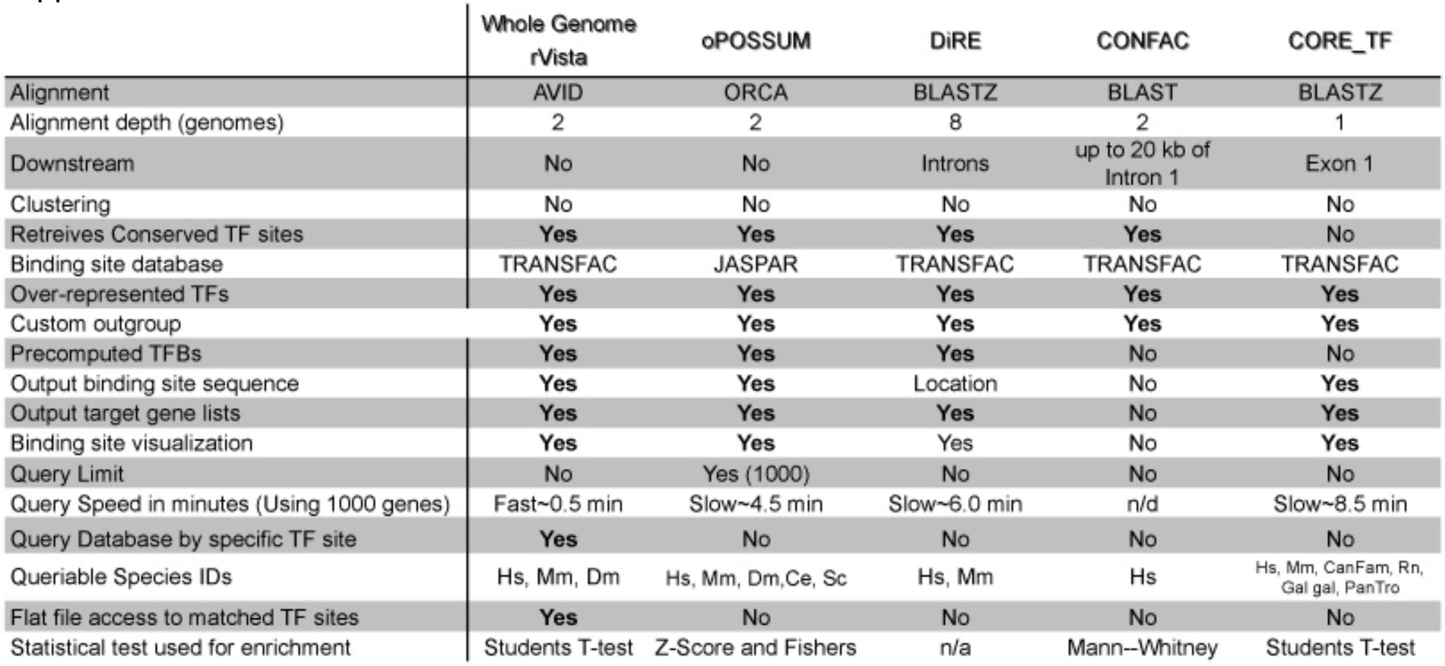
Supplemental Methods:

Genome assemblies for alignment are downloaded from UCSC Browser, along with gene annotations. A whole-genome alignment is performed using VISTA alignment pipeline (Dubchak, et al., 2009; Frazer, et al., 2004). For each pair of orthologous genes, 5,000 base pairs of upstream sequence are considered the promoter region. Each promoter region is searched for potential transcription factor binding sites using the MATCH tool and the TRANSFAC database (Release 2011.4) using default parameters. These binding sites are then filtered by conservation using rVISTA: Only those sites which are aligned between two selected species (in this case here, between mouse and human or between two *Drosophila* species) and are highly conserved within a 21 base pair sliding window around each site are kept (Loots and Ovcharenko, 2004; Loots, et al., 2002). The resulting binding sites are stored in a database, along with their locations and relative position within the promoter region. When a user queries the database with a list of reference species genes, binding sites upstream of the submitted genes are retrieved. For each transcription factor represented among these binding sites, a test using the binomial distribution is performed to determine whether the number of the binding sites for that factor in the user submission exceeds the expected number for a subset of that size, if sites were distributed randomly throughout all promoter regions in the genome. If the user has provided a background set, only those genes' promoter regions are used in the statistical test, rather than all promoter regions in the genome.

Supplemental Data:

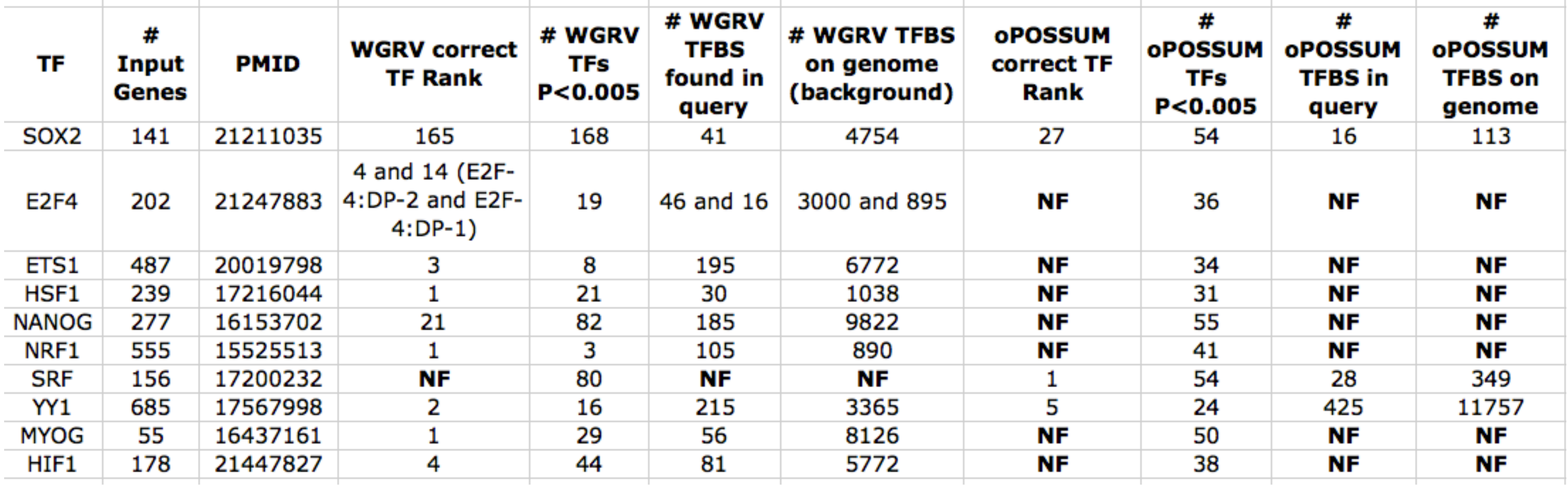
Supplemental Table 1



	Whole Genome rVista	oPOSSUM	DiRE	CONFAC	CORE_TF
Alignment	AVID	ORCA	BLASTZ	BLAST	BLASTZ
Alignment depth (genomes)	2	2	8	2	1
Downstream	No	No	Introns	up to 20 kb of Intron 1	Exon 1
Clustering	No	No	No	No	No
Retreives Conserved TF sites	**Yes**	**Yes**	**Yes**	**Yes**	No
Binding site database	TRANSFAC	JASPAR	TRANSFAC	TRANSFAC	TRANSFAC
Over-represented TFs	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Custom outgroup	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Precomputed TFBs	**Yes**	**Yes**	**Yes**	No	No
Output binding site sequence	**Yes**	**Yes**	Location	No	**Yes**
Output target gene lists	**Yes**	**Yes**	**Yes**	No	**Yes**
Binding site visualization	**Yes**	**Yes**	Yes	No	**Yes**
Query Limit	No	Yes (1000)	No	No	No
Query Speed in minutes (Using 1000 genes)	Fast~0.5 min	Slow~4.5 min	Slow~6.0 min	n/d	Slow~8.5 min
Query Database by specific TF site	**Yes**	No	No	No	No
Queriable Species IDs	Hs, Mm, Dm	Hs, Mm, Dm,Ce, Sc	Hs, Mm	Hs	Hs, Mm, CanFam, Rn, Gal gal, PanTro
Flat file access to matched TF sites	**Yes**	No	No	No	No
Statistical test used for enrichment	Students T-test	Z-Score and Fishers	n/a	Mann--Whitney	Students T-test

Supplemental Table 1-Anaylsis of key features of publically available bioinformatic tools that employ evolutionary conservation of coregulated gene promoters. n/d-not determined; n/a-not available.

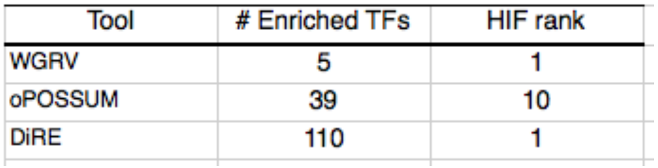
Supplemental Table 2



TF	# Input Genes	PMID	WGRV correct TF Rank	# WGRV TFs P<0.005	# WGRV TFBS found in query	# WGRV TFBS on genome (background)	oPOSSUM correct TF Rank	# oPOSSUM TFs P<0.005	# oPOSSUM TFBS in query	# oPOSSUM TFBS on genome
SOX2	141	21211035	165	168	41	4754	27	54	16	113
E2F4	202	21247883	4 and 14 (E2F-4:DP-2 and E2F-4:DP-1)	19	46 and 16	3000 and 895	**NF**	36	**NF**	**NF**
ETS1	487	20019798	3	8	195	6772	**NF**	34	**NF**	**NF**
HSF1	239	17216044	1	21	30	1038	**NF**	31	**NF**	**NF**
NANOG	277	16153702	21	82	185	9822	**NF**	55	**NF**	**NF**
NRF1	555	15525513	1	3	105	890	**NF**	41	**NF**	**NF**
SRF	156	17200232	**NF**	80	**NF**	**NF**	1	54	28	349
YY1	685	17567998	2	16	215	3365	5	24	425	11757
MYOG	55	16437161	1	29	56	8126	**NF**	50	**NF**	**NF**
HIF1	178	21447827	4	44	81	5772	**NF**	38	**NF**	**NF**

Supplemental Table 2-Evaluation of WGRV and oPOSSUM to predict TF when querying previously validated ChIP-Seq targets. Each database was queried with lists of previously published TF target genes established by chromatin immunoprecipitation studies that were randomly selected and downloaded from the PAZAR or Amadeus resources (Linhart, et al., 2008; Portales-Casamar, et al., 2007). Input lists and raw WGRV and oPOSSUM results are supplied as Supplemental Data File 1. NF-Not found. PMID-PubMed identification.

Supplemental Table 3

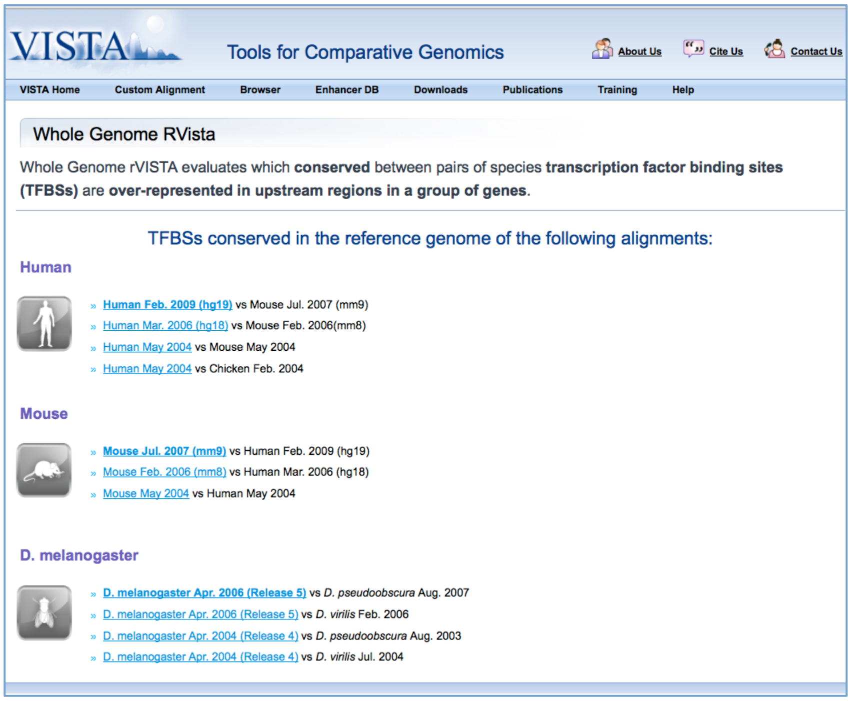


Tool	# Enriched TFs	HIF rank
WGRV	5	1
oPOSSUM	39	10
DiRE	110	1

Supplemental Table 3- Ranking of HIF and number of enriched TF binding sites (P<0.005) in 162 annotated probesets significantly upregulated (Fold>2, P<0.05) by hypoxia using WGRV, oPOSSUM and DiRE set to 2000 kb proximal promoter regions and P<0.005. Input lists and raw WGRV, oPOSSUM and DiRE results are supplied as Supplemental Data File 2.

Supplemental Figure 1

A



Whole Genome RVista

Whole Genome rVISTA evaluates which **conserved** between pairs of species **transcription factor binding sites (TFBSs)** are **over-represented in upstream regions in a group of genes**.

TFBSs conserved in the reference genome of the following alignments:

Human

- **Human Feb. 2009 (hg19)** vs Mouse Jul. 2007 (mm9)
- Human Mar. 2006 (hg18) vs Mouse Feb. 2006(mm8)
- Human May 2004 vs Mouse May 2004
- Human May 2004 vs Chicken Feb. 2004

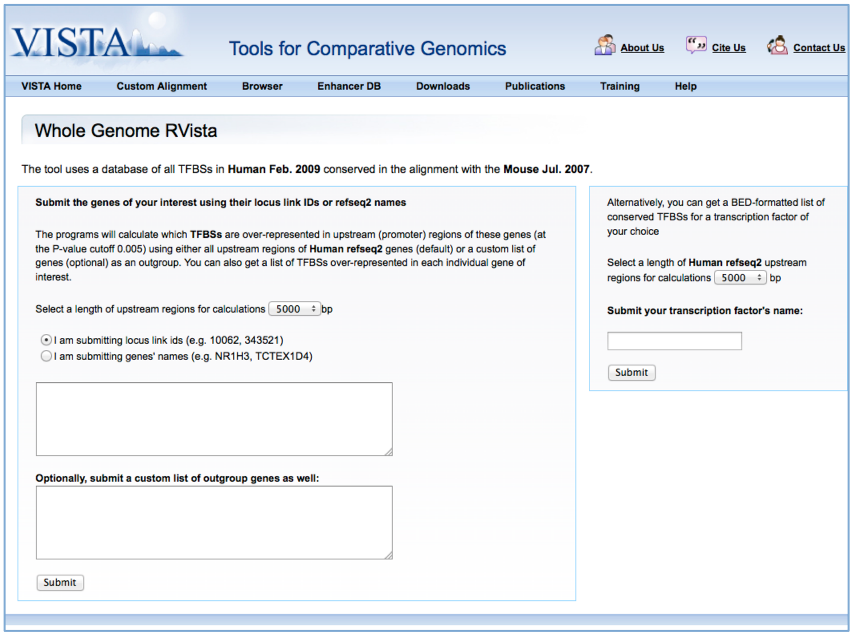
Mouse

- **Mouse Jul. 2007 (mm9)** vs Human Feb. 2009 (hg19)
- Mouse Feb. 2006 (mm8) vs Human Mar. 2006 (hg18)
- Mouse May 2004 vs Human May 2004

D. melanogaster

- **D. melanogaster Apr. 2006 (Release 5)** vs *D. pseudoobscura* Aug. 2007
- D. melanogaster Apr. 2006 (Release 5) vs *D. virilis* Feb. 2006
- D. melanogaster Apr. 2004 (Release 4) vs *D. pseudoobscura* Aug. 2003
- D. melanogaster Apr. 2004 (Release 4) vs *D. virilis* Jul. 2004

B



Whole Genome RVista

The tool uses a database of all TFBSs in **Human Feb. 2009** conserved in the alignment with the **Mouse Jul. 2007**.

Submit the genes of your interest using their locus link IDs or refseq2 names

The programs will calculate which **TFBSs** are over-represented in upstream (promoter) regions of these genes (at the P-value cutoff 0.005) using either all upstream regions of **Human refseq2** genes (default) or a custom list of genes (optional) as an outgroup. You can also get a list of TFBSs over-represented in each individual gene of interest.

Select a length of upstream regions for calculations 5000 bp

● I am submitting locus link ids (e.g. 10062, 343521)
○ I am submitting genes' names (e.g. NR1H3, TCTEX1D4)

Optionally, submit a custom list of outgroup genes as well:

Submit

Alternatively, you can get a BED-formatted list of conserved TFBSs for a transcription factor of your choice

Select a length of **Human refseq2** upstream regions for calculations 5000 bp

Submit your transcription factor's name:

Submit

Supplemental Figure 1-Screenshots of WGRV submission page http://genome.lbl.gov/cgi-bin/WGRVistaInputCommon.pl indicating A) comparative species databases B) page linked to submission page to enter gene IDs as either locus link or gene symbols. Page also enables user to select upstream region to be queried.

Supplemental References:

Dubchak, I., Poliakov, A., Kislyuk, A. and Brudno, M. (2009) Multiple whole-genome alignments without a reference organism, *Genome Res*, **19**, 682-689.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. and Dubchak, I. (2004) VISTA: computational tools for comparative genomics, *Nucleic Acids Res*, **32**, W273-279.

Linhart, C., Halperin, Y. and Shamir, R. (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets, *Genome Res*, **18**, 1180-1189.

Loots, G.G. and Ovcharenko, I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites, *Nucleic Acids Res*, **32**, W217-221.

Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I. and Rubin, E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites, *Genome Res*, **12**, 832-839.

Portales-Casamar, E., Kirov, S., Lim, J., Lithwick, S., Swanson, M.I., Ticoll, A., Snoddy, J. and Wasserman, W.W. (2007) PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation, *Genome Biol*, **8**, R207.