BIOINFORMATICS

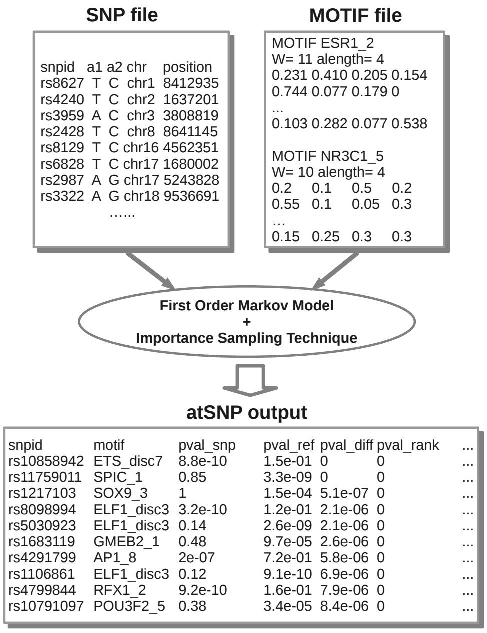
Vol. 00 no. 00 2005
Pages 1–28

Supplementary Materials for “atSNP: Transcription factor binding affinity testing for regulatory SNP detection”

Chandler Zuo [1,2,], Sunyoung Shin [1,2] and Sündüz Keleş,[1,2*]

[1]Department of Statistics, University of Wisconsin Madison.
[2]Department of Biostatistics and Medical Informatics, University of Wisconsin Madison.

1 OVERVIEW OF ATSNP



```
snpid   a1 a2 chr   position
rs8627  T  C  chr1  8412935
rs4240  T  C  chr2  1637201
rs3959  A  C  chr3  3808819
rs2428  T  C  chr8  8641145
rs8129  T  C chr16 4562351
rs6828  T  C chr17 1680002
rs2987  A  G chr17 5243828
rs3322  A  G chr18 9536691
              ……..
```

```
MOTIF ESR1_2
W= 11 alength= 4
0.231 0.410 0.205 0.154
0.744 0.077 0.179 0
...
0.103 0.282 0.077 0.538

MOTIF NR3C1_5
W= 10 alength= 4
0.2    0.1    0.5    0.2
0.55   0.1    0.05   0.3
...
0.15   0.25   0.3    0.3
```

```
snpid       motif        pval_snp   pval_ref pval_diff pval_rank   ...
rs10858942  ETS_disc7    8.8e-10    1.5e-01  0         0           ...
rs11759011  SPIC_1       0.85       3.3e-09  0         0           ...
rs1217103   SOX9_3       1          1.5e-04  5.1e-07   0           ...
rs8098994   ELF1_disc3   3.2e-10    1.2e-01  2.1e-06   0           ...
rs5030923   ELF1_disc3   0.14       2.6e-09  2.1e-06   0           ...
rs1683119   GMEB2_1      0.48       9.7e-05  2.6e-06   0           ...
rs4291799   AP1_8        2e-07      7.2e-01  5.8e-06   0           ...
rs1106861   ELF1_disc3   0.12       9.1e-10  6.9e-06   0           ...
rs4799844   RFX1_2       9.2e-10    1.6e-01  7.9e-06   0           ...
rs10791097  POU3F2_5     0.38       3.4e-05  8.4e-06   0           ...
```

Fig. 1. A flow chart describing atSNP analysis. The input SNP file contains the reference (a1) and the SNP (a2) alleles; however, when only dbSNP IDs are provided, atSNP acquires the necessary location and allele information using the R package `rsnps` (`http://cran.r-project.org/web/packages/rsnps/rsnps.pdf`). The motif file is in MEME motif format, one of the several allowed formats. atSNP uses a first order Markov model for generating random background sequences and importance sampling techniques for efficient p-value calculation. This atSNP output table contains the SNPs with the most significant affinity score changes for our example in Section 3 of the main paper. Each row provides in-depth SNP-motif pair information such as SNP ID (snpid), motif name (motif), p-value for the binding affinity with the SNP and the reference alleles (pval_snp, pval_ref), and the p-value for binding affinity change based on log-likelihood ratio and log-rank ratio (pval_diff, pval_rank).

2 IMPORTANCE SAMPLING ALGORITHMS

In this section, we describe the algorithms for computing and testing the affinity scores for each allele and change in affinity scores between the alleles. We code the four nucleotides by ‘A’-1, ’C’-2, ’G’-3, and ’T’-4. The reverse complement of nucleotide i is obtained by $5 - i$ in this coding scheme. Let W denote the $4 \times L$ position weight matrix for a motif of length L and $W(i, l)$ the entry for nucleotide i at position l with $\sum_{i=1}^{4} W(i, l) = 1$.

The affinity score calculation requires considering all possible nucleotide sequences of length L that overlaps the SNP position. Such a sequence must be located within a window of size $2L - 1$ around the SNP position. Let $\mathbf{x} = (x_1, x_2, \cdots, x_{2L-1})$ denote the nucleotides in this window. The binding affinity score of a subsequence $(x_s, x_{s+1}, \cdots, x_{s+L-1})$ is given by

$$C(\mathbf{x}, s) = \sum_{l=1}^{L} \log W(x_{l+s-1}, l). \tag{1}$$

Then, the affinity score of $\mathbf{x}$ is the maximum of the scores across all subsequences from both strands given by

$$C(\mathbf{x}) = \max\{C(T(\mathbf{x}), s) : T \in \{I, R\}, s = 1, 2, \cdots, L\},$$

where I and R are two strand operators with $I(\mathbf{x}) = \mathbf{x}$, $R(\mathbf{x}) = (5 - x_{2L-1}, 5 - x_{2L-2}, \cdots, 5 - x_1)$, i.e., the reverse complement sequence.

The binding affinity score definition in Eqn. (1) assumes that W describes a motif with a product multinomial distribution as in Grant *et al.* (2011); Chan *et al.* (2010), i.e., $W(i, l) \in [0, 1]$, and, therefore, $C(\mathbf{x}, s)$ represents the log-likelihood of the subsequence starting at position s under this model. If W is already a transformed version of the product multinomial model parameters, e.g., $W(i, l) \in \mathbb{R}$, then the affinity score simply corresponds to

$$C(\mathbf{x}, s) = \sum_{l=1}^{L} W(x_{l+s-1}, l). \tag{2}$$

The affinity tests of atSNP are based on Eqn. (1) by default; however, they can be modified to adapt Eqn. (2) by an exponential transformation of the entries of the PWM, i.e., by replacing $W(i, l)$ with $\exp(W(i, l))$. In the subsequent sections, we describe the p-value computation algorithms based on Eqn. (1). These algorithms readily provide the tests for Eqn. (2) once we apply the exponential transformation.

*To whom correspondence should be addressed.

2.1 Computing and testing allele-specific binding affinity scores

We assume that, under the null hypothesis that a subsequence overlapping the SNP comes from a genomic background distribution, the nucleotide sequences follow a stationary reversible first order Markov model with distribution $P(X_l = k) = \pi(k)$, $k = 1, \cdots, 4$, and transition probabilities $P(X_{l+1} = n|X_l = k) = p(k, n)$, $k, n = 1, \cdots, 4$. Under this model, the joint probability for sequence $\mathbf{x}$ is given by

$$f_{\mathcal{H}_0}(\mathbf{x}) = \pi(x_1) \prod_{l=1}^{2L-2} p(x_l, x_{l+1}). \tag{3}$$

Given an observed sequence $\mathbf{x}_0$, either from the reference or the SNP allele, atSNP computes the allele-specific p-value defined as the probability that affinity score of a sequence from the null background model is at least as large as $C(\mathbf{x}_0)$:

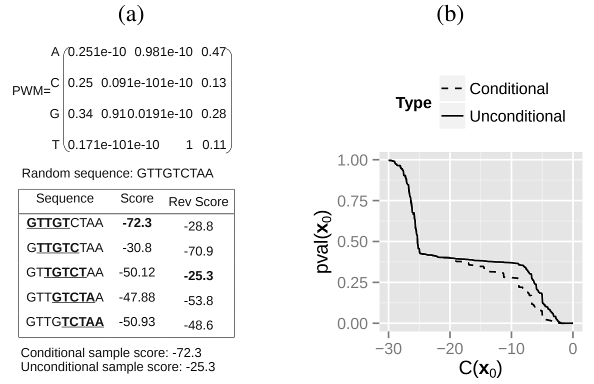
$$pval(\mathbf{x}_0) = P\{C(\mathbf{X}) \geq C(\mathbf{x}_0)|\mathbf{X} \sim f_{\mathcal{H}_0}\}, \tag{4}$$

where $\mathbf{X}$ is the random variable denoting the sequence of length $2L - 1$ overlapping the SNP. Note that this p-value corresponds to the whole sequence of length $2L - 1$ which includes all subsequences of length L that can overlap the SNP position. Another useful quantity is the so-called *conditional p-value* that can be calculated for a fixed subsequence of length L. The traditional algorithms, such as FIMO, that scan a sequence with PWMs calculate such p-values for each subsequence. Formally, we define the conditional p-values as follows. Given the observed sequence $\mathbf{x}_0$, we first find the location of the subsequence that best matches the PWM: $(T_0, s_0) = \arg\max\{C(T(\mathbf{x}_0), s) : T \in \{I, R\},\ 1 \leq s \leq L\}$. The conditional p-value is the probability for the score of a random sequence evaluated at this fixed location to be as large as the observed score $C(\mathbf{x}_0)$. This can be formulated as:

$$pval'(\mathbf{x}_0) = P\{C(T_0(\mathbf{X}), s_0) \geq C(\mathbf{x}_0)|\mathbf{X} \sim f_{\mathcal{H}_0}\}. \tag{5}$$

Before we describe the estimation algorithm for the p-values, we will discuss the differences between these two p-value types in Figure 2. Both quantities compare $C(\mathbf{x}_0)$ with the affinity scores from sequences randomly generated under the null model. Given a null sequence, the conditional p-value calculates the sample affinity score based on a fixed strand and location, while the p-value calculates the maximum affinity score based on all subsequences from both strands. As a result, the sample affinity scores corresponding to the p-values are at least as large as those for the conditional p-values. Therefore, p-values are always larger than the conditional p-values; however, they directly reflect the significance of the maximum affinity score for the observed sequence. We argue that, because we do not know the location of the subsequences that best match to the motifs, Eqn. (4) is more appropriate for calculating allele-specific significance. atSNP provides computation of both the p-values and the conditional p-values, thereby allows us to compare its accuracy with the conditional p-values from FIMO (Subsection 4.1).

Next, we describe the estimation algorithm. If we can simulate B sequences, $\mathbf{x}_1, \cdots, \mathbf{x}_B$, under the null distribution $f_{\mathcal{H}_0}$, an empirical estimator for the p-value is $\sum_{b=1}^{B} 1\{C(\mathbf{x}_b) \geq C(\mathbf{x}_0)\}/B$. We note that the p-value is just the probability of the event $\{C(\mathbf{X}) \geq C(\mathbf{x}_0)\}$. This is a rare event when p-values are small, and the naive Monte-Carlo simulation method requires a large number of simulations. The importance sampling technique addresses this problem by the following insight:



Sequence	Score	Rev Score
GTTGTCTAA	**-72.3**	-28.8
G**TTGTC**TAA	-30.8	-70.9
GT**TGTCT**AA	-50.12	**-25.3**
GTT**GTCTA**A	-47.88	-53.8
GTTG**TCTAA**	-50.93	-48.6

Conditional sample score: -72.3
Unconditional sample score: -25.3

Fig. 2. Difference between the p-value and the conditional p-value. (a) Computing the sample affinity scores based on a random sequence generated from the null hypothesis. In this example, we assume $T_0 = I$, $s_0 = 1$, and we have $C(T_0(\mathbf{x}), s_0) = -72.3 > C(\mathbf{x}) = -25.3$. For any random sequence $\mathbf{x}$, we have $C(T_0(\mathbf{x}), s_0) \geq C(\mathbf{x})$. As a consequence, the conditional p-value is always no less than the p-value. (b) The p-values and conditional p-values at different affinity scores.

$$\begin{aligned} pval(\mathbf{x}_0) =& E[1\{C(\mathbf{X}) \geq C(\mathbf{x}_0)\}|\mathbf{X} \sim f_{\mathcal{H}_0}] \\ =& E[1\{C(\mathbf{X}) \geq C(\mathbf{x}_0)\} \frac{f_{\mathcal{H}_0}(\mathbf{X})}{h(\mathbf{X})}|\mathbf{X} \sim h], \end{aligned} \tag{6}$$

where h is a sampling distribution under which the event $\{C(\mathbf{X}) \geq C(\mathbf{x}_0)\}$ occurs more often compared to $f_{\mathcal{H}_0}$. Motivated by the idea from Chan *et al.* (2010), we consider a sampling distribution by adding the exponents of the affinity score as weights to $f_{\mathcal{H}_0}$. First, we consider sampling a random sequence $\mathbf{X}$ and a motif matching position S from the following distribution:

$$g_\theta(\mathbf{x}, s) = \frac{f_{\mathcal{H}_0}(\mathbf{x}) \exp(\theta C(\mathbf{x}, s))}{H(\theta)}.$$

Here, θ is a tilting parameter and $H(\theta)$ is the normalizing constant. Because we put a weight of $\exp(\theta C(\mathbf{x}, s))$, when $\theta > 0$, we are more likely to get sequences with large affinity scores. Then, the sampling distribution for the sequence $\mathbf{X}$ is given by

$$h_\theta(\mathbf{x}) = \frac{\sum_{s=1}^{L} f_{\mathcal{H}_0}(\mathbf{x}) \exp(\theta C(\mathbf{x}, s))}{H(\theta)}. \tag{7}$$

A useful property for g_θ is:

$$E(C(\mathbf{X}, S)|(\mathbf{X}, S) \sim g_\theta) = \frac{d}{d\theta} \log H(\theta).$$

Since, under g_θ, a random sequence $\mathbf{x}$ tends to have a large affinity score for a subsequence starting from s, it is very likely that $C(\mathbf{x}, s) = C(\mathbf{x})$. In other words, if we simulate sequences under g_θ, then the expected value of $C(\mathbf{X})$ is approximately $\frac{d}{d\theta} \log H(\theta)$. Chan *et al.* (2010) suggested choosing θ such that $E(C(\mathbf{X})|g_\theta) \approx C(\mathbf{x}_0)$ for estimating p-value at $C(\mathbf{x}_0)$. Following this suggestion, we first group the scores from all SNPs into multiple ranges, and then for each range where the scores are close to c, we use the sampling distribution g_θ with θ set by solving $\frac{d}{d\theta}\log H(\theta) = c$.

Finally, the p-values can be estimated by

$$\widehat{pval}(\mathbf{x}_0) = \frac{1}{T}\sum_{t=1}^{T} 1\{C(\mathbf{x}_t) \geq C(\mathbf{x}_0)\}\frac{H(\theta)}{\sum_{s=1}^{L}\exp(\theta C(\mathbf{x}_t, s))}. \quad (8)$$

Similarly, the conditional p-value can be estimated using the same sampling distribution by

$$\widehat{pval'}(\mathbf{x}_0) = \frac{1}{B}\sum_{b=1}^{B} 1\{C(T(\mathbf{x}_b), s_b) \geq C(\mathbf{x}_0)\}\frac{H(\theta)}{\exp(\theta C(\mathbf{x}_b, s_b))}. \quad (9)$$

In summary, the p-values for all SNPs for a given PWM can be computed by this importance sampling scheme as follows:

1. Group the affinity scores into different sets $\mathcal{G}_1, \cdots, \mathcal{G}_K$ such that the scores within each set are close to each other and to c_k. K and representative score value c_k for each set k, $k = 1, \cdots, K$ are set as follows.

   a. Denote the number of SNPs by N. If $N \leq 20$, then each $\mathcal{G}_k$ is the singleton set of one score, and $K = N$.

   b. If $N > 20$, set $K = 20$, and $p_k = 1 - N^{-k(k+1)/[K(K+1)]}$ for $1 \leq k \leq K$.

   c. Set c_k as the $100 \times p_k$-th percentile of the observed scores of all SNPs across both alleles.

   d. Set $\mathcal{G}_k$ as the set of the scores in the interval $((c_{k-1} + c_k)/2, (c_k + c_{k+1})/2]$ for $2 \leq k \leq K-1$, $\mathcal{G}_1$ as the set of the scores in the interval $(-\infty, (c_1 + c_2)/2]$, $\mathcal{G}_K$ as the set of the scores in the interval $((c_{K-1} + c_K)/2, \infty)$. Set c_k as the representative score for $\mathcal{G}_k$.

2. For each set of scores $\{C(\mathbf{x}_0^i)\}$ in $\mathcal{G}_k$, $k = 1, \cdots, K$ with representative score c_k:

   a. Set $\theta : \frac{d}{d\theta}\log H(\theta) = c_k$. Calculate $H(\theta)$.

   b. To set the Monte-Carlo sample size B, first calculate B' as the integer part of $100(1 - p_k)/p_k$. If $B' > 10^5$, set $B = 10^5$; if $B' < 2000$, set $B = 2000$; otherwise, set $B = B'$.

   c. Simulate B Monte-Carlo samples $(\mathbf{x}_b, s_b)$ from the distribution g_θ. Compute $C(\mathbf{x}_b)$ and $\sum_{s=1}^{L} C(\mathbf{x}_b, s)$. Let $(T_b, s_b) = \arg\max\{C(T(\mathbf{x}_b), s) : T \in \{I, R\},\ 1 \leq s \leq L\}$.

   d. Estimate the p-value and the conditional p-value for each $\mathbf{x}_0^i \in \mathcal{G}_k$ by Eqns. (8) and (9).

The details for computing $H(\theta)$ and sampling from g_θ are discussed in Section 3.

2.2 Computing and testing binding affinity score change between alleles

We assume that the sequence of the reference allele $\mathbf{X}$ under the null distribution follows the first order Markov model in Eqn. (3). The SNP allele sequence differs from the reference allele sequence only by nucleotide x_L. We let $\mathbf{x}^a = (x_1, \cdots, x_{L-1}, x_L^a, x_{L+1}, \cdots, x_{2L-1})$ denote the sequence with the SNP allele and assume that

$$P(X_L^a = x_L^a | X_L = x_l) = \frac{1\{x_L^a \neq x_L\}}{3}.$$

Then, the joint distribution of $\mathbf{x}$ and $\mathbf{x}^a$ is given by

$$f^a(\mathbf{x}, \mathbf{x}^a) = \frac{1\{x_L \neq x_L^a\}}{3}\pi(x_1)\prod_{l=2}^{2L-1} p(x_{l-1}, x_l)1\{x_L \neq x_L^a\}.$$

For a given SNP-PWM pair, atSNP evaluates whether the SNP allele impacts the match to PWM significantly, either by disrupting a subsequence overlapping the SNP position with good binding affinity score or generating a subsequence with even better score. It computes two types of p-values corresponding to different test statistics. The first p-value, denoted by $pval_d$, assesses whether the change in the binding affinity scores of the two alleles is significantly different than what would be expected by chance and is given by

$$pval_d(\mathbf{x}_0, \mathbf{x}_0^a) = P\{|C(\mathbf{X}) - C(\mathbf{X}^a)| \geq |C(\mathbf{x}_0) - C(\mathbf{x}_0^a)|| \\ (\mathbf{X}, \mathbf{X}^a) \sim f^a\}.$$

The second p-value, denoted by $pval_r$ assesses whether the change in the ranks of the PWM matches of the subsequences with the reference and SNP alleles is significantly different than what would be expected by chance and is given by

$$pval_r(\mathbf{x}_0, \mathbf{x}_0^a) = P\{|\log(pval(\mathbf{X})) - \log(pval(\mathbf{X}^a))| \geq \\ |\log(pval(\mathbf{x}_0)) - \log(pval(\mathbf{x}_0^a))|| \\ (\mathbf{X}, \mathbf{X}^a) \sim f^a\},$$

where $pval$ follows the definition in Eqn. (4). $pval_d$, which compares the log of the likelihoods of the best subsequence matches to the PWM with the reference and SNP allele, is motivated by the likelihood ratio test framework and is easier to compute. However, we observed that since the binding affinity score difference is bounded, if the PWM has multiple highly conserved bases with probability of the relevant nucleotide occurrence close to 1, the p-value at the maximum score difference can still be insignificant. The rank test attenuates this problem since the maximum log rank ratio, $|\log(pval(\mathbf{X})) - \log(pval(\mathbf{X}^a))|$, is essentially unbounded. Figure 3 provides an example illustrating this difference.

Next, we introduce a few additional quantities to derive the importance sampling distribution. Let IW, a $4 \times L$ matrix, denote induced PWM with entries

$$IW(i, l) = \frac{W(i, l) + 1/4}{2}.$$

Let D denote another $4 \times L$ matrix with entries

$$D(i, l) = \exp\left(\frac{\sum_{j \neq i}(\log W(i, l) - \log W(j, l))}{3}\right).$$

In the sampling distribution, we assume that in addition to a subsequence of length L in the center, subsequences on the two

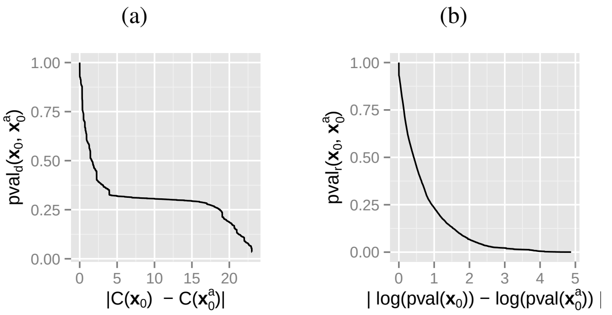


Fig. 3. Comparison between the score statistic- ($pval_d$) and rank-based ($pval_r$) p-values. (a) The score test p-values at varying score changes between the reference and the SNP allele. The p-value at the maximum possible score change is 0.0312. (b) The rank test p-values at varying log rank ratios.

ends follow the Markov model. With a slight abuse of the notation as $f(x_m, \cdots, x_n) = \pi(x_m)p(x_m, x_{m+1})\cdots p(x_{n-1}, x_n)$, we have

$$\begin{aligned} h^a_\theta(\mathbf{x}, \mathbf{x}^a) =& \frac{1\{x^a_L \neq x_L\}}{3H^a(\theta)} \sum_{s=1}^{L} \Bigg\{ f(x_1, \cdots, x_{s-1}) \\ & \Big[\prod_{1\le l\le L, l\neq L} IW(x_{s+l-1}, l) \Big] D(x_L, L-s+1)^\theta \\ & f(x_{L+s}, \cdots, x_{2L-1}) \Bigg\}, \end{aligned}$$

which marginalizes g^a_θ over s:

$$\begin{aligned} g^a_\theta(\mathbf{x}, \mathbf{x}_0, s) =& \frac{1\{x^a_L \neq x_L\}}{3H^a(\theta)} f(x_1, \cdots, x_{s-1}) \\ & \Big[\prod_{1\le l\le L, l\neq L} IW(x_{s+l-1}, l) \Big] D(x_L, L-s+1)^\theta \\ & f(x_{L+s}, \cdots, x_{2L-1}). \end{aligned}$$

The key points when simulating sequences for calculating change in binding affinity scores are (1) the sequence should have a subsequence matching to the PWM and (2) a change at base x_L of the SNP position will result in a large change in the affinity score. In g^a_θ, $\big[\prod_{l=1}^{L} IW(x_{s+l-1}, l)\big]$ weighs a length L subsequence starting from s, and $\log D(x_L, L-s+1)$ is the expected change in affinity score for this subsequence when x_L is changed. We also have

$$E_{g^a_\theta}[C(\mathbf{x}, s) - C(\mathbf{x}^a, s)] = \frac{d}{d\theta} \log H^a(\theta).$$

Therefore, to compute the p-value for an observed score change $|C(\mathbf{x}_0) - C(\mathbf{x}^a_0)|$, we can pick a value Δc close to $|C(\mathbf{x}_0) - C(\mathbf{x}^a_0)|$, and set θ by solving $\Delta c = \frac{d}{d\theta} \log H^a(\theta)$. For computing p-values for score changes at all SNPs, we implement the following algorithm:

1. Group the difference in affinity scores into different sets, $\mathcal{G}_1, \cdots, \mathcal{G}_K$, such that the scores within each set is close to each other and Δc_k, $k = 1, \cdots, K$.

   a. Set $p_k = 0.1, \cdots, 0.9, 0.91, 0.92, \cdots, 0.99$ for $1 \le k \le K = 18$.

   b. Set Δc_k as the $100 \times p_k$ percentile of the observed scored differences across all SNPs.

   c. Set $\mathcal{G}_k$ as the score set in the range $((\Delta c_{k-1} + \Delta c_k)/2, (\Delta c_k + \Delta c_{k+1})/2]$ for $2 \le k \le K-1$, $\mathcal{G}_1$ as the score set in the range $(-\infty, (\Delta c_1 + \Delta c_2)/2]$, $\mathcal{G}_K$ as the score set in the range $((\Delta c_{K-1} + \Delta c_K)/2, \infty)$.

2. For each set of score differences $\{|C(\mathbf{x}^i_0) - C(\mathbf{x}^{ai}_0)|\}$ in $\mathcal{G}_k$, $k = 1, \cdots, K$:

   a. Set $\theta : \frac{d}{d\theta} \log H^a(\theta) = \Delta c_k$. Calculate $H^a(\theta)$.

   b. To set the Monte-Carlo sample size B, first, calculate B' as the integer part of $100(1-p_k)/p_k$. If $B' > 10^5$, set $B = 10^5$; if $B' < 2000$, set $B = 2000$; otherwise, set $B = B'$.

   c. Simulate B Monte-Carlo samples $(\mathbf{x}_b, \mathbf{x}^a_b, s_b)$ from distribution g^a_θ. Compute $C(\mathbf{x}_b)$, $C(\mathbf{x}^a_b)$.

   d. Estimate the score test p-value for each pair $(\mathbf{x}^i_0,\ \mathbf{x}^{ai}_0)$ by

$$\begin{aligned} &\widehat{pval_d}(\mathbf{x}^i_0, \mathbf{x}^{ai}_0) = \\ &\quad \frac{1}{B} \sum_{b=1}^{B} 1\{|C(\mathbf{x}_b) - C(\mathbf{x}^a_b)| \ge |C(\mathbf{x}^i_0) - C(\mathbf{x}^{ai}_0)|\} \\ &\quad \cdot \frac{h^a_\theta(\mathbf{x}_b, \mathbf{x}^a_b)}{f^a(\mathbf{x}_b, \mathbf{x}^a_b)}. \end{aligned}$$

   e. For the rank tests p-values, first compute the allele-specific p-value for each Monte-Carlo observation by

$$pval_b = \frac{1}{B} \sum_{b'=1}^{T} 1\{C(\mathbf{x}_{b'}) \ge C(\mathbf{x}_b)\} \cdot \frac{h^a_\theta(\mathbf{x}_b, \mathbf{x}^a_b)}{f^a(\mathbf{x}_b, \mathbf{x}^a_b)}.$$

$$pval^a_b = \frac{1}{B} \sum_{b'=1}^{T} 1\{C(\mathbf{x}_{b'}) \ge C(\mathbf{x}^a_b)\} \cdot \frac{h^a_\theta(\mathbf{x}_b, \mathbf{x}^a_b)}{f^a(\mathbf{x}_b, \mathbf{x}^a_b)}.$$

   f. Estimate the rank test p-value for each pair $\mathbf{x}^i_0$, $\mathbf{x}^{ai}_0$ by

$$\begin{aligned} &\widehat{pval_r}(\mathbf{x}^i_0, \mathbf{x}^{ai}_0)) = \\ &\quad \frac{1}{B} \sum_{b=1}^{B} 1 \Bigg[\Bigg\{ \big| \log(pval_b) - \log(pval^a_b) \big| \ge \\ &\quad \big| \log(\widehat{pval}(C(\mathbf{x}^i_0))) - \log(\widehat{pval}(C(\mathbf{x}^{ai}_0))) \big| \Bigg\} \\ &\quad \cdot \frac{h^a_\theta(\mathbf{x}_b, \mathbf{x}^a_b)}{f^a(\mathbf{x}_b, \mathbf{x}^a_b)} \Bigg]. \end{aligned}$$

3 COMPUTATIONAL DETAILS

3.1 Details for allele-specific tests

To compute $H(\theta)$, we first note that $H(\theta) = \sum_{s=1}^{L} H_s(\theta)$, where

$$H_s(\theta) = \sum_{\mathbf{x}\in\{1,2,3,4\}^{2L-1}} \pi(x_1) \prod_{l=1}^{2L-2} p(x_l, x_{l+1}) \prod_{l=1}^{L} W(x_{l+s-1}, l)^{\theta}.$$

We use the recursive algorithm in Chan *et al.* (2010) to compute $H_s(\theta)$. Let V be a $4 \times (2L-1)$ matrix, with $V(i,l) = W(i, l - s + 1)^{\theta}$ for $l = s, \cdots, s + L - 1$ and the rest of the entries set as 1. Then,

$$H_s(\theta) = \sum_{\mathbf{x}\in\{1,2,3,4\}^{2L-1}} \pi(x_1) \prod_{l=1}^{2L-2} p(x_l, x_{l+1}) \prod_{l=1}^{2L-1} V(x_l, l), \tag{10}$$

can be computed by the following recursion:

$$\begin{aligned}
Q_s(i, 2L-1)) &= V(i, 2L-1),\ 1 \le i \le 4; && (11)\\
Q_s(i,l) &= V(i,l)\sum_{j=1}^{4} p(i,j) Q_s(j, l+1), && (12)\\
& \quad 1 \le l \le 2L-2,\ 1 \le i \le 4; &&\\
H_s(\theta) &= \sum_{i=1}^{4} \pi(i) Q_s(i,1). && (13)
\end{aligned}$$

Finally, $(\mathbf{X}, S) \sim g_\theta$ can be simulated as follows:

$$\begin{aligned}
P(S = s) &= \frac{H_s(\theta)}{H(\theta)}; && (14)\\
P(X_1 = x_1 | S = s) &= \frac{\pi(x_1) Q_s(x_1, 1)}{H_s(\theta)}; && (15)\\
P(X_l = x_l | X_{l-1} = x_{l-1}, S = s) &= \frac{p(x_{l-1}, x_l) Q_s(x_l, l)}{Q_s(x_{l-1}, l-1)}, && (16)\\
& \quad 2 \le l \le 2L-1. &&
\end{aligned}$$

3.2 Details for tests of change in affinity scores between the alleles

To compute $H^a(\theta)$, we first note that $H^a(\theta) = \sum_{s=1}^{L} H_s^a(\theta)$, where

$$\begin{aligned}
H_s^a(\theta) = &\sum_{\mathbf{x}\in\{1,2,3,4\}^{2L-1}} \Big\{ f(x_1, \cdots, x_{s-1}) f(x_{s+L}, \cdots, x_{2L-1})\\
&\Big[\prod_{1\le l\le L, l\ne L} IW(x_{l+s-1}, l)\Big] D(x_L, L - s + 1)^{\theta} \Big\}\\
= &\sum_{\{x_s, \cdots, x_{s+L-1}\}\in\{1,2,3,4\}^L} \Big\{ D(x_L, L - s + 1)^{\theta}\\
&\Big[\prod_{1\le l\le L, l\ne L} IW(x_{l+s-1}, l)\Big] \Big\}\\
= &\Big\{ \prod_{1\le l\le L, l\ne L-s+1} \Big[\sum_{i=1}^{4} IW(i,l)\Big] \Big\} \Big[\sum_{i=1}^{4} D(i, L - s + 1)^{\theta}\Big]\\
= &\sum_{i=1}^{4} D(i, L - s + 1)^{\theta}.
\end{aligned}$$

A sequence following g_θ^a can be simulated as follows.

$$\begin{aligned}
P(S = s) &= \frac{H_s^a(\theta)}{H^a(\theta)}, && (17)\\
P(X_l = x_l | S = s) &= \pi(x_l), && (18)\\
& \text{for} \quad l = 1, s + L, && (19)\\
P(X_l = x_l | S = s, X_{l-1} = x_{l-1}) &= p(x_{l-1}, x_l) && (20)\\
\text{for} \quad l = 2, \cdots, s-1, s+L+1, \cdots, 2L-1, & && (21)\\
P(X_l = x_l) &= IW(x_l, l - s + 1) && (22)\\
\text{for} \quad l = s, \cdots, L-1, L+1, \cdots, s+L-1, & && (23)\\
P(X_L = x_L) &= \frac{D(x_L, L - s + 1)^{\theta}}{H_s^a(\theta)} && (24)
\end{aligned}$$

4 NUMERICAL EVALUATIONS

In this section, we first compare the conditional p-values from atSNP with the p-values from FIMO (Grant *et al.*, 2011) to evaluate the accuracy of atSNP p-values that are based on importance sampling. Next, we compare the results for the evaluation of the binding affinity changes from atSNP and is-rSNP. We then apply atSNP's between allele affinity score change test to a set of rSNPs with known SNP-TF interactions from the ORegAnno database (Griffith *et al.*, 2008). All the analysis are based on hg19 version of the human genome.

4.1 Comparison with FIMO

To assess the computation accuracy of atSNP, we compared atSNP's conditional p-values with FIMO's p-values using the set of 26,100 SNPs from the Psychiatric Genomics Consortium (`http://www.med.unc.edu/pgc`) and the ENCODE-derived PWM for an arbitrarily chosen TF ATF3 [1] (Kheradpour and Kellis, 2013). Figure

[1] ATF3_GM12878_encode-Myers_seq_hsa_r1:MDscan#1#Intergenic.

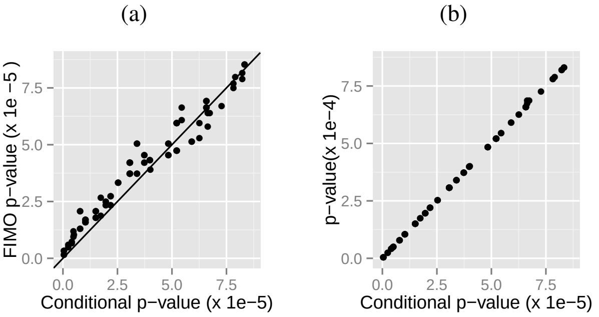


Fig. 4. (a) Comparison between FIMO's p-values and atSNP's conditional p-values. (b) Comparison between atSNP's conditional p-values and p-values.

4(a) compares FIMO p-values of all SNPs with a p-value less than 1e-4 (default threshold of FIMO [2]) with the conditional p-values from atSNP and indicates that the two sets of p-values agree well. Furthermore, for the SNPs with FIMO p-values larger than 1e-4, conditional p-values from atSNP were also larger than 1e-4. This suggests that our importance sampling algorithm is indeed speeding up the computations without sacrificing accuracy. Similar conclusions are obtained when we utilize other TFs instead of ATF3. Because the allele-specific affinity tests are an intermediate step in is-rSNP and are not included in the output, we were not able to compare their results with our conditional p-values.

We also compared atSNP p-values with its conditional p-values in Figure 4(b). We observe that the difference between the two p-value types are quite apparent at large affinity score values.

4.2 Comparison with is-rSNP

We used comparison with FIMO as a way of validating the accuracy of our importance sampling algorithm. Next, we compared the tests for affinity score changes between atSNP and is-rSNP. Since is-rSNP does not support batch processing large SNP sets[3], we compared atSNP and is-rSNP using one SNP from Section 4.1, namely rs9909429, as a representative case and utilized the PWMs from the JASPAR database[4]. is-rSNP reported the p-values adjusted by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). As a comparison, we adopted the same procedure to adjust rank test p-values and thresholded the adjusted p-values at 0.05 for both methods.

Table 1 lists the 16 motifs identified by atSNP and/or is-rSNP. Five of these motifs are identified as significantly affected by the SNP by both atSNP and is-rSNP. atSNP and is-rSNP assigns different significance on the effect of the SNP for the other 11 motifs. These discrepancies can be attributed to multiple factors. First, is-rSNP seems to compute the affinity score using Eqn. (2) even when the entries of the PWM are in the form of nucleotide probabilities, while atSNP applies the definition in Eqn. (1), which corresponds to log likelihood of the sequence when the PWM is in the form of nucleotide probabilities. Second, is-rSNP evaluates the change in the binding affinity by first scoring the reference and SNP allele versions of the subsequences overlapping the SNP and identifying the subsequence with the maximum affinity score, i.e., best match might be achieved with the reference or the SNP allele. Then, it compares the affinity scores of this subsequence with both the reference and the SNP allele. This approach overlooks the possibility that both the reference and SNP alleles may provide equally good matches to the PWM, albeit with subsequences starting at different positions. Figure 1 of the main text provides an illustrative example of this scenario. Here, is-rSNP chooses the subsequence starting at the 2nd position in the reference genome as the best match to the PWM. Then, it evaluates the binding affinity of this subsequence with the SNP allele and obtains a big change in the affinity score. However, as is visible from the logo, there is an almost equally good match to the PWM starting at the 1st position of the sequence with the SNP allele. Clustered degenerate binding sites are especially susceptible to these types of potential false positives (Zhang *et al.*, 2006). A third source of discrepancy is that is-rSNP assumes an independent multinomial model for the background distribution whereas atSNP accommodates dependency between consecutive positions motivated by the fact that modeling dependency between the positions of the background sequences improve motif detection (Thijs *et al.*, 2001).

We present the composite sequence logo plots comparing the reference and SNP alleles for all the commonly identified SNP-PWM pairs in Section 5.1, for SNP-PWM pairs only identified by atSNP in Section 5.2, and by is-rSNP in Section 5.3. We observe that all the commonly identified motifs have very good matching subsequences with either the reference or the SNP allele, and the SNP significantly impacts the binding affinity. Motifs prioritized only by one method typically have a number of mismatches to the motif consensus in their best matching subsequence around the SNP in addition to the mismatch at the SNP position. Overall, motifs prioritized by atSNP seem to have slightly better matching subsequences to the motif with either the reference or the SNP allele. On average, the proportions of positions that do not agree with the motif consensus are 0.14 and 0.27 for atSNP and is-rSNP, respectively. These proportions are obtained by counting the mismatches between the best matching subsequences and the most likely consensus sequences from the PWMs by discarding the degenearte positions that are on either edges of the PWM.

We further observe that many of the significant PWMs are very similar to each other (e.g., CN0007.1, CN0002.1, MA0139.1, PF0045.1 are variants of Ctcf PWM) indicating that the hypotheses evaluated within the multiple testing framework are far from independent. This suggests that the classical multiple testing procedures adopted by FIMO and is-rSNP, i.e., Benjamini-Hochberg FDR procedure (Benjamini and Hochberg (1995)) and Storey's q-value (Storey (2003)), can be overly conservative. One possible remedy for this is to adopt group false discovery rate procedure proposed by Hu *et al.* (2010). However, its implementation requires additional considerations such as appropriate grouping within the PWM libraries. For these reasons, atSNP currently does not support a built-in multiple testing adjustment method. We suggest using the commonly

[2] FIMO run without any thresholding did not complete within 24 hours.

[3] All versions of is-rSNP (1.0 and 2.0) can only analyze at most 20 SNPs at a time

[4] We used the latest is-rSNP version 2.0. and found that is-rSNP uses 2010 freeze of the JASPAR database. In order to make our results comparable, we also used this version of the JASPAR database in Sections 4.2 and 4.3.

Motif	Motif Info	$pval_r$-BH	$pval_{adj}$
rSNPs identified by both atSNP and is-rSNP			
CN0007.1	LM7	2.4e-4	1.8e-6
CN0002.1	LM2	3.6e-4	4e-6
MA0139.1	CTCF	4.5e-4	2.1e-5
PF0045.1	CCANNAGRKGGC	1.6e-3	5.7e-5
MA0055.1	MYF	0.044	9.9e-4
rSNPs identified only by atSNP			
PF0057.1	ACCTGTTG	0	1
CN0023.1	LM23	5.8e-4	1
PL0011.1	HLH-2::HLH-4	1.9e-3	1
PL0002.1	HLH-2::HLH-3	2.1e-3	1
CN0146.1	LM146	0.0032	0.63
CN0047.1	LM47	4.4e-3	0.383
rSNPs identified only by is-rSNP			
MA0322.1	INO4	0.128	0.042
PL0017.1	HLH-2::HLH-10	0.22	5.3e-3
CN0049.1	LM49	0.252	9.4e-3
CN0194.1	LM194	0.267	8.1e-3
CN0169.1	LM169	0.482	0.028

Table 1. rSNP interactions of SNP rs9909429 identified by atSNP and is-rSNP. '$pval_r$-BH' reports the rank test p-values of atSNP adjusted by the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) whereas '$pval_{adj}$' reports the BH adjusted p-values from is-rSNP.

adapted conservative procedures already available by `R` functions `stats::p.adjust` and `qvalue::qvalue`.

4.3 Validation using known rSNP-TF interactions

The ORegAnno database (Griffith *et al.*, 2008) lists 36 known rSNP-TF pairs. We analyzed each SNP with the PWMs of the JASPAR motifs (Mathelier *et al.*, 2013) that came from the TF family of each TF in the rSNP-TF pair. We used the TF families defined as the homolog clusters according to (Fulton *et al.*, 2009). For 9 of the SNPs, neither the reference nor the SNP allele matched the allele listed at the SNP location in hg19 version of the genome and, hence, we discarded these from the analysis. We further discarded 3 rSNPs for which the associated TF families were not included among the JASPAR motifs. For each of the remaining 24 SNPs, we ran atSNP against the JASPAR PWMs to identify the motif with the most significant regulatory effect both among the whole set of motifs and among motifs from the associated TF family (Table 2). atSNP successfully identified motifs from the ORegAnno-reported TF family for 20 of the SNPs based on the rank test p-value (at significance level of 0.05).

Table 2 shows the ranks for the top motifs from the ORegAnno-reported TF family among the entire set of JASPAR motifs. Because we evaluated the SNPs against the whole set of JASPAR motifs, we were able to calculate the ranking of the motifs from ORegAnno-reported TF families. For all the known 20 rSNPs, these motifs fall in the top 5% among the 1192 JASPAR motifs. However, none of the top ranked motifs is from the PWMs of the reported TF family. This does not indicate that atSNP results are inconsistent with the ORegAnno database; it is possible that these more significant SNP-TF interactions may not have been studied experimentally or cannot be further discriminated based on sequence alone. For example, for the rs2251746-GATA1 pair, the most significant score change within the GATA family is obtained with GATA3 PWM and ranked as the 35th most significant change among the whole set of JASPAR motifs (Figure 40(a)); however, when rs2251746 is evaluated against the whole set of PWMs in the JASPAR library, PWM for TOS8 is reported as exhibiting the most significant change in the affinity score (Figure 40(b)). When we visualize the sequence logo plots for these two PWMs, we observe that both changes seem significant, and rs2251746 is disrupting a match to the longer TOS8 motif.

atSNP provides a way to prioritize putative SNP-TF interactions and these interactions can further be filtered by other functional data such as ChIP-seq data of transcription factors from ENCODE or other consortia projects. We display the composite sequence logo plots for the SNP-TF pairs in Table 2 in Section 5.4. These plots directly illustrate how each SNP affects the binding pattern of the corresponding motif. For each SNP, the top motif in the library always has an almost perfect match to a sequence around the SNP location, while the SNP location is matched to a nucleotide that significantly changes the affinity score. Such patterns indicate strong *in silico* evidence for the regulatory effects.

We also analyzed these set of SNPs with is-rSNP (Table 3). Figure 5 displays the atSNP and is-rSNP ranks of the top motifs in the ORegAnno-reported TF family for each SNP across all the JASPAR PWMs. Overall, the median rank of the highest ranked motifs in the ORegAnno-reported TF family is 10.5 for atSNP and 20 for is-rSNP across all the JASPAR PWMs.

4.4 Run-time comparisons

Table 4 presents an illustrative summary of run time comparisons.

ORegAnno-reported		Top motif in the ORegAnno-reported TF family				Top motif in the JASPAR library			Dist
SNP ID	TF	Motif	Motif Info	$pval_r$	Rank	Motif	Motif Info	$pval_r$	
rs2569190	SP FAMILY	PB0075.1	SP4	8.6e-06	2	MA0381.1	SKN7	0	0.4436
rs763110	CEBPB	MA0102.1	CEBPA	2.4e-04	4	MA0327.1	MATA1	0	0.3223
rs12720461	ETS	MA0062.2	GABPA	3.7e-04	4	MA0275.1	ASG1	0	0.4372
rs28095	SP1/SP3	MA0079.1	SP1	6.6e-04	4	POL010.1	DCE_S_III	0	0.2911
rs712829	SP1	PB0075.1	SP4	7.1e-04	4	MA0373.1	RPN4	2.0e-04	0.427
rs13434811	YY1	PB0097.1	ZFP410	7.4e-04	9	MA0035.2	GATA1	2.3e-05	0.3081
rs16998970	YY1	PB0097.1	ZFP410	7.6e-04	10	CN0095.1	LM95	3.7e-05	0.2976
rs1800775	SP1/SP3	PB0025.1	GLIS2	0.0012	2	PF0056.1	GGGTGGRR	7.9e-04	0.4139
rs243865	SP1	MA0039.2	KLF4	0.0018	4	PF0082.1	CTGYNNCTYTAA	6.2e-04	0.3789
rs934345	TP53	MA0106.1	TP53	0.0022	11	MA0217.1	CAUP	0	0.4177
rs2333227	SP1	PB0096.1	ZFP281	0.0027	15	PL0002.1	HLH-2::HLH-3	4.1e-04	0.4154
rs213045	E2F2	MA0024.1	E2F1	0.0028	5	MA0334.1	MET32	3.6e-04	0.3927
rs1800590	SP1/SP3	PB0051.1	PLAGL1	0.0029	8	MA0381.1	SKN7	0	0.425
rs1862513	SP1/SP3	PB0096.1	ZFP281	0.0031	8	MA0366.1	RGM1	0	0.4312
rs2838769	TP53	MA0106.1	TP53	0.0034	11	MA0233.1	MIRR	0	0.4013
rs27646	SP1	MA0163.1	PLAG1	0.0039	12	MA0410.1	UGA3	0	0.3745
rs2227306	CEBPB	MA0102.2	CEBPA	0.0133	16	PF0172.1	TTGCWCAAY	0.0023	0.2776
rs1658728	TP53	MA0106.1	TP53	0.0164	41	PB0040.1	MAFB	4.4e-04	0.4148
rs2251746	GATA1	MA0037.1	GATA3	0.0259	35	MA0408.1	TOS8	9.4e-05	0.3306
rs2279744	SP1	MA0146.1	ZFX	0.0316	38	MA0185.1	DEAF1	0	0.3567
rs3761624	TP53	MA0106.1	TP53	0.0687	99	PF0106.1	CCGNMNNTNACG	3.3e-04	0.3426
rs268682	TP53	MA0106.1	TP53	0.0812	66	MA0260.1	CHE-1	0	0.4267
rs2232945	TP53	MA0106.1	TP53	0.0846	110	PF0134.1	CATRRAGC	0	0.3478
rs11836625	CREB1	MA0414.1	XBP1	0.1724	185	MA0130.1	ZNF354C	0	0.2894

Table 2. Affinity score change tests for the curated rSNP-TF pairs in the ORegAnno database (Griffith *et al.*, 2008) by atSNP. Columns 3-6 correspond to the top significant motif in the ORegAnno-reported TF family, while Columns 7-9 correspond to the top significant motif among all the 1192 motifs in the JASPAR database. 'Rank' is the rank of $pval_r$ for the top motif in the TF family across the whole motif library. 'Dist' is the L^2 distance between the top motifs in the TF family and in the whole library, normalized by the matrix size (R function implementing this distance is available at `http://www.stat.wisc.edu/~keles/Software/motif_distance.R`. When the two PWM have different sizes, 'Dist' is based on the submatrix of the larger PWM that minimizes the distance to the smaller PWM.

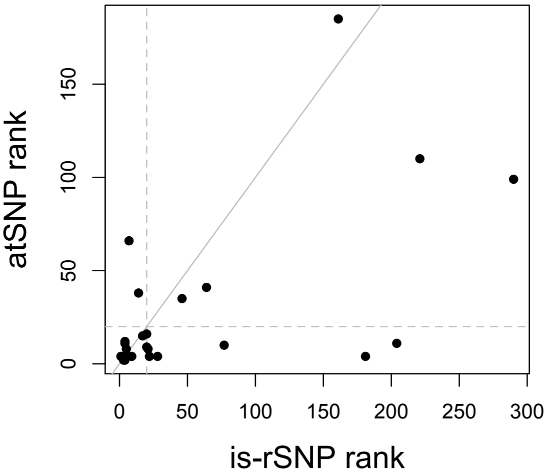


Fig. 5. atSNP and is-rSNP ranks of the top motifs in the ORegAnno-reported TF family for each SNP across all the JASPAR PWMs. Horizontal and vertical dashed lines mark rank 20.

SNP ID	ORegAnno-reported TF	Top motif in the ORegAnno-reported TF family			
		Motif	Motif Info	p-value	Rank
rs28095	SP1	MA0079.1	SP1	0	1
rs934345	TP53	MA0106.1	TP53	0	4
rs1800775	Glis2_1	PB0025.1	GLIS2	0	3
rs27646	Zfx	MA0146.1	ZFX	0	4
rs2569190	PLAG1	MA0163.1	PLAG1	1e-04	4
rs1862513	Osr2_1	PB0051.1	PLAGL1	1e-04	5
rs2333227	Zfp281_1	PB0097.1	ZFP410	1e-04	17
rs213045	E2F1	MA0024.1	E2F1	2e-04	5
rs243865	Egr1_1	PB0010.1	EGR1	2e-04	9
rs268682	TP53	MA0106.1	TP53	2e-04	7
rs13434811	Zfp410_1	PB0098.1	ZFP691	3e-04	20
rs2279744	PLAG1	MA0163.1	PLAG1	3e-04	14
rs2227306	Cebpa	MA0102.1	CEBPA	4e-04	20
rs1800590	Hic1_1	PB0029.1	HIC1	7e-04	21
rs12720461	ELF5	MA0136.1	ELF5	8e-04	28
rs712829	Zfp281_1	PB0097.1	ZFP410	9e-04	22
rs2251746	GATA2	MA0036.1	GATA2	0.0013	46
rs1658728	TP53	MA0106.1	TP53	0.0013	64
rs16998970	Zfp410_1	PB0098.1	ZFP691	0.0013	77
rs763110	XBP1	MA0414.1	XBP1	0.0038	181
rs2838769	TP53	MA0106.1	TP53	0.0059	204
rs2232945	TP53	MA0106.1	TP53	0.0065	221
rs11836625	CREB1	MA0018.2	CREB1	0.0066	161
rs3761624	TP53	MA0106.1	TP53	0.0095	290

Table 3. Affinity score change tests for the curated rSNP-TF pairs in the ORegAnno database (Griffith *et al.*, 2008) using is-rSNP. 'Rank' is the rank of 'p-value' for the top motif in the TF family across the whole motif library.

Method	# of SNPs	# of PWMs	# of cores	Total time	Time for reading in data	Time for writing data
atSNP	26,100	1	1	3m8s	26s	41s
atSNP	26,100	10	10	7m15s	25s	3m13s
atSNP	26,100	10	1	23m4s	25s	3m13s
FIMO	26,100	10	1	2h30m*		
atSNP	500	2,065	30	2h2m	2s	25s
atSNP	26,100	2,065	30	5h48m	26s	18m35s

Table 4. Run time evaluations of atSNP. *: only outputs results with p-value ≤ 0.1.

5 SUPPLEMENTARY FIGURES

5.1 Sequence logo plots for commonly identified SNP-PWM pairs in Table 1

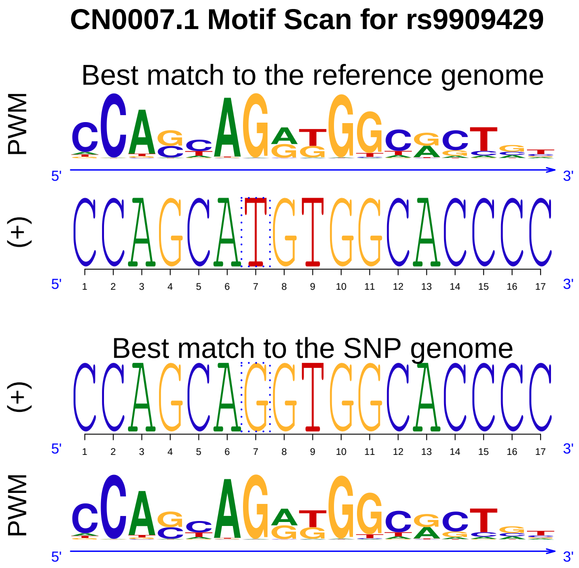


Fig. 6. Sequence logo plot for CN0007.1-rs9909429.

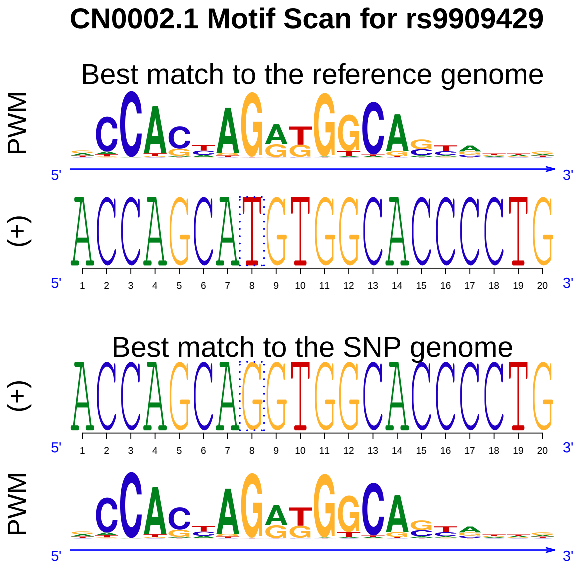


Fig. 7. Sequence logo plot for CN0002.1-rs9909429.

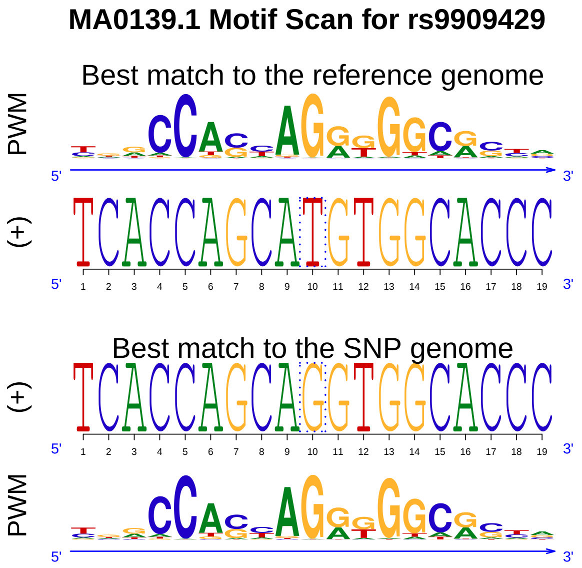


Fig. 8. Sequence logo plot for MA0139.1-rs9909429.

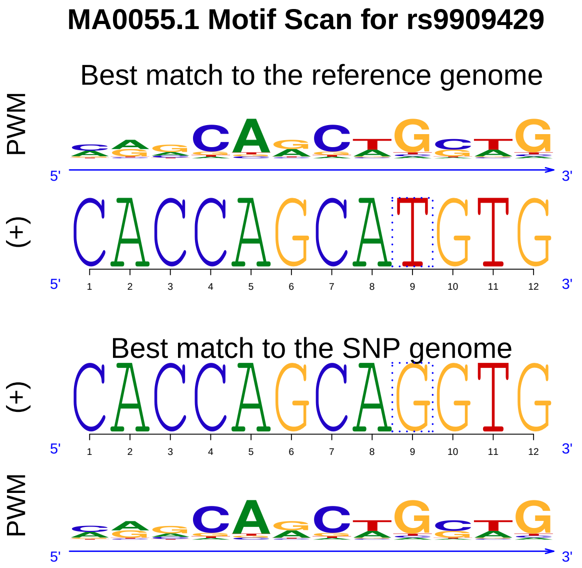


Fig. 10. Sequence logo plot for MA0055.1-rs9909429.

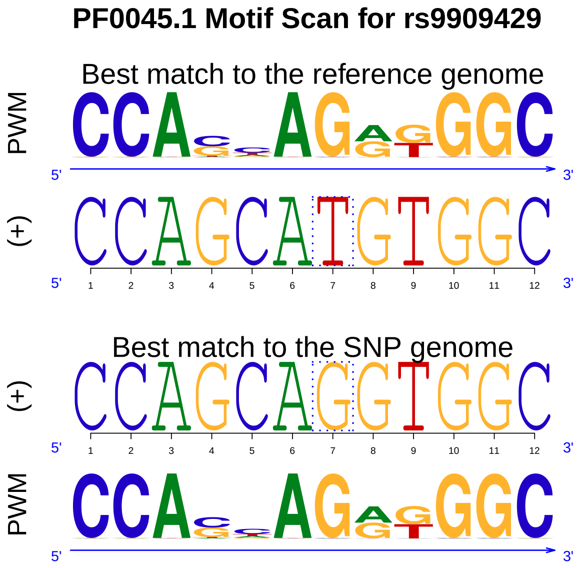


Fig. 9. Sequence logo plot for PF0045.1-rs9909429.

5.2 Sequence logo plots for SNP-PWM pairs identified only by atSNP in Table 1

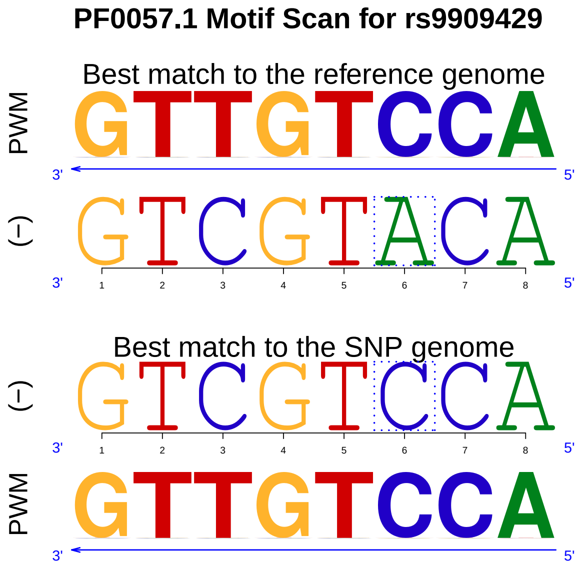


Fig. 11. Sequence logo plot for PF0057.1-rs9909429.

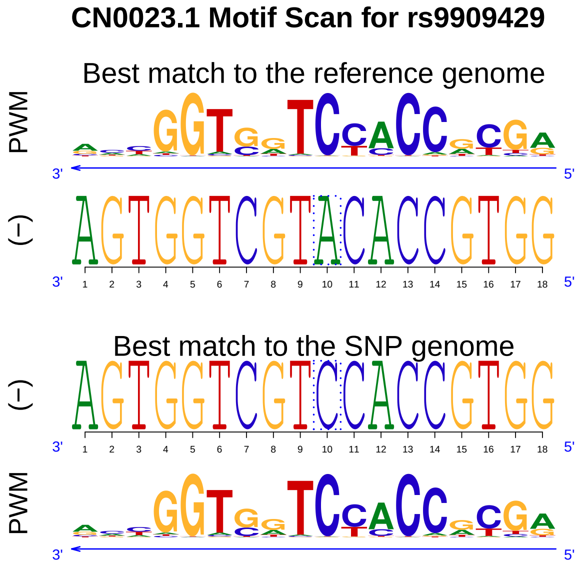


Fig. 12. Sequence logo plot for CN0023.1-rs9909429.

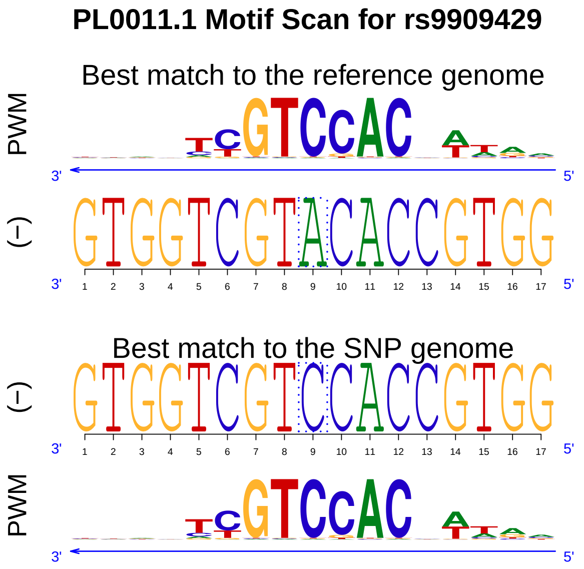


Fig. 13. Sequence logo plot for PL0011.1-rs9909429.

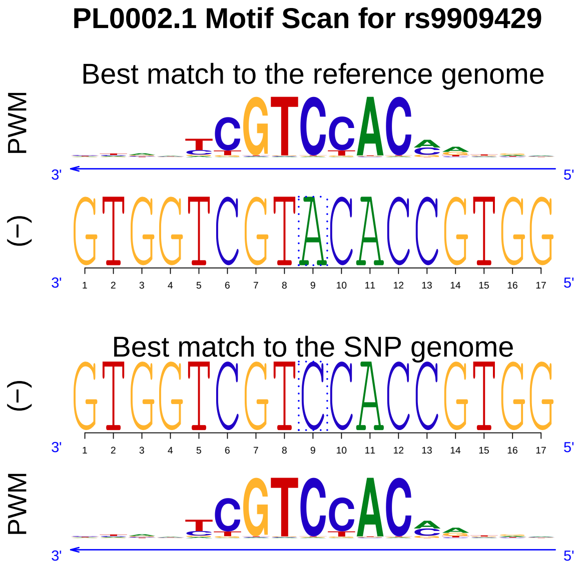


Fig. 14. Sequence logo plot for PL0002.1-rs9909429.

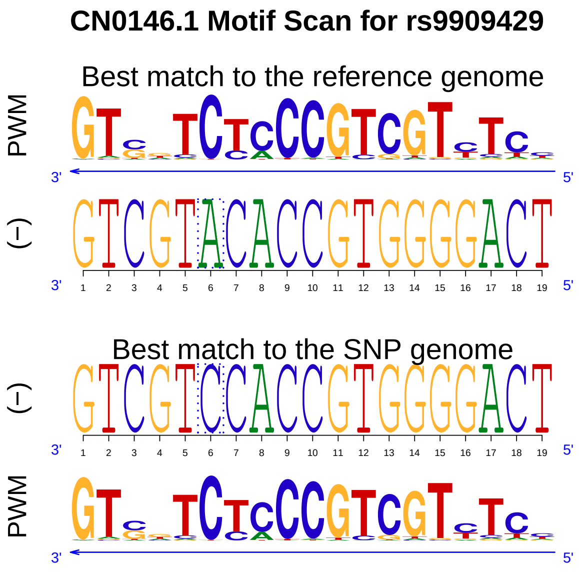


Fig. 15. Sequence logo plot for CN0146.1-rs9909429.

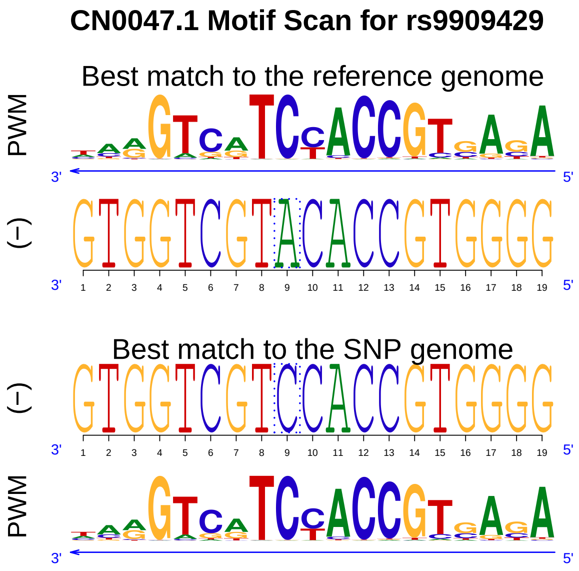


Fig. 16. Sequence logo plot for CN0047.1-rs9909429.

5.3 Sequence logo plots for SNP-PWM pairs identified only by is-rSNP in Table 1

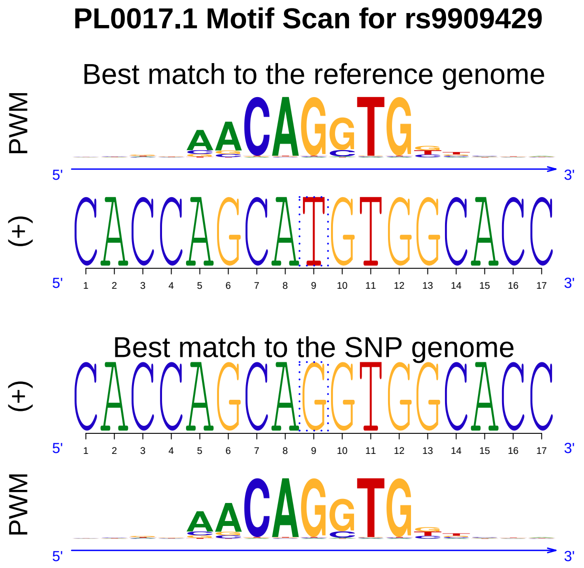


Fig. 17. Sequence logo plot for PL0017.1-rs9909429.

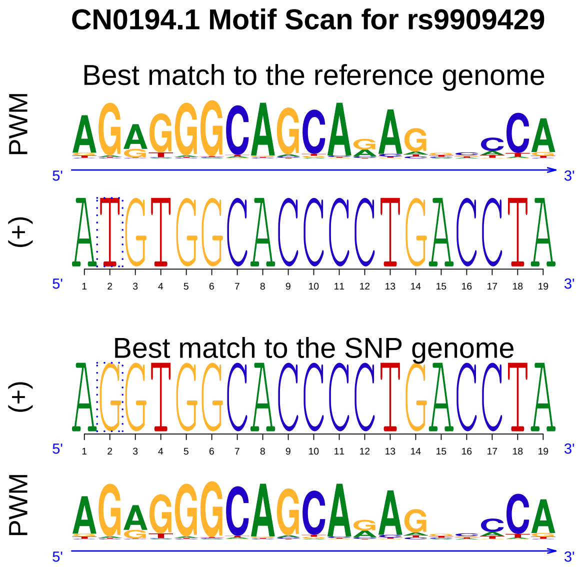


Fig. 18. Sequence logo plot for CN0194.1-rs9909429.

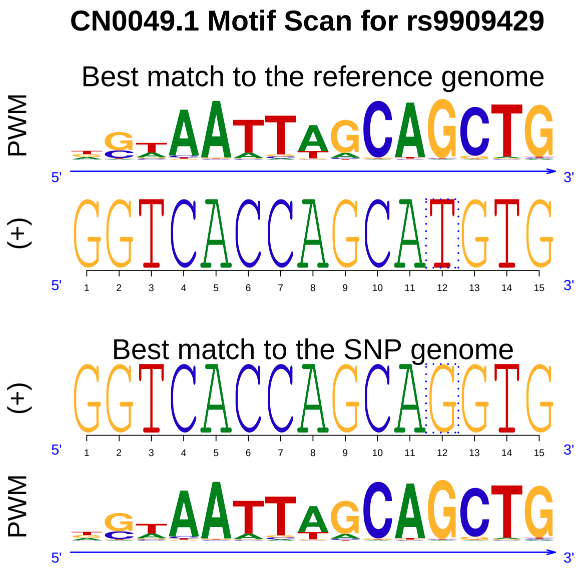


Fig. 19. Sequence logo plot for CN0049.1-rs9909429.

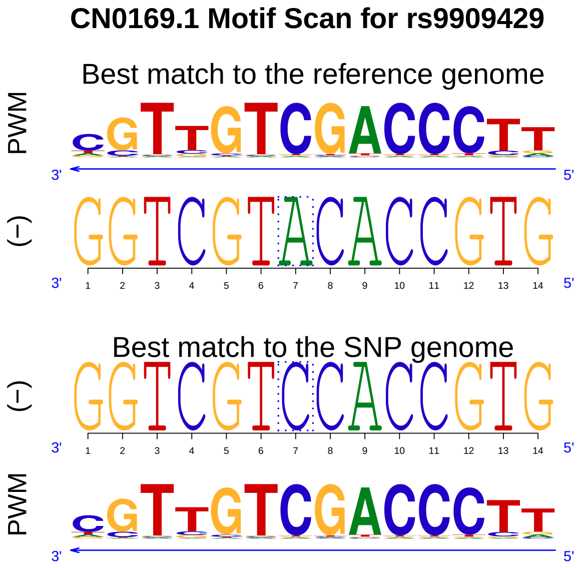


Fig. 20. Sequence logo plot for CN0169.1-rs9909429.

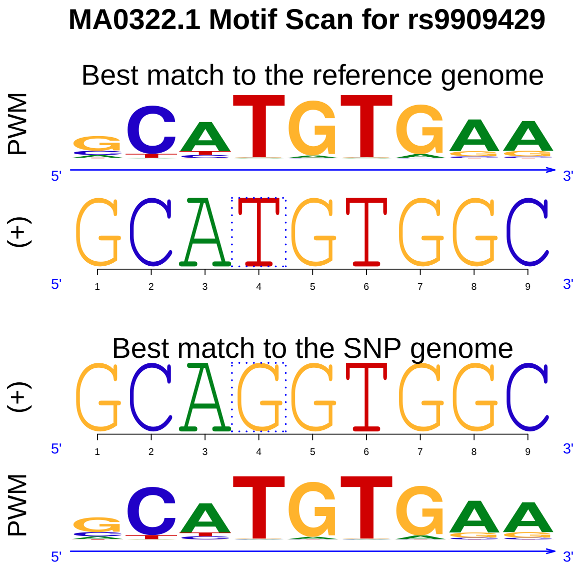
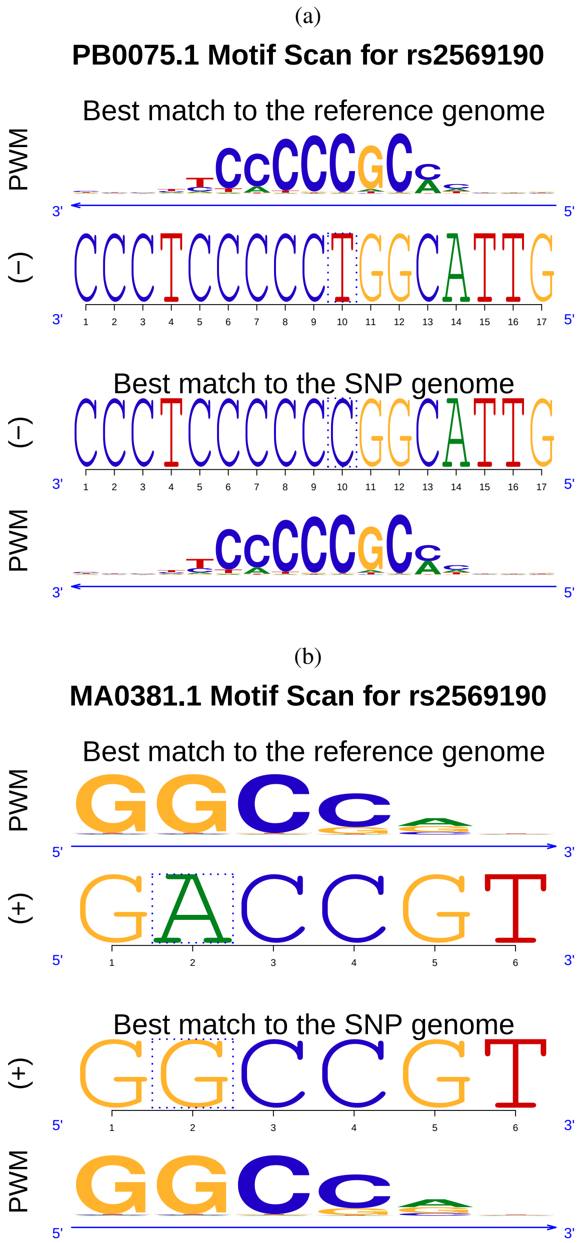

MA0322.1 Motif Scan for rs9909429
Best match to the reference genome
PWM
5'
3'
(+)
GCATGTGGC
1 2 3 4 5 6 7 8 9
Best match to the SNP genome
GCAGGTGGC
PWM

Fig. 21. Sequence logo plot for MA0322.1-rs9909429.

5.4 Sequence logo plots for SNP-PWM pairs in Table 2

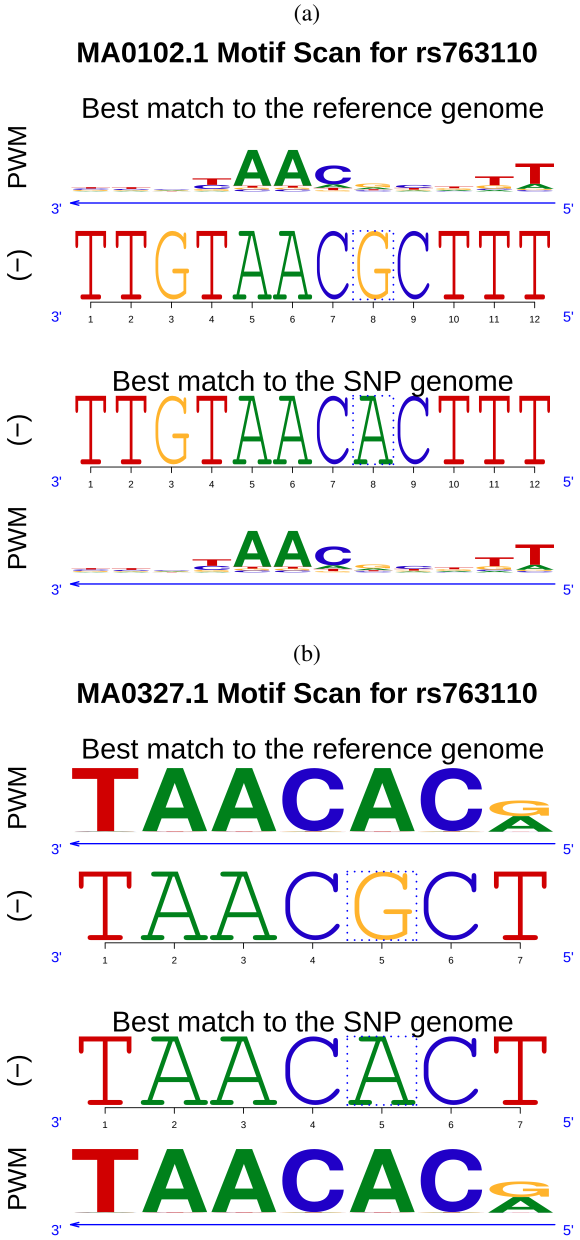

(a)

PB0075.1 Motif Scan for rs2569190

(b)

MA0381.1 Motif Scan for rs2569190

Fig. 22. Sequence logo plot for rs2569190 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

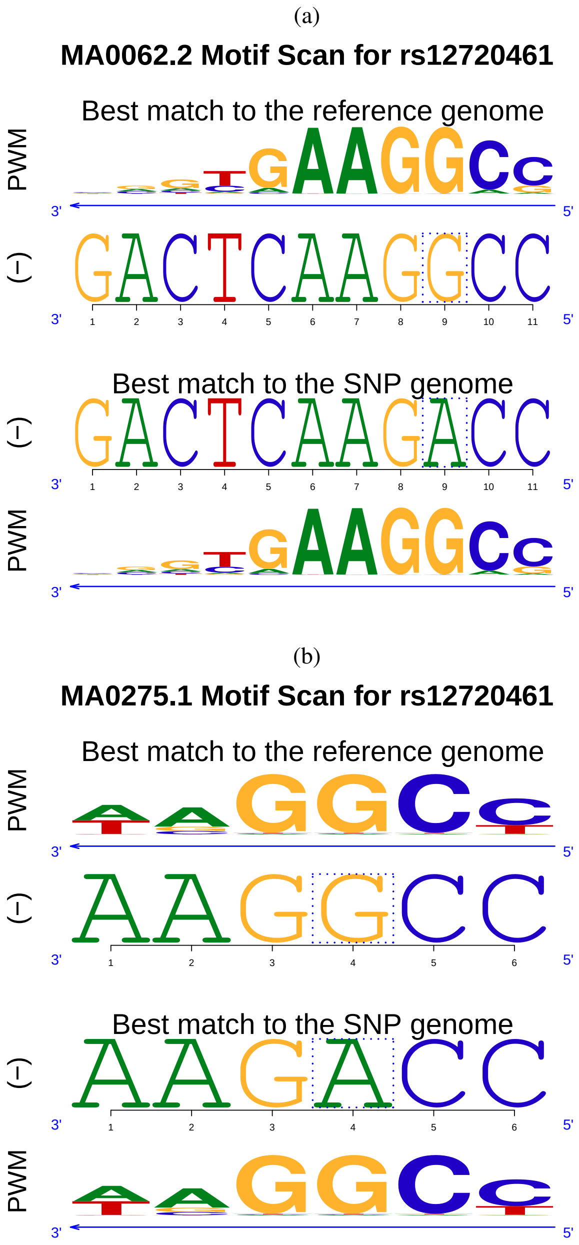

(a)

MA0102.1 Motif Scan for rs763110

(b)

MA0327.1 Motif Scan for rs763110

Fig. 23. Sequence logo plot for rs763110 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



(a)

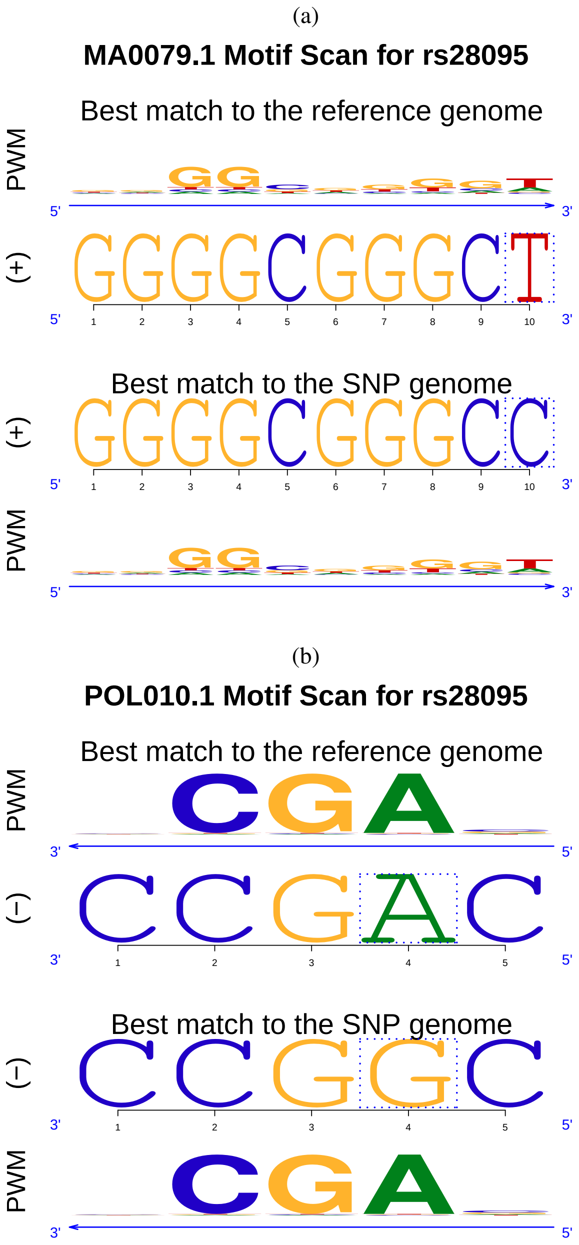
MA0062.2 Motif Scan for rs12720461

(b)

MA0275.1 Motif Scan for rs12720461

Best match to the reference genome
PWM
(−)
GACTCAAGGCC
Best match to the SNP genome
GACTCAAGACC
3'
5'
1
2
3
4
5
6
7
8
9
10
11
AAGGCC
AAGACC

Fig. 24. Sequence logo plot for rs12720461 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



(a)

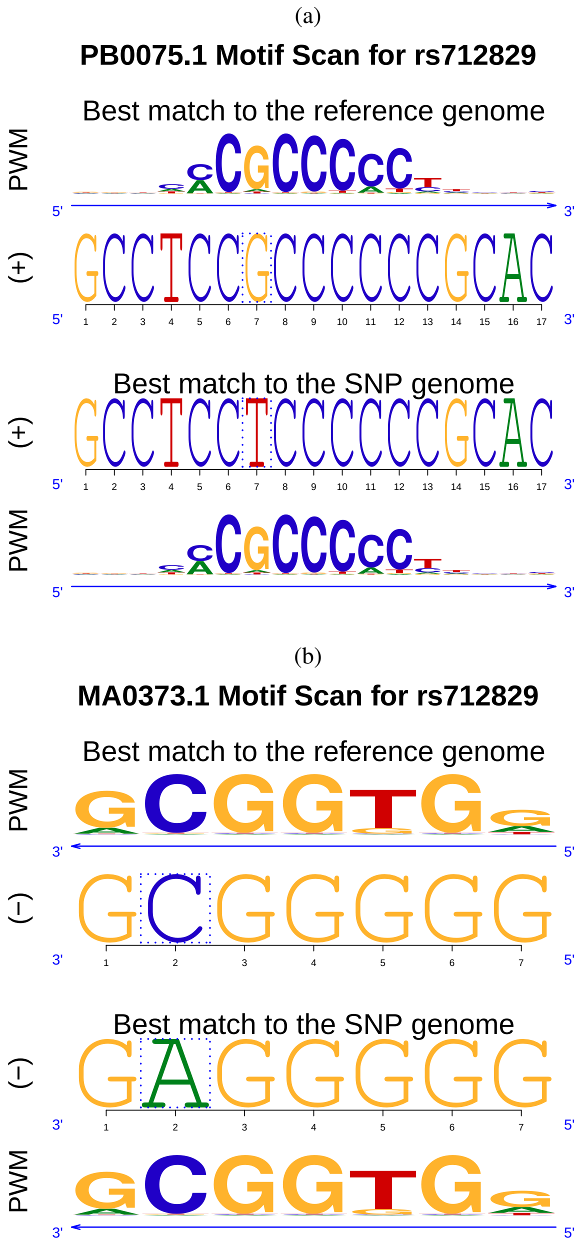
MA0079.1 Motif Scan for rs28095

(b)

POL010.1 Motif Scan for rs28095

Best match to the reference genome
PWM
(+)
GGGGCGGGCT
Best match to the SNP genome
GGGGCGGGCC
(−)
CCGAC
CCGGC
5'
3'
1
2
3
4
5
6
7
8
9
10

Fig. 25. Sequence logo plot for rs28095 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



(a)

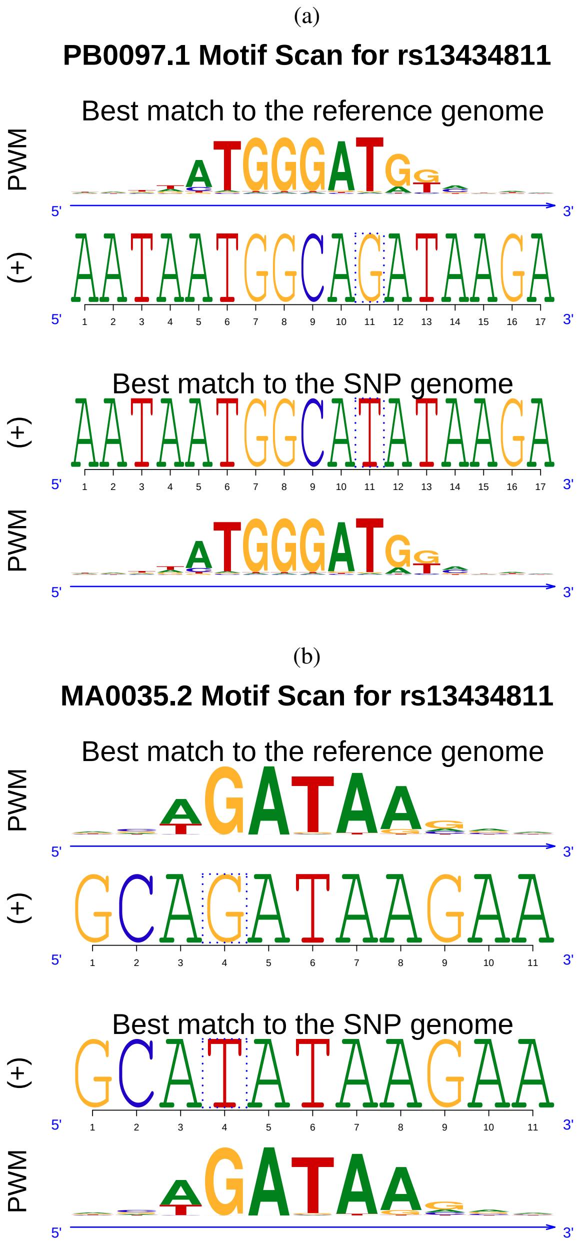
PB0075.1 Motif Scan for rs712829

(b)

MA0373.1 Motif Scan for rs712829

Best match to the reference genome
PWM
3'
5'
(−)
GCGGGGG
1 2 3 4 5 6 7
Best match to the SNP genome
(−)
GAGGGGG
PWM

Fig. 26. Sequence logo plot for rs712829 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



(a)

PB0097.1 Motif Scan for rs13434811

(b)

MA0035.2 Motif Scan for rs13434811

Best match to the reference genome
PWM
5'
3'
(+)
GCAGATAAGAA
1 2 3 4 5 6 7 8 9 10 11
Best match to the SNP genome
(+)
GCATATAAGAA
PWM

Fig. 27. Sequence logo plot for rs13434811 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

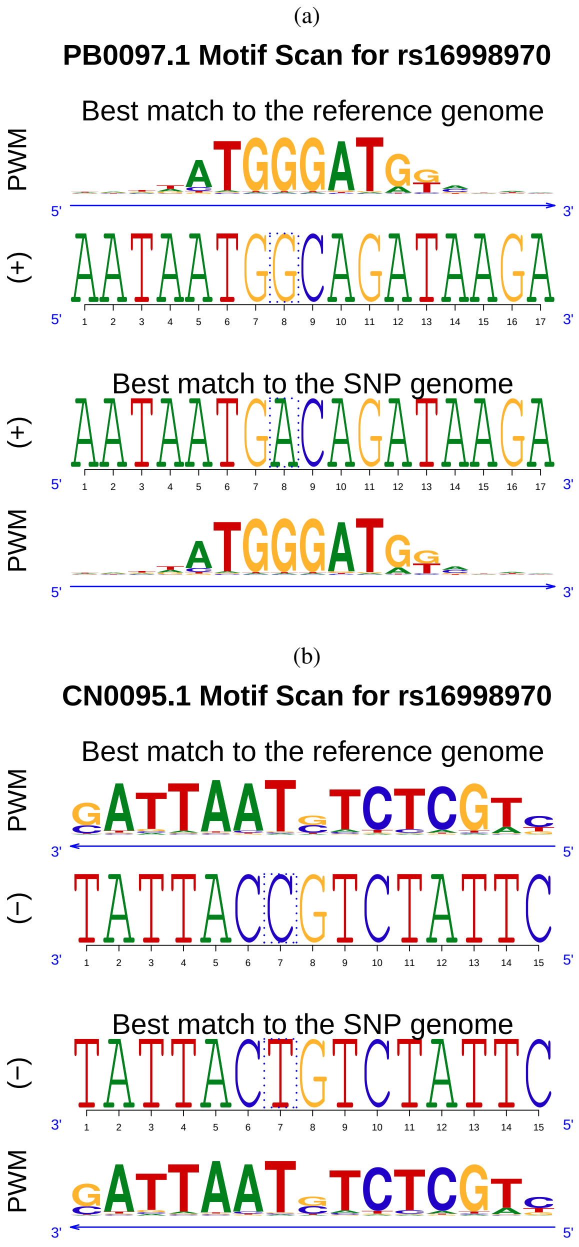


Fig. 28. Sequence logo plot for rs16998970 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

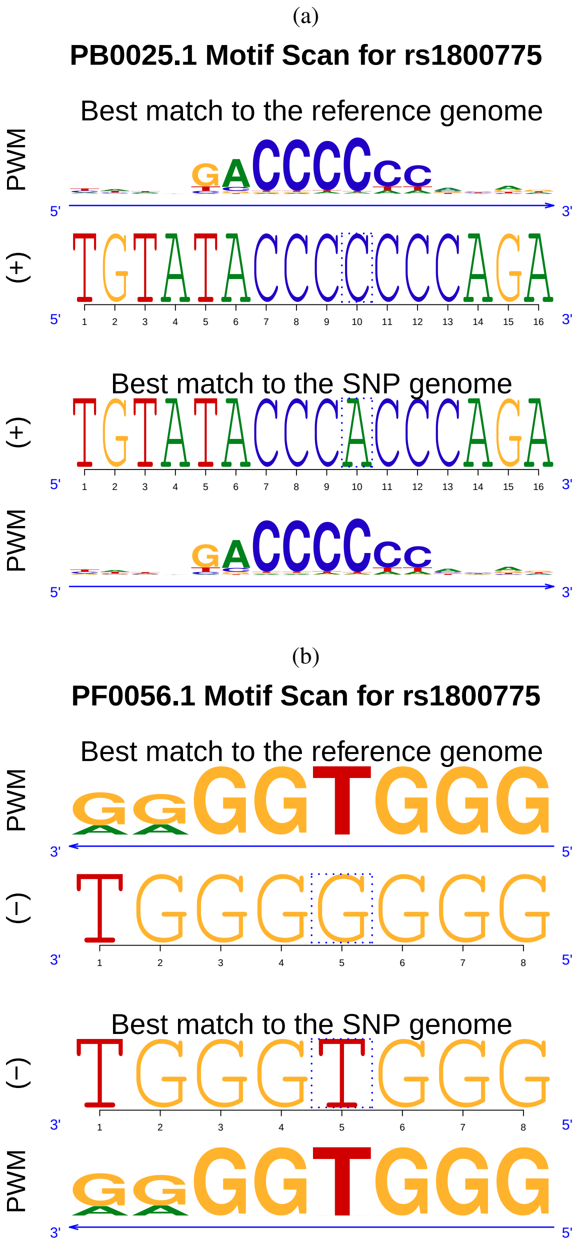
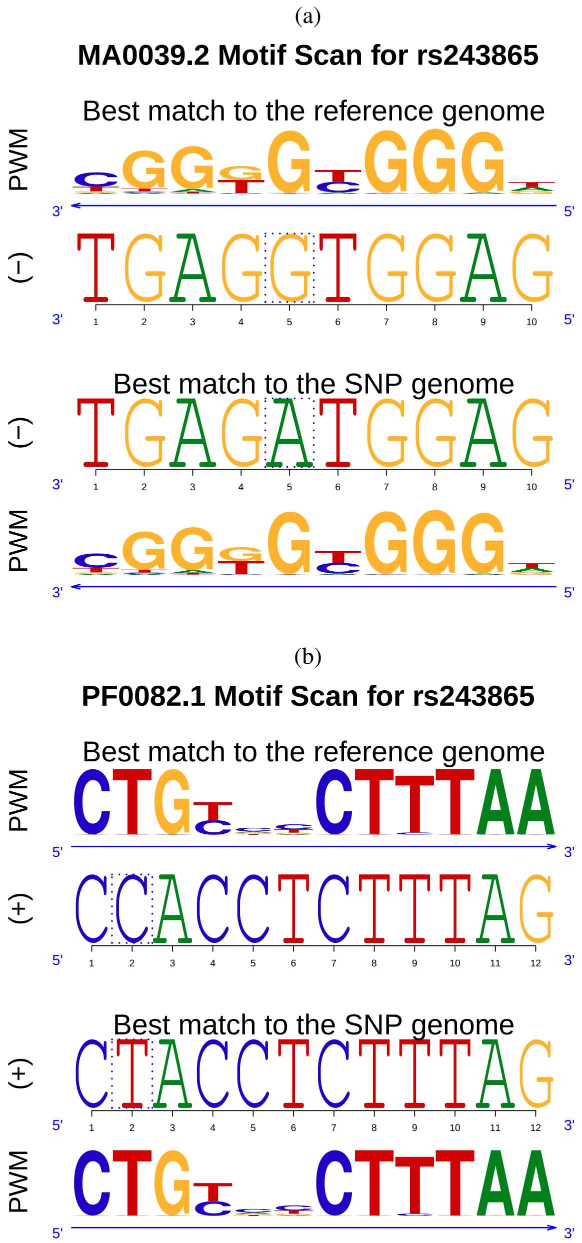

(a)
PB0025.1 Motif Scan for rs1800775
Best match to the reference genome
PWM
GACCCCCC
5'
3'
(+)
TGTATACCCCCCCAGA
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Best match to the SNP genome
(+)
TGTATACCCACCCAGA
PWM
(b)
PF0056.1 Motif Scan for rs1800775
Best match to the reference genome
PWM
GGGGTGGG
(−)
TGGGGGGG
1 2 3 4 5 6 7 8
Best match to the SNP genome
(−)
TGGGTGGG
PWM

Fig. 29. Sequence logo plot for rs1800775 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



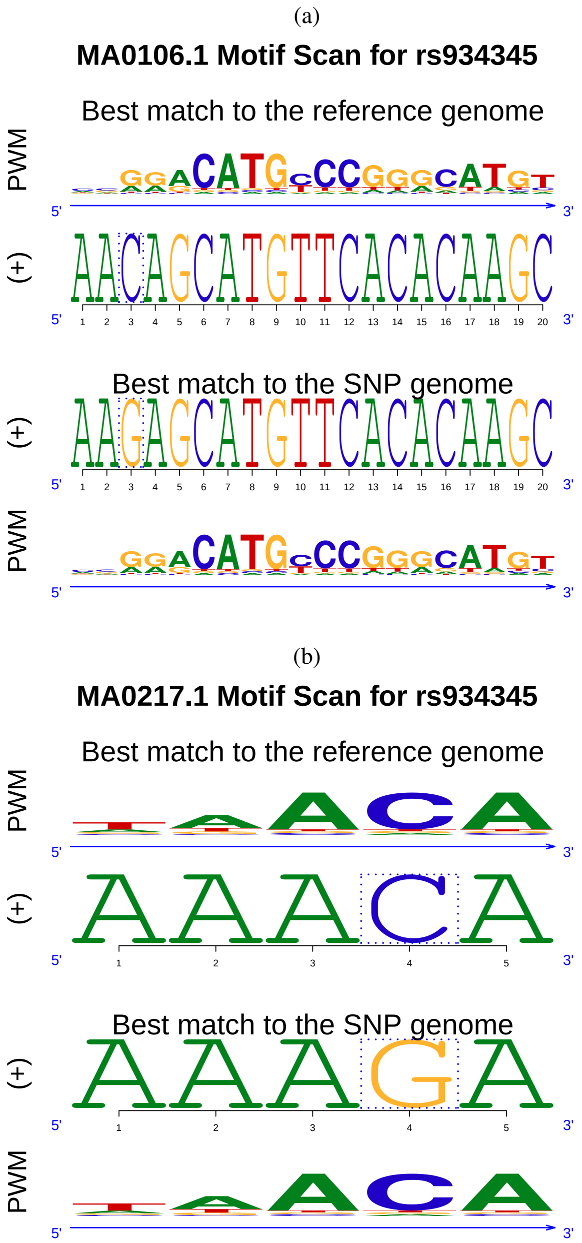
(a)

MA0039.2 Motif Scan for rs243865

(b)

PF0082.1 Motif Scan for rs243865

Fig. 30. Sequence logo plot for rs243865 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



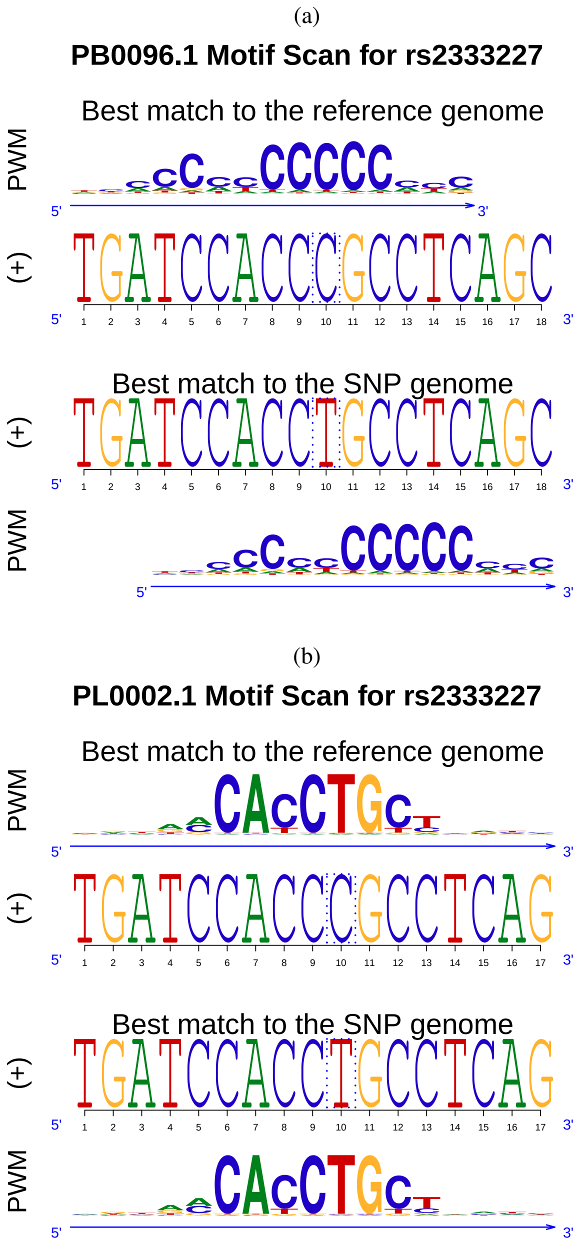
(a)

MA0106.1 Motif Scan for rs934345

(b)

MA0217.1 Motif Scan for rs934345

Fig. 31. Sequence logo plot for rs934345 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



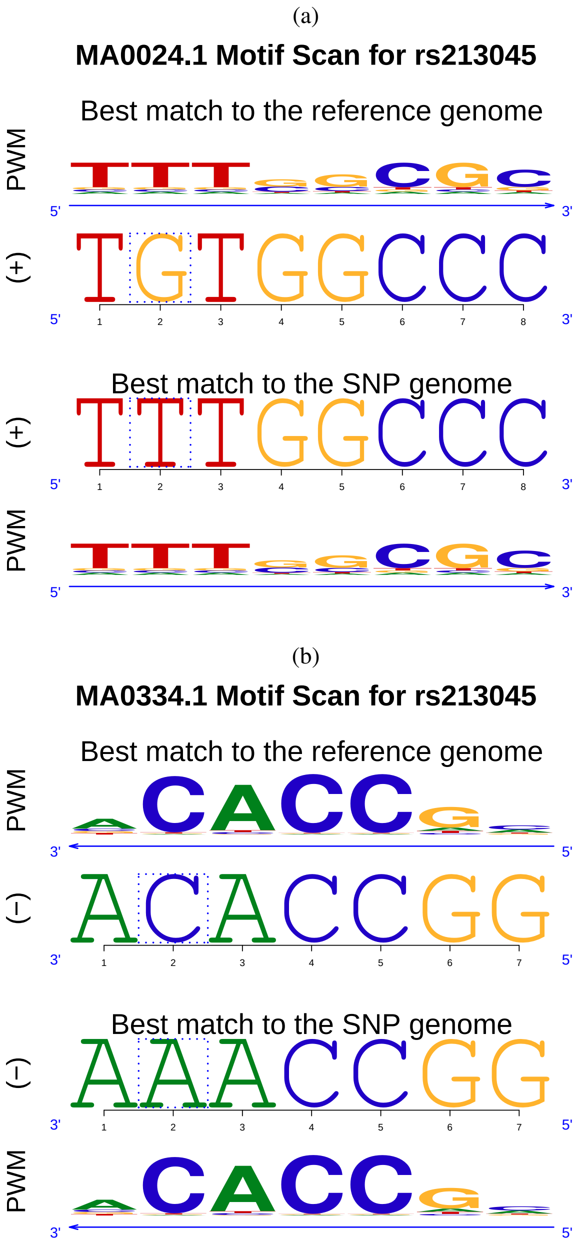
(a)

PB0096.1 Motif Scan for rs2333227

(b)

PL0002.1 Motif Scan for rs2333227

Fig. 32. Sequence logo plot for rs2333227 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



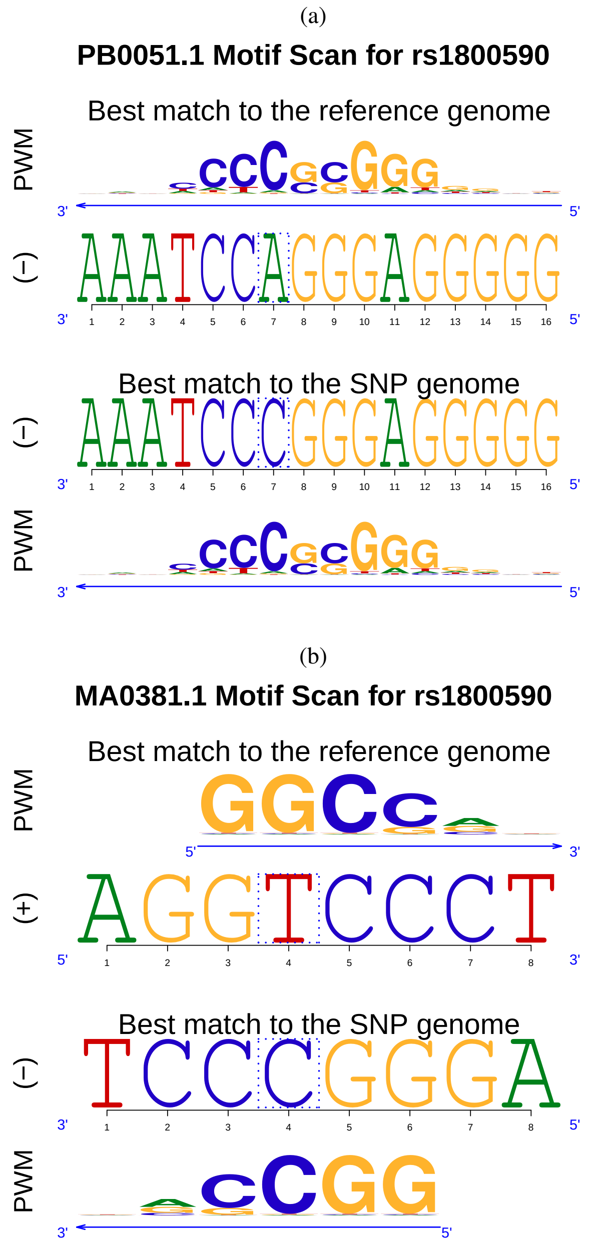
(a)

MA0024.1 Motif Scan for rs213045

(b)

MA0334.1 Motif Scan for rs213045

Fig. 33. Sequence logo plot for rs213045 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



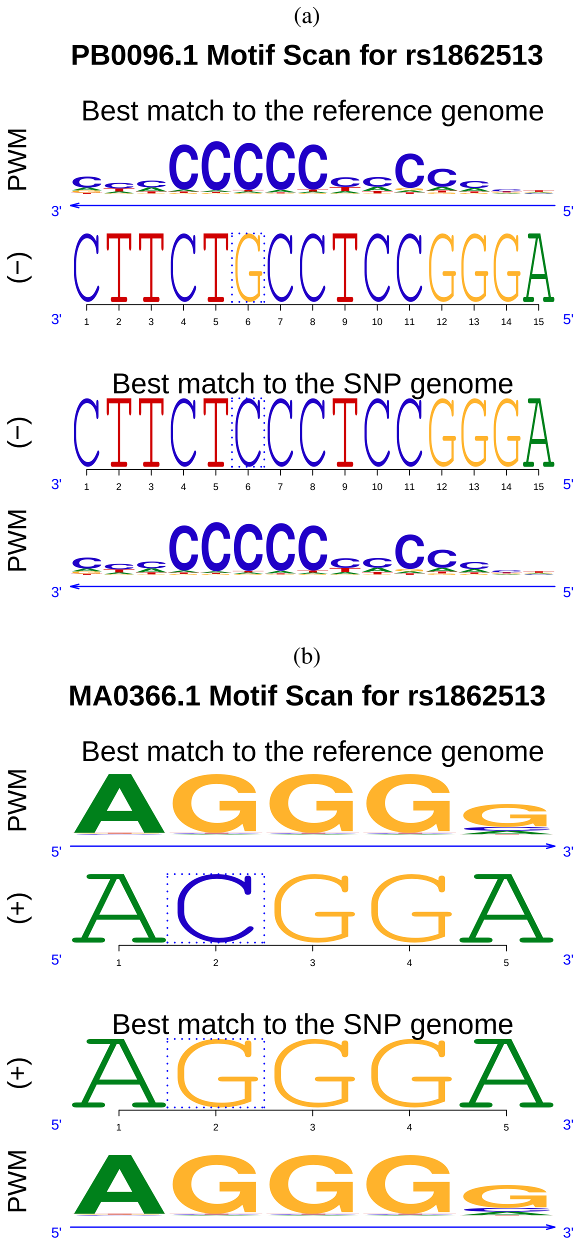
(a)

PB0051.1 Motif Scan for rs1800590

(b)

MA0381.1 Motif Scan for rs1800590

Fig. 34. Sequence logo plot for rs1800590 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.



(a)

PB0096.1 Motif Scan for rs1862513

(b)

MA0366.1 Motif Scan for rs1862513

Fig. 35. Sequence logo plot for rs1862513 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

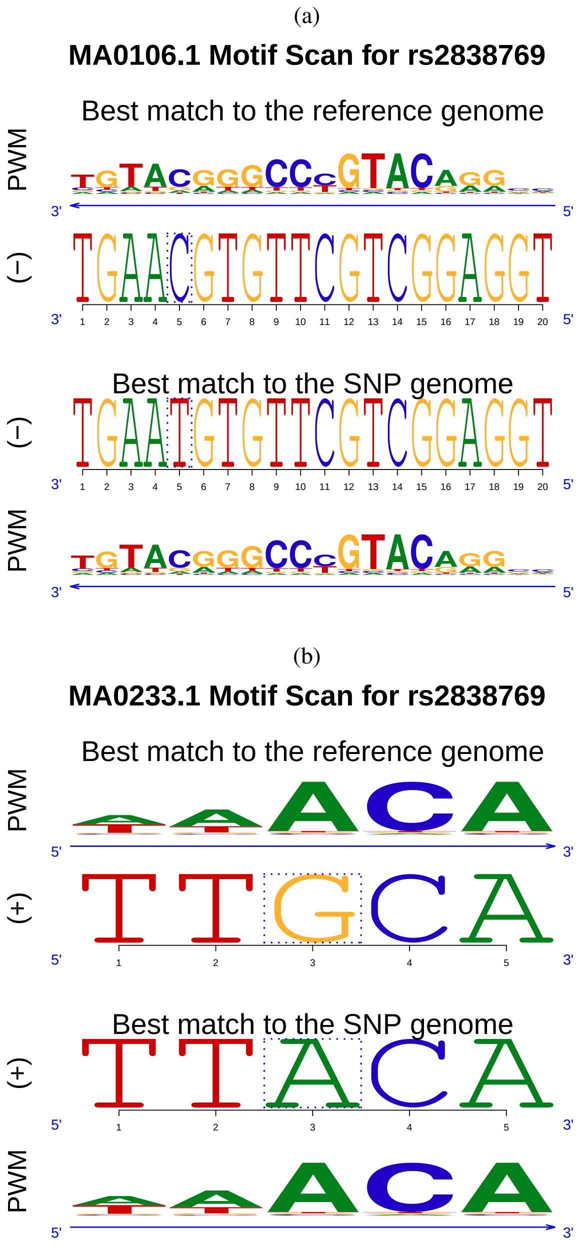


Fig. 36. Sequence logo plot for rs2838769 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

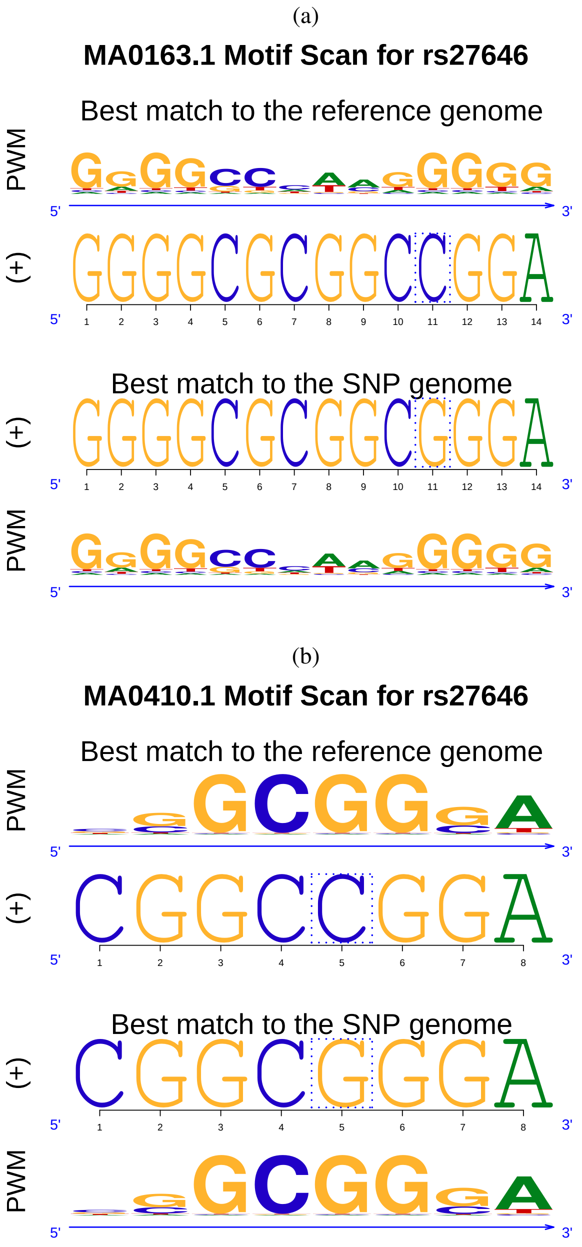


Fig. 37. Sequence logo plot for rs27646 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

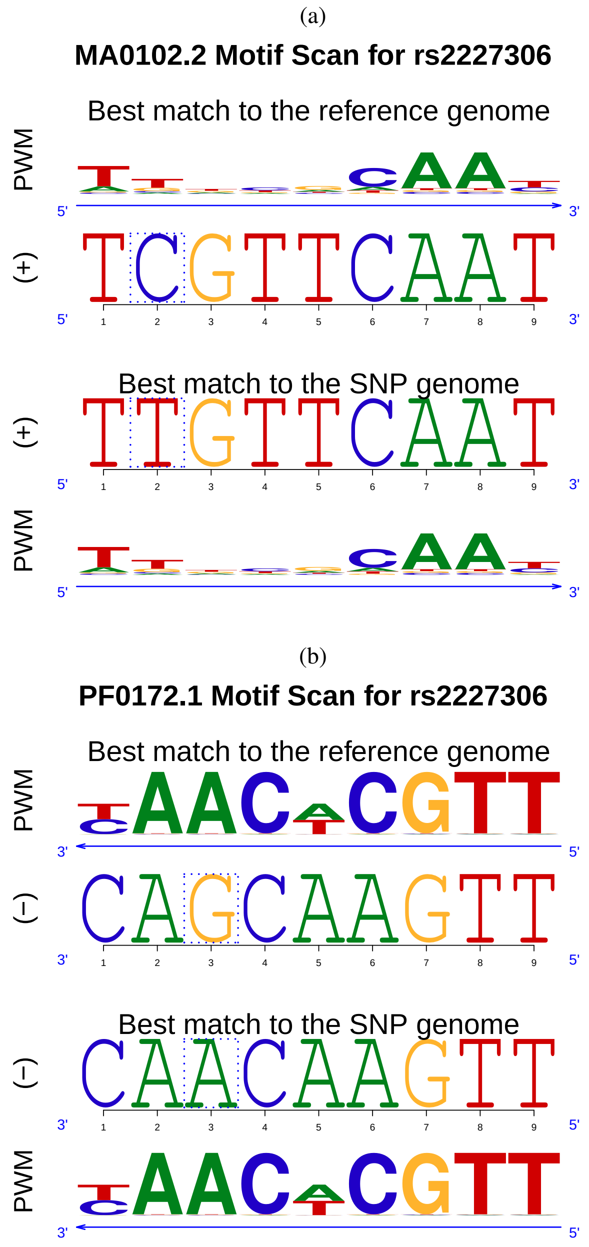


Fig. 38. Sequence logo plot for rs2227306 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

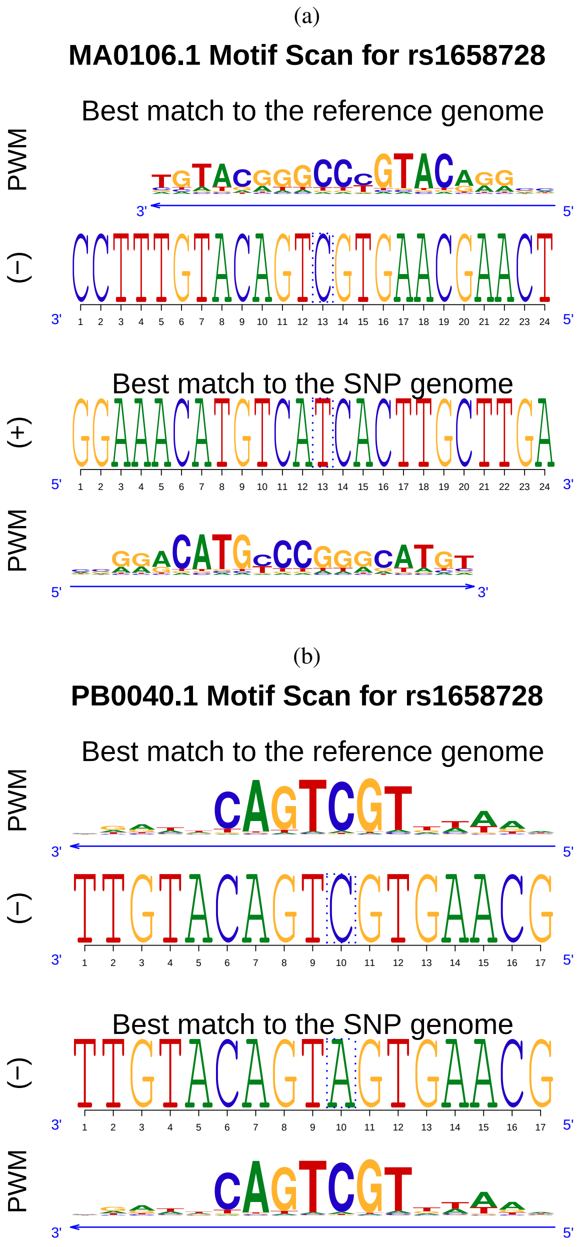


Fig. 39. Sequence logo plot for rs1658728 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

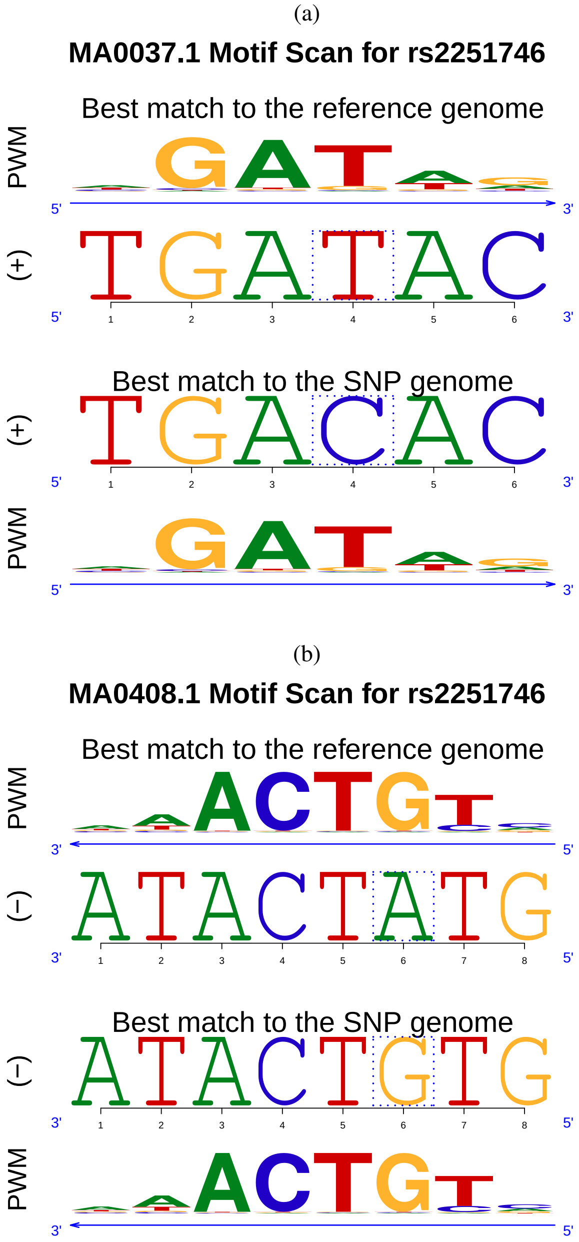


Fig. 40. Sequence logo plot for rs2251746 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

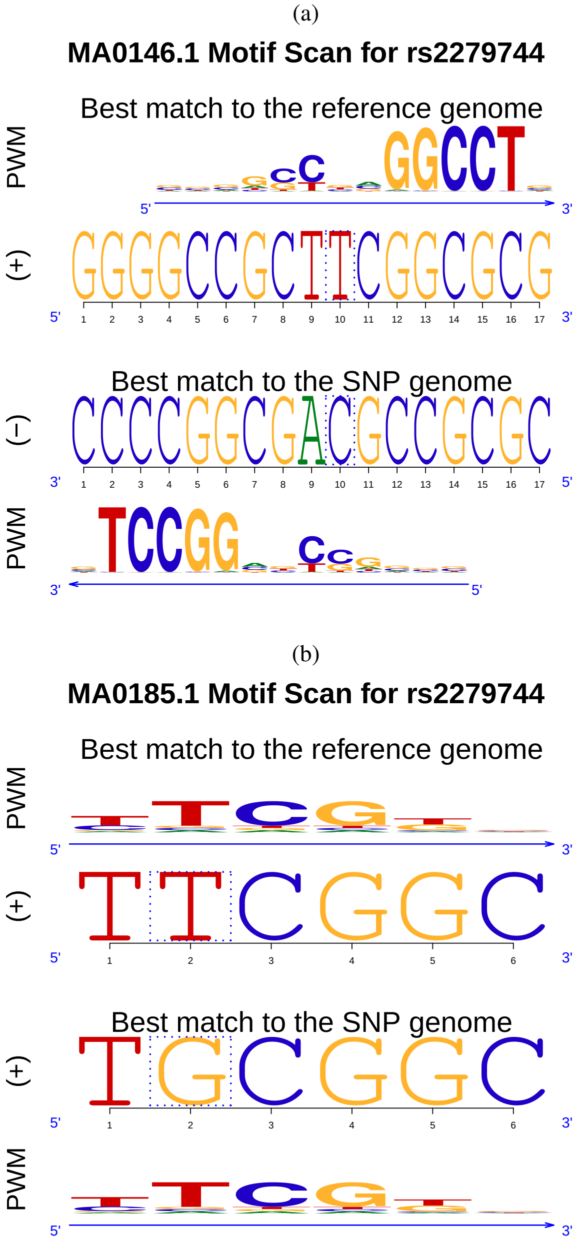


Fig. 41. Sequence logo plot for rs2279744 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

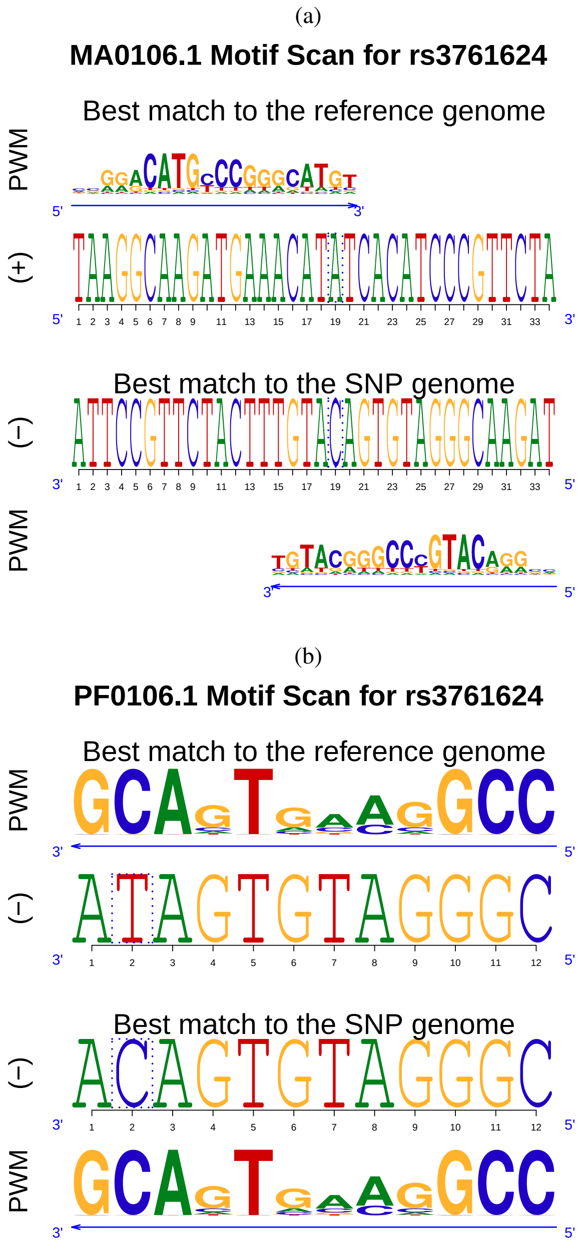


Fig. 42. Sequence logo plot for rs3761624 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

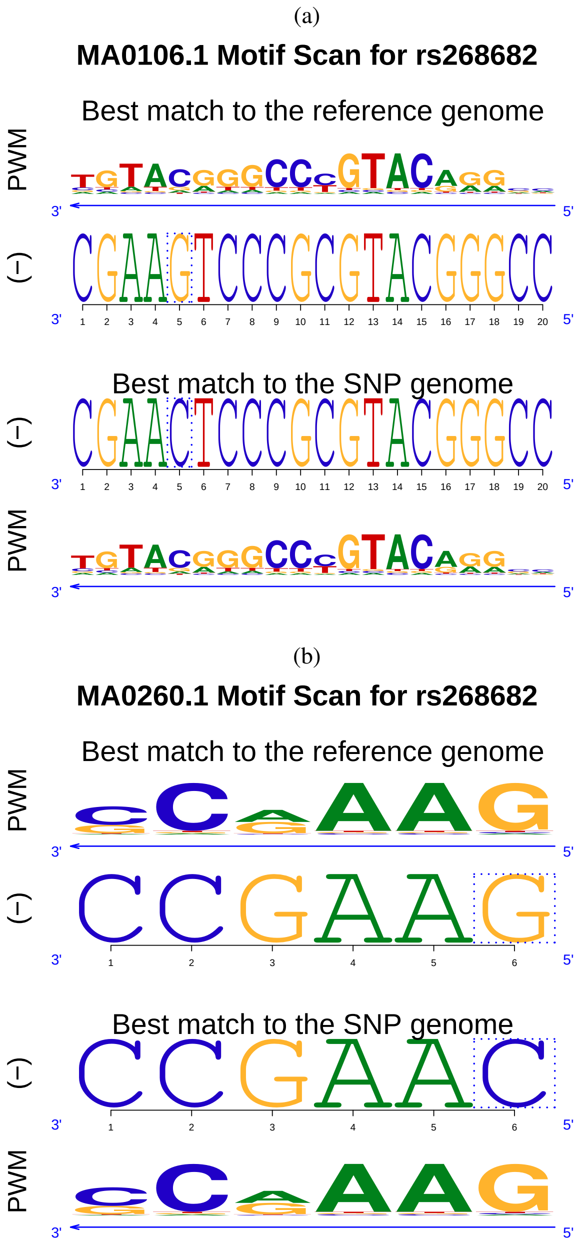
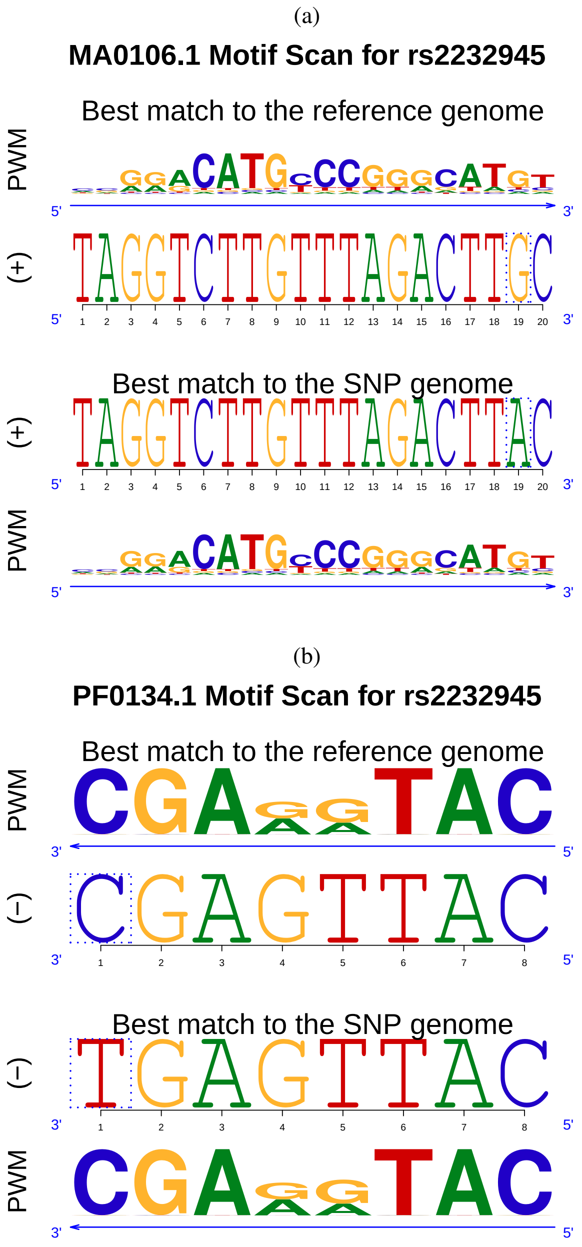


Fig. 43. Sequence logo plot for rs268682 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

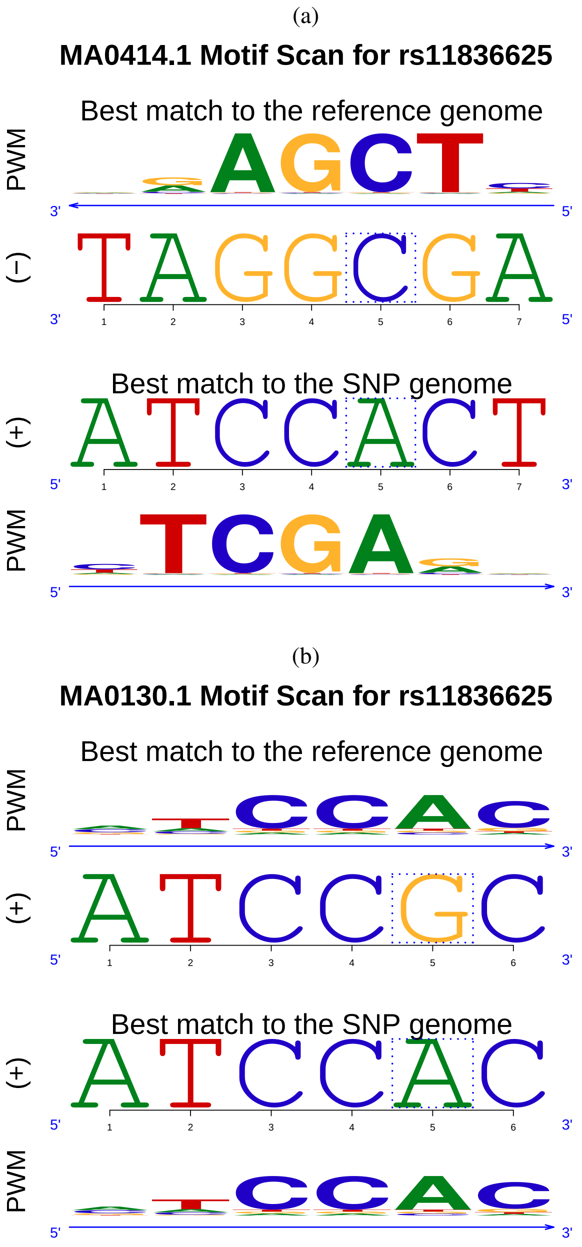

(a)

MA0106.1 Motif Scan for rs2232945

(b)

PF0134.1 Motif Scan for rs2232945

Fig. 44. Sequence logo plot for rs2232945 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.


(a)

MA0414.1 Motif Scan for rs11836625

(b)

MA0130.1 Motif Scan for rs11836625

Fig. 45. Sequence logo plot for rs11836625 with the motif of the top regulatory effect in (a) the reported TF family; and (b) the whole JASPAR library.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistics Society, Series B*, **57**, 289–300.

Chan, H. P., Zhang, N. R., and Chen, L. H. (2010). Importance sampling of word patterns in DNA and protein sequences. *Journal of Computational Biology*, **17**(12), 16971709.

Fulton, D., Sundararajan, S., Badis, G., Hughes, T., Wasserman, W., Roach, J., and Sladek, R. (2009). Tfcat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**.

Grant, C. E., Bailey, T. L., and Nobel, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **7**, 1017.

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M. C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S. M., Giardine, B., Hooghe, B., Loo, P. V., Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E., Donaldson, I. J., Robertson, A. G., Wadelius, C., Bleser, P. D., Vlieghe, D., Halfon, M. S., Wasserman, W. W., Hardison, R., Bergman, C. M., Jones, S. J. M., and the Open Regulatory Annotation Consortium (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research*, pages D107–D113.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, **105**(491), 1215–1227.

Kheradpour, P. and Kellis, M. (2013). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, **13**.

Mathelier, A., Zhao, X., Zhang, A., Parcy, F., Worsley-Hunt, R., Arenillas, D., Buchman, S., Chen, C., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*.

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *Ann. Statist.*, **31**(6), 2013–2035.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**(12), 1113–1122.

Zhang, C., Xuan, Z., Otto, S., Hover, J. R., McCorkle, S. R., Mandel, G., and Zhang, M. Q. (2006). A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Research*, **34**(8), 2238–2246.