The quantification of DNA motifs within flanking sequences of gene expression quantitative trait loci in humans: Laboratory Report

Jacqueline Kiewa

June 7, 2018

Contents

1 Introduction

An important complex trait that differs amongst individuals is the level of gene expression, obtained by measuring the amount of mRNA present in specific tissues. The class of variants that influence this trait has been labelled expression quantitative trait loci (eQTL). These variants are particularly important to Genome Wide Association (GWA) studies, since they can help to prioritize likely causal variants amongst the SNPs identified by the GWA study, as well as provide some insight into the biological mechanisms used by organisms to influence complex traits and diseases [Albert and Kruglyak, 2015]. Albert and Kruglyak [2015] note that the majority of QTL identified by human GWA studies occur in non-coding regions that are not in linkage disequilibrium (LD) with coding exons. It is therefore hypothesised that these QTL influence their targeted gene through regulation of its expression [Nica and Dermitzakis, 2013]. A further hypothesis is that an eQTL and its surrounding region will be enriched in biologically active DNA sequence motifs which will contribute in some way to gene expression regulation.

Lloyd-Jones et al. [2017] conducted an analysis using data from the Consortium for the Architecture of Gene Expression (CAGE), which comprises individual-level whole-blood expression and genotype data on 2,765 individuals. This analysis identified a total of 14,995 expression quantitative trait loci (eQTL), of which 11,204 were in cis. The aim of this research is to identify DNA motifs in the genomic regions surrounding the set of cis eQTL identified by Lloyd-Jones et al. [2017], following the hypothesis that flanking sequences of an eQTL will be enriched in recurring motifs of DNA that have a biological function, most likely involved in transcription.

1.1 Motif function

DNA sequence motifs are short, recurring sequences in DNA that are thought to have some biological function [D'haeseleer, 2006]. They may form binding sites for proteins such as transcription factors or nucleases, or they may provide signals for important regulatory processes such as methylation, or ribosome binding or mRNA processing [D'haeseleer, 2006]. Gene regulation is an important aspect of the eukaryotic genotype [Beckerman, 2005], changes in which are responsible for much of the phenotypic divergence between and within species [Stewart et al., 2012]. However, an understanding of the biological mechanisms by which this diversity regulates gene expression has been challenging [Pai et al., 2015, Gaffney, 2013]. Transcription constitutes the first and one of the most intensely regulated steps of gene expression [Zabidi and Stark, 2016], and a subset of DNA motifs in the region of eQTL are sequence binding sites for transcription factors (TFs). Much research has therefore focused on the identification of transcription factors that coincide with eQTL and are likely to enhance or restrict gene expression.

Once chromatin has become accessible, transcription can initiate within the core promoter site, which recruits RNA polymerase II and assembles the pre-initiation complex [Zabidi and Stark, 2016]. However, transcription will proceed at a very low basal level without the contribution of enhancers. Enhancers are DNA elements up to several hundred bp in length, and contain many short TF binding sites. Either individual or cooperative combinations of TFs are recruited to these sites and function as activators or repressors that regulate transcription from the target core promoter. Despite the abundance of research that has focussed on the relationship between eQTL and transcription factors, as a subset of DNA sequence motifs, TFs form just one piece in the complex jigsaw of gene regulation, with other regulatory processes also playing a part. A de novo motif search by Schor et al. [2017], for example, found a number of core motifs (not TF binding sites) that controlled the shape of promoter regions, resulting in changed transcription levels. Other systems of gene regulation include methylation of CpG sites, alternative splicing, and post-transcriptional effects such as interaction with miRNAs and polyadenylation [Gaffney, 2013]. Gaffney [2013] also observed that eQTL are often enriched in exons, and Kirsten et al. [2015] found that eQTL also coincide with non-coding RNAs and pseudogenes.

The following overview of potentially important motifs in DNA follows that of Boeva [2016].

Tandem repeats are repeated DNA sequence ordered in a head-to-tail fashion and include microsatellites, minisatellites, and satellite sequences. Short tandem repeats (STRs) may serve as binding sites for specific transcription factors, whilst longer satellite repeats can influence the

3D structure shaping of the genome. Interspersed repeats are similar sequences to STRs and are located throughout the genome and include transposable elements such as SINEs and LINEs.

AT-rich sequence are often located in gene promoters and play a role in transcription initiation. Approximately 24% of human genes contain an AT-rich sequence within the core promoter, with 10% containing a canonical TATA-box motif (TATAWAWR; W = A/T, R = A/G). In general, AT-rich DNA is easier to unwind than GC-rich DNA, since AT base pairing contains fewer bonds than GC base pairs, but this process is amplified by TATA binding protein which is recruited by the TATA-box.

The remaining 76% of human promoters that are GC-rich contain multiple binding sites of the transcriptional activator SP1 [Yang et al., 2007]. As much as 56% of human genes, including most housekeeping genes, possess CpG islands, i.e. 300-3000 bp GC-rich sequences around gene transcription start sites (TSS's) with a high density of CpG dinucleotides. CpG islands have a high methylation level, which is associated with transcriptional repression.

Splice sites are involved post initial transcription, when the RNA undergoes the process of splicing, during which introns are removed and the remaining exons are joined together. Generally this process is catalyzed by spliceosomes that recognize a donor site, which is almost invariably 'GU' at the 5' end of the intron; a branch site, which is an 'A' followed by a pyrimidine-rich tract near the 3' end of the intron; and an acceptor site, which is almost always 'AG' at the 3' end of the intron. A DNA mutation in a splice site may have a wide range of functional consequences often leading to a defective or truncated protein.

Micro RNA molecules (miRNA's) regulate the amount of protein at the post-transcriptional level. Micro RNA's form part of the RNA-Induced Silencing Complex (RISC). The function of miRNA in this complex is to bind to the 3' untranslated region of of messenger RNA (mRNA). A successful binding will lead to repression of the translation of mRNA into protein. Mutations in an miRNA target site can therefore disrupt miRNA repressive regulation, resulting in protein over expression.

1.2 Motif detection

DNA motif detection can be done experimentally or with high-throughput data analysis using computer algorithms. Identification of DNA motifs can be achieved through in vivo or in vitro experiments, which include, for example, ChIP-seq (in vivo) that uses actual TF binding events in particular biological conditions, such as cell type or treatment time point [Inukai et al., 2017]. In vitro approaches (e.g. SELEX) use artificially created DNA and are well suited for large-scale characterization of intrinsic TF binding sequence preferences [Inukai et al., 2017]. Results from a ChIP-seq experiment typically have a resolution of approximately 100 bp, and the set of sequences identified as a result of the experiment (ChIP-seq data) becomes the focus of a search for a single motif of the particular transcription factor used within the binding experiment, i.e. one motif that is the best match over all sequences identified within the experimental process. This process has been termed 'OOPS', or one occurrence per sequence [Zhang et al., 2016]. This consensus motif is putatively labelled as a TFBS for the bound transcription factor identified within the experimental data.

Alternatively, computational based motif detection searches for short lengths of DNA that have been conserved across sequences, but is not limited to one motif per sequence or to transcription factor binding sites (TFBS's). Zero, one, or multiple different motifs per sequence can be identified. The conservation of these motifs is thought to indicate biological significance. This significance is not necessarily that of a TFBS, but might be due to some other factor, as described in section 1.1 above.

As for single motif finding tools, multiple motif finding tools were first developed for finding transcription factors [Dassi and Quattrone, 2016], and many of them require ChIP-seq sequences as input. Later tools were developed to find other motifs, but again often focused on a particular section of DNA, such as promoters or enhancers [Boeva, 2016].

Das and Dai [2007] describe motif discovery as one of the most challenging problems in molecular biology and computer science, and formulate it as follows: given a set of sequences, find an unknown pattern that occurs frequently. If a pattern of m letters long appears exactly in every sequence, a simple enumeration of all m-letter patterns that appear in the sequences gives the solution. To find this pattern, in a set of t sequences of length n, we need to consider all $(n - m + 1)t$ possible starting positions or candidates for motifs.

The problem increases in complexity with the introduction of d, which is the number of point changes that will be accepted in any motif instance. This change might take the form of a simple change in nucleotide (e.g. from an A to a T) or of an insertion or a deletion. Since an exact match may not exist, the motif detection algorithm must consider all possible combinations of "words" of the given length across all sequences. Thus the problem becomes exponential with the number of sequences. For each combination, the detection algorithm must perform an alignment and calculate the probability that this combination of "words" are instances of the same motif. At the end of this process, the combination of "words" with the best probability will become the basis for a Position Weight Matrix which will be the accepted consensual motif for this group of sequences.

Given this level of complexity, the motif finding problem is an example of an NP-complete problem [Tran and Huang, 2014] (NP : nondeterministic polynomial time). A solution to the search problem can theoretically be found and verified in polynomial time by a nondeterministic algorithm which has the power of guessing correctly at every step [Dasgupta et al., 2006]. However, although such an algorithm theoretically exists, no polynomial time algorithm has yet been discovered for an NP-complete algorithm [Cormen et al., 2009]. Algorithms for solving this problem can be categorised as per [Sun et al., 2015], who grouped algorithms into exact algorithms, which achieve efficiency through organisational pre-processing such as a suffix tree to reduce the search space; and approximate algorithms, which use heuristic methods such as expectation maximisation to reduce processing time.

1.3 Conceptual example of a motif detection algorithm, and the Position Weight Matrix

This section will be used to describe one motif finding algorithm: the Gibbs algorithm [Lawrence et al., 1993]. A conceptual example will be provided, following the teaching videos provided by [Algoshareify, 2012].

The example begins with a set of sequences, $n = 10$ and $t = 4$, in which we are to find one example of a motif M of length $l = 7$ in each sequence:

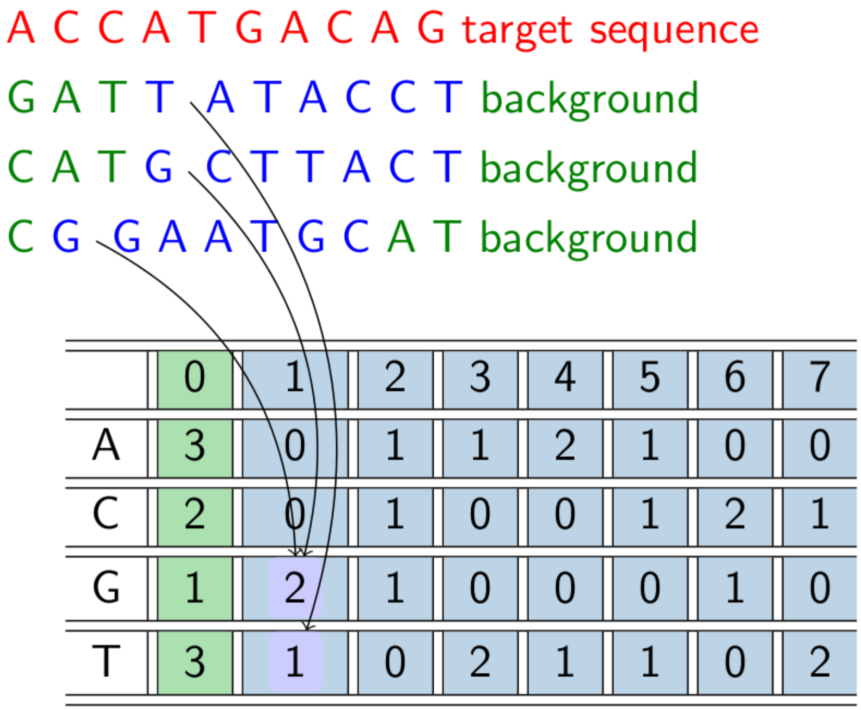
A C C A T G A C A G
G A T T A T A C C T
C A T G C T T A C T
C G G A A T G C A T

The algorithm works through repeated iterations of a single process. Each iteration ends with a choice of starting position for the motif in one of the sequences.

An iteration begins by setting one sequence aside as the target sequence. This sequence is chosen randomly, but in this example (see Figure 1, the first sequence (marked in red) will be used as the random sequence . All other sequences form the background. The background sequences are used to create a frequency table, which is then used to calculate the probability of the motif occurring at each possible position in the target sequence.

To construct the frequency table, first a motif position is chosen, at random, for each background sequence. In Figure 1 the background motifs are marked blue. Next the blue elements in the table are filled in according to the counts of the nucleotides for each position in the background motifs. The column headers in the table indicate the positions in the motifs. The green column (with the zero header) is filled in according to the counts of nucleotides in the non-motif background, irrespective of position. This creates the following scenario:

Pseudocounts are added to the counts in the table to eliminate zero's, and counts are turned into frequencies by dividing by the number of sequences (for the blue columns) or the total number of


A C C A T G A C A G target sequence
G A T T A T A C C T background
C A T G C T T A C T background
C G G A A T G C A T background

	0	1	2	3	4	5	6	7
A	3	0	1	1	2	1	0	0
C	2	0	1	0	0	1	2	1
G	1	2	1	0	0	0	1	0
T	3	1	0	2	1	1	0	2

Figure 1: Gibbs algorithm: creating the frequency table

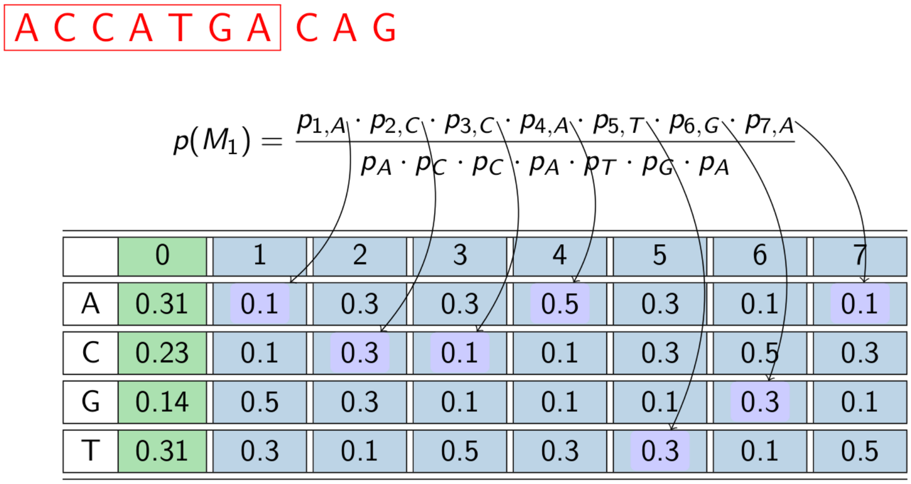
non-motif nucleotides (for the green column).

The second part of the iteration returns to the target sequence. For each possible motif starting position in this sequence, the probability of the resultant motif is worked out using the frequency table and the following equation:

$$p(M_1) = \frac{p_{1,A} \cdot p_{2,C} \cdot p_{3,C} \cdot p_{4,A} \cdot p_{5,T} \cdot p_{6,G} \cdot p_{7,A}}{p_A \cdot p_C \cdot p_C \cdot p_A \cdot p_T \cdot p_G \cdot p_A} \tag{1}$$

This procedure is illustrated in Figure 2. As illustrated in this figure, each element in the numerator is found by referincing the corresponding motif position (column) and nucleotide (row) in the table. Each element in the denominator is found by referencing the corresponding nucleotide in the zero column, irrespective of position. The lower table displays the values calculated for each of the four possible motif starting positions.

The lower table of motif probabilities is normalised, and then used to weight a random choice of motif start position for the target sequence. This choice of motif for the target sequence is



ACCATGA CAG

$$p(M_1) = \frac{p_{1,A} \cdot p_{2,C} \cdot p_{3,C} \cdot p_{4,A} \cdot p_{5,T} \cdot p_{6,G} \cdot p_{7,A}}{p_A \cdot p_C \cdot p_C \cdot p_A \cdot p_T \cdot p_G \cdot p_A}$$

	0	1	2	3	4	5	6	7
A	0.31	0.1	0.3	0.3	0.5	0.3	0.1	0.1
C	0.23	0.1	0.3	0.1	0.1	0.3	0.5	0.3
G	0.14	0.5	0.3	0.1	0.1	0.1	0.3	0.1
T	0.31	0.3	0.1	0.5	0.3	0.3	0.1	0.5

$p(M_1)$	$p(M_2)$	$p(M_3)$	$p(M_4)$
0.20	0.16	0.07	0.01

Figure 2: Gibbs algorithm: calculating the probability of the first motif start position

the culmination of this iteration. After hundreds of such iterations, depending on the number of sequences, motif positions will no longer be totally random, but most will have been weighted according to the procedure described above. After a certain number of iterations, a log likelihood of the probability of the current set of motifs is taken. If this log likelihood is better than the previous, it is retained; otherwise it is discarded and the algorithm returns to the motif positions for the previous log likelihood.

After many thousands of iterations the log likelihood is never better than the previous, and the algorithm will always return to the previous log likelihood. At this point convergence has been reached and a position weight matrix can be calculated for the set of motifs found.

The position weight matrix (PWM) itself is a useful tool that ensures that no information is lost with respect to the consensual motif sequence, since it records the frequency of each nucleotide in each position over all sequences (the position frequency matrix) before working out the probabilities for the final PWM. Another advantage is its convenient visualisation as a sequence logo, and a number of data bases (e.g. JASPAR [Mathelier et al., 2016]; Hocomoco [Kulakovskiy et al., 2018] and UniPROBE [Hume et al., 2015]) now store many hundreds of PWMs that match every known transcription factor. In fact the PWM *is* the motif, rather than the consensual sequence. The main disadvantage of the PWM is its assumption of independence of each position within the binding site, which may not always be true [Jayaram et al., 2016]. However, as Jayaram et al. [2016] points out, in practice, PWM models tend to perform as well as more complex models that attempt to include nucleotide interdependencies in their calculations, but which are often more prone to learning noise.

2 The Pilot Study

2.1 Introduction to the Pilot Study

Computationally efficient algorithms were required for this analysis and thus a possible set of highly-used algorithms were evaluated via a pilot study using a set of 201 base pair regions with the full set of 14,995 cis and trans eQTLs, which reduced to 13,930 eQTLs once duplicates were removed. The results from the pilot study were used to guide a larger investigation into the the frequencies of detected motifs across the cis sequences, and the identification of transcription factors with binding sites that match the final list of motifs and characterise the biological function of detected motifs. These analyses will give insight into the regulation of gene expression in regions of the genome that contribute to variation of gene expression in whole blood.

The general hypothesis is that if DNA motifs are detected in these sequences then we would expect them to have a common regulatory process across a subset of genes that harbour eQTL in the nearby region. This becomes a search for a conserved set of mechanisms across blood gene expression that acts through DNA motifs. This project sought to establish first whether it is indeed the case that DNA motifs can be detected in a robust manner.

2.2 Organisation of the data

The purpose of the pilot study was to investigate the possible methods for the identification of motifs associated with eQTLs identified by Lloyd-Jones et al. [2017]. Short target sequences of 201 base pairs, centred on the full set of each independent eQTL (both cis and trans) from the CAGE analysis, formed a set of 13,930 sequences for the pilot investigation. These sequences were constructed using the Genome Reference Consortium Human Build 37 (GRCh37)[Lander et al., 2001], also known as hg19. Using the UCSC Table Browser [Karolchik et al., 2004] as a reference, these sequences were then divided into four subsets according to their location in the genome: exons; introns; promoters; and intergenic. This division into smaller subsets reduced running time within each class. Table 1 provides the number of target sequences that overlapped with these classes. Where a sequence overlapped with more than one class, it was included in both groups; hence the figures add up to more than the 13,930 sequences.

Discriminative algorithms search for motifs that are overrepresented in target sequences compared to background sequences. Since such motifs may account for the increased expression of genes associated with the target eQTLs, a discriminative approach seems ideally suited to this project.It became necessary, therefore, to identify a background set of sequences. For this set, sequences were centred on a random group of 17,226 variants from the CAGE study that have no effect on gene expression. Sequences associated with these variants (null sequences) were also divided into exon, intron, promoter and intergenic subclasses. The number of null sequences in each class is also provided in Table 1.

Table 1: Comparison of target and background sequences across classes

Class	Exon	Intron	Promoter	Intergenic	Totals
Target Sequences	2131	7834	1231	5132	16328
Background Sequences	387	6886	256	10202	17731
Totals	2518	14720	1487	15334	34059

A comparison of the target and background sequence frequencies (see Table 1) found significant differences in the numbers for each class across target and background sequences ($\chi^2 = 3532.8$, $p < 0.00001$). Comparatively few of the null variants fell into exon or promoter sequences, but almost twice as many null variants compared to eQTLs fell into intergenic sequences. There was little difference between nulls and eQTLs with respect to the likelihood that they might be within an intron sequence.

2.3 Potential motif algorithm set

An initial literature review provided some insight into the types of algorithms that can be used to identify motifs. The OMICTools directory [Henry et al., 2014] was used to provide a list of motif-finding algorithms. The following criteria was used as a guide to the choice of motifs:

- Used a discriminative algorithm (comparing frequency of motif in target sequences compared to background sequences);
- Number of citations;
- Represented different approaches to the Planted Motif Problem;
- Ease of use (some programs were very difficult to install, or were no longer maintained and wouldn't run at all); and
- Appropriate for eQTL sequences: Some algorithms are written for specific sequence sets, such as promoters, and it is unclear whether they are generally applicable to other types of sequences. Where possible algorithms that are specifically described as broadly applicable were used.

Based on these criteria, six algorithms: DREME, HOMER, motifRG, STEME, BaMM!motif and DECOD were eventually chosen for comparison. A brief overview of each motif follows.

DREME (Discriminative Regular Expression Motif Elicitation) [Bailey, 2011] uses an enumerative algorithm, which is exact in the sense that it examines each possible 'word' in the sequence, which is very inefficient, with time increasing exponentially with the size of the motif and the number of sequences. DREME achieves efficiency by limiting the size of the motifs found, as well as a few shortcuts. Exact enumeration is used for seed motifs restricted to 3-8 base pairs in length, and significance established through comparison with the background sequence, using Fisher's Exact Test. The motif with the highest significance is used to create a set of non-overlapping occurrences in the set of sequences that are aligned to create a position-specific probability matrix. Following this step, the motif is 'erased' and the search repeated.

HOMER [Heinz et al., 2010a] begins by weighting background sequences to resemble the same GC-content distribution of the target sequences, since enrichment scores used in the motif detection algorithms might fail if the distribution of the length of GC content of the target sequences significantly differs from the background set. Input sequences are then "parsed" to gather all oligos of the desired motif length and read into an oligo table that records how many times each oligo occurs in the target and background sequences. These frequencies are used to calculate the relative enrichment of each oligo, using cumulative binomial distribution. The enrichment of longer oligos are calculated using the enrichment values of smaller oligos that make up the longer oligos, similar to the DREME method of combining seed significances. The most enriched oligos are then subjected to a local optimization algorithm which uses a probability matrix to score each oligo according to its similarity to the matrix. By decreasing the detection threshold, more oligos can be included, until an optimal enrichment is found. New probability matrices can then be created at different detection levels, and the matrix that results in the highest enrichment will be used to produce the final motif. This motif is then masked from the sequences and the whole process repeated to find the next motif [Heinz et al., 2010b].

MotifRG [Yao et al., 2014] begins by counting all k-mers, for a given motif length k, in all sequences. The motifRG algorithm uses logistic regression to fit a model with the best absolute z-value (which needs to be above an enrichment threshold). The chosen k-mer is used as a seed motif, which is then refined by stepped extension of a given number of nucleotides on each end of the seed at a time. Refinement ends when no further improvement in the z-score results after two successive steps. Small perturbations in the seed motif are then tried and accepted if the z-score is thereby improved, followed by the extension process. Since a small difference in the z-value may not be meaningful, a simple bootstrap test is done if the difference falls below a threshold. A default of 5 random samples of the sequences are used to calculate z-values for the new and original motifs.

These z-values are then subjected to a t-test to establish whether their difference is significant. Once the motif is established, it is masked from all sequences and the process begun again.

STEME (Suffix Tree EM for Motif Elicitation) [Reid and Wernisch, 2011] is modelled on the well established MEME algorithm [Bailey et al., 2009], which uses Expectation Maximisation (EM), and this short review begins with a brief description of this process. EM is an extremely robust algorithm but each iteration begins with a time-inefficient initialisation step that generates a matrix of each possible motif and fills in the probability of each base by counting the number of times each base occurs in each position in the set of all motifs and dividing by the number of times it could have occurred (= the number of motifs). The expectation step uses the matrix to calculate the probability of each motif, and then this probability is used to weight the count of bases in the matrix (the maximisation step). The expectation and maximisation (E and M) steps are repeated until convergence is reached and the matrix no longer changes. The EM algorithm is subject to local optimas, and the solution adopted by MEME, to rerun the algorithm with different initial starting positions, results in a running time that is quadratic to the number of sequences and is not practicable with large data sets. STEME modifies the EM algorithm by ignoring all W-mers (subsequences of length W) which have a probability less than some threshold. In addition, STEME uses a suffix tree to iterate over all W-mers. Once the initial matrix is built, it is the content rather than the position of the W-mer that is relevant. A suffix tree can then achieve efficiency over the EM algorithm in two ways: firstly, if any two W-mers are identical they do not have to be repeated; and secondly, partial evaluations are calculated for each step of the suffix tree and shared across W-mers below that node in the tree. In contrast, MEME evaluates every base in each W-mer individually. STEME was still the slowest of these algorithms to run, but its authors claim that it is one order faster than the MEME algorithm with the same data [Reid and Wernisch, 2011].

BaMM!motif (Bayesian Markov model or BaMM) [Siebert and Söding, 2016] addresses the issue that position weight matrices (PWMs) assume independence of each base in the motifs [Jayaram et al., 2016], which may not always be true. By using a Markov Model, the BaMM algorithm incorporates the probability of prior nucleotides into the final probability of any nucleotide in a particular position. BaMm begins with a 5th order Markov Model and uses expectation maximisation to estimate the transition and emission probabilities of these potential motifs. Probable motifs are lengthened using a Bayesian approach of multiplying the probability of the preceding markov model with the probability of the proposed additional nucleotide. Because the tables of probabilities produced by this algorithm have incorporated a higher order Markov modelling, the authors refer to them as BaMMs rather than as PWMs, although their layout is the same [Siebert and Söding, 2016].

DECOD (DECOnvolved Discriminative motif finder) [Huggins et al., 2011] begins with an enumeration step: given a user-specified motif length k, all k-mers are extracted from the target and background sequences and arranged into a table of frequencies of occurrence. However, after this step, no further analysis is conducted of the actual sequences - all subsequent analysis is of the k-mer counts table and PWMs are formed on the basis of frequency of occurrence in target compared to background sequences. This heuristic approach greatly speeds up the algorithm but bears the cost of losing the context of the k-mer. This loss of context may mean that overlapping k-mers are counted for the same motif more than once, leading to a convolved PWM that is inaccurate. To overcome this problem DECOD adopts a deconvolution process that removes k-mers that form a subset of the true motif [Huggins et al., 2011].

Using six motif-finding algorithms with four classes of sequences meant that many motifs were produced. Manual comparison of motifs across algorithms searching for similarities amongst results became very difficult.

The STAMP algorithm [Mahony and Benos, 2007] was designed for this purpose. It has two main functions: 1. To perform motif alignment and group motifs according to similarity; and 2. To search the user's database of choice for known transcription factors with similar motifs. STAMP provides a large number of user options:

1. Alignment of motifs can be according to Needleman-Wunsch (global) or Smith-Waterman (local) alignment methods;
2. These methods are based on column comparison scores calculated by one of five distance met-

rics: Pearson's correlation coefficient; Kullback-Leibler information content; sum of squared distances; average log-likelihood ratio; or average log-likelihood with a lower limit of -2.

3. Alignment can be gapped or ungapped with a variety of penalties imposed for gap-opening and gap-extension;
4. Users may choose to trim edges of motifs;
5. Motif multiple alignment strategies can be according to 'progressive profile alignment' or 'iterative refinement';
6. Two tree building algorithms are offered: an agglomerative method (UPGMA) and a divisive method based on a self-organising tree algorithm (SOTA);
7. Five databases are offered for motif matching.

2.4 Results of the pilot study

A disparate number of motifs were detected for each algorithm, which was a source of confusion. Table 2 outlines some of the source of this confusion.

The algorithms also varied in the statistical details provided to support their findings, as described in Table 3.

Running the algorithms on each subset of the data resulted in a set of motifs for each sequence subset. Within each subset, motifs were formatted as input for the STAMP algorithm, which performed motif alignment across all motif sets and grouped the motifs according to similarity. Default parameters were used: Pearson Correlation Coefficient for distance metrics; Smith-Waterman local alignment ungapped alignment: Iterative refinement multiple alignment; UPGMA tree building; and the JASPAR database [Sandelin et al., 2004] to identify those motifs that corresponded to a known transcription factor.

No strong differences emerged in the results from the four groups (exon, intron, promoter, and intergenic). Since the purpose of the pilot study was to evaluate the algorithms, rather than to

Table 2: Summary of motif detection algorithm properties.

Algorithm	**No. of motifs option**	**Entered**	**Result**
DREME	Required	15	Stops finding motifs once a minimum threshold is reached
HOMER	Optional	Used a default of 25	Finds 25 motifs for every length motif requested. Since 3 lengths were default, 75 motifs were returned
STEME	Required	20	Always finds the requested number of motifs, but often reports e-values above any useful threshold
DECOD	Required	20	Often does not find requested number of motifs, but no information supplied in the documentation to explain any thresholds
BaMM	No options	No default	No information provided re thresholds. Varying numbers of motifs resulted, but always with useful e-values
MotifRG	Optional	20	Has threshold minimum fraction of target/background sequences (=0.01) and minimum fold change of motif in target Vs background (=1.3)

Table 3: Comparison of provision of statistical details across algorithms

Algorithm	Enrichment Calculation	p-value/E Value provided	Other details
DREME	Fishers Exact Test	Provides p-values and associated e-values	
HOMER	Cumulative binomial distribution	Provides uncorrected p-values	Provides the number of target and background sequences that contain the motif
STEME	Not provided	Provides e-values	Provides target and background numbers within mass of log data (difficult to locate)
DECOD	z scores	Provides e-values	E-values only provided for the GUI-version, not for the terminal command
BaMM	Log-odds score	Provides e-values	Varying numbers of motifs produced with significant e-values. Documentation states that threshold is $p < 1.0$, but this is not reflected in results, which have very small E Values.
motifRF	Wald test	Not provided	

do an indepth study of motifs found and associated transcription factors, a brief summary will be provided of results from one group only: the exon set of sequences.

The six different algorithms returned a total of 770 motifs in the exon sequences, illustrated in the circular cladogram developed using Evolview tree viewing software [He et al., 2016]. Although the number of motifs means that the labelling on the cladogram is not easily distinguishable, the colouring provides an illustration of the number and spread of motifs found by each algorithm. In the exon sequences, HOMER found the most motifs, with good variability. STEME, DECOD and DREME also achieved a good spread, but reported relatively few motifs. However, in the case of DREME, this paucity of motifs seemed to be due to the use of the website rather than the downloaded version of the motif finding tool, since a later check using the downloaded version returned a large number of motifs. BaMM also ahieved a good spread of a large number of motifs. RGMotif did not perform well, with only a few motifs all clumped together, bearing no relationship to motifs found by other algorithms.

Using the JASPAR data base, the Stamp algorithm was able to match all of the motifs as potential

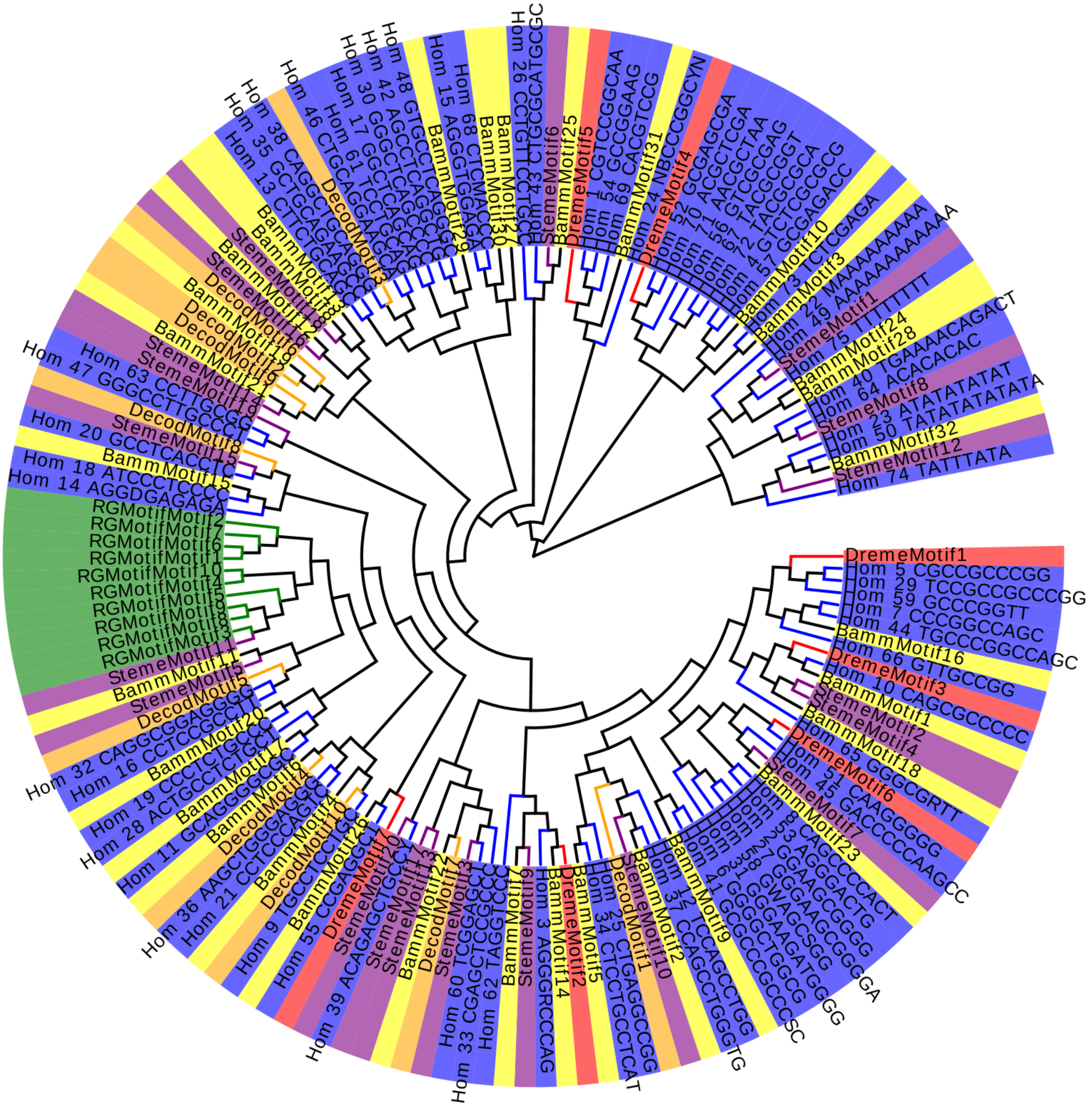
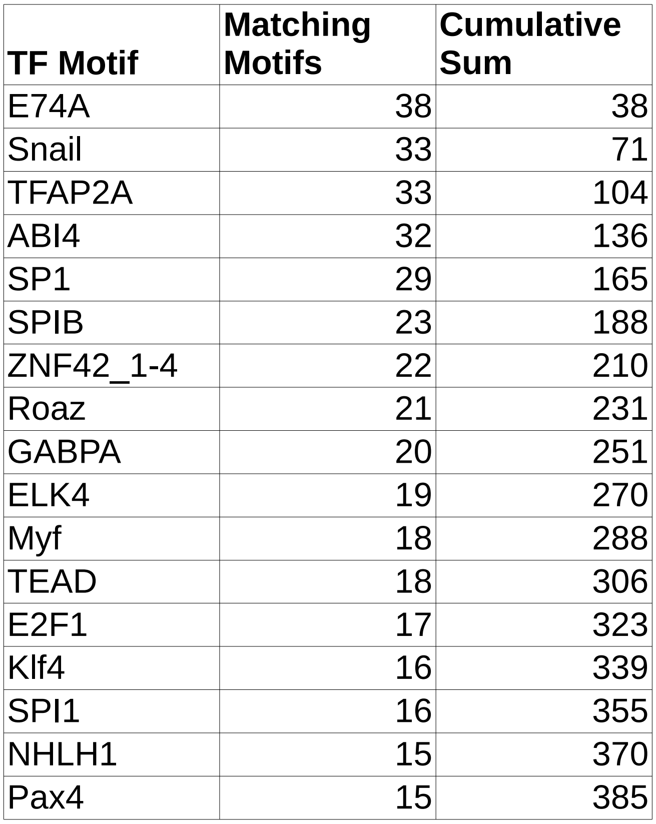


Figure 3: **Circular cladogram of 770 exon motifs from Evolview tree viewing software [?]** . The cladogram depicts motifs detected from DREME (red), HOMER (blue), BaMM (yellow), MotifRG (green), STEME (purple) and DECOD (orange).

binding sites for transcription factors. It is often the case that a single transcription factor has many binding sites. For this reason the 770 motifs were identified as binding sites for 99 transcription factors, with more than half the motifs identified as potential binding sites for a subset of just 17 transcription factors, as illustrated in Table 4.



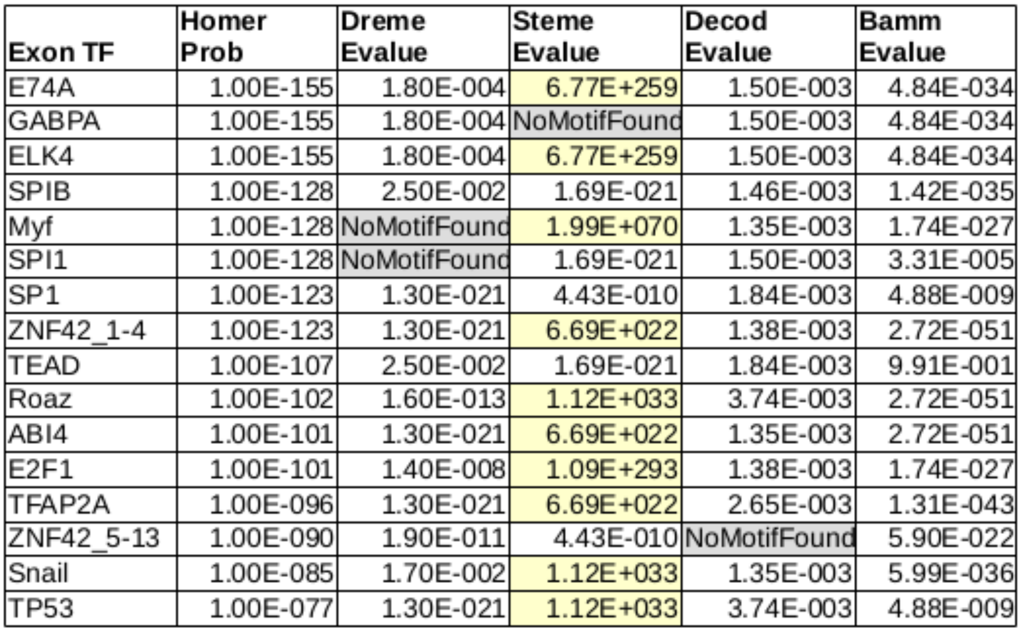
TF Motif	Matching Motifs	Cumulative Sum
E74A	38	38
Snail	33	71
TFAP2A	33	104
ABI4	32	136
SP1	29	165
SPIB	23	188
ZNF42_1-4	22	210
Roaz	21	231
GABPA	20	251
ELK4	19	270
Myf	18	288
TEAD	18	306
E2F1	17	323
Klf4	16	339
SPI1	16	355
NHLH1	15	370
Pax4	15	385

Table 4: **Summary of transcription factor (TF) matches to motifs detected in exonic sequences in the pilot study.**

It should be noted that the p-values for these matches were not subjected to any Bonferroni correction. In the larger study, this correction was implemented, resulting in a much smaller number of matched motifs and a large number of motifs unmatched to a transcription factor.

Enrichment values for motifs found in eQTL sequences compared to the null sequences are provided in Table 5. As illustrated, many values found by STEME (shaded yellow) are at odds with those found by other algorithms. As mentioned previously, RGmotif did not provide enrichment values and so is missing from this table.

No further analysis is provided of these results, as the purpose of the pilot study was not so much



Exon TF	Homer Prob	Dreme Evalue	Steme Evalue	Decod Evalue	Bamm Evalue
E74A	1.00E-155	1.80E-004	6.77E+259	1.50E-003	4.84E-034
GABPA	1.00E-155	1.80E-004	NoMotifFound	1.50E-003	4.84E-034
ELK4	1.00E-155	1.80E-004	6.77E+259	1.50E-003	4.84E-034
SPIB	1.00E-128	2.50E-002	1.69E-021	1.46E-003	1.42E-035
Myf	1.00E-128	NoMotifFound	1.99E+070	1.35E-003	1.74E-027
SPI1	1.00E-128	NoMotifFound	1.69E-021	1.50E-003	3.31E-005
SP1	1.00E-123	1.30E-021	4.43E-010	1.84E-003	4.88E-009
ZNF42_1-4	1.00E-123	1.30E-021	6.69E+022	1.38E-003	2.72E-051
TEAD	1.00E-107	2.50E-002	1.69E-021	1.84E-003	9.91E-001
Roaz	1.00E-102	1.60E-013	1.12E+033	3.74E-003	2.72E-051
ABI4	1.00E-101	1.30E-021	6.69E+022	1.35E-003	2.72E-051
E2F1	1.00E-101	1.40E-008	1.09E+293	1.38E-003	1.74E-027
TFAP2A	1.00E-096	1.30E-021	6.69E+022	2.65E-003	1.31E-043
ZNF42_5-13	1.00E-090	1.90E-011	4.43E-010	NoMotifFound	5.90E-022
Snail	1.00E-085	1.70E-002	1.12E+033	1.35E-003	5.99E-036
TP53	1.00E-077	1.30E-021	1.12E+033	3.74E-003	4.88E-009

Table 5: **TFs found by all algorithms with associated E-Values, ordered by DREME E-Values in pilot exon sequences.**

to analyse the motifs found, but rather, firstly, to demonstrate that the algorithms are capable of detecting motifs; and secondly, to provide a basis to evaluate the algorithms with respect to their suitability for the larger scale analysis.

The pilot study suggests that the two algorithms HOMER and BaMM!motif appear to have the most extensive coverage of motif binding sites. The recently published BaMM!motif algorithm is particularly impressive, given the fine tuning of its motifs, unlike HOMER, whose motifs are frequently clumped into broad clusters. Both these motif finding algorithms have high computational efficiency making them well suited to the large sequence analysis. DREME is also efficient, but fails to find some motifs that are found by every other algorithm. However, with further analysis, this proved to be due to the use of the Website version of the software during the pilot study. The use of the downloaded version led to motif discovery comparable to the Homer algorithm. DECOD was computationally time consuming and did not report many motifs. STEME took days to run, and reported E-Values at odds with those reported by the other algorithms. Finally, motifRG seemed to miss many motifs and did not report the significance of those it found, which were, additionally, clumped together, failing to align with any of the motifs found by other algorithms.

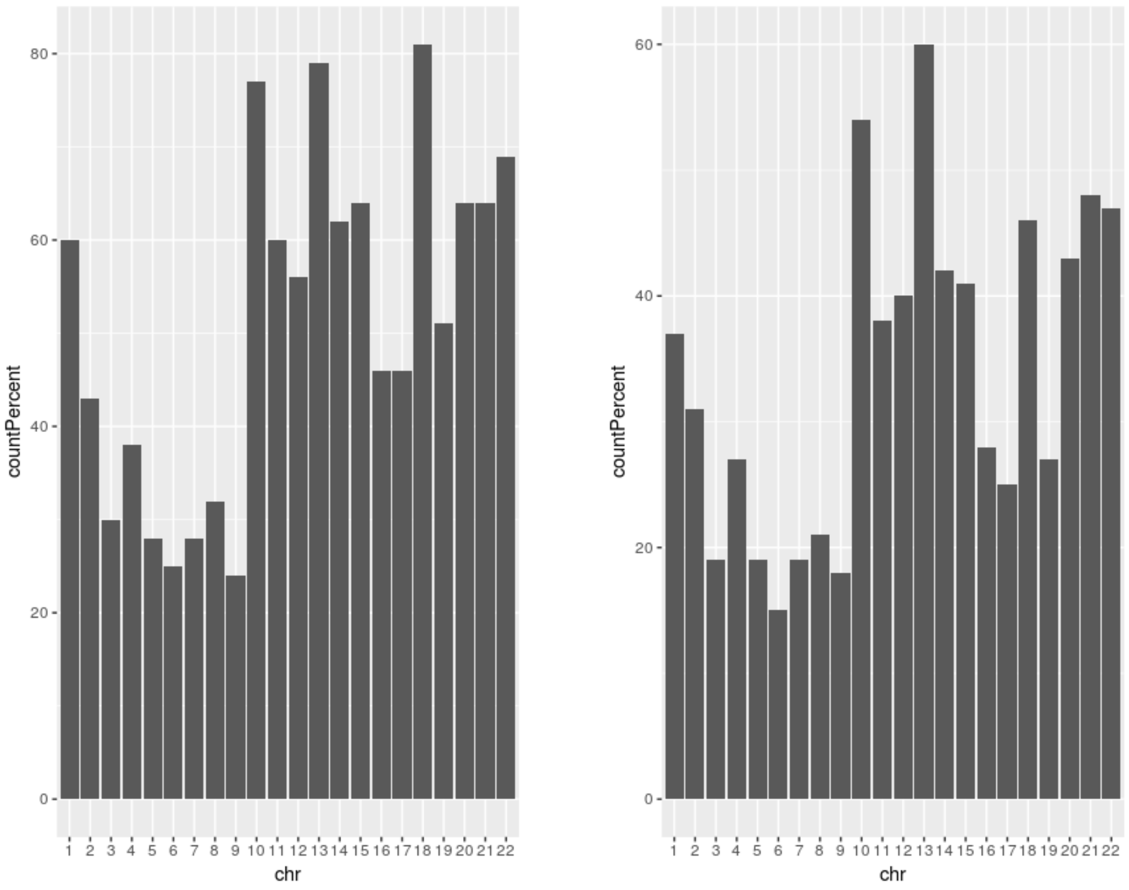
On the basis of this summary evaluation, the algorithms chosen for the main study were HOMER, BaMM! and DREME.

2.5 Moving from 200 kb sequences to larger sequences

The pilot data looked at small regions surrounding the eQTLs. The full analysis examined much longer sequences, which increased the computational challenge relative to the pilot study. An initial investigation was conducted to determine the most appropriate length of sequence with respect to the possibility of sequence overlap between eQTL and null sequences. The purpose of this investigation was to determine the degree of overlap between eQTL and null sequences. Substantial overlap reduces the ability of algorithms to discriminate between target and null sequences when searching for relative motif enrichment. Sequences of four different lengths: 200001 bases; 40001 bases; 10001 bases; and 2001 bases) (termed 200KB, 40KB, 10KB and 2KB sequences) were created, all centred on the eQTLs used in the pilot study. The same length sequences were also created for the variants with no effect (termed null sequences). The likelihood of overlap increases with the length of the sequence. A python program was written to isolate all overlaps. If an eQTL sequence had more than one overlap (e.g. with different null sequences overlapping the start and finish of the eQTL sequence), the largest overlap was used to calculate the size of the overlap. An R program was used to convert these overlaps into a histogram of the number of overlapping sequences as a percentage of all sequences within each chromosome.

Figure 4(a) provides a histogram (by chromosome) of the percentage of eQTL sequences that overlapped with null sequences for the 200 kb sequences. As illustrated, a majority of chromosomes had at least 40 percent of their sequences with some degree of overlap, with close to 80 percent of sequences in three chromosomes overlapping with null sequences. Figure 4(b) illustrates the percentage of sequences within each chromosome that overlapped by at least 50 percent of sequence length. Twenty percent of sequences in almost every chromosome fell into this category, and for chromosome 13, sixty percent of sequences had at least 50 percent overlap.

For 40KB sequences, the percentage of eQTL sequences that overlapped with null sequences by



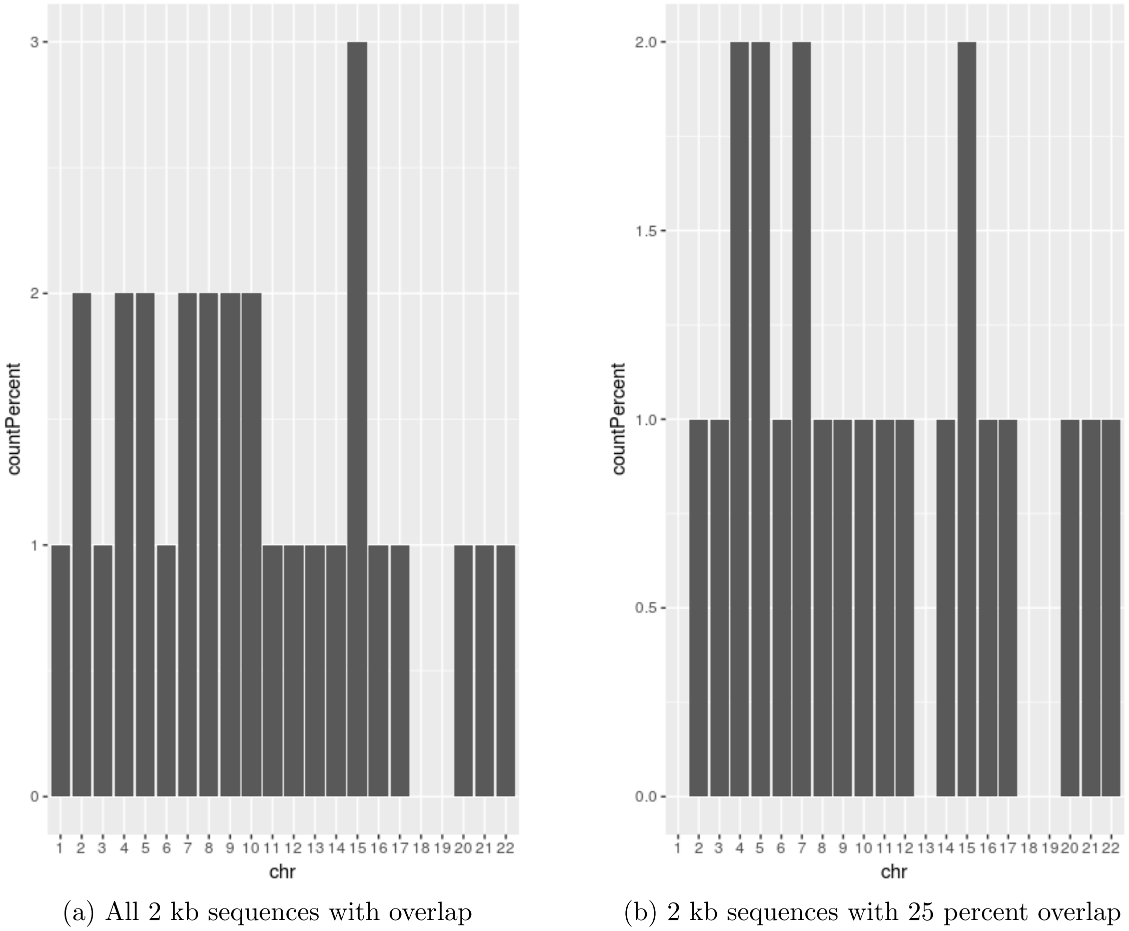
(a) All 200 kb sequences with overlap

(b) 200 kb sequences with 50 percent overlap

Figure 4: Summary of overlap percentage of eQTL and null sequences for 200 kb sequences with overlaps

chromosome, with no account taken on the size of the overlap was approximately 20 percent. Only one chromosome had more than 30 percent of sequences overlapping with null sequences, although all chromosomes had at least ten percent of their sequences with some overlap. Only one chromosome had more than 15 percent of its sequences with 50 percent overlap or more. For 10 kb sequences, the overlap is substantially reduced. A maximum of eight percent of sequences in any chromosome experience any overlap with on average four percent of the sequences having overlap greater than 25 percent on average across the genome.

Finally, Figure 5 illustrates the amount of overlap when sequence length is restricted to 1000 bases on either side of the variant (2 kb sequences). Figure 5(a) shows the percentage of all overlaps (of any length) within each chromosome, and Figure 5(b) shows the percentage of sequences that overlap by at least 25 percent. As shown in this figure, almost all chromosomes have only 1 or 2 percent overlapping sequences, and only 1 or 2 percent overlap by more than 25 percent.



(a) All 2 kb sequences with overlap

(b) 2 kb sequences with 25 percent overlap

Figure 5: 2 kb sequence overlap between null and eQTL sequences.

This analysis indicated that using sequences of length 2001 base pairs is a good compromise that both maximised computational efficiency and minimised sample overlap.

3 Large scale study

3.1 Materials and methods

The main study used HOMER [Heinz et al., 2010a], BaMM! [Siebert and Söding, 2016] and DREME [Bailey, 2011] motif algorithms to analyse 10,499 genomic sequences centred at the cis eQTL discovered in Lloyd-Jones et al. [2017]. The null sequences were generated using 2 kb sequences centred around a random sample of 10,499 null variants from the Lloyd-Jones et al. [2017] gene expression study.

As for the pilot study, all sequences were constructed using the hg19 reference genome [Lander et al., 2001]. An R program was written drawing on sequence information provided by the 'BSgenome.Hsapiens.UCSC.hg19' package [Team TBD, 2014]. Cis eQTLs that regulated more than one gene were repeated in the original 11,204 cis eQTLs, which resolved to 10,499 cis eQTLs once duplicates were removed. The same number of null sequences, centred on a random selection of non-eQTL variants, were also produced. The R program created bedfiles, which were used for HOMER annotation data, and fasta files, which were used as target and background by the motif finding algorithms.

The following sections detail the parameters that were used as input for each of the algorithms, and the output of the algorithm.

3.1.1 Detailed description of application of DREME, HOMER and BaMM!motif algorithms

DREME parameters and output

DREME runs from the terminal using the following command:

```
/usr/bin/dreme[options]
```

Options used were:

- the specified output directory (overwriting if necessary);
- the positive sequence file;
- the negative sequence file

The default motif E-value of 0.05 was used as the threshold.

As output, DREME provides two main files: the PWM file of motifs, and an html file listing all motifs as sequence logos, with both an E-value and an 'Unerased' E-value which is calculated without erasing sites of previously found motifs. The html file provides further information which includes the number of target and background sequences that contain the motif, together with the corresponding P-value and E-value (calculated by multiplying the P-value by the number of motifs found). The P-value was calculated for each motif using Fisher's Exact Test for enrichment of the motif in the positive sequences.

As part of the MEME suite, DREME provides opportunity to submit the motif(s) to other algorithms within the MEME suite. These other algorithms include: 'Tomtom', which will search for similar known motifs; 'MAST' and 'FIMO' which will search sequences for known motifs and provide enrichment details; 'GOMO' which will identify Gene Ontology terms for the motif; and 'SpaMo' which will search for possible transcription factor complexes.

HOMER parameters and output

HOMER can use both bedfiles and fasta sequences. With fasta sequences HOMER runs from the terminal using the following command:

```
findMotifs.pl <positive sequence file> fasta <output directory> -fasta <negative sequence file>
[options]
```

Options used were:

- -mask: mask motifs once found, as well as oligos immediately adjacent to the site that overlap with at least one nucleotide.

- -len 4,5,6,7,8,9,10,11,12,13,14 : find motifs of these lengths

The number of motifs per length requested (default=25), plus the length of the oligo seeds (default=1,2,3), were left at default.

An important time saving feature used by the HOMER algorithm is an initial scan of the sequences using HOMER's data bank of known motifs that match to transcription factors. Scanning for known PWMs is much faster than searching for motifs de novo, and any matches that are found can be masked from the sequences, thereby reducing the search space for the de novo search. All motifs found in this initial scan are placed in a folder named 'knownResults' as a set of sequence logos.

HOMER also normalises its oligo enrichment counts according to the 'ATGC' content of each oligo. A file with the extension '.autonorm.tsv' provides details of overall nucleotide content and the normalisation factor for each oligo.

The full set of motifs found by HOMER is provided in a file with the extension '.all.motifs'. In this file, the PWMs are organised firstly by length, secondly by p-value. The default number of motifs found per motif length is 25. Given that motifs were requested for fourteen different lengths i.e. 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14, the file contained $14 \cdot 25 = 350$ PWM's. Each PWM begins with a header such as:

```
>ATGATTCAATTACC 114-ATGATTCAATTACC 10.16267 -42.58457 0.0
T:666.0(6.34%),B:378.0(3.67%),P:1e-18
Tpos:1015.5,Tstd:559.1,Bpos:988.7,Bstd:749.1,StrandBias:0.1,Multiplicity:1.05
```

- ">" + Consensus sequence

- Motif name (usually the same as the consensus sequence)

- Log odds detection threshold, used to determine bound vs. unbound sites

- log P-value of enrichment
- 0.0 A place holder for backward compatibility
- Occurence Information separated by commas, including:
  - T: number of target sequences with motif, % of total of total targets
  - B: number of background sequences with motif, % of total background
  - P: final enrichment p-value
- Motif statistics separated by commas:
  - Tpos: average position of motif in target sequences (0 = start of sequences)
  - Tstd: standard deviation of position in target sequences
  - Bpos: average position of motif in background sequences (0 = start of sequences)
  - Bstd: standard deviation of position in background sequences
  - StrandBias: log ratio of + strand occurrences to - strand occurrences.
  - Multiplicity: The average number of occurrences per sequence in sequences with 1 or more binding sites.

HOMER also tries to match each de novo motif found using its data bank of transcription factors, and provides a number of files describing the statistical details of each matched motif with its possible transcription factor.

HOMER has an extensive set of webpages describing different ways in which its algorithms can be used. This includes a caution that only motifs that are 'very enriched' should be considered to be robust - HOMER uses a cumulative binomial distribution to calculate P-values and suggests a cut-off of $p < 1E - 50$ for robust significance. This cut-off was adopted throughout this project.

As described above, the motif header in the PWMs provides the average position of the motif together with its standard deviation. Further information can be recovered using another tool that is part of the HOMER software. The 'annotatePeakspl' command has options that can be used to recover general genomic information about the set of sequences, as well as the specific position of the motif in each sequence. An option that was used in this project is the '-hist' option, which provides data that can be used to create a graph of the distance of the motif from the centre of the sequence over the set of 10,499 sequences.

BaMM parameters and output

BaMM runs from the terminal using the following command:

```
BaMMmotif <output directory> <positive sequence file> [options]
```

Options used were:

- –negSequenceSet: the set of null sequences
- –reverseComp: use both strands of dna (this was default for both DREME and HOMER)
- –maxPValue: the maximum P-value of the PWMs (The maximum P-value was changed from the default of 1 to $p < 0.05$, but applying this option seemed to have no affect on the results obtained)

As output BaMM provides a set of matrices which the writers of the algorithm suggest be referred to as a set of 'BaMMs' rather than PWMs. Each 'BaMM' is preceded by a header listing the number of target sequences containing the motif and the E-value. This file follows the standardised MEME formatting so that it could be submitted directly to the MEME suite for further processing if desired.

In addition to the 'BaMM' file, the BaMM algorithm provides a PWM file created using a zero order markov model. It provides a 'MotifFile.txt' which lists each motif as a sequence, together with its E-Value and the number of target sequences containing the motif (this is the same information as is provided in the header to each 'BaMM'). A very long file (with the extension 'Pvals.txt') lists, for every motif, each sequence that contains the motif (as a sequence number), the start position of the motif in the sequence, the strand (+ or -) and the P-value of the individual motif in the sequence. P-values were calculated using the log-odds score calculated by dividing the probability of the motif in the positive sequences by the probability of the motif in the background sequences [Siebert and Söding, 2016]. Finally, another file with the extension 'sequence.txt' provides a key that links the sequence number listed in the 'Pvals.txt' file to the relevant eQTL provided in the original fasta file submitted to the algorithm.

3.1.2 Computer resources

All algorithms were run and computations performed on a personal computer (Dell Latitude E6320 with 8GB RAM; Intel Core i5-2520M CPU @2.50GHz x 4). Calculation of run times are provided as approximates, since the algorithms were run several times with different parameters and data. The approximate run times for the algorithms were as follows:

- HOMER: took approximately 6 hours and required no extra memory. Other tasks could be performed whilst HOMER ran in the background.
- DREME: took approximately 15 hours and required no extra memory, but other tasks were slowed down considerably.
- BaMM: took approximately 8 hours to produce almost all the data, but then an extra 24 hours to produce the final log likelihood. This was not necessary information, so the program was concluded manually (using 'Control+C') once all other files were produced. The program required an extra 32GB of virtual memory, and no other tasks could be performed whilst it was running.

3.1.3 Validation of algorithm results

To validate the results of the algorithms with respect to the null sequences, a number of checks were performed:

1. Two types of analysis were performed with respect to the null sequences:
   (a) The first analysis used the algorithms to check whether the null sequences contained any motifs that were enriched in the null sequences compared to the eQTL sequences.

It was expected that few, relatively unimportant, motifs would be found.

(b) The second analysis performed a random split of all null sequences into two groups, and used the algorithms to check whether enriched motifs might be found in the first group compared to the second group. It was expected that the two groups would be generally similar and that no motifs would be found.

2. A further check used the AME algorithm [Buske et al., 2010] to provide enrichment values for all motifs found by the DREME, HOMER and BaMM algorithms.

3.1.4 Analysis of the algorithm results

The motif files were formatted and subjected to the Stamp algorithm, which performed the following procedure:

1. Alignment of motifs
2. Creation of a cladogram to illustrate similarities of motifs
3. Matching of motifs to transcription factors using the JASPAR data base [Mathelier et al., 2016]

A number of programs (both in Python and R) were written and incorporated into scripts to further process these initial results, as described in Results. HOMER's annotation algorithm was also used to calculate average distance from the variant (both eQTL and null) across all motifs and sequences.

3.2 Results

Table 6 provides the results of all three algorithms with respect to the total number of motifs found that were enriched in eQTL sequences compared to null sequences, as well as the number of motifs that meet the stringent cut-off of $p < 1E - 50$.

A visual examination of the large number of HOMER motifs found that many of the smaller

Table 6: **Motifs enriched in eQTL sequences compared to null sequences**

All Motifs			Motifs With $p < 1E - 50$		
DREME	HOMER	BaMM	DREME	HOMER	BaMM
114	238	99	44	141	38

motifs were subsets of the larger motifs. Although masking prevents motif repetition during a single search, the masking is removed for a new search for a motif of a different length. After all motif subsets within the HOMER motifs were removed, the number of significant HOMER motifs was reduced from 141 to 123.

Before further analysis of these motifs was conducted, validation of algorithms was conducted using the null sequences. The first check was a search for algorithms that were enriched in the null sequences compared to the eQTL sequences. Results of this search are provided in Table 7:

Table 7: **Motifs enriched in null sequences compared to eQTL sequences**

All Motifs			Motifs With $p < 1E - 50$		
DREME	HOMER	BaMM	DREME	HOMER	BaMM
103	235	278	39	0	168

These results are unexpected in the case of DREME and BaMM algorithms. HOMER returned a large number of motifs, but none of them met the stringent cut-off threshold. DREME found approximately the same number of motifs in the null sequences as in the eQTL sequences, and BaMM found many more (168 significant motifs in the null compared to 38 in the eQTL sequences).

The BaMM results become more puzzling in the second check, which was of two randomly chosen groups of null sequences. It was expected that the algorithms should find no significant motifs in one of these groups compared to the other. Results are provided in Table 8

Table 8: **Motifs enriched in null Group A compared to null Group B**

All Motifs			Motifs With $p < 1E - 50$		
DREME	HOMER	BaMM	DREME	HOMER	BaMM
0	55	237	0	0	132

The results for both DREME and HOMER are as expected. DREME found no motifs at all, whilst HOMER found some motifs at a very low level of significance ($PValue > 1E - 8$). BaMM, however, found a large number of motifs. Given this result, the test was repeated for the BaMM algorithm, reversing the groups. BaMM found a similar number of motifs significantly enriched in Group B compared to Group A. Even more puzzling is the fact that the most highly enriched motif for Group A (CTACTAAAAATACAAAA) is also the most highly enriched motif for Group B. One possible explanation is that, although the BaMM algorithm is very successful at finding motifs, it fails in its ability to discriminate between groups.

Given this difficulty, the need to validate algorithm enrichment values became imperative. To this end, the AME algorithm [Buske et al., 2010] wihin the MEME suite was used to provide discriminative enrichment values for all significant motifs found by all algorithms. The AME algorithm is not able to perform de novo motif detection, but it can provide enrichment values for given motifs. Motifs found by AME to be enriched with a $P - value < 1E - 50$ were selected for further processing by the Stamp algorithm. Final significant motif figures are provided in Table 9.

Table 9: **Enriched according to AME selection**

prior AME Selection			After AME Selection		
DREME	HOMER	BaMM	DREME	HOMER	BaMM
44	123	38	43	119	28

As demonstrated in Table 9 , 1 of the DREME motifs, 5 of the HOMER motifs, and 10 of the BaMM motifs were discarded due to this process.

This final set of 190 motifs were formatted for Stamp processing and submitted to the Stamp algorithm. Figure 6 illustrates the alignment produced by Stamp as a circular cladogram, and Figure 7 provides a detail of this cladogram. This cladogram was produced using EvolView software [He et al., 2016].

The whole cladogram (Figure 6) is too small a scale to view the alignments clearly, but the colouring (DREME=red; HOMER=blue; BaMM=green) provides information about the spread of the motifs. The detail (Figure 7) indicates that close alignment was found for many motifs.

The Stamp algorithm also provided five possible transcription factor matches for every motif, as

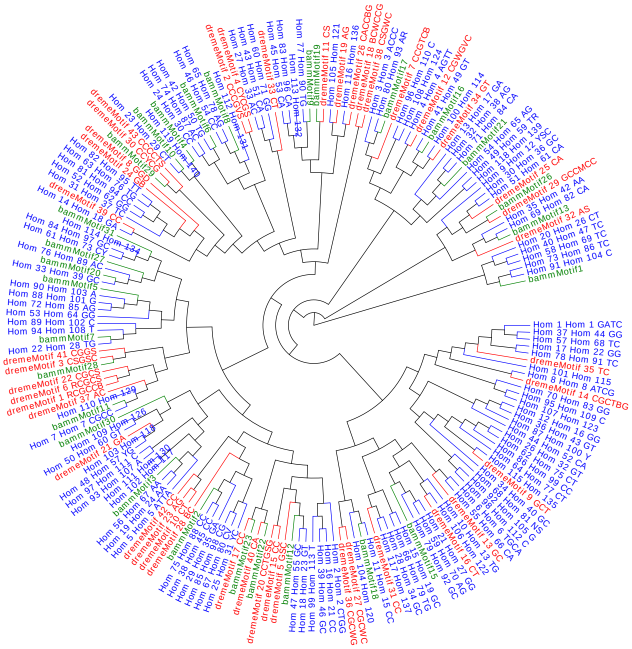


Figure 6: **Circular cladogram of Stamp motif alignment. blue: HOMER, red: DREME, green: BaMM**

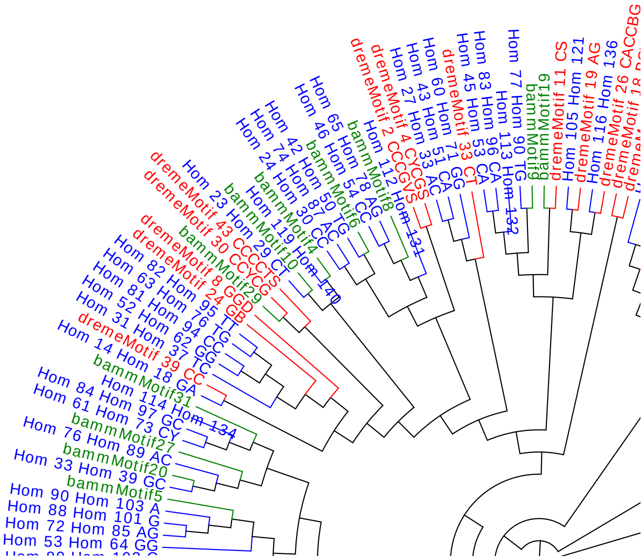


Figure 7: **Detail of Stamp motif alignment. blue: HOMER, red: DREME, green: BaMM**

illustrated in Table10. The

Table 10: **Transcription factor matches for Hom_1_Hom_1 motif**

>	Hom_1_Hom_1	
Snail	$1.28E - 005$	GATCACTTGA
Nkx2-5	$1.70E - 003$	TCAAGTGATC
Mycn	$1.85E - 003$	GATCACTTGA
Arnt	$1.87E - 003$	GATCACTTGA
Pax5	$1.92E - 003$	———TCAAGTGATC

Given that $190/cdot5 = 950$ transcription factors were provided, the match probability needed to be adjusted accordingly. Applying a Bonferroni correction to an initial $p < 0.05$ meant that $p < 5.0E - 5$ were required. For the TF matches provided in Table10, only the first (Snail; $p = 1.28E - 005$) was kept and the remainder were discarded.

Using $p < 5.0E - 5$ as the match threshold, 37 of the 190 motifs were matched to 28 TFs. Of these, 9 TFs were matched to motifs from more than one algorithm. Table 11 lists these nine TFs with best match p-values.

Table 11: **Transcription factors matched to motifs from at least 2 algorithms**

TF	Best HOMER Match	Best DREME Match	Best BaMM Match
IRF1	8.15E-008	No motif found	3.74E-005
Snail	1.41E-0.008	4.23E-011	No motif found
deltaEF1	2.94E-005	9.60E-007	No motif found
Fos	2.66E-006	No motif found	2.46E-005
ID1	1.50E-005	No motif found	6.61E-006
MEF2A	1.95E-005	No motif found	2.12E-008
Prrx2	9.64E-006	No motif found	3.63E-005
SQUA	6.27E-006	No motif found	3.71E-007
ESR1	4.96E-005	3.38E-005	No motif found

Of these transcription factors:

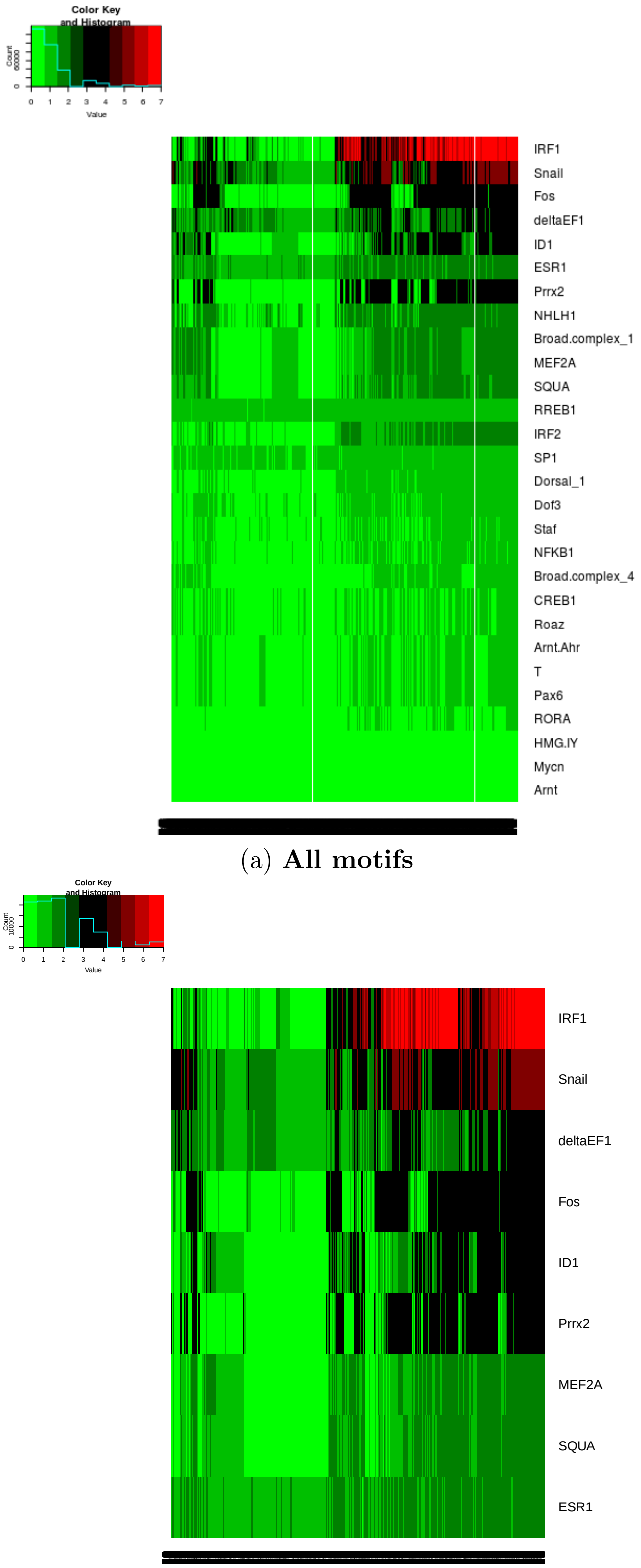
- IRF1 and deltaEF1 are both associated with the immune system
- Snail, MEF2A, and ESR1 play a role in cell proliferation and have associations with tumours
- Fos and ID1 also play a role in cell proliferation and differentiation
- Prrx2 also has a possible role in cellular proliferation

A heatmap was used to illustrate the abundance of motifs over all sequences. Figure 8 shows the abundance of all motifs over all sequences, as well as the abundance of motifs matching the 9 robust TFs over all sequences. As illustrated in Figure 8(b), a relatively large number of motifs that match IRF are present in more than half of the 10,499 sequences, but motifs matching Snail, DeltaEF1 and ESR1 are present in moderate to high numbers over all sequences. Data to create the heatmaps was accessed through the HOMER suite of software, which includes annotation data for motifs. This data includes the presence/absence of the motif on each sequence. Motifs found by both DREME and BaMM were formatted appropriately to be submitted for annotation by HOMER, as well as the HOMER motifs.

Although these top transcription factors bind to motifs that are enriched in eQTL sequences, this is not to say that they do not exist in null sequences at all. The motif annotation tool that is part of the HOMER software was used to detect the presence of the enriched motifs (those matching TFs) in the null sequences as well as the eQTL sequences. Figure 9 illustrates the relative abundance of motifs matching all TFs across both eQTL and null sequences, and figure 10 compares the number of motifs matching the robust nine TFs found by at least two algorithms.

A further tool offered by the HOMER motif annotation tools is the distance tool, which calculates average distance from the centre for each motif. This information was used to create graphs of average distance from the variant (both null and eQTL) for all motifs that bind to transcription factors. Figure 11 illustrates the extent to which binding motifs are clustered around either the null variant or the eQTL for all transcription factors.

The same tool was used to measure the distance from the centre of the motifs that were relatively enriched on the null sequences. Figure **??** demonstrates that these motifs are much more widely scattered than motifs that are enriched in the eQTL sequences, with no clustering around the null variant. The incidence of these motifs is also illustrated in Figure 13, which indicates that these motifs were far less abundant in both eQTL and null sequences than were the motifs enriched in eQTL sequences. Although the top transcription factor(SOX9) has an incidence of close to 25,000 binding sites over the 10,499 sequences, the graph quickly dwindles to smaller numbers.



(a) **All motifs**

(b) **Motifs matching robust TFs**

Figure 8: **Abundance of motifs over all sequences**

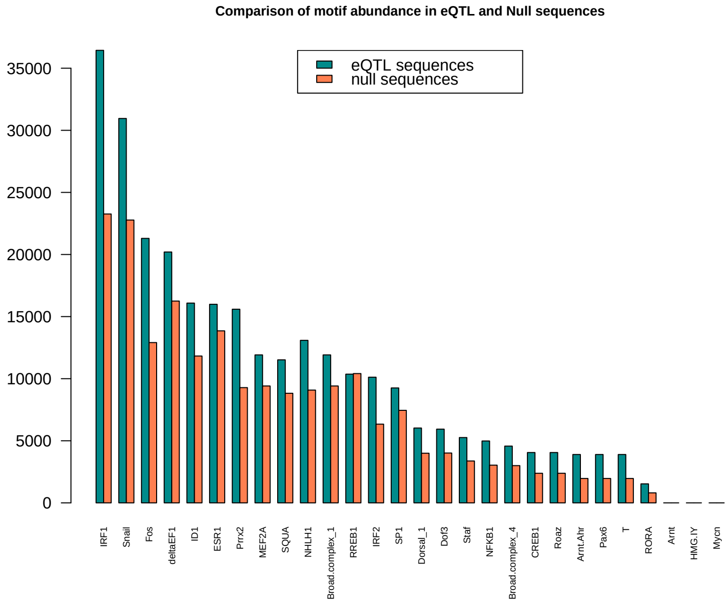


Figure 9: **Comparison of abundance of binding sites for all TFs within eQTL and null sequences**

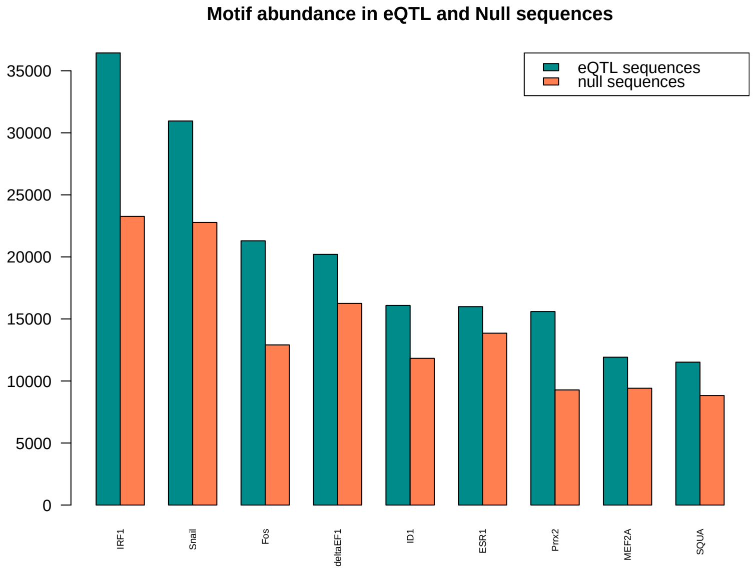


Figure 10: **Comparison of abundance binding sites for nine TFs within eQTL and null sequences**

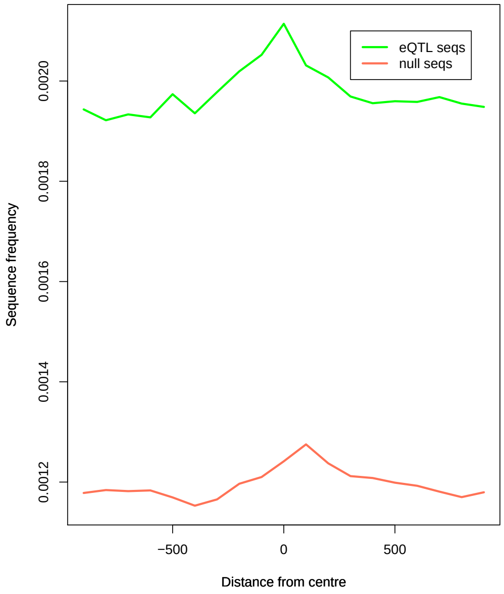


Figure 11: **Comparison of distance from the centre for all TF binding sites for eQTL and null sequences**

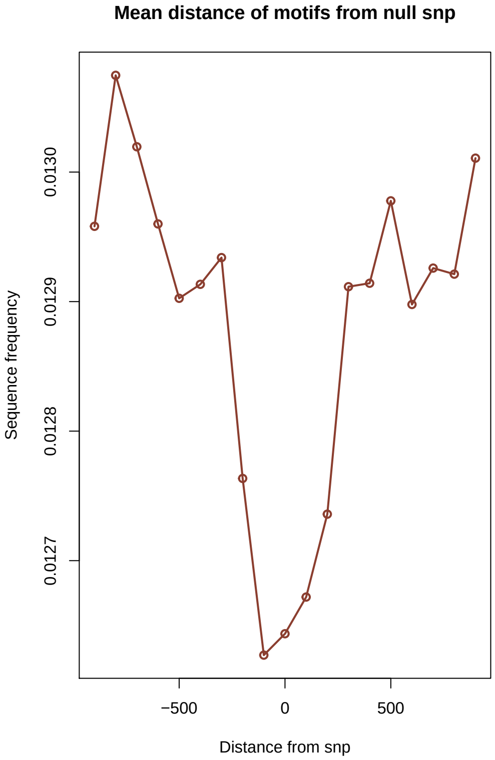


Figure 12: **Motifs enriched in null sequences:distance from the null variant**

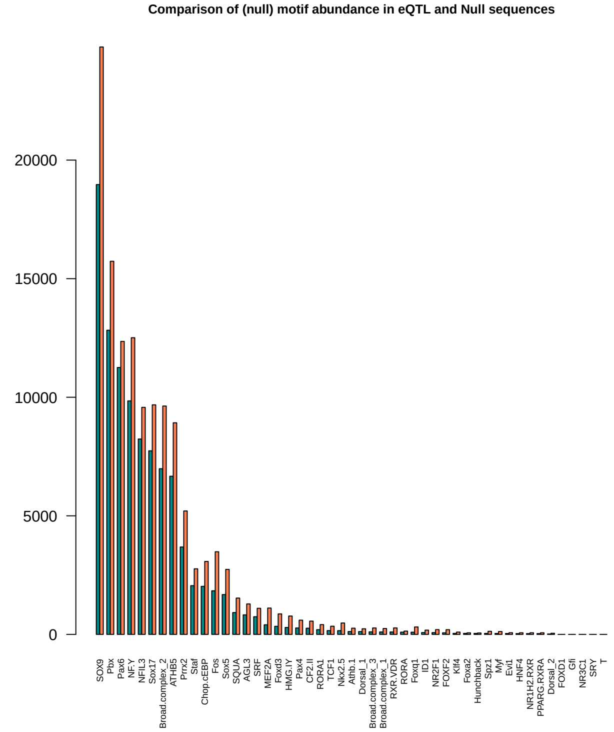


Figure 13: **Comparison of motifs enriched in null sequences across null and eQTL sequences**

Table 12: **Genomic position as a percentage of sequences**

	Exon	Intron	Promoter	TTS	Intergenic	3'UTR	5'UTR
eQTL seqs	2.8	51.5	6.2	4.8	28.1	4.5	0.6
null seqs	0.3	38.0	0.9	0.6	59.4	0.5	0.0

Given the strong focus of research on transcription factors as the major form of gene regulation, and the presence of many TFs in promoter regions, the HOMER annotation tool for genomic regions was used to compare the position of eQTL and null sequences throughout the genome. Table 12 provides the percentage of sequences found in the different regions for both sets of sequences.

Although the focus of this section of the paper has been on motifs that were strongly matched to transcription factors, these motifs only represent 19.5% of the data. No strong matches were found for 80.5% of motifs.

4 Discussion

The aim of this project was to identify, quantify, and characterize dna motifs that are enriched in sequences centred on eQTLs found in whole blood, compared with sequences centred on variants which have no effect on gene expression. Many algorithms exist which can find motifs de novo. Reviews of these motifs have failed to establish any “gold standard” and usually recommend using more than one algorithm to substantiate results. This study found that the HOMER algorithm was reliable and easily run on a normal laptop. Although it found many more motifs than either DREME or BaMM, its results were validated by the AME algorithm. It also offers a useful suite of software to annotate motifs. The usefulness of this algorithm is indicated by its increasing number of citations, which has shifted from 2,370 when the pilot study was written up to the present level of 2,754 citations. Although BaMM was a promising algorithm, its ability to discriminate between target and background sequences seems doubtful, and any results need to be validated using other algorithms. DREME is part of the well regarded MEME suite, and results indicate that it discriminates between target and background appropriately. DREME also found a substantial number of motifs that were validated by AME, although only four of these were strongly matched to TFs and only two motifs matched TFs that were also bound by motifs found by other algorithms.

The Stamp algorithm proved to be extremely useful in organising data. Its alignment tool provided a useful indication of the validity of different algorithms, in that it illustrated which algorithms found motifs that were similar. Its use of the JASPAR database to match motifs to TFs was a convenient tool that allowed easy identification of TFs strongly matched to the binding sites found through the motif detection algorithms.

The Bonferroni correction applied to the match p-values meant that some false negatives may have been introduced to the TF matches, but was necessary given the wealth of data. Strong matches were found for nine TFs with binding sites found by at least two algorithms. Of these TFs, two were directly connected with the immune system. Interferon Regulatory Factor 1 (IRF1) activates genes involved in both innate and acquired immune responses and DeltaEF1 inhibits the expression of Interleukin-2 gene expression, which regulates white blood cell activity **??**. All the other TFs were associated with regulation of cell proliferation. ESR1 in particular has been targeted for its role in breast cancer. Mutations in the binding domain of ESR1 have been implicated in hormone resistance and anti-estrogen therapies **??**

HOMER annotation tools were used to provide more detailed enrichment figures across the different binding sites. As illustrated in Figure 10, substantial numbers of these sites occur in the null sequences as well as the eQTL sequences. Figure 11 also indicates that the binding sites in the null sequences also have a slight tendency to clump towards the variant. However, as illustrated in this figure, the tendency of the variant to impact directly upon the binding site through proximity is more marked in the case of the eQTLs, and the combination of less abundance and increased distance may be responsible for lack of effect of the null variants. These factors are more exaggerated for motifs that are enriched in the null sequences; in fact Figure 12 indicates a strong tendency for these motifs to be found well distant from the variant.

The impact of the variant upon binding sites may also become more important with the location of the variant in promoters, exons, introns and 3'UTR in mRNA's. Given the higher incidence of eQTLs, compared to null variants, in all these sites, small perturbations in binding sites become more important. These locations underline the importance of the unmatched motifs, in that TFs tend to play a large part in promoter regions, but eQTLs in other regions such as introns or the 3'UTR point to other functions, such as splice sites or miRNA binding sites. Although a number of transcription factors were moderately abundant over all sequences, Figure 8 indicates broad bands of sequences in which these TFs were not in evidence. Figure 8 is only of the incidence of motifs that had been matched to TFs. A useful investigation would be of the genomic regions of these bands of sequences, as well as of the abundance of the unmatched motifs across the sequences. Other, unmatched binding sites might be more important within these bands of sequences.

One possible example is the TATAAT box, a well conserved sequence centered around 10 bp upstream of transcription initiation. This motif forms the binding site for a subunit of the RNA polymerase. Despite the high degree of conservation it extremely rare to find a promoter that matches this consensus exactly. The activity of the promoter is related to how well it matches the consensus sequence and so the activity of each gene can be fine tuned by how much its region deviates from the consensus. The affinity of a DNA binding site is typically correlated with how well the site matches the consensus sequence. Not all positions in a binding site are equally forgiving of mismatches and not all mismatches have the same effect [D'haeseleer, 2006]. One motif found by HOMER is the sequence GCATATTCTCAC, with its complementary strand CG$\boxed{\text{TATAAG}}$AGTG. An important immediate task in further research would be the investigation of the genomic regions in which these motifs fall.

5 Conclusion

This project succeeded in its aim of identifying algorithms that can reliably detect and organise motifs enriched in target sequences compared to background sequences. The application of these algorithms resulted in the identification of a number of transcription factors in the vicinity of eQTLs. The identified TFs are important in immune system responses as well as regulation of cell proliferation. Disruptions in the binding sites of these TFs are likely to affect genes particularly expressed in whole blood, a factor supportive of these TFs as part of the mechanism by which these eQTLs regulate the expression of genes in whole blood. Comparison with the same binding sites in sequences associated with null variants indicated that these TFs were less abundant in null sequences with less tendency to cluster in the proximity of the null variant.

Further investigation should identify the actual sequences that harbour large numbers of binding sites for TFs (as indicated by the red and black areas of the heatmaps in Figure 8) and identify the associated genes that are affected by these particular eQTLs. This might corroborate or contradict the likelihood of these transcription factors as part of the regulatory framework.

Transcription factor binding sites account for a very small percentage of the motifs detected by the algorithms, and, given the genomic regions of the eQTL sequences, it is likely that these motifs form binding sites for proteins other than transcription factors. Further research might focus on the sequences that show little evidence of abundant binding sites for TFs and the genomic regions in which these sequences are located, which might provide clues to the likely function of binding sites enriched in these regions.

References

Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

Algoshareify. Gibbs sampling (parts 1 - 4) - YouTube, 2012. URL `https://www.youtube.com/watch?v=cAtCTVVqCVU`.

Timothy L Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2):W202—-W208, 2009.

Martin Beckerman. Gene Regulation in Eukaryotes. *Molecular and Cellular Signaling*, pages 385–410, 2005.

Valentina Boeva. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. *Frontiers in Genetics*, 7:24, feb 2016. ISSN 1664-8021. doi: 10.3389/fgene.2016.00024. URL `http://journal.frontiersin.org/Article/10.3389/fgene.2016.00024/abstract`.

Fabian A. Buske, Mikael Bodén, Denis C. Bauer, and Timothy L. Bailey. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, 26 (7):860–866, apr 2010. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq049. URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq049.`

Thomas H Cormen, Charles Leisersen, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, Massachusetts, 3rd edition, 2009. URL `https://books.google.com.au/books?hl=en&lr=&id=aefUBQAAQBAJ&oi=fnd&pg=PR5&dq=cormen+2009&` `2009&f=false`.

Modan K Das and Ho-Kwok Dai. A survey of DNA motif finding algorithms. *BMC bioinformatics*, 8(7):S21, 2007.

Sanjoy Dasgupta, Christos H Papadimitriou, and Umesh Vazirani. *Algorithms*. McGraw-Hill, Inc., 2006.

Erik Dassi and Alessandro Quattrone. DynaMIT: the dynamic motif integration toolkit. *Nucleic acids research*, 44(1):e2—-e2, 2016.

Patrik D'haeseleer. What are DNA sequence motifs? *Nature biotechnology*, 24(4):423–425, 2006.

Daniel J Gaffney. Global properties and functional complexity of human gene regulatory variation. *PLoS genetics*, 9(5):e1003501, 2013.

Zilong He, Huangkai Zhang, Shenghan Gao, Martin J. Lercher, Wei Hua Chen, and Songnian Hu. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic acids research*, 44(W1):W236–W241, 2016. ISSN 13624962. doi: 10.1093/nar/gkw370.

Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, may 2010a. ISSN 10972765. doi: 10.1016/j.molcel.2010.05.004. URL `http://linkinghub.elsevier.com/retrieve/pii/S1097276510003667`.

Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Homer Software for motif discovery and next generation sequencing analysis, 2010b. URL `http://homer.ucsd.edu/homer/`.

Vincent J Henry, Anita E Bandrowski, Anne-Sophie Pepin, Bruno J Gonzalez, and Arnaud Desfeux. OMICtools: an informative directory for multi-omic data analysis. *Database*, 2014. doi: 10.1093/database/bau069. URL `http://omictools.com/`.

Peter Huggins, Shan Zhong, Idit Shiff, Rachel Beckerman, Oleg Laptenko, Carol Prives, Marcel H

Schulz, Itamar Simon, and Ziv Bar-Joseph. DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics*, 27(17):2361–2367, 2011.

Maxwell A Hume, Luis A Barrera, Stephen S Gisselbrecht, and Martha L Bulyk. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic acids research*, 43(Database issue):D117–22, jan 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1045. URL http://www.ncbi.nlm.nih.gov/pubmed/25378322 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4383892.

Sachi Inukai, Kian Hong Kock, and Martha L Bulyk. Transcription factor–DNA binding: beyond binding site motifs. *Current Opinion in Genetics & Development*, 43:110–119, 2017.

Narayan Jayaram, Daniel Usvyat, and Andrew C. R. Martin. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1298-9. URL https://link.springer.com/content/pdf/10.1186%2Fs12859-016-1298-9.pdf http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1298-9.

D. Karolchik, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(90001):493D–496, jan 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh103. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh103.

Holger Kirsten, Hoor Al-Hasani, Lesca Holdt, Arnd Gross, Frank Beutner, Knut Krohn, Katrin Horn, Peter Ahnert, Ralph Burkhardt, Kristin Reiche, Jorg Hackermuller, Markus Loffler, Daniel Teupser, Joachim Thiery, and Markus Scholz. Dissecting the genetics of the human transcriptome identifies novel trait-related trans -eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Human Molecular Genetics*, 24(16):4746–4763, aug 2015. ISSN 0964-6906. doi: 10.1093/hmg/ddv194. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddv194.

Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, Fedor A Kolpakov, and Vsevolod J Makeev. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, jan 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1106. URL `http://academic.oup.com/nar/article/46/D1/D252/4616875`.

Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland

Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, feb 2001. ISSN 0028-0836. doi: 10.1038/35057062. URL `http://www.nature.com/doifinder/10.1038/35057062`.

CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, and JC Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science (New York, N.Y.)*, 262(5131):208–214, 1993. doi: 10.1126/science.8211139.

Luke R Lloyd-Jones, Alexander Holloway, Allan McRae, Jian Yang, Kerrin Small, Jing Zhao, Biao Zeng, Andrew Bakshi, Andres Metspalu, Manolis Dermitzakis, and Others. The genetic architecture of gene expression in peripheral blood. *The American Journal of Human Genetics*, 100(2):228–237, 2017.

Shaun Mahony and Panayiotis V Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic acids research*, 35 (suppl_2):W253—-W258, 2007. doi: 10.1093/nar/gkm272. URL `https://watermark.silverchair.com/gkm272.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf`

Anthony Mathelier, Oriol Fornes, David J. Arenillas, Chih Yu Chen, Gr??goire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, Allen W. Zhang, Fran??ois Parcy, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1176.

Alexandra C Nica and Emmanouil T Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368(20120362), 2013. doi: 10.1098/rstb.2012.0362. URL `http://dx.doi.org/10.1098/rstb.2012.0362`.

Athma A Pai, Jonathan K Pritchard, and Yoav Gilad. The genetic and mechanistic basis for variation in gene regulation. *PLoS genetics*, 11(1):e1004857, 2015.

John E Reid and Lorenz Wernisch. STEME: efficient EM to find motifs in large data sets. *Nucleic acids research*, 39(18):e126—-e126, 2011.

A. Sandelin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001):91D–94, jan 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh012. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh012`.

Ignacio E Schor, Jacob F Degner, Dermot Harnett, Enrico Cannavò, Francesco P Casale, Heejung

Shim, David A Garfield, Ewan Birney, Matthew Stephens, Oliver Stegle, and Eileen E M Furlong. Promoter shape varies across populations and affects promoter evolution and expression noise. *Nature Genetics*, 49(4):550–558, 2017. ISSN 1061-4036. doi: 10.1038/ng.3791. URL http://www.nature.com/doifinder/10.1038/ng.3791.

Matthias Siebert and Johannes Söding. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, 44(13):6055–6069, 2016. ISSN 13624962. doi: 10.1093/nar/gkw521.

Alexander J. Stewart, Joshua B. Plotkin, Sridhar Hannenhalli, and Joshua B. Plotkin. Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(3):973–985, 2012. ISSN 00166731. doi: 10.1534/genetics.112.143370.

Chunxiao Sun, Hongwei Huo, Qiang Yu, Haitao Guo, and Zhigang Sun. An affinity propagation-based DNA motif discovery algorithm. *BioMed research international*, 2015.

Team TBD. BSgenome.Hsapiens.UCSC.hg19: Full genome sequences for Homo sapiens (UCSC version hg19), 2014.

Ngoc Tam L Tran and Chun-Hsi Huang. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biology Direct*, 9 (1):4, feb 2014. ISSN 1745-6150. doi: 10.1186/1745-6150-9-4. URL http://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-9-4.

Chuhu Yang, Eugene Bolotin, Tao Jiang, Frances M Sladek, and Ernest Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1):52–65, 2007.

Zizhen Yao, Kyle L. MacQuarrie, Abraham P. Fong, Stephen J. Tapscott, Walter L. Ruzzo, and Robert C. Gentleman. Discriminative motif analysis of high-throughput dataset. *Bioinformatics*, 30(6):775–783, mar 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt615. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt615.

Muhammad A. Zabidi and Alexander Stark. Regulatory Enhancer - Core-Promoter

Communication via Transcription Factors and Cofactors. *Trends in Genetics*, 32(12):801–814, 2016. ISSN 13624555. doi: 10.1016/j.tig.2016.10.003. URL http://dx.doi.org/10.1016/j.tig.2016.10.003.

Yipu Zhang, Ping Wang, and Maode Yan. An Entropy-Based Position Projection Algorithm for Motif Discovery. *BioMed research international*, 2016, 2016.