotypische verschillen binnen het tweelingpaar waargenomen in longitudinale vragenlijst gegevens die werden verzameld van 1 tot 21 jaar.

Deel III - Populatiegenetica: De genetische opmaak van Nederland

Het autochtone deel van de Nederlandse bevolking wordt over het algemeen als een genetisch homogene populatie gezien. In hoofdstuk 6 bekijken we hoe homogeen deze populatie daadwerkelijk is en worden patronen van genetische variatie in Nederland in kaart gebracht met een principale componenten analyse (PCA) op 500,000 *Single Nucleotide Polymorfismen* (SNPs). SNPs zijn vaak voorkomende genetische varianten en bestaan uit een verandering in een enkele nucleotide (DNA is opgebouwd uit vier verschillende nucleotiden met de nucleobasen adenine, thymine, guanine en cytosine, afgekort als A, T, G en C). SNP varianten in een populatie zijn verspreid over het hele genoom. Een PCA is een statistische methode die in een groot aantal gemeten variabelen (SNPs in dit geval) de grootste patronen van variatie samenvat in zogenaamde principale componenten (PCs). Deze analyses zijn uitgevoerd bij 4441 ongerelateerde Nederlandse individuen en 1014 ongerelateerde individuen uit 14 verschillende populaties uit de hele wereld (de 1000 Genomes dataset). Het filteren van SNPs die hoog met elkaar zijn gecorreleerd (oftewel hoog in linkage disequilibrium [LD] met elkaar zijn) had niet veel invloed op de PCs van de 1000 Genomes dataset (waarin de populaties relatief sterker van elkaar verschillen), maar zorgde er in de Nederlandse dataset voor dat de PCs aanzienlijk beter de genetische verschillen tussen Nederlanders oppikte. Het minimaliseren van LD verhoogde de correlaties tussen PCs en geografie binnen Nederland significant en resulteerde in drie PCs die genetische afkomst reflecteren: 1) de Noord-Zuid PC, die de verschillen

oppikt tussen de noordelijke provincies en de provincies ten zuiden van de drie grote rivieren (en in de Randstad zijn die noord-zuid verschillen meer met elkaar vermengd), 2) de Oost-West PC, die de verschillen tussen het noordoosten en de rest van het land reflecteert, en 3) de Midden-Strook PC, waaruit verschillen tussen een strook door het midden van het land en de rest van het land te zien zijn. De Noord-Zuid PC liet verscheidene overeenkomsten zien met Europese verschillen in genetische afkomst tussen noord en zuid: 1) een correlatie van .66 met de 1000 Genomes PC die Noord-Europa van Zuid-Europa van elkaar onderscheidt, 2) een significante correlatie met genoom-wijde homozygositeit (noorden = meer homozygoot), 3) een significante correlatie met lichaamslengte (noord = langer), 4) een signaal van selectiedruk op de SNP die bepalend is voor bruin/blauwe oogkleur (noord = meer blauwe ogen). De PCs konden ook gebruikt worden om SNPs te detecteren die onder selectiedruk hebben gestaan. Onder selectiedruk komen genetische varianten meer te verschillen tussen de Nederlandse subpopulaties dan het overgrote gedeelte van het genoom. Naast het signaal uit HERC2 (het gen verantwoordelijk voor bruin/blauw oogkleur, en het sterkste signaal) waren er nog 544 SNPs (uit 184 genen) die een signaal van selectiedruk suggereerden. Genen die een rol spelen in het brein waren significant oververtegenwoordigd in deze signalen.

In hoofdstuk 7 wordt de variatie van zeldzamer en moeilijker te meten genetische varianten verkend: indels (inserties en deleties < 20 baseparen) en grotere deleties (20 – 10,000 baseparen). Deze zijn in kaart gebracht met Next Generation Sequencing en zijn afkomstig uit een viertal Nederlandse biobanken in het *Genome of the Netherlands* (GoNL) project. Er zijn PCAs uitgevoerd op indels en op grotere deleties bij 490 ongerelateerde Nederlandse individuen met het doel variatie binnen Nederland in kaart te brengen. Indels lieten drie PCs zien die genetische afkomst reflecteren, en grotere deleties vijf. De indel PC die de meeste variatie verklaarde in genetische afkomst liet geografisch dezelfde Noord-Zuid distributie zien als de SNPs in hoofdstuk 6. De andere twee indel PCs lieten een geografische verdeling zien die sterk leek op de Oost-West SNP PC. De vijf PCs van de PCA op grotere deleties lieten significante maar lagere correlaties met de SNP PCs uit hoofdstuk 6 zien (significante correlaties tussen .16 en .30). Er was geen duidelijke geografische verdeling zien op de Nederlandse kaart (al lieten ze wel significante correlaties tussen .18 en .26 zien met geografie gebaseerd op geboorteplaats). PCs van grotere deleties zijn ook de enige PCs die geen significante correlatie binnen ouderparen laten zien. Het is nog niet duidelijk of de grotere deleties andere genetische afkomstverschillen oppikken, of dat de signalen zwakker zijn vanwege de relatief weinig betrouwbaar gemeten grotere deleties. Een aantal indel en deletie PCs lieten net als de Noord-Zuid SNP PC overeenkomsten zien met Europese Noord-Zuid verschillen (significante correlaties met de Europese Noord-Zuid PC

uit de 1000 Genomes dataset, homozygositeit, lichaamslengte en haarkleur).

Deel IV – Runs of Homozygosity: De invloed van het gedrag van (voor)ouders op de huidige genetische samenstelling

In hoofdstuk 8 worden opeenvolgende reeksen van homozygote varianten onderzocht, ook wel *Runs of Homozygosity* (ROHs) genoemd. Een stuk genoom is homozygoot als er voor dat stuk identieke kopieën van genetische varianten op beide chromosomen aanwezig zijn. Omdat dit betekent dat een kind van beide ouders dezelfde varianten heeft geerfd, is de proportie van het genoom dat uit ROHs bestaat (ook F_{roh} genoemd) groter naarmate ouders meer verwant zijn. F_{roh} laat een significante associatie met religiositeit zien, wat waarschijnlijk verklaard kan worden door demografische en historische factoren. De geografische verdeling van de Noord-Zuid PC in hoofdstuk 6 (het sterkste patroon van verschillen in genetische afkomst binnen Nederland) komt overeen met de geografische verdeling van de twee religieuze groepen in Nederland (katholieken en protestanten) die ongeveer vier eeuwen vrij stabiel is gebleven. In de huidige samenleving is dezelfde distributie nog steeds zichtbaar, maar met een toename van niet-religieuze individuen door de toenemende secularisatie in de afgelopen halve eeuw. De correlatie binnen ouderparen voor religie is zeer sterk (.73), in lijn met de hoge correlatie tussen ouders voor de Noord-Zuid PC (.56) en de historische documentatie van de partnerkeuze van katholieken en protestanten ("*twee geloven op één kussen, daar slaapt de duivel tussen*"). Post-hoc analyses wezen uit dat de associatie tussen religie en F_{roh} verklaard kan worden door het feit dat het niet-religieuze deel van de bevolking minder homozygote varianten heeft. Dit komt hoogstwaarschijnlijk omdat niet-religeuzen een minder beperkte partnerkeuze hebben dan katholieken en protestanten, waardoor genetische verschillen met hun partner groter kunnen zijn. In de niet-religieuze groep waren significant meer mensen met een klinische depressie, waardoor er een significante associatie leek te zijn tussen F_{roh} en depressie. Deze indirecte associatie verdwijnt na corrigeren voor religie. De PCs zoals beschreven in hoofdstuk 6 waren niet voldoende om te corrigeren voor deze oneigenlijke correlatie.

In hoofdstuk 9 wordt een significant verband tussen F_{roh} en opleidingsniveau gerapporteerd. Deze associatie is niet te wijten is aan causale ROHs die de kans verkleinen op een hoger opleidingsniveau, maar aan het gedrag van de (voor)ouders. F_{roh} in het nageslacht is veel sterker geassocieerd met het opleidingsniveau van hun ouders dan van het nageslacht zelf: ouders met een hoger opleidingsniveau hebben kinderen met een lagere F_{roh}. Voor ouders met een hoger opleidingsniveau was er een grotere afstand tussen hun eigen geboorteplaats en de geboorteplaats van hun kinderen of echtgeno(o)t(e). De afstand tussen de geboorteplaats van de vader en die van de moeder liet ook een significante associatie zien met F_{roh} (grotere

afstand = lagere F_{roh}). Als er gecorrigeerd wordt voor deze afstand, verdwijnt ook het significante verband tussen het (ouderlijke) opleidingsniveau en F_{roh}. Dit geeft aan dat de associatie tussen F_{roh} en (ouderlijke) opleidingsniveau ontstaat omdat hoger opgeleide ouders vaker en verder migreren. Hoger opgeleide ouders kiezen vaker een partner die ook hoger opgeleid is, waardoor de kans groter wordt op een partner die zelf ook meer mobiel is en uit een geografische regio komt met een andere genetische achtergrond. Een andere aanwijzing hiervoor is dat kinderen van hoger opgeleide ouders een lagere correlatie laten zien tussen de PCs die genetische afkomst reflecteren en geografie.

LIST OF PUBLICATIONS

First Author

Abdel Abdellaoui*, Karin J H Verweij*, Brendan P Zietsch: *No Evidence for Genetic Assortative Mating Beyond that due to Population Stratification*. Proceedings of the National Academy of Sciences of the Unites States of America (under review). **these authors contributed equally*

Abdel Abdellaoui, Jouke-Jan Hottenga, Gonneke Willemsen, Meike Bartels, Toos van Beijsterveldt, Erik A Ehli, Gareth E Davies, Andrew Brooks, Patrick F Sullivan, Brenda W J H Penninx, Eco J de Geus, Dorret I Boomsma: *Educational Attainment Influences Genetic Variation through Migration and Assortative Mating.* PLOS ONE (under review).

Abdel Abdellaoui, Jouke-Jan Hottenga, Xiangjun Xiao, Paul Scheet, Erik A Ehli, Gareth E Davies, James J Hudziak, Dirk J A Smit, Meike Bartels, Gonneke Willemsen, Andrew Brooks, Patrick F Sullivan, Johannes H Smit, Eco J de Geus, Brenda W J H Penninx, Dorret I Boomsma: *Association Between Autozygosity and Major Depression: Stratification Due to Religious Assortment*. Behavior Genetics 2013; 43:455-467.

Abdel Abdellaoui, Jouke-Jan Hottenga, Peter de Knijff, Michel G Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, Erik A Ehli, Yueshan Hu, Gareth E Davies, James J Hudziak, Patrick F Sullivan, Toos van Beijsterveldt, Gonneke Willemsen, Eco J de Geus, Brenda W J H Penninx, Dorret I Boomsma: *Population Structure, Migration, and Diversifying Selection in the Netherlands.* European Journal of Human Genetics 2013; 21:1277-1285.

Erik A Ehli*, **Abdel Abdellaoui***, Yueshan Hu, Jouke Jan Hottenga, Mathijs Kattenberg, Toos van Beijsterveldt, Meike Bartels, Robert R Althoff, Xiangjun Xiao, Paul Scheet, Eco J de Geus, James J Hudziak, Dorret I Boomsma*, Gareth E Davies*: *De Novo and Inherited CNVs in MZ Twin Pairs Selected for Discordance and Concordance on Attention Problem*. European Journal of Human Genetics 2012; 20(10):1037-43. **these authors contributed equally*

Abdel Abdellaoui, Marleen H M de Moor, Lot M Geels, Jenny H D A van Beek, Gonneke Willemsen, Dorret I Boomsma: *Thought Problems from Adolescence to Adulthood: Measurement Invariance and Longitudinal Heritability.* Behavior Genetics 2011; 42(1):19-29.

Abdel Abdellaoui, Meike Bartels, James J Hudziak, Patrizia Rizzu, Toos Cem van Beijsterveldt, Dorret I Boomsma: *Genetic Influences on Thought Problems in 7-Year-Olds: a Twin-Study of Genetic, Environmental and Rater Effects.* Twin Research and Human Genetics 2009; 11(6):571-8.

Contributing Author

Sanja Franic, Maria M Groen-Blokhuis, Conor V Dolan, Matthijs V Kattenberg, Xiangjun Xiao, Paul A Scheet, Erik A Ehli, Gareth E Davies, Sophie van der Sluis, **Abdel Abdellaoui**, Narelle K Hansell, Nicholas G Martin, James J Hudziak, Catharina E M van Beijsterveldt, Suzanne C Swagerman, Hilleke E Hulshoff Pol, Eco J C de Geus, Meike Bartels, H Hilger Ropers, Jouke-Jan Hottenga, Dorret I Boomsma: *Intelligence: Shared Genetic Basis between Mendelian Disorders and a Polygenic Trait*. European Journal of Human Genetics (under review).

Maria M Groen-Blokhuis, Christel M Middeldorp, Kees-Jan Kan, **Abdel Abdellaoui**, Catharina E M van Beijsterveldt, Erik A Ehli, Gareth E Davies, Paul A Scheet, Xiangjun Xiao, James J Hudziak, Jouke-Jan Hottenga, Psychiatric Genomics Consortium ADHD Working Group, Ben M Neale, Dorret I. Boomsma: *Attention Deficit Hyperactivity Disorder polygenic risk scores predict Attention Problems in a Population-Based Sample of Children*. Journal of the American Academy of Child & Adolescent Psychiatry 2014 (accepted).

Karin J H Verweij, **Abdel Abdellaoui**, Juha Veijola, Sylvain Sebert, Markku Koiranen, Matthew C Keller, Marjo-Riitta Järvelin, Brendan P Zietsch: *The Association of Genotype-Based Inbreeding Coefficient with a Range of Physical and Psychological Human Traits.* PLOS ONE 2014; e103102.

Laurent Francioli, Androniki Menelaou, Sara Pulit, Clara Elbers, Pier Palamara, Morris Swertz, Freerk van Dijk, Pieter Neerincx, Patrick Deelen, Itsik Pe'er, **Abdel Abdellaoui**, Wigard Kloosterman, Mannis van Oven, Martijn Vermaat, Mingkun Li, Jeroen Laros, Mark Stoneking, Peter de Knijff, Manfred Kayser, Jan Veldink, Leonard van den Berg, Heorhiy Byelas, Johan den Dunnen, Martijn Dijkstra, Najaf Amin, Joeri van der Velde, Jouke- Jan Hottenga, Jessica van Setten, Elisabeth van Leeuwen, Alexandros Kanterakis, Mathijs Kattenberg, Lennart Karssen, Barbera van Schaik, Jan Bot, Isaac Nijman, David van Enckevort, Hailiang Mei, Vyacheslav Koval, Kai Ye, Eric-Wubbo Lameijer, Matthijs Moed, Jayne Hehir-Kwa, Robert Handsaker, Shamil Sunyaev, Mashaal Sohail, Fereydoun Hormozdiari, Tobias Marschall, Alexander Schoenhuth, Victor Guryev, Eline Slagboom, Marian Beekman, Anton de Craen, Eka Suchiman, Albert Hofman, Cornelia van Duijn, Dorret I Boomsma, Gonneke Willemsen, Bruce Wolffenbuttel, Matthieu Platteel, Qibin Li, Yingrui Li, Yuanping Du, Ruoyan Chen, Hongzhi Cao, Ning Li, Sujie Cao, Jun Wang, Jasper Bovenberg, Cisca Wijmenga, Gert-Jan van Ommen, Steven Pitts, Shobha Potluri, Purnima Sundar, Paul de Bakker: *Whole Genomes of Dutch Families, their Population Structure and Demography.* Nature Genetics 2014; 46(8): 818-825.

Fred Wright, Patrick Sullivan, Andrew Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, **Abdel Abdellaoui**, Sandra Batista, Casey Butler, Guanhua Chen, Ting-Huei Chen, Wonil

Chung, David D'Ambrosio, Paul Gallins, Min Jin Ha, Jouke-Jan Hottenga, Shunping Huang, Rick Jansen, Mathijs Kattenberg, Jaspreet Kochar, Christel Middeldorp, Ani Qu, Andrey Shabalin, Jay A Tischfield, Laura Todd, Jung-Ying Tzeng, Gerard van Grootheest, Jacqueline Vink, Qi Wang, Wei Wang, Weibo Wang, Gonneke Willemsen, Zhaoyu Yin, Yi-Hui Zhou, Eco de Geus, Johannes Smit, Brenda W J H Penninx, Dorret I Boomsma: *Heritability and Genomics of Gene Expression in Peripheral Blood*. Nature Genetics 2014; 46(5): 430-437.

Wouter Peyrot, Yuri Milaneschi, **Abdel Abdellaoui**, Patrick F. Sullivan, Jouke-Jan Hottenga, Dorret I Boomsma, Brenda W J H Penninx: *The Effect of Polygenic Risk Scores on Depression is Increased in the Presence of Childhood Trauma*. British Journal of Psychiatry 2014; 205(2): 113-119.

Gitta H Lubke, Charles Laurin, Najaf Amin, Jouke-Jan Hottenga, Gonneke Willemsen, Gerard van Grootheest, **Abdel Abdellaoui**, Lennart C Karssen, Ben A Oostra, Cornelia M van Duijn, Brenda W J H Penninx, Dorret I Boomsma: *Genome-Wide Analyses of Borderline Personality Features.* Molecular Psychiatry 2014; 19(8): 923-929.

Robert A Power, Matthew C Keller, Stephen Ripke, **Abdel Abdellaoui**, Naomi R Wray, Patrick F Sullivan, MDD PGC Working Group, Gerome Breen: *A Recessive Genetic Model and Runs of Homozygosity in Major Depressive Disorder.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 2014; 165(2):157-166.

Charlotte Huppertz, Meike Bartels, Maria M. Groen-Blokhuis, Conor V. Dolan, Marleen H.M. de Moor, **Abdel Abdellaoui**, Catharina E.M. van Beijsterveldt, Erik A. Ehli, Jouke-Jan Hottenga, Gonneke Willemsen, Xiangjun Xiao, Paul Scheet, Gareth E. Davies, Dorret I. Boomsma, James J. Hudziak, Eco J.C. de Geus: *The Dopaminergic Reward System and Leisure Time Exercise Behavior: a Candidate Allele Study.* BioMed Research International 2014; Epub.

Gibran Hemani, Jian Yang, Anna Vinkhuyzen, Joseph Powell, Gonneke Willemsen, Jouke-Jan Hottenga, **Abdel Abdellaoui**, Massimo Mangino, Ana Valdes, Sarah Medland, Pamela Madden, Andrew Heath, Anjali Henders, Dale Nyholt, Eco de Geus, Patrik Magnusson, Erik Ingelsson, Grant Montgomery, Tim Spector, Dorret I Boomsma, Nancy Pedersen, Nicholas Martin, Peter Visscher: *Inference of the Genetic Architecture Underlying BMI and Height using 20,240 Sib-Pairs.* American Journal of Human Genetics 2013; 93(5):865-875.

Dorret I Boomsma, Cisca Wijmenga, Eline P Slagboom, Morris A Swertz, Lennart C Karssen, **Abdel Abdellaoui**, Kai Ye, Victor Guryev, Martijn Vermaat, Freerk van Dijk, Laurent C Francioli, Jouke Jan Hottenga, Jeroen F J Laros, Qibin Li, Yingrui Li, Hongzhi Cao, Ruoyan Chen, Yuanping Du, Ning Li, Sujie Cao, Jessica van Setten, Androniki Menelaou, Sara L Pulit, Jayne Y Hehir-Kwa, Marian Beekman, Clara C Elbers, Heorhiy

Byelas, Anton J M de Craen, Patrick Deelen, Martijn Dijkstra, Johan T den Dunnen, Peter de Knijff, Jeanine Houwing-Duistermaat, Vyacheslav Koval, Karol Estrada, Albert Hofman, Alexandros Kanterakis, David van Enckevort, Hailiang Mai, Mathijs Kattenberg, Elisabeth M van Leeuwen, Pieter B T Neerincx, Ben Oostra, Fernanodo Rivadeneira, Eka H D Suchiman, Andre G Uitterlinden, Gonneke Willemsen, Bruce H Wolffenbuttel, Jun Wang, Paul I W de Bakker, Gert-Jan van Ommen, Cornelia M van Duijn: *The Genome of the Netherlands: Design, and Project Goals.* European Journal of Human Genetics 2014; 22:221-227.

Cornelius A Rietveld, Sarah E Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, **Abdel Abdellaoui**, Arpana Agrawal, Eva Albrecht, Behrooz Z Alizadeh, Najaf Amin, John Barnard, Sebastian E Baumeister, Kelly S Benke, Lawrence F Bielak, Jeffrey A Boatman, Patricia A Boyle, Gail Davies, Christiaan de Leeuw, Niina Eklund, Daniel S Evans, Rudolf Ferhmann, Krista Fischer, Christian Gieger, Håkon K Gjessing, Sara Hägg, Jennifer R Harris, Caroline Hayward, Christina Holzapfel, Carla A Ibrahim-Verbaas, Erik Ingelsson, Bo Jacobsson, Peter K Joshi, Astanand Jugessur, Marika Kaakinen, Stavroula Kanoni, Juha Karjalainen, Ivana Kolcic, Kati Kristiansson, Zoltán Kutalik, Jari Lahti, Sang H Lee, Peng Lin, Penelope A Lind, Yongmei Liu, Kurt Lohman, Marisa Loitfelder, George McMahon, Pedro Marques Vidal, Osorio Meirelles, Lili Milani, Ronny Myhre, Marja-Liisa Nuotio, Christopher J Oldmeadow, Katja E Petrovic, Wouter J Peyrot, Ozren Polašek, Lydia Quaye, Eva Reinmaa, John P Rice, Thais S Rizzi, Helena Schmidt, Reinhold Schmidt, Albert V Smith, Jennifer A Smith, Toshiko Tanaka, Antonio Terracciano, Matthijs JHM van der Loos, Veronique Vitart, Henry Völzke, Jürgen Wellmann, Lei Yu, Wei Zhao, Jüri Allik, John R Attia, Stefania Bandinelli, François Bastardot, Jonathan Beauchamp, David A Bennett, Klaus Berger, Laura J Bierut, Dorret I Boomsma, Ute Bültmann, Harry Campbell, Christopher F Chabris, Lynn Cherkas, Mina K Chung, Francesco Cucca, Mariza de Andrade, Philip L De Jager, Jan-Emmanuel De Neve, Ian J Deary, George V Dedoussis, Panos Deloukas, Maria Dimitriou, Guðný Eiríksdóttir, Martin F Elderson, Johan G Eriksson, David M Evans, Jessica D Faul, Luigi Ferrucci, Melissa E Garcia, Henrik Grönberg, Vilmundur Guðnason, Per Hall, Juliette M Harris, Tamara B Harris, Nicholas D Hastie, Andrew C Heath, Dena G Hernandez, Wolfgang Hoffmann, Adriaan Hofman, Rolf Holle, Elizabeth G Holliday, Jouke-Jan Hottenga, William G Iacono, Thomas Illig, Marjo-Riitta Järvelin, Mika Kähönen, Jaakko Kaprio, Robert M Kirkpatrick, Matthew Kowgier, Antti Latvala, Lenore J Launer, Debbie A Lawlor, Terho Lehtimäki, Jingmei Li, Paul Lichtenstein, Peter Lichtner, David C Liewald, Pamela A Madden, Patrik KE Magnusson, Tomi E Mäkinen, Marco Masala, Matt McGue, Andres Metspalu, Andreas Mielck, Michael B Miller, Grant W Montgomery, Sutapa Mukherjee, Dale R Nyholt, Ben A Oostra, Lyle J Palmer, Aarno Palotie, Brenda WJH Penninx, Markus Perola, Patricia A Peyser, Martin Preisig, Katri Räikkönen, Olli T Raitakari, Anu Realo, Susan M Ring, Samuli Ripatti, Fernando Rivadeneira, Igor Rudan, Aldo Rustichini, Veikko

Salomaa, Antti-Pekka Sarin, David Schlessinger, Rodney J Scott, Harold Snieder, Beate St Pourcain, John M Starr, Jae Hoon Sul, Ida Surakka, Rauli Svento, Alexander Teumer, The LifeLines Cohort Study, Henning Tiemeier, Frank JA van Rooij, David R Van Wagoner, Erkki Vartiainen, Jorma Viikari, Peter Vollenweider, Judith M Vonk, Gérard Waeber, David R Weir, H-Erich Wichmann, Elisabeth Widen, Gonneke Willemsen, James F Wilson, Alan F Wright, Dalton Conley, George Davey-Smith, Lude Franke, Patrick JF Groenen, Albert Hofman, Magnus Johannesson, Sharon LR Kardia, Robert F Krueger, David Laibson, Nicholas G Martin, Michelle N Meyer, Danielle Posthuma, A Roy Thurik, Nicholas J Timpson, André G Uitterlinden, Cornelia M van Duijn, Peter M Visscher, Daniel J Benjamin, David Cesarini, Philipp D Koellinger: *GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment*. Science 2013; 340(6139):1467-71.

Gonneke Willemsen, Jacqueline M Vink, **Abdel Abdellaoui**, Anouk den Braber, Jenny H D A van Beek, Harmen H M Draisma, Jenny van Dongen, Dennis van 't Ent, Lot M Geels, Rene van Lien, Lannie Ligthart, Mathijs Kattenberg, Hamdi Mbarek, Marleen H M de Moor, Melanie Neijts, Rene Pool, Natascha Stroo, Cornelis Kluft, H Eka D Suchiman, P Eline Slagboom, Eco J C de Geus, Dorret I Boomsma: *The Adult Netherlands Twin Register: Twenty-Five Years of Survey and Biological Data Collection*. Twin Research and Human Genetics 2013; 16(1):271-281.

Pim van der Harst, Weihua Zhang, Irene Mateo Leach, Augusto Rendon, Niek Verweij, Joban Sehmi, Dirk S Paul, Ulrich Elling, Hooman Allayee, Xinzhong Li, Aparna Radhakrishnan, Sian-Tsung Tan, Katrin Voss, Christian X Weichenberger, Cornelis A Albers, Abtehale Al-Hussani, Folkert W Asselbergs, Marina Ciullo, Fabrice Danjou, Christian Dina, Tõnu Esko, David M Evans, Lude Franke, Martin Gögele, Jaana Hartiala, Micha Hersch, Hilma Holm, Jouke-Jan Hottenga, Stavroula Kanoni, Marcus E Kleber, Vasiliki Lagou, Claudia Langenberg, Lorna M Lopez, Leo-Pekka Lyytikäinen, Olle Melander, Federico Murgia, Ilja M Nolte, Paul F O'Reilly, Sandosh Padmanabhan, Afshin Parsa, Nicola Pirastu, Eleonora Porcu, Laura Portas, Inga Prokopenko, Janina S Ried, So-Youn Shin, Clara S Tang, Alexander Teumer, Michela Traglia, Sheila Ulivi, Harm-Jan Westra, Jian Yang, Jing Hua Zhao, Franco Anni, **Abdel Abdellaoui**, Antony Attwood, Beverley Balkau, Stefania Bandinelli, François Bastardot, Beben Benyamin, Bernhard O Boehm, William O Cookson, Debashish Das, Paul I W de Bakker, Rudolf A de Boer, Eco J C de Geus, Marleen H de Moor, Maria Dimitriou, Francisco S Domingues, Angela Döring, Gunnar Engström, Gudmundur Ingi Eyjolfsson, Luigi Ferrucci, Krista Fischer, Renzo Galanello, Stephen F Garner, Bernd Genser, Quince D Gibson, Giorgia Girotto, Daniel Fannar Gudbjartsson, Sarah E Harris, Anna-Liisa Hartikainen, Claire E Hastie, Bo Hedblad, Thomas Illig, Jennifer Jolley, Mika Kähönen, Ido P Kema, John P Kemp, Liming Liang, Heather Lloyd-Jones, Ruth J F Loos, Stuart Meacham, Sarah E Medland, Christa Meisinger, Yasin Memari, Evelin Mihailov, Kathy

Miller, Miriam F Moffatt, Matthias Nauck, Maria Novatchkova, Teresa Nutile, Isleifur Olafsson, Pall T Onundarson, Debora Parracciani, Brenda W Penninx, Lucia Perseu, Antonio Piga, Giorgio Pistis, Anneli Pouta, Ursula Puc, Olli Raitakari, Susan M Ring, Antonietta Robino, Daniela Ruggiero, Aimo Ruokonen, Aude Saint-Pierre, Cinzia Sala, Andres Salumets, Jennifer Sambrook, Hein Schepers, Carsten Oliver Schmidt, Herman H W Silljé, Rob Sladek, Johannes H Smit, John M Starr, Jonathan Stephens, Patrick Sulem, Toshiko Tanaka, Unnur Thorsteinsdottir, Vinicius Tragante, Wiek H van Gilst, L Joost van Pelt, Dirk J van Veldhuisen, Uwe Völker, John B Whitfield, Gonneke Willemsen, Bernhard R Winkelmann, Gerald Wirnsberger, Ale Algra, Francesco Cucca, Adamo Pio d'Adamo, John Danesh, Ian J Deary, Anna F Dominiczak, Paul Elliott, Paolo Fortina, Philippe Froguel, Paolo Gasparini, Andreas Greinacher, Stanley L Hazen, Marjo-Riitta Jarvelin, Kay Tee Khaw, Terho Lehtimäki, Winfried Maerz, Nicholas G Martin, Andres Metspalu, Braxton D Mitchell, Grant W Montgomery, Carmel Moore, Gerjan Navis, Mario Pirastu, Peter P Pramstaller, Ramiro Ramirez-Solis, Eric Schadt, James Scott, Alan R Shuldiner, George Davey Smith, J Gustav Smith, Harold Snieder, Rossella Sorice, Tim D Spector, Kari Stefansson, Michael Stumvoll, W H Wilson Tang, Daniela Toniolo, Anke Tönjes, Peter M Visscher, Peter Vollenweider, Nicholas J Wareham, Bruce H R Wolffenbuttel, Dorret I Boomsma, Jacques S Beckmann, George V Dedoussis, Panos Deloukas, Manuel A Ferreira, Serena Sanna, Manuela Uda, Andrew A Hicks, Josef Martin Penninger, Christian Gieger, Jaspal S Kooner, Willem H Ouwehand, Nicole Soranzo, John C Chambers: *Seventy-Five Genetic Loci Influencing the Human Red Blood Cell*. Nature 2012; 492(7429):369-375.

Paul Scheet, Erik A Ehli, Xiangjun Xiao, Catharina E M van Beijsterveldt, **Abdel Abdellaoui**, Robert R Althoff, Jouke Jan Hottenga, Gonneke Willemsen, Kelly A Nelson, Patricia E Huizenga, Yueshan Hu, Christopher I Amos, Meike Bartels, Maria M Groen-Blokhuis, Eco Jc de Geus, James J Hudziak, Gareth E Davies, Dorret I Boomsma: *Twins, Tissue, and Time: An Assessment of SNPs and CNVs.* Twin Research and Human Genetics 2012; 15(6): 737-745.

Jacqueline M Vink, Meike Bartels, Toos C E M van Beijsterveldt, Jenny van Dongen, Jenny H D A van Beek, Marijn A Distel, Marleen H M de Moor, Dirk J A Smit, Camelia C Minica, Lannie Ligthart, Lot M Geels, **Abdel Abdellaoui**, Christel M Middeldorp, Jouke Jan Hottenga, Gonneke Willemsen, Eco J C de Geus, Dorret I Boomsma: *Sex Differences in Genetic Architecture of Complex Phenotypes?* PLoS ONE 2012; 7(12):e47371.

Marijn A Distel, Irene Rebollo-Mesa, **Abdel Abdellaoui**, Catherine A Derom, Gonneke Willemsen, John T Cacioppo, Dorret I Boomsma: *Familial Resemblance for Loneliness.* Behavior Genetics 2010; 40(4):480-94.

DANKWOORD

Ik heb dit proefschrift aan heel veel mensen te danken; allereerst aan alle duizenden tot tientuizenden mensen die de lange vragenlijsten hebben ingevuld, en hun DNA hebben afgestaan: alle tweelingen en hun familieleden die bij het Nederlands Tweelingen Register zijn ingeschreven en alle NESDA en GoNL deelnemers. De tijd die jullie hier vrijwillig in hebben gestoken is van onschatbare waarde voor ons.

Ook mijn promotoren en co-promotoren ben ik erg dankbaar voor de kans die mij ze mij hebben geboden om dit proefschrift te schrijven en de uitstekende begeleiding daarbij. Dorret, bedankt voor deze kans, de altijd snelle en zorgvuldige feedback op mijn stukken en de betrokkenheid waarmee je altijd inging op mij als ik weer eens je kamer binnen kwam wandelen. Ook heel erg bedankt voor al die interessante projecten waar je mij bij hebt betrokken. Je hebt een waardevolle stempel achtergelaten op mijn wetenschappelijke loopbaan die ik de rest van mijn leven met trots bij me kan dragen. Brenda, het was een eer met jou en die indrukwekkende NESDA dataset te werken. Bedankt daarvoor en voor de altijd nuttige feedback en fijne samenwerking. Jouke-Jan, bedankt voor de interessante discussies, feedback en kritische blik op mijn stukken. Je was een goede sparring partner en zorgde ervoor dat ik altijd nog een paar keer goed moest nadenken over wat ik deed. Eco, het was een plezier jou als co-promotor te hebben. Bedankt voor het altijd frisse perspectief en grondige feedback op mijn stukken en voor het altijd klaarstaan als ik vragen had, hoe groot de stapel werk op je bureau ook was. En natuurlijk ook bedankt voor het altijd significant bijdragen aan de gezelligheid van de vrijdagmiddagborrels.

Ik wil ook graag al mijn collega's bij het Nederlands Tweelingen Register bedanken. Dirk en Michel, het was me een genoegen een kamer met jullie te delen de afgelopen vier jaar. Bedankt voor het gezelschap in de kamer, voor al die mooie discussies en al die foute humor. Bedankt alle (ex)-AIO's (Anouk, Lannie, Rene, Lot, Wouter, Michel, Sanja, Maria, Melanie, Janneke, Laura, Charlotte, Eveline, Jenny, Jenny, Nienke, Ineke, Diane, Nuno, Suzanne, Jorien, Camelia, Iryna) voor de goede sfeer en gezelligheid, de fijne samenwerking, de leuke uitjes en de leuke stukjes. Gonneke, Toos, Conor, Kees-Jan, Dennis, Harmen, Rene, Jacqueline, Meike en alle andere postdocs en senior collega's die ik regelmatig lastigviel met mijn vragen en altijd klaarstonden als je een ervaren persoon nodig had: bedankt voor jullie hulp, de fijne samenwerking en de gezelligheid. Michiel, Ellen, Michelle, Therese, Cyrina, en al het ander ondersteunend personeel: bedankt voor de uitstekende ondersteuning en dataverzameling, en natuurlijk ook voor alle gezellige praatjes tussendoor. Ook de ex-collega's (Lot, Rene, Marleen, Niels, Hannah, Ellen, Marijn) bedankt voor alle leuke tijden en jammer dat ik jullie niet meer dagelijks zie. Hamdi, thanks for the fun times on our trips to the US. Karin, bedankt voor de leuke

samenwerking en de opvang en leuke tijden in Australie. Ook de NESDA collega's die regelmatig over de vloer kwamen (Wouter, Rick, Yuri, Gerard): bedankt voor de fijne samenwerking en gezellige lunches. Natascha, bedankt voor al je hulp bij de praktische zaken, voor alle gezelligheid en voor het geweldige werk dat je doet voor de afdeling. Bedankt allemaal voor het opbouwen en staande houden van een zeer vruchtbare en ook hele gezellige afdeling. Ik heb het heel goed gehad met jullie de afgelopen jaren, en verheug me op de komende tijd!

I am also thankful for all my great colleagues abroad. Erik, Gareth, and Ryan: thanks for making me feel welcome in Sioux Falls, for your visits to Amsterdam, and of course for all the great lab work you do at Avera (this thesis would not have been possible without it). Patrick Sullivan, thanks for all the great feedback on my manuscripts and for a great job on the GODOT project. Peter Visscher, thanks for the all the great things you taught me during my visit. I could not have wished for a better preparation for my PhD than the six months I spent in your lab under your supervision. Nick Martin, thanks for your hospitality during my visit, and for the always useful feedback on my work and manuscripts. Brendan Zietsch, thanks for all the good times we had in Brisbane and in Amsterdam, and for the great collaborations.

I am grateful to the GoNL consortium and would like to thank everyone that contributed to a great working experience, especially Kai, Victor, Laurent, Jayne, Tobias, Alexander, Wigard, and the rest of the SV-group for the well-organized weekly calls. Ook de GoNL steering group (Cisca Wijmenga, Eline Slagboom, Cornelia van Duijn, Dorret Boomsma, Morris Swertz, Gert-Jan van Ommen en Paul de Bakker) wil ik hartelijk bedanken voor dit prachtige project en voor de kans die ze mij boden om eraan mee te werken.

Stackoverflow.com and all its users that made the programming part of my job so much easier also deserve many thanks. I would recommend anyone with programming as part of their job (or leasure time) to join this great community.

Jay, heel erg bedankt voor de topnotch behandeling van het artwork. Ik hoop dat je nog heel lang door blijft gaan met wat je doet. Farid, heel erg bedankt voor je hulp bij de organisatie van het afrondingsproces van dit traject.

Mijn broeders van andere moeders: Farid, Moh, Ruben, bedankt voor de afgelopen 20 jaar. Wie ik vandaag de dag ben heb ik voor een groot gedeelte aan jullie te danken.

Hetzelfde geldt natuurlijk voor mijn familie. Heel erg bedankt voor het warme nest mama, ba, Rachida, Zakarya, Maryam, Ghadija, Fatima en Mohammed.

Mama en ba, bedankt voor alle steun en liefde die jullie me gegeven hebben. Ik weet niet wat voor genen jullie mij gegeven hebben, maar ik ben er heel blij mee; ze hebben voor een aanzienlijk deel bijgedragen aan dit proefschrift.

Farah, bedankt voor de mooie tijden die we doorbrengen met elkaar en voor je begrip en steun als ik af en toe wat langer moest doorwerken dan de bedoeling was. Het meest dankbaar ben ik je natuurlijk voor het mooiste geschenk dat ik ooit heb gekregen: Liam, bedankt voor de nieuwe betekenis die je mijn leven geeft. Ik dacht dat ik in wetenschap de richting voor mijn leven had gevonden, en toen kwam jij en relativeerde alles. En op een of andere manier heeft die nieuwe betekenis aan mijn leven ook weer mijn motivatie voor dit proefschrift verhoogd. Ik heb je vandaag beloofd een kopie van dit proefschrift te geven waarmee je mag doen wat je wilt (inkleuren, opvouwen, mee voetballen), maar ik bewaar ook een kopie voor je voor als je dit kan lezen. <3

- Appie/Abdel/Abderrahman, 17 augustus 2014, Amsterdam