Statistical Genetics Group
Centre for Addiction and Mental Health
250 College Street, 1st Fl. Room R-32
Toronto, Ontario M5T 1R8

January 19, 2016

Institute of Genetics
Canadian Institutes of Health Research
160 Elgin Street, Room 10-501
Ottawa, Ontario K1A 0W9

To Whom it May Concern:

With the recent radical decrease in genotyping cost, genome wide association studies (GWAS) have become a commonly used tool for the hypothesis free discovery of disease associated loci. Loci identified through GWAS may be validated in biological systems and become potential targets of clinical intervention. Crucial to these studies, which incorporate million of simultaneous statistical tests with a complex depedency structure, is the decision whether to accept or reject the null hypothesis H_0 of no association. As the number of simultaneous statistical tests increases, so too does the number of false positives identified. This issue is denoted as the "multiple comparissons problem" and methods arising from this area are becoming critical to differentiating between statistical coincidence and biological relevance.

One such method which has been gaining traction among the statistical genetics community is "stratefied false disovery rate" (sFDR), which incorporates biologically relevant strata before correcting association P values to control the false discovery rate to an acceptable level. Despite much controversy, the applicability of this methodology to the human gneome remains and open question in the field. It is unknown whether or notthis method adequetely controls the false disovery rate in a system as complex as the human geome, and it is unknown whether or not this method is demonstrably superior to false disovery rate (FDR) in terms of true asocciated loci that it uncovers.

Under my supervision, Christopher Cole will investigate this open problem through a simulation study. Using the reference panel collected by the 1000 Genomes Consortium and the UK10K Consortium, Christopher will simulate novel human genomes while maintaining linkage disequilibrium (LD) structure with HAPGEN2. He will then investigate the ability of various multiple testing correction methodologies to differentiate between random noise and simulated biological signal. He will additionally research common scenarios where researchers use sFDR and FDR and examine how these situations could impact the methodology's accuracy. Christopher will also question under which circumstances (LD, minor allele frequence of variants, etc.) the statistical assumptions and ability to control the FDR are conserved.

Major goals for this study include

1. Developing a computational pipeline which simulates novel genomes and assigns a

disease phenotype in order to compare multiple testing comparison methodologies.
2. Study the efficacy of sFDR under various genomic environments
3. Comapare sFDR to other methodologies in scenarios often encountered by researchers
4. Develop open source tools for release to the research community.

This methodology may possibly identify more truly associated variants and fewer false disoveries, allowing genome wide association studies to identify more and better variants which could have clinical implications. This would allow the field to simultaenously become more reproducible and accurate. The proposed study would be both timely and important in this field, as the paradigm of hypothesis free genome wide association studies shifts to hypothesis testing with methodologies such as sFDR which incorporate biological information into decision making.

Christopher would be directly mentored by myself through weekly progress meetings. In these meetings, we will discuss his progress in the project, cover any topics that he may not be knowledgable about, and fix any logical problems that he may be having while designing his pipeline. Christopher will recieve mentoring on how to use high throughput cluster systems by the resident computer scientists at CAMH, as well as systems administrators on the cluster system. Additionally, Christopher will recieve mentorship from graduate students in the statistical genetics group on issues relating to programming, statistics, and genetics. Additionally, he will be exposed to current research in the field at weekly departmental meetings where graduate students present their research on a variety of issues ranging from the genetics of mental illness to biochemical research to purely statistical studies. He will have the opportunity to communicate and ask questions of these presenters, as well as having the opoportunity to present his own novel findings at the conclusion of his term. There are also seminars offered from a variety of other units at the university, and Christopher will have the ability to attend any number of seminars relating to his interests. Additionally, Christopher will participate in monthly calls with the Psychiatric Genetics Consortium and stay current on developments in the field. He will additionally join in *ad hoc* meetings with other research groups interested in collaboration. We will also have specific meetings with researchers interested in both the genetic and statistical portions of multiple corrections, upon which he will base his research plans.

Funding outside of the amount provided by CIHR shall be provided by the investigator at a minimum of $1250.

Cordially yours,

Jo Knight, PhD
Associate Professor, Department of
Psychiatry, University of Toronto
Joanne Murphy Professor in Behavioural
Science