Hae Kyung Im
12/15/2014

SCHOLARSHIP STATEMENT

(a) Past research in genomics:

After I joined the Biostatistics Laboratory at The University of Chicago, I became acquainted with the field of genomics through collaborations with Drs. Nancy Cox, Eileen Dolan, and Dan Nicolae. Shortly after the start of these collaborations, I discovered that my past scientific training in Physics and my statistical expertise gave me a unique perspective and allowed me to come up with novel ways to approach research questions. The following research projects exemplify the novelty of my contributions.

Intrinsic Cellular Proliferation. In the Pharmacogenomics of Anticancer Agents group (PAAR), we were interested in examining the genetic basis of drug response using lymphoblastoid cell lines. These cell lines are part of the HapMap and 1000 Genomes projects and offer a rich set of phenotypic and genotypic data. Despite the successes of Dolan lab in studying the genetic effect of drugs using lymphoblastoid cell lines, the validity of the model was being questioned because of many experimental artifacts that may affect the phenotypes. At the center of the criticism was the growth inhibition phenotype that was the basis of the drug response traits. I used a mixed effects model to come up with a novel phenotype, termed intrinsic growth, which took care of the experimental noise and was able to show that growth was under significant genetic control. In addition, I found that as much as 30% of the gene expression levels were associated with the intrinsic growth phenotype, which had been missed in previous studies because of the noisier phenotype used. This work was published in Plos Genetics (Im et al, 2012).

Genomic Privacy. Genomic privacy and data sharing are issues of relevance for the whole scientific community. Protecting the privacy of individuals who participate in a study has always been a top priority and it has been widely assumed that publishing summary results did not jeopardize privacy. In 2008, Homer et al found that in the case of genome wide association studies (GWAS), summary results such as allele frequencies for a large number of genetic variants can reveal whether a person participated in a study and the disease status of the individual. These results forced the NIH to withdraw most of the public access to GWAS study results. My collaborators were interested in sharing results from quantitative traits such as gene expression phenotypes, which provide critical information on the regulatory role of genetic variants. The question here was whether publishing regression coefficients from GWAS would also allow re-identification. I proved mathematically that re-identification based on regression coefficients was possible, provided an explicit method and computed its theoretical power as a function of sample size, number of markers, and false positive rate. In fact, I found that even the sign of the regression coefficients was enough to reveal a person's participation. This work was published in American Journal of Human Genetics (Im et al, 2012). Given its wide public relevance, this paper was featured on the University of Chicago Hospitals Science blog http://sciencelife.uchospitals.edu/2012/05/29/a-crack-in-the-safe-of-genomic-studies and has been routinely cited by papers on genomic privacy.

As my knowledge of the field evolved, I started to independently identify needs for which new methodologies had to be developed or borrowed and tailored from other fields. The following is an example of a prediction method I borrowed from Geostatistics and applied to the prediction of complex traits.

Poly-Omic Prediction of Complex Traits. Prediction of disease risk or treatment response is one of the pillars of personalized medicine. Although genome-wide association studies have discovered thousands of well-replicated polymorphisms associated with a broad spectrum of complex traits, the combined predictive power of these associations for any given trait is generally too low to be of clinical relevance. To address these issues, I proposed a systems approach to complex trait prediction, which leverages and integrates similarity in genetic and other high throughput molecular traits (omic data). The approach translates the omic similarity into phenotypic similarity using a method called Kriging, commonly used in Geostatistics. My method called OmicKriging emphasizes the use of a wide variety of systems-level data, such as those increasingly made available by comprehensive surveys of the genome, transcriptome and epigenome, for complex trait prediction. Application to clinical and cellular phenotypes shows the advantages of integrating multiple omic data in a collective manner. This work was published in Genetic Epidemiology (Wheeler et al 2014) with myself as the senior author.

(b) Past research in spatial statistics:

Statistical approximation of air quality model system. I developed fast approximations to some of the outputs of a computationally expensive chemical transport model called CMAQ (Community Multiscale Air Quality). By dramatically reducing the processing time, I was able to solve a large-scale inverse problem that corrected ammonia emissions estimates combining observed depositions and physical model output. The results were published in the Journal of Geophysical Research, a leading journal in Geophysics (Im et al 2005).

Semiparametric spectral density estimation. I developed a new class of covariance functions that shares the benefit of the widely used Matern class but is more flexible in the middle frequencies of the spectral domain. Applications to rainfall data and to numerous simulated datasets show that our semi-parametric model outperforms existing methods both in terms of predicted values and uncertainty estimations. This work was published in the Journal of the American Statistical Association (Im et al 2007), a leading journal in Statistics.

Space-time interpolation of temperature. As part of a study of the effect of air quality in asthma, it was found that temperature was a stronger predictor than particulate matter and ozone. Thus to improve a model of asthma incidence, I studied the spatiotemporal properties of air temperature in the Chicago area based on measurements from 10 observation sites over 20 years. I generated an interpolated map of the region, which borrows strengths from both spatially and temporally close data. This work was published in Environmetrics (Im et al 2009).

(c) Current research:

PrediXcan, regulatory mechanism driven gene-based test. Genome-wide association studies (GWAS) have identified thousands of variants robustly associated with complex traits. However, the biological mechanisms underlying these associations are, in general, not well understood. I proposed a gene-based association method called PrediXcan that directly tests the molecular mechanisms through which genetic variation affects phenotype. The approach estimates the component of gene expression determined by an individual's genetic profile and correlates the "imputed" gene expression with the phenotype under investigation to identify genes involved in the etiology of the phenotype. The genetically regulated gene expression is estimated using whole-genome tissue-dependent prediction models trained with reference transcriptome datasets. PrediXcan enjoys the benefits of gene-based approaches such as reduced multiple testing burden, more comprehensive annotation of gene function compared to that derived from single variants, and a principled approach to the design of follow-up experiments. Since no actual expression data are used in the analysis of GWAS data - only in silico expression - reverse causality problems are avoided. PrediXcan harnesses reference transcriptome data for disease mapping studies. The manuscript describing the method and application to Wellcome Trust Case Control Consortium data is currently under review in Nature Genetics as a companion to the GTEx main paper, under revision for Science. A draft can be found here: https://www.dropbox.com/s/gq913dfvx4q44aj/NG-sumitted-41531_0_merged_1415729631.pdf?dl=0

Database of prediction models – PredictDB. A key component of the PrediXcan is the prediction/imputation of gene expression traits based on whole genome variation. I am building a publicly available database of prediction models for a range of tissues. The database is currently online on the Open Science Data Cloud (https://imlab.uchicago.edu/page/predictdb-access).

Genetic architecture of expression traits across tissues. Using gene expression data for over 40 tissues from the GTEx consortium, I am investigating the cross tissue and tissue-specific genetic architecture of expression traits, which will guide the model selection for the prediction/imputation of the traits. More specifically, I am interested on the local (genetic variant near the gene, cis variants) and distal (genetic variants far from the gene, trans variant) heritability of expression traits. I am investigating another aspect of genetic architecture important for model selection: the sparsity/polygenicity of expression traits.

(b) Proposed and future research:

In the future, I will continue focusing on the development of methods and tools to facilitate the translation of genomic knowledge to improve health and the prevention and treatment of disease. In the near term, I will develop resources and methods to improve the prediction of complex traits as well as the biological interpretability of genetic discoveries.

GTEx collaboration. I will continue participation in the GTEx consortium. GTEx (Genotype-Tissue Expression) project is an NIH Common Fund project that aims to collect a comprehensive set of tissues from 900 deceased donors (for a total of about 20,000 samples) and to provide the scientific community a database of genetic associations with molecular traits such as mRNA levels, methylation status, telomere length, protein levels, DNAse hypersensitivity, among others. This trove of molecular data will allow us to test many hypotheses and motivate multiple methods to address important biological questions.

Beta cell expression modeling and diabetes dissection. I have established collaboration with Mark McCarthy and Anna Gloyn, world leaders in the genetics of diabetes and metabolic traits. This collaboration will allow us early access to the transcriptome levels of 800 beta cell samples, a very challenging tissue to access. The collection is made possible by contributions from a worldwide network of collaborators, which include Francis Collins and Leif Groop. We plan to develop prediction models of expression traits using beta cell samples and apply the PrediXcan method to diabetes related phenotypes. In addition to this discovery goal, there are multiple opportunities for method development for integration with GTEx and functional data from other consortia such as ENCODE.

MetaXcan: extension of PrediXcan to summary statistics. I am planning to generate methods and resource that will allow us to apply the PrediXcan method using only summary statistics. This will dramatically reduce the computational effort and the applicability of the approach, since summary statistics are much more readily available. We will be able to take advantage of massive meta-analysis results that are currently based on hundreds of thousands of individuals but will most likely grow to millions of individuals in the future.