Supplementary material for

Altered transcription factor binding events predict personalized gene expression and confer insight into functional cis-regulatory variants

Wenqiang Shi[1,2], Oriol Fornes[1], and Wyeth W. Wasserman[1,*]

[1] Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, 950 28th Ave W, Vancouver, BC V5Z 4H4, Canada

[2] Bioinformatics Graduate Program, University of British Columbia, 2329 West Mall, Vancouver, BC V6T 1Z4, Canada

* To whom correspondence should be addressed. Tel: +1 (604) 875-3812; Fax: +1 (604) 875-3819; Email: wyeth@cmmt.ubc.ca.

Supplementary Notes

RNA-seq data for 462 LCLs (individuals) were downloaded from the GEUVADIS project [1]. For 445 of them, genotype information was obtained from the 1000 Genomes Project [2]. Individuals covered 5 populations, including 89 North-Europeans from Utah (CEU), 92 Finns (FIN), 86 British (GBR), 91 Toscani (TSI) and 87 Yoruba (YRI). Because African subjects (YRI) differ substantially from the four European sets, they were excluded from the analysis. Reasons for exclusion included: 1) African individuals exhibit significantly more sequence variations than Europeans [2]; 2) they also exhibit more population-specific differentially expressed genes [1]; and 3) the reference DHS and TF binding events used in this study are derived from GM12878 cells (an LCL from a European individual).

Supplementary Figures

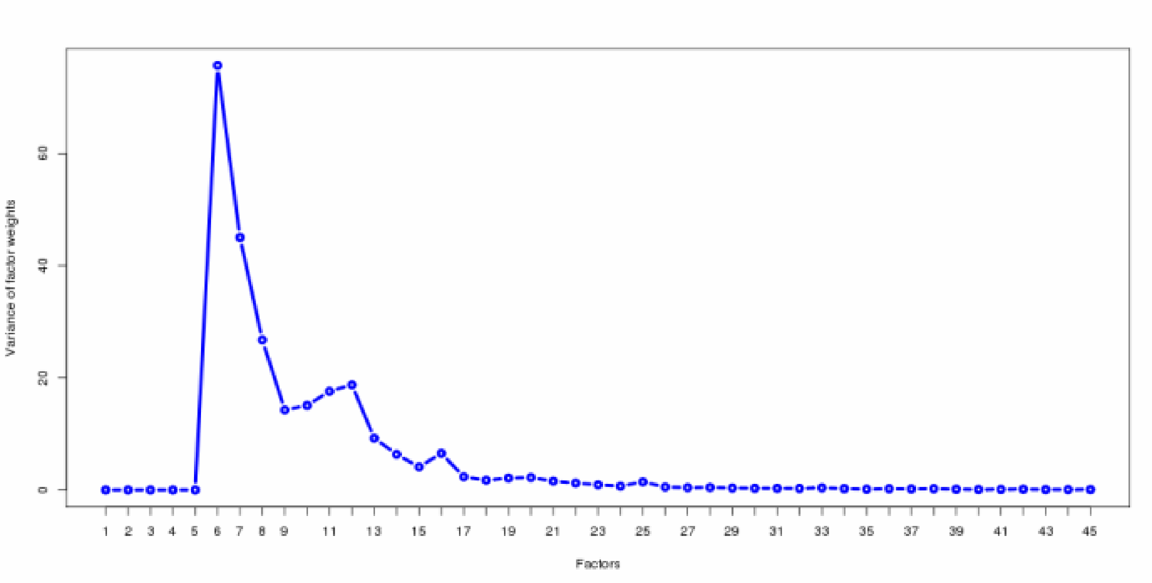

Figure S1. Determine the number of hidden factors in the expression data. We used PEER package to detect the impact of known covariates (4 population and one gender factors) and 40 potential hidden factors. The natural choice for the number of hidden factors is usually observed as the converged point in the factor variance plot [3]. We chose to remove first 27 factors (The first five known covariates and additional 22 hidden factors) in our expression data.

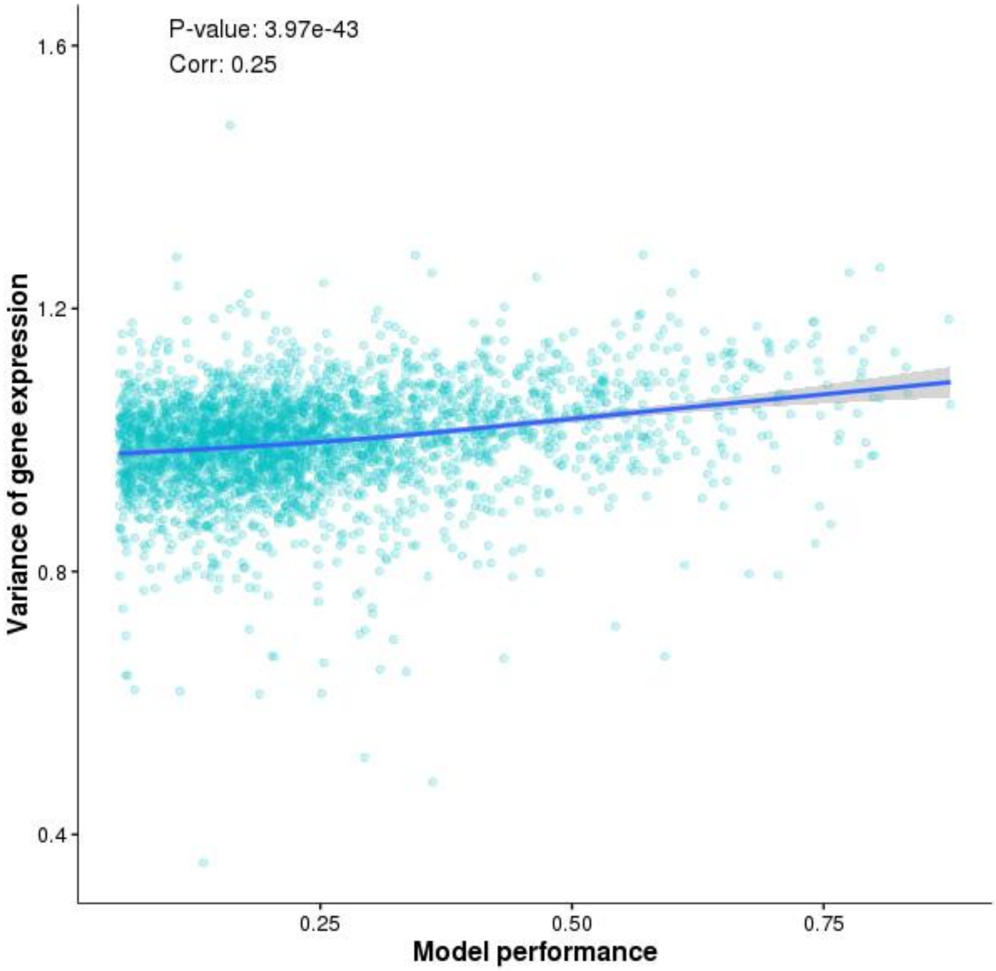

Figure S2. The performances of TF2Exp models are correlated with the variance of gene expressions. Each dot represents one predictable gene. The dot coordinates indicate TF2Exp model performance (x axis) and the variance of gene expression (y axis). We tested the correlation between the two axieses, and the spearman correlation coefficient and P-value are given on the plot. The blue line shows the general trends drawn by the locally weighted scatterplot smoothing method across all the dots.

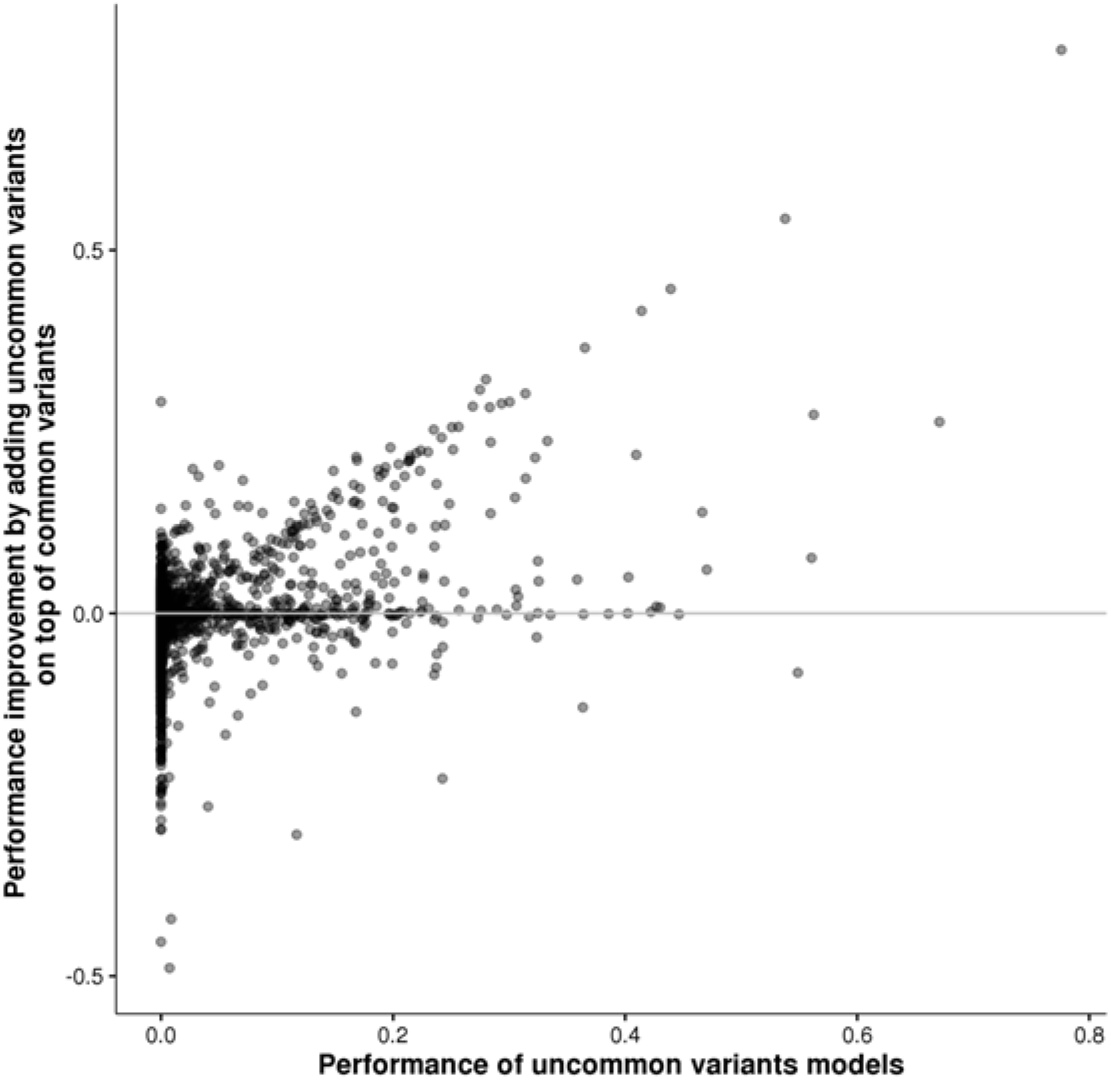

Figure S3. Uncommon variants improve the TF2Exp performance for a subset of genes

Each dot represents one predictable gene in the TF2Exp models. The contributions of uncommon variants were measured in two ways: 1) model performance when trained only using uncommon variants (x axis); 2) performance improvement after adding common variants on top of common variants (y axis).

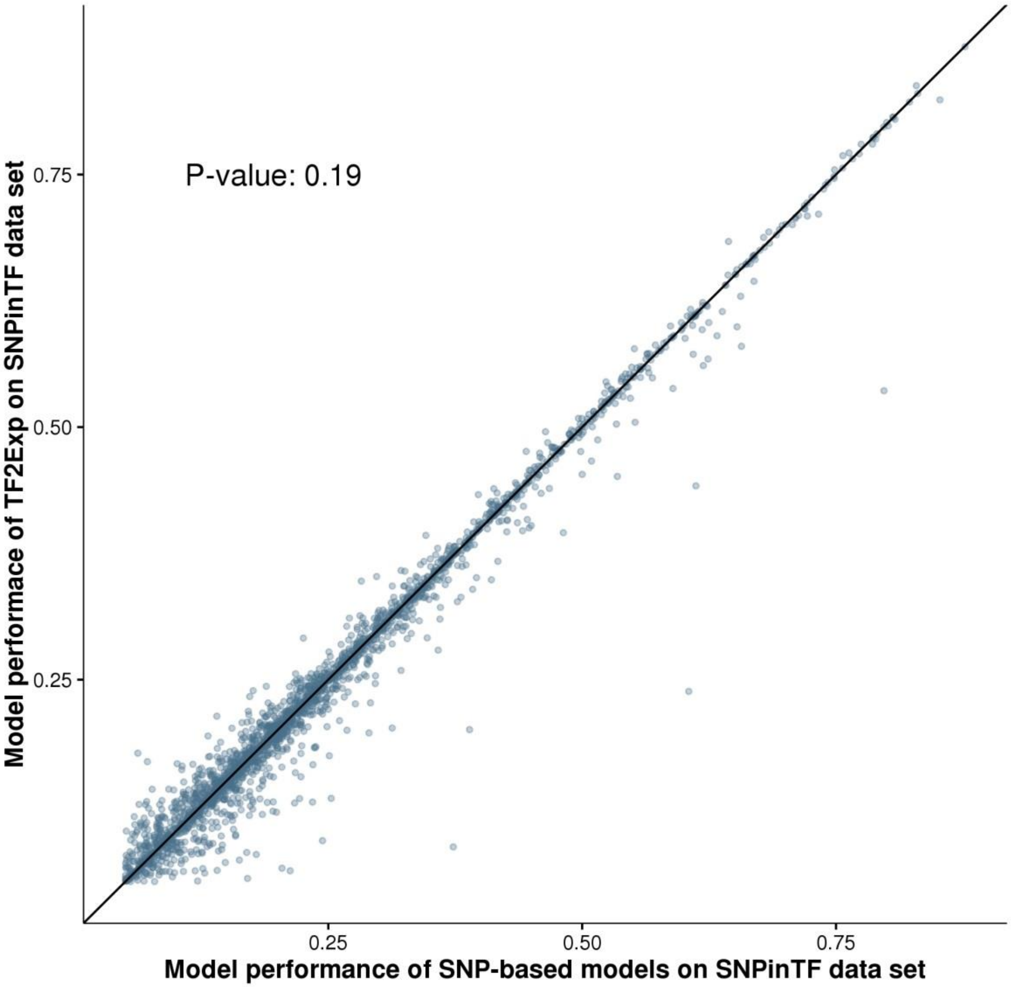

Figure S4. Performance comparison between TF2Exp and SNP-based models

Each dot represents an evaluated gene-model. The x, y coordinates are given by the cross-validation performances of the SNP-based and TF2Exp models, respectively, which were trained on SNPs in TF binding events associated with that gene (SNPinTF). The p-value reflects a non-significant difference in predictive power between two types of models (Wilcoxon signed-rank test).

References

1. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG *et al*: **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature* 2013, **501**(7468):506-511.
2. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA *et al*: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.
3. Stegle O, Parts L, Piipari M, Winn J, Durbin R: **Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses**. *Nature protocols* 2012, **7**(3):500-507.