Supplementary material for

**Gene expression models based on transcription factor binding events confer insight into functional *cis*-regulatory alterations**

Wenqiang Shi1,2,3, Oriol Fornes1, and Wyeth W. Wasserman1,\*

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

3 Beijing Institute of Microbiology and Epidemiology, Beijing, China

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

**Supplementary Notes**

**Quantifying gene expression from RNA-seq data**

RNA-seq data for 462 LCLs (individuals) were downloaded from the GEUVADIS project [[1](#_ENREF_1)]. For 445 of them, genotype information was obtained from the 1000 Genomes Project [[2](#_ENREF_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](#_ENREF_2)]; 2) they also exhibit more population-specific differentially expressed genes [[1](#_ENREF_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).

**Analyzing selected features using FANTOM5 data**

In the FANTOM5 project, an enhancer is associated with a gene based on the correlation of expression between the enhancer and the gene promoter across more than 800 tissues and cell types [[3](#_ENREF_3)]. For genes with selected events in distal regulatory regions, we counted the number of selected (and unselected) TF binding events in distal regulatory regions overlapping FANTOM5 enhancers associated to that gene [[3](#_ENREF_3)]. Individual gene statistics were aggregated, and the overall enrichment of selected features in enhancer regions was calculated using Fisher’s exact test.

**Validating the correlation between TF binding and gene expression in vivo**

For the key TF-binding events identified by TF2Exp models, we sought to validate whether the TF binding correlates with the gene expression in vivo. We obtained CTCF ChIP-seq mapped data (BAM files) for 45 LCLs [[4](#_ENREF_4)] and PU.1 for another set of 45 LCLs (38/45 overlap with CTCF LCLs) [[5](#_ENREF_5)] from the 358 LCLs in the original training data. For each TF binding event, the TF binding signal was quantified as the number of ChIP-seq reads in each ChIP-seq experiment using HOMER [[6](#_ENREF_6)]. TF binding signals were then normalized through multiple steps, including scaling by library size, averaging between replicates of each individual, converting to standard deviation units (standardization), performing quantile normalization, and removing batch effects by PEER [[7](#_ENREF_7)]. The resulting normalized data constitutes the in vivo TF binding signal for each TF binding event in each LCL.

We reserved the LCLs for which the extra ChIP-seq data was available as testing sets (one set for each of the two TFs). TF2Exp models were retrained on the non-testing LCLs, identifying 370 CTCF and 309 PU.1 TF binding events as key features for the subset of predictable genes. As less than 10% of in vivo TF binding events have been previously reported to show variable binding (greater inter-individual variance than intra-replicate variance) [[8](#_ENREF_8), [9](#_ENREF_9)], we anticipated that the majority of selected TF binding events in the testing cases would be invariable. To focus on potentially variable TF binding events (and to minimize multiple testing impacts), we restricted the analysis to the subset of TF binding events which exhibit a strong DeepSEA score variance (top 10% of all TF binding events), resulting in 83 CTCF and 72 PU.1 selected TF binding events. Then, we assessed the correlation between TF binding of the selected events and the associated gene expression in the two testing sets. Recognizing the small sample size, we estimated the minimum detectable correlation coefficient for the given testing size (n=45) at significance of 0.05 and power of 0.6 using the pwr package [[10](#_ENREF_10)].

**Rank the impact of eQTLs across genes**

To rank the impact of eQTLs across different TF-bound regions and genes, we normalized impact of the eQTL based on following formula:

$$impact=\frac{diff\_{DeepSEA}}{sd\_{TF-region}}\*Coef\_{TF-region}\*Performance \_{model}$$

where $diff\_{DeepSEA}$ is the score difference between two alleles of the eQTL, $sd\_{TF-region}$ is the standard deviation of TF alteration score for the overlapped region in the training samples, $Coef\_{TF-region}$ is the model coefficient of the overlapped region, and $Performance \_{model}$ is model performance of the gene.

**Evaluating the impact of noisy DeepSEA predictions on TF2Exp performance**

As some inaccurate DeepSEA predictions would add noise to the model feature space, we seek to test whether the noise would impact TF2Exp performance. As there is no ground truth datasets to tell the real binding signal, it’s hard to determine the magnitude of “noise” in the DeepSEA score. Alternatively, we created a “noisy” DeepSEA score set by extending the flanks of the original TF-bound regions (usually several hundred base pairs) to 1,100 bp centered at the peakMax position. Including variants in the 1,100 bp region introduces a mean relative change of -0.31% (standard deviation of 7.5%) on DeepSEA score across 78 TFs (Supplementary Table 2). Then, we test the impact of “noisy” DeepSEA scores on genes from chr1. We found that the model performance was quite similar between the two versions of DeepSEA score (Wilcoxon signed-rank test, p-value = 0.33, estimated median difference = 5.8e-4), suggesting that, overall, the noise from DeepSEA score would not have a dramatic impact on model performance.

**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 [[7](#_ENREF_7)]. We chose to remove first 27 factors (The first five known covariates and additional 22 hidden factors) in our expression data.



Figure S2. The length distribution of Hi-C fragments and proximal regulatory regions.



Figure S3. Distribution of distance between TSS of gene and its distal regulatory regions.



Figure S4. 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 S5. Feature effect sizes in proximal and distal regulatory regions. (A) Effect sizes of selected features decrease rapidly with respect to their distance to gene start positions. Each dot represents one selected feature (TF-binding event) of a predictable gene, and the coordinates indicate the feature distance to the gene start site (x axis) and the feature effect size (y axis) obtained in the TF2Exp model. Green contours indicate estimated dot density. Feature effect sizes are plotted separately for proximal regions (top panel) and distal regulatory regions (bottom panel). (B) Comparing the absolute feature effect sizes of selected TF-binding events at proximal and distal regulatory regions across the all the predictable genes. The labeled p-value indicates the significance of the difference between the two groups (Wilcoxon rank-sum test).



Figure S6. 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).



Figure S7. TF2Exp can quantify functional roles of SNPs in LD

Most TF binding events selected by TF2Exp models overlapped with the SNPs selected by SNP-based models for the same gene. A subset of the overlapped SNPs were in high LD (r2>0.9) with other SNPs in the same TF-bound region. Each dot depicts the absolute impact of a selected SNP by a SNP-based model (x axis) versus the absolute impact of its linked SNP, according to the TF2Exp model. The dashed line indicates two-fold impact of the linked SNPs compared with the selected SNPs. The plot includes the 1,002 linked pairs of SNPs.

Table S1. Average correlations between predicted (TF2Exp) and observed expression levels for 10 tissues from the GTEx project.

|  |  |  |
| --- | --- | --- |
| GTEx tissue | Number of individuals | Avg. Spearman’s correlation |
| adipose subcutaneous | 315 | 0.11 |
| artery tibial | 322 | 0.10 |
| esophagus mucosa | 286 | 0.11 |
| left ventricle heart | 228 | 0.09 |
| Lung | 307 | 0.11 |
| muscle skeletal | 403 | 0.08 |
| nerve tibial | 306 | 0.11 |
| sun-exposed skin | 345 | 0.09 |
| Thyroid | 315 | 0.10 |
| whole blood | 316 | 0.08 |

Table S2. DeepSEA score differences between called TF binding regions and 1,100 bp regions centered at peakMax positions. For each TF binding event, we calculate the DeepSEA score difference between the two versions and then, we normalized the score difference by the reference binding score of given TF binding region (i.e. relative score difference). For each TF, we calculate the mean score difference (score\_diff column), standard deviation of the score difference (diff\_sd), the mean relative score difference (rel\_diff) and the standard deviation of relative score difference (rel\_sd) across all binding events of that TF.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TF | score\_diff | diff\_sd | rel\_diff | rel\_sd |
| ATF2 | 5.84E-05 | 1.10E-02 | -1.71E-03 | 6.67E-02 |
| ATF3 | -1.23E-03 | 8.98E-03 | -1.95E-02 | 8.26E-02 |
| BATF | -2.99E-04 | 1.35E-02 | -5.14E-03 | 8.85E-02 |
| BCL11A | -3.47E-05 | 8.65E-03 | -6.55E-03 | 1.05E-01 |
| BCL3 | -2.13E-04 | 8.86E-03 | -1.82E-03 | 6.50E-02 |
| BCLAF1 | 4.84E-04 | 5.65E-03 | 2.57E-03 | 6.24E-02 |
| BHLHE40 | -4.99E-04 | 8.96E-03 | -6.27E-03 | 9.52E-02 |
| BRCA1 | 1.88E-04 | 1.33E-03 | 6.62E-04 | 3.42E-02 |
| CEBPB | -6.25E-04 | 1.03E-02 | -1.01E-02 | 8.83E-02 |
| c-Fos | 1.59E-03 | 6.97E-03 | 4.22E-03 | 4.43E-02 |
| CHD1 | -1.60E-04 | 3.88E-03 | -4.91E-03 | 7.83E-02 |
| CHD2 | -3.85E-04 | 1.38E-02 | -6.97E-03 | 8.02E-02 |
| c-Myc | 5.86E-04 | 6.40E-03 | 1.98E-03 | 3.07E-02 |
| COREST | 5.22E-05 | 2.39E-03 | -6.42E-04 | 7.35E-02 |
| CTCF | 3.76E-04 | 2.01E-02 | -4.61E-03 | 9.80E-02 |
| DNase | 4.28E-04 | 1.38E-02 | 9.52E-04 | 7.28E-02 |
| E2F4 | -1.30E-04 | 3.81E-03 | -1.06E-03 | 3.75E-02 |
| EBF1 | 6.86E-05 | 7.50E-03 | -5.64E-03 | 1.41E-01 |
| Egr-1 | 7.73E-05 | 1.02E-02 | -2.54E-03 | 6.74E-02 |
| ELF1 | 3.77E-04 | 1.28E-02 | -2.55E-04 | 5.17E-02 |
| ELK1 | -6.34E-04 | 7.52E-03 | -8.14E-03 | 7.25E-02 |
| ETS1 | 5.05E-04 | 7.58E-03 | 1.05E-03 | 6.84E-02 |
| EZH2 | -4.66E-05 | 1.48E-03 | -4.07E-03 | 7.13E-02 |
| FOXM1 | -2.10E-04 | 8.90E-03 | -3.10E-03 | 6.73E-02 |
| GABP | 1.66E-04 | 1.64E-02 | -2.23E-03 | 1.06E-01 |
| IKZF1 | 5.78E-04 | 7.65E-03 | 1.52E-03 | 5.33E-02 |
| IRF4 | -8.57E-04 | 1.06E-02 | -1.38E-02 | 1.16E-01 |
| JunD | -3.49E-05 | 4.15E-03 | -1.03E-02 | 1.49E-01 |
| Max | 5.64E-06 | 9.52E-03 | -4.21E-04 | 5.89E-02 |
| MAZ | -1.70E-04 | 8.82E-03 | -3.42E-03 | 6.61E-02 |
| MEF2A | -2.58E-04 | 7.83E-03 | -1.46E-02 | 1.43E-01 |
| MEF2C | -1.27E-04 | 5.79E-03 | 1.75E-03 | 9.44E-02 |
| MTA3 | -3.47E-04 | 9.61E-03 | -4.96E-03 | 8.06E-02 |
| Mxi1 | -5.07E-04 | 1.21E-02 | -5.41E-03 | 7.25E-02 |
| NFATC1 | -3.33E-04 | 5.14E-03 | -3.29E-03 | 5.35E-02 |
| NF-E2 | 3.24E-04 | 3.47E-03 | 3.02E-03 | 1.89E-02 |
| NFIC | -1.84E-04 | 1.21E-02 | -4.75E-03 | 7.53E-02 |
| NFKB | 2.48E-04 | 9.21E-03 | -3.16E-03 | 9.49E-02 |
| NF-YA | 1.00E-04 | 3.67E-03 | 4.77E-04 | 2.94E-02 |
| NF-YB | -1.48E-04 | 7.40E-03 | -2.54E-03 | 5.04E-02 |
| Nrf1 | 4.45E-04 | 1.20E-02 | -1.57E-03 | 6.78E-02 |
| NRSF | -1.90E-04 | 1.44E-02 | -3.61E-03 | 8.17E-02 |
| p300 | -6.59E-05 | 7.13E-03 | -8.57E-03 | 1.26E-01 |
| PAX5-C20 | -7.78E-05 | 8.62E-03 | -5.14E-03 | 8.50E-02 |
| PAX5-N19 | -1.69E-04 | 8.10E-03 | -8.16E-03 | 1.17E-01 |
| Pbx3 | 1.86E-04 | 6.52E-03 | -3.46E-03 | 7.76E-02 |
| PML | 6.14E-04 | 1.76E-02 | -8.44E-04 | 5.85E-02 |
| Pol2 | -2.30E-04 | 9.08E-03 | -7.58E-04 | 4.21E-02 |
| Pol3 | 6.05E-05 | 3.15E-04 | -4.40E-04 | 3.64E-02 |
| POU2F2 | 9.13E-04 | 1.20E-02 | 1.52E-03 | 7.46E-02 |
| PU.1 | -6.90E-04 | 1.49E-02 | -7.03E-03 | 1.09E-01 |
| Rad21 | 5.63E-04 | 1.71E-02 | -3.31E-03 | 9.62E-02 |
| RFX5 | 2.63E-04 | 3.80E-03 | -2.99E-03 | 3.73E-02 |
| RUNX3 | -7.05E-05 | 1.31E-02 | -2.92E-03 | 7.68E-02 |
| RXRA | 1.63E-05 | 1.64E-03 | 3.92E-04 | 8.61E-02 |
| SIN3A | 8.72E-04 | 1.84E-02 | 5.35E-04 | 7.80E-02 |
| SIX5 | -1.37E-03 | 1.70E-02 | 2.54E-04 | 7.58E-02 |
| SMC3 | 3.97E-04 | 1.82E-02 | -3.54E-03 | 1.37E-01 |
| SP1 | -3.42E-04 | 1.18E-02 | -2.76E-03 | 8.23E-02 |
| SRF | 7.26E-05 | 7.46E-03 | -2.68E-03 | 5.72E-02 |
| STAT1 | 9.01E-05 | 8.11E-04 | -1.66E-03 | 3.75E-02 |
| STAT3 | -3.49E-04 | 6.23E-03 | -1.38E-02 | 9.24E-02 |
| STAT5A | -7.36E-04 | 8.53E-03 | -6.83E-03 | 6.50E-02 |
| TAF1 | 4.50E-04 | 1.67E-02 | -4.79E-03 | 6.06E-02 |
| TBLR1 | 4.87E-05 | 1.02E-02 | -2.92E-03 | 7.70E-02 |
| TBP | 5.51E-04 | 1.28E-02 | -4.01E-03 | 1.02E-01 |
| TCF12 | 9.67E-05 | 9.32E-03 | -4.79E-03 | 7.97E-02 |
| TCF3 | 4.57E-05 | 8.62E-03 | -3.01E-03 | 6.62E-02 |
| TR4 | -4.49E-04 | 5.16E-03 | -1.34E-03 | 2.74E-02 |
| USF-1 | 4.90E-04 | 1.39E-02 | -1.25E-04 | 7.95E-02 |
| USF2 | 2.20E-04 | 8.10E-03 | 4.89E-03 | 9.18E-02 |
| WHIP | 1.94E-04 | 4.64E-03 | -3.14E-04 | 5.81E-02 |
| YY1 | 2.98E-04 | 1.33E-02 | -2.14E-03 | 7.36E-02 |
| ZBTB33 | 3.94E-04 | 3.81E-03 | -6.16E-04 | 1.24E-01 |
| ZEB1 | 3.73E-04 | 5.20E-03 | 3.72E-03 | 5.76E-02 |
| Znf143 | -9.47E-04 | 2.05E-02 | -8.71E-03 | 1.03E-01 |
| ZNF274 | -9.18E-04 | 4.68E-03 | -5.51E-03 | 2.46E-02 |
| ZZZ3 | -4.33E-05 | 7.08E-04 | 2.73E-03 | 3.59E-02 |
| mean | -7.60E-07 | 9.12E-03 | -3.14E-03 | 7.48E-02 |

**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. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T *et al*: **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**(7493):455-461.

4. Ding Z, Ni Y, Timmer SW, Lee BK, Battenhouse A, Louzada S, Yang F, Dunham I, Crawford GE, Lieb JD *et al*: **Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association**. *PLoS genetics* 2014, **10**(11):e1004798.

5. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A *et al*: **Population Variation and Genetic Control of Modular Chromatin Architecture in Humans**. *Cell* 2015, **162**(5):1039-1050.

6. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities**. *Molecular cell* 2010, **38**(4):576-589.

7. 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.

8. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE *et al*: **Variation in transcription factor binding among humans**. *Science* 2010, **328**(5975):232-235.

9. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV *et al*: **Extensive variation in chromatin states across humans**. *Science* 2013, **342**(6159):750-752.

10. Champely S: **pwr: Basic functions for power analysis**. *R package version* 2012, **1**(1).