Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

Motif-Raptor: A Cell Type-Specific and Transcription Factor Centric Approach for Post-GWAS Prioritization of Causal Regulators

Qiuming Yao[1,2,3], Paolo Ferragina[4], Yakir Reshef[5,6], Guillaume Lettre[7,8], Daniel E. Bauer[2,3,6,9,*], Luca Pinello[1,3,6,*]

[1]Department of Pathology, Massachusetts General Hospital, Charlestown, MA, USA, [2]Division of Hematology/Oncology, Boston Children's Hospital, [3]Harvard Medical School, Boston, MA, USA, [4]Department of Computer Science, University of Pisa, Pisa, Italy, [5]Department of Computer Science, Harvard University, [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA, [7]Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada, [8]Montreal Heart Institute, Montreal, Quebec, Canada, [9]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Genome-wide association studies (GWAS) have identified thousands of common trait-associated genetic variants but interpretation of their function remains challenging. These genetic variants can overlap the binding sites of transcription factors (TFs) and therefore could alter gene expression. However, we currently lack a systematic understanding on how this mechanism contributes to phenotype.

Results: We present Motif-Raptor, a TF-centric computational tool that integrates sequence-based predictive models, chromatin accessibility, gene expression datasets and GWAS summary statistics to systematically investigate how TF function is affected by genetic variants. Given trait associated non-coding variants, Motif-Raptor can recover relevant cell types and critical TFs to drive hypotheses regarding their mechanism of action. We tested Motif-Raptor on complex traits such as rheumatoid arthritis and red blood cell count and demonstrated its ability to prioritize relevant cell types, potential regulatory TFs and non-coding SNPs which have been previously characterized and validated.

Availability: Motif-Raptor is freely available as a Python package at:
https://github.com/pinellolab/MotifRaptor.

Contact: lpinello@mgh.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transcription factors (TFs) are DNA-binding proteins that recognize short DNA sequences and are critical for dynamic gene regulation (Lambert *et al.*, 2018). The rate of gene transcription is controlled by TFs in a cell type-specific fashion to regulate specific pathways and guide stages of development (Whyte *et al.*, 2013). TF binding sites have been characterized by *in-vitro* binding assays (e.g. HT-SELEX or PBM) as well as *in-vivo* through DNA foot-printing and chromatin immunoprecipitation (ChIP) (Lambert *et al.*, 2018). The TF-DNA binding pattern (often referred as a *TF motif*) can be described and predicted by several models. The simplest and most common is the Position Weight Matrix (PWM), a matrix that encodes the nucleotide preferences at each position of putative binding sites.

Several studies have reported that genetic variants can enhance or disrupt TF-DNA binding affinity (Wienert *et al.*, 2015; Weinhold *et al.*, 2014; De Gobbi *et al.*, 2006). Genome-wide association studies (GWAS) have uncovered thousands of genetic variants (SNPs) associated with complex traits or human disease (Buniello *et al.*, 2019). Despite these efforts, functional studies to prioritize potential causal variants have lagged behind (Gallagher and Chen-Plotkin, 2018), resulting in a limited interpretation of the underlying pathophysiology mechanisms connecting

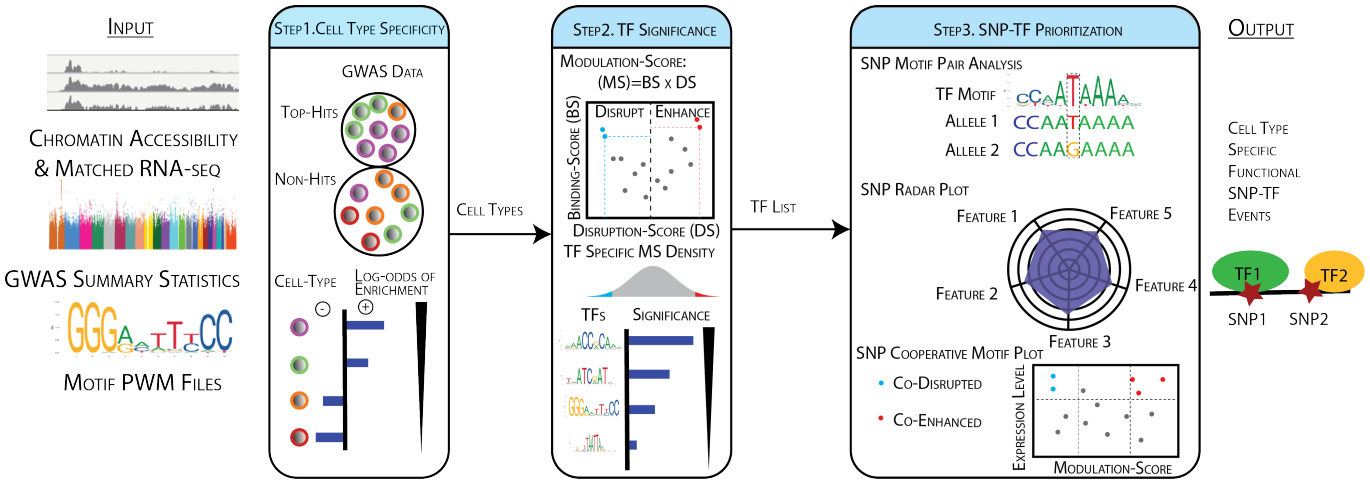

Figure 1. Summary of Motif-Raptor analysis workflow. Three steps are performed: (1) characterize relevant cell types based on the enrichment of phenotype associated SNPs in chromatin accessible sites, (2) find TFs with binding sites that are significantly modulated by genetic variants in these cell types and (3) identify and visualize individual TF-SNP regulation events.

variant to phenotype. A few missense SNPs can alter the function of a given TF by affecting its coding sequence, protein structure and therefore DNA binding capability, especially for Mendelian disease (Barrera *et al.*, 2016). For common diseases and complex traits, the great majority (>90%) of associated SNPs are in non-coding regions and mainly in DNase I~~-seq-based~~ hypersensitive sites. These SNPs correspond to functionally relevant non-coding regions such as enhancers and promoters (Maurano *et al.*, 2012). This observation suggests that chromatin state alterations and gene deregulation may be mediated by SNPs that modulate TF binding activities. In other words, genetic variants in these non-coding regions may perturb TF recognition sequences to enhance or disrupt TF-DNA binding events ultimately changing the downstream gene expression programs (Deplancke *et al.*, 2016). Even if single non-coding SNPs may only moderately alter binding sites and are underpowered to explain gene expression programs, statistics on a set of SNPs modulating common TF binding sites could be significant enough to reveal the convergent regulatory mechanism in complex traits. The method we present is based on this key idea.

Despite the fact that several approaches have been proposed to explore how TF binding sites could be affected by genetic variants, challenges remain. The next paragraphs provide a short summary and the rationale behind the development of Motif-Raptor. ~~**Supplementary Material Section 1 provides an extended discussion of these methods and their limitations, while the next paragraphs provide a short summary and the rationale behind the development of Motif-Raptor.**~~ First, current availability of ChIP-seq data unfortunately limit the utility of tools such as MMARGE (Link *et al.*, 2018), GERV (Zeng *et al.*, 2016), DeepSEA (Zhou and Troyanskaya, 2015), Basset (Kelley *et al.*, 2016), IMPACT (Amariuta *et al.*, 2019), RegulomeDB (Boyle *et al.*, 2012) and HaploReg4 (Ward and Kellis, 2012). In fact, these tools are extremely powerful and practical only when genome-wide maps of TF occupancy and/or chromatin marks in relevant cellular contexts are available. We therefore found a unique value proposition in developing a framework to accommodate scenarios in which only PWM models and gene expression data are available. Second, available models based on ChIP-seq or PWM data do not systematically provide a global ranking and the significance of the TFs based on all trait-associated variants, rather a per SNP scoring. In fact, current methods based on PWM and/or DNase I-seq data, such as Combined Annotation Dependent Depletion (CATO) (Maurano et al., 2015), CENTIPEDE (Pique-Regi et al., 2011; Moyerbrailean et al., 2016), Affinity Testing for regulatory SNPs (atSNP) (Zuo et al., 2015), do not provide a procedure to formally test the global effect of a set of GWAS variants on the set of overlapping TF binding sites. To solve this limitation we propose here a novel genome-wide statistic to prioritize putative causal TFs based on the entire set of binding sites and overlapping variants rather than single loci. Third, these methods do not consider linkage disequilibrium (LD) for the tagged loci by the GWAS-associated variants. This is important given that several non-causal SNPs have similar association scores as the true causal ones and that this potentially confound the analysis. In fact, these false positives can dilute our power of detecting the true mechanisms behind the causal variants. Our approach tries to account for this problem based on the two following strategies. By relying on cell type specific chromatin accessibility regions, we are already reducing the space of variants in each LD block. To implicitly account for local LD structure, we sample our background set of chromatin accessibility regions in close proximity of the regions that are specific for each cell type. With these strategies we mitigate the problem by specifically looking for variants within regions that are cell type specific. To our knowledge, among available tools only SLDP (Reshef *et al.*, 2017) overcomes this problem and offers a genome-wide significance score for each TF, based on the directional modulation of TF binding sites by SNPs. Another tool, GREGOR (Schmidt et al., 2015), also explicitly accounts for LD structure to assess the enrichment on sentinel SNPs in arbitrary genomic regions (for example to prioritize cell types based on cell type specific annotations). Supplementary Material Section 1 provides an extended discussion of these methods and their advantages and limitations while a formal comparison of Motif-Raptor with SLDP, GREGOR and other similar tools is presented in section 3.4 and Supplementary Material Section 4.

To address above limitations of current approaches, we developed Motif-Raptor, an approach that integrates TF binding motif databases, cell type-specific chromatin accessibility and gene expression to prioritize TFs whose function may be modulated by genetic variants associated with different traits. Motif-Raptor provides a cell type-specific TF-centric analysis with associated statistics, comprehensive reporting and visualization functionalities. This tool can facilitate the discovery and interpretation of the action of non-coding variants on key regulators of complex traits.

2 Methods

2.1 Motif-Raptor overview

Motif-Raptor takes as input: GWAS summary statistics for a given trait or disease, TF binding models (PWM), chromatin accessibility, and transcriptomic data. It produces as output a ranked list of the putative trait-associated TFs whose binding sites are modulated by genetic variants in relevant cell types. In addition, for each variant intuitive visualization to explore external annotations and the potential involvement of co-factors are offered. Briefly, the analysis performed by Motif-Raptor consists of three main steps: (1) characterize relevant cell and tissue types, (2) find significant TFs in these cell types and (3) identify and annotate critical TF-SNP regulation events (**Figure 1**). Motif-Raptor is an open-source command line utility built with Python and Cython to achieve both portability and efficiency.

2.2 Quantification of the effects of genetic variants on TF binding

The prerequisite to quantifying the effects of genetic variants at a given TF binding site, is to first assess the binding affinity of this TF, given a generic DNA sequence. Motif-Raptor implements a scoring procedure as proposed in Motif Occurrence Detection Suite (MOODS) (Korhonen *et al.*, 2017, 2009) however we specifically adopt different data structures (see section 2.3) to efficiently calculate genome-wide and threshold free scans of all the binding sites overlapping a set of SNPs in order to efficiently compute the modulation scores and the null models required to test its significance, as described below. Briefly, given a genomic target sequence, $S = \{S_i\}$ of length m, and a position weighted matrix, M for a given TF of length m, the matching score $M(i, S_i)$ represents the likelihood at position i $(1 \leq i \leq m)$ of observing the nucleotide $S_i \in \{A, T, C, G\}$. The binding score, BS is derived from $M(i, S_i)$ as a log-likelihood over the entire binding region from independent multinomial-distributed random variables. It is then corrected to account for genome-wide (or region-specific) nucleotide frequency, $B(S_i)$, as follows:

$$BS(S, M) = log \prod_{i=1}^{m} \frac{M(i,S_i)}{B(S_i)} = \sum_{i=1}^{m} (log(M(i, S_i)) - log(B(S_i))) \quad (1)$$

Based on this scoring procedure, we derive a disruption score, DS to model the potential effect of genetic variants on a given binding site as follows. Given a SNP within a target sequence, S assuming only two haplotypes, we will use S^{REF} and S^{ALT} to denote the two different alleles, i.e. reference and alternative, respectively. These alleles can be scored with equation (1), above. To make our scoring efficient, we restrict our computation onto a region, R of length $2m - 1$ centered around the target SNP. This enforces that any sequence, S corresponding to a putative binding site of length m and spanning this SNP is contained within R. We consider in each region, R for both the reference and the alternative allele, the best putative binding position, $1 \leq K \leq m$ using the following equation:

$$K = arg \max_{1 \leq k \leq m} (BS(S^{REF}_{k:k+m-1}, M), BS(S^{ALT}_{k:k+m-1}, M)) \quad (2)$$

The disruption score at the optimal binding position, K on this SNP is then defined as follows:

$$DS(S, M) = \Delta BS = BS(S^{ALT}_{K:K+m-1}, M) - BS(S^{REF}_{K:K+m-1}, M) \quad (3)$$

The sign and amplitude of the disruption score is informative on the directionality and strength of a putative TF-SNP modulation event. A positive score suggests that a SNP may enhance the binding affinity, while a negative score may reduce it. Considering that different TF binding motifs have various lengths and specificity, this binding score cannot be used directly, therefore we rescale it by considering the sequence with the highest binding affinity accordingly to this model, as follows:

$$BS_{max}(M) = \max_{S_i \in \{A,T,C,G\}} \sum_{i=1}^{m} (log(M(i, S_i)) - log(B(S_i))) \quad (4)$$

$$DS_{max}(M) = \max_{A_i \neq B_i \in \{A,T,C,G\}} |\, log\big(M(i, A_i)\big) - log\big(M(i, B_i)\big)\,| \quad (5)$$

After this operation, the binding and disruption scores are within [0,1] and [-1,1], respectively. Based on these two scores, we then define a space called B-D (Binding and Disruption). This space can be used to visualize and summarize the effect of TF-SNP events globally across factors and conditions. In this B-D space, we are interested in events that are close to the distal corners from the origin, i.e. (1,1) or (1,-1), since they represent strong binding and large modulation, mediated by genetic variants. We formalize this intuition by combining the two scores into a single score, the *modulation score* (MS) as follows:

$$MS(S, M) = \left(\frac{BS(S,M)}{BS_{max}(M)}\right) \times \left(\frac{DS(S,M)}{DS_{max}(M)}\right) \quad (6)$$

Intuitively, the modulation score represents the rectangular area spanned by the scaled disruption score and the scaled binding score. Large absolute modulation scores (distal corners from the origin in the B-D space) correspond to meaningful modulating events. However, to quantify whether a set of GWAS-associated SNPs are significantly disrupting a TF it is necessary to model the distribution of the modulation scores using an appropriate null model. We investigated if MS distributions can be modeled with parametric functions, however no distributions fit satisfactorily the observed data.

Therefore, we propose an estimation of a null model based on the central limit theorem that is complete, i.e. based on a complete enumeration of all the putative binding sites across the genome. In fact, owing to the efficient data structures and related algorithms proposed in this paper, Motif-Raptor can compute efficiently, genome-wide SNP-based binding scores and disruption scores for all available SNPs and TFs without using pre-determined scores, *p*-value cutoffs, or computationally intensive shuffling procedures. This is a key contribution for building exact null models, since the complete enumeration of all the endogenous binding sites and the estimation of their modulation by observed genetic variants cannot be performed efficiently with current available tools as discussed in the next sections.

Finally, to provide a ranked list of TFs we define a TF score based on the combination of its cell type specific expression and modulation score in chromatin accessible regions. This is an important step, given that several TFs share similar motifs but are expressed and work in different cellular contexts. For each cell type C and motif M the TF-score is defined as:

$$TF\text{-}Score(M, C) = EP(M, C) \times (1 - minFDR(M, C)) \quad (7)$$

This score is bounded in [0,1], $EP(M, C)$ is the expression percentile and $minFDR(M, C)$ the corrected p-value for the significance of the modulation score comparing the distribution of MS in cell type specific chromatin accessibility peaks with the genome wide distribution.

The efficient calculation of the MS genome-wide is presented in the next two sections and more details on the overall scoring and ranking procedure are presented in **Supplementary Material Section 2.**

2.3 Ultra-fast SNP-based genome-wide motif scanning

As discussed above, to build a null model for our proposed B-D space and modulation score, it is necessary to perform a complete enumeration of all putative TF binding sites in the genome and calculate their potential modulation by overlapping SNPs not associated with a phenotype. Fast tools such as FIMO (Grant *et al.*, 2011) and MOODS (Korhonen *et al.*, 2009) can be used to enumerate all binding sites of a TF motif in a reasonable time, however they are not designed to efficiently compute the modulation of binding affinity introduced by a set of SNPs (Zuo *et al.*, 2015). To filter putative false positives and/or improve computational efficiency,

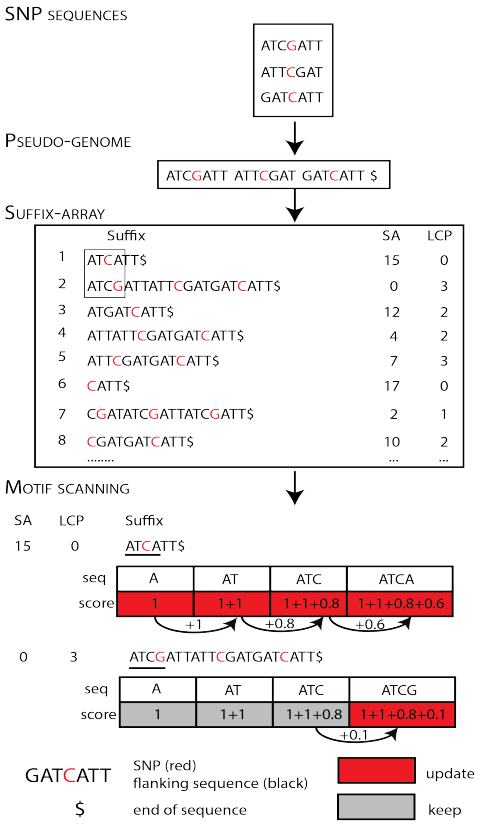

Figure 2. Data structures used to obtain an efficient scanning procedure in Motif-Raptor. Scanning and score calculation using a suffix array (SA) and the longest-common-prefix array (LCP).

these tools rely on a small pre-specified significance level (*p*-value or q-value), however this is problematic for two reasons. First, they may exclude weak but true binding events (i.e. false negatives) that could bias our estimation of the complete empirical distribution of the modulation scores. Second, it is neither practical nor reasonable to fix the same threshold for different TF motifs given that they generally have different lengths and complexities. atSNP (Zuo *et al.*, 2015), a genome-wide SNP-specialized tool solves the first problem by proposing a specialized procedure to estimate the *p*-value of binding efficiently. However, it does not provide an efficient threshold-free scanning technique.~~; a key requirement for a complete null model estimation as discussed in the results section (and shown in **Supplementary Figures 1** and **2**).~~

Motivated by above limitations, we developed a threshold-free algorithm to scan motifs and calculate their modulation scores (4)-(6) efficiently. Our algorithm avoids redundant calculations by using two key data structures, a suffix array (SA) and a longest-common-prefix (LCP) array (Holmes and Gusfield, 1999; Kasai *et al.*, 2001; Puglisi *et al.*, 2007; Gusfield, 1997) that index the genome. This allows detection and skipping repeated portions of the genome that are responsible for redundant calculations (**Figure 2**).

To develop and test our approach we used the human reference genome (hg19), 719 TF PWM models from the JASPAR2018 vertebrates database (Khan *et al.*, 2018) and the genetic variants from the 1000 Genomes project (phase3 Europeans) (Auton *et al.*, 2015). However, our approach is generalizable to any reference genome, TF motif database or genetic variants.

We first retrieve all SNPs and their flanking sequences from the genome. Suppose we have N SNPs and the flanking sequence around each SNP is of length $2m-1$ (block size). We fetch all blocks, S^{REF} and S^{ALT}, paste them back to create a single sequence, and record their original genome positions. In this way we create one unique long pseudo-genome having length $2\times(2m-1)\times N$, which we denote hereafter as P. Then SA and LCP array for the sequence P are constructed. SA stores the alphabetically sorted list of all suffixes of P and LCP stores, for each pair of adjacent suffixes in SA, the length of their longest common prefix. The construction of the SA and of the LCP array takes linear time in the total length of the pseudo-genome, hence $O(m*N)$ time (Kasai *et al.*, 2001; Puglisi *et al.*, 2007).

Technically, we scan these substrings according to their lexicographic order, dictated by the suffixes in SA, which they prefix. Let us assume that we are at iteration h of this scanning process and that we have inductively computed in an auxiliary array $BS[1:m]$, the binding scores between the motif M and the first m characters of the suffix of P which starts at $SA[h-1]$(recall that m is the length of the motif M). Clearly, $BS[m]$ is the binding score for the position $SA[h-1]$ in P against the entire motif M. Initially, the array BS is pre-filled with m zeroes. At the next iteration h, the algorithm needs to maintain the induction by computing the array $BS[1:m]$ for the first m characters of the next suffix $SA[h]$. Interestingly enough, this computation can take advantage of the current values stored in $BS[1:m]$ which refer to $SA[h-1]$ and of the value stored in $LCP[h]$. Indeed, if the length of the common prefix between $SA[h-1]$ and $SA[h]$ is $j=LCP[h]$, then updating of the binding scores can start from position, $j+1$ since we know that the binding scores in $BS[1:j]$ remain the same given that $SA[h-1]$ and $SA[h]$ share the first j characters, for which we have already computed the binding scores, $BS[1:j]$.

In detail, we have that for $j+1\le x\le m$, we can compute

$$BS[x]=BS[j]+\sum_{i=j+1}^{x}(log(M(i,S_i))-log(B(S_i))) \quad (8)$$

To map the binding score $BS[m]$ relative to an m-long substring of P, aligned with the motif M, back to the SNP site, we resolve the SNP position and the binding position using the following equations:

$$SNP\# = \lfloor SA[h]/(2m-1)\rfloor \quad (9)$$

$$binding_{site} = SA[h]\ mod\ (2m-1) \quad (10)$$

Not all m-long prefixes of P's suffixes are substrings of the original genome since P is constructed by concatenating blocks centered at SNP-sites. Therefore, we further expedite our calculations by excluding cross block sequences.

With these new data structures and score design, we can efficiently calculate binding and disruption scores genome-wide for each TF (and shown in **Supplementary Figure 1 and 2**). This allows to estimate complete null models for each TF and easily assess their significance as discussed in the next sections.

2.4 Efficient assessment of motif modulation score significance

In this section we introduce the procedures to assess the significance of a given TF based on a set of trait-associated SNPs and on the corresponding modulation scores.

Given a set of target SNPs (T), i.e. top ranked SNPs in GWAS summary statistics, we want to ask if the motif modulation score distribution on these target SNPs $\mathcal{D}_{MS}(T)$ is significantly different from the background SNPs $\mathcal{D}_{MS}(B)$ (genome-wide) or not. To assess the distribution difference between target and background sets we propose a non-parametric test with null hypothesis $E(\mathcal{D}_{MS}(T))=E(\mathcal{D}_{MS}(B))$. One might consider a naïve approach to implement this test based on a simple re-sampling procedure. However, this will require the generation of thousands of samples for hundreds of TF motifs; this approach is time consuming and thus impractical. We instead reasoned that, based on the central limit theorem (CTL), the distribution of the sample mean for B will converge to a normal

distribution. Enabled by the SA-based procedure described in section 2.3 we can efficiently calculate the genome wide population mean and standard deviation. It is therefore straightforward to derive a computationally efficient procedure to test the null hypothesis that $E(\mathcal{D}_{MS}(T)) = E(\mathcal{D}_{MS}(B))$ (Supplementary Figure 3B). In fact, the sample mean will have a normal distribution with mean $E(\mathcal{D}_{MS}(B))$ and variance $Var(\mathcal{D}_{MS}(B))/N_{re-sample}$ regardless of the underlying modulation score distribution. Based on this assumption we tested enhanced binding ($E(\mathcal{D}_{MS}(T))> E(\mathcal{D}_{MS}(B))$, disrupted binding ($(E(\mathcal{D}_{MS}(T))< E(\mathcal{D}_{MS}(B))$, or both. This identification of significant shift of modulation score distributions builds the foundation of characterizing significant TFs given a set of trait-associated SNPs. In fact, owing to the efficient data structures and related algorithms proposed in this paper, Motif-Raptor can compute efficiently, genome-wide SNP-based binding scores and disruption scores for all available SNPs and TFs without using pre-determined scores, p-value cutoffs, or computationally intensive shuffling procedures (see also **Supplementary Material Section 2.4**).

2.5 Main steps in running Motif-Raptor

The analysis performed by Motif-Raptor consists of three main steps as illustrated in **Figure 1**. In step 1, to characterize different cell type specific regulatory programs and regions, we collected expression and chromatin accessibility data for 83 distinct cell/tissue types from the ENCODE project (Davis *et al.*, 2018) (see **Supplementary Material Section 2.1** and **Supplementary Figure 3A**). Then for each trait, we partitioned the GWAS summary statistics into top hits and non-hits and applied an enrichment test in cell-type specific chromatin open regions to rank the most associated cell types (see **Supplementary Material Section 2.2-2.3**). In step 2, to rank and uncover potential causal TFs for each of the prioritized cell types obtained in step 1 we calculate the TF-score presented in section ~~2.3~~2.2 (and detailed in **Supplementary Material Section 2.4**).

Finally, in step 3, we provide several visualization strategies to explore single TF-SNP events. This includes a radar plot for each pair of TF-SNP to visualize the binding or SNP features, and two additional plots to

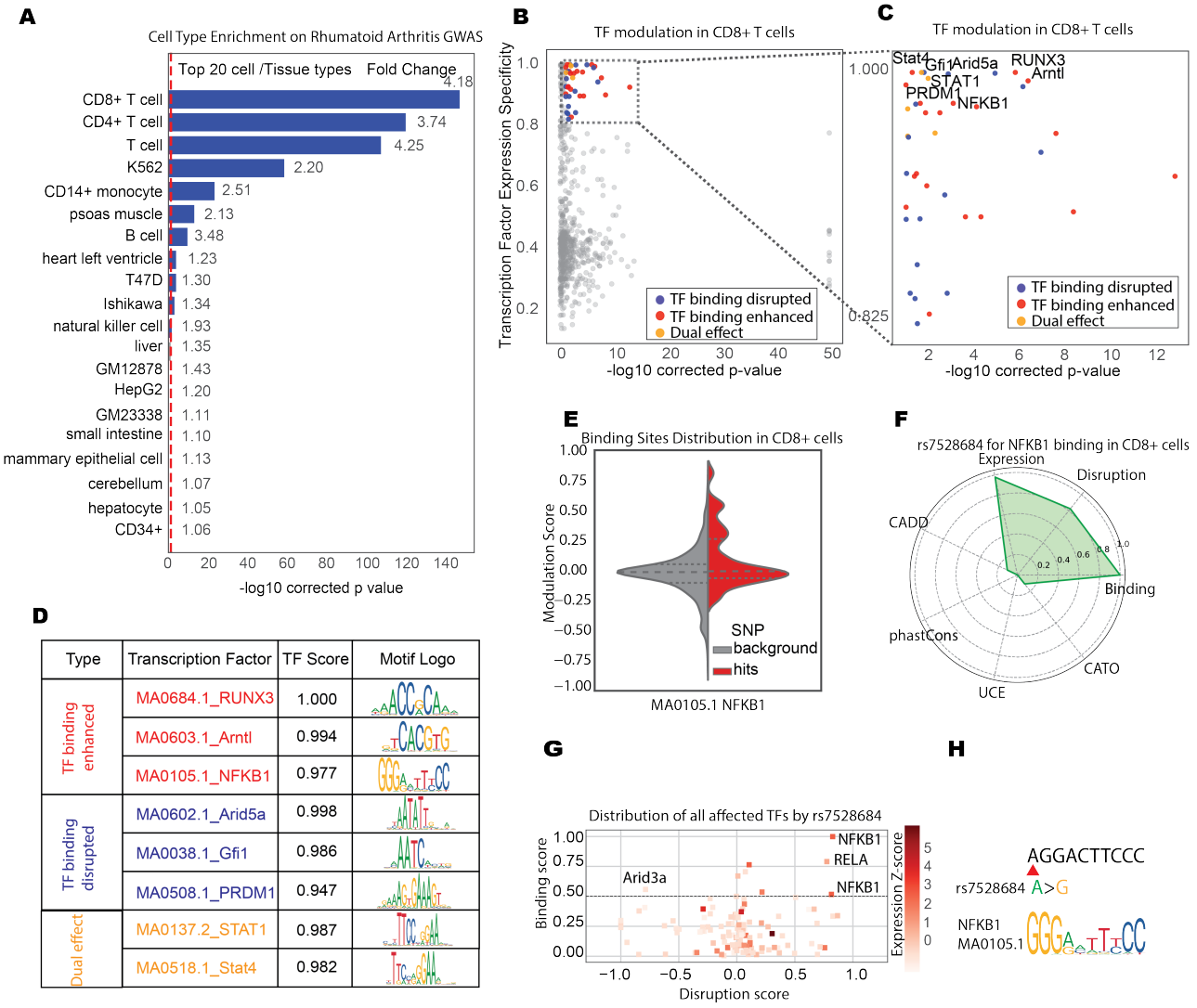

Figure 3. Cell type association and specific TF identification for Rheumatoid Arthritis. (A) The top associated cell types are CD8+ and CD4+ cells (blue bars represent p-values, and the red line represents the proposed cutoff at 5%); (B,C) Modulated TFs in CD8+ T cells and (D) Top 10 TFs ranked by the TF-score; (E) NFKB1 modulation score distribution in CD8+ T cells; (F) Radar plot for rs7528684; SNP centered features to assess its potential mechanism are presented and based on available annotations or models; (G) B-D plot for rs7528684 to investigate potential cooperative factors (the duplicated NFKB1 corresponds to two different PWM models i.e. MA0105.1 and MA0105.2); (H) rs7528684 is enhancing the binding of NFKB1 based on its PWM.

explore potential co-factors for given SNP and other modulated SNPs for a given TF (see **Supplementary Material Section 2.5**).

3 Results

3.1 Study on rheumatoid arthritis

Rheumatoid arthritis (RA) is a common autoimmune disease in humans. A GWAS study on RA explored >100,000 individuals, profiling ~10 million SNPs; this identified 98 potential causal genes on 101 risk loci (Okada *et al.*, 2014). In this study, 26,285 SNPs were significantly associated to this phenotype (p-value<$5x10^{-8}$). Although the activation of some transcription factors is critical in RA (Okamoto *et al.*, 2008), a systematic analysis of how associated genetic variants globally alter TF binding sites is lacking. Notably, ~25% (6494) of these SNPs overlap at least one DNase I-seq peak in the 83 cell types collected. Applying Motif-Raptor to these SNPs, we uncovered significant (corrected p<0.05) and relevant cell types, the top three being CD8+ T cells, CD4+ T cells, and T cells, with log-odds 4.18, 3.74, 4.25 (**Figure 3A**). Importantly, in support of our results, previous studies have shown the importance of T lymphocytes, B lymphocytes, dendritic cells, plasma cells, mast cells and osteoclasts in the RA synovium (Tran *et al.*, 2005; Matsumoto *et al.*, 2006).

Next, using the most significant cell type, CD8+ T cells, we identified 34 TF motifs of factors that have relatively high expression specific to this cell type (transcription factor expression levels with FPKM>2, expression percentile>0.8) with a statistically significant modulation of binding by genetic variants associated in RA as well as in cell type-specific DNase I-seq peaks (**Figure 3B,C and Supplementary Table 1**). The obtained motifs were ranked by combining expression specificity and FDR, as described in section 2.2 (**Figure 3D**). Based on these criteria, RUNX3, ARNTL, NFKB1, were the top 3 enhanced TF motifs, meaning they gained binding affinity from overlapping SNPs. ARID5A, GFI1, PRDM1 are top 3 disrupted TF motifs, meaning they lost binding affinity from overlapping SNPs. STAT4 and STAT1 show instead a dual behavior - they are both enhanced and disrupted. In fact, the distribution of the modulation score shows a bimodal distribution.

The role of GFI1, a transcriptional repressor (Kim *et al.*, 2014; Huang *et al.*, 2005; Pinello *et al.*, 2014), STAT4 (Remmers *et al.*, 2007; Korman *et al.*, 2008), and NFKB1 (Makarov, 2001; Simmonds and Foxwell, 2008; Liu *et al.*, 2017) in RA have been extensively studied. To explore the informative potential of the cell line ranking obtained in the first step, we performed a similar analysis on the CD4+ cell type, the second-most significant cell type associated with RA (**Supplementary Figure 4A,B**). We identified 47 TFs, 22 of which overlapped with the ones recovered in the CD8+ cell type (**Supplementary Figure 4C**). RUNX3, ARNTL, NFKB1, ARID5A, GFI1, STAT4 were also significant in CD4+ cells. Conversely, PRDM1, a known repressor in CD8+ cells, was specific to this cell type, suggesting it may play a distinct role in CD4+ cells (Rutishauser *et al.*, 2009; Fu *et al.*, 2017). Their putative target gene expression was plotted in (**Supplementary Figure 4D**), indicating NFKB1 may activate relatively lowly-expressed genes.

Enabled by the visualizations generated in the final step of Motif-Raptor, the opposite binding modulation of NFKB1 and ARID5A is evident (**Figure 3E, Supplementary Figure 5** and **6A**). Specifically, NFKB is primarily enhanced while ARID5A is primarily disrupted (**Figure 3E, Supplementary Figure 5A**). It is known that both NFKB1 and ARID5A promote autoimmunity and auto-inflammation response regulated closely through separate mechanisms (Puel and Casanova, 2018).

Next, we focused our attention on the SNPs with the largest effects based on the modulation scores and visual inspection of the B-D plots (with the caveat that this may lead to false positives since no significance is assigned to individual SNPs). For NFKB1, we recovered rs7528684, a SNP with a large disruption score that overlaps a site with a high binding affinity (**Figure 3F-H and Supplementary Figure 6A**). Several studies have reported that this SNP reflects the association of ~~FCRL3 with~~ RA (Jiang *et al.*, 2012; Eyre *et al.*, 2006; Zhao *et al.*, 2013) and increases the binding affinity of NFKB1 (Zhao *et al.*, 2013). Using a radar plot, Motif-Raptor integrates and visualizes several publicly available scores including phastCons, (Siepel et al., 2005), CADD (Combined Annotation Dependent Depletion), (Rentzsch et al., 2019), and CATO (Contextual Analysis of TF Occupancy), (Maurano et al., 2015). PhastCons estimates evolutionary conservation based on multiple alignments), CADD assesses the deleteriousness of SNPs and insertion/deletions variants, and CATO assesses how variants in accessible sites disrupt TFs. (**Figure 3F**). The SNP-specific B-D plot for rs7528684 allows us to uncover additional TFs that might cooperate with NFKB1, even if modulation of their binding site is not significant in a genome-wide context (**Figure 3G**). For example, RELA shows a strong binding and disruption for this SNP. RELA has been reported as an important partner of NFKB1 (Handel *et al.*, 1995; Makarov, 2001), forming a heterodimer as the NFKB complex in RA (Oeckinghaus and Ghosh, 2009). Contrasting this example, ARID3A is disrupted at the same SNP site, potentially suggesting competitive binding with NFKB1 dictated by the presence of a particular allele.

For ARID5A, we found that rs17425622 has the strongest disruption score (**Supplementary Figure 5A,B and 6A**), also a relatively high CADD score but low conservation based on the phastCons (**Supplementary Figure 5B**). The B-D plot for this SNP shows a potential cooperative effect between ARID5A and SRY proteins, despite modest expression (**Supplementary Figure 5C**).

Based on the output of Motif-Raptor, we also performed downstream analyses to explore the potential effects of the genetic variants that modulate the binding sites of NFKB1 and ARID5A on gene expression. We studied the relative gene expression level of the potential target genes (defined by proximal genes), regulated by of NFKB1 and ARID5A. NFKB1 may activate relatively lowly-expressed genes compared with ARID5A, which may repress already highly expressed genes (**Supplementary Figure 4D**). However, this trend is not general for other TFs as it depends on the nature of the TF (i.e. an activator or repressor). To explore potential gene ontology (GO) terms for these target genes, we performed a GO enrichment analysis (Mi *et al.*, 2013) (**Supplementary Figure 7**). The most enriched terms are related to antigen processing and presentation and T cell immune functions, providing a potential mechanism of action for these genetic variants.

Finally, we explored factors for which the binding sites are significantly modulated in CD8+ and CD4+ cells but with a modest or low cell type-specific expression (expression percentile from 25% to less than 80%) (**Supplementary Figure 8**). SP3, TCFL5 and TFDP1 show this pattern in both CD8+ and CD4+ cells. For instance, when we inspect the B-D plots and score distributions (**Supplementary Figure 8C,D**) from CD8+ cells, the statistical effect of the score shifting to enhance the TFDP1 binding events is quite dramatic. From the SNP-specific B-D plot for the most significant SNP rs1611742 we observed a possible interaction of TFDP1 with SP3 and factors from the E2F family (**Supplementary Figure 8E**). To support our hypothesis, previous studies have shown an *in vivo* cooperative binding of TFDP1 and the E2F family (Helin *et al.*, 1993; Wu *et al.*, 1995). Notably, TFDP1 stimulates E2F dependent transcription which correlates with IL-6 immune response in RA patients (Zhang *et al.*, 2018).

In contrast, we did not find prior literature supporting roles of SP3 and TCFL5 in RA.

3.2 Study on red blood cell count

Red blood cell (RBC) count is one of the blood indices that can reflect normal or dysregulated hematopoiesis. GWAS data for the RBC count was downloaded from a meta-analysis of 173,480 European ancestry individuals in three large-scale UK studies (Astle *et al.*, 2016). In this study 28,722 SNPs were significantly associated to the RBC count (p-value<$8.31x10^{-9}$) and ~23% (6710) of SNPs overlap with at least one DNase I-seq peak in the 83 cell types used.

As previously suggested in other studies (Reshef *et al.*, 2018), to better recover signals across the genome, we performed our analysis with and without removing the MHC (Major Histocompatibility Complex) region from chromosome 6 (hg19:chr6:28,477,797-33,448,354). This region contains several highly polymorphic genes important for the adaptive immune system that may mask signals not associated with immunity.

Using Motif-Raptor, we first identified the most highly-associated cell types based on the enrichment of genetic variants in cell type-specific chromatin accessibility peaks. The most significant cell type is K562 with 585 SNPs in unique peaks (log-odds 1.99,p-value<10^{-30} Fisher's exact test). Although K562 is a cancer cell line, it recapitulates some aspects of erythropoiesis. One example of this is the specific expression of the erythroid-specific master regulators (Pinello *et al.*, 2018; Ulirsch *et al.*, 2016). We also uncovered T47D cells, cerebellum cells, CD8+T cells, and trophoblast cells in the top 5 associated cell types (**Supplementary Figure 9A**).

Next, using the same strategy illustrated before (transcription factor expression levels with FPKM>2, expression percentile>0.8) we uncovered and ranked 26 TF motifs that are modulated in K562 cells [Supplementary Table 2]. The most enriched were TAL1-GATA1 and HLTF ($minFDR$ <0.1) (**Supplementary Figure 9B**). These two factors were also enriched when the MHC region was included, suggesting recovery of signals independent of including this region. The modulation score distributions for TAL1-GATA1 and HLTF show a negative shift (**Supplementary Figure 9 C,D and Supplementary Figure 6B**), suggesting that the genetic variants associated with the RBC count may disrupt the binding sites of these TFs. Tal1 and Gata1 have been previously reported as key master regulators for erythropoiesis (Cantor and Orkin, 2002; Ulirsch *et al.*, 2016). To our knowledge, the association of HLTF with this process is novel and further investigation may be prudent.

Next, we focused our attention on identifying the SNPs with the largest effects based on the inspection of the B-D plots and modulation score rankings. From the B-D plots, the top SNPs for HLTF and TAL1-GATA1 are rs10758656 and rs145910606 (**Supplementary Figure 6B**). As illustrated in the radar plot, the SNP rs10758656 for HLTF is instead predicted to be deleterious based on the CADD score. However, it has a low conservation (low phastCons and UCE scores), suggesting this element may be important only in the *human* RBC (**Supplementary Figure 9E**) trait. The SNP rs145910606 is evolutionarily conserved based on the phastCons and UCE scores. Its disruption is predicted to be deleterious based on the CADD score (**Supplementary Figure 9F**). The CATO score was not available for these two SNPs (this annotation is available only for a subset of SNPs and TF motifs).

The co-binding and therefore co-disruption of TAL1 and GATA1 (this motif may correspond to other GATA family members) are shown on the B-D plot and motif logos for both rs10758656 and rs145910606 (**Supplementary Figure 9G-J**). These SNPs significantly disrupt the GATA1 motif - rs10758656 is an A to G mutation and rs145910606 is a deletion overlapping the GATA sequence. Further inspection of the B-D plot for rs10758656 (**Supplementary Figure 9G**) shows a potential co-binding of HLTF with GATA family factors. Importantly, we validated these potential co-binding effects on reference alleles through available ChIP-seq data for TAL1, GATA1 and HLTF in K562 cells (Davis *et al.*, 2018; Pope *et al.*, 2014), and verified that both SNPs are in chromatin-accessible regions (**Supplementary Figure 10**). These SNPs correspond to true binding events; in both cases they are within overlapping strong peaks for GATA1 and TAL1, suggesting their co-binding at these locations. For rs10758656, HLTF shows a weak overlapping peak with the peaks of GATA1 and TAL1. We checked the enrichment of the ChIP-seq peaks overlapping GWAS hits in K562-specific open-chromatin regions. GATA1 shows significant enrichment in SNP hits (p-value=0.002, fold-change=1.69) as expected, but HLTF only shows significance when removing the overlapping peaks with GATA1 (p-value=0.04 fold-change=2.25). These results using ChIP-seq peaks not only serve as *in silico* validation of our prioritized TFs, but also indicate that HLTF may bring regulations independent of GATA1 among top ranked SNPs. The data were downloaded from the ENCODE portal (Davis *et al.*, 2018) with the following identifiers: ENCFF341UEE (TAL1) ENCFF389WLJ (GATA1) and ENCFF830OUU (HLTF).

3.3 Robustness and generalizability of Motif-Raptor across traits

To test the robustness of our procedure in recovering biological insights we performed additional analyses. First, we assessed the ability to recover relevant cell types and TFs as different numbers of SNPs were included corresponding to different thresholds for significance (up to a nominal p-value of 0.05). These results are presented in **Supplementary Material Section 4** and **Supplementary Figure 11-14**. Briefly we observed that the prioritized cell types and the significant TFs are still reported even when relaxing the proposed thresholds by the respective studies. However, it seems beneficial for the recovery of the TFs (at least for the traits we have analyzed) to relax the GWAS proposed threshold and include more SNPs. For example, for RBC, we can recover in addition to TAL1:GATA1 additional factors previously associated to erythroid biology in K562 (KLF1, several GATA1 motifs and GFI1B). This can be rationalized by the fact that our search space is limited to the chromatin accessible sites and by including more SNPs we may capture weaker signals that may still contribute to the global modulation of these factors in these regions while controlling for potential false positives.

Second to assess the quality of recovered cell types and TFs we compared the results for related and unrelated traits. The idea here is that related traits may share similar biological mechanisms and therefore cell types and TFs while unrelated traits different ones. To this end we selected lupus (from UKBB v3), since it is more related to rheumatoid arthritis (RA) and autoimmunity, and total cholesterol (Hoffmann TJ et.al, Nature Genetics, 2018) is not related to RA nor red blood cell count. Briefly, running Motif-Raptor on variants associated with lupus we can recover immune related cell types (specifically, T cells) as we have also observed for RA (**Supplementary Figure 11A** and **13A**) and identified overlapping immune related transcription factors, e.g. STAT1, GFI1, NFKB1 (**Supplementary Figure 11B** and **13B**). Running Motif-Raptor on total cholesterol, we identified HepG2 and Ishikawa as the top two associated cell lines (**Supplementary Figure 14A**). Notably, HepG2, a liver derived line is commonly used to study the effect of genes or variants on cholesterol levels (HDL/LDL). Interestingly for this cell type we recovered factors

associated with the insulin pathway (e.g. HNF1A, ISL2, and MAFG), a pathway previously associated to cholesterol levels (Gylling et.al, Journal of Lipid Research, 2010). These analyses suggest that our procedure is robust and can prioritize relevant cell types and factors that may be implicated in different traits and that it may be important to explore different significance thresholds to recover additional TFs for some traits.

3.4 Comparison of Motif-Raptor with available tools

We compared Motif-Raptor with existing tools to perform similar tasks even if they cover only a single step of our proposed pipeline.

First, we compared step 1 of Motif-Raptor, i.e. the recovery of cell types based on a set of SNPs in DHS sites, with GREGOR (Schmidt *et al.*, 2015) a tool that can explicitly account for LD structure but that requires in input sentinel SNPs. We observed highly concordant results for both GREGOR and Motif-Raptor for all the 4 traits considered (Supplementary Table 7).

Second, we compared step 2 of Motif-Raptor with tools to scan and recover TF binding sites (MOODS), calculate the disruption based on genetic variants (atSNP), and/or test the disruption significance per SNP (atSNP) or globally (SLDP). Based on the computational requirements, features of currently available tools and the obtained results, we believe that this comparison illuminates the motivation behind development of Motif-Raptor (**Supplementary Figures 1-2**).

Finally, to test the possibility of extending Motif-Raptor beyond TF PWM models, we considered for the calculation of the TF modulation score in step 2 precomputed DeepBind models (Alipanahi *et al.*, 2015). DeepBind characterizes TF binding based on a convolutional neural network. We were not able to justify the adoption of DeepBind models in place of PWMs in Motif-Raptor. For example, although we observed similar results for RBC traits in terms of recovering Gata1, in RA the proposed DeepBind model for NFKB1 was not recovered (**Supplementary Figures 15-18 and Supplementary Table 9**). Nonetheless, we believe this integration example illustrates the key steps necessary to extend Motif-Raptor to integrate more powerful models to predict TF binding affinity in addition to PWMs.

These comparisons and testing results are presented in detail in **Supplementary Material Section 4.**

4 Discussion

We have described Motif-Raptor, a computational toolkit to study the effect of genetic variants on transcription factor binding sites in non-coding regions. These variants are associated with traits and disease phenotypes. Motif-Raptor consists of three steps that allow users to find relevant cell types, cell type-specific regions and TFs, to inspect and annotate the SNPs with TF binding effects. Importantly, our analysis does not rely on ChIP-seq tracks since they are not readily available for several cell type/TFs but on the more broadly available TF PWM models. Owing to the efficient algorithmic design, it allows to compute genome-wide null models for each TF, and exhaustively explore and quantify the relationship between SNPs and putative TF binding sites. Motif-Raptor leverages not only chromatin accessibility, but also gene expression data to filter out false positives. Further, Motif-Raptor integrates well-established annotations to score individual SNPs based on their conservation and deleteriousness.

In running our proposed three-step procedure it is important to keep in mind the following considerations and limitations.

First, the removal of MHC region as proposed elsewhere (Reshef *et al.*, 2018) could significantly alter results. The application of this strategy depends on the phenotype. For example, SNPs associated with RA are known to be enriched in the MHC region and thus the inclusion of this region may be important to characterize single RA-relevant SNPs with strong effects (Weyand and Goronzy, 2000; Newton *et al.*, 2004). On the other hand, removal of the MHC region may be desired to capture residual signals elsewhere in the genome. As for the RBC count, the inclusion or exclusion of this region seems unimportant as shown by the presented results.

Second, our approach is helpful to investigate if cell type specific regulators may be significantly disrupted by trait- or disease- associated variants. However, our approach may miss mechanisms shared across different cell type/tissues.

Third, while TF motif models (PWMs) are available for hundreds of TFs, their quality is highly variable, and our analysis might be affected. In addition, multiple models may exist for the same TF. For these reasons in Motif-Raptor we use the non-redundant JASPAR database where the best motifs for each TF are selected and strict quality filters are applied by its curators. Also, these models assume independence between nucleotides, an assumption that is some cases an oversimplification (Korhonen *et al.*, 2017). However, future versions of this tool could be extended to incorporate more advanced and specialized motif models based on Support Vector Machines (SVM) (Mordelet *et al.*, 2013), SNP effect matrix (SEMpl) (Nishizaki *et al.*, 2020), or deep learning classifiers (Alipanahi *et al.*, 2015; Movva *et al.*, 2019).

Forth, our tool measures the significance of the effect of genetic variants on the entire set of binding events for a TF rather than individual SNPs in an efficient and scalable manner. Despite the overall association of these SNPs with phenotypes through modulation of TF binding site, the individual SNP effect on the direction of the phenotype needs to be examined case by case, depending on whether a given allele might increase or decrease the risk. However, we provide external scores/annotations and interactive plots for individual SNPs to help explore the most promising variant(s) for experimental validation.

Fifth, Motif-Raptor does not identify target genes and to assess potential involvement of recovered TFs in modulating the expression of putative target genes we used a simple but imperfect heuristic i.e. we mapped each modulated binding site to the closest gene and then averaged the expression of these genes. Imprecise regulatory element to gene assignment may lead to false positives and weaken the signal, therefore it may be important in the future to explore more sophisticated methods such as ABC (Fulco et al., 2019).

Finally, we want to reiterate that Motif-Raptor was designed to prioritize TFs and provides only informative plots for individual SNPs in its current implementation. However, future extensions or downstream analyses are required to assess the significance of individual SNPs and their putative target genes. Also Motif-Raptor doesn't implement any preprocessing or filtering step for the summary statistic files in input, therefore despite the assumption that these files are comprehensive and accurate, the biological insights and the prioritized factors we can recover may depend on the quality of these files.

In summary, Motif-Raptor is a computational toolkit to test the significance of the effects of genetic variants from GWAS analyses on transcription factor binding sites. We believe that its adoption will help the genomic community in prioritizing potential cell type-specific, causal variants from GWAS summary statistics and to generate important hypotheses and insights to the mechanisms of action of genetic variants in complex disease.

Acknowledgements

We thank people in the Pinello and Bauer labs for testing the software and for helpful discussions.

Funding

D.B. was supported by grants from the National Institute of Health (P01HL032262 and DP2HL137300). P.F. was supported in part by the European Integrated Infrastructure for Social Mining and Big Data Analytics (SoBigData++, EU Grant Agreement #871042). L.P. was supported by grants from the National Institute of Health (R00HG008399 and R35HG010717).

Conflict of Interest: none declared.

References

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Amariuta,T. *et al.* (2019) IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors. *Am. J. Hum. Genet.*, **104**, 879–895.

Astle,W.J. *et al.* (2016) The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, **167**, 1415-1429.e19.

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Barrera,L.A. *et al.* (2016) Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science (80-.).*, **351**, 1450–1454.

Boyle,A.P. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, **22**, 1790–1797.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

Cantor,A.B. and Orkin,S.H. (2002) Transcriptional regulation of erythropoiesis: An affair involving multiple partners. *Oncogene*, **21**, 3368–3376.

Davis,C.A. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

Deplancke,B. *et al.* (2016) The Genetics of Transcription Factor DNA Binding Variation. *Cell*, **166**, 538–554.

Eyre,S. *et al.* (2006) Association of the FCRL3 gene with rheumatoid arthritis: A further example of population specificity? *Arthritis Res. Ther.*, **8**.

Fu,S.H. *et al.* (2017) New insights into Blimp-1 in T lymphocytes: A divergent regulator of cell destiny and effector function. *J. Biomed. Sci.*, **24**.

Fulco,C.P. *et al.* (2019) Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*

Gallagher,M.D. and Chen-Plotkin,A.S. (2018) The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.*, **102**, 717–730.

De Gobbi,M. *et al.* (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science (80-.).*, **312**, 1215–1217.

Grant,C.E. *et al.* (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Gusfield,D. (1997) Algorithms on Stings, Trees, and Sequences. *ACM SIGACT News*, **28**, 41–60.

Handel,M.L. *et al.* (1995) Nuclear factor–kB in rheumatoid synovium. Localization of P50 and P65. *Arthritis Rheum.*, **38**, 1762–1770.

Helin,K. *et al.* (1993) Heterodimerization of the transcription factors E2F-1 and DP-1 leads to cooperative trans-activation. *Genes Dev.*, **7**, 1850–1861.

Holmes,S.P. and Gusfield,D. (1999) Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. *J. Am. Stat. Assoc.*, **94**, 989.

Huang,D.Y. *et al.* (2005) GATA-1 mediates auto-regulation of Gfi-1B transcription in K562 cells. *Nucleic Acids Res.*, **33**, 5331–5342.

Jiang,Y. *et al.* (2012) Meta-Analysis of 125 Rheumatoid Arthritis-Related Single Nucleotide Polymorphisms Studied in the Past Two Decades. *PLoS One*, **7**.

Kasai,T. *et al.* (2001) Linear-time longest-common-prefix computation in suffix arrays and its applications. In, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*., pp. 181–192.

Kelley,D.R. *et al.* (2016) Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.

Khan,A. *et al.* (2018) JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.

Kim,W. *et al.* (2014) Gfi-1 regulates the erythroid transcription factor network through Id2 repression in murine hematopoietic progenitor cells. *Blood*, **124**, 1586–1596.

Korhonen,J. *et al.* (2009) MOODS: Fast search for position weight matrix matches in DNA sequences. *Bioinformatics*, **25**, 3181–3182.

Korhonen,J.H. *et al.* (2017) Fast motif matching revisited: High-order PWMs, SNPs and indels. *Bioinformatics*, **33**, 514–521.

Korman,B.D. *et al.* (2008) STAT4: genetics, mechanisms, and implications for autoimmunity. *Curr. Allergy Asthma Rep.*, **8**, 398–403.

Lambert,S.A. *et al.* (2018) The Human Transcription Factors. *Cell*, **172**, 650–665.

Link,V.M. *et al.* (2018) MMARGE: Motif mutation analysis for regulatory genomic elements. *Nucleic Acids Res.*, **46**, 7006–7021.

Liu,T. *et al.* (2017) NF-κB signaling in inflammation. *Signal Transduct. Target. Ther.*, **2**.

Makarov,S.S. (2001) NF-KB in rheumatoid arthritis: A pivotal regulator of inflammation, hyperplasia, and tissue destruction. *Arthritis Res.*, **3**, 200–206.

Matsumoto,T. *et al.* (2006) Infliximab for rheumatoid arthritis in a patient with tuberculosis [16]. *N. Engl. J. Med.*, **355**, 740–741.

Maurano,M.T. *et al.* (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.*, **47**, 1393–1401.

Maurano,M.T. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-.).*, **337**, 1190–1195.

Mi,H. *et al.* (2013) Large-scale gene function analysis with the panther classification system. *Nat. Protoc.*, **8**, 1551–1566.

Mordelet,F. *et al.* (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. In, *Bioinformatics*.

Movva,R. *et al.* (2019) Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One*, **14**.

Moyerbrailean,G.A. *et al.* (2016) Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLoS Genet.*, **12**.

Newton,J.L. *et al.* (2004) A review of the MHC genetics of rheumatoid arthritis. *Genes Immun.*, **5**, 151–157.

Nishizaki,S.S. *et al.* (2020) Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics*, **36**, 364–372.

Oeckinghaus,A. and Ghosh,S. (2009) The NF-kappaB family of transcription

factors and its regulation. *Cold Spring Harb. Perspect. Biol.*, **1**.

Okada,Y. *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**, 376–381.

Okamoto,H. *et al.* (2008) Molecular aspects of rheumatoid arthritis: Role of transcription factors. *FEBS J.*, **275**, 4463–4470.

Pinello,L. *et al.* (2014) Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. *Proc. Natl. Acad. Sci. U. S. A.*, **111**.

Pinello,L. *et al.* (2018) Haystack: Systematic analysis of the variation of epigenetic states and cell-type specific regulatory elements. *Bioinformatics*, **34**, 1930–1933.

Pique-Regi,R. *et al.* (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.

Pope,B.D. *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402–405.

Puel,A. and Casanova,J.L. (2018) Arid5a makes the IL-17A/F-responsive pathway less arid. *Sci. Signal.*, **11**.

Puglisi,S.J. *et al.* (2007) A taxonomy of suffix array construction algorithms. *ACM Comput. Surv.*, **39**.

Remmers,E.F. *et al.* (2007) STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N. Engl. J. Med.*, **357**, 977–986.

Reshef,Y.A. *et al.* (2017) Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk.

Reshef,Y.A. *et al.* (2018) Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.*, **50**, 1483–1493.

Rutishauser,R.L. *et al.* (2009) Transcriptional Repressor Blimp-1 Promotes CD8+ T Cell Terminal Differentiation and Represses the Acquisition of Central Memory T Cell Properties. *Immunity*, **31**, 296–308.

Schmidt,E.M. *et al.* (2015) GREGOR: Evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*.

Simmonds,R.E. and Foxwell,B.M. (2008) Signalling, inflammation and arthritis: NF-κB and its relevance to arthritis and inflammation. *Rheumatology*, **47**, 584–590.

Tran,C.N. *et al.* (2005) Synovial biology and T cells in rheumatoid arthritis. *Pathophysiology*, **12**, 183–189.

Ulirsch,J.C. *et al.* (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*, **165**, 1530–1545.

Ward,L.D. and Kellis,M. (2012) HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**.

Weinhold,N. *et al.* (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.*, **46**, 1160–1165.

Weyand,C.M. and Goronzy,J.J. (2000) Association of MHC and rheumatoid arthritis HLA polymorphisms in phenotypic variants of rheumatoid arthritis. *Arthritis Res.*, **2**, 212–216.

Whyte,W.A. *et al.* (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.

Wienert,B. *et al.* (2015) Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. *Nat. Commun.*, **6**.

Wu,C.L. *et al.* (1995) In vivo association of E2F and DP family proteins. *Mol. Cell. Biol.*, **15**, 2536–2546.

Zeng,H. *et al.* (2016) GERV: A statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*, **32**, 490–496.

Zhang,R. *et al.* (2018) A critical role of E2F transcription factor 2 in proinflammatory cytokines-dependent proliferation and invasiveness of fibroblast-like synoviocytes in rheumatoid Arthritis. *Sci. Rep.*, **8**.

Zhao,S.X. *et al.* (2013) A Refined Study of FCRL Genes from a Genome-Wide Association Study for Graves' Disease. *PLoS One*, **8**.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.

Zuo,C. *et al.* (2015) AtSNP: Transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics*, **31**, 3353–3355.