Nucleic Acids Research Advance Access published June 18, 2007

Nucleic Acids Research, 2007, **1–8**
doi:10.1093/nar/gkm427

oPOSSUM: integrated tools for analysis of regulatory motif over-representation

Shannan J. Ho Sui[1,2], Debra L. Fulton[1,2], David J. Arenillas[1,3], Andrew T. Kwon[1,2] and Wyeth W. Wasserman[1,3,*]

[1]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, [2]Genetics Graduate Program and [3]Department of Medical Genetics, University of British Columbia, Vancouver BC, Canada

Received January 31, 2007; Revised April 13, 2007; Accepted May 10, 2007

ABSTRACT

The identification of over-represented transcription factor binding sites from sets of co-expressed genes provides insights into the mechanisms of regulation for diverse biological contexts. oPOSSUM, an internet-based system for such studies of regulation, has been improved and expanded in this new release. New features include a worm-specific version for investigating binding sites conserved between *Caenorhabditis elegans* and *C. briggsae*, as well as a yeast-specific version for the analysis of co-expressed sets of *Saccharomyces cerevisiae* genes. The human and mouse applications feature improvements in ortholog mapping, sequence alignments and the delineation of multiple alternative promoters. oPOSSUM2, introduced for the analysis of over-represented combinations of motifs in human and mouse genes, has been integrated with the original oPOSSUM system. Analysis using user-defined background gene sets is now supported. The transcription factor binding site models have been updated to include new profiles from the JASPAR database. oPOSSUM is available at http://www.cisreg.ca/oPOSSUM/

INTRODUCTION

Functional genomics research often generates lists of genes with observed common properties, such as coordinated expression. For many studies, a key challenge is the generation of relevant and testable hypotheses about the regulatory networks and pathways that underlie observed co-expression. Our strategy for elucidating regulatory mechanisms identifies over-represented sequence motifs that are present in the upstream regulatory regions of genes. The motifs may represent transcription factor binding sites (TFBSs) that have a role in regulating expression.

oPOSSUM (1) and oPOSSUM2 (2) were developed to identify over-represented, predicted TFBSs and combinations of predicted TFBSs, respectively, in sets of human and mouse genes. The user inputs a list of related genes, selects the TFBS profile set to be included in the analysis, and the algorithm determines which, if any, predicted TFBSs occur in the promoters of the set of input genes more often than would be expected by chance. Both analytic approaches rely on a database of aligned, orthologous human and mouse sequences, and the delineation of conserved regions within which TFBS predictions are analyzed. While the approach does not explicitly address uncharacterized transcription factors (TFs), the effective coverage is broadened by the fact that members within certain structural families of TFs can exhibit similarities in binding specificity. While intra-class similarity is not always the case, as exemplified by the zinc-finger family of TFs (3), the observation holds true for many TF families (4,5).

Here we describe the new release of the oPOSSUM system, which integrates the two previously developed applications, and has been expanded to accommodate new species (yeast and worms). It also includes new methods for orthology assignment, transcription start site (TSS) determination and sequence alignment.

MATERIALS AND METHODS

Over-representation analysis

oPOSSUM single site analysis (SSA). The oPOSSUM system for identifying over-represented TFBSs in sets of co-expressed genes first focused on SSA (1). Two scores were developed to assess over-representation, one at the TFBS occurrence level and the other at the gene level. The Z-score, based on the normal approximation to the binomial distribution, indicates how far and in what direction the number of TFBS occurrences deviates from

*To whom correspondence should be addressed. Tel: +1 604 875 3812; Fax: +1 604 875 3819; Email: wyeth@cmmt.ubc.ca the background distribution's mean. The second score, the Fisher exact test, indicates if the proportion of genes containing the TFBS is greater than would be expected by chance. TFBS predictions situated within overlapping alternative promoters are counted only once when calculating over-representation in human and mouse genes. For *Caenorhabditis elegans* genes in operons, TFBS predictions in the upstream region of the first gene in the operon apply to all genes in the operon.

oPOSSUM combination site analysis (CSA). TFBSs do not act in isolation to initiate the transcription process. Transcriptional regulation can be viewed as mediated by arrays of *cis*-regulatory sequences, termed *cis*-regulatory modules (CRMs), which are bound by multiple TFs. In oPOSSUM2, Huang *et al.* (2) address the detection of over-represented sets of TFBSs in the promoters of a set of co-expressed genes. In brief, the method reduces combinatorial complexity through an initial clustering step, which partitions similar TFBS profiles into groups, herein denoted 'TFBS classes', along with an analysis step to determine a TFBS class representative profile for each TFBS class, which is used to detect over-represented sets of TFBS classes. Since each distinct, over-represented set of detected TFBS classes, herein described as a 'TFBS class combination', implicates the over-representation of one or more underlying TFBS profile-specific combinations, each of these TFBS class combinations is expanded to all possible TFBS profile-specific combinations (for the indicated classes) and then all combinations are analyzed for over-representation. Furthermore, given that CRMs can contain locally dense clusters of TFBSs, the system also provides for the specification of an inter-binding site distance (IBSD) constraint to confine the number of TFBS combinations that are investigated. A scoring scheme, adopted from the Fisher exact test, utilizes two sets of TFBS (class or profile-specific) combination counts to compare the degree of their over-representation: (i) the number found in the promoters of the co-expressed gene set versus (ii) the number found in the promoters of genes in a background set (all genes in the database). TFBS combinations occurring in multiple alternative gene promoter regions are counted only once.

Species-specific databases

In addition to enhancements to the human/mouse oPOSSUM database, we introduce new species databases for studies of over-represented TFBSs in yeast and worms. While the SSA over-representation analysis remains the same for all species, differences in gene structure require that the construction of the underlying databases be particular to each species.

Human/mouse. Ambiguities in ortholog assignments and the definition of TSS positions are major challenges when performing alignments for a large proportion of human and mouse genes. We have expanded the human/mouse database through (i) the discrimination of potential orthologs from predicted paralogs based on upstream sequence similarity (Figure 1), and (ii) the delineation of alternative promoters for human and mouse genes (Figure 2) to address the alignment failure observed in previous database builds.

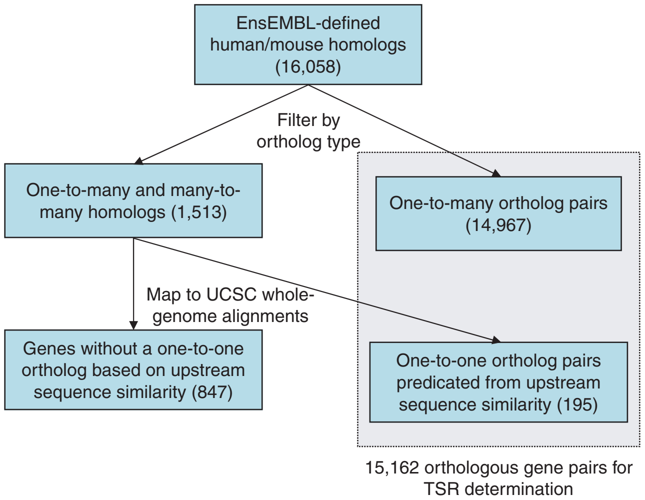

EnsEMBL-defined human/mouse homologs (16,058)
Filter by ortholog type
One-to-many and many-to-many homologs (1,513)
One-to-many ortholog pairs (14,967)
Map to UCSC whole-genome alignments
Genes without a one-to-one ortholog based on upstream sequence similarity (847)
One-to-one ortholog pairs predicated from upstream sequence similarity (195)
15,162 orthologous gene pairs for TSR determination

Figure 1. Determination of one-to-one orthologs for human and mouse genes. An initial set of homologs was downloaded from EnsEMBL v41 (30). All homologs annotated as 'one2one' are extracted. To select the closest putative ortholog pairs from homologs with 'one2many' or 'many2many' relationships, we check for upstream conservation using the whole-genome human–mouse alignments (6). We re-annotate unambiguously aligned homologs as putative one-to-one orthologs, adding 195 gene pairs to our set and bringing the total number of orthologs to 15 162.

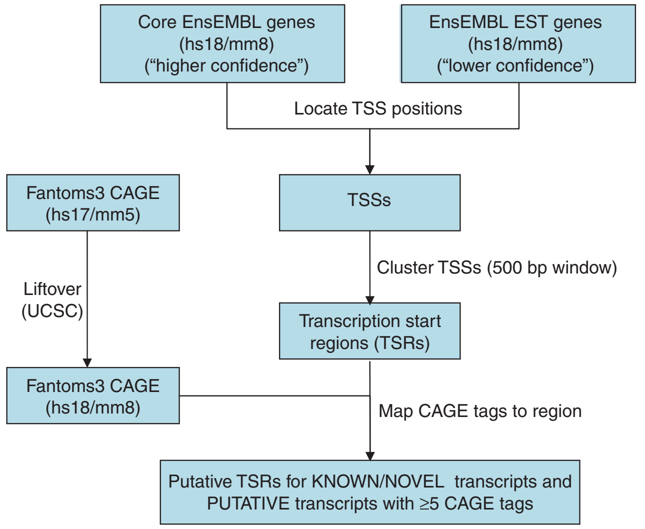

Core EnsEMBL genes (hs18/mm8) ("higher confidence")
EnsEMBL EST genes (hs18/mm8) ("lower confidence")
Locate TSS positions
Fantoms3 CAGE (hs17/mm5)
TSSs
Cluster TSSs (500 bp window)
Liftover (UCSC)
Transcription start regions (TSRs)
Fantoms3 CAGE (hs18/mm8)
Map CAGE tags to region
Putative TSRs for KNOWN/NOVEL transcripts and PUTATIVE transcripts with ≥5 CAGE tags

Figure 2. Identification of transcription start regions (TSRs) using a combination of EnsEMBL annotations and CAGE data. To improve our alignments, we determine putative alternative TSSs for the human and mouse genes. For each gene, the entire repertoire of transcripts from both EnsEMBL core genes and EST genes are retrieved. The TSSs for all transcripts are recorded, followed by a clustering step such that TSSs within 500 bp of one another are merged to form a transcriptional start region (TSR). For each TSR containing a transcript annotated as 'known' or 'novel', we accept the TSR as is. For TSRs based solely on EST gene transcripts, we require a minimum of 5 CAGE tags as evidence for transcription initiation.

While the inclusion of promoter comparisons for candidate ortholog assignment may be controversial, the impact is marginal as <1.3% of gene pairs were derived from this approach. This brings the total number of orthologs to 15 162. Despite improvements in EnsEMBL's ortholog prediction, this is only 1079 more orthologs than were present in our previous database build. Based on the small incremental increases in mapped orthologs, we may be nearing the upper bound for the number of genes in human and mouse that are truly orthologous and detectable by sequence conservation. Detailed descriptions of transcription start region (TSR) determination and the distribution of TSRs for human and mouse genes are available as Supplementary Data.

For each human/mouse orthologous pair, we determine the coordinates of the longest region from the UCSC genome alignments (6) spanning all transcripts plus an additional 10 kb of upstream sequence. The orthologous sequences are retrieved and re-aligned using ORCA, a pairwise global progressive alignment algorithm (1) to optimally align short, conserved blocks within longer global alignments. If possible, TSRs from human and mouse are paired in the alignment. We apply three dynamically computed and progressively more stringent conservation thresholds corresponding to the top 10, 20 and 30% of all 100-bp non-coding windows, each with a minimum percent identity of 70, 65 and 60%, respectively. Of the 15 162 orthologous gene pairs supplied as input to the oPOSSUM pipeline, 15 121 (99.7%) successfully align, and 15 027 (99.1%) have non-exonic conserved regions above 60% nucleotide identity. This is a significant improvement over the previous version of oPOSSUM.

Caenorhabditis elegans/*Caenorhabditis briggsae*. To facilitate transcriptional regulatory analysis of the numerous gene expression studies performed in *C. elegans*, we have implemented a worm version of oPOSSUM. While the database structure and pipeline procedure are very similar to that used for the human/mouse database, there are small modifications that allow for mapping of genes to their operons, as defined by Blumenthal *et al.* (7). In addition, nucleotide identity thresholds for conserved regions were reduced to 60, 55 and 50% for the top 10, 20 and 30% of non-coding windows, respectively, to account for the greater sequence divergence between *C. elegans* and *C. briggsae* compared to human and mouse. The set of orthologs for *C. elegans* and *C. briggsae* is defined by one-to-one InParanoid clusters (8) from WormBase (WS160) (9). After filtering overlapping genes, 10592 orthologous gene pairs (of which, 2140 genes are in operons) remain for alignment. Alignments are performed on the orthologous gene sequences plus 2 kb of upstream sequence (relative to the start codon) for *C. elegans*, and 4 kb of upstream sequence for *C. briggsae*. Annotations are not as mature for *C. briggsae*, and the longer upstream region aids in the alignment of the worm promoter sequences. Alternative promoters have not been considered in this first version; however, should CAGE data or other reliable means for annotating TSSs in worms become available, efforts will certainly be made to include them. Of the 10 592 worm orthologs, 9331 (88%) successfully align.

Yeast. The analysis of yeast promoters is simplified by the more compact nature of the yeast genome. This characteristic diminishes the requirement for comparative methods to reduce the search space and noise inherent in larger genomes. Computational methods using *S. cerevisiae* sequences alone have successfully been used to identify regulatory elements associated with known sets of related genes (10,11). We opted to exclude phylogenetic footprinting for yeast, and instead, select promoter sequences corresponding to the 5′ untranslated region 1000 bp immediately upstream of the start codon of each open reading frame (ORF). Note that for all applications, users have the option to further restrict the search space if they wish. The sequences were downloaded from the *Saccharomyces* Genome Database (12).

TFBS prediction

For the metazoan species, we search for matches to TFBS profiles contained in the JASPAR CORE and JASPAR PhyloFACTS database collections (13,14). Additionally, we include a set of profiles compiled for *C. elegans* TFs from literature review for Worm SSA (Table S2). Binding sites are predicted for the sequences using the TFBS suite of Perl modules for regulatory sequence analysis (15). A predicted binding site for a given TF model is reported if the site occurs in the promoters of both orthologs above a threshold PSSM score of 70% and at equivalent positions in the alignment. Overlapping sites for the same TF are filtered such that only the highest scoring motif is kept. The genomic location, profile score, motif orientation and local sequence conservation level of each TFBS match in orthologous genes are stored in the respective species databases. For *S. cerevisiae*, we compiled a collection of yeast-specific TFBS motifs from both the Yeast Regulatory Sequence Analysis (YRSA) system (16) and the literature (Table S3), and record the genomic location, profile score and motif orientation for each prediction.

Based on the observation that members of the same structural family of TFs often bind to similar sequences, plant and insect matrices are available for inclusion in the analysis. The MADS family of TFs is an excellent example of conservation of binding domains between plants and vertebrates (17,18), and there are numerous examples of conservation of binding domains across vertebrates, flies and worms. Thus, in cases where a profile for the TF of interest is not available in the database, oPOSSUM can still provide insights into the underlying regulation by suggesting a particular TF family that may be involved.

RESULTS AND DISCUSSION

Examples of applications

Each oPOSSUM component was validated on sets of reference genes. The results of all validations are available as Supplementary Data (Tables S4–S13). In the interest of space, selected examples are described for each system.

Human SSA. Wonsey and Follettie (19) performed a microarray analysis of genes that are transcriptionally regulated by FoxM1, a member of the forkhead family of TFs, using BT-20 cells that had been transfected with FoxM1 siRNA. They identified a set of 27 genes that were specifically regulated in cells transfected with FoxM1 siRNA (Table S4). The 27 Affymetrix UG144A identifiers were mapped to 27 EnsEMBL gene identifiers and submitted to Human SSA with default parameters. Of these, 22 genes had a unique mouse ortholog and were used in the oPOSSUM analysis. While a specific profile for FoxM1 is not present in JASPAR CORE, other members of the forkhead family were ranked in the top 10 highest scoring TFBS profiles (Table 1). There is also a known association between HNF4, the highest scoring TFBS profile, and the forkhead TF, FOXO1 in the regulation of gluconeogenic gene expression in hepatocytes (20), which may explain the over-representation of the HNF4 profile.

We previously identified over-represented Fos binding sites in a set of genes induced after transformation by c-Fos in rat fibroblast cells (1,21). We analyzed 160 orthologous genes from the original list of 252 induced genes (Table S7). This is a notable improvement over the previous version where only 98 genes were included in the oPOSSUM analysis. The Fos TFBS profile ranked second in the list of over-represented TFBSs (Table s7). Inspection of the results using the JASPAR PhyloFACTS profiles with default parameters illustrates how inclusion of this new set of profiles provides additional, meaningful information (Table 2). The highest ranked PhyloFACTS motif (TGANTCA) is noted by JASPAR as being most similar to the binding profile for AP-1, and the third highest scoring motif (TGASTMAGC) is most similar to the bZIP TF NF-E2. AP-1 complexes are comprised of Fos and Jun proteins, and the structurally related NF-E2 and AP-1 TFs bind similar sequence motifs (22).

Human CSA. The CSA was validated on a set of mouse skeletal muscle genes comprised of the union of the results of the microarray studies of Moran *et al.* (23) and Tomczak *et al.* (24) (Table S9). To avoid circularity, we removed muscle-specific genes used to generate the JASPAR binding site profiles for Mef-2, Myf, Sp-1, SRF and Tef. These factors occur in clusters in CRMs that contribute to skeletal muscle-specific expression (25). Table 3 lists the top five over-represented pairwise TFBS combinations for this set of genes, along with the JASPAR class each TF profile clustered to, and the Fisher score obtained for each pair. The five most over-represented

Table 1. oPOSSUM results for human FoxM1-regulated gene cluster

JASPAR CORE	TF Class	IC	Target gene hits	Background TFBS rate	Target TFBS rate	Z-score	Fisher score
HNF4	Nuclear	9.62	13	0.0054	0.0085	7.19	2.64*E*−02
Fos	bZIP	10.67	15	0.0111	0.0146	5.72	4.29*E*−01
Pbx	Homeo	14.64	5	0.0019	0.0033	5.57	3.10*E*−01
FOXI1	Forkhead	13.18	16	0.0153	0.0186	4.49	9.05*E*−02
RORA1	Nuclear Receptor	17.42	4	0.0020	0.0029	3.54	5.04*E*−01
TAL1-TCF3	bHLH	14.07	12	0.0052	0.0066	3.30	5.88*E*−02
Staf	Zn-Finger, C2H2	17.54	3	0.0014	0.0021	3.16	3.03*E*−01
Foxa2	Forkhead	12.43	13	0.0152	0.0174	3.04	4.83*E*−01
Foxd3	Forkhead	12.94	13	0.0172	0.0194	2.93	5.27*E*−01
TEAD	TEA	15.67	6	0.0028	0.0037	2.85	4.70*E*−01

Table 2. oPOSSUM results for c-Fos-regulated gene cluster

JASPAR PhyloFACTS	Similar to	IC	Target gene hits	Background TFBS rate	Target TFBS rate	Z-score	Fisher score
TGANTCA	AP-1	12.06	46	0.0011	0.0023	18.05	1.40*E*−04
GGGYGTGNY	–	14.18	82	0.0059	0.0083	15.64	4.98*E*−02
TGASTMAGC	NF-E2	16.60	43	0.0013	0.0024	15.64	1.19*E*−03
GGARNTKYCCA	–	17.13	44	0.0016	0.0026	12.54	1.11*E*−03
GGGAGGRR	MAZ	14.00	111	0.0171	0.0202	11.98	3.16*E*−01

Table 3. oPOSSUM results for skeletal muscle genes identified by Moran *et al.* and Tomczak *et al*

TF name (Class ID)	TF class name	TF name (Class ID)	TF class name	Score
MEF2A (class 4)	MADS	Myf (class 22)	bHLH	1.65*E*−06
MEF2A (class 4)	MADS	ZNF42_1–4 (class 25)	Zn-finger, C2H2	4.24*E*−06
Myf (class 22)	bHLH	SRF (class 1)	MADS	2.52*E*−05
SP1 (class 31)	Zn-finger, C2H2	SRF (class 1)	MADS	2.68*E*−05
Agamous (class 1)	MADS	MEF2A (class 4)	MADS	7.63*E*−05

pairs of TFBS profiles include combinations of Mef-2, SRF and Sp-1.

The inclusion of alternative promoters provides notable improvements in the Human SSA and Human CSA analyses. The same datasets were used to validate our previous and current human oPOSSUM analyses systems. Demarcation of additional promoter boundaries increases the signal in the discovery process, improving the signal for both over-represented single TFBSs and combinations of TFBSs in the gene sets analyzed.

Worm SSA. Worm SSA was tested on a set of well-characterized nematode muscle genes (Table S10) (26). Analysis of 1000 bp of upstream sequence, using the top 10% of conserved regions (minimum of 60% sequence identity), a matrix match threshold of 80% and the worm profiles, identified the putative muscle1 motif with a Z-score of 20.6 and a Fisher score <0.01 (Table 4). This is, however, somewhat circular, given that 19 of the 41 input genes were used to generate the putative muscle-specific worm profiles. Analysis using the JASPAR CORE profiles ranked SP1 and Su(H) within the top 10 scoring profiles (Table S10B). Studies in *Xenopus* and *Drosophila* provide evidence that MyoD triggers Notch signaling through Su(H) for muscle determination (27,28). Although SP1 has been implicated in muscle CRMs, it is a general TF involved in the expression of many different genes and binds to GC-rich motifs.

Yeast SSA. The yeast CLB2 gene cluster is comprised of 32 genes whose pattern of expression peaks at late G2/early M phase of the cell cycle (Table S11). Transcription of these genes is regulated by two TFs: FKH, which is a component of the TF SFF, and MCM1, a member of the early cell cycle box (ECB) binding complex. Analysis of 500 bp of upstream sequence using a matrix match threshold of 85% ranked ECB, MCM1 and FKH1 in the top five scoring TFBS profiles (Table 5), which is consistent with the literature (29).

Web server

The four oPOSSUM systems, Human SSA, Human CSA, Worm SSA and Yeast SSA, have been integrated into a user-friendly website at: http://www.cisreg.ca/oPOSSUM/. We recommend that users of the system begin with the SSA to quickly identify TFBSs that may be relevant to their input data sets. For sets of human and mouse genes, this can be followed with the CSA, which takes longer to process, but which can provide insights into TFBSs that may be acting in concert to regulate the set of genes.

The web implementation allows for analysis in default and custom modes. Default mode processing is faster as TFBS counts have been pre-calculated and stored for pre-defined conservation levels, matrix match thresholds and promoter lengths. In either mode, the user is required to select a species and to enter a list of gene identifiers (EnsEMBL, RefSeq, HGNC and Entrez Gene are supported for human). A number of options are available to specify the TFBS profile set to be used in the analysis. Finally, the conservation level, matrix match threshold and the promoter length can be varied. In the custom mode, users may define their own background set, which provides users with more control, but results in more variable processing speeds depending on the size of the background set and the parameters selected.

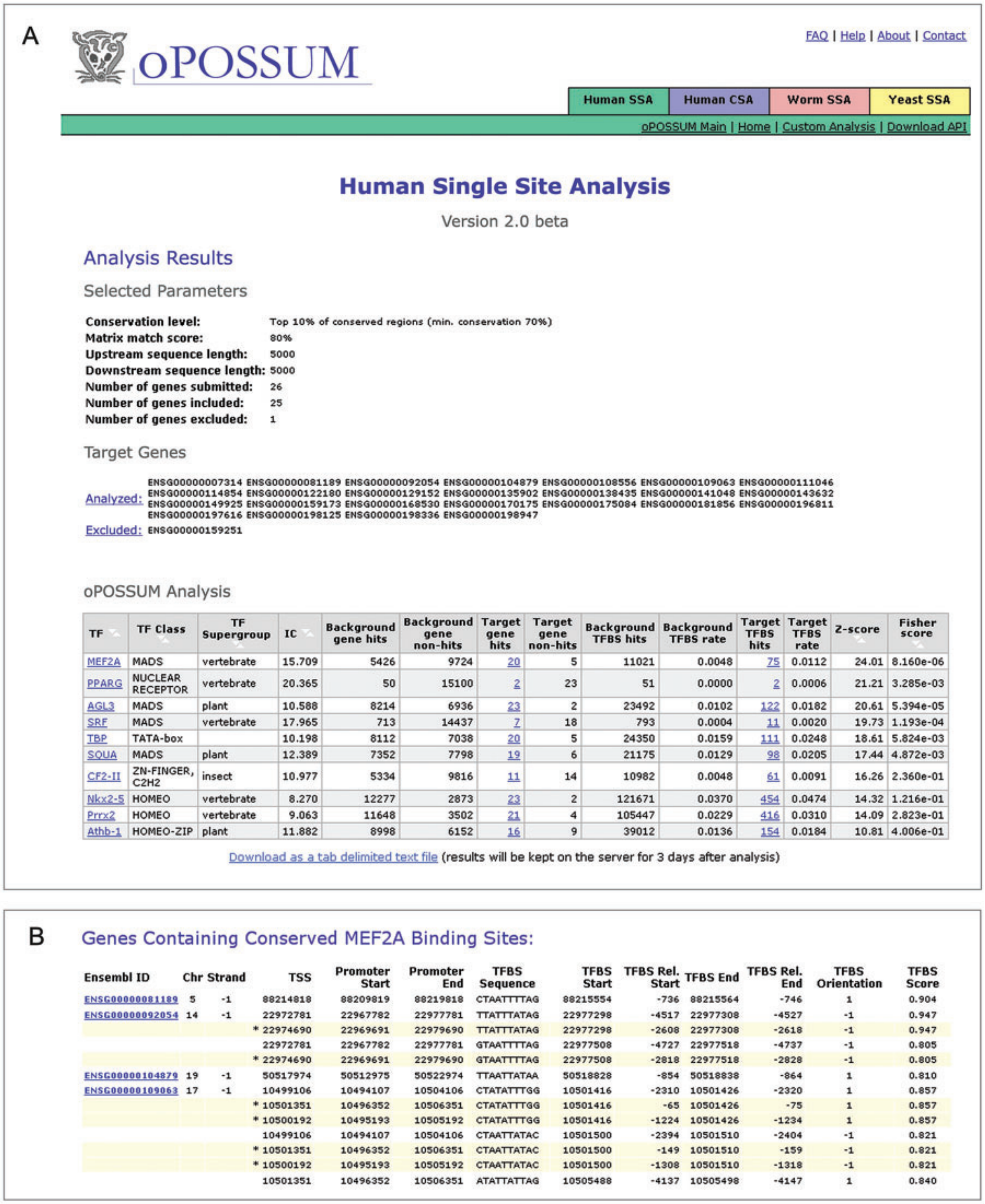
Upon submission, oPOSSUM SSA generates a summary of the input parameters, and produces a single table that ranks the over-represented TFBSs by descending Z-score. The table may be sorted by TF name, TF class, supergroup, information content (IC), Z-score and Fisher score (Figure 3A). Pop-up windows linked to each TFBS foreground count display the genes in which the putative site is located, the promoter region(s) for each gene, as well as the TFBS's co-ordinates and score (Figure 3B). TFBSs that occur in overlapping promoter regions are marked by an asterisk and highlighted in yellow. The TF names are linked to the JASPAR database for easy access to information regarding the binding

Table 4. oPOSSUM results for worm skeletal muscle genes using worm profiles

Worm	Status	IC	Target gene hits	Background TFBS rate	Target TFBS rate	Z-score	Fisher score
Muscle1	Putative	11.34	6	0.0025	0.0156	20.56	4.24*E*–04
Muscle2	Putative	11.97	4	0.0022	0.0089	11.19	1.39*E*–02
LIN-14	Putative	9.13	9	0.0143	0.0280	9.12	1.17*E*–01
Muscle3	Putative	16.67	4	0.0029	0.0064	5.02	6.96*E*–02

Table 5. oPOSSUM results for the yeast CLB2 gene cluster

YEAST	TF Class	IC	Target gene hits	Background TFBS rate	Target TFBS rate	Z-score	Fisher score
ECB	Unclassified	16.65	13	0.0019	0.0131	32.87	8.68*E*–09
MCM1	MADS	9.15	10	0.0073	0.0165	13.71	1.08*E*–02
FKH1	Forkhead	13.28	30	0.0305	0.0473	12.26	4.05*E*–02
CCA	Unclassified	16.93	3	0.0017	0.0040	7.08	2.02*E*–01
LYS14	C6_Zinc finger	17.02	6	0.0030	0.0053	5.20	9.41*E*–02


A
oPOSSUM
FAQ | Help | About | Contact
Human SSA
Human CSA
Worm SSA
Yeast SSA
oPOSSUM Main | Home | Custom Analysis | Download API
B

Human Single Site Analysis

Version 2.0 beta

Analysis Results

Selected Parameters

Conservation level:	Top 10% of conserved regions (min. conservation 70%)
Matrix match score:	80%
Upstream sequence length:	5000
Downstream sequence length:	5000
Number of genes submitted:	26
Number of genes included:	25
Number of genes excluded:	1

Target Genes

Analyzed: ENSG00000007314 ENSG00000081189 ENSG00000092054 ENSG00000104879 ENSG00000108556 ENSG00000109063 ENSG00000111046 ENSG00000114854 ENSG00000122180 ENSG00000129152 ENSG00000135902 ENSG00000138435 ENSG00000141048 ENSG00000143632 ENSG00000149925 ENSG00000159173 ENSG00000168530 ENSG00000170175 ENSG00000175084 ENSG00000181856 ENSG00000196811 ENSG00000197616 ENSG00000198125 ENSG00000198336 ENSG00000198947

Excluded: ENSG00000159251

oPOSSUM Analysis

TF	TF Class	TF Supergroup	IC	Background gene hits	Background gene non-hits	Target gene hits	Target gene non-hits	Background TFBS hits	Background TFBS rate	Target TFBS hits	Target TFBS rate	Z-score	Fisher score
MEF2A	MADS	vertebrate	15.709	5426	9724	20	5	11021	0.0048	75	0.0112	24.01	8.160e-06
PPARG	NUCLEAR RECEPTOR	vertebrate	20.365	50	15100	2	23	51	0.0000	2	0.0006	21.21	3.285e-03
AGL3	MADS	plant	10.588	8214	6936	23	2	23492	0.0102	122	0.0182	20.61	5.394e-05
SRF	MADS	vertebrate	17.965	713	14437	7	18	793	0.0004	11	0.0020	19.73	1.193e-04
TBP	TATA-box		10.198	8112	7038	20	5	24350	0.0159	111	0.0248	18.61	5.824e-03
SQUA	MADS	plant	12.389	7352	7798	19	6	21175	0.0129	98	0.0205	17.44	4.872e-03
CF2-II	ZN-FINGER, C2H2	insect	10.977	5334	9816	11	14	10982	0.0048	61	0.0091	16.26	2.360e-01
Nkx2-5	HOMEO	vertebrate	8.270	12277	2873	23	2	121671	0.0370	454	0.0474	14.32	1.216e-01
Prrx2	HOMEO	vertebrate	9.063	11648	3502	21	4	105447	0.0229	416	0.0310	14.09	2.823e-01
Athb-1	HOMEO-ZIP	plant	11.882	8998	6152	16	9	39012	0.0136	154	0.0184	10.81	4.006e-01

Download as a tab delimited text file (results will be kept on the server for 3 days after analysis)

Genes Containing Conserved MEF2A Binding Sites:

Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
ENSG00000081189	5	-1	88214818	88209819	88219818	CTAATTTTAG	88215554	-736	88215564	-746	1	0.904
ENSG00000092054	14	-1	22972781	22967782	22977781	TTATTTATAG	22977298	-4517	22977308	-4527	-1	0.947
			* 22974690	22969691	22979690	TTATTTATAG	22977298	-2608	22977308	-2618	-1	0.947
			22972781	22967782	22977781	GTAATTTTAG	22977508	-4727	22977518	-4737	-1	0.805
			* 22974690	22969691	22979690	GTAATTTTAG	22977508	-2818	22977518	-2828	-1	0.805
ENSG00000104879	19	-1	50517974	50512975	50522974	TTAATTATAA	50518828	-854	50518838	-864	1	0.810
ENSG00000109063	17	-1	10499106	10494107	10504106	CTATATTTGG	10501416	-2310	10501426	-2320	1	0.857
			* 10501351	10496352	10506351	CTATATTTGG	10501416	-65	10501426	-75	1	0.857
			* 10500192	10495193	10505192	CTATATTTGG	10501416	-1224	10501426	-1234	1	0.857
			10499106	10494107	10504106	CTAATTATAC	10501500	-2394	10501510	-2404	-1	0.821
			* 10501351	10496352	10506351	CTAATTATAC	10501500	-149	10501510	-159	-1	0.821
			* 10500192	10495193	10505192	CTAATTATAC	10501500	-1308	10501510	-1318	-1	0.821
			10501351	10496352	10506351	ATATTATTAG	10505488	-4137	10505498	-4147	1	0.840

Figure 3. (**A**) A screenshot of the output of the oPOSSUM Human SSA analysis, with TFBS profiles ranked by Z-score. The arrows allow the user to sort and re-order the results by Fisher score, TF name, TF class, TF supergroup or TF profile information content (IC). Each TF name links to a pop-up window displaying the TFBS profile information. (**B**) Pop-up window displaying genes that contain a particular TFBS (in this case, MEF2A; partial list shown), as well as the promoter coordinates associated with each gene, and the motif locations and scores. Sites in overlapping alternative promoters are highlighted for emphasis. Such sites are only counted once in the statistical analysis.

site profiles. The output for oPOSSUM CSA is similar, providing (i) a ranked list of over-represented TFBS class combinations, and (ii) a list of the most significant TFBS combinations (found in the set of expanded top-ranked class combinations).

Based on the underlying assumption of the statistics employed that DNA sequences are randomly generated, there is little reason to accept the calculated scores as accurate reflections of significance. Instead, as suggested in the original published description of the oPOSSUM algorithm, we recommend that the scores are best used as rankings rather than significance measures. For this reason, a multiple testing correction is not applied as it does not alter the relative ranks. Empirically, we determined that TFBS profiles with Z-scores $\geqslant 10$ and Fisher scores $\leqslant 0.01$ facilitate the identification of relevant TFBSs for our sets of reference genes (1). However, these are relatively stringent thresholds, and we encourage users to examine the scores of top-ranked TFBS profiles before applying any cutoffs.

We provide a consistent display for all four systems. However, there are slight differences between the systems, such as different parameters for selection on the input pages which are relevant for each species database and system. Also, due to the longer processing times required to compute combinations of TFBSs, Human CSA queues the analysis request on the server and emails the completed results to the user.

CONCLUSIONS

The oPOSSUM system is under continued development. Efforts are underway to allow users to submit custom TF profiles to be included in the analysis. An improved search method for nuclear hormone receptors, which typically contain two half sites separated by a variable length spacer, has been developed and will be included in a future release. We will continue to add TFBS profiles as they become available, with an emphasis on expanding the repertoire of worm TFBS profiles. We believe the oPOSSUM web server is and will continue to be a useful resource for inference of mechanisms of co-regulation based on observed co-expression.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Project funding provided through the Canadian Institutes of Health Research (CIHR) and Merck. A.T.K. and S.J.H.S. are supported by the National Science and Engineering Research Council of Canada (NSERC); D.L.F., S.J.H.S. and W.W.W. are supported by the Michael Smith Foundation for Health Research (MSFHR); and D.L.F. and W.W.W. are supported by the CIHR. The authors would like to thank Jonathan Lim and Miroslav Hatas for systems support. Funding to pay the Open Access publication charges were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Ho Sui,S.J., Mortimer,J.R., Arenillas,D.J., Brumm,J., Walsh,C.J., Kennedy,B.P. and Wasserman,W.W. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.*, **33**, 3154–3164.
2. Huang,S.S., Fulton,D.L., Arenillas,D.J., Perco,P., Ho Sui,S.J., Mortimer,J.R. and Wasserman,W.W. (2006) Identification of over-represented combinations of transcription factor binding sites in sets of co-expressed genes. In *Advances in Bioinformatics and Computational Biology*. Imperial College Press, London, UK, Vol. 3, pp 247–256.
3. Urnov,F.D. and Rebar,E.J. (2002) Designed transcription factors as tools for therapeutics and functional genomics. *Biochem. Pharmacol.*, **64**, 919–923.
4. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, REVIEWS001.
5. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
6. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
7. Blumenthal,T., Evans,D., Link,C.D., Guffanti,A., Lawson,D., Thierry-Mieg,J., Thierry-Mieg,D., Chiu,W.L., Duke,K. *et al.* (2002) A global analysis of Caenorhabditis elegans operons. *Nature*, **417**, 851–854.
8. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
9. Stein,L., Sternberg,P., Durbin,R., Thierry-Mieg,J. and Spieth,J. (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res.*, **29**, 82–86.
10. Zhang,M.Q. (1999) Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.*, **23**, 233–250.
11. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.
12. Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T. *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.*, **26**, 73–79.
13. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
14. Vlieghe,D., Sandelin,A., De Bleser,P.J., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.
15. Lenhard,B. and Wasserman,W.W. (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
16. Sandelin,A., Hoglund,A., Lenhard,B. and Wasserman,W.W. (2003) Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct. Integr. Genomics*, **3**, 125–134.
17. Alvarez-Buylla,E.R., Pelaz,S., Liljegren,S.J., Gold,S.E., Burgeff,C., Ditta,G.S., Ribas,D.P., Martinez-Castilla,L. and Yanofsky,M.F. (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc. Natl Acad. Sci. USA*, **97**, 5328–5333.
18. Shore,P. and Sharrocks,A.D. (1995) The MADS-box family of transcription factors. *Eur. J. Biochem.*, **229**, 1–13.
19. Wonsey,D.R. and Follettie,M.T. (2005) Loss of the forkhead transcription factor FoxM1 causes centrosome amplification and mitotic catastrophe. *Cancer Res.*, **65**, 5181–5189.
20. Lin,J., Tarr,P.T., Yang,R., Rhee,J., Puigserver,P., Newgard,C.B. and Spiegelman,B.M. (2003) PGC-1beta in the regulation of hepatic glucose and energy metabolism. *J. Biol. Chem.*, **278**, 30843–30848.
21. Ordway,J.M., Williams,K. and Curran,T. (2004) Transcription repression in oncogenic transformation: common targets of epigenetic repression in cells transformed by Fos, Ras or Dnmt1. *Oncogene*, **23**, 3737–3748.
22. Daftari,P., Gavva,N.R. and Shen,C.K. (1999) Distinction between AP1 and NF-E2 factor-binding at specific chromatin regions in mammalian cells. *Oncogene*, **18**, 5482–5486.
23. Moran,J.L., Li,Y., Hill,A.A., Mounts,W.M. and Miller,C.P. (2002) Gene expression changes during mouse skeletal myoblast

differentiation revealed by transcriptional profiling. *Physiol. Genomics*, **10**, 103–111.

24. Tomczak,K.K., Marinescu,V.D., Ramoni,M.F., Sanoudou,D., Montanaro,F., Han,M., Kunkel,L.M., Kohane,I.S. and Beggs,A.H. (2004) Expression profiling and identification of novel genes involved in myogenic differentiation. *FASEB J.*, **18**, 403–405.
25. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
26. GuhaThakurta,D., Schriefer,L.A., Waterston,R.H. and Stormo,G.D. (2004) Novel transcription regulatory elements in Caenorhabditis elegans muscle genes. *Genome Res.*, **14**, 2457–2468.
27. Rusconi,J.C. and Corbin,V. (1998) Evidence for a novel Notch pathway required for muscle precursor selection in Drosophila. *Mech. Dev.*, **79**, 39–50.
28. Wittenberger,T., Steinbach,O.C., Authaler,A., Kopan,R. and Rupp,R.A. (1999) MyoD stimulates delta-1 transcription and triggers notch signaling in the Xenopus gastrula. *EMBO J.*, **18**, 1915–1922.
29. Kreiman,G. (2004) Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res.*, **32**, 2889–2900.
30. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.