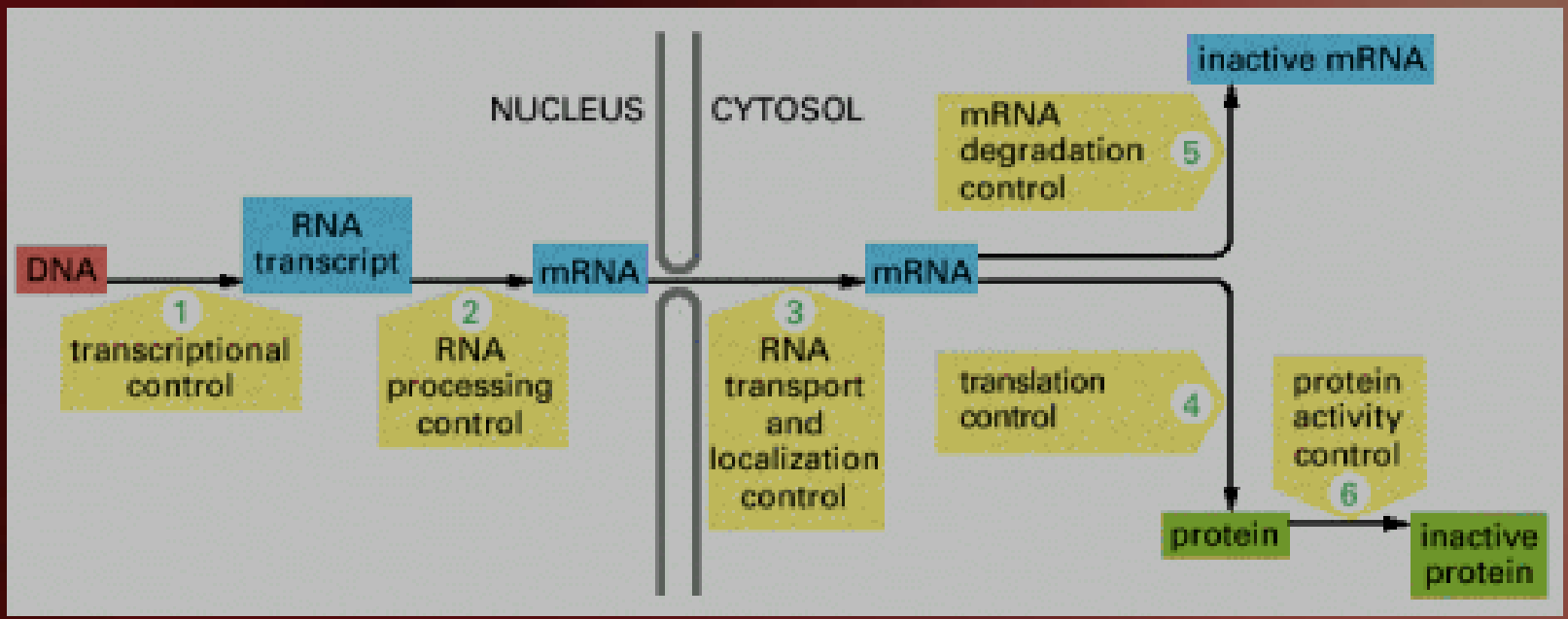
# ORTHOMINE - A dataset of Drosophila core promoters and its analysis

Sumit Middha
Advisor: Dr. Peter Cherbas

- Introduction
- Challenges and Motivation
- D melanogaster Promoter Dataset
  - Expanding promoter sequences
  - Merging the Datasets
- Analysis of D melanogaster Dataset
- Associating Flybase Transcript IDs
- Multiple fly species Dataset
  - Initial results with multiple drosophila Dataset
- Future Work

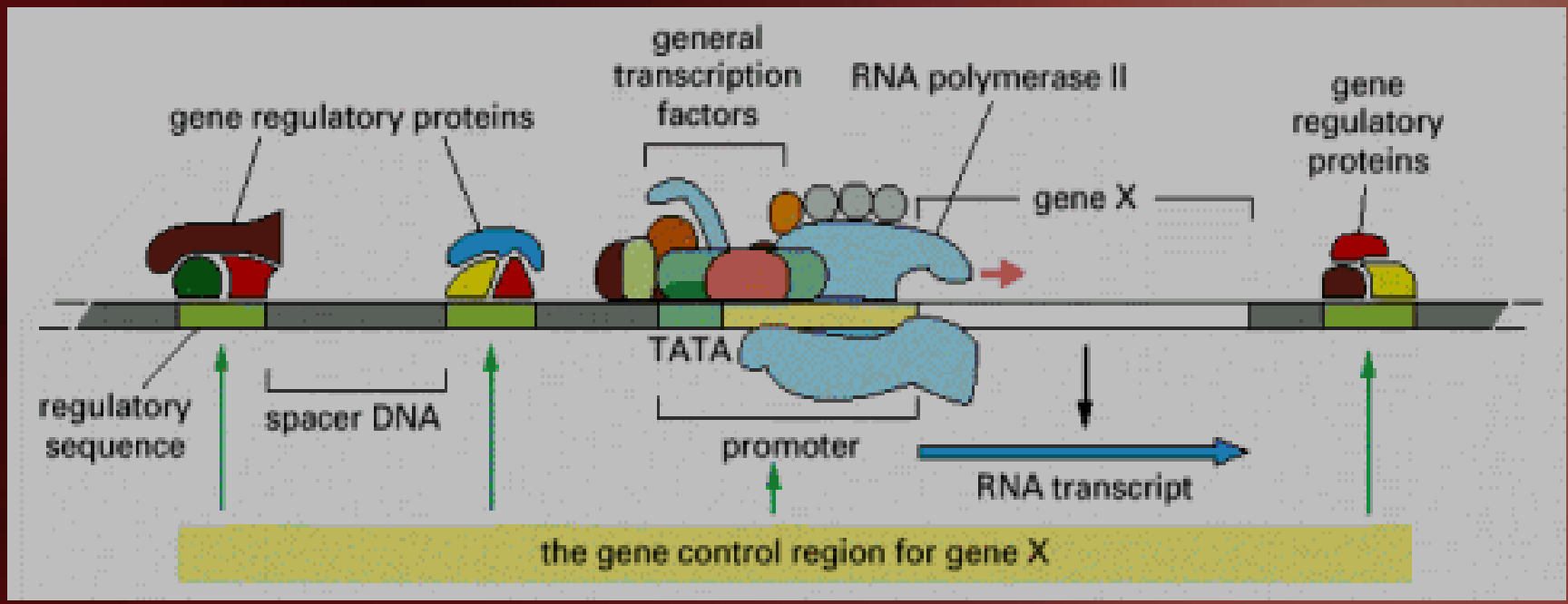
# INTRODUCTION

Stages of gene expression: Single gene multiple products



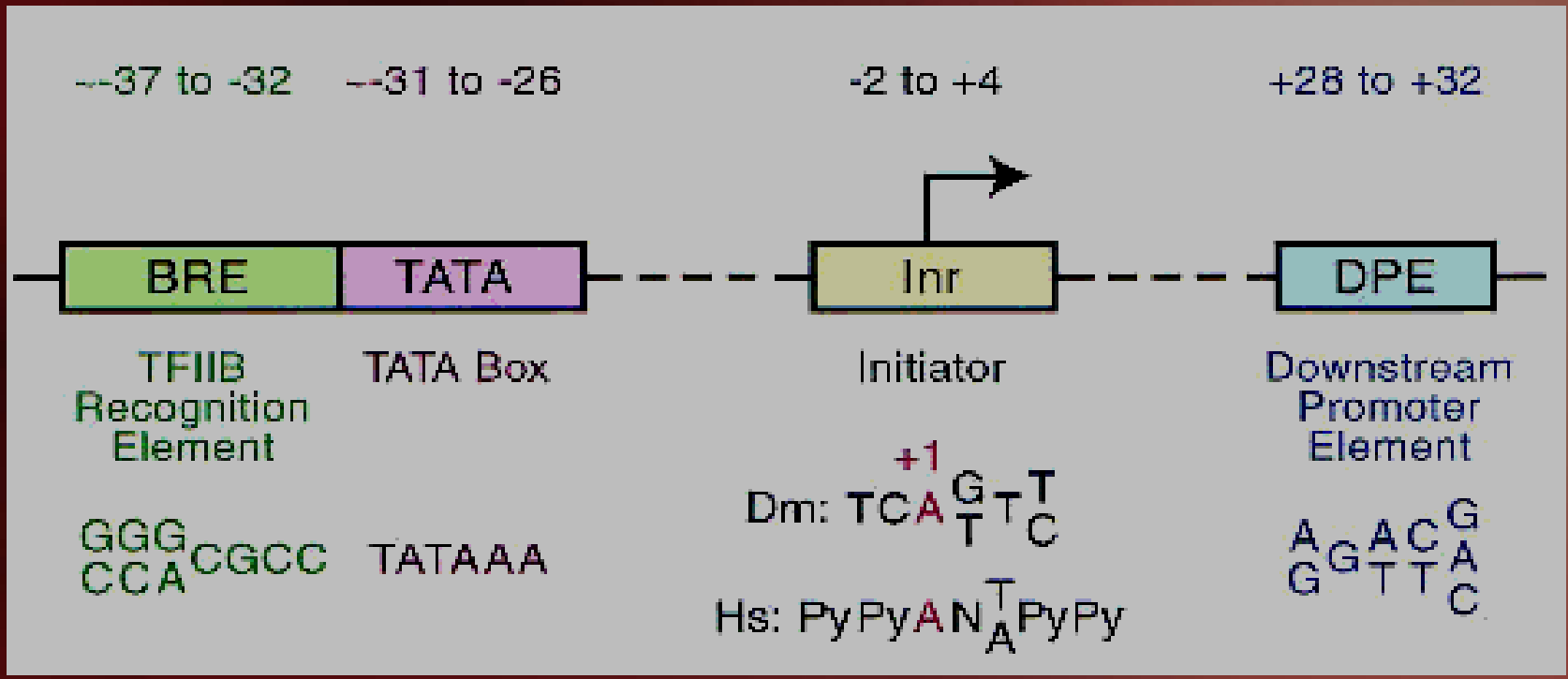
# INTRODUCTION

Transcriptional Control



# INTRODUCTION

## Drosophila Core Promoters




Above image edited from:
http://163.238.8.180/~davis/Bio_327/lectures/Transcription/TranscriptionOver.html

# Challenges and Motivation

- Known regulatory elements are small and degenerate
- Position of TSS is not accurately known
- Many regulatory elements are still unknown

Focus on assembling a dataset of promoter regions around the Transcription Start Site

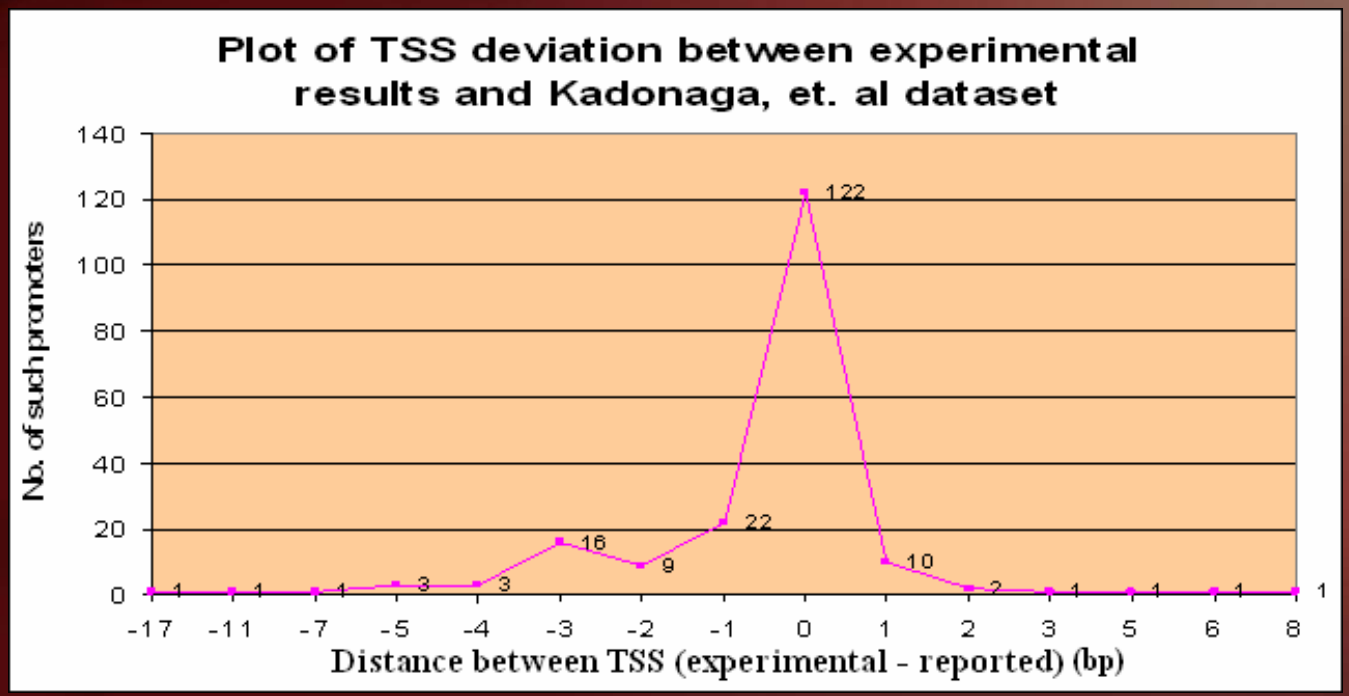
- Promoters for RNA polymerase II genes
- Experimentally determined TSS (+/-10bp)

# The Promoter Datasets

- Rubin, et al: 1941 promoter sequences; -250 to +50
- Kadonaga, et al: 205 promoter sequences; -47 to +45
- EPD: 1926 promoter sequences; -250 to +50

# Kadonaga, et al (UCSD) – 205 promoters

- Proposed adjustment from Inr consensus (*T-C-$A_{+1}$-G/T-T-C/T)* was nullified using original references
- Removed 9 promoters for which we could not confirm the TSS from the original papers



# Expanding promoter sequences

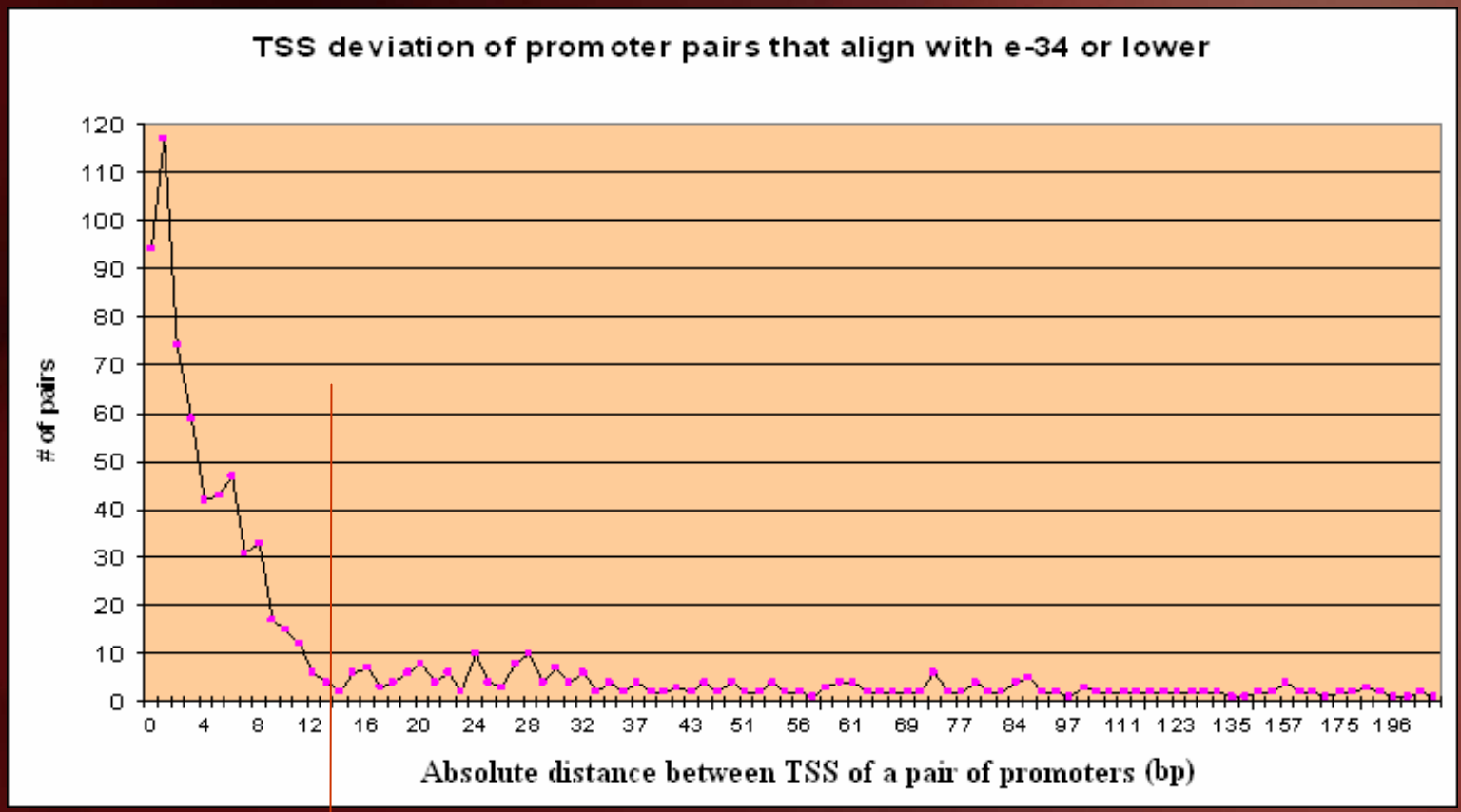
- Total **4060** promoters
- Extract genomic sequence for each promoter ( -250 to +100bp) – uniformity
- 80 Ambiguous/low scoring BLAST hits – removed

http://bioportal.cgb.indiana.edu/orthomine/extend_dataset_cases.htm

# Merging the datasets

- Overlaps – entries in datasets for the same promoter element
- Fasta all pairwise comparison
- E-value cutoff = 1e–34
- Max distance between TSS cutoff = 14bp
- Non Redundant Dataset – **3393** (1908 + 157 + 1328)

# Merging the datasets



- Introduction
- Challenges and Motivation
- D melanogaster Promoter Dataset
  - Expanding promoter sequences
  - Merging the Datasets
- Analysis of D melanogaster Dataset
- Associating Flybase Transcript IDs
- Multiple fly species Dataset
  - Initial results with multiple drosophila Dataset
- Future Work

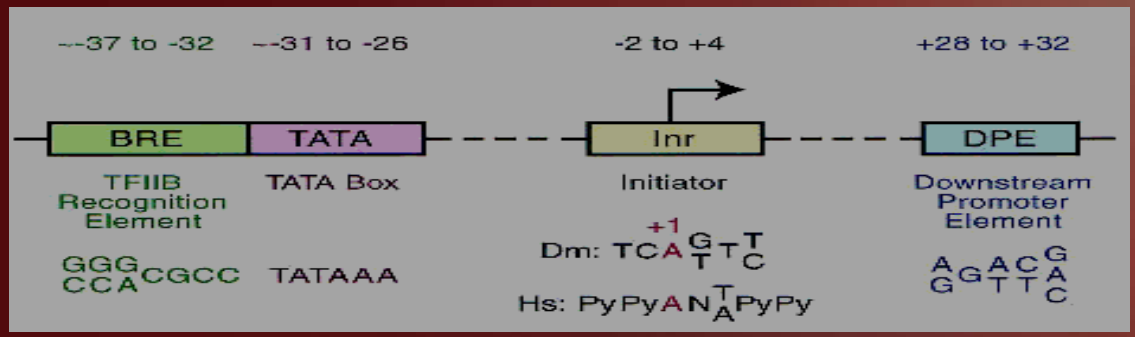
# Dataset Analysis – Word over-representation

- Cannot ignore position preference of regulatory motifs
- An analysis of the Inr-negative set showed over-expression of patterns in the same positions as the Inr, TATA and DPE should be, and could be possible *yet unknown* synonyms

Yogita Mantri
(MS Bioinformatics – IUB)

# Dataset Analysis – GO Clustering

- 1st Level of Biological Processes (7 clusters)
- No strong TCAGT motif in 'behavior' and 'regulation of biological processes' clusters
- Initial Results

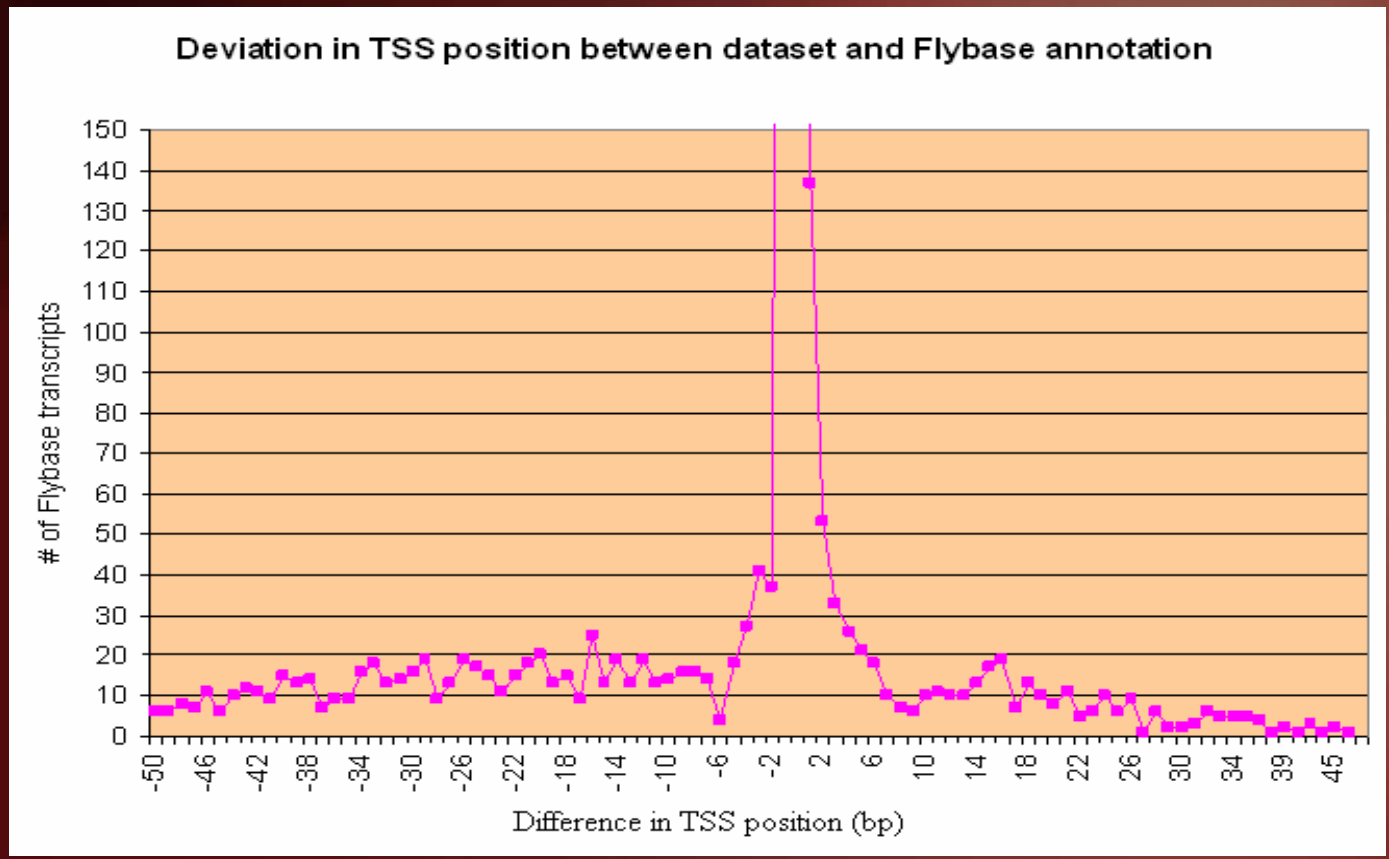


# Flybase Transcript ID – Promoter

- BLAST the non-redundant promoter dataset with the database (D mel chromosomes arms)
- Flybase rel4.1 – TSS coordinate
- TSS of dataset promoter within +/- 50 of TSS from rel4.1 – assign the transcript ID to promoter (2699)

# Flybase Transcript ID – Promoter (contd.)

| | |
|---|---|
| -5 | 18 |
| -4 | 27 |
| -3 | 41 |
| -2 | 37 |
| **-1** | **490** |
| **0** | **961** |
| **1** | **137** |
| 2 | 53 |
| 3 | 33 |
| 4 | 26 |
| 5 | 21 |
| 6 | 18 |

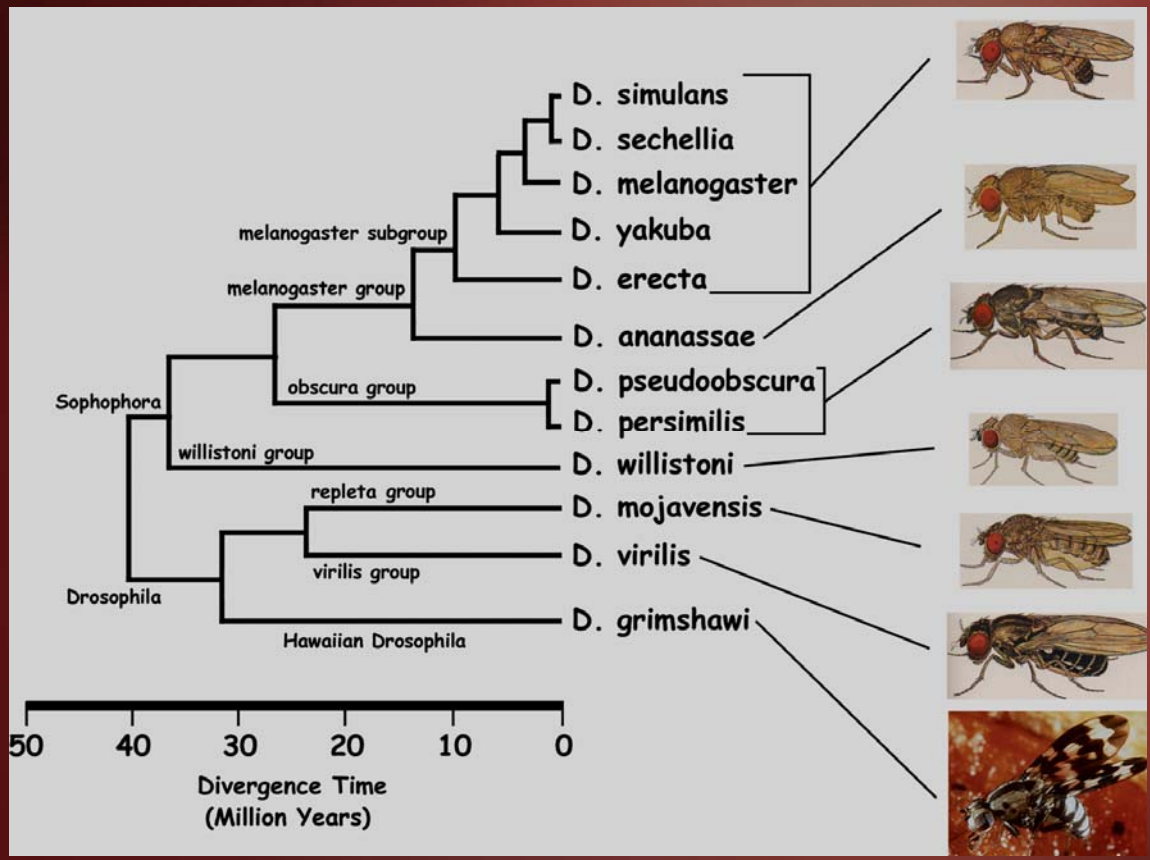


- Introduction
- Challenges and Motivation
- D melanogaster Promoter Dataset
  - Expanding promoter sequences
  - Merging the Datasets
- Analysis of D melanogaster Dataset
- Associating Flybase Transcript IDs
- Multiple fly species Dataset
  - Initial results with multiple drosophila Dataset
- Future Work

# Multiple Fly species Dataset

- D. ananassae
- D. erecta
- D. mojavensis
- D. pseudoobscura
- D. simulans
- D. virilis
- D. yakuba

Image from Flybase
http://bugbane.bio.indiana.edu:7151/



# Methods for obtaining Orthologs

1. Direct BLAST

2. Circular Method

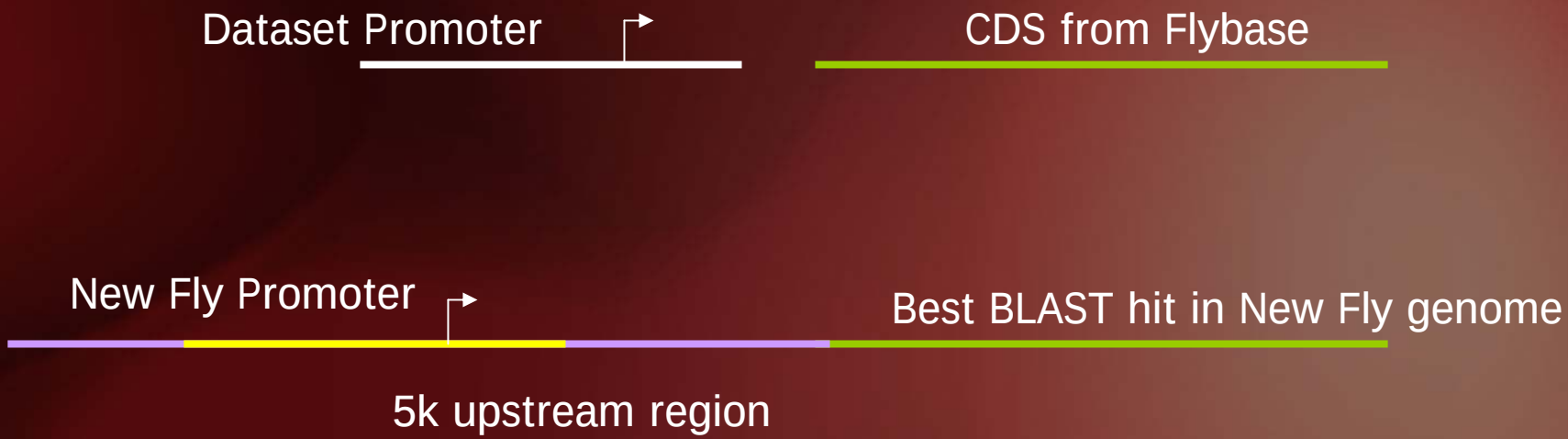
# Method 1 – Direct BLAST

- Short cut method: Simply BLAST the promoter sequence with new-fly database (chr/contig/scaffold)

# Method 2 – Circular Method

- Extract CDS using the Flybase gene ID for promoter
- BLAST CDS with new-fly database (contig/scaffold/chr)
- Take 5kbp region upstream of the hit
- BLAST with original D melanogaster promoter to get orthologous promoter in new-fly (same orientation here)

# Method 2 – Circular Method



# Compare the methods

Two orthologs match if:

- the start coordinates of the BLAST hit are within 50bp
  OR
- BLAST of the orthologous new fly promoters by 2 methods gives e-70 or lower hit (paralog)

# Direct BLAST vs Circular method

**7560 comparisons** of which **6890 matched**

| Dxxx | Matches | No Match |
|---|---|---|
| Dsim | 1651 | 20 |
| Dyak | 2192 | 27 |
| Dere | 2032 | 20 |
| Dann | 556 | 240 |
| Dpse | 291 | 197 |
| Dvir | 117 | 96 |
| Dmoj | 87 | 70 |

No match cases

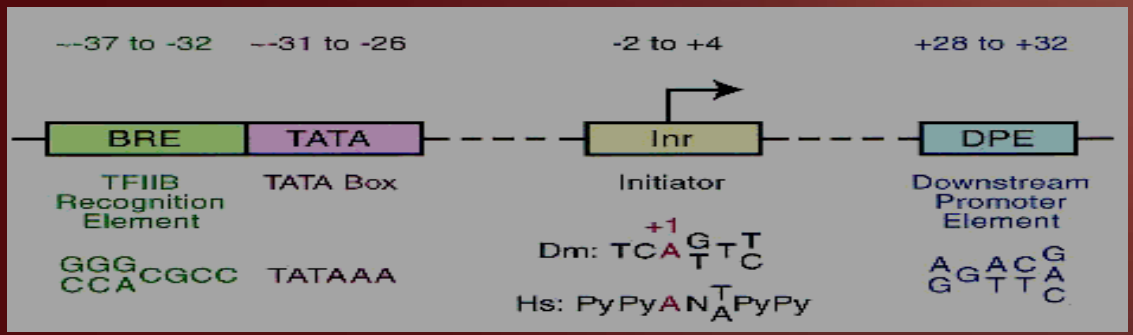
# Dataset Analysis – Multiple Drosophila

MEME motifs: Weak TCAGT and TCGAT (DRE) motifs in the D mojavensis

Ortholog sets motifs: Conserved Blocks

Clustalw analysis: High conservation in INR & TATA regions; conserved 5mers

GO + Clustal: TCATT is more conserved than TCAGT in 'reg of biological processes' and 'unknown function' genes



# Future Work

- More analysis of Datasets to draw further conclusions
- Further analysis of MEME and CLUSTAL results

# Acknowledgements


- Dr. Peter Cherbas
- Dr. Haixu Tang
- Dr. Sun Kim
- CGB
- The discussion groups

# References


- *Kutach, A.K. and Kadonaga, J.T,* 2000. "The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters", **Mol. Cell. Biol**. 20: 4754-4764

- *Ohler, U., Liao, G., Niemann, H., and Rubin, G.M,* 2002 "Computational analysis of core promoters in the Drosophila genome" **Genome Biol.** 3: research0087.1–research0087.12

- *Cherbas, L., and P. Cherbas.* 1993, "The arthropod initiator: the capsite consensus plays an important role in transcription", **Insect Biochem. Mol. Biol**. 23:81-90

- *Christoph D. Schmid, Viviane Praz, Mauro Delorenzi, Rouaïda Périer, Philipp Bucher, "*The Eukaryotic Promoter Database EPD: the impact of in silico primer extension", 82-85

# Questions

# Basic Statistics of Dataset (Extra)

1187550 basepairs in 3393 sequences

AT-content 59.05%, GC-content 40.95%

Mono nucleotides

| | % |
|---|---|
| **A** | 29.70% |
| **C** | 20.67% |
| **G** | 20.28% |
| **T** | 29.35% |

Di nucleotides

| | | second nucleotide | | | |
|---|---|---|---|---|---|
| | | **A** | **C** | **G** | **T** |
| **first nucl.** | **A** | 10.83% | 5.21% | 5.41% | 8.16% |
| | **C** | 6.45% | 4.08% | 4.97% | 5.10% |
| | **G** | 5.31% | 5.73% | 3.71% | 5.46% |
| | **T** | 7.02% | 5.59% | 6.14% | 10.53% |

# Ohler, Rubin, et al (Berkley) – 1941 promoters (Extra)

- aligning 5' expressed sequence tags (ESTs) from cap-trapped cDNA libraries to the genome
- stringent criteria concerning coverage and 5'-end distribution
- cautious approach for identifying TSSs - require 5' ends of the alignments of multiple, independent cap-selected cDNAs to lie in close proximity

  http://genomebiology.com/2002/3/12/research/0087

# Kadonaga, et al (UCSD) – 205 promoters (Extra)

- Drosophila Core promoter database
- TSS mapped by nuclease protection, primer extension, or multiple 5' RACE clones
- If reported start site overlaps a consensus Inr element, the central ‘A’ in the Inr consensus (T-C-A+1-G/T-T-C/T) was designated as the TSS

  http://mcb.asm.org/cgi/content/full/20/13/4754

  http://www-biology.ucsd.edu/labs/Kadonaga/DCPD.html

# The Eukaryotic Promoter Database 1926 promoters (Extra)

- Transcription by RNA polymerase II (protein coding genes)
- Assignment of TSS – experimental data (precision of +/- 5 bp or higher)
- Proposed adjustment from consensus – Ignored
- Minor discrepancies b/w publications – averaged
- Alternative promoters for same gene – min. 20bp apart

http://www.epd.isb-sib.ch/

# Multiple flies – AXT Alignments (Extra)

- UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html)
- BLASTZ alignments (Miller's lab PSU)
- AXT file

  0 chr19 3001012 3001075 chr11 70568380 70568443 – 3500

  TCAGCTCATAAATCACCTCCTGCCACAAGCCTGGCCTGGTCCCA

  TCTGTTCATAAACCACCTGCCATGACAAGCCTGGCCTGTTCCCAA
- Issue – Gapped aligment