Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.3389/fgene.2014.00401

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:13581255

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis

Bahram Namjou1,2*, Keith Marsolo2,3, Robert J. Carroll4, Joshua C. Denny4,5, Marylyn D. Ritchie6, Shefali S. Verma7, Todd Lingren2,3, Aleksy Porolito2,3, Beth L. Cobb1, Cassandra Perry7, Leah C. Kottyan1,2,8, Marc E. Rothenberg8, Susan D. Thompson1,2, Ingrid A. Holm9, Isaac S. Kohane10 and John B. Harley1,2,11

1 Center for Autoimmune Genomics and Etiology, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, USA
2 College of Medicine, University of Cincinnati, Cincinnati, OH, USA
3 Division of Biomedical Informatics, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, USA
4 Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA
5 Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA
6 Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA, USA
7 Division of Genetics and Genomics, The Pennsylvania State University, Philadelphia, PA, USA
8 Division of Allergy and Immunology, Department of Pediatrics, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, USA
9 Division of Genetics and Genomics, Department of Pediatrics, The Marston Center for Orphan Disease Research, Harvard Medical School, Boston Children’s Hospital, Boston, MA, USA
10 Children’s Hospital Informatics Program, Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA
11 U.S. Department of Veterans Affairs Medical Center, Cincinnati, OH, USA

Objective: We report the first pediatric specific Phenome-Wide Association Study (PheWAS) using electronic medical records (EMRs). Given the early success of PheWAS in adult populations, we investigated the feasibility of this approach in pediatric cohorts in which associations between a previously known genetic variant and a wide range of clinical or physiological traits were evaluated. Although computationally intensive, this approach has potential to reveal disease mechanistic relationships between a variant and a network of phenotypes.

Method: Data on 5049 samples of European ancestry were obtained from the EMRs of two large academic centers in five different genotyped cohorts. Recently, these samples have undergone whole genome imputation. After standard quality controls, removing missing data and outliers based on principal components analyses (PCA), 4268 samples were used for the PheWAS study. We scanned for associations between 2476 single-nucleotide polymorphisms (SNP) with available genotyping data from previously published GWAS studies and 539 EMR-derived phenotypes. The false discovery rate was calculated and, for any new PheWAS findings, a permutation approach (with up to 1,000,000 trials) was implemented.

Results: This PheWAS found a variety of common variants (MAF > 10%) with prior GWAS associations in our pediatric cohorts including Juvenile Rheumatoid Arthritis (JRA), Asthma, Autism and Pervasive Developmental Disorder (PDD) and Type 1 Diabetes with a false discovery rate < 0.05 and power of study above 80%. In addition, several new PheWAS findings were identified including a cluster of association near the NDFIP1 gene for mental retardation (best SNP rs10057309, \( p = 4.33 \times 10^{-7}, \text{OR} = 1.70, 95\% \text{CI} = 1.38 - 2.09 \)); association near PLCL1 gene for developmental delays and speech disorder (best SNP rs1595825, \( p = 1.13 \times 10^{-8}, \text{OR} = 0.65(0.57 - 0.76) \)); a cluster of associations in the IL5-IL13 region with Eosinophilic Esophagitis (EoE) [best at rs12653750, \( p = 3.03 \times 10^{-9}, \text{OR} = 1.73 95\% \text{CI} = (1.44 - 2.07) \)], previously implicated in asthma, allergy, and eosinophilia; and association of variants in GCKR and JAZF1 with allergic rhinitis in our pediatric cohorts [best SNP rs780093, \( p = 2.18 \times 10^{-5}, \text{OR} = 1.39, 95\% \text{CI} = (1.19 - 1.61) \]), previously demonstrated in metabolic disease and diabetes in adults.

Conclusion: The PheWAS approach with re-mapping ICD-9 structured codes for our European-origin pediatric cohorts, as with the previous adult studies, finds many previously reported associations as well as presents the discovery of associations with potentially important clinical implications.

Keywords: PheWAS, ICD-9 code, genetic polymorphism
INTRODUCTION
Phenome-wide association study (PheWAS) is a relatively new genomic approach to link clinical conditions with published variants (Denny et al., 2010). The concept, although not new, was originally applied to genomic research by the eMERGE (electronic MEDical Records and GEnomics) network, which is in a unique position to access tens of thousands of Electronic Medical Records (EMR) linked to ICD-9 codes in structured data. Multiple eMERGE PheWAS results have been published that primarily address adult cohorts (Denny et al., 2011, 2013). The phenotypic data used in PheWAS may include ICD-9 codes, epidemiologic data in health surveys, biomarkers, intermediate or quantitative traits (Pendergrass et al., 2011, 2013; Neuraz et al., 2013; Liao et al., 2014). By virtue of this inclusive approach, new hypotheses may be generated that provide insight into genetic architecture of complex traits. Challenges with PheWAS include multiple test corrections across the thousands of phenotypes tested and auto-correlation of some of the phenotypes. Nevertheless, novel robust insights have resulted from PheWAS, for example, genetic association findings with heart rate variability are notable (Ritchie et al., 2013).

PheWAS combines multiple phenotypes from previous GWAS, and identify common SNPs affecting different traits. In this study, we used this approach to evaluate whether known GWAS variants identified in adult diseases can be also identified in children using two EMR-linked pediatric datasets from eMERGE. PheWAS in pediatrics is particularly important because it not only assesses the effect of early age of onset on many established adult-GWAS loci, but also may provide insights into how a primary phenotype during child development develops into one or more diseases in adulthood. A priori, there are several reasons that in principle might make a pediatric PheWAS more challenging. These include the change in heritability with age for several traits (St Pourcain et al., 2014), the flux in the recommendations for pediatric monitoring for traits that are routinely measured in adults (Gidding, 1993; Klein et al., 2010) and the use of cross-sectional standardization rather than longitudinal standardization of developmental traits such as height (Tiiusla and Kantero, 1971).

To determine whether robust association signals would be present in the context of these challenges, we conducted the first PheWAS study in pediatrics on our available samples. We successfully translated 93,724 specific ICD-9 diagnostic codes into 1402 distinct PheWAS code groups and 14 major disease concept paths and evaluated 2481 previously published variants. After quality control, only 2476 genetic variants were analyzed in 539 diseases in the two pediatric sites. Finally we replicated 24 genetic variants and identified 14 new possible associations confirming our hypothesis. Our primary results highlight the utility of an EMR-based PheWAS approach as a new line of investigation for discovery of genotype-phenotype associations in pediatrics.

MATERIALS AND METHODS
STUDY SUBJECTS
Protocols for this study were approved by the Institutional Review Boards (IRBs) at the institutions where participants were recruited. All study participants provided written consent prior to study enrolment; consent forms were obtained at each location under IRB guidelines. Children and teens, aged through 19 years old were included. The EMR-linked pedi atric emerge cohorts consist of 4560 subjects from Cincinnati Children's Hospital Medical Center (CCHMC) and 1000 subjects from Boston Children's Hospital (BCH). Only those self-reported to have European ancestry were selected for this study (Table 1).

SNP PRIORITIZATION
We limit our investigation to particular genetic variants: First, we obtained the list of all previously published SNPs from different public domain databases including The National Human Genome Research Institute (NHGRI) catalog of published Genome-Wide Association Studies (http://www.genome.gov/gwastudies), Genetic Association of Complex Diseases and Disorders (GAD, http://geneticassociationdb.nih.gov), the UCSC Genome Browser database (UCSC, http://genome.ucsc.edu/), Online Mendelian Inheritance in Man (OMIM, http://www.omim.org/), and PharmGKB (pharmgkb, https://www.pharmgkb.org). After linking this collection to PubMed reference numbers, only those with at least one reported of positive associations were selected regardless of the previously observed p values or number of publications. In addition, all downloaded databases were current at the time of this submission. From the filtered variants, 2476 variants were available and assessed in our clean, post-imputation genotyping dataset for analysis.

GENOTYPING AND STATISTICAL ANALYSES
High throughput SNP genotyping was carried out previously in CCHMC and BCH using Illumina™ or Affymetrix™ platforms, as previously described (Namjou et al., 2013). Quality control (QC) of the data was performed before imputation. In each genotyped cohort, standard quality control criteria were met and single nucleotide polymorphisms (SNPs) were removed if (a) >5% of the genotyping data was missing, (b) out of Hardy-Weinberg equilibrium (HWED, p < 0.001) in controls, or a minor allele frequency (MAF) <1%. Samples with call rate <98% were excluded.

Recently all eMERGE cohorts have also undergone whole genome imputation. The details of these procedures are available in this issue of Frontiers in Genetics (Setia et al., 2014). Briefly, the imputation pipeline was implemented using IMPUTE2 program and the publicly available 1000-Genomes Project as the reference haplotype panel composed of 1092 samples (release version 2 from March 2012 of the 1000 Genomes Project Phase I, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521) (Howie et al., 2011). The eMERGE imputed data provided to us were already filtered, i.e., imputed data with a threshold of 0.90 for the genotype posterior probability and with a IMPUTE2 info score > 0.7 (Howie et al., 2011). Principle component analysis (PCA) performed to identify outliers and hidden population structure using EIGENSTRAT (Price et al., 2006). The first two principle components explained most of the variance and were retained and used as covariates during the association analysis in order to adjust for population stratification. In addition, 14 outlier samples were removed. To illustrate the overall inflation rate a phenotype with sufficient number of cases and
controls has been selected (autism) and the inflation of $\lambda = 1.03$ was obtained.

Next, from our prioritized SNP list mentioned above, 2481 variants were available. Five of these SNPs had a site-specific effect with either CCHMC or BCH ($p < 10^{-5}$ for the difference between sites) and were removed from final analyses. For each phenotype, logistic regression was performed between cases and control adjusted for two principal components using PLINK (Purcell et al., 2007). To investigate whether either the phenotype or the genotype has an effect on the outcome variable, we perform phenotypic and genotypic conditional analyses, controlling for the effect of a specific SNP or phenotype. After pruning of highly correlated SNPs ($r^2 > 0.5$), we used false discovery rate (FDR) methods to correct for multiple testing using the Benjamini–Hochberg procedure implemented in PLINK (Purcell et al., 2007). As a result of LD pruning 1828 independent variants were available [PLINK (Purcell et al., 2007)]. We also report previous associations using PheWAS and present current PheWAS map and PheWAS script written in R is available [http://phewisecatalog.org, (Carroll et al., 2014)]. In this study, subgroups of European cases with more than 20 samples were selected for PheWAS association study (539 subgroups) and the available published SNPs that passed quality controls were evaluated. The case cohorts for the two phenotypes of Juvenile Idiopathic Arthritis (JIA) and Eosinophilic Esophagitis (EoE) have been included as parts of larger phenotype specific studies (Rothenberg et al., 2010; Thompson et al., 2012; Hinks et al., 2013). The origin of all case records is presented in Table 1. In this study, Juvenile Onset Rheumatoid Arthritis (JRA) is identified by ICD-9 codes and designated as JRA; when the criteria for Juvenile Idiopathic Arthritis (JIA) were applied in the studies of others (Thompson et al., 2012), then this phenotype was referred to as JIA.

RESULTS
In this study only European ancestry was included in the analyses to avoid potential bias induced by ancestry. The demographic distribution of the European ancestry population under study (Table 2) had 93,724 specific ICD-9 diagnostic codes representing 1402 distinct PheWAS code groups and 14 major disease concept paths. The frequencies of concept path hierarchy of the ontology (Figure 1) show the neuropsychiatric concept path as the most frequent and neoplastic and infection paths as the least frequent.

Replication of existing associations using PheWAS
We compared SNPs with previous GWAS-reports and present association findings (FDR-q < 0.05) after corrected for population stratification and standard quality control (Table 2).

First, for the two phenotypes of JRA and EoE samples overlap largely with those previously reported phenotype specific GWAS study (Rothenberg et al., 2010; Thompson et al., 2012; Kottyan et al., 2014). We reproduced the major findings of those publications using different methodology. For JRA, association with PPVN22 is a consistent finding. As expected, we replicated a previous report of association of PPVN22 at non-synonymous coding SNP rs2476601 with this phenotype and with the same direction

---

Table 1 | The demographic distribution of the European ancestry population (CCHMC-BCH).

<table>
<thead>
<tr>
<th>Cohort names</th>
<th>#Europeans</th>
<th>M/F</th>
<th>Mean age (95%CI)</th>
<th>Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCHMC**</td>
<td>1228</td>
<td>758/470</td>
<td>7.32(7.03–7.62)</td>
<td>Illumina-610</td>
</tr>
<tr>
<td>Cytogenetics</td>
<td>609</td>
<td>373/236</td>
<td>218.73(216.73–216.73)</td>
<td>Illumina-Omni-1</td>
</tr>
<tr>
<td>EoE†</td>
<td>543</td>
<td>394/149</td>
<td>12.27(11.70–12.67)</td>
<td>Illumina-Omni-5</td>
</tr>
<tr>
<td>Cincinnati- control cohorts</td>
<td>673</td>
<td>329/344</td>
<td>13.50(13.25–13.84)</td>
<td>Illumina-Omni-6</td>
</tr>
<tr>
<td>Total</td>
<td>4268</td>
<td>2403/1865</td>
<td>11.52(11.16–11.91)</td>
<td></td>
</tr>
</tbody>
</table>

*BCH, Boston Children’s Hospital; **CCHMC, Cincinnati Children’s Hospital Medical center; †, Eosinophilic Esophagitis (EoE) cohorts; ‡, Juvenile Idiopathic Arthritis (JIA) cohorts (JIA). The details of platforms used have been described elsewhere (Namjou et al., 2013).
<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP Position</th>
<th>Gene</th>
<th>Minor/Control</th>
<th>OR</th>
<th>Description</th>
<th>Case/Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2476601 114377568</td>
<td>PTPN22</td>
<td>A0</td>
<td>0.87 (1.46–2.41)</td>
<td>JRA</td>
<td>272/3412</td>
</tr>
<tr>
<td>1</td>
<td>rs2476601 114377568</td>
<td>PTPN22</td>
<td>A0</td>
<td>2.78E-05</td>
<td>Thyroiditis</td>
<td>23/3571</td>
</tr>
<tr>
<td>1</td>
<td>rs2476601 114377568</td>
<td>PTPN22</td>
<td>A0</td>
<td>0.007</td>
<td>T1DM</td>
<td>47/3609</td>
</tr>
<tr>
<td>1</td>
<td>rs6679677 114303808</td>
<td>PTPN22</td>
<td>A0</td>
<td>1.92 (1.49–2.47)</td>
<td>JRA</td>
<td>272/3412</td>
</tr>
<tr>
<td>1</td>
<td>rs6679677 114303808</td>
<td>PTPN22</td>
<td>A0</td>
<td>2.00E-05</td>
<td>Thyroiditis</td>
<td>23/3571</td>
</tr>
<tr>
<td>1</td>
<td>rs6679677 114303808</td>
<td>PTPN22</td>
<td>A0</td>
<td>0.005</td>
<td>T1DM</td>
<td>47/3609</td>
</tr>
<tr>
<td>2</td>
<td>rs3771180 102953617</td>
<td>IL1RL1</td>
<td>T0</td>
<td>0.46 (1.19–1.96)</td>
<td>EoE or Food Allergy</td>
<td>599/2346</td>
</tr>
<tr>
<td>2</td>
<td>rs7574865 191964633</td>
<td>STAT4</td>
<td>A0</td>
<td>1.80 (1.22–2.49)</td>
<td>Eosinophilia</td>
<td>250/3344</td>
</tr>
<tr>
<td>3</td>
<td>rs78122814 85200034</td>
<td>CADM2</td>
<td>T0</td>
<td>0.96 (1.11–1.22)</td>
<td>Autism</td>
<td>601/1840</td>
</tr>
<tr>
<td>5</td>
<td>rs3806932 110405675</td>
<td>TSLP</td>
<td>G0</td>
<td>0.69 (0.59–0.80)</td>
<td>EoE</td>
<td>446/2586</td>
</tr>
<tr>
<td>5</td>
<td>rs272889 131665378</td>
<td>SLC22A4</td>
<td>A0</td>
<td>1.45 (1.22–1.84)</td>
<td>Atopic Dermatitis</td>
<td>298/3031</td>
</tr>
<tr>
<td>5</td>
<td>rs12653750 131971902</td>
<td>IL5-IL13</td>
<td>T0</td>
<td>0.41 (0.32–0.53)</td>
<td>JRA</td>
<td>272/3412</td>
</tr>
<tr>
<td>5</td>
<td>rs2516051 32569691</td>
<td>HLA-DRB1</td>
<td>A0</td>
<td>0.41 (0.32–0.53)</td>
<td>JRA</td>
<td>272/3412</td>
</tr>
<tr>
<td>5</td>
<td>rs2516049 32570400</td>
<td>HLA-DRB1</td>
<td>C0</td>
<td>0.36 (0.27–0.46)</td>
<td>JRA</td>
<td>272/3412</td>
</tr>
<tr>
<td>5</td>
<td>rs622137 32569852</td>
<td>HLA-DRB1</td>
<td>A0</td>
<td>0.41 (0.32–0.53)</td>
<td>JRA</td>
<td>272/3412</td>
</tr>
<tr>
<td>5</td>
<td>rs9388489 126698719</td>
<td>CENPW</td>
<td>G0</td>
<td>2.46 (1.58–3.34)</td>
<td>EoE or Food Allergy</td>
<td>599/2346</td>
</tr>
<tr>
<td>5</td>
<td>rs1490388 126835655</td>
<td>CENPW</td>
<td>T0</td>
<td>2.37 (1.95–2.86)</td>
<td>EoE</td>
<td>250/3344</td>
</tr>
<tr>
<td>9</td>
<td>rs7850258 100549013</td>
<td>FOXE1</td>
<td>A0</td>
<td>0.35 (0.15–0.78)</td>
<td>Thyroiditis</td>
<td>23/3571</td>
</tr>
<tr>
<td>9</td>
<td>rs1443338 140083416</td>
<td>MAI-RD2</td>
<td>T0</td>
<td>2.00 (1.29–3.83)</td>
<td>EoE or Food Allergy</td>
<td>599/2346</td>
</tr>
<tr>
<td>16</td>
<td>rs1292470 11887783</td>
<td>CLEC16A</td>
<td>A0</td>
<td>0.74 (0.64–0.84)</td>
<td>Asthma</td>
<td>499/33175</td>
</tr>
<tr>
<td>17</td>
<td>rs2290734 39096240</td>
<td>ITGB3</td>
<td>C0</td>
<td>1.53 (1.23–1.92)</td>
<td>Autism</td>
<td>601/1840</td>
</tr>
</tbody>
</table>

False discovery rate (FDR) < 0.05 was set for the threshold of significance. The calculated odds ratio was based on minor allele frequency and the coded alleles were shown. All positions were based on NCBI build 37 (NS not significant). The p-values and q-values are ordered based on chromosome and position.
of allele frequency, \((p = 9.10 \times 10^{-7}, \text{OR} = 1.87, 95\% \text{CI} 1.46 - 2.40)\). The SNP in proxy (rs6679677, \(r^2 = 1\)) also produced a similar result (Table 2). In our cohorts, variants in \(PTPN22\) are also associated with thyroiditis as well as Type 1 diabetes mellitus (T1DM), consistent with previous reports and despite low sample size (Table 2) (Plenge et al., 2007; Todd et al., 2007; Lee et al., 2011). From these three known associations of \(PTPN22\), i.e., JRA, T1DM, and thyroiditis, the largest magnitude of the association is with pediatric onset thyroiditis (Table 2, \(\text{OR} = 3.52\) 95%CI 1.84 – 6.75).

For JRA, multiple loci in the HLA region were also associated at the level of \(p < 10^{-12}\) including rs477515 and rs2516049 near \(HLA-DRB1\) (Table 2). Of note, the size effect of HLA related SNPs, were highest for those with coexisting uveitis (best SNP rs477515, \(\text{OR} = 6.5, 95\% \text{CI} = 2.73 – 15.68\) for the risk allele, Table 2). In addition, for JRA, another previously published association (rs12411988 in \(REP3\)) was also found and with the same size effect as previously described (\(\text{OR} = 1.53\)) (Table 2) (Thompson et al., 2012).

Furthermore, with regard to EoE traits, we also replicated previous major finding of association of SNP rs3806932 located at the vicinity of the \(TSLP\) gene at 3q22 region \((p = 5.59 \times 10^{-7}, \text{OR} = 0.69\) (95%CI 0.59 – 0.80]) in these cohorts (Table 2) (Rothenberg et al., 2010; Kottyan et al., 2014).

For asthma, the best PheWAS results were detected at 17q21 which includes \(GSDMB\) and has been previously reported to be associated specifically with childhood onset Asthma (Verlaan et al., 2009). In fact, the best associated SNP rs8067378 in our cohorts \((p = 3.13 \times 10^{-6}, \text{OR} = 1.37\) (1.19 – 1.57]), tags the asthma associated haplotype in which the allele-specific expression analyses for this haplotype has previously shown strong association with Asthma risk (Verlaan et al., 2009). There is strong support for this association from a cluster of variants in this neighborhood (Figure 2A).

The minor allele (T) of the intronic SNP rs7903146 in \(TCF7L2\) is one of the larger magnitude and more frequently identified associations in Type 2 diabetes mellitus (T2DM) and hyperlipidemia in many adult GWAS studies (Lyssenko et al., 2007; Huertas-Vazquez et al., 2008). In fact, the best PheWAS trait in our cohorts at this variant was also related to T2DM and hyperlipidemia as well, although our sample size was small. In this family of ICD-9 codes the best suggestive result was obtained for an abnormal glucose test with \((p = 0.001, \text{OR} = 2.00\) (95%CI 1.29 – 3.08)) (Table 2).

Specifically, for T1DM, in addition to the positive association with \(PTPN22\) mentioned above, additional published loci were confirmed and with relatively larger effect sizes (\(\text{OR} > 2\)) including known HLA-SNP rs660895 \((p = 7.85 \times 10^{-7}, \text{OR} = 2.73\) (95%CI 1.80 – 4.13)], as well as variants near \(CENPW\) that previously have been reported for this trait (Table 2) (Barrett et al., 2009).

**Other effects**

Several loci previously associated with autism and pervasive developmental disorders (PDD) (GWAS or copy number variations reports) including those at \(MACROD2\), \(ITGB3\), \(CADM2\), and \(GRIK2\) (Jamain et al., 2002; Weiss et al., 2006; Thomas et al., 2008; Anney et al., 2010) also provided evidence of association in our cohorts for these traits (Table 2). Variants in the \(FOXE1\) gene that have been previously associated with primary hypothyroidism and thyroiditis in adult eMERGE cohorts (Denny et al., 2011), produced a trend of association and consistent in

---

**FIGURE 1 | Frequency and distribution of 14 major ontology concept path categories from CCHMC/BCH European pediatric cohorts.**
directionality with thyroiditis in our pediatric cohorts despite low sample size (Table 2). No gene-gene interaction was evident between PTPN22 and FOXE1 for hypothyroidism in these data. Rs7574865 is a SNP in the third intron of the STAT4 that has been associated with SLE and related autoimmune diseases (Namjou et al., 2009). In these cohorts, pediatric onset lupus was under-represented (less than 20 cases), however, suggestive associations with wheeze and asthma were detected \( p = 0.004 \), OR 1.46 (95%CI = 1.11 – 1.92) (Table 2) with the same direction of the difference in allele frequency previously observed in autoimmune

---

**FIGURE 2** | Association results and signals contributing to Asthma, Eosinophilic Esophagitis, Mental Retardation, and Developmental Delays. SNPs are plotted by position in a 0.2 Mb window against association signals \((-\log_{10} P\)-value). For each trait, the most significant SNP is highlighted. Estimated recombination rates (from HapMap) are plotted in cyan to reflect the local LD structure. The SNPs surrounding the most significant SNP are color-coded to reflect their LD with identified SNP (taken from pairwise \( r^2 \) values from the HapMap CEU database, www.hapmap.org). Regional plots were generated using LocusZoom (http://csg.sph.umich.edu/locuszoom). (A) Cluster of the association effect for asthma at 17q21 near the gasdermin-B (GSDMB) gene. (B) Association signal for Eosinophilic Esophagitis at 5q31 (IL5-IL13 cluster region). (C) Cluster of association near the NDFIP1 gene for Mental Retardation traits. (D) Plot of association effects in the PLCL1 region for Developmental Delays-Speech Disorders.
traits. This possible association has also been reported in another study (Pykäläinen et al., 2005). Of note, in contrast to rheumatoid arthritis, the STAT4 association effect was weak for JRA in our cohorts (effect size = 1.12, p = 0.17). GWAS studies have linked Inflammatory Bowel disease (IBD) to a number of IL-23 pathway genes, in particular IL23R. The well-known coding variant in the IL23 receptor (rs11209026) also showed a trend toward association with IBD in our cohorts with the same allelic direction but due to low sample size (31 cases) it did not reach significance (FDR-q > 0.05) (Li et al., 2010) (data not shown).

**Novel findings from this PheWAS**

A number of potentially novel associations remained significant after the permutation procedure to assess the probability of the observed distribution with beta > 0.8 FDR-q < 0.05 (Table 3). Variants in the Glucokinase Regulator gene (GCKR) have been previously implicated in metabolic disease, diabetes and hypertriglyceridemia in adults (Bi et al., 2010; Onuma et al., 2010) and were mostly associated with allergic rhinitis in our pediatric cohorts [best SNP rs780093 p = 2.18 × 10^-5, $p_{(\text{perm})} = 8.06 \times 10^{-5}$, OR = 1.39, 95%CI = (1.19 - 1.61)] (Table 3), while no significant association was found for diabetes. Indeed, conditional analyses, controlling for diabetes related traits suggest that this is an independent effect ($p$-conditional = $6.75 \times 10^{-5}$). Another major regulatory locus for diabetes in adults, TCF7L2, was also associated with allergic rhinitis in our cohorts (Table 3) even after controlling for diabetes ($p$-conditional = $8.46 \times 10^{-5}$, for rs1635852). No significant gene-gene interaction was detected between these two loci or TCF7L2.

Variants in a cytokine cluster of the IL5-IL13 region, which is known to be associated with Asthma, Allergy, Atopic Dermatitis (AD) and Eosinophilia, produced a cluster of association with EoE in our cohorts [best SNP rs12653750, $p = 3.03 \times 10^{-9}$, $p_{(\text{perm})} = 1.00 \times 10^{-6}$, OR = 1.73 (1.44 - 2.07)] (Bottema et al., 2008; Granada et al., 2012). There is a cluster of significant variants in this neighborhood of chromosome 5 (5q31) associated with EoE (Figure 2B). In our cohorts, weaker associations can be detected for all allergy-related phenotypes with the association effect being the most impressive [p = 9.74 \times 10^{-5} (Table 2)]. However, conditional analyses and controlling for Asthma and Eosinophilia suggest that an independent effect still exists for EoE at this locus using EMR data (conditional $p = 9.74 \times 10^{-5}$). Moreover, no long distance linkage disequilibrium between rs3806932 in TSPI gene at 5q22 and rs20541 was detected in this population ($r^2 = 0.0002, D^2 = 0.02$).

We also observed association with AD within this cytokine cluster consistent with previous reports (Paternoster et al., 2011). However, the best associated SNP for AD (rs272889) was located at SLC22A4 in our population (Table 2). These two variants, rs272889 and rs12653750, were separated by more than 300kb with low linkage disequilibrium ($r^2 < 0.1$). A residual effect still exists for AD and rs272889 after controlling for EoE status or the rs12653750 variant that suggests a distinct effect ($p$-conditional = 0.002). Noteworthy, with regard to AD, another reported SNP (rs2897442) downstream of this cluster at KIF3A gene produced only a suggestive association ($p = 0.005$) in our cohort (data not shown).

---

**Table 3 | Novel PheWAS findings in CCHMC/BCH pediatric cohorts.**

<table>
<thead>
<tr>
<th>Description</th>
<th>Case/Control</th>
<th>Chr</th>
<th>SNP</th>
<th>Position</th>
<th>Gene</th>
<th>Minor allele</th>
<th>Case Control</th>
<th>OR</th>
<th>p value</th>
<th>p-permute</th>
<th>Case needed*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergic rhinitis</td>
<td>408/2754</td>
<td>2</td>
<td>rs780093</td>
<td>27730940</td>
<td>GCKR</td>
<td>T</td>
<td>0.47</td>
<td>2.94E-05</td>
<td>702.0E-06</td>
<td>1.21E-02</td>
<td>250</td>
</tr>
<tr>
<td>Allergic rhinitis</td>
<td>408/2754</td>
<td>2</td>
<td>rs780093</td>
<td>27742237</td>
<td>GCKR</td>
<td>T</td>
<td>0.47</td>
<td>2.94E-05</td>
<td>702.0E-06</td>
<td>1.21E-02</td>
<td>250</td>
</tr>
<tr>
<td>Allergic rhinitis</td>
<td>408/2754</td>
<td>2</td>
<td>rs780093</td>
<td>27818053</td>
<td>GCKR</td>
<td>C</td>
<td>0.43</td>
<td>6.59E-05</td>
<td>2.70E-05</td>
<td>1.00E-04</td>
<td>200</td>
</tr>
<tr>
<td>Allergic rhinitis</td>
<td>408/2754</td>
<td>7</td>
<td>rs8647</td>
<td>28180556</td>
<td>JAZF1</td>
<td>C</td>
<td>0.43</td>
<td>2.73E-05</td>
<td>1.00E-04</td>
<td>1.14E-01</td>
<td>100</td>
</tr>
<tr>
<td>Eosinophilic Esophagitis</td>
<td>446/2586</td>
<td>5</td>
<td>rs4143832</td>
<td>131862977</td>
<td>IL5-IL13</td>
<td>T</td>
<td>0.24</td>
<td>1.98E-06</td>
<td>4.70E-06</td>
<td>1.70E-05</td>
<td>200</td>
</tr>
<tr>
<td>Eosinophilic Esophagitis</td>
<td>446/2586</td>
<td>5</td>
<td>rs12653750</td>
<td>131971902</td>
<td>IL5-IL13</td>
<td>C</td>
<td>0.24</td>
<td>1.98E-06</td>
<td>4.70E-06</td>
<td>1.70E-05</td>
<td>200</td>
</tr>
<tr>
<td>Eosinophilic Esophagitis</td>
<td>446/2586</td>
<td>5</td>
<td>rs20541</td>
<td>131995964</td>
<td>IL5-IL13</td>
<td>T</td>
<td>0.24</td>
<td>1.98E-06</td>
<td>4.70E-06</td>
<td>1.70E-05</td>
<td>200</td>
</tr>
<tr>
<td>Mental retardation</td>
<td>297/1840</td>
<td>5</td>
<td>rs11167764</td>
<td>141479065</td>
<td>NDFIP1</td>
<td>A</td>
<td>0.29</td>
<td>1.33E-07</td>
<td>4.33E-07</td>
<td>2.00E-05</td>
<td>150</td>
</tr>
<tr>
<td>Suppurative otitis media</td>
<td>362/3082</td>
<td>1</td>
<td>rs10801047</td>
<td>191559356</td>
<td>NRXN3</td>
<td>C</td>
<td>0.66</td>
<td>4.76E-06</td>
<td>1.10E-04</td>
<td>1.10E-04</td>
<td>100</td>
</tr>
</tbody>
</table>

*P (permute): empirical permutation p values after case and control labels are permuted randomly (up to 1,000,000). All results were at the level of FDR-q < 0.05.

**Cases needed** refers to the estimated number of cases needed to achieve 80% power to detect an association at alpha = 0.05 given the identified odds ratio and the MAF in this population.
Because of the pleotropic effects between EoE and other allergy related traits, in addition to conditional analyses, we also found possible synergistic effects. One of the closely related phenotypes with EoE is the presence of food allergy. When we combined these two as a subgroup, two additional effects were identified. One cluster was in \textit{IL1RL1} that was previously associated with the related phenotype, i.e., allergy and asthma (best SNP rs3771180, \(p = 5.71 \times 10^{-5}\), \textit{Table 2}, Torgerson et al., 2011) and another was in \textit{CLEC16A}, previously associated with different autoimmune diseases [best SNP rs12942729, \(p = 3.34 \times 10^{-6}\) \textit{Table 2}], (Mells et al., 2011)] and was reported as a suggestive effect in recent GWAS study for EoE (Kottyan et al., 2014).

Variants near \textit{RGS} cluster of genes on chromosome 1, previously reported to be associated with IBD and other autoimmune diseases (Hunt et al., 2008; Esposito et al., 2010), were associated with susceptibility to infection, in particular suppurative otitis media [best SNP rs10801047, \(p = 1.61 \times 10^{-6}\), \(p_{\text{perm}} = 2.00 \times 10^{-6}\), \(OR = 1.77 95\%CI = 1.398 – 2.24\)].

New association signals have been detected near the \textit{NDFIP1} gene for mental retardation related traits. Variants near this gene that is expressed mostly in brain, were previously reported to be associated with IBD through an unknown mechanism and with a risk effect for major allele (SNP = rs11167764) (Franke et al., 2010). Instead, we found a risk effect for the minor allele (best SNP rs10057309, \(p = 4.33 \times 10^{-7}\), \(p_{\text{perm}} = 2.00 \times 10^{-6}\), \(OR = 1.702, 95\%CI = 1.38 – 2.09\) \textit{Table 3}). Similarly, cerebral palsy, which is linked to mental retardation, was also associated with this variant (\(p = 9.00 \times 10^{-4}\)). However, conditional analyses controlling for cerebral palsy suggest an independent effect for overall mental retardation (conditional \(p = 8.00 \times 10^{-4}\)). Furthermore, excluding the small number of samples with known chromosomal abnormalities (\(N < 40\)) did not affect this result. The overall cluster effect in this neighborhood for mental retardation bolsters the suspicion that an association is found here (\textit{Figure 2C}).

Additionally, for developmental delays of speech and language, a novel signal effect was detected in the \textit{PLCL1} gene at chromosome 2 [best SNP rs1595825, \(p = 1.13 \times 10^{-8}\), \(OR = 0.65 (0.57 – 0.76)\) \textit{Figure 2D, Table 3}]. Weaker associations (0.01 > \(p > 0.00001\)) were also detected for related neurologic phenotypes including abnormal movement, lack of coordination and epilepsy at this locus (data not shown).

\textit{NRXN3} polymorphisms that have been previously reported to be associated with substance dependence (Docampo et al., 2012), smoking behavior and attention related problems (Stoltenberg et al., 2011), were associated with depression in our pediatric cohorts \textit{Table 3}. Noteworthy, the major allele of our reported SNP (rs7141420) has been linked to obesity in adult cohorts (Berndt et al., 2013), while we found association with the minor allele for depression [\(p = 4.76 \times 10^{-5}\), \(OR = 1.78 (1.34 – 2.34)\), \textit{Table 3}]. Furthermore, rare micro-deletions in this gene were previously reported for Autism case reports but these rare variants are not available to assess in our genotyped cohorts (Vaags et al., 2012).

**DISCUSSION**

This first pediatric PheWAS finds 38 associations, 24 previously known phenotype-genotype associations in a pediatric population using EMR-linked eMERGE databases and identified 14 new possible associations at beta > 0.8 and FDR-q < 0.05. From analysis performed on EMR-linked data from 4268 European individuals, we successfully confirmed several major effects for phenotypes with moderate to large sample size, in particular for Asthma, Autism, and neurodevelopmental disease as well as several effects for Type 1 and Type 2 Diabetes (T1DM, T2DM) and Thyroiditis. Almost all of the significant phenotype associations were with common variants (MAF > 10\% \textit{Tables 2, 3}). In addition, we compared and verified the consistency of allele frequency of reported markers among cohorts, sample collection sites and with CEU-Hapmap data. Considering a desired power of 0.8, for variants at the fixed allele frequency of 10\% and size effect of 1.5 or above, 200 cases are sufficient to detect association at an alpha level of 0.05. Indeed, we have surpassed this level for most of our reported traits. In addition, for all reported phenotypes the control sample was at least two or three times larger than cases \textit{Tables 2, 3}. Importantly, since our control samples for each trait are an EMR-derived population and not healthy individuals, this large number of control samples provides minor allele frequencies consistent with hapmap-CEU frequencies for all of our reported variants.

The results for JRA and EoE depend upon previously published studies of these phenotypes. While the case samples are mostly identical, the control samples were substantially different. Consequently, we cannot refer to these particular findings as constituting confirmation and yet our results and different methodology support the previous reports.

In addition, we also identified several novel PheWAS findings for pediatric traits in particular for Allergic Rhinitis, Otitis Media, EoE, Mental Retardation, and Developmental Delays all with sufficient power (beta > 0.8) \textit{Table 3, Figures 2B–D}). This study, however, is underpowered to make discoveries for rare variants or uncommon traits. The power to detect a finding in PheWAS is determined by many factors, including sample size, risk allele frequency, effect size, model of inheritance, the effect of environment and the prevalence of a phenotype within the population.

Similar to previous studies, we also observed pleiotropy for a number of loci in particular \textit{PTPN22} for JRA, T1DM, and Thyroiditis, IL5 for Eosinophilia, Asthma, and EoE and \textit{NDFIP1} for Mental Retardation traits and Cerebral Palsy. These pleotropic effects are specifically expected to be due to underlying biologic correlations. On the other hand, we rarely observed simultaneous robust associations with multiple unrelated phenotypes that had sufficient power. Furthermore, one of the advantages of PheWAS studies is the ability to control the granularity of a database with regard to related phenotypes. For example, by combining two related phenotypes such as uveitis with JRA or food allergy with EoE, we were able to evaluate new subgroups and identify new loci responsible for shared underlying pathways that otherwise cannot be detected or require much larger sample sizes. Further studies with larger sample sizes would be useful to test and perhaps corroborate these findings.

Association of Allergic Rhinitis with loci responsible for diabetes in adults (\textit{GCKR-JAZF1}) may highlight a shared underlying mechanism. In fact, the connection between allergy and diabetes...
has been previously suggested in humans but cannot be explained by the Th1/Th2 paradigm (Dales et al., 2005). Moreover, in animal experiments, treating mice with mast cell-stabilizing agents reduced diabetes manifestations (Liu et al., 2009). It is also possible that in our pediatric cohorts we have under-diagnosed children who are diagnosed with diabetes which would appear in a later stage of development. In fact, GCKR is an inhibitor of glucokinase (GCK), a gene responsible for the autosomal dominant form of T2DM that usually develops later in life and in adulthood. Of note, neither of these two loci showed significant association with Body Mass Index (BMI) in our previous report with these data nor has the obesity link been established in adult studies (Namjou et al., 2013).

The novel association of a cytokine cluster in the IL5-IL13 region for the EoE trait is particularly interesting since anti-IL5 monoclonal antibodies have been recommended as a novel therapeutic agent for EoE and other eosinophilia–related traits (Corren, 2012). In general, both IL5 and IL13 play a major role for regulation of maturation, recruitment, and survival of eosinophils and the variant reported here has been previously associated with other allergic–related traits and with the same direction of allele frequency difference (Bottema et al., 2008; Esposito et al., 2010). The link between susceptibility to eosinophilic forms of T2DM that usually develops later in life and in adulthood. Of note, neither of these two loci showed significant association with Body Mass Index (BMI) in our previous report with these data nor has the obesity link been established in adult studies (Namjou et al., 2013).

Moreover, combining subgroups of patients with food allergy and EoE revealed two new loci that may explain shared etiology. Indeed, the connection between allergy and interleukin 1 receptor-like-1 (IL1R1) is already known (Torgerson et al., 2011). The ligand for IL1R1, IL-33, is a potent eosinophil activator (Bouffi et al., 2013). Interestingly, there is also a report of association of CLEC16A variants with allergy in large analysis with more than 50,000 subjects from 23andMe Inc. (Hinds et al., 2013). G-type lectin domain family 16, also known as CLEC16A, is mostly associated with autoimmune related traits and is highly expressed in B lymphocytes and natural killer cells. The molecular and cellular functions of CLEC16A are currently under investigation.

Our conditional analyses suggest an independent effect at the SLC22A4 gene for Atopic Dermatitis. This solute carrier family gene is predominantly expressed in CD14 cells and has an important role for elimination of many endogenous small organic cations as well as a wide array of drugs and environmental toxins. The associated SNP, rs272889, has been previously shown to be correlated with blood metabolite concentration (Suhre et al., 2011). Other variants in this gene were associated with IBD and Crohn's disease as well (Feng et al., 2009). Of note, a key substrate of this transporter is ergothioneine, a natural antioxidant, which Mammalia acquire exclusively from their food. Ergothioneine is a powerful antioxidant though its precise physiological purpose remains unclear.

Asthma is associated at the 17q21 in our cohorts (Figure 1). The best associated SNP, rs8067378, is known to function as a cis-regulatory variant that correlates with expression of the GSDMB gene (Verlaan et al., 2009). Variants in GSDMB have been shown to determine multiple asthma related phenotypes specifically in childhood asthma including associations with lung function and disease severity (Tulah et al., 2013). These gasdermin-family genes are implicated in the regulation of apoptosis mostly in epithelial cells and have also been linked to cancer; however, their actual function with respect to disease association remains unknown. The associated variants in this cluster are suspected to be regulatory SNPs that govern the transcriptional activity of at least three nearby genes (ZBPBP2, GSDMB, and ORMDL3) (Verlaan et al., 2009).

We confirmed several loci responsible for Autism and Pervasive Developmental Disease including MACROD2, ITGB3, CADM2, and GRIK2. ITGB3 has been known as a quantitative trait locus (QTL) for whole blood serotonin levels (Weiss et al., 2004, 2006). Serotonin is a monoamine neurotransmitter that has long been implicated in the etiology of Autism. In fact, about 30 percent of patients with autism have abnormal blood serotonin levels (Weiss et al., 2004). Similarly, GRIK2 is an ionotropic glutamate receptor associated with autism (Cook, 1990; Cook et al., 1997). CADM2 is a member of the synaptic cell adhesion molecule with roles in early postnatal development of the central nervous system (Thomas et al., 2008). The function of MACROD2 (previously c20orf133) is still largely unknown. For Autism that is more commonly seen in males, we found no significant gender effect for these loci.

Association of variants in the neighborhood of RGS cluster genes with suppressive otitis media is another novel finding. SNPs in this region have been previously linked to celiac disease, multiple sclerosis and other autoimmune diseases (Hunt et al., 2008; Esposito et al., 2010). The link between susceptibility to infection and autoimmunity has been long suggested given the fact that the level and regulation of RGS proteins in lymphocytes also significantly impact lymphocyte migration and function. In our pediatric cohort the number of patients with celiac disease was small (n = 23) and the association was not detected. Interestingly, one of the major risk variant for celiac disease, rs13151961 (KIAA1109), as well as known HLA variants, produced a tread toward association for celiac disease but did not pass the FDR threshold (data not shown).

Finally we also detected a novel association between mental retardation and the NDFIP1 gene (Figure 2C, Table 3). Of note, no effect was detected with Autism at this locus. Indeed, the
only other effect observed in this region was related to Cerebral Palsy (p = 9.00 × 10^{-4}) and, as mentioned above, an independent effect exists for Mental Retardation. The PheWAS code for mental retardation includes ICD-9 codes for mild, moderate and profound degrees of retardation as well as not-otherwise-specified (MR-NOS). Indeed, an additive correlation can also be detected when we score these subgroups according to severity excluding the MR-NOS subgroup (p = 3.00 × 10^{-4}). Larger sample size is necessary to fully elucidate this interesting effect.

The Nedd4 family-interacting protein 1 (Ndfip1) is an adaptor protein for the Nedd4 family of E3 ubiquitin ligases important for axon and dendrite development. In fact, cerebral atrophy is one of the main findings in Ndfip1 KO mice (Hammond et al., 2014).

Another neurodevelopmental association effect was observed in the vicinity of the Phospholipase C-Like 1 (PLCL1, PRIP-1) gene for overall Developmental Delays-Speech and Language Disorder (Table 3, Figure 2D). This gene which is expressed predominantly in brain, regulates the turnover of GABA-receptors, contributes to the maintenance of GABA-mediated synaptic inhibition, and has been implicated in several pathologies in animal models and human including epilepsy, bone density and cancer (Liu et al., 2008; Zhu et al., 2012). Finally, we also detected a link between Neuroxin-3 and early onset depression in this study (Table 3).

In fact, this gene has a major role in synaptic plasticity and function in the nervous system as a receptor and cell adhesion molecule.

In summary, by using the PheWAS approach and re-mapping the ICD-9 codes on our European ancestry pediatric cohorts we have been able to verify and confirm a variety of previously reported associations as well as discover new effects that potentially have clinical implications. Similar to adult PheWAS studies, our data also support the importance of this approach in pediatrics. We replicated known phenotype-genotype associations in a pediatric population using these EMR-linked eMERGE databases, and also noted a number of new possible associations that warrant additional study, especially including the relationship of PLCL1 to speech and language development and IL5-IL13 to EOe. Some of the limitations to the current PheWAS map include the fact that current map does not take into account specified (MR-NOS). Indeed, an additive correlation can also be detected when we score these subgroups according to severity as not-otherwise-specified (MR-NOS).

ACKNOWLEDGMENTS

We are grateful to the individuals who participated in this study. We thank the genotyping core facilities in both academic centers (CCHMC, BCH) and our colleagues who facilitated the genotyping and recruitment of subjects.

This work was supported by a grant from the National Human Genomic Research Institute: U101HG006828 with other NIH support (R37 A1024717, P01 A1083194, U19 A1066738, and P01 AR049084), the US Department of Veterans Affairs, the Campaign Urging Research For Eosinophilic Diseases (CURED) Foundation, as well as the Food Allergy Research Education (FARE) Foundation.

REFERENCES


November 2014 | Volume 5 | Article 401 | 11


Conflict of Interest Statement: The Guest Associate Editor Mariza De Andrade declares that, despite having collaborated with authors Bahram Namjou, Joshua C. Denny, Leah C. Kottyan, Marylyn D. Ritchie, and Shafali S. Verma, the review process was handled objectively and no conflict of interest exists. The Review Editor Andrew Skol declares that, despite having collaborated with author John B. Harley, the review process was handled objectively and no conflict of interest exists. Marc E. Rothenberg is a consultant for Immune Pharmaceuticals and has an equity interest. Marc E. Rothenberg has a royalty interest in reslizumab being developed by Teva Pharmaceuticals. Marc E. Rothenberg, John B. Harley, and Leah C. Kottyan are co-inventors of a patent application, being submitted by CCHMC, concerning the genetics of EoE. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 May 2014; accepted: 31 October 2014; published online: 18 November 2014.


This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

Copyright © 2014 Namjou, Marsolo, Denny, Ritchie, Verma, Lingren, Porollo, Cobb, Perry, Kottyan, Rothenberg, Thompson, Holm, Kohane and Harley. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.