A genome-wide cis-regulatory landscape in *Drosophila*

Charles Blatti[1] (blatti@illinois.edu), Majid Kazemian[1] (akazemi2@illinois.edu),
Susan Celniker[2], Scot Wolfe[3], Michael Brodsky[3], Saurabh Sinha[1]

1. MOTIVATIONS

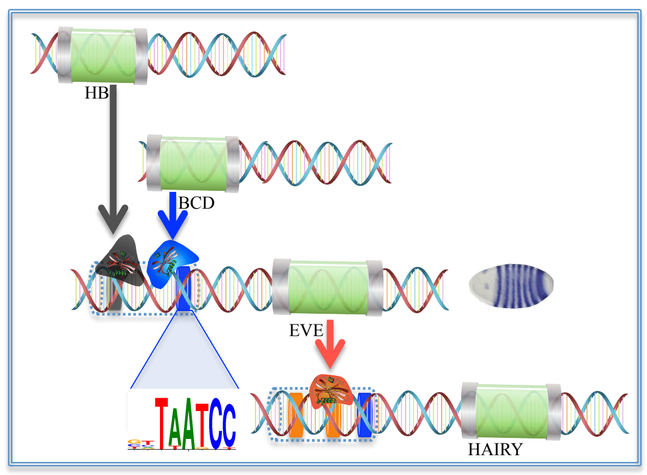
We can characterize a specific regulatory network using

- binding profiles of relevant transcription factors (TFs)
- expression pattern of the genes of the network
  - e.g. A/P segmentation in *Drosophila*

We have extensive databases of ~300 TF binding specificities and ~6000 gene expression patterns.

The goal is to leverage these two resources and annotate the regulatory landscape of many *Drosophila* developmental networks.

mod ENCODE & BDTNP* provide potential regulatory regions across various time points, but do not describe their network context.



* Thomas S, et al. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome Biol. 2011;12:R43.

2. COMPUTATIONAL BINDING PROFILES

Rather than limit ourselves to ~35 TFs with experimental binding profiles, we sought to computationally predict TF occupancy using the binding specificities of ~300 factors.

Approach	Average Correlation
STUBB	0.22
MS_STUBB	0.20
DHS	0.44
STUBB+DHS	0.51
MS_STUBB+DHS	0.31

We found the best substitute for TF binding profiles is scores generated by an HMM filtered by experimental Dnase I-hypersensitive site data (STUBB+DHS).

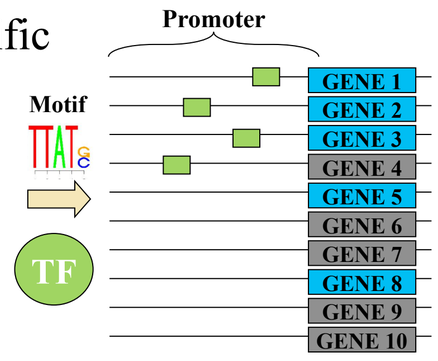
- Conservation information did not improve correlation.

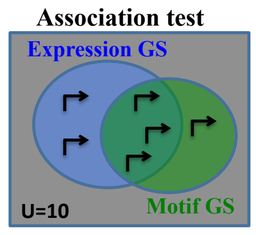
Our evaluation was based on the average correlation between the scores of the computational method and ChIP profiles for 2000 windows balanced between high affinity ChIP bound regions and background.

3. PREDICTING PRIMARY REGULATORS

To find primary factors that regulate specific expression in embryonic tissue/cell types:

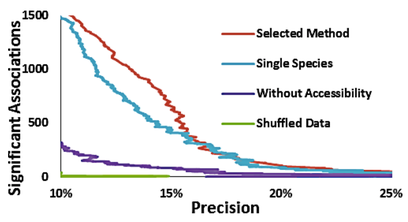
- ~300 TFs, ~200 expression terms
- find motif gene sets using computational profiles
- quantify associations using hyper-geometric test
- select associations where the regulating TF is annotated with the expression term ("consistent expression")





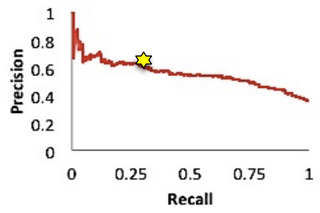
Evaluate on consistent expression

MS_STUBB+DHS is best



Evaluate on ChIP profiles

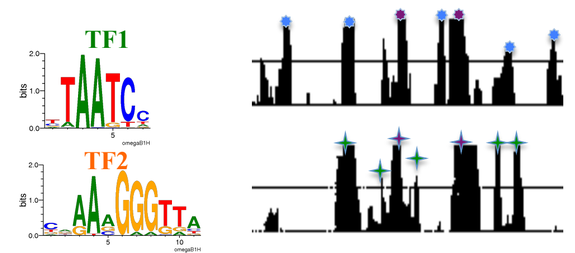
64% precision & 27% recall

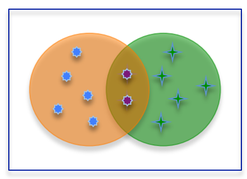


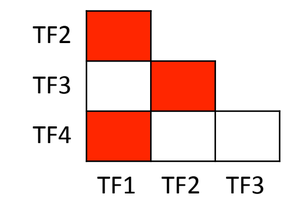
4. PREDICTING CO-ACTING REGULATORS

Gene expression in eukaryotes is often regulated by combinations of transcription factors.

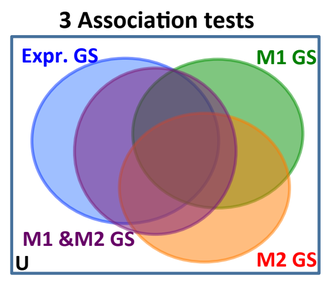
We identified "co-localized" pairs of TFs whose accessible top motif binding locations significantly overlap, suggesting possible physical interaction or combinatorial action.







We recover ~1600 associations of gene sets with co-localized TF pairs that are stronger than the associations with either individual TF.



5. MAP OF TISSUE TYPE DETERMINANTS

Overall, we found ~2700 interesting associations between our ~300 motifs (determinants) and ~200 expression terms.

- Average of 21 motifs per term and 10 terms per motif
- ~700 associations using a strict threshold on p-value
- ~1400 with relaxed threshold of 0.005 but requiring consistent TF expression
- ~650 associations from co-acting regulator analysis

We found many associations involving the central nervous system, which is known to have a high degree of cell type diversity.

Expression Pattern	#Motifs
ventral nerve cord primordium	110
embryonic brain	107
brain primordium	96
ventral nerve cord	87
cellular blastoderm	84
…	…
plasmatocytes anlage	1
embryonic cuprophilic cell	1
embryonic proventriculus intermediate layer	1

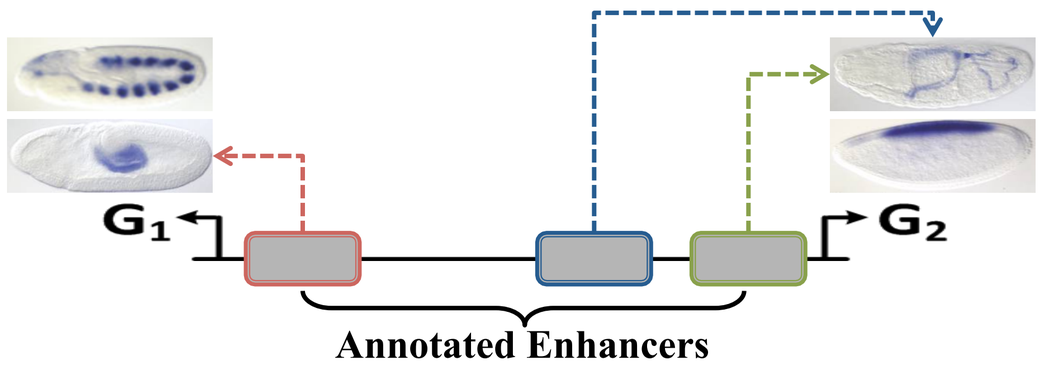
Motif	#EGS
Trl_FlyReg	83
z_FlyReg	48
Blimp-1_SANGER_5	45
Adf1_SANGER_5	44
toy_FlyReg	38
…	…
nau_SANGER_5	1
lola_SOLEXA_5	1
CG15601_SANGER_5	1

We selected TF-gene interactions from a subset of ~140 associations for experimental validation using TF mutations and RNAi knockdowns.

6. REGULATORY GENOME ANNOTATION

Given gene expression patterns, motif-expression term associations, and TF-gene interactions, the goal is

- to annotate candidate enhancers with the expression pattern of the gene that they regulate



To achieve this goal we:

- use DHS to identify potential enhancers
- find the motifs harbored by the enhancer
- assign one of the neighboring genes and its proper expression pattern using the motif scores

1- University of Illinois, Urbana, IL, 61801
2- Lawrence Berkeley National Laboratory, Berkeley, CA 94720
3- University of Massachusetts Medical School, Worcester, MA 01605