Published in final edited form as:

Birth Defects Res A Clin Mol Teratol. 2012 November ; 94(11): 908–911. doi:10.1002/bdra.23066.

Age of Onset and Effect Size in Genome-Wide Association Studies

A.J. Agopian, Lisa M. Eastcott, and Laura E. Mitchell*

Human Genetics Center, Division of Epidemiology, Human Genetics and Environmental Sciences, University of Texas School of Public Health, Houston, Texas

Abstract

BACKGROUND—Genome-wide association studies (GWAS) have identified many susceptibility loci for complex traits, but have not identified the majority of the genetic contribution to common diseases. We explored whether the magnitude of associations detected in GWAS and, therefore, the likelihood of detecting a significant association for a given sample size, is generally greater for childhood-onset traits (e.g., birth defects) than for traits with onset in adulthood.

METHODS—Data were obtained from the National Human Genome Research Institute Catalog of Published GWAS. Traits were categorized as having an average age of onset in childhood (<18 years, $n = 15$ traits), early adulthood (18–54 years, $n = 32$ traits), or late adulthood (≥ 55 years, $n = 31$ traits). The relationship between age of onset category and the magnitude of significant associations detected by GWAS was assessed using logistic regression.

RESULTS—Associations characterized by an odds ratio (OR) ≥ 1.5 were significantly more common for GWAS of childhood traits than for late adulthood-onset traits after adjustment for several covariates (adjusted OR, 2.55; 95% confidence interval, 1.37–4.73). Results in subgroup analyses using more stringent inclusion criteria (based on sample size, effect size, p value threshold for inclusion, and novel variant-trait associations) were similar.

CONCLUSIONS—These findings suggest that, on average, marker-trait associations detected in GWAS for traits with young onset may have a larger magnitude of effect than those for traits with adult onset. Therefore, GWAS for young-onset traits, such as birth defects, may be more likely than those for adult-onset traits to identify major genetic risk factors.

Keywords

age of onset; genome-wide association study; epidemiology; genetics; polymorphism; single nucleotide

INTRODUCTION

Genome-wide association studies (GWAS) have identified single nucleotide polymorphism (SNP)-trait associations (STAs) of varying magnitude. For example, the odds ratio for one statistically significant exfoliation glaucoma STA exceeds 20 (Thorleifsson et al., 2007), whereas the majority of significant STAs for type II diabetes are less than 1.2 (McCarthy and Zeggini, 2009). For birth defects, the magnitude of significant STAs identified by GWAS has been relatively strong (Birnbaum et al., 2009; Grant et al., 2009; Mangold et al., 2010; van der Zanden et al., 2011). These differences in STA effect sizes across GWAS likely reflect differences in sample sizes (i.e., larger samples have better power to detect smaller effects) as well as differences in the underlying genetic contribution to the analyzed traits. Although sample size issues for GWAS have received considerable attention (Burton et al., 2009; Spencer et al., 2009; Murcray et al., 2011), factors that would help to predict the extent to which a trait would be amenable to “genetic dissection” by GWAS have received relatively little consideration.

One characteristic of traits that might help to predict the genetic contribution to risk is age of onset. For traits with both early and late onset forms (e.g., breast cancer, Alzheimer disease), the relative contribution of genetic risk factors tends to be higher for the earlier than for the later onset form, presumably because affected individuals with late onset have had more time to experience environmental factors that influence disease risk. Using a similar rationale, it can be hypothesized that genetic risk factors will be relatively more important for traits that develop in early life (e.g., birth defects) than for traits that develop later in life, because the window during which risk might be influenced by environmental factors will be shorter for the former. If this hypothesis is correct, it would be expected that STAs detected in GWAS of conditions with childhood onset would generally be larger in magnitude than STAs detected in GWAS of conditions with adult onset. Because this would have practical implications for GWAS prioritization, we evaluated published GWAS from 2005 to 2010 to determine whether the magnitude of STAs detected in GWAS are associated with trait age of onset.

MATERIALS AND METHODS

Data Source and Exclusion Criteria

Data for this study were obtained from the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies (Hindorff et al., 2009, 2011). This catalog includes a regularly updated listing of STAs with $p < 10^{-5}$ from all published GWAS using at least 100,000 SNPs. The database includes p values and odds ratios (ORs) estimated from the largest sample size (e.g., combined analysis of discovery and replication samples) reported in each publication. For studies reporting multiple SNPs with $p < 10^{-5}$ in a given gene, only the SNP with the lowest p value is included in the catalog (Hindorff et al., 2011). Additional details of catalog development and curation are described elsewhere (Hindorff et al., 2009, 2011).

For this study, data were downloaded from the catalog on December 31, 2010 and included GWAS based on case-control, nested case-control, and family-based study designs

published through December 16, 2010. We included discrete traits (i.e., quantitative traits were excluded) that were directly related to health (e.g., traits such as hair color, and responses to a stimulus such as hepatitis C treatment were excluded) and were sufficiently defined in the catalog for age of onset to be clearly assigned (e.g., hematologic parameters and obesity were excluded). Results from meta-analyses (based on the title of the article) were not included, nor were SNPs for which an OR was not present in the catalog. In addition, to ensure comparability of STA ORs across studies, all studies were manually abstracted to determine whether the ORs reported in the catalog were estimated under a log-additive genetic model (e.g., versus a recessive model). Only SNPs evaluated under a log-additive model were included in analyses (data for 36/1114 STAs evaluated under a non-log-additive model were excluded).

Statistical Analysis

For this study, the magnitude of the association between a SNP and a trait was considered the “outcome” and the trait age of onset was considered the “exposure.” In initial analyses, the magnitude of the STAs was classified as being either moderate to strong (STA OR ≥ 1.5) or weak (STA OR, $1.0 < 1.5$), based on the estimated risk to heterozygotes (compared with the common homozygotes) under a log-additive model. This threshold was based on commonly accepted cut points for small effect sizes in GWAS (Manolio et al., 2008; Hindorff et al., 2009). Using average age of onset estimates derived from the literature, traits were classified as having a childhood (<18 years), early adulthood (18–54 years), or late adulthood (≥ 55 years) age of onset (or diagnosis) and as being more common (prevalence or incidence $>0.5\%$) or less common ($\leq 0.5\%$). When traits with early or late forms were labeled as such in the NHGRI Catalog (e.g., neonatal lupus vs. systemic lupus erythematosus), the different forms were considered to be different traits. Thus, each STA corresponded to a single age of onset category. The late adulthood onset category was used as the referent group for all analyses.

We determined the total number of eligible articles and the total number of significant STAs included in the catalog (defined by NHGRI as STAs with $p \leq 10^{-5}$; Hindorff et al., 2011) for each trait. In addition, we calculated the mean STA OR for each trait and for each age of onset category. The mean number of cases and mean minor allele frequency (MAF) across all associated SNPs were also determined for each of the three age-of-onset categories.

Crude logistic regression analyses were used to obtain estimates of unadjusted odds ratios for the association between age of onset and the magnitude of the STA. Multivariable logistic regression models, including variables that might confound the relationship between age of onset and the magnitude of the STA, were also developed. The multivariable models included parameters for: MAF (nine categories in 0.05 increments), total number of cases in the study sample (<2000 , 2000 – 3999 , 4000 – 5999 , ≥ 6000), total number of publications per trait, and trait prevalence or incidence ($>0.5\%$ vs. $\leq 0.5\%$). Tests for a linear increase in the magnitude of association across the three categories of age of onset were also performed. The logistic regression analyses included one observation per STA in the catalog. Thus, some publications were represented more than once in the analyses, because the publication reported STAs with $p \leq 10^{-5}$ for more than one gene. In addition, some specific STAs were

represented more than once in the main analyses, because the same STA was reported in more than one publication. However, analyses were repeated after removing replicated STAs (i.e., STAs in the catalog previously reported in another GWAS of the same trait), to compare results among novel STAs.

To assess the possibility that the results of the logistic regression analyses were driven by extreme STAs in a small number of studies, all analyses were repeated after excluding data from studies with less than 300 or more than 2000 cases and SNPs with $p < 10^{-100}$.

Furthermore, we repeated analyses comparing STAs with particularly strong effect sizes (i.e., >2.0) to those with weak effect sizes (i.e., <1.5). To consider a more conservative cutoff for genome-wide significance, analyses were also repeated among the subgroup of STAs with $p < 5 \times 10^{-8}$ (based on a Bonferroni correction for 1 million independent comparisons) in the total sample. All analyses were conducted using SAS (version 9.2, SAS Inc., Cary, NC).

RESULTS

Data were obtained from the NHGRI Catalog of Published Genome-Wide Association Studies for studies published through December 16, 2010. Data from 257 eligible articles, which in combination described associations between 1078 SNPs and 78 different phenotypic traits, were included in these analyses.

A summary of the 78 traits included in these analyses are presented in Supplemental Table 1. Data were available for 15 traits with onset in childhood (younger than 18 years); these data were derived from 34 articles and included information on 119 STAs. Data were also available for 32 traits with onset in early adulthood (18–54 years); these data were derived from 115 articles and included information on 606 STAs. In addition, data were available for 31 traits with onset in late adulthood (≥ 55 years); these data were derived from 108 articles and included information for 353 STAs. The references used to define trait age of onset and prevalence are also included in Supplemental Table 1.

The mean number of cases included in the GWAS was lowest for the traits with onset in childhood and highest for the traits with onset in late adulthood, whereas the mean MAF was similar across the three age-of-onset categories (Table 1). The proportion of STAs with moderate to strong effect sizes (i.e., STA OR ≥ 1.5) was 49.6% ($n = 59$) for childhood-onset traits, 20.0% ($n = 121$) for early-adult-onset traits, and 22.4% ($n = 79$) for late-adult-onset traits (Table 2). The corresponding figures for STAs with particularly strong effect sizes (i.e., STA OR ≥ 2.0) were 14.4% ($n = 17$), 6.4% ($n = 39$), and 9.4% ($n = 33$). Furthermore, the mean odds ratio for the STAs was 1.77 for traits with childhood onset, 1.42 for traits with onset in early adulthood, and 1.53 for traits onset in late adulthood onset (Supplemental Table 1).

In the unadjusted analysis of the association between age of onset and the magnitude of the STAs, moderate to strong STAs (i.e., STA OR ≥ 1.5) were significantly more likely for traits with childhood onset compared to those with onset in late adulthood (OR, 3.41; 95% CI, 2.20–5.29; Table 2). This association remained significant after adjustment for MAF,

total number of cases in the study sample, number of eligible GWAS publications (i.e., per trait), and trait prevalence (adjusted OR, 2.55; 95% CI, 1.37–4.73). The association between age of onset and the magnitude of the STAs was not significantly different for traits with early and late adult onset in either the unadjusted or the adjusted analyses (Table 2).

When analyses were repeated in the subset of STAs from studies with discovery samples of 300 to 2000 cases and excluding SNPs with $p < 10^{-100}$ ($n = 685$ SNPs), moderate to strong STAs (STA OR = 1.5) were significantly more common for traits with onset in childhood (adjusted OR, 3.37; 95% CI, 1.68–6.75) and early adulthood (adjusted OR, 1.78; 95% CI, 1.06–2.99) compared to traits with onset in late adulthood (Table 2). In this subset of the data, the test for a linear trend in the magnitude of the STA across the three age categories was statistically significant ($p < 0.0005$). Analyses were also repeated among the subset of novel STAs (i.e., STAs that had not been previously reported in the catalog for a given trait) and the subset of STAs with $p < 5 \times 10^{-8}$, and subgroup analyses were performed comparing STAs with particularly strong effect sizes (i.e., >2.0) versus weak effect sizes (i.e., <1.5). For all these analyses, results were similar to the main results described previously (data not shown).

DISCUSSION

This study explored published GWAS from 2005 to 2010 to determine whether the magnitude of STAs detected was associated with the age of onset of discrete traits. Our analyses of these published data suggest that GWAS of traits with early onset, such as birth defects, are more likely to detect moderate to strong STAs (OR = 1.5) than are studies of traits with onset in late adulthood. Although this finding is consistent with the hypothesis that genetic risk factors are relatively more important for traits with earlier onset than for traits with later onset, such a difference in the underlying genetic architecture of early and late onset conditions remains a matter of speculation. Nonetheless, this finding may have implications for prioritizing GWAS efforts. Specifically, our results suggest that, in general, GWAS of childhood-onset traits may be more successful at identifying major genetic risk factors than those focused on traits with onset in late adulthood. Furthermore, because study power is always a function of true effect size, our results suggest that the sample size required to detect STAs should be lower for GWAS of traits with onset in childhood than for traits with onset later in life.

The findings from this study should be interpreted with caution, in light of some potential limitations. One potential limitation was our assignment of conditions to discrete categories for age of onset. As many conditions develop over a fairly wide age range (e.g., schizophrenia, type 2 diabetes), the possibility of overlap between the age of onset categories exists. However, the effects of any such overlap should be minimal for the comparison between conditions with childhood (<18 years) and late adult (>55 years) onset. Furthermore, we assigned age of onset category for each trait based on the most common age of onset (e.g., Alzheimer disease was assigned to late adult onset), and we had to assume that the original GWAS studies used a reasonably homogeneous case definition and excluded cases that were suspected to have etiologies that varied (e.g. early-onset Alzheimer disease) from other cases. Another potential limitation is the range of conditions included in

each age of onset category. Our general conclusion, that GWAS of traits with onset in childhood are more likely to detect moderate to strong STAs (i.e., OR 1.5) compared to studies of traits with onset in late adulthood, might not be true for all diseases. For example, the mean STA ORs for some specific childhood-onset conditions were lower than those for some specific late adulthood-onset conditions (Supplemental Table 1). Determination of the extent to which our results are generalizable to specific conditions or disease categories will require the accumulation of GWAS data for a larger number of traits. Finally, although our main results were adjusted for several potential confounding factors, the possibility of residual confounding from measured and unmeasured potential confounders remains (e.g., differences in specific phenotype definitions, variance in the magnitude of STAs, study design, coverage of genotyping platforms). However, the results of our analyses were consistent across subanalyses designed to further control for potential confounders such as sample size (e.g., limiting analyses to discovery samples with $n = 300\text{--}2000$). Another potential limitation is that our analyses focused on main effects of single common variants and did not consider rare variants or pathway-based approaches.

Acknowledging these limitations, our analyses suggest that, in general, childhood-onset conditions, such as birth defects, may be better candidates for GWAS than are conditions with later onset. Additional studies will be required to determine the extent to which this generalization may hold within more homogenous disease categories and to determine whether there is a biologic basis for this observation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Grant sponsor: The National Institutes of Health; Grant numbers: R21 HL098844, R01 HD39195

References

- Birnbaum S, Ludwig KU, Reutter H, et al. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet.* 2009; 41:473–477. [PubMed: 19270707]
- Burton PR, Hansell AL, Fortier I, et al. Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol.* 2009; 38:263–273. [PubMed: 18676414]
- Grant SF, Wang K, Zhang H, et al. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J Pediatr.* 2009; 155:909–913. [PubMed: 19656524]
- Hindorf LA, MacArthur J, Wise A, et al. [Accessed March 16, 2011] A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies/
- Hindorf LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA.* 2009; 106:9362–9367. [PubMed: 19474294]
- Mangold E, Ludwig KU, Birnbaum S, et al. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet.* 2010; 42:24–26. [PubMed: 20023658]
- Manolio TA, Brooks LD, Collins FS, et al. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest.* 2008; 118:1590–1605. [PubMed: 18451988]

- McCarthy MI, Zeggini E. Genome-wide association studies in type 2 diabetes. *Curr Diab Rep.* 2009; 9:164–171. [PubMed: 19323962]
- Murcray CE, Lewinger JP, Conti DV, et al. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet Epidemiol.* 2011; 35:201–210. [PubMed: 21308767]
- Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* 2009; 5:e1000477. [PubMed: 19492015]
- Thorleifsson G, Magnusson KP, Sulem P, et al. Common sequence variants in the *LOXL1* gene confer susceptibility to exfoliation glaucoma. *Science.* 2007; 317:1397–1400. [PubMed: 17690259]
- van der Zanden LF, van Rooij IA, Feitz WF, et al. Common variants in *DGKK* are strongly associated with risk of hypospadias. *Nat Genet.* 2011; 43:48–50. [PubMed: 21113153]

Table 1

Characteristics of 257 GWAS Articles and 1078 SNPs by Trait Onset Age, 2005–2010

Characteristics	Mean (SD, range)		
	Less than 18 years (n = 34 articles, 119 SNPs)	18–54 years (n = 115 articles, 606 SNPs)	55 or more years (n = 108 articles, 353 SNPs)
Study characteristic			
No. of cases in total sample ^a	2760.4 (3455.4, 84–16506)	5988.8 (8315.4, 116–42542)	6694.0 (6109.6, 37–27036)
SNP characteristic			
Minor allele frequency	0.39 (0.23, 0.01–0.94)	0.44 (0.25, 0.01–0.97)	0.39 (0.24, 0.01–0.97)

^aTotal = discovery sample or the sum of the discovery and replication sample when the latter was available.

GWAS, genome-wide association study; SNP, single nucleotide polymorphism.

Table 2

Unadjusted and Adjusted Odds Ratios for the Association Between Trait Onset Age and STA^a Effect Size for 1078 SNPs From Previously Published GWAS, 2005–2010

	Weak association (OR < 1.5) n (%)	Moderate to strong association (OR 1.5) n (%)	Unadjusted OR (95% CI)	Adjusted OR ^b (95% CI)
Full sample (n = 1078 SNPs)				
Onset age (years)				
<18	60 (50.4)	59 (49.6)	3.41 (2.20–5.29)	2.55 (1.37–4.73)
18–54	485 (80.0)	121 (20.0)	0.87 (0.63–1.19)	1.19 (0.78–1.83)
55	274 (77.6)	79 (22.4)	1.0 (ref)	1.0 (ref)
Restricted sample^c (n= 685 SNPs)				
Onset age (years)				
<18	41 (51.9)	38 (48.1)	3.53 (2.06–6.03)	3.37 (1.68–6.75)
18–54	262 (73.60)	94 (26.40)	1.37 (0.93–2.01)	1.78 (1.06–2.99)
55	198 (79.20)	52 (20.80)	1.0 (ref)	1.0 (ref)

^aSNP-trait association.

^bAdjusted for minor allele frequency, total number of cases in study (e.g., discovery and replication samples combined), total number of eligible GWAS publications on condition, and trait prevalence or incidence (0.5 vs. 0.5%).

^cSubset of STAs from studies with discovery samples of 300–2000 cases and excluding SNPs with $p < 10^{-100}$.

STA, SNP trait association; SNP, single nucleotide polymorphism; GWAS, genome-wide association study; OR, odds ratio; CI, confidence interval.