Nils Blüthgen, Szymon M. Kiełbasa, and Hanspeter Herzel

Inferring Combinatorial Regulation of Transcription *in silico.*

Supplement

Table S1: Significantly associated biological processes (FDR<0.05) for the 14 sub-sets of the muscle related transcription factors that have at least one significantly associated process. (#genes: number of genes that have both a cluster of binding sites and are annotated with the GO-term; p-value: single test p-value obtained by the one sided Fisher's exact test; FDR: false discovery rate)

GO id	GO-term	# genes	p-value	FDR
	Set: SRF, TEF			
7517	muscle development	9	$3.26\ 10^{-6}$	0.0010
	Set: Sp-1, SRF, TEF			
7517	muscle development	16	$8.2\ 10^{-5}$	0.012
45449	regulation of transcription	129	$8.8\ 10^{-5}$	0.012
6351	transcription, DNA-dependent	131	$9.0\ 10^{-5}$	0.012
6355	regulation of transcription, DNA-d.	128	0.00011	0.012
6350	transcription	132	0.00015	0.012
7275	development	93	0.00015	0.012
9653	morphogenesis	60	0.00078	0.050
	Set: Myf, SRF			
6631	fatty acid metabolism	13	$7.52\ 10^{-5}$	0.034
	Set: Myf, SRF, TEF			
7517	muscle development	14	$5.5\ 10^{-6}$	0.0019
	Set: Myf,Sp-1, SRF			
6351	transcription, DNA-dependent	161	$3.7\ 10^{-5}$	0.011
45449	regulation of transcription	157	$6.7\ 10^{-5}$	0.011
6355	regulation of transcription, DNA-d.	156	$7.6\ 10^{-5}$	0.011
6350	transcription	161	0.00011	0.011
	Set: Mef-2, TEF			
30183	B-cell differentiation	5	$2.7\ 10^{-5}$	0.0096
30098	lymphocyte differentiation	6	$6.64\ 10^{-5}$	0.014
42113	B-cell activation	5	0.00035	0.047
30097	hemopoiesis	7	0.00045	0.047

Table S1 continued

GO id	GO-term	# genes	p-value	FDR
	Set: Mef-2,SRF			
7517	muscle development	16	$3.9\ 10^{-8}$	$1.1\ 10^{-5}$
30183	B-cell differentiation	5	$1.3\ 10^{-5}$	0.0023
30098	lymphocyte differentiation	6	$2.7\ 10^{-5}$	0.0038
42113	B-cell activation	5	0.00017	0.016
30097	hemopoiesis	7	0.00017	0.016
9887	organogenesis	36	0.00023	0.016
6936	muscle contraction	11	0.00048	0.026
9653	morphogenesis	38	0.00063	0.032
	Set: Mef-2,SRF,TEF			
7517	muscle development	15	$8.2\ 10^{-8}$	$2.5\ 10^{-5}$
30183	B-cell differentiation	5	$8.4\ 10^{-6}$	0.0015
30098	lymphocyte differentiation	6	$1.7\ 10^{-5}$	0.0026
9887	organogenesis	35	$9.1\ 10^{-5}$	0.0073
30097	hemopoiesis	7	0.00010	0.0073
42113	B-cell activation	5	0.00011	0.0073
9653	morphogenesis	37	0.00024	0.013
45191	regulation of isotype switching	2	0.00057	0.048
	Set: Mef-2,Sp-1,SRF			
30183	B-cell differentiation	6	$4.3\ 10^{-5}$	0.015
7517	muscle development	18	$8.5\ 10^{-5}$	0.015
42113	B-cell activation	7	0.00012	0.015
30098	lymphocyte differentiation	7	0.00030	0.038
	Set: Mef-2,Myf			
6941	striated muscle contraction	9	$3.1\ 10^{-5}$	0.013
6936	muscle contraction	19	0.00018	0.036
45190	isotype switching	3	0.00021	0.036
30097	hemopoiesis	10	0.00031	0.036
30098	lymphocyte differentiation	7	0.00037	0.036

Table S1 continued

GO id	GO-term	# genes	p-value	FDR
	Set: Mef-2,Myf,TEF			
6941	striated muscle contraction	9	$7.6\ 10^{-6}$	0.0025
7517	muscle development	18	$1.3\ 10^{-5}$	0.0025
6936	muscle contraction	18	$5.9\ 10^{-5}$	0.0074
45190	isotype switching	3	0.00013	0.018
30183	B-cell differentiation	5	0.00029	0.028
30097	hemopoiesis	9	0.00039	0.028
42113	B-cell activation	6	0.00044	0.028
	Set: Mef-2,Myf,SRF			
7517	muscle development	18	$2.7\ 10^{-6}$	0.00079
45190	isotype switching	3	$8.9\ 10^{-5}$	0.015
6936	muscle contraction	16	0.00016	0.015
30097	hemopoiesis	9	0.00017	0.015
30183	B-cell differentiation	5	0.00017	0.015
6941	striated muscle contraction	7	0.00023	0.015
42113	B-cell activation	6	0.00024	0.015
30098	lymphocyte differentiation	6	0.00052	0.032
	Set: Mef-2,Myf,SRF,TEF			
7517	muscle development	15	$2.5\ 10^{-5}$	0.0089
	Set: Mef-2,Myf,Sp-1,SRF,TEF			
7517	muscle development	20	$8.0\ 10^{-6}$	0.0027
6936	muscle contraction	20	$4.2\ 10^{-5}$	0.0080
42113	B-cell activation	7	0.00012	0.017

Table S2: Significantly associated biological processes (FDR<0.05) for the transcription factors CREB, CEBP, p50/p65, Sp-1, ETS and AP1. (#genes: number of genes that have both a cluster of binding sites and are annotated with the GO-term; p-value: single test p-value obtained by the one sided Fisher's exact test; FDR: false discovery rate)

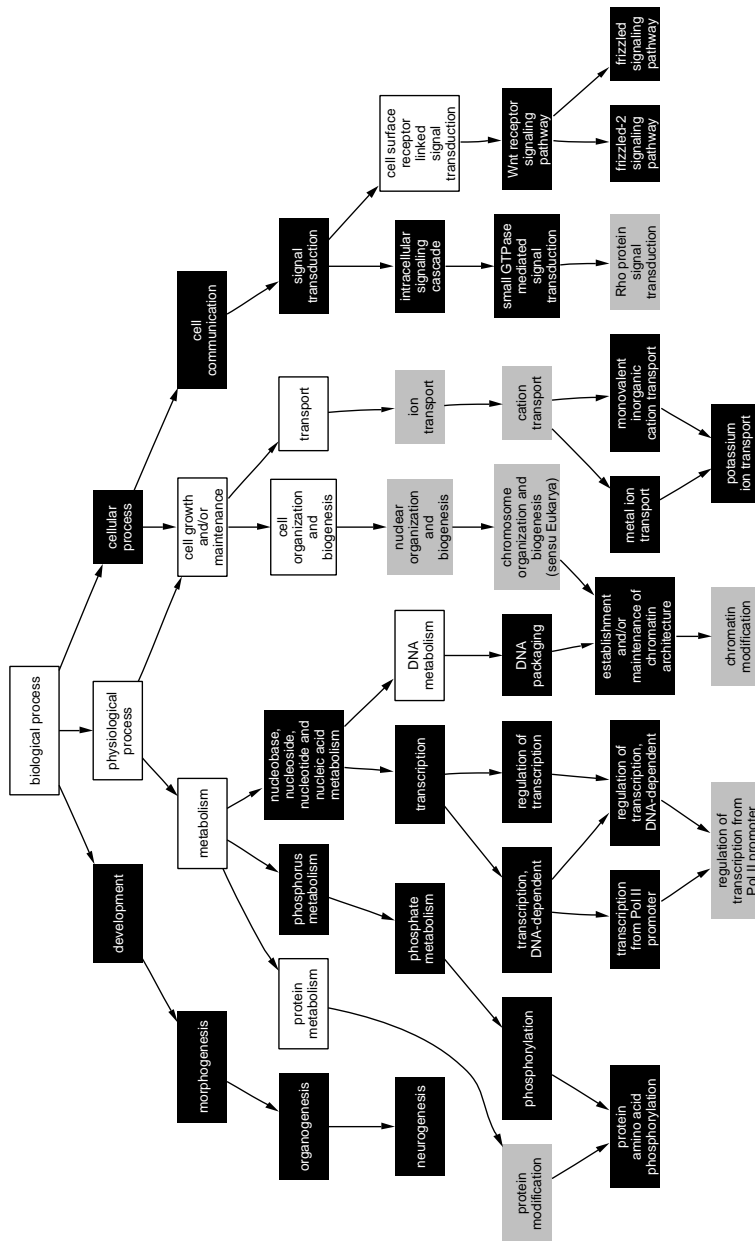| GO id	GO-term	# genes	p-value	FDR
6935	chemotaxis	30	$2.9 \ 10^{-5}$	0.0059
42330	taxis	30	$2.9 \ 10^{-5}$	0.0059
7154	cell communication	406	0.00012	0.016
6950	response to stress	120	0.00028	0.028
9607	response to biotic stimulus	131	0.00030	0.036
7155	cell adhesion	90	0.00050	0.042

Fig. S1. Black and grey boxes indicate the significantly associated biological processes (FDR<0.01 and 0.05, respectively) with the 10% of genes having the highest GC-content in their upstream regions.
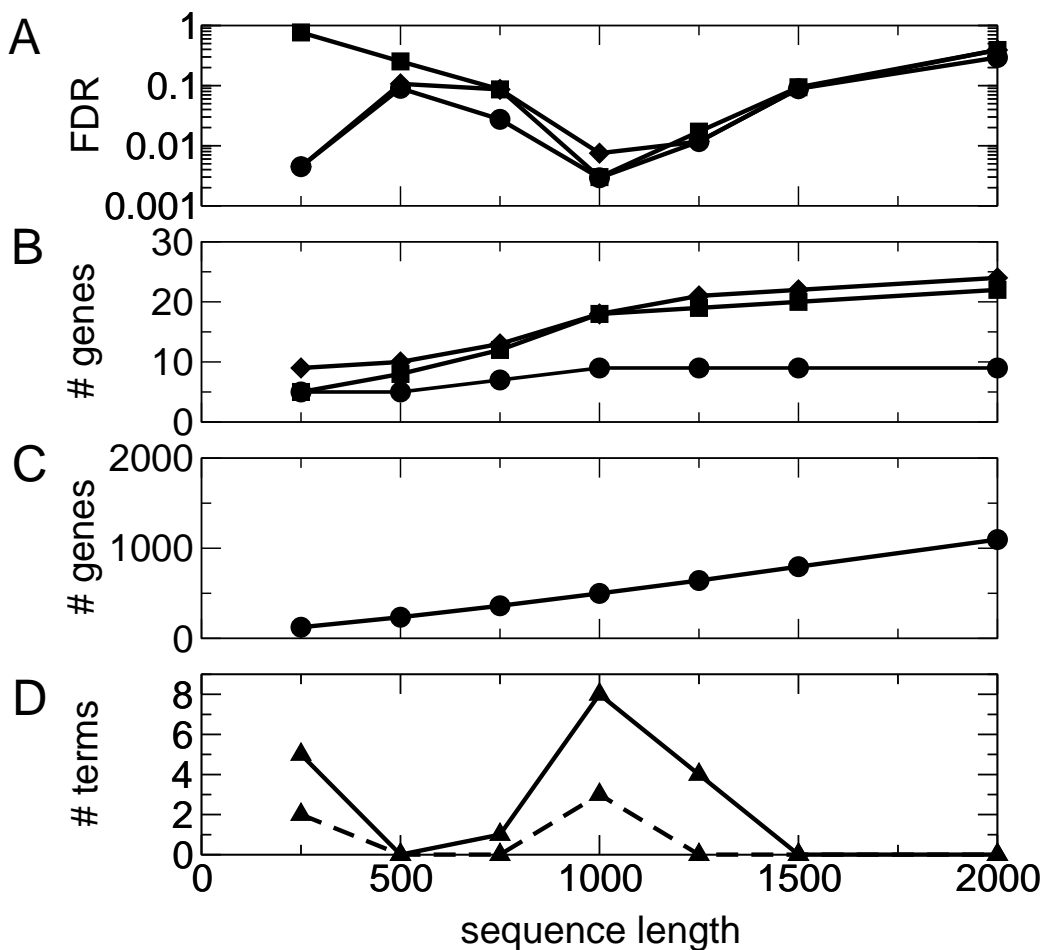
Fig. S2. The effect of varying the length of the upstream regions on the predictions for the set of Mef-2, Myf and TEF transcription factors. A: False Discovery Rate (FDR) of the terms *striated muscle contraction* (circles), *muscle contraction* (squares) and *muscle development* (diamonds). B: Number of genes where a cluster of binding sites has been found by cluster-buster annotated with the terms *striated muscle contraction* (circles), *muscle contraction* (squares) and *muscle development* (diamonds). C: Total number of genes with a cluster of binding sites predicted by cluster-buster. D: Number of significant terms with $FDR < 0.05$ (solid line) and $FDR < 0.01$ (dashed line).
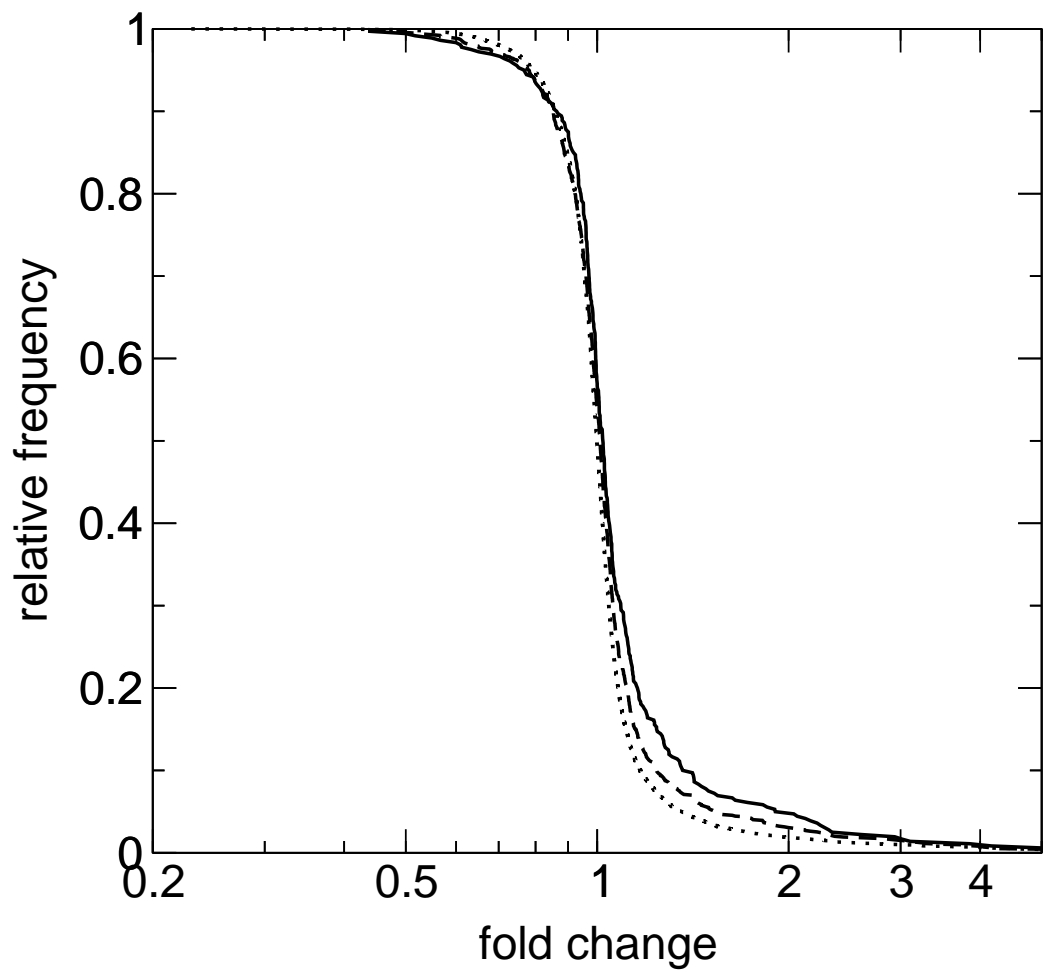
Fig. S3. Normalized cummulative histograms of mean fold changes in the microarray data set for all genes (dotted line), genes where Cluster-Buster detected a cluster of binding sites (dashed line), and after additional filtering with GO (solid line).