Quantitative phenotypes emerge everywhere in systems biology and biomedicine. They are of special interest in complex common diseases in which high individual variability makes difficult or impossible to separate cases into distinct categories. Fitting quantitative phenotypes from genome-wide data has been considered only recently [1].

We introduce the use of regularization methods in the prediction of quantitative traits. The implementation presented in this study is based on a complete pipeline for high-density genotype data. Available methods are standard Support Vector Regression (SVR) and L1L2 Regression [2], an evolution of the elastic net model [3]. They are tested on data from the Heterogeneous Stock Mice genotype/phenotype dataset [4]. With reference to a Reversible Jump Monte Carlo Markov Chain (MCMC) model [1], comparable or better accuracies are found on 2 quantitative phenotypes (12,112 SNPs and ~1500 samples).

### INTRODUCTION

The data analysis pipeline (Fig. A) is composed by a common preprocessing step and method-specific model selection protocols. The genotype data are encoded as 0, 1, or 2 (dominant homozygous, heterozygous, recessive homozygous respectively). Missing data (4.7% of the whole dataset) are randomly imputed with probability equal to the relative frequency of each allele at that locus in the population.

For SVR (Fig. A, bottom left), optimal parameters were found by grid search on 50% bootstrap (10 replicates) and then validated on 15 train/test bootstrap, as in [1]. Members of a family were assigned all either to the training or to the test set (interfamily sampling), thus avoiding information leakage due to very high genetic similarity between individuals in the same family.

For L1L2 (Fig. A, bottom right) we used a more sophisticated protocol, from the guidelines of the MAQC-II project [5]: given the 15 development-validation interfamily splits, model selection in the accuracy-stability space [6,7] was obtained by internal 10-Cross Validation (CV) on each of the dev data. The model defined by the optimal (\(\beta, \tau, \lambda\)) was then trained and evaluated on each dev/val split, obtaining estimates for accuracy and weighted SNPs lists. The L1L2 algorithm and its protocol are implemented within the mlpy Python package and run on the MPBA Linux cluster.

### RESULTS

- **Predictions**: Table B shows the squared correlation coefficients averaged over 15 splits for SVR (Gaussian kernel, \(\sigma=2.5\times10^{-2}\)), L1L2 and the reference MCMC. SVR leads on the MCH phenotype, while the three methods are equivalent on CD8+.

- **Selected features**: Lists of features (SNPs here), ranked by the regression weights are yielded by the pipeline. The SNPs selected by SVR and L1L2 overlap well with the genome regions identified by previous GWAS [5]. In Fig. C, for the CD8+ phenotype, SNP features are shown if present in at least 14/15 runs and in the top 10-percentile of the distribution of the absolute weights.

- **Stability of features**: Stability of features is crucial for reproducibility and identification of the most relevant biomarkers. We can compute stability through the Canberra distance [6,7], and select models having both good accuracy and stability. The accuracy-stability diagnostic plot in Fig. D for the L1L2 method and CD8+ shows that the same parameter set is optimal on each of the 15 runs (average on the 10 CV for 9 different parameter sets).

### METHODS AND PIPELINE

The data analysis pipeline (Fig. A) is composed by a common preprocessing step and method-specific model selection protocols. The genotype data are encoded as 0, 1, or 2 (dominant homozygous, heterozygous, recessive homozygous respectively). Missing data (4.7% of the whole dataset) are randomly imputed with probability equal to the relative frequency of each allele at that locus in the population.

For SVR (Fig. A, bottom left), optimal parameters were found by grid search on 50% bootstrap (10 replicates) and then validated on 15 train/test bootstrap, as in [1]. Members of a family were assigned all either to the training or to the test set (interfamily sampling), thus avoiding information leakage due to very high genetic similarity between individuals in the same family.

For L1L2 (Fig. A, bottom right) we used a more sophisticated protocol, from the guidelines of the MAQC-II project [5]: given the 15 development-validation interfamily splits, model selection in the accuracy-stability space [6,7] was obtained by internal 10-Cross Validation (CV) on each of the dev data. The model defined by the optimal (\(\beta, \tau, \lambda\)) was then trained and evaluated on each dev/val split, obtaining estimates for accuracy and weighted SNPs lists. The L1L2 algorithm and its protocol are implemented within the mlpy Python package and run on the MPBA Linux cluster.

### DATA

The dataset used in this study was made available on the website http://gscan.well.ox.ac.uk by the Wellcome Trust Center for Human Genetics (WTCHG) [4]. They include familiar, genotype and phenotype information from a population of 4 generations of heterogenous stock mice. Two quantitative phenotypes were used in this study: the percentage of CD8+ cells (CD8+) and the MCH phenotype information from a population of 4 generations of heterogeneous stock mice. WTCHG – Heterogeneous Stock Mice dataset, CD8+ phenotype

### REFERENCES


