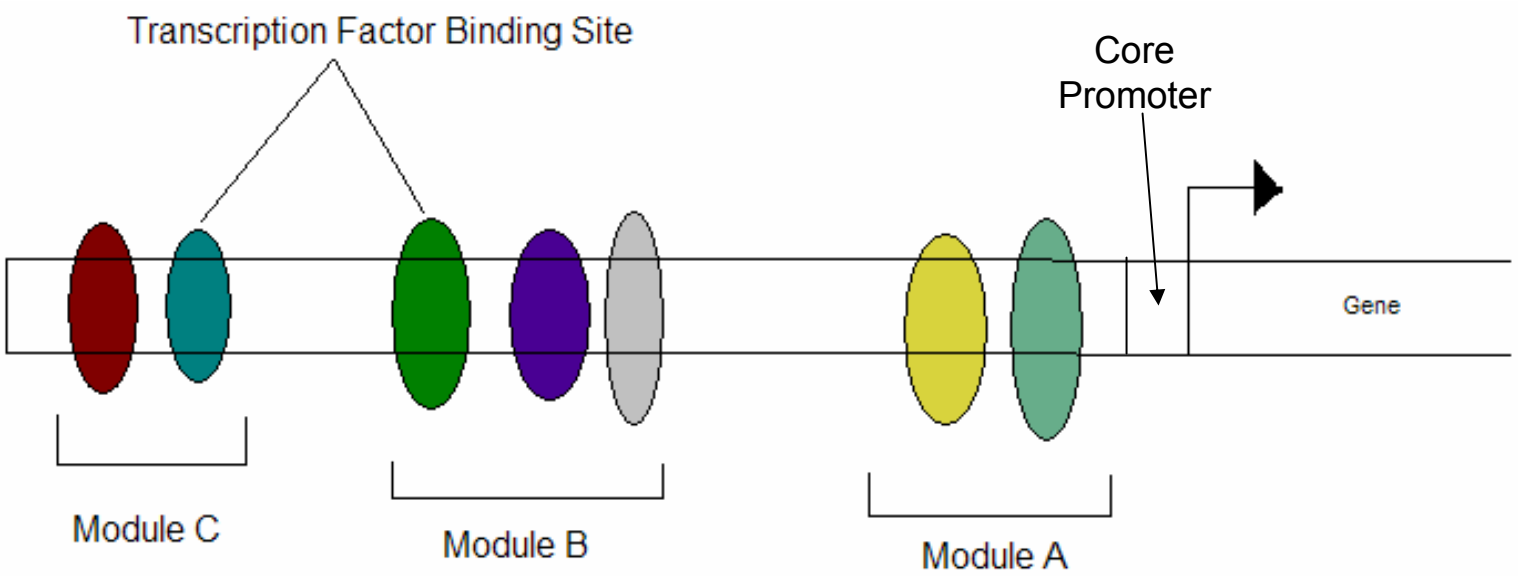
Evolutionary analysis of the well characterized *endo16* promoter reveals substantial variation within functional sites

Paper by: James P. Balhoff and Gregory A. Wray
Presentation by: Stephanie Lucas
Reviewed by: Marie Wilkening and Chi Zheng

Cis-Regulatory System

- Responsible for the regulation of the gene by enhancers and repressors
- Core Promoter- site at which the transcriptional machinery (RNA polymerase) binds
- Transcription Factor Binding Sites- areas in which protein (TFs) bind to regulate transcription
- Module- a fragment of the regulatory system that generates a part of the overall regulatory function- consists of more than one TFBS and the sequence between them



Here's the Problem....

- Changes within the regulatory region will affect gene expression which in turn is believed to affect the evolution of developmental and structural traits
- We do not understand the evolutionary processes that cause variation within these regions of genes
  - Since there's no genetic code for these sequences, it's hard to predict the consequences of a change in the nucleotide sequence

The ideal sequence: *endo16*

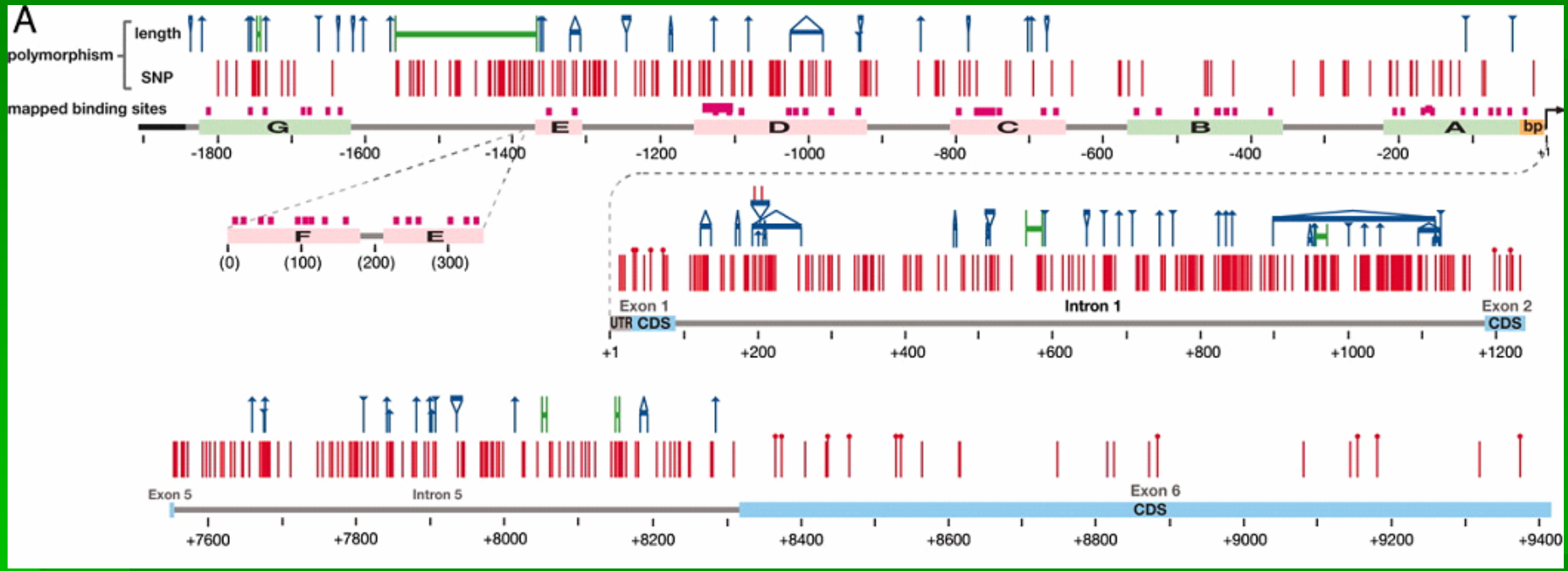
Endo16 is a gene found in *S. purpuratus* (purple sea urchin)

It is an extracellular protein believed to be involved in cell adhesion.

- Many of the binding sites have been mapped and many of the cis-regulation mechanisms are understood
- 56 TFBS (modules A and B) that are necessary for transcription to occur
- Six modules (A,B,C,D,“E-F”,G)

Distal modules

Proximal modules



Endo16 promoter and coding sequence until exon 6

Methods

- PCR/ cloning
- Sequence DNA
- Alignments done with CLUSTALX
  - Sectioned into subalignments for each module, TFBS, and non-functional sites
- Calculate nucleotide differences with DNASP

Overall levels of polymorphisms between different areas of a gene

- Expected levels of polymorphisms
  - Introns- highest
  - Cis-regulatory regions- intermediate
  - Exons- lowest
- Why?
  - Introns are believed to have no function- variation is more accepted
  - Cis-regulatory regions
    - some important areas that are necessary for transcription to occur- more conserved
    - less important areas- variation is more tolerated
  - Exons code for functional mRNA- variation is not tolerated due to change in potential protein structure and function

Table 1: Nucleotide diversity and fixed differences

Sequence partition	πper site
Entire promoter	0.040
All modules	0.041
All intermodules	0.039
All b.s.	0.049
All non-b.s.	0.037
All non-GCF1 b.s.	0.044
All GCF1 b.s.	0.062
Exon 1	0.009
Exon 2	0.029
Exon 6	0.006
Intron 1	0.028
Intron 5	0.060

- Generally, levels of polymorphism within *endo16* are consistent with the expected results.

π= average number of nucleotide differences between sequences

b.s.- binding sites
**S. purpuratus* vs. *S. droebachiensis*
┼Number of nucleotides excluding indels in population data

Observed levels of single nucleotide polymorphisms- promoter region only

- > 250 SNPs in the entire promoter
- Within the promoter region, modules B, C, D, and G surprisingly exhibit higher levels of polymorphisms within the binding sites compared to the non-binding sites.

Table 1: Nucleotide diversity and fixed differences in each module

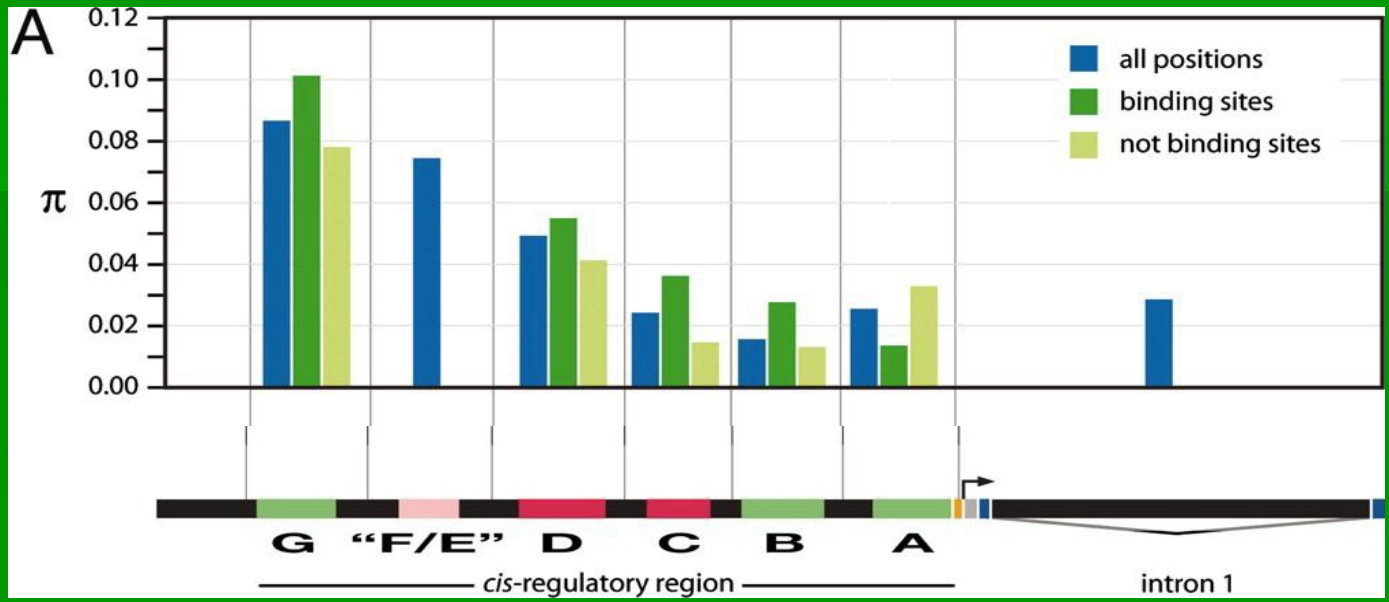
Sequence partition	π per site	θ per site	Fixed differences per site*	Length†
Module A	0.026	0.023	0.041	184
Module A b.s.	0.014	0.017	0.030	64
Module A non-b.s.	0.033	0.027	0.048	120
Module B	0.016	0.020	0.047	213
Module B b.s.	0.028	0.026	0.060	54
Module B non-b.s.	0.012	0.018	0.043	159
Module C	0.024	0.029	0.070	159
Module C b.s.	0.036	0.037	0.088	66
Module C non-b.s.	0.015	0.023	0.057	93
Module D	0.050	0.053	0.070	227
Module D b.s.	0.064	0.066	0.087	86
Module D non-b.s.	0.041	0.045	0.061	141
FE region	0.075	0.059	0.204	54
Module G	0.086	0.072	0.091	197
Module G b.s.	0.110	0.103	0.102	48
Module G non-b.s.	0.078	0.062	0.088	149

b.s.- binding sites
*_S. purpuratus_* vs. _S. droebachiensis_
┼Number of nucleotides excluding indels in population data

------With the exception of module A, b.s. are more polymorphic than non-b.s.

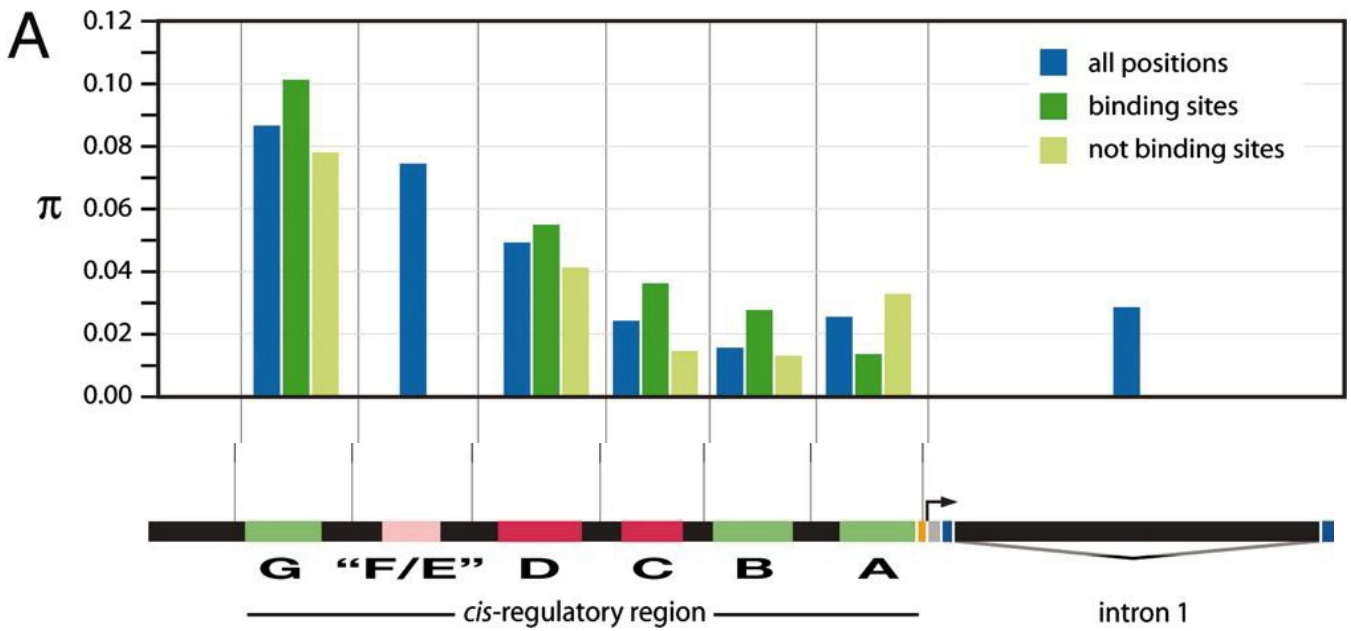
Observed levels of polymorphisms in promoter region

- Module A had lower levels of single nucleotide polymorphisms compared to non-binding sites in the module (expected)
- Conclusion: Module A is most conserved-
  - Function: to integrate the effects of all the other modules, making it very important.



Single Nucleotide Variation

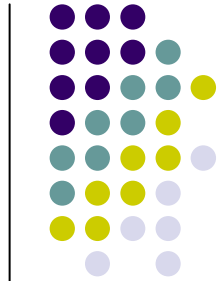
- Modules A-C: $\pi \leq 0.026$, distal regions have more than double this value, where π is the average fraction of nucleotide differences between sequences in each module
- Conclusion: proximal half of the cis-regulatory region under greater selective constraint than distal region



Comparison to close relative

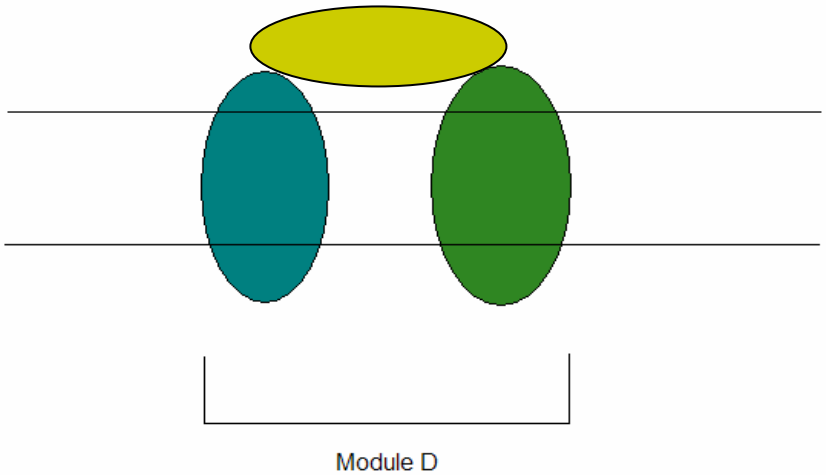
- This analysis was compared to *S. droebachiensis*, a close relative of the purple sea urchin (green).
- The conclusions drawn before were confirmed.

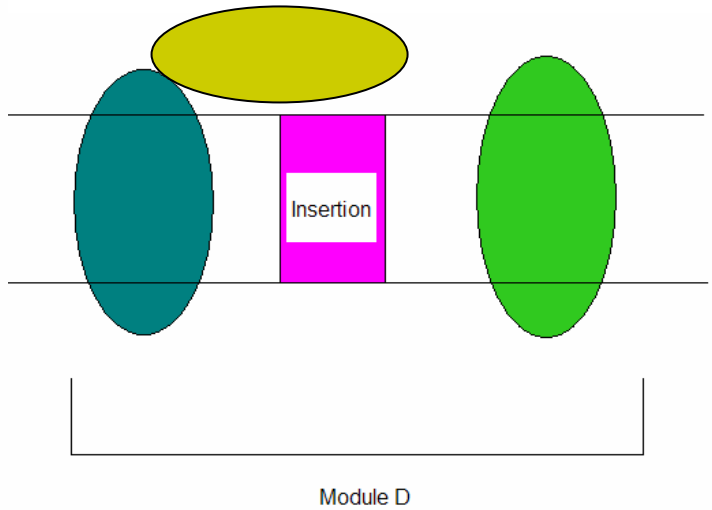
Indel (Length) Variation



- Indel- insertions/ deletions (can be more than on base pair)---it is common in promoter regions
- The longer the indel is in the promoter, the more disruptive it is to local protein interactions
- Developed a way to calculate the length variation so that they could compare its effects with other indels
  - This method weighs each indel differently depending one the size of the insertion/ deletion

Hypothesis for indel variation

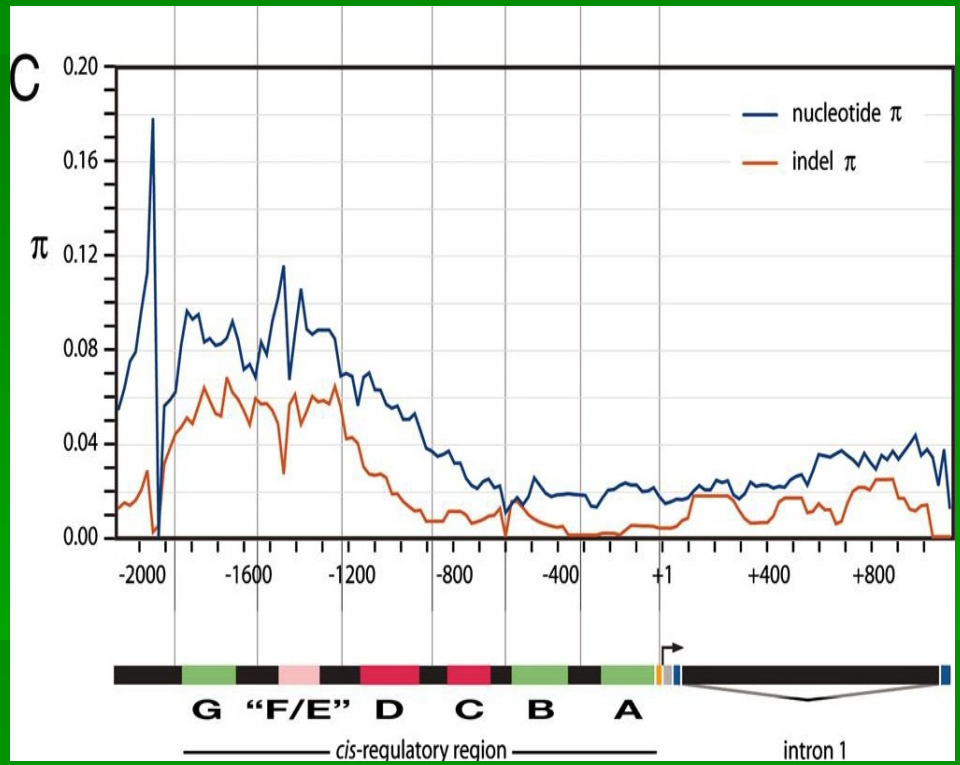




- Indels are most likely to be found in between modules rather than in between the TFBS within the module
- This is because the nucleotide sequence between the TFBS in a module may affect the binding of the TF

Observed indel variation

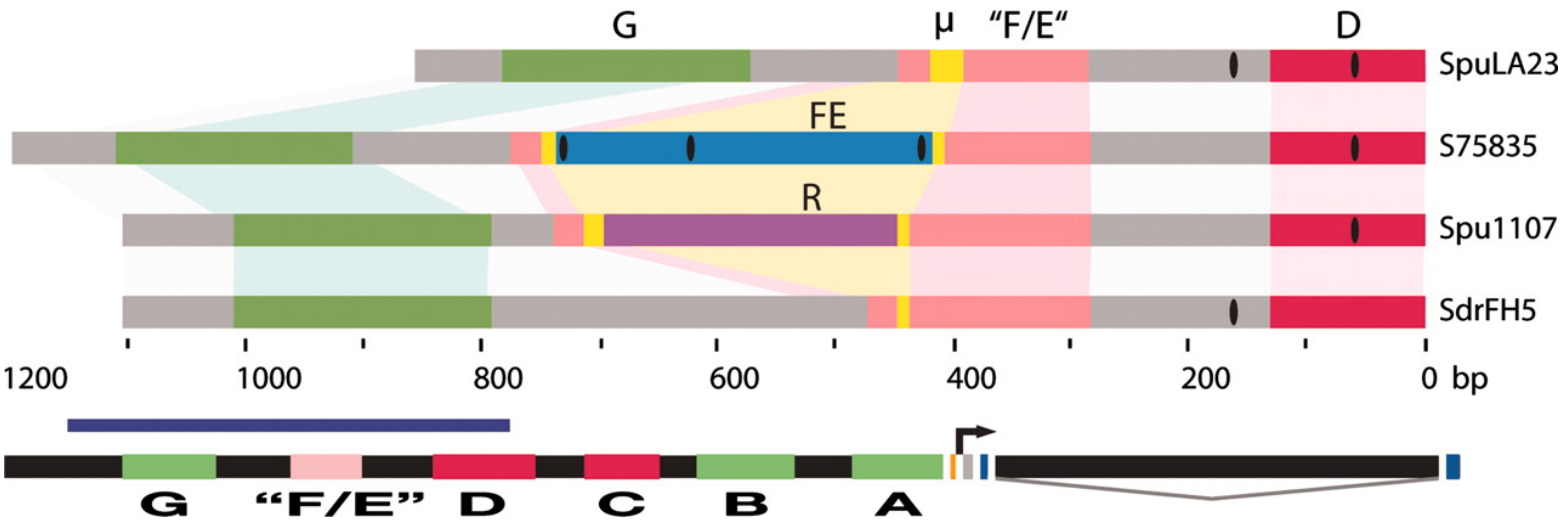
- > 40 indels in promoter
- Indels range from 1 to 340 bps
- Follows single nucleotide polymorphism pattern
- Distal modules are more variable than proximal modules- respond similarly to constraint placed on local sequences (like SNPs)



x-axis: bps (intervals of 300 for indels, and 30 for SNPs)

Inserted sequences

- Out of 70 individuals (140 alleles), 2 alleles were much longer than the other samples
- One was an unrelated sequence (Spu1107)
- One was similar to the F and E modules that aren't found in most *endo16* sequences (S75835)
  - 16 verified protein binding sites have been found in this inserted region of over 300 bps
  - Insertion of entire functional module



F and E modules

- Ectodermal repressors
  - Most binding sites are located in this inserted area
- Question- How is *endo16* translation repressed in many of the samples?
- Their lack of presence in most of the samples suggest that other sequences of the regulatory region have this function
  - 2 binding sites near the D module may be involved in ectodermal repression

Discussion and Conclusions

- Unexpectedly, within the promoter sequence, TFBS are more polymorphic than non-coding regions within a module
  - Due to physical binding of the protein, some nucleotides within the TFBS are more important to the binding of the protein
  - In order to compensate for a SNP in the TFBS, another SNP may be necessary (2:1)
- Module A is under selective constraint due to its importance in transcription
  - Function to act as “integrator of input”---

Discussion and Conclusions Cont'd

- Different modules within the promoter sequence have different levels of polymorphisms- proximal modules are more conserved than distal ones
  - Length polymorphisms at proximal modules may be caused by background selection (since it's so close to the gene)
- Modules E and F are the result of a recent insertion and are rare in *S. purpuratus*
  - Whole functional module inserted---this shows how evolution can dramatically affect the promoter region

References

- Arnone, Maria I.; and Davidson, Eric H. "The hardwiring of development: organization and function of genomic regulatory systems". Development 124 (1997): 1851-1864.
- Balhoff, James P. and Wray, Gregory A. "Evolutionary analysis of the well characterized *endo16* promoter reveals substantial variation within functional sites". PNAS. 102:24 (2005) 8591-8596.
- Miyata, Takashi; Yasunaga, Teruo; and Nishida, Toshiro. "Nucleotide sequence divergence and functional constraint in mRNA evolution". Genetics 17:12 (1980): 7328-7332.
- Yuh, Chiou-Hwa; Bolouri, Hamid; and Davidson, Eric H. "*Cis*-regulatory logic in the *endo16* gene: switching from a specification to a differentiation mode of control". Development 128 (2001) 617-629.