

Gene prioritization by genomic data fusion

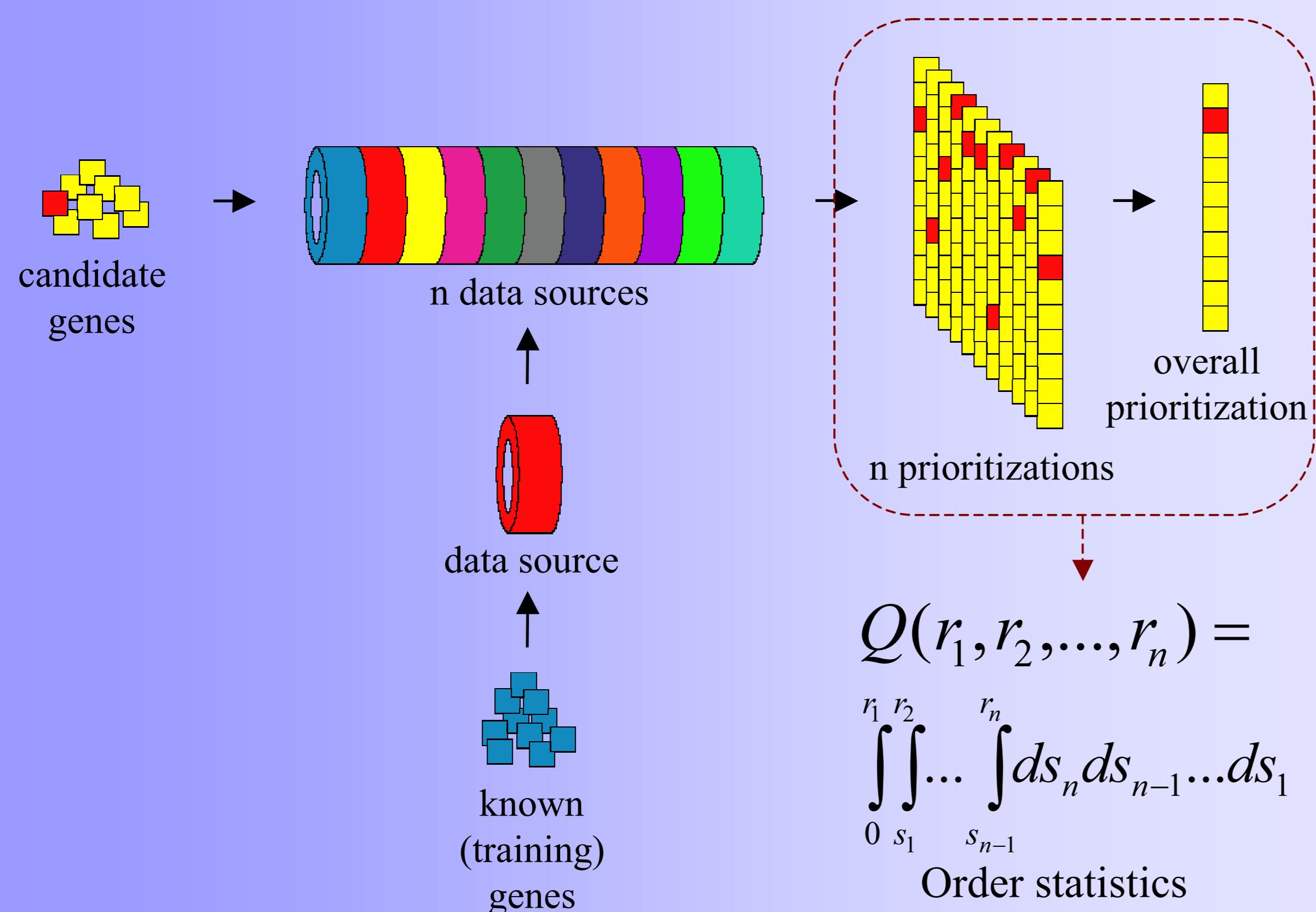
1. Introduction

Geneticists hunting for genes involved in specific diseases or processes are often confronted with relatively large lists of candidate genes. As only a limited number of genes can be further validated, there is a need to efficiently prioritize the candidate genes based on their likelihood to be involved in the disease or process. This requires the integration of large quantities of data from many different data sources. As the number and content of available information sources continue to increase, the need for an unbiased computational integration of all these data arises.

2. Objectives

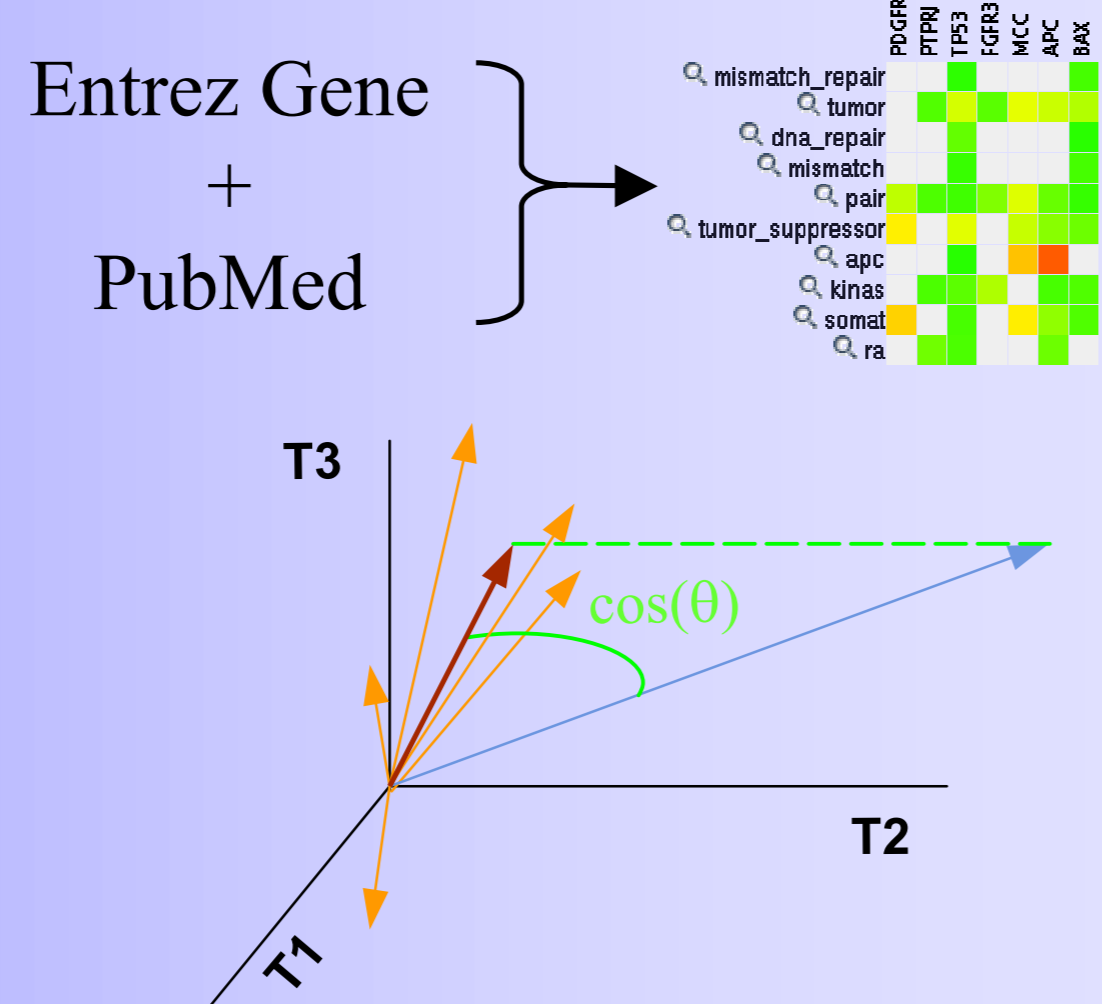
This work aims to design a computational method to prioritize any set of candidate genes, based on similarity to a set of "training genes", genes known to be involved in the disease or process, efficiently combining data from many heterogeneous data sources. We aspire to validate this method computationally, as well as *in vitro* and *in vivo*.

3. Genomic data fusion



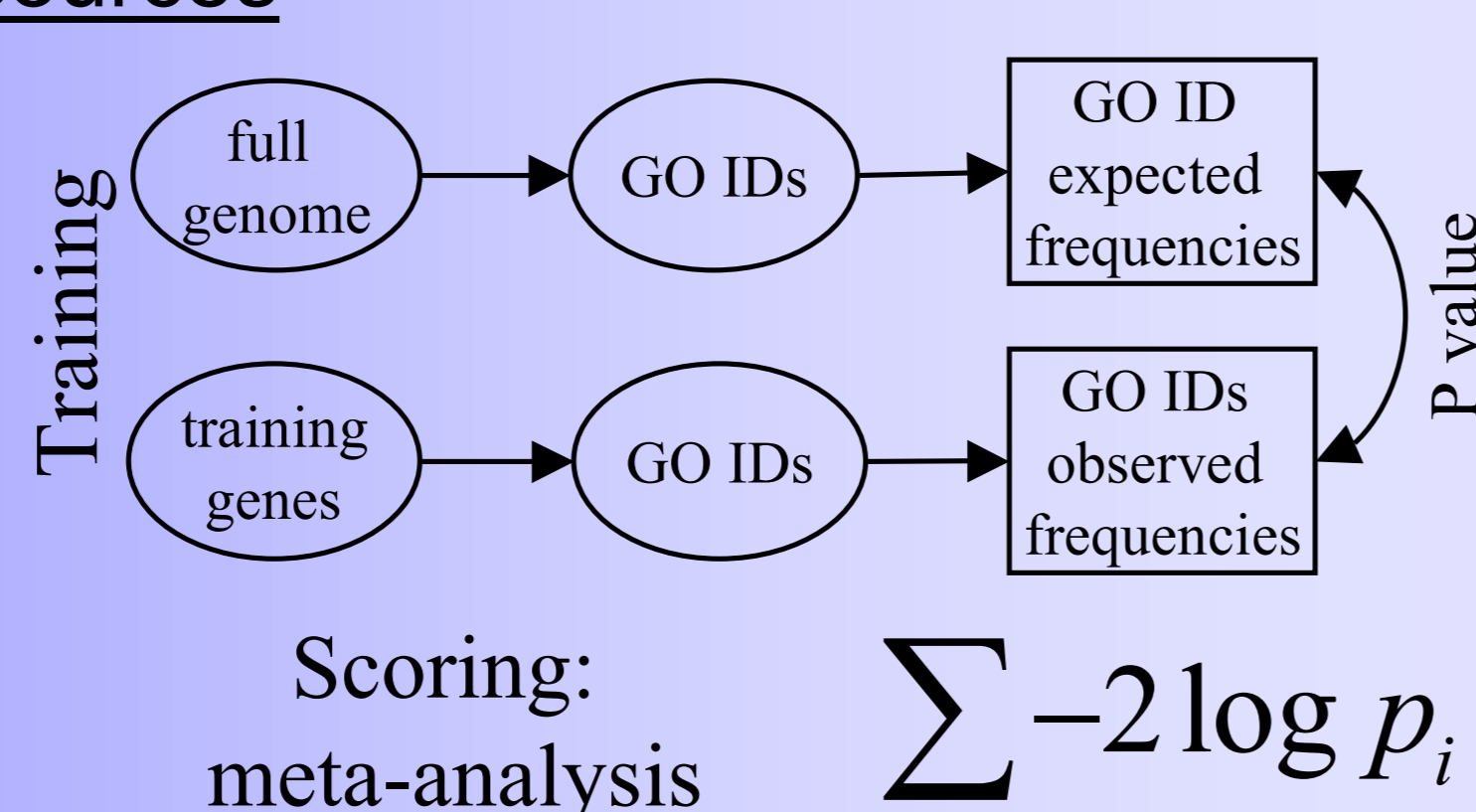
Vector based data sources

- Literature data
Vector of term weights
- Microarray data
Vector of normalized expression values
- Cis-regulatory motifs
Vector of transcription factor binding sites



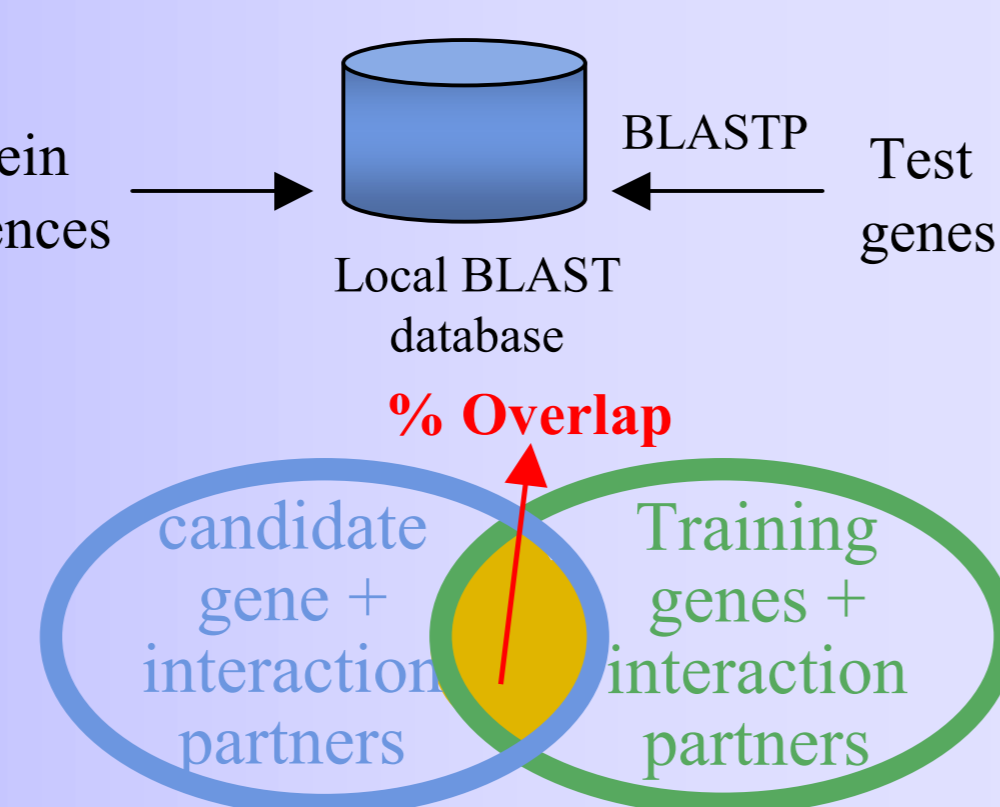
Attribute based data sources

- Gene ontology
- Protein domains
InterPro
- Pathways
KEGG
- Anatomical expression (EST)

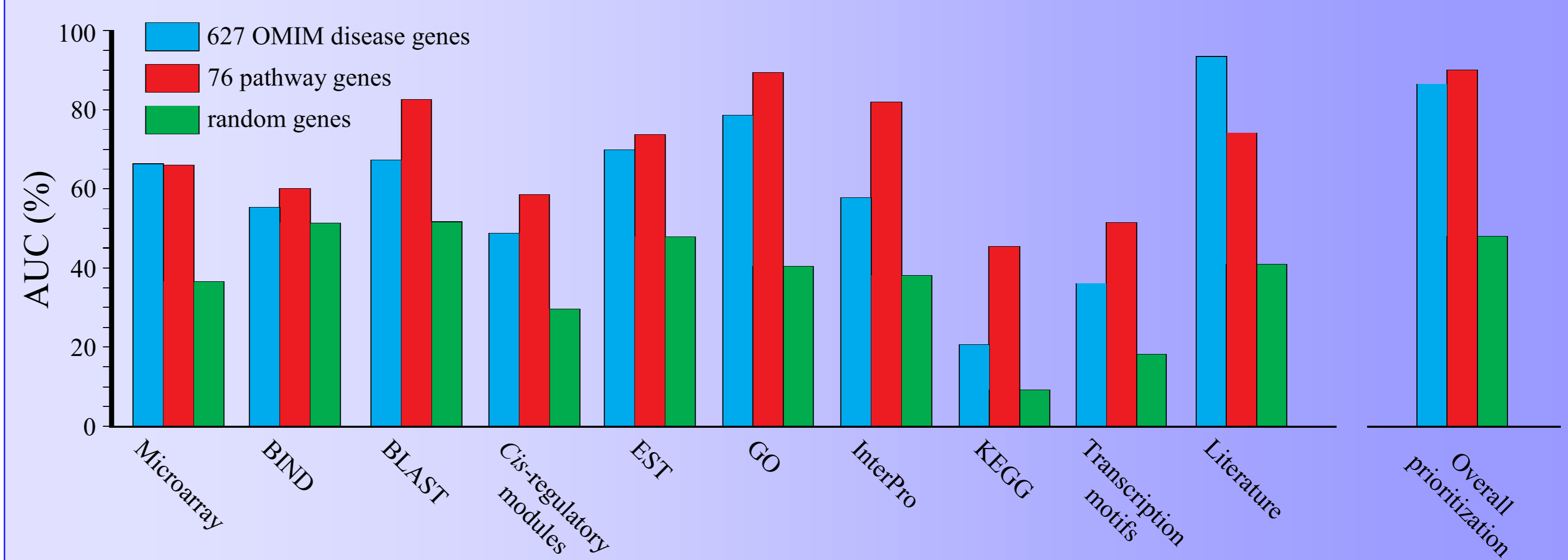
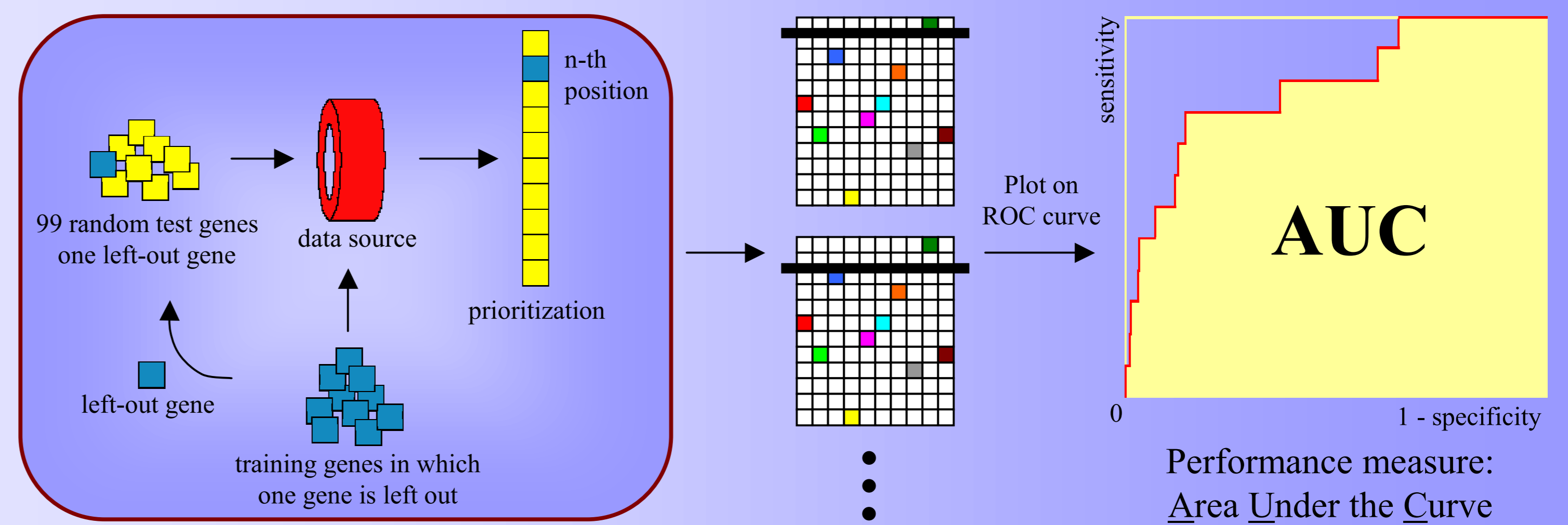


Other data sources

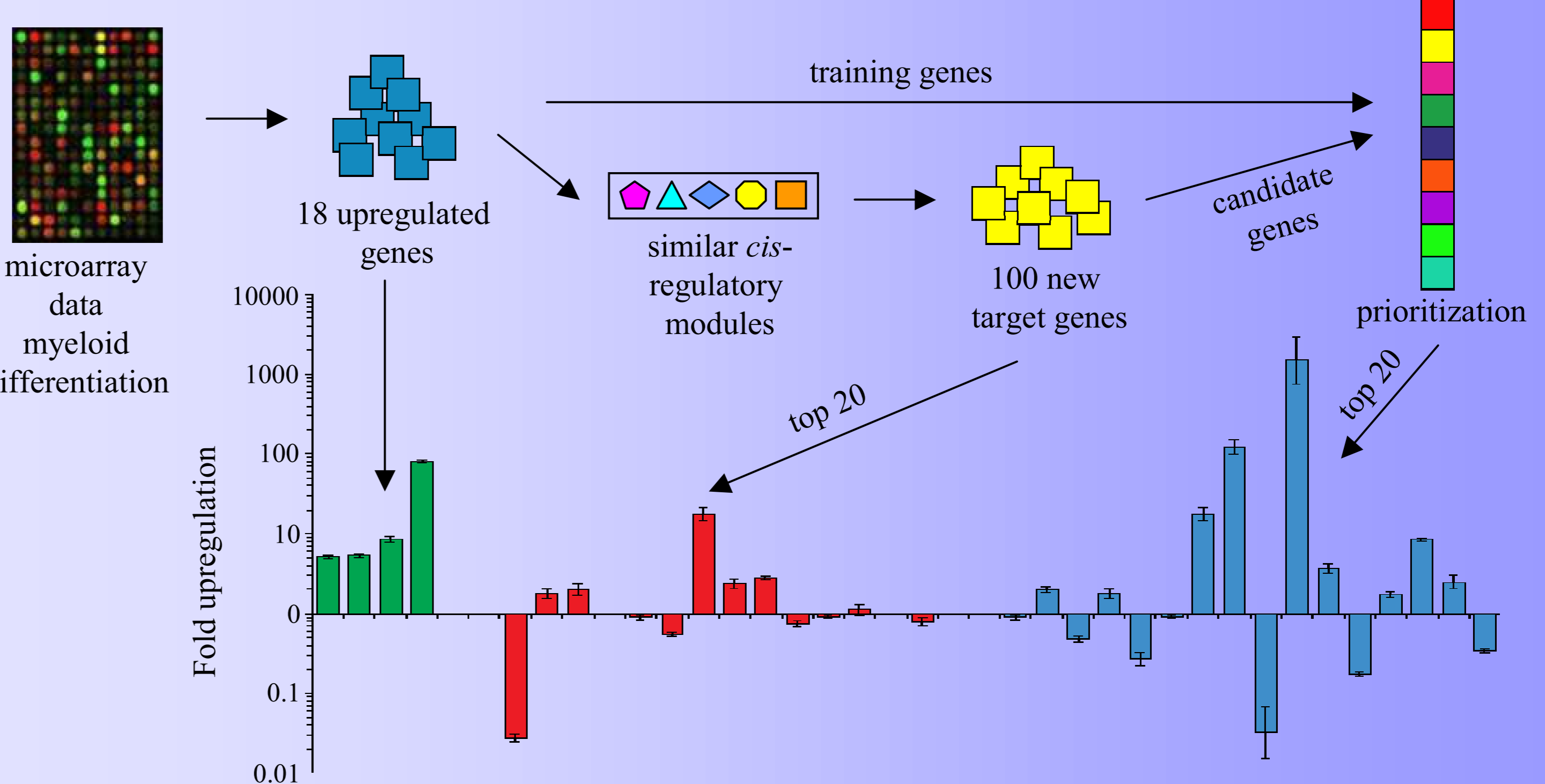
- BLAST
- Cis-regulatory modules
- Protein-protein interactions
BIND



4. Large scale cross-validation

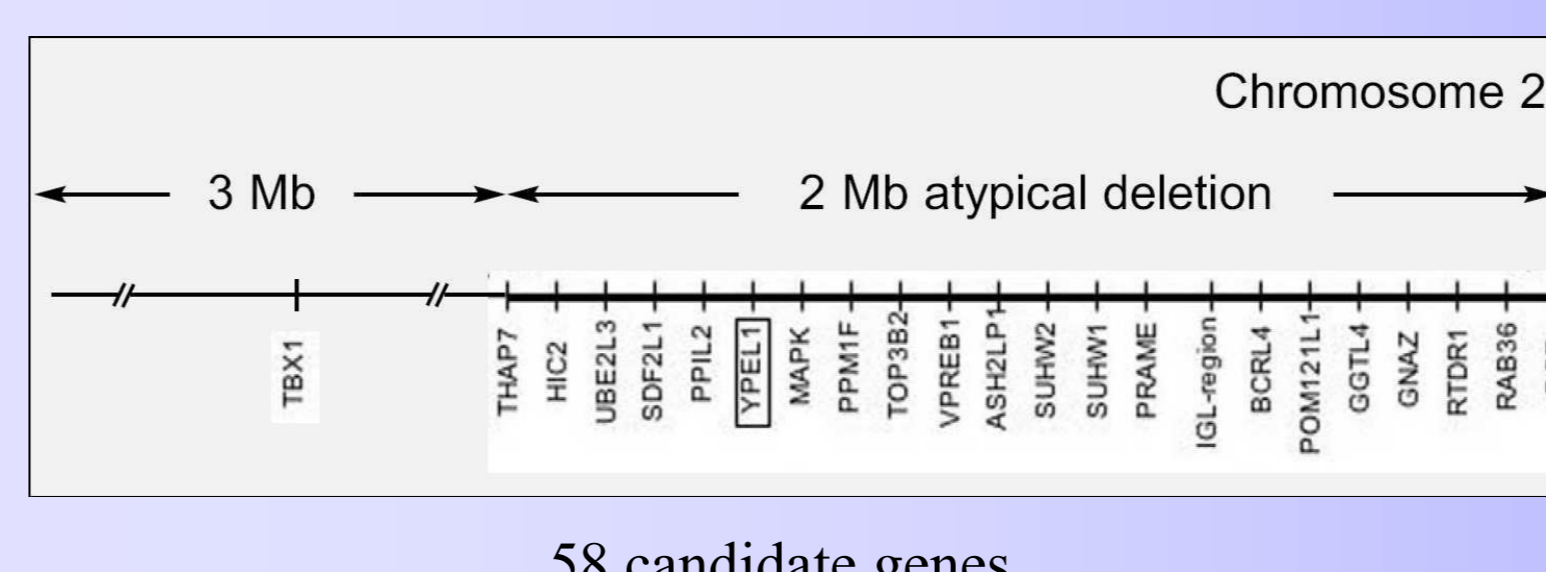


5. In vitro pathway case study



6. In vivo disease case study

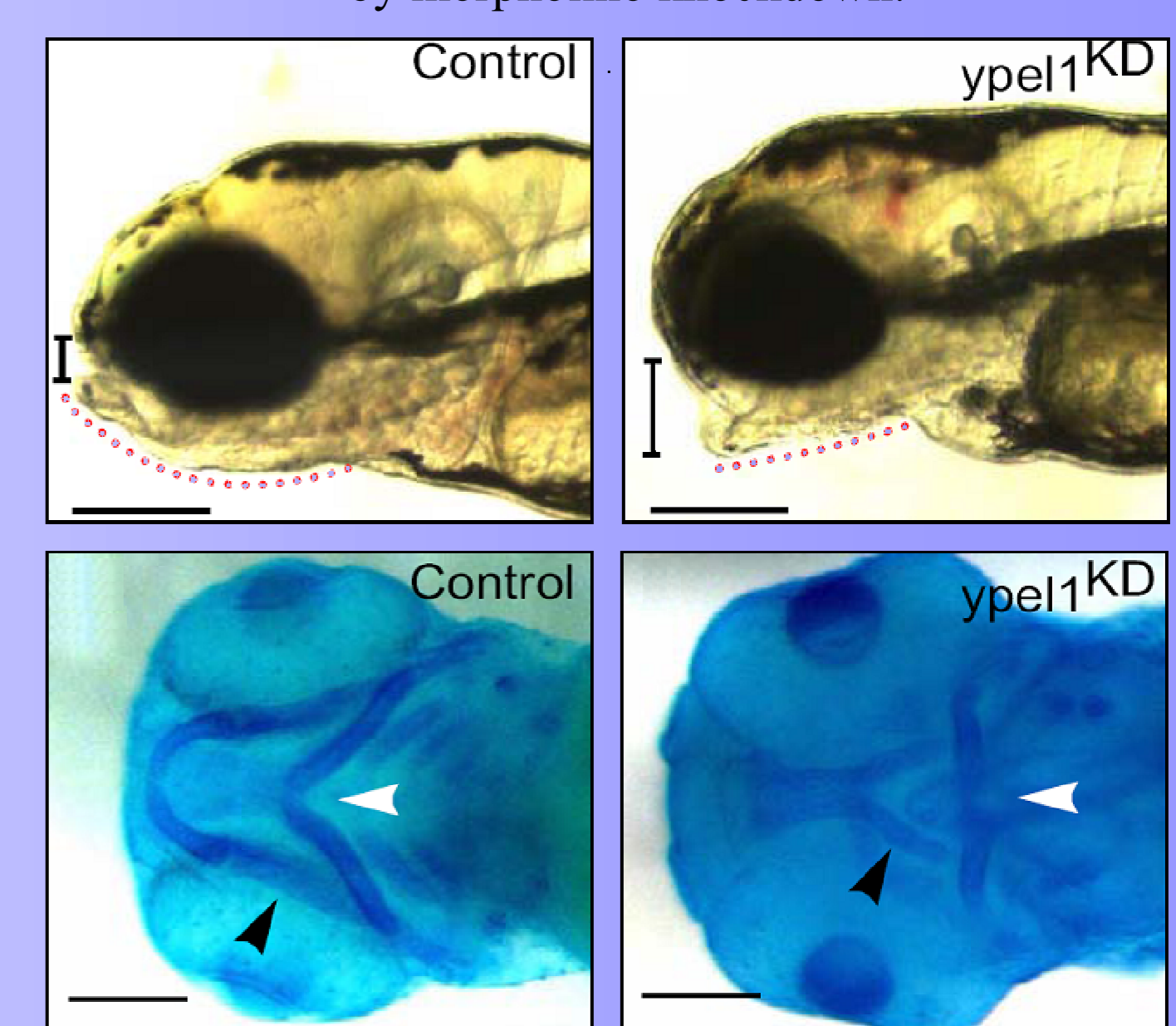
DiGeorge Syndrome: an atypical 2Mb deletion at 22q11:



Prioritization using 4 training sets modelling 4 different aspects of DiGeorge Syndrome identifies YPEL1 as a candidate gene

| Training sets used to prioritize TBX1 or YPEL1 | Rank assigned to TBX1 | Rank assigned to YPEL1 |
|--|-----------------------|------------------------|
| DGS (14) | 1 | 1 |
| Cardiovascular birth defects (14) | 1 | 3 |
| Cleft palate birth defects (9) | 1 | 2 |
| Neural crest genes (14) | 2 | 1 |
| Average rank | 1.25 ± 0.25 | 1.75 ± 0.48 |

In vivo functional validation of Ypel1 in zebrafish by morpholino knockdown:



7. Conclusions

- ENDEAVOUR is a novel method to prioritize a set of candidate genes based on similarity to a set of training genes
- Genomic data fusion: combination of multiple, heterogeneous data sources in one global prioritization
- Validation *in silico*, *in vitro* and *in vivo*
- Try ENDEAVOUR at: <http://www.esat.kuleuven.be/endeavour>

Reference: Aerts, S., Lambrechts, D., Maity, S., Van Lo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24: 537-544.