Principal Investigator

Dr. David Heckerman

Address

Microsoft, Microsoft Research, 1100 Glendon Avenue, PH1, Los Angeles CA90024, USA

Summary of research:

GWAS, LMMs, genetic similarity matrix, LD

Application Lay Summary:

1a: We aim to develop a GWAS/PheWAS algorithm that can realistically handle datasets reflecting over one million individuals. The class of model that we will explore are mixed models, which require the specification of a set of genetic markers to define the similarity between individuals. We will use the UK Biobank data to determine how many genetic markers (spread across the genome) should be used for this purpose.

1b: The stated aim of the UK Biobank is "improving the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses." One key step towards this goal is association analyses. Existing association analyses have pointed to the need to perform analyses with large sample sizes to uncover the subtle signals in genome-phenotype association. With large samples comes confounding and the need to use methods that can correct for confounding. Our study aims to show that linear mixed models can be used for these studies.

1c: We will try different numbers of SNPs to correct for confounding, and see how few are needed. The fewer the better, as our goal is to apply our mixed-model algorithm to datasets with greater than one million samples.

1d: We believe the number of SNPs that will be needed is on the order of 50k, so our experiment would require a dataset with at least 100k samples.