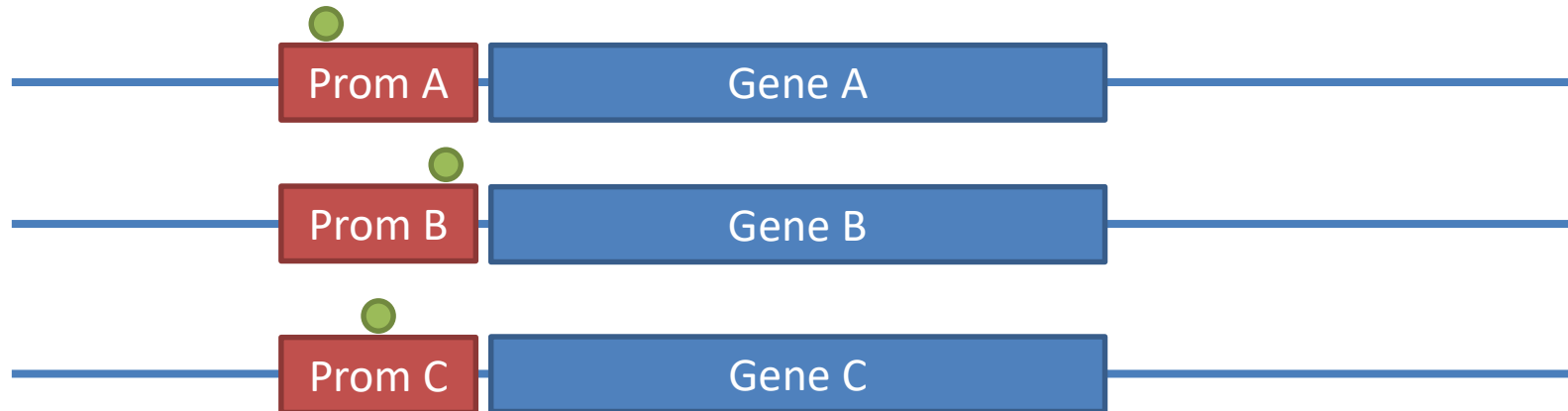
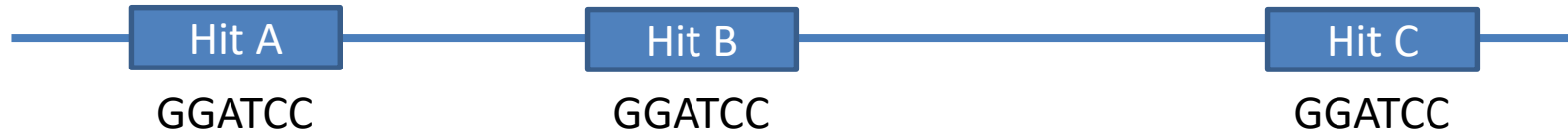


Motif Searching

Simon Andrews
simon.andrews@babraham.ac.uk
@simon_andrews

V2021-11

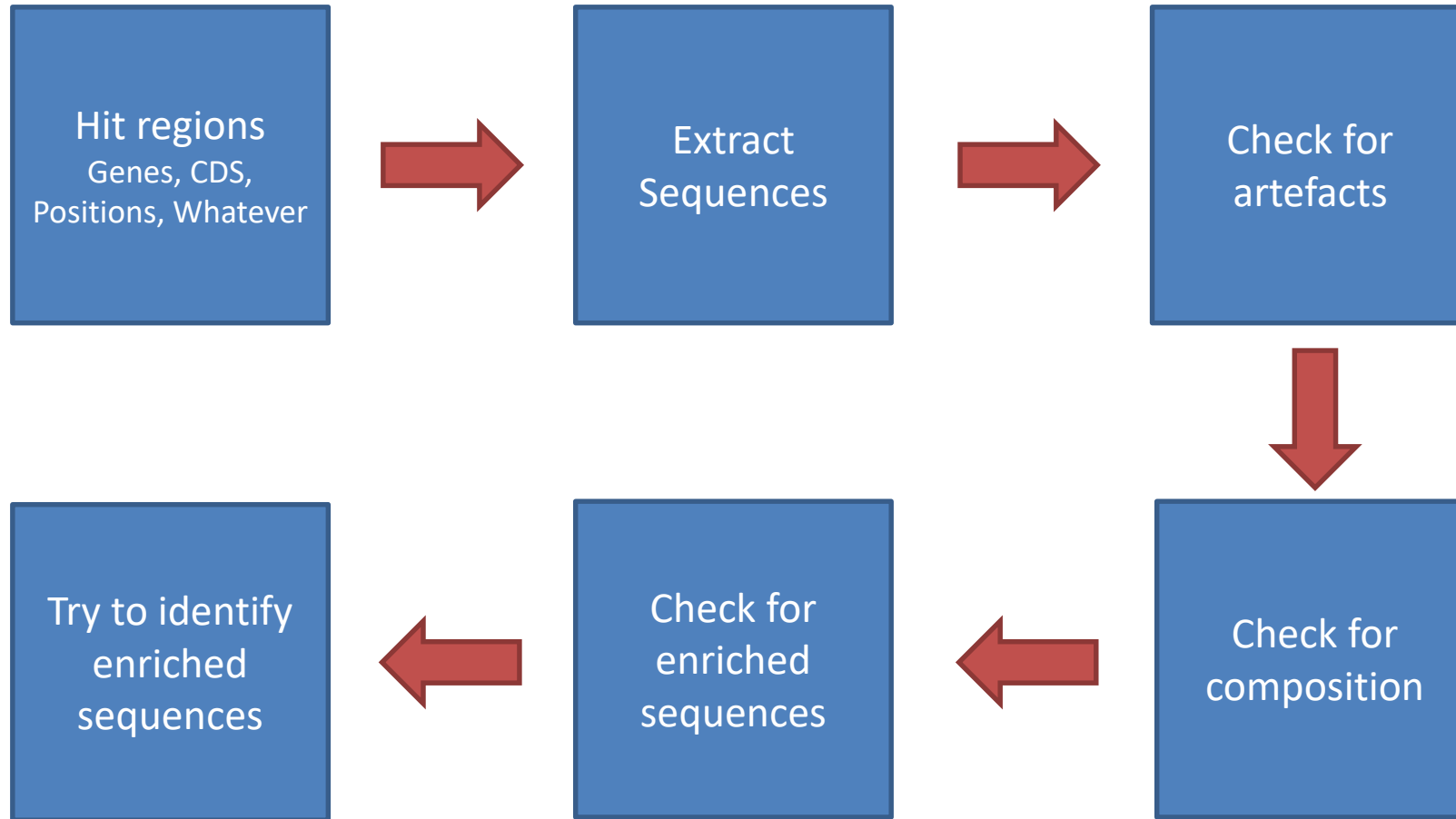
Rationale



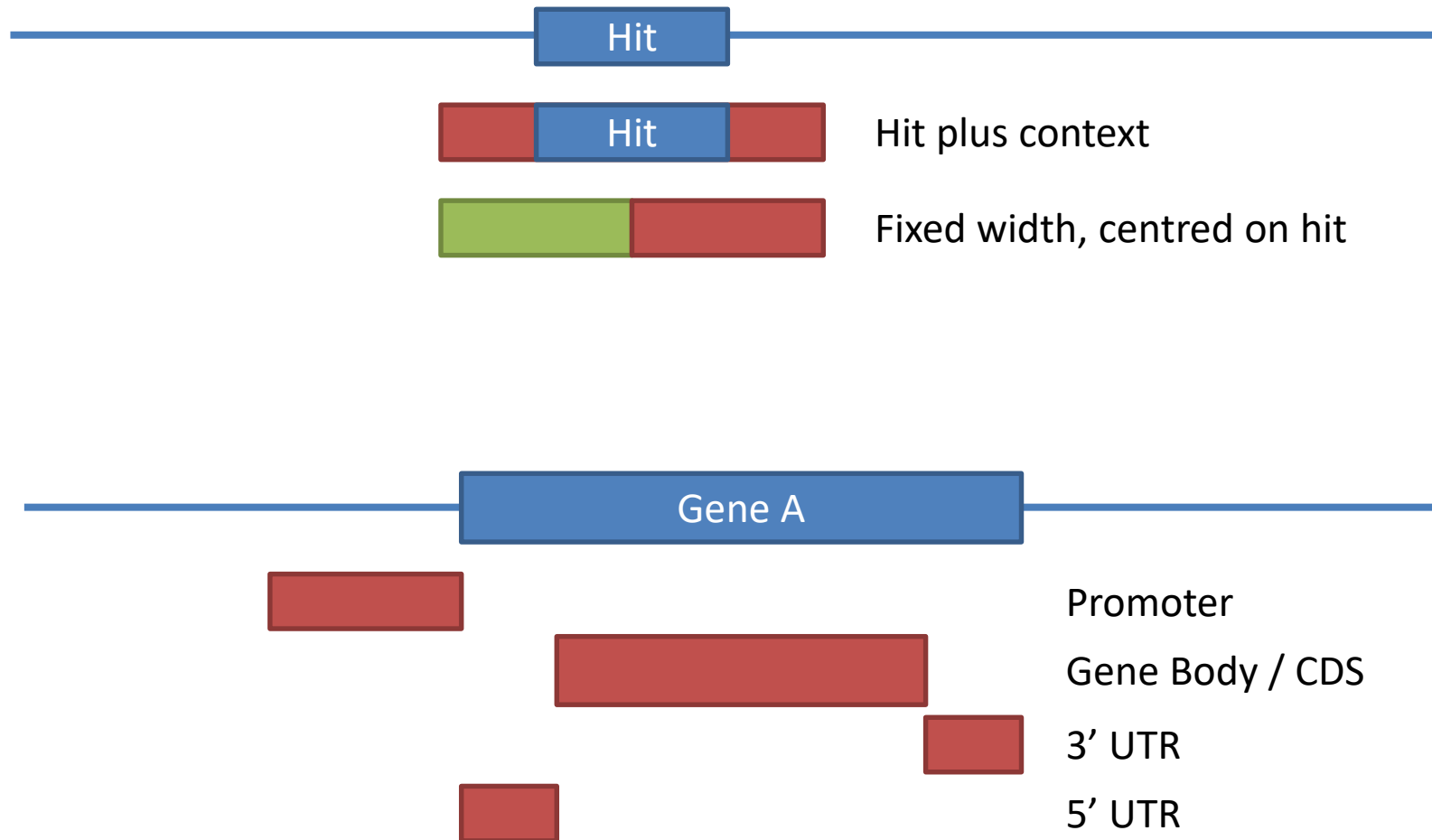
Basic Questions

- Does the sequence around my hits look unusual?
- Do specific sequences turn up more often than expected in my hits?
- If so, do the sequences look like any known functional sequence?

Basic Workflow



Deciding what to extract



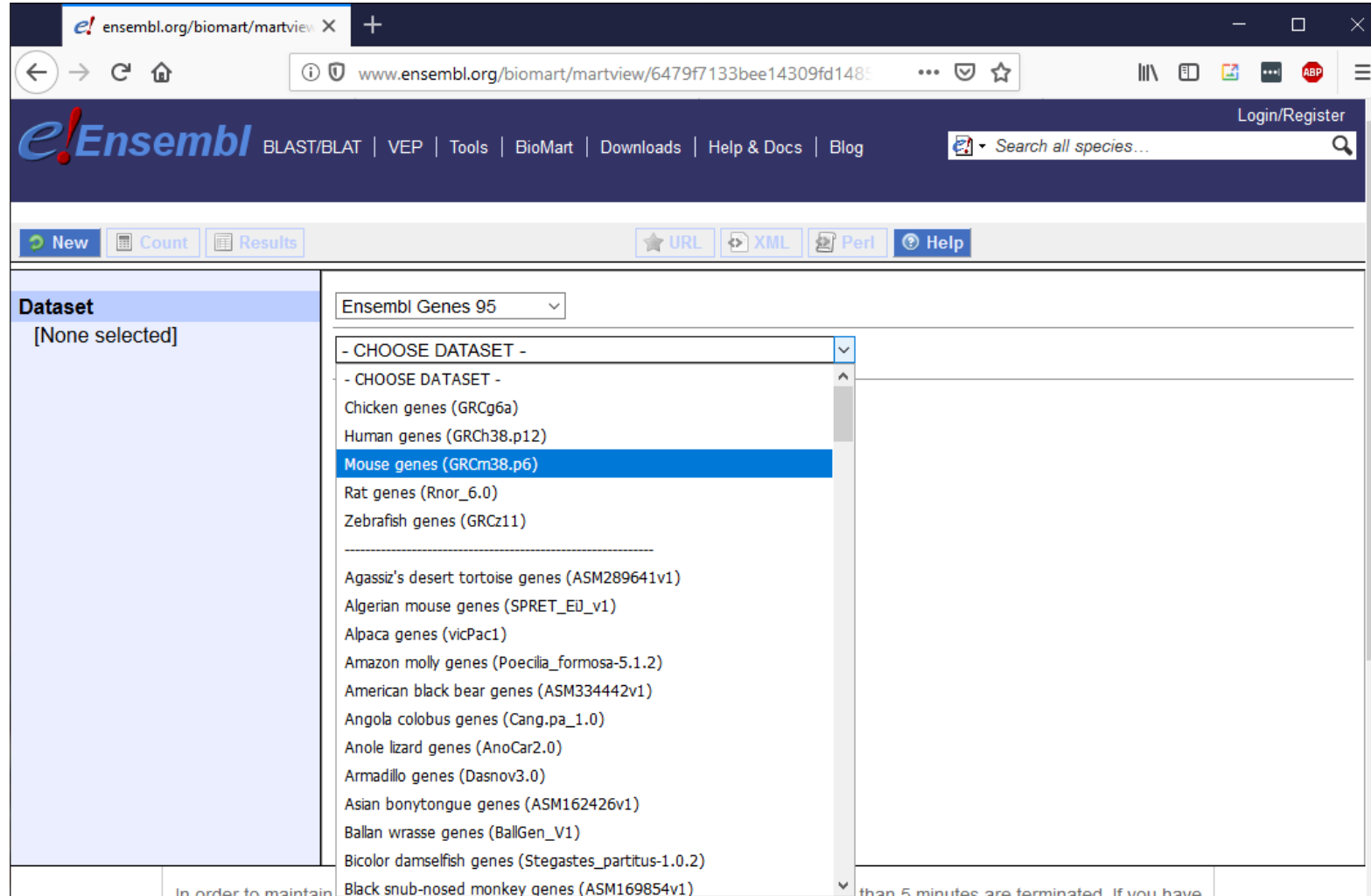
Extracting Sequence

- From positions
 - BEDTools
 - Genome Browsers*
 - Custom scripts

- From features
 - Genome Browsers*
 - BioMart

*not easily automatable for multiple sequences

BioMart – Selecting Assembly



The screenshot shows the Ensembl BioMart interface. The browser address bar displays www.ensembl.org/biomart/martview/6479f7133bee14309fd1485. The Ensembl logo is visible in the top left, and navigation links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog are in the top center. A search bar for species is on the top right. Below the navigation bar, there are buttons for 'New', 'Count', and 'Results'. A toolbar contains 'URL', 'XML', 'Perl', and 'Help' options. The main content area is divided into two sections. On the left, the 'Dataset' section shows '[None selected]'. On the right, a dropdown menu is open, displaying a list of datasets. The current selection is 'Ensembl Genes 95'. The dropdown list includes: '- CHOOSE DATASET -', '- CHOOSE DATASET -', 'Chicken genes (GRCg6a)', 'Human genes (GRCh38.p12)', 'Mouse genes (GRCm38.p6)', 'Rat genes (Rnor_6.0)', 'Zebrafish genes (GRCz11)', a separator line, 'Agassiz's desert tortoise genes (ASM289641v1)', 'Algerian mouse genes (SPRET_EU_v1)', 'Alpaca genes (vicPac1)', 'Amazon molly genes (Poecilia_formosa-5.1.2)', 'American black bear genes (ASM334442v1)', 'Angola colobus genes (Cang.pa_1.0)', 'Anole lizard genes (AnoCar2.0)', 'Armadillo genes (Dasnov3.0)', 'Asian bonytongue genes (ASM162426v1)', 'Ballan wrasse genes (BallGen_V1)', 'Bicolor damselfish genes (Stegastes_partitus-1.0.2)', and 'Black snub-nosed monkey genes (ASM169854v1)'. The 'Mouse genes (GRCm38.p6)' option is highlighted in blue.

<https://ensembl.org/biomart/martview>

BioMart – Specifying features

The screenshot shows the Ensembl BioMart interface. The browser address bar displays `ensembl.org/biomart/martview`. The Ensembl logo and navigation menu are visible at the top. The left sidebar shows the 'Dataset' as 'Mouse genes (GRCm38.p6)' and the 'Filters' tab is selected. The main content area is titled 'Please restrict your query using criteria below' and contains several filter sections:

- REGION:** (empty)
- GENE:**
 - Limit to genes (external references)... With CCDS ID(s) Only Excluded
 - Input external references ID list [Max 500 advised] Gene Name(s) [e.g. mt-Tp]
 - Gpr101
 - Fate1
 - Xlr3a
 - Cypt3
 - Limit to genes (microarray probes/probesets)... With AFFY MG U74A probe ID(s) Only Excluded
 - Input microarray probes/probesets ID list [Max 500 advised] AFFY MG U74A probe ID(s) [e.g. 96290_f_at]

BioMart – selecting seq region

The screenshot shows the Ensembl BioMart interface. The browser address bar displays `ensembl.org/biomart/martview/6479f7133bee14309fd1485`. The Ensembl logo and navigation links (BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, Blog) are visible at the top. A search bar contains the text "Search all species...".

The main interface is divided into several sections:

- Dataset:** Mouse genes (GRCm38.p6)
- Filters:** Gene Name(s) [e.g. mt-Tp]: [ID list specified]
- Attributes:** A list of attributes including Gene stable ID, Transcript stable ID, Flank (Gene), and Upstream flank []. The "Attributes" section is circled in red.
- Dataset:** [None Selected]

The central area contains a selection interface with the following elements:

- Instruction: "Please select columns to be included in the output and hit 'Results' when ready"
- Warning: "Missing non coding genes in your mart query output, please check the following [FAQ](#)"
- Radio button options: Features, Structures, Sequences, Variant (Germline), Homologues. The "Sequences" option is circled in red.
- Section: "SEQUENCES: Sequences (max 1)" with a diagram of a gene structure (exons and introns) and a red dashed line indicating a sequence region.
- Radio button options for sequence types: Unspliced (Transcript), Unspliced (Gene), Flank (Transcript), Flank (Gene), Flank-coding region (Transcript), Flank-coding region (Gene), 5' UTR, 3' UTR, Exon sequences, cDNA sequences, Coding sequence, Peptide. The "Flank (Gene)" option is circled in red.
- Section: "Upstream flank" with a checked checkbox and an input field containing "500".
- Section: "Downstream flank" with an unchecked checkbox and an empty input field.

BioMart – header info

The screenshot shows the Ensembl BioMart interface. The browser address bar is www.ensembl.org/biomart/martview/6479f7133bee14309fd1485. The left sidebar shows the dataset 'Mouse genes (GRCm38.p6)' and the 'Attributes' section with 'Gene name' selected. The main content area is titled 'Please select columns to be included in the output and hit 'Results' when ready'. It contains a message: 'Missing non coding genes in your mart query output, please check the following [FAQ](#)'. Below this, there are radio buttons for 'Features', 'Structures', 'Sequences' (selected), and 'Homologues'. There are also radio buttons for 'Variant (Germline)'. Under 'SEQUENCES:', there is a section for 'HEADER INFORMATION:' with a sub-section 'Gene Information' containing checkboxes for 'Gene name' (checked), 'Gene stable ID', 'Gene description', 'Source of gene name', 'Chromosome/scaffold name', 'Gene start (bp)', 'Gene end (bp)', 'Gene type', 'Ensembl Protein Family ID(s)', 'UniParc ID', 'UniProtKB/Swiss-Prot ID', and 'UniProtKB/TrEMBL ID'. There is also a 'Transcript Information' section with checkboxes for 'CDS start (within cDNA)', 'CDS end (within cDNA)', '5' UTR start', '5' UTR end', '3' UTR start', '3' UTR end', 'Transcript stable ID', 'Protein stable ID', 'Transcript type', 'Strand', 'Transcript start (bp)', 'Transcript end (bp)', 'Transcription start site (TSS)', and 'Transcript length (including UTRs and CDS)'. Finally, there is an 'Exon Information' section with checkboxes for 'CDS Length', 'CDS start', 'CDS end', 'Exon stable ID', 'Start phase', 'End phase', 'cDNA coding start', and 'cDNA coding end'.

BioMart - exporting

The screenshot shows the Ensembl BioMart interface. The browser address bar displays `www.ensembl.org/biomart/martview/6479f7133bee14309fd148:`. The Ensembl logo and navigation menu are visible at the top. The main content area is divided into a left sidebar and a right main panel.

Left Sidebar:

- Dataset:** Mouse genes (GRCm38.p6)
- Filters:** Gene Name(s) [e.g. mt-Tp]: [ID-list specified]
- Attributes:** Flank (Gene), Upstream flank [500], Gene name
- Dataset:** [None Selected]

Main Panel:

Export all results to: Unique results only

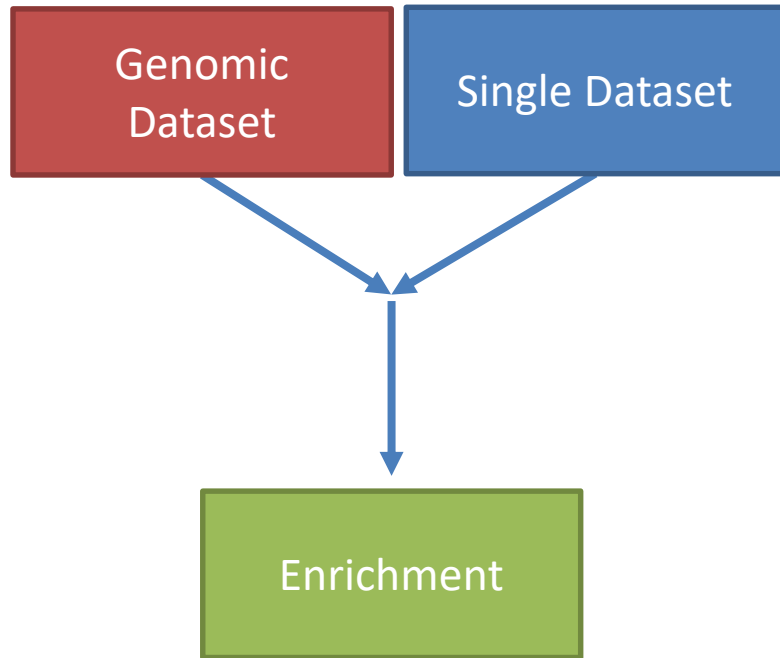
Email notification to:

View: rows as Unique results only

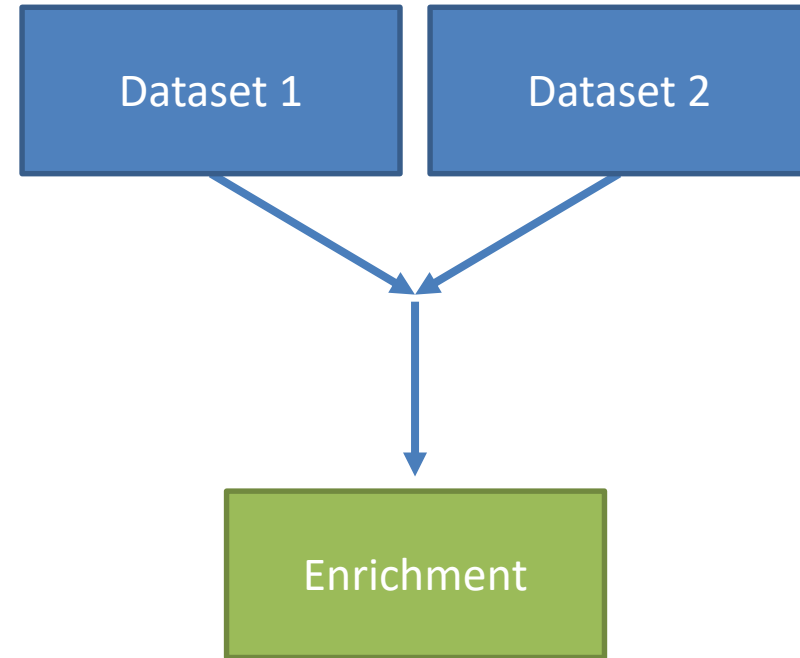
FASTA Output:

```
>Cnn1
TTC AAAAGAAATAAAGCTTTGCTGAAGTTCTCCTTTGTGCCAGCTTTCATACTGGGCACCC
TGGAGGTGACATTTCCCTCCCTGCCCTTTACTGCACCCTTGTGAGGCAAGCACAGTTGTTA
GCCCCTCTAGAGATTTGGCAATAGGGTCCCATAGAGGGGAAGGCTCTGTGTAGAGGGTTG
GTGAATGGAGGTCTGT CAGATCATCTTGCTATGCTAGGGGCTTGGGTGGGGGTGCAGCGG
TTCTGTCTTGGGACTGAAGAAGGAGATCAATGACCTCTGAGATAAGGAGTCTCAGAGACG
GGACAGTTGGCATGGGGAGAAGGGTGGAAAAGGGTGGGCTGTATGAGAACCCTCTCCC
AGAATAAGGCATT CAGCCCCCTAGGTGGAAA CAATGACACAGTCAGCTCCCAATACCAAG
GCTCTGACATCAGGAGGTGGGGGTGGCCAGAGTATGTGTGGGGTGCCACGCTCTTGGCA
GCCCCGTGGCCAATGGGAC
>Vrk3
GGCATGGCTCCGTGTCTCCGTACCTCAGCATGGCTGGCAACCCAAAGAACATATGCTTTC
CAGGATCCCAAGAAAACCCAACTCCTCATGAGGAGGGCTCCCTCAGAGTACAGGGGAAA
TTCTCCGAGGAAGTTCCGGTTCAGAACTTCTCCTCAAATTTATCCACCCAGGTAAC
TCCGCGCCAAATTCAGGACCCACCACCTGTGGTCTCCAACCCACTGAGCCCTCTAGGAAG
TAGAGAGGAAACAGATT CATGGCTAAGTGACACCAAGGTGGATGGCCACTACCCTGTT
ATAGGACTACTTTCACCTTCTACACCCGAAAATTCATTTGGAGGTGACAGCGGGTTGC
CATAGATACAGTTTGAAGGCAAACTGAGATTACAGAGAGAGCACAGAAGCAAGGAGAGT
TTAAAAATGACAAACAGGACCTTTGATTGGTGGCTGT CACGTATTTTCATCAAGATTGACGT
CAGACATGCGCAGTAGAAGG
>Tat
TTTTTCCTTGGAGTGCGGTTGAATTTTTTGTGGAGATTCCCATTTGTCCATAGCAAATCCCA
CAGCCACTAGACCAATAGCTCAAGCAACCTCTATCATTTCCCACTCCAGCAGCTCAGACT
```

Deciding on a comparison



Single Input Set



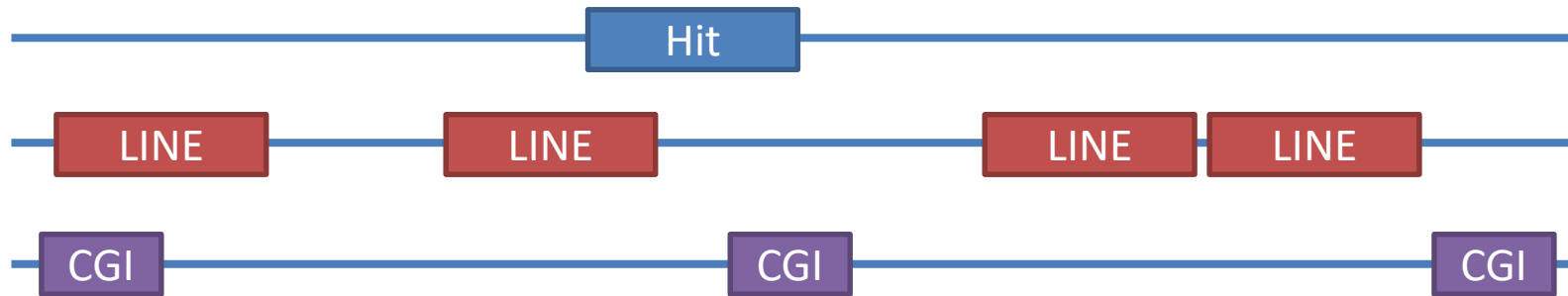
Double Input Set

Filtering list of hits



- High specificity
 - Quick run times
 - Potentially lower power
 - Highest hit artefacts
- More power
 - Long run times
 - More noise
- Don't need all hits to generate motif
 - Often better to have a clean sequence set
 - Remove sequences which look unusual

Artefacts



- Exclude common repeats
 - Simple repeats (poly-A, SerThr repeats etc)
 - Complex repeats (retroviral etc)
 - Exclude hits with repeats
 - Repeatmasked sequence
- Check composition
 - Analyse compositionally biased regions explicitly

Software

The MEME Suite
Motif-based sequence analysis tools

meme-suite.org



xxmotif.genzentrum.lmu.de/



lgsun.grc.nia.nih.gov/CisFinder/



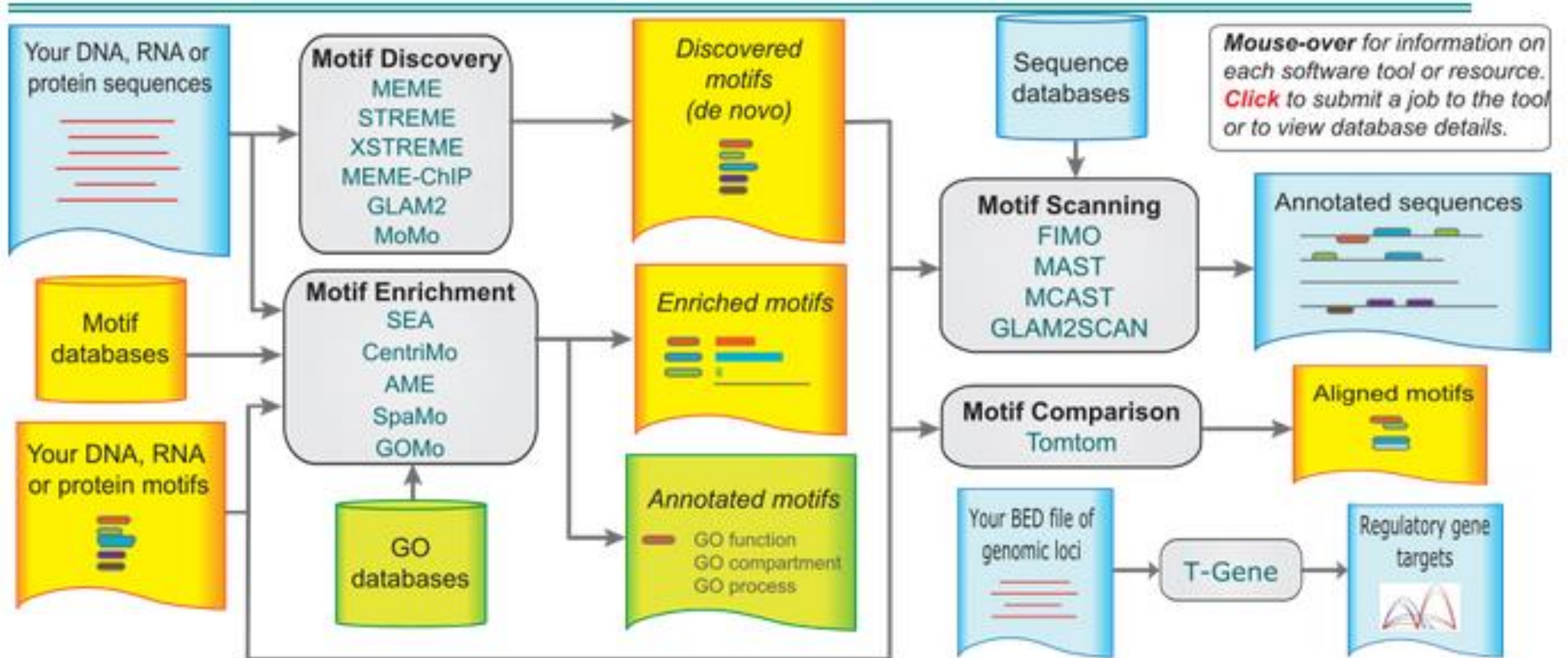
cb.utdallas.edu/cread/



HOMER

homer.salk.edu/homer/motif/

MEME Suite



MEME Motif Discovery

- MEME
 - Original motif enrichment program
 - PWM based motifs
 - Long ungapped motifs, sensitive search, slow!
- STREME/XSTREME
 - Short ungapped discriminatory motifs
 - STREME when you expect the motif to be positioned within your sequence (ie ChIP peaks)
 - XSTREME when you don't expect the motif to be positioned (eg Promoters)
 - Degeneracy based motifs
 - Quick!
- GLAM2
 - Gapped motifs

MEME Suite 4.10.1

MEME Multiple Em for Motif Elicitation
Version 4.10.1

Data Submission Form

Perform motif discovery on DNA, RNA or protein datasets.

Select the motif discovery mode

Normal mode Discriminative mode ?

Input the primary sequences

Enter sequences in which you want to find motifs. ?

Upload sequences No file selected.

Select the site distribution

How do you expect motif sites to be distributed in sequences? ?

Select the number of motifs

How many motifs should MEME find? ?

Input job details

(Optional) Enter your email address. ?

(Optional) Enter a job description. ?

Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.

Version 4.10.1 [Please send comments and questions to: meme-suite@uw.edu](#) [Powered by Opal](#)

[Home](#) [Documentation](#) [Downloads](#) [Authors](#) [Citing](#)

Main Parameters:

- Sequences (multi-fasta)
- Expected sites
- How many motifs to find

Advanced

- Custom background
- Negative set
- Motif size restriction

NB: Query size limited to 60kb

Local installations don't have this limit

Good Result

MEME
Multiple Em for Motif Elicitation

For further information on how to interpret these results or to get a copy of the MEME software please access <http://meme-suite.org>.

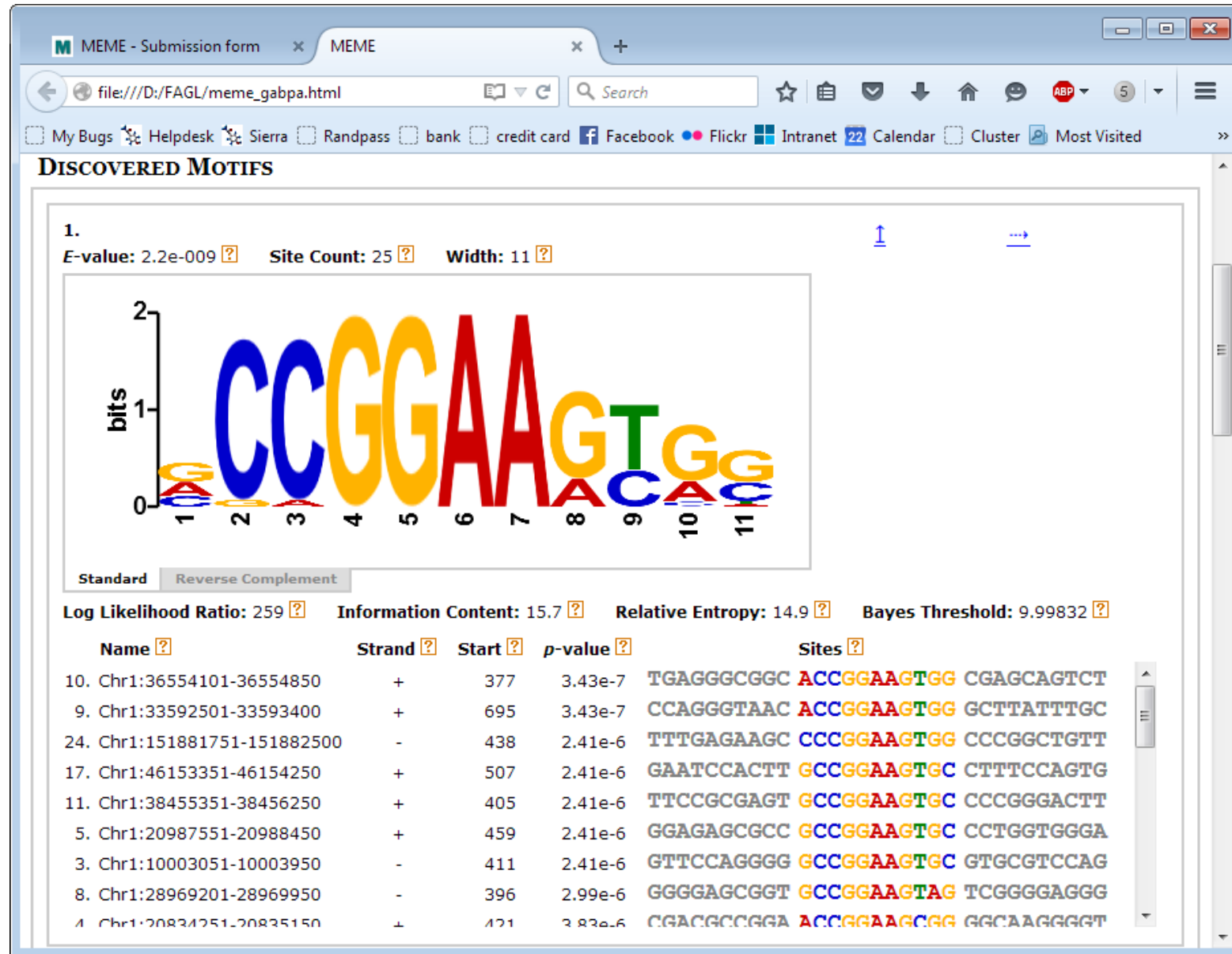
If you use MEME in your research, please cite the following paper:
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994. [\[pdf\]](#)

[DISCOVERED MOTIFS](#) | [MOTIF LOCATIONS](#) | [PROGRAM INFORMATION](#)

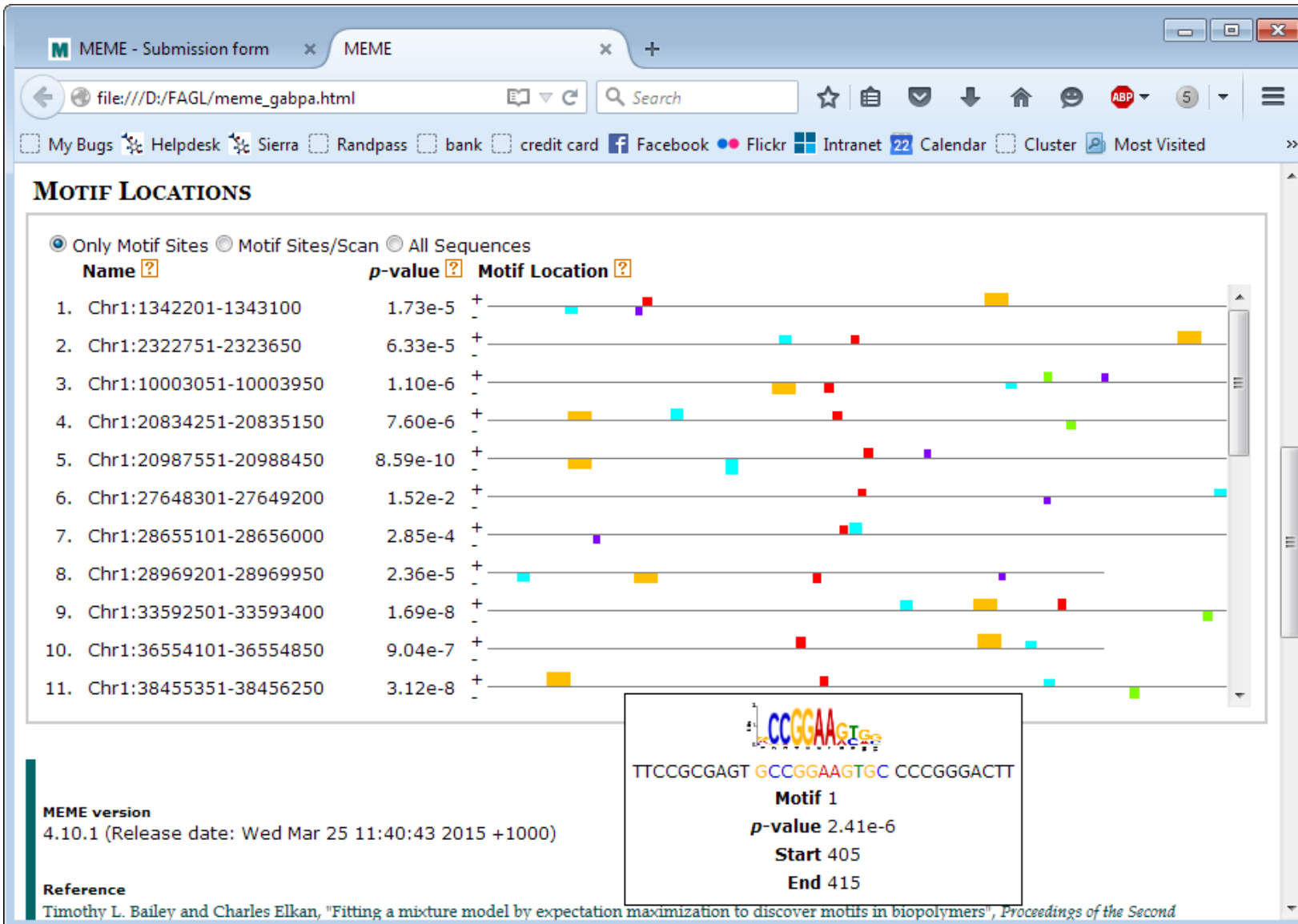
DISCOVERED MOTIFS

	Logo	E-value [?]	Sites [?]	Width [?]	More [?]	Submit/Download [?]
1.		2.2e-009	25	11	I	...
2.		7.5e+001	25	15	I	...
3.		2.3e+003	13	11	I	...
4.		2.4e+004	18	8	I	...
5.		4.0e+005	23	29	I	...

Good Result - Motif

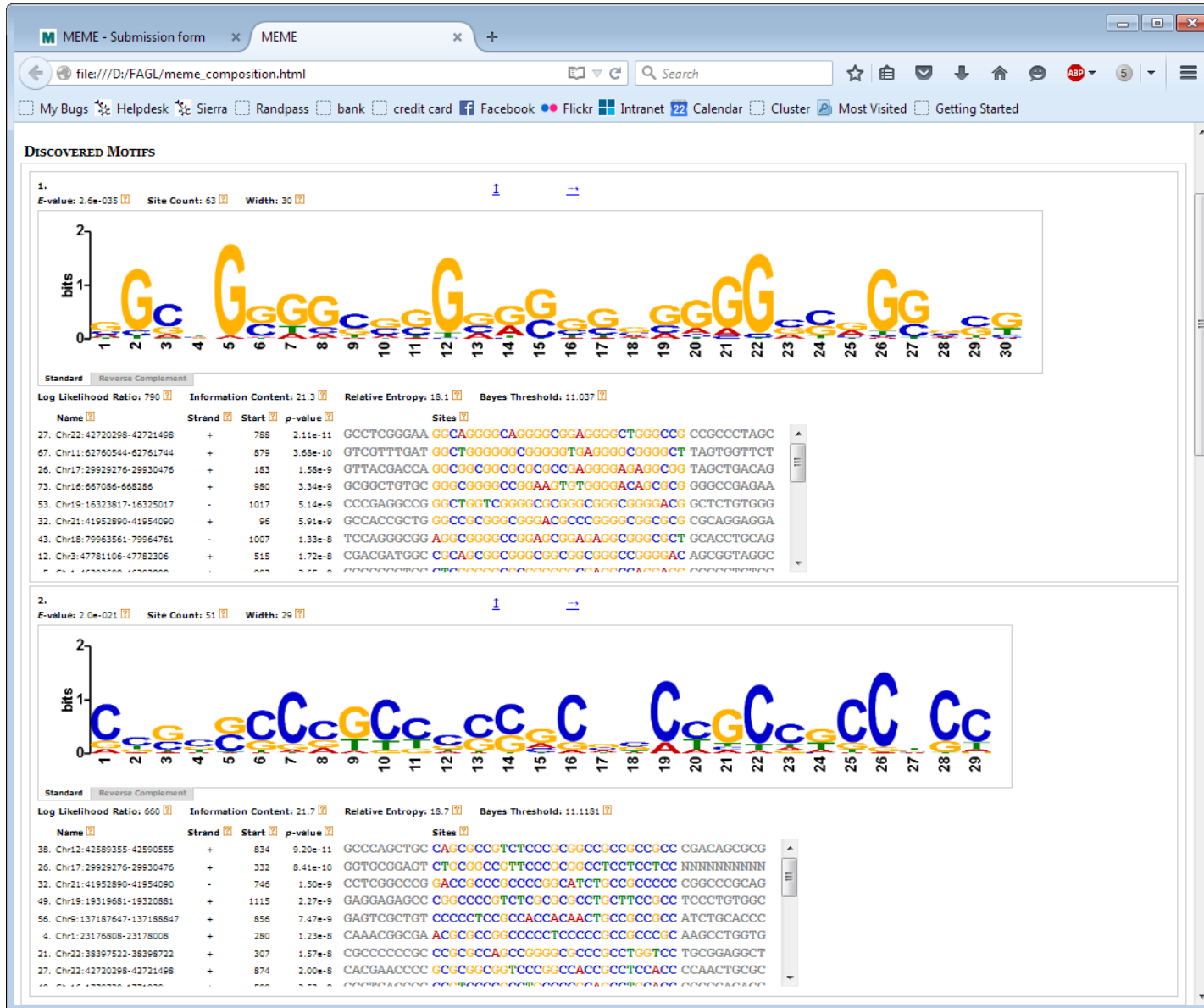


Good Result - Positioning



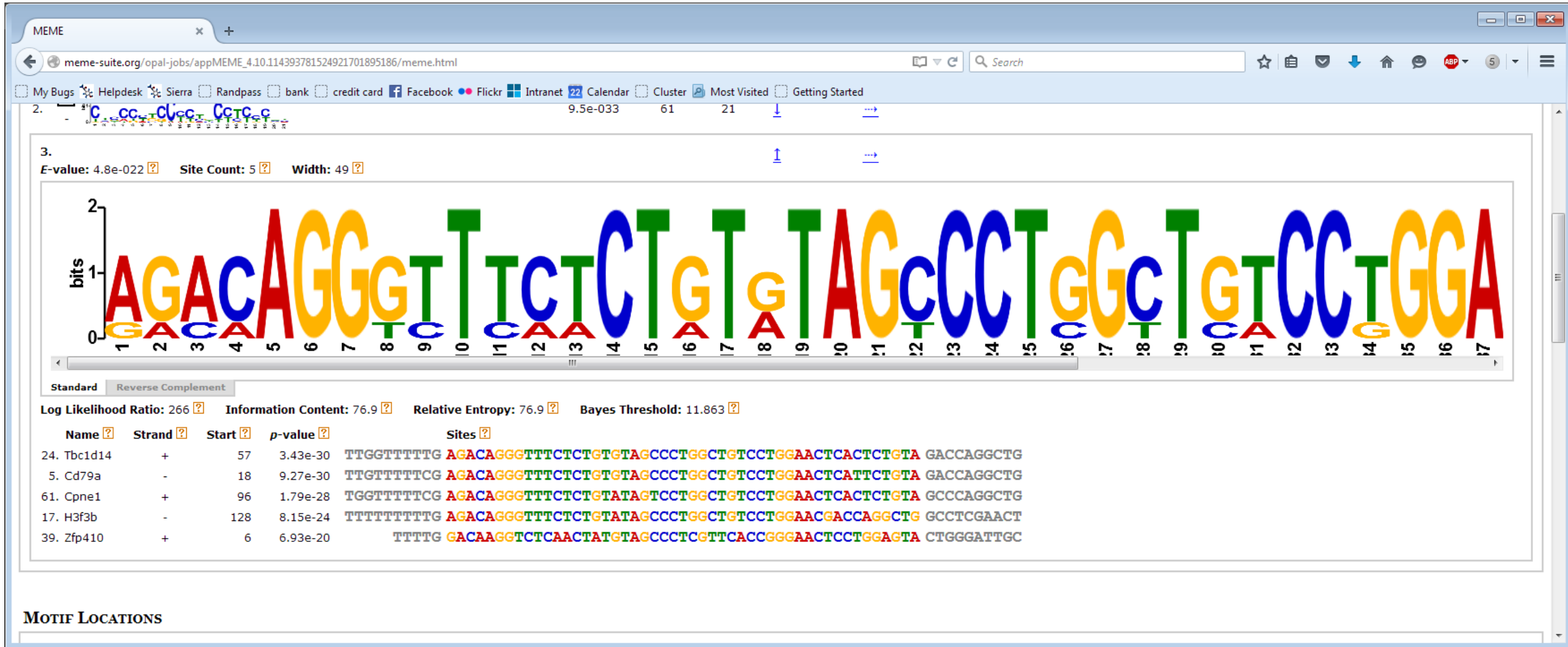
For 'peak' data, expect motifs to be roughly centred. For promoter data there may be no pattern.

Artefactual Result - Composition



MEME tends to favour long compositionally biased motifs
Real motifs can be further down the list

Artefactual Result - Duplication



Multiple transcripts with the same promoter
Overlapping regions


AME – Known motif search

- Quicker / easier than de-novo discovery
- Limited to characterised binding sites
- Can choose from common motif sources
- Good place to start

AME - Submission form

meme-suite.org/tools/ame

My Bugs Helpdesk Sierra Randpass bank credit card Facebook Flickr Intranet 22 Calendar Cluster



AME
Analysis of Motif Enrichment
Version 4.10.1

AME identifies **known** or **user-provided** motifs that are **relatively** enriched in your nucleotide sequences compared with shuffled sequences or your control sequences (sample output from sequences, control and motifs). AME treats motif occurrences the same, regardless of their locations within the sequences. See this [Manual](#) for more information.

MEME Suite 4.10.1

- ▶ Motif Discovery
- ▼ Motif Enrichment
 - CentriMo
 - AME
 - SpaMo
 - GOMo
- ▶ Motif Scanning
- ▶ Motif Comparison
- ▶ Manual
- ▶ Guides & Tutorials
- ▶ Sample Outputs
- ▶ File Format Reference
- ▶ Databases
- ▶ Download & Install
- ▶ Help
- ▶ Alternate Servers
- ▶ Authors & Citing
- ▶ Recent Jobs

Data Submission Form

Perform standard (non-local) motif enrichment analysis.

Select the type of control sequences to use

Shuffled input sequences User-provided control sequences [?](#)

Input the primary sequences

Enter the nucleotide sequences in which you want to find enriched motifs. [?](#)

Upload sequences Browse... No file selected. [?](#)

Input the motifs

Select a [motif database](#) or enter the motifs you wish to test for enrichment. [?](#)

Multi-organism DNA [?](#)

Vertebrates (In vivo and in silico) [?](#)

Input job details

(Optional) Enter your email address. [?](#)

(Optional) Enter a job description. [?](#)

▶ **Advanced options**

Note: if the combined form inputs exceed 80MB the job will be rejected.

Version 4.10.1 [Please send comments and questions to: meme-suite@uw.edu](#) Powered by Opal

[Home](#) [Documentation](#) [Downloads](#) [Authors](#) [Citing](#)

Input the motifs

Select a [motif database](#) or enter the motifs you wish to test for enrichment. [?](#)

Multi-organism DNA

User supplied

Type in motifs

Upload motifs

Databases (select category)

Multi-organism DNA

JASPAR DNA

JASPAR (REDUNDANT) DNA

CIS-BP Single Species DNA

CISBP-RNA Single Species RNA

RNA

miRBase Single Species microRNA

ARABIDOPSIS (Arabidopsis thaliana) DNA

ECOLI (Escherichia coli) DNA

FLY (Drosophila melanogaster) DNA

HUMAN (Homo sapiens) DNA

MALARIA (Plasmodia falciparum) DNA

MOUSE (Mus musculus) DNA

WORM (Caenorhabditis elegans) DNA

JASPAR CORE (2014)

Databases

JASPAR CORE (2014)

JASPAR CORE (2014) vertebrates

JASPAR CORE (2014) fungi

JASPAR CORE (2014) insects

JASPAR CORE (2014) nematodes

JASPAR CORE (2014) plants

JASPAR CORE (2014) urochordates

JASPAR PHYLOFACTS

JASPAR FAM

JASPAR POLII

JASPAR CNE

JASPAR SPLICE

AME Result

No additional detail

Could check for positional
Bias with CentriMo

Beware similar motifs
from different factors

AME results

meme-suite.org/opal-jobs/appAME_4.10.11439385738671-528977

AME
Analysis of Motif Enrichment

For further information on how to interpret these results or to get a copy of the MEME software please access <http://meme-suite.org>.

If you use AME in your research, please cite the following paper:
Robert McLeay and Timothy L. Bailey, "Motif Enrichment Analysis: A unified framework and method evaluation", *BMC Bioinformatics*, 11:165, 2010, doi:10.1186/1471-2105-11-165.
[full text](#)

[ENRICHED MOTIFS](#) | [INPUT FILES](#) | [PROGRAM INFORMATION](#)

ENRICHED MOTIFS

Fixed partition size: number of primary sequences (99)

Sequence motif score: avg_odds
Background model source file: motif input file
Background model frequencies: 0.25,0.25,0.25,0.25
Total pseudocount added to a motif column: 0.25

Statistical test: Wilcoxon rank-sum test
Ranksum method: quick
Threshold p -value for reporting results: 0.05
Number of multiple tests for Bonferroni correction: #Motifs \times #PartitionsTested = 205 \times 1 = 205

Logo	Database	ID	Name	p -value	Adjusted p -value
	JASPAR CORE 2014 vertebrates	MA0592.1	ESRRA	5.49e-10	1.13e-7
	JASPAR CORE 2014 vertebrates	MA0528.1	ZNF263	5.26e-7	1.08e-4
	JASPAR CORE 2014 vertebrates	MA0160.1	NR4A2	2.73e-6	5.59e-4
	JASPAR CORE 2014 vertebrates	MA0149.1	EWSR1-FLI1	3.37e-6	6.90e-4
	JASPAR CORE 2014 vertebrates	MA0141.2	Esrrb	4.99e-6	1.02e-3
	JASPAR CORE 2014 vertebrates	MA0512.1	Rxra	9.82e-6	2.01e-3



STREME discovers **ungapped** motifs (recurring, fixed-length patterns) that are **enriched** in your sequences or **relatively enriched** in them compared to your control sequences (sample output from sequences). See this [Manual](#) or this [Tutorial](#) for more information.

MEME Suite 5.4.1

▼ Motif Discovery

- MEME
- STREME
- XSTREME
- MEME-ChIP
- GLAM2
- MoMo
- DREME (deprecated)

► Motif Enrichment

► Motif Scanning

► Motif Comparison

► Gene Regulation

► Manual

► Guides & Tutorials

► Sample Outputs

► File Format Reference

► Databases

► Download & Install

► Help

► Alternate Servers

► Authors & Citing

► Recent Jobs

← Previous version 5.3.3

Data Submission Form

Perform discriminative motif discovery in sequence datasets (including in very **large** datasets). The sequences may be in the DNA, RNA or protein alphabet, or in a custom alphabet.

Select the type of control sequences to use

- Shuffled input sequences
- User-provided sequences ?

Select the sequence alphabet

Use sequences with a standard alphabet or specify a custom alphabet. ?

- DNA, RNA or Protein
- Custom

Input the sequences

Enter the sequences in which you want to find motifs. ?

?

Input the control sequences

STREME will find motifs that are enriched relative to these sequences. ?

?

Convert DNA sequences to RNA?

- Convert DNA to RNA ?

Input job details

(Optional) Enter your email address. ?

(Optional) Enter a job description. ?

► Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.



Sensitive, Thorough, Rapid, Enriched Motif Elicitation

For further information on how to interpret these results please access <https://meme-suite.org/meme/doc/streme.html>.

To get a copy of the MEME software please access <https://meme-suite.org>.



Details

Train Positives	Train Negatives	Score	Test Positives
69 / 81 (85.2%)	2 / 83 (2.4%)	7.2e-031	6 / 8 (75.0%)

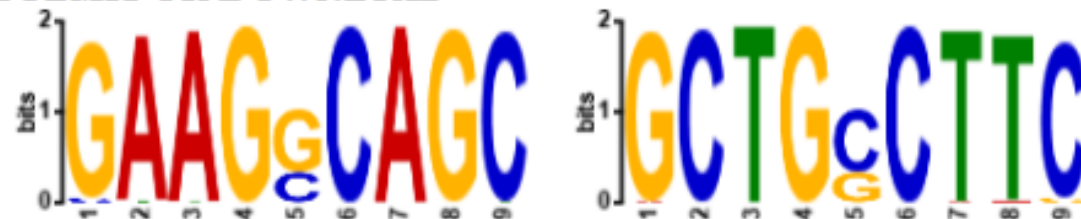


3-CCCTGGGGCGS



1.0e+000 4.0e+000

SUBMIT OR DOWNLOAD



Submit Motif Download Motif Download Logo

Submit to program

- Tomtom Find similar motifs in published libraries or a library you supply.
- FIMO Find motif occurrences in sequence data.
- MAST Rank sequences by affinity to groups of motifs.
- GOMo Identify possible roles (Gene Ontology terms) for motifs.
- SpaMo Find other motifs that are enriched at specific close spacings which might imply the existence of a complex.

Submit

Cancel



For further information on how to interpret these results or to get a copy of the MEME software please access <http://meme-suite.org>.

If you use TOMTOM in your research, please cite the following paper:
 Shobhit Gupta, JA Stamatoyannopolous, Timothy Bailey and William Stafford Noble, "Quantifying similarity between motifs", *Genome Biology*, 8(2):R24, 2007. [\[full text\]](#)

[QUERY MOTIFS](#) | [TARGET DATABASES](#) | [MATCHES](#) | [PROGRAM INFORMATION](#)

QUERY MOTIFS

[Next Top](#)

Name ?	Alt. Name ?	Preview ?	Matches ?	List ?
GCCTCTAA	DREME		3	MA0503.1 (Nkx2-5) , MA0122.1 (Nkx3-2) , MA0504.1 (NR2C2)

TARGET DATABASES

[Previous](#) [Next](#) [Top](#)

Database ?	Number of Motifs ?	Motifs Matched ?
JASPAR_CORE_2014 Vertebrates.meme	205	3

MATCHES TO QUERY MOTIF GCCTCTAA (DREME)

[Previous](#) [Next](#) [Top](#)

Summary ?	Alignment ?
<p>Name MA0503.1</p> <p>Alt. Name Nkx2-5</p> <p>Database JASPAR_CORE_2014_Vertebrates.meme</p> <p>p-value 0.0152266</p> <p>E-value 3.12146</p> <p>q-value 1</p> <p>Overlap 8</p> <p>Offset 1</p> <p>Orientation Normal</p>	

Motif Searching Exercise