Supplementary Material
MEME-ChIP: motif analysis of large DNA datasets

Philip Machanick and Timothy L. Bailey

March 18, 2011

1 Introduction

This supplement provides further detail of how to run the MEME-ChIP web service, including further detail of the example reported in the paper, and further examples illustrating differences in the type of output. We provide more detail on repeating the given example, and describe tools we use including design choices in setting the tools' parameters.

2 Availability and Repeating an Example

The features described here are available in MEME from version 4.6.0 onwards, which can be obtained following the "Downloads" link at `http://meme.nbcr.net`. The downloadable software includes all the command line tools invoked by the web service. You can also download the databases for use with the MEME suite. The file `motif_databases_current.tgz` contains the JASPAR CORE 2009 database used in examples we describe here. The example available on the web service (as the DNA sequences example accessible from the MEME-ChIP form, illustrated in Figure 1) is the Klf1 data set [4] we briefly describe in the paper, with sequences only containing Ns, arising from repeat masking [3], removed (reducing the number of sequences from 945 to 940). To run that example, go to the MEME web service at `http://meme.nbcr.net/`, click on any of the links to the MEME-ChIP service, click on the text **Sample DNA Input Sequences**, select all the text, copy it, return to the input form and paste in the sequences. Add your email address as required on the form and click **Start search**.

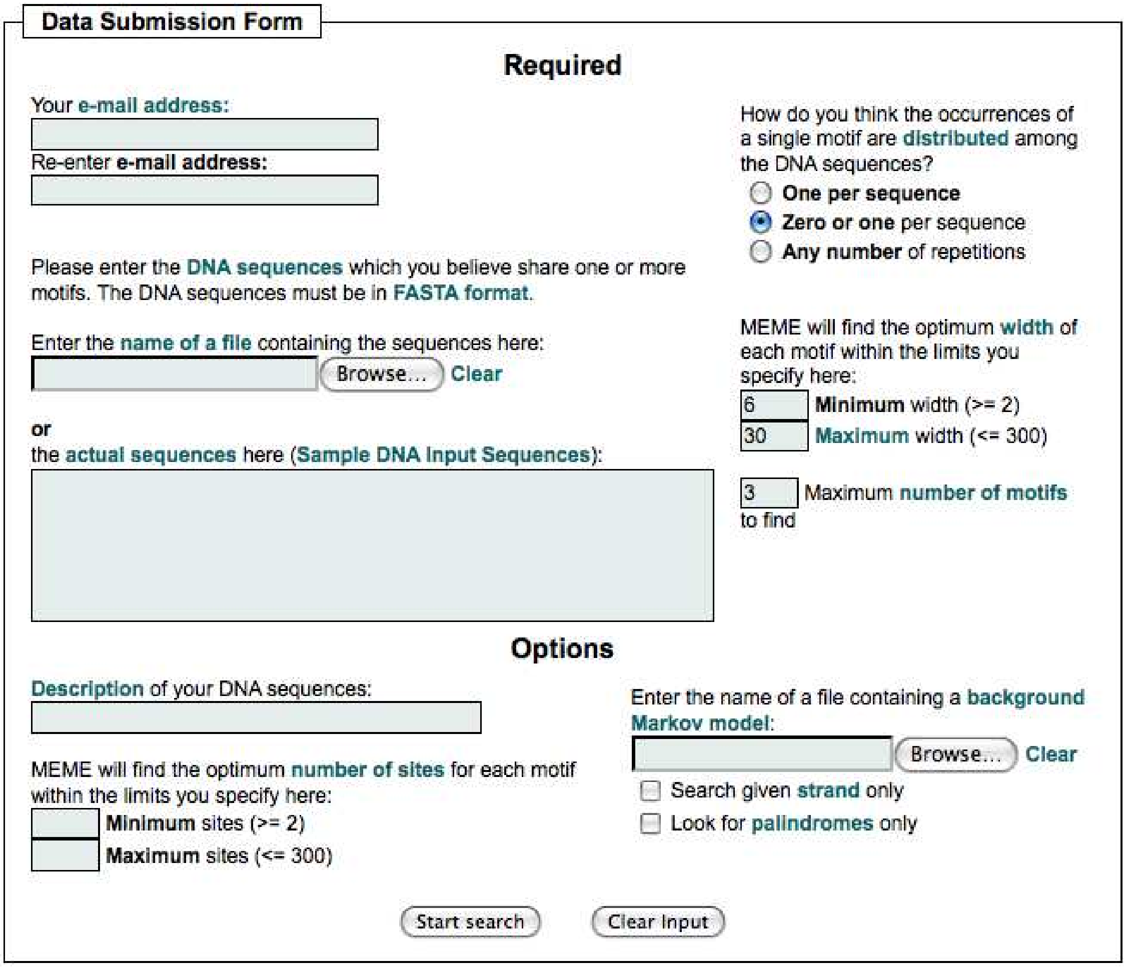
Total run time should be about an hour, depending on load on the server.

3 Tools

We present more detail of the tools used here, including design choices for parameters. To illustrate how the tools work in a real example, we described how they apply to the Klf1 data set described in the paper. It is possible to recreate the results either by rerunning on the web server, or by using the command line tools if you have installed a recent MEME distribution.

3.1 Design Choices

Since MEME has run time $O(n^2)$, as described in the paper, we limit data to MEME to 600 sequences, randomly sampled without replacement if the original data contains more than this number of sequences. For most of the other tools (with the exception of MAST, which uses the entire data set), we use all the sequences, but trimmed to the central 100 bases. The logic for trimming the sequences to the central 100 bases is that with good peak calling we expect the binding site for the chipped TF to be near the centre of each sequence that contains a binding site. Making the sequences wider than needed adds noise. While peak calling methods are highly variable [2], we expect enough common data sets to have sufficiently accurate peak calling for focusing on the central 100 bases to be useful.



Data Submission Form

Required

Your e-mail address:

Re-enter e-mail address:

Please enter the DNA sequences which you believe share one or more motifs. The DNA sequences must be in FASTA format.

Enter the name of a file containing the sequences here:

or
the actual sequences here (Sample DNA Input Sequences):

How do you think the occurrences of a single motif are distributed among the DNA sequences?

○ One per sequence
● Zero or one per sequence
○ Any number of repetitions

MEME will find the optimum width of each motif within the limits you specify here:

6 Minimum width (>= 2)
30 Maximum width (<= 300)

3 Maximum number of motifs to find

Options

Description of your DNA sequences:

MEME will find the optimum number of sites for each motif within the limits you specify here:

Minimum sites (>= 2)
Maximum sites (<= 300)

Enter the name of a file containing a background Markov model:

□ Search given strand only
□ Look for palindromes only

Figure 1: **The MEME-ChIP input form.** You can set a few options for MEME, but not for the other tools. Above the text box for typing in "actual sequences", you can click on the text "Sample DNA Input Sequences" for the Klf1 data set.

We configure the defaults for MEME to find relatively long motifs, if they exist, with a default upper width of 30. A user can alter MEME settings, in the same way as with the existing MEME web form, which we retain as a separate web service. We expect that most users who will make significant variations on MEME parameters will use the older input form, since the default settings here are designed to complement the other tools in the MEME-ChIP suite.

3.2 Detail of Main Example

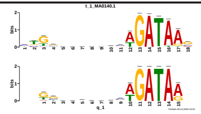
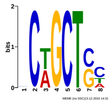
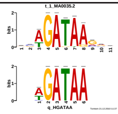
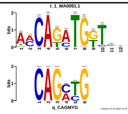
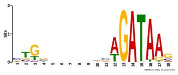
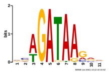
algorithm	*first motif*	*second motif*
MEME		
DREME		
AME		

Figure 2: **Two most significant motifs found by the MEME, DREME and AME algorithms in the SCL ChIP-seq data.** For MEME and DREME motifs, the motif Logo (bottom) is shown aligned with the most similar JASPAR motif Logo (top) if the similarity is significant ($E \leq 0.05$).

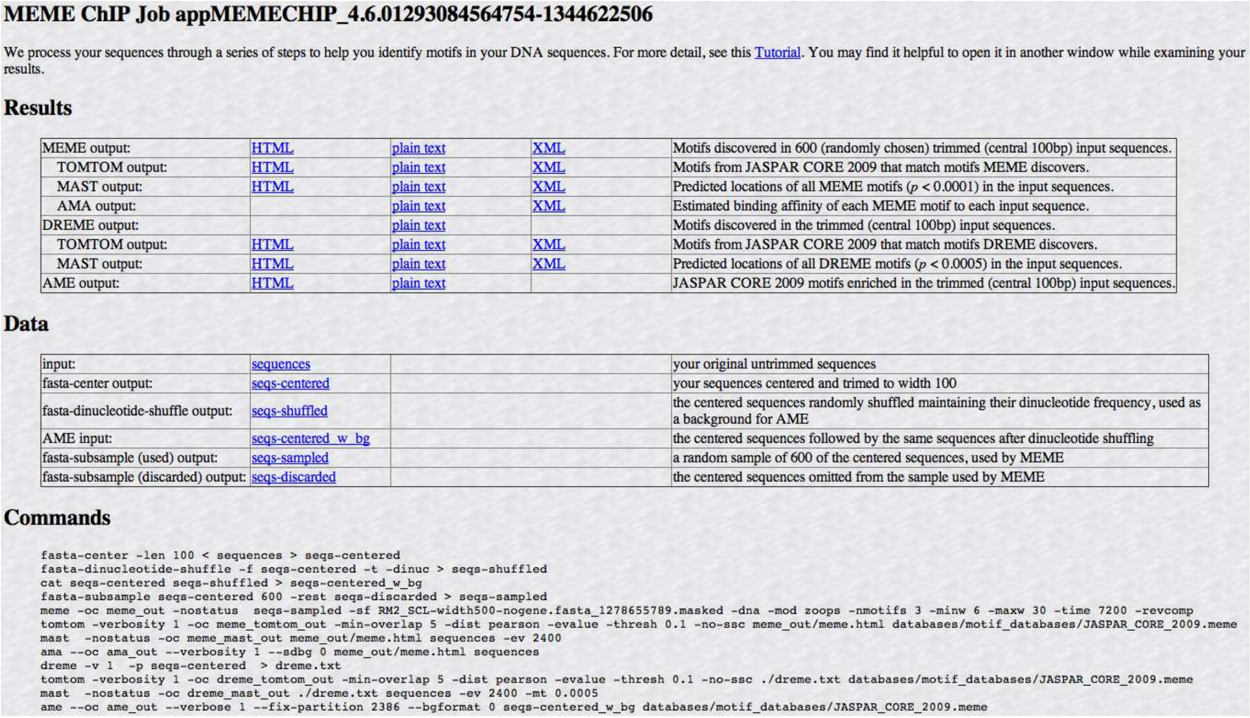
To demonstrate the functionality of MEME-ChIP we use it to analyze the ChIP-seq peak regions reported by [1] for SCL (also called Tal1), a key regulator of erythropoeisis. We use the UCSC genome browser to extract repeat-masked 500 bp regions surrounding each of the 2400 binding peaks defined in Supplementary Table 1 of [1]. MEME finds three statistically significant motifs, DREME finds nine. The subsequent analysis by TOMTOM shows that two out of the three MEME motifs significantly match known vertebrate DNA-binding motifs in the JASPAR CORE database, as do six of the nine DREME motifs. The AME motif enrichment analysis reports enrichment of 15 vertebrate motifs from JASPAR CORE and confirms many of the MEME/DREME–TOMTOM predictions. (Complete results are available at `http://meme.nbcr.net/meme/doc/examples/memechip_example_output_files`.)

The most significant motif discovered by MEME (MEME $E \approx 10^{-163}$) is a 15 bp motif that is characteristic of binding by a complex involving SCL and GATA-1 and TOMTOM identifies this motif as being significantly similar (TOMTOM $E \approx 10^{-21}$) to JASPAR motif MA0140.1 (Fig. 2, row 1, column 1). This JASPAR motif is also reported by AME to be most significantly enriched in the input sequences (AME $p \approx 10^{-230}$). DREME is unable to find this motif since it limits its search to motifs with widths up to 8 bp. The most significant motif found by DREME (DREME $E \approx 10^{-253}$) closely matches the JASPAR GATA-1 motif MA0035.2 (TOMTOM $E \approx 0.004$, Fig. 2, row 2, column 1). This is also the second most significant motif reported by AME.

MEME and DREME also discover variants of the well-known E-box (`CANNTG`) binding motif of SCL (Fig. 2, column 2). In both cases this is the second most significant motif found ($E \approx 10^{-40}$ and $p \approx 10^{-34}$ for MEME and DREME motifs, respectively). Neither motif exactly matches the JASPAR E-box motif (MA0091.1), but the similarity of the DREME motif according to TOMTOM is almost statistically significant ($E \approx 0.07$). The AME analysis reports this motif as being the third most significantly enriched vertebrate motif (AME $p \approx 10^{-19}$).

MEME-ChIP executes the following commands in creating the output in this example:

```
fasta-center -len 100 < sequences > seqs-centered
fasta-dinucleotide-shuffle -f seqs-centered -t -dinuc > seqs-shuffled
cat seqs-centered seqs-shuffled > seqs-centered_w_bg
fasta-subsample seqs-centered 600 -rest seqs-discarded > seqs-sampled
meme -oc M_out seqs-sampled -dna -mod zoops -nmotifs 3 -minw 6 -maxw 30 -time 7200 -revcomp
tomtom -oc M_TT_out -min-overlap 5 -dist pearson -evalue -thresh 0.1 -no-ssc M_out/meme.html DB
mast -oc M_mast_out M_out/meme.html sequences -ev 2400
ama --oc ama_out --sdbg 0 M_out/meme.html sequences
```



MEME ChIP Job appMEMECHIP_4.6.01293084564754-1344622506

We process your sequences through a series of steps to help you identify motifs in your DNA sequences. For more detail, see this Tutorial. You may find it helpful to open it in another window while examining your results.

Results

MEME output:	HTML	plain text	XML	Motifs discovered in 600 (randomly chosen) trimmed (central 100bp) input sequences.
TOMTOM output:	HTML	plain text	XML	Motifs from JASPAR CORE 2009 that match motifs MEME discovers.
MAST output:	HTML	plain text	XML	Predicted locations of all MEME motifs ($p < 0.0001$) in the input sequences.
AMA output:		plain text	XML	Estimated binding affinity of each MEME motif to each input sequence.
DREME output:		plain text		Motifs discovered in the trimmed (central 100bp) input sequences.
TOMTOM output:	HTML	plain text	XML	Motifs from JASPAR CORE 2009 that match motifs DREME discovers.
MAST output:	HTML	plain text	XML	Predicted locations of all DREME motifs ($p < 0.0005$) in the input sequences.
AME output:	HTML	plain text		JASPAR CORE 2009 motifs enriched in the trimmed (central 100bp) input sequences.

Data

input:	sequences		your original untrimmed sequences
fasta-center output:	seqs-centered		your sequences centered and trimed to width 100
fasta-dinucleotide-shuffle output:	seqs-shuffled		the centered sequences randomly shuffled maintaining their dinucleotide frequency, used as a background for AME
AME input:	seqs-centered_w_bg		the centered sequences followed by the same sequences after dinucleotide shuffling
fasta-subsample (used) output:	seqs-sampled		a random sample of 600 of the centered sequences, used by MEME
fasta-subsample (discarded) output:	seqs-discarded		the centered sequences omitted from the sample used by MEME

Commands

```
fasta-center -len 100 < sequences > seqs-centered
fasta-dinucleotide-shuffle -f seqs-centered -t -dinuc > seqs-shuffled
cat seqs-centered seqs-shuffled > seqs-centered_w_bg
fasta-subsample seqs-centered 600 -rest seqs-discarded > seqs-sampled
meme -oc meme_out -nostatus  seqs-sampled -sf RM2_SCL-width500-nogene.fasta_1278655789.masked -dna -mod zoops -nmotifs 3 -minw 6 -maxw 30 -time 7200 -revcomp
tomtom -verbosity 1 -oc meme_tomtom_out -min-overlap 5 -dist pearson -evalue -thresh 0.1 -no-ssc meme_out/meme.html databases/motif_databases/JASPAR_CORE_2009.meme
mast  -nostatus -oc meme_mast_out meme_out/meme.html sequences -ev 2400
ama --oc ama_out --verbosity 1 --sdbg 0 meme_out/meme.html sequences
dreme -v 1  -p seqs-centered  > dreme.txt
tomtom -verbosity 1 -oc dreme_tomtom_out -min-overlap 5 -dist pearson -evalue -thresh 0.1 -no-ssc ./dreme.txt databases/motif_databases/JASPAR_CORE_2009.meme
mast  -nostatus -oc dreme_mast_out ./dreme.txt sequences -ev 2400 -mt 0.0005
ame --oc ame_out --verbose 1 --fix-partition 2386 --bgformat 0 seqs-centered_w_bg databases/motif_databases/JASPAR_CORE_2009.meme
```

Figure 3: **The MEME-ChIP output report.** Note the list of commands at the end allowing reproduction of the outputs on the command line. The output report includes explanatory text and a pointer to a tutorial.

```
dreme -p seqs-centered  > dreme.txt
tomtom -oc dreme_TT_out -min-overlap 5 -dist pearson -evalue -thresh 0.1 -no-ssc ./dreme.txt DB
mast -oc dreme_mast_out ./dreme.txt sequences -ev 2400 -mt 0.0005
ame --oc ame_out --verbose 1 --fix-partition 2386 --bgformat 0 seqs-centered_w_bg DB
```

These commands are as reported by the web service as part of its summary of outputs, as we illustrate in Figure 3. The only change needed to run the commands as reported by the web service is the original file name is renamed as `sequences`; in the above we have also removed options that do not change the results (they reduce inessential outputs) for brevity. We also abbreviate some of the names to shorten command lines. To recreate the example, you need to download the `JASPAR_CORE_2009.meme` file (as we describe in §2 and replace "`DB`" its name (with path if necessary). Finally, we remove the `-sf` flag from the MEME line, which documents the original file name but has no effect on the output, and the restriction on MEME execution time, `-time 7200`.

Each command has the following effect, and all the outputs are available by clicking on links in the **Results** and **Data** sections of the output summary:

- `fasta-center` – trims the sequences to a maximum width of 100 centered on the original sequence; we exclude any sequences that only contain Ns from the output of `fasta-center`.
- `fasta-dinucleotide-shuffle` – produces a dinucleotide-shuffled version of the centred sequences; we join these two files in with the centered sequences first then the shuffled sequences using `cat`.
- `fasta-subsample` – stores a random sample of up to 600 sequences (without replacement) of the original sequences in `seqs-sampled`; we save the unused sequences in `seqs-discarded` but we do not use them.
- `meme` – finds up to 3 motifs of width anything from 6 to 30 in the centred, sampled data. Each motif need not be found in every sequence (using the `zoops`, zero occurrences per sequence, moodel), and MEME searches both strands (`-revcomp`). We further process MEME outputs as follows:

  - `tomtom` – finds motifs in the JASPAR CORE database that match the motifs found by MEME.
  - `mast` – identifies locations in all of the sequences (without any trimming or sampling) that contain each motif MEME finds. We use `-ev 2400`, where the number is the total number of sequences, to ensure that matches in all sequences are reported, even those that are not statistically significant.
  - `ama` – calculates the binding affinity of each motif MEME finds to the centred sequences (the entire set, not the sample MEME uses).

- `dreme` – uses a regular expression-based search to find short motifs in the full centered data set. We further process DREME outputs as follows:
  - `tomtom` – finds motifs in the JASPAR CORE database that match the motifs found by DREME. DREME at time of writing does not generate logos for its motifs, so it is also useful to view this TOMTOM output to visualise the motifs.
  - `mast` – identifies locations in all of the sequences (without any trimming or sampling) that contain each motif DREME finds. We set a high threshold of 0.0005 for reporting a match for DREME, since DREME aims to find short motifs.

- `ame` – finds motifs in the JASPAR CORE database that are enriched in the sequences (the centered sequences without sampling). AME runs in fixed partition mode, with the foreground set the centred sequences and the background the dinucleotide-shuffled version of the centred sequences. Note the number `2400` on the command line, indicating the partition between the foreground and background. AME finds a p-value for a match of each motif in the database to each sequence, and counts those that are below a given threshold. It uses a Fisher exact test to calculate an overall p-value for each motif for the sequences set.

To run a different example on the command line, the numbers used in the `mast` and `ame` command lines can be obtained using the `getsize` utility included in the MEME distribution, which reports statistics on a FASTA file, including the number of sequences.

4 Comparing Results of Multiple Examples

In this section we examine outputs from Klf1 and Gata1 data sets, compare them with SCL, and show how we could arrive at similar conclusions to those of our earlier study of these data sets [4]. We do not present this as a new scientific finding, since it would be more convincing if we made a discovery that was not previously known to us, but present this example as indicative of how the MEME-ChIP suite can be used by biologists.

First, in Figure 4, we present a summary of outputs from DREME (all of those from Gata1 and Klf1; and the first nine from SCL). We highlight motifs that look similar, to illustrate how a biologist could use MEME-ChIP to look for patterns in data sets that may be related. In this case, we have highlighted three examples of similar-looking motifs that occur across all three examples, and one that we find in two of the examples. We use light blue shading to highlight GATA motifs, light green to highlight E-box motifs, pink to highlight CACC motifs, and an outline to highlight the final common motif.

Next, we examine motifs found by MEME across the three data sets. In Figure 5 we present the two motifs with lowest E-value that MEME finds in the three data sets.

References

[1] Mira T Kassouf, Jim R Hughes, Stephen Taylor, Simon J McGowan, Shamit Soneji, Angela L Green, Paresh Vyas, and Catherine Porcher. Genome-wide identification of TAL1's functional targets: Insights into its mechanisms of action in primary erythroid cells. *Genome Res*, 8:1064–1083, Jun 2010.

[2] Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(11s):S22–S32, 2009.

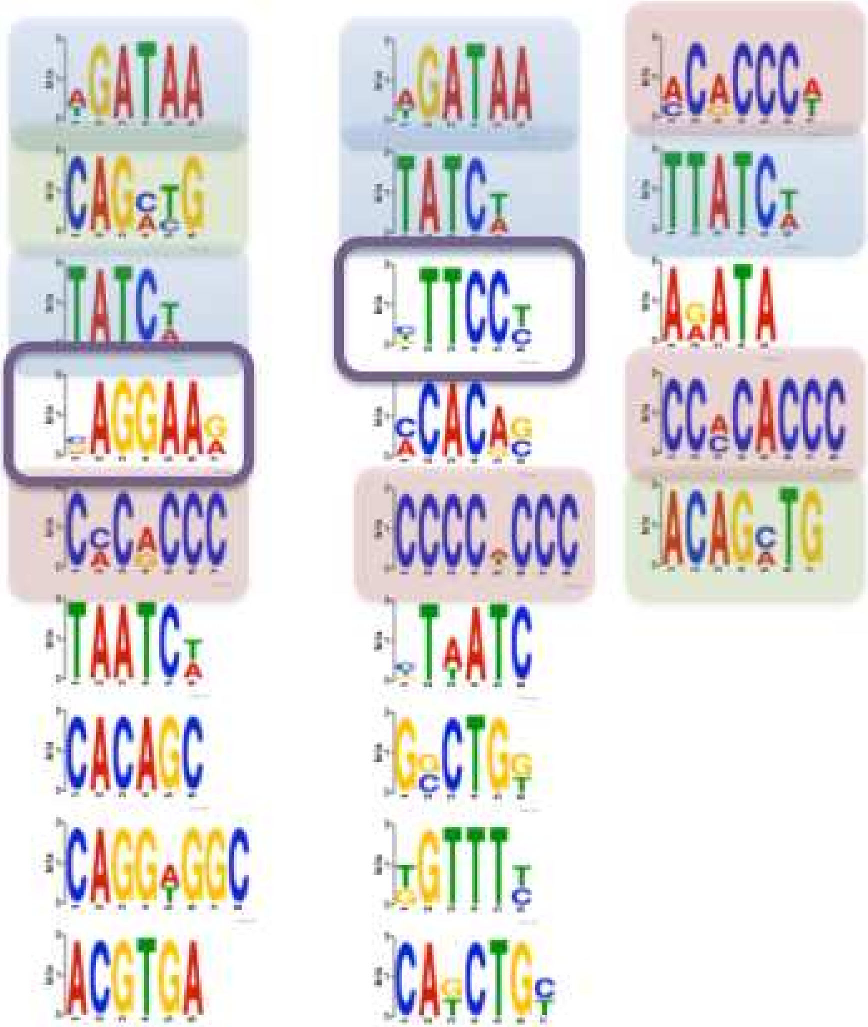


Figure 4: **Comparison of (left to right) SCL, Gata1 and Klf1 DREME motifs.** We highlight four different sets of similar-looking motifs. These motifs are all of the outputs for Gata1 and Klf1, and the first nine from SCL.

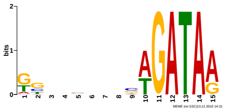
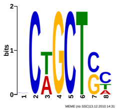
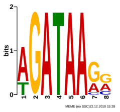
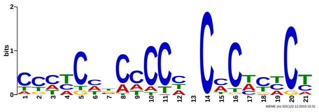
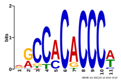
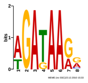
Data set	**Motif 1**	**Motif 2**
SCL		
E-value (sites)	2.2×10^{-164} (307)	1.6×10^{-40} (299)
Gata1		
E-value (sites)	3.1×10^{-68} (167)	1.8×10^{-21} (75)
Klf1		
E-value (sites)	2.6×10^{-237} (232)	1.2×10^{-32} (156)

Figure 5: **Two most significant motifs found by MEME in the SCL, Gata1 and Klf1 ChIP-seq data.** For each motif, we also give the E-value and the number of sites (sequences) out of 600 in which it is found.

[3] A. Smit, R. Hubley, and P. Green. RepeatMasker. Available at http://www.repeatmasker.org, ????

[4] Michael R Tallack, Tom Whitington, Wai Shan Yuen, Elanor N Wainwright, Janelle R Keys, Brooke B Gardiner, Ehsan Nourbakhsh, Nicole Cloonan, Sean M Grimmond, Timothy L Bailey, and Andrew C Perkins. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res*, 20(8):1052–1063, Aug 2010.