A functional strategy to characterize expression Quantitative Trait Loci

Elena Grassi(1)[†] , Elisa Mariella(2), Mattia Forneris(3), Federico Marotta(2), Marika Catapano(4), Ivan Molineris(5) and Paolo Provero(2,6)

(1) Department of Computer Science, University of Turin, Turin, Italy
(2) Department of Molecular Biotechnology and Health Sciences, Molecular Biotechnology Center, University of Turin, Turin, Italy
(3) Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
(4) Department of Medical and Molecular Genetics, King's College, London, UK
(5) Candiolo Cancer Institute - IRCCS, Candiolo, Italy
(6) Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute - IRCCS, Milan, Italy

[†] Email: elisa.mariella@unito.it

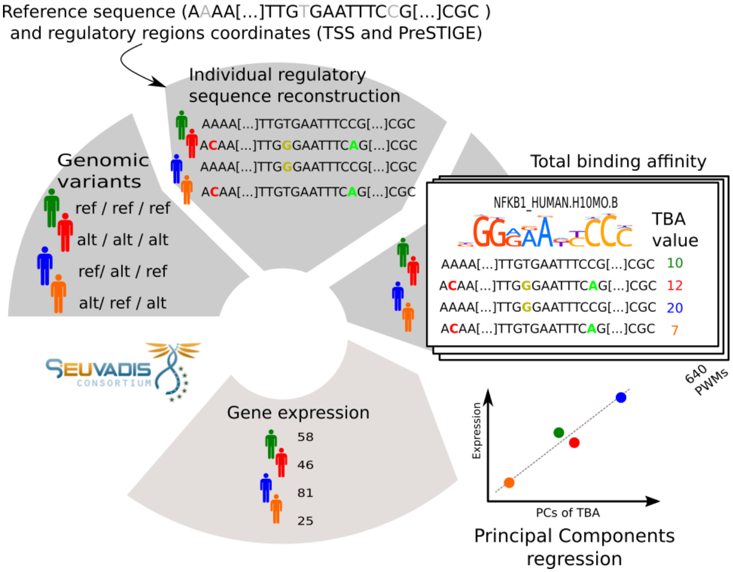
Motivation
Thousands of common genetic variants have been associated with common diseases by Genome Wide Association Studies (GWAS). However, the functional interpretation of GWAS hits is usually non trivial, especially because most of them lay outside the coding genome. These non-coding variants presumably exert their effect by altering gene expression levels, therefore expression Quantitative Trait Loci (eQTL) studies represent an important step in understanding their functional relevance and identifying target genes. We propose a new strategy to detect eQTLs taking into account the combined effect of genetic variants within regulatory regions and leveraging the idea that changes in gene expression often pass through the alteration of transcription factors (TFs) binding.

Methods
Our approach is based on the concept of Total Binding Affinity (TBA), that is a measure of the propensity of a TF to bind any DNA sequence, and its implementation requires a large dataset in which coupled whole-genome sequencing data and gene expression data are available for several individuals. First of all, genetic variants are used, together with the reference sequence of human genome, for the reconstruction of the sequence of regulatory regions in all the individuals. Then, on all these sequences we compute TBA scores for several TFs. Finally, we correlate the TBA profile of a regulatory region with the gene expression level of a target gene doing a principal component regression. This strategy was initially applied to 344 individuals and 22,125 genes from the GEUVADIS dataset. For each gene we fitted an independent TBA model for each linked local and distal regulatory region. Local regulatory regions were defined as the region spanning 1,500 bp upstream and 500 bp downstream from each transcription start site. In addition, we were able to associate at least one distal regulatory region to 4,291 genes exploiting the enhancer–gene interactions delineated by the PreSTIGE tool in lymphoblastoid cells.

For the computation of TBA scores, we used the HOCOMOCO database that includes 640 positional weight matrices for human TFs.

Results

Our TBA-based inference globally detected 3,781 significant genes, more than those found by the traditional univariate eQTL approach. In particular none genetic variant was individually associated with the expression variation of 1,543 genes for which instead at least one TBA model was significant. For each significant gene further "univariate TBA models" were fitted independently for each PWM on each linked regulatory region whose TBA model was significant. In this way we obtained, for each gene on each associated regulatory region, a list of putative TFs whose binding variation affects the gene expression in a sequence-dependent way. To validate these predictions, we used a recent systematic evaluation of binding QTLs (bQTLs) carried out in lymphoblastoid cells for 5 TFs (JUND, NF-KB, PU.1, POU2F1 and STAT1). Specifically, we fitted, for each of the five TFs, a logistic model in which the independent variable is the presence of a bQTL in a regulatory region and the regressors are the length of the region and the significance of the appropriate PWM in the univariate TBA model. For all five TFs, the coefficient of the TBA term was positive, as expected, and for two of them (NF-KB and STAT1) it was statistically significant. The illustrated results support the validity of the TBA-model in revealing associations between regulatory variants and gene expression and driving the formulation of mechanistic hypotheses for the gene expression variation pinpointing the most relevant TFs. Now would like to further develop this idea performing TBA – based Transcriptome Wide Association Studies (TWAS). At least in principle, exploiting the TBA can give some advantages with respect to standard TWAS, in particular it would permit considering also the effect of rare variants that cannot be measured in traditional eQTL studies.


Reference sequence (AAAA[...]TTGTGAATTTCCG[...]CGC)
and regulatory regions coordinates (TSS and PreSTIGE)
Individual regulatory sequence reconstruction
AAAA[...]TTGTGAATTTCCG[...]CGC
ACAA[...]TTGGGAATTTCAG[...]CGC
AAAA[...]TTGGGAATTTCCG[...]CGC
ACAA[...]TTGTGAATTTCAG[...]CGC
Genomic variants
ref / ref / ref
alt / alt / alt
ref/ alt / ref
alt/ ref / alt
SEUVADIS CONSORTIUM
Total binding affinity
NFKB1_HUMAN.H10MO.B
TBA value
AAAA[...]TTGTGAATTTCCG[...]CGC 10
ACAA[...]TTGGGAATTTCAG[...]CGC 12
AAAA[...]TTGGGAATTTCCG[...]CGC 20
ACAA[...]TTGTGAATTTCAG[...]CGC 7
640 PWMs
Gene expression
58
46
81
25
Expression
PCs of TBA
Principal Components regression