Investigating the Functional Association of Genes Sharing Cis-Acting Motifs in Maize Using Bioinformatics and Gene Ontology Analysis

Mary Youngberg, Becca Sayad, Gabriela Limeres and Dr. Hank W. Bass

FSU Department of Biological Science

FSU UNDERGRADUATE RESEARCH OPPORTUNITY PROGRAM
CENTER FOR UNDERGRADUATE RESEARCH & ACADEMIC ENGAGEMENT

Abstract:

The research hypothesis for our experiment is that genes that share cis-regulatory elements also share a biological function. We have hundreds of cis-regulatory elements (short DNA motifs) in our model genetic organism, maize (*Zea mays*), but none of them are assigned a biological function. To test our hypothesis, we carried out the following steps: (1) intersected the DNA motif family with genes to make a list of genes for each motif and (2) used the gene list to conduct a Gene Ontology (GO) analysis with the AgriGO database. Gene ontologies, such as "development" or "disease response" indicate genetic pathways or biological functions. This research is important because the motifs may identify genes that respond to similar signals and environmental cues. Specifically, we first chose a motif family with at least 1000 locations (mode of actions peaks of motifs). We obtained the genomic locations for these motifs as BED files. We used the bioinformatics program, DeepTools Intersect, to generate a list of genes with each motif family. From this preliminary GO analysis, we saw that many of the motif-specific genes were enriched for certain pathways. For instance, for the genes intersecting DNA motif family "dym63", we observed enrichment for GO categories: PROCESS *cellular response to stimulus* (GO:0051716, p-value 2.40E-45), PROCESS *reproductive process* (GO:0022414, p-value 1.30E-40), and PROCESS *organelle organization* (GO: 0006996, p-value 3.20E-35). This indicates that dym63 regulates genes in these pathways. Among them, *reproductive process* is consistent with the motifs having originated from developing earshoot, a reproductive tissue. Some GO categories appear to be detected too frequently. We are developing an independent test for significance to identify what might be false positives from the GO analysis. The results of this study will help us understand which gene motifs regulate which biological pathways. Knowing this can inform genetic strategies for crop improvement.

Introduction:

- Genes often contain cis-acting motifs in their regulatory regions that play a crucial role in coordinating gene expression.
- These motifs serve as binding sites for transcription factors, which regulate gene activity in response to various signals.
- Genes involved in shared biological pathways and processes frequently share common motifs, suggesting a functional relationship.

Challenges and Limitations:

- Technical challenges:
  - Converting file formats can be manual and time-consuming, increasing the risk of errors.
- Biological limitations:
  - While motif databases exist, they do not always provide direct evidence of functional relationships between genes sharing motifs.

This research lays the groundwork for future studies on cis-regulatory factors in gene expression, enhancing our understanding of genetic regulation in maize and other plant species.

Intersection Procedure:

Collected two source/input BED files:

DNA motif locations

om001.bed-om140.bed and dym01.bed-dym75.bed

Gene list

ZmB73v5_Genes300.bed

Shell Script Code URL

1. Used AI to build a "shell script" `ListUniqueGeneIDsPerMoMo.sh` (which does the following):
   a. intersect the genes with the motifs
   b. convert the intersected genes output to gene names only
   c. Removed duplicated gene names
   d. Make a gene list file for the input motif (e.g.`dym63_interset_geneIDs_namesOnly_unique_1200.txt`) [*1200 indicates that 1200 genes were on the list]
2. Output from # 1 is the input we put into AgriGo:
   a. Input Gene Lists into AgriGo [Figure 2]
   b. Copy results table back to Google Sheet to log the "hits" [Figure 3]

Figure 1: Flowchart of the Bioinformatic intersections to generate gene lists from motifs

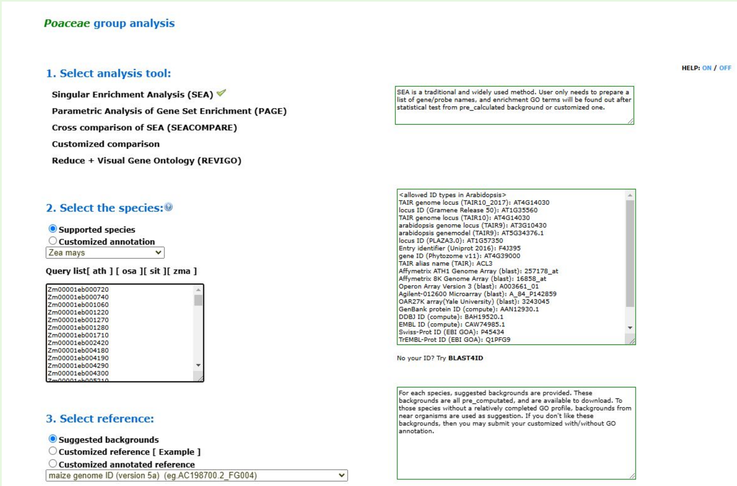
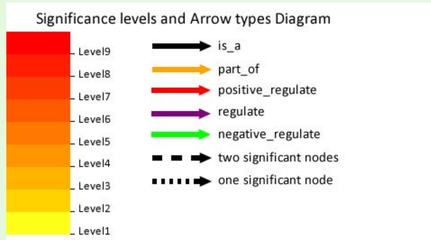


Figure 2: Displays the inputs of AgriGo. Utilizing a singular enrichment analysis, with Zea mays as our species, we input the output from #1 into the query list and used maize genome ID (version 5a) and our background. After running through AgriGo, we obtained a table [Figure 2] of the results.

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0009987	P	cellular process	742	12319	1.80E-58	5.00E-55
GO:0051716	P	cellular response to stimulus	78	204	2.40E-45	3.30E-42
GO:0042221	P	response to chemical stimulus	104	445	1.60E-43	1.50E-40
GO:0022414	P	reproductive process	53	79	1.30E-40	8.80E-38
GO:0044237	P	cellular metabolic process	591	9649	3.30E-40	1.80E-37
GO:0016043	P	Cellular component organization	131	869	1.30E-36	5.80E-34
GO:0006996	P	organelle organization	92	440	3.20E-35	1.20E-32
GO:0048519	P	negative regulation of biological process	35	39	2.80E-30	9.40E-28

Figure 3: Displays the table we received on AgriGo of shared biological functions between the motifs and gene list. It is sorted by GO term and p-value. Due to our false positives, the range of interest for us included any GO term that had a p-value above 30.



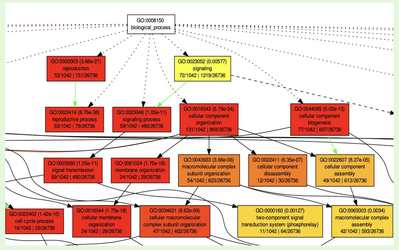


Figure 4 and 5: Significance chart of significance diagrams (from AgriGO) and significance chart of dym63

Conclusions and Future Directions:

Completed:

- Generated 215 GeneLists
- tested ~20 with AgriGo with motifs that had over 1,000 genes.
- obtained enriched GO terms for dym63 and several others.

Conclusions:

- different motif families were associated with different lists of genes
- GO analysis revealed a lot of shared enriched ontologies

Future Directions:

- Test for detection of "false positives" from the GO enrichment
  - False positives are common to every gene list
  - We will mine all the GO output tables for uncommon GO terms
- Identify other GO enrichment analysis website for maize

References:

MOA-seq: Savadel, S.D., et al. (2021) *PLoS genetics*, 17(8), e1009689.

DeepTools: Ramírez, F., et al. (2016) *Nucleic Acids Research*.

AgriGO: Tian, T., et al. (2017)., *Nucleic Acids Res*. 45(W1):W122–W129.

Resources:

- http://www.genomaize.org/ genome browser with motif locations and genes
- https://www.maizegdb.org/ maize genome data repository
- https://you.com/?chatMode=default combination of LLMs used to build shell scripts