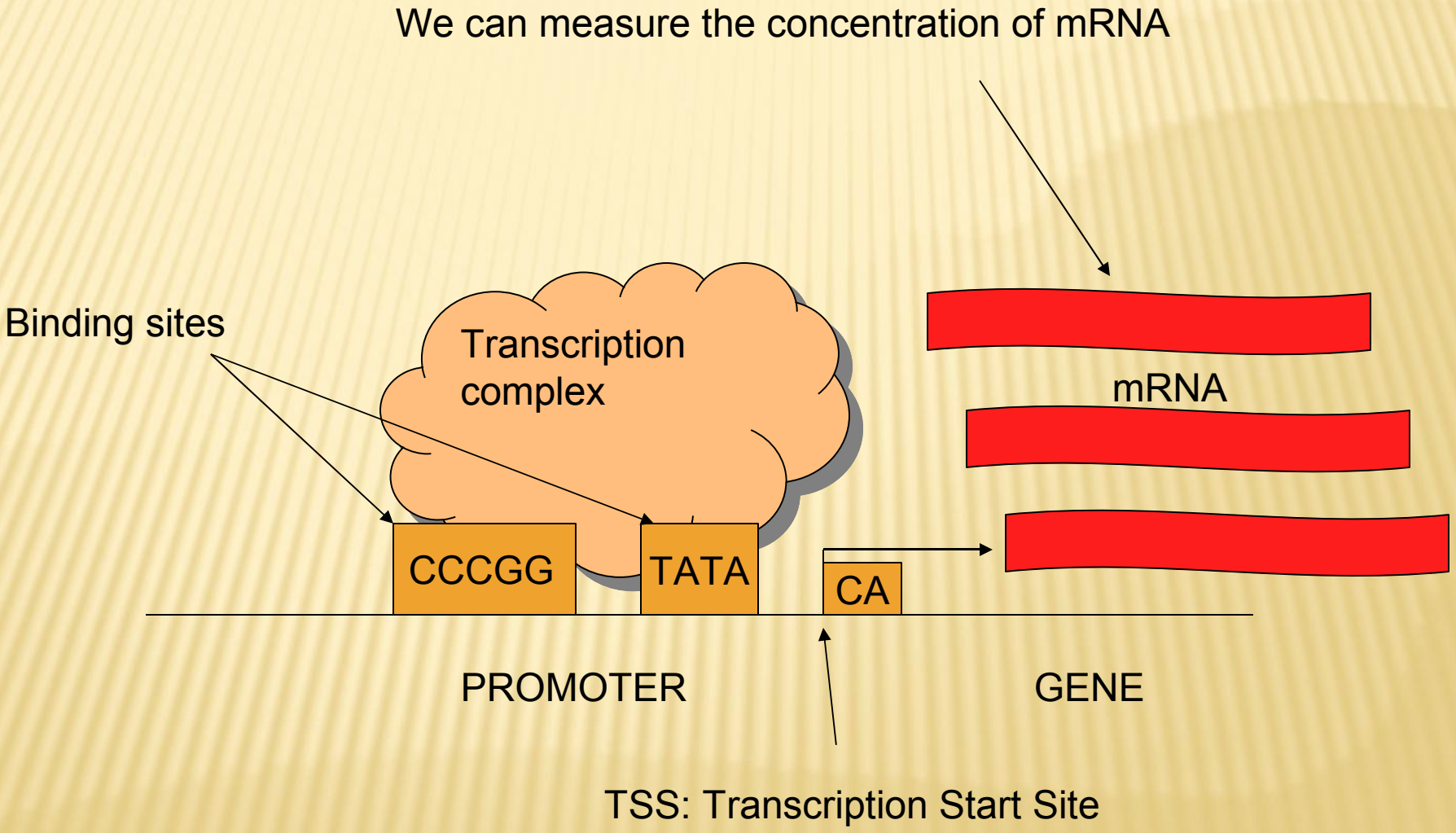
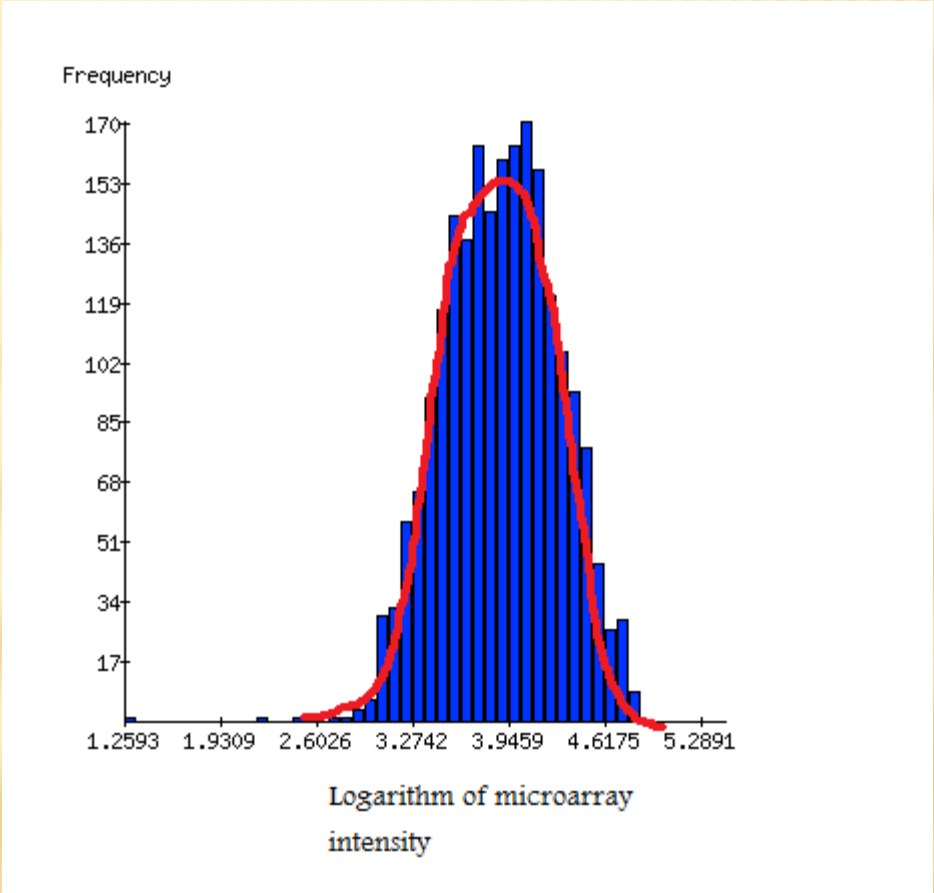
Part II

# FINDING REGULATORY ELEMENTS

# What can we measure?



# HOW DOES THE DISTRIBUTION OF INTENSITIES LOOK?



# COMBINE EXPRESSION INFORMATION WITH INFORMATION ABOUT BINDING SITES

To assess significance of influence of each motif in a given position, we compute its Z-score as:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

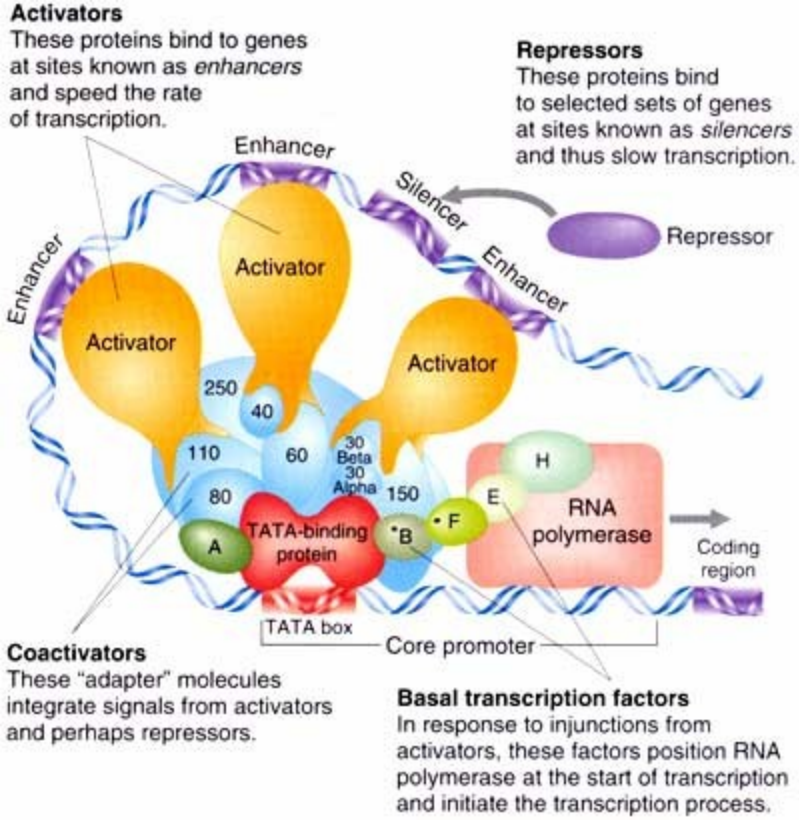
$\bar{X}$ is the average expression of genes with the motif
$\sigma$ is the standard deviation of expression intensity the population
$\mu$ is the mean expression intensity of the population

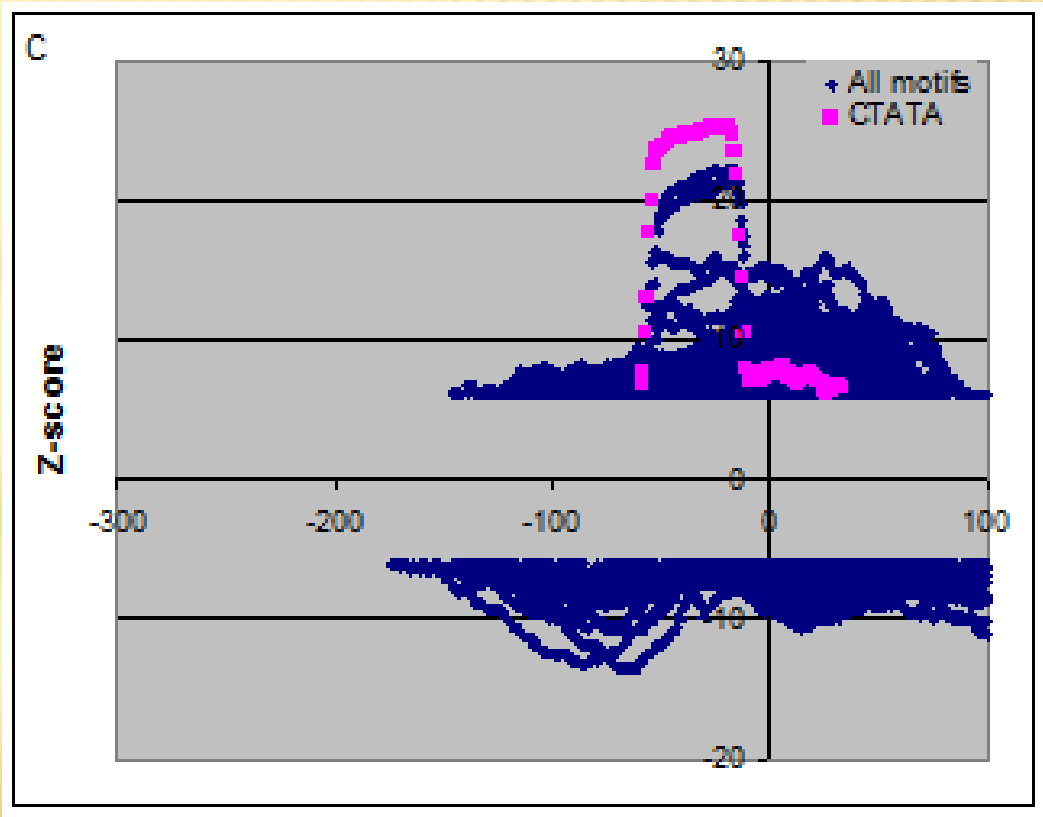
Motifer/MotiferGOLD

# COMPONENTS OF SUCCESS

- Determine the amount of mRNA
  - Gene expression experiments
  - Publicly available collections of expression data grows exponentially
  - Need to manually sift through the data to group similar experiments
- Find the promoter region
  - Multiple transcripts per gene
  - Position – specific motifs

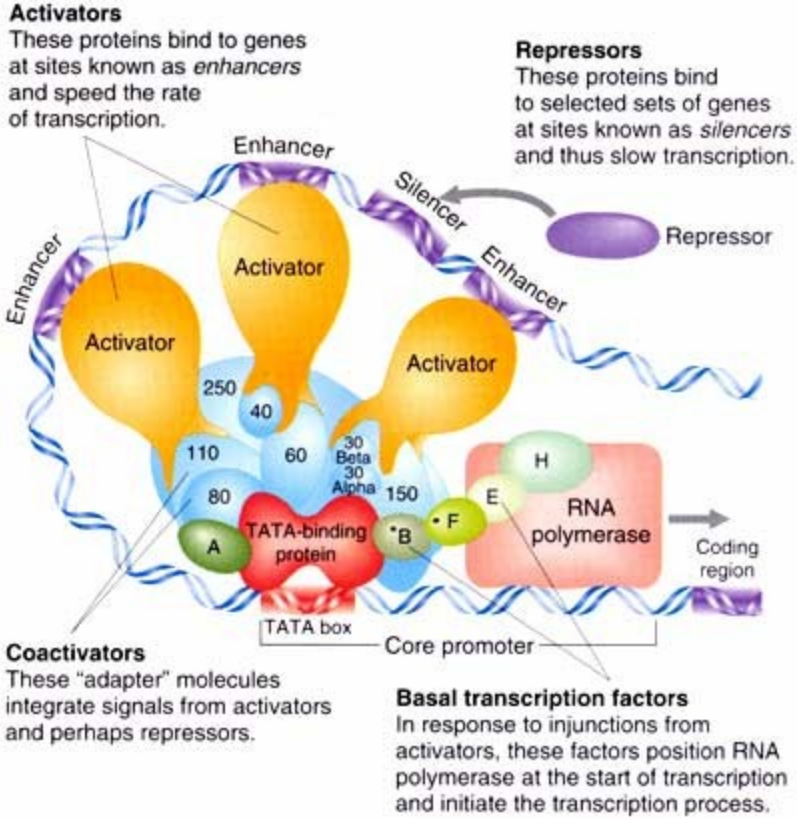
# SOME RESULTS: TATA-BOX

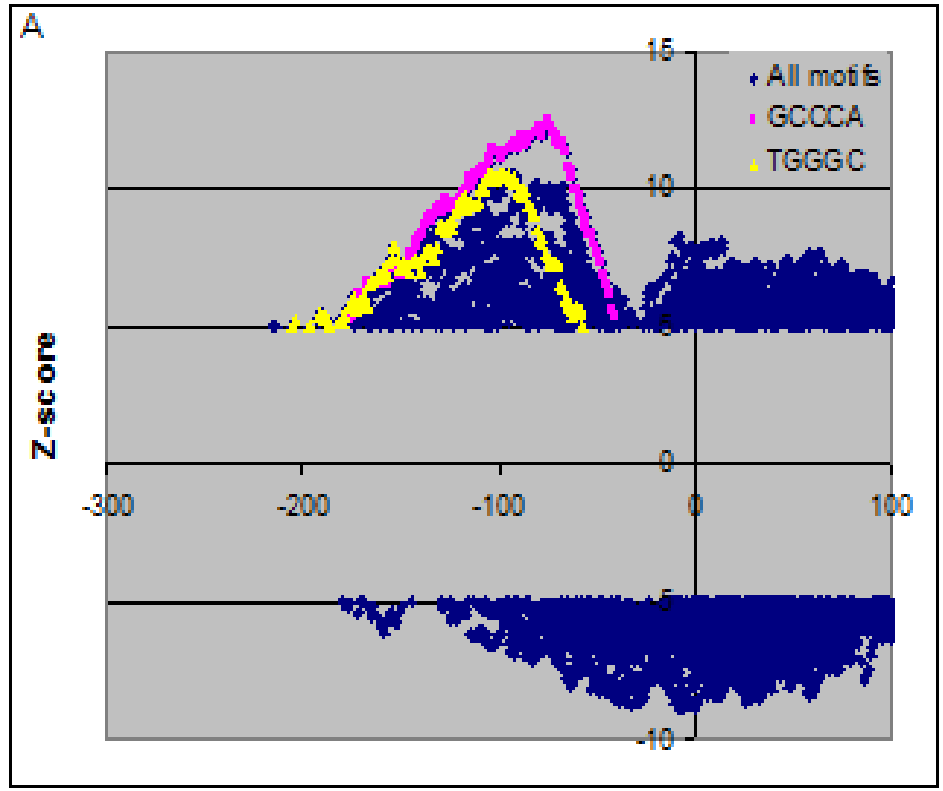



Expression variability is a standard deviation of log intensities across different tissues and treatments

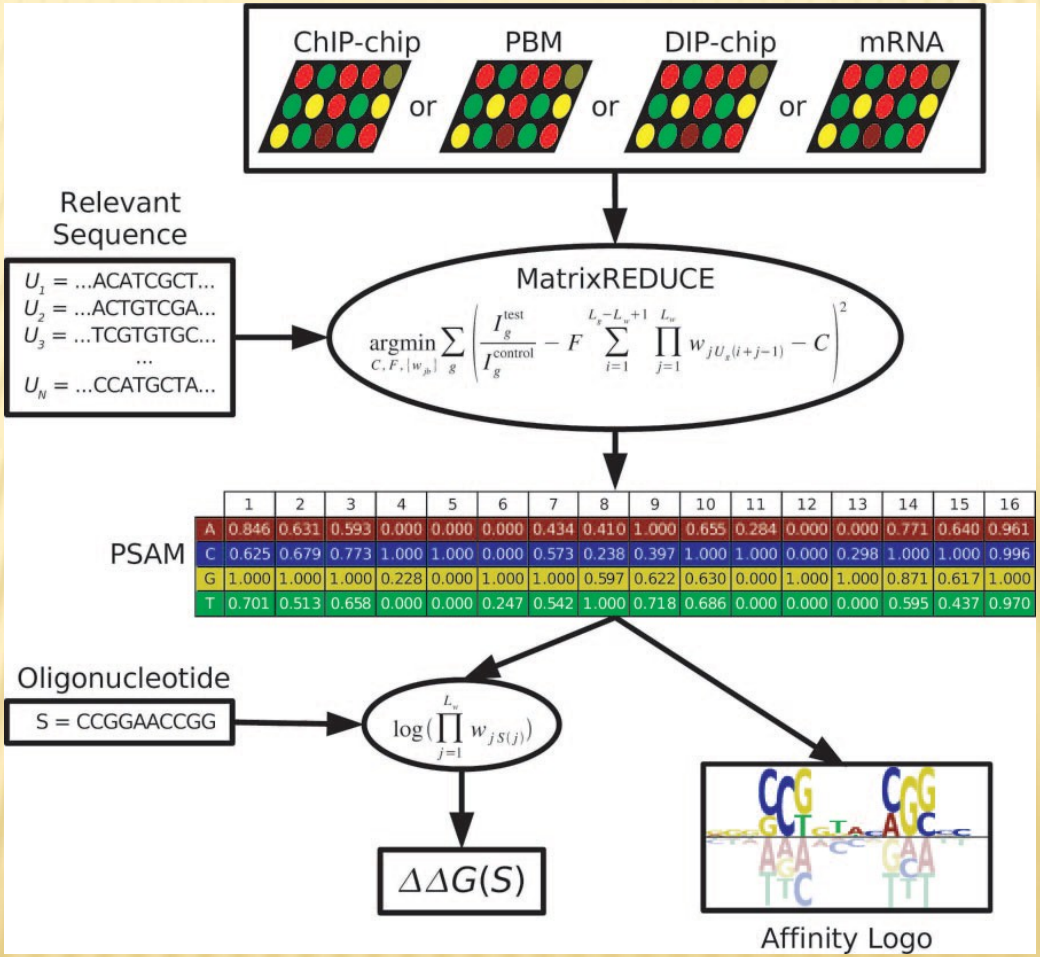
# SOME RESULTS: ENHANCERS




Expression strength is an average log intensity across different tissues/treatments

# REDUCE/MATXRIXREDUCE

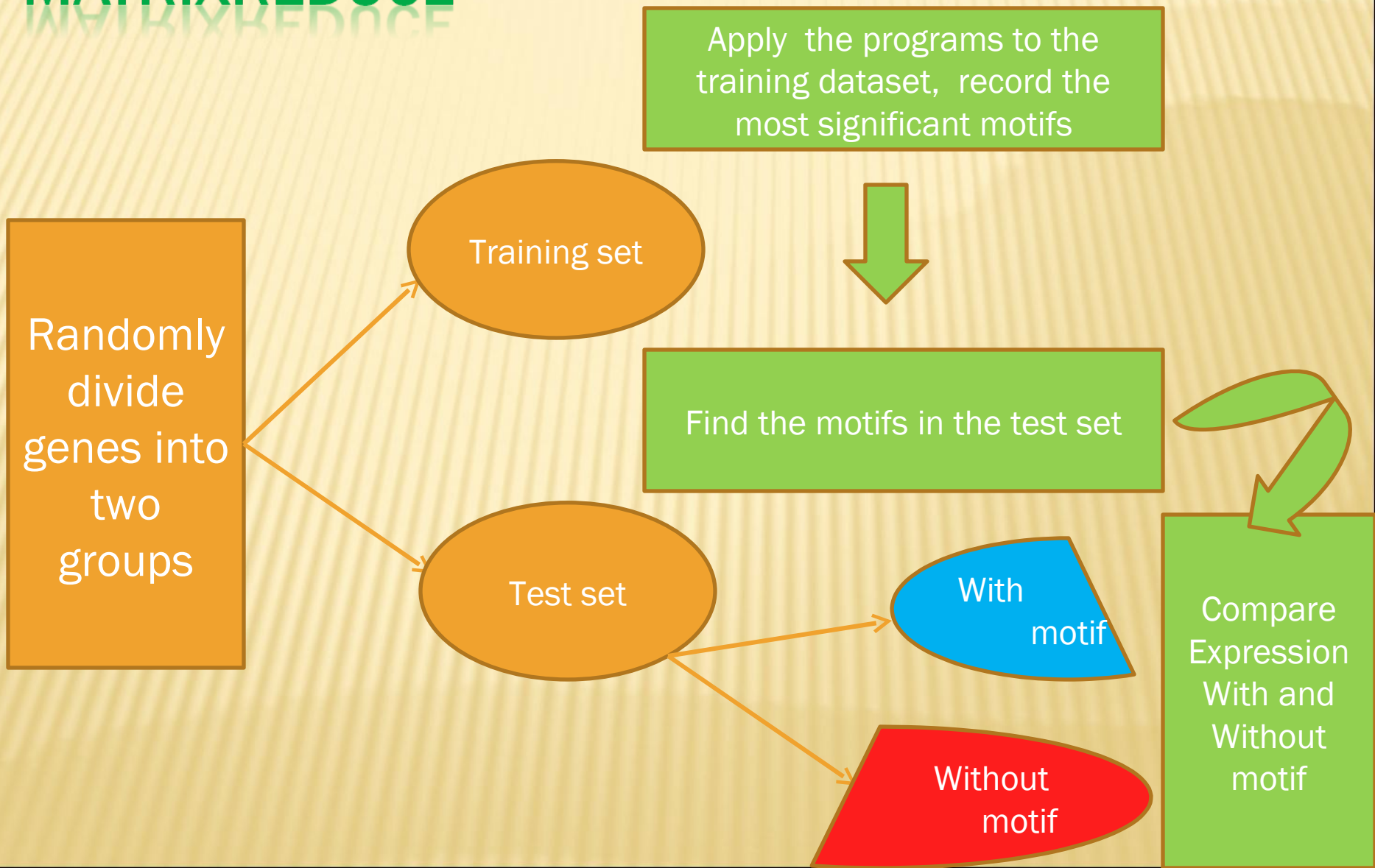



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.846 | 0.631 | 0.593 | 0.000 | 0.000 | 0.000 | 0.434 | 0.410 | 1.000 | 0.655 | 0.284 | 0.000 | 0.000 | 0.771 | 0.640 | 0.961 |
| C | 0.625 | 0.679 | 0.773 | 1.000 | 1.000 | 0.000 | 0.573 | 0.238 | 0.397 | 1.000 | 1.000 | 0.000 | 0.298 | 1.000 | 1.000 | 0.996 |
| G | 1.000 | 1.000 | 1.000 | 0.228 | 0.000 | 1.000 | 1.000 | 0.597 | 0.622 | 0.630 | 0.000 | 1.000 | 1.000 | 0.871 | 0.617 | 1.000 |
| T | 0.701 | 0.513 | 0.658 | 0.000 | 0.000 | 0.247 | 0.542 | 1.000 | 0.718 | 0.686 | 0.000 | 0.000 | 0.000 | 0.595 | 0.437 | 0.970 |

MatrixREDUCE does not take into account positions of motifs relative to transcription start site.

Float et al. 2006

# BENCHMARKING: COMPARISON OF MOTIFER TO MATRIXREDUCE

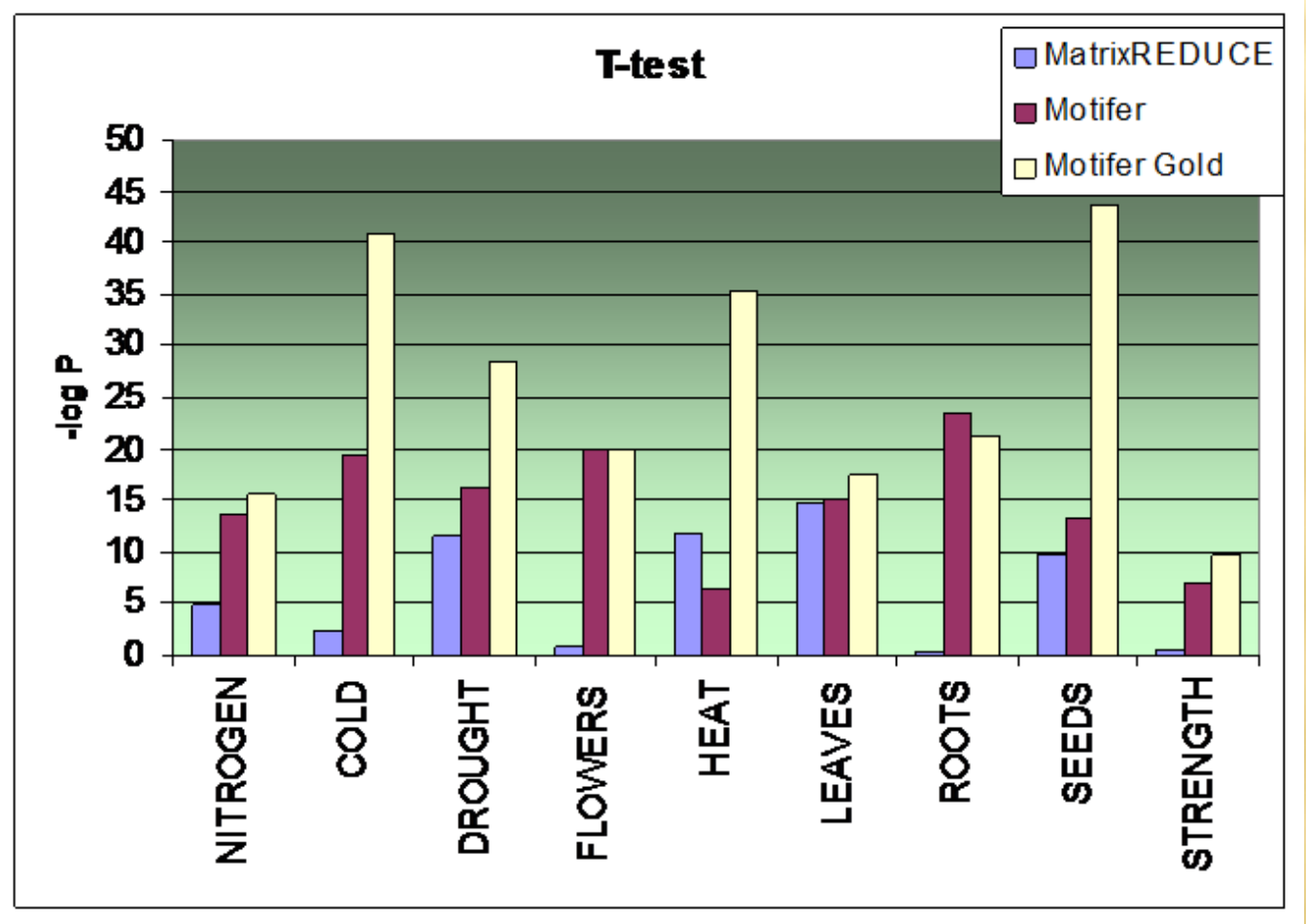


# COMPARISON: T-TEST

We test the null hypothesis that the mean expression of genes that contain the motif is equal to the mean expression of genes without the motif.

$$t_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } \sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

# RESULTS




Performance of MotiferGold is consistently better than MatrixReduce

# FUTURE PLANS

- Improve TSSer
  - Add additional sources of data
    - Known binding sites
    - Distribution of lengths of untranslated region
    - Tiling array data
    - Homology data
  - Incorporate alternative TSS, splice variants
- Develop pipeline for functional genome annotation

# CONCLUSIONS

- Simple statistical methods can lead to discoveries
- More sophisticated methods may underperform in case of noisy datasets

# REFERENCES


- [Molina05]: Molina, C. and Grotewold, E. (2005), Genome wide analysis of Arabidopsis core promoters, BMC Genomics 6(1): 25.
- [Nature 07] The ENCODE Project Consortium, (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Vol. 447| 14, June 2007| doi:10.1038/nature05874
- [Ohler02] U. Ohler et.al., (2002) Computational analysis of core promoters in Drosophila genome, Genome Biology, Vol.3,N. 12.
- [Berendzen06]: Berendzen, K. et al., (2006) *Cis*-motifs upstream of the transcription and translation initiation sites and effectively revealed by their positional disequilibrium in eukaryote genome
- [Wakaguri08] Wakaguri, H., et al. DBTSS: database of transcription start sites, progress report 2008, Nucleic Acids Research, 2008, Vol. 36, Database issue D97–D101
- [Wang07] Wang, H. C. and D. A. Hickey (2007) Rapid divergence of codon usage patterns within the rice genome. BMC Evol Biol 7 Suppl 1: S6s using frequency distribution curves, BMC Bioinformatics 2006, 7:522
- [Sugahara01] Y. Sugahara et al., (2001) Comparative evaluation of 5′-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries, Gene, Volume 263, Issues 1-2, 24 January 2001, Pages 93-102.
- [Solovyev03] Solovyev, V.V. and Shahmuradov, I.A. (2003) PromH: promoters identification using orthologous genomic sequences. Nucleic Acids Research, 2003, Vol. 31, No. 13.