# Disease-gene discovery by integration of 3D gene expression and transcription factor binding affinities

[Piro et al. *Bioinformatics* 29(4):468–475, 2013]

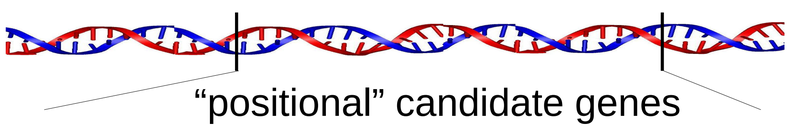

Rosario M. Piro[1,2], Ivan Molineris[3], Ferdinando Di Cunto[3], Roland Eils[1,2] and Rainer König[2,4,5]

[1] Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, BioQuant, University of Heidelberg, Germany
[2] Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[3] Department of Genetics, Biology and Biochemistry and Molecular Biotechnology Center, University of Turin, Torino, Italy
[4] Center for Sepsis Control and Care, University Hospital Jena, Germany
[5] Hans-Knöll-Institute (HKI), Jena, Germany


The computational evaluation of candidate genes for hereditary disorders is a non-trivial task (Piro and Di Cunto, 2012). We have shown previously that spatially mapped, i.e. 3D, gene expression data from the mouse brain can be successfully used to prioritize candidate genes for human Mendelian disorders of the central nervous system (Piro *et al*., 2010). We improve our approach two-fold: i) we show that condition-independent TF binding affinities of the candidate genes' promoters are relevant for disease-gene prediction and integrate them with our previous method; and ii) we define a novel similarity measure—termed *Relative Intensity Overlap (RIO)*—for both 3D gene expression patterns and binding affinity profiles that better exploits their disease-relevant information content. Finally, we present an extensive leave-one-out cross validation and predict promising candidates for disorders of unknown molecular basis that are characterized by mental retardation.

## The problem:

Many mapped disease loci remain "orphan", i.e. the disease-associated genes within the loci have not yet been identified.
>1000 loci in OMIM (Hamosh *et al.*, 2002)



"positional" candidate genes

*Which of the candidate genes are more likely to be involved in a given disorder?*

## Common approaches:

comparison of candidates with known disease "reference" genes to predict a functional relationship
- ⊳ functional annotation
- ⊳ protein-protein interactions
- ⊳ co-expression
- ⊳ intrinsic gene/protein properties
- ⊳ multiple sources

(many data sources are biased towards well-characterized candidates)

## Our goal:

Can promoter-TF binding affinities help in candidate gene prioritization?
- ⊳ advantage: unbiased; computed from promoter sequences
- ⊳ disadvantage: context-independent (static)

How can we best exploit their information?
- ⊳ new similarity measure

Integration with our previous approach
- ⊳ 3D gene expression patterns

## What is total binding affinity (TBA)?

Traditional promoter analysis: PWM scans 

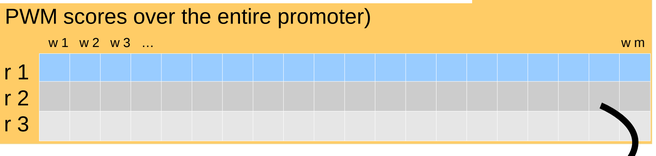
Problem: ignores relevant low-affinity binding sites

Our approach: (see Foat *et al*.,2006; Molineris *et al*., 2011)
likelihood that a promoter *r* can be bound by a TF, estimated from *r*'s TBA $a_{rw}$ for the TF's PWM *w*:

$$a_{rw} = log \sum_{i=1}^{L-l} max \left( \prod_{j=1}^{l} \frac{P(w_j, r_{i+j})}{P(b, r_{i+j})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j})}{P(b, r_{i+j})} \right)$$

(sum of PWM scores over the entire promoter)

for each promoter:
TBA profile for 130
Jaspar Core PWMs




## Co-affinity / co-expression:

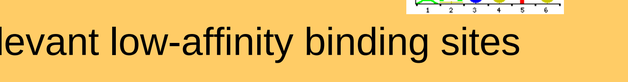
New similarity measure: *relative intensity overlap (RIO)*
- overlap by multiplication of intensities $I$
- relative to max. possible overlap
- for both normalized 3D expression profiles and z-transformed TBA profiles
- better than Pearson correlation coefficient on our datasets

$$\mathrm{RIO}(a,b) = \frac{\sum_{xyz} \left( I^a_{xyz} \times I^b_{xyz} \right)}{\sum_{xyz} \left( max(|I^a_{xyz}|, |I^b_{xyz}|) \right)^2}$$

## Spatially mapped (3D) gene expression:

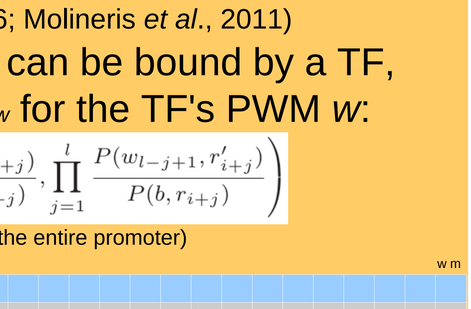
Traditional gene expression data carry little spatial information; may be a drawback for tissues with a high degree of spatial organization (e.g. brain):

Traditional co-expr.:



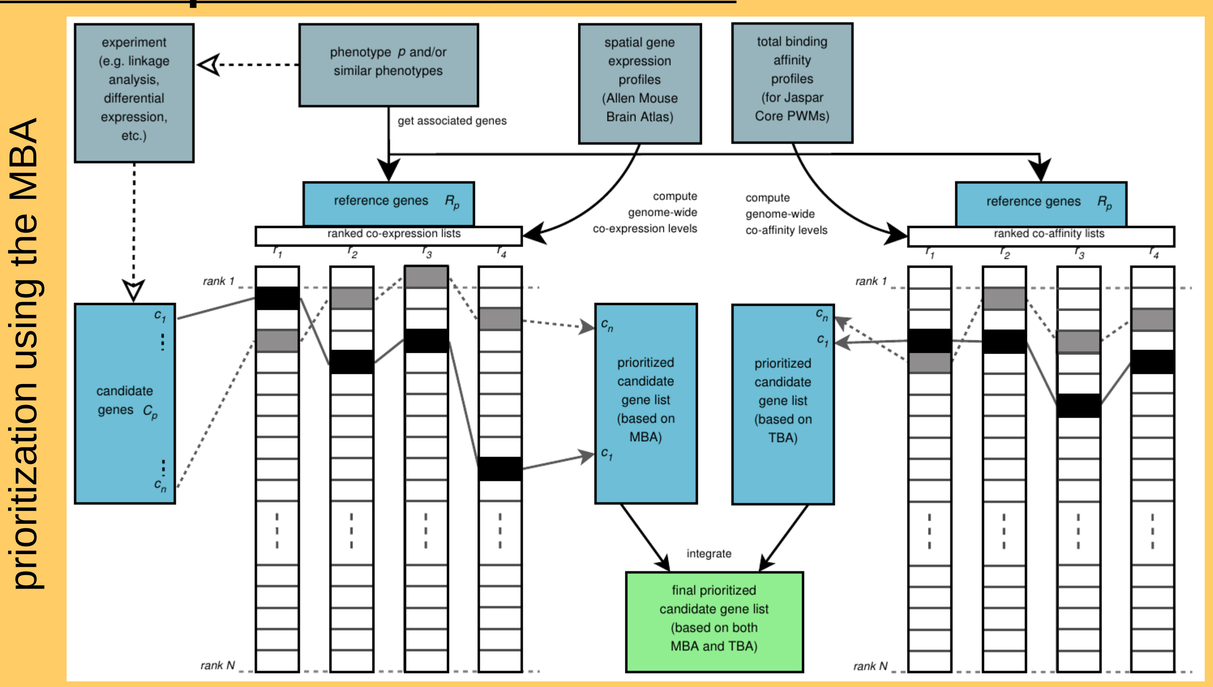

3D gene expression profiles of the mouse brain atlas (MBA):

(Allen Brain Atlas; Lein et al., 2007)




3D coexpression predicts XLMR genes (Piro et al., 2010)

## Candidate prioritization / workflow:




1) prioritize the candidates according to their co-expression with the reference genes (MBA)
2) prioritize the candidates according to their co-affinity with the reference genes (using TBAs)
3) integration: final ranking by means of a adapted generalized noisy-OR gate (Díez, 1993).

## Leave-one out cross validation:

(using hundreds of known gene-disease associations from OMIM); AUC ≈ 0.81

| Sim. | *N* | $C_p$ | *g*–*p* pairs | Ranked first | | | Ranked 1st–3rd | | | Ranked 1st–10th | | | Ranked≤10% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Obs. | E. | *P*-value | Obs. | E. | *P*-value | Obs. | E. | *P*-value | Obs. | E. | *P*-value |
| PCC | 50 | 73.4 | 756 | 22 | 10 | 8.53e-04 | 59 | 31 | 2.19e-06 | 157 | 103 | 3.73e-08 | 119 | 76 | 4.82e-07 |
| PCC | 100 | 136.3 | 805 | 16 | 6 | 3.81e-04 | 39 | 18 | 5.81e-06 | 97 | 59 | 1.25e-06 | 123 | 81 | 1.63e-06 |
| PCC | 200 | 253.3 | 808 | 9 | 3 | 5.40e-03 | 26 | 10 | 6.81e-06 | 62 | 32 | 8.01e-07 | 126 | 81 | 4.40e-07 |
| PCC | 400 | 439.3 | 808 | 8 | 2 | 6.19e-04 | 17 | 6 | 5.93e-05 | 38 | 18 | 3.20e-05 | 131 | 81 | 2.90e-08 |
| RIO | 50 | 73.4 | 756 | **38** | 10 | 9.78e-12 | **97** | 31 | 3.33e-23 | **230** | 103 | 2.20e-33 | **174** | 76 | 6.31e-26 |
| RIO | 100 | 136.3 | 805 | 16 | 6 | 3.81e-04 | **58** | 18 | 7.85e-15 | **166** | 59 | 6.18e-34 | **206** | 81 | 5.63e-37 |
| RIO | 200 | 253.3 | 808 | **20** | 3 | 1.70e-10 | **50** | 10 | 8.09e-21 | **140** | 32 | 8.48e-49 | **262** | 81 | 2.03e-68 |
| RIO | 400 | 439.3 | 808 | **16** | 2 | 1.16e-10 | **43** | 6 | 1.89e-24 | **124** | 18 | 1.70e-62 | **315** | 81 | 1.81e-105 |

## Prediction results for mental retardation (MR):

Applied to several orphan loci for MR syndromes: Many of the high ranking candidates are already known to be involved in some distinct MR syndrome or other neurological disorder.

Alopecia/mental retardation syndrome 1 (APMR1; OMIM %203650; 3q26.3-q27.3)
- Some promising new candidates including, for example:
  - ***DVL3****:* interacts with Shank, which is involved in several neuronal disorders including other MR syndromes (Saupe *et al*., 2011)

## Conclusions.

We showed that static, condition-independent binding affinities for transcription factors can be used to successfully prioritize or rank candidate genes for hereditary disorders. We have integrated this approach into our previous method based on 3D gene expression patterns from the mouse brain and further improved it through a new similarity measure for both 3D gene expression patterns and TBA profiles.
We applied the new approach to orphan loci for disorders characterized by mental retardation and found several promising candidates.

## References


Díez FJ (1993) Parameter adjustment in Bayes networks. The generalized noisy-OR gate. In: Proc. 9th *Conf. Uncertainty in Artificial Intelligence*, pp. 99-105.

Foat BC *et al*. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**(14):e141–e149.

Hamosh A *et al*. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. *Nucleic Acids Res* **30**:52–55.

Lein ES *et al*. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**:168–176.

Molineris I *et al*. (2011) Evolution of promoter affinity for transcription factors in the human lineage. *Mol. Biol. Evol.* **28**(8):2173–2183.

Piro RM *et al*. (2010) Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR. *Bioinformatics* **26**(18):i618–i624.

Piro RM, Di Cunto F (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.* **279**(5):678–696.

Saupe J *et al*. (2011) Discovery, structure-activity relationship studies, and crystal structure of nonpeptide inhibitors bound to the Shank3 PDZ domain. *ChemMedChem* **6**:1411-1422.