CAGEd-oPOSSUM: motif enrichment analysis from CAGE-derived TSSs

David J. Arenillas[1], Alistair R.R. Forrest[2,3], Hideya Kawaji[2,3,4], Timo Lassmann[2,3,5], The FANTOM Consortium, Wyeth W. Wasserman[1,*], and Anthony Mathelier[1,*]

[1] Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, BC Children's Hospital, Department of Medical Genetics, University of British Columbia, Vancouver, BC V5Z 4H4, Canada.
[2] Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Kanagawa 230-0045, Japan
[3] RIKEN Omics Science Center
[4] RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako
[5] Telethon Kids Institute, The University of Western Australia, Subiaco, WA 6008, Australia

[*] Corresponding authors: WWW: wyeth@cmmt.ubc.ca, AM: anthony.mathelier@gmail.com

Running title: CAGEd-oPOSSUM: motif enrichment from CAGE TSSs

Abstract

Summary: With the emergence of large-scale Cap Analysis of Gene Expression data sets from individual labs and the FANTOM consortium, one can now analyze the *cis*-regulatory regions associated with gene transcription at an unprecedented level of refinement. By coupling transcription factor binding site (TFBS) enrichment analysis with CAGE-derived *cis*-regulatory regions, CAGEd-oPOSSUM can identify TFs that act as key regulators of genes involved in specific mammalian cell and tissue types. The webtool allows for the analysis of CAGE-derived transcription start sites (TSSs) either (i) provided by the user or (ii) selected from ~1,300 mammalian samples from the FANTOM5 project with pre-computed TFBS predicted with JASPAR TF binding profiles. The tool can help power insights into the transcriptional regulation of genes through the study of the specific usage of TSSs within specific cell types and/or under specific conditions.

Availability and implementation: The CAGEd-oPOSUM web tool is implemented in Perl, MySQL, and Apache and is available at http://cagedop.cmmt.ubc.ca/CAGEd_oPOSSUM. The source code is freely available from GitHub at https://github.com/wassermanlab/CAGEd-oPOSSUM.

Contact: dave@cmmt.ubc.ca, wyeth@cmmt.ubc.ca, and anthony.mathelier@gmail.com.

Supplementary information: Supplementary Figures are available on the journal website. Supplementary Data is available at http://cagedop.cmmt.ubc.ca/CAGEd_oPOSSUM/archived_results/.

Introduction

The Cap Analysis of Gene Expression (CAGE) technology [1] has revolutionized our capacity to analyze the *cis*-regulatory regions that are active in cell types and tissues at specific time-points. Through the FANTOM5 project [2,3], 889 human and 389 mouse samples (from primary cells, tissues, and cancer cell lines) have been subjected to CAGE experiments to highlight active transcription start sites (TSSs, derived from CAGE peaks) and their usage in mammals [4]. Over $1x10^6$ human and $6x10^5$ mouse CAGE peaks were identified, providing both the precise location of TSSs and a quantitative measure of transcriptional activity in the studied samples. It provides the scientific community with an unprecedented opportunity to analyze the *cis*-regulatory regions acting upon the transcription of RNAs from the identified TSSs.

The regulation of gene transcription is driven by the binding of transcription factor (TF) proteins to the DNA at *cis*-regulatory regions. Most of these TFs bind to the DNA in a sequence-specific manner at TF binding sites (TFBSs). Classically, TFBSs recognized by a TF are modeled using TF binding profiles, which capture the prefered DNA motifs bound by the TF [5]. Tools such as oPOSSUM [6–8] and HOMER [9] provide the ability to perform TF binding profile over-representation analyses on sets of *cis*-regulatory regions to infer the likely TFs acting upon these regions (for instance, gene promoter regions). Specifically, the tools look for the enrichment of motifs in a set of foreground sequences compared to a set of background ones.

By combining sample-specific CAGE-derived *cis*-regulatory regions with TF binding profile enrichment analysis, it now becomes possible to predict the TFs that are likely involved in the regulation of active genes associated with TSSs. One key advantage of the CAGE technology is its capacity to capture the specific usage of multiple TSSs (and therefore promoters) associated with the same gene in different samples. We hereby introduce the CAGEd-oPOSSUM tool to perform TF binding profile enrichment analysis from CAGE data.

This work is part of the FANTOM5 project. Data downloads, genomic tools, and co-published manuscripts are summarized here http://fantom.gsc.riken.jp/5/.

Usage and implementation

Through a web interface, CAGEd-oPOSSUM provides functionalities to determine the enrichment of TF binding profiles in a given set of user selected CAGE-derived regulatory regions by comparing the frequency of predicted TFBSs in the regions versus their frequency in background regions.

CAGE derived *cis*-regulatory region selection

The user can either provide a set of CAGE peak regions as a BED-formatted file, supply a set of FANTOM5 CAGE peak identifiers [2], or select FANTOM5 samples from which CAGE peaks will be selected according to expression level characteristics. To facilitate ease of selection in the latter case, an expandable tree of FANTOM5 samples, based on the corresponding FANTOM5 sample ontology [2] derived from cell types [10], anatomical [11], and disease ontologies [12], is displayed. Specifically, an ontology tree is provided from which the user may select one or more

FANTOM5 samples. The user then selects 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 which 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 *cis*-regulatory regions. This selected set of CAGE peaks can be further filtered based on proximity to specific genes, 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 [2].

A similar procedure is used to select a background set of CAGE peaks. As an additional option for the background, the user may choose to use a random set of regions that is automatically generated to match the %GC composition and length characteristics of the foreground CAGE peak regions using the HOMER software [9].

TFBS prediction and enrichment analysis

Once the set of CAGE peaks are determined, CAGEd-oPOSSUM will extract *cis*-regulatory regions around the corresponding TSSs (for instance using 500bp up- and downstream). Note that overlapping regions are merged before the prediction of TFBSs using BEDTools [13] to avoid counting the same TFBS multiple times. A set of TF binding profiles is then selected from the JASPAR database [14] or provided by the user to predict TFBSs in the *cis*-regulatory regions. By default, two statistical tests are employed to determine significance as previously described in oPOSSUM3 [8]: (1) a Z-score based on normal approximation to the binomial distribution that measures the change in the relative number of TFBSs in the foreground region set compared with the background set, and (2) a Fisher score based on a one-tailed Fisher exact probability which assesses the number of regions with a TFBS in the foreground set versus the background set. CAGEd-oPOSSUM outputs a result table consisting of a ranked list of TF binding profiles based on the enrichment scores. As stated above, the enrichment analysis is performed using oPOSSUM3 statistical tests by default. CAGEd-oPOSSUM additionally allows the user to perform motif enrichment analysis through the HOMER stand-alone package (specifically using the findMotifsGenome.pl script) [9] as a complementary search.

Precomputation of TFBSs for time efficiency

As the prediction of TFBSs can be time-consuming on a large set of *cis*-regulatory regions, a precomputation was performed from all FANTOM5 CAGE peaks using TFBS profiles from the 2016 release of the JASPAR database [14]. The resulting predicted TFBSs were stored in a MySQL database. Briefly, 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 *cis*-regulatory regions. The genomic sequences corresponding to these *cis*-regulatory regions were extracted from Ensembl [15] and scanned with the TF binding profiles. TFBSs were predicted where the corresponding position weight matrix relative score was above 80% (as in [8]). All FANTOM5 CAGE peak positions, samples, and expression level data along with the computed merged regions and predicted TFBSs were stored in the database. This precomputation allows for the fast analyses of FANTOM5 samples using

JASPAR TF binding profiles as the predicted TFBSs are fetched directly from the database to compute the statistical enrichment scores.

If the user provides CAGE peak coordinates or TF binding profiles, a similar process of extracting flanking regions, merging overlapping regions, retrieving the corresponding sequences, and scanning these sequences for TFBSs, as described above to build the database, is executed in real time.

Examples of application

As case examples, we tested CAGEd-oPOSSUM on three input FANTOM5 data sets from human primary cell and tissue samples: liver, adult pool1 (FANTOM5 identifier FF:10018-101C9); CD19-positive B-cells, donor 1, 2, and 3 (FF:11544-120B5, FF:11624-122B4, and FF:11705-123B4); and testis, adult, pool 1 and 2 (FF:10026-101D8 and FF:10096-102C6). Default analysis parameters were used to perform TF binding profile enrichment analyses with both oPOSSUM and HOMER for each of these data sets. Specifically, a CAGE peak relative expression level above 1 was used to select TSSs 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, and all JASPAR 2016 CORE vertebrate profiles with a minimum information content of 8 bits were used to predict TFBSs with a relative score of at least 85% for the oPOSSUM analysis.

The most enriched profile predicted by both oPOSSUM (using the Fisher scores accounting for the number of *cis*-regulatory 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 [16]. Using the three samples corresponding to CD19-positive B-cells, the most enriched profiles (from both oPOSSUM 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-cells development such as GABPA [17], ETS1 [18], PU.1/SPI1 [19], and SPIB [19]. In 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-cells survival [20]. Finally, from the testis samples, CAGEd-oPOSSUM predicted RFX-related factors as the most enriched profiles (Supplementary Figure 3 and Supplementary Data). RFX TFs have already been described to be important in testis during spermatogenesis [21–23]. These results highlight the accuracy of CAGEd-oPOSSUM to predict key master regulators driving specific expression in biological samples.

Conclusion

Building upon the large-scale availability of CAGE experiments and established tools (oPOSSUM3 and HOMER) for motif enrichment analysis, we introduced a new computational tool, CAGEd-oPOSSUM, to predict TFs regulating transcriptional events in specific cell types and tissues. CAGEd-oPOSSUM will empower the community with the capacity to highlight specific TFs that may act as key transcriptional regulators in their biological samples of interest when CAGE experiments are available.

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. 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.

Acknowledgments

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 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.

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

1. Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T, Katayama S, Kojima M, et al.: Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Research 2011 May 19;21:1150–1159.

2. The FANTOM Consortium and the RIKEN PMI and CLST (dgt): A promoter-level mammalian expression atlas. Nature 2014 Mar 27;507:462–470.

3. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al.: An atlas of active enhancers across human cell types and tissues. Nature 2014 Mar 26;507:455–461.

4. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al.: Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology 2015;16:22.

5. Stormo GD: Modeling the specificity of protein-DNA interactions. Quant Biol 2013 Jun;1:115–130.

6. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, et al.: oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Research 2005 Jun 2;33:3154–3164.

7. Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT, Wasserman WW: oPOSSUM: integrated tools for analysis of regulatory motif over-representation. Nucleic Acids Research 2007 May 8;35:W245–W252.

8. Kwon AT, Arenillas DJ, Hunt RW, Wasserman WW: oPOSSUM-3: Advanced Analysis of Regulatory Motif Over-Representation Across Genes or ChIP-Seq Datasets. G3&#58; Genes|Genomes|Genetics 2012 Sep 1;2:987–1002.

9. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al.: Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell 2010 May 28;38:576–589.

10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000 May;25:25–29.

11. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA: Uberon, an integrative multi-species anatomy ontology. Genome Biol 2012;13:R5.

12. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007 Nov;25:1251–1255.

13. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010 Mar 15;26:841–842.

14. Mathelier A, Fornes O, Arenillas DJ, Chen C, Denay G, Lee J, et al.: JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucl Acids Res 2015 Nov 3;gkv1176.

15. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al.: Ensembl 2015. Nucl Acids Res 2015 Jan 28;43:D662–D669.

16. Babeu J-P, Boudreau F: Hepatocyte nuclear factor 4-alpha involvement in liver and intestinal inflammatory networks. World J Gastroenterol 2014 Jan 7;20:22–30.

17. Xue H-H, Bollenbacher-Reilley J, Wu Z, Spolski R, Jing X, Zhang Y-C, et al.: The transcription factor GABP is a critical regulator of B lymphocyte development. Immunity 2007 Apr;26:421–431.

18. Eyquem S, Chemin K, Fasseu M, Chopin M, Sigaux F, Cumano A, et al.: The development of early and mature B cells is impaired in mice deficient for the Ets-1 transcription factor. Eur J Immunol 2004 Nov;34:3187–3196.

19. Sokalski KM, Li SKH, Welch I, Cadieux-Pitre H-AT, Gruca MR, DeKoter RP: Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia. Blood 2011 Sep 8;118:2801–2808.

20. Prendes M, Zheng Y, Beg AA: Regulation of developing B cell survival by RelA-containing NF-kappa B complexes. J Immunol 2003 Oct 15;171:3963–3969.

21. Wolfe SA, van Wert J, Grimes SR: 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. J Cell Biochem 2006 Oct 15;99:735–746.

22. Morotomi-Yano K, Yano K, Saito H, Sun Z, Iwama A, Miki Y: Human regulatory factor X 4 (RFX4) is a testis-specific dimeric DNA-binding protein that cooperates with other human RFX members. J Biol Chem 2002 Jan 4;277:836–842.

23. Wolfe SA, Vanwert JM, Grimes SR: Transcription factor RFX4 binding to the testis-specific histone H1t promoter in spermatocytes may be important for regulation of H1t gene transcription during spermatogenesis. J Cell Biochem 2008 Sep 1;105:61–69.