

# Quantitative Genetics Essay

Title: The current *state of the art* in gene hunting studies



Mark Patrick Roeling  
[mp.roeling@psy.vu.nl](mailto:mp.roeling@psy.vu.nl)  
student id: 2023806

course Quantitative Genetics 2010-2011, #470733  
prof. dr. D.I. Boomsma  
dr. M. de Moor

Period: October, 2010

Words: 1.780

## The current *state of the art* in gene hunting studies

Over the past years, there has been an enormous increase in genomic discoveries involving complex, non-Mendelian, diseases. Over 100 loci for as many as 40 common diseases are robustly identified and replicated in genome-wide association (GWA) studies. GWA studies make use of high throughput genotyping technologies to assay hundreds of thousands of common genetic variants with the purpose of finding genetic linkage of a quantitative trait with implications to uncover the genetic basis of quantitative phenotypic variation (Mackay, Stone, & Ayroles, 2009). This revolutionary approach allows the interrogation of the entire genome at high-resolution level in numerous unrelated individuals, constrained by prior hypotheses regarding genetic association with disease (Hirschorn & Daly, 2005). The aim of GWA studies has been two-fold: (1) identifying markers that can be used to predict individual disease risk by full description of the susceptibility architecture of major biomedical traits; and (2) highlighting the molecular pathways underlying disease in order to translate the findings into clinical practice (e.g. by providing potential targets for therapy) (Hardy & Singleton, 2009; McCarthy et al., 2009).

The typical GWA study comprises four stages. First, selection of a large number of individuals with the disease or quantitative trait of interest and an adequate control group. Second, DNA isolation, genotyping, and data review to guarantee quality. Third, statistical tests for associations between the SNPs passing quality thresholds and the disease/trait, often by use of a Manhattan plot. Finally, replication of identified associations in an independent population sample or examination of functional implications experimentally. Replication experiments can result in the combination of three outcomes: selected loci show clear and unequivocal association with disease, show no association signal whatsoever, or show an association with disease that is not of sufficient magnitude to pass a predetermined statistical threshold (Hardy & Singleton, 2009).

Three major advances in human genetics have enabled GWA studies: the publication of the human genome, the characterisation of common human genetic variation by the HapMap project<sup>1</sup>, and the development of SNP(Single Nucleotide Polymorphism)-arrays (Strachan & Read, 2011). Fundamental in GWA studies is the linkage of Quantitative Trait Loci (QTL) to a marker locus. Respectively, individuals with a different marker locus genotype will have different mean values of the quantitative trait. The most common markers are SNPs, polymorphic insertions or deletions (indels) and simple sequence repeats (microsatellites) (Mackay, Stone, & Ayroles, 2009).

### Study designs

An association study uses recombination, but from a historical attenuation. The population used in association mapping is removed from its progenitors from many generations. This results in the shuffling of the initial haplotypes through recombination in a randomly mating population. The effect of this shuffling is to uncouple all but the most tightly linked markers from the causal locus. The causal locus can be localized with precisions as only the most tightly linked markers will predict the organismal phenotype.

The most published GWA study design to date makes use of the case-control design. Here, the allele frequencies in patients with the disease of interest are compared to those in a disease free comparison group (McCarthy et al., 2009; Strachan & Read, 2011). This design is prone to many assumptions to prevent substantial biases and spurious associations (see Table 1). Case control studies can produce valid results that, especially for rare diseases, may not be obtainable in any other way (Pearson & Manolio, 2008). Also, the manner in which the controls are ascertained has implications for the power of the study and for sample size

---

<sup>1</sup> International HapMap Consortium. A haplomap type of the human genome. *Nature*, 2005; 437; 1299-1320.

(McCarthy et al., 2009). Cohort studies depend on the collection of extensive baseline information in a large number of participants who are then observed to assess the incidence of participants who are then observed to assess the incidence of disease in subgroups. This implies the subjects are under observation during development, allowing a direct measure of risk with fewer biases compared to case-control studies. Considerations in this design involve the large sample size that is required, the length of the follow-up, and the control of confounding trait-variation. The trio design includes the affected case participant and both of his/her parents. Here, only the offspring is phenotyped, but all three members are genotyped.

**Table 1. Genome Wide Association Study designs**

	<b>Case-control</b>	<b>Cohort</b>	<b>Trio</b>
Assumptions	Case and control participants are drawn from the same population Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified Genomic and epidemiologic data are collected similarly in cases and controls Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls	Participants under study are more representative of the population from which they are drawn Diseases and traits are ascertained similarly in individuals with and without the gene variant	Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents
Advantages	Short time frame Large numbers of case and control participants can be assembled Optimal epidemiologic design for studying rare diseases	Cases are incident (developing during observation) Direct measure of risk Fewer biases than case-control studies Continuum of health related measures available in population samples not selected for presence of disease	Controls for population structure: immune to population stratification Allows check for Mendelian inheritance patterns in genotyping quality control Logistically simpler for studies of children's conditions Does not require phenotyping of parents
Disadvantages	Prone to a number of biases including population stratification Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases Overestimate relative risk for common diseases	Large sample size needed for genotyping if incidence is low Expensive and lengthy follow-up Existing consent may be insufficient for GWA genotyping or data sharing Requires variation in trait being studied Poorly suited for studying rare diseases	May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset Highly sensitive to genotyping error

From: Pearson and Manolio, 2008

The frequency with which an allele is transmitted to an affected offspring from heterozygous parents is then estimated (Spielman, McGinnis, & Ewens, 1993). Despite advantages on the level of unsusceptibility to population stratification or genetic differences between case and

control participants, the trio design is challenged by its sensitivity to small degrees of genotyping error (see Table 1).

#### Many variants with few effects

With the publication of numerous GWA studies, different complications have been identified regarding to pleiotropy, recombination and epistatic interactions (Mackay, Stone, & Ayroles, 2009). Recent discussion relates to the issue whether continuing to increase sample size is beneficial in relation to the costs and whether the GWA approach is appropriate in complex traits in the perspective of subject ascertainment (McCarthy et al., 2009). Because true linkage signals have to be separated from noise, researchers have to make use of high thresholds that needs to be exceeded by a marker before a signal is accepted as a likely disease-causing candidate. These high thresholds reduce the problem of false positives but are also responsible for the loss of true disease markers with small effects. By increasing the sample size, statistical noise from non-associated markers can be dialed down in order to filter out smaller effect disease genes (Pearson & Manolio, 2008). Several recent papers confirmed larger sample sizes are related to higher statistical power. Allen et al. (2010) comprised 183,727 individuals to identify >180 common genetic variations to be associated with adult height, explaining 10% of the phenotypic variation in height. Despite the relative high percentage of explained phenotypic variance, the variants that have been found explain a small fraction of the overall genetic contribution (Goldstein, 2009). The differentiation between the explained phenotypic variance by GWA studies and the heritability estimates postulated the argument that heritability may be overestimated, or measured inaccurately or even wrong (Manolio et al., 2009). Many SNPs are required to find the remaining variants in a given trait. GWA studies rely on the *common disease, common variant* hypothesis which suggests that genetic influences on many common diseases will be at least partly attributable

to a limited number of allelic variants present in more than 1-5% of the population (Collins et al., 1997). Rare disease causing variants are unlikely to be detected with this approach (Pearson & Manolio, 2008). Common variants typically only increase risk by 10 to 70%. Although several complex traits have been studied extensively, less than 20% of the heritable variance is explained (e.g. Wellcome Trust Case Control Consortium, 2007). Moreover, many reported genes are not directly associated with protein expression or regulation. Instead of loci mapping to recognizable proteins in open reading frames, they modulate the RNA-product by alteration of transcription or translational efficiency (Hardy & Singleton, 2009). Another approach was suggested by Yang et al. (2010) in which a regression framework was used considering all 294,831 SNPs simultaneously, together accounting for 45% of the variation. The authors write: "There are two logical explanations for the failure of validated SNP associations to explain the estimated heritability: either the causal variants each explain such a small amount of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped". While demonstrating that many variables are highly correlated with height, the authors are also concluding that there must be thousands of weakly penetrant causes. Feeding the gesture that the other 55% of the variance is to be found in variants that have not been typed in the phenotype of interest (Lambert, 2010). Increasing studies in sample size postulates the problem of ascertainment. Luckily, for many psychiatric conditions this has been standardized (Ku et al., 2010). In complex behavioral traits (e.g. autism), the phenotype is often defined according to two principles: narrow and broad (Benayed et al., 2009; Folstein & Rosen-Sheidley, 2001; Gharani et al., 2006). This approach seems beneficial as the narrow clinically defined (affected/unaffected) phenotypes are not necessarily appropriate in genetic studies.

## Prospects for the future

Although only a small amount of the phenotypic variance has been explained by GWA studies and only common variants are reported, they have contributed substantially in our understanding of disease mechanisms (Hirschhorn, 2009). Common variants can identify large numbers of loci that implicate biologically relevant genes and pathways (Allen et al., 2010) but contribute very little to individual disease risk predictions over existing clinical makers for most common diseases (Goldstein, 2009). In the future, these common variants with very little (subthreshold) contributions can be picked up by large scale meta-analysis. Besides presenting new loci for height, Allen et al. (2010) also conducted a 46-studies GWA-data meta-analysis to recover reported variants associated with adult height that, until then, remained largely irrelevant. Rapid progress is being made through the publication of meta-analyses that suggest many more common variants conferring a risk of disease will be identified soon (Kraft & Hunter, 2009). Several sources agree on a substantial role for rarer variants with substantially larger effect sizes in the near future. Variants with the largest individual effects on disease will tend to be rare as the result of natural selection prohibiting increases in disease-associated variants (Goldstein, 2009). The application of Copy Number Variants (CNV) to study functional impact of rare variants in autism has been reported a success by Pinto et al. (2010). While the technology for GWA studies has been well established and reliable, the survey of rare variants is only just becoming feasible (Mackay, Stone, Ayroles, 2009).

As sequencing technology becomes more affordable, targeted gene sequencing studies and GWA studies using new chips targeting variant throughout the genome at lower frequencies will come to rise soon. Complete genome sequencing of numerous patients and controls will be applied to seek for rare variants explaining individual differences. The detection of small effects by sequencing could yield a long list of rare disease-associated

variants explaining more phenotypic variance compared to common variants. Also, sequencing allows for the identification of mutations and epistatic interactions. Altogether these findings could be of much more use for individual risk prediction (Kraft & Hunter, 2009). New statistical models and applications will be required to allow interaction of underlying genetic associations, integrate environmental and epigenetic inheritance patterns, and test for causal relations between genes and quantitative traits instead of correlations. How fast the genetic interplay between common and rare genetic variants may be unravelled will ultimately depend on the cooperation of institutes around the world. Finally, the quest for an integration of common and rare genetic effects together with environmental influences to the phenotypic variation relies on the availability of technological and biometrical applications.

## References

- Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., ... Hirschhorn, J.N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *published online*.
- Benayed, R., Gharani, N., Rossman, I., Mancuso, V., Lazar, G., Kamdar, S. et al. (2005). Support for the homeobox transcription factor gene ENGRAILED 2 as an autism spectrum disorder susceptibility locus. *Am J Hum Genet*, *77*, 851–868.
- Collins, F.S., Guyer, M.S., & Chakravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science*, *278* (5343), 1580-1581.
- Folstein, S., & Rosen-Sheidley, B. (2001). Genetics of autism: complex aetiology for a heterogeneous disorder. *Nature Review Genetics*, *2*, 943-955.
- Gharani, N., Benayed, R., Mancuso, V., Brzustowicz, L.M., Millonig, J.H. (2004). Association of the homeobox transcription factor, ENGRAILED 2, 3, with autism spectrum disorder. *Mol Psychiatry*, *9*, 474–484.
- Goldstein, D.B. (2009). Common genetic variation and human traits. *New England Journal of Medicine*, *360*, (17), 1696 – 1698.
- Hardy, P., & Singleton, A. (2009). Genomewide Association Studies and Human Disease. *New England Journal of Medicine*, *360*, (17), 1759 - 1768.
- Hirschhorn, J.N. (2009). Genomewide Association Studies – Illuminating Biological Pathways. *New England Journal of Medicine*, *360*, (17), 1699 – 1701.
- Kraft, P., & Hunter, D.J. (2009). Genetic Risk Prediction – Are we there yet? *New England Journal of Medicine*, *360*, (17), 1701 - 1703.
- Ku, C.S., Loy, E.Y., Pawitan, Y., & Chia, K.S. (2010). The pursuit of genome-wide association studies: where are we now? *Journal of human genetics*, *55*, 195-206.

Lambert, C. (2010). Missing heritability and the future of GWAS. Retrieved online from <http://blog.goldenhelix.com/?p=228>

Mackay, T.F.C., Stone, E.A., & Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Review Genetics*, *10*, 565-577.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., ... Visscher, P.M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747-753.

McCarthy, M.I., Abecasis, C.R., Cardon, L.R., Goldstein, D.B., Little, J., ... Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Review Genetics*, *9*, 356-369.

Pearson, T.A., & Manolio, T.A. (2008). How to interpret a Genome-wide Association Study. *JAMA*, *299* (11), 1335 – 1344.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., ... Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, *466*, 368 – 372.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., ... Loos, R.J.F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *686*.

Spielman, R.S., McGinnis, R.E., & Ewens, W.J. (1993). Transmission test for linkage disequilibrium. *American Journal of Human Genetics*, *52*, 3, 506-516.

Strachan, T., & Read, A. (2011). *Human Molecular Genetics*, 4<sup>th</sup> edition. New York: Garland Science.

Welcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature*, *447*, 661 – 678.

### Unclear in the following article:

Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., ... Hirschhorn, J.N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*.

- At the moment it is unclear for me how this GWAS may provide more clues about the identity of the functional genes at each locus. If so, why haven't other large GWA studies previously attenuated this.
  - ➔ Clear, to keep it short: by the increase in sample size and the application of meta-analysis. I found the answer in class and in the book of Strachan and Read.
- What is the GRAIL text-mining algorithm?
  - ➔ Grail stands for gene relationships across implicated loci and is a technique that can be used to prioritize SNPs to follow up replication or functional studies. In a "fuzzy" region, Grail can clear out the noise by using a simple text mining algorithm to ascertain the degree of connectivity among the associated genes. It looks at the similarity of the vectors of words pulled from PubMed abstracts which mention the gene of interest.
- Is there a specific reason why the regression approach of Yang et al. (2010) has not been applied?
  - ➔ The Yang et al. Approach was very unconventional and new. This study has an entirely other design but could be replicated in the same way the study of Yang was conducted. However, the Allen study already explained a (in genetic terms) major proportion of the phenotypic variance. And SNP variants as Yang mentions in his article are not likely to be associated with a specific phenotype.

### Review of Essay:

Looking back at the essay now, 2 weeks later, I'm overall satisfied with the content. The definition, goal, stages, study design and complications are discussed providing a short but informative overview. However, some parts have been deleted (e.g. the parts concerning Linkage analysis). Furthermore, a small contribution has been written with regard to the recent success of meta-analysis in the discovery of SNPs and functional biomarkers. Moreover, missing heritability is addressed. Although I believe this essay contains the basic information that should be embodied in a short essay, many more items could be addressed. Such as the role of ascertainment bias, recombination, epistatic interactions, and methodological considerations. Due to the limitation in the potential length of the essay (it had to be short) these topics are only shortly discussed. Also, due to the high number of recent publications concerning GWA studies, this paper lacks scientific relevance and new information. Finally, if this essay was meant to get me updated in GWA studies, that goal has been fulfilled.