Statistical and computational techniques assisting in the identification of potential enhancer elements for the genes expressed in the γ neurons of *Drosophila*

Raymond A. Ford, Ming-Ying Leung, Kyung-An Han

The University of Texas at El Paso

February 28, 2015

Overview

1. Background
2. Methods
3. Results

Drosophila melanogaster

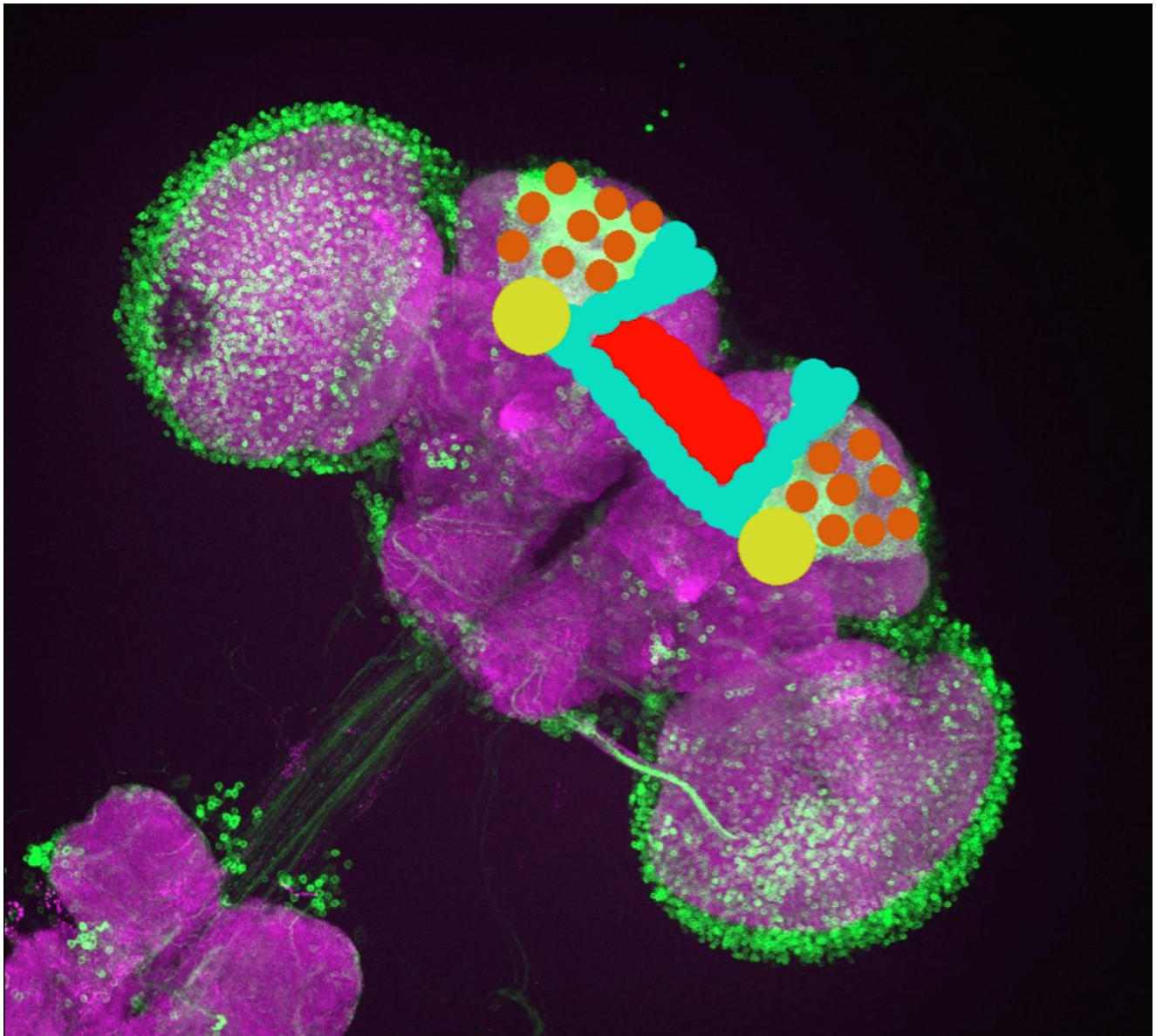
0: Also known as the fruit fly

1: Nervous system comparable to humans

2: Fully sequenced genome

3: Mushroom body contains three distinct structures: α, β, and γ lobes

The fruit fly's brain



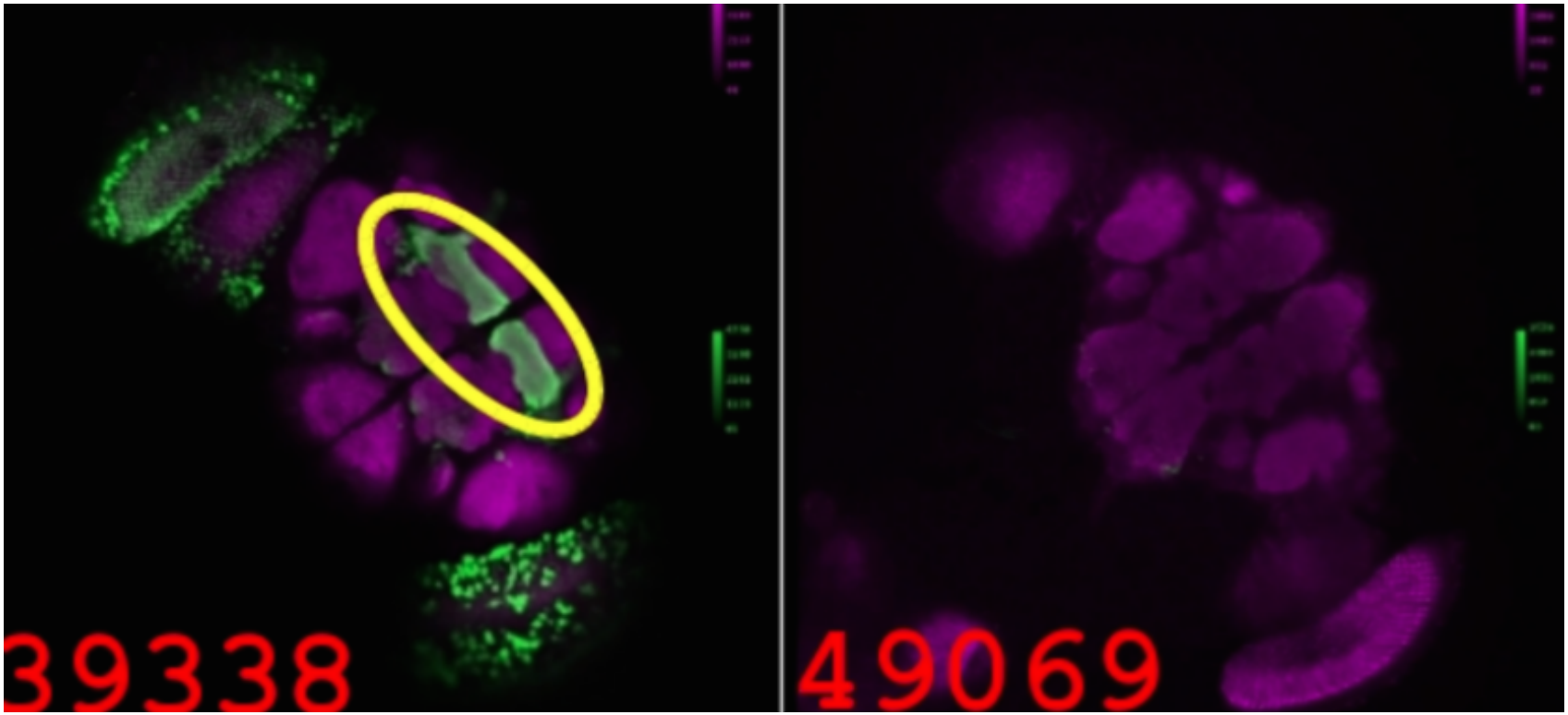
Significance of γ neurons

0: Dopamine and octopamine

1: Learning and memory

2: Neuropsychiatric disorders (ADHD, autism, etc.)

γ^+ versus γ^-



Expression data videos

0: $\sim$ 1240 videos viewed

1: Notes about website layout were made

2: Other information that may be important noted

Automated data collection

0: Scripts were written to automate data collection from $\sim 35,000$ unique websites

1: Data collection required javascript interaction with several sources

2: GUI application created to enable efficient access to data

Automated data collection

Home | Browse | Search | Order | Fees | Accounts | Fly Food | Supplies | Import Permits

Janelia GAL4 Stocks

Bloomington Drosophila Stock Center INDIANA UNIVERSITY

All Janelia Farm GAL4 stocks (a.k.a. "GMR_Brain_exp_1" or "Rubin GAL4" lines) currently available at Bloomington are listed here. These lines express GAL4 under the control of defined sequence fragments from either flanking non-coding or intronic regions of associated genes. In general, these drivers express GAL4 in more restricted patterns than observed with standard enhancer traps.

Expression Data - Data on the expression patterns generated by these GAL4 lines in the adult brain and ventral nerve cord, in third instar imaginal discs, and in the embryonic CNS can be found at Janelia's FlyLight site. Links to the data for individual lines are provided in the "Expression data" column below. The data presented at FlyLight are the work of the Janelia Farm FlyLight Project Team and the laboratories of Gerald M. Rubin, Richard S. Mann and Christopher Q. Doe.

Please see the **Janelia lexA page** for important information on potential expression differences between insertions carrying lexA or GAL4 under the control of the same sequence fragment, but inserted in different attP sites.

Fragment information - The fragment reports linked to below contain a GBrowse view of the fragment location. The FlyBase report on the collection can be found **here**. Files containing construct information, including primers and sequence coordinates for the fragments in the constructs, can be downloaded as an **Excel file** or a **csv file**.

Stock/Insertion information - All the constructs were inserted by site-specific recombination into the attP2 site at 68A4 on 3L. The line used by Janelia Farms for balancing and stocking these strains is available as BDSC Stk#36305 w[1118]; Dr[1] e[1]/TM3, Sb[1]

Download a stocklist - A list of Janelia stocks available from the BDSC with their stock numbers can be downloaded as an **Excel file** or a **csv file**. Stock numbers can be cut and pasted from these files into our ordering page.

- This collection was constructed by the methods described in Pfeiffer et. al., 2008.
Please cite those who generated and analyzed these materials when publishing your own work with these and other Stock Center stocks.

Go back to the **main GAL4** page.

BEWARE: we will be discarding ~1150 Janelia GAL4 lines in 2014. If you wish to use or screen these lines, please order them by **August 31, 2014**. See the **Janelia cull** page for the list of affected lines.

IMPORTANT! Please read cautionary information on the Janelia GAL4 stocks here.

	Stk #	Insertion	Associated gene	Expression data	Fragment report	Verified?
1	46537	P{GMR64D02-GAL4}attP2	14-3-3zeta	expression	FBsf0000166214	
2	46560	P{GMR65F02-GAL4}attP2	14-3-3zeta	expression	FBsf0000166330	
3	49613	P{GMR65G01-GAL4}attP2	14-3-3zeta	expression	FBsf0000166341	verified
4	47708	P{GMR65G03-GAL4}attP2	14-3-3zeta	expression	FBsf0000166343	
5	46566	P{GMR65H08-GAL4}attP2	14-3-3zeta	expression	FBsf0000166360	
6	38826	P{GMR52D03-GAL4}attP2	5-HT1A	expression	FBsf0000165354	
7	38843	P{GMR52G04-GAL4}attP2	5-HT1A	expression	FBsf0000165390	
8	49583	P{GMR53B03-GAL4}attP2	5-HT1A	expression	FBsf0000165423	verified
9	38870	P{GMR53C03-GAL4}attP2	5-HT1A	expression	FBsf0000165434	
10	38873	P{GMR53C10-GAL4}attP2	5-HT1A	expression	FBsf0000165441	
11	50425	P{GMR53D01-GAL4}attP2	5-HT1A	expression	FBsf0000165444	

Automated data collection

Home FlyLight

Expression Patterns of GAL4 Driver Lines

Rubin Lab Truman Lab Doe Lab Mann Lab

HHMI janelia farm research campus

R64D02

Fly Core ID	31081	# residues	3184
Gene	CG17870, 14-3-3zeta, FBgn0004907, FBgn0010635, FBgn0019723, FBgn0023038, FBgn0046306, FBgn0064146	Coordinates	2R: 5988621-5991805
Rearray plate	GR.64	Left primer	caccgcccagagtgccttgtacctatgtg
Robot ID	1125510	Right primer	ccacacagttcgatatgcccctcaa
Transformant type	Homozygous	Strand	+1
Well	D02		

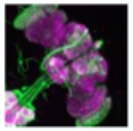
Brain (GAL4)

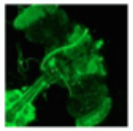
Vector	pBPGUw
Landing site	attP2-3L:68A4
Age	Day 1-5
Gender	Female
Reporter	JFRC2-10XUAS-IVS-mCDGFP

Download LSM file

It is possible to download a limited number of full confocal stacks (~500 MB LSM files) from this web server using the button to the left. Those users requiring large numbers of confocal stacks should send a request to: gal4-gen1@janelia.hhmi.org.

Projection Pattern projection Translation

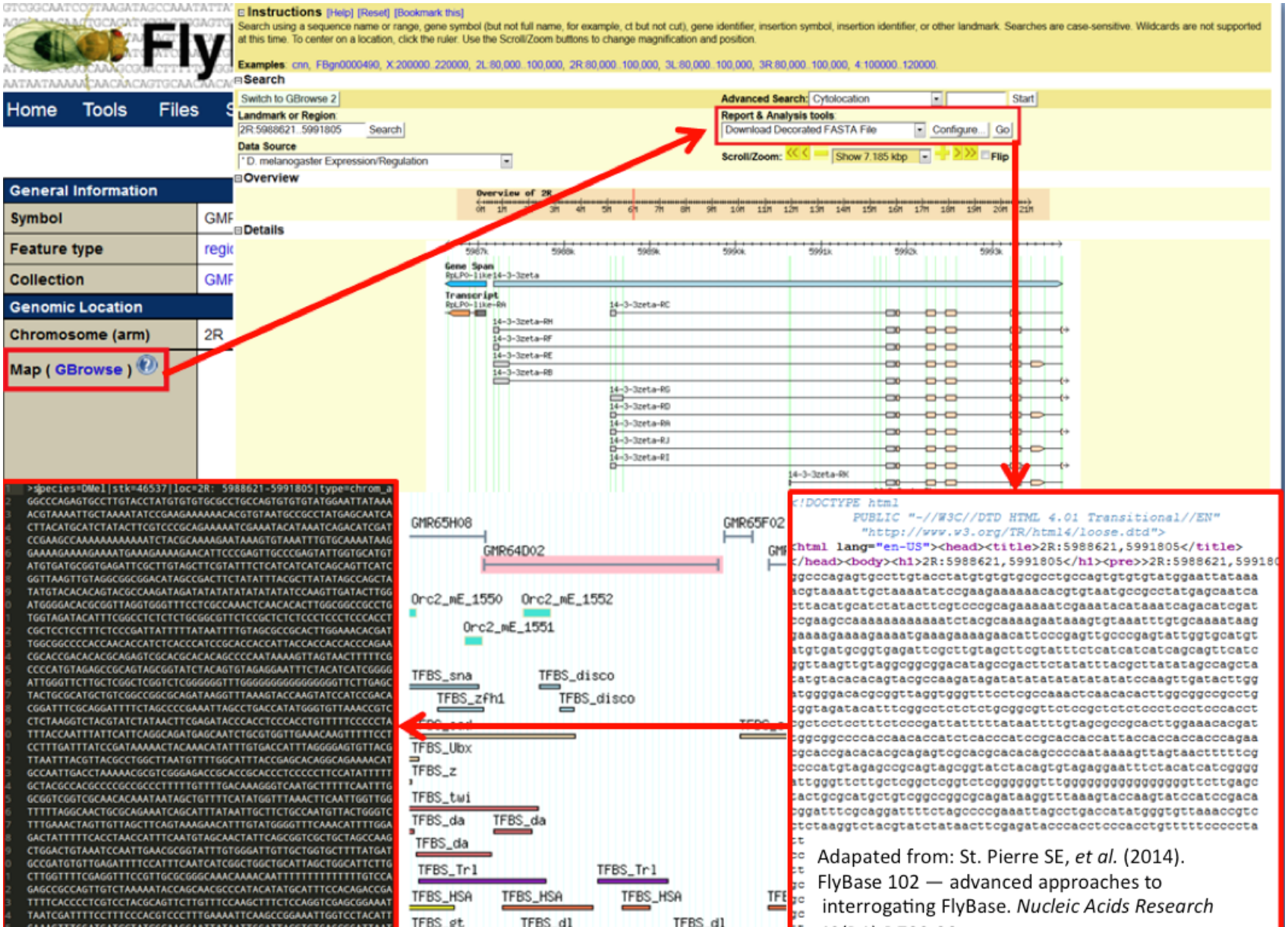




Display movie

Maximum intensity projections of the full stack are shown in two versions, one with the reference channel and one without. To generate the translation movie each of the ~1 µm slices of the original stack was turned into a movie frame. The MPEG-4 compression algorithm was used to reduce file size. During the production of the movies, the image data in each frame was contrast optimized to improve the ability to see weak signals. A calibration bar is included in each frame, which displays the maximum and minimum intensities in the original image. The calibration bars should be used when judging the strength of an expression pattern. A series of ~10 µm sub-stack maximum intensity

Automated data collection



Adapated from: St. Pierre SE, *et al.* (2014). FlyBase 102 — advanced approaches to interrogating FlyBase. *Nucleic Acids Research*

DNA motif discovery

- Six different DNA motif discovery tools used
  - YMF
  - MEME
  - Consensus
  - Weeder
  - Gibbs Motif Sampler
  - MDscan

DNA motif discovery (YMF)

University of Washington
Computer Science & Engineering

YMF 3.0: Finds short motifs in DNA sequences What is YMF ? FAQ

CSE Home | YMF Home | Send Mail

Motif size 6
Maximum of 0 spacers in middle
Maximum of 2 degenerate symbols (R,Y,W,S)
Organism -- Select One -- create own organism
User-created organisms None created so far can't find your organism ?
Paste Sequences (*) in FastA Format (See example)
Processing is faster if sequences are equi-length and unmasked.
Or Upload a FastA file (*): Browse...
SUBMIT

* Total uploaded sequence data should be < 10000 characters
News: YMF is now being hosted on a new server. If you encounter any problems using this web server, please let us know at sinhas@cs.uiuc.edu

Motifs in session

none

DNA motif discovery (MEME)

MEME Suite Menu
- Submit A Job
- Documentation
- Downloads
- User Support
- Alternate Servers
- Authors
- Citing

MEME
Multiple Em for Motif Elicitation

Version 4.9.1

Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (motif) for each pattern it discovers.

Data Submission Form

Required

Your e-mail address:

Re-enter e-mail address:

How do you think the occurrences of a single motif are distributed among the sequences?
- One per sequence
- Zero or one per sequence
- Any number of repetitions

Please enter the sequences which you believe share one or more
motifs. The sequences may contain no more than 60000 characters
total total in any of a large number of formats.

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

6 Minimum width (>= 2)

50 Maximum width (<= 300)

Enter the name of a file containing the sequences here:

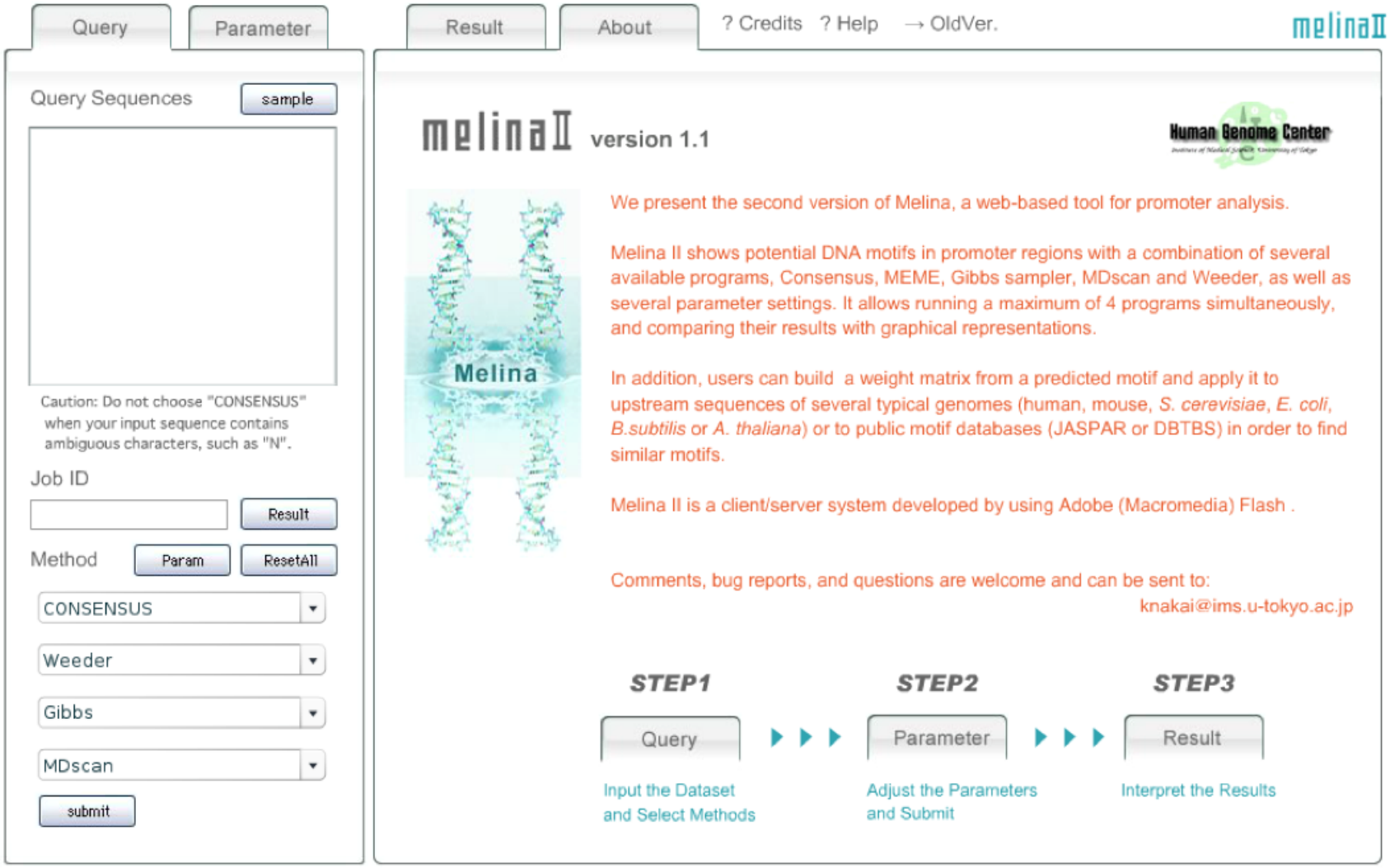
Browse... Clear

3 Maximum number of motifs to find

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

Options

DNA motif discovery (Melina-II)



Statistical methods

0: All discovered motifs ranked

1: Fisher's exact test and χ^2 test used to identify statistical significance

2: Statistical significance established with a Bonferroni correction

DNA motif ranking

$S(m) :=$ Score of motif

$$S(m) := \frac{\#\ \gamma^+ \text{ lines with motif in seq.}}{\#\text{ of } \gamma^+ \text{ lines}} \Big/ \frac{\#\text{ all lines with motif in seq.}}{\#\text{ of all lines}}$$

Statistical significance tests

Counts are summarized in the form of a 2×2 contingency table for each motif as follows.

	1	0
γ^+	x_1	$n_1 - x_1$
γ^-	x_2	$n_2 - x_2$

With test statistics given by:

$$\text{Test Statistic} := \begin{cases} x_1 & \text{Fisher's exact test} \\ \sum_{i=1}^{2} \frac{(x_i - n_i\hat{\theta})^2}{n_i\hat{\theta}(1-\hat{\theta})} & \chi^2 \textit{test} \end{cases}$$

Bonferroni correction

Let X denote the number of significant results obtained—due to chance alone—when testing a set of k hypotheses simultaneously, with a given significance level of α. Then

$$\begin{aligned}\mathbb{P}(X \geq 1 \mid \alpha) &= 1 - \mathbb{P}(X = 0) \\ &= 1 - (1 - \alpha)^k.\end{aligned}$$

As k increases, so too does $\mathbb{P}(X \geq 1)$. One way to address this issue is to only declare results of a test to be significant if the p-value is less than α/k.

Numerical example of Bonferroni correction

Take $k = 8$ and $\alpha = 0.05$, then $\mathbb{P}(X \geq 1) \doteq 0.34$.

Our new level of significance is given by $\alpha = 0.05/8 = 0.00625$.

Results

- 6,931 FASTA files created
- 160 DNA motifs discovered and ranked
- 8 DNA motifs identified to be of interest
- 10 contingency table runs performed on those 8 DNA motifs
- GUI application developed that assists in accessing relevant data

DNA motif

Sites ?

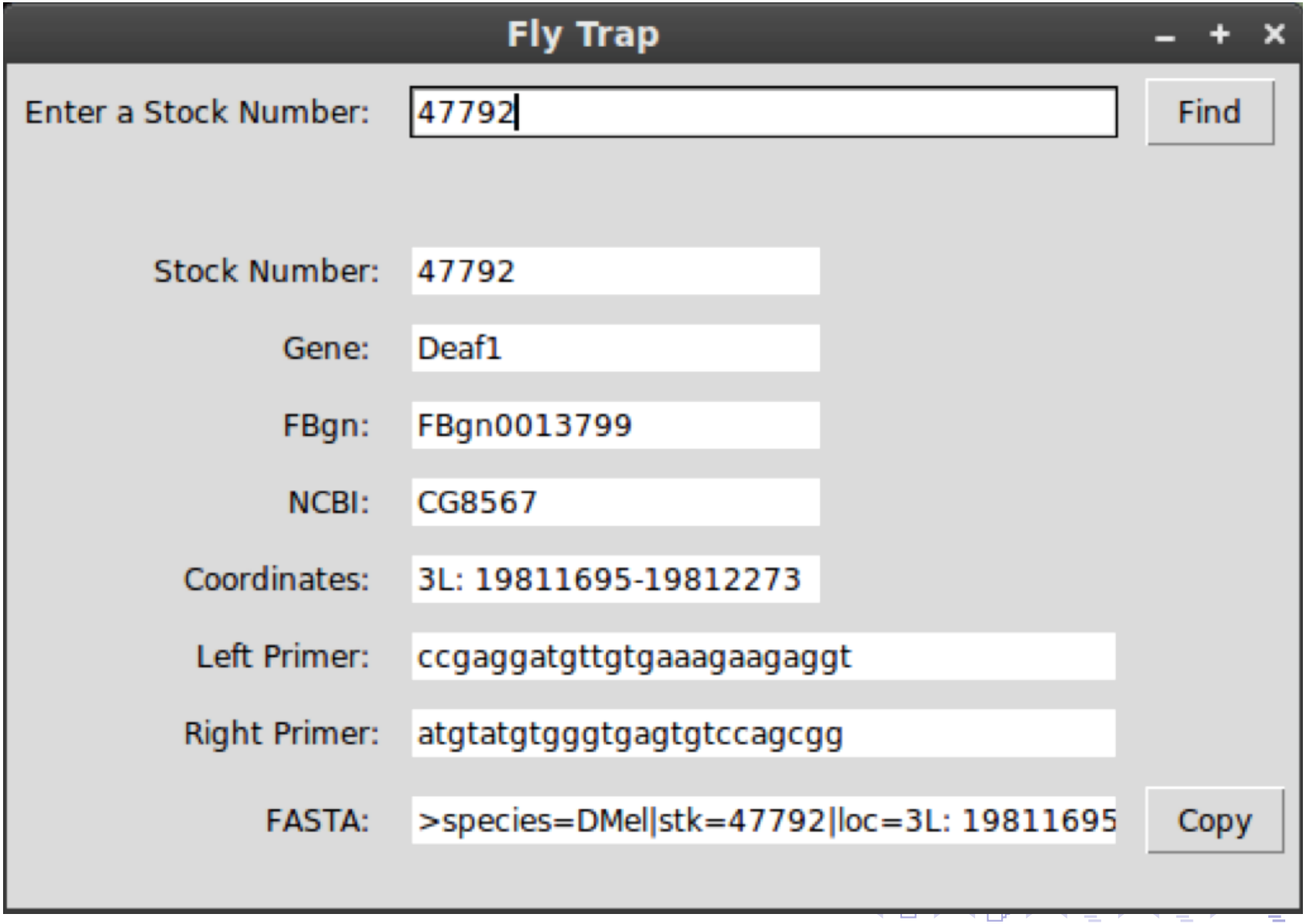
CCGTTTACCA GTGTGTGTGTGC GAGCGTTCCA
GTGTGTGTGT GTGTGTGTGTGC GCGAGTGTGT
CTCTCCCTTT GTGTGTGTGAGC CCGCCGCCTC
TGCTGGGTGT GTGTGTGTGTGT TTGTCTGGGT
CCTTGTGGTT GTGTGTGTATGC CGGCCGCTTC
GTATAACTCT GTGTGTGTATGC ATAATCATCA
TTATGTGCGC GTGTGTGTGTTT GTTTTCGGTT
AAAAAGTCAA GTGTGGGTAAGA CGGGAAAAAT
GAAAACCTCT GTGTGTGGGGGA AAATGGAACA

Test results from one run

	Test			
	χ^2		Fisher's Exact	
Motif	TS	*p*-value	TS	*p*-value
ATCGCTTCAA	3.1682	0.0751	3.0000	0.1182
TSCMTSCYCCYC	4.3077	0.0379	4.0000	0.0558
GGCACCGG	3.1682	0.0751	3.0000	0.1182
YSCMKSCYMCTS	8.0000	0.0047	7.0000	0.0051
WGTGTGTGTATG	3.1682	0.0751	3.0000	0.1182
AAGGTCKT	4.3077	0.0379	4.0000	0.0558
CTCGTATC	3.1682	0.0751	3.0000	0.1182
YCSTTTYTSYCT	3.1682	0.0751	3.0000	0.1182

Ambiguous base codes		
K := G or T	S := C or G	Y := C or T
M := A or C	W := A or T	

GUI application



Future work

0: Attempt to replicate results

1: Perform an additional five DNA motif discovery runs

2: Process and rank new motifs

3: Access statistical significance of motifs

4: If similar, then send information for wet lab

Acknowledgements

Mentors

Prof. Ming-Ying Leung Department of Mathematical Sciences

Prof. Kyung-An Han Department of Biological Sciences

Carolina Guerra (Alum.) Bioinformatics Program

Funding This work is supported in part by the NSF grant DUE-0926721 and NIH-NIMHD grant 5G12MD007592.