U01: O'Connell - High-performance mixed model toolset for integrative omics analysis of big data

The recent large scale production of whole genome sequence and other multi-omics in TOPMed and other projects calls for parallel development of comprehensive, powerful and flexible toolset capable of large data management, analysis and integration. Mega/integrated analyses are essential to fully utilize these data to elucidate the complexity of the biological mechanisms and advance our understanding of complex trait biology to drive precision medicine. TOPMed estimates that the VCF for 60,000 subjects will contain 400M variants and require 100TB of space, and much of our current genetic analysis toolset does not scale up to these data sizes.

For rare variant analysis, mixed model mega analysis is more powerful than meta-analysis as mega analysis can include additional random effects to account for genetic relatedness between all subjects and cross-study phenotypic, genetic and environmental heterogeneity. However cross-study mega analysis within the mixed model is still an uncharted territory. We believe mega analysis will spur more creative analysis approaches provided the needed toolsets are available. In cloud computing “time is money”, and new approaches are required to solve structural differences in resource allocation and data access compared to local computing. MMAP (Mixed Models for Analysis of Pedigrees/Populations) is robust mixed model software that already published mixed model analysis on a sample size of 90,000 that included dominance variance and developed a cloud-efficient version of mixed model rare variant analysis.

The goal of this proposal is to further expand and improve this toolset to deliver to the research community a flexible, versatile, and comprehensive cross-platform mixed model toolset scalable to efficient local and cloud analysis of large WGS and omics data. We plan to implement several new features in our toolset including: 1) Efficient binary genotype file format for optimal storage of terabyte VCF genotypes. 2) Large-scale modeling of non-additive variation such as dominance, X- lined, mitochondrial and epistasis. 3) Optimized rare variant analysis with flexible integration of annotation and variant weighting resources. 4) Optimized expression/epigenome-wide association (EWA) analysis. 5) Comprehensive multi-omics integration into the mixed model as fixed and random effects. 6) Development of a multi-omics simulation software to guide systems biology modeling. 7) Integrating mixed model equations for prediction from animal breeding.

This proposal will deliver the research community an analysis toolset that will push research boundaries well beyond additive SNP association to a space filled with complex biological fixed and random effects models integrating the full spectrum of multi-omics data. We plan to develop a multi-omics simulation tool to better understand the complex evolutionary processes that shape the complex trait landscape. Our toolset will be extensively shaped by collaboration with TOPMed working groups to meet analysis priorities and develop analysis plans. Our toolset will surely evolve in novel and unexpected directions in response to new ideas and challenges as we dive deeper into this unique data set.

NHLBI Program officer: Cashell Jaquish

PI:	Jeff O'Connell
Award Type:	U01 NHLBI TOPMed Program: Integrative Omics Approaches for Analysis of TOPMed Data (RFA-HL-17-011)
Award number:	U01 HL137181-01
Start Year:	2017

Source URL (modified on 06/10/2019 - 4:31pm):https://topmed.nhlbi.nih.gov/awards/2429