

Discovering motifs in ranked list of DNA sequences

Aleksander Sudakov

14.05.07

Source

Eden, E., Lipson, D., Yogev, S., Yakhini, Z. 2007.
Discovering motifs in ranked lists of DNA sequences.
Plos Comput. Biol. Vol. 3, No. 3, e39
[doi:10.1371/journal.pcbi.0030039](https://doi.org/10.1371/journal.pcbi.0030039)

DRIM – Discovery of Rank Imbalanced Motifs
TF BS – transcription factor binding site

Data and motifs

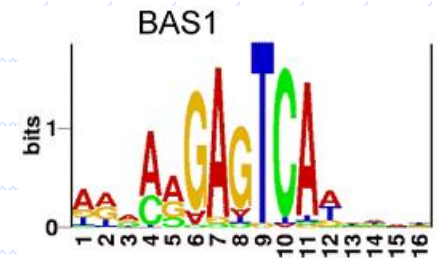
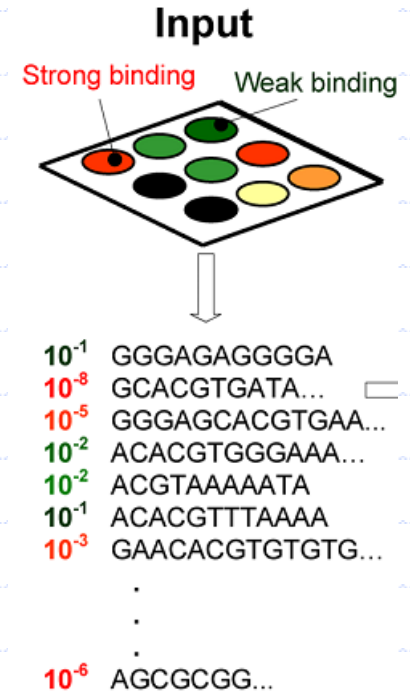
Available

Ranked data

- TF binding signal
- CpG methylation signal
- Expression data

Find

- Target and background set
- Motif k-mer model with symbols above IUPAC



Challenges

In TFBS motif enrichment discovery

- 1) Cutoff
- 2) P-value
- 3) Multiple motifs in single promoter
- 4) False motif discovery in randomly generated data

mHG

Minimum hyper-geometric score

$$\text{Prob}(X = b) = \text{HG}(b; N, B, n) = \frac{\binom{n}{b} \binom{N-n}{B-b}}{\binom{N}{B}}.$$

$$\text{Prob}(X \geq b) = \text{HGT}(b; N, B, n) = \sum_{i=b}^{\min(n, B)} \frac{\binom{n}{i} \binom{N-n}{B-i}}{\binom{N}{B}}$$

$$\lambda = \lambda_1, \dots, \lambda_N \in \{0, 1\}^N \quad b_n(\lambda) = \sum_{i=1}^n \lambda_i$$

$$\text{mHG}(\lambda) = \min_{1 \leq n \leq N} \text{HGT}(b_n(\lambda); N, B, n)$$

Cutoff

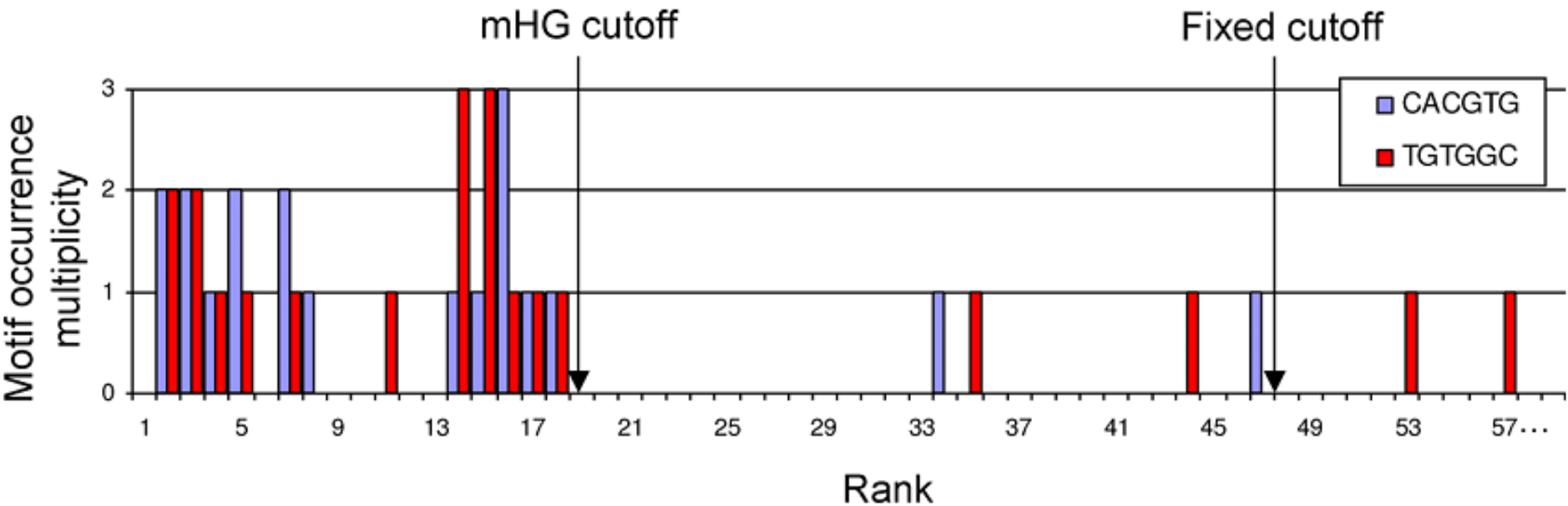


Figure 7. Motif Occurrences in the Top 59 (of ~6,000) Promoters That Were Ranked According to Met32 Binding Signal
 A comparison is made between the data-driven mHG cutoff and the arbitrary fixed cutoff. It can be seen that the motifs are significantly more enriched when the list is partitioned using the mHG cutoff.
 doi:10.1371/journal.pcbi.0030039.g007

Exact p-value of mHG score

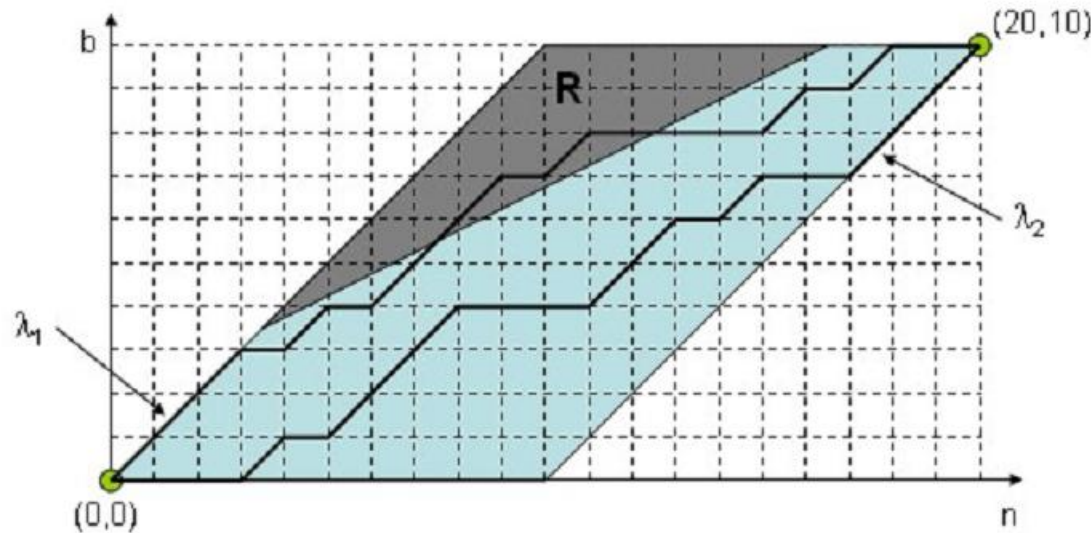
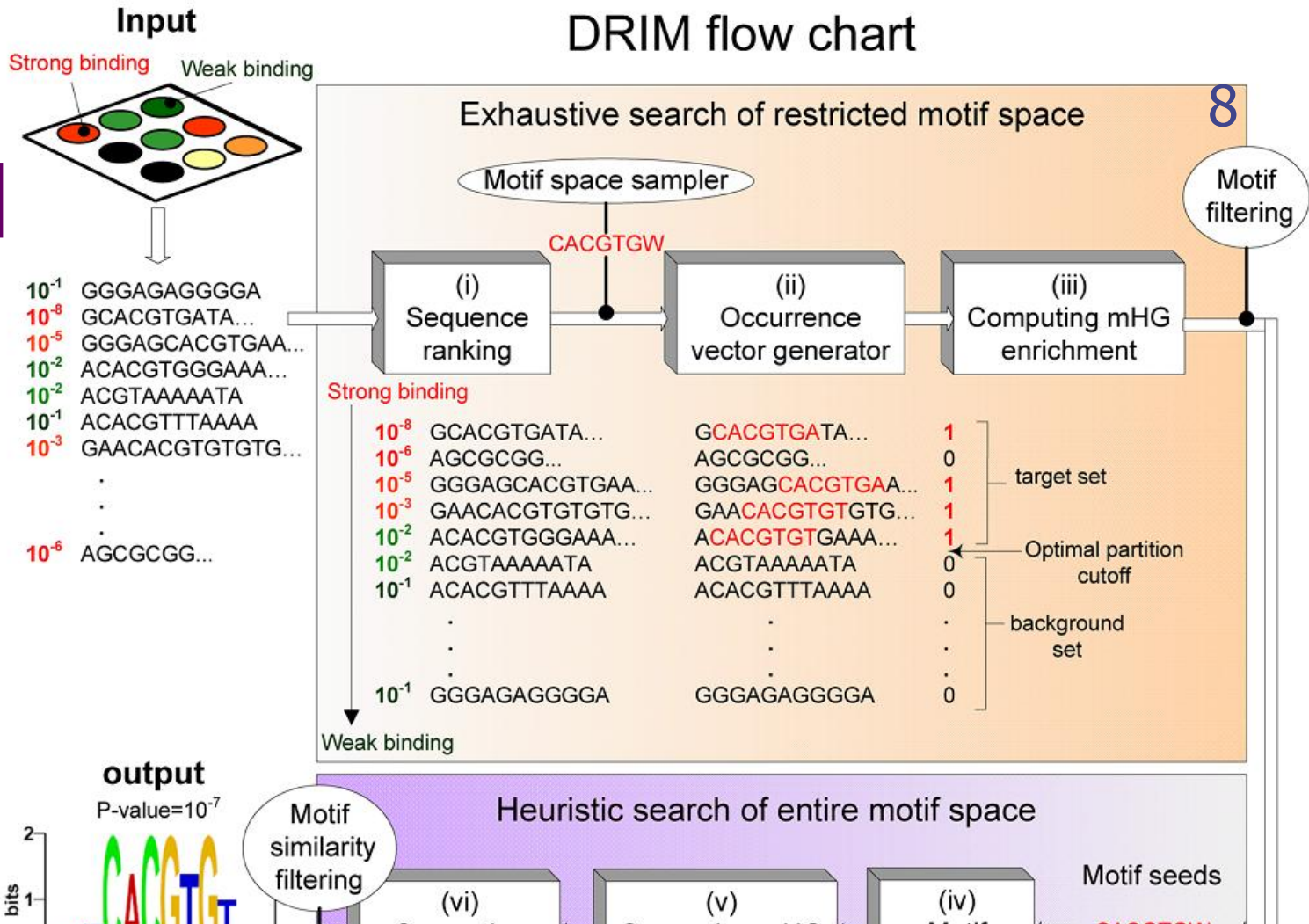


Figure 8: Two-dimensional grid used for calculating mHG p-value. In this example $N = 20$, $B = 10$, $p = 0.1$. Light shaded area describes all attainable values of n and b . Dark shaded area describes the subset R : all values of n and b for which $\text{HGT}(b; N, B, n) \leq p$. Two $(0,0) \rightarrow (N, B)$ paths are depicted, representing the binary label vectors $\lambda_1 = \{1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0\}$ and $\lambda_2 = \{0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1\}$. The path λ_1 traverses R , demonstrating that $\text{mHG}(\lambda_1) \leq p$. The path λ_2 does not traverse R , demonstrating that $\text{mHG}(\lambda_2) > p$.

DRIM

DRIM flow chart



DRIM receives a list of DNA sequences as input and a criterion by which the sequences should be ranked, for example, TF binding signals as measured by ChIP ChIP-chip:

- (i) The sequences are ranked according to the criterion.
- (ii) A “blind search” is performed over all the motifs that reside in the restricted motif space (in this study the restricted motif space contains ~100,000 motifs, see Methods, The DRIM software). For each motif an occurrence vector is generated. Each position in the vector is the number of motif occurrences in the corresponding sequence, (the figure shows the vector for the motif CACGTGW).
- (iii) The motif significance is computed using the mHG scheme, and the optimal partition into target and background sets in terms of motif enrichment is identified. The promising motif seeds are passed as input to the heuristic motif search model and the rest are filtered out.
- (iv,v) The motif seeds are expanded in an iterative manner (the mHG is computed in each lap), until a local optimum motif is found.
- (vi) The exact mHG *p*-value of the motif is computed. If it has a *p*-value < 10⁻³, then it is predicted as a true motif (the choice of this threshold is explained in Results, Proof of principle). The output of the system is the motif representation above IUPAC, its PSSM, mHG *p*-value, and optimal set partition cutoff.

DRIM

Discovery of Rank Imbalanced Motifs

1) Restricted motif space $S \sim 10^5$ motifs:

$$S_1 = \{A, C, G, T, R, W, Y, S, N\}^7$$

$$S_2 = \{A, C, G, T\}^3 N^{3-25} \{A, C, G, T\}^3$$

2) mHG enrichment calculation

3) Motif expansion by heuristic search

Motif similarity filtering

Pick most significant motif and discard all similar and overlapping motifs

$(\alpha, \beta \in \{A, C, G, T, R, Y, W, S, M, K, H, B, V, D, N\})$

$$\delta(\alpha, \beta) = 1 - 2 \frac{|\alpha \cap \beta|}{|\alpha| + |\beta|}$$

1. For identical symbols $\delta(\alpha, \beta) = 0$ (i.e. identical symbols have a distance of 0) ;
2. For disjoint symbols $\delta(\alpha, \beta) = 1$ (e.g. $\{A\}$ vs. $\{C\}$ or $W = \{A, T\}$ vs. $S = \{C, G\}$) ;
3. $0 \leq \delta(\alpha, \beta) \leq 1$ (e.g. $\alpha = \{A, C, G\}$, $\beta = \{A, G\} \rightarrow \delta(\alpha, \beta) = 0.2$) .

$$D(a, b) = \sum_{i=1}^{\min(n,m)} \delta(a_i, b_i) + w_1 u,$$

Multi-mHG

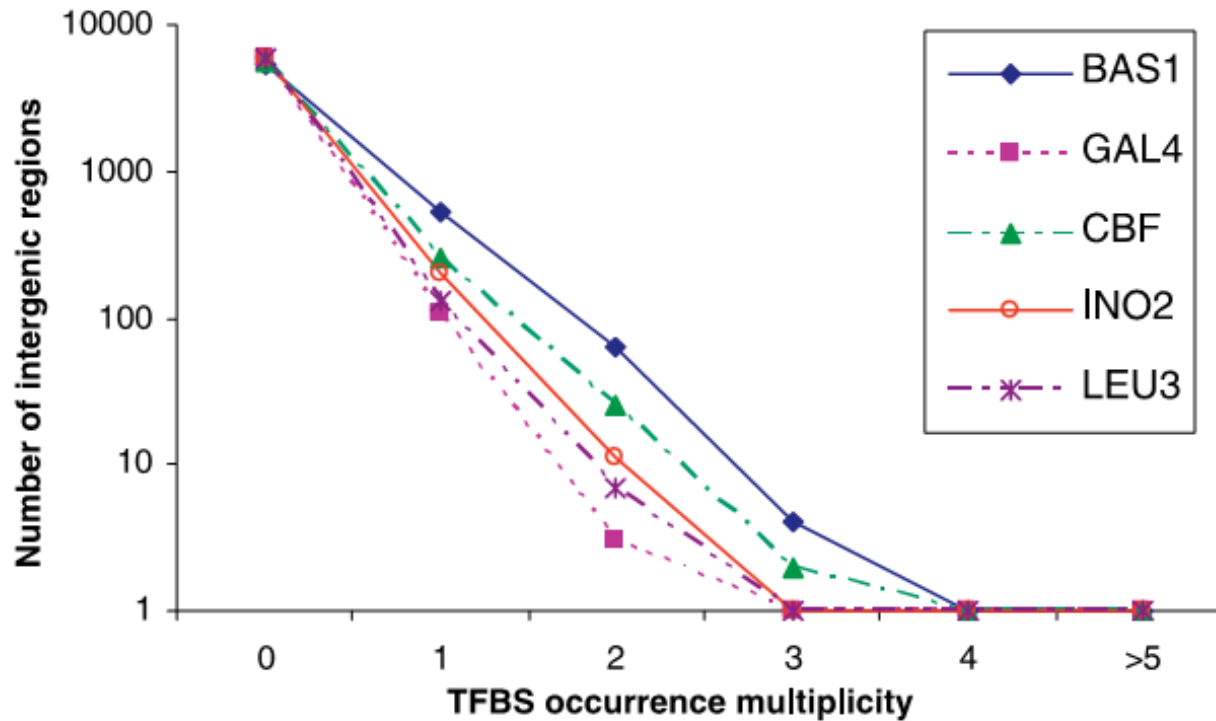


Figure 9. The Distribution of TFBS Occurrence Multiplicities per Intergenic Region in *S. cerevisiae* Is Shown for Five TFs Whose TFBS Motif Was Experimentally Verified

Note that the y-axis is logarithmic. It can be seen that in most instances the TFBS appears in either zero, one, or two copies per intergenic region.
doi:10.1371/journal.pcbi.0030039.g009

Proof of principle

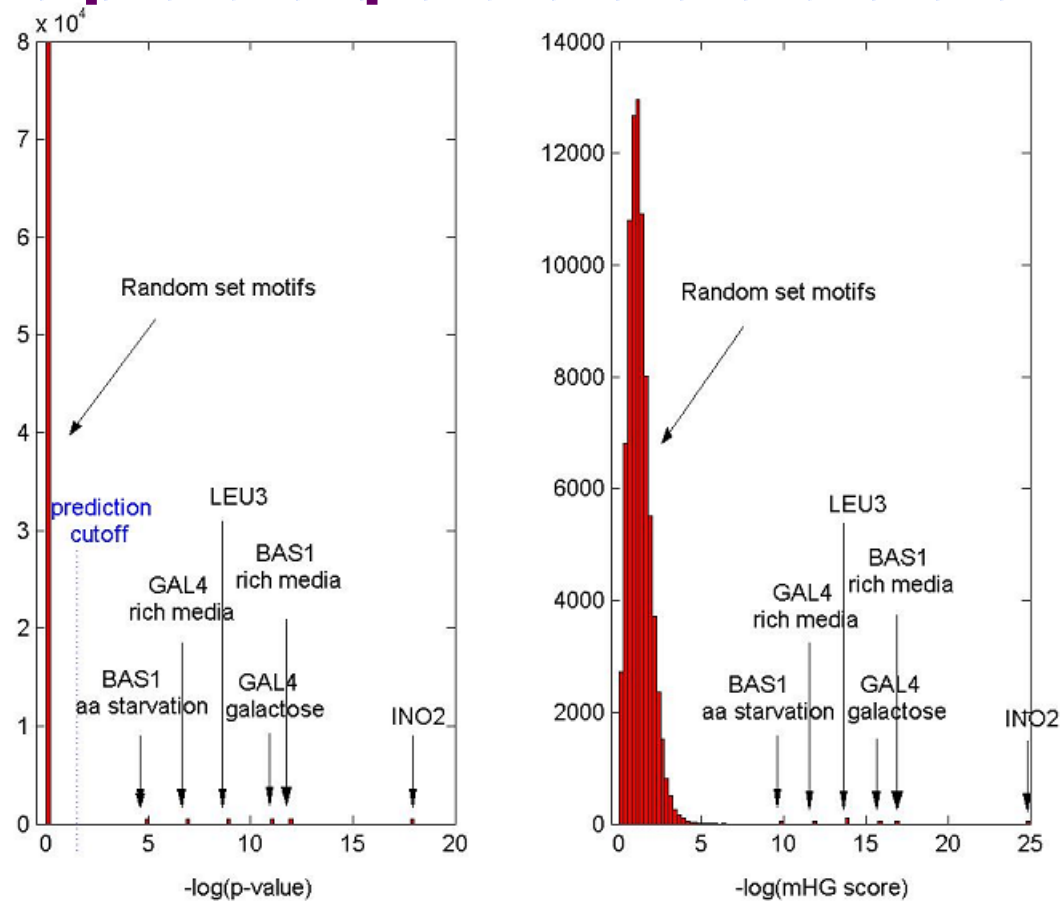


Figure S3: Comparison of mHG score and p-value distributions for motifs in randomly ranked sequences with those of true TFBS motifs in ranked lists derived from the corresponding ChIP-chip assays. $\sim 100,000$ motifs were scanned in 400 randomly ranked genomic sequences, and their corresponding corrected p-value (a) and mHG score (b) were recorded. The corrected p-values involves two levels of multiple test corrections: correction on the number motifs that were tested and correction for the multiple cutoffs that are tested as part of the mHG optimization process. None of the tested motifs had a corrected p-value $< 10^{-3}$. DRIM was applied on the ChIP-chip data of 5 TFs and the mHG scores and corrected p-values of the true TFBS motifs (as previously determined experimentally) were recorded. In all instances the true TFBS motifs were predicted with p-values that were several orders of magnitude more significant than the best random set motif p-value.

Application

Yeast ChIP-chip

Human cancer CpG methylation

Human ChIP-chip

Expression data: ranking by over-, under-
or differential expression

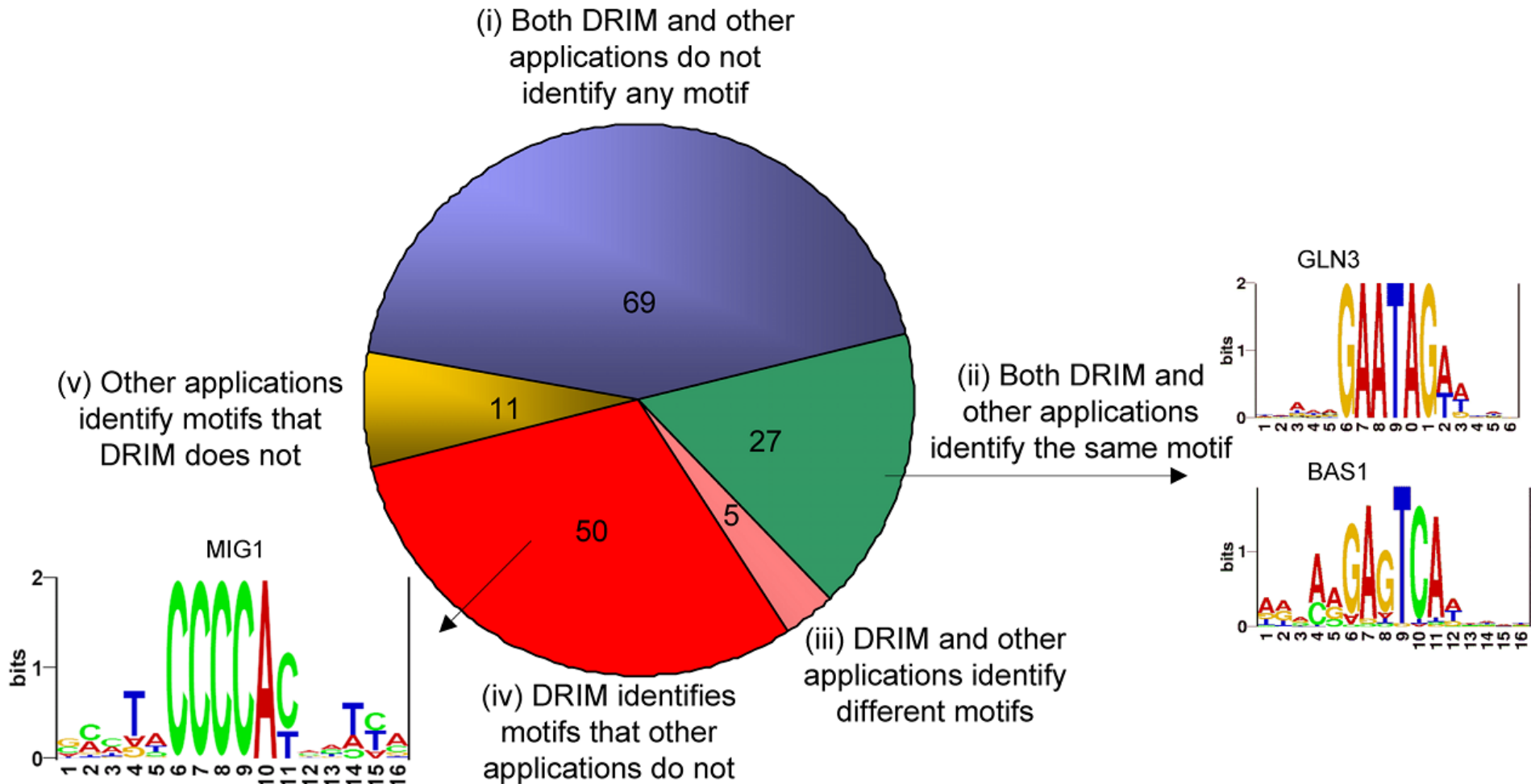


Figure 2. Comparison between Predictions of DRIM and Published Predictions of Six Other Methods and Conservation Data as Reported in [25]. Overall, out of 162 unique TFs, DRIM identified significant motifs for 82 TFs with p -value $< 10^{-3}$. Out of the 162 TFs, DRIM and the other applications agree on 96 TFs: 27 TFs for which a similar motif was found and 69 TFs for which no significant motifs were found. There are five TFs for which the motifs predicted by DRIM and other applications differ; 11 for which the other applications identified motifs that DRIM did not; and 50 for which DRIM identified a motif that the other applications did not (for details see Tables S2 and S3). Sequence logos were generated using the *RNA Structure Logo* software [56].
doi:10.1371/journal.pcbi.0030039.g002

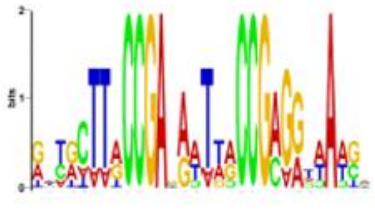
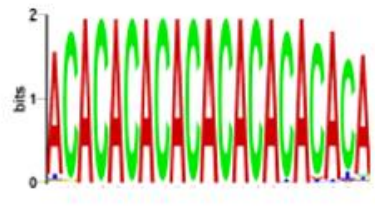
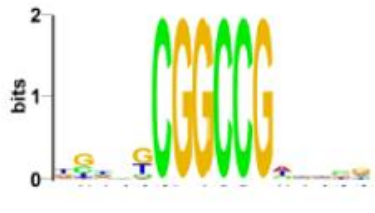


| TF | Motif | p-value |
|--|--|---|
| Aro80 (YPD and SM) |  | 10^{-11} |
| ARR1 (YPD), GCR2 (YPD), IME4 (YPD), ACE2 (YPD), AFT2 (H ₂ O ₂), MAL33 (H ₂ O ₂), SFP1(H ₂ O ₂) |  | $10^{-6}, 10^{-5},$ $10^{-19}, 10^{-8},$ $10^{-21}, 10^{-27},$ 10^{-4} |
| IME1 (H ₂ O ₂) |  | 10^{-6} |
| Met4 (YPD and SM), Met31 (YPD), Met32 (YPD and SM) |  | $10^{-5},$ $10^{-4},$ 10^{-8} |
| Met4 (YPD and SM), Met31 (YPD), Met32 (YPD and SM) |  | $10^{-3},$ $10^{-3},$ 10^{-3} |

Figure 3. Examples of TFs for Which DRIM Identifies Novel Motifs

We further investigated these motifs and show evidence of their biological function. YPD, H₂O₂, and SM denote the ChIP-chip experimental conditions [25] in which the motifs were identified.

doi:10.1371/journal.pcbi.0030039.g003

Table 1. Enriched Motifs Associated with CpG Methylation in Four Human Cancer Cell Lines and Comparison to Motifs in Regions Bound by the Polycomb Complex

| Cell Line | CpG Methylation Motif | Number of Experiments | Average <i>p</i> -Value | Notes | Polycomb Complex Motif |
|-----------|------------------------|-----------------------|-------------------------|-----------------------|------------------------|
| Caco-2 | SSCCCCANG ^a | 4 | <10 ⁻¹⁰ | Novel prediction | Yes [41,44] |
| Caco-2 | CNGCTGC ^a | 3 | <10 ⁻⁵ | Novel prediction | Yes [41] |
| Caco-2 | GAGGGA | 2 | <10 ⁻⁴ | In agreement with [2] | |
| Caco-2 | DGAGAGV | 2 | <10 ⁻⁴ | Novel prediction | Yes [41,43,44] |
| Carcinoma | CA repeat | 2 | <10 ⁻⁷⁹ | Novel prediction | Yes [41,42] |
| PC3 | CA repeat | 1 | <10 ⁻⁷ | Novel prediction | Yes [41,42] |
| PC3 | GGGGTNCC ^a | 1 | <10 ⁻⁶ | In agreement with [2] | Yes [44] |
| PC3 | ACACNCAC | 2 | <10 ⁻¹⁰ | In agreement with [2] | |
| PC3 | GCTGC | 2 | <10 ⁻⁵ | Novel prediction | Yes [41] |
| PC3 | RGCCAA | 2 | <10 ⁻⁴ | Novel prediction | |
| Polyp | CA repeat | 2 | <10 ⁻⁵⁸ | Novel prediction | Yes [41,42] |
| Polyp | CNNGCGCC ^a | 3 | <10 ⁻¹³ | Novel prediction | Yes [44] |
| Polyp | GCTGCNBB | 2 | <10 ⁻⁶ | Novel prediction | Yes [41] |

Number of Experiments corresponds to the number of replicate experiments of the same cell line in which the same motif was independently identified. The CA repeat motifs have a variable length.

Polycomb Complex Motif denotes motifs that appear in regions bound by the Polycomb complex [41,42,44].

^aMotifs that have G-C content >66%. Their enrichments are partially attributed to the G-C content bias that is found in the CpG methylation data.

doi:10.1371/journal.pcbi.0030039.t001