A Machine Learning Pipeline for Phenotype Prediction from Genotype Data

Giorgio Guzzetta, Giuseppe Jurman, **Cesare Furlanello**

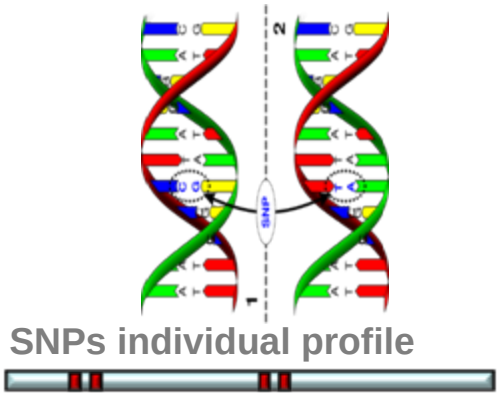
FONDAZIONE BRUNO KESSLER

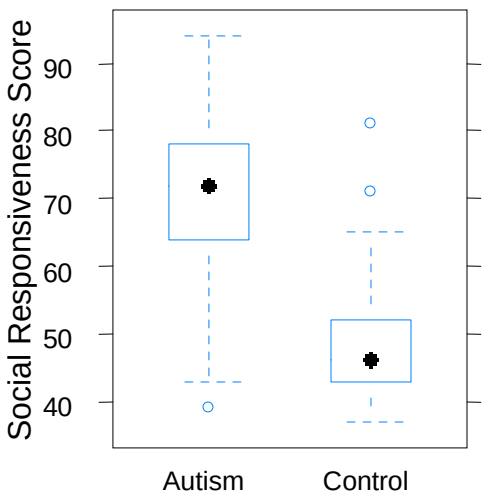
FBK-MPBA: Predictive Models for Biomedicine and Environment

http://mpba.fbk.eu

Funds from: Hiperdart FP7 and BICG AIRC

General Framework / 1

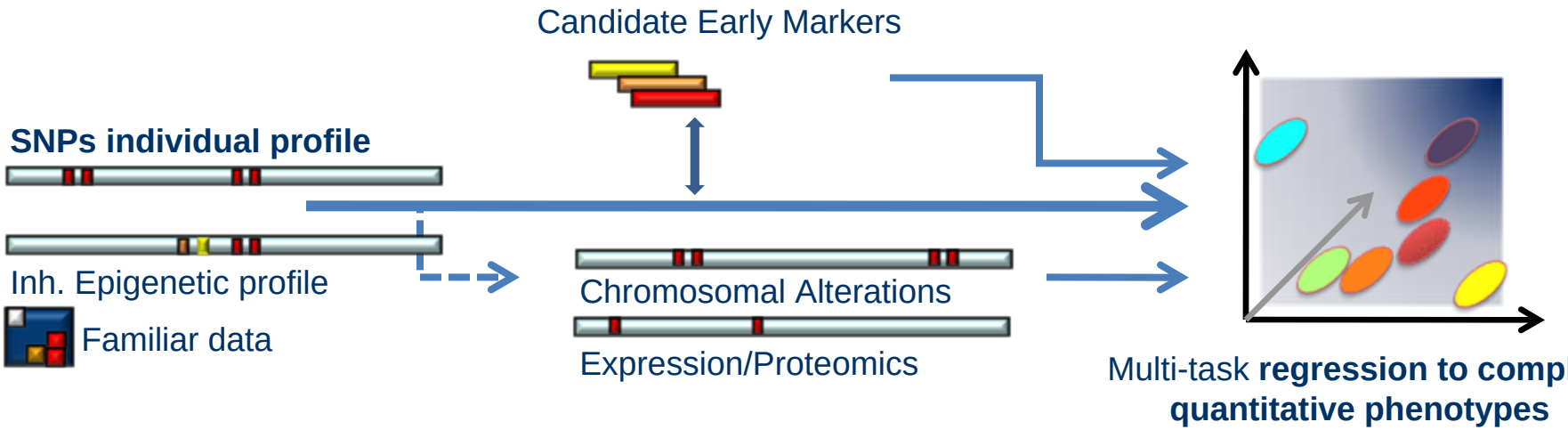




Fitting Quantitative Phenotypes from Genotype

- Quantitative phenotypes emerge everywhere in systems biology and biomedicine
- Complex common diseases with high individual variability and heterogeneous symptoms (no easy categorization among affected or affected/unaffected cases)
- Pre-condition for studying trajectories
- Combined approaches: estimate from molecular data the parameters needed to improve infectious disease modeling (e.g susceptibility)

General Framework / 2



Example: Autism Spectrum Disorders are measured by semi-structured diagnostic assessments and sets of clinical tests. No easy categorization. Variable response to treatment. Evidence of different individual trajectories.

Also: molecular basis of response to pharmacological treatment in Major Depression

FONDAZIONE BRUNO KESSLER

Reference Data and Study

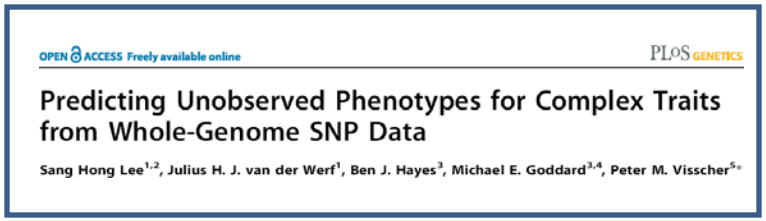
Reference study: A Monte Carlo Markov Chain (MCMC) model for **predicting quantitative traits** from genome-wide SNP data (Lee et al., 2008)

GSCAN Public mice dataset from the Wellcome Trust Center for Human Genetics (http:/gscan.well.ox.ac.uk)

•Familiar, phenotype and genotype (biallelic)

IN THIS STUDY

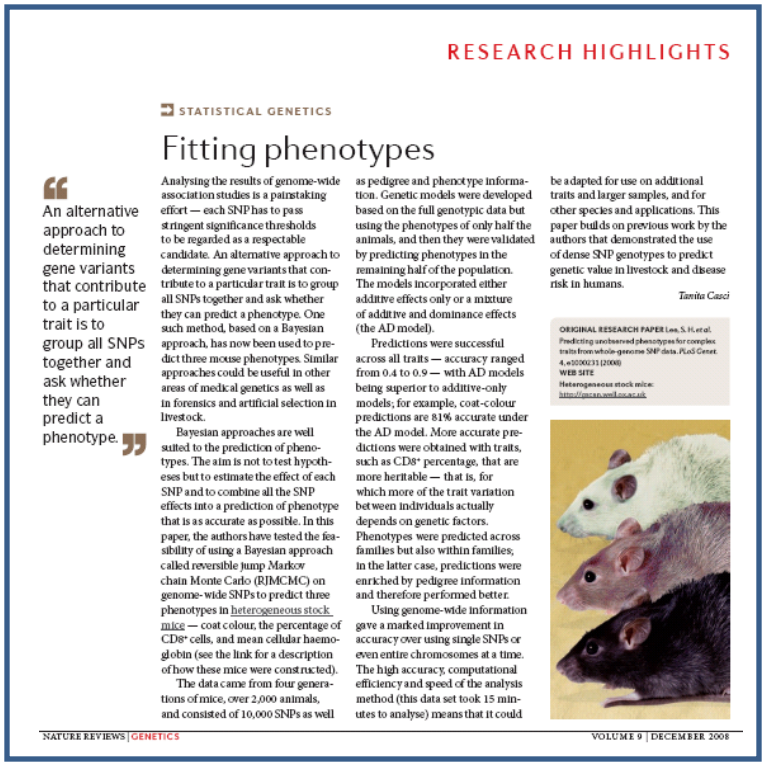
- **2 Quantitative phenotypes:**
  a. % of CD8+ cells (**CD8**)
  b. Mean Cell Haemoglobin (**MCH**)
- **# samples:**
  - 1,521 (CD8)
  - 1,591 (MCH)
- **# features:** 12,113



OPEN ACCESS Freely available online PLoS GENETICS

Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data

Sang Hong Lee[1,2], Julius H. J. van der Werf[1], Ben J. Hayes[3], Michael E. Goddard[3,4], Peter M. Visscher[5]*



RESEARCH HIGHLIGHTS

STATISTICAL GENETICS

Fitting phenotypes

An alternative approach to determining gene variants that contribute to a particular trait is to group all SNPs together and ask whether they can predict a phenotype.

Analysing the results of genome-wide association studies is a painstaking effort — each SNP has to pass stringent significance thresholds to be regarded as a respectable candidate. An alternative approach to determining gene variants that contribute to a particular trait is to group all SNPs together and ask whether they can predict a phenotype. One such method, based on a Bayesian approach, has now been used to predict three mouse phenotypes. Similar approaches could be useful in other areas of medical genetics as well as in forensics and artificial selection in livestock.

Bayesian approaches are well suited to the prediction of phenotypes. The aim is not to test hypotheses but to estimate the effect of each SNP and to combine all the SNP effects into a prediction of phenotype that is as accurate as possible. In this paper, the authors have tested the feasibility of using a Bayesian approach called reversible jump Markov chain Monte Carlo (RJMCMC) on genome-wide SNPs to predict three phenotypes in heterogeneous stock mice — coat colour, the percentage of CD8+ cells, and mean cellular haemoglobin (see the link for a description of how these mice were constructed).

The data came from four generations of mice, over 2,000 animals, and consisted of 10,000 SNPs as well as pedigree and phenotype information. Genetic models were developed based on the full genotypic data but using the phenotypes of only half the animals, and then they were validated by predicting phenotypes in the remaining half of the population. The models incorporated either additive effects only or a mixture of additive and dominance effects (the AD model).

Predictions were successful across all traits — accuracy ranged from 0.4 to 0.9 — with AD models being superior to additive-only models; for example, coat-colour predictions are 81% accurate under the AD model. More accurate predictions were obtained with traits, such as CD8+ percentage, that are more heritable — that is, for which more of the trait variation between individuals actually depends on genetic factors. Phenotypes were predicted across families but also within families; in the latter case, predictions were enriched by pedigree information and therefore performed better.

Using genome-wide information gave a marked improvement in accuracy over using single SNPs or even entire chromosomes at a time. The high accuracy, computational efficiency and speed of the analysis method (this data set took 15 minutes to analyse) means that it could be adapted for use on additional traits and larger samples, and for other species and applications. This paper builds on previous work by the authors that demonstrated the use of dense SNP genotypes to predict genetic value in livestock and disease risk in humans.

Tamita Cassi

ORIGINAL RESEARCH PAPER Lee, S. H. et al. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet. 4, e1000231 (2008)
WEB SITE
Heterogeneous stock mice: http://gscan.well.ox.ac.uk

Background/1

Experimental populations (e.g. mice)

R Package	Domain	Features	Methods
R/qtl	Experimental crosses	•QTL mapping •Genotyping errors identification •Single-QTL genome scans •Two-QTL, two-dimensional genome scans	• HMM for dealing with missing data • Interval mapping (EM algorithm) • Haley-Knott regression • Multiple imputation
HAPPY	Heterogeneous stocks	•Ancestral haplotype reconstruction •QTL fine mapping	•Dynamic programming •Linear regression •ANOVA
bqtl	Inbred crosses, recombinant inbred lines	•QTL mapping	•Likelihood-based •Bayesian techniques

Background/2

Natural populations (e.g. humans)

Software	Domain	Features	Methods
GENEHUNTER	Sib-pair	•Linkage analysis	•Haseman-Elston regression (traditional and EM) •Maximum likelihood
MERLIN	Any	•Linkage analysis •Linkage disequilibrium adjustment	•Gene flow trees •Sham regression
LOKI	Any	•Segregation analysis •Linkage analysis	•Monte Carlo Markov Chain
PLINK	Any	•Association •Epistasis tests	•Standard linear regression

Hoggart et al 2008: *"Testing one SNP at a time does not fully realise the potential of genome-wide association studies to identify multiple causal variants, which is a plausible scenario for many complex diseases."*

Marchini et al 2009: *"The inconclusive findings identified with this study reflect the status of the field of autism genetics and suggest that classical approaches such as linkage SNP association and CNV analysis and association analyses alone may not be sufficient to deal with the genetic and phenotypic heterogeneity seen in autism."*

Multivariate and ML approaches

1. **Regularization**
   - **This study (method)**: l1/l2 , n=12000 x p=1600 (mice), various encodings, Intra/interfamily effects fully managed by the experimental plan (DAP)
   - CMU SailingLab (W84): heterogeneous multitask.
2. **Kernel-based**
   - **This study (baseline):** epsilon-SVR : R/Libsvm implementation, with linear loss function, Kernels: Gaussian, Linear, poly, custom kernels.
   - Gonzalez-Recio et al., 2008 & 2009: kernel regression with 'ad hoc' kernel n=3500 x p=400 samples (broilers)
3. **Bayesian**
   - **This study (reference):** Lee et al., 2008, MCMC (Reversible Jump), Encoding: (-1,0,1); Dominance, Intra/interfamily effects
   - de los Campos et al., 2009: Bayesian regression coupled with LASSO, largest study n=11000 x p=1900 (mice), Strong pedigree effect
   - Gonzalez-Recio et al., 2009: Bayesian regression, Dominance effect and interaction terms included, Hypotheses on priors

l1-l2 Regularization

1. **Basis**

   RLS regression model with embedded feature selection (De Mol et al 09)

   - A variant of the elastic net model (Zou&Hastie 05)
   - **Optimization problem**: $\beta_{l1/2} = \arg\min_{\beta} \frac{1}{n}\left\|Y - X\beta\right\|_2^2 + \tau\left\|\beta\right\|_1 + \mu\left\|\beta\right\|_2^2$

   (β: regression weights; **Y**: observed output;
   X: input data matrix; μ, τ: regularization parameters)

   μ and τ modulate the selection of the features:
   - μ preserves correlation among selected features
   - τ enforces the sparsity of the solution

2. **Algorithm**

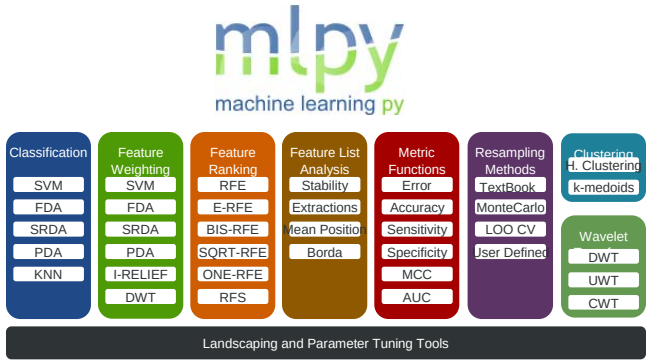
   - **Feature selection step**: by Iterative Soft Thresholding
   - **RLS Step**: At each step correction of the weight bias by a RLS regression on selected features (RLS parameter: λ)

   Note: Additive-only genetic model. But (Hill et al. 08) show that most variance is of an additive type even when dominance effects are present.

De Mol, C. et al. A regularized method for selecting nested groups of relevant genes from microarray data. J Comp Biol (2009)
Zou, H., Hastie, T. Regularization and variable selection via the elastic net. J R Statist Soc (2005)
Hill, W.G., et al. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genetics (2008)

l1-l2 regularization

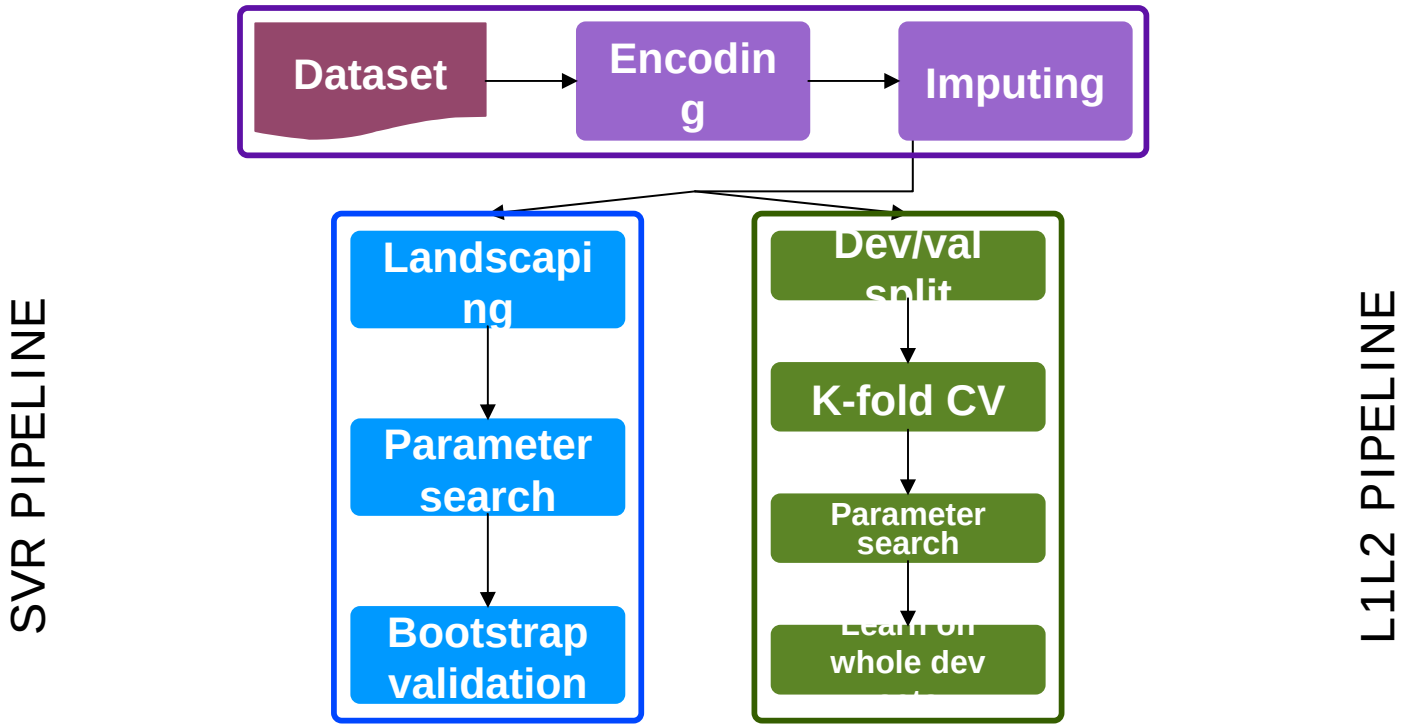
3. **A New Implementation**
   - l1-l2 with double optimization implemented in Python/NumPy, now a component of the `mlpy` package (https://mlpy.fbk.eu), using its functions for data import, handling and cross-validation. Parallelized for HPC. Extensively tested on 550K SNPs input features.
4. **Previous Results**
   - DeMol et al., 2009: classification from real predictions on gene expression data of leukaemia (n=72, p=7,000), lung cancer (n=181, p=12,000) and prostate cancer (n=102, p=12,000)
   - Fardin et al., 2009: a signature for hypoxia in gene expression data, neuroblastoma cell lines (n=18, p=50,000), classification from real predictions.



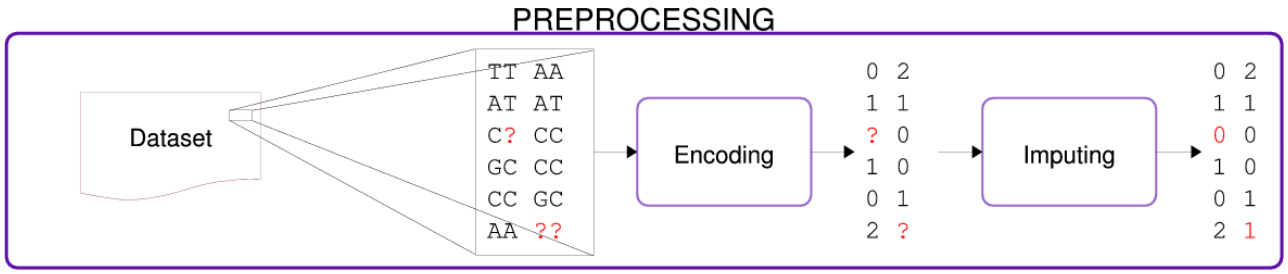
Fardin, P. et al. The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines. BMC Genomics (2009).

Data Analysis Protocol: overview

PREPROCESSING



Preprocessing



Encoding

•Linear regression approaches need a numerical representation for ternary SNP data (dominant homozygous – AA, heterozygous – Aa, or recessive homozygous – aa)

•Options tested:

- 0, 1, 2 (biological interpretation: # of alleles deviating from dominant)
- -1, 0, 1
- relative frequency of each allelic class at each locus over the sample population

•No significant variation in predictive power between 0, 1, 2 and -1, 0, 1

•Predictions worsen with the frequency-based encoding

•Final choice: **0, 1, 2** representation (classical in literature)

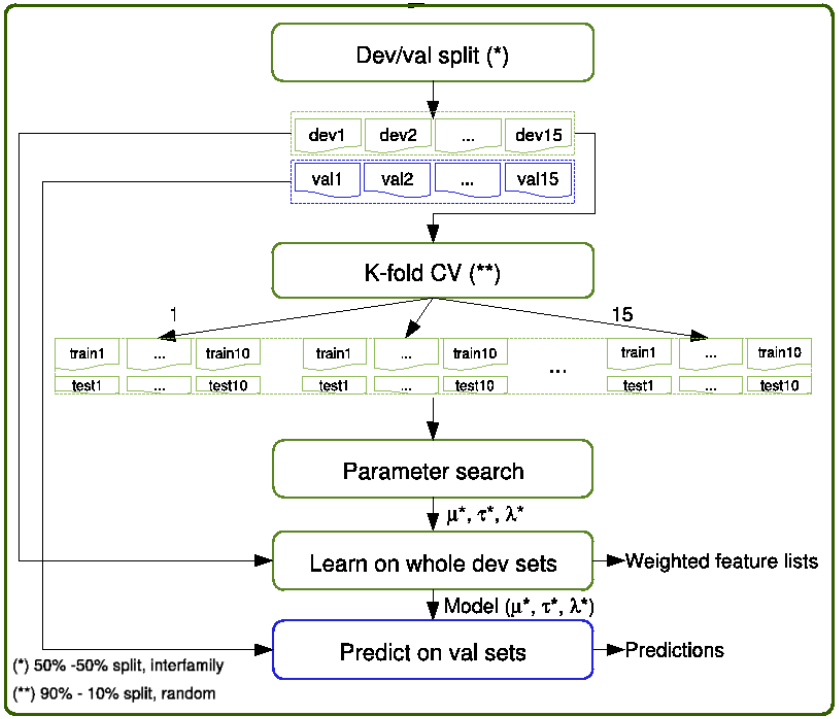
Imputing

•Few missing data in the GSCAN dataset (4.7%): **random imputation** with probability equal to the relative frequency of each allele at that locus in the population.

L1L2 workflow

Protocol inspired by the **MAQC-II project guidelines** (Shi et al., 2008)

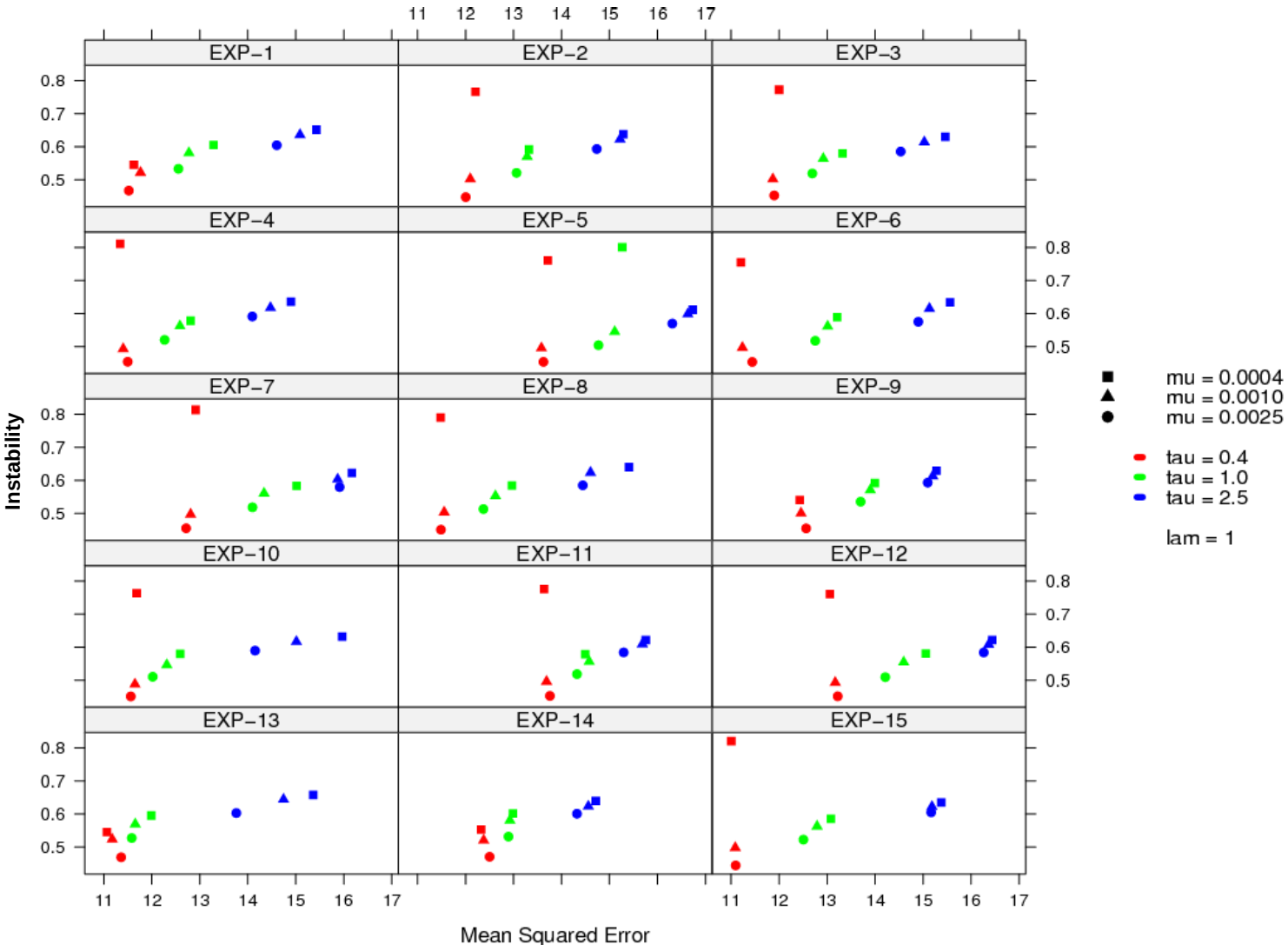
- Split dataset in **15 development / validation sets** (50% bootstrap **interfamily** re-sampling)
- For each experiment, run a **10-fold CV** on the development set
- Select optimal parameter set (μ^*, τ^*, λ^*) according to average prediction performances (accuracy, instability) across the 10 test sets
  - **Accuracy**: MSE or R^2
  - **Instability**: **Canberra distance** (Jurman et al., 2008) of lists selected in the 10 test sets ranked by abs weights
- Learn the 15 development sets using (μ^*, τ^*, λ^*)
- Assess results on the 15 validation sets



Shi, L. et al. Reproducible and reliable microarray results through quality control: Good laboratory proficiency and appropriate data analysis practices are essential. Curr Opin Biotechnol (2008)

Jurman, G. et al. Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics (2008)

Model selection



FONDAZIONE BRUNO KESSLER

Prediction accuracy

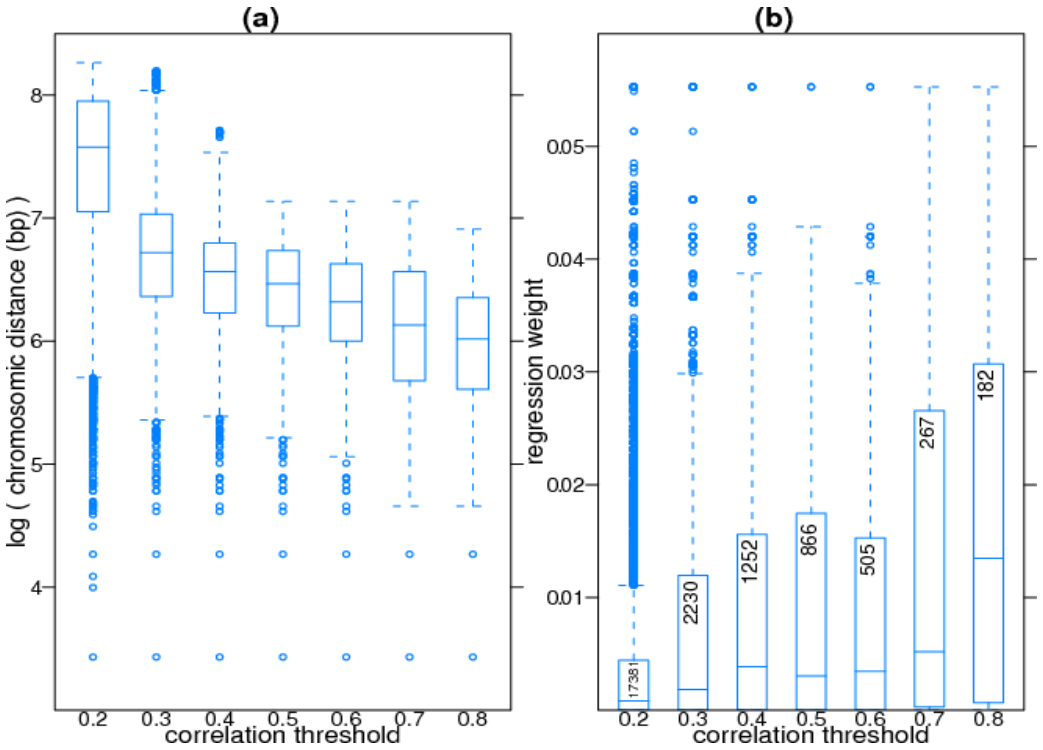
- Regression R for SVR, L1L2 and the reference MCMC model for each phenotype
- Standard deviation in parentheses

Method	CD8+	MCH
SVR	0.551 (0.05)	0.379 (0.06)
L1L2	0.559 (0.06)	0.323 (0.045)
MCMC - additive	0.51 (0.05)	0.33 (0.06)
MCMC – additive+dominance	0.56 (0.06)	0.33 (0.09)

NB: Interfamily experiments. Members of a family were assigned all either to the training or to the test set, thus avoiding information leakage due to very high genetic similarity between individuals in the same family.

Top-k Analysis

- We define **top-ranked SNPs** those in the top 10-percentile of the distribution of the absolute weights in at least 14/15 exps
- It is well-known that correlated, functionally important variables may be discarded or poorly ranked in feature selection methods
- We define **top-correlated SNPs** those having an absolute Pearson correlation coefficient with at least one top-ranked SNP higher than 0.8
- In our analyses, top-correlated SNPs are:



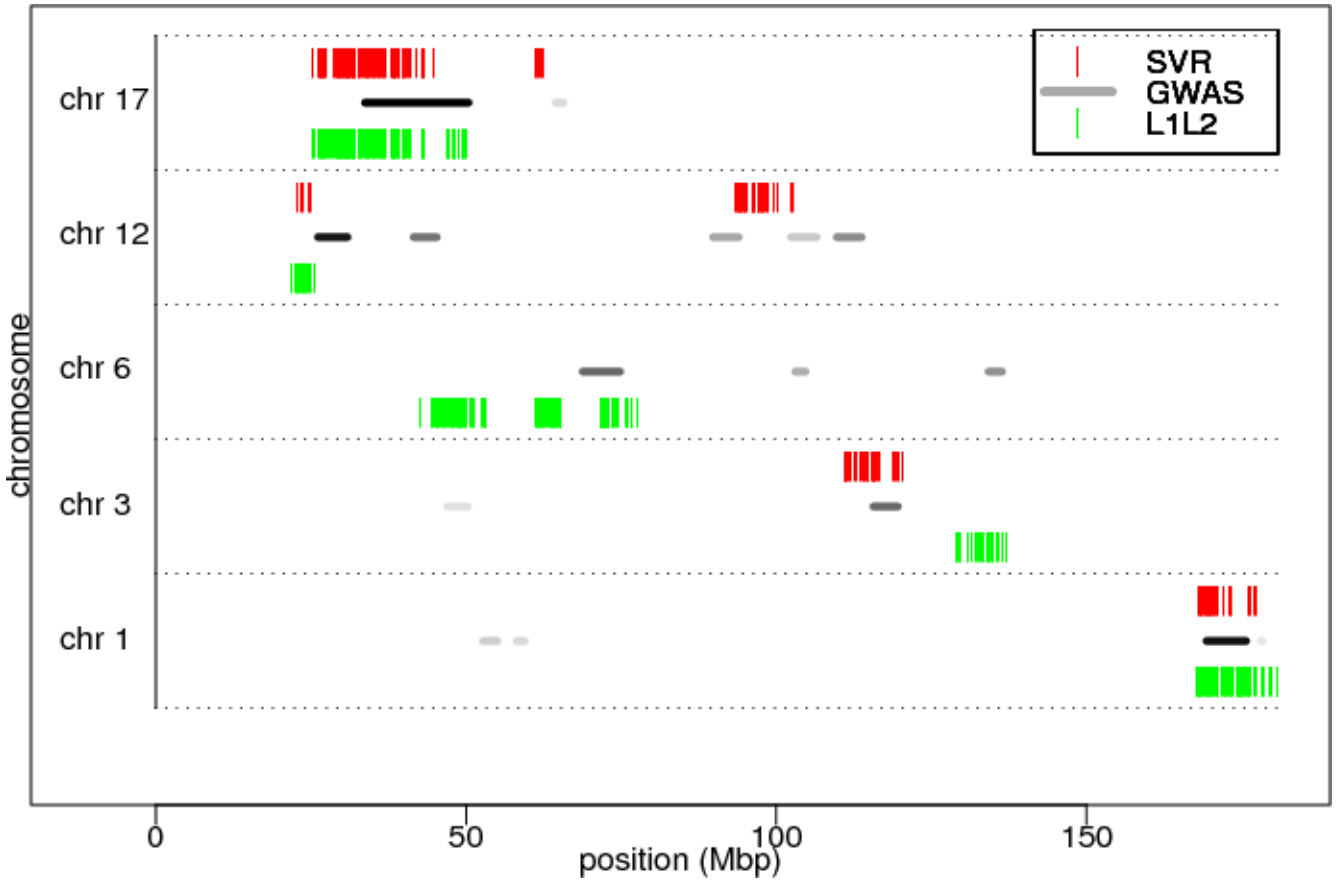
(a) Clustered around the reference top-ranked SNPs: the median distance smaller for higher correlation levels

(b) Highly ranked on average: the median absolute regression weight larger for higher correlation levels

For CD8: from 41 top-ranked to 182 top-correlated SNPs

FONDAZIONE BRUNO KESSLER

Results: feature selection



Pooling top-ranked and top-correlated SNPs, both methods select many SNPs in **candidate loci** previously identified by GWAS (Valdar et al., 2006)

Valdar, W. et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature Genetics (2006)

Conclusions

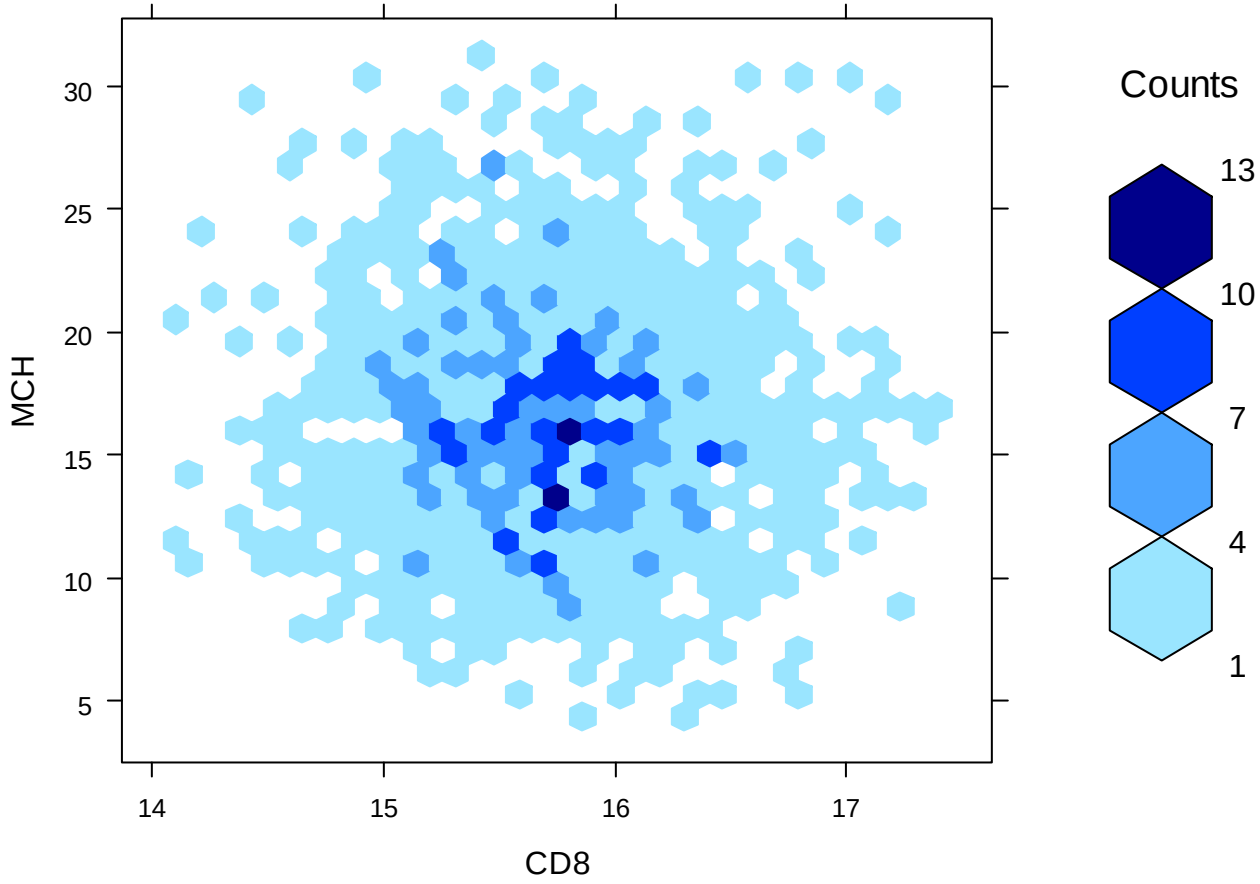
Fitting quantitative phenotypes from high-throughput data

•**L1/L2 with double optimization** can be used for building regression models and extract biomarkers in large scale GWAS studies.
•Apply carefully within an experimental Data Analysis protocol (DAP) to avoid bias. Still one module within a pipeline: each component a source of variability – see MAQC-II studies
•Use stability and biological hypothesis to detect potential markers

•**Applications:**

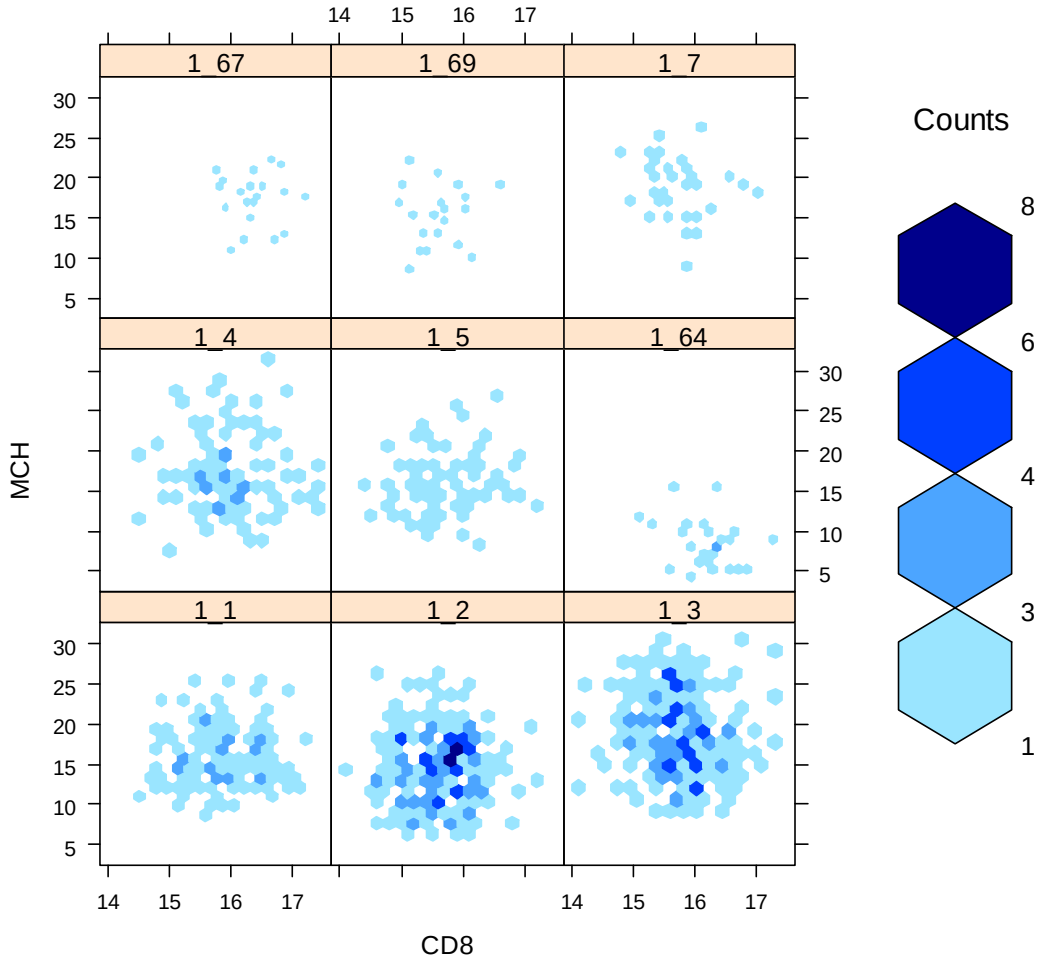
- Effective in complex common diseases with high individual variability (e.g. neurogenomics): trajectories.
- Parameter estimation for infectious disease modeling
- Gene expression as a quantitative trait: identify traits predictive of gene networks and gene expression regulation (eQTL)
- Survival analysis
- May be used for classification purposes where classifiers fail

Supplementary Material/1



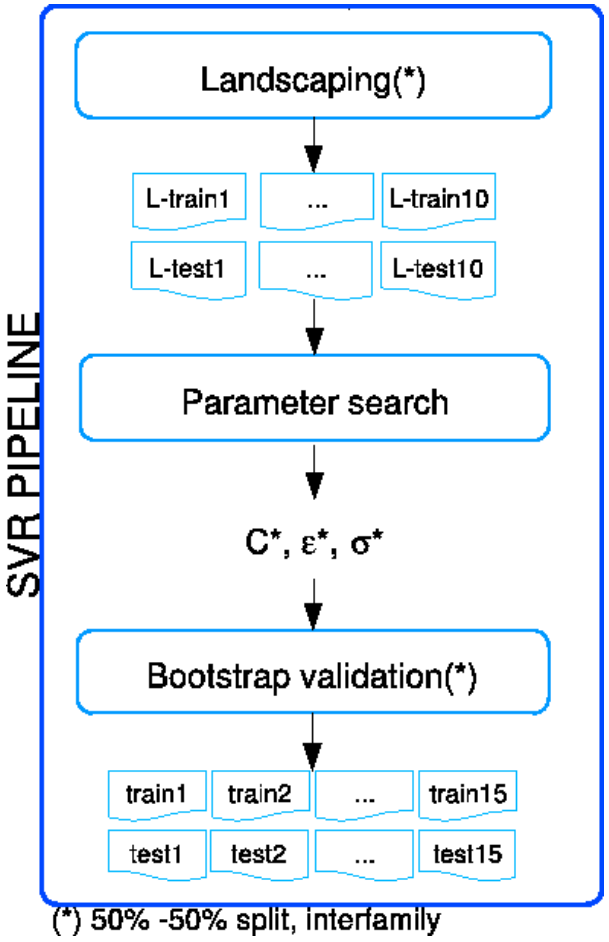
- GSCAN dataset

Supplementary Material/1



- GSCAN dataset
  most numerous families

Supplementary Material/2



SVR – Model selection
Protocol inspired from (Lee et al., 2008)

Landscaping

•For each parameter set (C, ε, σ), train the SVM on 10 training sets (50% bootstrap re-sampling)

•**Interfamily sampling**: members of a family were assigned all either to the training or to the test set

- avoid information leakage due to very high genetic similarity among individuals in the same family.

•Parameter space explored with a **grid search**

•Select the optimal parameter set (C*, ε*, σ*) as the one having the best average R^2 on the test sets

Bootstrap validation

•Train the SVM with the optimal parameter set (C*, ε*, σ*) on 15 training sets 50% bootstrap re-sampling

•Evaluate predictions on the test sets