**A data-driven approach for genotype imputation and phasing**

**Benjamin Chu,** Rory Wasiolek, Eric Sobel, Hua Zhou, Kenneth Lange

**Background**: Currently, popular methods for genotype imputation exploit a pool of haplotype reference panels. These methods are based on hidden Markov models (HMM) that assume an underlying genetic architecture. A handful of data-driven techniques exist, but they are computationally intensive and do not support phasing.

**Methods:** Taking a data-driven perspective, we revisit the problem of genotype imputation and phasing from a pool of haplotype reference panels.

**Results:** By casting the problem as a least-squares minimization problem, we efficiently invoke BLAS level-3 routines and demonstrate that the computational bottleneck of our algorithm reduces to searching for minimum entries of small symmetric matrices. Our Julia implementation is 5-10x faster than the fastest alternative, requires less memory, and has comparable imputation quality to existing methods. We also demonstrate that datasets violating the common genetic model suffer poor imputation quality with HMM-based methods, and that our approach is robust to these scenarios.

**Significance:** We present the first data-driven technique that supports both genotype imputation and phasing. Our implementation is the fastest among its competitors. Along with the OpenMendel platform, our software offers easy-to-use, feature-rich interactive sessions for general-purpose data processing, data manipulation, and result visualizations, enabling exploratory data science.