frontiers in Genetics

METHODS
published: 29 May 2018
doi: 10.3389/fgene.2018.00189

Removing Background Co-occurrences of Transcription Factor Binding Sites Greatly Improves the Prediction of Specific Transcription Factor Cooperations

Cornelia Meckbach[1]*, *Edgar Wingender*[1] and *Mehmet Gültas*[1,2,3]

[1] Institute of Bioinformatics, University Medical Center Göttingen, Georg-August-University Göttingen, Göttingen, Germany, [2] Department of Breeding Informatics, Georg-August University Göttingen, Göttingen, Germany, [3] Center for Integrated Breeding Research (CiBreed), Georg-August University Göttingen, Göttingen, Germany

Today, it is well-known that in eukaryotic cells the complex interplay of transcription factors (TFs) bound to the DNA of promoters and enhancers is the basis for precise and specific control of transcription. Computational methods have been developed for the identification of potentially cooperating TFs through the co-occurrence of their binding sites (TFBSs). One challenge of these methods is the differentiation of TFBS pairs that are specific for a given sequence set from those that are ubiquitously appearing, rendering the results highly dependent on the choice of a proper background set. Here, we present an extension of our previous PC-TraFF approach that estimates the background co-occurrence of any TF pair by preserving the (oligo-) nucleotide composition and, thus, the core of TFBSs in the sequences of interest. Applying our approach to a simulated data set with implanted TFBS pairs, we could successfully identify them as sequence-set specific under a variety of conditions. When we analyzed the gene expression data sets of five breast cancer associated subtypes, the number of overlapping pairs could be dramatically reduced in comparison to our previous approach. As a result, we could identify potentially cooperating transcriptional regulators that are characteristic for each of the five breast cancer subtypes. This indicates that our approach is able to discriminate specific potential TF cooperations against ubiquitously occurring combinations. The results obtained with our method may help to understand the genetic programs governing specific biological processes such as the development of different tumor types.

Keywords: transcription factor (TF), TF cooperations, sequence-set specific TF cooperations, background correction, TF co-occurrences

OPEN ACCESS

Edited by:
Alexandre V. Morozov, Rutgers University, The State University of New Jersey, United States

Reviewed by:
Vladimir B. Teif, University of Essex, United Kingdom
Hauke Busch, Universität zu Lübeck, Germany

***Correspondence:**
Cornelia Meckbach
cornelia.meckbach@bioinf.med.uni-goettingen.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *16 February 2016*
Accepted: *08 May 2018*
Published: *29 May 2018*

Citation:
Meckbach C, Wingender E and Gültas M (2018) Removing Background Co-occurrences of Transcription Factor Binding Sites Greatly Improves the Prediction of Specific Transcription Factor Cooperations. Front. Genet. 9:189. doi: 10.3389/fgene.2018.00189

1. INTRODUCTION

Transcription factors (TFs) are a special class of cellular proteins that are essential for controlling different genetic programs such as adaption to the environment, immune response, organogenesis or embryonic development by regulating gene expression. The human genome encodes roughly 1500–2000 different TFs which bind to short degenerate DNA motifs, known as transcription factor

binding sites (TFBSs). In higher organisms, the binding of TFs occurs in a specific combination within DNA regulatory regions (promoters as well as distal elements, such as enhancers) to form purposive dimers or higher order complexes to activate or repress their target genes. Due to the fact that eukaryotic DNA is packed in chromatin, TFs show additionally competing or cooperative DNA binding with chromatin associated proteins (Teif and Rippe, 2010). Besides this, based on the co-occurence of their TFBSs TFs exert functional cooperations which play an important role in the regulation of the different genetic programs in mammals (Boyer et al., 2005; Hu and Gallo, 2010; Neph et al., 2012). Today, it is well-known that the selection of cooperation partners for TFs depends on their biological functions, e.g., cell cycle control, cell homeostasis, or cell differentiation in different cell types. As a result of these properties, TFs change their partners to specify their functions according to the cellular context.

In the last decade, a various number of computational methods for the identification of cooperating TFs has been proposed (Hu et al., 2007; Van Loo and Marynen, 2009; Girgis and Ovcharenko, 2012; Ha et al., 2012; Sun et al., 2012; Deyneko et al., 2013; Nandi et al., 2013; Jankowski et al., 2014; Navarro et al., 2014; Meckbach et al., 2015; Wu and Lai, 2016; Spadafore et al., 2017). Among these methods, predicting the putative TFBSs in the sequences under study and building a meaningful quantification measure of the cooperation between two TFs are two essential steps to make the predictions successful. Based on these steps, different strategies/ideas have been used for the identification of cooperating TF pairs such as the TFBS co-occurrences of cooperative pairs are more often than expected by chance and have significantly closer distances. In this context, several methods such as statistical methods like the hypergeometric test, clustering approaches, randomized occurrence frequency model (OF_r) or Markov models have been developed (Hu et al., 2007; Chuang et al., 2009; Girgis and Ovcharenko, 2012; Ha et al., 2012; Mysickova and Vingron, 2012; Sun et al., 2012; Nandi et al., 2013; Jankowski et al., 2014; Lai et al., 2014; Navarro et al., 2014; Spadafore et al., 2017).

Employing a comprehensive performance evaluation study on the prediction results of those methods, Lai et al. (2014) have shown that the success rates of different approaches strongly depend on the corresponding evaluation criteria. This finding is also supported by our results, which we have presented in Meckbach et al. (2015). However, the predictions of almost all of these methods suffer from many types of obstacles that might occur as a result of high background like common regulatory programs between cell types and the environmental components in their regulatory sequences like GC content or nucleotide composition - indicating the ratio of the constituent monomer units/bases- as well as the noise effect of false positive putative TFBSs. Hence, such obstacles lead into background co-occurrence of TFBSs and consequently the results of a certain method are often highly overlapping for different sequence sets. Zeidler et al. (2016) have clearly demonstrated this problem in their study for detection of stage-specific TF pairs in a time series data set during heart development. To overcome this problem, they have further applied Markov clustering algorithm (MCL) (Dongen, 2000) to the pairs predicted by MatrixCatch methodology (Deyneko et al., 2013). Although several negligible TF cooperations could be eliminated, the application of MCL algorithm in this context is only based on the observed frequencies of TFBSs and does not consider the sequence specific environmental components. Consequently, the results of this approach seem to be conservative and not sequence set specific, yet.

To deal with this problem to some extent, we applied in our previous study the average product correction (APC) theorem (Dunn et al., 2008) in order to determine for each TFBS pair their background co-occurrence resulting from their possibly false positive TFBS predictions in the entire sequence set under study. Although, with respect to APC theorem, the background noise effect of false positive TFBSs could be successfully eliminated in the detection of significant TF pairs, the power and functionality of APC theorem appears to be insufficient to handle the remaining obstacles for the identification of sequence-set specific TF cooperations. In order to overcome the missing point of PC-TraFF workflow (Meckbach et al., 2015), we propose in this study an efficient approach that accurately quantifies the level of background co-occurrence of two TFBSs considering different types of obstacles (mentioned above) in the sequences under study. For this purpose, by preserving the (oligo-) nucleotide composition of the sequences of interest, we create a sufficient number of new shuffled sequence sets and based on these sets the background co-occurrence of a TFBS pair is measured. This process ensures that TF cooperations, which are very sensitive regarding the context of nucleotides and the distance of their binding sites, will become remarkable small background-values in comparison to common (ubiquitously occurring) TF pairs. These ubiquitously occurring TF pairs are often found as significant for different sequence sets and are less susceptible to the behavior of their binding sites in the set of sequences. Consequently, removal of this background leads to the separation of sequence set-specific TF pairs from the common ones.

To demonstrate the performance and functionality of our proposed approach, we analyzed a simulation data set as well as five breast cancer subtype-associated gene sets, and present the results step by step by providing comparative analysis. These data sets have been chosen because the importance of cooperating TF pairs have been well-studied in Meckbach et al. (2015).

Terminology

For the sake of simplicity, we adapt the terminology of our previous paper (Meckbach et al., 2015). In doing so, each match of a position weight matrix (PWM) with a segment of genomic DNA is called a (potential) *transcription factor binding site* (TFBS). TFBSs are represented by names of their corresponding PWMs. The PWMs of TRANSFAC (Wingender, 2008) used in this report are denoted with their TRANSFAC identifiers, the structure of which is: *V*$*factorname_version*, where "V$" indicates that the PWM is representing a TFBS of a vertebrate TF. *factorname* refers to the TF name, while there are more than one PWM representing the binding motif of a certain factor, *version* is required for the unambiguous identification of the PWM. TFBS pairs refer to co-occurring TFBSs. It is important to note that we

cannot make any statement about the kind of interaction such co-occurrence may be associated with (cooperativity, synergistic or antagonistic interaction etc.). The term cooperation refers to any kind of functional cooperation and/or physical interaction between the constituents of the predicted TFBS pairs.

2. RESULTS AND DISCUSSION

In this study, we introduce an extension of our previous methodological approach PC-TraFF for the separation of sequence-set specific cooperating transcription factors based on the co-occurrence of their binding sites from common ones. The overall workflow of our approach comprises two parts. First, the original PC-TraFF algorithm is used in order to predict significant TFBS pairs in a set of sequence where PC-TraFF provides for each significant TFBS pair t_a and t_b a pointwise mutual information score $\mathbb{PMI}_{pc}^{APC}(t_a; t_b)$. Thereby, the minimal and maximal distance threshold for two TFBSs to form a pair is set to 5 and 20 bp, respectively, in order to provide a proper comparison to the original PC-TraFF-results.

Second, in order to separate PC-TraFF significant TFBS pairs into the two groups of sequence-set specific and common (generally important) combinations, we apply our extension approach. For this purpose, out of the sequences of interest, a sufficiently large number of background sets is created by shuffling the original sequences, whereby the general nucleotide composition of the sequences as well as the core of the putative TFBSs are maintained. For all these background sets, the original PC-TraFF algorithm is applied to calculate $\mathbb{PMI}_{pc}^{APC}$-values between all TFBS pairs. Afterwards, using these values the level of average background cooperation, which is defined as $AVG\left(\mathbb{PMI}(t_a; t_b)\right)$-value, between two TFs based on their binding sites over all sets of background sequences is calculated. The subtraction of $AVG\,(\mathbb{PMI})$-values from their initial $\mathbb{PMI}_{pc}^{APC}$-values results in the separation of sequence-set specific pairs from the common co-occurrences. To this end, we additionally introduced a factor $\alpha \in [-1, 1]$ to enlarge/reduce the effect of the subtracted background level by linearly influencing the subtracted average value $AVG\left(\mathbb{PMI}(t_a; t_b)\right)$. If $\alpha = 1$, the $2\times AVG\left(\mathbb{PMI}(t_a; t_b)\right)$-value is subtracted from the initiate $\mathbb{PMI}_{pc}^{APC}$-value, $\alpha = 0$ results simply in the subtraction of the observed $AVG\left(\mathbb{PMI}(t_a; t_b)\right)$ value, while an α-value of -1 results in the original PC-TraFF predictions. Thus, α enlarges/reduces the level of the subtracted background and is thereby influencing the number of identified specific pairs. However, our results suggest that the impact of α on the number of specific pairs strongly depends on the individual sequence sets and appears not to be linear (e.g., see **Figure 1**) although the factor itself has a linear influence on the subtracted background level.

It is important to note that the Results section of this study mainly considers the influence of our proposed extension approach on the cooperating TFs identified by the PC-TraFF algorithm. Researchers, who are interested in the biological functions of individual TF cooperations, are kindly referred to the original PC-TraFF paper (Meckbach et al., 2015).

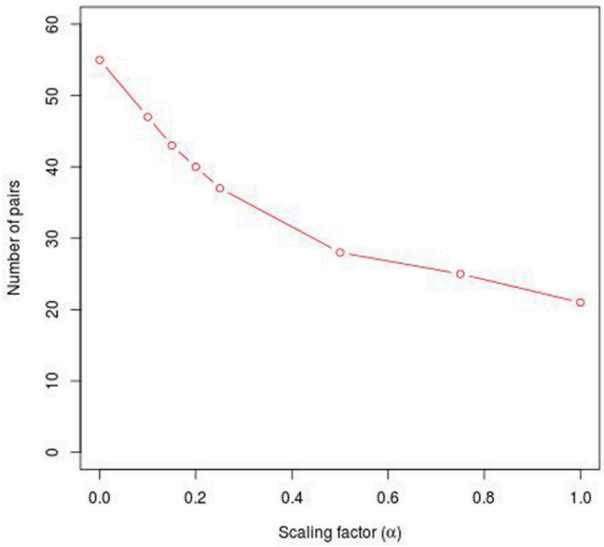

FIGURE 1 | Number of specific TFBS pairs for the synthetic sequence set in dependence on different α-values. The synthetic sequence set consists of 200 sequences of length 1000 bps, each of these sequences contains artificially inserted binding site pairs (V$IRF1_01 - V$USF_01) for the cooperation between transcription factors IRF1 and USF1 with a minimal distance of 5 bp and a maximal distance of 20 bp. The α-value linearly influences the subtracted background level (e.g., $\alpha = 0$ results in the subtraction of the $AVG\left(\mathbb{PMI}(t_a; t_b)\right)$ value, $\alpha = 1$ indicates the subtraction of the $2\times AVG\left(\mathbb{PMI}(t_a; t_b)\right)$-value) .

TABLE 1 | Total number of specific TFBS pairs for the simulation data set using different α-values.

α-value	Rank of artificially inserted pair	Total number of pairs found
$\alpha = -1$	18	58
$\alpha = 0$	16	55
$\alpha = 0.1$	15	47
$\alpha = 0.15$	14	43
$\alpha = 0.2$	12	40
$\alpha = 0.25$	11	37
$\alpha = 0.5$	6	28
$\alpha = 0.75$	6	25
$\alpha = 1$	5	21

The rank according to z-score *indicates the position of the inserted pair. The scaling factor $\alpha = -1$ indicates the significant TFBS pairs identified by the original* PC-TraFF *algorithm.*

2.1. Analysis of Simulation Data

Analyzing the sequences in the simulation data set, the original PC-TraFF algorithm identified 58 TFBS pairs as significant ($\alpha = -1$), where the artificially inserted binding site pair of the cooperating transcription factors IRF1 and USF1 is on position 18 according to *z-score* ranking. However, applying our extension approach to the results of PC-TraFF, only three of the 58 significant pairs were determined as common ones (see **Table 1**) based on the calculated background co-occurence of TFBSs ($\alpha = 0$). This rather low number of common pairs indicates that in a unspecific sequence set, the quantification

of correct background could be difficult which, in the worst case, may cause that sequence-set specific cooperations cannot be separated from common ones. To overcome this problem, the consideration of the scaling factor α is important. **Figure 1** shows the influence of α on the results. Although a variety of pairs are eliminated by means of different scaling factors, the inserted pair has been identified as sequence-set specific for each α-value. Considering the *z-score* ranking of TFBS pairs, the position of the inserted pair is rising with an increasing α-value (see **Table 1**). It has to be noted that the inserted binding sites are also matched by other PWMs, resulting in a variety of additional artificially arising TFBS pairs that consequently appear to be specific for the given sequence set.

2.2. Analysis of Breast Cancer Subtype Associated Promoter Sequences

Applying the original PC-TraFF algorithm to each BRC-subtype associated promoter sequences, we observed: (i) 62 TFBS pairs for *Luminal A*; (ii) 63 pairs for *Luminal B*; (iii) 68 pairs for *Basal-like*; (iv) 49 pairs for *Normal-like*; and (v) 62 pairs for *ErbB2 over-expressing* data set as significant. A comparison between these pairs shows that there are several pairs found as significant for more than one BRC-subtype (see **Figure 2A**), although the promoter sequences in all subtypes are unique (not overlapping). The reason of these overlapping pairs could be due to the same origin of the data and common regulatory programs which interfere with the identification of BRC-subtype specific TF cooperations.

To reveal the BRC-subtype specific TF cooperations, we additionally applied our extension approach using different α-values to these significant pairs. The results of this analysis indicate that the scaling factor α dramatically influences the number of sequence-set specific TFBS pairs. For example, on average 90% of the significant pairs have been determined as sequence-set specific by setting $\alpha = 0$, and 66% or 35% of significant pairs are assigned as sequence-set specific by setting $\alpha = 0.2$ or $\alpha = 0.5$, respectively (**Figure 3**). Further, **Figure 3** shows that, the influence of the scaling factor α is not consistent between the different sequence sets. While the number of specific TFBS pairs detected for *Luminal A* promoter sequences is dramatically decreasing and finally, 1% of all significant pairs have been determined as specific, the number of specific pairs for *ErbB2 over-expressing* promoter sequences has only slightly decreased in accordance with the increment of α-value and in

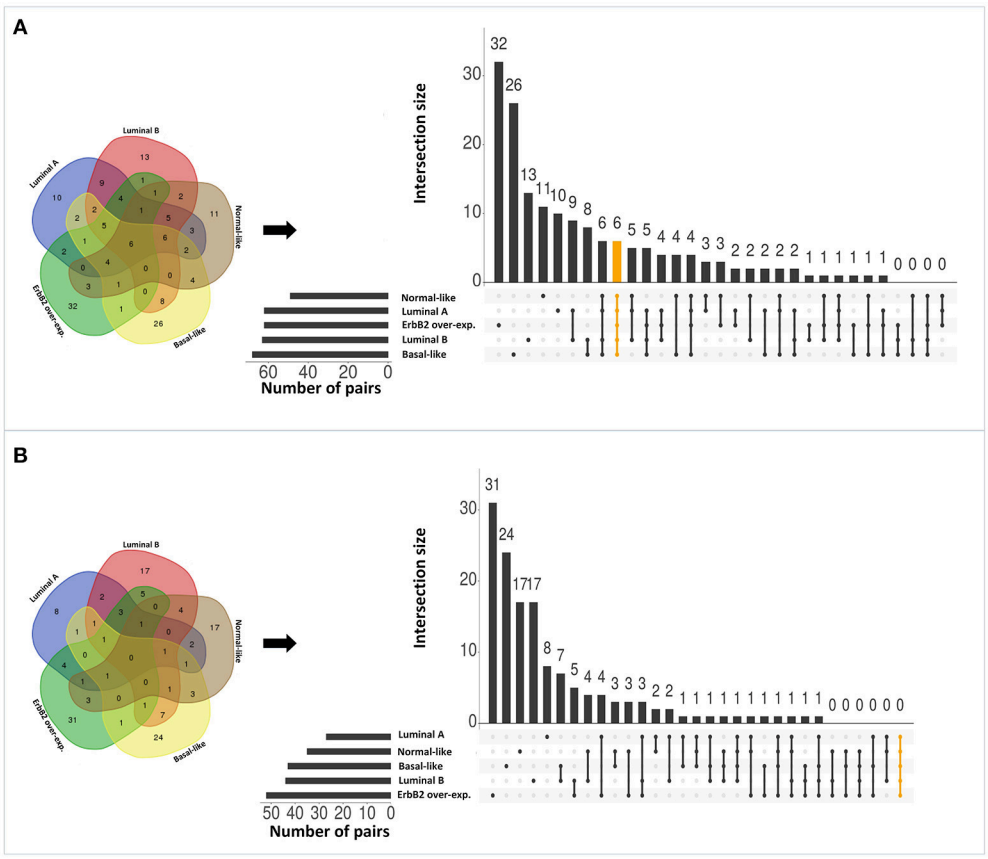

FIGURE 2 | Number of significant TFBS pairs of five BRC-subtypes and their overlap represented in Venn diagrams and in matrix layouts using UpSet technique (Conway et al., 2017). Dark circles in the matrix layout indicate subtypes that are part on the intersection. Orange lines highlight the intersection between all BRC-subtypes. **(A)** Pairs identified by the original PC-TraFF version. **(B)** Sequence-set specific pairs determined by our extension approach using a scaling factor $\alpha = 0.2$.

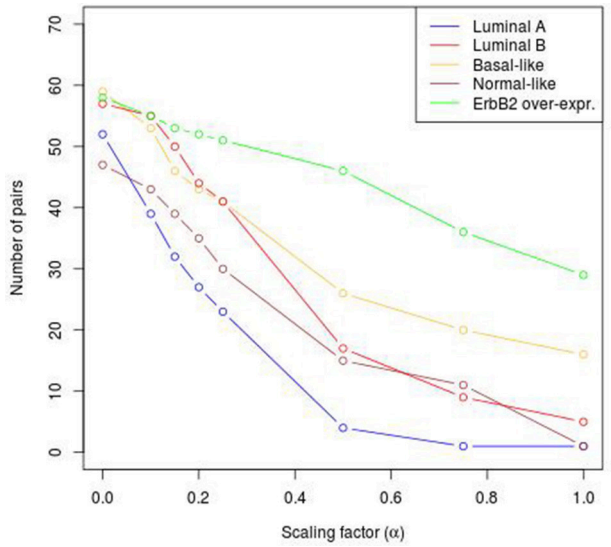

FIGURE 3 | Number of sequence-set specific pairs found in the promoter sequences of differentially expressed genes of five BRC-subtypes depending on the α-value. The α-value linearly influences the subtracted background level (e.g., $\alpha = 0$ results in the subtraction of the mean, $\alpha = 1$ indicates the subtraction of the $2\times$ *AVG* $(\mathbb{PMI}(t_a; t_b))$-value).

TABLE 2 | Pairs that were identified as significant by PC-TraFF algorithm ($\alpha = -1$) for different BRC-subtypes but are specific solely for a certain subtype using an α-value of 0.2 for the background correction.

Specific for subtype	TFBS pairs	Significant in subtypes
Normal-like	V$CEBPB_02 - V$HMGIY_Q6	*Basal-like, Luminal A, Luminal B, Normal-like*
	V$ELK1_02 - V$CETS1P54_01	*Basal-like, Luminal A, Luminal B, Normal-like*
	V$CEBPB_02 - V$CEBP_Q2	*ErbB2 over-expressing, Luminal B, Normal-like*
	V$NFKB_Q6 - V$SP1_Q4_01	*Luminal A, Normal-like*
	V$EGR_Q6 - V$AHRHIF_Q6	*Basal-like, Normal-like*
	V$GR_Q6_01 - V$PR_Q2	*ErbB2 over-expressing, Normal-like*
Luminal B	V$CETS1P54_01 - V$AHRHIF_Q6	*Luminal A, Luminal B, Normal-like*
	V$E2F_Q3_01 - V$PEBP_Q6	*Luminal A, Luminal B*
	V$MYCMAX_B - V$AHRHIF_Q6	*Basal-like, Luminal A, Luminal B*
	V$NFKB_Q6 -V$E2F_Q3_01	*Luminal A, Luminal B*
	V$NFKB_Q6 -V$AHRHIF_Q6	*Luminal A, Luminal B*
	V$CETS1P54_01- V$CP2_02	*Luminal A, Luminal B*
	V$CETS1P54_01 -V$MYCMAX_B	*Basal-like, Luminal A, Luminal B, Normal-like*

an extreme case ($\alpha = 1$) 47% of significant pairs in this subtype are assigned as specific. In addition, **Figure 2** depicts in detail for $\alpha = 0.2$ the differences between significant and specific pairs for any BRC-subtype. By considering the sequence-set specific pairs, it is remarkable that like in the original PC-TraFF analysis, the *Luminal A* promoter sequence set has the lowest number of unique pairs (eight), and *ErbB2 over-expressing* promoter sequences have the largest number of unique TFBS pairs. The intersection of all BRC-subtypes specific pairs is zero.

Interestingly, after applying our extension approach, there are more sequence-set specific unique pairs for *Normal-like* and *Luminal B* subtypes (**Figure 2B**) than significant unique pairs (**Figure 2A**). For *Normal-like* data set, there are 11 significant and 17 specific unique pairs. In particular, six pairs that were identified in the original PC-TraFF analysis for several subtypes are determined to be solely sequence-set specific for *Normal-like* subtype. For example, the pairs (V$CEBP_02 – V$HMGIY_Q6) and (V$ELK1_02 – V$CETS1P54_01) are significant for four different breast cancer subtypes or the pair (V$CEBPB_02 – V$CEBP_Q2) is significant in the originial PC-TraFF version for three BRC-subtypes, but they are sequence-set specific only for *Normal-like* subtype (for details see **Table 2**).

For *Luminal B* subtype, 13 pairs were uniquely identified as significant by the original PC-TraFF algorithm and 17 pairs were uniquely assigned as specific. In this case, seven pairs that were common in the original PC-TraFF analysis have been determined to be sequence-set specific only for *Luminal B* subtype. Further, three of the unique significant pairs (V$MYB_Q5_01 – V$MAF_Q6_01, V$NFKB_Q6 – V$CP2_02, V$HMGIY_Q6 – V$MAF_Q6_01) were assigned as common co-occurences according their negative $\mathbb{PMI}^{specific}$-values.

Besides this, there are further six pairs identified by the original PC-TraFF algorithm as significant for all five BRC-subtypes, but they are assigned to be specific only for some of these subtypes (for details see **Figure 2** and **Table 3**). For example the TFBS pair (V$CEBPB_02 – V$STAT6_01) indicating the cooperation between the transcription factors CEBPB and STAT6 can still be found in the sequence-set specific pairs of *Luminal A, Luminal B* and *Basal-like* subtypes. In contrast, the pairs (V$MYCMAX_B – V$E2F_Q3_01) and (V$STAT6_01 – V$HMGIY_Q6) have been determined as specific only for *Basal-like* and *Normal-like* promoter sequence sets, respectively.

Finally, we built up cooperation networks based on the significant TFBS pairs, where the nodes refer to TFBSs and edges to predicted co-occurrences and thus, to cooperations between them, in order to demonstrate in an exemplary way the comparative analysis between the results of our extension approach and those of the original PC-TraFF algorithm. The cooperation network based on PC-TraFF significant TFBS pairs for *Luminal A* subtype (see **Figure 4**) consists of 33 nodes and 62 edges. Reducing the network by only considering sequence-set TFBS pairs results in the elimination of 7 nodes and 35 edges. Consequently, the remaining part of the network is built up of 26 nodes with their 27 sequence-set specific cooperations (edges). It is remarkable that some TFBSs that serve as hubs in the original network are still hub nodes in the reduced network but show a lower number of neighboring nodes (e.g., V$CETS1P54_01, V$MYB_Q5_01, and V$HMGIY_Q6). On the other side, there are some highly connected nodes of the original network that are missing in the specific pair network. For example the degree

TABLE 3 | TFBS pairs, which were identified as significant by original PC-TraFF algorithm for all five BRC-subtypes but were determined as specific only in certain subtypes.

TFBS pair	Specific for subtype(s)	Pairs documentation
V$CETS1P54_01 - V$ETS_Q4	*ErbB2 over-expressing, Luminal A*	BioGRID, TransCompel®
V$MYCMAX_B - V$E2F_Q3_01	*Basal-like*	TransCompel®
V$CEBPB_02 - V$STAT6_01	*Luminal A, Luminal B, Basal-like*	TransCompel®
V$STAT6_01 - V$HMGIY_Q6	*Normal-like*	-
V$CETS1P54_01 - V$NFKB_Q6	*Luminal A, Normal-like, Basal-like*	TransCompel®
V$AP1_Q2_01 - V$AP1_Q4_01	*Luminal A, Luminal B, ErbB2 over-expressing*	BioGRID, TransCompel®

The last column indicates the databases that document the evidence for these pairs. For this purpose, we used TRANSCompel® (Kel-Margoulis et al., 2002) and BioGRID interaction database (Chatr-aryamontri et al., 2014), which contain experimentally proven pairs.

of V$NFKB_Q6 or V$AHRIF_Q6 decreases from six neighbors to one neighbor and V$SP1_Q4_01 is totally missing in the network of specific pairs. The node representing the binding site V$SMAD_Q6_01 lost just one of its neighbors in this network and thereby, it is among the 25% nodes of highest degree.

A closer look at the cooperation network of significant TFBS pairs identified for the *Basal-like* data set discloses that 43 out of 68 significant pairs have been assigned to be sequence-set specific based on our extension approach with a scaling factor $\alpha = 0.2$ (see **Figure 5A**). Setting $\alpha = 0.5$ for this analysis leads to elimination of the vast majority of the pairs and consequently 16 pairs have been determined to be specific in the promoter sequences of *Basal-like* subtype (see **Figure 5B**). A comparison between cooperation networks of *Luminal A* and *Basal-like* subtypes suggests that by considering the same scaling factor our extension approach has more influence on significant pairs found for *Luminal A* data set than those found for *Basal-like* data set. The reason for this finding might be that *Basal-like* data set is more specific than *Luminal A* data set regarding to transcriptional regulation. Thus, the level of background co-occurrence of TFBSs resulting from common regulatory programs seems to be remarkable higher in *Luminal A* data set than those of *Basal-like* data set.

3. METHODS

3.1. Data Sets

In order to assess the effectiveness of our approach and to present a detailed comparison with the results of original PC-TraFF algorithm, we analyzed in this study the data sets that have already been reported in Meckbach et al. (2015). The first data set is a simulation data set consisting of 200 sequences with the length of 1000 bps. Each of these sequences contains artificially inserted binding site pairs (V$IRF1_01 - V$USF_01) for the cooperation between transcription factors IRF1 and USF1 with a minimal distance of 5 bp and a maximal distance of 20 bp. For the two inserted binding sites we used the consensus sequences given by the position weight matrices V$IRF1_01 and V$USF_01, respectively.

The second data set is a breast cancer (BRC) gene set determined by Sorlie et al. (2003) and taken from Joshi et al. (2012). The genes have been identified based on their differential mRNA expression behavior in cancer cells and are grouped according to their expression pattern into the five molecular breast cancer-associated subtypes: Luminal A, Luminal B, Normal-like, ErbB2 over-expressing and Basal-like using hierarchical clustering (Sorlie et al., 2003). Our analysis is based on the promoter sequences of the associated genes. The number of genes as well as their corresponding promoter sequences (−500 bp to +100 bp relative to the transcription start site defined by Joshi et al. (2012) in each subtype are given in **Table 4**. It can be seen that the BRC-subtype data sets differ in the number of genes and consequently in the number of promoter sequences. For example, *Luminal A* gene set appears to be the largest set by consisting of 86 promoter sequences and in turn, the set *ErbB2 over-expressing* is the smallest sequence set by owning 15 promoter sequences (see **Table 4**). Such differences are important and make it possible to demonstrate the functionality of our extension approach for different sequence-set sizes.

The Methods section of this study comprises two main parts. First, we review our previous work PC-TraFF (Meckbach et al., 2015) so that the readers have sufficient background information to understand the proposed extension in the PC-TraFF workflow. After that, we present our proposed extension approach for the separation of sequence-set specific TF cooperations from common (generally important) ones.

Previous Work: Introduction to PC-TraFF

PC-TraFF is an information theory based method that uses the pointwise mutual information ($\mathbb{PMI}$) for the identification of potentially cooperating transcription factors according to their binding site pattern in a set of sequences. The algorithm of PC-TraFF comprises six phases and provides for each TFBS-pair t_a and t_b a $\mathbb{PMI}_{pc}(t_a, t_b)$-value based on their distances and frequencies in the sequences, under study.

The overall workflow of PC-TraFF can be briefly given as:

3.1.1. Phase 1: Construction and Filtering of the TFBS-Sequence Matrix

In the first step we predict all transcription factor binding sites (TFBSs) in a set of sequences by applying Match™ program (Kel et al., 2003) using the profile parameters and the position weight matrix (PWM) library specified in Deyneko et al. (2013). The PWMs are taken from TRANSFAC database (Wingender, 2008).

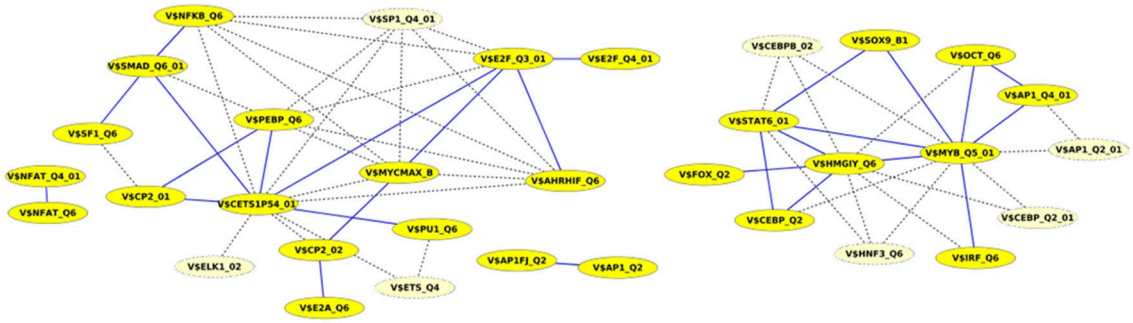

FIGURE 4 | Cooperation network according to PC-TraFF significant TFBS pairs for *Luminal A* gene set. The nodes represent TFBSs identified by the indicated PWMs. Edges represent their potential cooperation based on observed co-occurrences. After applying our extension approach: while blue edges correspond to the sequence-set specific cooperations ($\alpha = 0.2$), the common cooperations are shown by dashed lines. The nodes with light yellow color indicate TFBSs that are involved in common TF cooperations, but not in the specific pairs.

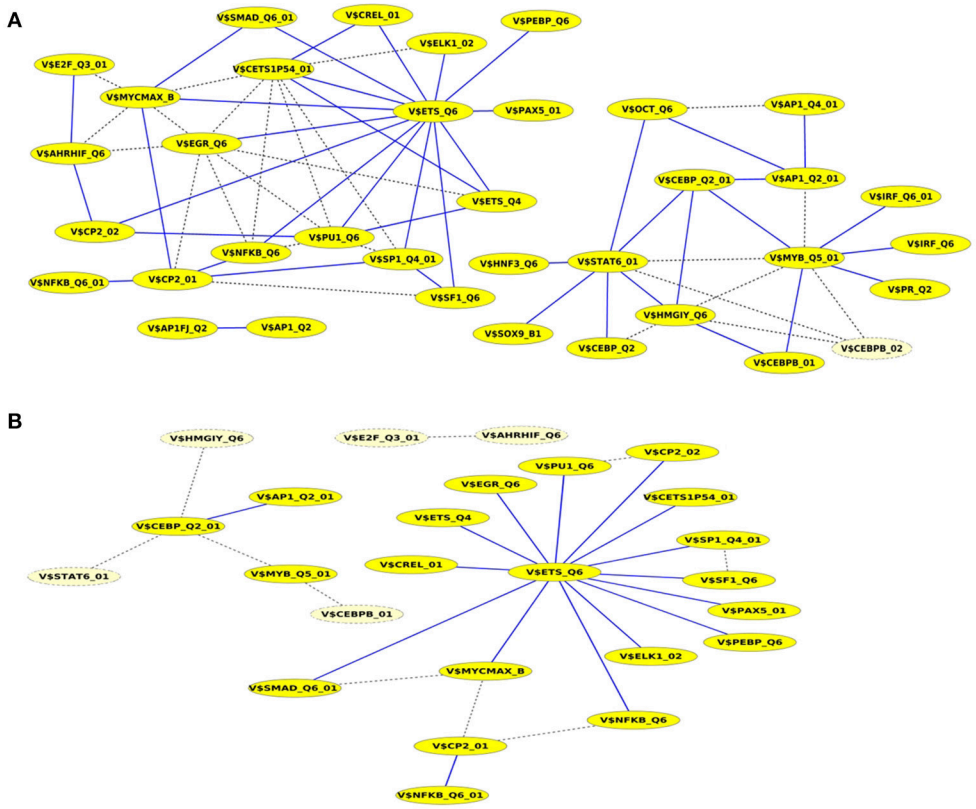

FIGURE 5 | Cooperation network according to PC-TraFF significant TFBS pairs for *Basal-like* gene set. The nodes represent TFBSs identified by the indicated PWMs. Edges represent their potential cooperation based on observed co-occurrences. After applying our extension approach: while blue edges correspond to the sequence-set specific cooperations for **(A)** $\alpha = 0.2$ and **(B)** $\alpha = 0.5$, the common (generally important) cooperations are shown by dashed lines. The nodes with light yellow color indicate TFBSs that are involved in common TF cooperations, but not in the specific pairs.

Based on the observed frequencies of TFBSs in the sequences under study a TFBS-sequence matrix $\mathbb{M}$ is constructed (see **Figure 6**). In $\mathbb{M}$, the row-names are presented by the IDs of the sequences and columns refer to the names of PWMs used in Match^{TM} algorithm for the prediction of putative TFBSs. An entry $x_{i,j}$ in $\mathbb{M}$ is the frequency of a putative TFBS t_j ($j = 1, .., n$, where n is the number of PWMs) identified by PWM j in sequence s_i ($i = 1, ..., m$, where m is the number of sequences under study). After that, columns of $\mathbb{M}$ are filtered in order to reduce the effect of highly over- or underrepresented TFBSs.

TABLE 4 | The number of genes and promoter sequences for the BRC-associated subtypes.

BRC subtypes	Number of genes	Number of promoter sequences
Luminal A	78	86
Luminal B	55	57
Normal-like	23	27
Basal-like	28	31
ErbB2 over-expressing	13	15

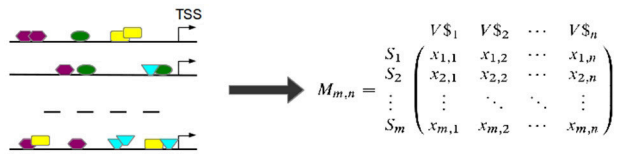

FIGURE 6 | TFBSs are identified for each sequence in the set (left). Afterwards, the TFBS frequencies are stored in a TFBS-sequence matrix $\mathbb{M}$ where an entry $x_{i,j}$ is the number of occurrences of TFBS t_j in sequence s_i. (TSS stands for "transcription start site").

3.1.2. Phase 2: Identification of Important TFBSs in Each Sequence

In order to identify important TFBSs for each sequence, we calculate the pointwise mutual information $\mathbb{PMI}(s_i; t_j)$ for each sequence s_i and TFBS t_j pair based on the frequencies of observed TFBSs in each sequence.

$$\mathbb{PMI}(s_i; t_j) = log_2 \frac{p(s_i, t_j)}{p(s_i)p(t_j)},$$

where $p(s_i, t_j)$ is the probability of a TFBS t_j to occur in sequence s_i. It is calculated as

$$p(s_i, t_j) = \frac{f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \quad (1)$$

where f_{ij} is the frequency of TFBS t_j in sequence s_i. $p(s_i)$ and $p(t_j)$ are the marginal probabilities and are calculated as

$$p(s_i) = \frac{\sum_{j=1}^{n} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \quad (2)$$

A TFBS t_j is regarded to be important for sequence s_i if the corresponding $\mathbb{PMI}(s_i, t_j) > 0$. In the following analysis steps, for each sequence only the important TFBSs are considered.

3.1.3. Phase 3: Filter to Avoid Overlaps

Overlapping TFBSs of the same type are filtered in a way that the TFBS survives which is closer to TSS in order to avoid the overestimation of these repetitive binding sites (see **Figure 7A**) and thereby to consider only these TFBSs that appear to be more functional (Whitfield et al., 2012).

3.1.4. Phase 4: Construction of TFBS Pairs

TFBS pairs are identified according to the distance of their centers (see **Figure 7B**). Two TFBSs can form a pair if their distance satisfies the pre-defined minimal and maximal thresholds.

3.1.5. Phase 5: Weighted Cumulative Pointwise Mutual Information

The weighted cumulative pointwise mutual information $\mathbb{PMI}_{pc}(t_a; t_b)$ of two putative TFBSs t_a and t_b is calculated as follows:

$$\mathbb{PMI}_{pc}(t_a; t_b) = \sum_{s \in S} w_s \cdot p(t_a, t_b) \cdot log_2 \frac{p(t_a, t_b)}{p(t_a) \cdot p(t_b)}, \quad (3)$$

where $p(t_a, t_b)$, $p(t_a)$ and $p(t_b)$ are the joint and marginal probabilities of TFBSs t_a and t_b, respectively. Further, w_s refers to the weight of a sequence s and is calculated based on the number of TFBS pairs N_s in s divided by the total number of TFBS pairs in the entire set of sequences S.

$$w_s = \frac{N_s}{\sum_{s_i \in S} N_{s_i}} \quad (4)$$

3.1.6. Phase 6: Background Noise Reduction of TFBSs Using Average Product Correction

To this end, using the average product correction (APC) theorem proposed by Dunn et al. (2008), the $\mathbb{PMI}_{pc}(t_a; t_b)$ scores have been adjusted:

$$\mathbb{PMI}_{pc}^{APC}(t_a; t_b) = \mathbb{PMI}_{pc}(t_a; t_b) - \frac{\mathbb{PMI}_{pc}(t_a; \overline{t_x}) \cdot \mathbb{PMI}_{pc}(t_b; \overline{t_x})}{\overline{\mathbb{PMI}_{pc}}} \quad (5)$$

where $\mathbb{PMI}_{pc}(t_a; \overline{t_x})$ is the mean $\mathbb{PMI}_{pc}$ of t_a to all other TFBSs in the sequences, and $\overline{\mathbb{PMI}_{pc}}$ is the mean $\mathbb{PMI}_{pc}$ value over all TFBS pairs.

The resulting $\mathbb{PMI}_{pc}^{APC}$ values are transformed into z-scores and only those pairs are considered to be significant that have a z-score ≥ 3.

Separation of Sequence Set Specific TF Cooperations From the Common Ones

According to their TFBS motifs, some TF cooperations are noticeable sensitive to the context of nucleotides - regarding the order and positions of nucleotides in sequences - in comparison to common TF cooperations, which are often found as significant for different sequence sets.

In order to separate such sequence-set specific significant TFBS pairs from the common (general important) significant pairs, we propose the following approach: The uShuffle algorithm (Jiang et al., 2008) is used to shuffle the nucleotides within each sequence by setting k-mers' size = 3. Thereby, not only the single nucleotide counts of each sequence are maintained but also the triplet counts and thus, the core of TFBSs. By repeating this

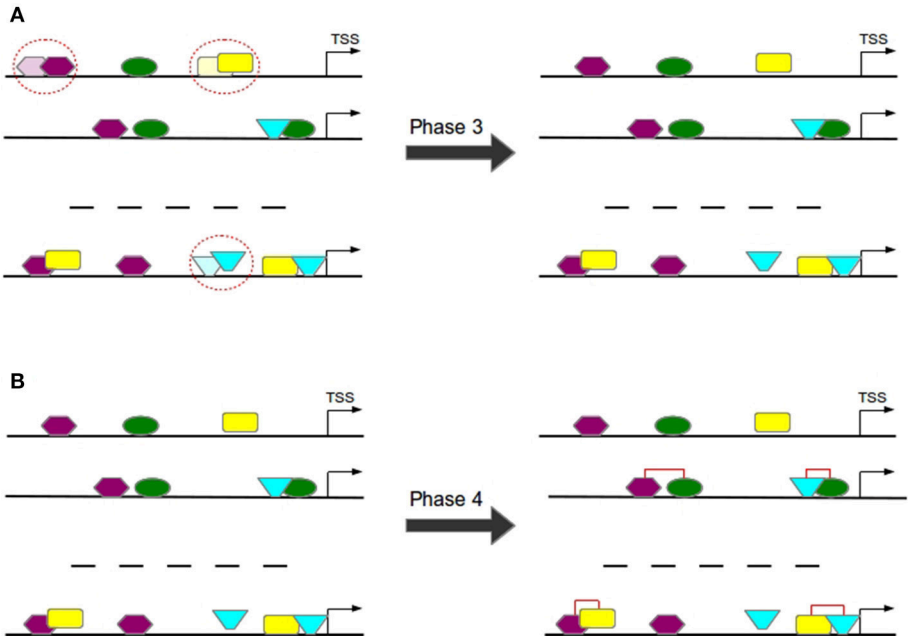

FIGURE 7 | In Phase 3 overlapping TFBSs of the same type are filtered by removing that TFBS having a larger distance to TSS **(A)**. In Phase 4 TFBS pairs are formed according to the distance of their centers **(B)**.

shuffling process several times, a sufficient number of randomly generated sequence sets (e.g., 1000) is created.

Second, employing the Match^{TM} algorithm for each set of shuffled sequences, the putative binding sites of TFs in these sequences are predicted. Third, applying PC-TraFF algorithm, new $\mathbb{PMII}_{pc}$-values for every TFBS pair in each randomly generated sequence set are calculated. Fourth, based on these $\mathbb{PMII}_{pc}$-values of each pair t_a and t_b, we define the average $\mathbb{PMII}$-value, $AVG\left(\mathbb{PMII}(t_a; t_b)\right)$ as

$$AVG\left(\mathbb{PMII}(t_a; t_b)\right) = \frac{1}{l}\sum_{i=1}^{l} \mathbb{PMII}_{pc}^{APC}(t_a; t_b)_i, \qquad (6)$$

where l is the number of randomly generated sequence sets.

After that, the $AVG\left(\mathbb{PMII}(t_a; t_b)\right)$-value of binding sites t_a and t_b is subtracted from their initial significant $\mathbb{PMII}_{pc}^{APC}(t_a; t_b)$ -value as

$$\mathbb{PMII}^{specific}(t_a; t_b) = \mathbb{PMII}_{pc}^{APC}(t_a; t_b) - \left[(1+\alpha) \times AVG\left(\mathbb{PMII}(t_a; t_b)\right)\right], \qquad (7)$$

where $\alpha \in [-1, +1]$ is a preassigned real number for monitoring the influcene of this process on the significant TFBS pairs. It can easily be seen that $\alpha = -1$ results in the original PC-TraFF analysis. By setting α =0 the average $AVG\left(\mathbb{PMII}(t_a; t_b)\right)$ is subtracted from the original $\mathbb{PMII}_{pc}^{APC}(t_a; t_b)$ value whereas an $\alpha \geq 0$ leads to a stronger effect of the subtraction and thus, a more strict selection process. However, for the proper application of this process the determination of an upper bound for α is crucial in order to avoid the overestimation of the efficacy of $AVG\left(\mathbb{PMII}(t_a; t_b)\right)$-values (background level) on the separation of sequence-set specific pairs from common ones. By systematically analyzing different values, we established that +1 is the most convenient upper bound for α.

A positive $\mathbb{PMII}^{specific}(t_a; t_b)$-value of binding sites t_a and t_b identified in the promoter sequences of a certain sequence set suggests that the binding of the related TF pair is strongly sequence context dependent. In contrast, a $\mathbb{PMII}^{specific}(t_a; t_b)$-value ≤ 0 indicates that the cooperations of corresponding TFs could have a general importance for the controlling of genetic programs.

4. CONCLUSIONS

Depending on their biological functions as well as cellular context, TFs specify the selection of cooperation partners in many ways for different cell types. However, the existing algorithms often focus on the identification of all predictable TF cooperations without distinguishing between sequence-set specific and common, i.e., ubiquitously occurring TF cooperations. To address this limitation, we propose in this study an approach that extends our previous method PC-TraFF in order to assign its predictions into two main categories: sequence-set specific and common (generally important) ones. For this aim, we estimated the background co-occurrence of any TF pair by preserving the nucleotide composition and the core of TFBS motifs in the sequences of interest. To maintain the core of TFBS motifs, we set the *k-mers*'size = 3 in the randomly shuffled new sets of sequences. It can be seen that,

while an increase in *k-mers*'size could lead to increment of background co-occurrence of TFBSs, a decrease in *k-mers*'size could in turn result in the reduction of background level of TF pairs. In order to assess the effectiveness of our extension approach, we analyzed promoter sequences of five different breast cancer-associated subtypes. The results show that the cooperating pairs identified by original PC-TraFF algorithm were considerably overlapping between the subtypes. Applying our extension approach, we could successfully separate sequence-set specific pairs from common ones and thereby reducing the number of overlapping pairs. Further, when we applied our extension approach of the original PC-TraFF algorithm to a simulation data set with varying α-values and, thus, different background levels, we could demonstrate that the cooperating TF pair was consistently identified as a sequence-set specific pair. The scaling parameter α is useful to extend or reduce the level of the subtracted background. Thereby, the influence of α itself is not linear but highly depending on the sequence set and thus on the respective background. Starting with an α-value of 0.2 we recommend to slightly increase α in order to assess the effect of α on the given data set and in doing so, to get the desired ratio between sensitivity and specificity. In summary, the proposed extension approach can successfully be applied for the distinction of sequence-set specific TF cooperations from common ones which are identified as generally important for different data sets.

AVAILABILITY OF DATA AND ALGORITHM

The extension of PC-TraFF is freely accessible at http://pctraffpro.bioinf.med.uni-goettingen.de/. All data sets and results of this paper are available from the corresponding author on request.

AUTHOR CONTRIBUTIONS

CM and MG developed the model and conducted computational analyses. EW interpreted the results and adjusted the model together with CM and MG. CM and MG conceived of and managed the project and wrote the final version of the manuscript. All authors read and approved the final manuscript.

FUNDING

CM was funded by ExiTox2 (Förderkennzeichen: 031L0120B) of the BMBF (German Ministry of Education and Research).

ACKNOWLEDGMENTS

We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

REFERENCES

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956. doi: 10.1016/j.cell.2005.08.020

Chatr-aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., et al. (2014). The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 43, D470–D478. doi: 10.1093/nar/gku1204

Chuang, C.-L., Hung, K., Chen, C.-M., and Shieh, G. S. (2009). Uncovering transcriptional interactions via an adaptive fuzzy logic approach. *BMC Bioinformatics* 10:400. doi: 10.1186/1471-2105-10-400

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). Upsetr: an r package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Deyneko, I., Kel, A., Kel-Margoulis, O., Deineko, E., Wingender, E., and Weiss, S. (2013). MatrixCatch - a novel tool for the recognition of composite regulatory elements in promoters. *BMC Bioinformatics* 14:241. doi: 10.1186/1471-2105-14-241

Dongen, S. (2000). *Graph Clustering by Flow Simulation.* PhD thesis, University of Utrecht, Netherlands.

Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24, 333–340. doi: 10.1093/bioinformatics/btm604

Girgis, H., and Ovcharenko, I. (2012). Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. *BMC Bioinformatics* 13:25. doi: 10.1186/1471-2105-13-25

Ha, N., Polychronidou, M., and Lohmann, I. (2012). COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. *PLoS ONE* 7:e52055. doi: 10.1371/journal.pone.0052055

Hu, Z., and Gallo, S. M. (2010). Identification of interacting transcription factors regulating tissue gene expression in human. *BMC Genomics* 11:49. doi: 10.1186/1471-2164-11-49

Hu, Z., Hu, B., and Collins, J. (2007). Prediction of synergistic transcription factors by function conservation. *Genome Biol.* 8:R257. doi: 10.1186/gb-2007-8-12-r257

Jankowski, A., Prabhakar, S., and Tiuryn, J. (2014). TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics* 15:208. doi: 10.1186/1471-2164-15-208

Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics* 9:192. doi: 10.1186/1471-2105-9-192

Joshi, H., Nord, S. H., Frigessi, A., Børresen-Dale, A.-L., and Kristensen, V. N. (2012). Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes. *BMC Genomics* 13:199. doi: 10.1186/1471-2164-13-199

Kel, A., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E. (2003). MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 31, 3576–3579. doi: 10.1093/nar/gkg585

Kel-Margoulis, O., Kel, A., Reuter, I., Deineko, I., and Wingender, E. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* 30, 332–334. doi: 10.1093/nar/30.1.332

Lai, F.-J., Jhu, M.-H., Chiu, C.-C., Huang, Y.-M., and Wu, W.-S. (2014). Identifying cooperative transcription factors in yeast using multiple data sources. *BMC Syst. Biol.* 8:S2. doi: 10.1186/1752-0509-8-S5-S2

Meckbach, C., Tacke, R., Hua, X., Waack, S., Wingender, E., and Gültas, M. (2015). PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinformatics* 16:400. doi: 10.1186/s12859-015-0827-2

Mysickova, A., and Vingron, M. (2012). Detection of interacting transcription factors in human tissues using predicted DNA binding affinity. *BMC Genomics* 13(Suppl 1):S2. doi: 10.1186/1471-2164-13-S1-S2

Nandi, S., Blais, A., and Ioshikhes, I. (2013). Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors. *Nucleic Acids Res.* 41, 8822–8841. doi: 10.1093/nar/gkt578

Navarro, C., Lopez, F. J., Cano, C., Garcia-Alcalde, F., and Blanco, A. (2014). CisMiner: Genome-wide *in-Silico* cis-regulatory module prediction by fuzzy itemset mining. *PLoS ONE* 9:e108065. doi: 10.1371/journal.pone.0108065

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150, 1274–1286. doi: 10.1016/j.cell.2012.04.040

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8418–8423. doi: 10.1073/pnas.0932692100

Spadafore, M., Najarian, K., and Boyle, A. P. (2017). A proximity-based graph clustering method for the identification and application of transcription factor clusters. *BMC Bioinformatics* 18:530. doi: 10.1186/s12859-017-1935-y

Sun, H., Guns, T., Fierro, A. C., Thorrez, L., Nijssen, S., and Marchal, K. (2012). Unveiling combinatorial regulation through the combination of ChIP information and *in silico* cis-regulatory module detection. *Nucleic Acids Res.* 40:e90. doi: 10.1093/nar/gks237

Teif, V. B., and Rippe, K. (2010). Statistical-mechanical lattice models for protein-DNA binding in chromatin. *J. Phys. Condens Matter* 22:414105. doi: 10.1088/0953-8984/22/41/414105

Van Loo, P., and Marynen, P. (2009). Computational methods for the detection of cis-regulatory modules. *Brief. Bioinform.* 10, 509–524. doi: 10.1093/bib/bbp025

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., et al. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* 13:R50. doi: 10.1186/gb-2012-13-9-r50

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* 9, 326–332. doi: 10.1093/bib/bbn016

Wu, W.-S., and Lai, F.-J. (2016). Detecting cooperativity between transcription factors based on functional coherence and similarity of their target gene sets. *PLoS ONE* 11:e0162931. doi: 10.1371/journal.pone.0162931

Zeidler, S., Meckbach, C., Tacke, R., Raad, F., Roa, A., Uchida, S., et al. (2016). Computational detection of stage-specific transcription factor clusters during heart development. *Front. Genet.* 7:33. doi: 10.3389/fgene.2016.00033

Conflict of Interest Statement: EW is head of geneXplain GmbH, the company that maintains and distributes the TRANSFAC database.

The other authors declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.