Supplementary Note

A gene-based association method for mapping traits using reference transcriptome data

Eric R. Gamazon[1,2,9], Heather E. Wheeler[3,9], Kaanan P. Shah[1,9], Sahar V. Mozaffari[4], Keston Aquino-Michaels[1], Robert J. Carroll[5], Anne E. Eyler[6], Joshua C. Denny[5], GTEx Consortium[7], Dan L. Nicolae[1,4,8], Nancy J. Cox[1,2,4], and Hae Kyung Im[1]

[1]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL

[2]Division of Genetic Medicine, Vanderbilt University, Nashville, TN

[3]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL

[4]Department of Human Genetics, University of Chicago, Chicago, IL

[5]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

[6]Department of Medicine, Vanderbilt University, Nashville, TN

[7]A full list of members and affiliations appears in the Supplementary Note.

[8]Department of Statistics, University of Chicago, Chicago, IL

[9]These authors contributed equally to this work.

Correspondence to:

Hae Kyung Im, Ph.D.

haky@uchicago.edu

Section of Genetic Medicine

Department of Medicine

The University of Chicago

Chicago, IL 60637

Single-variant results for SNPs in PrediXcan gene models

As expected, the genes associated with autoimmune diseases (RA and CD) each contained multiple SNPs that are individually associated with disease risk (Supplementary Table 1). Thus, the identified disease gene associations are consistent with the single-variant meta-analysis results. Interestingly, in many cases, we detect these associations with much smaller sample sizes. Furthermore, our gene-based results allow for more direct biological interpretation compared to individual SNPs.

The PrediXcan associations for BD and HT have not been observed before using traditional single-variant GWAS. The association between predicted expression of *PTPRE* and BD is further supported by single variant meta-analysis results from the Psychiatric Genetics Consortium (PGC)[31]. Supplementary Table 1 shows the meta-analysis p-values for each SNP included in the predictor of *PTPRE*. While none of the SNPs is individually genome-wide significant, 10 out of 23 are nominally associated with BD disease risk in the PGC meta-analysis ($p<0.05$). Follow-up studies of this disease association are necessary, but our analysis in combination with existing results suggests *PTPRE* may be an excellent BD candidate gene. Furthermore, this result highlights the advantage of our gene-based approach that combines information across SNPs, each of which many only contribute nominally to disease risk and therefore remain below the detection limits of single-variant analyses.

Supplementary figure legends

Supplementary Figure 1. Comparison of 10-fold cross-validated predictive performance between all tested methods (LASSO, elastic net with $\alpha=0.5$, top SNP,

polygenic score at several p-value thresholds) in the DGN whole blood cohort. Predictive performance was measured by the R^2 between predicted (GReX) and observed expression.

Supplementary Figure 2. Comparison of 10-fold cross-validated predictive performance of elastic net in different starting SNP sets (4.6M 1000 Genomes Project (TGP) SNPs, 1.9 M HapMap Phase II SNPs, 300K WTCCC genotyped SNPs) in the DGN whole blood cohort. Predictive performance was measured by the R^2 between predicted (GReX) and observed expression.

Supplementary Figure 3. Prediction performance of elastic net in GTEx tissues. Using whole blood prediction models trained in DGN, we compared predicted levels of expression with observed levels from nine tissues of the GTEx pilot project. The observed squared correlation between predicted and observed gene expression levels, R^2, is plotted against the null distribution of R^2.

Supplementary Figure 4. Comparison of prediction performance between local- and distal- based prediction models. Using whole blood prediction models trained in DGN, we compared predicted levels of expression with observed levels in GTEx whole blood. Local predictors were generated using elastic net on SNPs within 1Mb of each gene and distal predictors include any *trans*-eQTLs outside this region with a linear regression $p<10^{-5}$. The observed (y-axis) squared correlation between predicted and observed gene expression levels, R^2, is plotted against the null distribution of R^2(x-axis).

Supplementary Figure 5. PrediXcan results in WTCCC. Q-Q plot of the association p-values from the PrediXcan analysis of 6 remaining WTCCC diseases using expression levels imputed from the DGN whole blood. The red line in each panel shows the null expected distribution of p-values and the blue line represents the bonferroni corrected genome-wide significance threshold. For each disease, the top 3 genes that exceed the bonferroni significance threshold are labeled. The diseases shown are (a) rheumatoid arthritis, (b) Crohn's disease, (c) bipolar disorder, (d) coronary artery disease, (e) hypertension, and (f) type 2 diabetes.

Supplementary Figure 6. PrediXcan results in WTCCC. Plot of the association p-values based on genomic position from the PrediXcan analysis of 6 remaining WTCCC diseases using expression levels imputed from the DGN whole blood. The blue line in each panel represents the bonferroni corrected genome-wide significance threshold. For each disease, the top 3 genes that exceed the bonferroni significance threshold are

labeled. The diseases shown are (a) rheumatoid arthritis, (b) Crohn's disease, (c) bipolar disorder, (d) coronary artery disease, (e) hypertension, and (f) type 2 diabetes.

Supplementary Figure 7. Enrichment of known disease genes. Each plot shows the null expected distribution for the number of genes expected to fall below a p-value threshold of 0.01. The null distribution was derived via 10,000 random permutations. The large point on the horizontal axis of each plot shows the observed number of previously known disease genes that fall below the p-value threshold. The diseases shown are (a) rheumatoid arthritis, (b) Crohn's disease, (c) bipolar disorder, (d) coronary artery disease, (e) hypertension, and (f) type 2 diabetes.

Supplementary table legends

Supplementary Table 1. Meta-Analysis p-values for SNPs in predictors of top PrediXcan results. For each of the genes that reached genome-wide significance in our analysis, we looked up the meta-analysis p-values for the SNPs that are included in each of the DNG whole blood predictors. For comparison, we also include the p-value from the single variant analysis of the WTCCC only data.

Acknowledgements

Grants

The project described was supported in part by Award Number K12CA139160 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health,
Pharmacogenetics of Anticancer Agents Research (PAAR) Group (NIH/NIGMS grant UO1GM61393),
PGRN Statistical Analysis Resource (U19 HL065962),
Genotype-Tissue Expression project (GTeX) (R01 MH101820 and R01 MH090937),
University of Chicago DRTC (Diabetes Research and Training Center; P30 DK20595, P60 DK20595),
The Conte Center for Computational Neuropsychiatric Genomics (P50MH094267),
"Integrated GWAS of complex behavioral and gene expression traits in outbred rats" P50DA037844

KS was supported in part by the Training in Emerging Multidisciplinary Approaches to Mental Health and Disease Grant (T32MH020065),
HEW was supported in part by the National Research Service Award F32CA165823.
JCD was supported by the NIH Grant U01 GM092691.

GTEx data
The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc.(HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v3.p1.

WTCCC data
This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available fromwww.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

DGN data
NIMH Study 7 (GenRED I) - Data and biomaterials were collected in six projects that participated in the National Institute of Mental Health (NIMH) Genetics of Recurrent Early-Onset Depression (GenRED) project. From 1999-2003, the Principal Investigators and Co-Investigators were: New York State Psychiatric Institute, New York, NY, R01 MH060912, Myrna M. Weissman, Ph.D. and James K. Knowles, M.D., Ph.D.; University of Pittsburgh, Pittsburgh, PA, R01 MH060866, George S. Zubenko, M.D., Ph.D. and Wendy N. Zubenko, Ed.D., R.N., C.S.; Johns Hopkins University, Baltimore, MD, R01 MH059552, J. Raymond DePaulo, M.D., Melvin G. McInnis, M.D. and Dean MacKinnon, M.D.; University of Pennsylvania, Philadelphia, PA, RO1 MH61686, Douglas F. Levinson, M.D. (GenRED coordinator), Madeleine M. Gladis, Ph.D., Kathleen Murphy-Eberenz, Ph.D. and Peter Holmans, Ph.D. (University of Wales College of Medicine); University of Iowa, Iowa City, IW, R01 MH059542, Raymond R. Crowe, M.D. and

William H. Coryell, M.D.; Rush University Medical Center, Chicago, IL, R01 MH059541-05, William A. Scheftner, M.D., Rush-Presbyterian.

NIMH Study 18 - Data and biomaterials were obtained from the limited access datasets distributed from the NIH-supported “Sequenced Treatment Alternatives to Relieve Depression” (STAR*D). STAR*D focused on non-psychotic major depressive disorder in adults seen in outpatient settings. The primary purpose of this research study was to determine which treatments work best if the first treatment with medication does not produce an acceptable response. The study was supported by NIMH Contract # N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is NCT00021528.

NIMH Study 52 (GenRED II) – Data and biomaterials in this release were collected in six projects that participated in the National Institute of Mental Health (NIMH) Genetics of Recurrent Early-Onset Depression (GenRED) project (1999-2009). The Principal Investigators and Co-Investigators were: New York State Psychiatric Institute, New York, NY, R01 MH 060912, Myrna M. Weissman, Ph.D.; Johns Hopkins University, Baltimore, MD, R01 MH059552, J. Raymond DePaulo, M.D., and James B. Potash, M.D., M.P.H.; University of Pennsylvania, Philadelphia, PA (1999-2005), and Stanford University (2006-2009), R01 MH61686, Douglas F. Levinson, M.D. (GenRED coordinator); University of Iowa, Iowa City, IW, R01 MH059542e, Raymond R. Crowe, M.D., and William H. Coryell, M.D.; Rush University Medical Center, Chicago, IL, R01 MH059541-05, William A. Scheftner, M.D.; and University of Pittsburgh, Pittsburgh, PA (1999-2003), R01 MH060866, George S. Zubenko, M.D., Ph.D., and Wendy N. Zubenko, Ed.D., R.N., C.S.

NIMH Study 88 -- Data was provided by Dr. Douglas F. Levinson. We gratefully acknowledge the resources were supported by National Institutes of Health/National Institute of Mental Health grants 5RC2MH089916 (PI: Douglas F. Levinson, M.D.; Co-investigators: Myrna M. Weissman, Ph.D., James B. Potash, M.D., MPH, Daphne Koller, Ph.D., and Alexander E. Urban, Ph.D.) and 3R01MH090941 (Co-investigator: Daphne Koller, Ph.D.).

Computing resources

This work made use of the Open Science Data Cloud (OSDC) which is an Open Cloud Consortium (OCC)-sponsored project. This work was supported in part by grants from Gordon and Betty Moore Foundation and the National Science Foundation and major contributions from OCC members like the University of Chicago.

https://www.opensciencedatacloud.org/

Grossman RL, Greenway M, Heath AP, Powell R, Suarez R, Wells W, White KP, Atkinson M, Klampanos I, Alvarez H, Harvey C and Mambretti J, The Design of a

Community Science Cloud: The Open Science Data Cloud Perspective. (2012) doi:10.1109/SC.Companion.2012.127

This work made use of the Bionimbus Protected Data Cloud (PDC), which is a collaboration between the Open Science Data Cloud (OSDC) and the IGSB (IGSB), the Center for Research Informatics (CRI), the Institute for Translational Medicine (ITM), and the University of Chicago Comprehensive Cancer Center (UCCCC). The Bionimbus PDC is part of the OSDC ecosystem and is funded as a pilot project by the NIH.

https://www.bionimbus-pdc.opensciencedatacloud.org/

Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, Bandlamudi C, McNerney ME, White KP and Grossman RL, Bionimbus: A Cloud for Managing, Analyzing and Sharing Large Genomics Datasets. J Am Med Inform Assoc (2014) doi:10.1136/amiajnl-2013-002155

GTEx Consortium Members

cancer Human Biobank (caHUB)

Biospecimen Source Sites (BSS)

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, *National Disease Research Interchange, Philadelphia, PA*
Richard Hasz, *Gift of Life Donor Program, Philadelphia, PA*
Gary Walters, *LifeNet Health, Virginia Beach, VA*
Nancy Young, *Albert Einstein Medical Center, Philadelphia, PA*
Laura Siminoff (ELSI Study), Heather Traino, Maghboeba Mosavel, Laura Barker, *Virginia Commonwealth University, Richmond, VA*
Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, *Roswell Park Cancer Institute, Buffalo, NY*
Susan Sullivan, Justin Bridge, *Upstate New York Transplant Service, Buffalo, NY*

Comprehensive Biospecimen Resource (CBR)

Scott Jewell, Dan Roher, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Melissa Hanson, Anthony Watkins, Brian Smith, *Van Andel Institute, Grand Rapids, MI*

Pathology Resource Center (PRC)

Leslie Sobin, James Robb, *SAIC-Frederick, Inc., Frederick, MD*
Phillip Branton, *National Cancer Institute, Bethesda, MD*
John Madden, *Duke Univeristy, Durham, NC*
Jim Robb, Mary Kennedy, *College of American Pathologists, Northfield, IL*

Comprehensive Data Resource (CDR)

Greg Korzeniewski, Charles Shive, Liqun Qi, David Tabor, Sreenath Nampally, *SAIC-Frederick, Inc., Frederick, MD*

caHUB Operations Management

Steve Buia, Angela Britton, Anna Smith, Karna Robinson, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, *SAIC-Frederick, Inc., Frederick, MD*
Kenyon Erickson, *Sapient Government Services, Arlington, VA*

Laboratory, Data Analysis, and Coordinating Center (LDACC)

Kristin Ardlie, Gad Getz, co-PIs; David S. DeLuca, Taylor Young, Ellen Gelfand, Daniel MacArthur, Manolis Kellis, Yan Meng, *The Broad Institute of Harvard and MIT, Inc., Cambridge, MA*

Brain Bank

Deborah Mash, PI; Yvonne Marcus, Margaret Basile, *University of Miami School of Medicine, Miami, FL*

Statistical Methods Development

Jun Liu, co-PI, *Harvard University, Boston, MA, USA*
Jun Zhu, co-PI; Zhidong Tu, *Mt Sinai School of Medicine, New York, NY*

Nancy Cox, Dan Nicolae, co-PIs; Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, *University of Chicago, Chicago, IL*

Jonathan Pritchard, PI; Matthew Stephens, Timothèe Flutre, Xiaoquan Wen, *University of Chicago, Chicago, IL*

Emmanouil T. Dermitzakis, co-PI; Tuuli Lappalainen, *University of Geneva, Geneva, Switzerland*
Roderic Guigo, co-PI; Jean Monlong, Michael Sammeth, *Center for Genomic Regulaton, Barcelona, Spain*
Daphne Koller, co-PI; Alexis Battle, Sara Mostafavi, *Stanford University, Palo Alto, CA*
Mark McCarthy, co-PI; Manuel Rivas, Andrew Morris, *Oxford University, Oxford, United Kingdom*

Ivan Rusyn, Andrew Nobel, Fred Wright, Co-PIs; Andrey Shabalin, *University of North Carolina - Chapel Hill, Chapel Hill, NC*

US National Institutes of Health

NCBI dbGaP

Mike Feolo, Steve Sherry, Jim Ostell, Nataliya Sharopova, Anne Sturcke, *National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD*

Program Management

Leslie Derr, *Office of Strategic Coordination (CommonFund), Office of the Director, National Institutes of Health, Bethesda, MD*
Eric Green, Jeffery P Struewing, Simona Volpi, Joy Boyer, Deborah Colantuoni, *National Human Genome Research Institute, Bethesda, MD*
Thomas Insel, Susan Koester, A Roger Little, Patrick Bender, Thomas Lehner, *National Institute of Mental Health, Bethesda, MD*
Jim Vaught, Sherry Sawyer, Nicole Lockhart, Chana Rabiner, Joanne Demchok, *National Cancer Institue, Bethesda, MD*