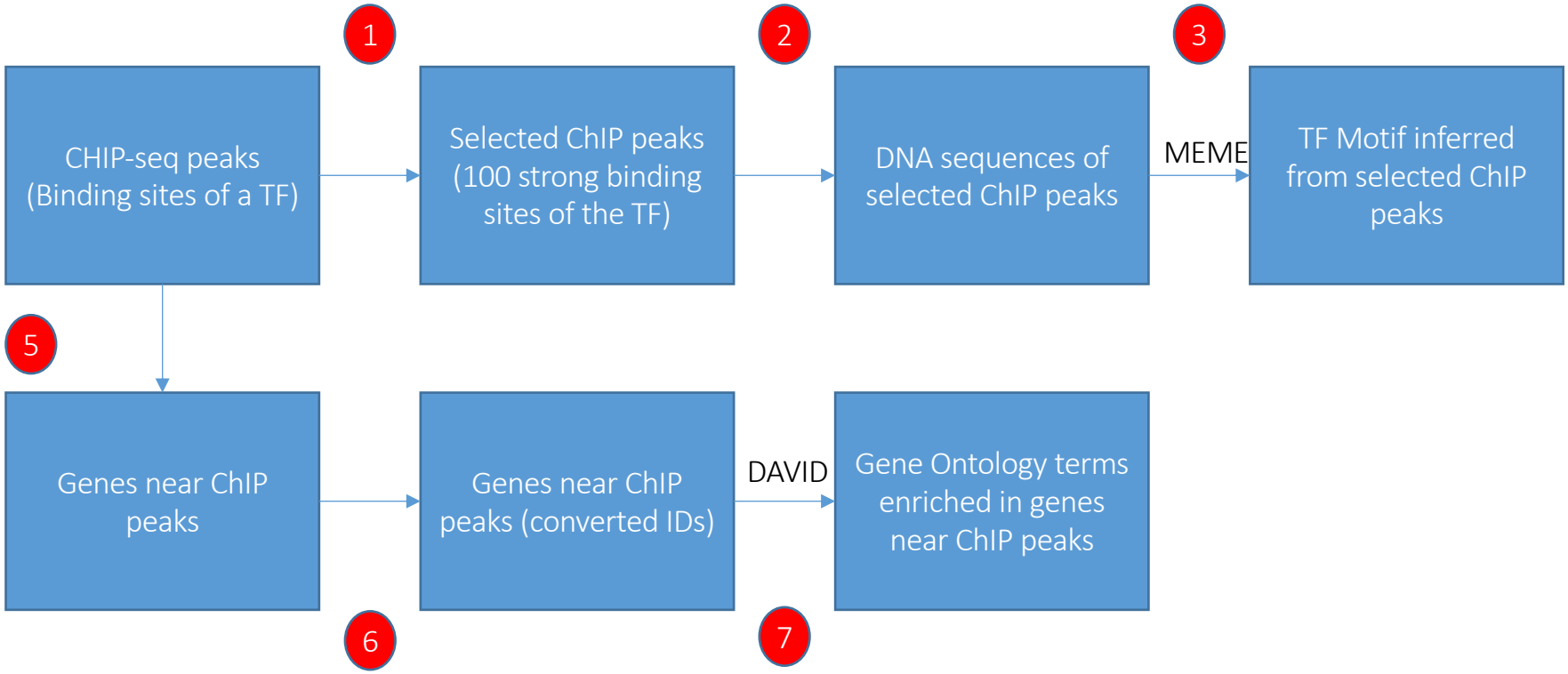
Regulatory Genomics Lab

Saurabh Sinha

PowerPoint by Saba Ghaffari
Edited by Shayan Tabe Bordbar

In this lab, we will do the following:.

- Use command line tools to manipulate a ChIP track for BIN TF in D. Mel.
- Subject peak sets to MEME suite.
- Compare MEME motifs with Fly Factor Survey motifs for BIN TF.
- Subject peak set to a gene set enrichment test.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks
5
Genes near ChIP peaks
Genes near ChIP peaks (converted IDs)
DAVID
Gene Ontology terms enriched in genes near ChIP peaks
6
7

Step 0A: Start the VM

- Follow instructions for starting VM. (This is the Remote Desktop software.)

- The instructions are different for UIUC and Mayo participants.

- Instructions for UIUC users are here: http://publish.illinois.edu/compgenomicscourse/files/2020/06/SetupVM_UIUC.pdf

- Instructions for Mayo users are here:

http://publish.illinois.edu/compgenomicscourse/files/2020/06/VM_Setup_Mayo.pdf

Step 0B: Accessing the IGB Biocluster

Open **Putty.exe**

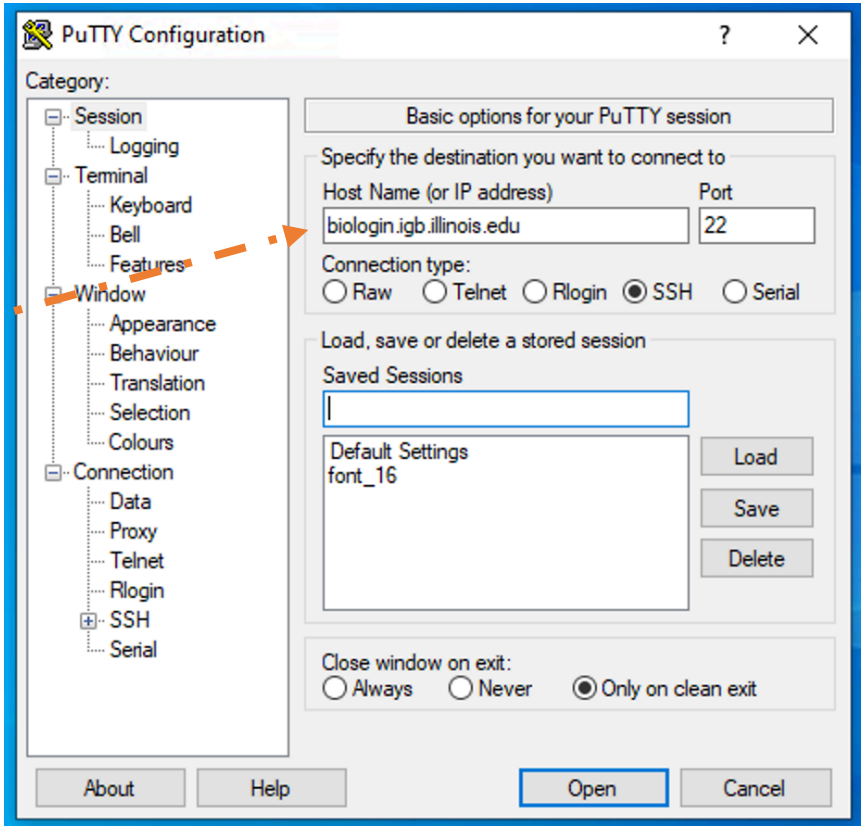
In the **hostname** textbox type:

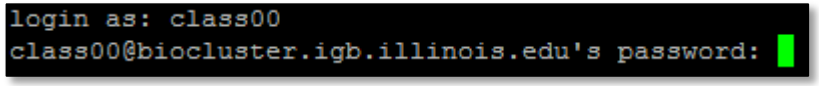
`biologin.igb.illinois.edu`

Click **Open**

If popup appears, Click **Yes**

Enter login credentials assigned to you; example, user **class00**.





Now you are all set!

In PuTTY

Step 0C: Lab Setup

The lab is located in the following directory:

`/home/classroom/mayo/2020/06_Regulatory_Genomics/`

Following commands will copy a shell script -designed to prepare the working directory- to your home directory. Follow these steps to copy and then submit the script as a job to biocluster:

```
$ cd ~/

# Note ~ is a symbol in Unix paths referring to your home directory

$ cp /home/classroom/mayo/2020/06_Regulatory_Genomics/src/prep-directory.sh ./

# Copies prep-directory.sh script to your working directory.

$ sbatch prep-directory.sh

# submits a job to biocluster to populate your home directory with necessary
files

$ squeue –u <userID> # to check the status of the submitted job
```

Step 0D: Working directory: data

Navigate to the created directory for this exercise and look what data folder contains.

```
$ cd 06_Regulatory_Genomics
$ ls
# output should be:
# data results src
$ ls data/
# BIN_Fchip_s11_1000.gff
# dm3.fasta
# flygenes_vm.bed
```

Name	Description
BIN_Fchip_s11_1000.gff	ChIP peaks for BIN transcription factor in GFF format
dm3.fasta	*Drosophila Melanogaster* genome
flygenes_vm.bed	Coordinates of all *Drosophila* genes in BED format

Step 0E: Working directory: scripts

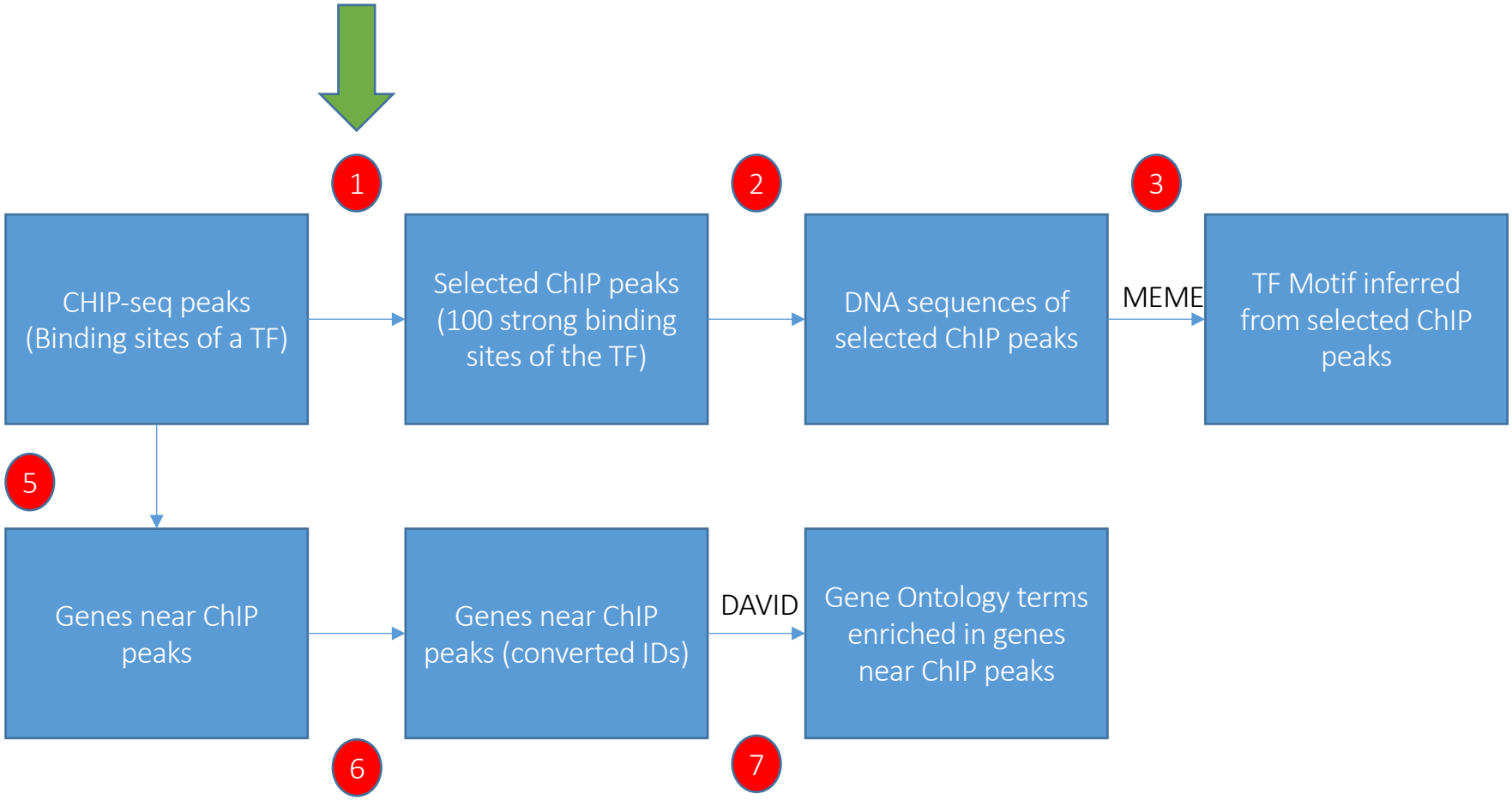
Navigate to the directory containing the scripts and look what's inside.

```
$ cd src

$ ls *.sh

# lists the scripts to be used in this lab:

# get_closest_genes.sh  get_sequence.sh  get_top100.sh
```

Computational Prediction of Motifs

In this exercise, after performing various file manipulations, we will use the MEME suite to identify a motif from the top 100 ChIP regions.

Subsequently, we will compare our predicted motif with the experimentally validated motif for BIN at Fly Factor Survey.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks
5
Genes near ChIP peaks
Genes near ChIP peaks (converted IDs)
DAVID
Gene Ontology terms enriched in genes near ChIP peaks
6
7

Step 1: Obtain the top 100 strongest ChIP peaks

- We will use “sort” command, to sort the peaks based on their score and then take the top 100 peaks.
- Use the following line to get get the top 100 chip peaks from the original ChIP gff file.

```
$ cd ~/06_Regulatory_Genomics/src/

$ head ~/06_Regulatory_Genomics/data/BIN_Fchip_s11_1000.gff

$ sbatch  get_top100.sh

# OUTPUT in ~/06_Regulatory_Genomics/results/Top_100_peaks.gff
```

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (get_top100.sh) is supposed to do in more detail.

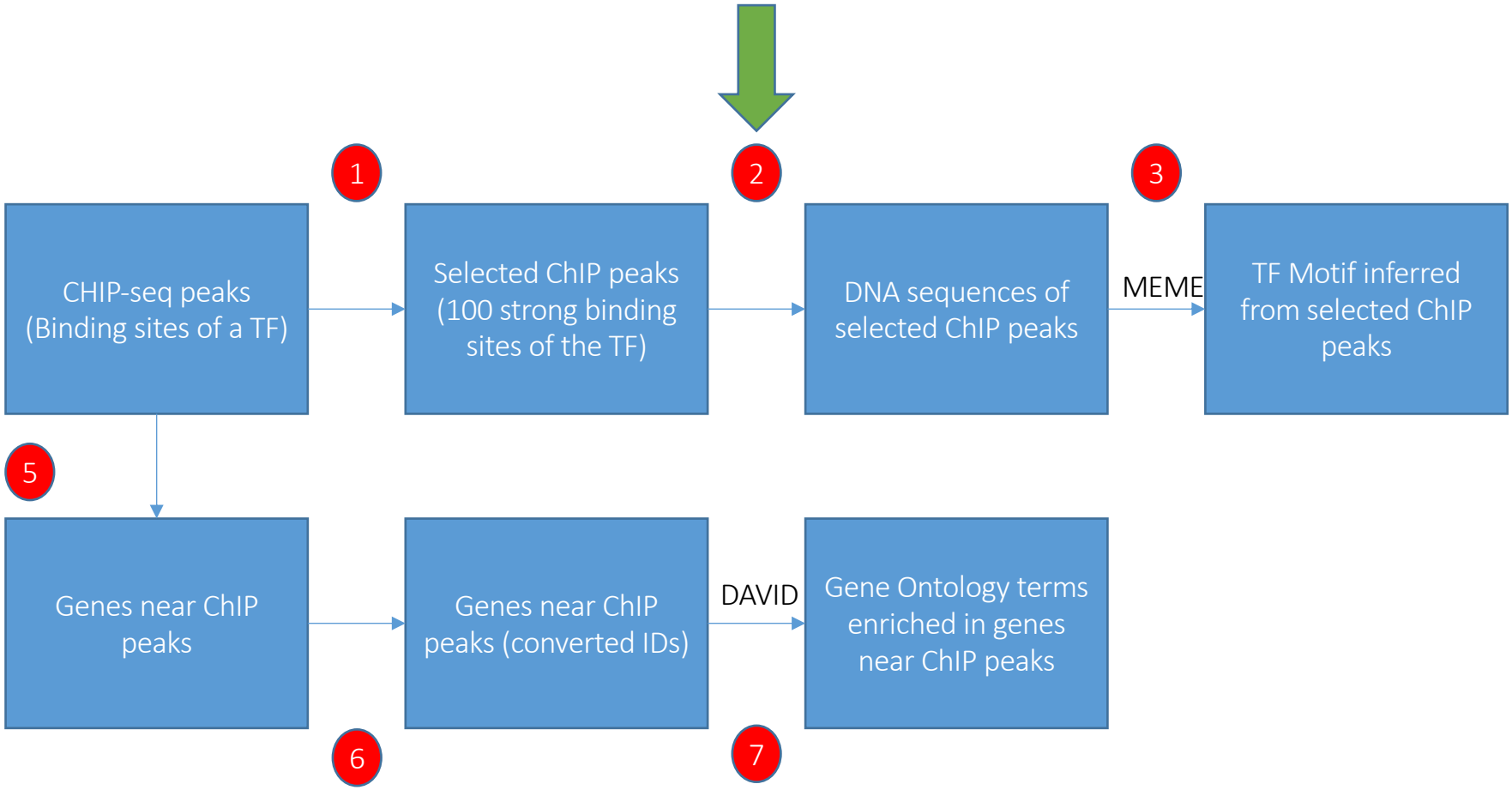
What's inside the **get_top100.sh** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J getTop100
#SBATCH -o getTop100.%j.out
#SBATCH -e getTop100.%j.err
#SBATCH -p classroom


# this is our input (gff)
export TOBESORTED=../data/BIN_Fchip_s11_1000.gff


sort -k 6,6nr  $TOBESORTED | head -100 > ../results/Top_100_peaks.gff
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory), run on the 'classroom' nodes, and name the job 'getTop100'

Create shortcut name for input ChIP peak file in GFF format.

Use Linux sort command to sort the file based on the numeric score stored in the 6th column of the gff file (ChIP score). [–k flag introduces the column to be sorted by. '**nr**' notes that we desire a **n**umeric sort in **r**everse order.] Output is directed to (>) Top_100_peaks.gff file.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks
5
Genes near ChIP peaks
Genes near ChIP peaks (converted IDs)
DAVID
Gene Ontology terms enriched in genes near ChIP peaks
6
7

Step 2: Extract DNA sequence of Top 100 ChIP Regions

We will use a “getfasta” tool from “bedtools” toolkit to get the DNA sequence for the top 100 ChIP peaks.

Usage:

Please do not try to Run the commands in the first box.
This is just to explain the arguments to bedtools getfasta

```
$ bedtools getfasta [options]      -fi <genome_file_name> > \

# specifies the path to the genome sequence in FASTA format

                                   -bed <file_name.bed>

# specifies the path to coordinates of input regions in (BED/GFF/VCF) # formats
```

Script get_sequence.sh uses Bedtools getfasta to get the sequence corresponding to peaks stored in Top_100_peaks.gff. Run the following command:

```
$ cd ~/06_Regulatory_Genomics/src/

$ sbatch get_sequence.sh

# OUTPUT in ~/06_Regulatory_Genomics/results/BIN_top_100.fasta

$ squeue -u <userID>
```

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (get_sequence.sh) is supposed to do in more detail.

What's inside the **get_sequence.sh** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J get_sequence
#SBATCH -o get_sequence.%j.out
#SBATCH -e get_sequence.%j.err
#SBATCH -p classroom

# load the tool environment
module load BEDTools

# this is our input (dm genome in fasta format)
export GENOME_DM3_FASTA=../data/dm3.fasta
export INPUT_CHIP=../results/Top_100_peaks.gff
export OUTPUT_NAME=../results/BIN_top_100.fasta

# use bedtools
bedtools getfasta -fi $GENOME_DM3_FASTA -bed $INPUT_CHIP | fold -w 60 > $OUTPUT_NAME
```

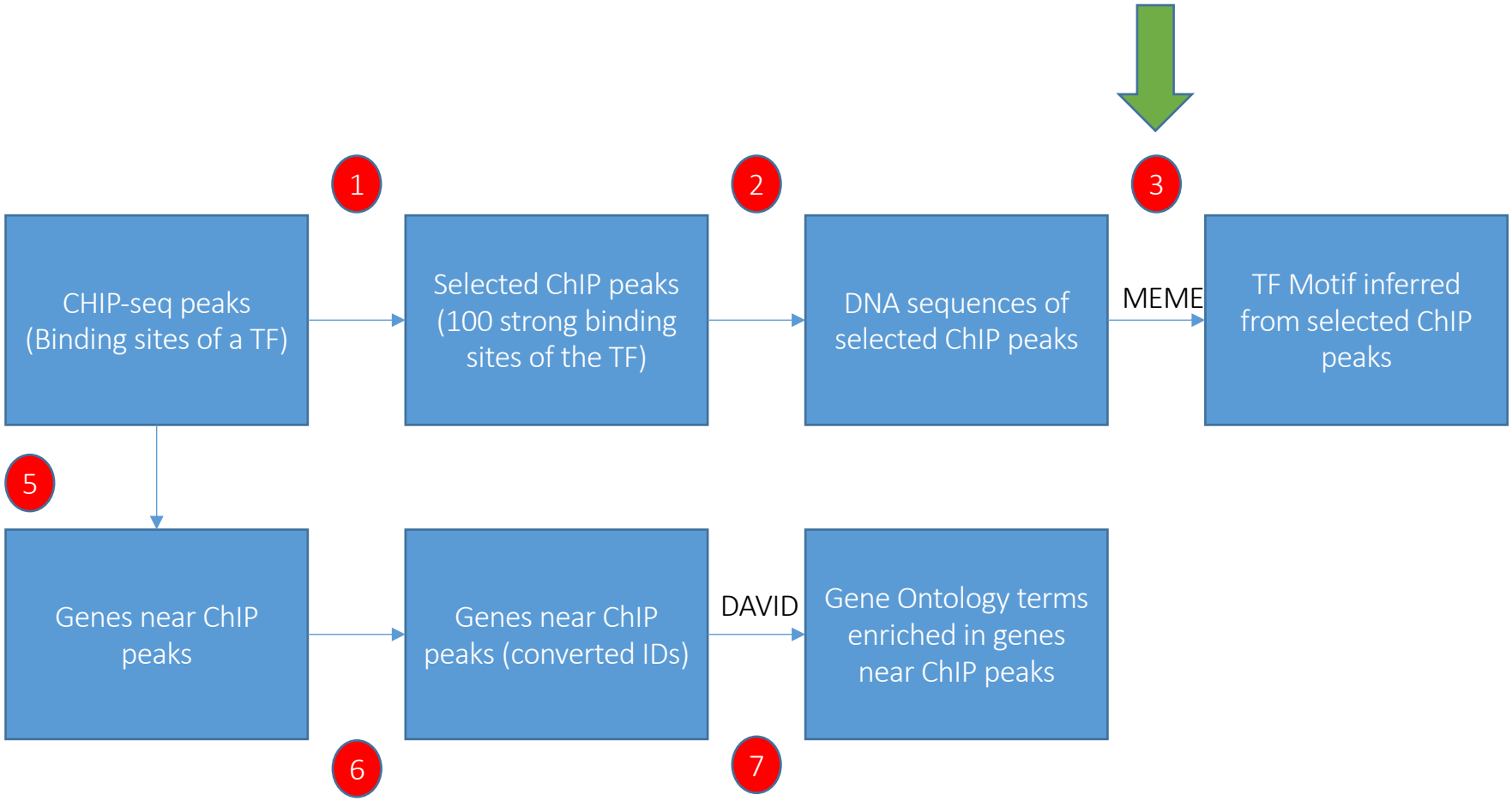
Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'get_sequence'

Load the software. We use a tool called 'BEDTools' to work with peak files.

Create shortcut names for input genome, input ChIP peak file and output FASTA file.

run 'bedtools getfasta' to get the DNA sequence in dm3.fasta genome corresponding to coordinates contained in Top_100_peaks.gff
fold –w 60 ensures that the width of lines in the output file does not exceed 60 characters.
results are directed to (>) BIN_top_100.fasta

Note that output of get_sequence.sh (BIN_top_100.fasta) has already been copied to the VM to be used in the next step.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks
5
Genes near ChIP peaks
Genes near ChIP peaks (converted IDs)
DAVID
Gene Ontology terms enriched in genes near ChIP peaks
6
7

Local Files (for UIUC users)

For viewing and manipulating the files needed for this laboratory exercise, denote the path C:\Users\IGB\Desktop\VM on the VM as the following:

[course_directory]

We will use the files found in:

[course_directory]\06_Regulatory_Genomics

Local Files (for mayo clinic users)

For viewing and manipulating the files needed for this laboratory exercise, denote the path **C:\Users\Public\Desktop\datafiles** on the VM as the following:

[course_directory]

We will use the files found in:

[course_directory]\06_Regulatory_Genomics

On Desktop

Step 3: Submit to MEME

DO NOT RUN THIS NOW. MEME TAKES A VERY LONG TIME.

In this step, we will submit the sequences to MEME

Go to the following address on your VM internet browser:

http://meme-suite.org/tools/meme

You can find BIN_top_100.fasta in the following directory on the VM:

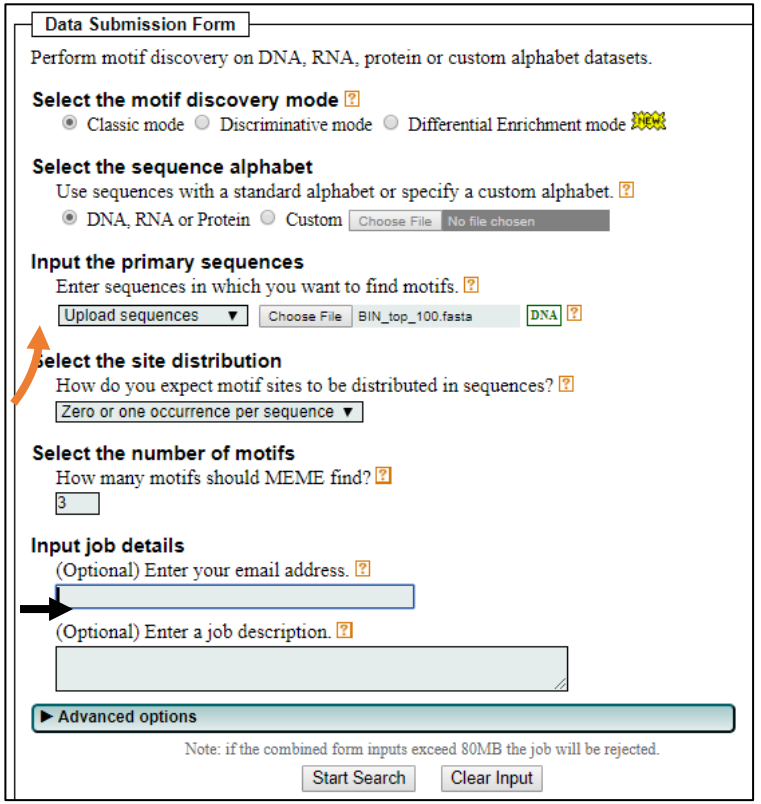
[course_directory]\06_Regulatory_Genomics\BIN_top_100.fasta

Upload your **sequences file** here

Enter **your email address** here.

Leave other parameters as default.

Click "Start Search".



Step 3A: Analyzing MEME Results

Go to the following web address: (You will receive notification email from MEME.

The webpage contains a summary of MEME's findings.

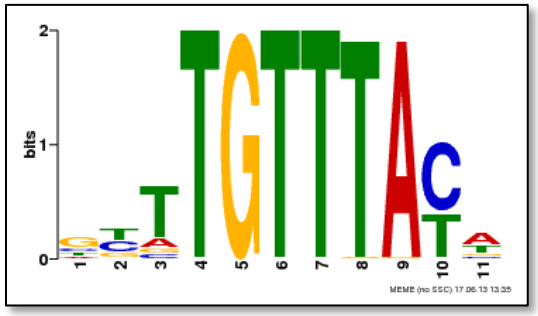
It is also available in the following directory:

```
[course_directory]\06_Regulatory_Genomics\MEME.html
```

Let's investigate the top hit.

Step 3B: Analyzing MEME Results

To the right is a LOGO of our predicted motif, showing the per position relative abundance of each nucleotide



At the bottom are the aligned regions in each of our sequences that helped produce this motif. As the p-value increases (becomes less significant) matches show greater divergence from our LOGO.

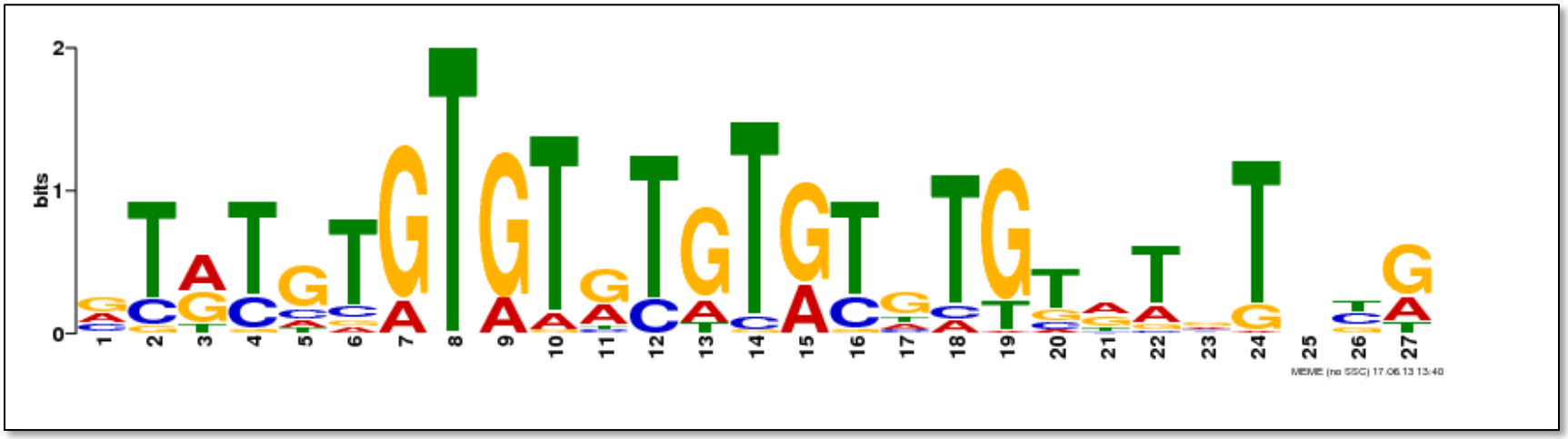
MOTIF LOCATIONS

Only Motif Sites | Motif Sites+Scanned Sites | All Sequences | Download PDF | Download SVG

	Name	*p*-value	Motif Locations
1.	chr3R:14813999-14814500	4.63e-5	
2.	chr2L:2445699-2446200	3.60e-2	
3.	chr3R:26607599-26608100	6.68e-8	
4.	chr2R:7896299-7896800	4.36e-5	
5.	chr2R:2676399-2676900	7.14e-2	
6.	chr2R:2871699-2872200	1.96e-4	
7.	chr2R:9451549-9452050	5.55e-4	
8.	chr3R:17214099-17214600	1.17e-1	
9.	chr2L:18336599-18337100	1.16e-4	
10.	chrX:7057299-7057800	3.58e-3	
11.	chr3L:14359199-14359700	1.50e-1	
12.	chr2R:9112349-9112850	1.04e-3	
13.	chr3R:4854499-4855000	1.03e-6	
14.	chr3R:17222349-17222850	1.88e-5	

Step 3C: Analyzing MEME Results

Other predicted motifs do not seem as plausible.



Step 4A: Comparison with Experimentally Validated Motif for BIN

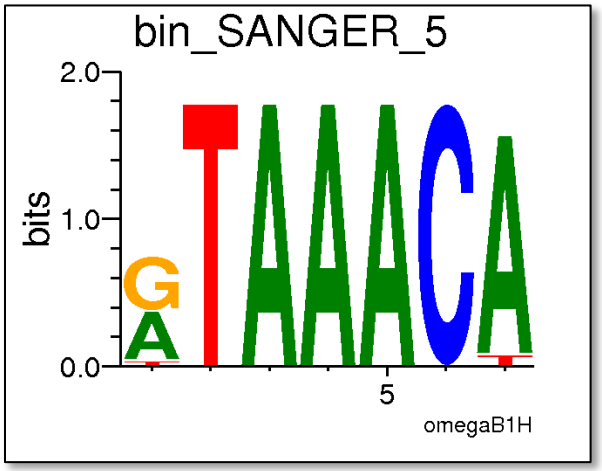
FlyFactorSurvey is a database of TF motifs in Drosophila Melanogaster.

Use the internet browser on your VM to go to the following link to view the motif for BIN:

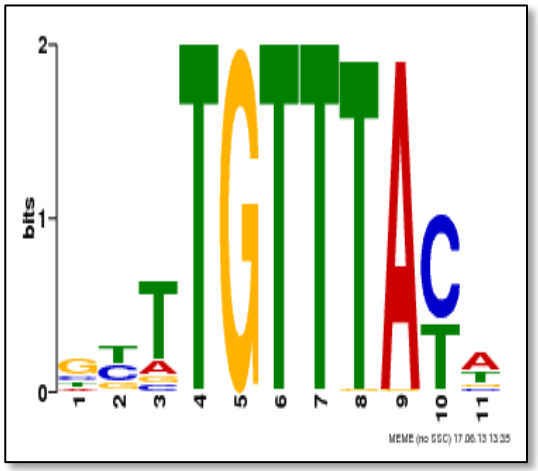
http://pgfe.umassmed.edu/ffs/TFdetails.php?FlybaseID=FBgn0045759

Step 4B: Comparison with Experimentally Validated Motif for BIN

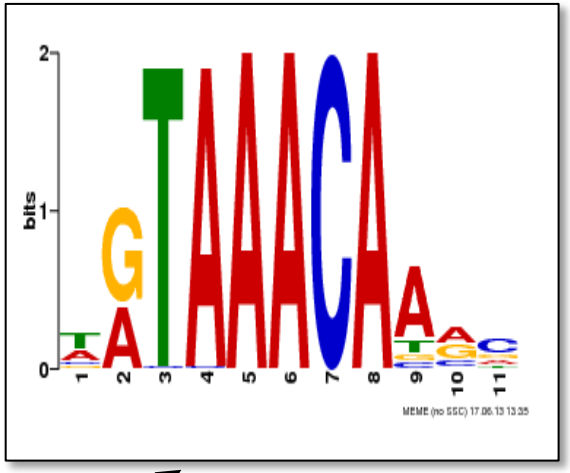
Actual BIN Motif



Best MEME Motif



Best MEME Motif Reverse Complemented

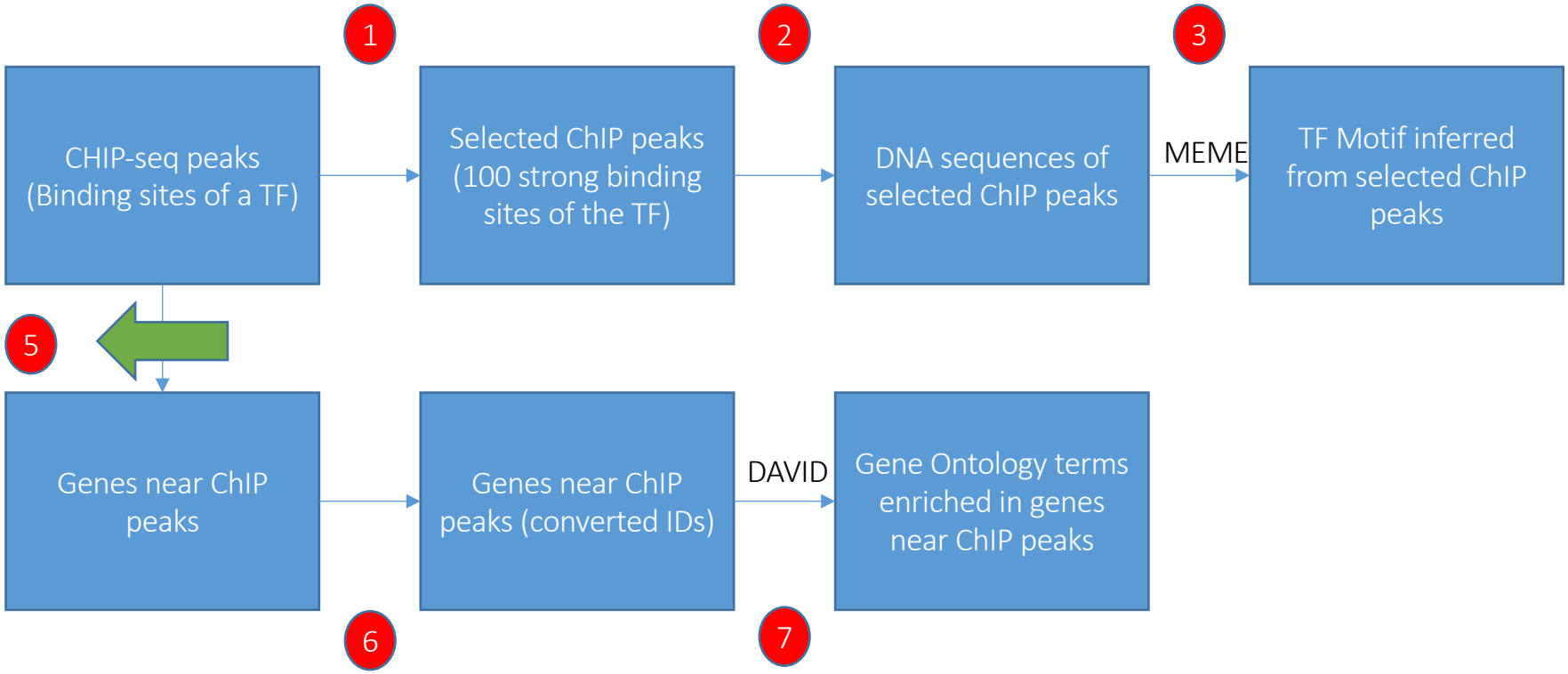


There is strong agreement between the actual motif and the reverse complement of MEME's best motif. This indicates MEME identified the BIN motif from the top 100 ChIP regions for this TF.

Gene Set Enrichment Analysis

In this exercise, we will extract the nearby genes for each one of the ChIP peaks for BIN.

We will then subject the nearby genes to enrichment analysis tests on various Gene Ontology gene sets utilizing DAVID.

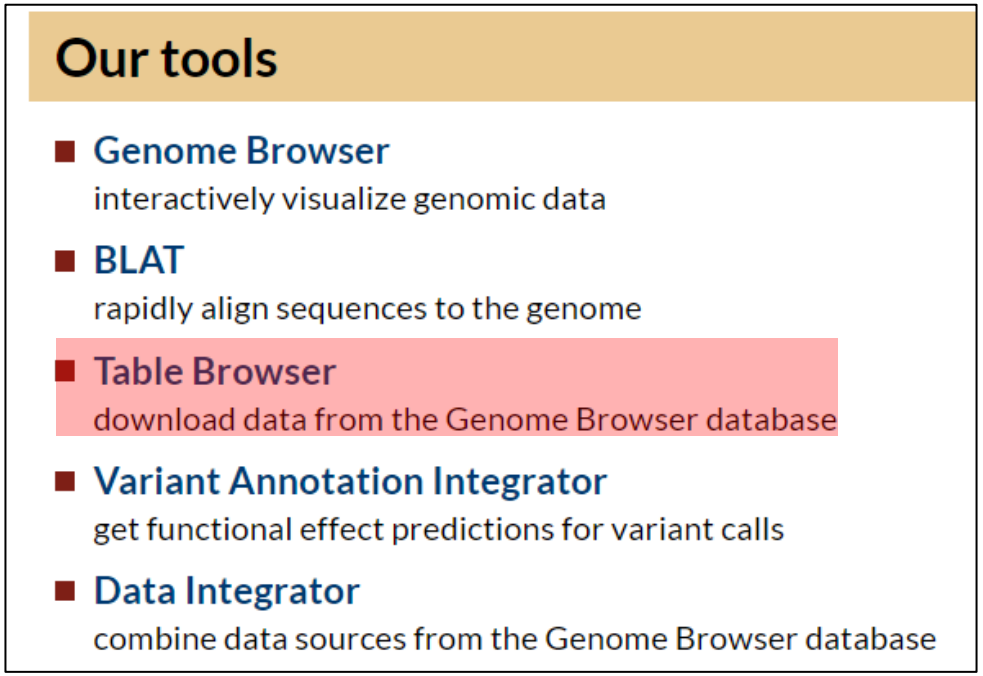

1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks
5
Genes near ChIP peaks
Genes near ChIP peaks (converted IDs)
DAVID
Gene Ontology terms enriched in genes near ChIP peaks
6
7

On Desktop

Step 5A: Acquire Nearby Genes

In this step, we will acquire all genes in Drosophila Melanogaster using UCSC Main Table Browser. Go to the following address using your VM internet browser:

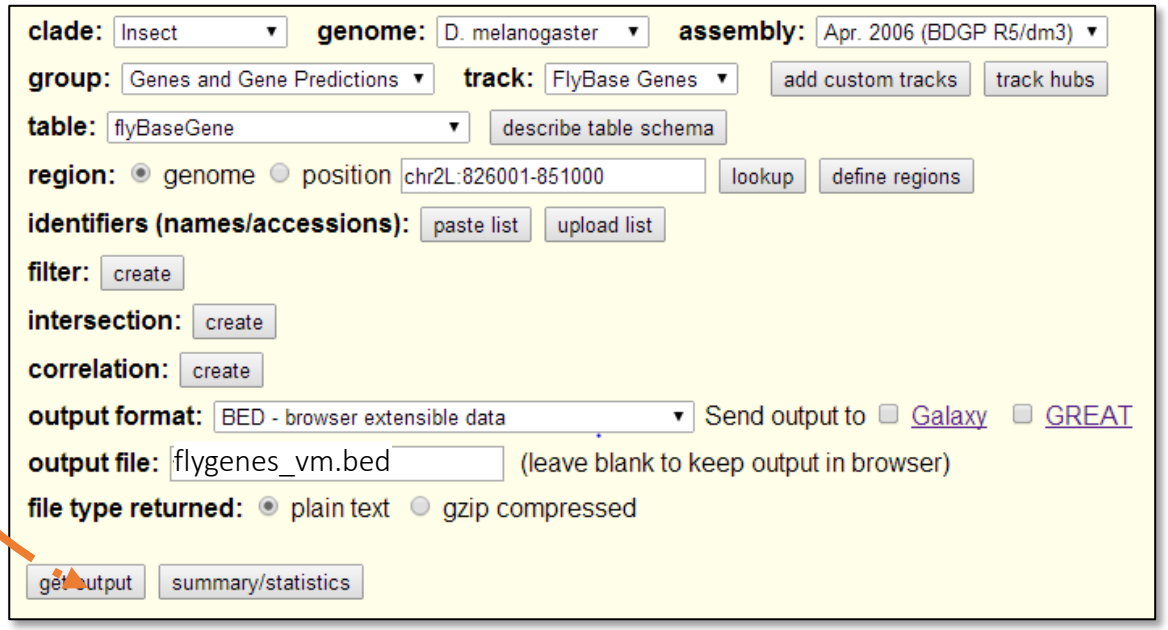
https://genome.ucsc.edu/

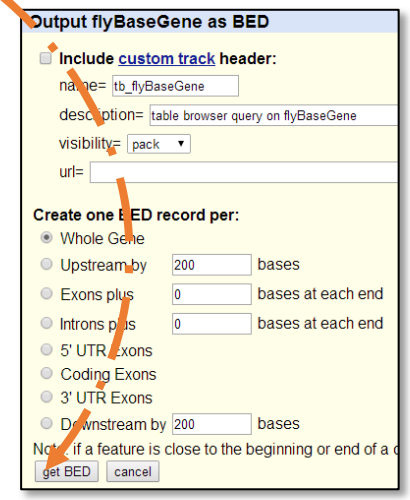


Step 5B: Acquire Nearby Genes

Ensure the following settings are configured.

Click **get output** and then **get BED.**





- Output of this exercise will be stored in VM Downloads directory.
- Note that the output of this exercise (flygenes_vm.bed) has already been copied to the following directory on biocluster for convenience:

`~/06_Regulatory_Genomics/data/flygenes_vm.bed`

In PuTTY

Step 5C: Acquire Nearby Genes

We will use a “closest” tool from “bedtools” toolkit to get the closest non-overlapping genes to the BIN ChIP peaks.

Usage:

Please do not try to Run the commands in the following box. This is just to explain the arguments to bedtools closest

```
$ bedtools closest [options]   -a <file_name> > \

# specifies the path to chip peak file in BED format

                               -b <file_name>

# specifies path to the BED file containing the coordinates for the

# feature of interest (i.e. genes in this case).
```

Step 5C: Acquire Nearby Genes

Script get_closest_genes.sh uses Bedtools closest to get name of the genes closest to ChIP peaks stored in BIN_Fchip_s11_1000.gff

All gene names and their corresponding coordinates are stored in flygenes_vm.bed which has been copied here from the output of exercise 5B.

Run the following command:

```
$ cd ~/06_Regulatory_Genomics/src/

$ sbatch get_closest_genes.sh

# OUTPUT in ~/06_Regulatory_Genomics/results/cg_transcript.txt

$ squeue –u <userID> # to check the status of the submitted job
```

Note that output of get_closest_genes.sh (cg_transcript.txt) has already been copied to the following directory on your VM for convenience.
[course_directory]\06_Regulatory_Genomics\cg_transcript.txt

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (get_closest_genes.sh) is supposed to do in more detail.

What's inside the **`get_closest_genes.sh`** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J get_closest_genes
#SBATCH -o get_closest_genes.%j.out
#SBATCH -e get_closest_genes.%j.err
#SBATCH -p classroom
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'get_closest_genes'

```
# load the tool environment
module load BEDTools
module load bedops
```

Load toolkits 'BEDTools' and 'bedops'

```
# this is our input (dm genome in fasta format)
export INPUT_CHIP_GFF=../data/BIN_Fchip_s11_1000.gff
export INPUT_CHIP_BED=../results/BIN_Fchip_s11_1000_sorted.bed
export FLYGENE_BED=../data/flygenes_vm.bed
export FLYGENE_BED_SORTED=../results/flygenes_vm_sorted.bed
export OUTPUT_NAME=../results/cg_transcript.txt
```

Create shortcut names for input ChIP peak files, input gene files, and output file.

```
gff2bed < $INPUT_CHIP_GFF > $INPUT_CHIP_BED
```

convert gff to bed format. Using 'gff2bed' tool from BEDOPS toolkit

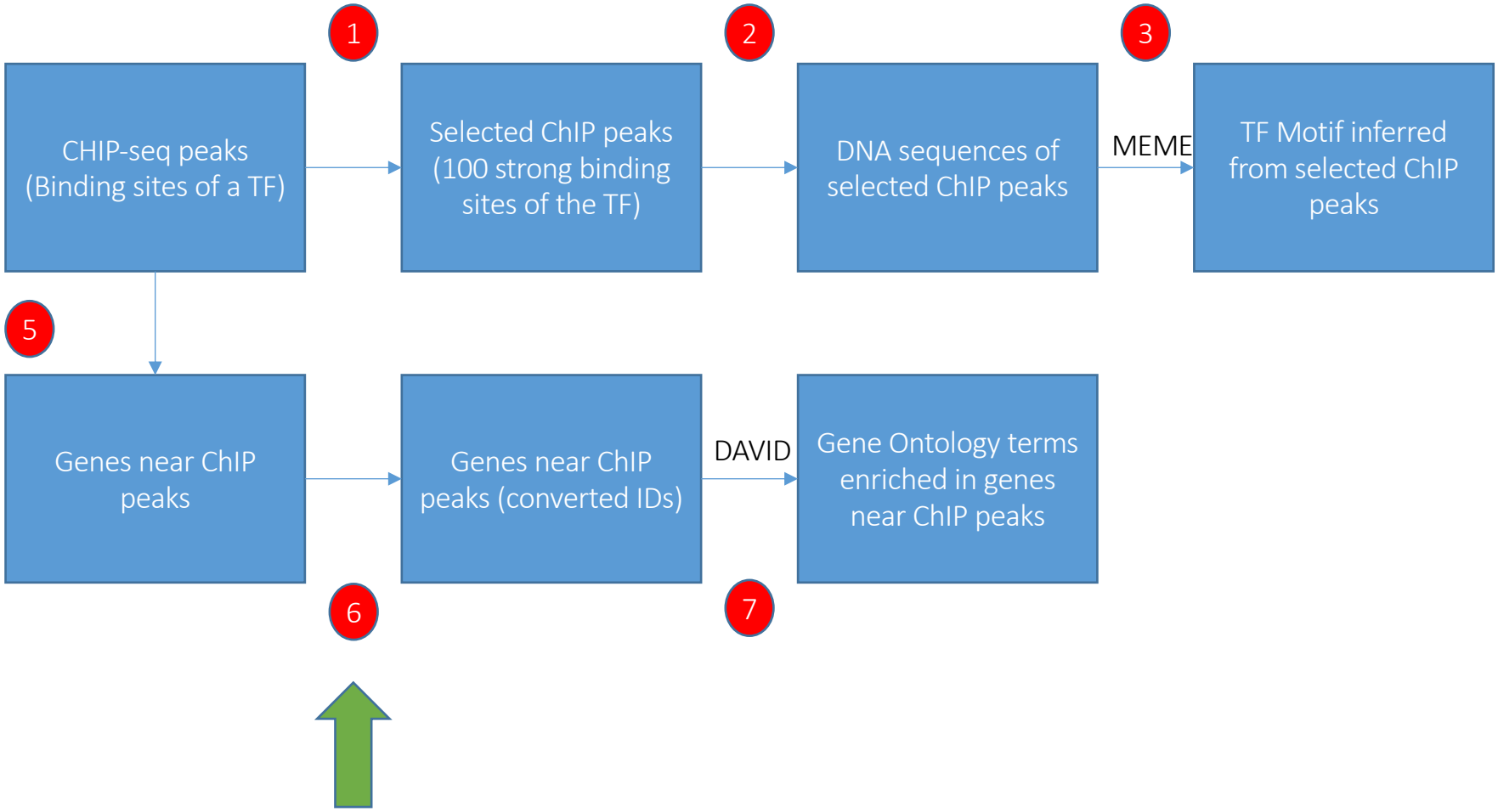
```
sortBed -i $FLYGENE_BED > $FLYGENE_BED_SORTED
```

inputs bed files to "bedtools closest" should be sorted based on genomic coordinates. 'sortBed' from bedtools does this.

```
bedtools closest -io -t first -a $INPUT_CHIP_BED -b $FLYGENE_BED_SORTED | cut -f 14 > $OUTPUT_NAME
```

'bedtools closest' finds the closest feature in -b to each line in -a
-io flag can be used in order to avoid overlapping features.
-t flag can be used to determine the action when there are ties. Can be one of 'first', 'all', or 'last'
'cut' is a Linux command used to extract the 14th column (-f 14) of the output, which contains gene names.

Exit putty by either closing the window or typing ‘exit’ in the command prompt.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks
5
Genes near ChIP peaks
Genes near ChIP peaks (converted IDs)
DAVID
Gene Ontology terms enriched in genes near ChIP peaks
6
7

On Desktop

Step 6A: Convert IDs

The enrichment tool we will use doesn't accept genes in this format.

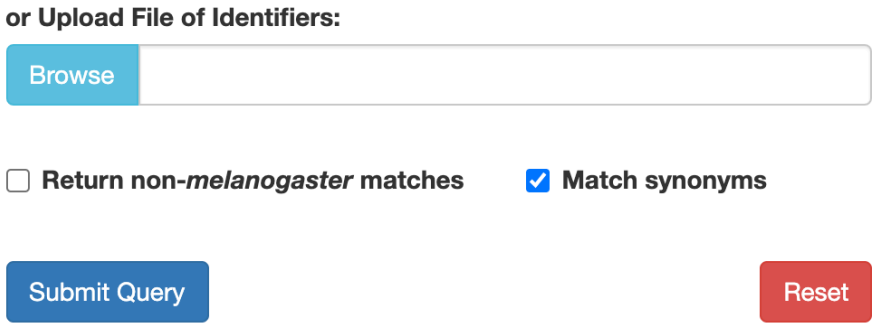
We will use the FlyBase ID converter to convert these transcript ids into FlyBase transcript ids.

On Desktop

Step 6A: Convert IDs

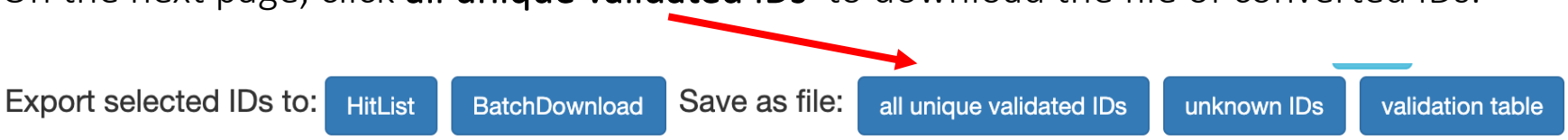
You can find a copy of cg_transcript.txt in the following location on the VM:
`[course_directory]\06_Regulatory_Genomics\cg_transcript.txt`
Go to the following link on your VM internet browser:

https://flybase.org/convert/id

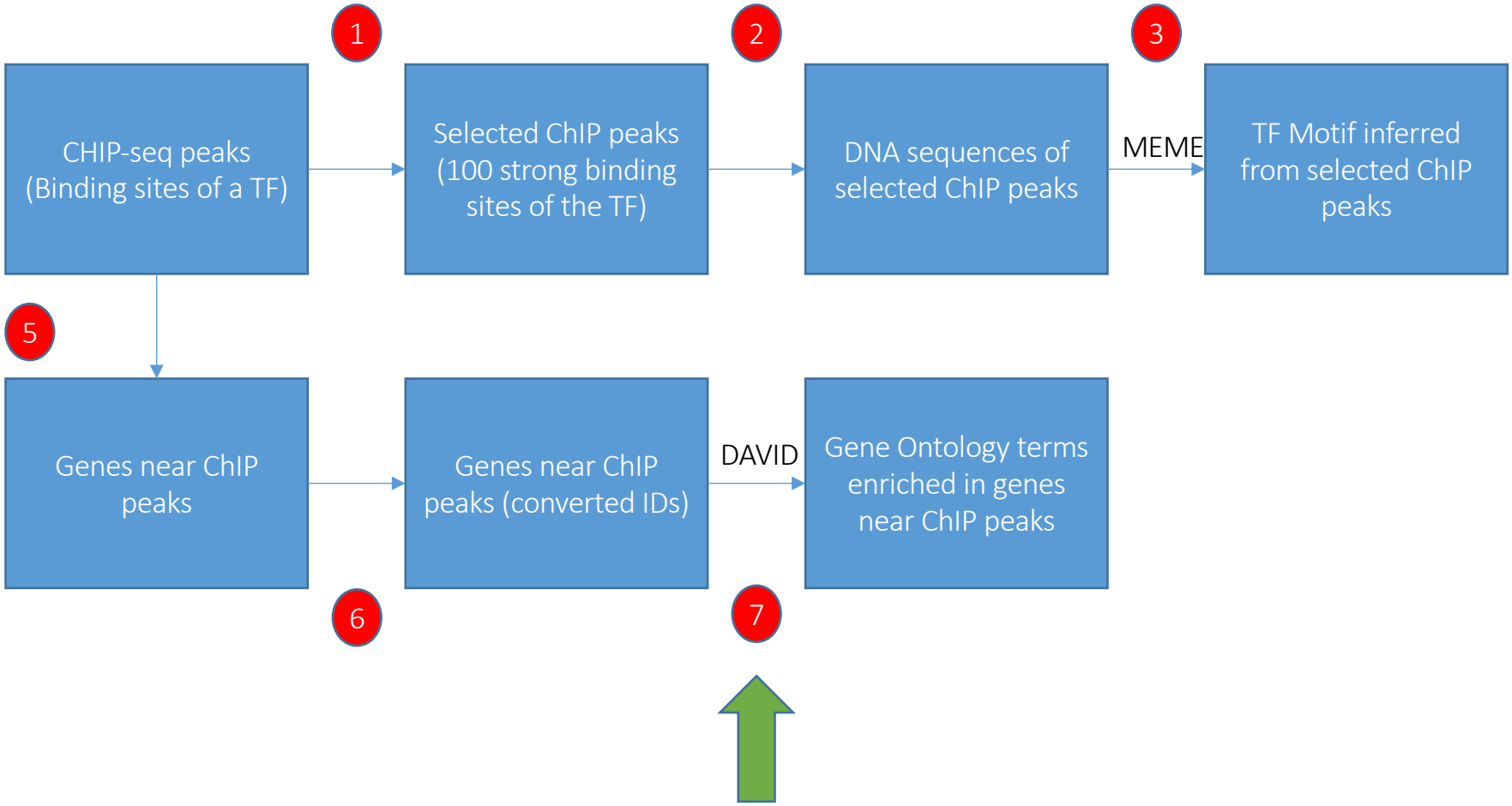


- Click Browse
- Navigate to the location of **cg_transcript.txt** and click **Open**
- Click **Submit Query**

On the next page, click **all unique validated IDs** to download the file of converted IDs.



Note that the downloaded file is named "FlyBase_IDs.txt" and will be in the Downloads folder.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks
5
Genes near ChIP peaks
Genes near ChIP peaks (converted IDs)
DAVID
Gene Ontology terms enriched in genes near ChIP peaks
6
7

Step 7A: Gene Set Enrichment - DAVID

With our correct ids of transcripts of genes near ChIP peaks, we now wish to perform a gene set enrichment analysis on various gene sets.

A tool that allows us to do this from a web interface is **DAVID** located at the following address (use your VM internet browser to go to this link):

https://david-d.ncifcrf.gov/summary.jsp

Step 7B: Gene Set Enrichment - DAVID

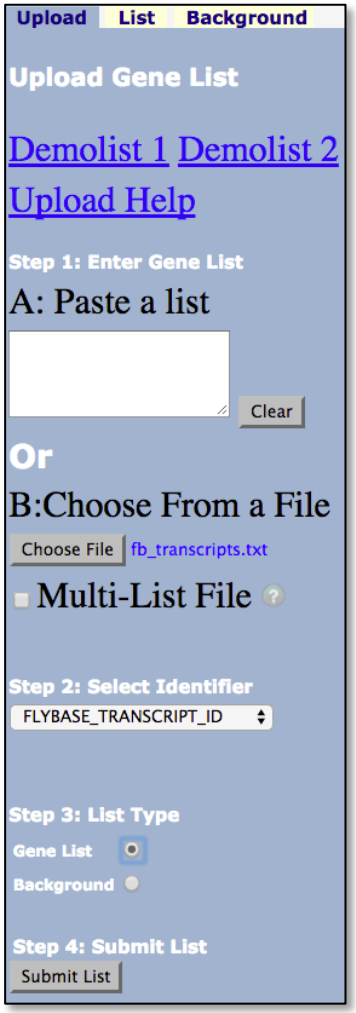
We will perform a Gene Set Enrichment Analysis on our transcript list (gene list) and see what GO categories are enriched in this set.

Analyze the gene list with **Functional Annotation Tool**

- Click **Choose File** and select "FlyBase_IDs.txt" from Downloads folder.
- If you were not able to download FlyBase_IDs.txt in the previous step:

Note that a copy of "FlyBase_IDs.txt" has already been copied to the following directory, you can instead use that file in this step: [course_directory]\06_Regulatory_Genomics\

- Under **Select Identifier** select **FLYBASE_TRANSCRIPT_ID.**
- Under **Step 3: List Type** check **Gene List**.
- Click **Submit List**.



Step 7C: Gene Set Enrichment - DAVID

On the next page, select **Functional Annotation Chart**.

Our gene set seems to be enriched in the **transcription regulator activity** Go term

This is consistent with the activity of **BIN** transcription factor in the literature:

https://flybase.org/reports/FBgn0045759#gene_ontology_section_sub

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_MF_FAT	transcription regulator activity	RT		58	10.3	1.0E-7	5.0E-5
	GOTERM_BP_FAT	tissue morphogenesis	RT		30	5.3	2.8E-7	4.2E-4
	GOTERM_BP_FAT	regulation of transcription, DNA-dependent	RT		49	8.7	8.8E-6	6.5E-3
	GOTERM_BP_FAT	epithelium development	RT		26	4.6	1.2E-5	5.6E-3
	GOTERM_BP_FAT	regulation of RNA metabolic process	RT		52	9.2	1.5E-5	5.3E-3
	GOTERM_BP_FAT	morphogenesis of an epithelium	RT		25	4.4	1.7E-5	4.9E-3
	GOTERM_BP_FAT	regulation of transcription	RT		58	10.3	2.1E-5	5.1E-3
	GOTERM_BP_FAT	embryonic development ending in birth or egg hatching	RT		24	4.3	4.6E-5	9.7E-3
	GOTERM_BP_FAT	negative regulation of cell differentiation	RT		13	2.3	4.9E-5	9.0E-3
	GOTERM_MF_FAT	transcription factor activity	RT		34	6.0	4.9E-5	1.2E-2
	SP_PIR_KEYWORDS	DNA binding	RT		17	3.0	5.1E-5	1.1E-2
	GOTERM_BP_FAT	cell morphogenesis	RT		37	6.6	6.0E-5	9.8E-3
	SP_PIR_KEYWORDS	dna-binding	RT		36	6.4	6.6E-5	7.2E-3
	GOTERM_BP_FAT	embryonic development via the syncytial blastoderm	RT		23	4.1	6.6E-5	9.7E-3