ERRATUM

[SARAH A. PENDERGRASS, SHEFALI S. VERMA, EMILY R. HOLZINGER, CARRIE B. MOORE, JOHN WALLACE, SCOTT M. DUDEK, WAYNE HUGGINS, TERRIE KITCHNER, CAROL WAUDBY, RICHARD BERG, CATHERINE A. MCCARTY, and MARYLYN D. RITCHIE (2012) NEXT-GENERATION ANALYSIS OF CATARACTS: DETERMINING KNOWLEDGE DRIVEN GENE-GENE INTERACTIONS USING BIOFILTER, AND GENE-ENVIRONMENT INTERACTIONS USING THE PHENX TOOLKIT. Biocomputing 2013: pp. 147-158, doi: 10.1142/9789814447973_0015]

]

"This corrects the above-titled article. There was an error in the case-control label for a subset of samples. This was corrected and analyses were re-run. The thrust of the results and discussion did not change, but these results are more precise and corrected."

!

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?

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.

!

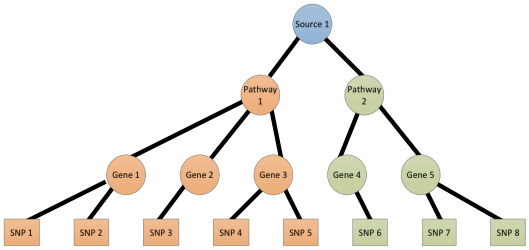

Figure 1. Simplified model for one Biofilter 1.0 database source with 2 pathways, 5 genes, and 8 SNPs

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 527,953 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.

Table 3. Number of Resulting Models for Each Implication Score Cutoff

Implication Index Cutoff	Number of Models
4	603,032
5	337,113
6	57,376
7	2479

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 via likelihood ratio test (LRT) of the full versus

reduced models, using logistic regression, where the full model was: SNP1 + SNP2 + SNP1*SNP2 and the reduced model was: SNP1 + SNP2 for all of the pairwise sets of SNPs generated by Biofilter with an implication index of 6. For the GxE (SNPxE) models, the same methods were employed using PLATO; however the full model was: SNP1 + ENV1 + SNP1*ENV1 and the reduced model was: SNP1 + ENV1 for all the possible unique SNPxE pairs, from the set of 527,936 SNPs and the PhenX variables described earlier in methods. Again, the outcome was case control status for cataracts. 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 with LRT p-values $\leq 1 \times 10^{-4}$, a total of 782 models exhibited an interaction with a p-value $\leq 1 \times 10^{-4}$ 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 "Weekend Sun Hours Last Decade" a survey question asking "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?" with the SNP rs6447541, located in an intron of *GABR1* on chromosome 4, with an association LRT p-value of 2.35×10^{-8}, was the most significant interaction found when compared to the other 12 PhenX measurements we used in our GxE analysis.

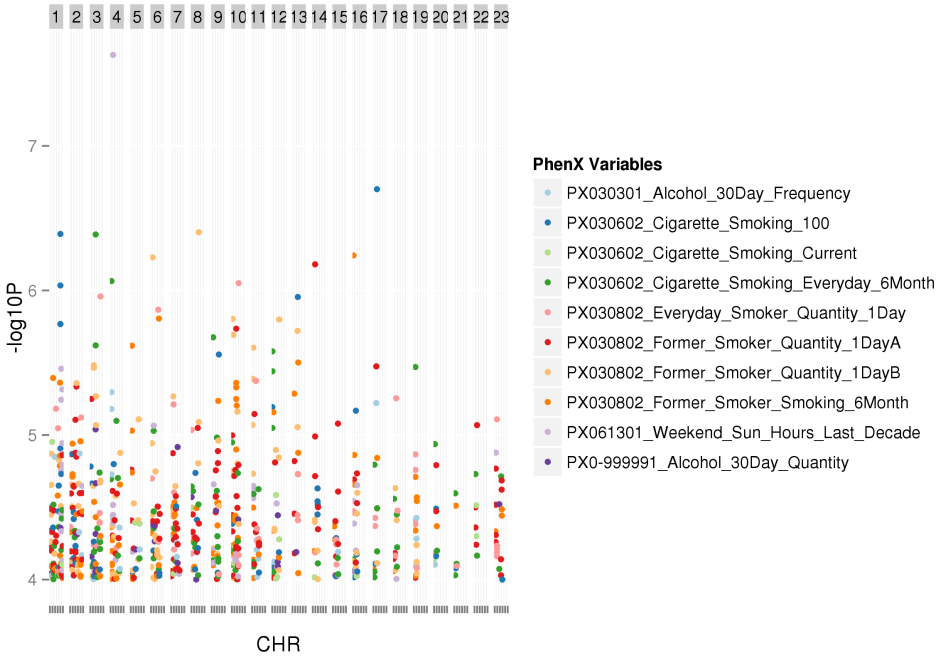

Figure 2. Manhattan plot of the association results for the GxE interaction models. Displayed are the results for the 12 PhenX measures that had interaction p-values < 1×10^{-4}. Two PhenX variables did not have an interaction p-value less than 1×10^{-4}.

Table 4. Five most significant results for each PhenX measurement, sorted by chromosome and position

RSID	PhenX Variable	Chr:BP	P-value	Gene
rs7529518	PX030602_Cigarette_Smoking_100	1:200718422	$1.71x10^{-6}$	*CAMSAP2*
rs2292097	PX030602_Cigarette_Smoking_100	1:200843768	$9.23x10^{-7}$	*GPR25**
rs10800745	PX030602_Cigarette_Smoking_100	1:200849676	$4.06x10^{-7}$	*GPR25**
rs11117581	PX061301_Weekend_Sun_Hours_Last_Decade	1:216613761	$5.70x10^{-6}$	*USH2A**
rs11117582	PX061301_Weekend_Sun_Hours_Last_Decade	1:216622208	$3.48x10^{-6}$	*USH2A**
rs10495409	PX061301_Weekend_Sun_Hours_Last_Decade	1:238255679	$4.84x10^{-6}$	*MTND5P18*
rs607949	PX030602_Cigarette_Smoking_Current	1:43695708	$1.12x10^{-5}$	*WDR65*
rs581503	PX030802_Former_Smoker_Smoking_6Month	1:61329593	$4.03x10^{-6}$	*NFIA**
rs11803470	PX030301_Alcohol_30Day_Frequency	1:95117783	$1.42x10^{-5}$	*ABCD3**
rs2587695	PX030802_Former_Smoker_Quantity_1DayA	2:120321817	$4.62x10^{-6}$	*PCDP1*
rs262302	PX030301_Alcohol_30Day_Frequency	2:180931461	$1.34x10^{-5}$	*CWC22**
rs5994737	PX030602_Cigarette_Smoking_Current	22:33804804	$3.07x10^{-5}$	*LARGE*
rs3846094	PX030602_Cigarette_Smoking_Everyday_6Month	3:101601159	$2.40x10^{-6}$	*NFKBIZ **
rs11720478	PX030602_Cigarette_Smoking_Everyday_6Month	3:101637806	$4.10x10^{-7}$	*NFKBIZ**
rs12495970	PX030802_Everyday_Smoker_Quantity_1Day	3:194928138	$1.10x10^{-6}$	*XXYLT1*
rs11735349	PX030602_Cigarette_Smoking_Everyday_6Month	4:16506826	$8.61x10^{-7}$	*LDB2*
rs157606	PX030301_Alcohol_30Day_Frequency	4:16699795	$5.05x10^{-6}$	*LDB2*
rs283018	PX030301_Alcohol_30Day_Frequency	4:16740168	$6.61x10^{-6}$	*LDB2*
rs6447541	PX061301_Weekend_Sun_Hours_Last_Decade	4:47215939	$2.35x10^{-8}$	*GABRB1*
rs16888770	PX030802_Former_Smoker_Smoking_6Month	5:21586180	$2.41x10^{-6}$	*GUSBP1*
rs13183503	PX030602_Cigarette_Smoking_Current	5:81515885	$4.04x10^{-5}$	*ATG10*
rs9376419	PX030802_Everyday_Smoker_Quantity_1Day	6:139801295	$1.36x10^{-6}$	*TXLNB*
rs3798756	PX030802_Former_Smoker_Smoking_6Month	6:152529260	$1.56x10^{-6}$	*SYNE1*
rs3094549	PX030802_Former_Smoker_Quantity_1DayB	6:29355148	$5.91x10^{-7}$	*OR12D2**
rs4712006	PX061301_Weekend_Sun_Hours_Last_Decade	6:52245415	$8.61x10^{-6}$	*PAQR8*
rs3889488	PX030802_Former_Smoker_Quantity_1DayB	8:141544748	$3.96x10^{-7}$	*AGO2*
rs6987670	PX030602_Cigarette_Smoking_Current	8:9883177	$3.18x10^{-5}$	*MSRA**
rs10968388	PX030602_Cigarette_Smoking_Everyday_6Month	9:28210699	$2.11x10^{-6}$	*LINGO2*
rs9783135	PX030802_Everyday_Smoker_Quantity_1Day	10:129937722	$8.90x10^{-7}$	*MKI67**
rs12360020	PX030802_Former_Smoker_Quantity_1DayB	10:15264322	$1.57x10^{-6}$	*FAM171A1*
rs2820100	PX030802_Former_Smoker_Quantity_1DayA	10:84491173	$1.84x10^{-6}$	*NRG3*
rs6592528	PX030802_Everyday_Smoker_Quantity_1Day	11:73377350	$4.22x10^{-6}$	*PLEKHB1**
rs4944859	PX030802_Everyday_Smoker_Quantity_1Day	11:73424135	$4.22x10^{-6}$	*RAB6A*
rs7977795	PX030802_Former_Smoker_Quantity_1DayB	12:132096632	$1.59x10^{-6}$	*SFSWAP**
rs7972947	PX030602_Cigarette_Smoking_Everyday_6Month	12:2170433	$2.64x10^{-6}$	*CACNA1C*
rs775474	PX030602_Cigarette_Smoking_Current	12:70075933	$2.60x10^{-5}$	*BEST3*
rs680711	PX030602_Cigarette_Smoking_100	13:101814804	$1.11x10^{-6}$	*NALCN*
rs4772995	PX030802_Former_Smoker_Smoking_6Month	13:109410933	$3.15x10^{-6}$	*MYO16*
rs7983958	PX030802_Former_Smoker_Quantity_1DayB	13:96473682	$1.90x10^{-6}$	*UGGT2*
rs1957480	PX030802_Former_Smoker_Quantity_1DayA	14:44397890	$6.59x10^{-7}$	*X10IF4BP1**

rs11644531	PX030802_Former_Smoker_Smoking_6Month	16:6008824	$5.72x10^{-7}$	*RBFOX1**
rs8075882	PX030301_Alcohol_30Day_Frequency	17:55469362	$6.01x10^{-6}$	*MSI2*
rs1443269	PX030802_Former_Smoker_Quantity_1DayA	17:55894564	$3.35x10^{-6}$	*MRPS23**
rs9911607	PX030802_Former_Smoker_Quantity_1DayA	17:55895539	$3.35x10^{-6}$	*MRPS23**
rs7210514	PX030602_Cigarette_Smoking_100	17:67793814	$1.99x10^{-7}$	*KCNJ16**

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 is within or nearest to (*indicates nearest gene is listed)

3.2. *GxG Results*

The top Biofilter 1.0 derived GxG models are presented in Table 5. A total of 13 models had an LRT p-value $< 1 \times 10^{-4}$ and full model p-value < 0.01. A total of 9 genes were in the thirteen models. Of these models, the most significant was for a model with *SOS1*, which encodes a guanine nucleotide exchange factor for RAS proteins, and *FYN*, which is a member of the protein-tyrosine kinase oncogene family.

Table 5. The 13 SNPxSNP models with an interaction p-value $< 1 \times 10\text{-}4$ after association testing of the Biofilter derived pairwise models.

SNP1	Gene 1	SNP2	Gene 2	Interaction P-value
rs2888586	*SOS1*	rs706885	*FYN*	$1.29x10^{-6}$
rs2888586	*SOS1*	rs17072912	*FYN*	$2.14x10^{-6}$
rs2888586	*SOS1*	rs11964650	*FYN*	$2.97x10^{-6}$
rs2888586	*SOS1*	rs9372313	*FYN*	$6.32x10^{-6}$
rs17446875	*CDH2*	rs6121791	*CDH4*	$2.64x10^{-5}$
rs9384805	*FYN*	rs11017910	*DOCK1*	$2.67x10^{-5}$
rs11083252	*CDH2*	rs6121791	*CDH4*	$4.39x10^{-5}$
rs13135792	*KIT*	rs10515074	*PIK3R1*	$4.74x10^{-5}$
rs631428	*COL4A1*	rs3803231	*COL4A2*	$6.67x10^{-5}$
rs613116	*COL4A1*	rs3803231	*COL4A2*	$6.99x10^{-5}$
rs17704348	*FYN*	rs4751282	*DOCK1*	$8.85x10^{-5}$
rs17446875	*CDH2*	rs1110359	*CDH4*	$8.85x10^{-5}$
rs809193	*FYN*	rs11594969	*DOCK1*	$9.64x10^{-5}$

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 1 for SNPs rs2292097 and rs7529518, 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. 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 be 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.

Acknowledgments

This work was supported by the following grants: U19 HL0659625, R01 LM010040, U01 HG006389

References

1. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, et al. (2012) Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. Journal of the American Medical Informatics Association : JAMIA 19: 225-234.

2. Waudby CJ, Berg RL, Linneman JG, Rasmussen LV, Peissig PL, et al. (2011) Cataract research using electronic health records. BMC Ophthalmol 11: 32.

3. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD (2005) Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Personalized Medicine 2: 49-79.

4. Pathak J, Pan H, Wang J, Kashyap S, Schad PA, et al. (2011) Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects. AMIA Summits Transl Sci Proc 2011: 41-45.

5. Michael R, Bron AJ (2011) The ageing lens and cataract: a model of normal and pathological ageing. Philos Trans R Soc Lond B Biol Sci 366: 1278-1292.

6. Hammond CJ, Duncan DD, Snieder H, de Lange M, West SK, et al. (2001) The heritability of age-related cortical cataract: the twin eye study. Invest Ophthalmol Vis Sci 42: 601-605.

7. Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput: 368-379.

8. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29-34.
9. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619-622.

10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25: 25-29.

11. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405-420.

12. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, et al. (2010) NetPath: a public resource of curated signal transduction pathways. Genome Biol 11: R3.

13. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39: D698-704.

14. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40: D857-861.

15. Stover PJ, Harlan WR, Hammond JA, Hendershot T, Hamilton CM (2010) PhenX: a toolkit for interdisciplinary genetics research. Curr Opin Lipidol 21: 136-140.

16. Abraham AG, Condon NG, West Gower E (2006) The new epidemiology of cataract. Ophthalmol Clin North Am 19: 415-425.

17. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, et al. (2011) Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet Chapter 1: Unit1 19.

18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559-575.

19. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

20. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, et al. (2010) Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. Pac Symp Biocomput: 315-326.

21. Wickham H (2009) ggplot2: elegant graphics for data analysis: Springer New York.

22. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) 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 106: 9362-9367.