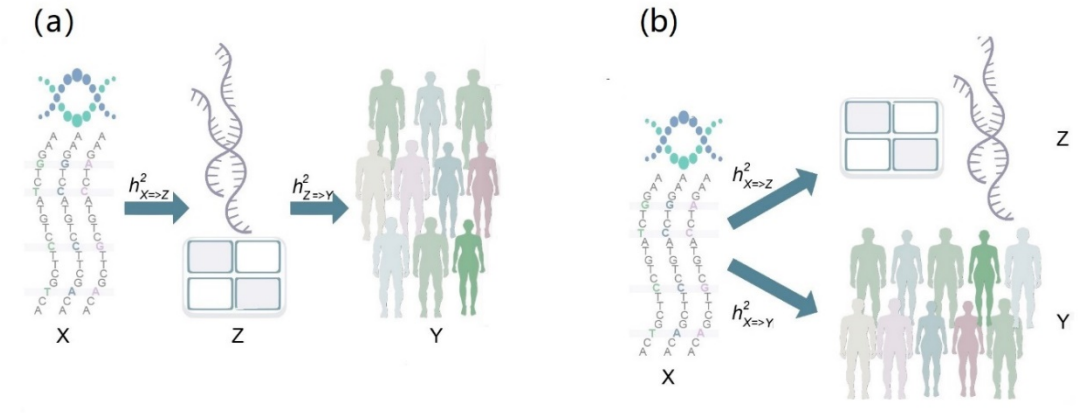
Disentangling genetic feature selection and aggregation in transcriptome-wide association studies

Chen Cao[1], Pathum Kossinna[1], Devin Kwok[2], Qing Li[1], Jingni He[1], Liya Su[3], Xingyi Guo[4], Qingrun Zhang[1,2,*], Quan Long[1,2,5,6,*]

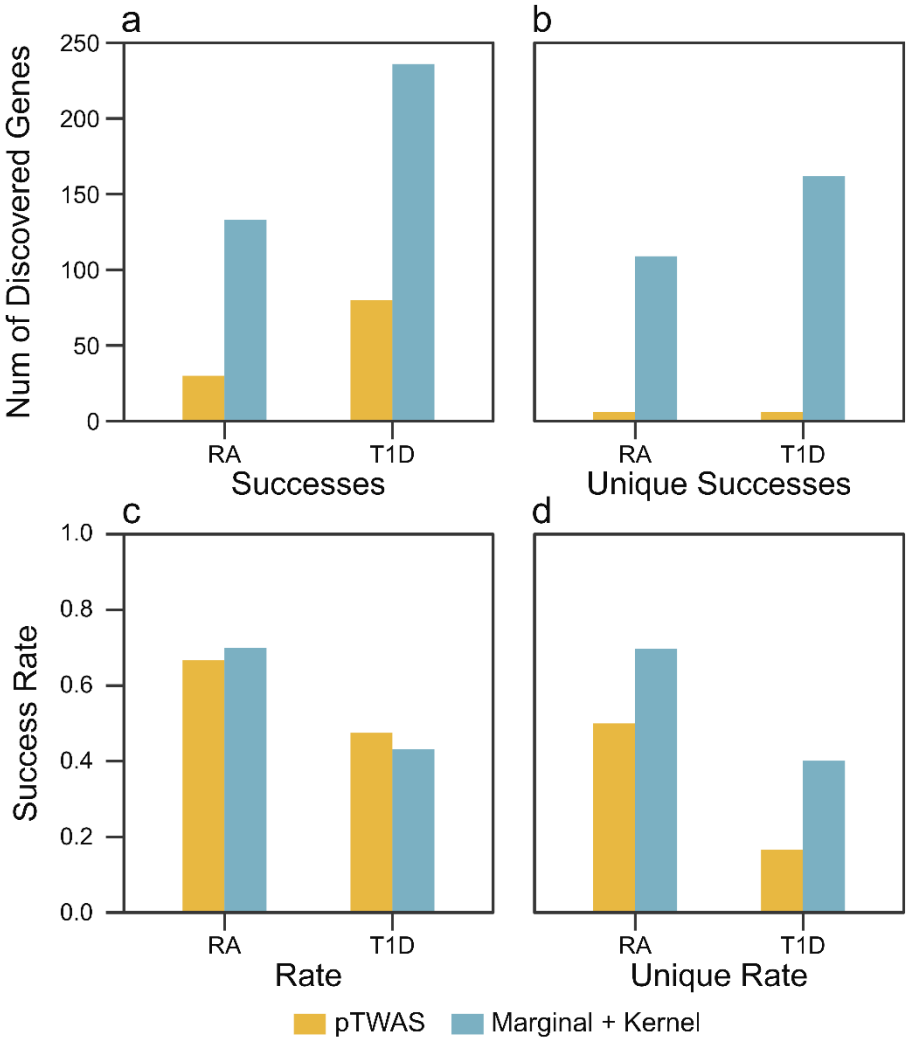
[1]Department of Biochemistry & Molecular Biology, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB T2N 4N1, Canada.
[2]Department of Mathematics & Statistics, University of Calgary, Calgary, AB T2N 1N4, Canada.
[3]Department of Pathology, Anatomy and Cell Biology, Thomas Jefferson University, Philadelphia, PA 19107, USA.
[4]Division of Epidemiology, Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN 37203, USA.
[5]Department of Medical Genetics, University of Calgary, Calgary, AB T2N 4N1, Canada.
[6]Hotchkiss Brain Institute, O'Brien Institute for Public Health, University of Calgary, Calgary, AB T2N 4N1, Canada.

*To whom correspondence should be addressed. Tel: +1 403-220-5580; Email: quan.long@ucalgary.ca
Correspondence may also be addressed to Qingrun Zhang. Tel: +1 403-220-2833; Email: qingrun.zhang@ucalgary.ca

Supplementary Figure S1: Pleiotropy v.s. casualty. Two commonly assumed scenarios are simulated. **(a)** Under causality, the genotype $\boldsymbol{x}$ is causal to gene expression $\boldsymbol{z}$, and in turn expression is causal to phenotype $\boldsymbol{y}$, resulting in dependence of $\boldsymbol{y}$ on both $\boldsymbol{z}$ and $\boldsymbol{x}$. (b) Under pleiotropy, the genotype $\boldsymbol{x}$ is independently causal to phenotype $\boldsymbol{y}$ and expression $\boldsymbol{z}$. As such, the phenotype and expression are not causal to each other.



Supplementary Figure S2: Comparison of PTWAS and marginal effect-based feature selection in WTCCC data. The two disentangled protocols PTWAS (left) and Marginal + Kernel (right) are compared on two WTCCC diseases (T1D and RA). Both protocols use kernel-based feature aggregation, but have different feature selection methods. **(a)** Number of discovered genes (successes) which are reported as disease-associated in DisGeNET. **(b)** Number of discovered genes (successes) discovered exclusively by one of the two protocols under comparison. **(c)** Proportion (success rate) of all discovered genes which are validated by DisGeNET. **(d)** Proportion (success rate) of genes discovered exclusively by each protocol which are validated by DisGeNET.



Literature support for discovered genes

For rheumatoid arthritis (RA), the majority of significant genes detected by Marginal + Kernel are within the Major Histocompatibility Complex (MHC) region. The MHC shows associations with almost all known autoimmune diseases as well as many inflammatory and infectious diseases (TROWSDALE AND KNIGHT 2013). Of the five most significant genes discovered by Marginal + Kernel (*HLA-DRB1*, *LY6G5B*, *HLA-DMA*, *FKBPL* and *HLA-DQB1-AS1*), three of them are from the human leukocyte antigen (HLA) gene family, all of which are located in the MHC. Genes from the HLA region have been reported as the most powerful disease risk predictors for RA (VAN DRONGELEN AND HOLOSHITZ 2017). In addition, there exist significant differences in the expression of *FKBPL* in some tissues for rheumatoid arthritis (HUFFMAN *et al.* 2017). The majority of the five most significant genes reported by the other three protocols under comparison are also supported by existing literature (**Supplementary Table S6**).

For type 1 diabetes (T1D), the majority of significant genes detected by Marginal + Kernel are also located in the MHC region. Of the five genes reported by Marginal + Kernel as most strongly associated with T1D (*TAP1*, *HLA-DQA1*, *CFB*, *HLA-DQB1*, and *HLA-DRB5*), three are from the HLA region (*HLA-DQA1*, *HLA-DQB1* and *HLA-DRB5*). The HLA region accounts for approximately half of the familial aggregation of T1D (NOBLE AND VALDES 2011). In particular, polymorphisms of class II HLA genes encoding DQ, DR, and to a lesser extent DP are the major genetic determinants of T1D (NOBLE AND VALDES 2011). *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB5* are all subunits of DQ and DR. *TAP1*, which Marginal + Kernel reports as the most significant gene polymorphism, has evidence of association with T1D (LI *et al.* 2014). The majority of the five most significant genes reported by the other three protocols in this study are also supported by previous literature (**Supplementary Table S6**).

A limited number of significant genes were identified for the five remaining diseases in WTCCC: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), type 2 diabetes (T2D), and hypertension (HT) (**Supplementary Table S3**). For example, transcription factor 7-like 2 (*TCF7L2*) is the most important susceptibility gene for type 2 diabetes identified by Marginal + Kernel (VILLAREAL *et al.* 2010). Among the four protocols, only Marginal + Kernel was able to identify *TCF7L2*. Another example is the gene *IRGM*, which plays an important role in the pathogenesis of Crohn's disease and is recognized as an independent major CD susceptibility locus from several previous studies (PRESCOTT *et al.* 2010; BASKARAN *et al.* 2014). Marginal + Kernel is the only method which detected *IRGM* as significantly associated with CD. Supporting literature PMIDs are listed in **Supplementary Table S6**.

Reference

Baskaran, K., S. Pugazhendhi and B. S. Ramakrishna, 2014 Association of IRGM gene mutations with inflammatory bowel disease in the Indian population. PLoS One 9**:** e106863.

Huffman, K. M., R. Jessee, B. Andonian, B. N. Davis, R. Narowski *et al.*, 2017 Molecular alterations in skeletal muscle in rheumatoid arthritis are related to disease activity, physical inactivity, and disability. Arthritis Res Ther 19**:** 12.

Li, Y. Y., W. Gao, S. S. Pang, X. Y. Min, Z. J. Yang *et al.*, 2014 TAP1 I333V gene polymorphism and type 1 diabetes mellitus: a meta-analysis of 2248 cases. J Cell Mol Med 18**:** 929-937.

Noble, J. A., and A. M. Valdes, 2011 Genetics of the HLA region in the prediction of type 1 diabetes. Curr Diab Rep 11**:** 533-542.

Prescott, N. J., K. M. Dominy, M. Kubo, C. M. Lewis, S. A. Fisher *et al.*, 2010 Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. Hum Mol Genet 19**:** 1828-1839.

Trowsdale, J., and J. C. Knight, 2013 Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet 14**:** 301-323.

van Drongelen, V., and J. Holoshitz, 2017 Human Leukocyte Antigen-Disease Associations in Rheumatoid Arthritis. Rheum Dis Clin North Am 43**:** 363-376.

Villareal, D. T., H. Robertson, G. I. Bell, B. W. Patterson, H. Tran *et al.*, 2010 TCF7L2 variant rs7903146 affects the risk of type 2 diabetes by modulating incretin action. Diabetes 59**:** 479-485.