Research Statement

Hae Kyung Im, PhD

My research goal is to utilize my statistical tools and modeling strategies to answer biomedical questions and make discoveries that can be translated to enhance health and improve disease prevention and treatment.

Current work in statistical genetics

Polyomic 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, transcriptomic or other omics-level data. The approach translate 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 show the advantages of integrating multiple omic data in a collective manner. A manuscript is currently under review and can be found on Arxiv. An R package, OmicKriging, implements the method and is publicly available.

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. A question we wanted to answer was whether the proliferation rate when no drug is applied was a mere environmental confounder or there was an intrinsic measure of growth with a genetic component. I used a mixed effects model to come up with a novel phenotype termed intrinsic growth, ran a genomewide association study to show genetic control of the phenotype. I also found a strong correlation with over 30% of gene expression as well as reliable population and gender differences. 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 all GWAS study results. We 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 that reidentification 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 was published in American Journal of Human Genetics (Im et al, 2012).

Past work in spatial statistics

For a geophysics project, 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).

For a statistical method project, I have developed a new class of covariance functions that shares the benefit of the widely used Matérn 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.

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 Environmentrics (Im et al 2009).

Collaborations

I am currently an analyst for several genetic and genomic consortia

- T2D-GENES consortium (Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples)
- Pharmacogenomics of Anticancer Agents Research group (PAAR)
- Genetics of Tissue Expression (GTEx).

These collaborations allow me to contribute to the discovery of genetic variations that affect disease, response, and their regulatory role. In addition, they allow me to keep current and

to identify the most pressing methodological needs in the field.

My experience at the Biostatistic Laboratory at the University of Chicago gave me first hand exposure to a wide range of biomedical studies, which helped me gain a broader perspective of the field. I collaborated on the design and analysis of several biomedical projects and acquired extensive experience in genome wide association studies, power analysis and sample size estimation, survival analysis, categorical data analysis, mixed and longitudinal regression models. I was the lead statistician for several projects in the Breast Spore and the Head and Neck Spore applications as well as many other grant applications. I was member of the Internal Scientific Advisory Panel (ISAP) for the University of Chicagos Institute for Translational Medicine (CTSA) where I performed the statistical review of numerous proposals for 2.5 years.

Future direction

Recently, I received the Institute for Translational Medicines Paul Calabresi Career Development in Clinical Oncology Award. This will cover at least 75% of my salary for 2 years and give me the opportunity to build an extramurally funded research program under the guidance of stellar mentors such as Nancy Cox, Ron Thisted, and Funmi Olopade. Dan Nicolae and Matthew Stephens have also committed to guide me in my path to become an independent biomedical investigator.

In addition, I have established an international collaboration with the Cross Consortia Pleiotropy group. This group is composed of analysts from large international genetic consortia. This collaboration allowed me to get access to GWAS summary results from inflammatory markers with sample sizes exceeding 10K with a granularity that would not be otherwise available. I plan to keep expanding my network of collaborators with whom I can start cutting edge disease genomic collaborations.

Overall, 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 methods and resources to improve the biological interpretability of genetic discoveries as described below.

Biological dissection of complex traits

More specifically, I will develop methods that correlate genetically predicted biomarkers with complex traits to dissect the underlying biology of the traits. Figure 1 shows a schematic of the approach. Building on the polyomic prediction methods I developed, I will generate predictive models for a range of biomarkers (such as inflammatory markers, metabolic traits, gene expression levels, etc.) and make them available to the research community.

These predictive models will allow us to explore the biology of diseases by taking advantage of the large number of disease and related trait GWAS and sequencing studies available through repositories such as dbGaP. Association between these genetically predicted biomarkers and disease risk can be quantified even though for most studies the actual biomarker levels are not available. A significant correlation of these markers with a given

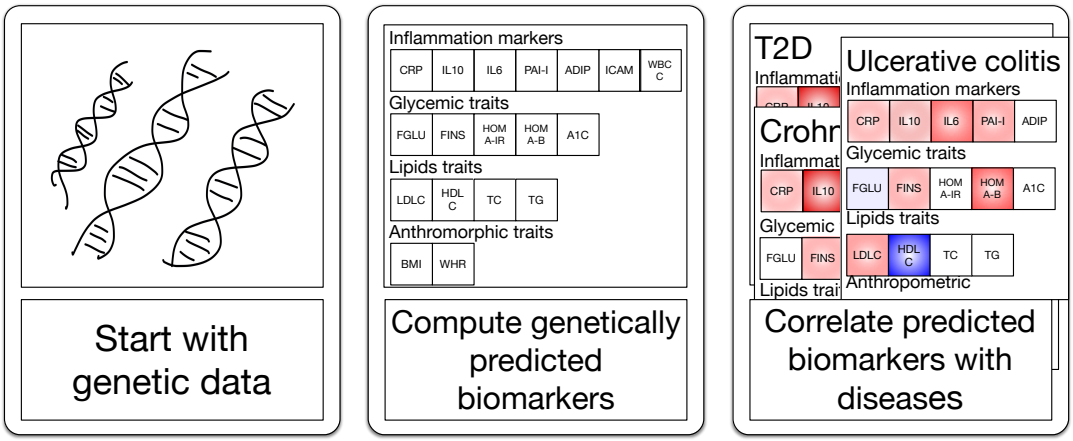

Figure 1: **Schematics of the dissection approach.**

disease will suggest a role a biological mechanism that can be followed up with experimental studies.

The rationale behind the approach is that individuals can be considered to have been 'randomized' to certain level of lifelong 'exposure' (determined by their genetic score) to a biomarker. My approach examines the possible effect of this 'exposure' on disease risk. This is similar to ideas underlying the so called mendelian randomization (MR). The main difference is that my goal is to generate new hypotheses by exploring a large number of biomarkers whereas MR tries to demonstrate a causal link between a given biomarker and disease by using genetic variants as instrumental variables.

Finally, I plan to build software, web resources and virtual machines to facilitate the implementation of this methods to the rest of the research community.