Supplementary Information

Supplementary Text

CAGE-derived genomic regions selection

The user can select an expression level threshold by specifying either the relative expression (corresponding to expression specificity) or a combination of the raw tag count and the normalized number of tags per million. CAGE peaks that meet or exceed the provided expression characteristics in any of the selected set of FANTOM5 samples are retrieved from the underlying database to be used for defining the foreground set of genomic regions. This selected set of CAGE peaks can be further filtered based on proximity to specific genes (using a 500 bp maximal distance between CAGE peaks and Ensembl transcription start sites (Cunningham et al., 2015) for user-supplied peaks or using the peak-to-gene associations for FANTOM5 peaks from (The FANTOM Consortium, 2014)), overlap with user-supplied regions of interest (such as ChIP-seq peak regions in a BED-formatted file), or prediction as true TSSs by the TSS classifier in (The FANTOM Consortium, 2014) for FANTOM5 CAGE-peaks.

TF binding profile clusters

Clusters of JASPAR (version 2016 (Mathelier et al., 2015)) TF binding profiles were computed using the *matrix-clustering* tool from RSAT (version 2015 (Medina-Rivera et al., 2015; Castro-Mondragon et al., in preparation)). We used the *average* linkage method with the *Ncor* similarity metric to compute the clusters with the following thresholds: *Ncor=0.55* and *cor=0.75*. The tool computed 136 clusters of TF binding profiles. CAGEd-oPOSSUM uses these clusters to combine TFBSs predicted from TF binding profiles in the same cluster and to compute the enrichment scores associated to the clusters as implemented in the TFBS Cluster Analysis of oPOSSUM3 (Kwon, Arenillas, Hunt, & Wasserman, 2012).

Precomputation of TFBSs for time efficiency

The flanking regions of 2,000 bp were applied to each FANTOM5 CAGE peak and overlapping regions were merged to create a set of maximal spanning, non-overlapping CAGE-derived genomic regions. The genomic sequences were extracted from Ensembl (Cunningham et al., 2015) and scanned with the TF binding profiles from JASPAR (Mathelier et al., 2016). TFBSs were predicted where the corresponding position weight matrix relative score was above 80% (as in oPOSSUM3 (Kwon et al., 2012)).

Examples of application

The default parameters used for the three case examples were as follows. A CAGE peak relative expression level of at least 1 was used to select transcription start sites specific to the samples.

Flanking regions of 500 bp upstream and downstream were extracted. Background sequences matching the %GC composition and length of selected regulatory regions were generated with HOMER (Heinz et al., 2010). All JASPAR 2016 CORE vertebrate profiles (Mathelier et al., 2016) with a minimum information content of 8 bits were used to predict TFBSs with a relative score of at least 85% for the oPOSSUM3 analysis.

The most enriched profile predicted by both oPOSSUM3 (using the Fisher scores accounting for the number of genomic regions containing at least one predicted TFBS) and HOMER is associated with the HNF4A TF for the liver sample (Supplementary Figure 1 and Supplementary Data). The HNF4A TF is a well-characterized TF involved in the regulation of several biological functions in liver (Babeu & Boudreau, 2014). Using the three samples corresponding to CD19-positive B-cells, the most enriched profiles (from both oPOSSUM3 and HOMER) are associated with ETS-related factors (Supplementary Figure 2 and Supplementary Data). Several of these ETS-related factors profiles are associated with TFs already known to be critical for B-cell development such as GABPA (Xue et al., 2007), ETS1 (Eyquem et al., 2004), PU.1/SPI1 (Sokalski et al., 2011), and SPIB (Sokalski et al., 2011). Of the top scoring TFs, RELA is the only non ETS-related factor predicted; it is known to regulate the development of B-cells and has a critical role in the regulation of B-cell survival (Prendes, Zheng, & Beg, 2003). Finally, from the testis samples, CAGEd-oPOSSUM identified RFX-related factors as the most enriched profiles with both oPOSSUM3 and HOMER (Supplementary Figure 3 and Supplementary Data). RFX TFs have already been described to be important in testis during spermatogenesis (Morotomi-Yano et al., 2002; Wolfe, van Wert, & Grimes, 2006; Wolfe, Vanwert, & Grimes, 2008).

Acknowledgements

We thank Miroslav Hatas for systems support and Dora Pak for management support, Wenqiang Shi and Chih-Yu Chen for comments on the manuscript, Chih-Yu Chen and Andrew Kwon for testing the software and providing comments, and all members of the Wasserman lab for insightful discussions. We thank Jaime Castro-Mondragon, Morgane Thomas-Chollier, and Jacques van Helden for helpful discussions and for providing the unpublished JASPAR 2016 TF binding profile clustering results using the *matrix-clustering* tool of RSAT. We thank Terry Meehan and Chris Mungall for support with the FANTOM5 ontology. We would like to thank all members of the FANTOM5 consortium for contributing to generation of samples and analysis of the data set and thank GeNAS for data production.

Funding

The Wasserman lab has been funded through the Genome Canada Large Scale Research Grant 174CDE. Funding has been provided to AM and WWW by the British Columbia Children's Hospital Foundation and the Child and Family Research Institute, Vancouver, Canada. AM was supported by the University of Oslo through the Centre of Molecular Medicine Norway (NCMM), which is a part of the Nordic European Molecular Biology Laboratory (EMBL) partnership. FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) to YH, Grant from MEXT for the RIKEN Preventive Medicine and Diagnosis Innovation Program to YH, Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to YH and Grant from MEXT

to the RIKEN Center for Life Science Technologies.

Authors' contributions

AM and WWW were responsible for project conception and oversight. DJA implemented the CAGEd-oPOSSUM web tool. TL was responsible for tag mapping. HK managed the data handling. ARRF was responsible for FANTOM5 management and its concept. DJA, WWW, and AM wrote the manuscript.

References

Babeu, J.-P., & Boudreau, F. (2014). Hepatocyte nuclear factor 4-alpha involvement in liver and intestinal inflammatory networks. *World Journal of Gastroenterology*, *20*(1), 22–30. http://doi.org/10.3748/wjg.v20.i1.22

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., … Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research*, *43*(D1), D662–D669. http://doi.org/10.1093/nar/gku1010

Eyquem, S., Chemin, K., Fasseu, M., Chopin, M., Sigaux, F., Cumano, A., & Bories, J.-C. (2004). The development of early and mature B cells is impaired in mice deficient for the Ets-1 transcription factor. *European Journal of Immunology*, *34*(11), 3187–3196. http://doi.org/10.1002/eji.200425352

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589. http://doi.org/10.1016/j.molcel.2010.05.004

Kwon, A. T., Arenillas, D. J., Hunt, R. W., & Wasserman, W. W. (2012). oPOSSUM-3: Advanced Analysis of Regulatory Motif Over-Representation Across Genes or ChIP-Seq Datasets. *G3; Genes|Genomes|Genetics*, *2*(9), 987–1002. http://doi.org/10.1534/g3.112.003202

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C., Denay, G., Lee, J., … Wasserman, W. W. (2015). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *44*(D1), D110–D115. http://doi.org/10.1093/nar/gkv1176

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., … van Helden, J. (2015). RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Research*, *43*(W1), W50–W56. http://doi.org/10.1093/nar/gkv362

Morotomi-Yano, K., Yano, K., Saito, H., Sun, Z., Iwama, A., & Miki, Y. (2002). Human regulatory factor X 4 (RFX4) is a testis-specific dimeric DNA-binding protein that cooperates with other human RFX members. *The Journal of Biological Chemistry*, *277*(1), 836–842. http://doi.org/10.1074/jbc.M108638200

Prendes, M., Zheng, Y., & Beg, A. A. (2003). Regulation of developing B cell survival by RelA-containing NF-kappa B complexes. *Journal of Immunology (Baltimore, Md.: 1950)*, *171*(8), 3963–3969.

Sokalski, K. M., Li, S. K. H., Welch, I., Cadieux-Pitre, H.-A. T., Gruca, M. R., & DeKoter, R. P. (2011). Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia. *Blood*, *118*(10), 2801–2808. http://doi.org/10.1182/blood-2011-02-335539

The FANTOM Consortium. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470. http://doi.org/10.1038/nature13182

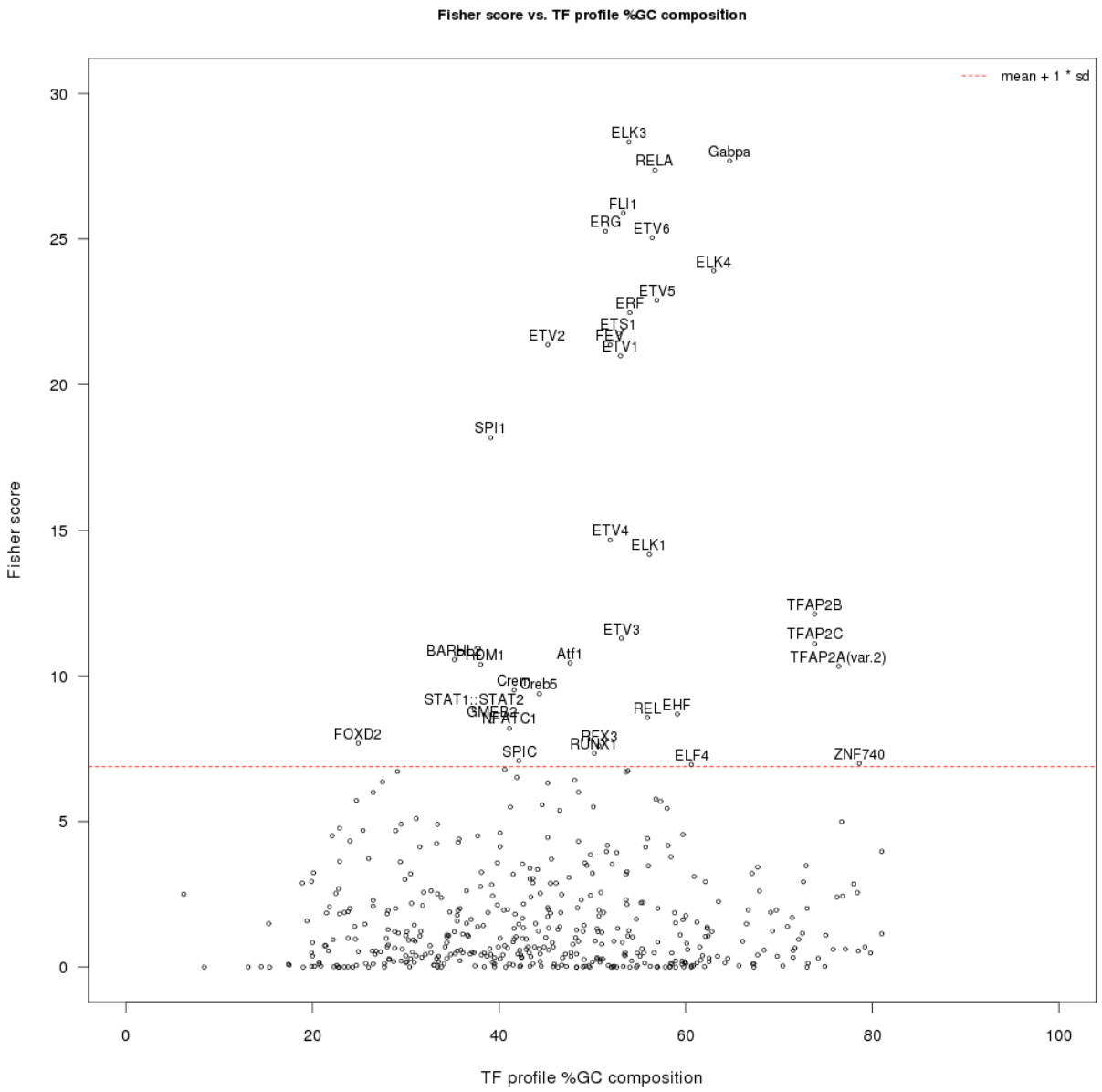
Wolfe, S. A., van Wert, J., & Grimes, S. R. (2006). Transcription factor RFX2 is abundant in rat testis and enriched in nuclei of primary spermatocytes where it appears to be required for transcription of the testis-specific histone H1t gene. *Journal of Cellular Biochemistry*, *99*(3), 735–746. http://doi.org/10.1002/jcb.20959

Wolfe, S. A., Vanwert, J. M., & Grimes, S. R. (2008). Transcription factor RFX4 binding to the testis-specific histone H1t promoter in spermatocytes may be important for regulation of H1t gene transcription during spermatogenesis. *Journal of Cellular Biochemistry*, *105*(1), 61–69. http://doi.org/10.1002/jcb.21793

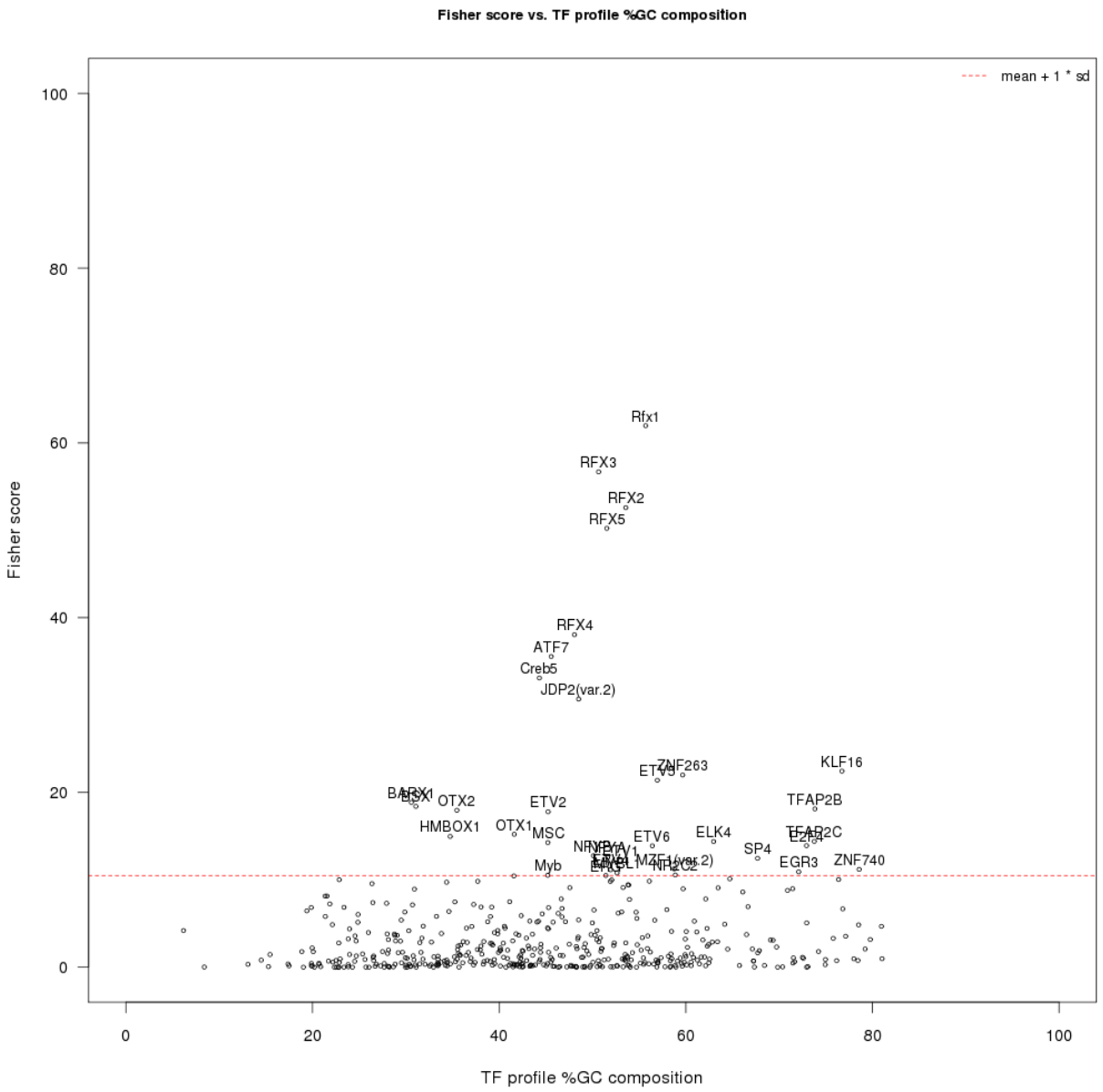
Xue, H.-H., Bollenbacher-Reilley, J., Wu, Z., Spolski, R., Jing, X., Zhang, Y.-C., … Leonard, W. J. (2007). The transcription factor GABP is a critical regulator of B lymphocyte development. *Immunity*, *26*(4), 421–431. http://doi.org/10.1016/j.immuni.2007.03.010

Supplementary Figures

Supplementary Figure 1. TF binding profiles over-representation results on the liver FANTOM5 sample (FF:10018-101C9). For each TF binding profile analyzed, we plot the Fisher-score on the y-axis along with the %GC composition of the TF binding profile on the x-axis. The name of the TFs associated with the profiles are provided when the Fisher-score is above a defined threshold (plot as a dashed red line). The threshold corresponds to *mean* + 1 * *sd* where *mean* and *sd* correspond to the average mean and standard deviation, respectively, of the distribution of all Fisher-scores. We note that the most over-represented profile is associated with the HNF4A TF, a known regulator in liver.



Supplementary Figure 2. TF binding profiles over-representation results on CD19-positive B cells, donor 1, 2 & 3 (FF:11544-120B5, FF:11624-122B4, and FF:11705-123B4) FANTOM5 samples. For each TF binding profile analyzed, we plot the Fisher-score on the y-axis along with the %GC composition of the TF binding profile on the x-axis. The name of the TFs associated with the profiles are provided when the Fisher-score is above a defined threshold (plot as a dashed red line). The threshold corresponds to *mean* + 1 * *sd* where *mean* and *sd* correspond to the average mean and standard deviation, respectively, of the distribution of all Fisher-scores. We note that ETS-related factors are the most enriched profiles.



Supplementary Figure 3. TF binding profiles over-representation results on testis FANTOM5 samples (FF:10026-101D8 and FF:10096-102C6). For each TF binding profile analyzed, we plot the Fisher-score on the y-axis along with the %GC composition of the TF binding profile on the x-axis. The name of the TFs associated with the profiles are provided when the Fisher-score is above a defined threshold (plot as a dashed red line). The threshold corresponds to *mean* + 1 * *sd* where *mean* and *sd* correspond to the average mean and standard deviation, respectively, of the distribution of all the Fisher-scores. We note that profiles associated with RFX TFs are the most enriched profiles.