Year 1: Research approach and accomplishments

The main goals of this project are to identify expression quantitative trait loci (eQTL) using high-throughput expression measurements from breast epithelial cells of normal (undiseased) women and use the results to assign function to genomic loci (SNPs) that have been previously implicated in genome wide association studies (GWAS), prioritize SNPs for further validation, establish network models that suggest potential causal relationships among SNPs and downstream phenotypes, and test some of these predictions in mouse models subjected to selected environmental agents.

Our main challenge is that the signal in our expression data is weak. During We have focused attention on expression data from a total of 62 breast tissue samples from young women without a history of breast cancer. Genomic data for each subject includes SNP genotype calls from the Illumina Omni 1M chip and quantitative normalized expression values from the Affymetrix GeneChip Human Gene 1.0 ST Array. Our data include expression levels for 25,927 probes and genotype calls for 1,016,423 SNPs. We have applied several analysis strategies as we persevered in efforts to extract biological insights from our data.

Publication of a report by Ghoussaini et al. in Nature Genetics encouraged us to consider an additional set of SNPs[1]. Ghoussaini et al. shared a list of 74 SNPs that had shown evidence of association with breast cancer status in UK citizens. Specifically, each of the 74 SNPs 1) was genotyped in one or both of two UK breast cancer GWAS, UK2 and British Breast Cancer Study (BBCS) and 2) was associated with breast cancer status (with $p < 0.0001$) in at least one of the two UK GWAS. We were intrigued by the opportunity to use our existing data in efforts to understand the biological roles of the 74 SNPs from Ghoussaini et al.'s report. We hypothesized that these 74 SNPs could be exhibiting associations with breast cancer status due to their potential impacts on a cell's gene expression profile. In other words, we wanted to ask if any of the 74 SNPs were expression quantitative trait loci (eQTL). To address this question, we developed an analysis pipeline to identify and to characterize SNP-expression trait associations.

Before December 31, 2012, we will apply this pipeline to identify cis and trans eQTL among 57 breast cancer-associated SNPs that NHGRI has tabulated in the GWAS Catalog[2].

Year 1: professional development accomplishments

Before December 31, 2012, Fred will have completed all items from the first-year training grant checklist. He convened two committee meetings and will complete Surgical sciences 812 on December 22, 2012. Fred presented at the toxicology seminar in September 2012, the NIEHS training grant poster session in May 2012, and the CIBM retreat poster session in September 2012. We recently (10 December 2012) submitted a F32 proposal to NIH. We eagerly await reviewer feedback.

Year 2 plans

We would like to request a second year of funding from February 1, 2013 to January 31, 2014 to aid Fred's continued development as a biomedical scientist. One of our foci for year 2 is the development of methods for sample size planning in post-GWAS eQTL studies. We emphasized this in our F32 proposal. Our approach involves hierarchical mixture models to accommodate latent and measured variables that impact gene expression. Michael Newton has extensive experience in mixture models, and he will lend expertise to Fred as Fred develops methods for sample size planning purposes. We will use both our pilot genomic data from 62 women and simulation studies to evaluate our sample size needs for eQTL studies.

During year 2, we also will develop a mechanisms for data sharing and for computer code sharing. One possibility is to submit our data and code as R packages for distribution via Bioconductor's repository[3].

Fred is eager to submit a first-author manuscript. This manuscript will include findings from our pilot genomics data, our analysis pipeline, and an overview of the data resource that we are creating. We are happy to provide additional details, to address any questions, and to discuss our project.

Literature cited

1. Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dennis, J., Wang, Q., Humphreys, M. K., Luccarini, C., Baynes, C. & others. Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nature genetics* **44,** 312–318 (2012).
2. Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, Manolio TA. *A Catalog of Published Genome-Wide Association Studies.* at <http://www.genome.gov/gwastudies/>
3. Gentleman, R. C., Carey, V. J., Bates, D. M. & others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5,** R80 (2004).