



cordially invites you to an

Interdisciplinary Seminar

with

Dr. Scott Emrich

on

“Integrating diverse data for improved computational genomics”

Tuesday, October 17, 2017

3:30-5 p.m.

Reception & refreshments at 3 p.m.

Hallam Auditorium, Room 206
1122 Volunteer Boulevard



Dr. Scott Emrich will join the faculty of the University of Tennessee in January in the Department of Electrical Engineering and Computer Science. He is currently the Director of Bioinformatics at University of Notre Dame with a primary appointment in Computer Science and Engineering and a courtesy appointment in Biology. He obtained a BS in Biology and Computer Science from Loyola College Maryland and a 2007 PhD in Bioinformatics and Computational Biology from Iowa State University. He has published over 78 peer-reviewed publications including venues such as *Science* (4, 2 covers), *PNAS* (3), *Nature* and *Genome Research*. Most of his genomics-focused work at Notre Dame involved his twelve PhD students (9 grads) in a leadership role. Two of his students co-won interdisciplinary submissions for both the Assemblethon2 and the DREAM regulatory prediction challenge. He has four active awards from the NIH.

Abstract: Genomics-driven analysis of many important species, which we have called “non-models,” remains challenging. My group is funded by the NIH to computationally leverage and enhance newer higher-throughput sequencing with domain expert-provided metadata including biological phenotypes (e.g., drug resistance in malaria), protein folding experiments, and karyotype data from mosquitoes. For this talk I will focus on updates to our 2016 ACM BCB paper, which has been recently published in IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). Previous computational approaches for imputation of missing genotype data have relied on a linear order of markers and a genotype panel, both of which are not common in non-models. We address this limitation with our ADDIT (Accurate Data-Driven Imputation Technique) approach, which is composed of two data integration-focused algorithms: a non-model variant that employs statistical inference, and a model organism variant that better leverages reference data using a supervised learning-based approach. I will show that ADDIT is more accurate, faster and requires less memory than state-of-the-art methods using model (human) and non-model (maize, apple, grape) datasets. I will conclude with emerging “-omics” results from a few other current projects, most involving either a mosquito vector or the disease they transmit. The methods we are developing combine traditional sequence-based analysis with existing and new models, mostly network-based, through collaborations to best integrate diverse data types. I will discuss some early successes and challenges, with the latter hopefully a start for new discussions in Knoxville come January.