Published in final edited form as:

Pac Symp Biocomput. 2013 ; : 147–158.

NEXT-GENERATION ANALYSIS OF CATARACTS: DETERMINING KNOWLEDGE DRIVEN GENE-GENE INTERACTIONS USING BIOFILTER, AND GENE-ENVIRONMENT INTERACTIONS USING THE PHENX TOOLKIT*

Sarah A. Pendergrass,

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 503 Wartik Lab, University Park, PA 16802, USA sap29@psu.edu

Shefali S. Verma,

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab, University Park, PA 16802, USA szs14@psu.edu

Emily R. Holzinger,

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab, University Park, PA 16802, USA, Emily.R.Holzinger@vanderbilt.edu

Carrie B. Moore,

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab, University Park, PA 16802, USA ccb12@psu.edu

John Wallace,

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab, University Park, PA 16802, USA, jrw32@psu.edu

Scott M. Dudek,

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab, University Park, PA 16802, USA, sud23@psu.edu

Wayne Huggins,

RTI International, Research Triangle Park, NC, USA, whuggins@rti.org

Terrie Kitchner,

Marshfield Clinic, Marshfield, WI, USA, Kitchner.Terrie@mcrf.mfldclin.edu

Carol Waudby,

Marshfield Clinic, Marshfield, WI, USA, WAUDBY.CAROL@mcrf.mfldclin.edu

Richard Berg,

Marshfield Clinic, Marshfield, WI, USA, Berg.Richard@mcrf.mfldclin.edu

Catherine A. McCarty, and

Essential Rural Health, Duluth, MN, USA, CMcCarty@eirh.org

Marylyn D. Ritchie

*This work supported by the following grants: U19 HL065962S, R01 LM010040, U01 HG006389

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab, University Park, PA 16802, USA, Marylyn.ritchie@psu.edu

Abstract

Investigating the association between biobank derived genomic data and the information of linked electronic health records (EHRs) is an emerging area of research for dissecting the architecture of complex human traits, where cases and controls for study are defined through the use of electronic phenotyping algorithms deployed in large EHR systems. For our study, 2580 cataract cases and 1367 controls were identified within the Marshfield Personalized Medicine Research Project (PMRP) Biobank and linked EHR, which is a member of the NHGRI-funded electronic Medical Records and Genomics (eMERGE) Network. Our goal was to explore potential gene-gene and gene-environment interactions within these data for 529,431 single nucleotide polymorphisms (SNPs) with minor allele frequency > 1%, in order to explore higher level associations with cataract risk beyond investigations of single SNP-phenotype associations. To build our SNP-SNP interaction models we utilized a prior-knowledge driven filtering method called Biofilter to minimize the multiple testing burden of exploring the vast array of interaction models possible from our extensive number of SNPs. Using the Biofilter, we developed 57,376 prior-knowledge directed SNP-SNP models to test for association with cataract status. We selected models that required 6 sources of external domain knowledge. We identified 5 statistically significant models with an interaction term with p-value < 0.05, as well as an overall model with p-value < 0.05 associated with cataract status. We also conducted gene-environment interaction analyses for all GWAS SNPs and a set of environmental factors from the PhenX Toolkit: smoking, UV exposure, and alcohol use; these environmental factors have been previously associated with the formation of cataracts. We found a total of 288 models that exhibit an interaction term with a p-value 1×10^{-4} associated with cataract status. Our results show these approaches enable advanced searches for epistasis and gene-environment interactions beyond GWAS, and that the EHR based approach provides an additional source of data for seeking these advanced explanatory models of the etiology of complex disease/outcome such as cataracts.

1. Introduction

DNA biobanks coupled to electronic health records (EHR) have become a valuable resource for investigating the genetic architecture of complex traits, as EHR contain a wide array of medical information including billing codes and clinical laboratory measurements, often yielding a large sample size. Through carefully defining phenotypes, and using deployable algorithms that combine multiple sources of information in the EHR, cases and controls can be defined for association studies, such as defining age-related cataract cases and controls [1, 2]. The Marshfield Personalized Medicine Research Project Biobank (Marshfield PMRP) and linked EHR, used for the study described herein, is one such resource [3]. The Marshfield PMRP is a member of the NHGRI-funded electronic Medical Records and Genomics (eMERGE) Network, a network of similar Biobanks coupled with EHR based data [4].

Cataracts are a leading cause of blindness globally [5], and are believed to arise from a combination of age, environmental factors, and heritable factors [6]. Thus, understanding the genetic etiology of cataracts, coupled with the effect of environment as a modifier, could have a profound impact on human health. For our study, algorithms proven for age-related cataract case identification [2] were deployed in the Marshfield PMRP EHR to identify 2580 cataract cases and 1367 controls, with further study details presented in Table 1. A total of 529,431 single nucleotide polymorphisms (SNPs) were available after PMRP genotyping

coupled with quality control filtering and selection for SNPs with a minor allele frequency > 1%.

Single SNP-phenotype associations are a dominant study design used in most genome-wide association studies (GWAS), however, more complex models that include interactions may more accurately describe the relationship between genetic variation and complex outcomes. Investigating all gene by gene (GxG), and in extension, all SNP by SNP (SNPxSNP) pairwise models is possible depending on the number of SNPs that have been genotyped. Unfortunately, the multiple hypothesis testing burden and risk of Type I error is inflated when investigating all pairwise models. A different approach can be used, utilizing prior biological knowledge methods directing model development. Thus, to investigate more complex models beyond single SNP-Phenotype associations for the Marshfield PMRP cataract dataset, we used the prior knowledge accessible through Biofilter 1.0 (a new implementation of Biofilter after the original description in [7]) to direct the investigation of pairwise GxG interaction models based on the following resources: the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8], Reactome [9], Gene Ontology (GO) [10], the Protein families database (Pfam) [11], NetPath [12], Biological General Repository for Interaction Datasets (BioGrid) [13], and the Molecular INTERaction Database (MINT) [14]. Using the Biofilter, we developed 57,376 prior-knowledge directed SNPxSNP models to test for association with cataract status.

In addition, for this study we investigated gene-environment interactions (GxE), as there are clearly known environmental exposures that increase cataract risk, and when incorporated into analyses, may provide new models for the contribution of both environment and genetic architecture to cataracts. The Marshfield PMRP collected standardized Phenotypes and eXposures (PhenX) measures as a member of the PhenX Real-world, Implementation, SharingING (PhenX RISING) project. PhenX has the goal of defining standard phenotypic measures through a framework of measurement protocols via a web-based toolkit [15]. Environmental exposures such as smoking, sun exposure, and alcohol use, have been associated with increased cataract rates [16]. Thus we used 12 PhenX defined environmental exposures to investigate GxE interactions for the Marshfield PMRP cataract data focused on smoking, UV exposure, and alcohol use measures.

Through integrating EHR data, advanced bioinformatics tools, and PhenX, we can pursue advanced searches for epistasis and gene-environment interactions in genome-wide studies of common disease.

2. Methods

2.1. Marshfield EHR and Age-Related Cataract Case Identification

The Marshfield PMRP is a population based biobank with ~20,000 subjects, aged 18 years and older, enrolled in the Marshfield Clinic healthcare system in central Wisconsin [3]. DNA, plasma, and serum samples are collected at the time the enrollee completes a written informed consent document, with allowance for ongoing access to the linked medical records. PMRP participants also complete questionnaires, including responses regarding smoking history, occupation, and diet.

To identify cataract surgery cases aged 50 years and older within the PMRP, Current Procedural Terminology (CPT) codes in the Marshfield Clinic EHR were used. A research coordinator manually abstracted additional information to identify the eye affected, the type and severity of the cataract, and the level of visual acuity prior to the cataract surgery. This was also done to remove any cases with non-age related cataracts.

To identify individuals with diagnosed cataracts but without surgery, and to identify the type of cataract, International classification of diseases, 9th revision (ICD-9) codes and CPT codes were used, coupled with Natural Language Processing (NLP) and Intelligent Character Recognition (ICR) of free-text in the EHR. NLP and exclusion criteria were used to identify individuals with congenital and traumatic cataracts for omission from the study. Further details of the identification of cataract cases and controls and the efficacy of the EHR defined phenotyping can be found in Waudby et al., 2011 [2].

All total, the procedures used on the EHR identified 2580 cataract cases and 1367 controls in the Marshfield PMRP data.

2.2. Genotyping

The eMERGE network and the Center for Inherited Disease Research (CIDR) at Johns Hopkins university performed the genotyping of the Marshfield PMRP samples, using the Illumina Human660W-Quad v1 A platform with total of 560,635 SNPs and 96,731 intensity-only probes. Bead Studio version 3.3.7 was used by CIDR for the genotyping calls. The total cohort genotyped included 3947 samples from the Marshfield PMRP, 21 blind duplicates, and 85 HapMap controls. The HapMap concordance rate was 99.8% and the blind duplicate reproducibility rate was 99.99%. For quality control and data cleaning the eMERGE quality control (QC) pipeline developed by the eMERGE Genomics Working Group [17] was used. Any SNPs with a minor allele frequency > 1%, SNP call rate > 99%, Sample Call Rate > 99% were used in further analysis. After QC and allele frequency filtering using PLINK [18], a total of 529,431 SNPs were used for further analysis.

2.3. PhenX

The standardized phenotypic and environmental consensus measures for Phenotypes and eXposures (PhenX) [15] were used to capture the environmental variables used in this study. The PhenX Toolkit (<https://www.phenxtoolkit.org/>) offers high-quality, well-established, standard measures of phenotypes and exposures for use in epidemiological studies.

The Marshfield PRMP is part of the PhenX RISING consortium, which is comprised of seven groups funded by the National Human Genome Research Institute (NHGRI) and the Office of Behavioral and Social Sciences Research (OBSSR) to incorporate PhenX (<https://www.phenxtoolkit.org/>) measures into existing population-based genomic studies.

For this initiative, Marshfield PRMP subjects with GWAS data who were alive with known, non-institutionalized addresses and who had given consent for re-contact were mailed a 32-page self-administered questionnaire that contained 35 PhenX measures across a range of phenotypic domains including alcohol and tobacco use questions (McCarty et al. 2012, *in preparation*).

For this study, we considered 12 of these measures, shown in Table 2.

2.4. BioFilter 1.0

For the SNPxSNP analysis, Biofilter 1.0 was used. Biofilter has been upgraded from the initial Biofilter 0.5 [7], with the addition of more data sources, improved the handling of data, and the development of an eternal database for prior knowledge called the Library of Knowledge Integration (LOKI). Biofilter 1.0 and LOKI are freely available for non-commercial research institutions. For full details see: <http://ritchielab.psu.edu/ritchielab/software>

Biofilter 1.0 utilizes prior biological knowledge through accessing the data of several publically available biological information databases, all compiled within the LOKI database developed specifically for Biofilter. The data sources selected for Biofilter contain information on networks, connections, and/or pathways that establish relationships between genes and gene products. Biofilter is a “gene based” approach, thus all the region information (such as genes) and position information (such as SNPs) are mapped to genes within LOKI.

The following sources that are compiled within LOKI were used for the Biofilter model building: the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8], Reactome [9], Gene Ontology (GO) [10], the Protein families database (Pfam) [11], and NetPath [12], Biological General Repository for Interaction Datasets (BioGrid) [13], and the Molecular INTERaction Database (MINT) [14]. The database source used in LOKI solely for the purpose of mapping SNPs to genes is the National Center for Biotechnology (NCBI) dbSNP [19] database.

The following process was used within Biofilter 1.0 to develop the SNPxSNP models used in prior knowledge directed association testing. Figure 1 shows a simplified example of how the Biofilter 1.0 model generation process works. First, the input list of SNPs are mapped to genes within Biofilter. Next, comprehensive pairs of genes that are all terminal leaves of the graph for Pathway 1 in Source 1, and Pathway 2 in Source 1 are generated, only for genes that contain SNPs in the input list of SNPs.

Implication scores are used in Biofilter to give each pairwise model a “score” indicating how many sources have that connected pair of genes represented, the higher the implication score, the more sources have indicated a connection between a pair of genes. The implication index is a measure of the number of data sources providing evidence of an interaction, a sum of the number of data sources supporting each of the two genes and the connection between them. In the case of our simplified example, for Genes 1–5, that all contain SNPs within the input list, the following pairwise Gene-Gene models would result, each with an implication score of 1:

Gene1 – Gene 2
 Gene1 – Gene 3
 Gene 2 – Gene 3
 Gene 4 – Gene 5

This process continues through all other sources used for Biofilter. Each time a pairwise combination of genes is found in another source (such as the pair Gene1-Gene2), the implication score for that pairwise model will be increased by 1. Lastly, the G-G models are broken into all pairwise combinations of SNPs across the genes, *within P1 or P2*. The SNP-SNP models would look like the following:

SNP1-SNP3
 SNP1-SNP4
 SNP1-SNP5
 SNP2-SNP3
 SNP2-SNP4
 SNP2-SNP5
 SNP3-SNP5
 SNP3-SNP4
 SNP6-SNP7
 SNP6-SNP8

This same process was used within Biofilter 1.0 to develop the SNPxSNP models used for our prior knowledge directed association testing. First, the 529,431 SNPs were mapped to their corresponding genes. Next, the genes corresponding to the SNPs of the dataset were

mapped to the gene-relationship graphs for each LOKI source used. After this mapping process, gene pairs were exhaustively generated for each occurrence of two genes within a single pathway and single source. Implication scores were calculated for the pairwise models. After the gene-gene models were developed in Biofilter, the models were divided into exhaustive SNP-SNP pairs for association testing.

Table 3 indicates the number of models that were found at each implication score cutoff. An implication index cutoff of 4 actually incorporates all possible pairwise models for all SNPs we had for this study, a total of 603,032 models. We found an implication score cutoff of 6 resulted in a balance between a large group of models for exploration (57,376 models), but still maintained a very computationally feasible set of associations to investigate, limiting our type 1 error rate more than using all exhaustive pairs of SNP-SNPs or some of the less stringent implication score cutoffs. With a requirement for an implication index of 6, as we had in this study, the gene-gene relationship or known interaction had to be found in nearly all of the data sources we used within LOKI.

2.5. Statistical Analysis

For the SNPxSNP models generated through the use of Biofilter, PLATO [20] was used to determine the significance of the interaction term (via a t-test), and the significance of the overall model (via an F test). The full model was: $\text{SNP1} + \text{SNP2} + \text{SNP1} * \text{SNP2}$, for all of the pairwise sets of SNPs generated by Biofilter with an implication index of 6.

For the GxE (SNPxE) models, the full model was: $\text{SNP1} + \text{ENV1} + \text{SNP1} * \text{ENV1}$, for all the possible unique SNPxE pairs, from the set of 529,431 SNPs and the PhenX variables described earlier in methods. Again, the outcome was case control status for cataracts. PLINK [18] was used, and results were maintained for further inspection that had an P-interaction term < 0.05 . The GGPlot2 [21] package in R was used for Figure 2.

3. Results

3.1. GxE Results

Figure 2 shows a Manhattan plot of the association results for the PhenX GxE models that had interaction p-values $< 1 \times 10^{-2}$, a total of 288 models exhibited an interaction term with a p-value $< 1 \times 10^{-4}$ associated with cataract status. The top five GxE interaction results for each PhenX measure are also presented in Table 4, sorted by chromosome to highlight results similar across SNPs and regions for multiple PhenX measures. The measurement “PX030802 Former Smoker Smoking 6Month” a survey question asking “Have you ever smoked cigarettes every day for at least 6 months?” with the SNP rs2058131, near the genes *ZNF471* and *ZFP28* on Chromosome 19, that had an association interaction term p-value of 2.72×10^{-7} , was the most significant interaction term p-value found when compared to the other 12 PhenX measurements we used in our GxE analysis.

3.2. GxG Results

The top Biofilter 1.0 derived GxG models are presented in Table 5. A total of 5 models had both a p-value for the interaction term and the overall model p-value < 0.05 . A total of 7 genes were in the five models. Of the five models, the most significant was for a model with *EGF*, which is epidermal growth factor, and *EGFR*, which codes for the cell surface receptor that binds to epidermal growth factor.

4. Discussion

The results presented herein are an exploration of the use of multiple novel approaches for investigating gene and phenotype associations within EHR based data. We performed an analysis with PhenX derived measures, seeking GxE interaction models for the Marshfield Cataract data set. The majority of the significant interactions were found for smoking related measures. We did find some highly correlated PhenX measures with significant interactions for SNPs within similar regions, such as the results on chromosome 6 for SNPs rs6899720 and rs12528760, for smoking related phenotypes. Through searches in the NCBI catalog [22], as well as the National Center for Biotechnology (NCBI) dbSNP [19], these two SNPs, as well the SNP in our most significant GxE model, did not show previous GWA level significant associations for any phenotypes.

We also performed an exploratory analysis with Biofilter 1.0, an updated and improved implementation of the originally published Biofilter. The results are intriguing, and provide the basis for hypotheses that can be investigated further, highlighting how Biofilter results have a biological context that provide additional information for resulting models. For instance, the most significant model contained *EGF* and *EGFR*, which are known to have a biological interaction as the gene product of *EGFR* is a receptor for the gene product of *EGF*. Epidermal growth factor is found in human tears [23], and ocular effects have been found after the administration of EGFR inhibitors administered to patients [24]. Interestingly, three of the models that passed our significance cutoff contained two of the same genes, *FYN*, a member of the protein-tyrosine kinase oncogene family implicated in cell growth, and *DOCK1*, dedicator of cytokinesis 1. These models as a whole implicate genes related to cell growth, the cell cycle, and epidermal growth.

We are currently developing Biofilter 2.0 which will include additional database sources and allow for the use of other position and region based information beyond SNPs and genes, such as copy number variation (CNV) data, evolutionary conserved regions, and regulatory regions, allowing users to incorporate additional sources of prior knowledge as well as utilize other sources of genetic variation measurement data, with a more user-friendly interface.

Our results provide more complex models for an association between genetic variation and cataract outcome, moving beyond the more standard SNP-phenotype associations. The models found we intend to investigate further and warrant additional investigation of the environment and genetic variables contributing to these more complex models. These bioinformatics approaches can be used with other datasets, to expand the investigation of the relationship between genetic architecture and phenotypic outcome. With these approaches that consider the complexity of the data and harness the power of novel bioinformatics tools, we will elucidate the missing heritability of complex traits.

References

1. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *Journal of the American Medical Informatics Association : JAMIA*. 2012; 19:225–234. [PubMed: 22319176]
2. Waudby CJ, Berg RL, Linneman JG, Rasmussen LV, Peissig PL, et al. Cataract research using electronic health records. *BMC Ophthalmol*. 2011; 11:32. [PubMed: 22078460]
3. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Medicine*. 2005; 2:49–79.

4. Pathak J, Pan H, Wang J, Kashyap S, Schad PA, et al. Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects. *AMIA Summits Transl Sci Proc.* 2011; 2011:41–45. [PubMed: 22211178]
5. Michael R, Bron AJ. The ageing lens and cataract: a model of normal and pathological ageing. *Philos Trans R Soc Lond B Biol Sci.* 2011; 366:1278–1292. [PubMed: 21402586]
6. Hammond CJ, Duncan DD, Snieder H, de Lange M, West SK, et al. The heritability of age-related cortical cataract: the twin eye study. *Invest Ophthalmol Vis Sci.* 2001; 42:601–605. [PubMed: 11222516]
7. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput.* 2009:368–379. [PubMed: 19209715]
8. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 1999; 27:29–34. [PubMed: 9847135]
9. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 2009; 37:D619–D622. [PubMed: 18981052]
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics.* 2000; 25:25–29. [PubMed: 10802651]
11. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins.* 1997; 28:405–420. [PubMed: 9223186]
12. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* 2010; 11:R3. [PubMed: 20067622]
13. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* 2011; 39:D698–D704. [PubMed: 21071413]
14. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 2012; 40:D857–D861. [PubMed: 22096227]
15. Stover PJ, Harlan WR, Hammond JA, Hendershot T, Hamilton CM. PhenX: a toolkit for interdisciplinary genetics research. *Curr Opin Lipidol.* 2010; 21:136–140. [PubMed: 20154612]
16. Abraham AG, Condon NG, West Gower E. The new epidemiology of cataract. *Ophthalmol Clin North Am.* 2006; 19:415–425. [PubMed: 17067897]
17. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet.* 2011 Chapter 1: Unit1 19.
18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics.* 2007; 81:559–575. [PubMed: 17701901]
19. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001; 29:308–311. [PubMed: 11125122]
20. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, et al. Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. *Pac Symp Biocomput.* 2010:315–326. [PubMed: 19908384]
21. Wickham, H. *ggplot2: elegant graphics for data analysis.* New York: Springer; 2009.
22. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America.* 2009; 106:9362–9367. [PubMed: 19474294]
23. Ohashi Y, Motokura M, Kinoshita Y, Mano T, Watanabe H, et al. Presence of epidermal growth factor in human tears. *Invest Ophthalmol Vis Sci.* 1989; 30:1879–1882. [PubMed: 2788149]
24. Saint-Jean A, Sainz de la Maza M, Morral M, Torras J, Quintana R, et al. Ocular Adverse Events of Systemic Inhibitors of the Epidermal Growth Factor Receptor: Report of 5 Cases. *Ophthalmology.* 2012

Figure 1.
Simplified model for one Biofilter 1.0 database source with 2 pathways, 5 genes, and 8 SNPs

Figure 2.

Manhattan plot of the association results for the GxE interaction models. Displayed are the results for the 10 PhenX measures that had interaction p-values $< 1 \times 10^{-2}$, two PhenX measures did not have an interaction p-value less than 1×10^{-2}

Table 1**Marshfield Cataract Study Description**

Age	50
Ancestry	European American
Total Sample Size	3947
# Controls	1367
# Cases	2580
After QC & Allele Frequency Cutoff	
# Controls	1364
# Cases	2576
% Women Controls	59
% Men Controls	41
% Women Cases	59
% Men Cases	41

Table 2

The PhenX measures used for this study

PhenX Measure	Survey Question
PX030301 Alcohol 30Day Frequency	During the past 30 days, on how many days did you drink one or more drinks of an alcoholic beverage?
PX030301 Alcohol 30Day Quantity	During the past 30 days, how many drinks did you usually have each day?
PX030602 Cigarette Smoking 100	Have you smoked at least 100 cigarettes in your entire life?
PX030602 Cigarette Smoking Current	Do you now smoke cigarettes every day, some days, or not at all?
PX030602 Cigarette Smoking Everyday 6Month	Have you EVER smoked cigarettes EVERY DAY for at least 6 months?
PX030802 Everyday Smoker Quantity 1Day	On the average, about how many cigarettes do you now smoke each day?
PX030802 Someday Smoker Days 1Month	On how many of the past 30 days did you smoke cigarettes?
PX030802 Someday Smoker Quantity 1Day	On the average, on those days, how many cigarettes did you usually smoke each day?
PX030802 Former Smoker Smoking 6Month	Have you EVER smoked cigarettes EVERY DAY for at least 6 months?
PX030802 Former Smoker Quantity 1DayA	When you last smoked every day, on average how many cigarettes did you smoke each day?
PX030802 Former Smoker Quantity 1DayB	When you last smoked fairly regularly, on average how many cigarettes did you smoke each day?
PX061301 Weekend Sun Hours Last Decade	On a typical weekend day in the summer, about how many hours did you generally spend in the mid-day sun in the past ten years?

Table 3

Number of Resulting Models for Each Implication Score Cutoff

Implication Index Cutoff	Number of Models
4	603032
5	337113
6	57376
7	2479

Table 4

Five most significant GxE results for each PhenX measurement, sorted by chromosome and gene

Chr	BP	RSID	PhenX variable	P-value	Gene
1	38802669	rs4568792	Former Smoker Smoking 6 Month	4.67×10^{-6}	
1	233192767	rs7412124	Alcohol 30 Day Quantity	0.00022	
2	52547408	rs6726893	Everyday Smoker Quantity 1 Day	0.00047	
3	69335296	rs12494107	Someday Smoker Days 1 Month	0.019	<i>FRMD4B</i>
3	999833201	rs13091236	Alcohol 30 Day Quantity	0.00028	<i>ST3GAL6</i>
3	100016640	rs13059624	Alcohol 30 Day Quantity	0.00025	<i>DCBLD2</i>
3	100092139	rs13084692	Alcohol 30 Day Quantity	0.00028	<i>DCBLD2</i>
4	8423988	rs747580	Former Smoker Quantity 1 Day A	2.13×10^{-5}	<i>ACOX3</i>
4	8480717	rs2631731	Former Smoker Quantity 1 Day A	8.40×10^{-6}	<i>ACOX3</i>
4	37295333	rs2048257	Former Smoker Smoking 6 Month	8.29×10^{-6}	<i>RELL1</i>
4	42991353	rs17457584	Alcohol 30 Day Quantity	2.64×10^{-5}	
4	53149495	rs346005	Alcohol 30 Day Frequency	1.92×10^{-6}	
4	114190823	rs1026975	Cigarette Smoking 100	1.0×10^{-5}	<i>ANK2</i>
5	123324046	rs2250107	Everyday Smoker Quantity 1 Day	0.00049	
5	123343048	rs2546839	Everyday Smoker Quantity 1 Day	0.00044	
6	4098136	rs653674	Former Smoker Quantity 1 Day A	2.53×10^{-5}	
6	66776318	rs6899720	Cigarette Smoking Current	8.87×10^{-6}	
6	66776318	rs6899720	Cigarette Smoking Everyday 6 Month	3.83×10^{-6}	
6	66900502	rs12528760	Cigarette Smoking Current	7.93×10^{-6}	
6	66900502	rs12528760	Cigarette Smoking Everyday 6 Month	4.36×10^{-6}	
6	135370485	rs6929661	Cigarette Smoking Everyday 6 Month	6.78×10^{-6}	<i>HBS1L</i>
6	135376293	rs1014021	Cigarette Smoking Everyday 6 Month	5.74×10^{-6}	<i>HBS1L</i>
6	135407511	rs6569990	Cigarette Smoking Everyday 6 Month	5.26×10^{-6}	<i>HBS1L</i>
6	170414666	rs3012437	Cigarette Smoking 100	1.36×10^{-5}	<i>LOC285804</i>
7	97704227	rs3735258	Former Smoker Quantity 1 Day A	3.01×10^{-5}	<i>DKFZP434B0335</i>
7	144435178	rs10254774	Cigarette Smoking 100	1.26×10^{-5}	

Chr	BP	RSID	PhenX variable	P-value	Gene
9	2449444	rs1006575	Former Smoker Smoking 6 Month	1.39×10^{-6}	
9	6256440	rs2026991	Cigarette Smoking Current	1.22×10^{-5}	
9	79252094	rs10116050	Alcohol 30 Day Frequency	1.09×10^{-5}	<i>GNAI4</i>
9	79255890	rs4745639	Alcohol 30 Day Frequency	6.74×10^{-6}	<i>GNAI4</i>
10	72815798	rs4747150	Alcohol 30 Day Frequency	9.41×10^{-6}	
12	96988908	rs11109339	Former Smoker Quantity 1 Day A	1.22×10^{-5}	
13	59162465	rs1379518	Former Smoker Quantity 1 Day B	0.0011	<i>DIAPH3</i>
14	36400549	rs1325530	Weekend Sun Hours Last Decade	1.10×10^{-5}	<i>SLC25A21</i>
15	20596568	rs3812923	Cigarette Smoking Current	7.53×10^{-6}	<i>NIPAI</i>
15	22652931	rs752873	Alcohol 30 Day Frequency	6.23×10^{-6}	<i>SNRPB</i>
15	48790334	rs10519284	Someday Smoker Days 1 Month	0.018	<i>SPPL2A</i>
15	48796548	rs12898588	Someday Smoker Days 1 Month	0.018	<i>SPPL2A</i>
15	48797199	rs7165492	Someday Smoker Days 1 Month	0.018	<i>SPPL2A</i>
15	59933890	rs17238096	Someday Smoker Quantity 1 Day	0.022	<i>VPS13C</i>
16	50786246	rs2245948	Former Smoker Quantity 1 Day B	0.0019	
16	50794864	rs7200614	Former Smoker Quantity 1 Day B	0.0019	
16	64857219	rs233546	Someday Smoker Days 1 Month	0.018	
16	76582490	rs6564494	Former Smoker Quantity 1 Day B	0.0021	
16	83259294	rs4783043	Former Smoker Quantity 1 Day B	0.0020	
17	2685811	rs9747501	Everyday Smoker Quantity 1 Day	0.00055	<i>GARNL4</i>
17	40022727	rs16970865	Someday Smoker Quantity 1 Day	0.021	
17	40035928	rs9904409	Someday Smoker Quantity 1 Day	0.020	
17	40057582	rs10451262	Someday Smoker Quantity 1 Day	0.019	
18	74848306	rs612829	Former Smoker Smoking 6 Month	3.52×10^{-6}	<i>SALL3</i>
19	61733570	rs2058131	Former Smoker Smoking 6 Month	2.72×10^{-7}	
20	45927980	rs8121494	Cigarette Smoking Current	6.86×10^{-6}	
21	33511140	rs9978523	Everyday Smoker Quantity 1 Day	0.00059	
22	22564245	rs738807	Weekend Sun Hours Last Decade	5.14×10^{-6}	

Chr	BP	RSID	PhenX variable	P-value	Gene
22	225644493	rs5751759	Weekend Sun Hours Last Decade	2.79×10^{-6}	
22	22565198	rs4822443	Weekend Sun Hours Last Decade	3.21×10^{-6}	<i>MIF</i>
22	22567862	rs2000466	Weekend Sun Hours Last Decade	4.33×10^{-6}	<i>MIF</i>
22	26292517	rs723184	Cigarette Smoking 100	1.63×10^{-5}	
22	26298208	rs5762257	Cigarette Smoking 100	1.57×10^{-5}	
22	32107269	rs5998902	Someday Smoker Quantity 1 Day	0.022	<i>LARGE</i>

Table abbreviations:
Chr = Chromosome
BP = Base pair location of SNP
RSID = SNP ID
P-Value = P-value of the interaction
Gene = Gene symbol of gene the SNP is within (blank if not within a gene)

Table 5

The 5 SNPxSNP models with an interaction p-value < 0.05 and a total model p-value < 0.05 after association testing of the Biofilter derived pairwise models. Presented in Table 5 are the effect coefficients for the main effects (β_1 and β_2) and the interaction term (β_3), as well as the P-value for the interaction term (P_{ixn}) and the P-value for the model (P_{mod}).

Gene 1	SNP1	Allele /MAF	Gene 2	SNP2	Allele /MAF	β_1	β_2	β_3	P_{ixn}	P_{mod}
EGF	rs2298999	0.41	EGFR	rs17172446	0.23	0.039	0.087	-0.086	3.37×10^{-6}	3.93×10^{-5}
LCP2	rs2338872	0.40	VAV1	rs12979659	0.37	0.036	0.058	-0.072	8.28×10^{-6}	6.18×10^{-5}
FYN	rs1327200	0.34	DOCK1	rs10829597	0.40	-0.084	-0.049	0.067	4.53×10^{-5}	2.56×10^{-5}
FYN	rs1327200	0.34	DOCK1	rs2050305	0.37	-0.080	-0.046	0.066	7.56×10^{-5}	4.24×10^{-5}
DOCK1	rs9418709	0.47	SRC	rs6063022	0.15	0.019	0.061	-0.083	5.74×10^{-5}	0.00036