Automatic identification of experimental conditions relevant to a specific trait

Stavros Makrodimitris
s.makrodimitris@tudelft.nl

TU Delft

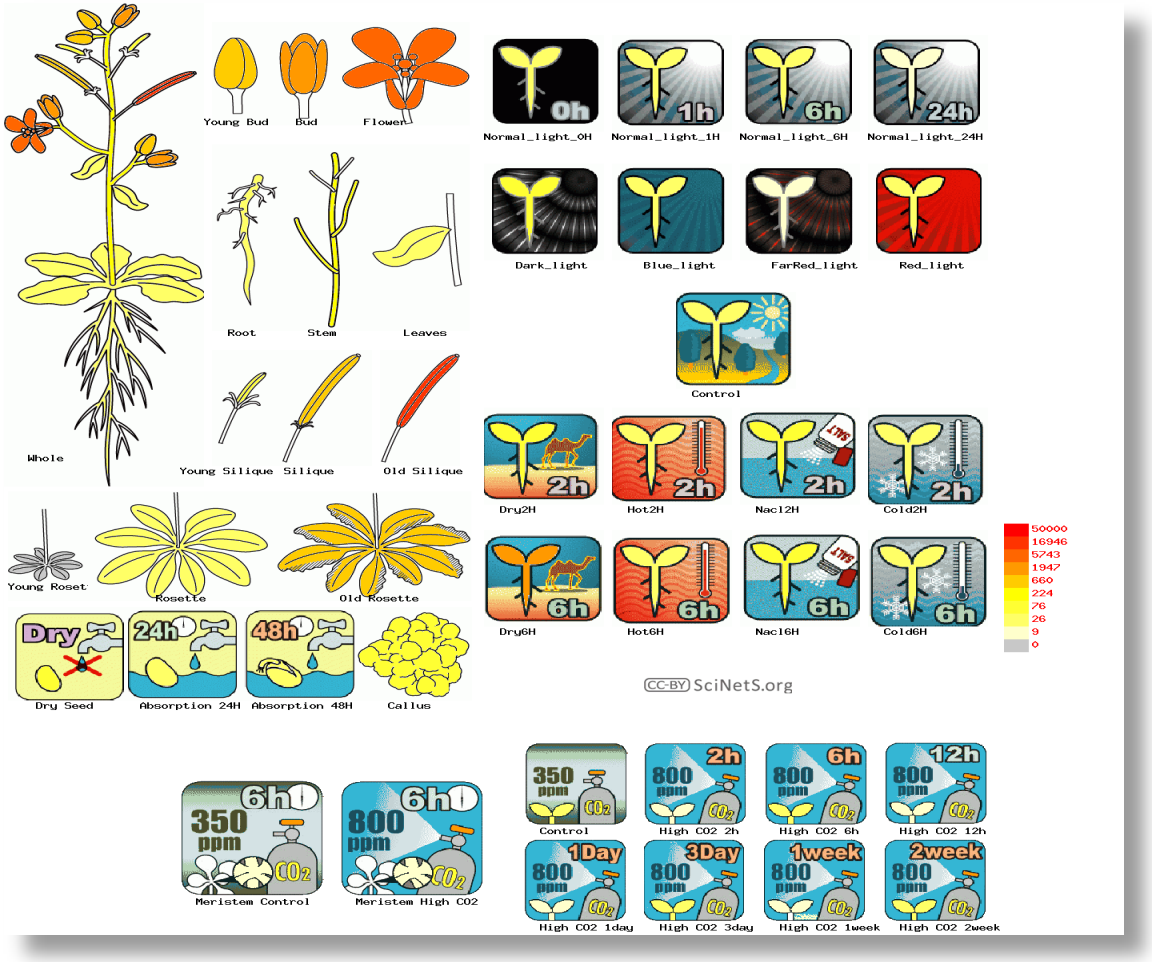
KeyGene

Candidate Gene Discovery

- Which genes affect a given trait?
- For all traits
- Genome-wide gene function prediction
- Experiments are accurate, but difficult

Can we use public data instead?

Many functions, Many conditions



Co-expression → Similar Function

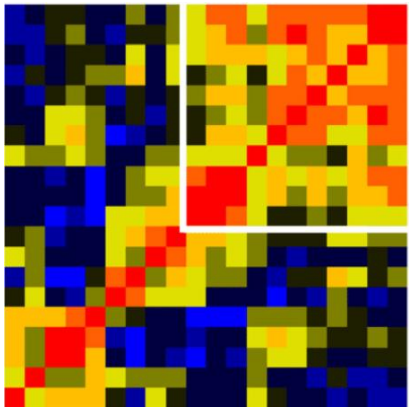
GEO

Samples

Genes

Genes

Genes

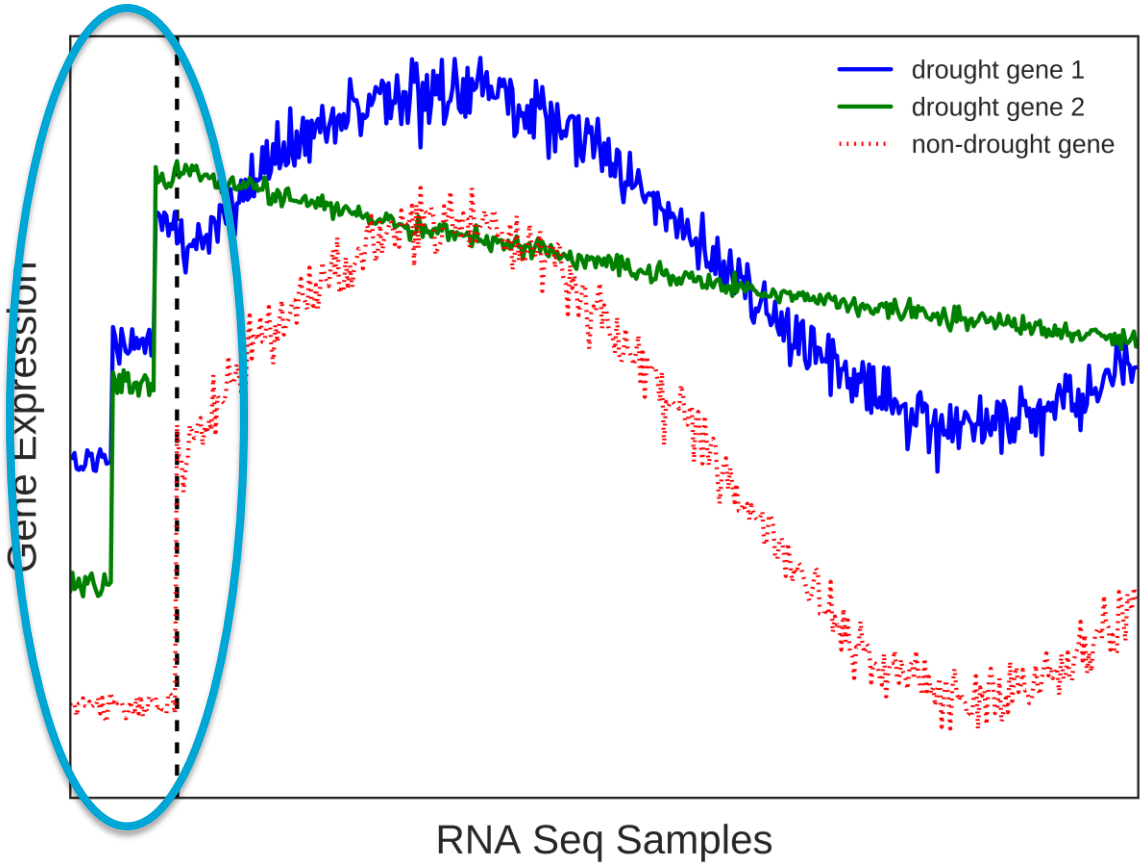


Known genes for any function

New Candidate Genes

TU Delft KeyGene

Most conditions are irrelevant!



Manual Sample Selection is Hard

- Poor meta-data
- Not obvious what might be relevant

Let the computer decide!

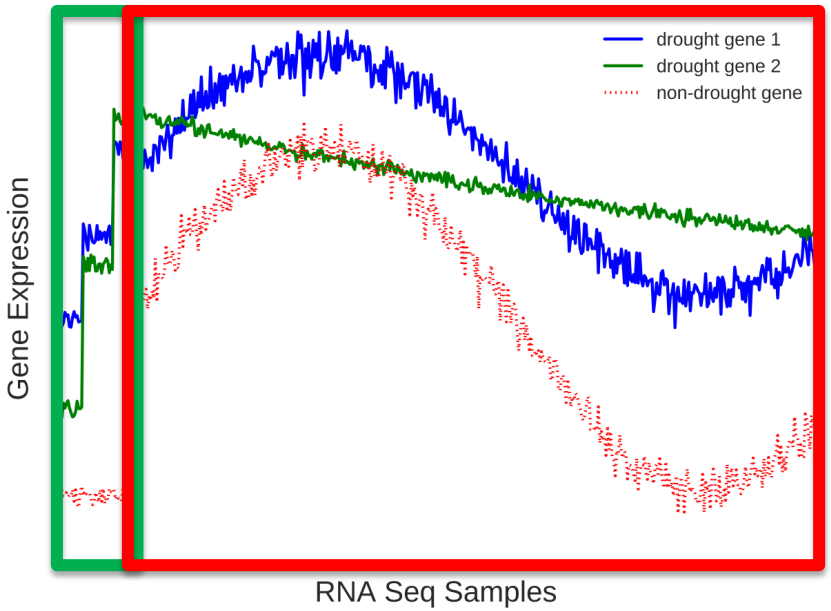
Score samples based on relevance

$$Coex(x_i x_j) = \sum_m x_{im} x_{jm}$$

$$Coex(x_i x_j, w) = \sum_m w_m x_{im} x_{jm}$$

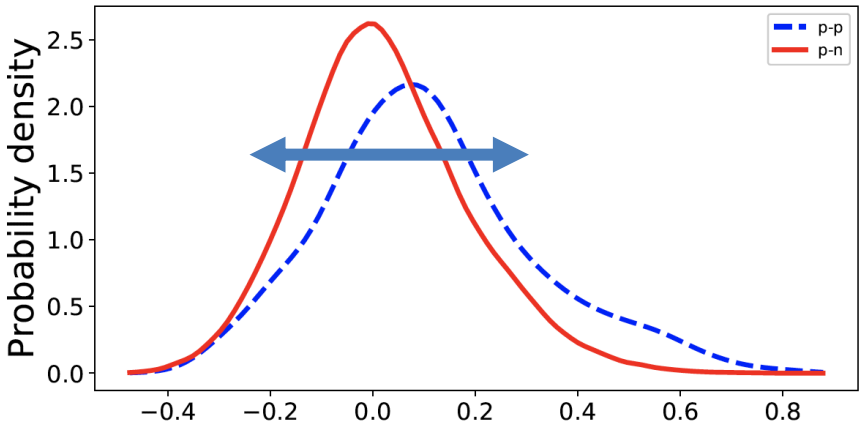
w = large

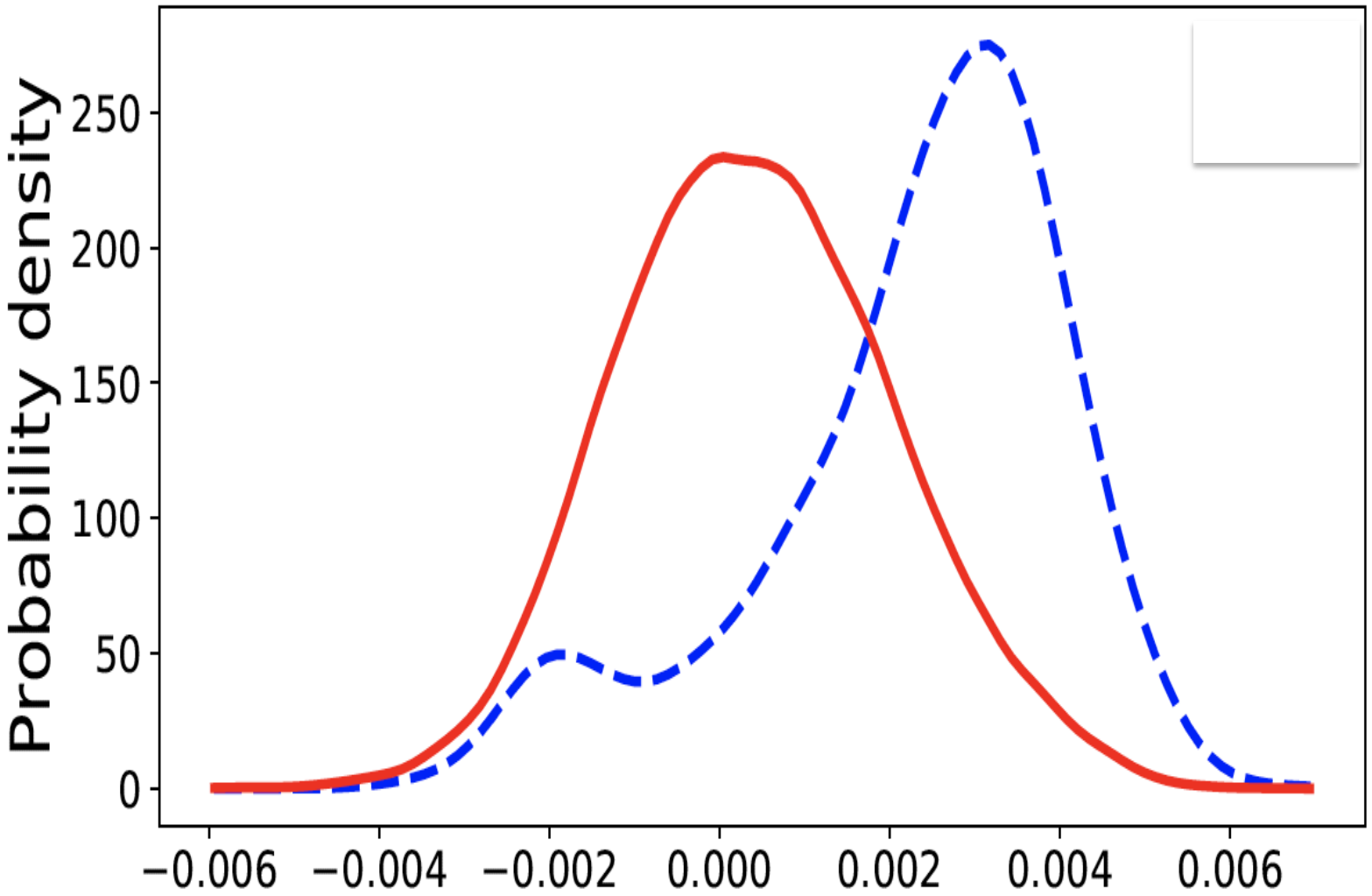
w = 0



Automatic Sample Selection

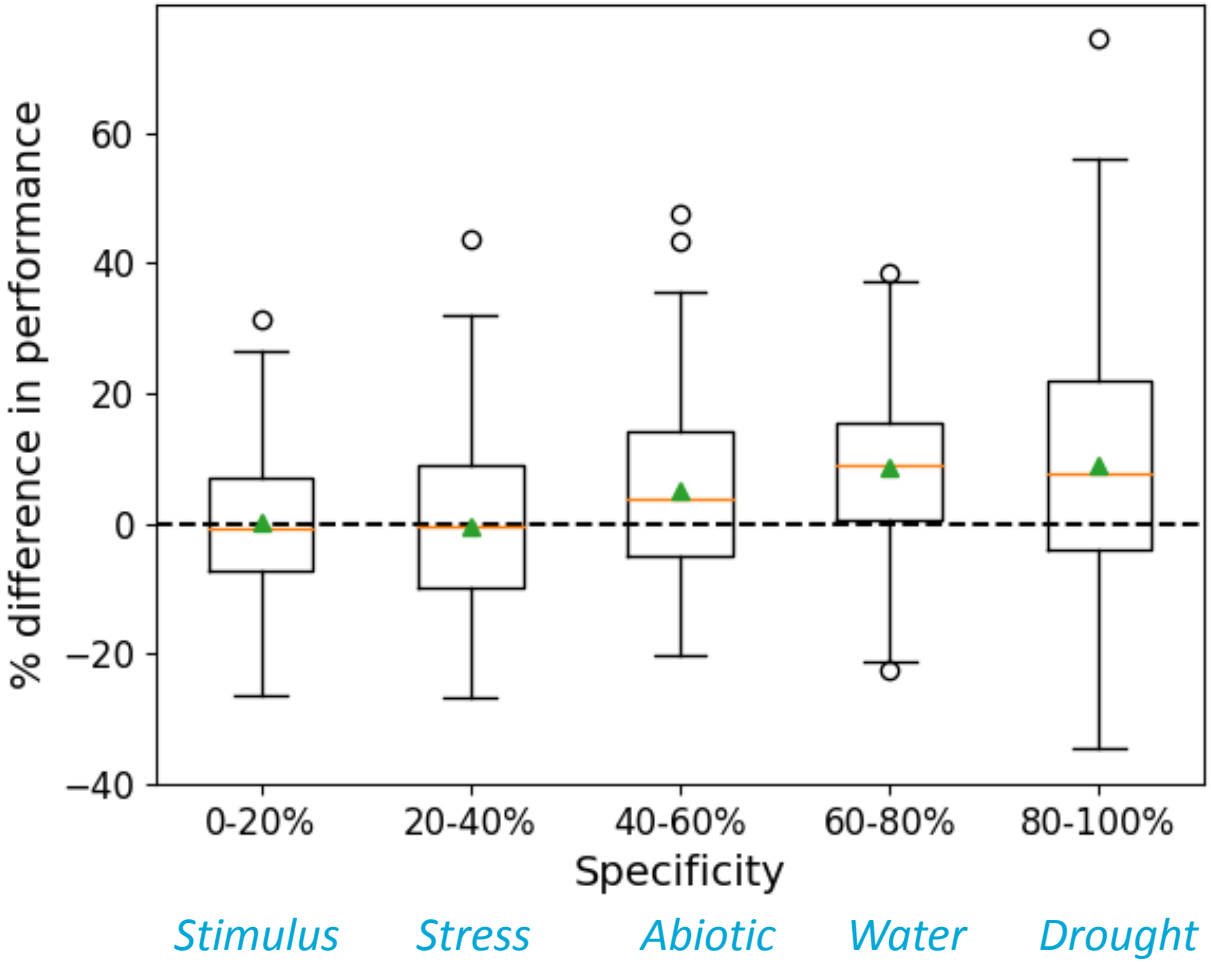
- ML model learns the w's
- Genes that have the function should:
  - Co-express with each other
  - Not co-express with other genes





TU Delft KeyGene

Results - Arabidopsis thaliana



Summary - MLC

- Trait-specific candidate gene discovery
- From a large collection of samples
- Learns from the data
- Recovers relevant samples
- Best at discovering specific knowledge

Questions?

Acknowledgements

Global Engage

Marcel Reinders
Roeland van Ham

Contact
s.makrodimitris@tudelft.nl
@sta_makro