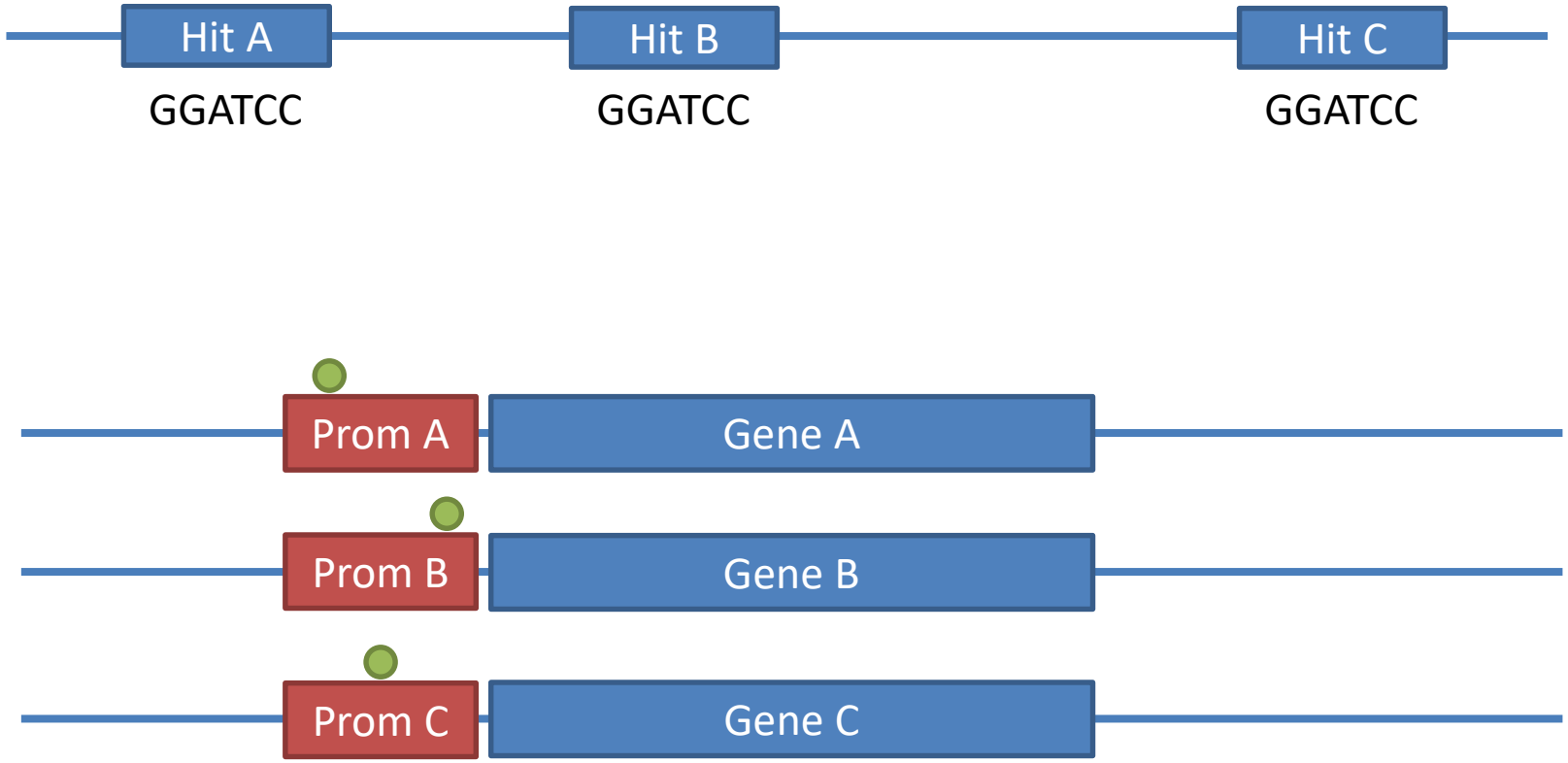
Motif Searching

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

V2021-11

Babraham Bioinformatics

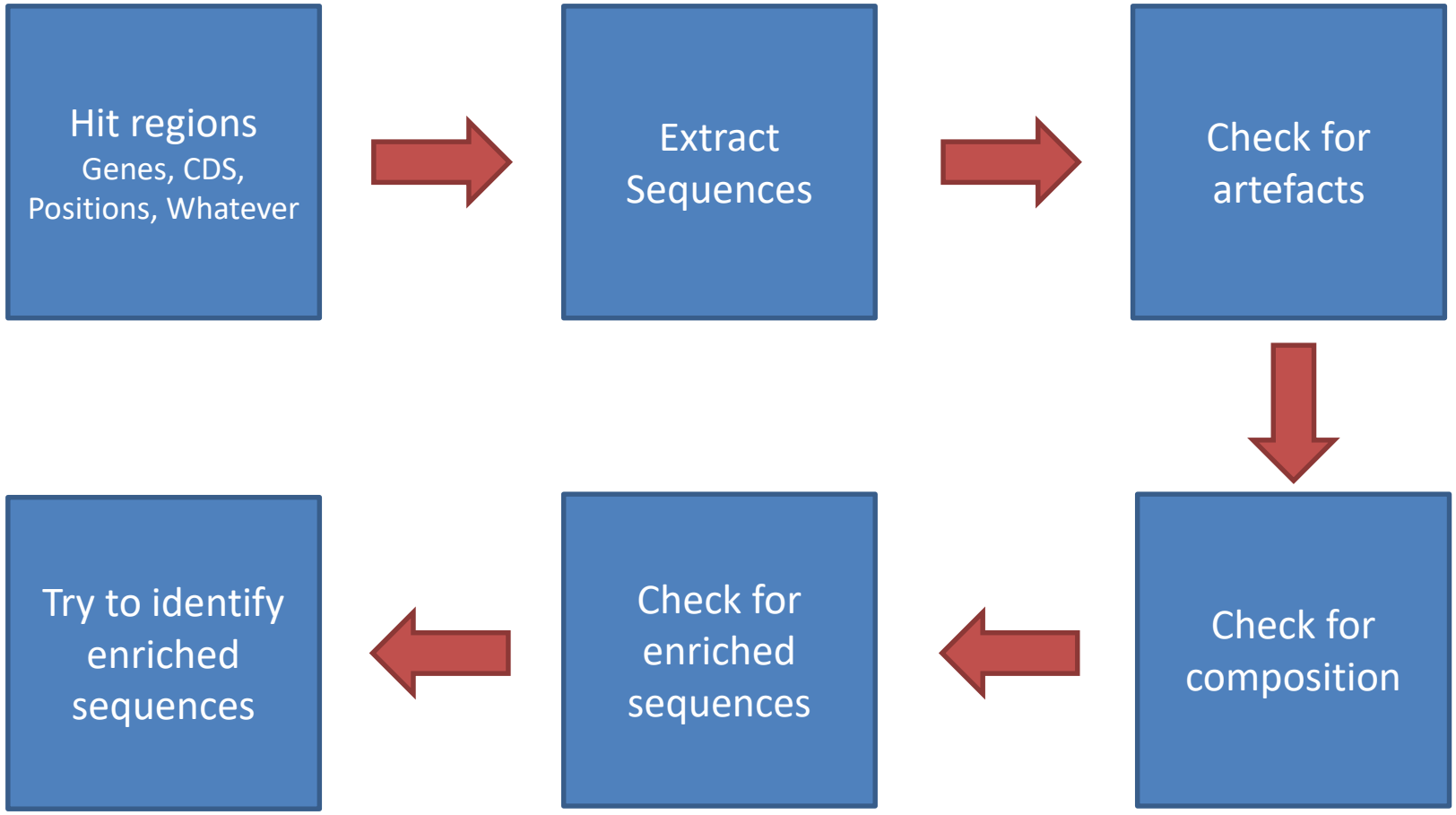
Rationale


Hit A
GGATCC
Hit B
GGATCC
Hit C
GGATCC
Prom A
Gene A
Prom B
Gene B
Prom C
Gene C

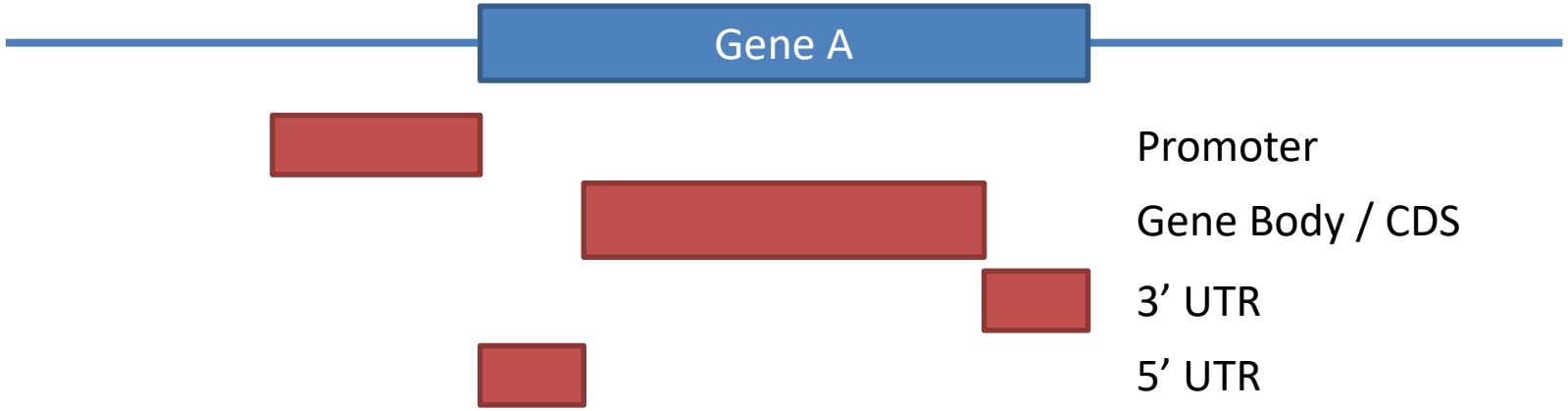
Basic Questions

- 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



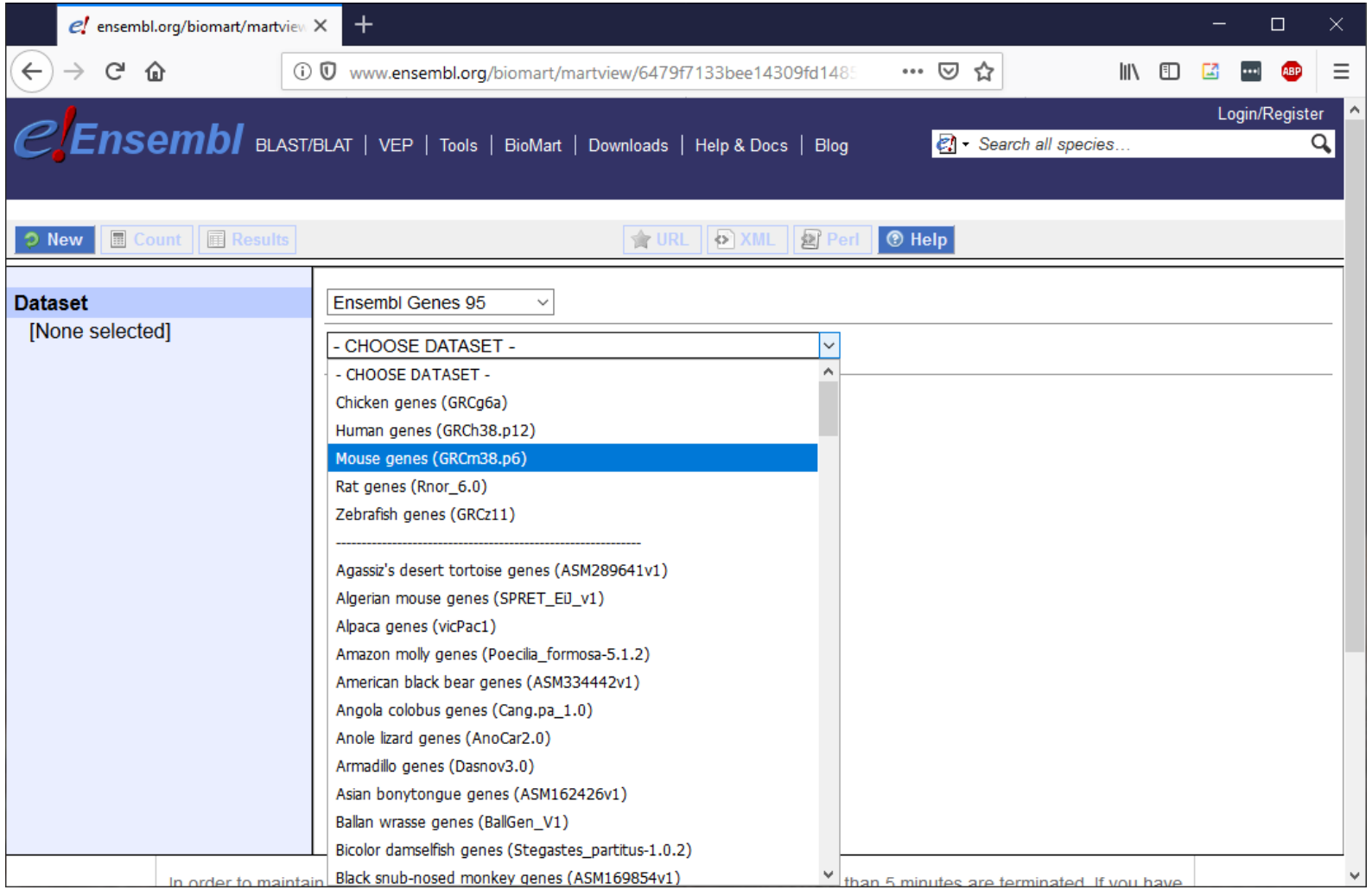
Extracting Sequence

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

- From features
  - Genome Browsers*
  - BioMart

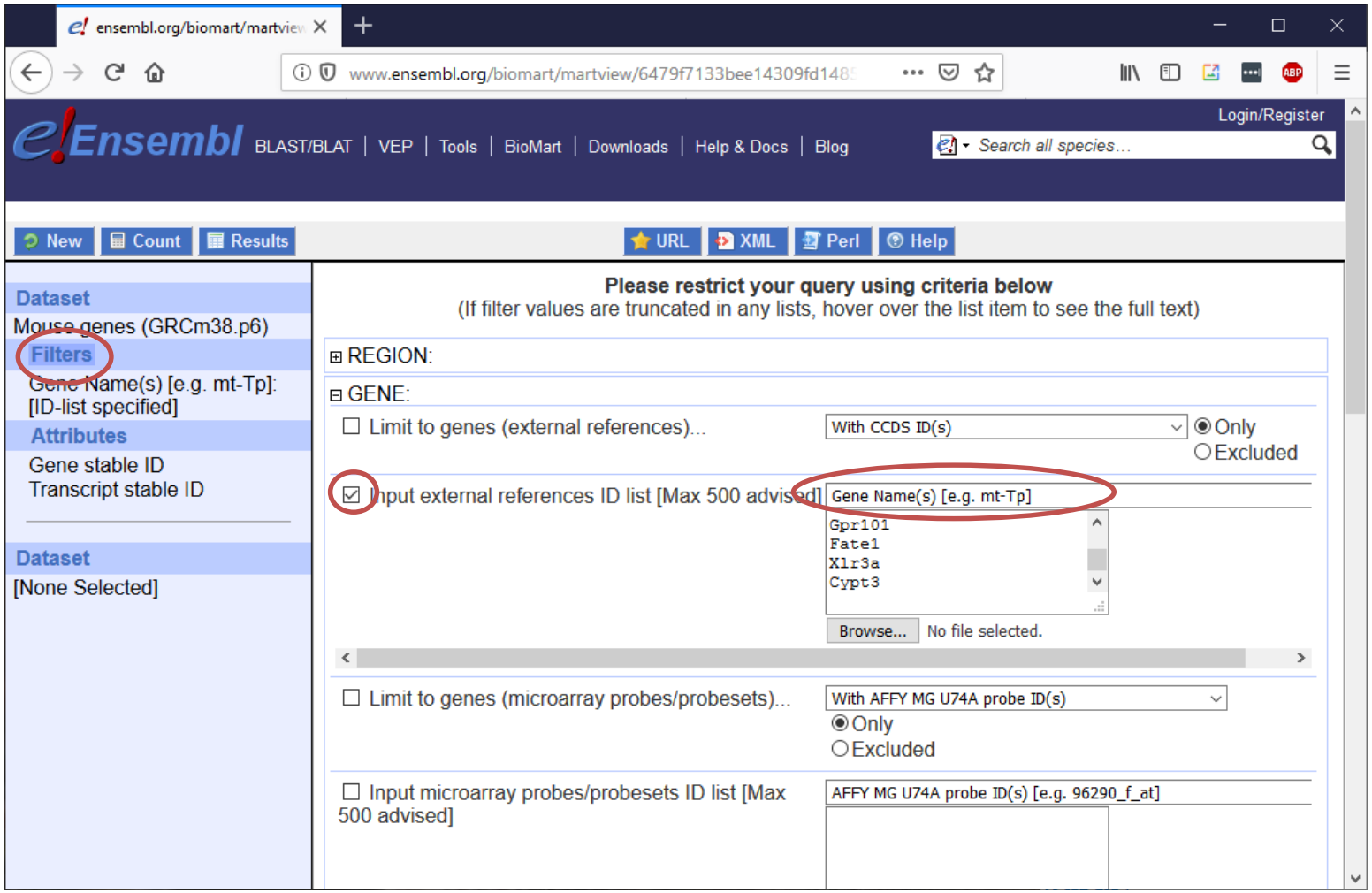
*not easily automatable for multiple sequences

BioMart – Selecting Assembly

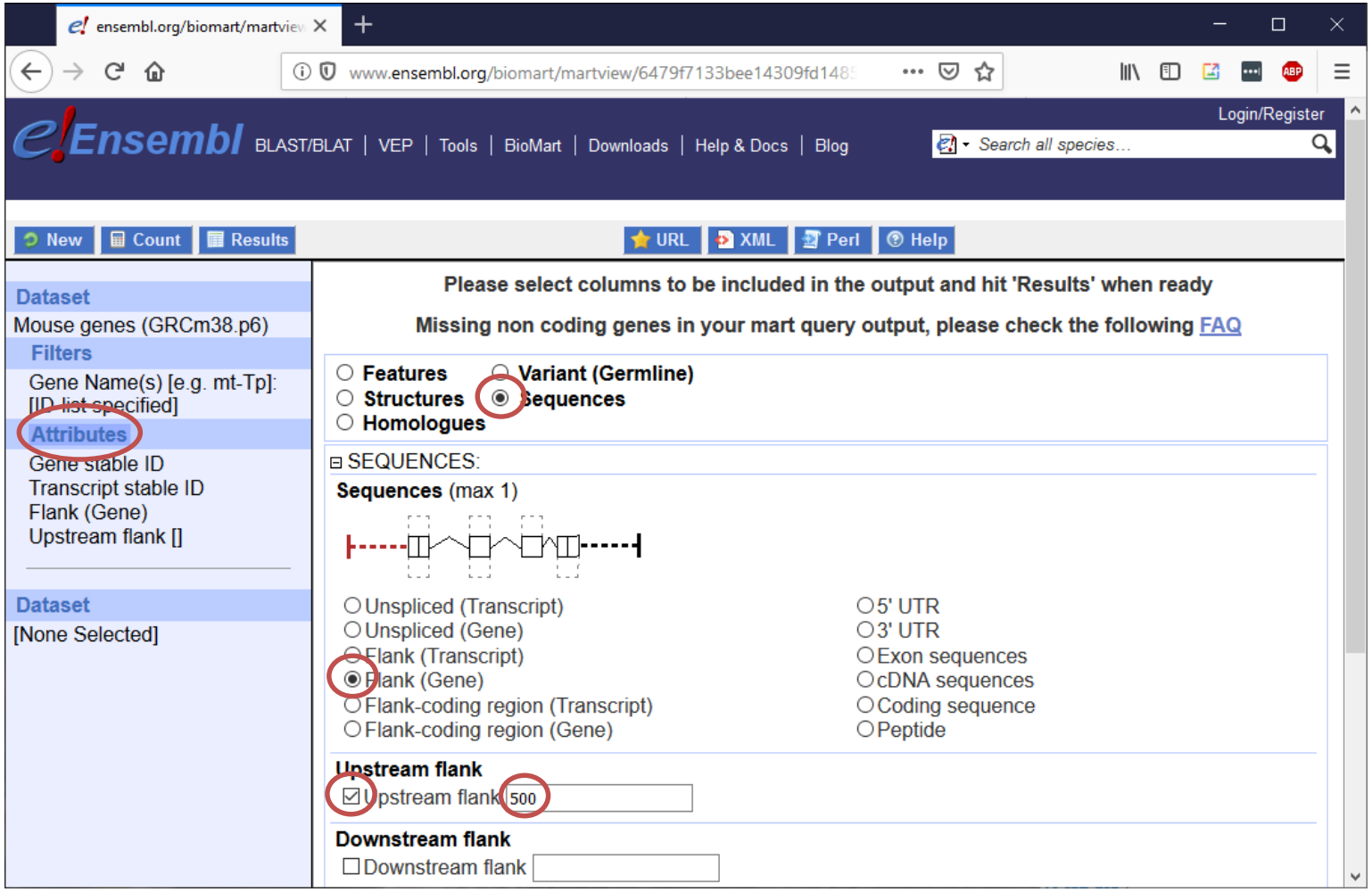


https://ensembl.org/biomart/martview

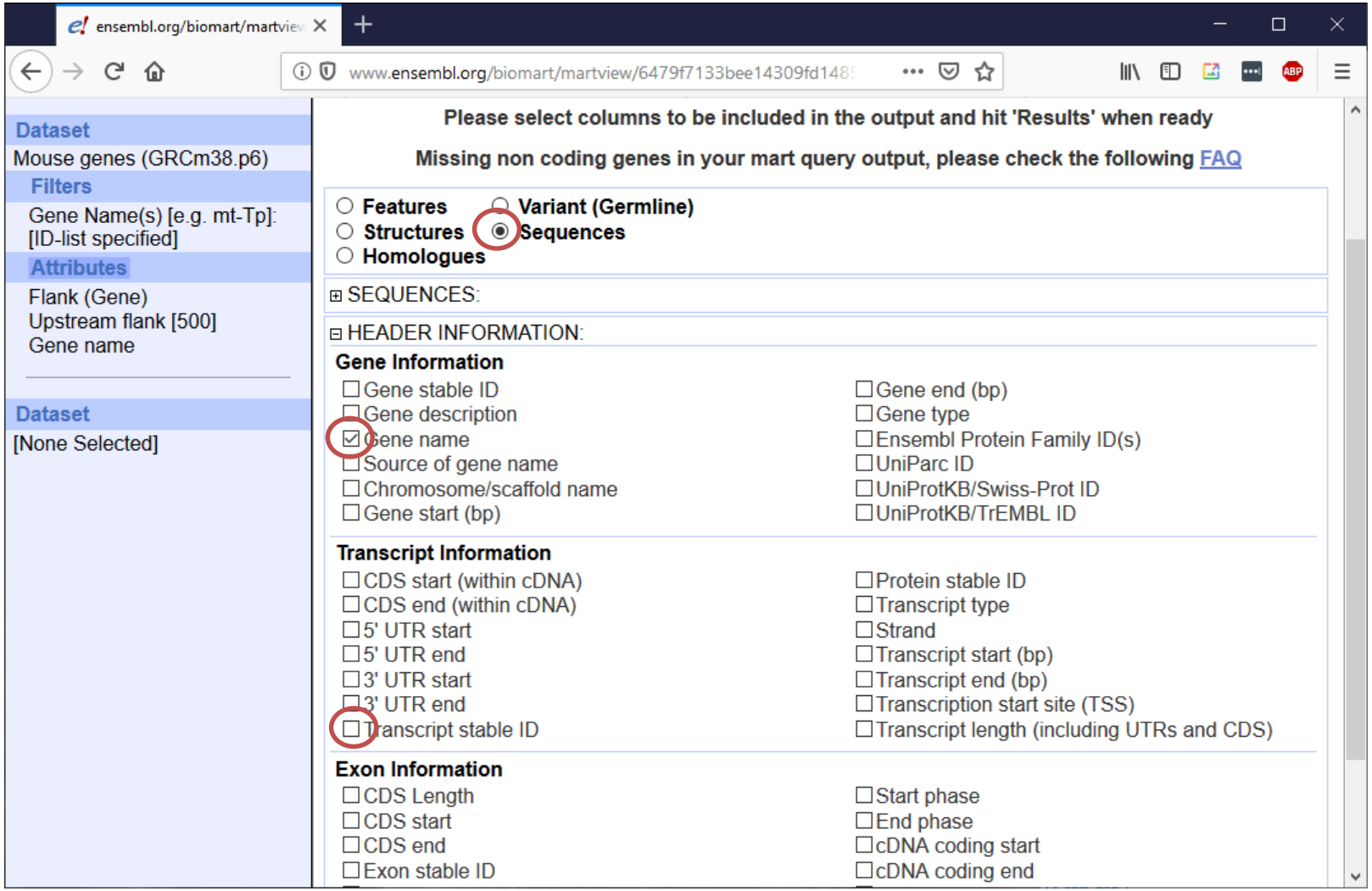
BioMart – Specifying features



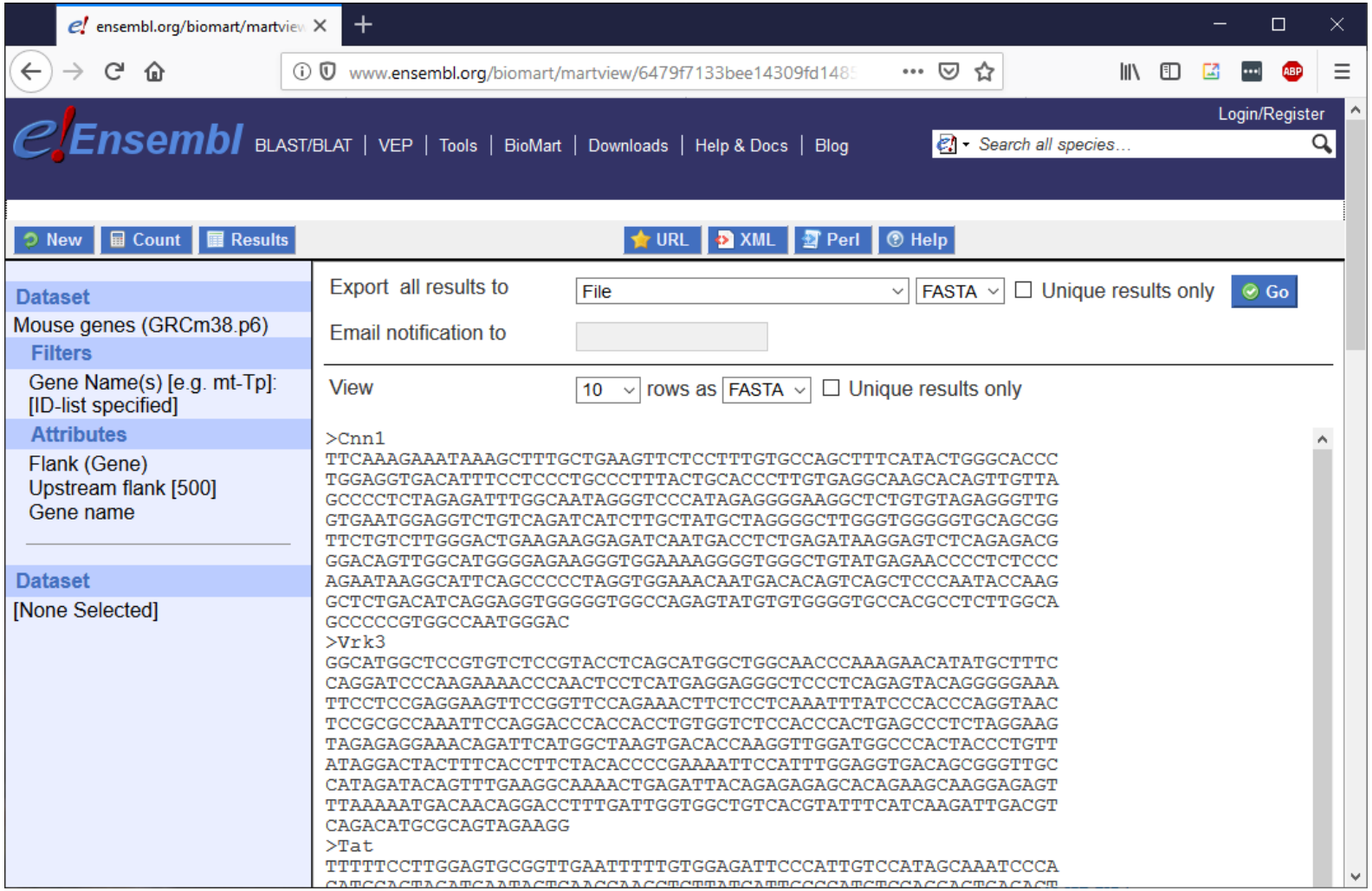
BioMart – selecting seq region



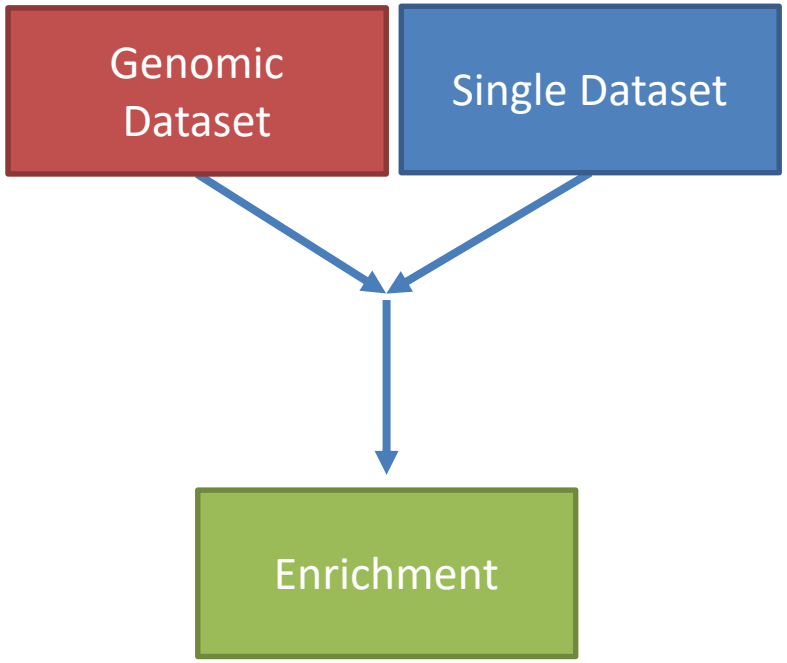
BioMart – header info



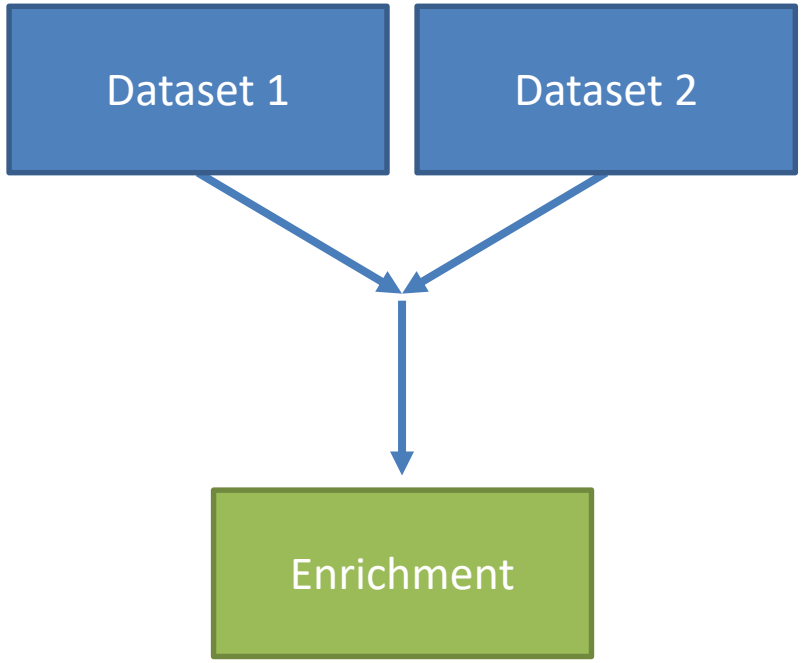
BioMart - exporting



Deciding on a comparison



Single Input Set



Double Input Set

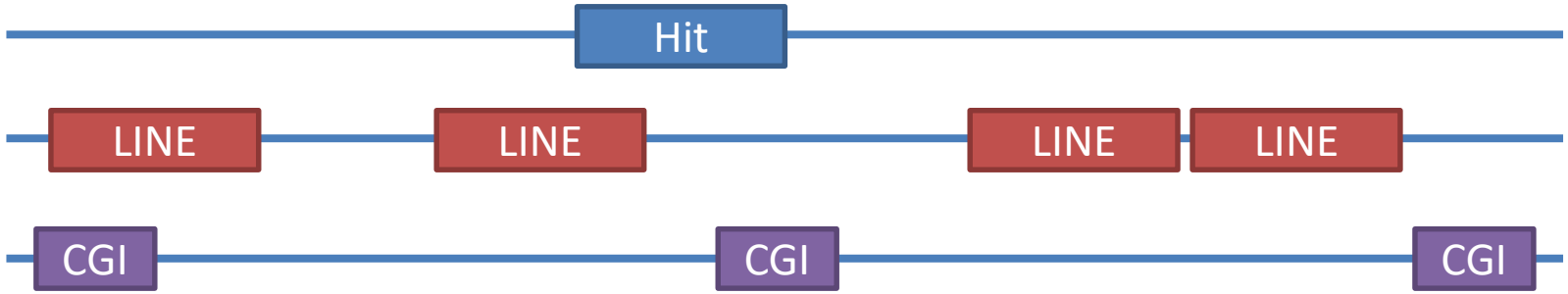
Filtering list of hits

- 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

xxmotif.genzentrum.lmu.de/

CisFinder

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

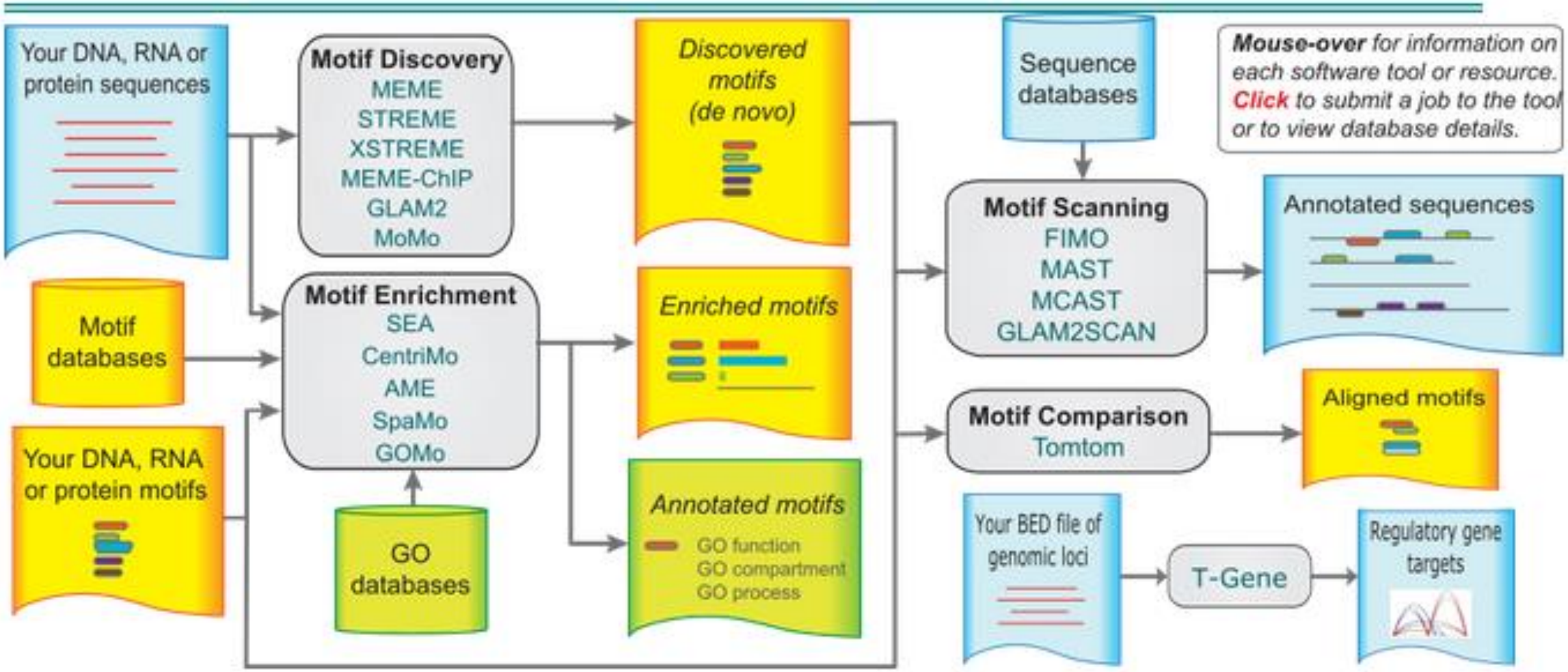
CREAD

cb.utdallas.edu/cread/

HOMER

homer.salk.edu/homer/motif/

MEME Suite

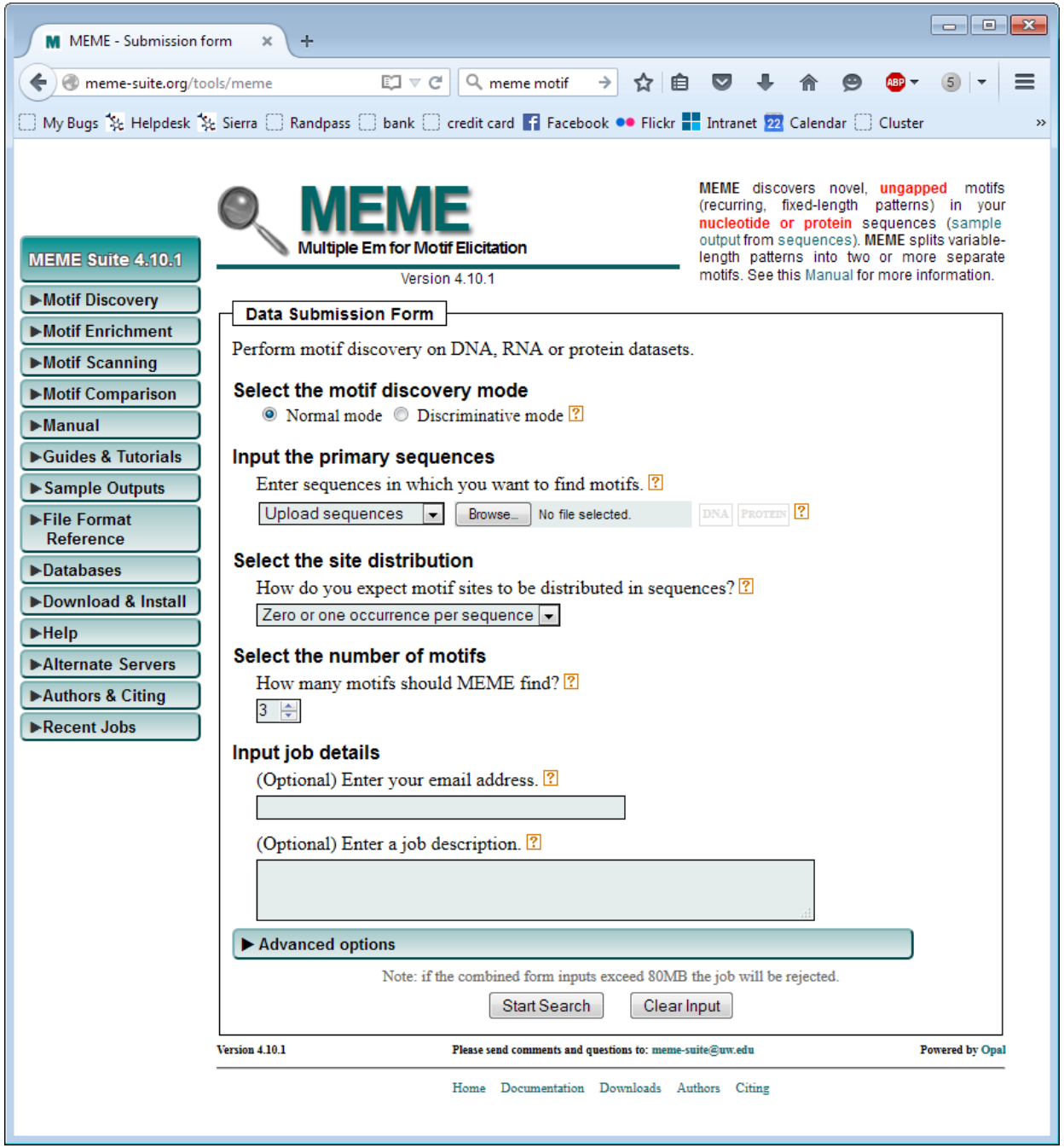

Your DNA, RNA or protein sequences
Motif Discovery
MEME
STREME
XSTREME
MEME-ChIP
GLAM2
MoMo
Discovered motifs (de novo)
Sequence databases
Mouse-over for information on each software tool or resource. Click to submit a job to the tool or to view database details.
Motif Scanning
FIMO
MAST
MCAST
GLAM2SCAN
Annotated sequences
Motif databases
Motif Enrichment
SEA
CentriMo
AME
SpaMo
GOMo
Enriched motifs
Motif Comparison
Tomtom
Aligned motifs
Your DNA, RNA or protein motifs
GO databases
Annotated motifs
GO function
GO compartment
GO process
Your BED file of genomic loci
T-Gene
Regulatory gene targets

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



Main Parameters:

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

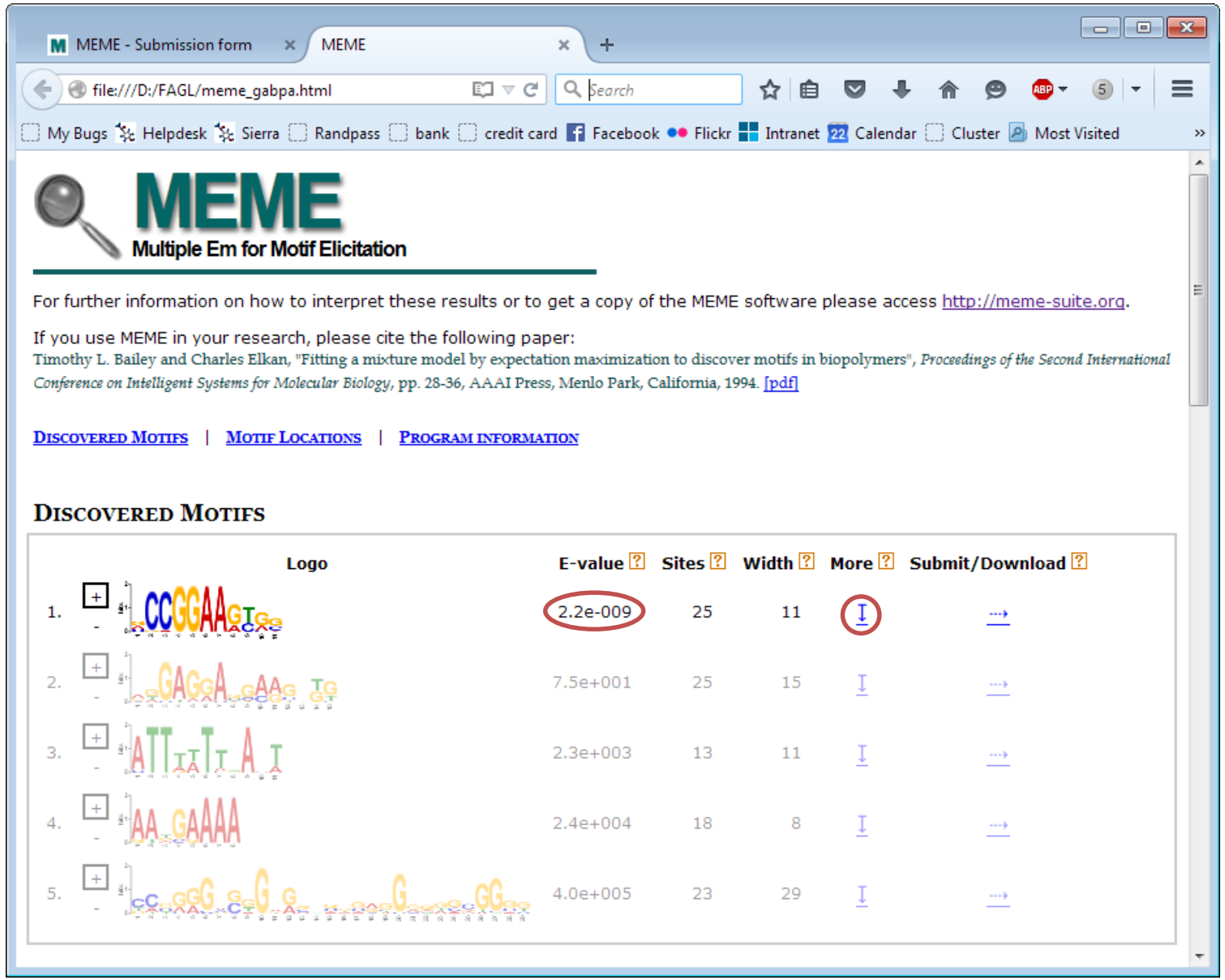
Advanced

- Custom background
- Negative set
- Motif size restriction

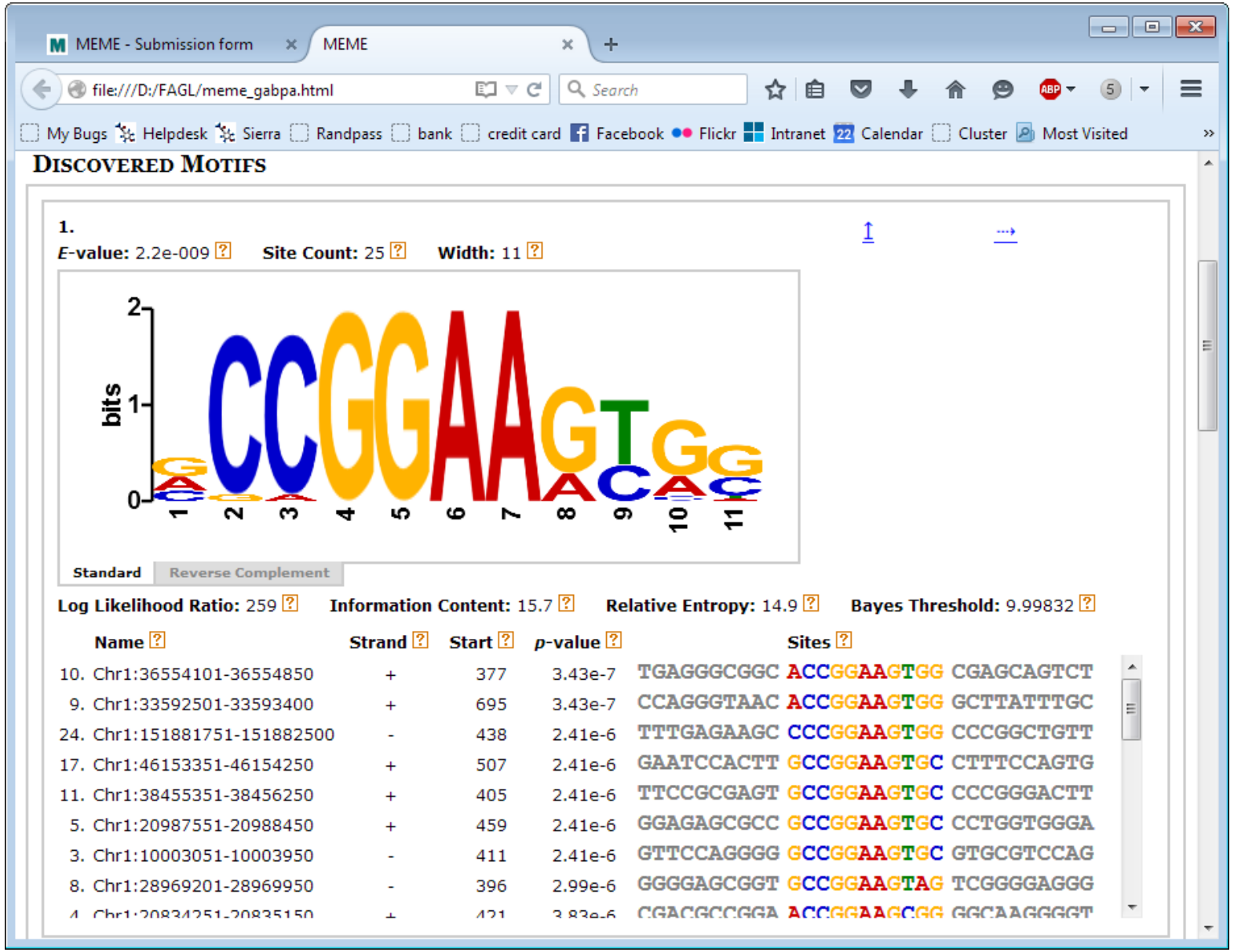
NB: Query size limited to 60kb

Local installations don’t have this limit

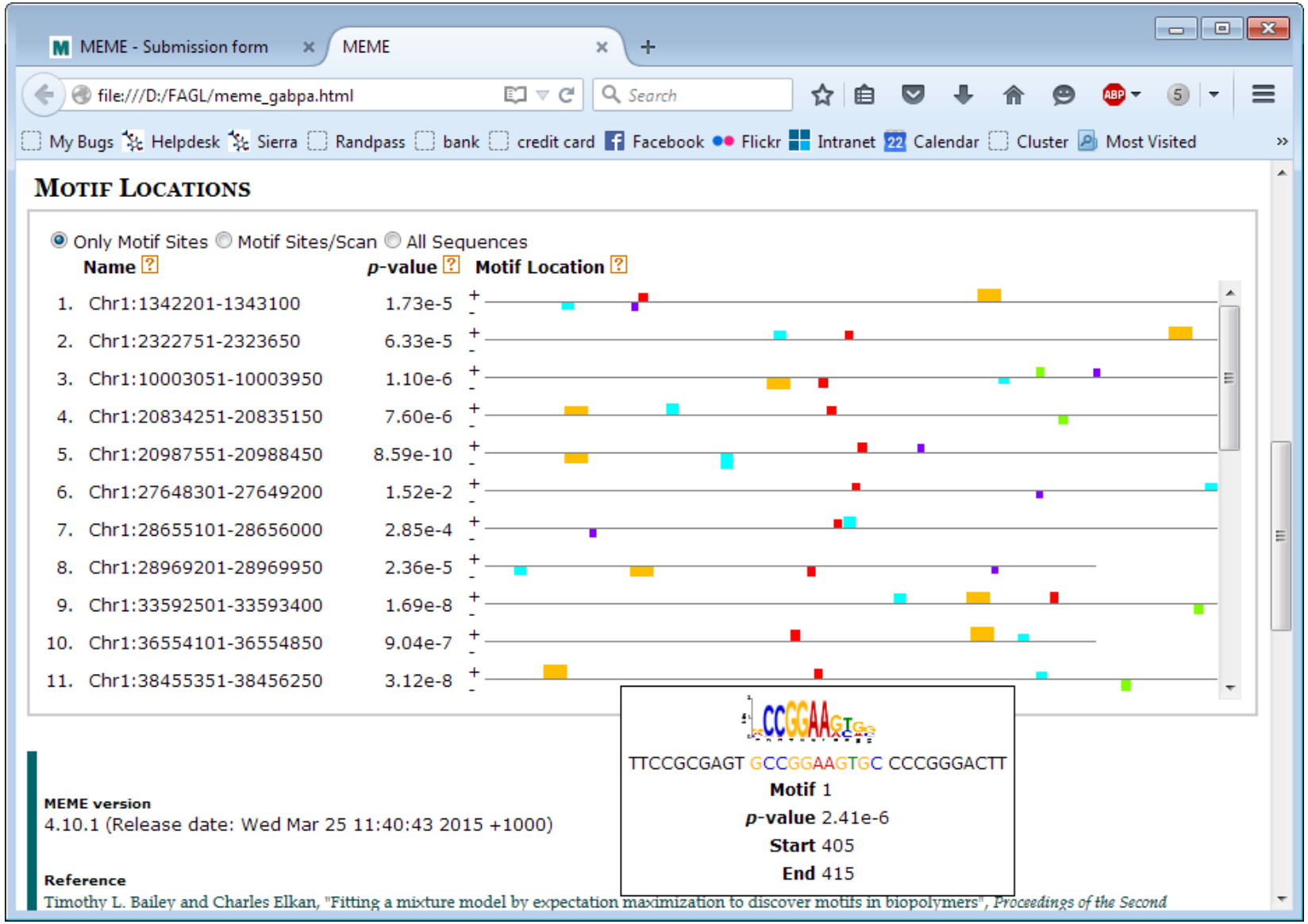
Good Result


MEME - Submission form
MEME
file:///D:/FAGL/meme_gabpa.html
My Bugs Helpdesk Sierra Randpass bank credit card Facebook Flickr Intranet Calendar Cluster Most Visited
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
2. 7.5e+001 25 15
3. 2.3e+003 13 11
4. 2.4e+004 18 8
5. 4.0e+005 23 29

Good Result - Motif

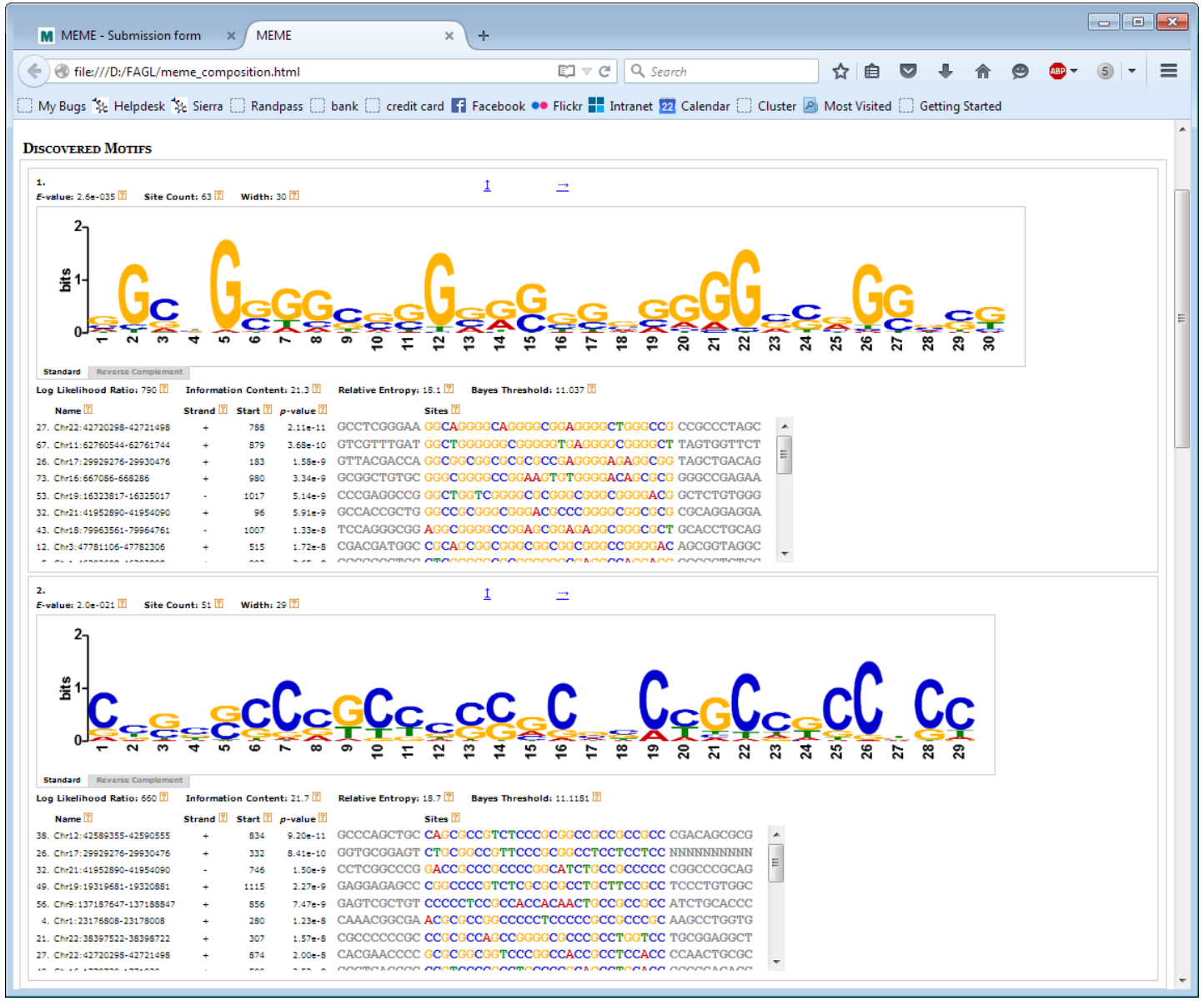


Good Result - Positioning



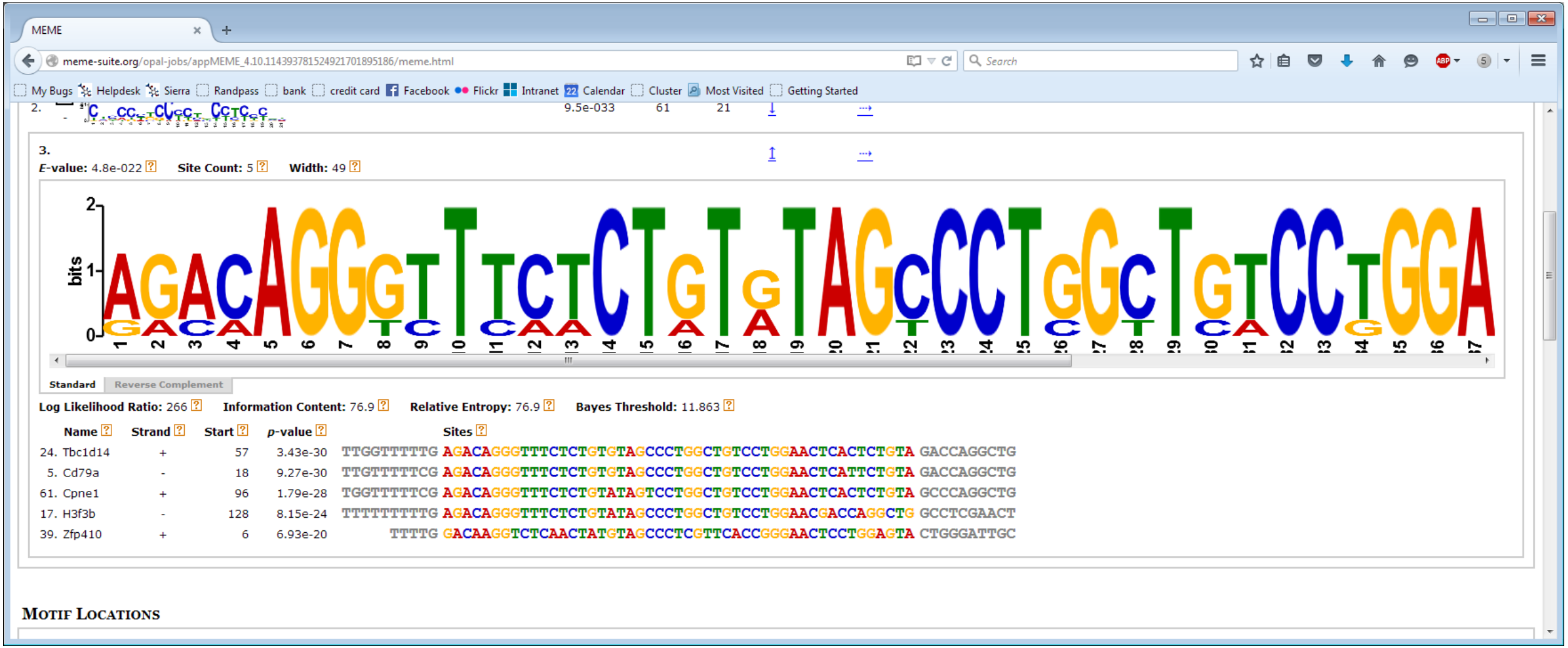
For ‘peak’ data, expect motifs to be roughly centred. For promoter data there may be no pattern.

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 | 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. | DNA | PROTEIN

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.

Start Search | Clear Input

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

My Bugs | Helpdesk | Sierra | Randpass | bank | credit card | Facebook | Flickr | Intranet | Calendar | Cluster | Most Visited | Getting Started

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 × #PartitionsTested = 205 × 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 - Submission form

https://meme-suite.org/meme/tools/stre 90%

STREME

Sensitive, Thorough, Rapid, Enriched Motif Elicitation

Version 5.4.1

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 Browse... No file selected.

Input the sequences

Enter the sequences in which you want to find motifs.

Upload sequences Browse... positive_cgi_set.txt DNA

Input the control sequences

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

Upload sequences Browse... negative_cgi_set.txt DNA

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.

Start Search Clear Input

Version 5.4.1 Please send comments and questions to: meme-suite@uw.edu Powered by Opal

Home Documentation Downloads Authors Citing

STREME Results

https://meme-suite.org/meme//opal-jobs/appSTREME_5.4.1163733537444312168004 90%

STREME

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 Positi
69 / 81 (85.2%)	2 / 83 (2.4%)	7.2e-031	6 / 8 (75.0

3-CCCTGGGGCGS			1.0e+000	4.0e+0

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

DREME - Submission form | DREME | TOMTOM

meme-suite.org//tools/tomtom

My Bugs | Helpdesk | Sierra | Randpass | bank | credit card | Facebook | Flickr | Intranet | Calendar | Cluster | Most Visited | Getting Started

Tomtom

Motif Comparison Tool

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

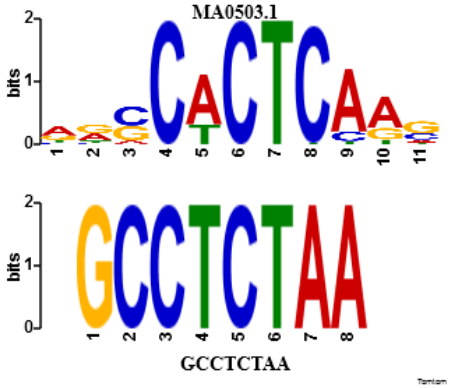
MATCHES TO QUERY MOTIF GCCTCTAA (DREME)

Previous Next Top

Summary

Name	MA0503.1
Alt. Name	Nkx2-5
Database	JASPAR_CORE_2014_vertebrates.meme
***p*-value**	0.0152266
***E*-value**	3.12146
***q*-value**	1
Overlap	8
Offset	1
Orientation	Normal

Alignment



Motif Searching Exercise

Babraham Bioinformatics