ECCB2010

0101010101111110100101001010101000101011110000101101001010111001101011

Computational Biology Unit
CBU@MBC
Molecular Biotechnology Center
Università di Torino

Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR

***Rosario M. Piro**, Ivan Molineris, Ugo Ala, Paolo Provero, and Ferdinando Di Cunto*

rosario.piro@unito.it

Molecular Biotechnology Center
and Dep. Genetics, Biology and Biochemistry
University of Torino, Italy

ACGTCAGCTAGCTAGCTAGTCAGTCGATCGATGTGTACATGCAATCGTAGCTAGCTAGCTAGCTA
TGCATGCAACTGATGCATGCATGCTGATCGATGCATGCT ROSARIO.PIRO@UNITO.IT CTAGC

Motivation

- Linkage analysis can help to identify disease-associated loci, but:

Motivation

- Linkage analysis can help to identify disease-associated loci, but:
  - often hundreds of *positional candidate genes*

Motivation

- Linkage analysis can help to identify disease-associated loci, but:
  - often hundreds of *positional candidate genes*
  - Online Mendelian Inheritance in Man (OMIM): currently over 1000 “orphan” disease loci for phenotypes with **unknown molecular basis** but mapped locus

Motivation

- Linkage analysis can help to identify disease-associated loci, but:
  - often hundreds of *positional candidate genes*
  - Online Mendelian Inheritance in Man (OMIM): currently over 1000 “orphan” disease loci for phenotypes with **unknown molecular basis** but mapped locus
  - next generation sequencing techniques?
    - large-scale re-sequencing of Chr. X in 208 XLMR families indicated 3 novel (and several known) disease genes, but in most cases the genetic cause could not be reliably determined, although many mutations (even truncating) were found. (*Tarpey et al., Nat. Genet., 2009*)

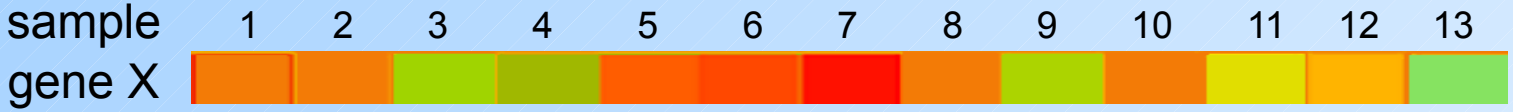
Motivation

- Linkage analysis can help to identify disease-associated loci, but:
  - often hundreds of *positional candidate genes*
  - Online Mendelian Inheritance in Man (OMIM): currently over 1000 “orphan” disease loci for phenotypes with **unknown molecular basis** but mapped locus
  - next generation sequencing techniques?
    - large-scale re-sequencing of Chr. X in 208 XLMR families indicated 3 novel (and several known) disease genes, but in most cases the genetic cause could not be reliably determined, although many mutations (even truncating) were found. (*Tarpey et al., Nat. Genet., 2009*)

=> Computational ***disease gene prioritization***

3D expression data

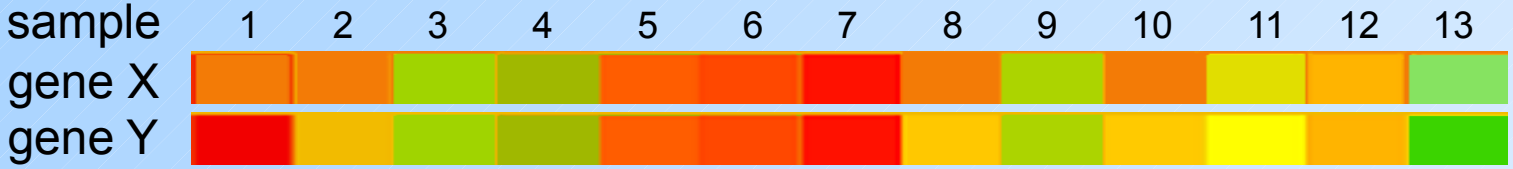
- “Traditional” high-throughput expression data



ROSARIO.PIRO@UNITO.IT

3D expression data

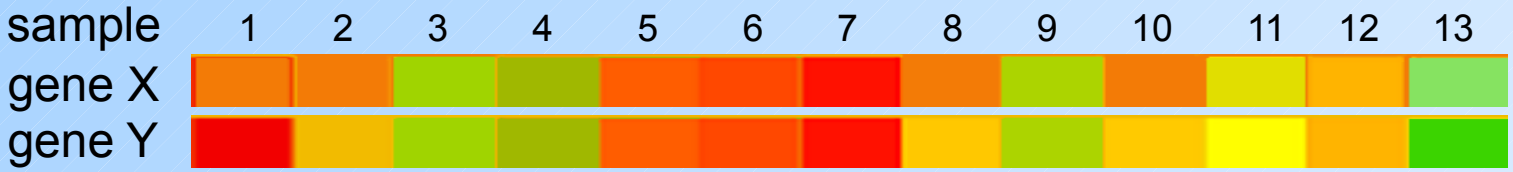
- “Traditional” high-throughput expression data



ROSARIO.PIRO@UNITO.IT

3D expression data

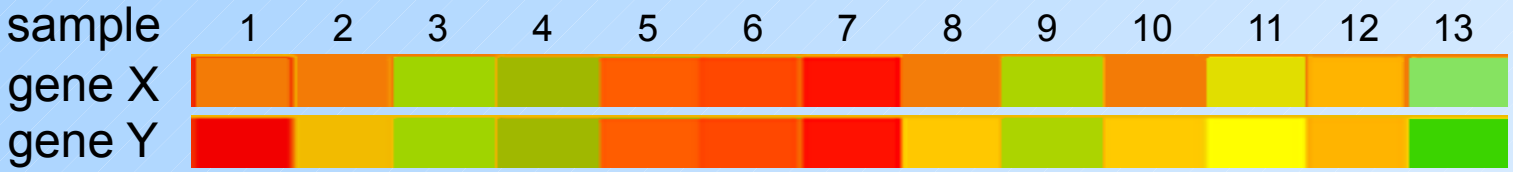
- “Traditional” high-throughput expression data



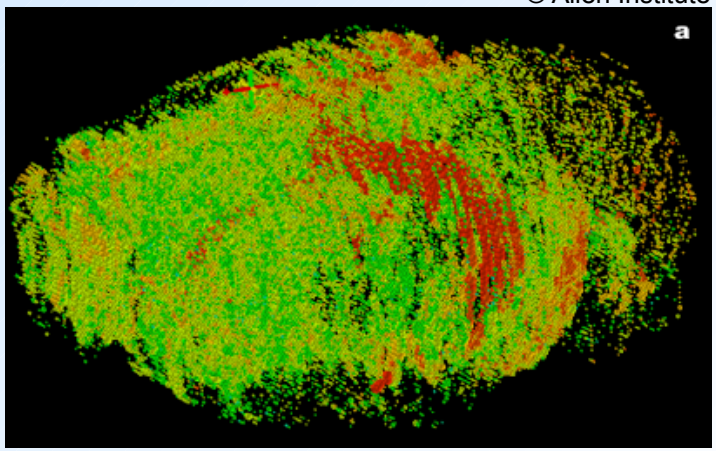
  - carry **limited spatial information**, no precise 3D localization
  - have a **sparse coverage** (arbitrary, not-consecutive positions)
  - problem for tissues/organs with a high degree of spatial organization, e.g. the central nervous system (CNS)

3D expression data

- "Traditional" high-throughput expression data



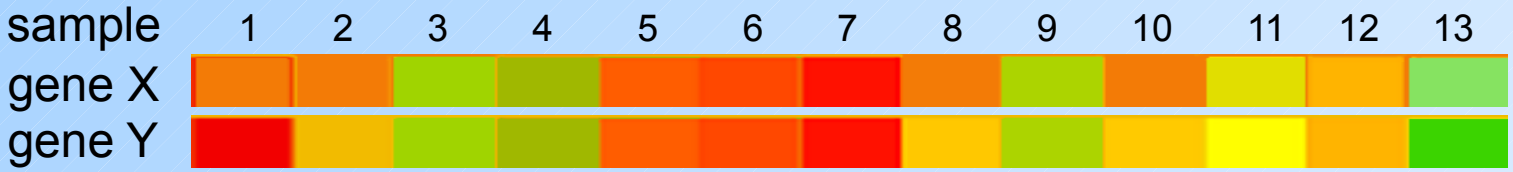
  - carry **limited spatial information**, no precise 3D localization
  - have a **sparse coverage** (arbitrary, not-consecutive positions)
  - problem for tissues/organs with a high degree of spatial organization, e.g. the central nervous system (CNS)



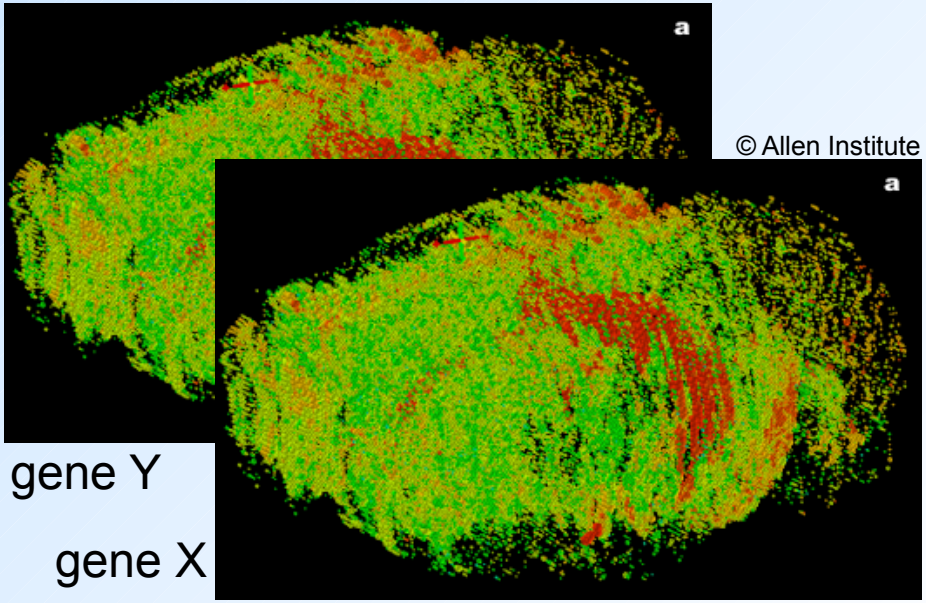
gene X

3D expression data

- “Traditional” high-throughput expression data



  - carry **limited spatial information**, no precise 3D localization
  - have a **sparse coverage** (arbitrary, not-consecutive positions)
  - problem for tissues/organs with a high degree of spatial organization, e.g. the central nervous system (CNS)

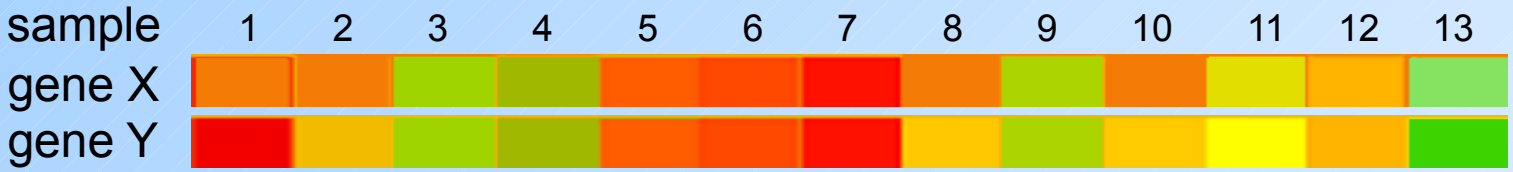


gene Y

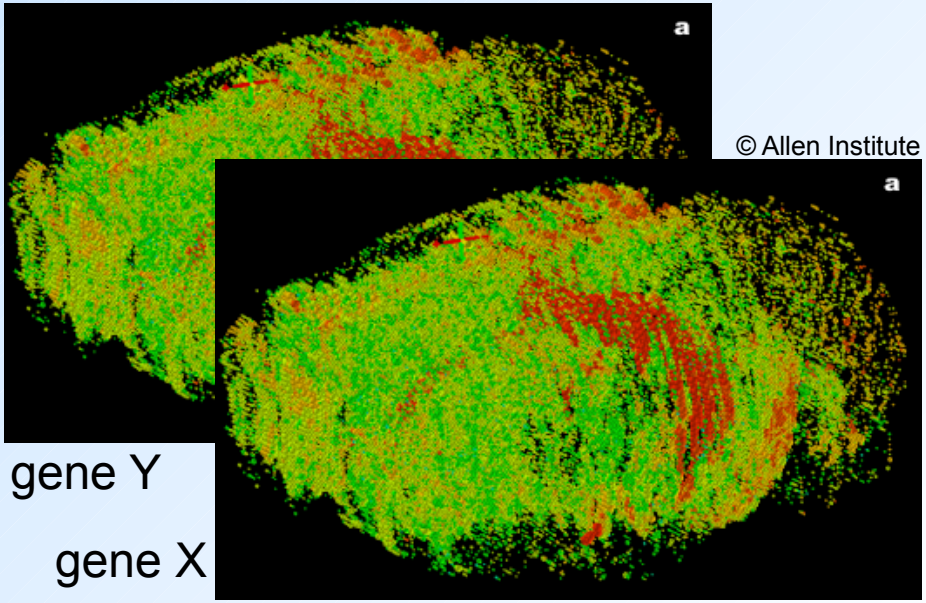
gene X

3D expression data

- "Traditional" high-throughput expression data

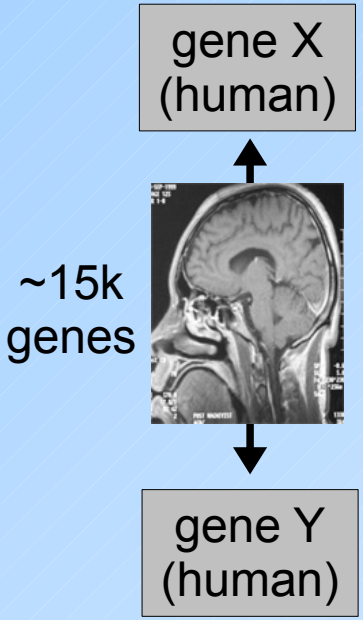


  - carry **limited spatial information**, no precise 3D localization
  - have a **sparse coverage** (arbitrary, not-consecutive positions)
  - problem for tissues/organs with a high degree of spatial organization, e.g. the central nervous system (CNS)
- Allen (Mouse) Brain Atlas
  (*Lein et al., Nature, 2007*)
  - http://www.brain-map.org/
  - ~20k genes (~18k with Entrez ID)
  - *in situ* hybridization (ISH)
  - covers the entire mouse brain
  - expression levels for "voxels" (cubes) of 200 µm side length



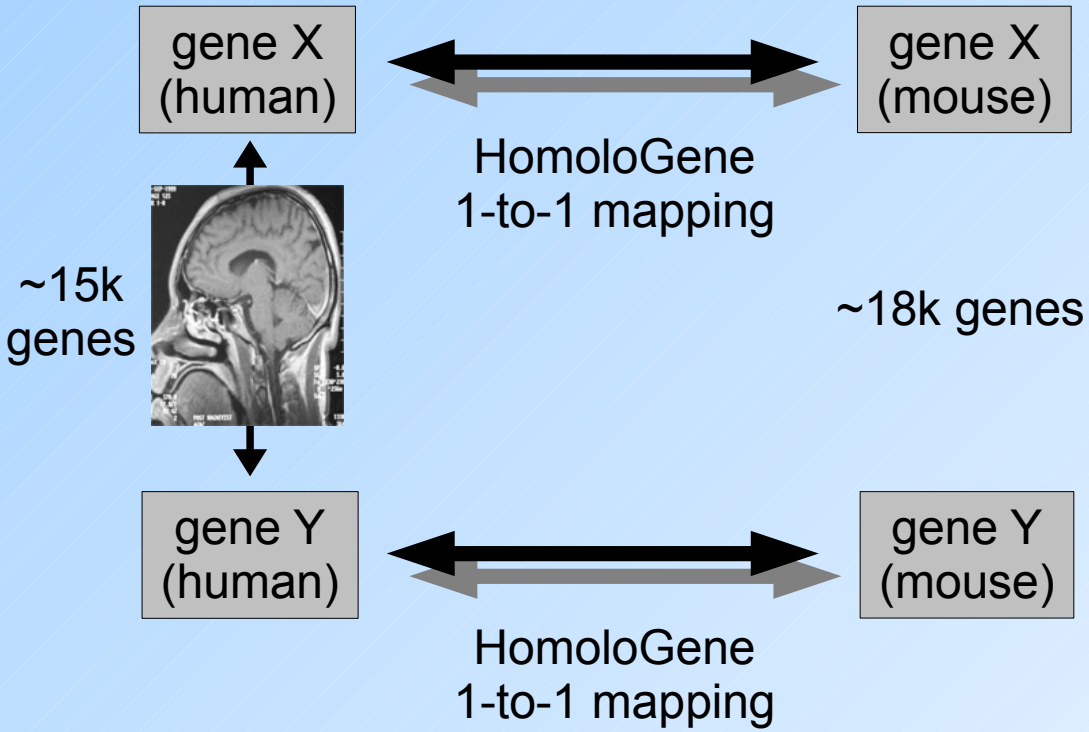
Mapping to human genes

- We use mouse expression data also for **human Mendelian disorders**



Mapping to human genes

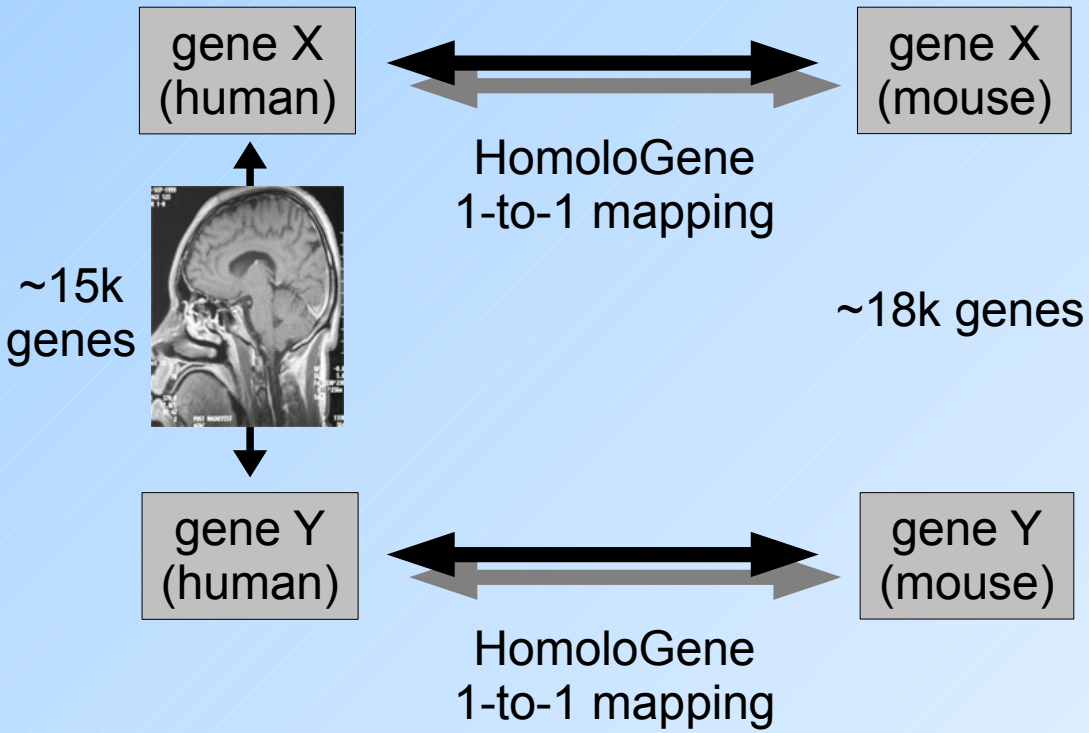
- We use mouse expression data also for **human Mendelian disorders**

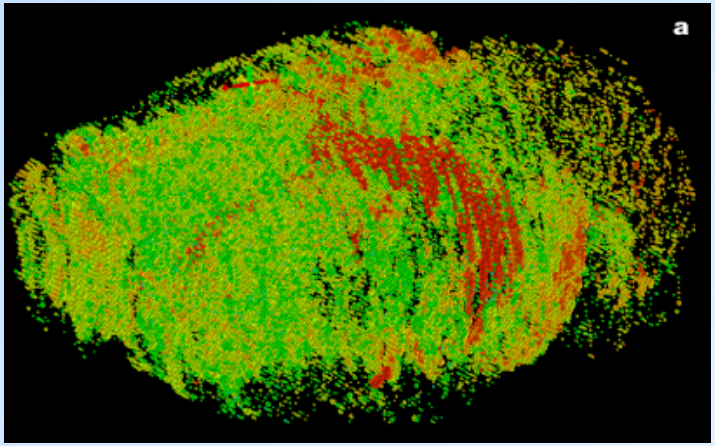


ROSARIO.PIRO@UNITO.IT

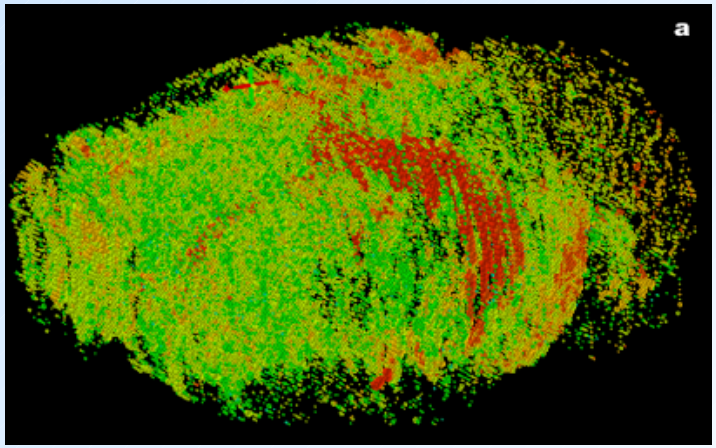
Mapping to human genes

- We use mouse expression data also for **human Mendelian disorders**





© Allen Institute



Prioritization approach (1)

- *Step 1 – select "reference genes"*
  - **genes already known** to be involved in the given phenotype or in similar phenotypes (MimMiner; *van Driel et al., Eur. J. Hum. Genet., 2006*)

Prioritization approach (1)

- *Step 1 – select "reference genes"*
  - **genes already known** to be involved in the given phenotype or in similar phenotypes (MimMiner; *van Driel et al., Eur. J. Hum. Genet., 2006*)
    - the method can be applied also to OMIM phenotypes of so far unknown molecular basis

Prioritization approach (1)

- *Step 1 – select "reference genes"*
  - **genes already known** to be involved in the given phenotype or in similar phenotypes (MimMiner; *van Driel et al., Eur. J. Hum. Genet., 2006*)
    - the method can be applied also to OMIM phenotypes of so far unknown molecular basis
- *Step 2 – for each reference gene g:*
  - build **ranked co-expression lists** by ranking all other genes according to their (Pearson) correlation with *g*

ROSARIO.PIRO@UNITO.IT

Prioritization approach (1)

- *Step 1 – select "reference genes"*
  - **genes already known** to be involved in the given phenotype or in similar phenotypes (MimMiner; *van Driel et al., Eur. J. Hum. Genet., 2006*)
    - the method can be applied also to OMIM phenotypes of so far unknown molecular basis
- *Step 2 – for each reference gene g:*
  - build **ranked co-expression lists** by ranking all other genes according to their (Pearson) correlation with *g*
- *Step 3 – for each positional candidate c:*
  - determine the **candidate's (relative) ranks** in the ranked co-expression lists of all reference genes

ROSARIO.PIRO@UNITO.IT

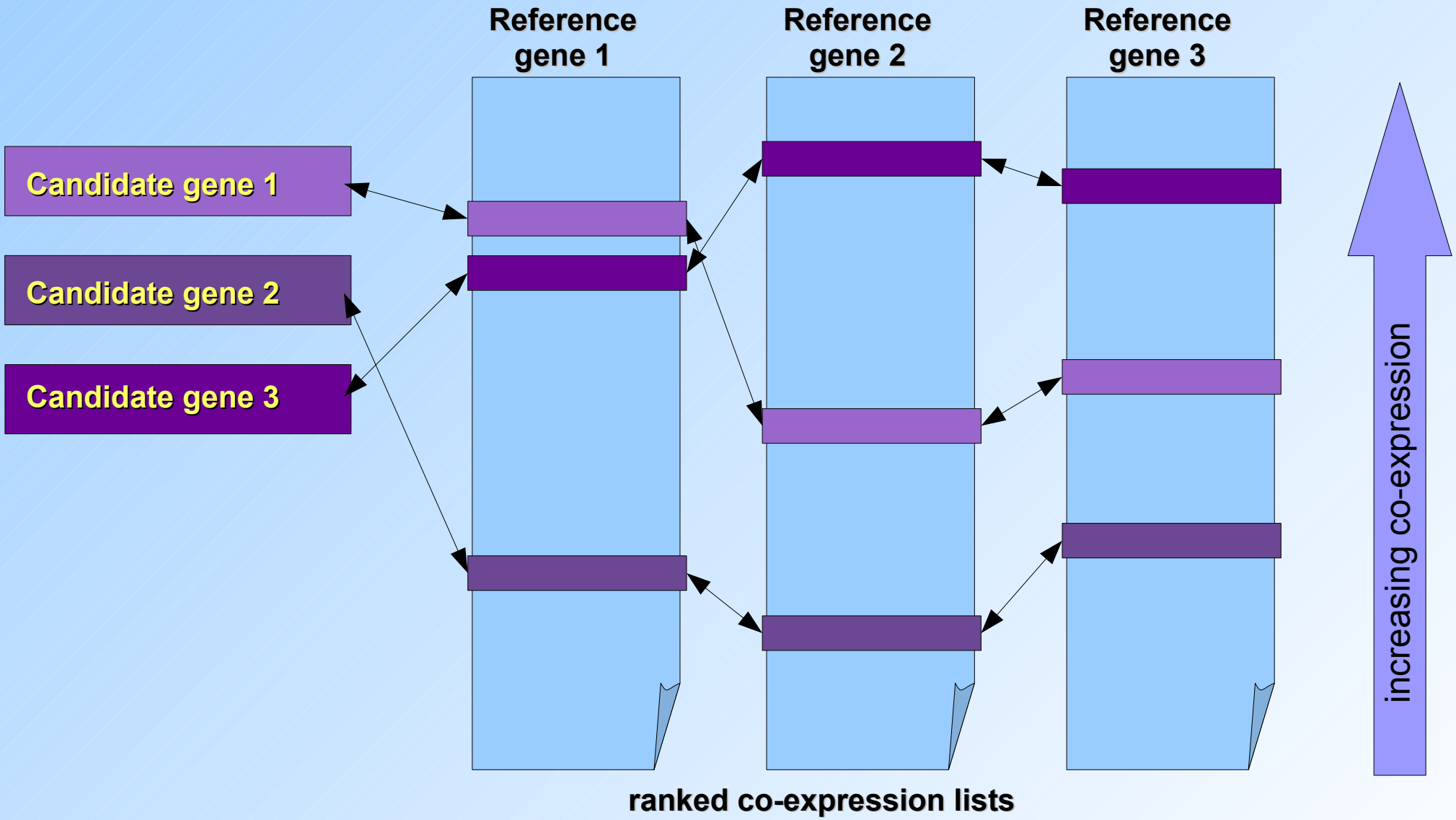
Prioritization approach (1)

- *Step 1 – select "reference genes"*
  - **genes already known** to be involved in the given phenotype or in similar phenotypes (MimMiner; *van Driel et al., Eur. J. Hum. Genet., 2006*)
    - the method can be applied also to OMIM phenotypes of so far unknown molecular basis
- *Step 2 – for each reference gene g:*
  - build **ranked co-expression lists** by ranking all other genes according to their (Pearson) correlation with *g*
- *Step 3 – for each positional candidate c:*
  - determine the **candidate's (relative) ranks** in the ranked co-expression lists of all reference genes
- *Step 4 – **score the candidates** (rank product)*

ROSARIO.PIRO@UNITO.IT

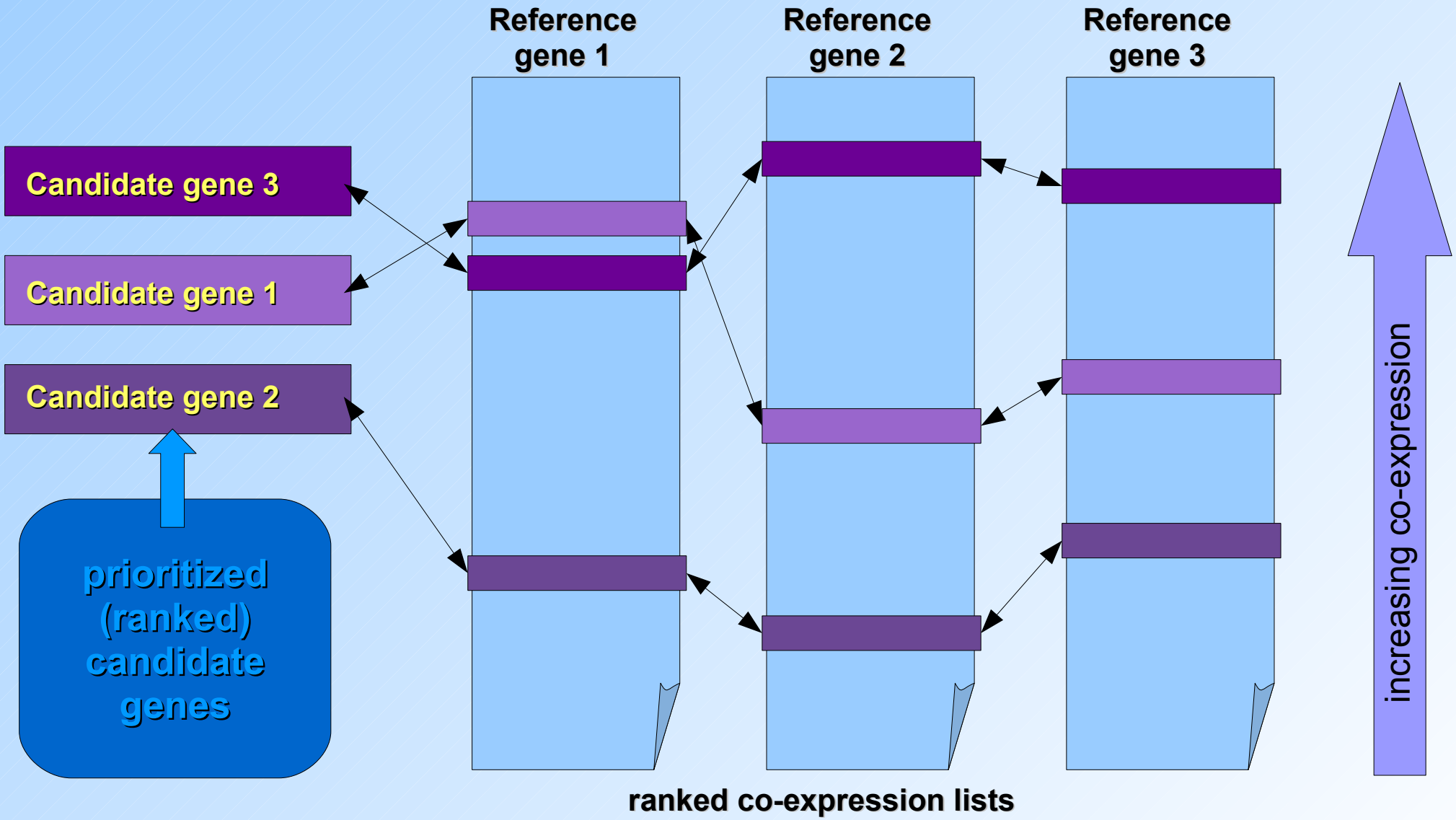
Prioritization approach (2)

Computational Biology Unit
CBU@MBC
Molecular Biotechnology Center
Università di Torino



Prioritization approach (2)

Computational Biology Unit
CBU@MBC
Molecular Biotechnology Center
Università di Torino



Leave-one-out validation

- Large-scale leave-one-out validations for ...
  - **… known CNS-related gene-phenotype associations** from different databases
    - Mouse Genome Database (MGD)
    - Online Mendelian Inheritance in Man (OMIM)
  - **… different sizes of loci**

 ROSARIO.PIRO@UNITO.IT

Leave-one-out validation

- Large-scale leave-one-out validations for ...
  - **… known CNS-related gene-phenotype associations** from different databases
    - Mouse Genome Database (MGD)
    - Online Mendelian Inheritance in Man (OMIM)
  - **… different sizes of loci**

- Results suggest that the method can be applied for both mouse and human phenotypes
  - Example: for large loci with up to 401 genes the phenotype-causing genes rank 1st-10th significantly more often than expected by chance
    - mouse phenotypes: $P = 4.79e\text{-}09$
    - human disorders: $P = 4.84e\text{-}06$

ROSARIO.PIRO@UNITO.IT

Case study: XLMR (1)

- Challenging, complex disorder:
  ~90 genes on Chr. X are known to be involved in some form of intellectual disability
  (*Cécz et al., Trends Genet., 2009*)

Case study: XLMR (1)

- Challenging, complex disorder:
  ~90 genes on Chr. X are known to be involved in some form of intellectual disability
  (*Cécz et al., Trends Genet., 2009*)

- A similar number probably remains to be identified
  (*Cécz et al., Trends Genet., 2009*)

Case study: XLMR (1)

- Challenging, complex disorder:
  ~90 genes on Chr. X are known to be involved in some form of intellectual disability
  (*Cécz et al., Trends Genet., 2009*)

- A similar number probably remains to be identified
  (*Cécz et al., Trends Genet., 2009*)

- Recent re-sequencing study of the exome of Chr. X
  (*Tarpey et al., Nat. Genet., 2009*)
  - 208 affected families
  - mutations in 3 novel (and several known) XLMR genes
  - many other mutations (also truncating), but in for most of the cases the genetic cause could not be reliably determined

ROSARIO.PIRO@UNITO.IT

Case study: XLMR (2)

- Evaluation:
  - candidate genes: all re-sequenced genes
  - reference genes: only genes known to be involved in similar phenotypes (none of the candidates from Chr. X)
  - can we “predict” known disease genes **pretending XLMR to be a phenotype of unknown molecular basis**?

Case study: XLMR (2)

- Evaluation:
  - candidate genes: all re-sequenced genes
  - reference genes: only genes known to be involved in similar phenotypes (none of the candidates from Chr. X)
  - can we “predict” known disease genes **pretending XLMR to be a phenotype of unknown molecular basis**?

- Prediction:
  - reference genes: XLMR genes
  - candidate genes: all other re-sequenced genes
  - can we find **promising novel candidates** for XLMR?

Case study: XLMR (2)

- Evaluation:
  - candidate genes: all re-sequenced genes
  - reference genes: only genes known to be involved in similar phenotypes (none of the candidates from Chr. X)
  - can we “predict” known disease genes **pretending XLMR to be a phenotype of unknown molecular basis**?

- Prediction:
  - reference genes: XLMR genes
  - candidate genes: all other re-sequenced genes
  - can we find **promising novel candidates** for XLMR?

- Important: evaluation and prediction have completely distinct sets of reference genes!

ROSARIO.PIRO@UNITO.IT

Case study: XLMR - Results

Computational Biology Unit
C B U @ M B C
Molecular Biotechnology Center
Università di Torino

Evaluation

1	BRWD3
2	IRAK1
3	SYP
4	BIRC4
5	MAGED1
6	MORF4L2
7	ZNF280C
8	SYN1
9	CXorf6
10	ATP6AP2
11	HCFC1
12	PJA1
13	NGFRAP1
14	FAM50A
15	HUWE1
16	GRIA3
17	PIGA
18	OGT
19	GNL3L
20	WDR40C

ROSARIO.PIRO@UNITO.IT

Case study: XLMR - Results

Evaluation

1	BRWD3
2	IRAK1
3	SYP
4	BIRC4
5	MAGED1
6	MORF4L2
7	ZNF280C
8	SYN1
9	CXorf6
10	ATP6AP2
11	HCFC1
12	PJA1
13	NGFRAP1
14	FAM50A
15	HUWE1
16	GRIA3
17	PIGA
18	OGT
19	GNL3L
20	WDR40C

- successful "re-discovery" of known **XLMR genes**

Case study: XLMR - Results

Computational Biology Unit
C B U @ M B C
Molecular Biotechnology Center
Università di Torino

Evaluation

1	BRWD3
2	IRAK1
3	SYP
4	BIRC4
5	MAGED1
6	MORF4L2
7	ZNF280C
8	SYN1
9	CXorf6
10	ATP6AP2
11	HCFC1
12	PJA1
13	NGFRAP1
14	FAM50A
15	HUWE1
16	GRIA3
17	PIGA
18	OGT
19	GNL3L
20	WDR40C

- successful "re-discovery" of known **XLMR genes**

Prediction

1	MORF4L2
2	PJA1
3	ZNF280C
4	MAGED1
5	MAGEE1
6	BIRC4
7	GRIPAP1
8	CXorf6
9	GNL3L
10	FAM50A
11	PGRMC1
12	GPM6B
13	IRAK1
14	HCFC1
15	PIGA
16	RPS4X
17	REPS2
18	ARMCX2
19	DRP2
20	MED14

Case study: XLMR - Results

Computational Biology Unit
C B U @ M B C
Molecular Biotechnology Center
Università di Torino

Evaluation

2	IRAK1
4	BIRC4
5	MAGED1
6	MORF4L2
7	ZNF280C
9	CXorf6
11	HCFC1
12	PJA1
13	NGFRAP1
14	FAM50A
17	PIGA
18	OGT
19	GNL3L
20	WDR40C

- successful "re-discovery" of known **XLMR genes**

Prediction

1	MORF4L2
2	PJA1
3	ZNF280C
4	MAGED1
5	MAGEE1
6	BIRC4
7	GRIPAP1
8	CXorf6
9	GNL3L
10	FAM50A
11	PGRMC1
12	GPM6B
13	IRAK1
14	HCFC1
15	PIGA
16	RPS4X
17	REPS2
18	ARMCX2
19	DRP2
20	MED14

ROSARIO.PIRO@UNITO.IT

Case study: XLMR - Results

Computational Biology Unit
CBU@MBC
Molecular Biotechnology Center
Università di Torino

Evaluation

2	IRAK1
4	BIRC4
5	MAGED1
6	MORF4L2
7	ZNF280C
9	CXorf6
11	HCFC1
12	PJA1
13	NGFRAP1
14	FAM50A
17	PIGA
18	OGT
19	GNL3L
20	WDR40C

- successful "re-discovery" of known **XLMR genes**
- very strong **overlap** between evaluation and prediction (P=6.62x10^{-14})

Prediction

1	MORF4L2
2	PJA1
3	ZNF280C
4	MAGED1
5	MAGEE1
6	BIRC4
7	GRIPAP1
8	CXorf6
9	GNL3L
10	FAM50A
11	PGRMC1
12	GPM6B
13	IRAK1
14	HCFC1
15	PIGA
16	RPS4X
17	REPS2
18	ARMCX2
19	DRP2
20	MED14

ROSARIO.PIRO@UNITO.IT

Case study: XLMR - Results

Computational Biology Unit
CBU@MBC
Molecular Biotechnology Center
Università di Torino

Evaluation

2	IRAK1
4	BIRC4
5	MAGED1
6	MORF4L2
7	ZNF280C
9	CXorf6
11	HCFC1
12	PJA1
13	NGFRAP1
14	FAM50A
17	PIGA
18	OGT
19	GNL3L
20	WDR40C

- successful "re-discovery" of known **XLMR genes**
- very strong **overlap** between evaluation and prediction (P=6.62x10^{-14})
- Tarpey et al. have found **mutations** of highly conserved residues

Prediction

1	MORF4L2
2	PJA1
3	ZNF280C
4	MAGED1
5	MAGEE1
6	BIRC4
7	GRIPAP1
8	CXorf6
9	GNL3L
10	FAM50A
11	PGRMC1
12	GPM6B
13	IRAK1
14	HCFC1
15	PIGA
16	RPS4X
17	REPS2
18	ARMCX2
19	DRP2
20	MED14

ROSARIO.PIRO@UNITO.IT

Limitations

- Proof-of-concept study
  - no extensive comparison with other data sources
  - no optimization
    - alternative scoring functions instead of the rank product?

Limitations

- Proof-of-concept study
  - no extensive comparison with other data sources
  - no optimization
    - alternative scoring functions instead of the rank product?
- Spatial information not fully exploited
  - alternative measures for the similarity of expression profiles? use HRC instead of the Pearson correlation coefficient? (HRC = histogram-row-column; *Liu et al., BMC Sys. Biol., 2007*)

Limitations

- Proof-of-concept study
  - no extensive comparison with other data sources
  - no optimization
    - alternative scoring functions instead of the rank product?
- Spatial information not fully exploited
  - alternative measures for the similarity of expression profiles? use HRC instead of the Pearson correlation coefficient? (HRC = histogram-row-column; *Liu et al., BMC Sys. Biol., 2007*)
- Spatial 3D expression data remain an exception
  - Human Brain Atlas (Allen Institute) to be completed in 2013
    - our work can be considered a pioneer study towards a direct application of human 3D expression data
  - so far only “normal” brain tissue; some potentially interesting applications would require disease conditions
    - “differential spatial expression”?

Summary

- We have shown that spatially mapped (3D) gene expression data can be successfully exploited for candidate gene prioritization
  - mouse CNS-related phenotypes
  - human CNS-related Mendelian disorders

ROSARIO.PIRO@UNITO.IT

Summary

- We have shown that spatially mapped (3D) gene expression data can be successfully exploited for candidate gene prioritization
  - mouse CNS-related phenotypes
  - human CNS-related Mendelian disorders

- Application to XMLR:
  - we successfully “re-discovered” known XLMR genes
  - the robustness of our prediction results led us to suggest some promising novel candidates

ROSARIO.PIRO@UNITO.IT

Acknowledgments

- Team at CBU@MBC:
  - **Ugo Ala**
  - **Ivan Molineris**
  - **Paolo Provero**
  - **Ferdinando Di Cunto**
  - Christian Damasco
  - Antonio Lembo
  - Eva Pinatel
  - Elena Grassi
  - Mario Giacobini
- Special thanks to
  - Han Brunner
  - Martin Oti

  Nijmegen Centre for Molecular Life Sciences (The Netherlands)

Advertisement

Computational Biology Unit
CBU@MBC
Molecular Biotechnology Center
Università di Torino

Poster

I-35

 ROSARIO.PIRO@UNITO.IT

Questions?

01010101011111010010100101010100010101111000010110100101011100110101 1

Computational Biology Unit
CBU@MBC
Molecular Biotechnology Center
Università di Torino

ACGTCAGCTAGCTAGCTAGTCAGTCGATCGATGTGTACATGCAATCGTAGCTAGCTAGCTAGCTA
TGCATGCAACTGATGCATGCATGCTGATCGATGCATGCT

ROSARIO.PIRO@UNITO.IT

CTAGC

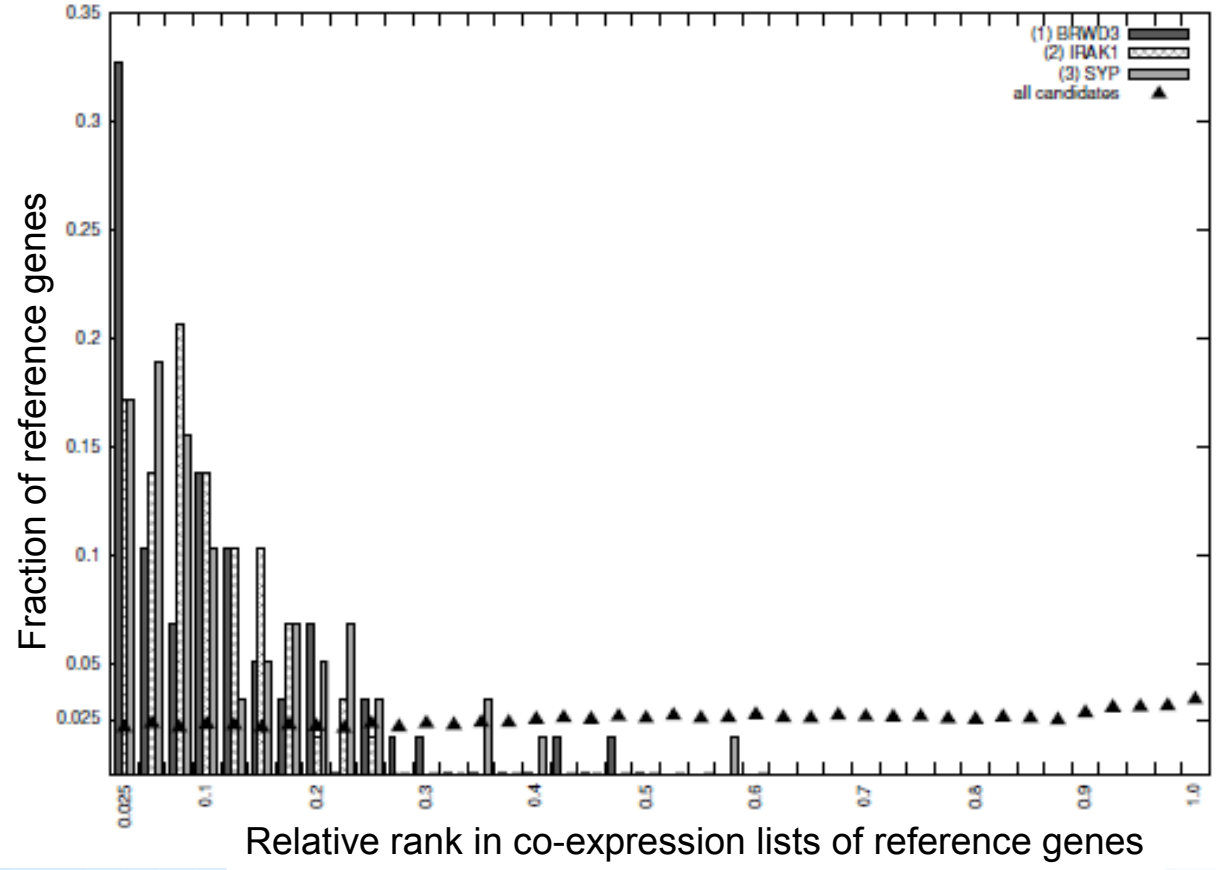
Case study: XLMR - Results

Computational Biology Unit
C B U @ M B C
Molecular Biotechnology Center
Università di Torino

Evaluation

1	BRWD3
2	IRAK1
3	SYP
4	BIRC4
5	MAGED1
6	MORF4L2
7	ZNF280C
8	SYN1
9	CXorf6
10	ATP6AP2
11	HCFC1
12	PJA1
13	NGFRAP1
14	FAM50A
15	HUWE1
16	GRIA3
17	PIGA
18	OGT
19	GNL3L
20	WDR40C

- successful "re-discovery" of known **XLMR genes**



ROSARIO.PIRO@UNITO.IT