Defining the phenotype

Combining phenotypes for discovery of novel genetic associations

Joanna von Berg

ISBN: 978-94-6458-311-3

Design and layout: Joanna von Berg

Printing: Ridderprint | www.ridderprint.nl

Defining the phenotype

Combining phenotypes for discovery of novel genetic associations

Het fenotype definiëren

Het combineren van fenotypes om nieuwe genetische associaties te ontdekken

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 15 juni 2022 des middags te 12.15 uur

door

Joanna von Berg

geboren op 22 januari 1992
te Voorburg

Promotor:
Prof. dr. E.P.J.G. Cuppen

Copromotoren:
Dr. J. de Ridder
Dr. S.W. van der Laan

Beoordelingscommissie:
Prof. dr. G. Pasterkamp
Prof. dr. ir. H.M. den Ruijter
Prof. dr. M.J.T. Reinders
Prof. dr. D.L. Oberski PhD
Dr. Y.M. Ruigrok

The work described in this thesis has been realised with financial support of the U.S. National Institutes of Health (R01NS100178). Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

What's in a name?

A rose by any other name would smell as sweet

William Shakespeare

Table of contents

Introduction
The relationship between variation in the genome and our individual traits

The genome

The human genome consists of 3.2 billion building blocks called nucleotides that are chained together in 23 different chromosomes. Each nucleotide can be one of four molecular units: adenine, thymine, cytosine, or guanine (indicated by the letters A, T, C and G respectively). Your genotype refers to your individual nucleotide sequence, and your genotype at a specific position refers to the nucleotide ("letter") at that position in your genome. Specific regions of the genome can be used as a blueprint to make proteins, which fulfil most of the functions in our body. These regions, and the regions around it that regulate how often the protein is made, are called genes. The majority of our genome – 99.5 % – is identical to the genome of any other person. [1] Because the genome is so large, the small proportion that varies still amounts to millions of nucleotides.

Genomic variation

The most common type of genetic variation is when one nucleotide at a specific position varies; this is called a Single Nucleotide Polymorphism, or SNP (pronounced "snip"). Other genetic variants that are often observed are small insertions and deletions - where a part of the genomic sequence is seemingly missing or added in some people. Whether this is called an insertion or a deletion depends on the reference genome .

The first reference genome was composed in 2003 and is based on the genomes of tens of people. [2] Larger genetic variants are mainly observed in cancer, where the genome in the tumor is often rearranged, although they can also be seen in the general population. An important distinction with for instance cancer genetics is that SNPs describe genetic variants that we carry with us from conception and do not change throughout our life (also called germline variation), while the genome in a tumor has often changed drastically from acquired mutations.

Different models of SNP-phenotype effects

Each trait that varies among people is a phenotype. Height, for example, is a continuous phenotype. Whether you can roll your tongue is a binary phenotype; its value is true or false. A very common type of binary phenotype are disease phenotypes; if someone has the disease, they are a case, otherwise they are a control. Binary (disease) phenotypes are thus also called case control phenotypes.

If we want to study the effect of a SNP on a certain trait, we should recall that each chromosome and thus each SNP has two copies. Each copy can have one of two options: the reference nucleotide (also called reference allele, indicated with ref) or the alternative nucleotide (alt). This means that there are three options for the combination of alleles: ref-ref, ref-alt or alt-alt. Recessive SNPs only lead to a phenotype if the effect allele is present in both copies, and dominant SNPs lead to a phenotype if any or both copies contain the effect allele. Others affect a phenotype additively; each additional copy of the effect allele increases the risk of getting a disease, for instance. This thesis describes additive SNP effects.

In the Mendelian inheritance model, one genetic variant leads to one predictable phenotype. [3] However, most common diseases are complex traits that are associated with many genetic variants, as well as non-genetic factors. Each associated genetic variant is assumed to increase disease risk by a very small amount. If an individual has many genetic and non-genetic risk factors, the chance that they will get the disease (their disease risk) is greater than that of someone with less risk factors. [4] The variation of a trait in the population is not always associated with genetic variation. Some traits – for example which language you speak – are acquired, and not influenced by our genome. The relative contribution of genetic variation to phenotypic variation is the heritability. It is 0 for acquired traits, and 1 (or 100%) for traits whose variation can be explained by genetic variation alone. Heritability does not describe how much of a certain trait is influenced by genetics: if environmental variance is very low, the relative contribution of genetic variation to the phenotypic variation is high. Still, an environmental change can have a large effect on the phenotype. [5], [6]

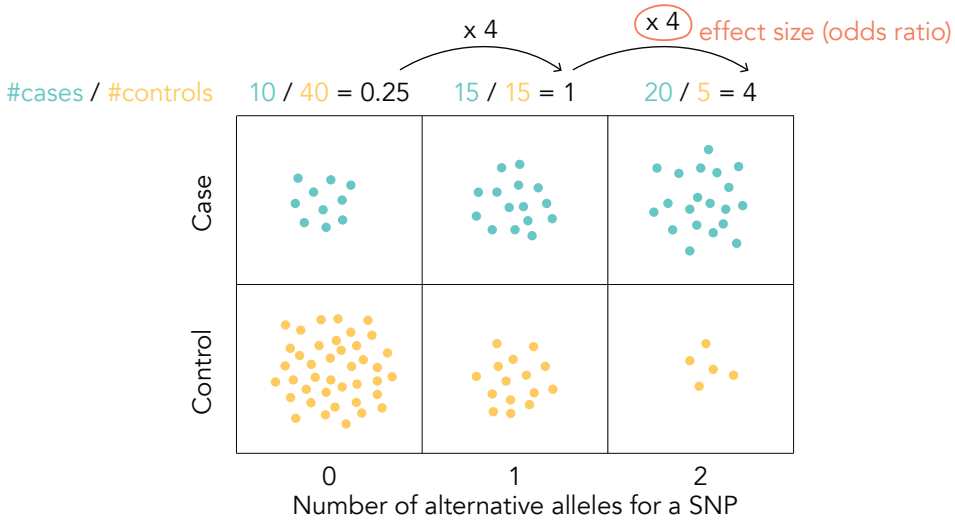


Figure 1. An example of case-control analysis at a SNP. Each dot is a person: blue for cases (top row) and yellow for controls (bottom row). The columns indicate the number of alternative alleles each person has (0 = ref-ref, 1 = ref-alt or alt-ref, 2 = alt-alt). The ratio of cases to controls (the odds of ischemic stroke) is indicated above each column. The odds ratio is the ratio of odds at consecutive columns.

Genome Wide Association Studies

To find genetic variants that are associated with a phenotype, we investigate each SNP for a relationship between the number of alternative alleles and the phenotype values. This is called a Genome Wide Association Study (GWAS). In figure 1 we see an example of a GWAS where we are looking for SNPs that are associated with higher risk of ischemic stroke . We describe the genotypes by counting the number of alternative alleles someone has; 0, 1 or 2. How much the risk increases or decreases with each additional copy of the alternative allele can be calculated from the odds ratio: the ratio of cases/controls ratio for two consecutive columns. In figure 2 we want to find SNPs that are associated with the age at onset of ischemic stroke. We have to analyse the data differently, because age at onset is a continuous value. The slope of the linear regression line tells us how much younger – on average – people get an ischemic stroke if they have an additional copy of the alternative allele.

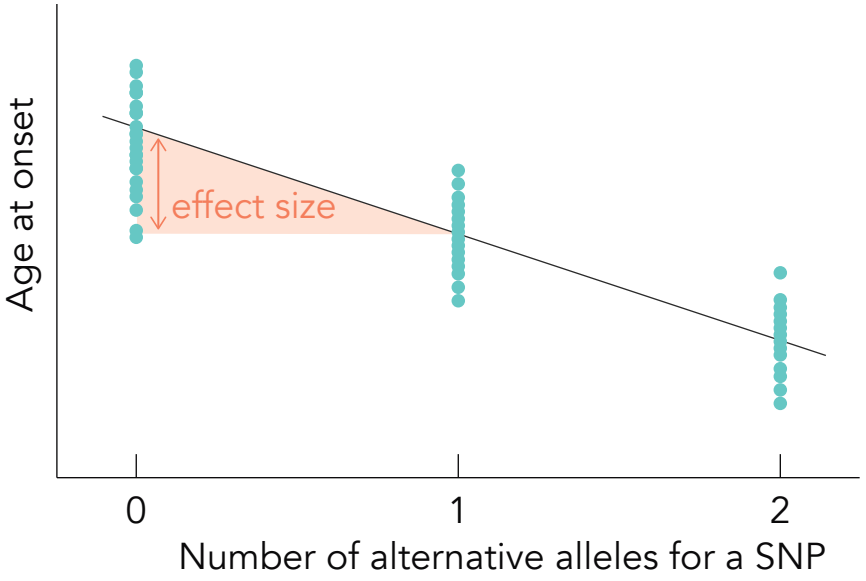


Figure 2. An example of linear regression of age at onset at a SNP. Each blue dot is a person. The x-axis indicates the number of alternative alleles they have (0 = ref-ref, 1 = ref-alt or alt-ref, 2 = alt-alt), the y-axis indicates how old they were when they got a stroke. The blue sloping line is the regression line; this line describes the observed data the best. The slope tells us the effect size of this SNP on age at onset; for each additional copy of the alternative allele, how much earlier do we expect to get a stroke?

Before 2010 most genetic studies were candidate gene studies. One or a few candidate genes are then chosen based on previous biological knowledge. When the first reference genome was published and computing power was enough to analyse millions of SNPs, the first GWAS could be carried out. GWAS provides a 'hypothesis-free' approach by considering SNPs across the whole genome, which contrasts it with the candidate gene studies. The latter are based on previous knowledge, but can be biased because the genes have to be chosen by the researchers themselves and thus are poorly replicated. [7] The first GWAS was published in 2007 and described SNP associations with seven common diseases. [8] Since then, thousands of new SNP associations have been described. [9] Translating the statistical findings from GWAS into actionable results that can be used in public health and healthcare has proven difficult for most complex diseases. This is not unexpected because most complex diseases are

polygenic; a lot of different SNPs are associated, implying that there are many biological pathways involved. Finding a treatment that interacts with any one of these pathways will probably not completely treat the disease as it does not affect the other pathways. Still, GWAS can help us understand more about the biological mechanisms of complex traits and thus bring us a step closer to a future treatment. [10], [11]

P-value versus effect size

When we test for association between a SNP and a phenotype, we use the effect size and standard error to calculate a p-value. P-values tell us how likely it is to observe an effect that is at least as large as what we observed by chance, in the situation that there is actually no effect. Observing a very large effect size if the true effect is zero is still possible, it's just unlikely. Conversely, a small effect size is usually not significantly different from zero (unless we use a very big sample). It can however still come from an effect that is different from zero. We are always reasoning from the situation of no effect (we call this the null hypothesis). A p-value alone should not be used to conclude anything about the relevance of a SNP on a phenotype. Considering the size of the observed effect together with a p-value can help us decide whether a SNP is interesting or not. A GWAS in a large group of people might find a SNP with a significant effect on height that is very small. Maybe people with that SNP are 0,001 cm taller, on average, than people without that SNP. Even though the effect is very significant, an effect this small is not very interesting.

Statistical power

As explained above, a p-value tells us the probability of observing an effect that is as large or larger than what we observed if the null hypothesis is true. [12] If we test one SNP, we set the p-value threshold α to 0,05; if the p-value is lower than α we reject the null hypothesis because there is enough evidence for association. Because of the definition of a p-value, the probability of incorrectly rejecting the null hypothesis if it were actually true is equal to α. Statistical power

tells us the probability of correctly rejecting the null hypothesis if the alternative hypothesis is actually true. [13] This depends on the alternative hypothesis and thus on the effect size and variance that we observed. If the effect size is very large and variance is small, statistical power is greater. If the effect size is small or variance is large, statistical power is lower. Statistical power is determined by a number of factors. Some can not be influenced, like allele frequency. Sample size is often considered because it can (relatively) easy be influenced. [14]–[16] Increasing the sample size – while keeping all other factors constant – will increase the statistical power to discover smaller effect sizes, so we generally want the sample to be as large as possible. More precise phenotype measurements – while keeping all other factors constant – also increase statistical power. [17] In reality these factors are often not independent. In chapter 1 we describe a situation where increased phenotype precision leads to a lower sample size, and discuss the implications for ischemic stroke GWAS. A bigger sample size does not necessarily lead to less precise phenotypes. However, in the context of limited funding it is often a question of measuring a select number of phenotypes on a bigger sample or measuring a large number of phenotypes on a smaller sample. [17] In most cases, the former option will be chosen, resulting in less phenotype precision.

Association does not equal causation

This thesis rests on observational data, which have been collected and used afterwards to infer associations between genetic variation at a certain locus and a phenotype. Using only observational data we cannot say anything about causality; an association between a SNP and lung cancer for instance might be due to a direct effect on a lung cancer pathway. It might also be influenced by a confounder – a factor that is associated with both the SNP and the phenotype – like smoking. Some SNPs are more common in one genetic ancestry than in others. If the case group consists mainly of people with ancestry A and the control group consists of people with ancestry B, the SNPs that are more common in ancestry A will be associated with being a

case in this GWAS whether they're actually associated with the disease or not. These two situations can be prevented to a certain extent, by correction of known confounders and by careful study design.

Another complicating factor is that SNPs are not inherited independently of each other. When an egg or sperm cell (a germ cell) is created during meiosis, it gets approximately half of the parent's DNA. Which half is determined by recombination, a process in which the genome breaks at certain places, and recombines to create a new combination of genetic material. This new chromosome is transferred to a germ cell. Genome breaks do not occur uniformly across the genome, and therefore certain SNPs are more often inherited together. This correlation structure of SNPs is called linkage disequilibrium (LD); a number ranging from 0 to 1 that tells us how often we see a pair of SNPs together. Because of the non-random breaking of the genome during meiosis, there are LD blocks with groups of SNPs that are often seen together. If we find an association with a SNP, we know that any of the SNPs in its LD block could be responsible for the association. The fact that some SNPs in the LD block do show significant associations and others do not can usually be explained by differences in frequency; if a SNP is more common it is easier for a real association to become significant.

To narrow down the set of SNPs that are causal for the association, we can use the exact LD values for each pair of SNPs in the LD block and the exact association strengths for each SNP (this is called fine mapping). Fine mapping results in a set of SNPs that is 95 % likely to contain the causal SNP, under the assumption that there is only one causal SNP in the genomic region of interest. Note that I do not mean causal in the traditional sense of the word as that is not possible with observational data. Fine mapping is helpful if we want to follow-up each associated SNP with lab experiments. We can for instance change a SNP in a model organism or cell culture and see if there is indeed an effect of the phenotype. These experiments are expensive and labor intensive, and a reduction from tens or hundreds of SNPs to a few likely 'causal' SNPs can make a big difference.

Functional interpretation of GWAS hits

Wet lab experiments, as described in the previous paragraph, are labor intensive and time consuming but are able to show whether a genetic variant causes a phenotype. Computational analyses can not distinguish causality from mere association but they can give insight into their function. If a SNP is located in a gene, previous knowledge on that gene can be used to gain some insight into a potential function for the genetic variant. Especially if a SNP is located in an exon – the part of a gene that is translated into protein – and changes the amino acid composition of the protein (a non-synonymous variant), we can predict what that might do to the function of the protein. Once we have mapped a SNP to a gene, we can also use the pathways that the protein is involved in to learn more about the SNP's function. However, most SNPs are not located in a gene (intergenic variants), which makes functional interpretation difficult. Fortunately we can use acquired knowledge about genomic domains to get an idea of a SNP's mechanism of action. For instance, an intergenic region that is very well conserved across species likely has an important gene regulatory function. SNPs in this region can thus be expected to have an effect on expression of the genes controlled by this region. [18]

Ischemic stroke

In the first two chapters of this thesis, we present the results of GWAS of ischemic stroke. If blood flow to the brain is blocked, there is loss of oxygen (ischemia). If the blockage was very temporary, this is called a Transient Ischemic Attack (TIA). If it lasts longer, it is an ischemic stroke (IS, also called cerebrovascular infarction or cerebrovascular accident). Another type of stroke occurs when the loss of oxygen in the brain is caused by rupture of an artery; a hemorrhagic stroke. Ischemic strokes occur four times as often as hemorrhagic strokes. The loss of oxygen during a stroke can lead to severe disability or death. Ischemic stroke is a complex trait with heritability estimated at 38%. There are three IS subtypes that are commonly distinguished by physicians. Cardioembolic stroke is assumed to be caused by a blood clot (thrombus) that forms

in the heart and travels to the brain. Large artery stroke is assumed to be caused by build-up of an atherosclerotic plaque that obstructs the carotid arteries. Finally, small vessel stroke is the least well understood subtype, where the small vessels in the brain itself are blocked.

To date, tens of SNPs have been associated with increased risk to get an IS, or any of the specific subtypes.[19]–[23] Identifying new SNP associations can help pinpoint biological processes that play a role in the origin of an ischemic stroke. It ¬could also help identify people with higher genetic risk, so they and their physician can manage modifiable risk factors earlier. [24] In chapters 1 and 2 we present the GWAS results of different phenotype definitions for ischemic stroke and identify new SNP associations.

Contributions of this thesis

Different ways of defining the phenotype lead to different results

A phenotype can be defined using different measurements. Height can be described as a continuous phenotype - someone's height in centimetres - and it can be described as a binary phenotype - someone can be short or tall.

In chapter 1, we consider different methods that can diagnose the subtype of an ischemic stroke case. These methods do not always agree with each other. For each ischemic stroke subtype, we asked each method which of the individuals have the subtype in question and used those as cases. We also used two new phenotype definitions: the group of people that are diagnosed with the subtype in question by at least one of the methods; the union, and the group of people that are diagnosed with the subtype in question by all methods; the intersect. By definition the union is bigger than the intersection, which theoretically increases statistical power. However, the intersection is stricter. If there are some people that do not actually have a certain subtype, they might be diagnosed by some but probably not by all methods. The intersect will then contain people with a higher theoretical confidence in the phenotype.

Previously, we used the example of the binary phenotype 'ability to roll your tongue'. We can determine whether you can roll your tongue during childhood and we know that this will not change later. This is not always true for binary disease phenotypes. Young people who do not have a disease – and thus are a control – can be diagnosed with that disease when they are older and become a case. On the other hand, someone who has a disease will never become a control. This means that we are more confident about the cases than we are about the controls. Ischemic strokes generally happen at older age, and they are common; the chance that someone will get an ischemic stroke during their life is 18%. [25] That means that there are relatively many people in the control group who will get an ischemic stroke later in their life. These 'future cases' also carry some risk SNPs, and because we are analysing them as controls this can make it more difficult to pick up the difference in allele frequency at those risk SNPs. Because we do not know which controls will get a stroke later, it is difficult to correct for this problem in a case-control study. Instead, we use a different phenotype that is related to ischemic stroke risk: the age at onset. We usually assume that people with a lower age at onset had a bigger genetic risk, because they had less time to acquire other risk factors (like high blood pressure). In chapter 2 we describe the results of a GWAS of age at onset of ischemic stroke. We analysed only cases, as only they had an ischemic stroke and thus an age at onset. We hypothesized that there might be SNPs that make you more likely to get an ischemic stroke earlier. These SNPs should also be associated with increased risk of getting an ischemic stroke. We found one SNP that was not previously described in case-control GWAS of ischemic stroke. This can mean that this SNP is also related to increased risk, but the previous case-control GWAS did not have enough statistical power to identify it. It can also mean that the association that we found is biased by something that we did not correct for. We know that this SNP is associated with earlier age at death. [26] If it leads to earlier death through a mechanism that is independent from ischemic stroke, this could lead to a bias; older people are less likely to have this SNP, because people with the

risk allele die earlier. However, it is not clear whether the association with earlier death is independent from ischemic stroke risk. The SNP changes an amino acid in a protein that is involved in lipid metabolism and has previously been associated with increased risk for a number of cardiovascular phenotypes. This biological knowledge makes it plausible that the SNP is associated with increased risk for ischemic stroke, which subsequently leads to an association with earlier death because having a stroke increases the risk of dying earlier. Currently, we cannot distinguish these two possible scenarios. Future studies of age at onset phenotypes should be aware of this potential bias.

The association of a phenotype with a specific SNP can be different in different groups of people. For instance, researchers have found three genetic variants that are associated with migraine risk in women but not in men. [27] We have also stratified our case group in women and men and did a GWAS of age at onset in both groups. We see that the SNP we identified has the largest effect in women: each additional copy of the alternative allele is associated with 1.6 years earlier onset of ischemic stroke. In men the effect of this SNP is not significantly different from zero. That could also explain why this association has not been described in previous case-control GWAS; they did not do sex-stratified analyses and thus would not have found a sex-specific association.

Genetic variants with an effect on more than one phenotype: pleiotropy

Many researchers have shared GWAS results in large databases that are publicly accessible. Not only does this mean that we do not have to redo the same analysis, but it also enables us to use these results for further research. We know that some SNPs have an effect on multiple phenotypes: they are pleiotropic. Pleiotropic SNPs can give insight into the biological processes that are involved in a phenotype; if we already understand the mechanism by which a SNP affects one phenotype and find out that that SNP also has an effect on another phenotype, the same mechanism might be

involved in the other phenotype as well. Identifying pleiotropic SNPs can help us understand which traits share underlying mechanisms. For example, SNPs with an effect on multiple auto-immune diseases might point at specific biological processes that can be assumed to play a role in the immune system. On the other hand, SNPs with an effect on multiple traits that are not phenotypically similar could indicate biological processes that have a more general function.

In chapter 3 we give an overview of existing methods that use GWAS results to find pleiotropic SNPs. We describe four methods and show that two methods do not identify SNPs with an effect on two or more traits but on one or more traits. This means that these methods will also call a SNP pleiotropic if it has an effect on only one trait.

In chapter 4 we introduce a new method - PolarMorphism - that can be used to find SNPs with a shared effect on any number of traits, based on their GWAS results. PolarMorphism is based on the notion that the effect of a SNP can be described with Cartesian coordinates but also with polar coordinates. In Cartesian coordinates, the x-coordinate is the effect on the first trait and the y-coordinate is the effect on another trait. In polar coordinates, each SNP is described by its distance from the origin and its angle with the x-axis. The distance gives its overall effect, which can be trait-specific or shared by the traits. The angle indicates how trait-specific or shared it is. After all, if a SNP is very specific for trait x, it has a large x-coordinate and a small y-coordinate. The angle is then zero. A SNP that is very specific for trait y has an angle of 90 degrees (or 0.5 * pi, which is the same angle given in radians instead of degrees). If a SNP has a shared effect, the angle is close to 45 degrees or 0.25 * pi. This gives us a way to measure 'sharedness' of SNPs; first we use the distance to determine which SNPs have a large enough overall effect, then we use the angle to determine which SNPs are shared.

References

[1] E. S. Lander et al., "Initial sequencing and analysis of the human genome," Nature, vol. 409, no. 6822, pp. 860–921, Feb. 2001.

[2] International HapMap 3 Consortium et al., "Integrating common and rare genetic variation in diverse human populations," Nature, vol. 467, no. 7311, pp. 52–58, Sep. 2010.

[3] G. Mendel, "Versuche über Pflanzen-Hybriden," Züchter, vol. 13, no. 10–11, pp. 221–268, Oct. 1941.

[4] S. Wright, "An analysis of variability in number of digits in an inbred strain of guinea pigs," Genetics, vol. 19, no. 6, pp. 506–536, Nov. 1934.

[5] E. E. Maccoby, "Parenting and its effects on children: on reading and misreading behavior genetics," Annu. Rev. Psychol., vol. 51, no. 1, pp. 1–27, 2000.

[6] P. M. Visscher, W. G. Hill, and N. R. Wray, "Heritability in the genomics era--concepts and misconceptions," Nat. Rev. Genet., vol. 9, no. 4, pp. 255–266, Apr. 2008.

[7] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn, "A comprehensive review of genetic association studies," Genet. Med., vol. 4, no. 2, pp. 45–61, Mar. 2002.

[8] Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," Nature, vol. 447, no. 7145, pp. 661–678, Jun. 2007.

[9] J. MacArthur et al., "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," Nucleic Acids Res., vol. 45, no. D1, pp. D896–D901, Jan. 2017.

[10] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery," Am. J. Hum. Genet., vol. 90, no. 1, pp. 7–24, Jan. 2012.

[11] P. M. Visscher et al., "10 Years of GWAS Discovery: Biology, Function, and Translation," Am. J. Hum. Genet., vol. 101, no. 1, pp. 5–22, Jul. 2017.

[12] R. L. Wasserstein and N. A. Lazar, "ASA statement on statistical significance and p-values," in The Theory of Statistics in Psychology, Cham: Springer International Publishing, 2020, pp. 1–10.

[13] P. C. Sham and S. M. Purcell, "Statistical power and significance testing in large-scale genetic studies," Nat. Rev. Genet., vol. 15, no. 5, pp. 335–346, May 2014.

[14] R. D. Ball, "Designing a GWAS: power, sample size, and data structure," Methods Mol. Biol., vol. 1019, pp. 37–98, 2013.

[15] W. Jiang and W. Yu, "Power estimation and sample size determination for replication studies of genome-wide association studies," BMC Genomics, vol. 17 Suppl 1, no. S1, p. 3, Jan. 2016.

[16] W. Jiang and W. Yu, "Erratum to: Power estimation and sample size determination for replication studies of genome-wide association studies," BMC Genomics, vol. 18, no. 1, p. 73, Jan. 2017.

[17] J. L. Gage, N. de Leon, and M. K. Clayton, "Comparing genome-wide association study results from different measurements of an underlying phenotype," G3 (Bethesda), vol. 8, no. 11, pp. 3715–3722, Nov. 2018.

[18] D. Polychronopoulos, J. W. D. King, A. J. Nash, G. Tan, and B. Lenhard, "Conserved non-coding elements: developmental gene regulation meets genome organization," Nucleic Acids Res., vol. 45, no. 22, pp. 12611–12624, Dec. 2017.

[19] S. L. Pulit et al., "Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study," Lancet Neurol., vol. 15, no. 2, pp. 174–184, 2016.

[20] J. F. Meschia et al., "Stroke Genetics Network (SiGN) study:

design and rationale for a genome-wide association study of ischemic stroke subtypes," Stroke, vol. 44, no. 10, pp. 2694–2702, Oct. 2013.

[21] M. Traylor et al., "Genetic variation at 16q24.2 is associated with small vessel stroke," Ann. Neurol., vol. 81, no. 3, pp. 383–394, Mar. 2017.

[22] J. F. Meschia et al., "Genomic risk profiling of ischemic stroke: results of an international genome-wide association meta-analysis," PLoS One, vol. 6, no. 9, p. e23161, Sep. 2011.

[23] NINDS Stroke Genetics Network (SiGN) and International Stroke Genetics Consortium (ISGC), "Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study," Lancet Neurol., vol. 15, no. 2, pp. 174–184, Feb. 2016.

[24] M. Dichgans, N. Beaufort, S. Debette, and C. D. Anderson, "Stroke genetics: Turning discoveries into clinical applications," Stroke, vol. 52, no. 9, pp. 2974–2982, Aug. 2021.

[25] E. J. Benjamin et al., "Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association," Circulation, vol. 135, no. 10, pp. e146–e603, Mar. 2017.

[26] L. C. Pilling et al., "Human longevity: 25 genetic loci associated in 389,166 UK biobank participants," Aging , vol. 9, no. 12, pp. 2504–2520, Dec. 2017.

[27] H. Choquet et al., "New and sex-specific migraine susceptibility loci identified from a multiethnic genome-wide meta-analysis," Commun Biol, vol. 4, no. 1, p. 864, Jul. 2021.

Chapter 1

Alternate approach to stroke phenotyping identifies a genetic risk locus for small vessel stroke

Joanna von Berg[1], Sander W. van der Laan[2], Patrick F. McArdle[3], Rainer Malik[4], Steven J. Kittner[5], Braxton D. Mitchell[3], Bradford B. Worrall[6*], Jeroen de Ridder[1,7*], Sara L. Pulit[1,8,9*] * these authors contributed equally to this work

1.Genetics, Centre for Molecular Medicine, University Medical Centre Utrecht, Utrecht, The Netherlands 2. Laboratory of Clinical Chemistry & Hematology, Division Laboratories, Pharmacy, and Biomedical Genetics, University Medical Center Utrecht, University Utrecht, Utrecht, NL 3. Division of Endocrinology, Diabetes, and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. 4. Institute for Stroke and Dementia Research (ISD), University Hospital, LMU Munich, Munich, Germany. 5. Department of Neurology, Veterans Affairs Maryland Healthcare System, and University of Maryland School of Medicine, Baltimore, MD, USA. 6. Departments of Neurology and Public Health Sciences, University of Virginia, Charlottesville, VA, USA. 7. Oncode institute, Utrecht, The Netherlands 8. Big Data Institute, Li Ka Shing Center for Health Information and Discovery, Oxford University, Oxford, UK. 9. Program in Medical and Population Genetics, Broad Institute, Boston, MA, USA.

This chapter is based on a manuscript that has been published in the European Journal of Human Genetics. Scan this QR code to go to the published manuscript, and to access the supplemental information:

DOI: 10.1038/s41431-020-0580-5

Introduction

Stroke is one of the primary causes of death worldwide and causes ~1 in every 20 deaths in the United States [1]. Eighty-seven percent of all strokes are ischemic, caused by a blockage of blood flow to the brain [1]. Ischemic stroke (IS) tends to affect those older than 65 years old and has several known risk factors, including type 2 diabetes, hypertension, and smoking. However, the affected population is extremely heterogeneous in terms of age, sex, ancestral background, and socioeconomic status.

Ischemic strokes themselves are also heterogeneous in terms of clinical features and presumed mechanism. The majority of IS are typically grouped into three subtypes: cardioembolic stroke (CES), typically occurring in people with atrial fibrillation; large artery stroke (LAS), caused by eroded or ruptured atherosclerotic plaques in arteries; and small vessel stroke (SVS), caused by a blockage of one of the small vessels in the brain. These subtypes also seem to be genetically distinct: genome-wide association studies (GWAS) in ischemic stroke have identified single-nucleotide polymorphisms (SNPs) that primarily associate with a specific IS subtype [2]. To date, GWAS have identified 20 loci associated with ischemic stroke, of which 9 appear to be specific to an IS subtype [2]. Furthermore, the subtypes also have varying SNP-based heritabilities (estimated at 16%, 12% and 18% for CES, LAS and SVS respectively [3]), indicating that the phenotypic variation captured by genetic factors varies across the subtypes. Improved genetic discovery can help further elucidate the underlying biology of ischemic stroke as well as potentially help identify genetically high-risk patients who could be candidates for earlier clinical interventions.

While neurologists and researchers agree on the delineation of ischemic stroke into these three primary categories (CES, LAS and SVS), several subtyping systems are currently used to assign a subtype to an ischemic stroke patient. The most commonly used approach is a questionnaire based on clinical knowledge that was originally developed for the Trial of Org 10172 in Acute Stroke Treatment

(TOAST) [4]. TOAST was designed for implementation in the clinic and has also been used as subtyping system in the majority of stroke GWAS. More recently, researchers have developed a second subtyping system: the Causative Classification System for Stroke (CCS) [5], a decision model based on clinical knowledge that also incorporates imaging data. There are two outputs of CCS: CCS Causative (CCSc), which assigns one subtype to each patient based on the presumed cause of the stroke; and CCS phenotypic (CCSp), which allows for multiple subtype assignments and incorporates the confidence of the assignment. Previous work indicates that TOAST and CCS have moderate, but not high, concordance in assigning subtypes in patients: agreement is lowest in SVS ($\kappa = 0.56$) and highest in LAS ($\kappa = 0.71$) [6]. Notably, both subtyping systems still place more than one third of all samples into a heterogeneous 'undetermined' category. [6]

Determining a patient's subtype is difficult and prone to misclassification [7], but critical to genetic discovery in ischemic stroke, as demonstrated by the prevalence of subtype-specific association signals. If a group of cases is comprised of phenotypically heterogeneous samples with different underlying genetic risk, power to detect a statistically significant association at a truly associated SNP is reduced (Fig 1). In contrast, a case definition that captures a more phenotypically homogenous group of cases would improve the chances of detecting genetic variants that associate with disease. Therefore, we used the TOAST, CCSc and CCSp subtype assignments to define two new phenotypes per subtype: the intersect, for which an individual must be assigned the same subtype across all three subtyping systems; and the union, for which an individual must be assigned that subtype by at least one of the subtyping systems. Analyzing the union potentially improves power for locus discovery due to its larger sample size, but at the cost of a more potential mis-classification. In contrast, analyzing the intersect may improve power for genetic discovery by generating a phenotype that is less prone to mis-classification, despite a smaller sample size.

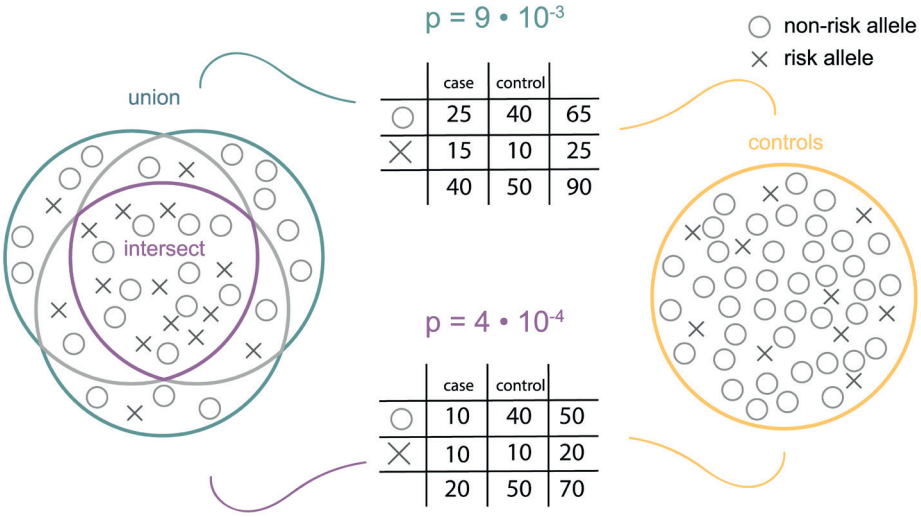


Figure 1. Hypothesized benefit of using the intersect, at a SNP associated with ischemic stroke. Circles indicate the protective allele, and crosses the risk allele. Using a chi-square test (visualized with contingency tables), the measured effect is stronger with a group of cases that is more homogeneous but smaller (intersect, purple) than with a group of cases that is less strictly defined but is larger (union, teal).

Here, we perform GWAS with the union and intersect phenotypes for each primary IS subtype to investigate whether these newly-defined phenotypes indeed improve our ability to detect genetic risk factors for ischemic stroke. We find heritability estimates to be highest in the intersect phenotype for all subtypes. We also find stronger effects at known associations for the intersect compared to the union, and we validate a previously suspected association with the CAMK2D locus in small vessel stroke through GWAS of the intersect phenotype.

Fig 1. Hypothesized benefit of using the intersect, at a SNP associated with ischemic stroke. Circles indicate the protective allele, and crosses the risk allele. Using a chi-square test (visualized with contingency tables), the measured effect is stronger with a group of cases that is more homogeneous but smaller (intersect, purple) than with a group of cases that is less strictly defined but is larger (union, teal).

Results

Genome-wide association study data processing

To investigate how redefining stroke phenotypes improves our ability to detect SNPs associated with ischemic stroke, we employed the SiGN dataset. Data processing of the SiGN dataset, including quality control and imputation, has been described in detail elsewhere. [8] Briefly, the dataset includes 13,930 IS cases and 28,026 controls of primarily European descent. Cases and controls were genotyped separately (with the exception of a small number of cohorts) and on various Illumina arrays and then merged together into case-control groups matched for genotyping array and sample ancestry (via principal component analysis). For the cases, phenotype definitions based on one or more of the CCSc, CCSp and TOAST subtyping systems are available (Table 1).

We began our analyses by running genome-wide association studies for all phenotype definitions in all subtypes, including our intersect and union definitions. We ran all GWAS using a linear mixed model implemented in BOLT-LMM (Supplemental Figure 2). [9] To take into account any residual population stratification and other batch effects, we included the first 10 principal components and sex as covariates in these analyses (Table S2).

	CES	LAS	SVS	undetermined	total
CCSc	3000	1565	2262	4574	11401
CCSp	3608	2449	2419	718	9194
TOAST	3333	2318	2631	3479	11761
Intersect	2219	1328	1548	not tested	5095
Union	4502	3495	3480	not tested	11477
Sym. diff.	2283	2167	1932	not tested	6382

Table 1. Case counts for the different phenotype definitions in the three subtypes. The control group is always the same group of 28,026 individuals

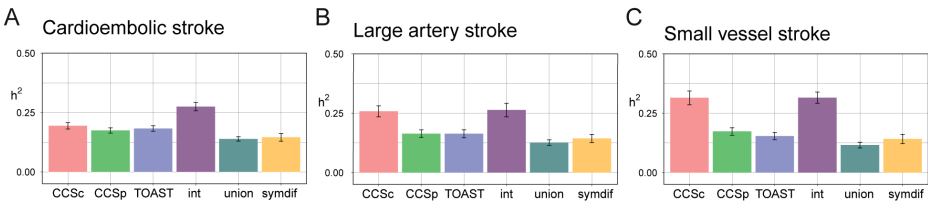


Figure 2. Intersect is the most heritable phenotype. Heritabilities on the liability scale for the six case definitions. int = intersect, symdif = symmetric difference. Bars indicate the standard error. Note that intersect has a relatively high standard error, due to its lower sample size. (A) In cardioembolic stroke, intersect is significantly more heritable than all other phenotype definitions (p-values for the difference between intersect and all others 3.6e-03 or lower). (B) In large artery stroke and (C) small vessel stroke, intersect is significantly more heritable than all other phenotype definitions except CCSc (p-values for the difference between intersect and all others except CCSc, 2.7e-03 or lower in LAS, 6.1e-07 or lower in SVS). P-values for heritability differences determined by t-test (see Table S5). See Table S4 for numerical values of heritabilities and standard errors.

Because the intersect by definition is contained in the union, one additional GWAS for each subtype was run to enable a truly independent comparison of intersect with the symmetric difference (the union minus the intersect). This study focuses on the balance in statistical power between a high sample size and a clean phenotype. Therefore, this sensitivity analysis was only done for the comparison between the two most extreme case definitions: the union and the intersect. The symmetric difference is not suited as a phenotype by itself.

Genetic variance in a strictly defined case group explains a higher proportion of phenotypic variance

To estimate how much of the variation in a particular phenotype can be explained by genetic variation, we calculated the heritability (h2) of each phenotype using BOLT-REML, assuming an additive model of effect sizes over all SNPs. We estimated heritability in each of the available phenotypes: the subtypes as defined by TOAST, CCSc, CCSp, the union, and the intersect. We found that the intersect yields a higher h2 estimate than the union in all ischemic stroke subtypes

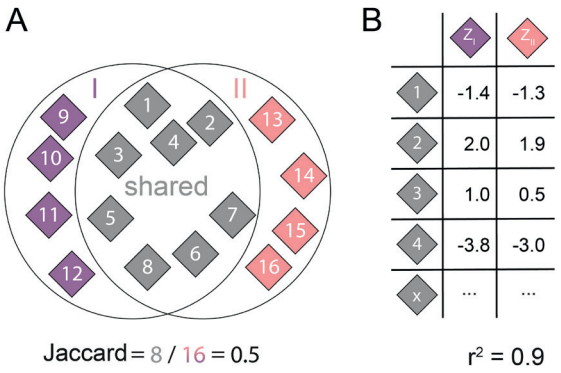


Figure 3. Graphical explanation of overlap analysis. (A) At a certain absolute z-score threshold Z, all SNPs that have a z-score lower than -Z or higher than +Z in GWAS I are determined (SNPs 1-8 and 9-12). Next, all SNPs that have a z-score lower than -Z or higher than +Z in GWAS II are determined (SNPs 1-8 and 13-16). The number of shared significant SNPs is divided by the union of significant SNPs to calculate the Jaccard index. (B) We also calculate the Pearson correlation of the z-scores of the shared SNPs.

(Fig 2, Table S3). For instance, in CES, h2 of union is 0.139 ± 0.009 and h2 of intersect is 0.275 ± 0.017. We additionally found that the second highest heritability in large artery and small vessel stroke was in CCSc (h2-LAS = 0.258 ± 0.023 and h2-SVS = 0.315 ± 0.029), which assigns only one subtype to each case. The heritabilities for CCSc, CCSp and TOAST were not significantly different from one another in cardioembolic stroke (Table S4), indicating that each original subtyping system is capturing approximately the same proportion of genetic risk

Different phenotype definitions represent genetically distinct phenotypes

While heritability gives an estimation of how much variation in a phenotype can be attributed to genetic factors, it does not show how different two phenotypes are from one another (i.e., two phenotypes can have the same heritability and yet be genetically distinct from each other). We therefore evaluated the overlap in significant SNPs for all pairwise combinations of phenotypes for which we performed a GWAS, where high proportions of shared SNPs between two phenotypes indicate genetic similarity. At multiple significance cutoffs,

we assessed overlap of significant SNP sets using two complementary similarity measures: the Jaccard index, which measures the ratio of overlapping SNPs (those are significant in both analyses) in the total set of SNPs that are significant in either analysis; and the Pearson correlation of the z-scores of the overlapping SNPs in both analyses (Fig 3). Significance is defined here as an absolute z-score that is higher than the selected z-score threshold (where SNPs can have an effect size < -Z or > +Z). A high Jaccard index indicates that two phenotypes share many of their associated SNPs, while a low Jaccard index means that the phenotypes have distinct genetic architecture. Correlation pertains only to the shared SNPs and indicates if they have similar directionality and magnitude of effect in both analyses.

In order to assess the results of the overlap analyses and their meaning with respect to the ischemic stroke phenotypes, we also performed these analyses between the phenotype definitions and an unrelated GWAS of educational attainment to obtain a null reference (Fig S3).

In cardioembolic stroke (Fig 4, first panel), the Jaccard index for all combinations with intersect decreases with more extreme z-scores to J≈0.2-0.3 while the correlation increases quickly to approach r2=1 at Z≈2.5, indicating that a relatively small group of SNPs is significant in both analyses with correlating z-scores, that gets increasingly smaller and stronger correlating. These findings indicate that the stricter the significance threshold is, the fewer shared SNPs there are between any two phenotypes, but that those shared SNPs have more concordant effect sizes. In large artery stroke (Fig S4) and small vessel stroke (Fig S5) the trend is similar, albeit with lower Jaccard indices and correlations, suggesting that there is a set of associated SNPs for each subtype that is found by all phenotype definitions. In all subtypes, when compared to symmetric difference, the intersect is the most genetically distinct phenotype. This confirms that if we combine symmetric difference and intersect, as in the union, we increase phenotypic heterogeneity and thereby decrease the likelihood of detecting a genome-wide significant signal.

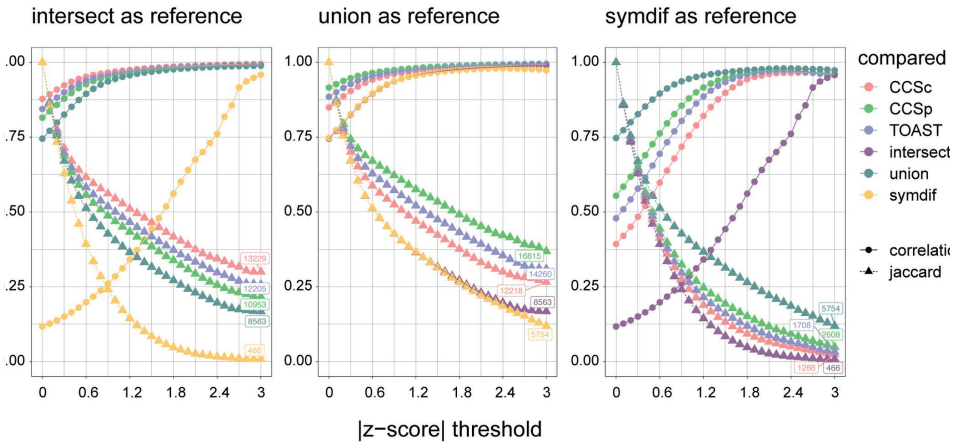


Figure 4. Different phenotype definitions capture different genetic risk factors. Overlap analysis in cardioembolic stroke. Similarity on the y-axis denotes either correlation (circles) or Jaccard index (triangles). The absolute z-score threshold is plotted on the x-axis. Numbers indicate the number of shared SNPs at Z = 3. (A) pairwise comparisons with intersect (B) pairwise comparisons with union (C) pairwise comparisons with symmetric difference.

Fig 4 shows pairwise comparisons only; to investigate if there is one group of SNPs that is significant in all analyses, we also calculated overall Jaccard index: the size of the intersect of SNPs that are significant in all 5 phenotypes (excluding symmetric difference, which we use for sensitivity testing only), divided by the size of their union. The overall Jaccard index (Fig 5) confirms what was suggested by the pairwise overlap analyses: there is a small set of SNPs that is shared across all phenotype definitions, albeit slightly smaller than the pairwise overlapping sets. The Jaccard index is relatively low at higher significance thresholds, indicating that there is also a substantial set of SNPs that is unique to each phenotype definition. Thus, we do find different associated SNPs to ischemic stroke subtypes depending on how exactly the subtype status is defined, but there are some concordant SNPs that are found by all case definitions, regardless of sample size or phenotype homogeneity.

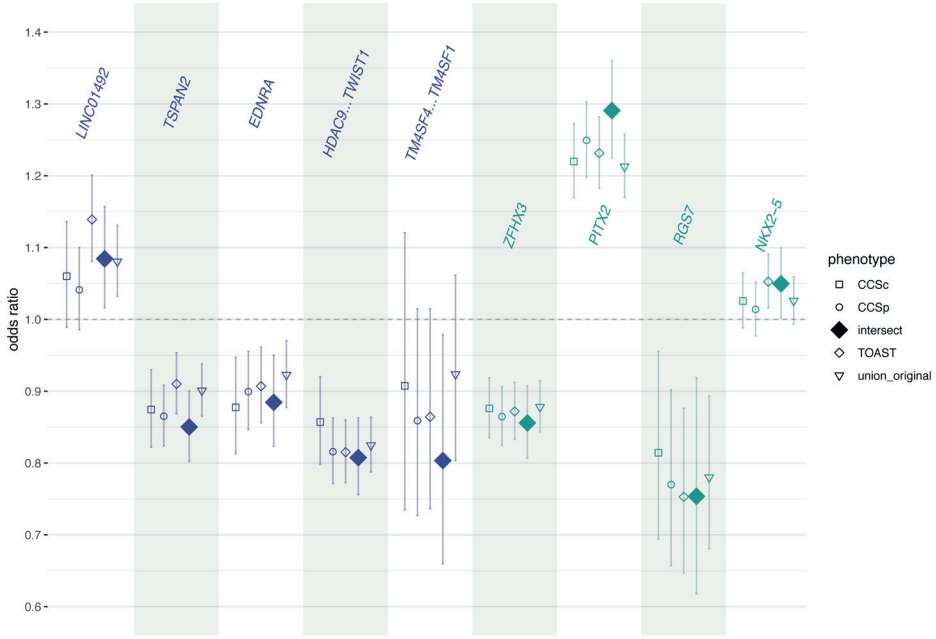


Figure 5. Intersect most often shows the strongest effect at previously identified subtype-specific associations. Odds ratios for the five LAS-associated SNPs (in purple) and the four CES-associated SNPs (in teal) in the five phenotype definitions. The dotted line indicates an OR of 1 (no effect). Error bars indicate the 95 % confidence interval. Intersect show the strongest effect at 5 of the 9 SNPs (binomial p = 0.0196).

Intersect shows the largest effect at previously known associations

A recent GWAS (MEGASTROKE) in 67,162 TOAST-subtyped cases and 454,450 controls identified 32 loci (22 novel) associated to stroke (either ischemic stroke or intracerebral hemorrhage) and its subtypes [2]. Four of the 32 loci associate to CES, five to LAS, and none to SVS. We investigated the potential to find stroke-associated loci in our redefined phenotypes, with a sample that is 4 to 7 times smaller than MEGASTROKE. To this end, we compared the odds ratios for the 9 known subtype-specific loci in our five phenotype definitions, see Fig 6. In cardioembolic stroke, the intersect phenotype consistently shows the strongest effect. In large artery

stroke, intersect shows the strongest effect as well, except at the LINC01492 locus. The fact that we find stronger associations at known loci by using the intersect indicates that this phenotypically more homogeneous phenotype is better suited as a phenotype in GWAS.

Besides comparing the ORs at subtype-specific signals, we also compared ORs at all stroke-associated loci (including any stroke, any ischemic stroke, cardioembolic stroke and large artery stroke), see Fig S1. We found that intersect shows the strongest odds ratio 30 times out of 96, (binomial p = 0.010), indicating that odds ratios derived from the intersect phenotype are indeed stronger than the ORs in the other phenotypes more often than expected by chance.

A stricter phenotype definition finds a new associated locus to small vessel stroke

Our analyses revealed 5 new loci (2 for SVS and 3 for CES, Table 2) which we validated using data from MEGASTROKE (based on the summary statistics of MEGASTROKE with the SiGN cohort removed, to ensure independence), while correcting for multiple testing per stroke subtype.

For SVS one variant (rs10029218) in the CAMK2D locus (Table 2, Figure S5), was found in the intersect analysis, and replicated in the trans-ancestry analysis of MEGASTROKE. The other SVS associated variant (rs11065979) in the SH2B3-BRAP-ALDH2 locus was found in the CCSp analysis, and replicated in both the trans-ancestry analysis and the European analysis (Table 2, Figure S5). For the 3 CES loci, only one variant (rs3790099, in the GNAO1 gene, found in the CCSp analysis) was replicated in the Europeans-only analysis (Table 2, Figure S5). In a meta-analysis of a) the MEGASTROKE GWAS without the SiGN cohort and b) the SiGN GWAS for these three SNPs, we found consistent direction of effect in both studies and a lower p-value (Table S5).

Previously, one other locus was reported to associate solely with SVS (16q24 [11]). Here, by applying stricter phenotyping, we identify 4p12 as a novel SVS locus. In general, despite the low sample size

Locus	SNP	Chr	A1	A2	Analysis	P-value	Freq1	OR	Beta	SE
CAMK2D	rs10029218	4	A	G	SVS-intersect	1,20E-8	0,12	1,27	0,02	0,00
					SVS-rep-EUR	2,46E-2	0,11	1,11	0,10	0,05
					SVS-rep-TRANS	5,98E-3	0,13	1,12	0,11	0,04
SH2B3 - BRAP - ALDH2	rs11065979	12	T	C	SVS-CCSp	9,40E-9	0,42	1,13	0,01	0,00
					SVS-rep-EUR	7,62E-3	0,43	1,08	0,08	0,03
					SVS-rep-TRANS	9,29E-3	0,41	1,08	0,07	0,03
PFH20	rs11697087	20	A	G	CES-intersect	3,20E-9	0,09	1,26	0,02	0,00
					CES-rep-EUR	1,55E-2	0,09	1,10	0,10	0,04
					CES-rep-TRANS	4,76E-2	0,09	1,07	0,07	0,03
5:114799266	rs2169955	5	T	C	CES-CCSc	3,90E-8	0,57	0,90	-0,01	0,00
					CES-rep-EUR	1,48E-2	0,56	0,95	-0,05	0,02
					CES-rep-TRANS	2,22E-2	0,56	0,96	-0,04	0,02
GNAO1	rs3790099	16	C	G	CES-CCSp	4,90E-8	0,85	0,87	-0,02	0,00
					CES-rep-EUR	2,97E-4	0,84	0,89	-0,11	0,03
					CES-rep-TRANS	1,10E-2	0,77	0,94	-0,07	0,03

Table 2. Summary statistics for the new genome-wide significant SNPs. Per locus, the SiGN GWAS is in the first row, in the format 'subtype-phenotype'. In the other rows, results in MEGASTROKE are shown with 'subtype-rep-EUR' for the Europeans-only analysis, and with 'subtype-rep-TRANS' for the trans-ancestry analysis. A1 = Allele 1, A2 = Allele 2, Freq1 = frequency of allele 1, OR = odds ratio, Beta = coefficient, SE = standard error. NB, Beta and SE of SiGN GWAS and MEGASTROKE GWAS are not comparable since they come from linear and logistic regression, respectively. The ORs are comparable. We did a Bonferroni correction: for SVS, □ = 0.0125 and for CES, □ = 0.00625. Replication p-values below the threshold are indicated in bold. Two SNPs (rs2169955:C>T and rs62379973:C>G, in CES-CCSc) that are relatively close (260 kb) on chromosome 5 were in two different clumps, even though they are in LD (r2 = 0.52, D' = 0.87, in a CEU population [10]) because the distance is just above the threshold (250 kb). Because they are in LD, and just a little farther apart than 250 kb, they were considered to be from the same locus and only the strongest association was kept (rs2169955:C>T).

as compared to MEGASTROKE, we find stronger associations in the intersect GWAS, likely due to the clearer separation of cases and controls. This further supports the claim that, in the absence of a 'gold standard' phenotype, taking the intersect of all subtyping systems yields a better suited phenotype for GWAS.

Discussion

To help uncover genetic associations with ischemic stroke that as yet have gone undetected, we defined new ischemic stroke phenotypes based on three existing subtyping systems (CCSc, CCSp, and TOAST). Specifically, we studied the intersect and union of these subtyping systems, for all ischemic stroke subtypes. The intersect results in a smaller number of available cases but potentially results in less misclassification due to agreement between subtyping systems. The union is potentially more heterogeneous, but results in a larger available group of cases. We find that the largest proportion of phenotypic variance explained by SNPs is in the intersect phenotype. Further, our overlap analyses show that, for each subtype, the phenotype definitions each have a unique set of significantly associated SNPs, but that there is also a small set of SNPs that is shared among all definitions, with concordant direction of effect and similar trend in magnitude of effect. We also show that the cases that are in the union but not in the intersect, are genetically distinct from the intersect-cases, implying that the union is a combination of phenotypically heterogeneous cases. With an effective sample size that is 4 to 7 times as small as in MEGASTROKE, we find stronger associations (i.e., higher ORs and lower p-values) at known loci by using the intersect (compared to the other phenotype definitions studied here). This indicates that the intersect yields more net power to detect associations due to its stricter definition, despite its lower sample size, and is thus better suited as a phenotype in GWAS.

We identify a previously sub-threshold association with a SNP in an intron of the CAMK2D locus in small vessel stroke by using the intersect, further demonstrating the utility of this phenotype in GWAS.

CAMK2D expresses a calcium/calmodulin-dependent protein kinase [12]; out of all tissues tested in gTEX, the two tissues with the highest expression are both in brain [13]. The CAMK2D locus was found to also associate with the P-wave [14], an electrocardiographic property that is implicated in atrial fibrillation, a trait that is associated with cardioembolic stroke [3]. Given that the association replicates in an independent dataset, and the protein is expressed in brain, functional follow-up of this locus might give more insight in disease mechanisms of stroke. Additionally, we find the SH2B3 - BRAP - ALDH2 locus to be associated with small vessel stroke. rs11065979 is an eQTL of ALDH2 (aldehyde dehydrogenase 2) [13]. ALDH2 is involved in ethanol metabolism; it converts one of the products, ethanal, into acetic acid. The allele that is associated with higher expression of this enzyme, is associated with lower incidence of small vessel stroke . ALDH2 is mainly expressed in liver, but it's also expressed in brain. [13] Previous work has shown an association between higher expression of ALDH2 and lower incidence of stroke in rats. [15] SH2B3 and BRAP are minimally expressed in brain, compared to the other tissues. [13] We also show an association between the GNAO1 locus and cardioembolic stroke. The protein product of this locus constitutes the alpha subunit of the Go heterotrimeric G-protein signal-transducing complex. [12] It is highly expressed in brain, and while its function is not completely clear, defects in the protein are associated with brain abnormalities. [16] Although this alternate approach to phenotyping has resulted in new associations with two ischemic stroke subtypes, the causality of these loci remains uncertain and warrants further studies.

Phenotype definition is an often-encountered issue in complex trait genetics, as diagnosing and subtyping methodologies can vary and even be contentious within disease areas. Further, phenotype labels are often broad definitions for cases that can be highly heterogeneous when their underlying genetic risk is examined. For example, most psychiatric diseases are also complex and phenotypically heterogeneous, lacking clear and robust diagnostic criteria. In an editorial, the Cross-Disorder Phenotype Group of the Psychiatric

GWAS Consortium states: "We anticipate that genetic findings will not map cleanly onto current diagnostic categories and that genetic associations may point to more useful and valid nosological entities". Our findings here further support this statement, showing that while the original subtyping systems might be useful for diagnosing individual patients, stricter criteria are needed for genetic studies.

Methods

The SiGN dataset

The Stroke Genetics Network (SiGN) Consortium composed a dataset consisting of 14,549 ischemic stroke cases. [17] The control group consists primarily of publicly available controls drawn from three large multi-ancestry cohorts. Descriptions of the contributing case and control cohorts have been published previously. [18] Cases and controls have been genotyped on a variety of Illumina arrays, and nearly all cases (~90%) have been subtyped using both TOAST [4] and CCS [19]. All newly-genotyped cases for the latest GWAS are available on dbGAP (accession number phs000615.v1.p1). A previous genome-wide association study has been done on the separate TOAST and CCS subtypes. [18] In this work, we use the same 28,026 controls from this previous GWAS, as well as the 13,930 ischemic stroke cases of European and African ancestry. A third group of cases and controls, primarily comprised of individuals who identify as Hispanic and residing in the United States, has been excluded due to data sharing restrictions. All data processing has been previously described. [18] All genotyping data was generated using human genome build hg19.

Genome-wide association studies in BOLT-LMM

We ran all GWAS in BOLT-LMM [9], which implements a linear mixed model (LMM). BOLT-LMM implements a Leave-One-Chromosome-Out (LOCO) scheme, so that the genetic relationship matrix (GRM) is built on all chromosomes except the chromosome of the variant being tested. Linear mixed models have been demonstrated to improve power in GWAS while correcting for

structure in the data [20]. In addition to the GRM, we included the first 10 principal components as fixed effects. We used the following approximation to convert the effect estimates from BOLT-LMM (on the observed scale) to effect estimates on the liability scale:

$$log(OR) = \frac{\beta}{\mu(1-\mu)}$$

where μ is the case fraction. [21] For each subtype, the intersect, union and symmetric difference of the original subtyping systems were used as phenotypes in separate GWAS. The original subtyping systems were also used as a phenotype in three additional GWAS per subtype to serve as a point of reference. All ischemic stroke cases that do not belong to the case definition under study were left out of the analysis. The same group of controls is used in all analyses. Association testing was done on all imputed SNPs with a minimum minor allele frequency of 1%. See Supplementary Table 8 in [22] for simulations of type 1 error inflation of BOLT-LMM in datasets with unbalanced case-control ratios. In the GWAS discussed here, case fractions range from 0.05 to 0.14 which means that at variants with MAF >1%, there is no significant inflation of type 1 error rates. Those SNPs that show a large frequency difference (>15%) across the populations in 1000 Genomes were removed (see the methods in [18] for details on how this list of SNPs was compiled). See Fig S2 for QQ-plots (stratified by imputation quality (INFO-score) and by minor allele frequency) and Manhattan-plots. The genomic inflation factor (lambda) varies between 1.0 and 1.1 for cardioembolic stroke and large artery stroke, and between 1.0 and 1.2 for small vessel stroke. We observed a relatively high inflation factor of 1.2 in only the imputed SNPs with a minor allele frequency lower than 5%. Therefore, summary statistics for these SNPs were removed from downstream analyses.

Heritability estimation in BOLT-REML

To estimate the heritability of the six phenotype definitions for each subtype, we used BOLT-REML [23]. BOLT-REML calculates heritability

from the SNPs included in the GRM, and these SNPs must be genotyped (and not imputation dosages). We therefore based our estimates on only genotyped SNPs. Furthermore, we excluded the MHC on chromosome 6, and chromosomal inversions on chromosomes 8 and 17 using PLINK 1.9 [24]. See Table 3 for more information. We filtered on various quality control measures, by passing the following flags to PLINK: --mind 0.05 --maf 0.10 --geno 0.01 --hwe 0.001. Additionally we pruned SNPs at an LD (r2) threshold of 0.2 (--indep-pairwise 100 50 0.2). We used the first 10 principal components and sex (determined by presence of XX or XY chromosomes) as covariates. To convert the heritabilities from the observed scale (as if the binary data, coded as 0-1, were continuous) to the liability scale (converting the heritabilities of the observed binary trait to the heritabilities of the underlying, unobserved, continuous liability of the trait), Dempster et al derived a formula that takes into account the prevalence of the disease in the population [25]. In the case of ascertained case-control traits, where the population prevalence is not equal to the study prevalence, this has to be taken into account as well [26]:

$$\hat{h_l^2} = \frac{K^2(1-K)^2}{P(1-P)\phi(t)^2}\hat{h_o^2}$$

where $\hat{h_l^2}$ is the heritability on the liability scale, K is the population prevalence, P is the study prevalence, t is the liability threshold, and $\hat{h_o^2}$ is the heritability on the observed scale. To test for significant difference between the estimated heritabilities, we performed an independent t-test.

Table 3. Genomic regions removed before heritability estimation

Chromosome	Start (Mb)	End (Mb)	Name
6	25.8	36.0	MHC
8	6.0	16.0	inversion
17	40.0	45.0	inversion

Overlap analysis

We first calculated z-scores using the following formula: z = beta / se, where beta is the effect size of the SNP and se is the standard error of the beta estimate. The z-scores thus have unit standard error, but we did not standardize them to zero mean (as is the conventional method for calculating z-scores) to maintain the original direction of effect. To assess overlap between two GWAS, we calculated the Jaccard index [27], which is the ratio of a) the number of SNPs significant in both analyses, to b) the number of SNPs significant in either analysis (i.e., the size of the intersect divided by the size of the union of the sets of significant SNPs). The index is a number between 0 and 1: it is 0 if the two sets of significant SNPs do not have any SNPs in common, and it is 1 if the two sets of significant SNPs completely overlap. We additionally calculated, within the set of SNPs that are significant in both analyses, the Pearson's correlation of the z-scores in the two GWAS to check the concordance of direction and size of effect in the two analyses being compared. Significance was defined as a z-score that is more extreme than an absolute z-score threshold z (varied from 0 until 3, in increments of 0.1). At the most extreme z-score threshold ($z > 3$ or $z < -3$), the absolute number of SNPs that are significant in both analyses is indicated in the plot. As a null comparator, these overlap analyses were also performed with GWAS results from a study of educational attainment in 1.1 million individuals [28] downloaded from EMBL-EBI's GWAS catalog. [29] The educational attainment study contains 10,098,325 SNPs, the SiGN study contains 10,156,805 SNPs. The overlap analysis was only done on the SNPs that are present in both datasets: the size of this overlapping set is 7,822,831 SNPs. For the overall comparisons per subtype, we considered all five GWAS. At each z-score threshold, we calculated the overall Jaccard index: the ratio (range between 0 and 1) of the number of SNPs significant in all five analyses to the number of SNPs significant in any analysis. See Fig 3 for a graphical explanation of this method.

Look-up of Megastroke loci in the union and intersect GWAS

Recently, the MEGASTROKE consortium completed the largest GWAS in ischemic stroke and its subtypes [2]. From this GWAS, we extracted the index SNPs of each genome-wide significant locus in each subtype. We then looked up these SNPs in our GWAS to compare effect sizes, resulting in 15 ORs per SNP (for each of the phenotype definitions in each of the subtypes). See Table S6 for the summary statistics of these look-ups. If the reference allele in MEGASTROKE was not identical to the reference allele in SiGN, the inverse of the odds ratio (1/OR) was taken. We counted how often the intersect showed the most extreme odds ratio, out of all 96 ORs (15 ORs per SNP, for the 32 SNPs that were reported in MEGASTROKE). To determine the probability of the number of times intersect was most extreme, under the null hypothesis that all phenotype definitions are just as likely to show the most extreme OR, we performed a binomial test in R[30].

Replication of new genome-wide hits in MEGASTROKE

To assess all genome-wide significant loci instead of the individual SNPs, we performed clumping in PLINK 1.9 [24] (http://pngu.mgh.harvard.edu/purcell/plink/). We used all SNPs significant at α = 1x10-5 as index SNPs. We generated clumps for all other SNPs closer than 250 kb to the index SNP and in LD with the index SNP (r2 > 0.05). We kept clumps if the p-value of the index SNP was lower than 5x10-8. From the genome-wide significant clumps, only the unique ones were kept (some clumps significantly associated to multiple case definitions). In the case of duplicates, the summary statistics for the analysis with the lowest p-value were kept. Ambiguous SNPs were removed, and if the reference allele in MEGASTROKE was not identical to the reference allele in SiGN, we calculated the inverse of the odds ratio (1/OR). This resulted in a list of 14 unique SNPs. We checked for SNPs that are not in a locus that had already been reported as an associated locus in MEGASTROKE, resulting in a list of 5 new SNPs (2 for SVS and 3 for CES), which we looked up for replication. To this end, we ran the MEGASTROKE GWAS again (European and trans-ancestry analysis per

subtype using TOAST [31]) without the SiGN cohort, to ensure summary statistics independent from the discovery GWAS. We set Bonferroni p-value thresholds to adjust for the number of SNPs looked-up for the phenotype in question, and for the number of GWAS it was looked up in (2, for the European and trans-ancestry analyses). We did a meta-analysis of the MEGASTROKE GWAS without SiGN, and the SiGN GWAS, for the 3 replicating SNPs only (Table S5). We performed meta-analysis in METAL [32], with the inverse-variance weighting scheme.

Acknowledgements

S.L.P. is supported by Veni Fellowship 016.186.071 from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO). J.v.B. is supported by R01NS100178 from the National Institute of Health. SWvdL is funded through grants from the Netherlands CardioVascular Research Initiative of the Netherlands Heart Foundation (CVON 2011/B019 and CVON 2017-20: Generating the best evidence-based pharmaceutical targets for atherosclerosis [GENIUS I & II]), ERA-CVD 'druggable-MI-targets' (grant number: 01KL1802), and Fondation Leducq 'PlaqOmics'. JdR is supported by a Vidi Fellowship (639.072.715) from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO).

References

1. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. Circulation. 2017;135: e146–e603.

2. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet. 2018;50: 524–537.

3. Pulit SL, Weng L-C, McArdle PF, Trinquart L, Choi SH, Mitchell BD, et al. Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes. Neurology Genetics. Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology; 2018;4: e293.

4. Adams HP, Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke. 1993;24: 35–41.

5. Arsava EM, Ballabio E, Benner T, Cole JW, Delgado-Martinez MP, Dichgans M, et al. The Causative Classification of Stroke system: an international

reliability and optimization study. Neurology. 2010;75: 1277–1284.

6. McArdle PF, Kittner SJ, Ay H, Brown RD Jr, Meschia JF, Rundek T, et al. Agreement between TOAST and CCS ischemic stroke classification: the NINDS SiGN study. Neurology. 2014;83: 1653–1660.

7. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9: 356–369.

8. Pulit SL, McArdle PF, Wong Q, Malik R, Gwinn K, Achterberg S, et al. Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. Lancet Neurol. Elsevier; 2016;15: 174–184.

9. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47: 284–290.

10. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31: 3555–3557.

11. Traylor M, Malik R, Nalls MA, Cotlarciuc I, Radmanesh F, Thorleifsson G, et al. Genetic variation at 16q24.2 is associated with small vessel stroke. Ann Neurol. 2017;81: 383–394.

12. Acids research N, 2016. UniProt: the universal protein knowledgebase. academic.oup.com. 2016; Available: https://academic.oup.com/nar/article-abstract/45/D1/D158/2605721

13. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv Biobank. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA; 2015;13: 311–319.

14. Christophersen IE, Magnani JW, Yin X, Barnard J, Weng L-C, Arking DE, et al. Fifteen Genetic Loci Associated With the Electrocardiographic P Wave. Circ Cardiovasc Genet. 2017;10. doi:10.1161/CIRCGENETICS.116.001667

15. Guo J-M, Liu A-J, Zang P, Dong W-Z, Ying L, Wang W, et al. ALDH2 protects against stroke by clearing 4-HNE. Cell Res. 2013;23: 915–930.

16. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet. 2007;80: 588–604.

17. Meschia JF, Arnett DK, Ay H, Brown RD Jr, Benavente OR, Cole JW, et al. Stroke Genetics Network (SiGN) study: design and rationale for a genome-wide association study of ischemic stroke subtypes. Stroke. 2013;44: 2694–2702.

18. NINDS Stroke Genetics Network (SiGN), International Stroke Genetics Consortium (ISGC). Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. Lancet Neurol. 2016;15: 174–184.

19. Ay H, Furie KL, Singhal A, Smith WS, Gregory Sorensen A, Koroshetz WJ. An evidence-based causative classification system for acute ischemic stroke. Ann Neurol. 2005;58: 688–697.

20. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46: 100–106.

21. BOLT-LMM v2.3.2 User Manual [Internet]. 11 Jun 2018 [cited 25 Jan 2019]. Available: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/#x1-5200010.2

22. Loh P-R, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed model association for biobank-scale data sets [Internet]. 2017. doi:10.1101/194944

23. Loh P-R, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, et al. Contrasting genetic architectures of schizophrenia and other complex diseases

using fast variance-components analysis. Nat Genet. 2015;47: 1385–1392.

24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81: 559–575.

25. Dempster ER, Lerner IM. Heritability of Threshold Characters. Genetics. 1950;35: 212–236.

26. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011;88: 294–305.

27. Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. 1901.

28. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018;50: 1112–1121.

29. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45: D896–D901.

30. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available: https://www.R-project.org/

31. Malik R, Rannikmäe K, Traylor M, Georgakis MK, Sargurupremraj M, Markus HS, et al. Genome-wide meta-analysis identifies 3 novel loci associated with stroke. Ann Neurol. 2018;84: 934–939.

32. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans [Internet]. Bioinformatics. 2010. pp. 2190–2191. doi:10.1093/bioinformatics/btq340

Chapter 2
GWAS of age at onset of disease can identify novel associations, but is potentially biased by associations with earlier death

Joanna von Berg[1,2], Patrick F. McArdle[3], Paavo Häppölä[7], Charles Kooperberg[8], SiGN consortium, Finngen , Women's Health Initiative, Steven J. Kittner[4,5], Braxton D. Mitchell[3,4], Jeroen de Ridder[1,2], Sander W. van der Laan[6]

1. Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands, 2. Oncode Institute, Utrecht, The Netherlands, 3. Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA 4. Geriatric Research and Education Clinical Center, VA Maryland Health Care System, Baltimore, MD, USA 5. Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA 6. Central Diagnostics Laboratory, Division Laboratories, Pharmacy, and Biomedical Genetics, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands 7. FinnGen 8. Women's Health Initiative

This chapter is based on a manuscript that is in preparation. Scan this QR code to access the supplemental information:

Introduction

Disease-related genetic variants are often discovered by comparing their allele frequencies in cases - who have the disease - and controls - who do not have the disease. Case-control genome-wide association studies (GWAS) have uncovered innumerable genetic associations with risk of disease. For ischemic stroke (IS), which is one of the main causes of death[1], around 40 risk-increasing single-nucleotide polymorphisms (SNPs) have been identified[2,3].

Linear regression of the age at onset (AAO) of disease can be employed as an alternative to case-control analysis. In case-control studies, the effect that is measured is an odds ratio: how much does risk increase with each additional copy of the associated allele? We generally assume that a lower AAO of a complex disease corresponds with a higher genetic risk contribution. Thus a case-control analysis restricted to lower ages at onset would find the same risk alleles as an unrestricted analysis, with larger effect sizes. If we use age at onset itself as the phenotype, the effect that is measured is a linear coefficient: how much earlier does a stroke occur with each additional copy of the associated allele, given one has had an ischemic stroke?

Age at onset has been employed for genetic association studies in stroke before. Traylor et al [4] used LTSOFT [5] to estimate liabilities based on age at onset, which they used as a GWAS phenotype. They identified an association at the MMP12 locus. Liability estimates can increase power for variant discovery because they incorporate covariates directly into the phenotype instead of correcting for them in logistic regression. [6] However, interpretation of SNP effect sizes on liability is difficult; the liability can not be observed in real life and it does not directly relate to risk. The effect size of AAO-related SNPs is expressed in years, and is more interpretable. Yet, to date, only 32 AAO traits are registered in the GWAS catalog (EFO_0004847). [7] To our knowledge, no GWAS of AAO of ischemic stroke has been published before.

IS does not occur equally often in women and men. [1] This could mean that the same biological mechanisms contribute more to IS onset in one sex than in another, which would result in sex-differential effect sizes for the SNPs related to these mechanisms. It is also possible that there are vastly different mechanisms that lead to IS in women and men, which would result in sex-specific effects. For most complex traits no sex-stratified GWAS have been performed, meaning that possible sex-specific effects will remain unidentified unless different sexes are analysed separately [8].

In this study we performed a two-stage genome-wide association study of ischemic stroke AOO. Stage 1 consisted of a GWAS of AOO in 10,857 stroke cases from SiGN [9], with a look-up of all associated SNPs (p-value below a liberal threshold of 5e-6) in FinnGen [10] in Stage 2. We perfomed a sex-combined and sex-stratified analysis. In a meta-analysis of Stages 1 and 2, we identified a variant in the ApoE locus, encoding the APOE4 allele, that was significantly associated (rs429358:T>C, meta p-value = 2.4e-8, beta = -1.63 years) with earlier onset in women. As this SNP is not associated with risk of stroke per se, and has previously been associated with stroke AOO [11], we hypothesized that the association with earlier AOO may reflect a co-occuring association of this variant with earlier death. [12] To test this hypothesis, we performed a simulation study whereby we simulated loci that are associated with overall mortality but not stroke risk and show that loci with an approximately two-fold increased risk in mortality via mechanisms not related to IS, would display this pattern.

Results

A missense variant in ApoE is associated with age at ischemic stroke in cases of European ancestry

The complete SiGN dataset consists of 14,549 IS cases from several cohorts. [9] Inclusion criteria for the cohorts for the current study were: complete information on the age at onset of IS, and no restrictions on age of cases for inclusion in the cohort. The latter criterion was used to prevent spurious associations due to possible differences in genetic background between cohorts with different age distributions.

Meta-analysis (in SiGN and FinnGen) of the effect sizes on AAO in women of European ancestry resulted in a genome-wide significant association at the Apolipoprotein E (ApoE) locus on chromosome 19 (rs429358:A>B, meta pmeta-value: 2.4x10e-8, beta = -1.63 years ± 0.29, Table 1, Table S2, Figure S1, Figure S2). This association is replicated in the Women's Health Initiative (p-value = 76.,097x10e-5) at an alpha of 0.05. Conditional analysis indicated no secondary associated SNPs at this locus. The top hit is a missense variant in the ApoE gene that changes the amino acid at the 112th position of the protein from a cysteine to an arginine. This changes the protein confirmation, and together with another SNP, rs7412:C>T, this SNP is used to determine an individual's ApoE isoform.

	A2 freq	INFO	Beta	SE	P-value	95% confidence interval		N	Mean age
SiGN - XX	0.13	0.96	-1.78	0.40	9.70E-06	-2.56	-1.00	4679	71.6
FinnGen - XX	0.17	1	-1.47	0.43	5.70E-04	-2.3	-0.63	3416	67.8
meta-analysis	0.15	0.96	-1.63	0.29	2.40E-08	-2.11	-1.15		
WHI - XX	0.13	WGS	-1.02	0.26	6.97E-05	-1.52	-0.52	3415	76.6

Table 1. Summary statistics of rs429358 in women, in the SiGN, Finngen and WHI cohorts as well as in the meta-analysis results of SiGN and FinnGen. Further details of the SNP and of the baseline characteristics of each cohort can be found in tables S1 to S4. A2 freq = frequency of the alternative allele, INFO = imputation quality, N = sample size.

Genetic effect sizes tend to be larger in women

ApoE rs429358 is associated with stroke AOO in both men and women, although the magnitude of association is stronger in women than men (Figure 1, two-sided unequal variances t-test, p = 4.3x10-4).

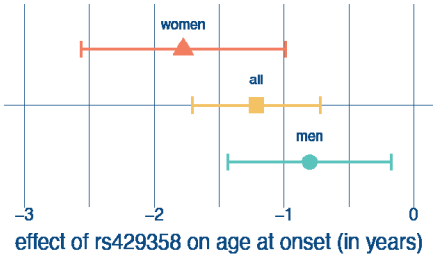


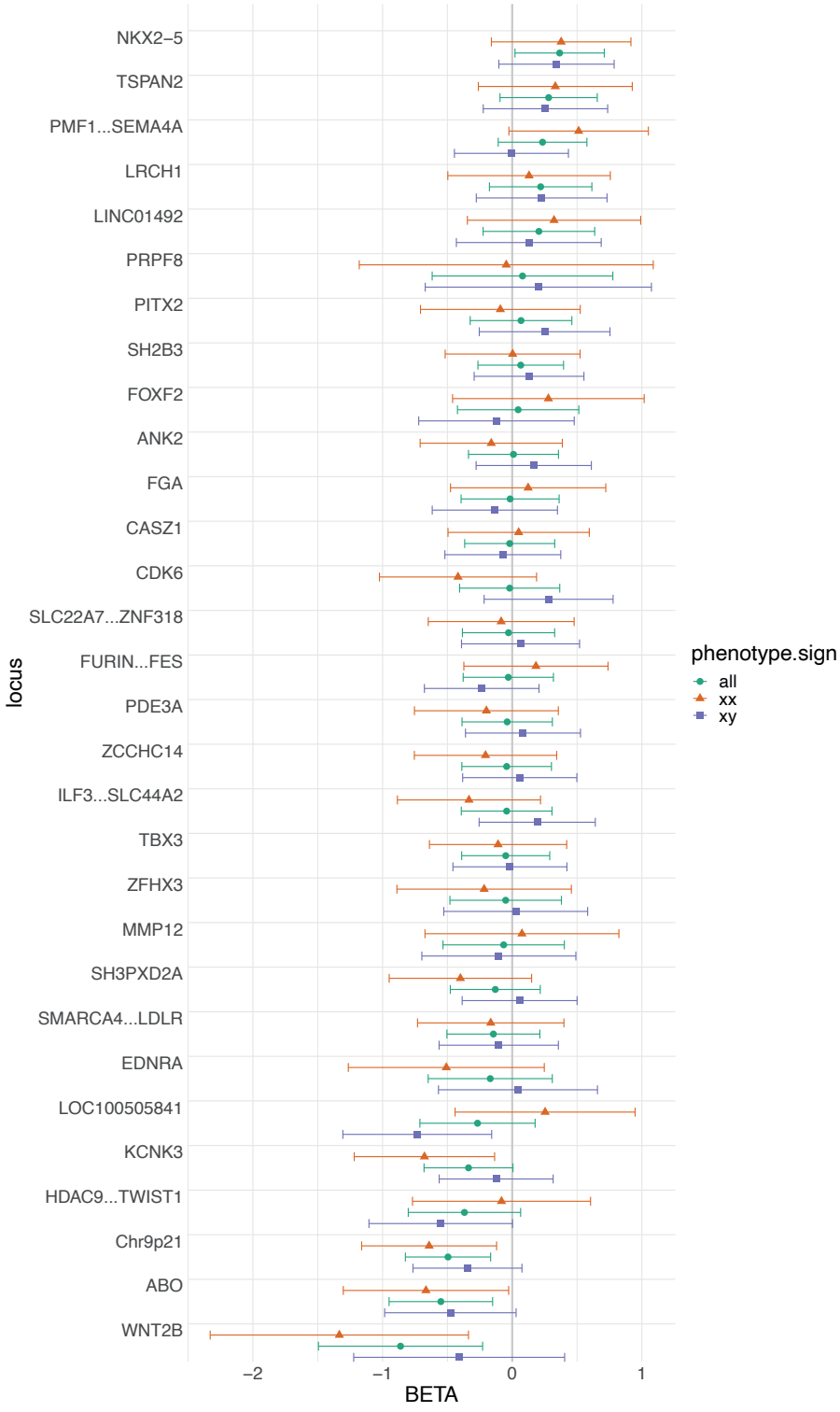
Fig 1. The ApoE locus exhibits sex-differential effects. Effect sizes beta (in years), for rs429358; 95 % confidence intervals are indicated by error bars.

We looked up previously known risk loci for stroke in our AAO analyses (Figure 2, Table S5). Overall, when there is a sex-differential effect, the strongest effect seems to be observed in women. The most extreme example is the WNT2B locus, which is nominally significant in women.

The ApoE locus is associated with stroke age at onset but not with risk of ischemic stroke

ApoE rs429358 is not associated with risk of ischemic stroke in MEGASTROKE [2] (OR = 1.00; 95% CI: 0.96-1.03 ; p = 0.77), nor is it associated with any stroke subtype (Figure 3, Table S6). The APOE locus shows the interesting pattern of being associated with AAO but not stroke risk (all 95% CI contain the null, OR = 1.0).

Fig 2 (next page). Some stroke risk loci show a sex-differential effect on age at onset. For each SNP in a list of loci that were previously shown to associate with binary risk of any stroke phenotype (composed by Malik et al [2]), we looked up the summary statistics in the AAO analyses. The effect sizes (beta) in years are ordered by their absolute effect size in 'all'; 95 % confidence intervals are indicated by error bars.


NKX2–5
TSPAN2
PMF1...SEMA4A
LRCH1
LINC01492
PRPF8
PITX2
SH2B3
FOXF2
ANK2
FGA
CASZ1
CDK6
SLC22A7...ZNF318
FURIN...FES
PDE3A
ZCCHC14
ILF3...SLC44A2
TBX3
ZFHX3
MMP12
SH3PXD2A
SMARCA4...LDLR
EDNRA
LOC100505841
KCNK3
HDAC9...TWIST1
Chr9p21
ABO
WNT2B
locus
−2
−1
0
1
BETA
phenotype.sign
all
xx
xy

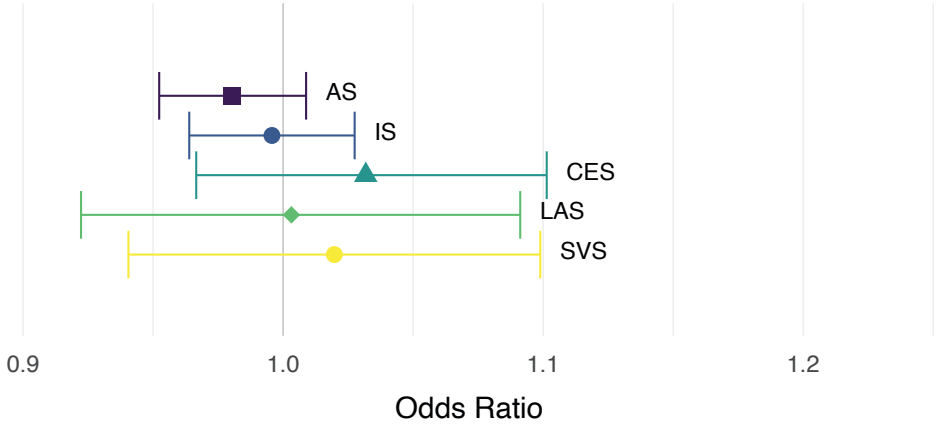


Fig 3. ApoE rs429358 is not associated with risk of all stroke, all ischemic stroke, cardioembolic stroke, large artery stroke, or small vessel stroke. Effect sizes for rs429358 from MEGASTROKE analysis of European ancestry [2] (performed with all SiGN cases left out); 95 % confidence intervals are indicated by error bars. any stroke (AS), any ischemic stroke (IS), cardioembolic stroke (CES), large artery stroke (LAS), and small vessel stroke (SVS). Any stroke includes IS and hemorrhagic stroke, and IS includes all subtypes (CES, LAS and SVS) and cases who could not be subtyped. The effect sizes are shown as odds ratios.

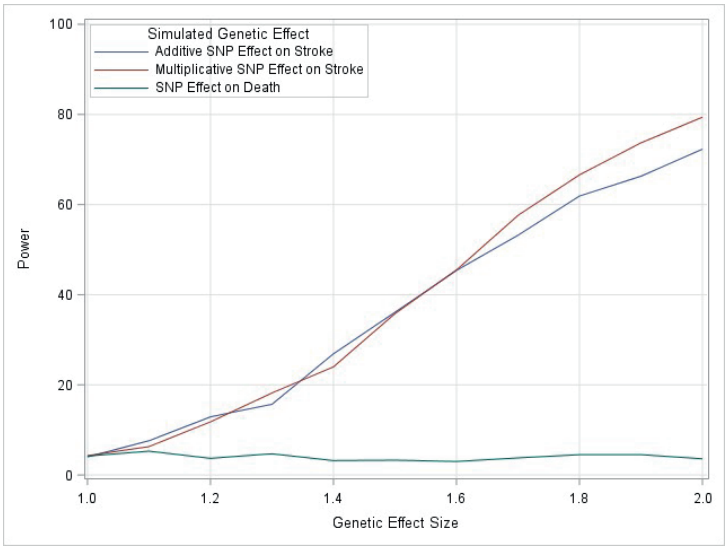


Fig 4. Estimated power curves for the association between genotype and case control status.

A variant that has an effect on age at death but not on ischemic stroke could lead to a confounded effect on age at onset of ischemic stroke

An association of ApoE rs429358 with stroke AOO has been reported previously [11] and this variant has also been associated with longevity [13] and with age of parental death [12]. These observations, coupled with the association we observed between ApoE rs429358 and stroke AOO , albeit genome-wide significant in women only, prompted us to investigate whether the stroke AOO association could be a manifestation of a more general association of ApoE rs429358 with mortality.

To evaluate this possibility, we performed a simulation study to explore the potential for this type of selection bias. Specifically, we simulate a population of individuals who were followed from birth until death based on age-specific mortality rates obtained from the Social Security Administrations Actuarial Life Tables [14]. Birthdates for the simulated subjects were randomly drawn between 1 January 1900 and 1 January 2020. Each individual was assigned a genotype for three SNPs, $Geno_{StrokeMult}$, $Geno_{StrokeAdd}$ and $Geno_{death}$. $Geno_{StrokeMult}$ and $Geno_{StrokeAdd}$ increase risk of ischemic stroke only, and $Geno_{death}$ increases risk of death only, but not through IS. Stroke was assumed to increase the risk of death as a function of the time since the event. We performed association analyses for each simulated SNP and two phenotypes: logistic regression of case-control status, and linear regression of age at onset. See Methods for simulation details.

As expected, both simulated loci influencing risk of stroke were identified via a case-control design (1000 cases and 1000 controls) (Fig 4). Whether they were associated with AAO depended on the functional form of risk. Loci with a relative increase in risk were not associated with AAO at all, but those with an additive increase in risk saw proportionally more stroke at early ages and thus the risk allele was associated with a lower AAO of IS. The simulated locus that was associated with mortality via mechanisms

unrelated to stroke was not associated with stroke risk. However, that locus was associated with AAO. A locus with a two-fold increase in mortality would display an association with a ~1.5 year decrease in age at onset, an effect size similar to that identified for the APOE locus in our GWAS. This means that it is possible that this association is biased by an association with earlier death, if that association is indeed independent of ischemic stroke.

Discussion

We found that APOE rs429358, encoding the ApoE-E4 haplotype, is associated with earlier ischemic stroke in women. In addition to having been associated with stroke AOO in a previous candidate gene study [11]], this SNP is also associated with increased risk of Alzheimer's disease, numerous cardiovascular traits, as well as with earlier mortality and lower AAO of Alzheimer's disease [15]. APOE rs429358 has not been associated with increased risk of ischemic stroke or any of its subtypes in previous GWAS. The different haplotypes of ApoE, defined by rs429358 and rs7412, have been associated with increased risk of ischemic stroke in a number of studies (while a few other studies did not find an association) [16,17]. In these studies, the different haplotypes were compared with each other, while most GWAS employ an additive model. A non-linear relationship between ApoE haplotype and risk of ischemic stroke could explain the absence of a GWAS association with risk of ischemic stroke.

However, if we assume that the pathway underlying ApoE's association with mortality does not influence ischemic stroke risk, it would also be possible to find an association with age at onset that is confounded by its effect on longevity. If people are more likely to die earlier with a specific genotype, then we are less likely to find that genotype in older people. We investigated this scenario in a simulation, and conclude that this would lead to effect sizes that are comparable to the ones we found in our GWAS. It is therefore possible that the association is confounded by an association with earlier death.

We can however not conclude that the association must be caused by a selection bias, even if it were possible. The most important assumption that we made in our simulation study, is that the variant does not have an effect on ischemic stroke. In reality, this is not known. Rather, we know that Apo-lipoprotein E is involved in lipid metabolism and atherosclerosis [18], rendering it plausible that it could have an effect on risk and/or onset of ischemic stroke. That would mean that genetic variation in ApoE is associated with risk of ischemic stroke, but the association has not been found yet. This could also be due to a sex-specific or sex-differential effect on risk. A sex-stratified GWAS on risk of ischemic stroke could help answer these questions. Sex differences in plasma concentration of ApoE have been described previously. In most studies, women are found to have higher ApoE concentrations. [19]

While we have studied only AAO of IS, the AAO of other complex diseases might also be associated with genetic variation. A recent preprint studied the AAO of a large number of diseases in UK Biobank. [20] Overall, they observed a negative genetic correlation between susceptibility and AAO of a certain disease; higher genetic risk is associated with lower AAO. Concordant with our findings for IS, they find that some AAO variants are also associated with increased risk while others seem to be solely associated with disease onset. A study of 530 complex traits in the UK biobank showed that sexual differences in genetic architecture are widespread. [21] Notably, they show that not stratifying on sex results in missed associations with genetic variants. This was the case in one third of binary traits and almost all continuous traits. Information on age at onset and sex is usually available in large GWAS datasets; stratifiying on sex as well as also analysing AAO should be considered in addition to case-control analysis. However, the choice for age at onset analyses - especially for phenotypes that are generally diagnosed at older age - should be carefully considered given a potential selection bias through associations with earlier death.

Disclosures

JvB is funded through R01NS100178 from the U.S. National Institutes

of Health. SJK, BDM are supported by the U.S. National Institutes of Health, NINDS R01NS100178 and R01NS105150. SJK is additionally supported by the Department of Veterans Affairs RR&D N1699-R and BX004672-01A1. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government. JdR is supported by a Vidi Fellowship (639.072.715) from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO). SWvdL is funded through grants from the Netherlands CardioVascular Research Initiative of the Netherlands Heart Foundation (CVON 2011/B019 and CVON 2017-20: Generating the best evidence-based pharmaceutical targets for atherosclerosis [GENIUS I&II]) and the Interuniversity Cardiology Institute of the Netherlands (ICIN, 09.001). We are thankful for the support of the ERA-CVD program 'druggable-MI-targets' (grant number: 01KL1802) and the Leducq Fondation 'PlaqOmics'. Dr. Sander W. van der Laan has received Roche funding for unrelated work.

Acknowledgements

We would like to thank Sara Pulit for supervision and mentoring during initiation of this study.

Data availability

Genotypes and phenotypes for the cases are available on dbGAP under accession number: phs000615.v1.p1.

Methods

The SiGN dataset

The cohorts included in this study are: ASGC, BASICMAR, BRAINS, EDIN, GASROS, GCNKSS, GOTEBURG, GRAZ, ISGS, KRAKOW, LEUVEN, LUND, MALMO, MCISS, MIAMISR, MUNICH, NOMAS, OXVASC, STGEORGE, SWISS, WUSTL. Human genome build hg19 was used as a reference.

Genome-wide linear regression in BOLT-LMM

When using BOLT-LMM [22], one needs to specify a subset of - genotyped or hard called - SNPs that should be used to build the genetic relationship matrix (GRM). PLINK 1.9 was used to hardcall and subset imputed SNPs. Only imputed SNPs with an uncertainty lower than 0.2 were kept: this means that calls with a highest probability lower than 0.8 were removed. Only genotyped SNPs with a genotyping rate of 95% or higher were kept (--geno 0.05). The minimum minor allele frequency (MAF) for the SNPs that are to be analysed was set to 5%, and the maximum percentage missing per SNP was set to 5%. We included study group as a covariate. Study groups were composed by the authors of [3]similar ancestry and genotyping platforms. After consideration of the QQ and Manhattan plots (Fig S1 and Fig S2) we decided to filter all SNPs on a minimum MAF of 5% and a minimum INFO score of 0.8 for further analysis. We performed this analysis on the following case groups: all cases, all male cases (as determined by XY chromosomes), and all female cases (as determined by XX chromosomes). These groups were further stratified in European and African ancestry samples. The sample sizes for African ancestry were too low to be analyzed.

Validation: meta-analysis and replication in external data

We looked up the SNPs that had a p-value lower than 5e-6 in any of the three analyses (all, XX, XY) in FinnGen, and performed meta-analysis in METAL (inverse variance weighted approach). We used the traditional p-value threshold of 5e-8 to determine if the meta-analysis results were significant. We looked-up the SNP with a p-value lower than 5e-8 in women in the Women's Health Initiative [23]. Baseline characteristics for the replication datasets can be found in supplemental tables S3 and S4.

Conditional analysis

We used GCTA COJO [24] to investigate whether there were additional associated SNPs at the discovered loci. We used the stepwise model selection procedure (--cojo-slct) and used the imputed genotype data (converted to hardcall, same as described for the GWAS) as input (--bfile).

Simulations

Data Generating Model

The data generating model is presented in Figure S3. Pseudo-men and women were simulated drawing a date of birth at random from 1 January 1900 to 1 January 2020. Each pseudo-individual was followed over the course of 120 years or until their death, whichever came first. At birth, genotypes were assigned at three loci each having a minor allele frequency of 10%. Two genotypes, GenoStrokeMult and GenoStrokeAdd incurred a risk on stroke only, and the other, GenoDeath incurred a risk on death via an unspecified pathway independent of stroke. The annual stroke risk was a function of sex, age and genotypes given by:

The genetic effect of the genotype, γ_{effect}, was simulated from 1.0 to 2.0 in increments of 0.1. An initial stroke event was drawn from

$$Annual\ stroke\ risk(p_{stroke} \mid sex,\ age,\ Geno_{StrokeMult}, Geno_{StrokeAdd}) = (0.01/(1\ +\ exp(-0.1 * (age - 60)))) * (0.95 * sex) * (\gamma_{effect} * Geno_{StrokeMult}) + (1 - \gamma_{effect}) * 0.003 * Geno_{StrokeAdd}$$

a Bernoulli distribution with probability given that the subject had not died previously and had not previously experienced a stroke. If the binomial draw indicated a stroke at that age, an exact date of stroke was randomly drawn from a uniform distribution of days in that year. Baseline annual risk of death was taken from the Social Security Administrations Actuarial Life Tables [14]. The mortality effect of GenoDeath was simulated using the same range of parameters, γ_{effect}, and was a function of age. The relative increase in risk was assumed to be close to null at young ages and then increased over the lifetime until a pre-specified risk ratio. Stroke was assumed to increase the risk of death as a function of the time since the event, given by

The resulting annual mortality risk was given by

$$Stroke\ Relative\ Risk(Stroke_{RR}) = 1 + (2/(exp(YearsSinceStroke)/10))$$

The given data generating model resulted in observations with 7 features: date of birth, sex,

$$AnnualMortalityRisk(P_{death}|sex, age, Geno_Death, YearsSinceStroke) = Base\ Risk_{sex,age} * (1 + ((\gamma_{effect} - 1))/(1 + exp(-0.1 * (age - 60))) * Geno_{death}) * Stroke_{RR}$$

$Geno_{StrokeMult}$, $Geno_{StrokeAdd}$, $Geno_{death}$, date of stroke, and date of death. Random draws of pseudo-individuals were made from the data generating model who were (1) alive as of 1 JAN 2020 and (2) over the age of 18 on that date until 1000 cases (defined as having a stroke prior to 1 JAN 2020) and 1000 controls (defined as never having a stroke or having a stroke after 1 Jan 2020) were drawn. Each simulation scenario was replicated 1000 times to make robust estimates of the mean of estimated parameters and standard errors. The simulation study was performed by using SASv9.4.

Genotypic Models

Two genotypic models were simulated. The first modeled a constant relative risk over the lifespan, given by γ_{effect} and parameterized as a risk ratio. The second modeled a constant additive risk over the lifespan given by a function of γ_{effect} as shown above. This model simulated a larger relative effect at younger ages than at older ages. It has been hypothesized that some genetic loci may have a disproportionate effect on stroke risk at younger ages versus older ages, and thus genetic contributors to stroke risk may be easier to find7. For example, when γ_{effect} = 1.1, the early onset locus had a relative risk of 1.6 at age 30, 1.1 at age 50 and 1.04 at age 70. This allows for a test of the ability of age at onset analyses to identify loci that have a larger relative effect early in life rather than later.

Target Parameter

Given the above data generating model, it is trivial to determine the age at stroke for each pseudo-individual (date of stroke – date of birth). The target parameter was defined as the difference in the age of stroke between genotypes among cases.

Estimates of this target parameter were made using linear regression

controlling for sex to approximate a common GWAS strategy.

Genotypes were coded as 0,1,2 to estimate the additive genetic

$$Target\ Parameter(\theta) = E((AAO|reference\ genotype)) - E((AAO|alternative\ genotype))$$

model. Models were run for each of the simulated loci separately.

References

1. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. Circulation. 2017;135: e146–e603.

2. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet. 2018;50: 524–537.

3. Pulit SL, McArdle PF, Wong Q, Malik R, Gwinn K, Achterberg S, et al. Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. Lancet Neurol. 2016;15: 174–184.

4. Traylor M, Mäkelä K-M, Kilarski LL, Holliday EG, Devan WJ, Nalls MA, et al. A novel MMP12 locus is associated with large artery atherosclerotic stroke using a genome-wide age-at-onset informed approach. PLoS Genet. 2014;10: e1004469.

5. Zaitlen N, Lindström S, Pasaniuc B, Cornelis M, Genovese G, Pollack S, et al. Informed Conditioning on Clinical Covariates Increases Power in Case-Control Association Studies. PLoS Genet. 2012;8. doi:10.1371/journal.pgen.1003032

6. Robinson LD, Jewell NP. Some Surprising Results about Covariate Adjustment in Logistic Regression Models. Int Stat Rev. 1991;59: 227–240.

7. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012.

8. Khramtsova EA, Davis LK, Stranger BE. The role of sex in the genomics of human complex traits. Nat Rev Genet. 2019;20: 173–190.

9. Meschia JF, Arnett DK, Ay H, Brown RD Jr, Benavente OR, Cole JW, et al. Stroke Genetics Network (SiGN) study: design and rationale for a genome-wide association study of ischemic stroke subtypes. Stroke. 2013;44: 2694–2702.

10. FinnGen project. In: FinnGen [Internet]. [cited 4 Feb 2022]. Available: https://www.finngen.fi/en/for_researchers

11. Lagging C, Lorentzen E, Stanne TM, Pedersen A, Söderholm M, Cole JW, et al. APOE ?4 is associated with younger age at ischemic stroke onset but not with stroke outcome. Neurology. 2019;93: 849–853.

12. Pilling LC, Kuo C-L, Sicinski K, Tamosauskaite J, Kuchel GA, Harries LW, et al. Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. Aging . 2017;9: 2504–2520.

13. Deelen J, Beekman M, Uh H-W, Helmer Q, Kuningas M, Christiansen L, et al. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. Aging Cell. 2011;10: 686–698.

14. Lew EA. Actuarial contributions to life table analysis. Natl Cancer Inst Monogr. 1985;67: 29–36.

15. Mahley RW. Apolipoprotein E: from cardiovascular disease to neurodegenerative disorders. J Mol Med . 2016;94: 739–746.

16. Wei LK, Au A, Menon S, Griffiths LR, Kooi CW, Irene L, et al. Polymorphisms

of MTHFR, eNOS, ACE, AGT, ApoE, PON1, PDE4D, and Ischemic Stroke: Meta-Analysis. J Stroke Cerebrovasc Dis. 2017;26: 2482–2493.

17. Belloy ME, Napolioni V, Greicius MD. A Quarter Century of APOE and Alzheimer's Disease: Progress to Date and the Path Forward. Neuron. 2019;101: 820–838.

18. 8702059 Polymorphisms related to lipid metabolism predictive of atherosclerosis: Apob, apocii, apoe, apoaiv: Philippe M Frossard, Philippe M Frossard assigned to Biotechnology Research Partners Ltd. Biotechnol Adv. 1987;5: 405.

19. Phillips NR, Havel RJ, Kane JP. Sex-related differences in the concentrations of apolipoprotein E in human blood plasma and plasma lipoproteins. J Lipid Res. 1983;24: 1525–1531.

20. Feng Y-CA, Ge T, Cordioli M, FinnGen, Ganna A, Smoller J, et al. Findings and insights from the genetic investigation of age of first reported occurrence for complex disorders in the UK Biobank and FinnGen. medRxiv. 2020; 2020.11.20.20234302.

21. Bernabeu E, Canela-Xandri O, Rawlik K, Talenti A, Prendergast J, Tenesa A. Sexual differences in genetic architecture in UK Biobank. doi:10.1101/2020.07.20.211813

22. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47: 284–290.

23. Study TWHI, Others. Design of the Women's Health Initiative clinical trial and observational study. Control Clin Trials. 1998;19: 61–109.

24. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44: 369–75, S1–3.

25. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. PLoS Genet. 2015;11: e1005378.

Chapter 3
Identification of pleiotropic SNPs from GWAS summary statistics; a literature review

Michelle ten Dam & Joanna von Berg

Introduction

Genetic variation is assumed to play a role in nearly all human diseases (Lindee, 2000). Genome-Wide Association Studies (GWAS) aim to identify associations between common single nucleotide polymorphisms (SNPs) and a phenotype, which can give insight into the biological mechanisms involved in a phenotype. Thousands of significant SNP associations have been accumulated from single-trait GWAS. Databases like the GWAS catalog [1] can be used to lookup associations of one SNP with multiple traits: a phenome-Wide Association Study (pheWAS). A SNP that is associated with more than one trait is called pleiotropic.

Pleiotropy is widespread in the human genome. An association analysis between millions of SNPs and hundreds of traits found that almost ten percent of SNPs were associated with more than one trait [2]. Pleiotropic SNPs have been identified for many phenotype combinations. In many cases, the traits are known to be biologically related; pleiotropic SNPs have been identified for several psychiatric phenotypes [3] and different types of cancers [4]. However, pleiotropic SNPs have also been described for seemingly unrelated diseases; for prostate cancer and type 2 diabetes [5], schizophrenia and Human Immunodeficiency Virus (HIV) infection [6], and Alzheimer's disease and lung cancer [7]. This could mean that those SNPs are involved in a biological process with a more general function that happens to be involved in both traits. It could also mean that the studied traits are more biologically related than was previously known and might have a common etiology. Identifying more pleiotropic SNPs can thus transform our current classification of diseases [8]. Pleiotropy of druggable genetic targets can help predict adverse treatment effects [8] as well as identify new diseases that could be treated with existing drugs [9]. Pleiotropy can also be leveraged for more accurate risk prediction [10]%. Finally, methods like Mendelian Randomization (MR) rely on the assumption that there is no direct effect of the instrumental variable SNPs on both exposure and outcome. [11]

Pleiotropy methods can be used to indicate whether some of the instrumental variable SNPs are pleiotropic so they can be removed.

The straightforward approach to identify pleiotropic SNPs is to take the intersect of significant SNPs for each trait. This will identify some pleiotropic SNPs but not all, as both GWAS needed to be sufficiently powered to identify this SNP. Additionally, the identified SNPs do not have a measure of pleiotropy indicating how shared they are. Several methods for identification of pleiotropic SNPs from GWAS summary statistics have been published recently. Summary statistics are smaller in size, less privacy-sensitive than individual-level data, and are often publicly available online. We will discuss the four most prominent methods of the last few years: PLACO (Pleiotropic Analysis under COmposite null hypothesis) [5], Primo (Package in R for Integrative Multi-Omics association analysis) [12], PLEIO (Pleiotropic Locus Exploration and Interpretation using Optimal test) [13], and HOPS (HOrizontal Pleiotropy Score) [14]. We show that PLEIO and HOPS do not identify SNPs that have an effect on two or more traits in question, but those that have an effect on one or more traits. Trait-specific SNPs - with an effect on only one of the traits - will thus also be called pleiotropic. While it might be interesting in some cases to identify SNPs with an effect on at least one trait, prospective users of these methods should be aware that they do not identify pleiotropic SNPs. PLACO uses a frequentist approach and tests whether the product of trait-specific effect sizes is different from zero. Primo uses a Bayesian approach, estimating the posterior probability of a SNP being associated with all traits.

Defining pleiotropy

Different types of pleiotropy

Different types of pleiotropy can be identified (see Figure 1). Horizontal (or biological) pleiotropy is when a SNP directly affects multiple traits. Vertical (or mediated) pleiotropy is when a SNP directly affects only one of the traits, but correlation between the traits leads to an association

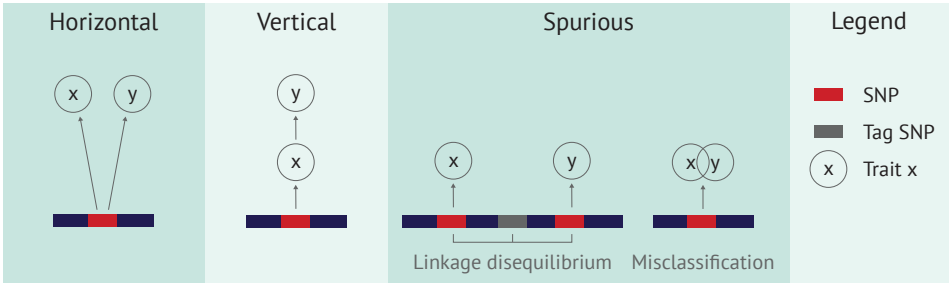


Figure 1. Visualization of horizontal, vertical and spurious pleiotropy, respectively. A horizontally pleiotropic SNP has an effect on all traits under consideration. A vertically pleiotropic SNP has an effect on only one of the traits, but because the traits are correlated it is also associated with the other trait. A SNP can seem pleiotropic because it is in linkage disequilibrium with two SNPs that each individually have an effect on a trait. Misclassification of individuals can also give rise to a seemingly pleiotropic effect.

of the SNP to both traits. There is some disagreement on the boundaries between horizontal and vertical pleiotropy. For example, if a gene produces a product that has two distinct functions and thereby exerts pleiotropy, most would call this horizontal pleiotropy. However, when that gene product has only *one* function and affects two different phenotypes, some would call this vertical pleiotropy [15]\ while others call this horizontal pleiotropy [16]. Finally, spurious pleiotropy can arise from bias in measuring the association between SNPs and traits [17]. For example, one marker SNP can be associated with two or more traits due to that marker being in Linkage Disequilibrium (LD) with another SNP that directly affects one of the traits and yet another SNP that directly affects another trait. The marker SNP seems to be pleiotropic, while in reality neither the marker SNP nor the nearby linked SNPs are pleiotropic. Distinguishing spurious pleiotropy due to LD and real pleiotropy from summary statistics is only possible with fine mapping approaches, which can be applied after identification of pleiotropic SNPs. Another source of spurious pleiotropy is the misclassification of traits. If certain symptoms are shared by two diagnoses, someone with these overlapping symptoms can be given either diagnosis. As a result, the genetic associations for these diagnoses will be highly correlated. Finally, shared controls and ascertainment bias (participant recruitment in a specific disease field) can also cause spurious pleiotropy [16].

Generally, the type of pleiotropy you want to consider will depend on your research question. If you are interested in SNPs that directly affect multiple traits to gain insight in the underlying biological mechanisms, you only want to identify horizontal pleiotropy. However, if you want to use pleiotropy to improve prediction of a trait, distinguishing between horizontal and vertical pleiotropy might not be essential. After all, a SNP increases or decreases risk regardless of the mechanism.

Pleiotropy mechanisms

SNPs may exert a pleiotropic effect through a distinct effect in various tissues, also called adoptive pleiotropy [18]. This could explain how seemingly unrelated traits can be associated with the same SNPs. Pleiotropic proteins have a higher number of protein-protein interactions [19], and pleiotropic genes are often expressed in more tissues. [2] These observations support the omnigenic model, which postulates that a SNP with an effect in a specific tissue affects all diseases that are modulated through that tissue by a small amount. [20]

Pleiotropy versus genetic correlation

Genetic correlation (r2) is defined as the correlation of SNP effect sizes on two traits [17]. Traits can be genetically correlated because they often occur together; if someone is taller, they generally weigh more as well. GWAS of height and weight would result in very similar effect sizes. Non-biological factors like sample overlap between the two GWAS can inflate the r2 estimate. LDSC [21] or HDL [22] can be used to obtain an r2 estimate that is not biased by sample overlap. An important observation is that genetic correlation leads to overall correlation of effect sizes, also in those SNPs with no effect on any of the traits. SNPs that do have an effect on any of the traits can influence r2 estimates; if they are very pleiotropic they can inflate r2, and if they are very trait-specific they can deflate r2. Therefore it is generally recommended to only use the subset of SNPs with no effect on any of the traits for r2 estimation. Also note that pleiotropic effects between traits can be present without genetic correlation, as pleiotropy is a SNP-specific metric and genetic correlation is a genome-wide metric. [21]

Overview of the methods

To find pleiotropic SNPs, we could simply count the number of traits a SNP is significantly associated with and correct for multiple testing. However, stringent multiple testing corrections could conceal relevant pleiotropic SNPs [23]. Additionally, this does not correct for genetic correlation and sample overlap, and no pleiotropy statistic is provided using this approach. Therefore, methods have been developed that systematically evaluate pleiotropy in the genome for a selected group of traits, using GWAS summary statistics. Summary statistics are often publicly available as they do not contain privacy sensitive information, in contrast to individual-level data. [24], [25] We will compare the methods on several topics, describing the statistical and practical similarities and differences. Prospective users can use the flowchart in figure 4 to guide the choice of method.

	Alternative hypothesis	Max p	Decorrelation?	LD correction	Statistic
PLACO	Effect on more than one trait	2	Yes	-	product(z)
Primo	Effect on more than one trait	-	No	Conditional analysis	-
HOPS	Effect on one or more than one trait	-	Yes	LD-score regression	Pm = (100/p) * sqrt(sum(z^2))
PLEIO	Effect on one or more than one trait	-	No	-	LRT statistic:

Table 1. Comparison of the methods. p = number of traits.

Hypotheses

PLACO can only be used to analyse two traits. The method tests whether the effect on each trait is significantly different from zero using the product of effect sizes. If the product is not different from zero, this means that the effect size for either or both traits is zero. Only if the product is different from zero is there a significant effect on both traits.

Primo is a Bayesian method that estimates the posterior probability of a SNP coming from a certain association pattern. An association pattern indicates which traits a SNP is associated with. For two traits, the possible association patterns are 00, 01, 10, and 11; indicating

association with neither trait, only trait 2, only trait 1, or both traits. The posterior probability of the association pattern where the SNP associates with all traits can be used for pleiotropy identification (11 in the two-trait example). The posterior probabilities are estimated based on parameters *theta_j* and *pi*, denoting the proportion of non-null SNPs for each trait and the proportion of SNPs following the association pattern of interest. *theta_j's* could be estimated from the proportion of SNPs that are significantly associated with each trait. *Pi* could be estimated from the proportion of SNPs that are significantly associated with both traits (to identify pleiotropic SNPs).

HOPS uses two different metrics: Pm is the square root of the sum of squared effect sizes, normalized for the number of traits. Pn is the number of traits that a SNP is nominally significant for. They use a chi distribution to test Pm and a binomial distribution to test Pn. Since Pm is a normalized version of the Euclidean distance, a large Pm can arise from a moderate effect on both traits in the same or opposite direction, or from a large effect on one trait. These two possibilities can not be distinguished with Pm alone and therefore HOPS is not able to discriminate pleiotropy from trait-specific effects.

PLEIO tests whether each SNP comes from a multivariate normal distribution MVN(0, $\tau_i^2\Omega$) where $\tau_i^2 = 0$ for SNPs with no effect and $\tau_i^2 > 0$ for associated SNPs. PLEIO essentially tests a version of the Euclidean distance as well. The difference with HOPS is that they first decorrelate the data, while PLEIO models the correlation in the hypothesis test. A SNP can be significantly different from a multivariate normal distribution with zero mean without being pleiotropic. Therefore, PLEIO is also not able to discriminate pleiotropy from trait-specific effects.

LD correction

Spurious pleiotropy can be caused by LD between causal SNPs and the marker SNP. To only consider SNPs that are truly pleiotropic for several traits, GWAS summary statistics analysis results need to be corrected for LD. HOPS uses LD-score regression to correct for the inflation of their statistic due to the total LD of a certain SNP with all

other SNPs (the LD-score). While this does correct for inflation, this does not take the explicit LD into account. To be able to identify which SNPs are truly pleiotropic, fine mapping approaches need to be applied on the results of the methods presented here.

Decorrelation

HOPS and PLACO decorrelate z-scores before calculating their statistics, while Primo and PLEIO explicitly model the estimated covariance matrix in the null distribution. All methods estimate the covariance matrix directly from the GWAS effect size estimates. This covariance matrix constitutes correlation coming from correlated 'real' (not estimated) effect sizes, and correlated error terms. The former is what we usually think of as genetic correlation, the latter arises from sample overlap and cryptic relatedness. Using the covariance matrix without distinguishing these two components has two consequences: 1) vertical pleiotropy will not be observed, and 2) the effect sizes are corrected for (potentially unknown) sample overlap and cryptic relatedness.

Inflation of the test statistic by large trait-specific SNP effects

Both PLACO and HOPS are sensitive to strong effect sizes as they use a product of z-scores and the sum of squared z-scores, respectively. If a SNP is associated with one trait and the other trait is highly polygenic, then that SNP is also likely to be associated with the other trait. HOPS corrects for polygenicity, yielding a polygenicity corrected and score. PLACO is sensitive to strong effect sizes as well, and will interpret a SNP with a strong effect on one trait and a small effect on the second trait, as very pleiotropic. PLACO does not correct for this, but the authors advise to remove SNPs with a large effect on any of the traits before analysis. This unavoidably means that the potentially most interesting SNPs - those with a large effect - are not analysed.

Test statistic and association strength

The choice of method may depend on the type of desired output. HOPS, for example, outputs a P_n and P_m score for every SNP.

For PLACO, a significant p_{PLACO}-value indicates that the SNP is pleiotropic to the two traits that PLACO analyzed (PLACO can only analyze a maximum of two traits). Primo outputs posterior probabilities for each possible trait association pattern, enabling the user to test multiple null hypotheses. PLEIO returns three different statistics (tau, PLEIO and LS statistic) and one p-value. HOPS and PLACO both return a statistic and a p-value per SNP.

Application to real datasets and validation of the methods

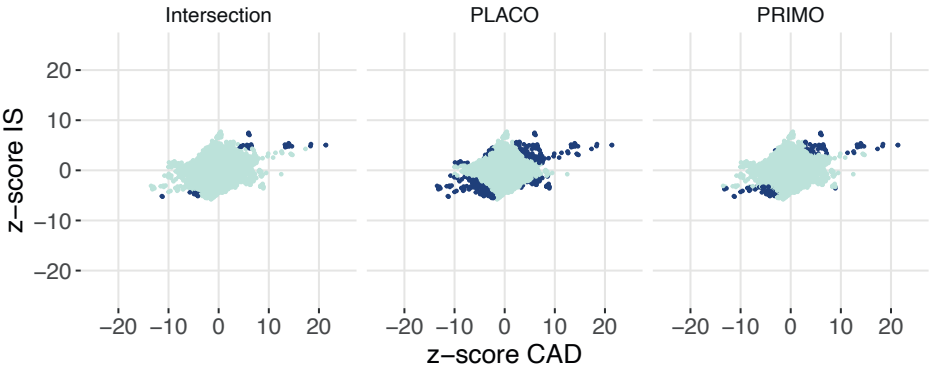


Figure 2. Not all multivariate methods distinguish between shared and trait-specific effects.

Not all methods have been applied to real datasets. We ran the methods on the same data (ischemic stroke and coronary artery disease): intersection, HOPS without polygenicity correction (HOPS-noPC), HOPS with polygenicity correction (HOPS-PC), PLACO, PRIMO, and PLEIO.

In figure 2 we plotted the z-scores for ischemic stroke and coronary artery disease, and colored the SNPs if they are identified as pleiotropic by a certain method (posterior probability > 0.80 for PRIMO, and q-value < 0.05 for the other methods). As expected from the null hypotheses, we see that Intersection, PLACO and PRIMO identify shared SNPs with an effect on both traits, and do not identify trait-specific SNPs as pleiotropic. PLEIO and HOPS

on the other hand identify all SNPs with a significant multivariate effect as pleiotropic, even if they are not shared. This is clearly visible for PLEIO and HOPS-noPC, and less so for HOPS-PC since it seems to identify only shared SNPs. However, the polygenicity correction of HOPS [14] does not take into account whether the larger value is driven by a shared or trait-specific effect.

Figure 3. Number of pleiotropic loci found versus the speed of each method. PLACO and Intersection are all Pareto optimums: no other method is able to identify more pleiotropic SNPs while also being faster.

The three methods that only identify pleiotropic SNPs were further analysed; Intersection, PRIMO and PLACO. We compared the individual SNPs and loci that were identified as pleiotropic by each method. 7 loci were found by all methods. Intersection did not find any additional SNPs. PLACO found 47 loci, PRIMO found 21 loci. 9 of where were identified by both PLACO and PRIMO. We measured the runtime of each method. Intersection was the fastest, with 0.1344 seconds needed for 1e5 SNPs. After that was Primo, with 1.1680 seconds. PLACO was the slowest with 162.7986 seconds (2.71331 minutes). In figure 3 we see the number of pleiotropic loci that were found by each method versus the speed of generating results (in number of input SNPs per second).

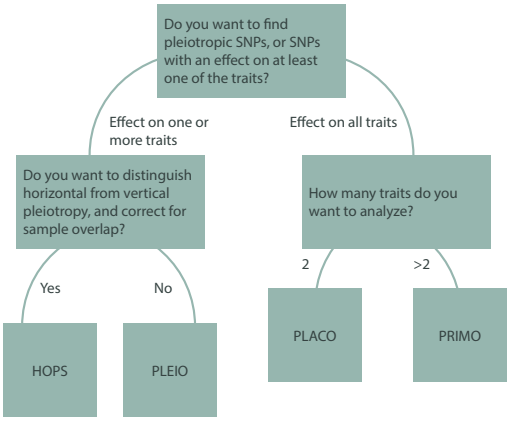


Figure 4. Flow chart

Discussion

The methods reviewed in this paper employ different approaches to identify pleiotropic SNPs from GWAS summary statistics. Only PLACO and Primo test for an effect on two or more traits; PLEIO and HOPS test for an effect on one or more traits, allowing trait-specific SNPs to be identified as pleiotropic as well. This approach can be interesting in certain studies, for instance as an alternative to meta-analysis of GWAS of the same phenotype in different samples. Nevertheless, SNPs that have an effect on at least one of the traits are not pleiotropic according to the definition that we have used. Of course, as long as both method developers and users clearly define what their goal is, terminology is not as important. We do want to warn potential users of these methods, especially if they want to use them for Mendelian randomization (MR). One of the assumptions for the validity of genetic variants used as instrumental variables in MR is that they are not horizontally pleiotropic for the exposure and outcome traits. If PLEIO or HOPS are used to answer this question, all SNPs that have an association with either exposure or outcome will be flagged as pleiotropic and thus invalid, leaving no SNPs for the MR analysis. PLACO identified the most pleiotropic SNPs, at the expense of a longer runtime than PRIMO. However, PLACO can not

be used to find SNPs with a shared effect on more than two traits. If the desired number of traits is higher than two, PRIMO is currently the only method that truly identifies shared SNPs. Of the pleiotropy methods only PLACO corrects for correlation between traits, resulting in SNPs that are horizontally and not vertically pleiotropic. PRIMO did not implement a decorrelation procedure, but explicitly models the correlation in its statistical tests. Note that the user can choose to decorrelate summary statistics themselves before applying any method.

In single-trait GWAS population stratification arises from over-representation of one ancestry in cases or controls, which can lead to spurious association of SNPs that have different allele frequencies in different ancestries. This cannot be corrected after the GWAS is performed, and should therefore be checked before including GWAS for pleiotropy analysis. Population stratification can also emerge when two GWASs performed on two different ancestries are used in pleiotropy analysis. SNP effects can differ between ancestries. Pleiotropy analysis of GWAS of different traits performed in different ancestries might therefore miss significant associations or overestimate them. This could lead to unexpected effects in applied medicine. Additionally, pleiotropy analysis of different traits analysed in different ancestries raises the question: to whom do the results apply? To only one ancestry, both, or none? Therefore, as pleiotropy analyses cannot (yet) correct for population stratification, it is important to combine GWAS from comparable ancestries.

None of the methods discussed in this paper account for epistasis; the phenomenon where the effect of SNP on a trait depends on the genetic background of the individual due to interaction between genetic elements [26]. GWASs that have tested interaction effects for multiple phenotypes could be used as input for the methods described here to identify SNP interactions that have a shared effect on multiple traits. It has not been shown whether it is possible to determine interaction effects from summary statistics, without individual-level data. Assessing the feasibility of such an approach could be an interesting direction for future research.

References

[1] A. Buniello et al., "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," Nucleic Acids Res., vol. 47, no. D1, pp. D1005–D1012, Jan. 2019.

[2] K. Watanabe et al., "A global overview of pleiotropy and genetic architecture in complex traits," Nat. Genet., Aug. 2019.

[3] T. Otowa et al., "Meta-analysis of genome-wide association studies of anxiety disorders," Mol. Psychiatry, vol. 21, no. 10, p. 1485, Oct. 2016.

[4] R. E. Graff et al., "Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts," Nat. Commun., vol. 12, no. 1, p. 970, Feb. 2021.

[5] D. Ray and N. Chatterjee, "A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between Type 2 Diabetes and Prostate Cancer," PLoS Genet., vol. 16, no. 12, p. e1009218, Dec. 2020.

[6] Q. Wang, R. Polimanti, H. R. Kranzler, L. A. Farrer, H. Zhao, and J. Gelernter, "Genetic factor common to schizophrenia and HIV infection is associated with risky sexual behavior: antagonistic vs. synergistic pleiotropic SNPs enriched for distinctly different biological functions," Hum. Genet., vol. 136, no. 1, pp. 75–83, Jan. 2017.

[7] Y.-C. A. Feng et al., "Investigating the genetic relationship between Alzheimer's disease and cancer using GWAS summary statistics," Hum. Genet., vol. 136, no. 10, pp. 1341–1351, Oct. 2017.

[8] S. Sivakumaran et al., "Abundant pleiotropy in human complex diseases and traits," Am. J. Hum. Genet., vol. 89, no. 5, pp. 607–618, Nov. 2011.

[9] T. A. O'Mara, J. Batra, and D. Glubb, "Editorial: Establishing genetic pleiotropy to identify common pharmacological agents for common diseases," Front. Pharmacol., vol. 10, p. 1038, Sep. 2019.

[10] R. Maier et al., "Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder," Am. J. Hum. Genet., vol. 96, no. 2, pp. 283–294, Feb. 2015.

[11] G. Hemani, P. Haycock, J. Zheng, T. Gaunt, and B. Elsworth, "TwoSampleMR: Two Sample MR functions and interface to MR Base database," R package version 030, 2018.

[12] K. J. Gleason, F. Yang, B. L. Pierce, X. He, and L. S. Chen, "Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits," Genome Biol., vol. 21, no. 1, p. 236, Sep. 2020.

[13] C. H. Lee, H. Shi, B. Pasaniuc, E. Eskin, and B. Han, "PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics," Am. J. Hum. Genet., vol. 108, no. 1, pp. 36–48, Jan. 2021.

[14] D. M. Jordan, M. Verbanck, and R. Do, "HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases," Genome Biol., vol. 20, no. 1, p. 222, Oct. 2019.

[15] G. P. Wagner and J. Zhang, "The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms," Nat. Rev. Genet., vol. 12, no. 3, pp. 204–213, Mar. 2011.

[16] N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller, "Pleiotropy in complex traits: challenges and strategies," Nat. Rev. Genet., vol. 14, no. 7, pp. 483–495, Jul. 2013.

[17] W. van Rheenen, W. J. Peyrot, A. J. Schork, S. H. Lee, and N. R. Wray, "Genetic correlations of polygenic disease traits: from theory to practice," Nat. Rev. Genet., vol. 20, no. 10, pp. 567–581, Oct. 2019.

[18] J. Hodgkin, "Seven types of pleiotropy," Int. J. Dev. Biol., vol. 42, no. 3, pp. 501–505, 1998.

[19] S. Ittisoponpisan, E. Alhuzimi, M. J. E. Sternberg, and A. David, "Landscape of pleiotropic proteins causing human disease: Structural and system biology insights," Hum. Mutat., vol. 38, no. 3, pp. 289–296, Mar. 2017.

[20] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An Expanded View of Complex Traits: From Polygenic to Omnigenic," Cell, vol. 169, no. 7. pp. 1177–1186, 2017.

[21] B. K. Bulik-Sullivan et al., "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies," Nat. Genet., vol. 47, no. 3, pp. 291–295, Mar. 2015.

[22] Z. Ning, Y. Pawitan, and X. Shen, "High-definition likelihood inference of genetic correlations across human complex traits," Nat. Genet., pp. 1–6, Jun. 2020.

[23] S. Hackinger and E. Zeggini, "Statistical methods to detect pleiotropy in human complex traits," Open Biol., vol. 7, no. 11, p. 170125, Nov. 2017.

[24] D. N. Paltoo et al., "Data use under the NIH GWAS data sharing policy and future directions," Nat. Genet., vol. 46, no. 9, pp. 934–938, Sep. 2014.

[25] C. Uhlerop, A. Slavković, and S. E. Fienberg, "Privacy-preserving data sharing for genome-wide association studies," J. Priv. Confid., vol. 5, no. 1, pp. 137–166, 2013.

[26] J. Domingo, P. Baeza-Centurion, and B. Lehner, "The causes and consequences of genetic interactions (epistasis)," Annu. Rev. Genomics Hum. Genet., vol. 20, no. 1, pp. 433–460, Aug. 2019.

Chapter 4
PolarMorphism enables discovery of shared genetic variants across multiple traits from GWAS summary statistics

Joanna von Berg[1,2], Michelle ten Dam[1], Sander W. van der Laan[3], Jeroen de Ridder[1,2]

1. Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands 2. Oncode Institute, Utrecht, The Netherlands 3. Central Diagnostics Laboratory, Division Laboratories, Pharmacy, and Biomedical genetics, University Medical Center Utrecht, Utrecht, the Netherlands

This chapter is based on a manuscript that has been accepted in Bioinformatics. Scan this QR code to go to the preprint, and to access the supplemental information:

DOI: 10.1101/2022.01.14.476302

Introduction

Genetic variation in the genome partly explains phenotypic differences between individuals. Genome-wide association studies (GWAS) aim to identify the specific genetic variants (usually single nucleotide polymorphisms, SNPs) that are associated with phenotypic variation. Over the past decades, GWAS have led to the discovery of thousands of SNP-trait associations [1], [2].

From these discoveries we know that some SNPs can influence multiple traits; i.e. they are pleiotropic [3]. Pleiotropy is widespread in the human genome. An association analysis between millions of SNPs and hundreds of traits found that almost ten percent of SNPs were associated with more than one trait [4]. Moreover, pleiotropic SNPs have been identified for many trait combinations. In many cases, the traits are known to be biologically related; pleiotropic SNPs have been identified for several psychiatric phenotypes [5] and different types of cancers [6]. However, pleiotropic SNPs have also been described for seemingly unrelated diseases; for instance for prostate cancer and type 2 diabetes [7], schizophrenia and Human Immunodeficiency Virus (HIV) infection [8], and Alzheimer's disease and lung cancer [9]. This could mean that those SNPs are involved in a biological process with a more general function. It could also mean that the studied traits are more biologically related than was previously known and might have a common etiology. Identifying more pleiotropic SNPs can thus transform our current classification of diseases [10].

Pleiotropy analysis can also be useful to identify pleiotropic SNPs in druggable genetic targets, which can help predict adverse treatment effects [10] as well as identify diseases that could be treated with existing drugs [11]. Moreover, pleiotropy can be leveraged for more accurate risk prediction [12]I. Finally, methods like Mendelian Randomization (MR) rely on the assumption that there is no direct effect of the SNPs used on both exposure and outcome [13]. Since pleiotropy methods can be used to indicate whether some SNPs are pleiotropic, they can be used to filter these SNPs.

It should be noted that analysis of similarity between traits can also be done using genetic correlation, but this answers a different question. Genetic correlation gives the overall - genome-wide - correlation of effect sizes. Pleiotropic SNPs have a shared effect regardless of the genetic correlation and may tag a specific biological pathway or process rather than describing a general relationship between two traits. If traits are correlated and often co-occur in individuals, then any SNP that affects trait X will also be associated with trait Y, even if it does not directly affect trait Y. These SNPs are not actually pleiotropic because they are only directly associated with one trait. For this reason, to identify pleiotropic SNPs it is not sufficient to take the intersection of SNPs that are associated with both traits. Even if the traits are uncorrelated, intersecting SNP-sets is not an optimal approach; both GWASs need to be sufficiently powered to discover the pleiotropic SNP. Moreover, SNPs that are found with this approach lack an important feature: we know that they are shared but we do not know how shared they are and if this might be statistically significant.

Recently, a few methods that aim to identify pleiotropic SNPs have been described. HOPS [14] and PLEIO [15] both identify a SNP as shared if it is associated with at least one of the traits of interest. Problematically, SNPs with an effect on only one trait will thus also be identified and cannot readily be differentiated from truly pleiotropic SNPs. Two other methods, PLACO [7] and PRIMO [16], identify a SNP as shared if it is associated with all traits of interest. PLACO can only be used for identification of SNPs that are shared by two traits. Moreover, we will show that PLACO has a high computational burden. PRIMO, on the other hand, only identifies a subset of the pleiotropic SNPs that PLACO finds.

Here, we present PolarMorphism, a new approach to identify pleiotropic SNPs that is more efficient, identifies the same number of pleiotropic SNPs as PLACO, but can be applied to more than two traits. This enables the identification of SNPs that have an effect on numerous traits, and possibly play a role in more general biological processes. PolarMorphism is based on a transformation of the trait-

specific effect sizes x and y to polar coordinates r (radius, the distance from the origin) and θ (theta, the angle with the x-axis). As a result, r is a measure of overall effect and θ a measure of sharedness, which can be used for downstream significance analysis and SNP ranking.

PolarMorphism enables construction of a trait network showing which traits share SNPs. From SNP-specific networks we observe that most SNPs are associated with traits within one trait domain. We find one SNP - rs495828 in the ABO locus - that is associated with traits across 7 trait domains. We show that analysis of more than two traits is more powerful than the intersection of pairwise results of those same traits. We provide PolarMorphism as an R package on Github under the MIT license: https://github.com/UMCUGenetics/PolarMorphism.

Methods

Overview of PolarMorphism

We aim to identify pleiotropic SNPs from GWAS summary statistics using an approach that can be routinely applied to combinations of two or more traits. After obtaining summary statistics with effect size beta and standard error SE, we calculate z-scores (beta / SE) per SNP. PolarMorphism can be applied on any number of traits, but here we explain the application to two traits. Analyzing more than two traits requires a slightly different approach (see the methods for a full description) but leverages the same principles.

Our aim is to identify horizontally pleiotropic SNPs. Therefore we first perform a decorrelating transform to attenuate vertical pleiotropy resulting from genetic correlation. Given summary statistics for trait x and y, we calculate a covariance matrix, and use this to apply decorrelation or whitening (see methods for details) yielding decorrelated summary statistic vectors $\vec{\hat{x}}$ and $\vec{\hat{y}}$. Next the trait-specific vectors $\vec{\hat{x}}$ and $\vec{\hat{y}}$ are used to calculate polar coordinates r_i (the distance from the origin) and θ_i (the angle with the x-axis, ranging from 0 to 2π). For SNPs that are specific to trait X, θ_i is close to 0 or 2π. For SNPs that are specific to trait Y, θ_i is close to 1/2 π

or 1 1/2 π. For SNPs that are shared, θ_i is approximately 1/4 π for concordant direction of effect and 1 1/4 π for opposite direction of effect. Each quadrant of the x,y plot only differs in direction of effect in the original GWAS. To simplify further analysis we use the fourfold transform of θ (θ_{trans}), which folds the quadrants on top of each other (equivalent to using the absolute values of the z-scores) and then stretches the angles so they still describe a full circle (Figure 1).

To assess significance of sharedness, we separately test the distance r and angle θ_{trans}. Under a null hypothesis of no overall effect, r is the square root of a sum of squared normally distributed variables with mean 0. We thus use a central χ distribution to calculate p-values for r (equivalent to using a χ^2 distribution to test r^2). The alternative hypothesis of this test is that SNP i affects at least one of the traits, which is insufficient to determine pleiotropy. Under a null hypothesis of trait-specific effect, θ_{trans} is equal to 0. To calculate p-values for θ_{trans} we use a von Mises distribution with concentration parameter κ. We show that κ depends on r (see supplemental methods). Estimates of κ from simulations under the null hypothesis are included in the R package. These are used to establish one p-value per SNP. The alternative hypothesis of the second test is that SNP i has a pleiotropic rather than a trait-specific effect.

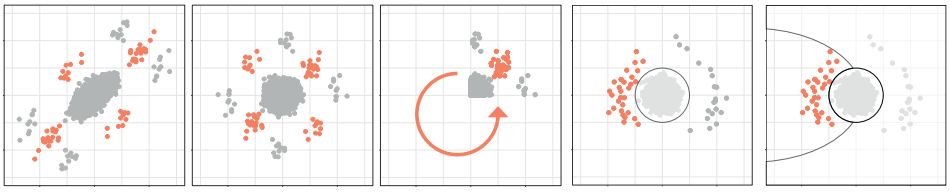


Figure 1. Overview of the method for 2 traits. Orange indicates true pleiotropic SNPs, grey indicates SNPs that are either trait-specific or do not have any effect. Z-scores for each trait are plotted on each axis and the data is decorrelated. Cartesian coordinates are transformed to polar coordinates. The absolute values of the z-scores are calculated and the angle is multiplied by 4. After subsetting on SNPs with a significant distance, we calculate p-values for the angle.

PolarMorphism for two traits

PolarMorphism works on uncorrelated, standardized data. z_x and z_y are vectors of length m containing the z-scores of SNPs 1 to m for trait x and trait y, respectively. We calculate polar coordinates r and θ per SNP i: r is the distance from the origin, and θ is the angle of the vector from the origin to the point ($z_{x,i}$, $z_{y,i}$).
$r_i = \sqrt{(z_{x,i})^2 + (z_{y,i})^2}$ and $\theta_i = tan^{-1}(z_{y,i}/z_{x,i})$

We first test whether r comes from a central chi distribution with degrees of freedom equal to the number of traits p. The chi distribution describes the distribution of the square root of the sum of squared normally distributed variables. The distribution of p-values from this test is used to calculate q-values, which are FDR-corrected p-values [17]. For all SNPs that have an effect, we want to know whether that effect is shared. We perform a four-fold transform of θ that 'folds' all quadrants of the Cartesian plane on top of each other and stretches it to make sure the angles can take any value on the circle [18]: $\theta_{trans} = 4 * \theta \mod 2\pi$. The von Mises distribution describes angular data. It takes into account that θ = 0 is equal to θ = 2π. It has two parameters: θ_{mu} is the mean value, and kappa (κ) is a concentration parameter that is similar to the inverse of the variance. θ_{mu} is zero under the null hypothesis of trait-specific effect. See the Supplementary methods for a description of how we obtained estimates for κ. Using the distribution of the observed r p-values for the distances of all SNPs, and the fact that p-values follow a uniform distribution under the null hypothesis, the false discovery rate (FDR) for each SNP can be calculated. This q-value gives the FDR if this SNP and all SNPs with a lower p-value would be called significant. We keep the SNPs that show a significant overall effect (r q-value < 0.05) and use the distribution of observed θ p-values for these SNPs to calculate θ q-values. We filter on θ q-value < 0.05 to obtain SNPs that are significantly shared (FDR < 0.05).

PolarMorphism for more than two traits

The distance of a SNP i in more than two dimensions is a straightforward extension of the distance in two dimensions:

$$r_i = \sqrt{\sum_{j=1}^{p} z_{i,j}^2}$$

Where $z_{i,j}$ is the z-score of SNP i for trait j. Describing the orientation of a SNP for p traits involves calculating the corresponding p-dimensional hyperspherical coordinates. This gives an additional angle for each added trait. Fortunately, this problem can be simplified. We define $\vec{X}_i$ as the vector from the origin of the p-dimensional sphere to an observed SNP, and $\vec{\mu}_i$ as the vector from the origin to the expected position of the SNP under the null hypothesis of trait-specific effect, along one of the axes. The goal is to determine the angular difference between $\vec{X}_i$ and $\vec{\mu}_i$. We choose $\vec{\mu}_i$ such that it lies along the axis that is closest to $\vec{X}_i$. In other words, we construct $\vec{\mu}_i$ as a vector with zeros for each coordinate except for the coordinate with the highest absolute value for the SNP under consideration. We set the length of $\vec{\mu}_i$ equal to the length of $\vec{X}_i$ (the distance r), so the only non-zero value of $\vec{\mu}_i$ is set to r. The two vectors of interest always lie in a 2-dimensional plane, regardless of the number of traits p. The dot product of the vectors is a scalar and is equal to:

$$\vec{\mu} \cdot \vec{X}_i = r_\mu r_x cos(\theta)$$

, therefore

$$\theta = cos^{-1}(\vec{\mu} \cdot \vec{X}_i / r^2)$$

Which can be rewritten as

$$\theta = cos^{-1}(\sum_{j=1}^{p} \mu_j x_j / \sum_{j=1}^{p} x_j^2)$$

This angle should be normalized so the maximum value is always π, regardless of p. The angle is maximal if all coordinates of a SNP have the same value (which we will call x). Recall that $\vec{\mu_i}$ has zeros for all coordinates but one. If θ is maximal, we can rewrite the expression for θ as:

$$\theta(p) = cos^{-1} \sum_{j=1}^{p} \mu_j x / \sum_{j=1}^{p} x_j^2$$

$$= cos^{-1}(((p-1)(0 \cdot x) + r \cdot x)/px^2) = cos^{-1}(\sqrt{p}/p)$$

The final correction factor with which the angles should be multiplied can then be obtained by dividing 2π by the result of this formula.

To test the significance of r, we use the same procedure as for two traits. In this case the degrees of freedom is equal to the number of traits p. To assign significance levels to the angle θ, we use the von Mises-Fisher distribution, which is an extension of the von Mises distribution. The probability density function of the von Mises Fisher distribution is given by:

Where C is a normalization constant, κ is the concentration

$$f = C \cdot exp(\kappa \vec{\mu} \cdot \vec{X})$$

parameter, $\vec{\mu_i}$ is the unit vector of the expected direction and $\vec{X_i}$ is the observed unit vector (i.e. the vector of the SNP divided by its length to get unit length). The inner product can be rewritten as where θ is the angle between the expected and observed vectors:

$$f = C \ exp(\kappa \ cos(\theta))$$

Functions to obtain the probability density function and the normalization constant C are implemented in the vMF package in R [19]. To obtain a cumulative density function the probability density function needs to be integrated. The definite integral for can not be defined using elementary functions. However, the exponent has the following series representation:

$$f = C\ exp(\kappa\ cos(\theta)) = C\ \sum_{j=0}^{\infty}((\kappa\ cos(\theta))^j/j!)$$

The integral is then equal to:

$$F = C\ \int\sum_{j=0}^{\infty}((\kappa\ cos(\theta))^j/j!) = C\ \sum_{j=0}^{\infty}\int((\kappa\ cos(\theta))^j/j!)$$

The term (as a function of the iterator j) does have an indefinite integral:

$$\int((\kappa\ cos(\theta))^j/j!) =$$
$$= -\frac{cot(\theta)\ \ abs(sin(\theta))\ \ (\kappa\ cos(\theta))^j\ \ hypergeo(1/2,(j+1)/2,(j+3)/2,cos^2(\theta))}{gamma(j+2)}$$

where cot is the cotangent function, hypergeo is the hypergeometric function and gamma is the gamma function. We implemented the summation so that it stops when the last added term is smaller than a user-defined value (called 'tol' in our R package). We use the hypergeo package for the hypergeometric function [20].
The values for κ as a function of p that we derived for p = 2 still apply here, because θ still describes a two-dimensional angle.

Simulated data generation

To estimate the false positive rate (FPR) of PolarMorphism we used the R package simplePHENOTYPES [21] to simulate GWAS data for two traits with horizontally pleiotropic SNPs and SNPs that are specific to each of the traits (49317 SNPs for each of the three categories, approximately 10% of the total number of SNPs), a genetic correlation of 0.8, and heritability of 0.6 for each trait. This was repeated 100 times. As input to the package we used genetic data from the HD genotype chip from phase 3 of the 1000 genomes dataset [22]. We included only individuals with non-Finnish European ancestry to keep the linkage disequilibrium (LD) as homogeneous as possible while maintaining a decent sample size (N = 549 individuals). We used bcftools [23] to include these samples and variants with allele frequency higher than 0.05 or lower than 0.95. We further filtered

the variants to include only high-confidence SNPs, using the list of SNPs with pre-computed LD-scores from the LD-score method [24]. The output of simplePHENOTYPES can readily be used as input for BOLT-LMM [25], with which we performed a GWAS of each instance of simulated data. The resulting summary statistics were used as input for PolarMorphism. To determine FPR for the angle , we considered the fraction of ground-truth trait specific SNPs in our simulated data with , as these SNPs would (falsely) be considered pleiotropic in our method.

To estimate the FPR of the distance r, we permuted the phenotypes as pairs. This ensures that the correlation between the traits remains but no association between genotype and phenotype should exist beyond what is expected under the null hypothesis of no effect. Each of the 100 instances of simulated data was permuted once. We again performed GWAS in BOLT-LMM and ran PolarMorphism. To determine FPR for the significance threshold on r we determine the fraction of all SNPs with , as these SNPs would (falsely) be considered SNPs with a – pleiotropic or trait-specific – effect.

The mean estimated on the non-permuted data is 0.060 (SD = 0.001). On the permuted data, the mean estimated is 0.050 (SD = 0.0003) and the mean estimated is 0.060 (SD = 0.001). Boxplots of the distribution of both FPRs can be found in Figure S1.

Preprocessing the summary statistics

We used publicly available summary statistics for the 41 traits shown in Table 1, encompassing a range of mostly cardiovascular phenotypes with relatively large sample sizes enabling biological interpretation of pleiotropic SNPs within a specific disease context. Data were obtained from the sources provided in Supplemental Table 2, which also contains references to the respective papers they were described in. We aligned reference and alternative allele across all traits, and filtered using the list of high-confidence SNPs provided with the LDSC software.[24] We divide effect sizes by their standard error to obtain z-scores. We calculate the covariance matrix on the subset of SNPs that do not have a large overall effect. To this end,

the covariance is calculated only on SNPs that have a mahalanobis distance smaller than 5. We use the ZCA-cor whitening method in the 'whitening' package in R [26], to decorrelate the data while ensuring that the x and y components of the transformed z-scores maximally correlate with the x and y components of the original z-scores.

Inferring relationships between traits from pleiotropic SNPs

For all trait pairs, we ran PolarMorphism and clumped the significant SNPs with Plink, using the q-values instead of p-values (--clump-kb 5000000, --clump-p1 0.05, --clump-p2 0.05, --clump-r2 0.2) [27]. We make an adjacency matrix from the number of shared loci per trait combination and use this to construct a graph using the igraph package in R [28]. We did the same per SNP to obtain SNP-specific networks. To create domain networks from the trait networks we draw an edge between domain A and B if an edge exists between any trait of domain A and any trait of domain B.

Gene set enrichment analysis in DEPICT

We changed the following settings from the default: association_pvalue_cutoff: 0.05 to accommodate for the fact that we use q-values instead of p-values. We performed gene set enrichment using the default gene sets provided by the DEPICT authors, but only considered gene sets from gene ontology [29], REACTOME [30], KEGG[31] and the PPI networks as defined by the DEPICT authors using the InWeb database [32] for further analysis.

Inferring relationships between traits from genetic correlation

To infer relationships between traits from genetic correlation, we ran LDSC[24] using the GenomicSEM [33] package in R. We calculated p-values from the correlation coefficients and their standard errors using the pnorm function in R, and used a Bonferroni corrected p-value threshold of 6.4*10-5 to correct for 780 trait combinations tested. For this purpose, we made an adjacency matrix from the genetic correlation for each trait combination and used this to make a graph using the igraph package in R.[28]

Comparison with other methods

Intersection refers to the straight-forward approach of finding shared SNPs: take the intersection of the SNPs that were significant for trait X and those that were significant for trait Y. We used the R package for HOPS (HOrizontal Pleiotropy Score) [14] We used our pre-processed z-scores (whitened). We ran HOPS both with and without polygenicity correction and used only the Pm p-values. We used the command line tool written in Python for PLEIO (Pleiotropic Locus Exploration and Interpretation using Optimal test) [15]. We used z-scores (not whitened and not corrected for LD-score) and supplied the sample sizes of the original GWAS. We used the R package for PRIMO (Package in R for Integrative Multi-Omics association analysis) [16]. We used PRIMO based on p-values. For the alt_props parameter (the expected proportion of SNPs that follow the alternative hypothesis per trait) we supplied the proportion of SNPs that were significant for trait 1 (q-value < 0.05) over all SNPs, idem for trait 2 (q-value < 0.05). We supplied c(2,2) for the dfs parameter. We used the R package for PLEIO (pleiotropic analysis under composite null hypothesis) [15]. We used whitened z-scores (not corrected for LD-score). We used the VarZ function to calculate the covariance matrix and supplied that, with the z-scores, to the placo function.

To assess how many loci were found by each method, we LD-pruned the significantly shared SNPs. For each method and for each locus, we checked if any of the SNPs in that locus were also found by another method. If that was the case, we gave that locus the same identifier in each method. Afterwards, we determined the loci that were found by all methods and those that were found by only one or a subset of the methods. We ran Intersection, HOPS (with polyenicity correction), PRIMO, PLACO, and PolarMorphism on the same data while supplying a dataframe with an increasing number of rows. For the Intersection method we added q-value calculation from the original GWAS p-values and a filtering step on both q-values to make it a fair comparison with the other methods. All five methods are written

in R, therefore we timed them in R using the tictoc package [34]. Running the software in the terminal could have a different runtime, but this does allow us to compare the runtimes among the methods.

Results

Defining pleiotropy

Pleiotropy can be identified in different ways ([3], [35] and Figure 2). Horizontal pleiotropic SNPs directly affect multiple traits. Vertical (or mediated) pleiotropic SNPs directly affect one of the traits, but dependence between the traits leads to an association with both traits. The difference between horizontal and vertical pleiotropy is particularly important in the context of Mendelian randomization (MR). With MR, the causal effect of an exposure (e.g. smoking) on an outcome (e.g. lung cancer) can be determined. Genetic variants that are associated with the exposure are used as so-called 'instrumental variables'. One important assumption is that these variants only have an effect on the outcome through the exposure. In other words, that they are vertically pleiotropic and not horizontally. Horizontally pleiotropic SNPs - which have a direct effect on both smoking and lung cancer - violate this assumption and should therefore not be used as instrumental variables in MR [36]. The final pleiotropy type is spurious

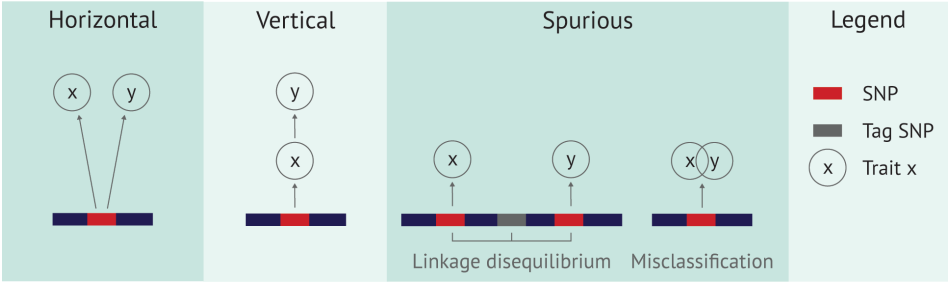


Figure 2. Visualization of horizontal, vertical and spurious pleiotropy, respectively. A horizontally pleiotropic SNP has an effect on all traits under consideration. A vertically pleiotropic SNP has an effect on only one of the traits, but because the traits are correlated it is also associated with the other trait. A SNP can seem pleiotropic because it is in linkage disequilibrium with two SNPs that each individually have an effect on a trait. Misclassification of individuals can also give rise to a seemingly pleiotropic effect.

pleiotropy, which can arise from bias in measuring association [37]. For example, one marker SNP can be associated with two or more traits due to that marker being in linkage disequilibrium (LD) with another SNP that directly affects one of the traits and yet another SNP that directly affects another trait. The marker SNP seems to be pleiotropic, while in reality neither the marker SNP nor the nearby linked SNPs are pleiotropic. Determining whether the same SNP is likely causal for both traits is only possible with colocalization approaches [38]. Another source of spurious pleiotropy is misclassification of traits. If certain symptoms are shared by two diagnoses, individuals with these overlapping symptoms can be given either diagnosis. As a result, the genetic associations for these diagnoses will be highly correlated. Finally, shared controls and ascertainment bias (participant recruitment in a specific disease field) can also cause spurious pleiotropy [39].

Inferring relationships between traits from pleiotropic SNPs

We applied PolarMorphism to all pairwise combinations of 41 traits from different trait domains (Table 1). The resulting pleiotropy network is shown in Figure 3. Herein, traits are nodes and the edge weights indicate the number of pleiotropic SNPs discovered by PolarMorphism. The contribution of each SNP to the edge weights is weighted by the inverse of the total number of traits it is associated with, in order to account for the effect that SNPs affecting many traits probably tag a biological process with a general function. Sharing such a SNP is less meaningful than sharing a SNP with an effect on only some traits.

The resulting pleiotropy network is densely connected (512 out of 820 possible edges), supporting earlier descriptions of widely occurring pleiotropy among traits [4], [39]W The lipid domain (HDL, LDL, TG and TC) and blood pressure domain (DPB, SBP and PP) each form a fully connected subgraph. SBP has the highest number of edges (degree), sharing SNPs with 37 of the 41 traits. ALS, which shares SNPs with 5 traits, has the lowest degree. Global analysis of the pleiotropy network thus readily reveals general characteristics of traits and trait domains.

Table 1 (next Page) . Trait domains and trait abbreviation as used in the figures.

Domain name	trait abbreviation	trait name
anthropomorphic	BMI	body mass index
	Height	height
cancers	PrCa	prostate cancer
	BC	breast cancer
cardiac traits	AF	atrial fibrillation
	HF	heart failure
	NICM	non-ischemic cardiomyopathy
cardiovascular	CAC	coronary artery calcification
	CAD	coronary artery disease
	cIMT	carotid intima-media thickness
	Plaque	presence of carotid plaque
immune	IBD	irritable bowel disease
	Asthma	asthma
lipids	HDL	high-density lipoprotein
	LDL	low-density lipoprotein
	TC	triglycerides
	TG	total cholesterol
neurodegenerative disease	AD	Alzheimer's disease
	ALS	Amyotrophic lateral sclerosis
	PD	Parkinson's disease
pressures	DBP	Diastolic blood pressure
	SBP	Systolic blood pressure
	PP	Pulse pressure
psychiatric / psychological	ASD	Autism spectrum disorder
	BIP	bipolar disorder
	DS	depressive symptoms
	EA	educational attainment
	IQ	intelligence quotient
	MDD	major depressive disorder
	Neuroticism	neuroticism
	SWB	subjective well being
	Insomnia	insomnia
smoking	EvrSmk	ever smoker
	FrmrSmk	former smoker
	logOnset	log of age at onset of smoking
	CpD	cigarettes per day
stroke	AS	any stroke (hemorrhagic or ischemic)
	IS	ischemic stroke
	CES	cardio-embolic stroke
	LAS	large artery stroke
	SVS	small vessel stroke

Analyzing the pleiotropy network in more detail, we find that most SNPs are associated with traits within one or across two trait domains (51% and 43%, respectively). We observe one SNP that is associated with traits across 7 trait domains: rs495828, a SNP in the ABO gene, which is ubiquitously expressed across many tissues and cell types [40]. For each trait domain, we determine how many SNPs only have associations within that domain (we call these single domain SNPs), and calculate the percentage of the total number of SNPs that were identified for that domain. We find that the psychiatric traits have the highest percentage of single domain SNPs; one third of all SNPs that are shared with a psychiatric trait are only associated with psychiatric traits. The smoking traits have the lowest percentage of single domain SNPs, suggesting that most smoking-associated variants tag general biological processes rather than smoking-specific processes.

A comparison with genetic correlation

Genetic correlation (r_g) is the correlation of SNP effect sizes on two traits [37]. Non-biological factors like sample overlap between the two GWAS can inflate the r_g estimate. LDSC [24] or HDL [41] can be used to obtain an r_g estimate that is not biased by sample overlap. Genetic correlation leads to overall correlation of effect sizes, also in those SNPs with no effect on any of the traits. SNPs that do have an effect can influence estimates; if they are very pleiotropic they can inflate r_g, and if they are very trait-specific they can deflate . r_g Therefore it is generally recommended to only use the subset of SNPs with no effect on any of the traits for r_g estimation. Also note that pleiotropic effects between traits can be present without genetic correlation, as pleiotropy is a SNP-specific metric and genetic correlation is a genome-wide metric [24].

To assess whether genetic correlation provides the same insight into trait relationships as pleiotropy, we built a network based on genetic correlation. The resulting network is sparse (138 out of 780 possible edges) and only partially overlaps with the pleiotropy network. Figure 4 shows separate subnetworks for edges that exist in both the genetic correlation network and the pleiotropy network or in only one of the two. In total, 416 trait pairs share at least one pleiotropic SNP, but are

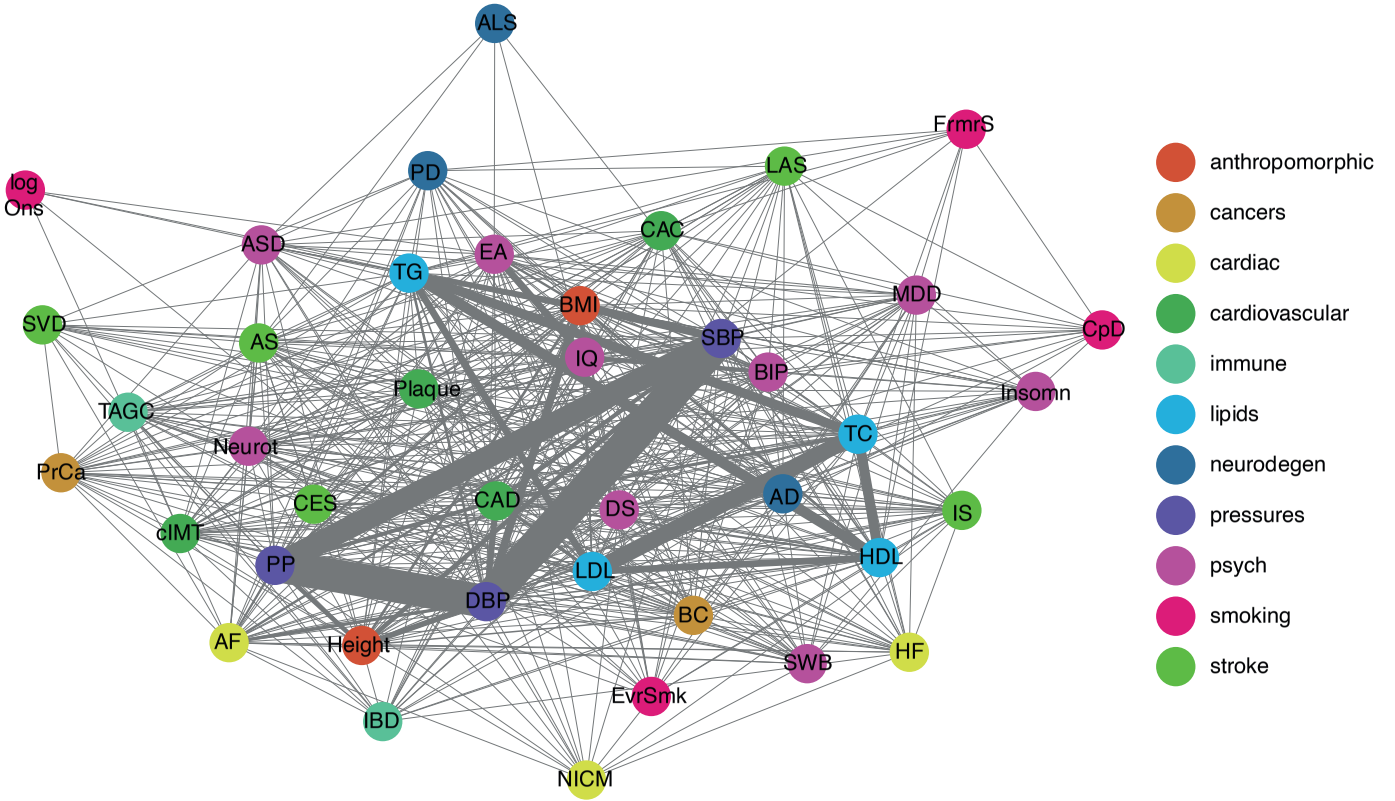


Figure 3. Trait network based on pleiotropic SNPs. Pairwise PolarMorphism results for 41 traits. Pleiotropic SNPs were defined as having an r q-value > 0.05 and a theta q-value > 0.05. Clumping was performed based on theta q-values and linkage disequilibrium. See methods for details. The thickness of the lines (network edges) indicates how many loci are shared between two traits (network nodes). Colored by disease domain.

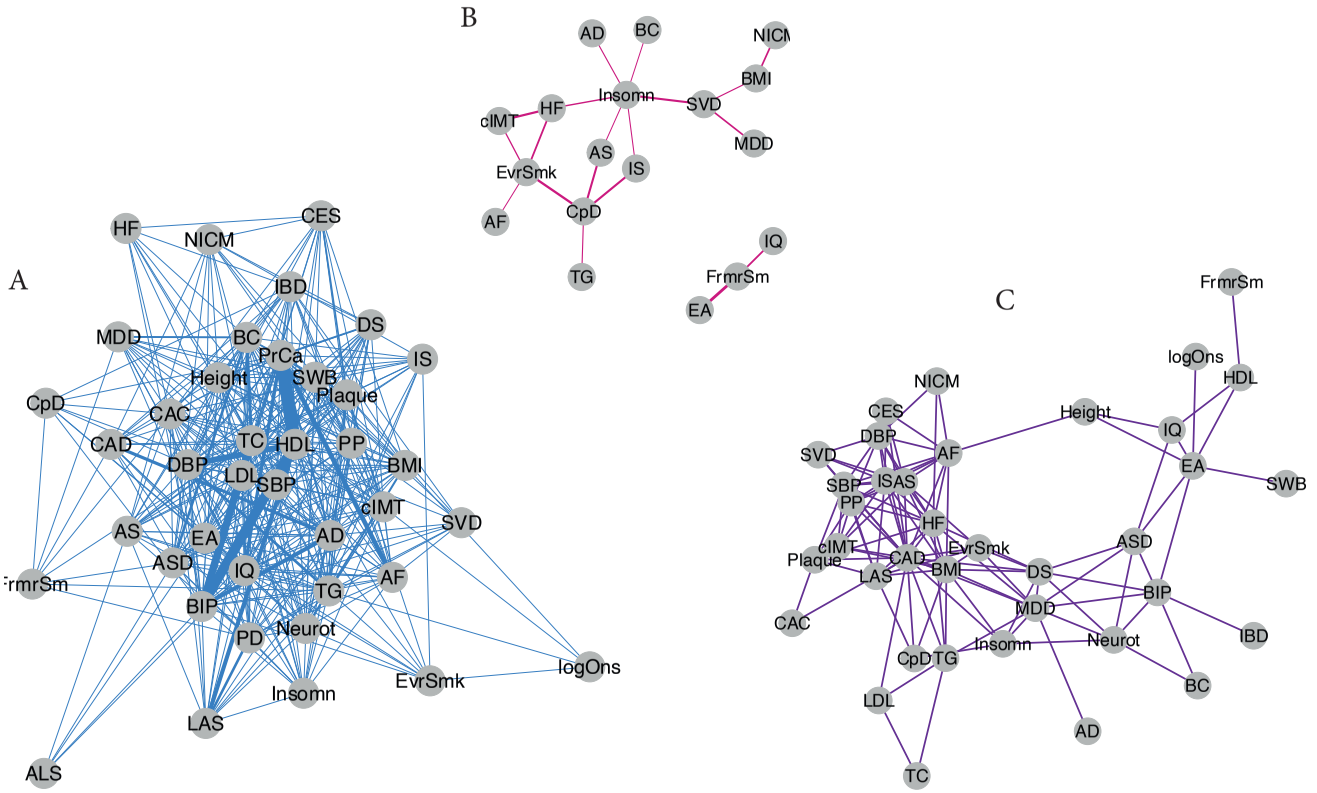


Figure 4. a) Edges denote trait pairs that share pleiotropic SNPs but are not genetically correlated. b) Edges denote trait pairs that are genetically correlated but do not share pleiotropic SNPs. c) Edges denote trait pairs that are genetically correlated and share pleiotropic SNPs.

not genetically correlated (Figure 4a). This situation can arise if there are only a few SNPs that are shared but the rest of the genetic architecture of the traits is independent. It is also possible that some shared SNPs have the same direction of effect in both traits while other shared SNPs have an opposite direction of effect, thereby averaging out . Seven trait pairs are genetically correlated, but do not share any SNPs that are horizontally pleiotropic (Figure 4b). Each SNP that is associated with one of the traits is more likely to also be associated with the other, because of the overall r_g [36]. After decorrelation, vertically pleiotropic SNPs will not be identified by PolarMorphism. 96 trait pairs are genetically correlated and share horizontally pleiotropic SNPs (Figure 4c). These are traits that share a number of vertically pleiotropic SNPs, leading to a higher r_g, as well as some horizontally pleiotropic SNPs. Our results seem to indicate that two traits are more likely to share at least one pleiotropic SNP than they are to be genetically correlated.

The stroke domain

The stroke domain consists of any stroke (AS); its subtype ischemic stroke (IS); and its subtypes cardioembolic stroke (CES), large artery stroke (LAS) and small vessel stroke (SVS). The three IS subtypes are generally believed to have different etiologies [42]–[44], and previous efforts have resulted in tens of subtype-specific associations [45]–[48]. In line with this, our analysis does not reveal any shared SNPs. It should be noted that shared SNPs have been described before for LAS and SVS and for LAS and CES [45]h. However, SNPs at these loci were low-confidence and therefore not included in our analysis (see methods for details).

Given the lack of shared SNPs among the IS subtypes, we investigated which other traits share SNPs with each of the IS subtypes. To that end we looked at the subnetwork composed of the IS subtypes and their direct neighbors (Figure 5). Our analysis reveals that six traits (CAD, DBP, Plaque, PP, SBP, TC) share SNPs with all IS subtypes. This indicates that all ischemic stroke subtypes are associated with biological pathways with a possible effect on blood pressure and lipids. CES shares most pleiotropic SNPs with atrial fibrillation (AF), which is believed to be its main cause [43]. LAS, which is thought to

arise from atherosclerotic plaques in the carotid arteries that rupture or block blood flow [48], shares most SNPs with cIMT - a proxy for the extent of carotid atherosclerosis. SVS, which is thought to have a cardiovascular origin like the other IS subtypes [49], shares most SNPs with CAD. Notably, it also shares many SNPs with Alzheimer's and Parkinson's disease. This might indicate that many of the SNPs that are associated with risk of small vessel stroke also influence risk of neurodegenerative disease. Note that the edges LAS-HDL, SVS-AD, SVS-PD and SVS-Plaque were only found in the pleiotropy network and not in the genetic correlation network. This indicates that pleiotropic SNPs harbor information that is complementary to genome-wide correlation measures. Furthermore, zooming in on one trait domain shows how PolarMorphism can be employed to gain more detailed insight in trait relationships than the general patterns that can be gathered from the complete network.

Joint analysis of more than two traits identifies more pleiotropic SNPs than pairwise analyses of the same traits

PolarMorphism can be used to find SNPs that are shared by any number of traits. A SNP with a small effect on each trait might not be identified in univariate or even pairwise analysis, but could be if more traits are included. We therefore investigated whether analysis of three or more traits is indeed more powerful than the combined results from pairwise analyses of those same traits. Pairwise analyses of the lipid domain (HDL, LDL, TC, TG) identifies 186 shared loci. Analysis of all four traits together identifies 1029 shared loci. 180 loci are found by both approaches.

To explore whether the increased number of loci is biologically relevant, we perform gene set enrichment analysis in DEPICT [50] on the significant loci from the pairwise analyses and the significant loci from the joint analysis. In order to get the relevant genes for each locus, we perform clumping using DEPICT's default settings. Hence the number of DEPICT loci differs from the loci that we identified (108 pairwise loci, 496 joint loci, see Tables S4 and S6). The pairwise results are enriched for 12 gene sets (Table S5) whereas the joint results are enriched for 85 gene sets (Table S7). Moreover,

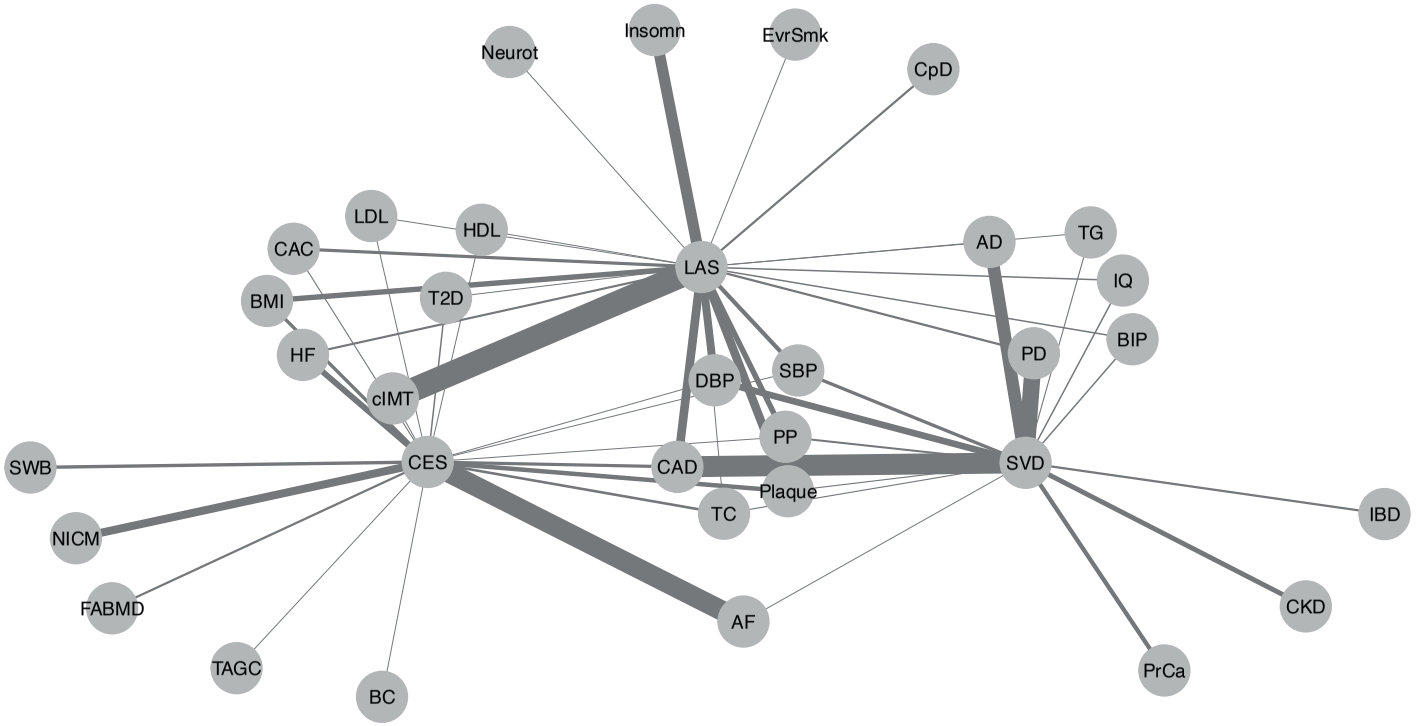


Figure 5. Trait network of the IS subtypes and their direct neighbors, based on the weighted full network as described earlier. Only edges between any of the IS subtypes and any other trait are drawn; in other words, edges between two nodes shown here that do not include an IS subtype, are not drawn.

the loci revealed by the joint analysis result in enrichments that are more significant: 85 of the 95 gene sets that are significant in either analysis are more significant in the joint analysis, and 2 of the 2 gene sets that are significant in both analyses are more significant in the joint analysis. Moreover, considering the 10 genes with the highest z-score for membership of these gene sets, we find that the genes implied by the joint analysis have a higher likelihood of gene set membership (see the DEPICT paper for a detailed explanation [50]), thus resulting in more coherent gene sets. For instance, the joint analysis identifies the LDLR (LDL receptor) gene, which has a high membership likelihood for the REACTOME "metabolism of lipids and lipoproteins" gene set. The pairwise analysis does not identify LDLR, making this gene set less enriched. These results show that joint pleiotropy analysis of multiple traits yields more biologically relevant insights compared to pairwise analysis of those same traits.

Runtime increases marginally with the number of traits analyzed

To assess how the runtime scales with the number of traits analyzed, we picked all traits that were affected by the most pleiotropic SNP, rs495828: AS, BC, CAD, CES, DBP, HDL, HF , IS, LDL, T2D, TAGC, and TC. In this order, we picked the first p traits and timed PolarMorphism (see Figure 6). Runtime increases slightly with larger p, but the effect is small. There is a large difference between p = 2 and p > 2 because we use different approaches if p > 2 (see methods).

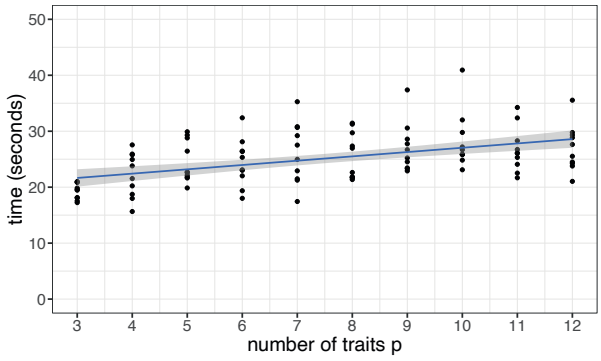


Figure 6. Runtime scales with the number of traits p. The number of traits p ranges from 3 to 12. The slope of the regression line is 0.75 (se = 0.13).

Comparison with other methods

To compare PolarMorphism to existing methods, we ran: PolarMorphism, intersection, PLACO, and PRIMO on a selection of traits (IS and myocardial infarction). We compared the individual SNPs and loci that were identified as pleiotropic by each method. Four loci are found by all methods. Intersection does not identify more than those four loci. PLACO and PolarMorphism both find 21 loci (19 of which are identical), PRIMO finds 13 loci that were also identified by PLACO and PolarMorphism. PLACO and PolarMorphism use a fundamentally different approach to identify pleiotropy: whereas PLACO tests if the effect for both traits is not equal to zero, PolarMorphism first tests whether the overall effect (distance) is different than expected and then tests the sharedness of a SNP.

We timed each method from cleaned input data (already in memory, timing done in R) to results. The number of pleiotropic loci that were found by each method and the speed of generating results (in number of input SNPs per second) are provided in Table 2. These data show that PLACO does not identify more loci than PolarMorphism and is slower.

Table 2. Comparison of methods. HOPS and PLEIO were not run because they use a pleiotropy definition that includes single-trait SNPs. Furthermore, PLACO can only be applied to two traits simultaneously.

	Decorrelation?	Pleiotropic loci found	speed (1k SNPs/second)
PolarMorphism	Yes	21	63
PLACO	Yes	21	0.61
Primo	No	13	86
HOPS	Yes	-	-
PLEIO	No	-	-

Discussion

We have developed a new method that identifies pleiotropic SNPs with an effect on multiple traits. PolarMorphism can be used on combinations of two or more traits. It uses GWAS summary statistics and corrects for correlation in effect sizes arising from genetic

correlation or potential sample overlap. The potential applications of PolarMorphism include a) identifying SNPs that are shared between traits within a trait domain to learn more about the domain-wide biological processes, b) identifying SNPs that are shared among a diverse set of traits to find general biological processes and c) using the identified SNPs to inform new trait ontologies. As an example, we apply PolarMorphism to a set of traits from different domains.

The network analyses indicate that there are no trait domains that only share SNPs within the domain. We observe that most SNPs are associated with traits within one or across two trait domains. We zoomed in on the stroke domain, which has very little domain-specific SNPs. This may mean that the stroke traits are associated with general SNPs or that the stroke traits do not share many biological pathways. Each ischemic stroke subtype shares more SNPs with non-stroke traits than with the other ischemic stroke subtypes. Note that these networks are heavily influenced by the choice of included traits. Conclusions drawn about the networks in this study are therefore not necessarily general, as each trait could share SNPs with a number of traits that were not included. Future applications of PolarMorphism to a diverse set of traits will result in a more complete and precise overview of pleiotropy across the genome and across phenotypes.

We compared PolarMorphism with similar methods. PolarMorphism identifies more pleiotropic SNPs than the standard intersection method and than PRIMO. PLACO identifies the same number of pleiotropic loci as PolarMorphism. However, PolarMorphism finished analysis of 1 million SNPs in less than 20 seconds (compared to >25 minutes for PLACO), making analysis of many trait combinations feasible. Furthermore, PLACO can only be used to analyze two traits together while PolarMorphism can analyze a theoretically unlimited number of traits. A five-fold increase in the number of identified pleiotropic loci for the lipid domain indicates that analyzing more than two traits is much more powerful than combined results from the respective pairwise analyses.

Acknowledgements

The authors thank René Eijkemans for helpful discussions.

Funding

JvB is supported by R01NS100178 from the National Institute of Health. JdR is supported by a Vidi Fellowship (639.072.715) from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO). SWvdL is funded through grants from the Netherlands CardioVascular Research Initiative of the Netherlands Heart Foundation (CVON 2011/B019 and CVON 2017-20: Generating the best evidence-based pharmaceutical targets for atherosclerosis [GENIUS I&II]). We are thankful for the support of the ERA-CVD program 'druggable-MI-targets' (grant number: 01KL1802), the EU H2020 TO_AITION (grant number: 848146), and the Leducq Fondation 'PlaqOmics'.

Conflict of Interest: none declared.

References

[1] P. M. Visscher et al., "10 Years of GWAS Discovery: Biology, Function, and Translation," Am. J. Hum. Genet., vol. 101, no. 1, pp. 5–22, Jul. 2017.

[2] A. Buniello et al., "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," Nucleic Acids Res., vol. 47, no. D1, pp. D1005–D1012, Jan. 2019.

[3] A. B. Paaby and M. V. Rockman, "The many faces of pleiotropy," Trends Genet., vol. 29, no. 2, pp. 66–73, Feb. 2013.

[4] K. Watanabe et al., "A global overview of pleiotropy and genetic architecture in complex traits," Nat. Genet., Aug. 2019.

[5] T. Otowa et al., "Meta-analysis of genome-wide association studies of anxiety disorders," Mol. Psychiatry, vol. 21, no. 10, p. 1485, Oct. 2016.

[6] R. E. Graff et al., "Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts," Nat. Commun., vol. 12, no. 1, p. 970, Feb. 2021.

[7] D. Ray and N. Chatterjee, "A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between Type 2 Diabetes and Prostate Cancer," PLoS Genet., vol. 16, no. 12, p. e1009218, Dec. 2020.

[8] Q. Wang, R. Polimanti, H. R. Kranzler, L. A. Farrer, H. Zhao, and J. Gelernter, "Genetic factor common to schizophrenia and HIV infection is associated with risky sexual behavior: antagonistic vs. synergistic pleiotropic SNPs enriched for distinctly different biological functions," Hum. Genet., vol. 136, no. 1, pp. 75–83, Jan. 2017.

[9] Y.-C. A. Feng et al., "Investigating the genetic relationship between Alzheimer's disease and cancer using GWAS summary

statistics," Hum. Genet., vol. 136, no. 10, pp. 1341–1351, Oct. 2017.

[10] S. Sivakumaran et al., "Abundant pleiotropy in human complex diseases and traits," Am. J. Hum. Genet., vol. 89, no. 5, pp. 607–618, Nov. 2011.

[11] T. A. O'Mara, J. Batra, and D. Glubb, "Editorial: Establishing genetic pleiotropy to identify common pharmacological agents for common diseases," Front. Pharmacol., vol. 10, p. 1038, Sep. 2019.

[12] R. Maier et al., "Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder," Am. J. Hum. Genet., vol. 96, no. 2, pp. 283–294, Feb. 2015.

[13] G. Hemani, P. Haycock, J. Zheng, T. Gaunt, and B. Elsworth, "TwoSampleMR: Two Sample MR functions and interface to MR Base database," R package version 030, 2018.

[14] D. M. Jordan, M. Verbanck, and R. Do, "HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases," Genome Biol., vol. 20, no. 1, p. 222, Oct. 2019.

[15] C. H. Lee, H. Shi, B. Pasaniuc, E. Eskin, and B. Han, "PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics," Am. J. Hum. Genet., vol. 108, no. 1, pp. 36–48, Jan. 2021.

[16] K. J. Gleason, F. Yang, B. L. Pierce, X. He, and L. S. Chen, "Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits," Genome Biol., vol. 21, no. 1, p. 236, Sep. 2020.

[17] J. D. Storey, "The positive false discovery rate: a Bayesian interpretation and the q-value," Ann. Stat., vol. 31, no. 6, pp. 2013–2035, Dec. 2003.

[18] L. Landler, G. D. Ruxton, and E. P. Malkemper, "Circular data in biology: advice for effectively implementing statistical procedures," Behav. Ecol. Sociobiol., vol. 72, no. 8, p. 128, Jul. 2018.

[19] E. A. Houndetoungan, "An R package for fast sampling from von Mises fisher distribution." [Online]. Available: https://nbviewer.jupyter.org/github/ahoundetoungan/vMF/blob/master/doc/vMF.pdf. [Accessed: 20-Oct-2021].

[20] R. K. S. Hankin, "Numerical Evaluation of the Gauss Hypergeometric Function with the hypergeo Package," R J., 2015.

[21] S. B. Fernandes and A. E. Lipka, "simplePHENOTYPES: SIMulation of pleiotropic, linked and epistatic phenotypes," BMC Bioinformatics, vol. 21, no. 1, p. 491, Oct. 2020.

[22] 1000 Genomes Project Consortium et al., "A global reference for human genetic variation," Nature, vol. 526, no. 7571, pp. 68–74, Oct. 2015.

[23] H. Li et al., "The Sequence Alignment/Map format and SAMtools," Bioinformatics, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.

[24] B. K. Bulik-Sullivan et al., "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies," Nature Genetics, vol. 47, no. 3. pp. 291–295, 2015.

[25] P.-R. Loh et al., "Efficient Bayesian mixed-model analysis increases association power in large cohorts," Nat. Genet., vol. 47, no. 3, pp. 284–290, Mar. 2015.

[26] A. Kessy, A. Lewin, and K. Strimmer, "Optimal Whitening and Decorrelation," Am. Stat., vol. 72, no. 4, pp. 309–314, Oct. 2018.

[27] S. Purcell et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," Am. J. Hum. Genet., vol. 81, no. 3, pp. 559–575, Sep. 2007.

[28] G. Csárdi, T. Nepusz, and E. M. Airoldi, "Statistical network

analysis with igraph." Berlin: Springer.[Google Scholar], 2016.

[29] M. A. Harris et al., "The Gene Ontology (GO) database and informatics resource," Nucleic Acids Res., vol. 32, no. Database issue, pp. D258-61, Jan. 2004.

[30] A. Fabregat et al., "The Reactome Pathway Knowledgebase," Nucleic Acids Res., vol. 46, no. D1, pp. D649–D655, Jan. 2018.

[31] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," Nucleic Acids Res., vol. 28, no. 1, pp. 27–30, Jan. 2000.

[32] K. Lage et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," Nat. Biotechnol., vol. 25, no. 3, pp. 309–316, Mar. 2007.

[33] A. D. Grotzinger et al., "Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits," Nat Hum Behav, vol. 3, no. 5, pp. 513–525, May 2019.

[34] S. Izrailev, "tictoc: Functions for timing R scripts as well as implementations of Stack and List structures (R package version 1.0)." 2014.

[35] A. L. Tyler, F. W. Asselbergs, S. M. Williams, and J. H. Moore, "Shadows of complexity: what biological networks reveal about epistasis and pleiotropy," Bioessays, vol. 31, no. 2, pp. 220–227, Feb. 2009.

[36] S. Burgess et al., "Guidelines for performing Mendelian randomization investigations," Wellcome Open Res, vol. 4, p. 186, 2019.

[37] W. van Rheenen, W. J. Peyrot, A. J. Schork, S. H. Lee, and N. R. Wray, "Genetic correlations of polygenic disease traits: from theory to practice," Nat. Rev. Genet., vol. 20, no. 10, pp. 567–581, Oct. 2019.

[38] C. Wallace, "Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses," PLoS Genet., vol. 16, no. 4, p. e1008720, Apr. 2020.

[39] N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller, "Pleiotropy in complex traits: challenges and strategies," Nat. Rev. Genet., vol. 14, no. 7, pp. 483–495, Jul. 2013.

[40] L. J. Carithers et al., "A novel approach to high-quality postmortem tissue procurement: the GTEx project," Biopreserv. Biobank., vol. 13, no. 5, pp. 311–319, 2015.

[41] Z. Ning, Y. Pawitan, and X. Shen, "High-definition likelihood inference of genetic correlations across human complex traits," Nat. Genet., pp. 1–6, Jun. 2020.

[42] H. Ay et al., "A computerized algorithm for etiologic classification of ischemic stroke: the Causative Classification of Stroke System," Stroke, vol. 38, no. 11, pp. 2979–2984, Nov. 2007.

[43] S. L. Pulit et al., "Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes," Neurology Genetics, vol. 4, no. 6. p. e293, 2018.

[44] R. Malik and M. Dichgans, "Challenges and opportunities in stroke genetics," Cardiovasc. Res., vol. 114, no. 9, pp. 1226–1240, Jul. 2018.

[45] R. Malik et al., "Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes," Nat. Genet., vol. 50, no. 4, pp. 524–537, Apr. 2018.

[46] M. Traylor et al., "A novel MMP12 locus is associated with large artery atherosclerotic stroke using a genome-wide age-at-onset informed approach," PLoS Genet., vol. 10, no. 7, p. e1004469, Jul. 2014.

[47] M. Traylor et al., "Genetic variation at 16q24.2 is associated with small vessel stroke," Ann. Neurol., vol. 81, no. 3, pp. 383–394, Mar. 2017.

[48] M. Dichgans, S. L. Pulit, and J. Rosand, "Stroke genetics: discovery, biology, and clinical applications," Lancet Neurol., Apr. 2019.

[49] S.-H. Lee, "Cerebral Small Vessel Disease," Stroke Revisited: Pathophysiology of Stroke. pp. 61–79, 2020.

[50] T. H. Pers et al., "Biological interpretation of genome-wide association studies using predicted gene functions," Nat. Commun., vol. 6, p. 5890, Jan. 2015.

General discussion

Contributions of this thesis

In this thesis I have shown that using different phenotype definitions and combining phenotypes can lead to discovery of new genetic associations in existing data. In chapter 1 we investigate the balance between stricter phenotype definitions and sample size. In chapter 2 we use the age at onset of ischemic stroke, a continuous phenotype that is complementary to the more commonly used case-control phenotype. We find an association of genetic variation in the ApoE locus with a lower age at onset in women but not in men.

We simulated a genetic variant that is associated with mortality and earlier death, but does not have a direct effect on ischemic stroke risk. Such a variant would be found in an analysis of age at onset of ischemic stroke, and not in a case-control analysis of risk. However, it is possible that this variant has an effect on ischemic stroke risk that has not been identified yet, possibly because it is sex-specific. The GWAS described in the first two chapters were performed in a relatively small sample. Further analysis in more individuals could identify more associated variants. Nonetheless, these studies do show that it is worthwhile to spend time on precisely defining the phenotype before a GWAS is performed. The literature review in chapter 3 enables a fair comparison of four previously published methods that aim to identify pleiotropic SNPs, as well as a comparison with our own method PolarMorphism. Our finding that two of these methods do not identify truly pleiotropic SNPs – but rather, SNPs with an effect on at least one of the traits under consideration - is valuable for prospective users of these methods. PolarMorphism, presented in chapter 4, is an efficient alternative for the methods described earlier. The code (implemented as a software package in R) is freely accessible on Github, enabling others to use it easily.

Limitations and starting points for future research

Genetic effects do not happen in isolation

We often call all phenotypic variation that cannot be explained by genetics 'environmental effects'. Some factors that are known to influence cardiovascular disease risk are commonly known health-related factors like high blood pressure, diet choices, and presence of other diseases (comorbidity). Socio-economic factors like income and educational attainment are also known to affect cardiovascular risk. Most of these factors are themselves associated with genetic variation. It is likely that many of these factors not only have an additional influence on a phenotype, but that the exact effect of our genotype changes depending on these factors; that there is an interaction between our genotype and our circumstances. If you have an auto-immune disease, genetic variation in immune-related pathways might have a bigger effect then someone who does not have an auto-immune disease. Studies of genetic and non-genetic effects should be incorporated more often to enable finding interactions between the two.

Large scale biobanks

In 2010 the UK biobank finished collecting data of 500,000 people; genetic data as well as numerous phenotypes ranging from blood pressure to how many cups of coffee they drink each day. [1] The UK biobank has proven to be a wealthy data source for research; almost 1,500 papers have been published about it. It is accessible to many researchers, provided their study proposal is approved by the ethical committee. Such a large and relatively homogeneous sample from one country makes it easier to study genotype-phenotype associations. Even correcting for confounders is possible (this can be difficult in other data sets, as we often do not have information on all confounders for all people in our data). However, the gain in statistical power from larger samples comes at a cost: most phenotypes are not very clearly defined. To give an example: we explored whether we could use the UK biobank

data to replicate our GWAS of age at onset of ischemic stroke, but the closest phenotype that was available was 'age stroke diagnosed' which included hemorhagic strokes in addition to ischemic strokes. Another reason why this phenotype was less useful to us, is that the UK biobank population is relatively young, and have not been followed for longer than 10 years. Therefore, there are not many stroke cases reported.

Another thing to be wary of is overfitting. Overfitting refers to the situation when a (prediction) model has too many parameters to fit on a small dataset and ends up fitting some of the parameters on noise. If the same model is applied to another dataset, it fails to perform because it learned the dataset-specific patterns or noise from the first dataset. We know how to deal with this if we are using data for one project. In the case of UK biobank however, thousands of people have been working with the same dataset for many projects. Some might be following up on results that have been generated with UK biobank data and using the same data to do additional analyses. Especially in those situations overfitting is a realistic problem, and researchers should use another dataset if possible.

Sex-specific effects

Sex-specific implies that a pathway that is tagged by a SNP is only relevant in one sex. Sex differences mean that the same pathway is relevant in multiple sexes, but to a different extent. In some cases, this could have a biological origin, hormone differences for instance. But in other cases, sociological differences could play a role. In biomedical research we usually consider sex, a biological phenotype that exists on a spectrum. Most people fall on either side and are male or female. Some people exist along the spectrum: intersex individuals.

Gender describes how we see ourselves. It also exists on a spectrum, with most people identifying as male or female, and some as non-binary (neither male or female) or a-gender (not strongly identifying with any gender). If someone's sex and gender align, they are cisgender. If someone's sex is not the same as someone's gender, they are transgender. Historically, most biomedical research has been done

in cisgender men. The past decades more attention has been given to the inclusion of women in trials and the study of diseases that are more common in women than in men. However, sex and gender diversity beyond men and women has largely been ignored. Most of the time, in genetic studies at least, only cisgender people are included. One of the reasons could be, that it is difficult to disentangle sex from gender. If we want to look at sex-specific effects, we generally split the group based on sex chromosomes (one group for people with XX-chromosomes and one for people with XY-chromosomes). To ensure that each group is as homogeneous as possible, we further narrow down to cisgender men and cisgender women. We usually do not have enough samples for other sexes or genders, or this information is not asked. While there are practical reasons to do this, it might not be the only way. In fact, one of the approaches to disentangle sex effects from gender effects could be to study transgender individuals who receive hormone replacement. However, this is a vulnerable community and research like this should be setup carefully.

Studying man-woman differences certainly is not enough to capture the full breadth of sex and gender diversity that exists among humans, but it is a step up from only studying biology in cisgender men. Everyone deserves the same health outcome, but sex- and gender-sensitive research and healthcare are necessary to ensure this equality.

Genetic ancestry diversity

The human reference genome is based on the genomes of only tens of people, and 70% was obtained from one individual who, as was found out later, had a high risk of diabetes. [2], [3] This begins to illustrate the lack of diversity in the reference genome, which is in theory just as useful as a reference as any individual's genome. The genetic ancestries of the individuals who are included in the human reference genome are largely unknown, although we know at least one person of African American ancestry was included. Thus, the reference genome is very biased to the few people who were included, but it is not necessarily very biased to European ancestry only. This problem

does show up in virtually all other aspects of genetic research: most people in GWAS datasets are white and of European ancestry. This can lead to less power to discover disease-associated SNPs that are common in non-European ancestries and rare in European ancestry, which further increases the existing racial healthcare gap. Furthermore, this has had and continues to have a trickle-down effect: if the big, well-powered GWAS are done in European ancestry individuals and we want to follow-up on that research, we should probably also restrict our study to European ancestry if we want to compare our outcomes with the previous study. The same goes for secondary analyses like LD-score regression, whose authors have made the LD-scores for European ancestry publicly available for download but not those for other ancestries. They did that because they understandably wanted to only show one example in their paper and did not calculate the LD-scores for other ancestries.[4] Pertaining to the work described in chapter 3, with the current methodology we have to restrict to one genetic ancestry because the LD-structure can vary considerably across ancestries. I chose to include only summary statistics of GWAS done in European ancestry because I wanted to be able to use the most well-powered GWAS. I do realize that by doing that I am upholding the status quo. I can not change the system on my own, but if more well-powered GWAS in non-European ancestries are available we can at least try to not make it worse. Luckily, more attention has been given to this problem recently [5] and biobanks in Taiwan [6], Africa [7,8], and many more countries and continents now exist.

The environmental cost of computation

A bioinformatician in a higher-income country has the possibility to use computers to analyse data in a fraction of the time they would have needed to do the analysis by hand. In contrast to wet lab researchers, computational researchers do not see resources being used. The hardware components are maintained in a location that they do not usually see, and there is a constant influx of the only resource needed: electricity. When we send a heavier task to the

computing cluster, it might take a longer time to complete but we are not aware of the energy requirements of our analysis. There are a few concrete steps that any researcher can take to make a change: a) Consider if you really need to run this analysis on all the data. b) Do you need to request all this memory? The energy cost depends on the amount of memory requested, not on the amount of memory that is actually used by your algorithm. c) Have you finished your analysis? Share the data! Not only does data sharing help advance science by enabling other researchers to use your results for new research. It also means that others do not have to run the same analysis to get the same results again, thereby saving the environmental costs of running the software each time someone uses your publicly shared data. You can read more about this topic on http://green-algorithms.org and read the accompanying paper. [9]

Final remarks

Over the past decades, GWAS methodology has developed to a point where large-scale analyses of millions of variants and ten thousands of individuals are routinely performed. Technological advances will almost certainly improve runtime and memory efficiency even more in the future. At present we can already gain a potential increase in statistical power by carefully defining our phenotype of interest, as we show in chapter one and two. As well as potentially improving the likelihood of identifying associated variants, different phenotypes can uncover associations with different facets of a disease. The discovery of pleiotropic variants from GWAS summary statistics (beyond those variants that are genome-wide significant in all GWAS of interest) is a relatively recent possibility. The engagement with the PolarMorphism preprint – it has been downloaded 346 times and mentioned by 24 twitter accounts in almost one month – shows that there is interest and demand for an efficient method for identification of pleiotropic variants from summary statistics. PolarMorphism and similar methods have the potential to uncover new trait relationships and shared underlying biology in the near future.

References

[1] C. Sudlow et al., "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," PLoS Med., vol. 12, no. 3, p. e1001779, Mar. 2015.

[2] S. Ballouz, A. Dobin, and J. A. Gillis, "Is it time to change the reference genome?," Genome Biol., vol. 20, no. 1, p. 159, Aug. 2019.

[3] R. Chen and A. J. Butte, "The reference human genome demonstrates high risk of type 1 diabetes and other disorders," Pac. Symp. Biocomput., pp. 231–242, 2011.

[4] B. K. Bulik-Sullivan et al., "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies," Nat. Genet., vol. 47, no. 3, pp. 291–295, Mar. 2015.

[5] H. Carress, D. J. Lawson, and E. Elhaik, "Population genetic considerations for using biobanks as international resources in the pandemic era and beyond," BMC Genomics, vol. 22, no. 1, p. 351, May 2021.

[6] J.-C. Lin, C.-T. Fan, C.-C. Liao, and Y.-S. Chen, "Taiwan Biobank: making cross-database convergence possible in the Big Data era," Gigascience, vol. 7, no. 1, Jan. 2018.

[7] P. Tindana et al., "Developing the science and methods of community engagement for genomic research and biobanking in Africa," Glob. Health Epidemiol. Genom., vol. 2, no. e13, 2017.

[8] P. Tindana et al., "Engaging research ethics committees to develop an ethics and governance framework for best practices in genomic research and biobanking in Africa: the H3Africa model," BMC Med. Ethics, vol. 20, no. 1, p. 69, Oct. 2019.

[9] L. Lannelongue, J. Grealey, and M. Inouye, "Green Algorithms: Quantifying the carbon footprint of computation," Adv. Sci. (Weinh.), vol. 8, no. 12, p. 2100707, Jun. 2021.

Nederlandse samenvatting
Het verband tussen variatie in ons genoom en onze individuele eigenschappen

Genetische variatie

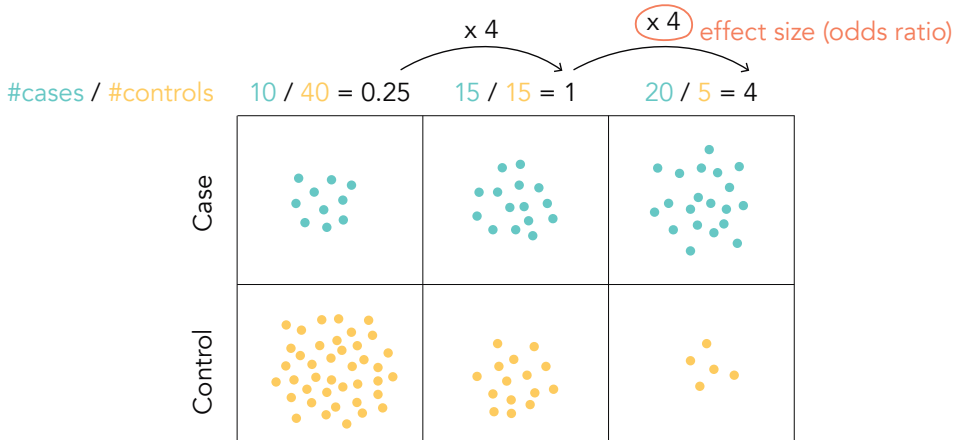
Ons genoom bestaat uit 3,2 miljard bouwstenen. Deze bouwstenen, nucleotiden, zijn aan elkaar geregen in 23 verschillende chromosomen. Elke nucleotide kan vier verschillende moleculaire vormen aannemen: adenine, thymine, cytosine, of guanine (afgekort tot A, T, C en G). De precieze volgorde van nucleotiden in iemands genoom noemen we het genotype. Het genotype van een willekeurig persoon is voor 99,5 % identiek aan het genotype van een willekeurig ander persoon. Toch verschillen mensen onderling in eigenschappen zoals lengte, haarkleur en risico op bepaalde ziekten. Deze variatie kan voor een deel worden verklaard door variatie in bijvoorbeeld dieet, leeftijd, en of iemand rookt of niet. Verschillen in genotype - genetische variatie - kunnen vaak ook een deel van de variatie in een eigenschap verklaren. Hoeveel van de variatie in een eigenschap verklaard kan worden door genetische variatie noemen we de erfelijkheid: een percentage tussen 0 en 100 %. Erfelijkheid is een moeilijk begrip, wat ik duidelijker zal maken door eeneiige tweelingen als voorbeeld te nemen. Eeneiige tweelingen hebben precies hetzelfde genoom, maar ervaren verschillende externe factoren tijdens hun leven. Als we zien dat de kans op een herseninfarct toeneemt als de tweeling-sibbe (de andere helft van de tweeling) een herseninfarct heeft gehad en afneemt als die geen herseninfarct heeft gehad, dan weten we dat de erfelijkheid van herseninfarct hoger is dan 0. Een plek in het genoom waar we veel verschillende nucleotiden zien in de populatie noemen we een Single Nucleotide Polymorphism (SNP, uitgesproken als ‘snip’).

Genoom-wijde Associatie Studies (GWAS)

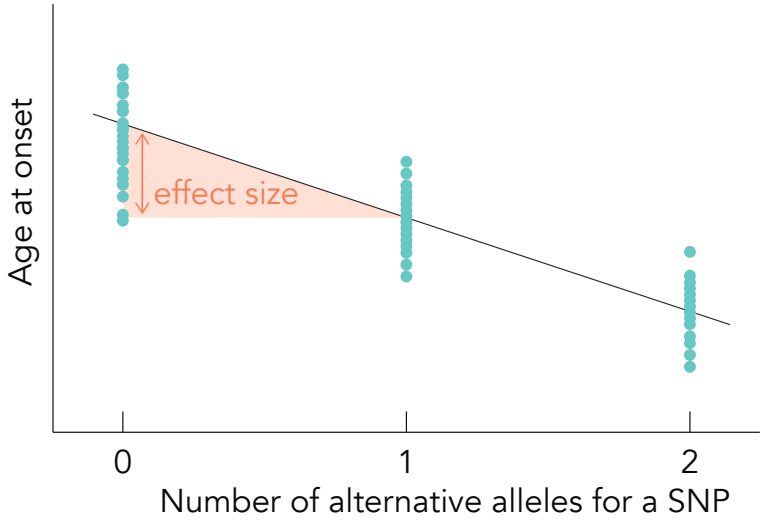
Als we SNPs willen vinden die de variatie in een eigenschap kunnen verklaren, hebben we van een groep mensen de volgende informatie nodig: het genotype, en de waarde van de eigenschap die we willen onderzoeken. Zie figuur 1 en 2 voor twee voorbeelden van zo’n studie (een GWAS). In figuur 1 zijn we op zoek naar SNPs die een verband hebben met een hoger risico op een herseninfarct. Per SNP doen we het volgende: mensen die geen herseninfarct hebben gehad (die

noemen we controls) zetten we in de onderste rij, en mensen die wel een herseninfarct hebben gehad (de cases) in de bovenste rij. De kolommen zijn opgedeeld op basis van het genotype voor de SNP in kwestie. Omdat we twee versies hebben van elk chromosoom hebben we ook twee versies van elke SNP. Er zijn dus drie opties: beide SNPs hebben het referentie allel, een heeft het referentie en de andere het alternatieve allel, of beide SNPs hebben het alternatieve allel. Een andere manier om dit te beschrijven, is het aantal alternatieve allelen te tellen; respectievelijk is dat 0, 1 of 2. Als de verhouding cases/controls per kolom groter wordt als er een extra alternatief allel bijkomt, dan wil dat zeggen dat deze SNP geassocieerd is met een hoger risico op het krijgen van een herseninfarct. Hoeveel groter het risico wordt kunnen we zien door de odds ratio te berekenen; we berekenen de verhouding cases/controls per kolom (respectievelijk 0.25, 1, en 4). De odds ratio geeft aan hoeveel groter de cases/controls verhouding wordt als er een extra alternatief allel bijkomt; in dit voorbeeld is de odds ratio 4 (want 1/0.25 en 4/1 is 4). De odds ratio geeft de effectgroottte van een SNP voor binaire case-control fenotypes.

In figuur 2 zijn we op zoek naar SNPs die een verband hebben met een grotere lengte. Deze data moeten we anders analyseren omdat lengte een continue eigenschap is; het kan in principe elke waarde aannemen die hoger is dan 0 en we kunnen niet oneindig veel groepjes mensen maken. Nu maken we voor elke SNP een grafiek; op de horizontale as noteren we weer het aantal alternatieve allelen voor de SNP, op de verticale as noteren we de lengte in meters. Elk persoon krijgt een stipje in de grafiek, op de plek van diens genotype en lengte. Hierna tekenen we een lijn die het verband tussen genotype en lengte het beste weergeeft. Deze lineaire regressielijn kunnen we opschrijven in een formule: lengte = helling*genotype + intercept. Wij zijn geïnteresseerd in de waarde van de helling; die geeft aan hoeveel de lengte toeneemt voor elk extra alternatieve allel. De helling beta geeft de effectgrootte van de SNP voor continue fenotypes.



Figuur 1. Een voorbeeld van case-control analyse voor een SNP. Elke stip is een persoon: blauw voor de cases (in de bovenste rij) en geel voor de controls (in de onderste rij). De kolommen geven het aantal alternatieve allelen aan dat ieder persoon heeft (0 = ref-ref, 1 = ref-alt of alt-ref, 2 = alt-alt). De ratio cases tot controls – de odds om herseninfarct te krijgen – is aangegeven boven elke kolom. De odds ratio is de ratio van de odds in opeenvolgende kolommen: in dit geval is de odds ratio 4, wat betekent dat het risico op herseninarct viervoudig toeneemt met elke bijkomende kopie van het alternatieve allel.



Figuur 2. Een voorbeeld van lineaire regressie van age at onset voor een SNP. Elke blauwe SNP is een persoon. The x-as geeft het aantal alternatieve alleles aan dat ieder persoon heeft (0 = ref-ref, 1 = ref-alt of alt-ref, 2 = alt-alt). De y-as geeft aan hoe oud ze waren toen ze een herseninfarct kregen. De schuine lijn is de regressielijn; deze lijn beschrijft de trend in de data het best. De helling van de lijn geeft de effectgrootte van deze SNP op age at onset; hoeveel eerder krijgen mensen gemiddeld eerder een herseninfarct met elk bijkomende kopie van het alternatieve allel?

Om te bepalen of we een verband misschien door toeval hebben gevonden, doen we een statistische test die ons een p-waarde geeft. Als de p-waarde laag genoeg is, kunnen we concluderen dat het waarschijnlijk geen toevallige vinding is. We kunnen echter nooit zeker weten of dit een oorzakelijk verband is zonder andere experimenten te doen.

Herseninfarct

Als de bloedtoevoer naar of in het brein is geblokkeerd, ontstaat er een plaatselijk gebrek aan zuurstof (ischemie). Als de blokkade maar kort duurde, heet dit een Transient Ischemic Attack (TIA). Als het langer duurde, heet het een herseninfarct (ook wel cerebrovasculair infarct of cerebrovasculair accident, CVA, genoemd). Het gebrek aan zuurstof in het brein kan leiden tot ernstige beperkingen of overlijden. Herseninfarct is deels erfelijk; ongeveer 38% van de variatie in voorkomen van herseninfarct kan verklaard worden door genetische variatie.

Er zijn drie subtypen van herseninfarct: cardio-embolisch herseninfarct wordt veroorzaakt door een bloedprop die in het hart vormt en via de bloedbaan naar het brein reist. Grote-vaten herseninfarct wordt veroorzaakt door aderverkalking die een van de twee halsslagaders blokkeert. Bij een kleine-vaten herseninfarct zijn de kleine vaten in het brein zelf geblokkeerd. Op dit moment zijn er tientallen SNPs gevonden die een verband hebben met een hoger risico op het krijgen van een herseninfarct of een van de subtypes. Als we meer SNPs kunnen vinden met een verband met herseninfarct, kunnen we meer leren over de biologische processen die een rol spelen in het ontstaan van een herseninfarct. Het zou ook kunnen helpen bij het identificeren van mensen met een hoger genetisch risico, zodat zij samen met hun arts ervoor kunnen zorgen dat hun aanpasbare niet-genetische risico zo laag mogelijk blijft. In hoofdstukken 1 en 2 bespreken we de GWAS resultaten voor verschillende fenotype-definities voor herseninfarct en vinden daarmee nieuwe SNPs.

Een andere fenotype-definitie leidt tot een ander resultaat

In hoofdstuk 1 gaan we op zoek naar SNPs die geassocieerd zijn met het risico op het krijgen van een van de subtypes van herseninfarct. Hier lopen we echter tegen een probleem aan; er zijn meerdere methodes om een herseninfarct-patiënt te diagnosticeren met een subtype. Deze drie methodes zijn het niet altijd met elkaar eens, dus we wisten niet zeker hoe we het fenotype moesten definiëren. We hebben voor elke subtype vijf verschillende fenotype-definities gebruikt en met elkaar vergeleken om te bepalen welke het beste werkt: de drie originele methodes, de intersect (doorsnede) en de union (vereniging). De intersect diagnosticeert iemand alleen met een bepaald subtype als alle originele methodes dat deden. De union diagnosticeert iemand met een bepaald subtype als tenminste een van de originele methodes dat deed. Dit betekent dat de intersect-cases een kleinere groep zijn dan de union-cases. En een kleinere groep betekent dat we minder kans hebben om een effect te ontdekken, tenzij het een heel groot effect is. Theoretisch is de intersect wel een strengere definitie, die twijfelgevallen eruit filtert en alleen die mensen diagnosticeert die vrij zeker een bepaald subtype hebben. Uit onze resultaten blijkt, dat de theoretisch striktere intersect in veel gevallen beter geschikt lijkt om te gebruiken als fenotype in een GWAS. Dit geeft aan dat het voor herseninfarct belangrijker lijkt te zijn dat we heel zeker weten wat iemands fenotype is en dat het minder belangrijk is om een zo groot mogelijke groep mensen te analyseren. Met deze nieuwe fenotype-definitie vinden we zelfs een SNP waarvan we nog niet wisten dat die een verband had met herseninfarct.

In hoofdstuk 2 gebruiken we een continu fenotype om SNPs te vinden die gerelateerd zijn aan herseninfarct. In plaats van het binaire case-control fenotype analyseren we de leeftijd waarop iemand een herseninfarct kreeg (de age at onset, of leeftijd bij aanvang). We splitsen hier de groep ook op in mannen en vrouwen, omdat veel hart- en vaatziekten sekse-specifieke risicofactoren hebben en dat wellicht

ook geldt voor age at onset. We vinden een SNP die geassocieerd is met een 1.6 jaar eerder herseninfarct in vrouwen. Dezelfde SNP is niet geassocieerd met een eerder herseninfarct in mannen.

Genetische varianten die geassocieerd zijn met meerdere fenotypes: pleiotropie

Tegenwoordig delen veel onderzoekers de resultaten van GWAS in online databases die toegankelijk zijn voor iedereen. Dat zorgt er niet alleen voor dat anderen geen analyses hoeven te doen die al gedaan zijn, maar ook dat de GWAS-resultaten gebruikt kunnen worden voor vervolgonderzoek. Moleculair biologen kunnen bijvoorbeeld in het lab gaan uitzoeken of de GWAS SNPs een direct effect hebben op het fenotype. Door deze databases weten we ook, dat sommige SNPs een associatie hebben met meerdere fenotypes; dit noemen we pleiotropie. Pleiotrope SNPs kunnen ons meer leren over de biologische mechanismen die betrokken zijn bij een fenotype; als we al iets weten over het onderliggende proces waarmee een SNP van invloed is op een fenotype en die SNP blijkt ook een effect te hebben op een ander fenotype, dan is dat datzelfde proces misschien ook betrokken bij het andere fenotype. Pleiotrope SNPs kunnen ons ook helpen om verbanden te leggen tussen fenotypes die misschien niet vaak samen voorkomen maar wel onderliggende processen delen.

In hoofdstuk 3 geef ik een overzicht van vier recent gepubliceerde methoden die SNPs kunnen identificeren die gedeeld zijn door meerdere fenotypes, door gebruik te maken van openbaar beschikbare GWAS-resultaten. Er bleken twee methoden te zijn die SNPs identificeren met een associatie met een of meer fenotypes, in plaats van alleen SNPs die een associatie hebben met méér dan een fenotype. Dat betekent dat deze methoden een SNP met een effect op maar een fenotype ook pleiotroop noemen, terwijl ze dat niet zijn. De andere twee methoden verschillen in aanpak maar geven vergelijkbare resultaten. De ene methode vindt meer pleiotrope SNPs dan de andere, maar de andere is sneller.

In hoofdstuk 4 introduceer ik een nieuwe methode die wij hebben ontwikkeld. PolarMorphism gebruikt GWAS-resultaten van meerdere fenotypes om pleiotrope SNPs te identificeren. PolarMorphism gebruikt een andere aanpak om de data te analyseren dan de eerder beschreven methodes. PolarMorphism is vele malen sneller dan de methode die in hoofdstuk 3 de meeste pleiotrope SNPs kon vinden, zonder daarbij minder SNPs te vinden. Ook kon deze andere methode maximaal twee fenotypes tegelijk analyseren, terwijl PolarMorphism een theoretisch oneindig aantal fenotypes kan analyseren.

Concluderend, in dit proefschrift heb ik aangetoond dat het anders definiëren of combineren van fenotypes kan leiden tot de ontdekking van nieuwe genetische associaties in bestaande data. De GWAS uit de eerste twee hoofdstukken zijn uitgevoerd in een relatief kleine groep mensen, en het is aannemelijk dat een analyse in een grotere groep mensen nog meer SNPs zal vinden. Beide studies tonen echter wel aan dat het loont om meer aandacht te besteden aan het precies definiëren van het fenotype voordat men een GWAS doet. Het literatuur-review beschrijft de overeenkomsten en verschillen tussen gepubliceerde methoden voor de identificatie van pleiotrope SNPs. Dit hoofdstuk maakt het mogelijk om onze methode PolarMorphism te vergelijken met bestaande methoden en zo op waarde te schatten. Onze bevinding dat twee gepubliceerde methoden niet doen wat ze claimen te doen is waardevol voor de wetenschappelijke gemeenschap. Ten slotte kan PolarMorphism gebruikt worden door andere onderzoekers, omdat we de code hebben gedeeld als softwarepakket dat gebruikt kan worden in de programmeertaal R.

Dankwoord

Dankwoord

Het is zover: mijn proefschrift is af. Aan het begin van mijn PhD kon ik me niet voorstellen dat ik ooit op dit punt zou komen. En eerlijk gezegd kon ik me dat ook de afgelopen weken nog vaak niet voorstellen. Dat het toch zover is gekomen heb ik te danken aan de volgende mensen.

First, I want to thank all people who gave consent for their data to be used. Without you I would not have been able to do any of the research presented here.

Geachte **Gerard Pasterkamp**, **Hester den Ruijter**, **Ynte Ruigrok**, **Daniel Oberski** en **Marcel Reinders**, Beste **leescommissie**: zonder jullie tijd en moeite had dit proefschrift niet beoordeeld kunnen worden. Ik wil jullie bedanken voor alle energie die jullie hebben besteed aan het lezen van mijn – toen nog niet zo netjes opgemaakte – proefschrift.

Jeroen, het eerste waar ik je voor wil bedanken is je vertrouwen. Zonder precies te hoeven begrijpen 'hoe ik werk', vertrouwde je erop dat mijn idee om een dag per twee weken vrij te nemen goed uit zou pakken. Toen ik het absurde idee kreeg om SNP effect sizes uit te drukken in polaire coordinaten liet je me mijn gang gaan, ook toen ik er steeds dieper in dook en met (in ons veld) zeldzaam gebruikte statistiek weer naar boven kwam. Als ik ooit minder motivatie had of – wat vaker voorkwam – even totaal het overzicht kwijt was, wist ik dat ik na een meeting met jou weer aan de slag zou kunnen met een duidelijk doel voor ogen. Na de zoveelste versie van het PolarMorphism paper wilde ik vooral mijn thesis zo snel mogelijk afmaken, dat paper zou later wel komen. Toch heb je me over kunnen halen om het paper te submitten en daarna mijn thesis af te maken. En daar ben ik je erg dankbaar voor, ik had het inderdaad echt niet leuk gevonden om een 'oude versie' van PolarMorphism te moeten verdedigen.

Sander, ook al waren we het niet altijd direct eens (en had ik altijd nog wel wat vragen of opmerkingen voor het zover was), uiteindelijk kwamen we toch op dezelfde conclusie uit. Ik heb veel van jou geleerd: specifieke ('een beller is sneller' ; al dacht ik eerst van niet met mijn bel-aversie, toch heb je eigenlijk wel gelijk) en minder specifieke dingen. Bijvoorbeeld dat je heel veel gewoon kunt vragen aan anderen, en dat er meer mogelijk is dan ik soms denk. Ik heb geleerd hoe waardevol een netwerk kan zijn, en ben dankbaar dat ik een deel van het jouwe heb mogen ontmoeten. Bijvoorbeeld tijdens de statgen meeting, het groepje mensen dat jij bij elkaar had verzameld dat allemaal 'iets met GWAS' deed in het UMC.

Sara, compared to the full length of my PhD you have only supervised me for a short time, but it doesn't feel that way. You were there at the very beginning and watched me – and helped me – grow as a scientist. I still remember that you wanted me to be 'a bucket of knowledge', how I should be able to explain exactly what I was working on, how the methods I used work, and of course what the null hypothesis is. I've definitely come a long way, but I think I'm there now. I want to thank you for your committed supervision.

I want to thank the De Riddertjes, my former colleauges in the De Ridder group. It's strange to finish your contract during a global pandemic, not having known that that one CMM borrel in early 2020 would be your last. I felt so supported by all of you, you made me feel comfortable at work. Even during the pandemic when we were working from home full-time I felt connected to our group, because of all the (serious and less serious) chat conversations on Slack. East siders Emmy, Joanna Wolthuis, Joske, Liting, Marleen, Roy, and Myrthe: thanks for making me feel at home in the office. Amin, Adrien, Alessio, Alexandra, Brent, Carlo, Ivo, Luca, Marc: thanks for the gezellige retreats, BBQ's and other lab activities. Joep: even though it's been a while since you left our lab, you were an important part of it, and you always made time to help others. Thanks to you, I dared to start up InDesign and Illustrator. I actually don't know what this thesis would have looked like if I had never done that.

I also want to thank my new colleagues of **the Kemmeren group** at the Prinses Maxima Centrum. It is a little scary to start a postdoc in a new field, but you have made me feel very welcome. I feel like I can always ask one of you for help. It doesn't matter whether it's about science, practical matters, technical things, or a recipe for vegan carrot cake.

Renée Verdiesen, bedankt voor de fijne samenwerking aan jouw project en het kopje koffie dat we buiten dronken tijdens de pandemie. Ik weet het nog goed, het was zo fijn om weer 'gewoon een leuke collega' te spreken.

Imogen Morris, even though the TRIB3 SNP did not replicate in external data and our collaboration ended, I want to thank you for taking the time to explain all the difficult immune cells to me ;-) I enjoyed the few meetings we had, it was really cool to combine your molecular biology experience with my statistical background.

René Eijkemans, bedankt voor de tijd die u nam om feedback te geven op het PolarMorphism project.

I want to thank all **co-authors** for the collaborations on the papers, which are now part of this thesis.

I want to thank everyone who has been part of the **statgen meeting** for the useful discussions. It was nice to get to know some people who were also doing GWAS.

Ik wil graag de studenten bedanken die ik heb mogen begeleiden: **Marten**, **Susanne** en **Michelle**. Ik ben erg blij met de literatuur-reviews die jullie schreven, het scheelde mij tijd en moeite dat ik de artikelen zelf niet hoefde uit te zoeken. Daarnaast vond ik het heel leuk om jullie steeds zelfstandiger en kritischer te zien worden. En Michelle, jouw literatuur-review is de basis geweest van hoofdstuk 3 en heeft een bijdrage geleverd aan hoofdstuk 4. Bedankt voor de leuke discussies over je verslag, vooral tijdens het thuiswerken door de pandemie was het echt iets om naar uit te kijken.

The **RSG board 2018-2020**: **Nila**, **David**, **Alexandra** and **Liting**. Thanks for the collaboration, I was happy to be part

of this team when we were busy organizing something. We organised quite a few well-visited events before corona hit, and I think even our online BioSB PhD retreat was gezellig.

En dan dat andere bestuur, waarmee ik tien jaar geleden de Utrechtse Scheikundige Studentenvereniging "PROTON" (ja dat moet met hoofdletters) draaiende hield: **Frans**, **Anne-Eva**, **Petra**, **Stijn** en **Leonie**. Bedankt dat jullie er toen waren en bedankt dat jullie er nog steeds zijn. Onze weekendjes weg zijn net als vroeger maar dan nog chiller, omdat we elkaar steeds langer kennen en stiekem misschien omdat we steeds ouder worden.

Frans en **Stijn** maken, samen met **Annelies** en **Robin**, ook deel uit van mijn D&D-groepje: ik ben blij dat ik jullie daardoor regelmatig zie. Heel fijn om na werk even bij te praten en dan lekker met heel andere dingen bezig te zijn, zoals "Zullen we hier uitrusten of is dat toch niet zo slim in deze donkere grot vol goblins?" of "Zal ik Robin healen of red ie het zelf wel nadat hij al een paar rondes voor death saves aan het rollen is?"

Anne-Eva, bij jou kan ik altijd terecht voor wijze raad. Of het nou om vriendschappen gaat, of om een nieuwe zonnebril. Dankzij jouw kennis van regels en richtlijnen rondom de PhD en werk in het algemeen voel ik me zekerder om voor mezelf op te komen als dat nodig is. En verder is het gewoon heel erg fijn om met je af te spreken. Er gaat vaak net iets te lang voorbij voor we elkaar weer zien, maar daarna weet ik weer wat ik miste.

Sarah, thank you for always being so enthusiastic! Maybe it's the Dutch and German words sprinkled in your sentences, but whenever I get a message from you I get a smile on my face. Now that the worst of the pandemic is (hopefully) over, we can actually experience that we both work on de Uithof and can get a coffee or lunch together ;-)

Papa en **mama**, jullie hebben me niet zoveel gezien de afgelopen tijd. Eerst corona, en toen die thesis. Maar die was bijna af! Maar toch nog even dit… En nog even dit paper afmaken. En het design moet nog. Het duurde allemaal langer dan ik had gedacht, en

het kostte vooral veel meer energie, waardoor ik niet zomaar even met de trein heen en weer naar Elst kon. Nu kunnen jullie tenminste eindelijk zien waar ik al die tijd mee bezig was.

Sergio en Yvo, ook jullie hebben me niet zoveel gezien de laatste maanden. Maar gelukkig konden we een dagje naar de Efteling (waar ik nog bezig was met een poster afmaken en opsturen… op mijn mobiel). Het was leuk om iets met zijn drieen te doen, misschien kunnen we dat vaker doen nu Yvo in Utrecht woont. En dan beloof ik dat ik geen poster zal maken… op mijn mobiel. En Yvo, bedankt voor je feedback op het ontwerp van de cover!

Joske, je gelooft nooit hoe laat ik dit stukje van het dankwoord schrijf. Geheel in jouw stijl ben ik tot een uur 's nachts opgebleven om dit te schrijven. Onze verschillende bedtijden zijn maar een klein voorbeeld van hoe verschillend we zijn. Maar het werkt blijkbaar goed: een nerveus, snel overprikkeld persoon die door schade en schande heeft geleerd heel georganiseerd en gestructureerd te leven, en de social butterfly die liever niet teveel plant en oplaadt in grote groepen. Ik heb door jou zoveel mensen leren kennen, en ben iets minder bang om zomaar op een groep af te stappen. Naast onze vriendschap die een goede afleiding was van de PhD-stress, kon ik ook over die PhD-stress heel goed bij jou uitrazen. Je begreep precies wat ik bedoelde, omdat je anderhalf jaar geleden door precies dezelfde fase was gegaan. Bedankt voor je steun, en alvast bedankt voor de steun die je me als paranimf rondom de verdediging ongetwijfeld zult geven.

Myrthe, toen je net in onze groep begon als postdoc kende ik je helemaal niet goed. Het viel me op hoe geinteresseerd je was in iedereen. Toen we samen het afscheidsfeest voor Sara en Joep organiseerden, leerde ik je kennen als iemand die niet alleen goed was in de praktische kanten van organiseren, maar vooral ook in de communicatie met mensen. Tegenwoordig spreken Joske, jij en ik nog steeds regelmatig af voor een kopje koffie en ik vind het dus heel tof dat jullie mijn paranimfen zijn.

Ruben, ik moest en zou per se vandaag mijn proefschrift afmaken en daarom schrijf ik dit in een stil huis terwijl jij al slaapt (voor de oplettende lezer, het is nu kwart over een 's nachts). De afgelopen maanden heb ik vaak mijn thesis op nummer een gezet, soms samen met en soms voor jou. Bedankt voor de momenten dat je me de ruimte gaf om nog even te programmeren of nog even te schrijven of nog even aan mijn thesis te werken. En ook heel erg bedankt voor de momenten dat je dat niet deed. Dat je me van achter de computer en voor de TV sleepte. Bedankt voor alles wat je in het huishouden doet, als ik een goede week heb. En voor alle extra taken die je doet in de andere weken. En bedankt dat je er altijd bent, samen met jou kan ik volgens mij alles aan. Zoals met een noise-cancelling koptelefoon lijkt het allemaal net iets minder hard als jij bij me bent.

About the author

Joanna von Berg was born on the 22nd of January 1992 in Voorburg, The Netherlands. Joanna grew up in Zoetermeer and attended the first year of high school (tweetalig VWO) at het Alfrink College in 2004. After moving to Elst (Gelderland), they completed high school at het Overbetuwe College in Bemmel in 2010. During their bachelor Chemistry at Utrecht University they focused on biochemistry and structural biology courses and also took a few courses from the computer science curriculum. In their master's programme Molecular and Cellular Life Sciences they took general biology and bioinformatics courses. They did two internships: one in the group of Loes Kroon-Batenburg in Utrecht, and one in the group of Hölger Fröhlich in Bonn, Germany. In 2017 Joanna started their PhD in the group of Jeroen de Ridder, under co-supervision of Sara Pulit (2017-2018) and Sander W. van der Laan (2018-2021).

From a young age Joanna has enjoyed being involved in organizations, starting with the 'leerlingenraad' (students council) in high school. During the last year of their bachelor's, Joanna was a full-time board member of the U.S.S. Proton, the association of chemistry students at Utrecht University. They wrote for and edited the Chemograaph, a magazine published by the U.S.S. Proton. From 2018 until 2020, they were a board member of RSG Netherlands, the Dutch Regional Student group of the ISCB (International Society of Computational Biology).