Thomas J Balmat

Research Computing | Duke University | thomas.balmat@duke.edu | 919-820-1057

Recent Projects

Interactive Cross-Phenotype Analysis of GWAS Database (iCPAGdb)
Department of Molecular Genetics and Microbiology, Duke University

- iCPAGdb employs a web app to accept external genome wide association study (GWAS) data and identify traits with significant association with common genotypes appearing in both the uploaded data and in a selected internal GWAS data set (the idea is to define potential compound traits X+Y, with common genotype association, where X and Y are identified in distinct studies)
- When compared with the previous generation of computational solutions, iCPAGdb improves memory 153 efficiency with built-in functions connecting to SQL GWAS and LD proxy databases and 154 improves computational efficiency and speed by utilizing multiple CPUs. For the 155 NHGRI-EBI GWAS Catalog, the growth of GWAS findings and improvements of iCPAGdb over the previous version of CPAG led to a 27.7-fold increase in direct cross-157 phenotype associations and a 47.7-fold increase in indirect cross-phenotype 158 associations.
- An algorithm that implements Fisher's exact test is employed for identifying pleiotropic effects (possible association of a genotype with traits from two GWAS studies being compared)
- Web app developed using R, Shiny Server, SQL, and various R packages (Shiny, ggplot2, DT)
- Web site: http://cpag.oit.duke.edu
- Paper: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-021-00904-z
- Additional technical information: https://github.com/tbalmat/iCPAGdb/blob/main/v1.1/iCPAGdbWebApp.pdf

Duke High Performance Computing Cluster R Consultation
Research Computing, Duke University

- Assist faculty, post-doc, and PhD student researchers in improving performance of R scripts in high data volume, high demand computational settings
- Adapt sequential algorithms to parallel by defining disjoint problem or data subsets
- Implement parallel algorithms using functions from the parallel package (mcmapply, parApply)
- Improve computational efficiency using parallel methods, Rcpp, and openMP
- Implement very high efficiency matrix operations on GPUs, by integrating custom Cuda algorithms with R
- Develop Slurm scripts for batch submission of R scripts
- Assess computational performance to specify optimal core and memory resource requests
- Compile and install R packages, by identifying necessary operating system resources
- Document of common computing cluster R configuration procedures to be used by researchers
- Compute Cluster info: https://oit-rc.pages.oit.duke.edu/rcsupportdocs/about/

Brain Computer Interface (BCI) Project
Pratt School of Engineering, Duke University

- BCI is a clinical project that uses participant electro encephalogram (EEG) signal data, combined with concurrent P300 speller signals (the P300 speller is a visual panel device used to generate characters and symbols based on input excitation signals) for use by amyotrophic lateral sclerosis (ALS) communication researchers
- Develop R scripts to interpret files containing BCI2000 source files, containing over fifty signal streams, and generate concurrent graphs of signals, equipment states, and selected speller symbols
- Develop parallel R scripts to convert approximately 10,000 propriety BCI2000 signal files (https://www.bci2000.org/mediawiki/index.php/Technical_Reference:BCI2000_File_Format) to the open European Data Format (EDF, https://www.edfplus.info/specs/edfplus.html)
- Assist researchers interpret EDF file contents for use in machine learning algorithms designed to improve the accuracy of P300 speller results (improvement in software selection of characters and symbols being those intended by study participant)
- Paper

Urea Cycle Disorder (UCD) Project
School of Nursing, Duke University

- The UCD project employed a clinical data set to be explored in a visual, ad-hoc manner for identification of associations between participant demographics, physiology, metabolic signals, cognitive performance, treatments, and UCD diagnosis signals
- Member researchers from Children's Hospital, Baylor University Medical Center, University of Nebraska Medical Center, and Duke University
- Develop Neo4j graph database to contain a clinical data set representing over one hundred Urea Cycle Disorder patients along with demographic, diagnostic, anatomical, and treatment codes defined by the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)
- Develop graph database queries to relate study participants by features specified in real time by clinical researchers (up to one hundred features at up to five levels of relational depth)
- Develop a graphical interface to compose and execute, in real time, graph queries as a researcher explores and learns relationships of interest (app developed in R, using the Shiny, visNetwork, and Neo4j packages)

Innovation Project
Fuqua School of Business, Duke University

- The innovation project seeks to measure the effect of venture capital funding, team characteristics, and political ideology on performance of organizations that research, develop, and patent innovative products
- Develop a large SQL database (366,000 organiztions, 900,000 employees, 164,000 funding rounds, and 85,000 voter registration records) to efficiently query characteristics of organizations, team member education, funding rounds, sources of investment, patent awards, and related performance covariates
- Implement queries to report various organization and team composition and performance characteristics as requested by senior researchers

Methane Reduction Target Project
Nicholas School of the Environment, Duke University

- The purpose of the project is to model change in important socio-environmental variables such as ozone levels, surface air temperature, premature death rates, and solar energy with respect to hypothetical change in methane emissions, CO2 emissions, and ambient aerosol concentrations
- Design and implement a virtual machine (VM) to host an interactive Shiny app that combines public researcher input with local data and models to graph and report effects of emissions on global and regional environments
- Web site: http://shindellgroup.rc.duke.edu/

U.S. Court of Appeals Decision Study
Duke University Law School

- The Appeals Project studies the effect of variables such as homogeneity of panel (ideal point correlation), influential judge phenomenon, gender effect, inter-district differences, and historical trends
- Develop MySQL database to contain data (sourced from LexisNexis) on (enumerate decisions, panels, judges, historical period)
- Develop queries (SQL) and graphs (R) to identify and expose patterns in the data for basic data integrity analysis, panel composition, historical patterns, and for legal theory hypothesis testing
- Fit logistic regression models to model proportion decisions by Shepard's codes and internally developed case category identifiers
- Text analysis to identify phrases by leading and trailing word groupings