Welcome to Genome Enhancer 2.0 BETA!

Genome Enhancer can scan any NCBI Genbank formatted genome for clusters of DNA motifs ranging in size from 3-500 bp. It uses a novel algorithm that scans genomes about 50X faster than our original Fly Enhancer program (Markstein et al., PNAS 99:763-768, 2002). Currently Genome Enhancer is offered in a line-command mode and will eventually be available with a web-interface.

If you would like to run Genome Enhancer with a genome not currently supported, please email us with your request.

Note: This software is being distributed pre-publication for academic research purposes only. We will make the source code available under the General Public License (http://www.gnu.org/copyleft/gpl.html) upon publication in a peer-reviewed journal. If you would like to use this program for commercial purposes, please contact us.

Contacts:
Michele Markstein: mmarkstein@opengenomics.org
Peter Markstein: peter@insilicolabs.org

REQUIREMENTS

Use this guide once you have unzipped the enhancer program and placed it on your desktop:

Your Downloaded File:

e.g. G5enhancer_fl G4enhancer_fl

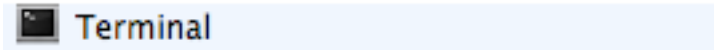
In this example mac_enhance

The contents of your folder should look like this:

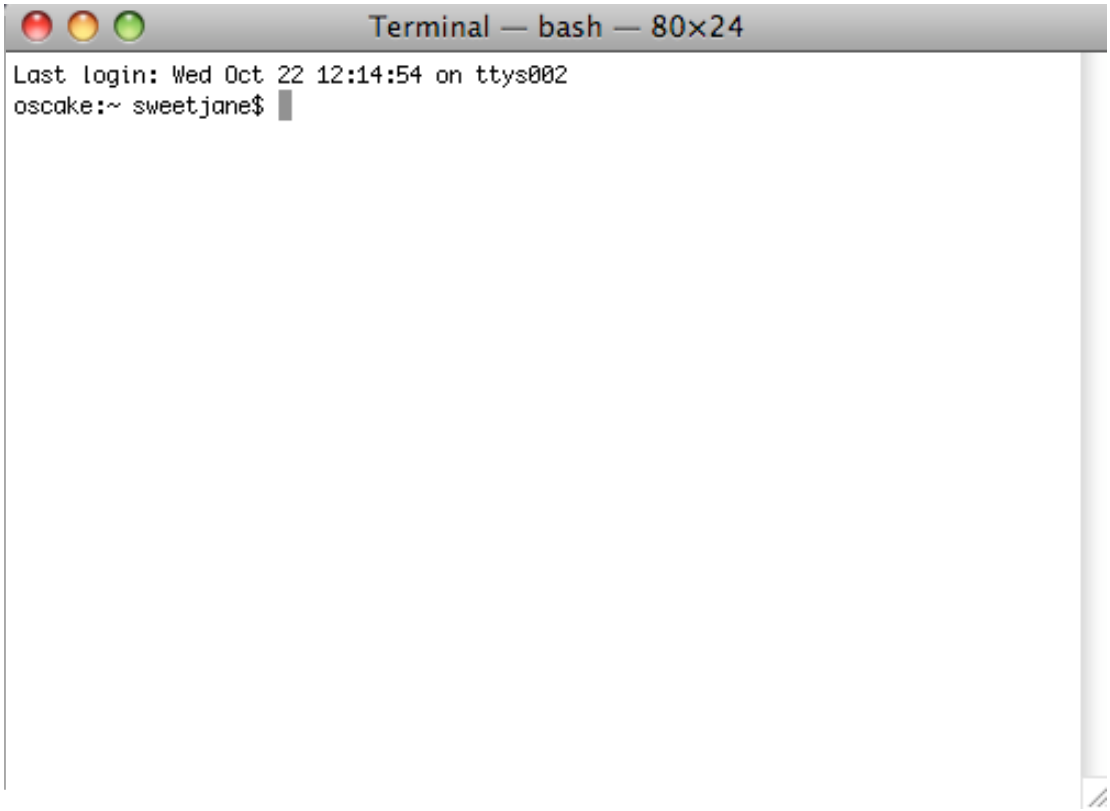
Name	Date Modified	Size	Kind
datasize.txt	Feb 6, 2007, 12:52 AM	4 KB	Plain text
enhancer	Today, 11:32 AM	36 KB	Unix Executable File
excep...ns.txt	Feb 6, 2007, 12:52 AM	8 KB	Plain text
genome.txt	Feb 6, 2007, 1:03 AM	28.2 MB	Plain text
master.txt	Feb 6, 2007, 1:05 AM	4 KB	Plain text
utrinfo.txt	Feb 6, 2007, 1:05 AM	308 KB	Plain text
xallnames.txt	Feb 6, 2007, 1:06 AM	788 KB	Plain text
xchrdata.txt	Feb 6, 2007, 1:06 AM	392 KB	Plain text
xcontigs.txt	Feb 6, 2007, 1:06 AM	4 KB	Plain text

GUIDE TO THE LINE-COMMAND VERSION OF GENOME ENHANCER

1. Double-click on the Terminal application.
This application comes with all Macs and is stored in the Applications Folder, within the Utilities Folder.



You will then see a window that looks something like this:



This window is the interface for the line-command version of Genome Enhancer.

2. Use the **cd** (“change directory”) command to open the genome enhancer folder that you placed on the desktop.

In this example, the program is in a folder called mac_enhancer on the Desktop. To get to the folder you need to “change directories” to the mac_enhancer folder in the Desktop folder:

Type cd Desktop/filename

e.g.
Type cd Desktop/G4enhancer_fly
Type cd Desktop/G5enhancer_fly
Type cd Desktop/PowerBook_enhancer_worm

…depending on the folder that you downloaded

3. Use the **ls** (list) command to verify that you have all the files

Type ls

```
Last login: Wed Oct 22 12:14:54 on ttys002
oscake:~ sweetjane$ cd Desktop/mac_enhancer
oscake:mac_enhancer sweetjane$ ls
datasize.txt    exceptions.txt  master.txt      xallnames.txt   xcontigs.txt
enhancer        genome.txt      utrinfo.txt     xchrdata.txt
oscake:mac_enhancer sweetjane$
```

you should see these 9 files

4. Use the **./** command to run the enhancer program

In this example we are running genome enhancer with the Drosophila melanogaster genome. Genome enhancer works the same way with all genomes.

Type ./enhancer

```
oscake:mac_enhancer sweetjane$ ./enhancer
```

Genome enhancer will next take a few seconds to read in the genome (fly in this case) and will then print:

```
Welcome to Genome Enhancer 2.0 (beta)
A log of this run will be found in file 10-22-08-46591.txt
Your genome has 118377130 nucleotides
```

Genome enhancer automatically creates a log file for you, with a unique name based on the date and number of seconds that have passed on the day that you started the program.

5. Type in search parameters

The enhancer program will ask you a series of questions, to acquire the same information that was on the web-version:

-The motif sequences that you want to search
-The minimum # of motifs required in a cluster
-Whether you want to impose boolean conditions (optional)
-The maximum window size, in terms of base pairs, in a cluster

In this example, let's look through the Drosophila genome for clusters that contain 3 or more Dorsal sites clustered in 400 bp. This example does not make use of the boolean option—you can see examples of boolean conditions in part 10.

The program will first ask you to type in the motifs to search for. When you are finished, just hit return. In this example, we typed in two motifs, for the Dorsal transcription factor, and then hit the return key when it asked for a third motif. Notice, we used the IUPAC symbol W to specify A or T and M to specify A or C. The program accepts all IUPAC symbols.

We did not specify a Boolean condition in this search, so it looked for clusters containing all combinations of the two motifs we entered.

```
Type in motif A=:GGGWWWWCCM
Type in motif B=:GGGWDWWWCCM
Type in motif C=:

Enter minimum number of motifs required in a cluster: 3
Enter boolean condition (hit return for none):

Enter window size (max #bp in cluster): 400
```

6. Your results are instantly printed!

Your result are printed in two places:

(1) the terminal window where you are running enhancer; and

(2) a text file, automatically generated genome enhancer; in this example, 10-22-08-46591.txt

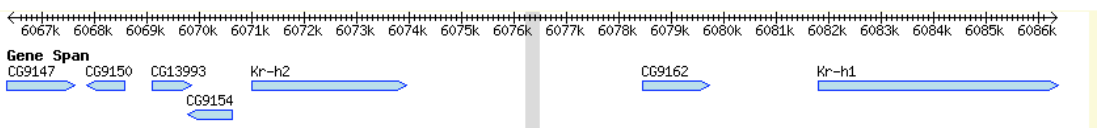
Let's look at the first cluster found.
In the terminal window you see:

```
Cluster 1 length 259 bp
MOTIFS B*B*A* at positions *0,*67,*249
On chromosome 2L at position 6076224 (2L:NT_033779.3)
        2262 bp dnstr of Kr-h2(+),
        1971 bp upstr of CG9162(+)
```

In the file you see the same information, plus the cluster:

```
Cluster 1 length 259 bp
MOTIFS B*B*A* at positions *0,*67,*249
On chromosome 2L at position 6076224 (2L:NT_033779.3)
        2262 bp dnstr of Kr-h2(+),
        1971 bp upstr of CG9162(+)
GGGAATTTCC Ctctagcaaa tatggtttcc actcaaatgc ctttaaatcg gttgaaacgc
gctcaaaGGG AATTTCCCtc caattggatt caatgggcat gacataaccg gctgtgagtc
cgcatgtgcg tgtgtgtgcg tgtgttggag tacaagtgct tccatttatt tatttattta
ttttttttt tggctgtggc cagtgacccc ttggttaacg gtatccatag tattatgtaa
gttgcctggG GGATTTCCC
```

This 259 bp cluster contains motif B at positions 0 and 67 and motif A at position 249 within the cluster. The cluster occurs between the genes Kr-h2 and CG9162, which both occur in the (+) 5'—>3' direction along the chromosome. You can verify this by blasting the cluster against the appropriate genome as shown below, using Blast at Flybase.org:



Cluster 1

7. At the end of the cluster list, genome enhancer tells you the total number of clusters identified by your search. In this case:

```
Clusters found: 15
```

8. The gene list option:
You are given the option to have all the genes associated with the identified clusters printed as a list in the file.

This feature is for systems biology projects. It generates a list of all genes associated with your clusters, which you can then compare with other lists. For Fly Enhancer, the program also prints the synonyms of each gene.

Type **y** to generate the list

```
type y to have all genes (and synonyms) printed as a long list in the file
10-22-08-46591.txt or just hit return: y
```

The list will appear after the clusters in the saved file:

```
Here is a list of all genes and their synonyms mentioned in this run.

Kr-h2
CG9159
Kr-h
CG9159
Kruppel homolog
CG9162
CG9162
eya
CG9554
Eya
cli
CLIFT
EY2-1
CG9554
```

the list continues in the actual file...

8. Decide whether to run the program again or quit

```
Can now search the genome again with the same or different motifs
Last time you searched for:
A: GGGWWWWCCM
B: GGGWDWWWCCM
Hit return to keep the motifs, enter c to Change motifs,
or q to Quit the program:
```

9. Additional information—The IUPAC code:

code	description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

:

10. Additional information—Boolean conditions

If you want specific numbers of each motif to be required in the cluster, use Boolean conditions. Here are some examples:

If you want every cluster to contain
at least 2 instances of motif A,
type **2A**

```
Enter boolean condition (hit return for none):
2A
```

If you want every cluster to contain
at least 2 instances of motif A and 3 instances of motif B,
type **2A and 3B**

```
Enter boolean condition (hit return for none):
2A and 3B
```

If you want every cluster to contain
at least 1 instances of motif A or 5 instances of motif B,
type **1A or 5B**

```
Enter boolean condition (hit return for none):
1A or 5B
```

If you want every cluster to contain
at least 2 instances of motif A and no instances of motif B,
type **2A and not B**

```
Enter boolean condition (hit return for none):
2A and not B
```

Genome Enhancer 2.0
vs. The web-interface for Fly Enhancer

Note: we are working on a web-implementation for Genome Enhancer 2.0.

There are many advantages offered by this line-command version of Genome Enhancer that were not available with the original web-interface. In short they are:
1. you can enter an unlimited number of motifs to search
2. you can see all the results, even when thousands of clusters are returned.
3. for searches in D. melanogaster, you get a list of all the associated genes and their synonyms which is useful when comparing datasets obtained by other means such as by microarray.

There are also important differences between this line-command version of Genome Enhancer and the web-interface:

1. The line-command version does not merge overlapping clusters, but instead counts each cluster distinctly. This means that you may get more clusters reported with the line-command program than you did with the web version.

2. The web version counted sites that overlapped one another whereas the line-command version looks at each nucleotide only once. This means that for sequences that overlap themselves, you'll get fewer spurious hits with the line-command version. For example, if you're looking for the SMAD sequence "GCCG" consider this sequence:

GCCGCCGA

The web version would report two smad sites:
GCCGCCGA
GCC**GCCGA**

The line-command version will report only one smad site—the first one it encounters in the string:

GCCGCCGA