**Research Goal**

The overarching goal of this study is to discover novel genes harboring genetic variants that contribute to diseases. Traditional genome-wide association studies have interrogated the common allelic architecture for correlation with disease endpoints using statistical approaches designed to detect additive effects. We hypothesize that, for both theoretical1-3 and empirical reasons4-14, new disease-predisposing genes can be identified through the use of study designs that are highly-powered to discover compound heterozygous effects and linkage signals in regions harboring multiple rare disease-susceptibility alleles. With the very large data sets employed, visualization of these approaches is highly critical. Hence, the study will also have the goal of producing user-friendly software that both analyzes and visualizes statistical and genetic results across the genome.

**Specific Aims**

**Specific Aim 1:** To identify a set of diseases in the new set of 8,400 PMRP samples with exome beadchip data. The hypothesis is that we can accurately extract phenotypic information from the EHR to generate a set of case/control sets from which we can evaluate signals identified from the initial study on the same phenotypes.

**Specific Aim 2**: To determine chromosomal regions shared identical-by-descent across the set of samples using haplotype phasing techniques applied to the exome array data. We hypothesize that using the exome beadchip array data, coupled with a localized haplotype clustering Viterbi algorithm, we can both determine haplotype phase across the sample sets examined and determine which chromosomal regions are shared by virtue of being identical-by-descent among all pairwise comparisons.

**Specific Aim 3:** To determine which chromosomal regions replicate a disease-linked signal across the original 10,000 sample set to the new 8,400 sample set. We hypothesize that disease-predisposing haplotypes exist in the PMRP population for the diseases studied; and that through statistically examining the enrichment of shared identity-by-descent chromosomal regions among disease-affected individuals in comparison to control groups, we can identify those regions.

**Specific Aim 4:** To evaluate the statistical significance of compound heterozygosity using putative functional variants at each gene in the exome across the set of clinical traits. We hypothesize that gene-based recessive and/or compound heterozygous effects (on both homologs) contribute to the risk of the clinical traits examined in this study.

**Background**

Genome-wide association studies have been successful at identifying numerous common alleles underlying complex diseases, producing thousands of replicated SNPs.15,16 However, several studies have demonstrated that much of the phenotypic variation attributable to inherited factors has yet to be accounted for.17,18 Substantial literature has been put forth speculating on the sources of this missing heritability.19-21

Several years ago, we speculated that disease genetics models consistent with 1) the observed patterns in genome-wide association studies, 2) the site frequency spectra, and 3) linkage results in multiplex families and affected sibling pairs, would favor high allelic heterogeneity models composed of rare variants of moderately strong penetrances. We further hypothesized that rare alleles, in *trans* configuration, compromising the function of a gene may also enjoy a high degree of plausibility. If these models were indeed underlying complex disease traits, then screening common alleles for additive effects would have inadequate power to reveal a reasonably large fraction of the heritability for typical sample sizes. Monte Carlo simulations and analytic calculations performed by our group and others show that the types of statistical analyses needed to identify causal variants and genes under these models would include those that test for linkage signals in extended kinships and those that explicitly examine compound heterozygosity.

The PMRP is a biobank of 20,000 participants with stored DNA, plasma and sera with accompanying electronic health records, the individuals of which have been drawn from a genetically homogeneous population in Central Wisconsin. In 2012-2013, Dr. Schrodi contributed externally-derived funds and worked with the MCRF core laboratory, Terrie Kitchner and a team of researchers at the University of Michigan to select 10,000 PMRP individuals and subject their DNAs to the Illumina exome beadchip. This genotyping array was designed to have excellent coverage of the exome down to fairly rare allele frequencies. As this genotyping array covers rare variants and the population studied is an extended kinship, this sample set is well-suited for linkage analyses to identify novel disease genes. Additionally, given that this sample set also directly targets the exome, the sample set is also highly germane for a compound heterozygosity scan.

Drs. Maadooliat, Schrodi and Guo have recently performed both types of studies on this initial sample set of 10,000 individuals. We initially developed phenotype algorithms for 13 diseases (cases and controls) using combinations of laboratory test results and ICD-9 codes. These algorithms were implemented by Brent Olson. Previously, we have found that, for numerous diseases, simply using diagnostic codes was insufficiently accurate in identifying cases and controls. Additionally, we explored the impact of this misclassification on genetic association studies, demonstrating a strong inflation in type I and II error rates for marginal levels of misclassification.22 This was the impetus behind supplementing ICD-9 codes with laboratory test data or other medical measurements where available. Subsequently, we investigated both the shared chromosomal regions in those diseases and subjected those data to exome-wide compound heterozygosity scans. These efforts have produced several highly interesting findings with statistical support. That said, replication of these findings in the 8,400 sample set is critical for determining which of our initial findings are true positives. Doing so, will not only generate novel genes involved in the pathogenesis of these diseases but will also validate the identity-by-descent and compound heterozygosity approaches. Neither Dr. Schrodi nor Dr. Maadooliat has research funds. Therefore we cannot support a programmer analyst applying the previously implemented algorithms to the new set of 8,400 PMRP samples to determine which of these individuals has each of these 13 diseases and which can serve as controls.

**Significance**

The discovery of genes involved in the molecular pathogenesis of complex diseases carries a high degree of scientific impact through several avenues: 1) Elucidating the genes underlying diseases provides insight into the biological networks involved in the disease, often revealing new pathways involved in these traits; 2) Disease genes or those interacting with disease genes can serve as targets for therapeutic development. Highly targeted therapies often are efficacious while minimizing adverse reactions; and 3) Disease variants can serve as features in diagnostic panels. For example, we have previously conducted experiments resulting in the discovery of genetic variants residing in the IL-23/IL-17 axis, strongly suggesting the actions of Th17 signaling plays a critically important role in autoinflammatory disease susceptibility.23-27 This knowledge complemented parallel efforts to apply anti-p40 and anti-IL17 monoclonal antibodies as a treatment to remediate psoriasis, Crohn’s disease and ankylosing spondylitis.28-35

In addition to the impact of the resulting biological findings from this investigation, the study will show how these two novel statistical/genetic approaches (compound heterozygosity screening and case/control identity-by-descent mapping) can be applied to large-scale genotype data. As numerous studies have generated GWAS data sets across a wide-variety of clinically-relevant traits, the re-analysis of these datasets in ways highly-powered to detect other plausible disease models is greatly anticipated by the human genetics community. Moreover, user-friendly visualization and analysis tools applied to these data sets are currently not available. This study will 1) test two aspects of exome-wide genotype association data—the impact of compound heterozygosity and linkage signals detected by shared chromosomal regions–both previously understudied in the community; and 2) fund the development of an interactive software package to visualize shared chromosomal regions across the genome and display statistical results for association and linkage of those regions with disease states.

**Preliminary Studies**

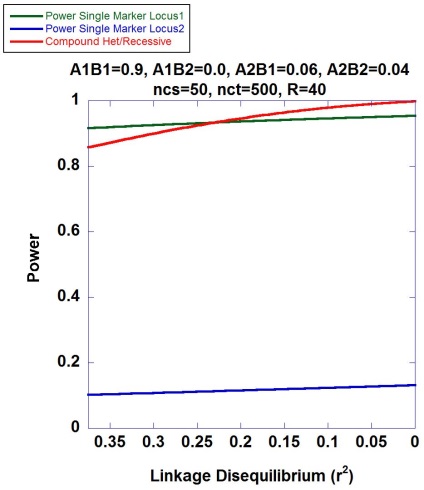
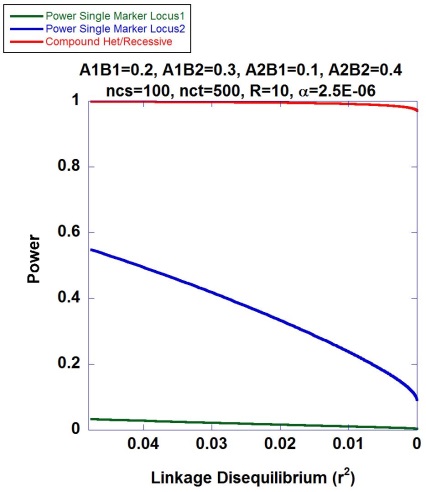
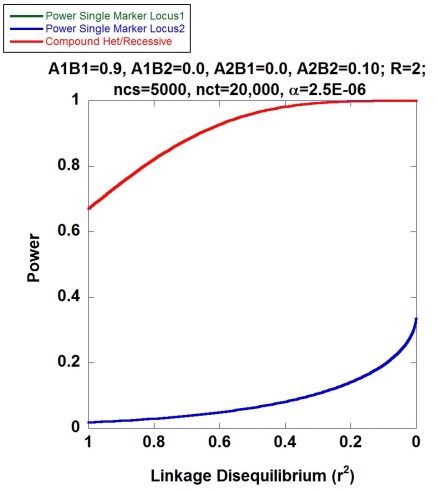
Drs. Maadooliat, Schrodi and Guo have completed several preliminary aspects of the study, including 1) power calculations of statistical test of compound heterozygosity under a wide spectrum of the parameter space in comparison with orthodox statistical tests used in GWAS, 2) quality control analyses using principal components analysis on the 8,400 PMRP samples with exome beadchip genotype data, 3) the development of a preliminary software tool that performs basic analyses of identity-by-descent shared chromosomal region analyses and visualization across the genome, 4) haplotype phasing of all individuals from the initial phase of PMRP samples genotyped on the exome beadchip (n=10,000) and subsequent exhaustive determination of pairwise identity-by-descent regions, 5) conducting an exome-wide compound heterozygote scan for iron overload in the initial phase of PMRP samples, and 6) conducting an exome-wide scan of shared chromosomal regions in rheumatoid arthritis cases compared to controls using the initial phase of PMRP samples.

***Power calculations***

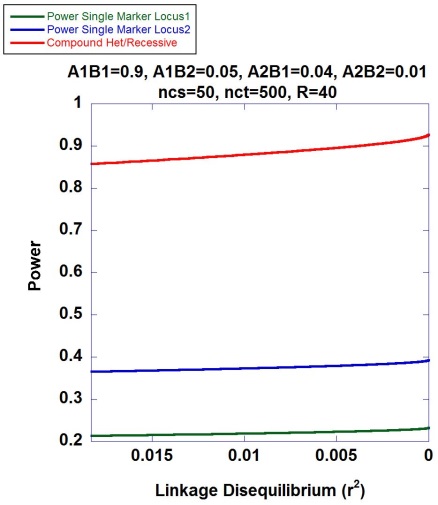
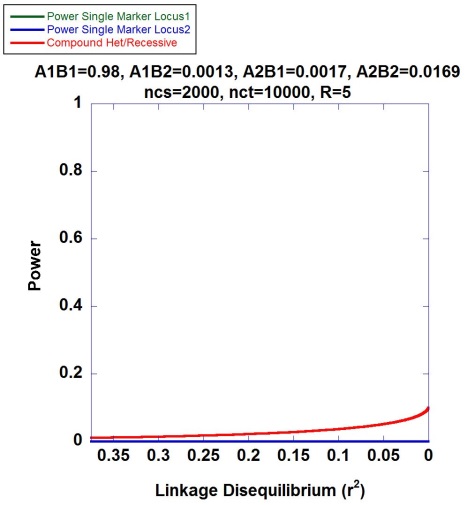
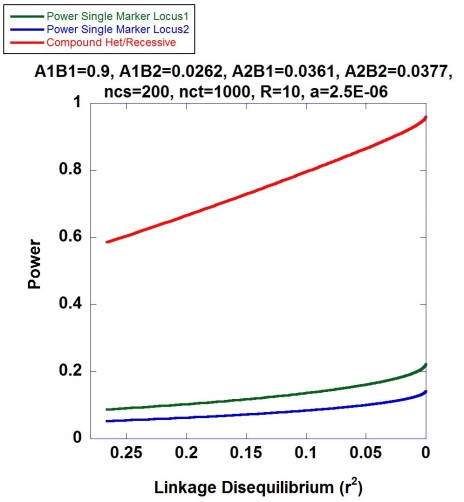
To investigate the power of tests specifically designed to detect disease association signals from compound heterozygous effects, we developed a two-locus model and performed analytic calculations. The motivation for doing so was to evaluate the performance of statistical approaches specifically designed to interrogate models of recessive diplotype effects compared to standard methods of analyzing genome-wide association studies. As previously discussed, recessive diplotype effects are plausible from both theoretical and experimental study results. We sought to investigate whether recessive diplotype effects would remain hidden to standard methods while new statistical tests for compound heterozygous configurations would reveal such effects. The model evaluated for the power calculations is a two-site model with two alleles segregating at each site. We further stipulated that one of the two alleles at each site has functional consequences with regard to a binary trait, whereas the other allele is wildtype. This results in four haplotypes and ten diplotypes. We classified these diplotypes into two classes: those with one or both haplotype(s) free of either differentially functional allele (i.e., at least one haplotype with both wildtype alleles), and those with both haplotypes carrying at least one differentially functional allele. These two classes of diplotypes were assigned two different trait penetrances. We evaluated a log-likelihood ratio test designed to test for recessive diplotype effects and compared the results to the standard Armitage trend test applied separately to genotypes at each of the sites. Both tests were calibrated by setting both penetrances equal to the trait prevalence, reflecting the null model of stochastic independence between the trait and diplotype configuration, and calculating the value of the test statistic that reflects the *a priori* significance level (=2.5E-06 across the calculations). Case and control numbers were parameters. The disease model consisting of haplotype frequencies and unequal penetrances is then applied to calculate statistical power. The Haldane recombination model was applied to generate a series of haplotype sets as summarized by different linkage disequilibrium values and Hardy-Weinberg equilibrium is used to generate diplotypes in the general population. A large proportion of the parameter space was explored and across almost all of the space, the power of the recessive diplotype log-likelihood ratio test exceeded that of either single marker tests. **Figure 1** shows power for the log-likelihood ratio test for numerous recessive diplotype disease models. The initial haplotype frequencies (prior to recombination) are shown in each panel along with a relative risk parameter (the ratio of the two penetrances) and the sample sizes. These results clearly show that many recessive diplotype models may remain recondite following the application of traditional GWAS statistical analyses.

**Figure 1: Power curves comparing the log-likelihood ratio test compared to single marker tests**

**Panel 1a Panel 1b Panel 1c**

****

**Panel 1d Panel 1e Panel 1f**

**** **** 

***Statistical Methodology Development***

WES and WGS-based data are now available in public databases.36,37 Using empirical sequencing data allows one to recapitulate realistic genetic variation patterns for use in theoretical studies, Bayesian analyses and power calculations. For better understanding the design and analyses of data stemming from compound heterozygosity scans and shared chromosomal analyses applied to mapping disease genes, we have derived scaling factors which transform sequence data from general population samples and produces generalized disease models for a recessive diplotype mode of inheritance.

are the individual haplotypes that confer functional effects, collectively they are

are the individual haplotypes that are wildtype, collectively they are

and are the frequencies of the and haplotypes, respectively

and are the frequencies of the and haplotype classes, respectively

indicates disease phenotypes; indicates control phenotypes

is the disease prevalence

and are the penetrances of the wildtype and pathogenic diplotypes, respectively

Scaling factors:

We have also started to develop the statistical machinery for using unphased data for the purpose of conducting recessive diplotype analyses without necessitating haplotype phasing. It is known that determining gametic phase of rare alleles is error-prone. Under panmixis, rare variants are also typically in linkage equilibrium as they arise through mutation and occur on an existing haplotype in a population at random.38,39 Hence, assuming equal probabilities of a compromised function variant falling on either homologous chromosome, then the probability that an individual has a compromised function diplotype is , where is the number of functional variants that the individual carries at the gene in question. Therefore, one can populate a 2x2 contingency table testing the association of classified diplotypes with a binary disease status with the following values:

is the number of cases with compromised function diplotypes

is the number of controls with compromised function diplotypes

is the number of cases with wildtype diplotypes

is the number of controls with wildtype diplotypes

is the number of diploid cases

is the number of diploid controls

is the number of cases carrying no compromised function alleles

is the number of controls carrying no compromised function alleles

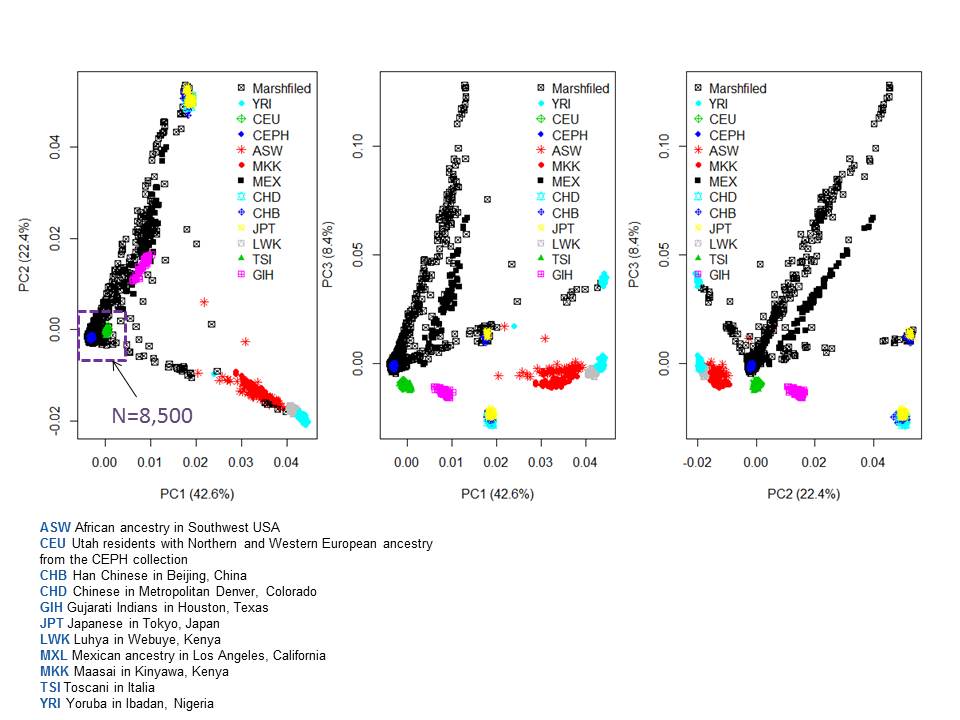
The log likelihood statistic for the contingency table is:

, , ,

***Quality Control***

We have commenced our initial statistical processing of the exome genotype data from the 8,400 individuals. As ancestral differences in samples studied can produce inflated type I and type II error rates through confounding by population stratification, a critically important step is to either remove individuals with outlier genetic backgrounds and/or use genetic background signals as covariates in statistical models to test for association with disease. Hence, we performed dimensional reduction through principal components analysis40,41 using the set of exome beadchip genotypes on the 8,400 PMRP sample set along with HapMap samples42 from known ancestries. **Figure 2** displays PC1 vs. PC2, PC1 vs. PC3, and PC2 vs. PC3 from this analysis. Using the first two principal components, we found that the large majority of the 8,400 PMRP individuals (>98%) were tightly clustered indicating a genetically homogeneous population, consistent with historical data, self-reported ethnicity data, and previous genetic studies on the PMRP. Therefore, we anticipate the confounding effects by population stratification to be minor. Additionally, this result suggests that the genetic architecture of diseases in this population will exhibit enhanced signal for some disease-predisposing alleles, while other predisposing alleles from external studies may be absent in our population. That is, the disease heritability will be concentrated in a subset of susceptibility genes.

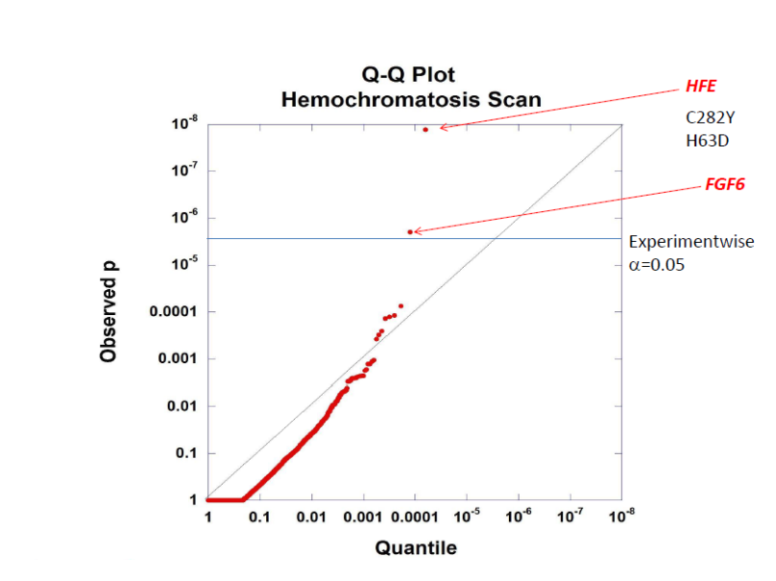
**Figure 2: PCA Results for the 8,400 PMRP Individuals and HapMap Samples**

******

***Compound Heterozygosity Scan for Iron Overload***

Drs. Guo, Schrodi and Maadooliat have completed the exome-wide, gene-based scans of compound heterozygosity for several diseases using the set of 10,000 PMRP samples with exome beadchip genotype data. The disease studied the most extensively in this first phase was hemochromatosis.43-46 Drs. Mazza and Epperla aided in the clinical aspects of iron overload conditions. PMRP individuals with exome beadchip data were selected as iron-overload cases on the basis of percent transferrin saturation laboratory values (i.e., the ratio of serum iron to transferrin iron-binding capacity) exceeding 48% and having two or more instances of ICD-9 codes indicating the diagnosis of hemochromatosis: 275.0 (iron metabolism disorder, excluding anemia), 275.01 (hereditary hemochromatosis), 275.03 (unspecified hemochromatosis), and/or 275.09 (other iron metabolism disorders). Using a principal components analysis of the exome genotyping data, individuals considered genetic background outliers were excluded from the study. Quality control measures for the genotyping data were employed as previously described.47 Putative functional variants included GWAS-significant markers, variants annotated as nonsynonymous, or residing in 3’UTR, 5’UTR or splice site regions. The next step in the study was to determine gametic phase for each individual at each gene. The localized haplotype-cluster model algorithm implemented in the software Beagle was used to accomplish this.48 All individuals in the 10,000 sample set were haplotype phased using Beagle, blinded to phenotype data. Following phasing, the first two steps to conducting our compound heterozygote scans were to define the genomic regions of each gene and to identify the putative functional variants within each of these regions. All cases and controls were placed in one of two categories for each gene. If a subject carried at least one putative functional allele on each homolog at a particular gene, they were deemed as having a diplotype of functional consequence; if they did not, then they were placed in the wildtype diplotype category. This binary variable was used to populate a contingency table against case/control status for each gene and statistical evaluation of the contingency table was performed by a two-tailed Fisher’s exact test, testing the null hypothesis of stochastic independence between the frequency of putative functional diplotypes and case/control status. In all, 18k genes were evaluated for hemochromatosis. The quantile-quantile plot shown in **Figure 3** reveals the strong association of *HFE*, a well-established positive control for iron overload disorders.49-53 This finding provides proof-of-principle evidence for our study design and analyses. Additionally, *FGF6* also reached experiment-wise statistical significance. Both findings also exceed a generalized family-wise significance level.54

*FGF6* encodes for the fibroblast growth factor 6, a protein critically important in numerous cellular functions including myoblast proliferation, angiogenesis, embryonic development, tissue repair and tumor invasion.55-57 Importantly, Fgf6 binds heparin—an inhibitor of hepcidin expression.58-60 Hepcidin is a small polypeptide primarily manufactured in hepatocytes and serves as a potent regulatory hormone for iron metabolism.61-63 Hepcidin blocks ferroporin-mediated iron transport in intestinal enterocytes and macrophages, thereby promoting cellular iron sequestration.64 Variants in *HAMP*, the hepcidin-encoding gene, and other hepcidin-regulating proteins are known causes of juvenile hemochromatosis.62,65,66

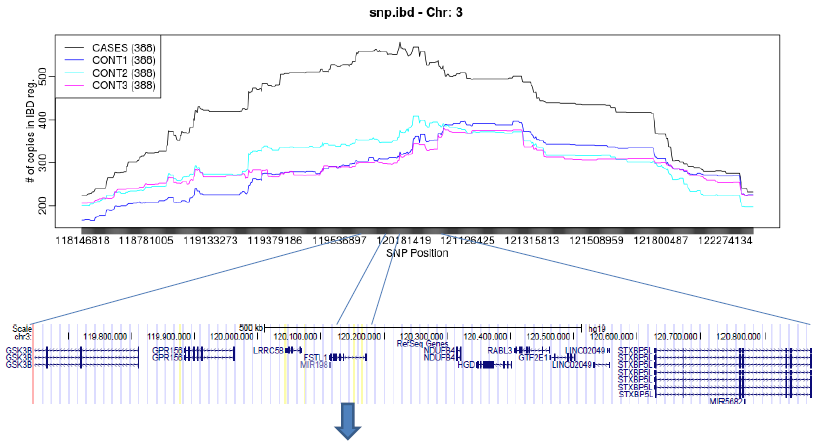
****

**Figure 3: Q-Q Plot of Compound Het Scan for Iron Overload Overload**

***Identity-by-Descent Mapping of Rheumatoid Arthritis***

A large fraction of the source population of the PMRP—the Central Wisconsin population—is an extended kinship. Drs. Maadooliat and Schrodi have utilized the latent relatedness structure of this cohort to conduct a disease gene mapping study by identifying genomic regions that exhibit increased chromosomal sharing, identical by descent, in cases compared to controls. This approach has been used previously and is designed to discover disease genes segregating multiple susceptibility alleles, particularly rare alleles.67,68 Importantly, this method enables one to capture both linkage and association signals. This method is promising on theoretical grounds as the site frequency spectra is decidedly enriched for rare variants in outbred and expanding populations. Dr. Maadooliat has started the development of a software package that uses phased data and pairwise identity-by-descent regions as input, performs processing of those data, and displays the results in an interactive visualization tool. This tool has been applied to rheumatoid arthritis cases and three independent sets of controls, all derived from the 10,000 PMRP individuals with exome array genotype data. Phasing and pairwise identity-by-descent calls were obtained through Beagle.48 **Figure 4** displays a region on chr 3q13.33 exhibiting a statistically significant excess of chromosomal sharing among rheumatoid arthritis cases (388 individuals) compared to all three control sets (each of 388 individuals without any autoimmune disease within their electronic health record). Under the peak departure between pairwise IBD sharing among cases compared to controls is the *FSTL1* gene.

**Figure 4. Excess of Shared Chromosomal Regions for Rheumatoid Arthritis Compared to Controls**

****

***FSTL1***

*FSTL1* encodes for the follistatin-like protein 1. The protein is expressed in synovial fluid, monocyotes, lymph nodes and peripheral blood. Experimental studies have shown interaction of Fstl1 with the LPS-binding CD14/TLR4 receptor complex on macrophages, neutrophils and dendritic cells. Interesting, anti-Fstl1 antibodies have been observed in synovial fluid and sera of rheumatoid arthritis patients69, leading these authors to suggest that Fstl1 is involved in systemic rheumatic disease molecular pathophysiology. Moreover, the *Mus* ortholog of Fstl1 is upregulated in the collagen-induced arthritis mouse model over controls.70 Drs. Maadooliat, Schrodi and Guo have started a collaboration with Drs. Hirsch and Chaly at the University of Iowa. Drs. Hirsch and Chaly are conducting ongoing functional research in Fstl1 biology and its role in rheumatoid arthritis pathobiology.71-74 Replication of this and other findings from the initial shared chromosomal region scan in the 8,400 PMRP sample set will provide additional evidence for involvement in rheumatoid arthritis susceptibility.

**Research Design and Methodology**

Our team has successfully completed several initial analyses on the initial set of 10,000 exome-genotyped PMRP samples resulting in several novel findings including the experiment-wide significant association of functional diplotypes at *FGF6* with hemochromatosis and significant enrichment in chromosomal linkage signals over *FSTL1* for rheumatoid arthritis compared to controls. Both of these investigations have devised and applied novel statistical genomics techniques that are designed to reveal genetic effects expected to remain obscured to traditional GWAS analyses. In addition, we have also developed an initial software package that analyzes and visualizes shared chromosomal regions identical-by-descent. To test for replication of these findings, our goal is to use the set of 8,400 PMRP samples recently subjected to genotyping via the Illumina exome beadchip. As neither PI of this proposal has access to research funds to conduct this replication work, we are applying for funds to support these efforts.

**Specific Aim 1:** To identify a set of diseases in the new set of 8,400 PMRP samples with exome beadchip data. The hypothesis is that we can accurately extract phenotypic information from the EHR to generate a set of case/control sets from which we can evaluate signals identified from the initial study on the same phenotypes.

To address Specific Aim 1, we will work with a research programmer analyst to apply the previously developed phenotyping algorithms (applied to the set of 10,000 exome genotyped PMRP individuals several years ago) to the set of 8,400 PMRP individuals recently genotyped on the Illumina exome beadchip. The diseases studied include:

Rheumatoid arthritis, including seropositive rheumatoid arthritis as a separate subgroup

Ankylosing spondylitis

Psoriatic arthritis

Relapsing-remitting multiple sclerosis

Autoimmune thyroid disease

Systemic sclerosis

Iron metabolism disease, overload

Premature acute myocardial infarction

Endometriosis

Super morbid obesity

Obsessive compulsive disorder

Extreme hyperlipidemia

Chronic lymphocytic leukemia

Additionally, algorithms to define control sets for each of the above diseases were developed by Dr. Schrodi and previously implemented by Brent Olson. We propose that a research programmer analyst in data analytics at MCRI re-run the algorithms on the 8,400 PMRP samples, thereby defining the clinically-relevant phenotypes listed above.

**Specific Aim 2**: To determine chromosomal regions shared identical-by-descent across the set of samples using haplotype phasing techniques applied to the exome array data. We hypothesize that using the exome beadchip array data, coupled with a localized haplotype clustering Viterbi algorithm, we can both determine haplotype phase across the sample sets examined and determine which chromosomal regions are shared by virtue of being identical-by-descent among all pairwise comparisons.

To commence our genetic analyses of the 8,400 samples, we will filter out individuals having genetic background data indicating poor clustering with the primary—most frequent—subpopulation. A combination of principal components analysis and the application of a population genetics Bayesian classifier yielding posterior probabilities of classification to subpopulations will be used. The ***Quality Control*** section above describes the results of the PCA. Individuals and SNPs/SNVs failing to meet genotyping quality measures have previously been removed from these data. Following the determination of a genetically homogeneous set of individuals, we will apply the localized haplotype clustering Viterbi algorithm as implemented in the software package Beagle to determine gametic phase across the entire set of genetically homogeneous individuals using their exome genotype data. The same software will then be used to concurrently determine pairwise shared chromosomal regions identical-by-descent. The code to run these algorithms on large sets of exome genotype data using the High Performance Cluster at MCRI has previously been written by Dr. Maadooliat and applied to the initial set of 10,000 exome-genotyped PMRP samples.

**Specific Aim 3:** To determine which chromosomal regions replicate a disease-linked signal across the original 10,000 sample set to the new 8,400 sample set. We hypothesize that disease-predisposing haplotypes exist in the PMRP population for the diseases studied; and that through statistically examining the enrichment of shared identity-by-descent chromosomal regions among disease-affected individuals in comparison to control groups, we can identify those regions.

Following the determination of all pairwise shared chromosomal regions, identical-by-descent from **Specific Aim 2**, we will utilize the preliminary software that Dr. Maadooliat has written to display and analyze case and control shared chromosomal regions across the genome. Test statistics including a Fisher’s exact test comparing the level of sharing between case and control sets will be calculated in this computational framework and displayed separately for each chromosome. A permutation routine will also be applied to these data to account for the stochastic dependence in observations based on pairwise comparisons. A preliminary version of this permutation routine has previously been implemented by Dr. Maadooliat to obtain appropriate p-values accounting for dependency within regions that appear to house compelling signals. For each disease, the results from this permutation will be used to identify regions with a high degree of statistical support for linkage. Following, theoretical results from Skol et al.75 and Sha et al.76, we will conduct a joint analysis across the results from both phases of the exome genotyped sets. Additional analyses using age and sex as covariates in a regression model may also be performed if preliminary analyses necessitate this approach. Drs. Maadooliat, Schrodi and Guo will also further work on developing the software to conduct these analyses and visualize the results.

**Specific Aim 4:** To evaluate the statistical significance of compound heterozygosity using putative functional variants at each gene in the exome across the set of clinical traits. We hypothesize that gene-based recessive and/or compound heterozygous effects (on both homologs) contribute to the risk of the clinical traits examined in this study.

Using the set of gametic phased 8,400 PMRP samples, we will conduct a scan for compound heterozygosity association with the defined diseases. As with the initial phase of these analyses on the 10,000 samples, putative functional variants will include GWAS-significant markers, variants annotated as nonsynonymous, or residing in 3’UTR, 5’UTR or splice site regions. For regions with an excess of very rare variants, we will use the approach described in the previous methodology development section using unphased data to estimate the number of individuals with recessive diplotypes within cases and controls. Additional analyses using age and sex as covariates in a logistic regression model will be performed. To assess the significance level for results in the replication set, we will employ a permutation routine. Similar to the analyses to address **Specific Aim 3**, we will take a joint analysis approach to combine statistical evidence across both sample sets (10,000 and 8,400).

**Timeline**

**Budget**

**Budget Justification**

Research Programmer Analyst. The research programmer analyst will apply the disease algorithms to the set of 8,400 individuals (reduced in number after removing the QC and genetic background outliers).

Robert Strenn. Mr. Strenn will merge the disease indicator variables with exome genotype data.

Dr. Shicheng Guo, Co-Investigator. For Years 1 and 2, Dr. Guo will work on data management, quality control, exome data analysis, application of novel methods to these data, biological interpretation of the results, and aid in manuscript preparation.

Dr. Mehdi Maadooliat, Co-Principal Investigator. Dr. Maadooliat will provide expertise on the statistical design and data analysis of the study in Year1. In Year 2, Dr. Maadooliat’s focus will be on statistical method development, programming and manuscript writing.

Dr. Steven Schrodi, Co-Principal Investigator. For Years 1 and 2, Dr. Schrodi will coordinate the various activities of the study, aid in the development of novel methods, aid with exome data analysis, aid in interpretation of results and prepare one or more manuscripts.

**Literature Cited**

1. Singleton AB, Hardy J, Traynor BJ, Houlden H (2010) Towards a complete resolution of the genetic architecture of disease. Trends Genet 26(10):438-442.
2. Schrodi SJ (2016) Reflections on the field of human genetics: A call for increased disease genetics theory. Front Genet 7:106.
3. Sanjak JS, Long AD, Thornton KR (2017) A model of compound heterozygous, loss-of-function alleles is broadly consistent with observations from complex-disease GWAS datasets. PLoS Genet 13(1):e1006573.
4. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133(1):1-9.
5. Gong S, Ware JS, Walsh R, Cook SA (2014) NECTAR: a database of codon-centric missense variant annotations. Nucleic Acids Res 42(Database issue):D1013-D1019.
6. MacArthus DG, Balasubramanian S, Frankish A, Huang N, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335(6070):823-828.
7. Zou J, Valiant G, Valiant P, Karczewski K, et al. (2016) Quantifying unobserved protein-coding variants in human population provides a roadmap for large-scale sequencing projects. Nat Commun 7:Article number 13293.
8. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305(5685):869-872.
9. Andreoletti G, Shakhnovich V, Christenson K, Coelho T, et al. (2017) Exome analysis of rare and common variants within the NOD signaling pathway. Sci Report 7:46454.
10. Khetarpal SA, Schjoldager KT, Christoffersen C, Raghavan A, et al. (2016) Loss of function of GALNT2 lowers high-density lipoproteins in humans, nonhuman primates, and rodents. Cell Metab 24(2):234-245.
11. Singh T, Kurki MI, Curtis D, Purcell SM, et al. (2016) Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat Neurosci 19(4):571-577.
12. Adam R, Spier I, Zhao B, Kloth M, et al. (2016) Exome sequencing identifies biallelic MSH3 germline mutations as a recessive subtype of colorectal adenomatous polyposis. Am J Hum Genet 99(2):337-351.
13. Hague S, Rogaeva E, Hernandez D, Gulick C, et al. (2003) Early-onset Parkinson’s disease cause by a compound heterozygous DJ-1 mutation. Ann Neurol 54(2):271-274.
14. Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, et al. (2009) Multicenter analysis of glucocerebrosidase mutations in Parkinson’s disease. N Engl J Med 361:1651-1661.
15. <https://www.ebi.ac.uk/gwas/>
16. Visscher PM, Brown MA, McCarthy MI, Yang J. (2012) Five years of GWAS discovery. Am J Hum Genet 90:7-24.
17. Manolio TA, Collins FS, Cox NJ, Goldstein DB, et al. (2009) Finding the missing heritability of complex diseases. Nature 461(7265):747-753.
18. Lee SH, Wray NR, Goddard ME, Visscher PM. (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88(3):294-305.
19. Gibson G (2012) Rare and common variants: Twenty arguments. Nat Rev Genet 13(2):135-145.
20. Zuk O, Hechter E, Sunyaev SR, Lander ES. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109(4):1193-1198.
21. Locke AE, Kahali B, Berndt SI, Justice AE, et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518:197-206.
22. Schrodi SJ (2017) The impact of diagnostic code misclassification on optimizing the experimental design of genetic association studies. J Healthc Eng 2017:7653071.
23. Begovich AB, Schrodi SJ, Leppert M, Krueger G, Cargill M (2006) A genome-wide association study of putative functional SNPs leads to the identification of two psoriasis loci – IL12B and IL23R – in 3 independent white North American sample sets. ASHG Abstract 163.
24. Cargill M, Schrodi SJ, Chang M, Garcia VE, et al. (2007) A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. Am J Hum Genet 80(2):273-290.
25. Garcia VE, Chang M, Brandon R, Li Y, et al. (2008) Detailed genetic characterization of the interleukin-23 receptor in psoriasis. Genes Immun 9(6):546-555.
26. Schrodi SJ. (2008) Genome-wide association scan in psoriasis: new insights into chronic inflammatory disease. Expert Rev Clin Immunol 4(5):565-571.
27. Nair RP, Duffin KC, Helms C, Ding J, et al. (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41(2):199-204.
28. Krueger GG, Langley RG, Leonardi C, Yeilding N, et al. (2007) A human interleukin-12/23 monoclonal antibody for the treatment of psoriasis. N Engl J Med 356:580-592.
29. <http://www.clinicaltrials.gov/ct2/show/study/NCT00955279>
30. <http://www.clinicaltrials.gov/ct2/show/NCT01009086>
31. <http://www.clinicaltrials.gov/ct2/show/results/NCT00265122>
32. <http://www.clinicaltrials.gov/ct2/show/NCT01945086>
33. <http://www.clinicaltrials.gov/ct2/show/NCT00771667>
34. <http://www.clinicaltrials.gov/ct2/show/NCT01647152>
35. <http://www.clinicaltrials.gov/ct2/show/NCT01330901>
36. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68-74.
37. Lek M, Karczewski KJ, et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285-291.
38. Pritchard J (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124-137.
39. Thornton KR, Foran AJ, Long AD (2013) Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genet 9(2):e1003258.
40. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2(11):559-572.
41. Purcell S, Neale B, Todd-Brown K, Thomas L, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81(3):559-575.
42. International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-862.
43. Adams PC, Barton JC (2007) Haemochromatosis. Lancet 370:1855-1860.
44. Andrews NC, Schmidt PJ (2007) Iron homeostasis. Annu Rev Physiol 69:69-85.
45. Andrews NC (1999) Disorders of iron metabolism. N Engl J Med 341(26):1986-1995.
46. Waldvogel-Abramowski S, Waeber G, Gassner C, Buser A, et al. (2014) Physiology of iron metabolism. Transfus Med Hemother 41(3):213-221.
47. Fritsche LG, Igl W, Bailey JN, Grassmann F, et al. (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet 48(2):134-143.
48. Browning SR, Browning BL. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81(5):1084-1097.
49. Cartwright GE, Skolnick M, Amos DB, Edwards CQ, Kravitz K, Johnson A (1978) Inheritance of hemochromatosis: linkage to HLA. Trans Assoc Am Physicians 91:273-281.
50. Edwards CQ, Griffen LM, Dadone MM, Skolnick MH, Kushner JP (1986) Mapping the locus for hereditary hemochromatosis: localization between HLA-B and HLA-A. Am J Hum Genet 38(6):805-811.
51. Jazwinska EC, Lee SC, Webb SI, Halliday JW, Powell LW (1993) Localization of the hemochromatosis gene close to D6S105. Am J Hum Genet 53(2):347-352.
52. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, et al. (1996) A novel MHC class-I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13(4):399-408.
53. Jazwinska EC, Cullen LM, Busfield F, Pyper WR, et al. (1996) Haemochromatosis and HLA-H. Nat Genet 14(3):249-251.
54. Schrodi SJ (2016) The use of multiplicity corrections, order statistics and generalized family-wise statistics with application to genome-wide studies. PLoS One 11(4):e0154472.
55. Zhang X, Ibrahimi OA, Olsen SK, Umemori H, et al. (2006) Receptor specificity of the fibroblast growth factor family. J Biol Chem 281(23):15694-15700.
56. Armand AS, Lecolle S, Launay T, Pariset C, et al. (2004) IGF-II is up-regulated and myofibres and hypertrophied in regenerating soleus of mice lacking FGF6. Exp Cell Res 297(1):27-38.
57. Armand AS, Laziz I, Chanoine C (2006) FGF6 in myogenesis. Biochim Biophys Acta 1763(8):773-778.
58. Li Y, Sun C, Yates EA, Jiang C, et al. (2016) Heparin binding preference and structures in the fibroblast growth factor family parallel their evolutionary diversification. Open Biol 6(3):150275.
59. Guerrini M, Hricovini M, Torri G (2007) Interaction of heparins with fibroblast growth factors: conformational aspects. Curr Pharm Des 13(2):2045-2056.
60. Poli M, Girelli D, Campostrini N, Maccarinelli F, et al. (2011) Heparin: a potent inhibitor of hepcidin expression in vitro and in vivo. Blood 117:997-1004.
61. Nicolas G, Viatte L, Lou D-Q, Bennoun M, et al. (2003) Constitutive hepcidin expression prevents iron overload in a mouse model of hemochromatosis. Nat Genet 34:97-101.
62. Roetto A, Papanikolaou G, Politou M, Alberti F, et al. (2003) Mutant antimicrobial peptide hepcidin is associated with severe juvenile hemochromatosis. Nat Genet 33:21-22.
63. Rossi E (2005) Hepcidin—the iron regulatory hormone. Clin Biochem Rev 26(3):47-49.
64. Gulec S, Anderson GJ, Collins JF (2014) Mechanistic and regulatory aspects of intestinal iron absorption. Am J Physiol Gastrointest Liver Physiol 307(4):G397-G409.
65. Core AB, Canali S, Babitt JL (2014) Hemojuvelin and bone morphogenetic protein (BMP) signaling in iron homeostasis. Front Pharmacol 5:104.
66. Babitt JL, Huang FW, Wrighting DM, Xia Y, et al. (2006) Bone morphogenetic protein signaling by hemojuvelin regulates hepcidin expression. Nat Genet 38(5):531-539.
67. Browning SR, Thompson EA (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190:1521-1531.
68. Grant GR, Manduchi E, Cheung VG, Ewens WJ (1999) Significance testing for direct identity-by-descent mapping. Ann Hum Genet 63(Pt 5):441-451.
69. Tanaka M, Ozaki S, Osakada F, Mori K, et al. (1998) Cloning of follistatin-related protein as a novel autoantigen in systemic rheumatic diseases. Int Immunol 10(9):1305-1314.
70. Miyamae T, Marinov AD, Sowders D, Wilson DC, et al. (2006) Follistatin-like protein-1 is a novel proinflammatory molecule. J Immunol 177(7):4758-4762.
71. Chaly Y, Hostager B, Smith S, Hirsch R (2014) Follistatin-like protein 1 and its role in inflammation and inflammatory diseases. Immunol Res 59(1-3):266-272.
72. Wilson DC, Marinov AD, Blair HC, Bushnell DS, et al. (2010) Follistatin-like protein 1 is a mesenchyme-derived inflammatory protein and my represent a biomarker for systemic-onset juvenile rheumatoid arthritis. Arthritis Rheum 62(8):2510-2516.
73. Chaly Y, Marinov AD, Oxburgh L, Bushnell DS, Hirsch R (2012) FSTL1 promotes arthritis in mice by enhancing inflammatory cytokine/chemokine expression. Arthritis Rheum 64(4):1082-1088.
74. Chaly Y, Fu Y, Marinov A, Hostager B, et al. (2014) Follistatin-like protein 1 enhances NLRP3 inflammasome-mediated IL-1B secretion from monocytes and macrophages. Eur J Immunol 44(5):1467-1479.
75. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38(2):209-213.
76. Sha Q, Zhang Z, Zhang S (2011) Joint analysis for genome-wide association studies in family-based designs. PLoS One 6(7):e21957.