Project period 09/01/1991 to 1/31/2020

Over the 29 years of this award we have developed methods for the analysis of data on related individuals to make inferences about the epidemiology of traits with a substantial genetic component. From the initial R01 award in 1991, there were 4 R01 4-year renewals; the award then became an R37 MERIT award in 2008, and renewed as an R37 for a final 5 years in 2013. Over the years, there has been a major transformation in computing power and technology, and a much greater transformation in the type and availability of genetic data, but gene localization and gene-effect estimation have remained a constant challenge for complex traits.

In the 1990s, data were polymorphic microsatellite (STR) loci, and we developed Markov chain Monte Carlo (MCMC) methods for joint segregation and linkage analysis from data on extended pedigrees and for estimating parameters of genetic models. Using, for that date, large numbers of genetic markers, methods were applied to Mendelian disorders and quantitative traits (lipid levels). They were used to study environmental and genetic risk factors for coronary heart disease, and applied in pedigrees segregating Alzheimer's disease.

Around 2000, marker data at increasing densities became available, MCMC methods were further developed to handle more complex traits and more complex pedigrees. Allelic association and multilocus haplotypes were studied at both the population and pedigree level, with methods addressing map accuracy, recombination heterogeneity and genetic interference. Applications now included not only Alzheimer's disease and cardiovascular disease, but also alcoholism and prostate cancer.

As SNP marker data became increasingly available (2003-7), and more complex diseases with both environmental and genetic components such as hypertension and psychiatric illness were studied, new MCMC methods were needed to overcome the statistical and computational constraints. Considering both Bayesian, and likelihood-based joint segregation and linkage analysis, the focus of inference changed to estimation of haplotypic identity by descent (IBD), and the used of this inferred IBD in linkage detection and gene localization.

These methods were further developed in 2007-12, with a continuing focus on gene localization, trait model estimation, haplotype analysis, and genetic map analysis, using data on extended pedigrees. Improved MCMC methods allowed inference of gene descent using a dense genome screen of markers, and methods were developed to use this inferred descent on joint multilocus linkage and segregation analyses of discrete and quantitative complex trait phenotypes. Methods for conditional inference, re-simulation and permutation approaches, and multiple-testing corrections were developed, assessed and applied. In addition to assessing the impact of multi-SNP haplotypes, copy-number variants, marker-map and model uncertainties, linkage disequilibrium (LD) became a

major focus, with assessment of the impact of LD on MCMC-based methods of haplotype inference, location-specific IBD estimates, and lod score analyses.

Finally, in 2013-18, methods continued to address new types of genetic marker data and trait data, including cardiovascular, neurological, and behavioral phenotypes. Rather than marker-location IBD, inference of IBD segments of genome allowed the use of increasingly dense marker data, and permitted the combination of within-pedigree and between-pedigree IBD information. Combining population and pedigree data has been a major focus: through the inference of location-specific IBD we can analyze both data on individuals of known pedigree, and individuals for whom the pedigree relationships are unknown. We can infer not only cryptic relatedness but also the precise segments of shared genome.

IBD inferred from marker data (whether in pedigrees or in populations), can be used in far more computationally efficient trait analyses, allowing the use of multiple trait models, and hence studies of trait model robustness, and significance of linkage findings. We have considered gene interactions (epistasis) and interactions of genes and environment. Additionally, MCMC-based inference of IBD provides new methods for imputation of genotypes and haplotypes at dense locations, incorporating LD and next-generation sequence data into the analysis of trait data on pedigrees.

Throughout the period of the award, starting in the 1990s, we have developed software incorporating novel methods, and made it available to practitioners, along with documentation and tutorial examples. Out first full version of MORGAN 2.0 was released in November 1997, and 20 versions later, the final MORGAN 2.9 was released in November 2008, with tutorials and examples release in 2009. MORGAN-3 was a fully revamped structure, separating the concept of trait loci from that of traits, and allowing for more complex trait architectures. MORGAN 3.0 was released in March 2008, and 12 releases later the final MORGAN 3.4 was released in December 2017, with online and download tutorial and examples in 2019.

The methods and software that we have developed over almost 30 years continue to be used by researchers working in the genetic epidemiology of complex traits. Earlier methods informed new development of later methods, and more recent methods will inform future research.