Prof. Dr. Rosario M. Piro
FB Mathematik und Informatik

Freie Universität Berlin

CONFLUENCE 2017 – Noida, India

Computational medicine: identification of causes of genetic disorders

Rosario M. Piro – r.piro@fu-berlin.de

Freie Universität Berlin

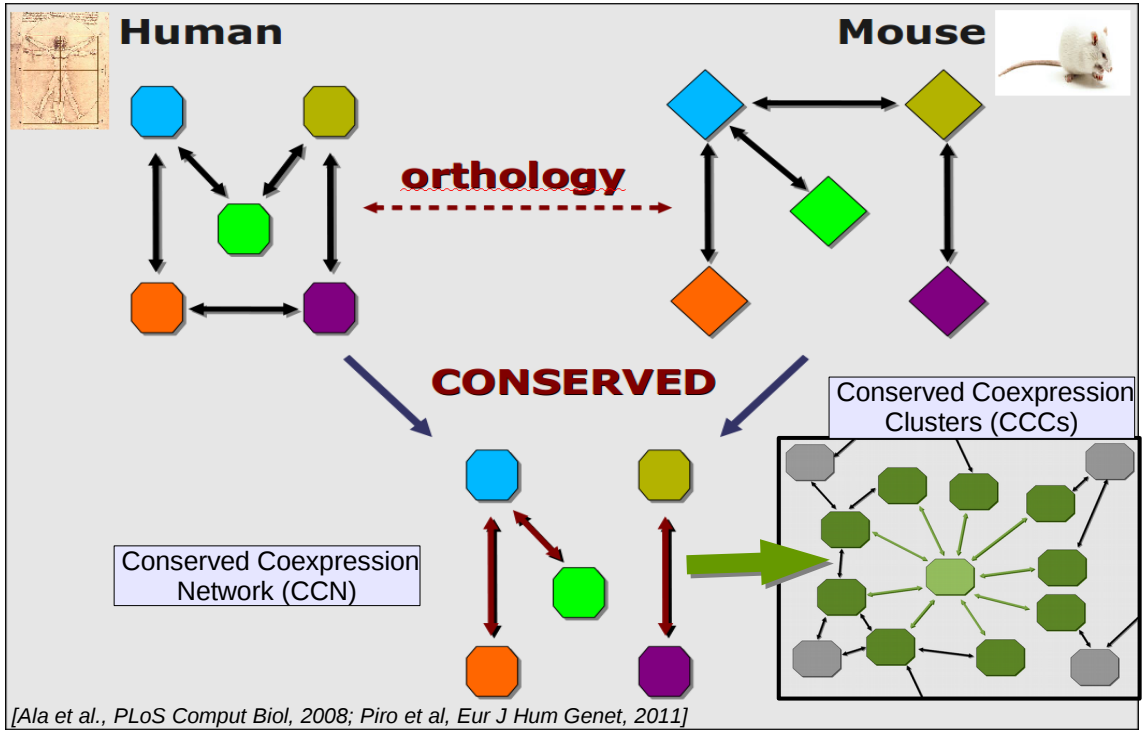
Prof. Dr. Rosario M. Piro
FB Mathematik und Informatik

Freie Universität Berlin

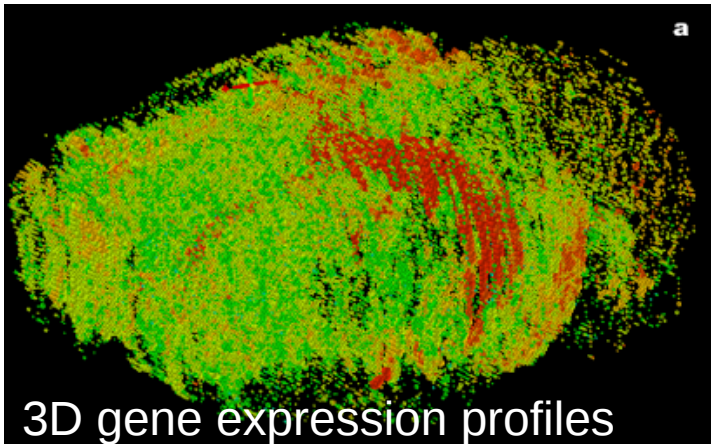
Computational Neuropathology

Identification of possible molecular causes of hereditary disorders

- Development of computational methods for “disease gene prediction“
- Focus on diseases affecting the central nervous system
  - Intellectual disability
  - Alzheimer's disease
  - ...



© Allen Institute



Prof. Dr. Rosario M. Piro
FB Mathematik und Informatik

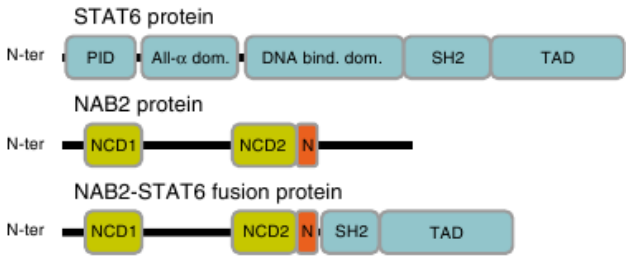
Freie Universität Berlin

Computational Oncology

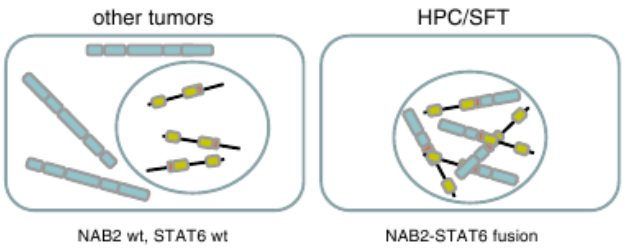
Identification of possible molecular causes of cancer

- Analysis of next generation sequencing (NGS) to identify genetic and genomic alterations, e.g.,
  - NAB2-STAT6 fusion gene in hemgiopericytomas of the central nervous system (can be used for diagnostic purpose)
  - Single point mutations in cancer-related genes

[Schweizer et al., Acta Neuropathol, 2013]



- Including (but not limited to) brain tumors



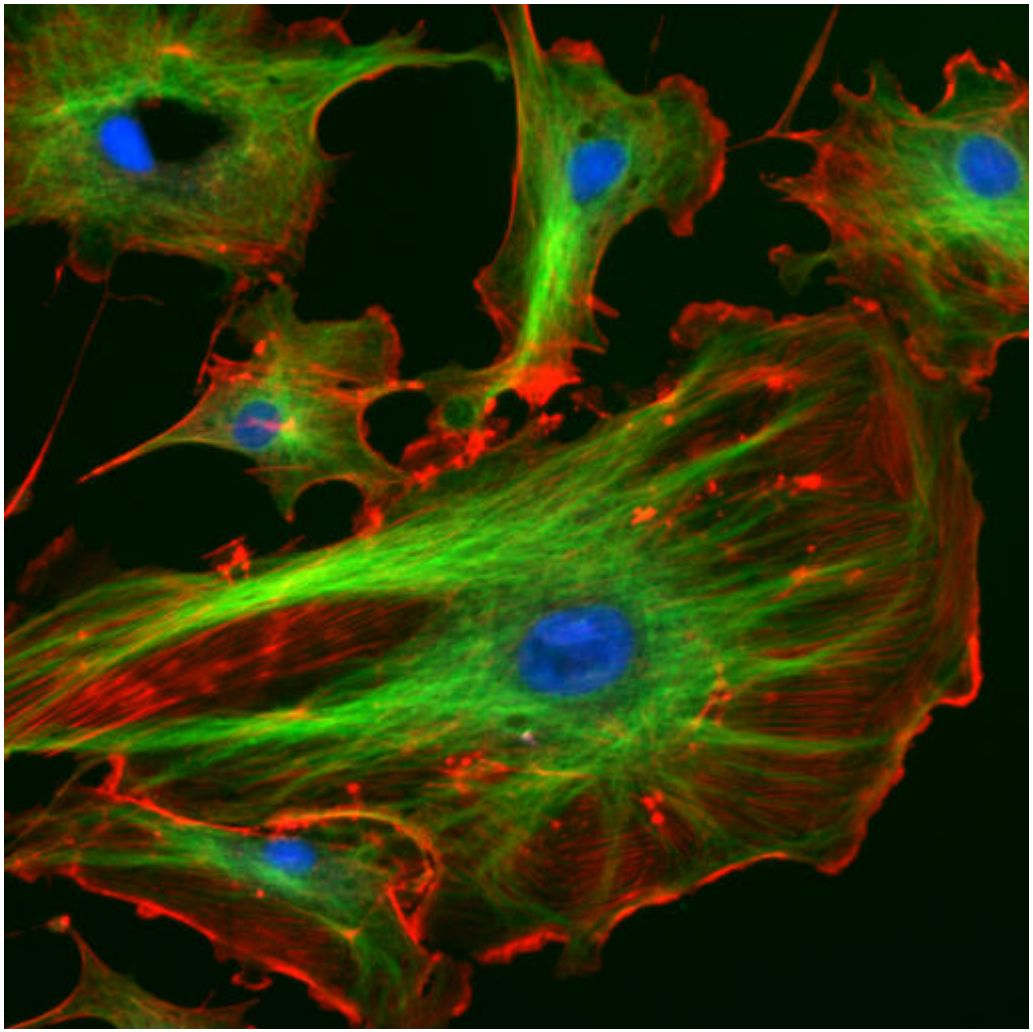
Freie Universität Berlin

Part I

Some molecular biology ...

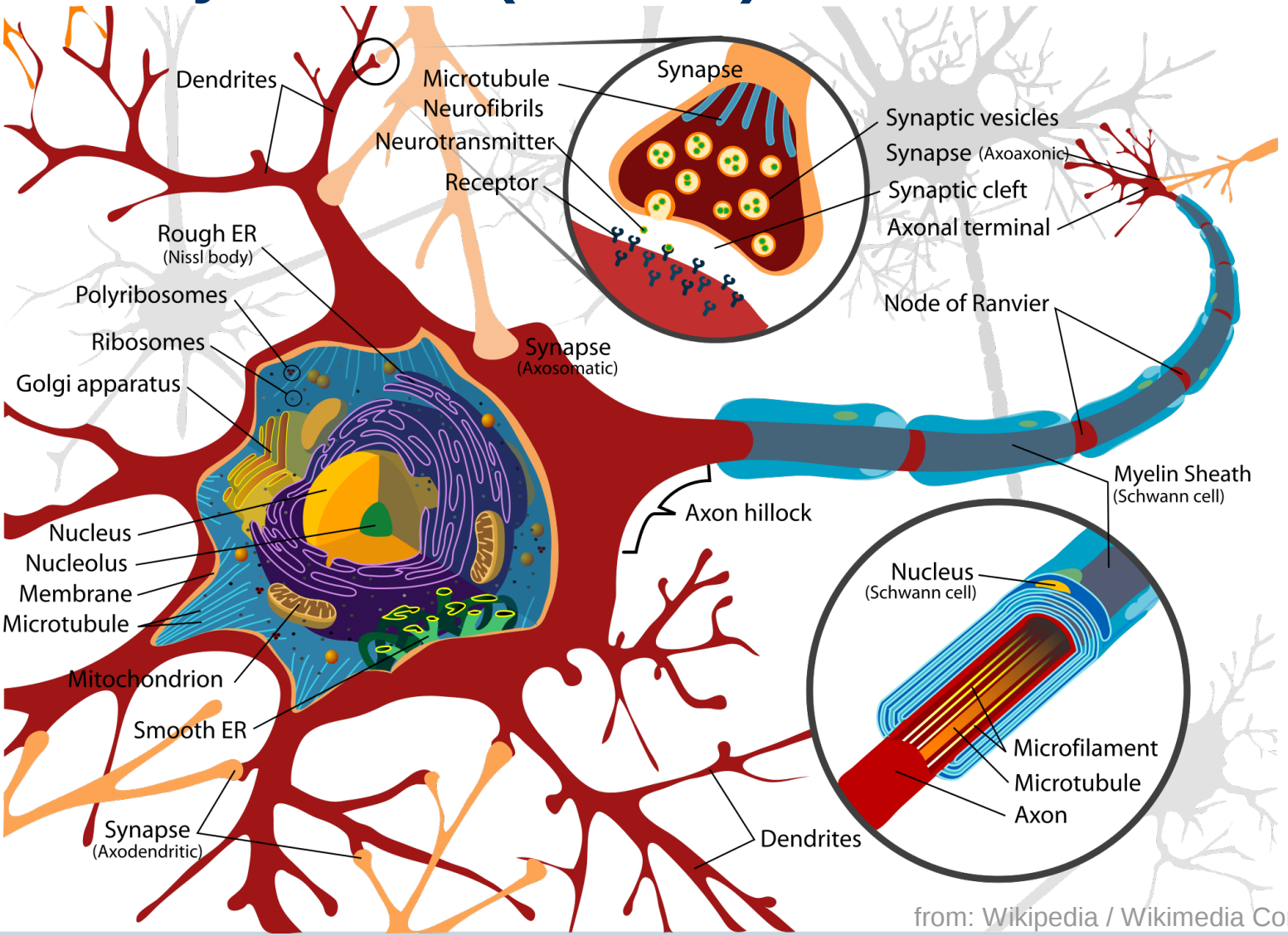
Basic notions about genes and their expression,
and proteins and their function

The eukaryotic cell



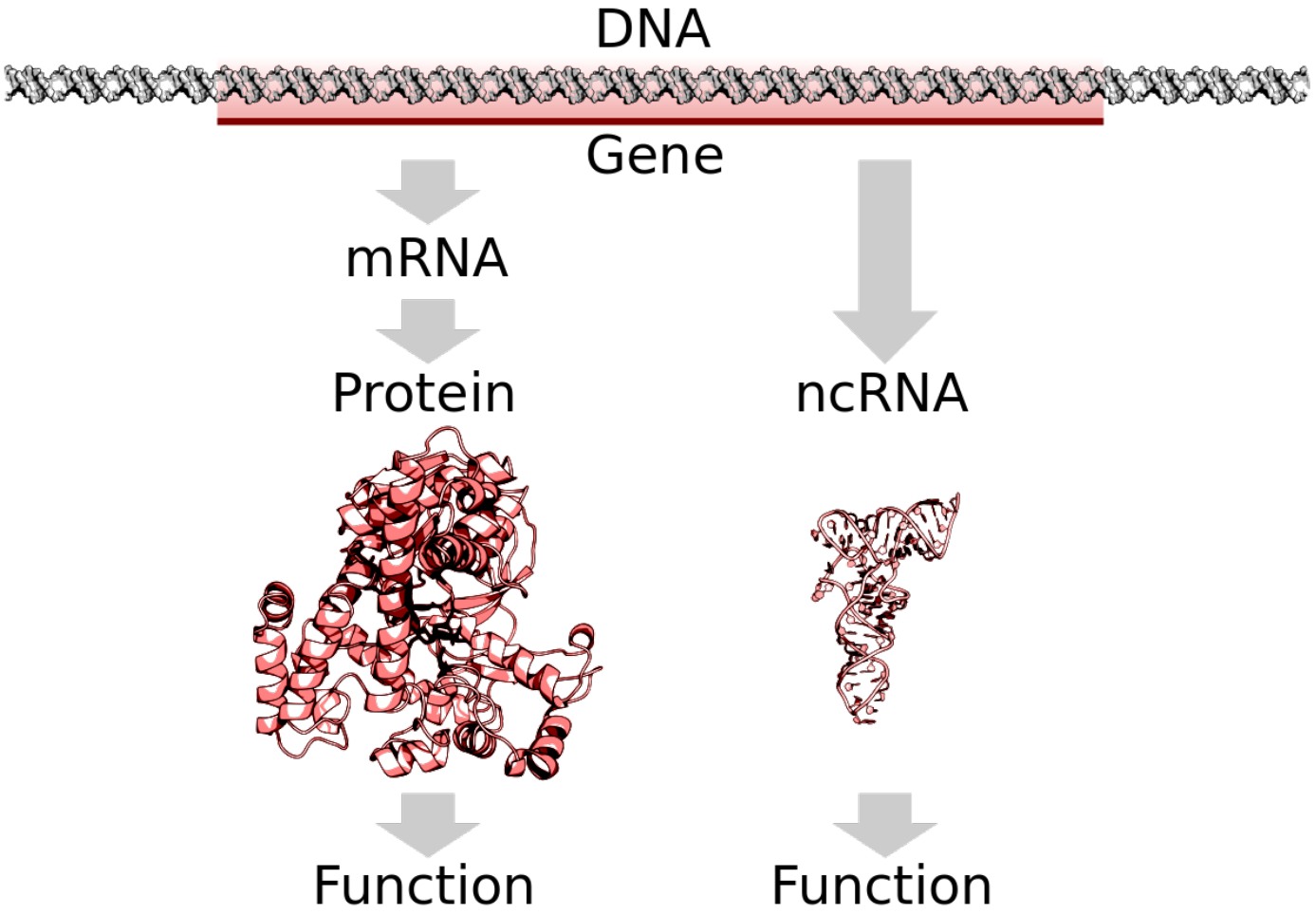
from: Wikipedia / Wikimedia Commons

The eukaryotic cell (neuron)



from: Wikipedia / Wikimedia Commons

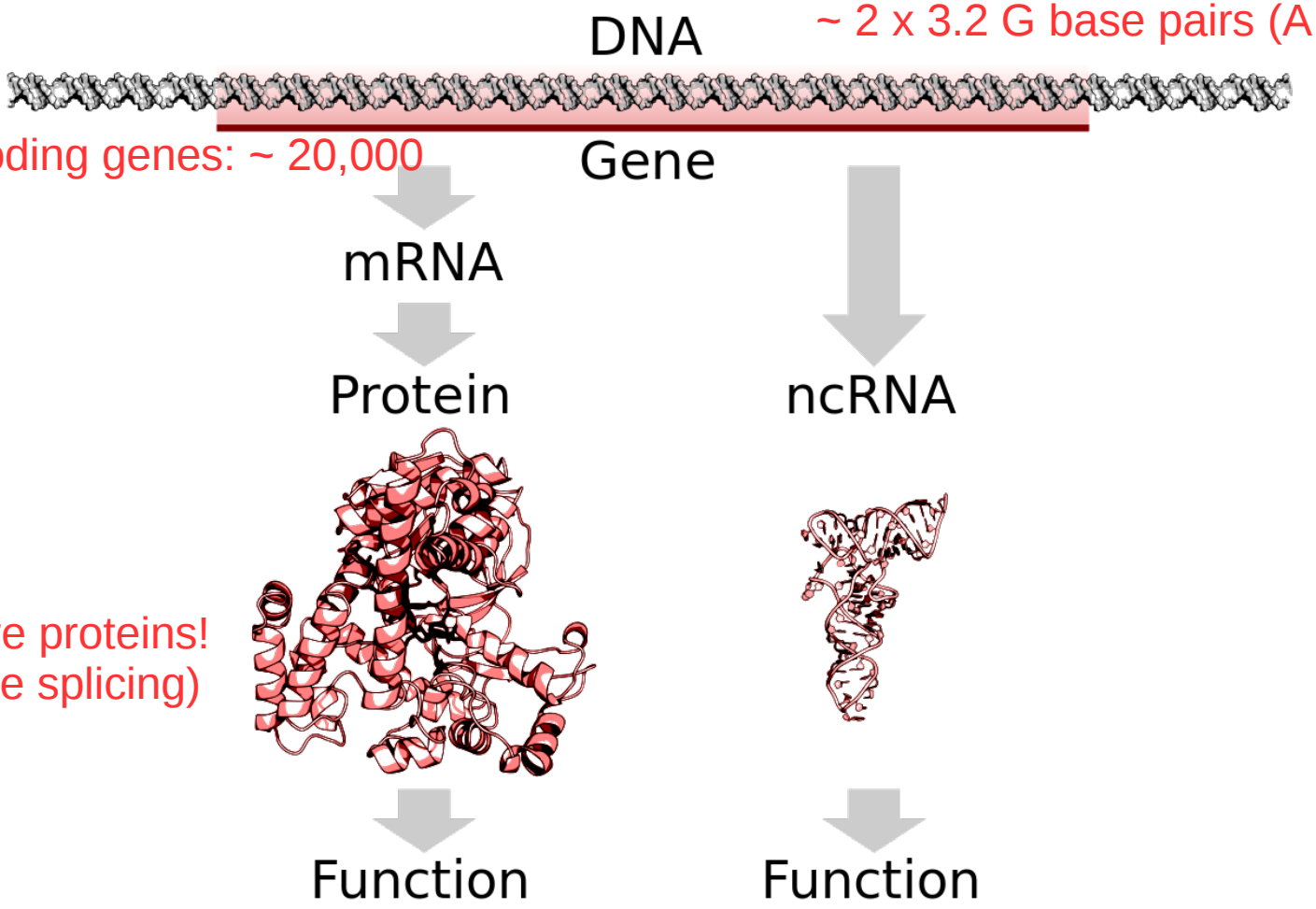
Genes



from: Wikipedia / Wikimedia Commons

Genes

Human genome:
~ 2 x 3.2 G base pairs (A,C,G,T)

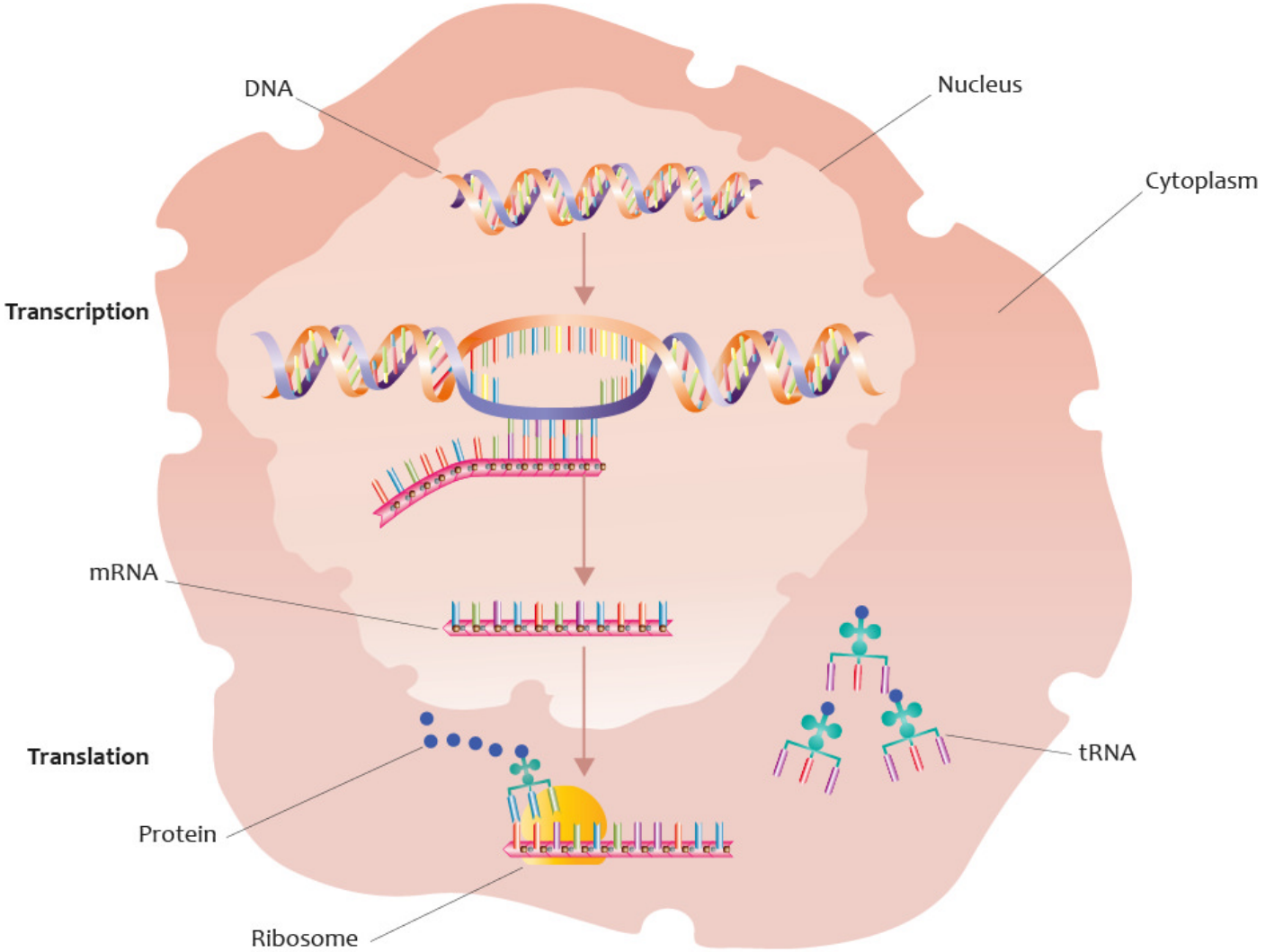


Protein coding genes: ~ 20,000

Many more proteins!
(alternative splicing)

from: Wikipedia / Wikimedia Commons

From gene to protein



from: Wikipedia / Wikimedia Commons

From gene to protein

GTGCATCTGACTCCTGAGGAGAAG

CACGTAGACTGAGGACTCCTCTTC

DNA

(transcription)

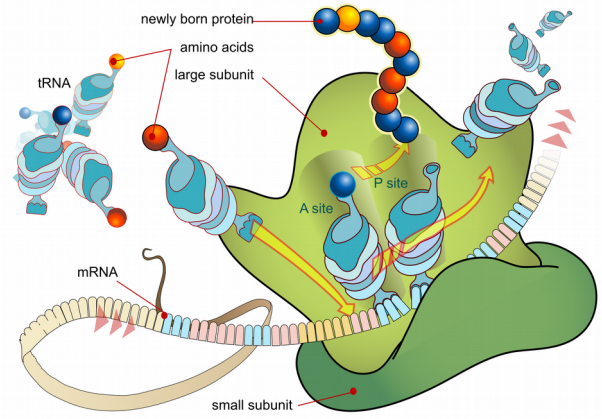
GUGCAUCUGACUCCUGAGGAGAAG

RNA

(translation)

V H L T P E E K

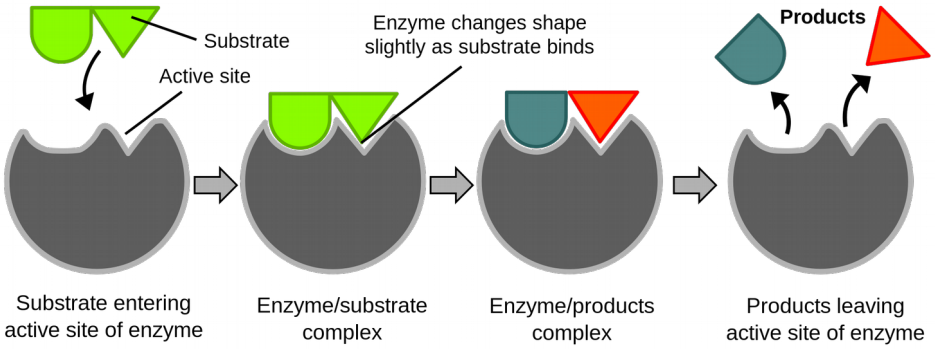
protein



from: Wikipedia / Wikimedia Commons

Protein function (examples)

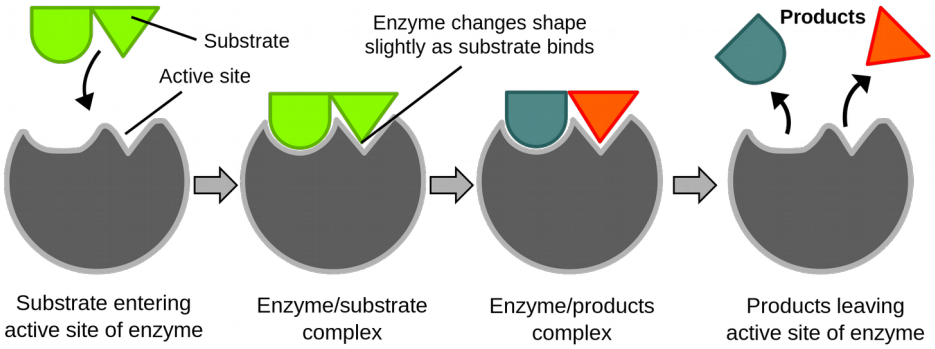
enzymes



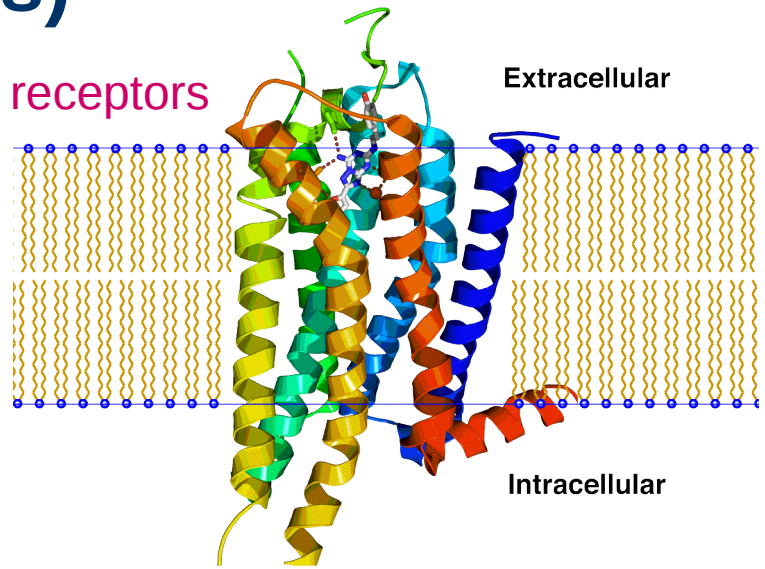
from: Wikipedia / Wikimedia Commons

Protein function (examples)

enzymes



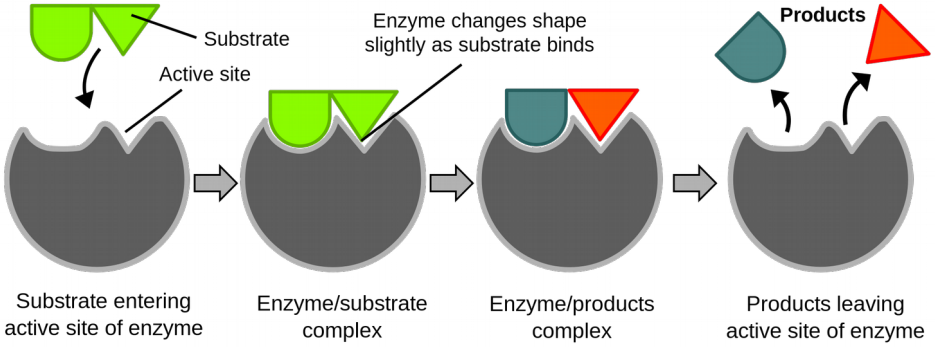
receptors



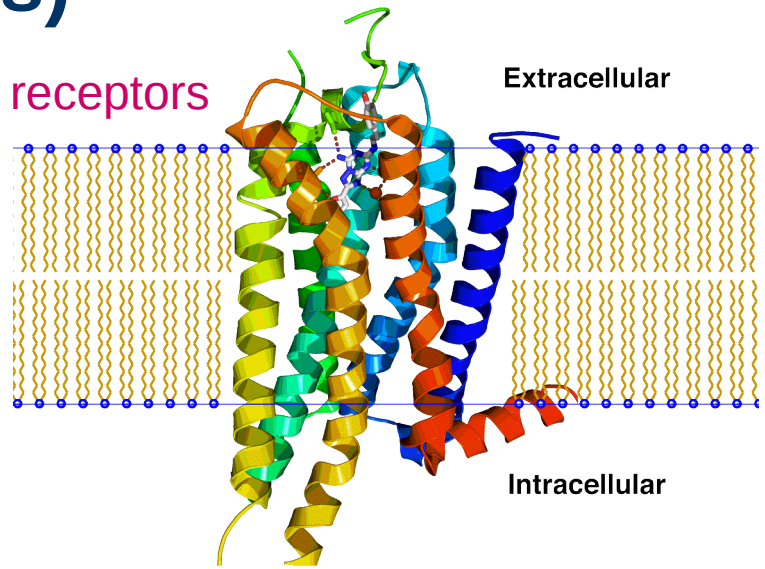
from: Wikipedia / Wikimedia Commons

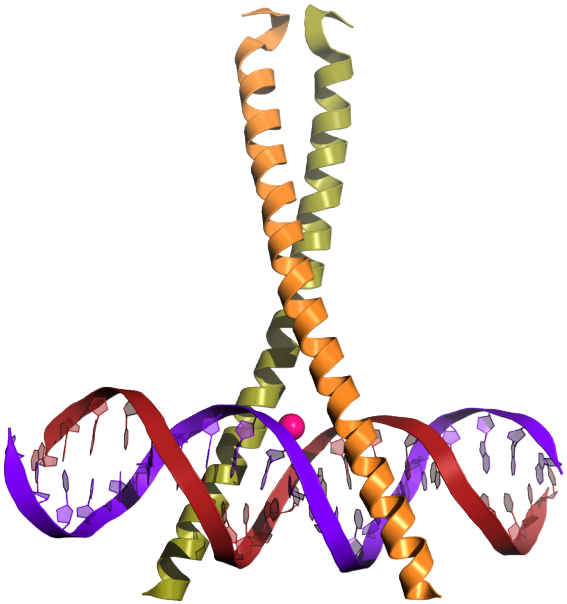
Protein function (examples)

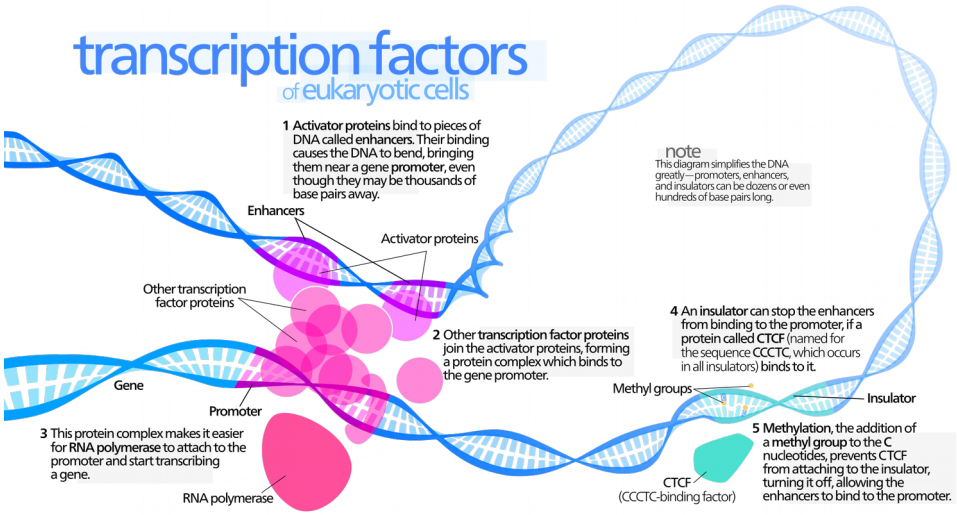
enzymes



receptors

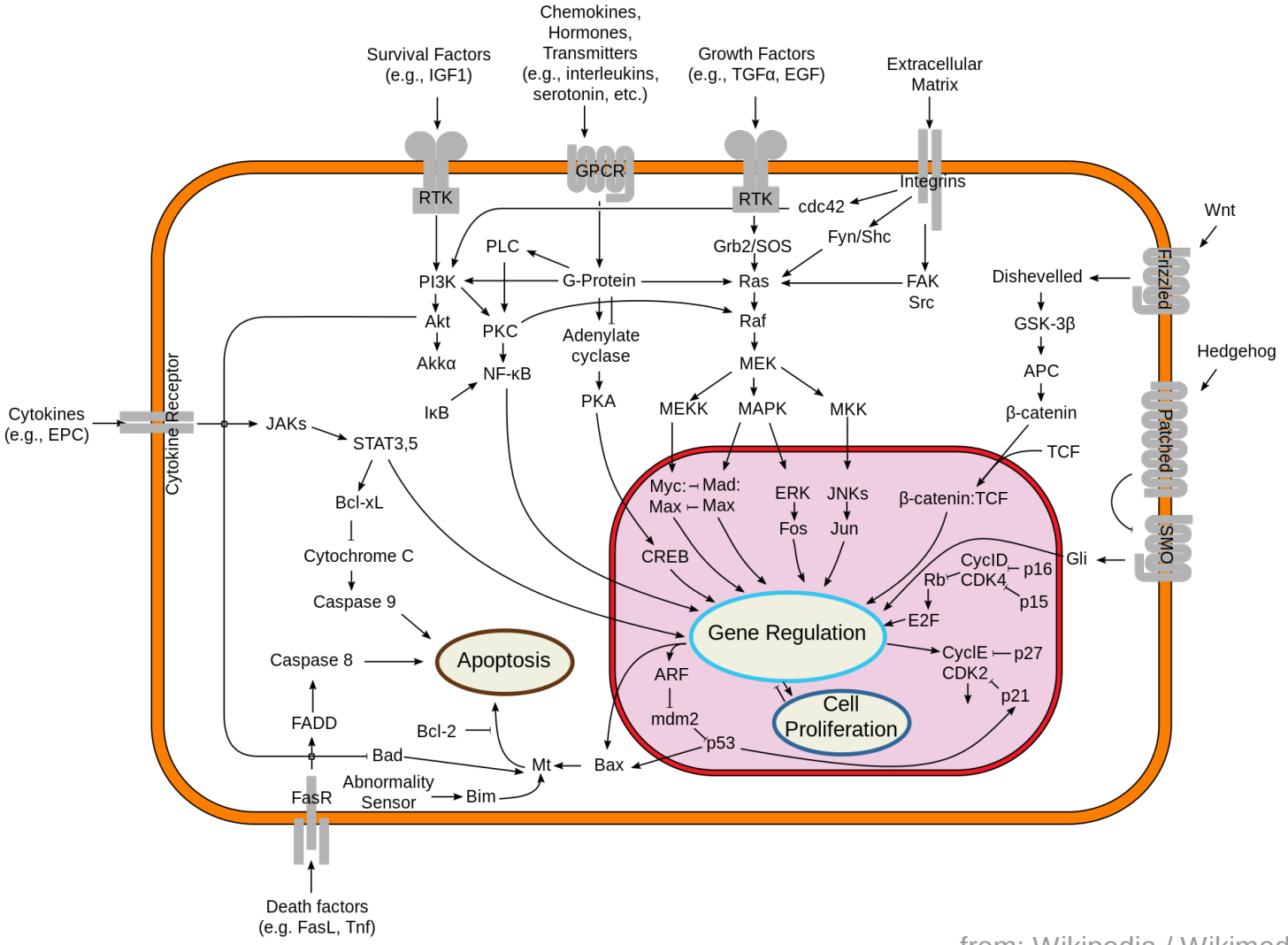






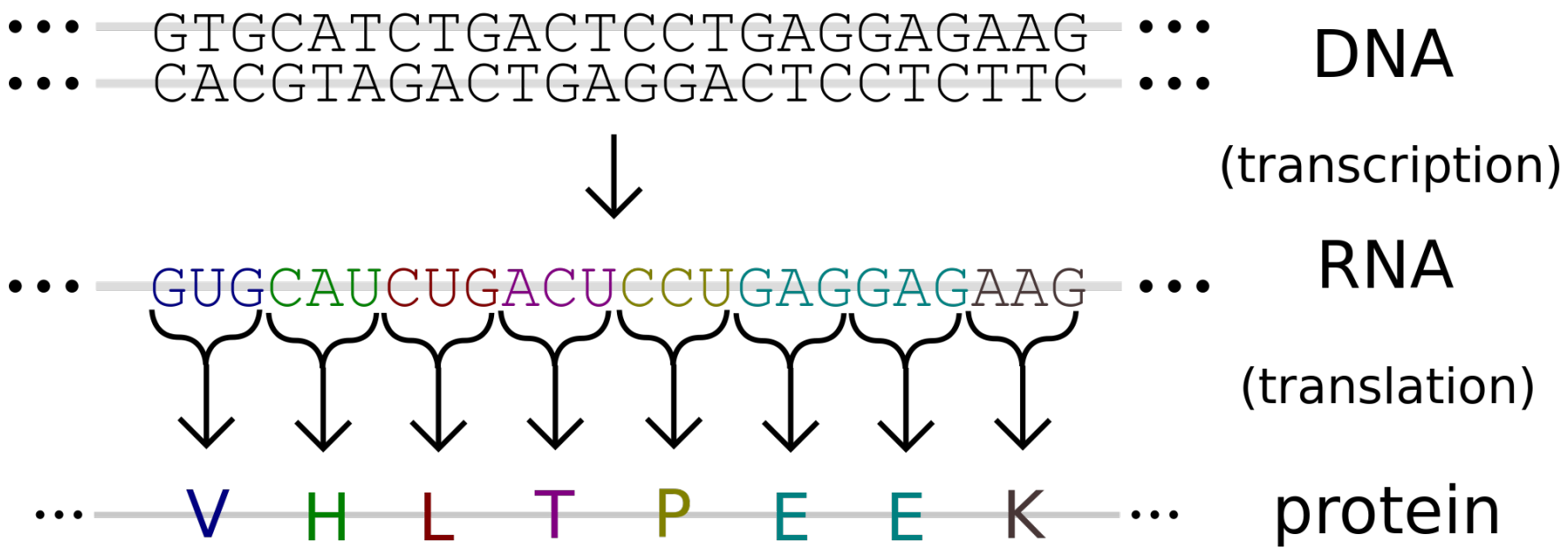
from: Wikipedia / Wikimedia Commons

Signal transduction and gene regulation



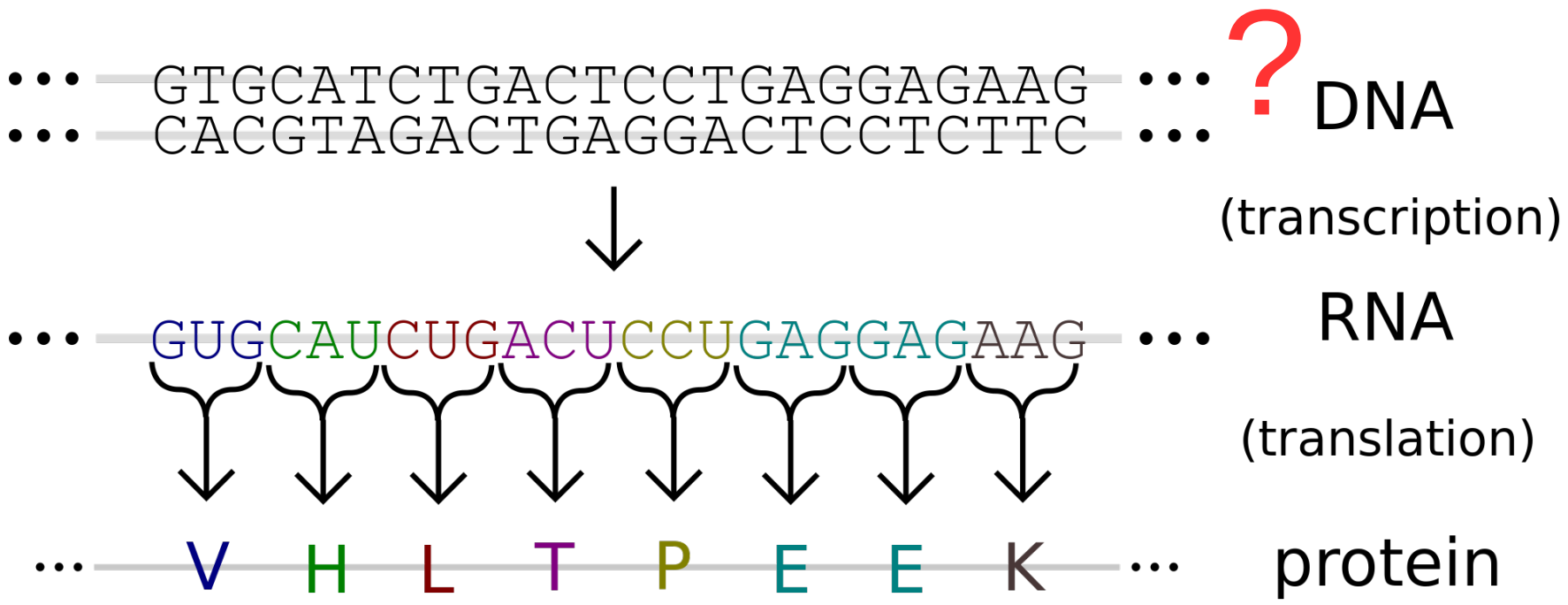
from: Wikipedia / Wikimedia Commons

Gene expression



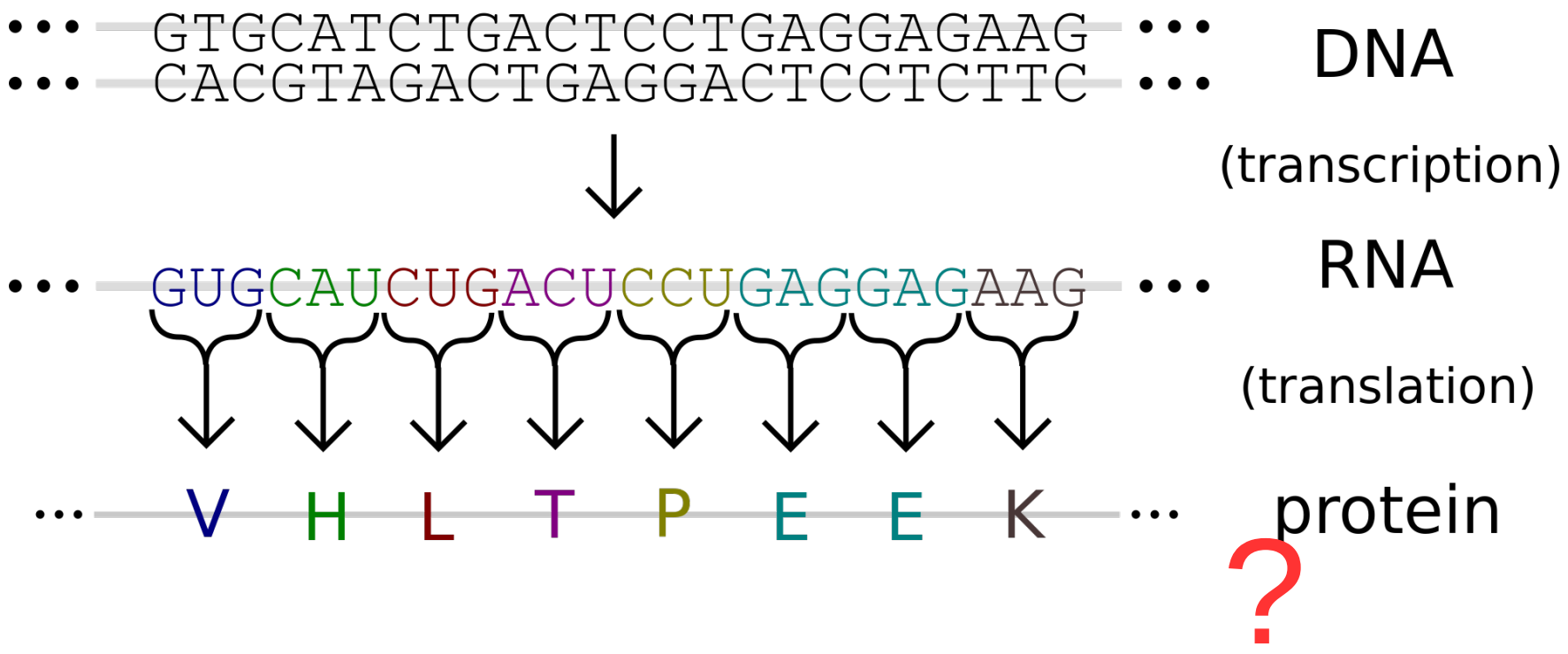
from: Wikipedia / Wikimedia Commons

Gene expression



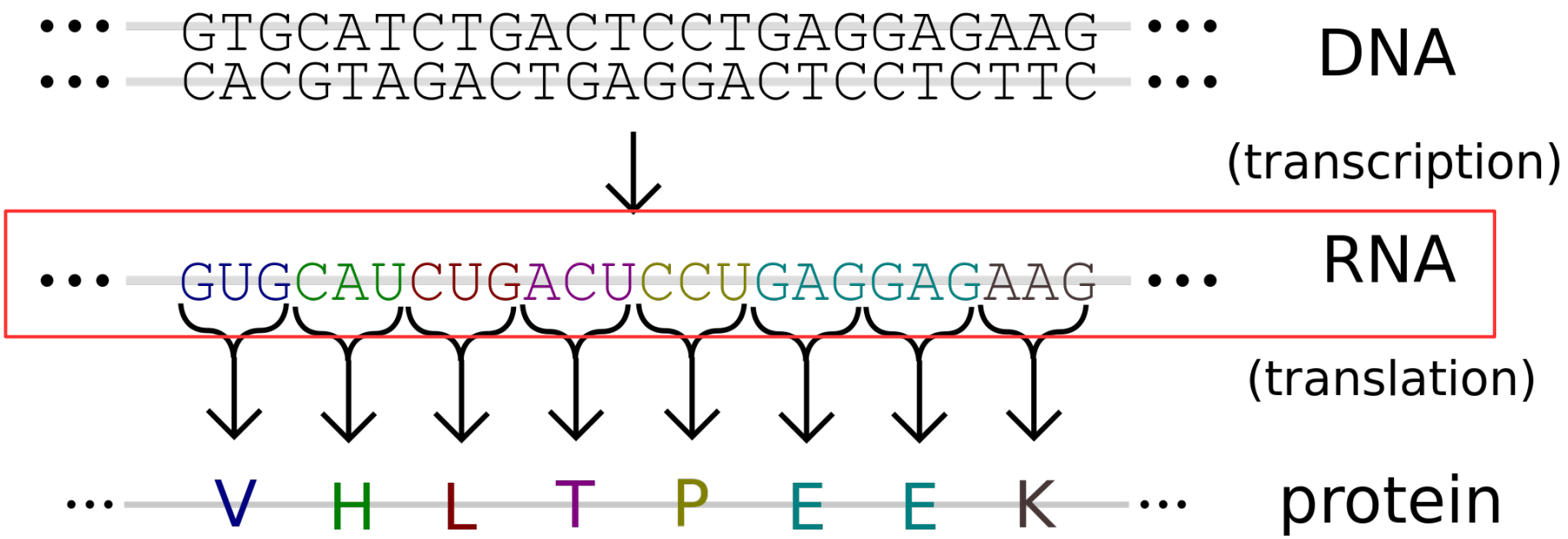
from: Wikipedia / Wikimedia Commons

Gene expression



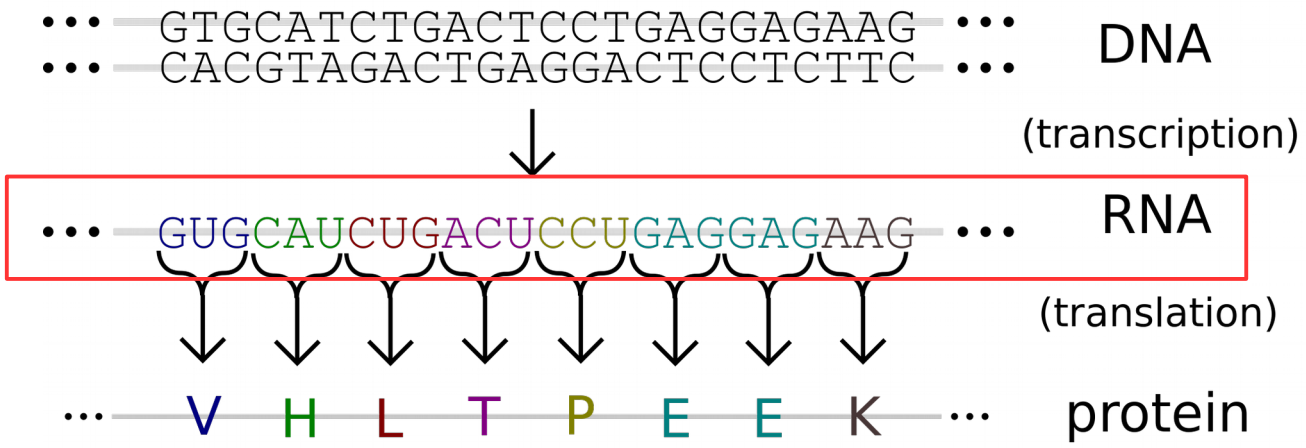
from: Wikipedia / Wikimedia Commons

Gene expression



from: Wikipedia / Wikimedia Commons

Gene expression



	S1	S2	S3	S4	...									S*n*
G1	1.5	4.7	7.2	1.9	0.4	...								
G2	2.0	6.8	7.1	1.5	8.1	...								
G3	7.8	3.1	0.5	0.8	6.3	...								
G4	...													
...														
G*m*														

m x *n* gene expression matrix

Part II

Computational Neuropathology

Prediction of genes associated with genetically determined disorders of the central nervous system (CNS)
Focus on mental retardation

Disease gene prediction: motivation

Disease gene prediction for hereditary disorders

Disease gene prediction: motivation

Disease gene prediction for hereditary disorders



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

Identified chromosomal regions often contain hundreds of “positional candidate genes”

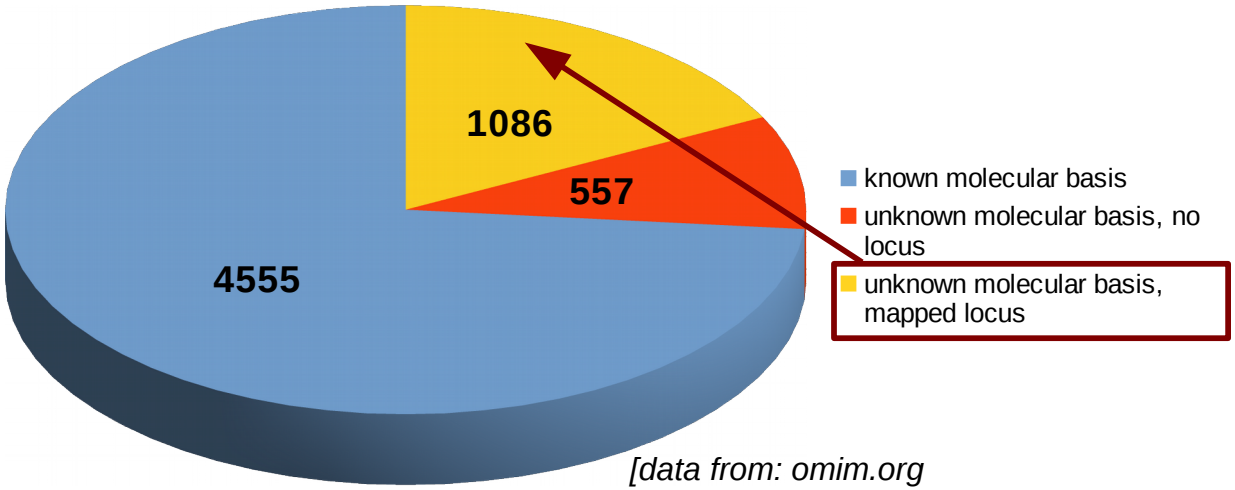
Disease gene prediction: motivation

Disease gene prediction for hereditary disorders



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

Identified chromosomal regions often contain hundreds of “positional candidate genes”



[data from: omim.org
Online Mendelian Inheritance in Man (OMIM), Sept. 27, 2015]

Disease gene prediction: general principles

“Guilt by assosciation“:

- Which candidate genes have a functional link to known disease genes?

Disease gene prediction: general principles

“Guilt by assosciation“:

- Which candidate genes have a functional link to known disease genes?

Inferring functional relationships:

- Functional annotations, shared protein domains, known homology, ...
- Protein-protein interaction networks
- Co-expression (correlated expression levels, e.g. in a set of patients or tissues)

Disease gene prediction: general principles

"Guilt by assosciation":

- Which candidate genes have a functional link to known disease genes?

Inferring functional relationships:

- Functional annotations, shared protein domains, known homology, ...
- Protein-protein interaction networks
- Co-expression (correlated expression levels, e.g. in a set of patients or tissues)

	S1	**S2**	**S3**	**S4**	**...**									**S*n***
G1	1.5	4.7	7.2	1.9	0.4	...								
G2	2.0	6.8	7.1	1.5	8.1	...								
G3	7.8	3.1	0.5	0.8	6.3	...								
G4	...													
...														
G*m*														

m x *n* gene expression matrix

Gene co-expression networks

	S_1	S_2	S_3
G_1	43.26	40.89	5.05
G_2	166.6	41.87	136.65
G_3	12.53	39.55	42.09
G_4	28.77	191.92	236.56
G_5	114.7	79.7	99.76
G_6	119.1	80.57	114.59
G_7	118.9	156.69	186.95
G_8	3.76	2.48	136.78
G_9	32.73	11.99	118.8
G_{10}	17.46	56.11	21.41

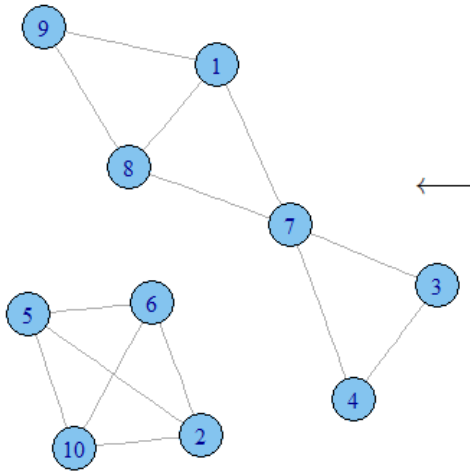
Gene expression values

$|r(G_i, G_j)|$

Pearson correlation

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	1.00	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	1.00	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	1.00	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	1.00	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	1.00	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	1.00	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	1.00	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	1.00	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	1.00	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	1.00

Similarity (Co-expression) score



	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

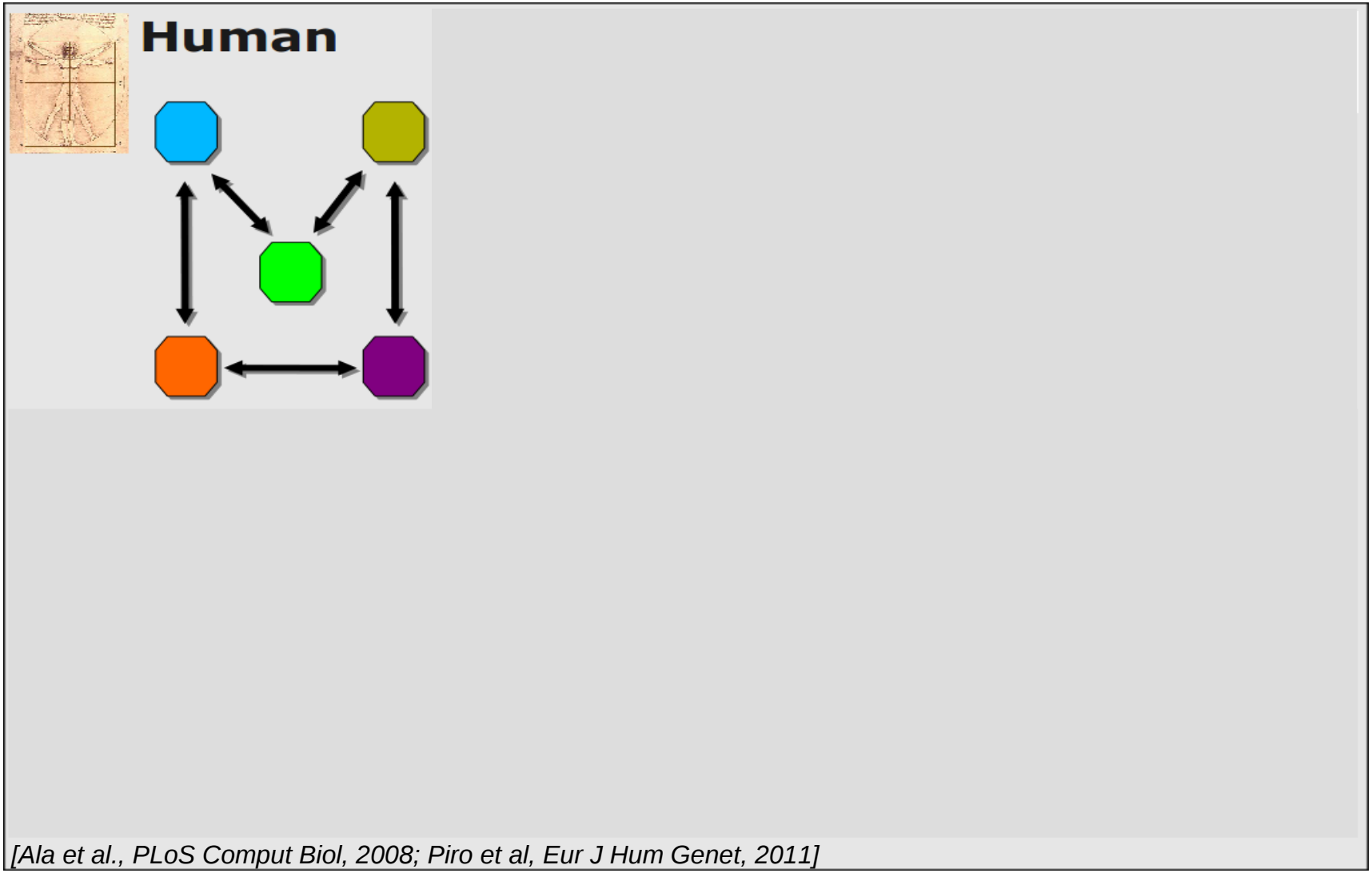
$|r(G_i, G_j)| >= 0.8$

Significance threshold

Network adjacency matrix

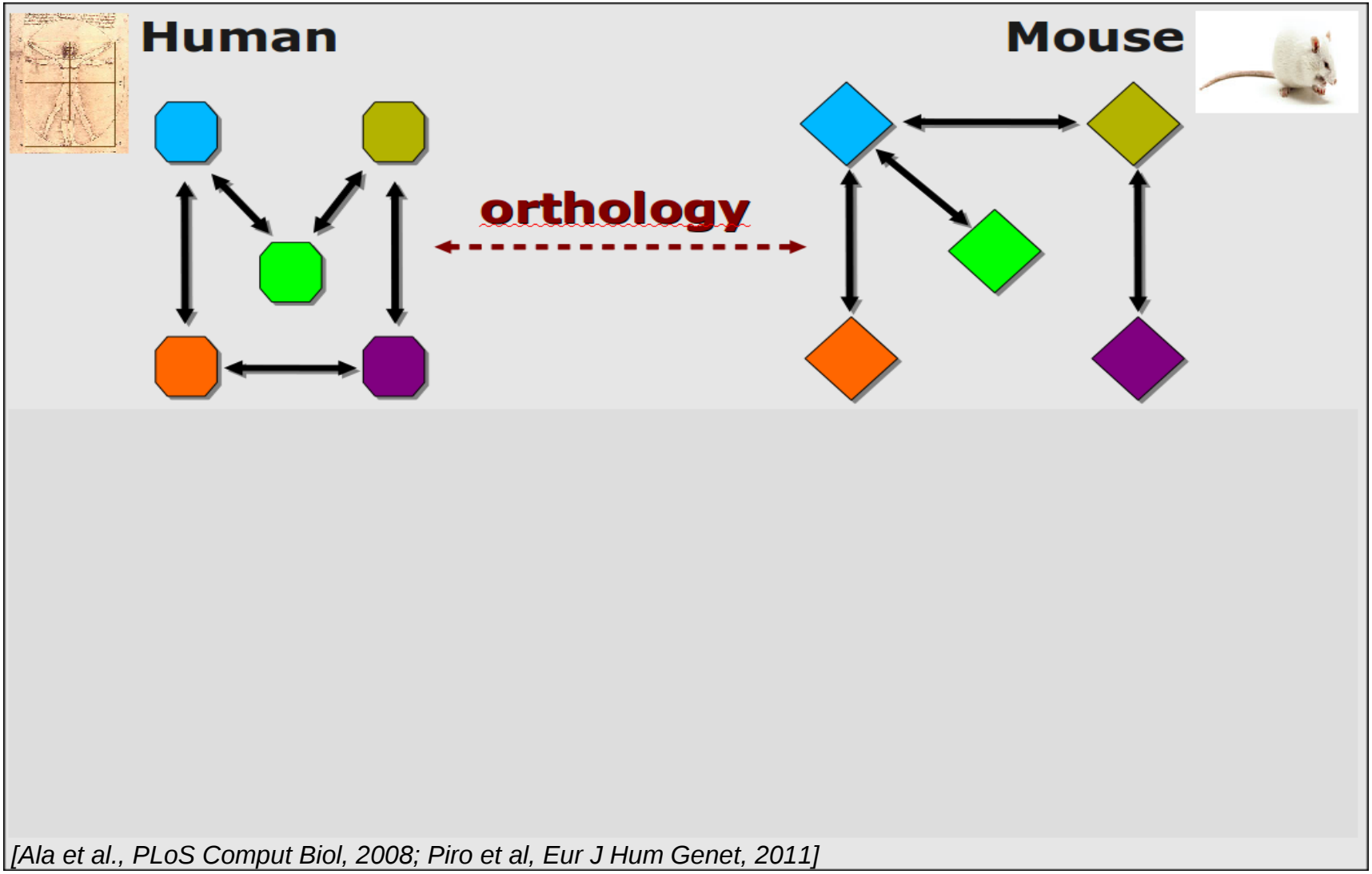
from: Wikipedia / Wikimedia Commons

Conserved coexpression



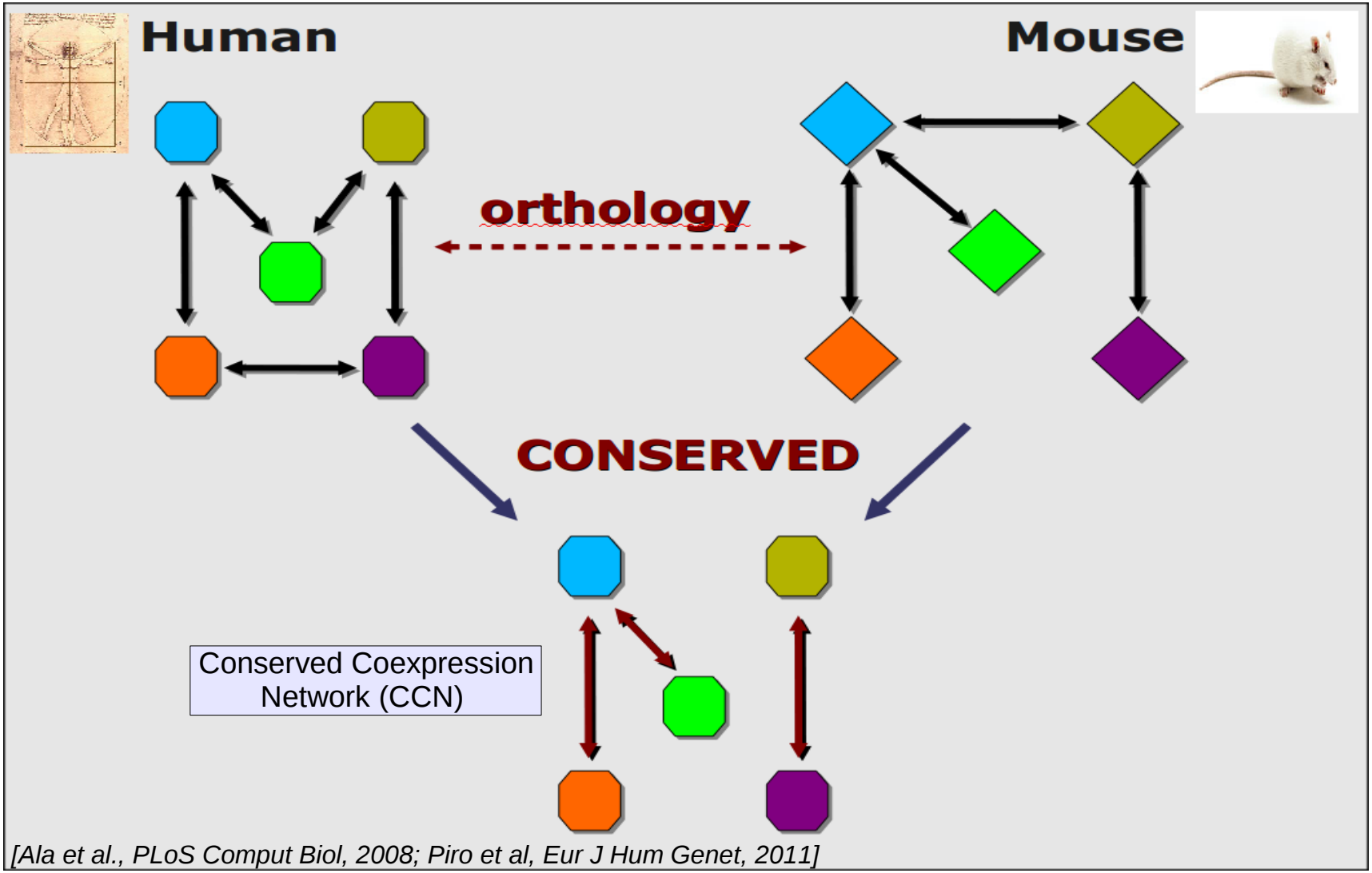
[Ala et al., PLoS Comput Biol, 2008; Piro et al, Eur J Hum Genet, 2011]

Conserved coexpression



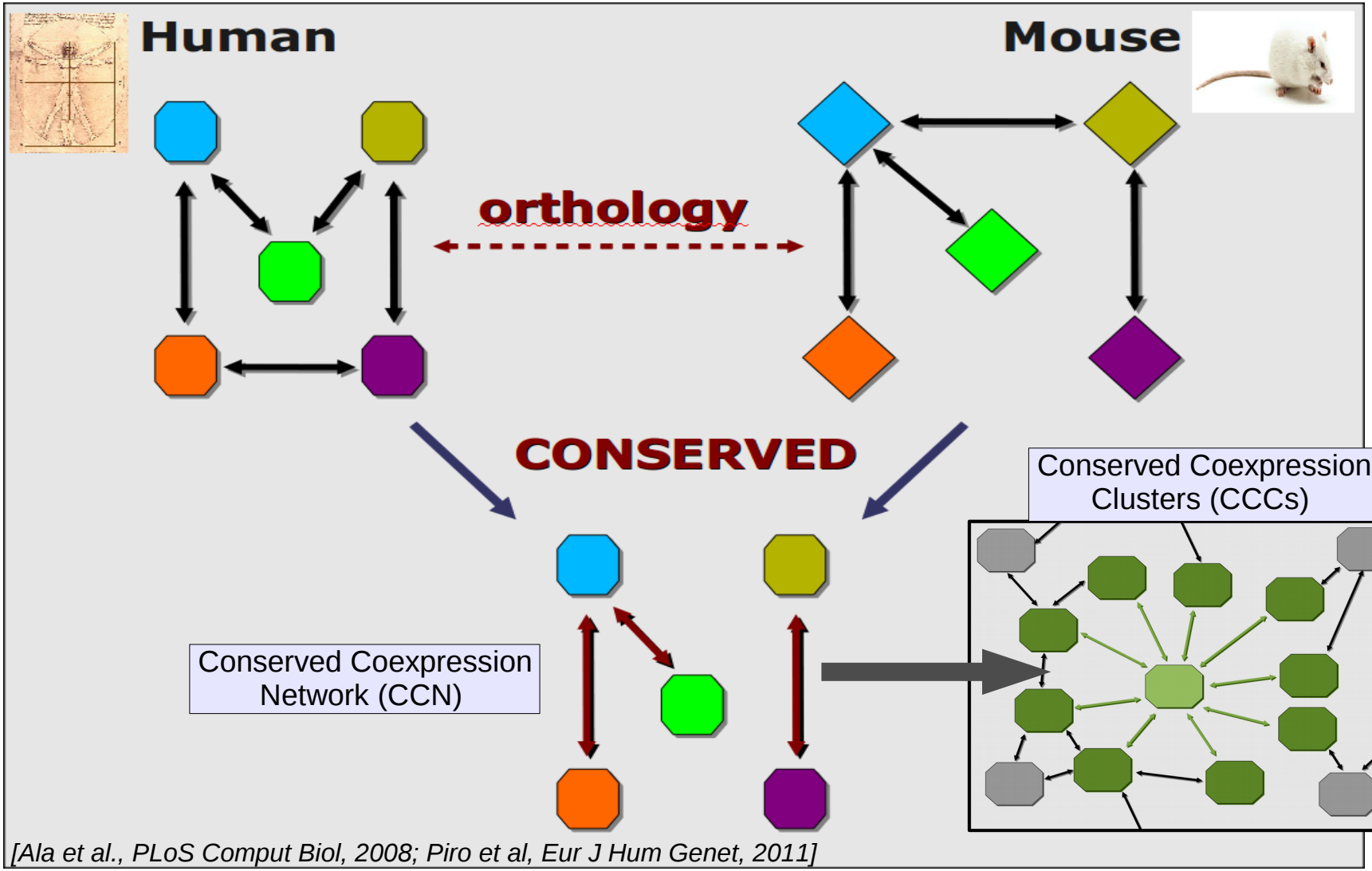
[Ala et al., PLoS Comput Biol, 2008; Piro et al, Eur J Hum Genet, 2011]

Conserved coexpression



[Ala et al., PLoS Comput Biol, 2008; Piro et al, Eur J Hum Genet, 2011]

Conserved coexpression

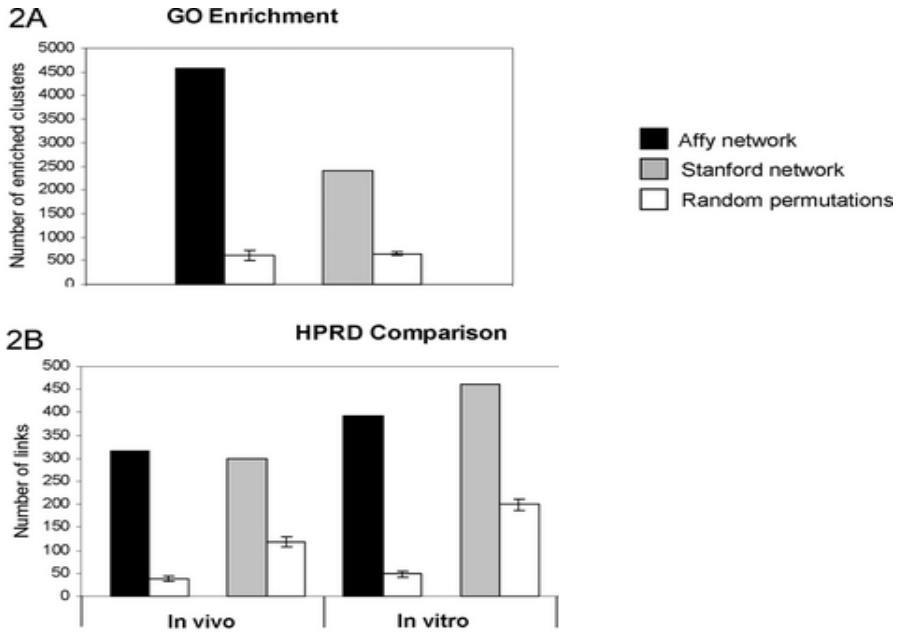


Conserved coexpression

Does conserved coexpression reflect biologivally meaningful information?

Clusters are enriched in

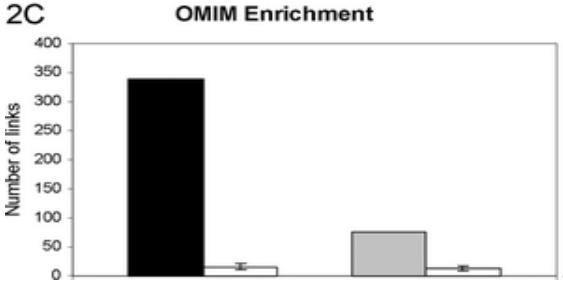
- shared Gene Ontology annotations (e.g., biological processes)



- known protein-protein interactions

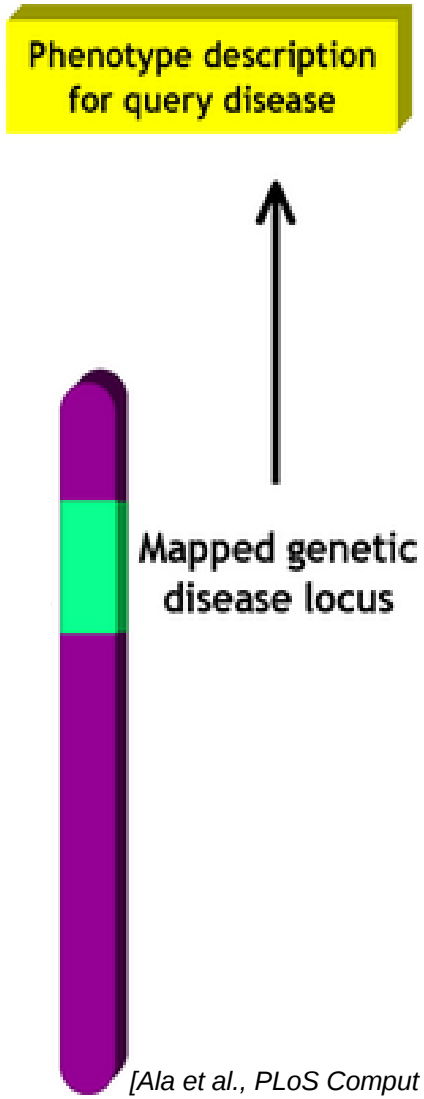
Network edges are enriched in

- gene pairs involved in similar disease phenotypes



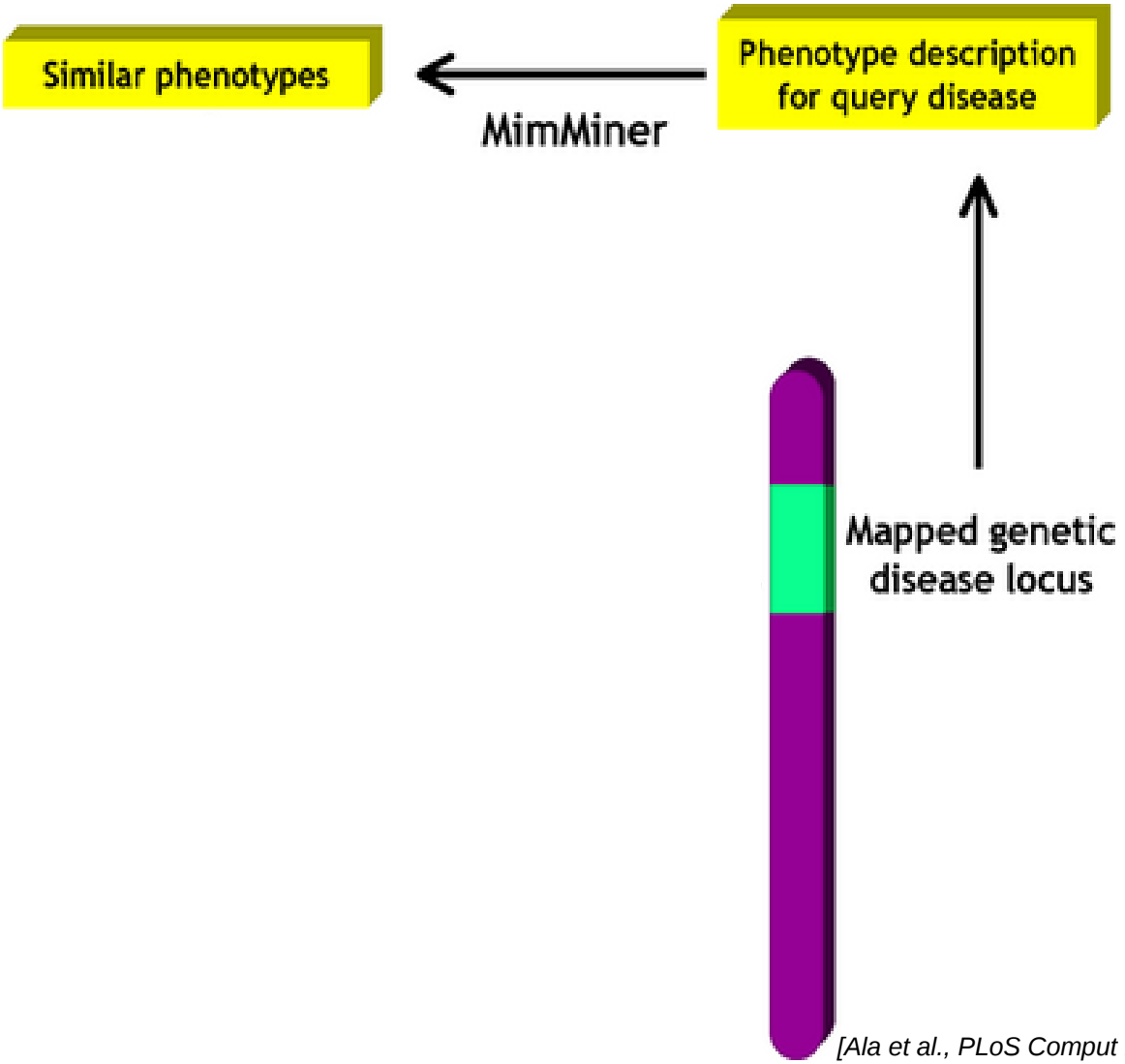
[Ala et al., PLoS Comput Biol, 2008]

Conserved coexpression: prediction



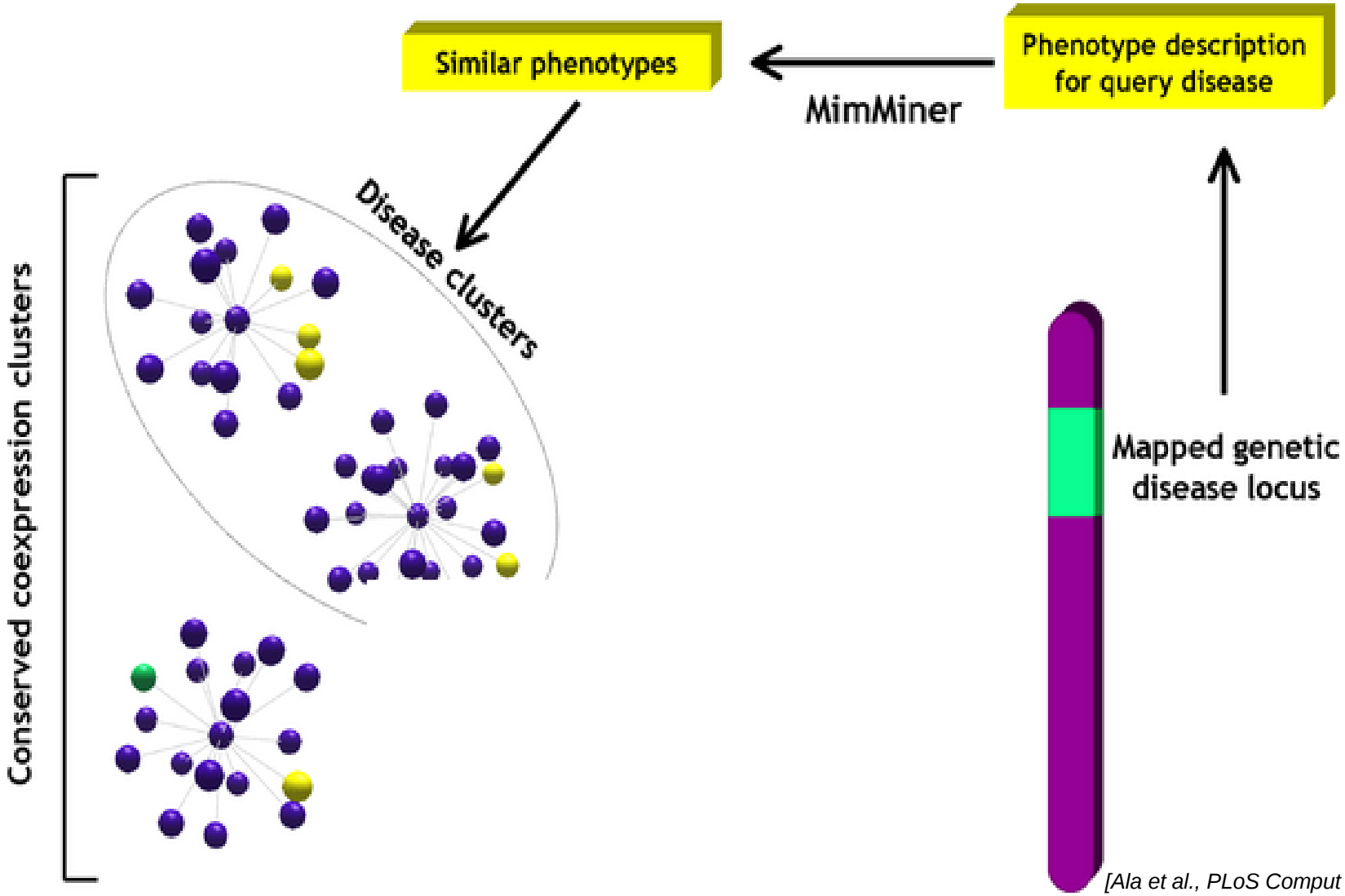
[Ala et al., PLoS Comput Biol, 2008]

Conserved coexpression: prediction



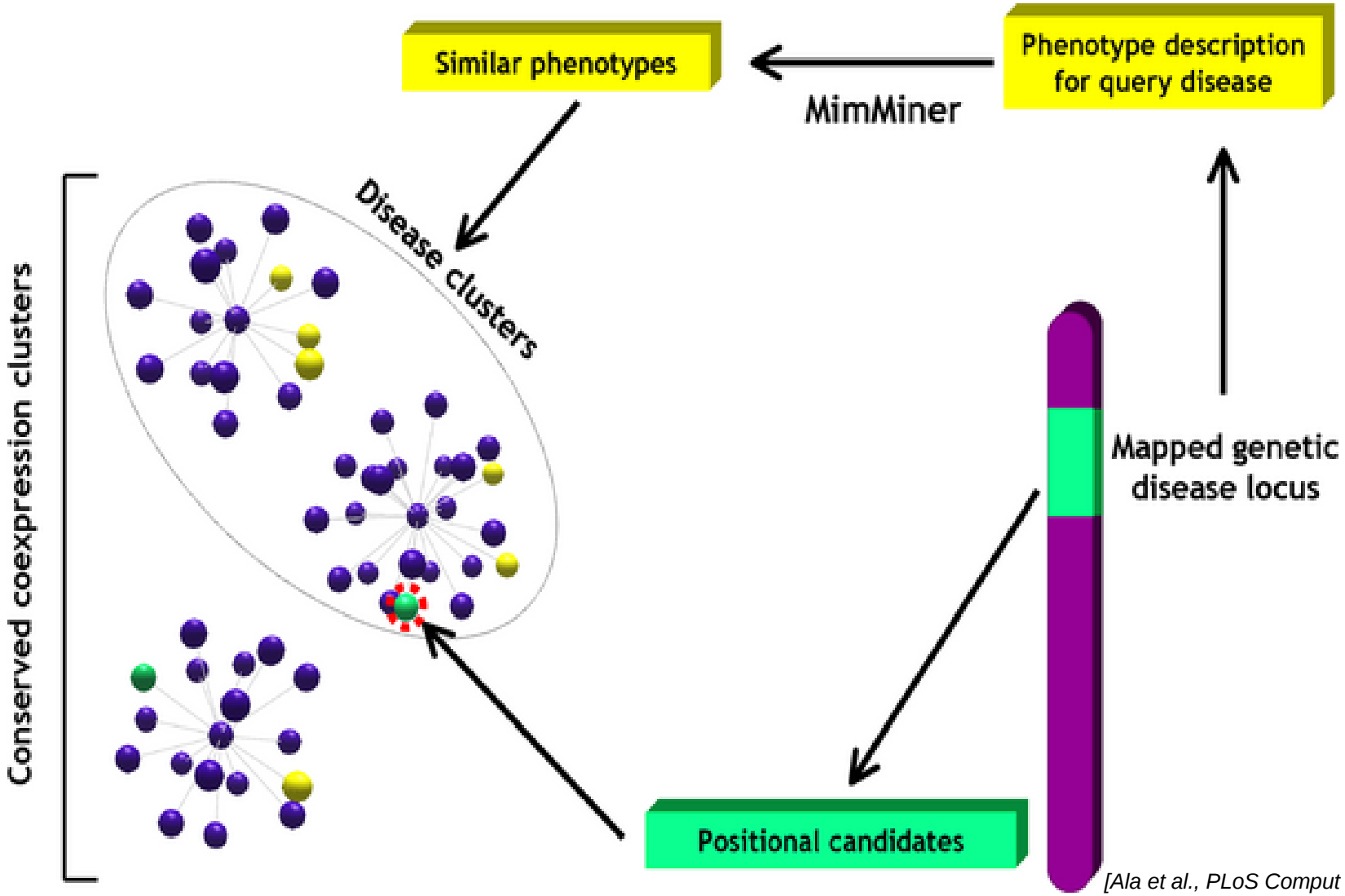
[Ala et al., PLoS Comput Biol, 2008]

Conserved coexpression: prediction



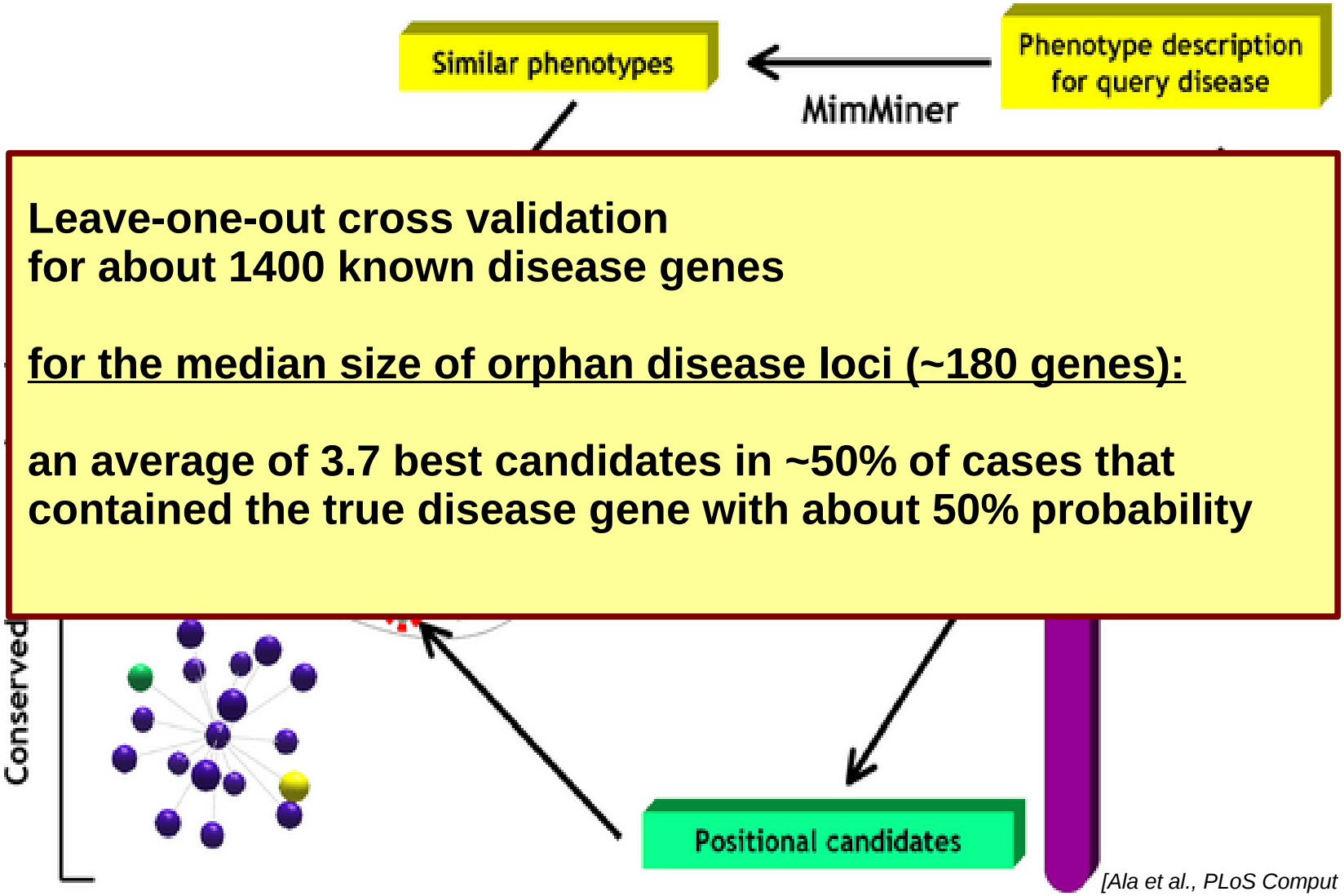
[Ala et al., PLoS Comput Biol, 2008]

Conserved coexpression: prediction



[Ala et al., PLoS Comput Biol, 2008]

Conserved coexpression: prediction



Leave-one-out cross validation
for about 1400 known disease genes

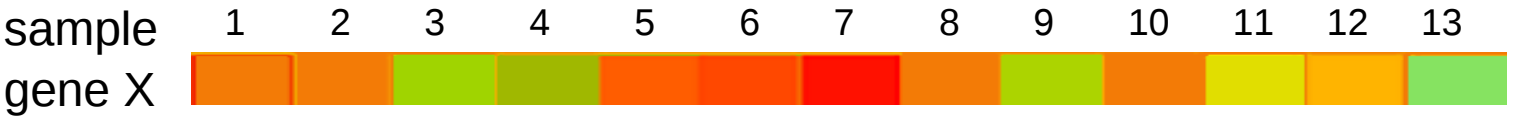
<u>for the median size of orphan disease loci (~180 genes):</u>

an average of 3.7 best candidates in ~50% of cases that contained the true disease gene with about 50% probability

[Ala et al., PLoS Comput Biol, 2008]

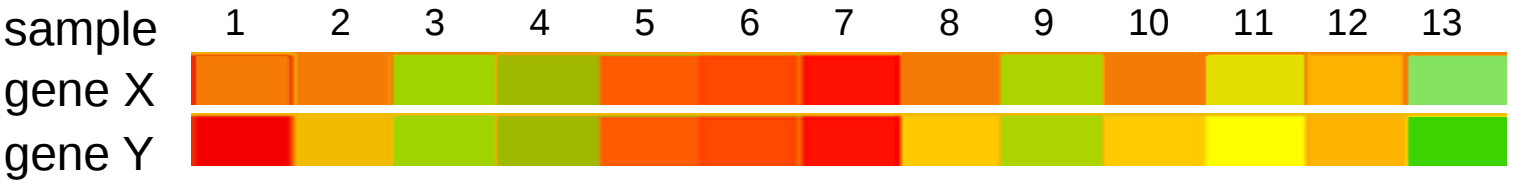
Spatial (3D) gene expression profiles

“Traditional” high-throughput expression data



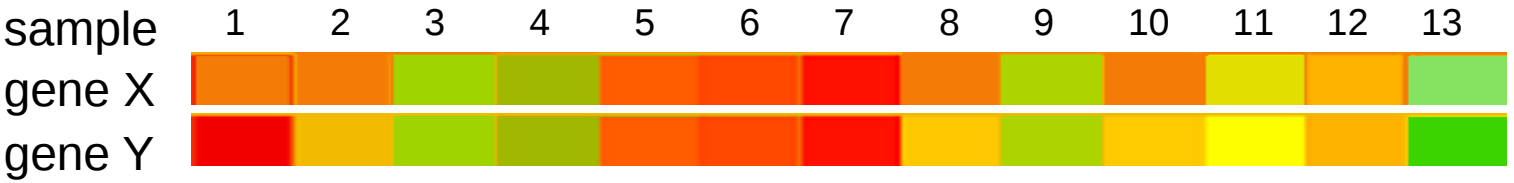
Spatial (3D) gene expression profiles

“Traditional” high-throughput expression data



Spatial (3D) gene expression profiles

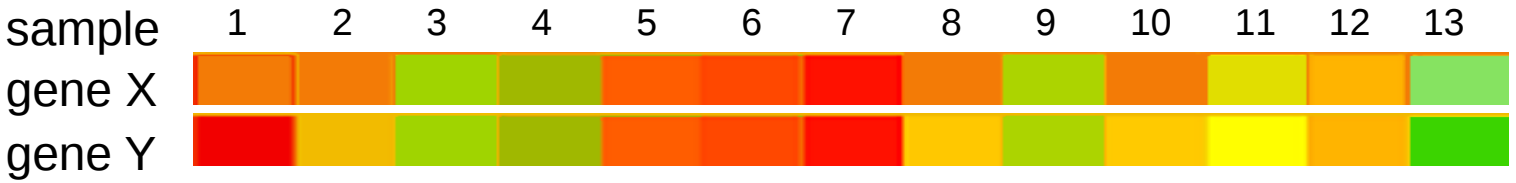
“Traditional” high-throughput expression data



- 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)

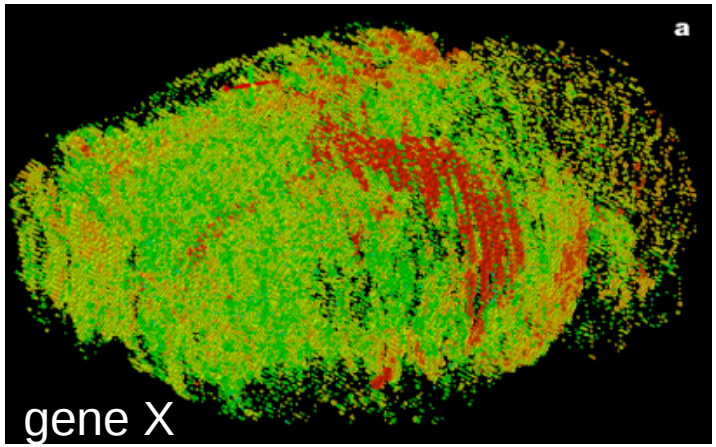
Spatial (3D) gene expression profiles

“Traditional” high-throughput expression data



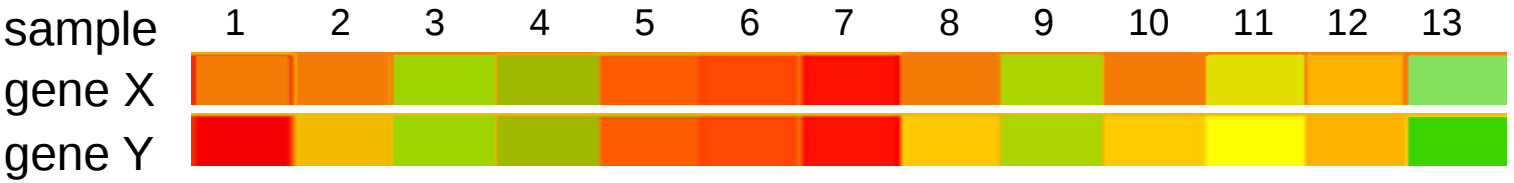
- carry **limited spatial information**, no precise 3D localization
- have a **sparse coverage** (arbitrary, not-consecutive positions)
- problem for tissues/organs with a high degree of spatial organization, e.g. the central nervous system (CNS)

© Allen Institute



Spatial (3D) gene expression profiles

"Traditional" high-throughput expression data

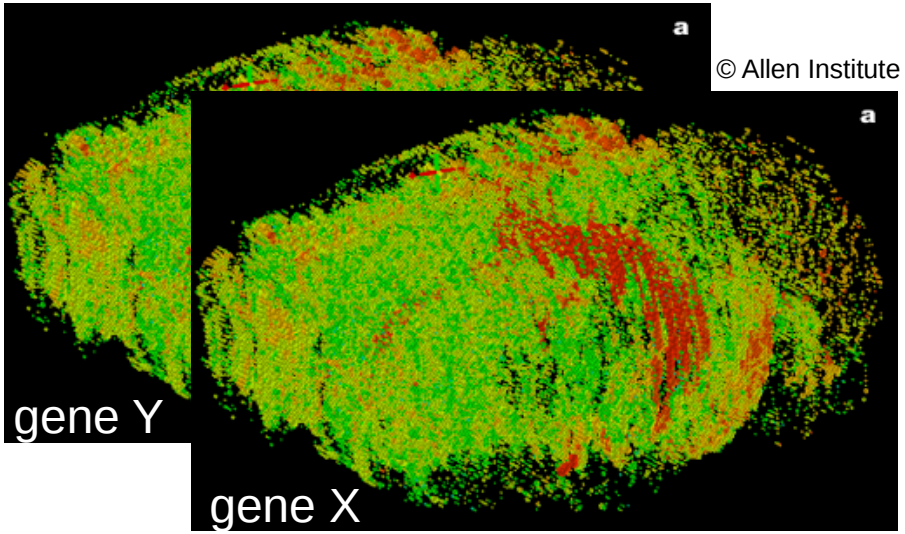


- carry **limited spatial information**, no precise 3D localization
- have a **sparse coverage** (arbitrary, not-consecutive positions)
- problem for tissues/organs with a high degree of spatial organization, e.g. the central nervous system (CNS)

Allen (Mouse) Brain Atlas *[Lein et al. 2007]*

- ~20k genes
- *in situ* hybridization (ISH)
- covers the entire mouse brain
- very high resolution

(Human Brain Atlas: microarrays)



© Allen Institute

Measuring 3D co-expression

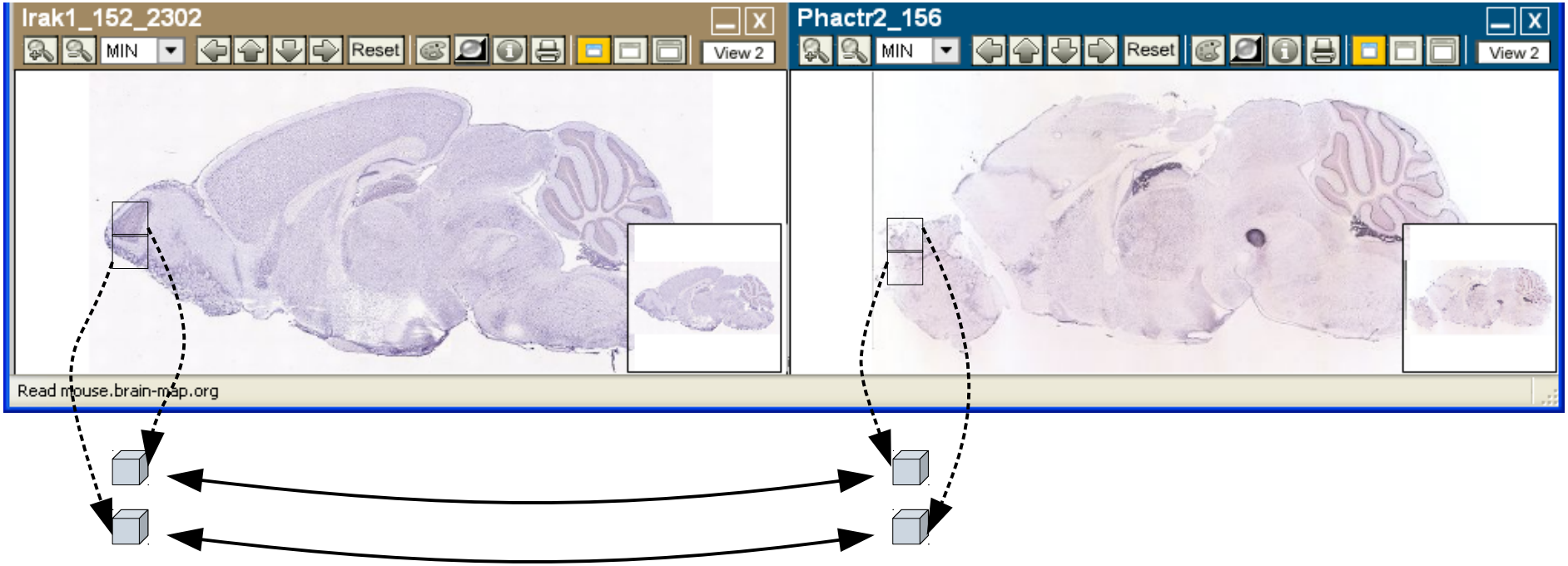
Use of numerical 3D expression patterns



Measuring 3D co-expression

Use of numerical 3D expression patterns

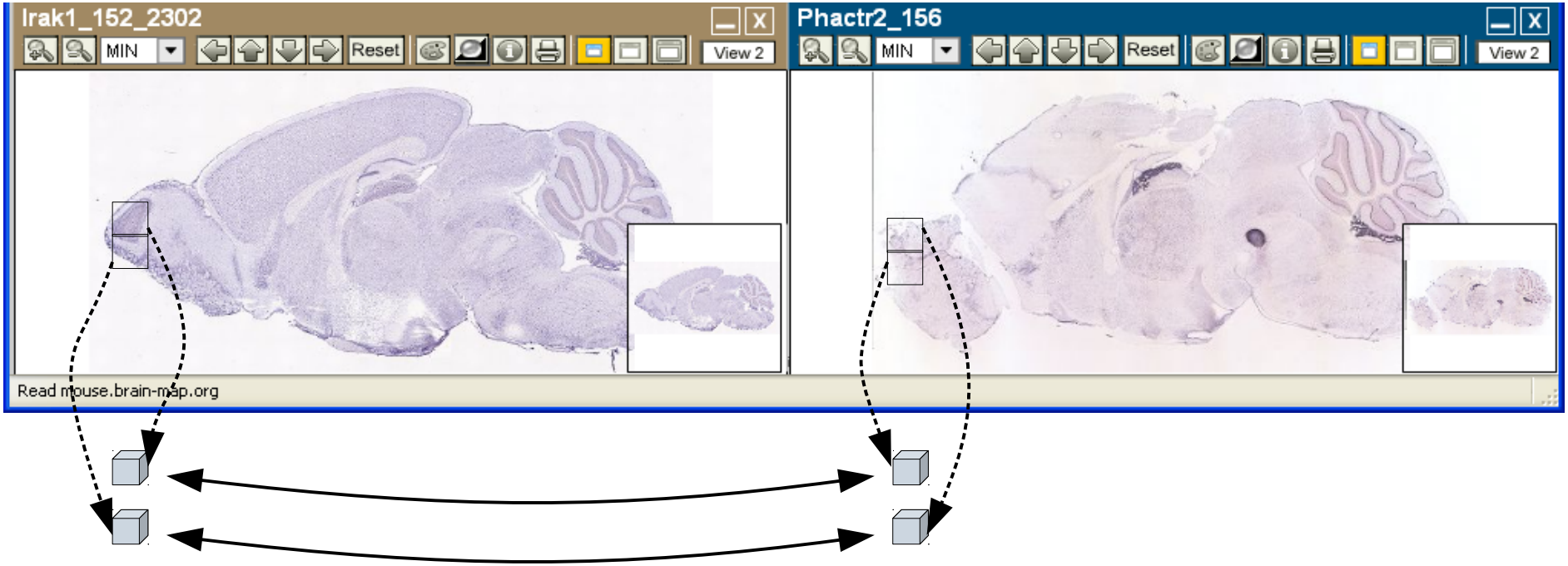
- expression levels for the entire brain divided in evenly spaced “voxels” (cubes; 3D pixels) of 200 µm side length



Measuring 3D co-expression

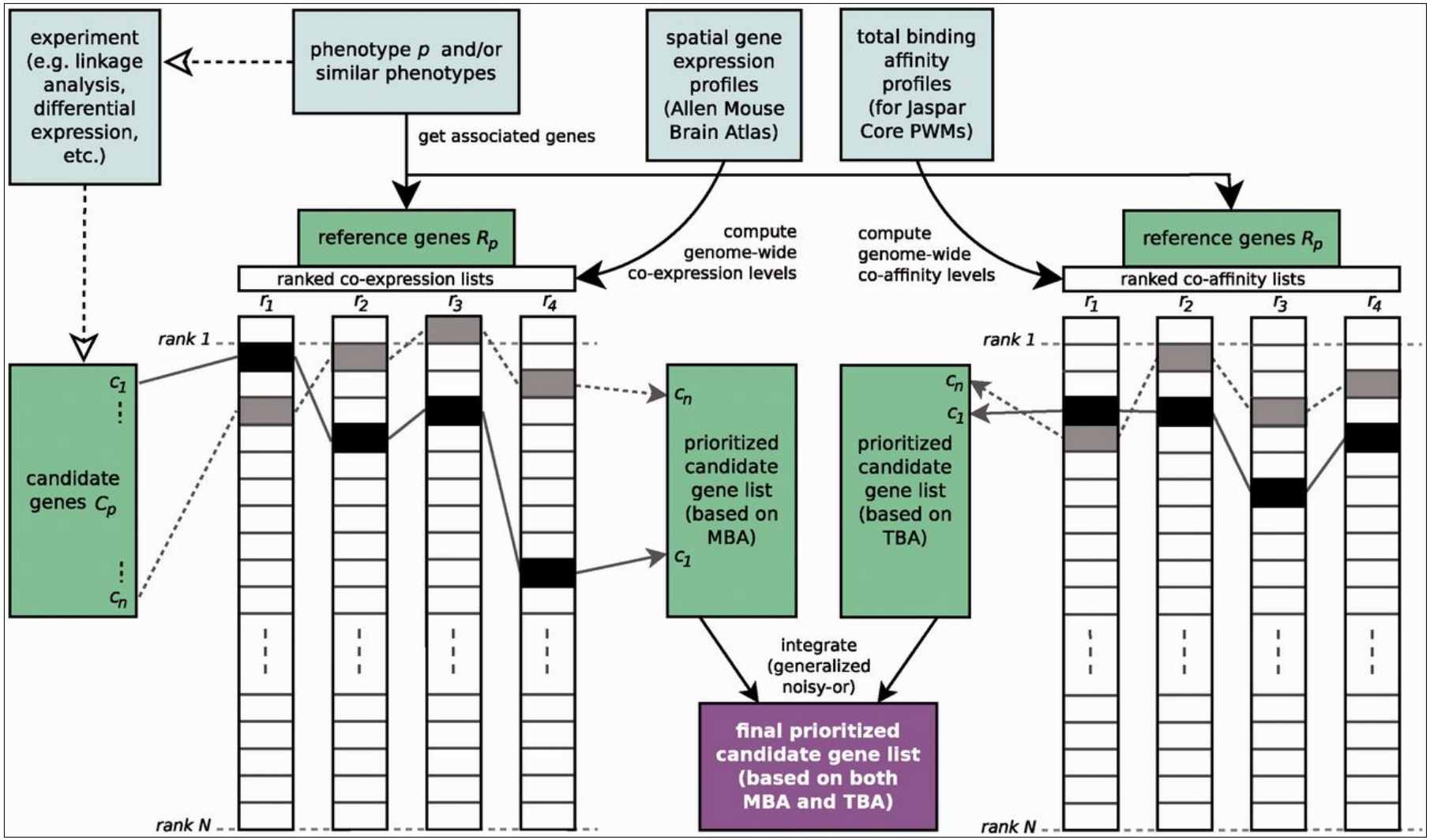
Use of numerical 3D expression patterns

- expression levels for the entire brain divided in evenly spaced “voxels” (cubes; 3D pixels) of 200 µm side length



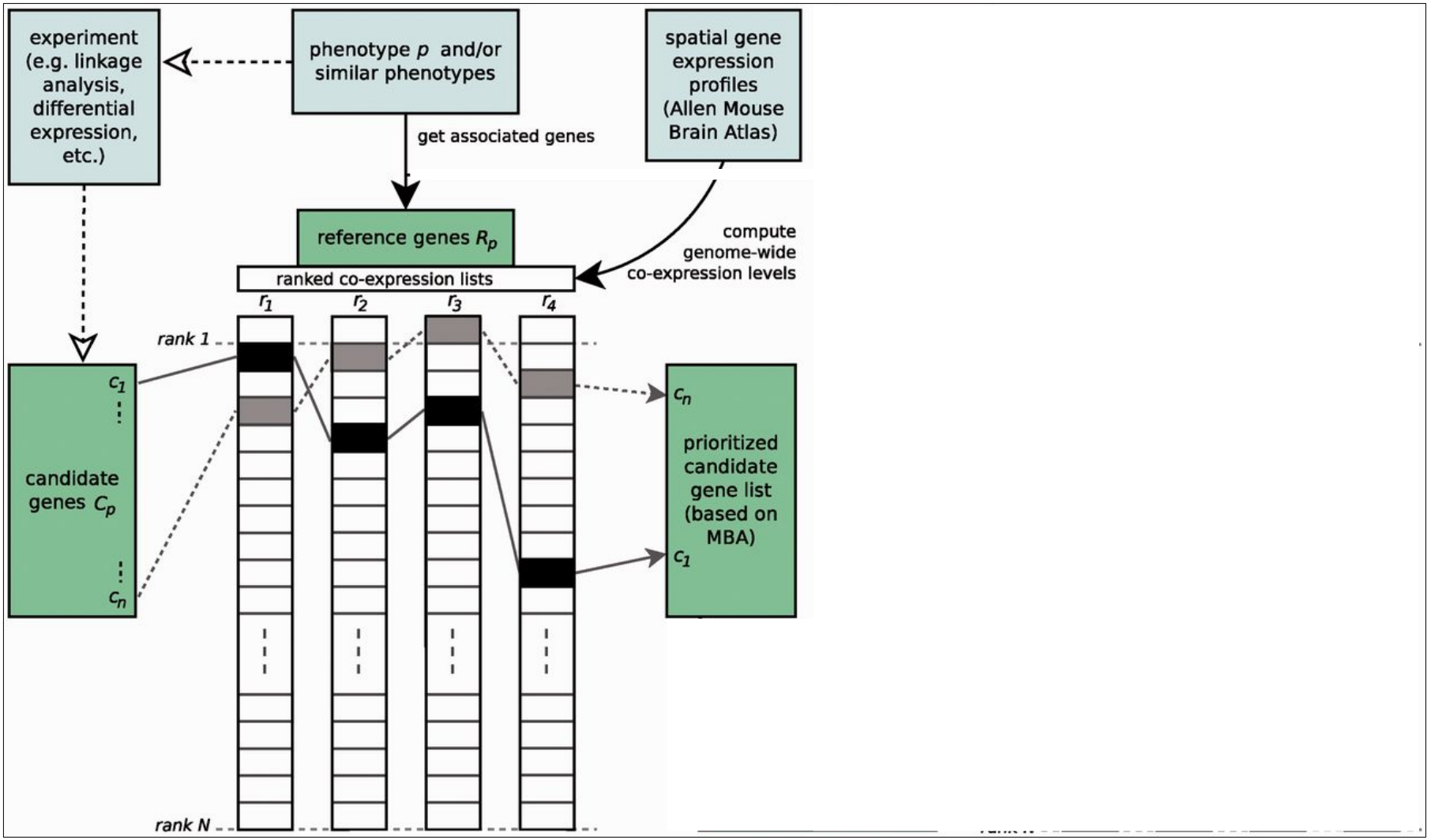
Pearson correlation coefficient (PCC)
Relative Intensity Overlap (RIO) *[Piro et al., Bioinformatics, 2013]*

Method for disease gene prioritization



[Piro et al., Bioinformatics, 2013]

Method for disease gene prioritization



[Piro et al., Bioinformatics, 2013]

Example

Case study:
X-linked mental retardation (XLMR)

Validation

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

Validation: autosomal MR genes as reference genes

[Piro et al., Bioinformatics, 2010]

Validation

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 "rediscovery" of known **XLMR genes**

Validation: autosomal MR genes as reference genes

[Piro et al., Bioinformatics, 2010]

Freie Universität Berlin

Validation

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 "rediscovery" of known **XLMR genes**

Validation: autosomal MR genes as reference genes

Prediction: XLMR genes as reference 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

[Piro et al., Bioinformatics, 2010]

Freie Universität Berlin

Validation

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 "rediscovery" of known **XLMR genes**

Validation: autosomal MR genes as reference genes

Prediction: XLMR genes as reference 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

[Piro et al., Bioinformatics, 2010]

Validation

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 "rediscovery" of known **XLMR genes**

very strong **overlap** between evaluation and prediction (P=6.62x10^{-14})

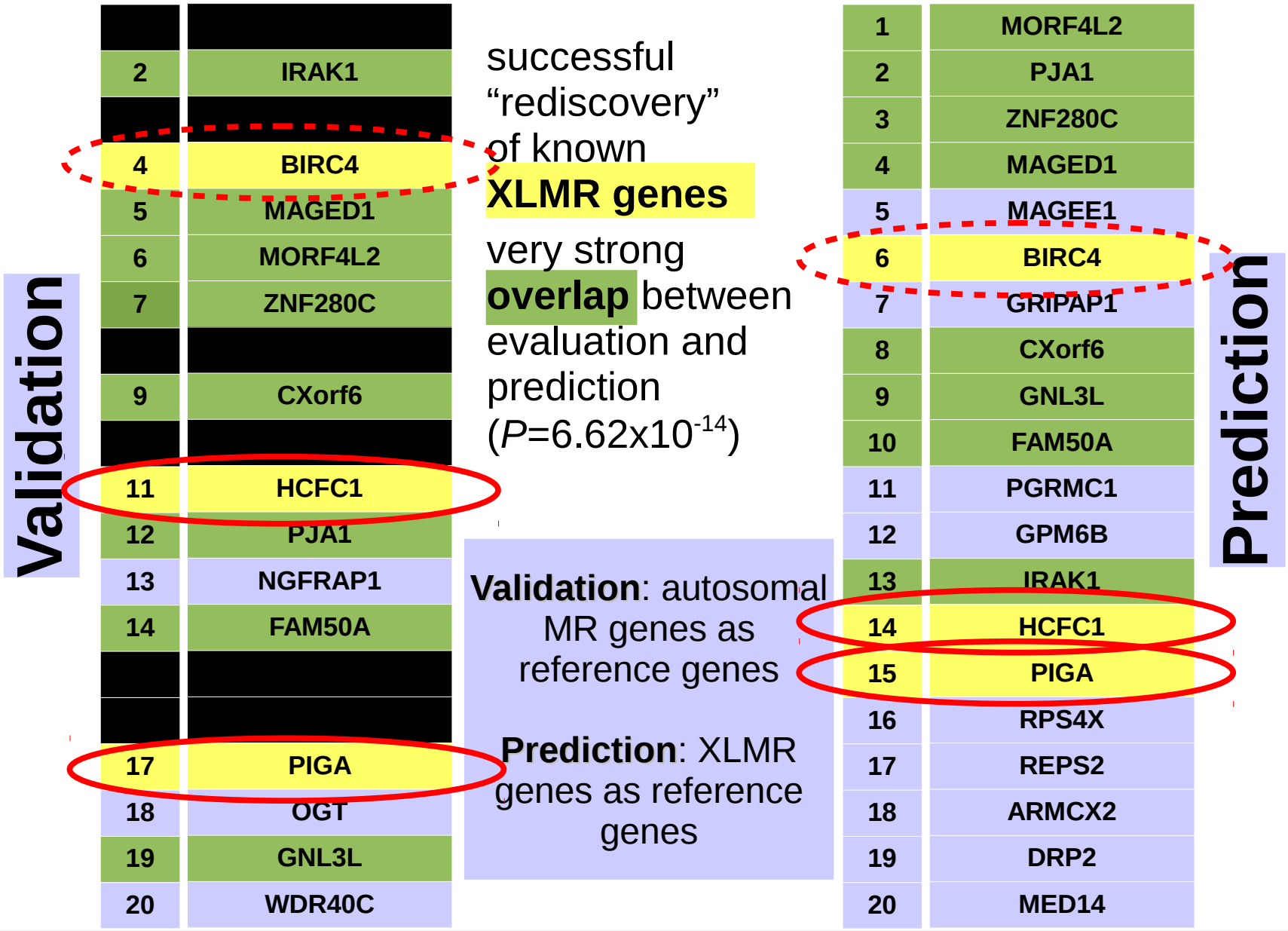
Validation: autosomal MR genes as reference genes

Prediction: XLMR genes as reference genes

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

Prediction

[Piro et al., Bioinformatics, 2010]

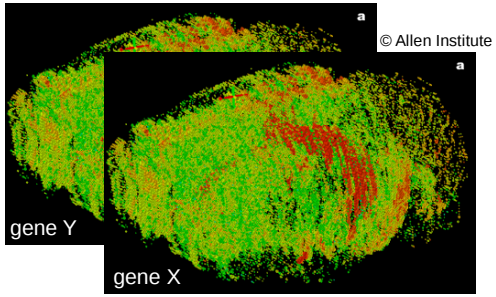


[Piro et al., Bioinformatics, 2010]

Similarity measure for 3D expression profiles

Normalization of expression levels (“intensities”) for each “3D image”

- ranging from 0 to 1



© Allen Institute

Relative Intensity Overlap (RIO):

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

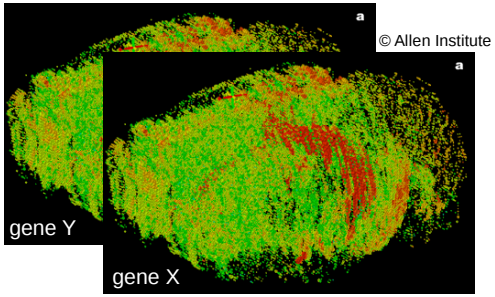
I^a_{xyz} = intensity of gene *a* at position *x,y,z*

[Piro et al., Bioinformatics, 2013]

Similarity measure for 3D expression profiles

Normalization of expression levels (“intensities”) for each “3D image”

- ranging from 0 to 1



© Allen Institute

Relative Intensity Overlap (RIO):

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

I^a_{xyz} = intensity of gene *a* at position *x,y,z*

Intensities *a*, *b* 0 0

Contribution to RIO 0

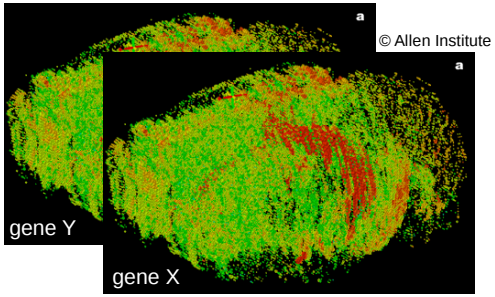
Penalization 0

[Piro et al., Bioinformatics, 2013]

Similarity measure for 3D expression profiles

Normalization of expression levels (“intensities”) for each “3D image”

- ranging from 0 to 1



© Allen Institute

Relative Intensity Overlap (RIO):

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

I^a_{xyz} = intensity of gene *a* at position *x,y,z*

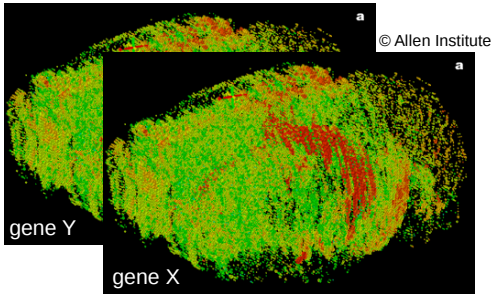
Intensities *a*, *b*	0	0	1	1
Contribution to RIO	0		1	
Penalization	0		0	

[Piro et al., Bioinformatics, 2013]

Similarity measure for 3D expression profiles

Normalization of expression levels (“intensities”) for each “3D image”

- ranging from 0 to 1



© Allen Institute

Relative Intensity Overlap (RIO):

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

I^a_{xyz} = intensity of gene *a* at position *x,y,z*

Intensities *a*, *b*	0	0		1	1		0	1
Contribution to RIO		0			1			0
Penalization		0			0			1

[Piro et al., Bioinformatics, 2013]

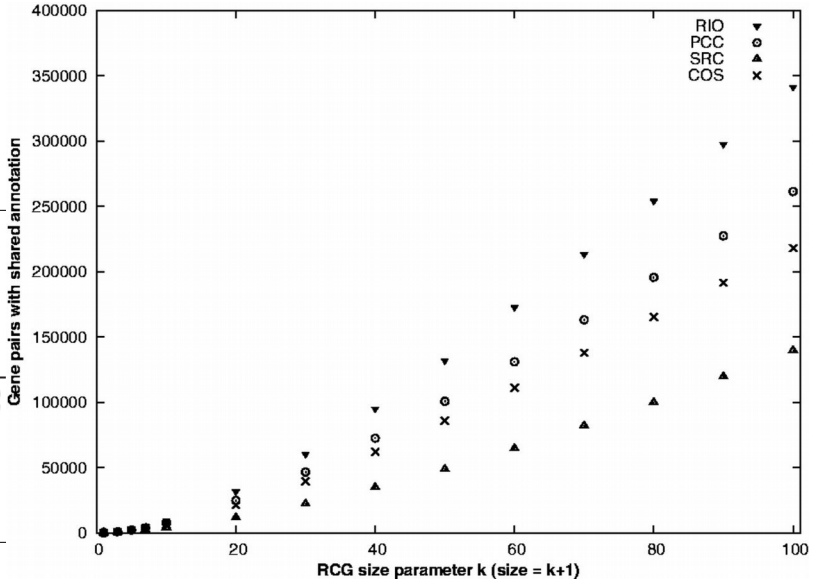
Similarity measure for 3D expression profiles

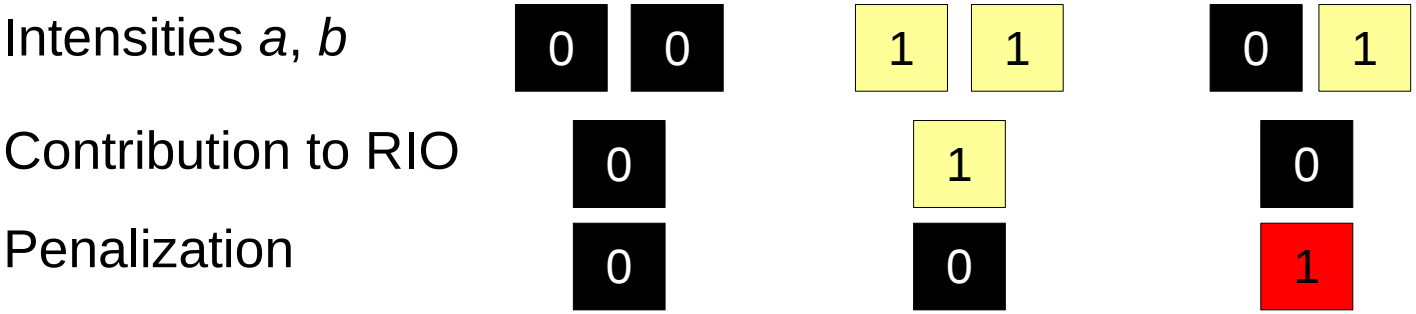
Normalization of expression levels (“intensities”) for each “3D image”

- ranging from 0 to 1

Relative Intensity Overlap (RIO):

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



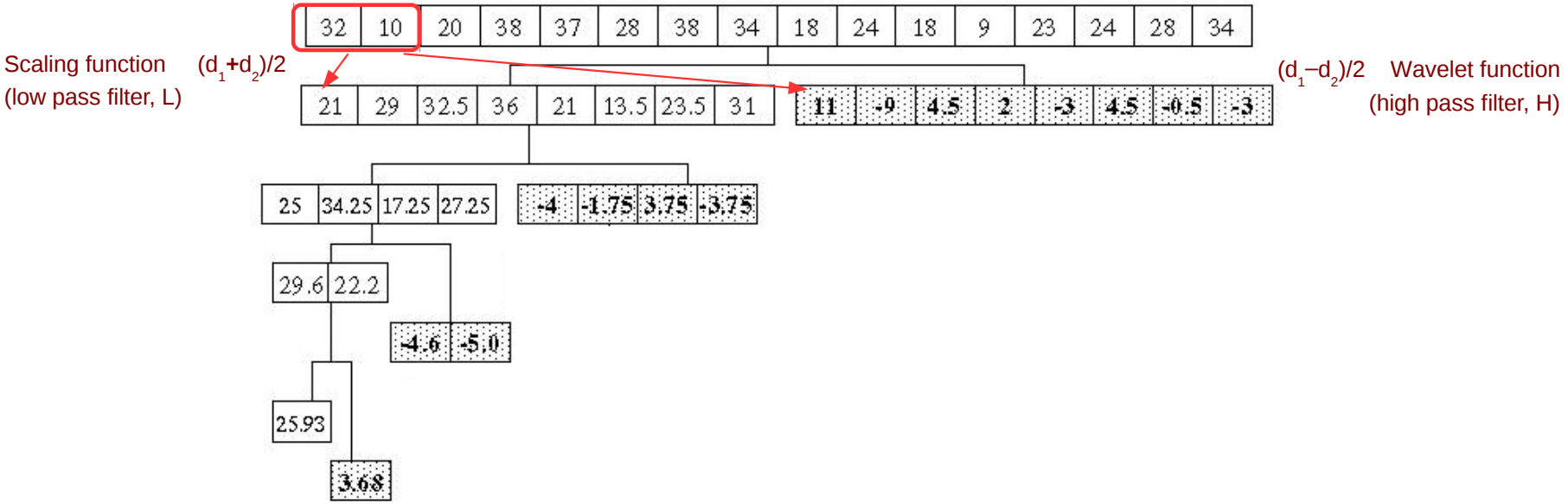


[Piro et al., Bioinformatics, 2013]

Can we better exploit the spatial information?

Techniques from image processing

- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



Can we better exploit the spatial information?

Techniques from image processing

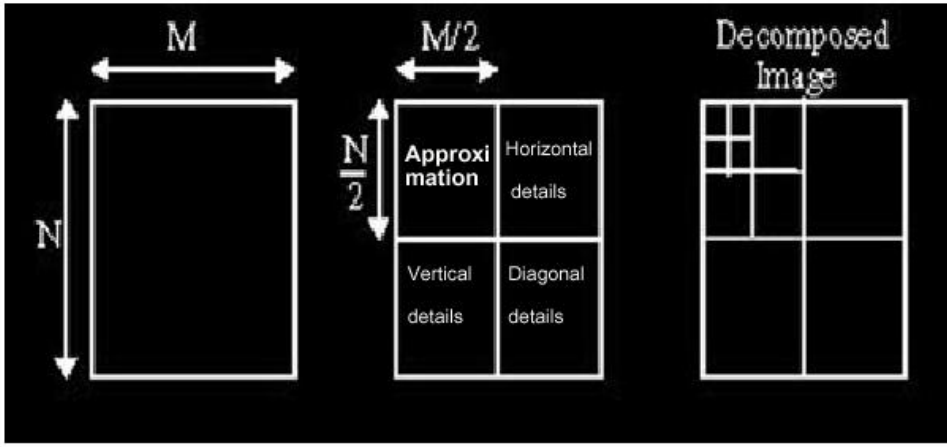
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales

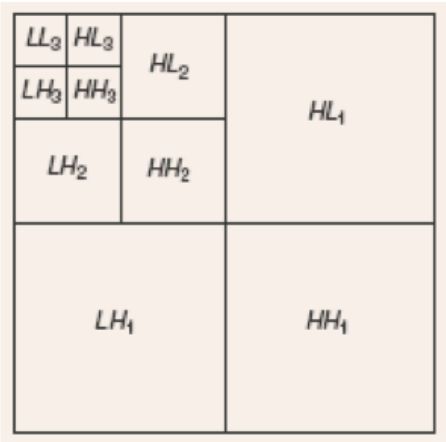
32	10	20	38	37	28	38	34	18	24	18	9	23	24	28	34

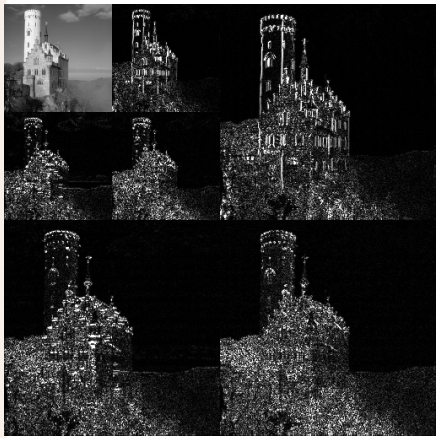
Scaling function (low pass filter, L) $(d_1+d_2)/2$

25.93	3.68	-4.6	-5.0	-4	-1.75	3.75	-3.75	11	-9	4.5	2	-3	4.5	-0.5	-3

$(d_1-d_2)/2$ Wavelet function (high pass filter, H)







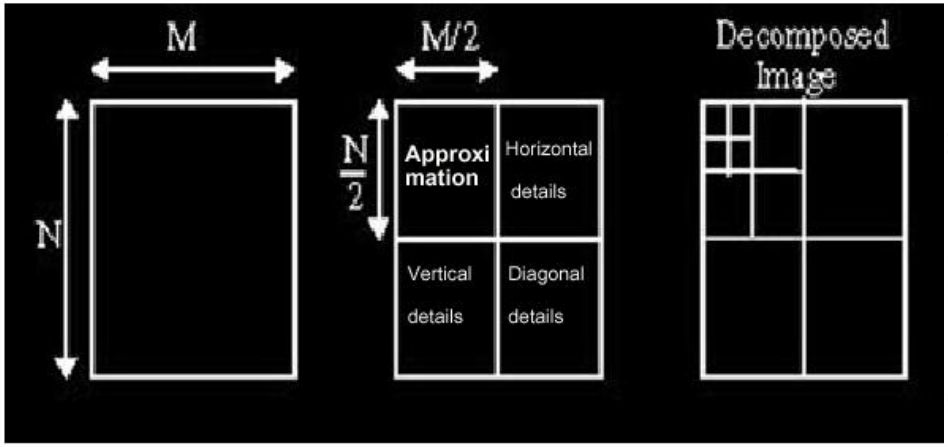
Can we better exploit the spatial information?

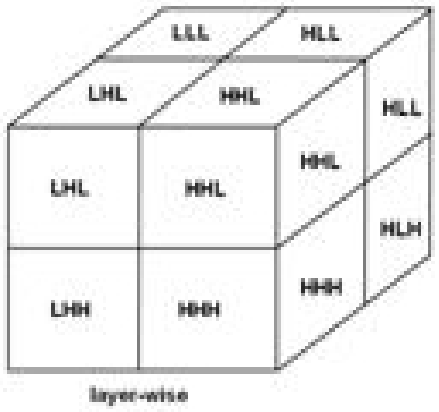
Techniques from image processing

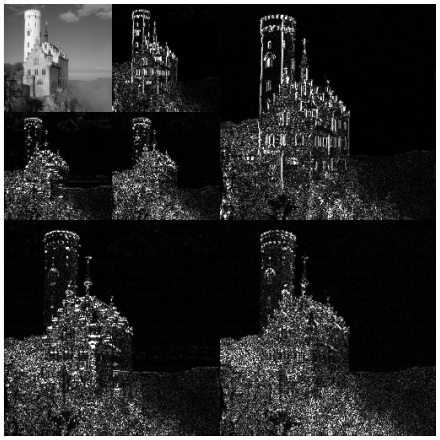
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales

	32	10	20	38	37	28	38	34	18	24	18	9	23	24	28	34
Scaling function (low pass filter, L) $(d_1+d_2)/2$	25.93	3.68	-4.6	-5.0	-4	-1.75	3.75	-3.75	11	-9	4.5	2	-3	4.5	-0.5	-3

$(d_1-d_2)/2$ Wavelet function (high pass filter, H)







Part III

Computational Oncology

KLF4 in secretory meningiomas

NAB2-STAT6 in meningeal hemangiopericytomas and solitary fibrous tumors

Next generation sequencing: applications

DNA sequencing (DNA-seq)

- Whole genome sequencing (WGS)
- Whole exome sequencing (WES)

RNA / transcriptome sequencing (RNA-seq)

Next generation sequencing: applications

DNA sequencing (DNA-seq)

- Whole genome sequencing (WGS)
- **Whole exome sequencing (WES)**

RNA / transcriptome sequencing (RNA-seq)

Epigenetics

- Bisulfite sequencing (to idenfiy DNA methylation)
- Histone modifications

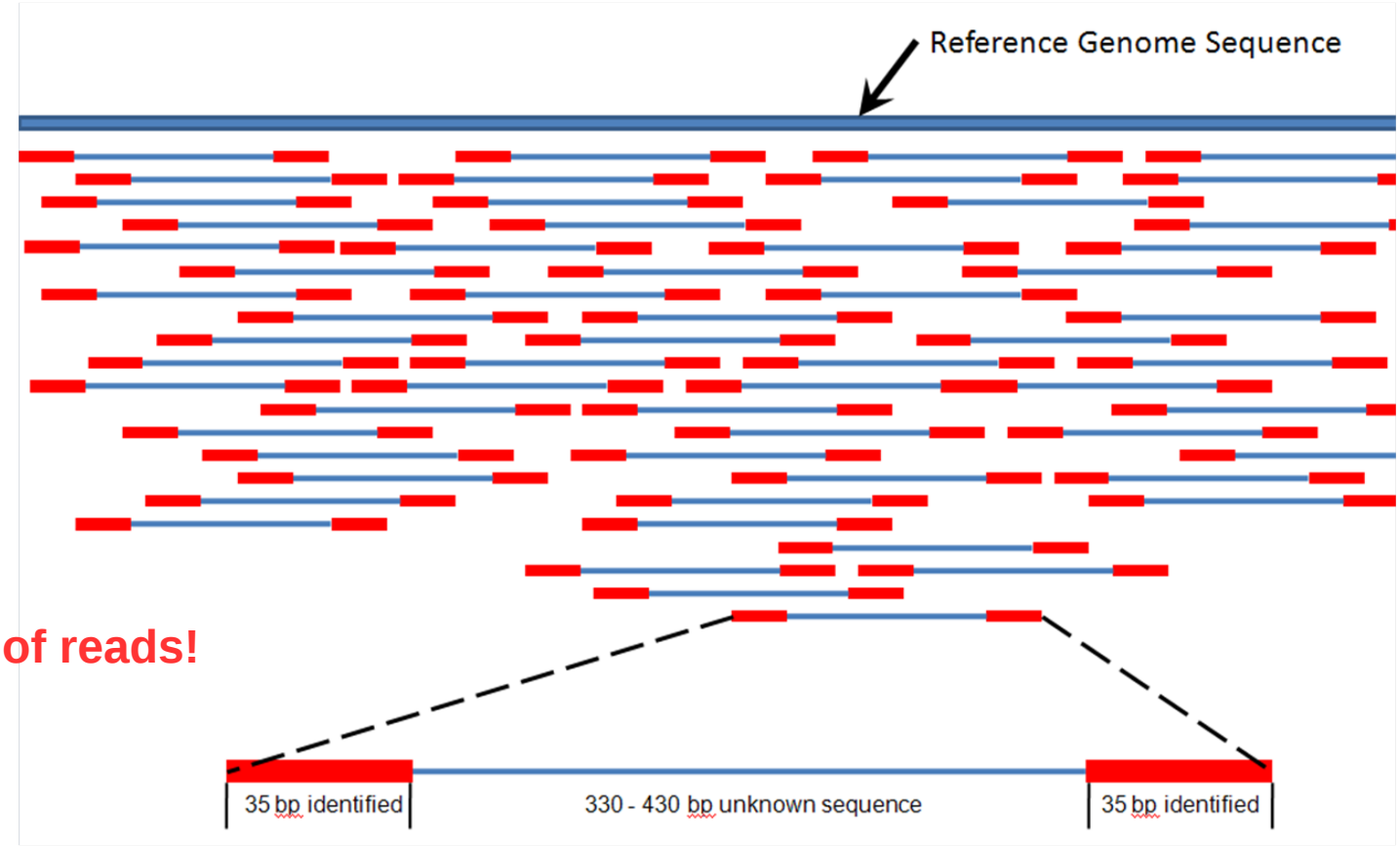
Transcription factor binding

- ChIP-seq (chromatin imunoprecipitation followed by DNA sequencing)

DNA sequencing: basic idea

“Shotgun“ sequencing: DNA broken into random fragments

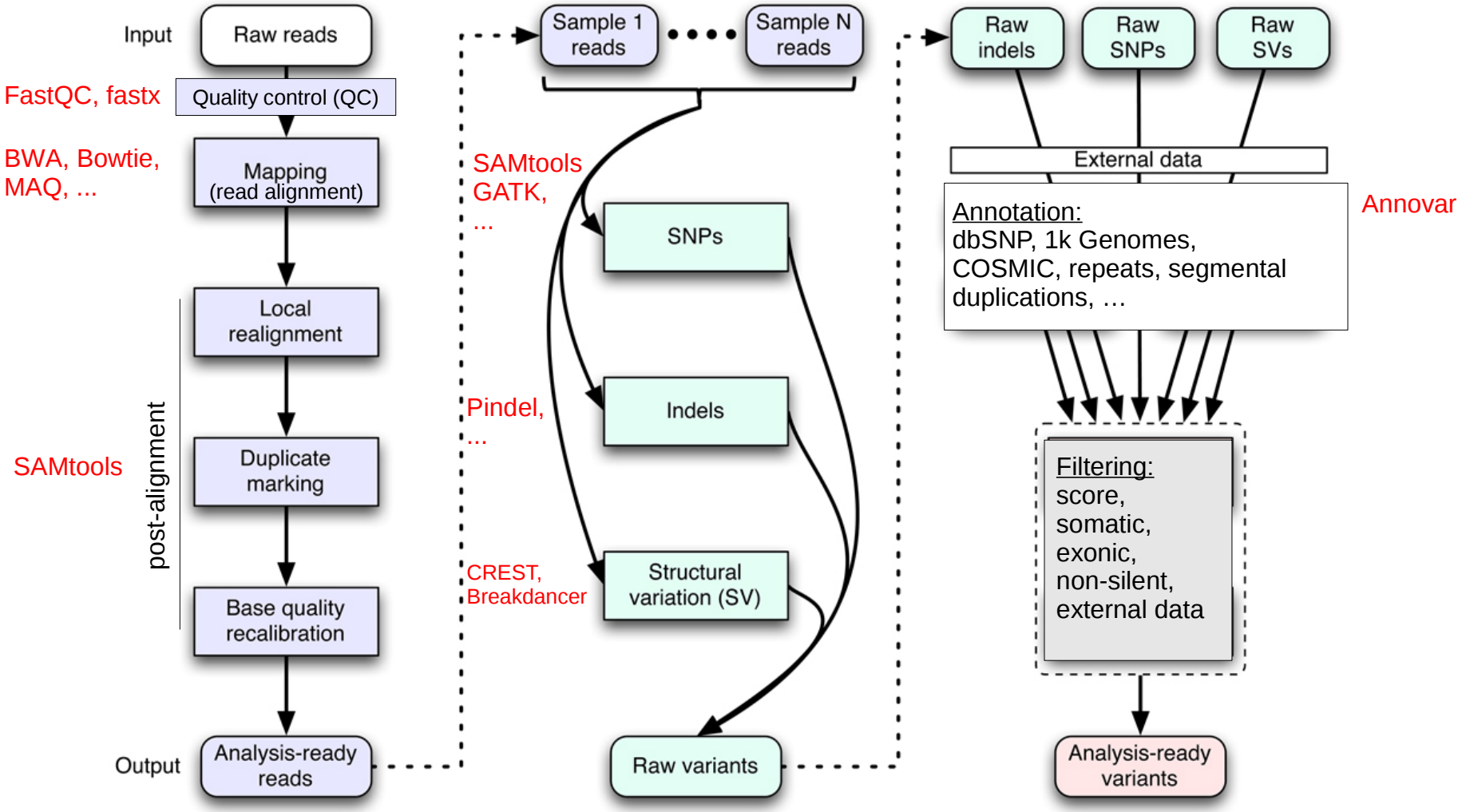
Sequencing of short “reads“ at both ends of these fragments (“paired-end seq.“)



millions / billions of reads!

from: Wikipedia / Wikimedia Commons

NGS data anlysis workflow

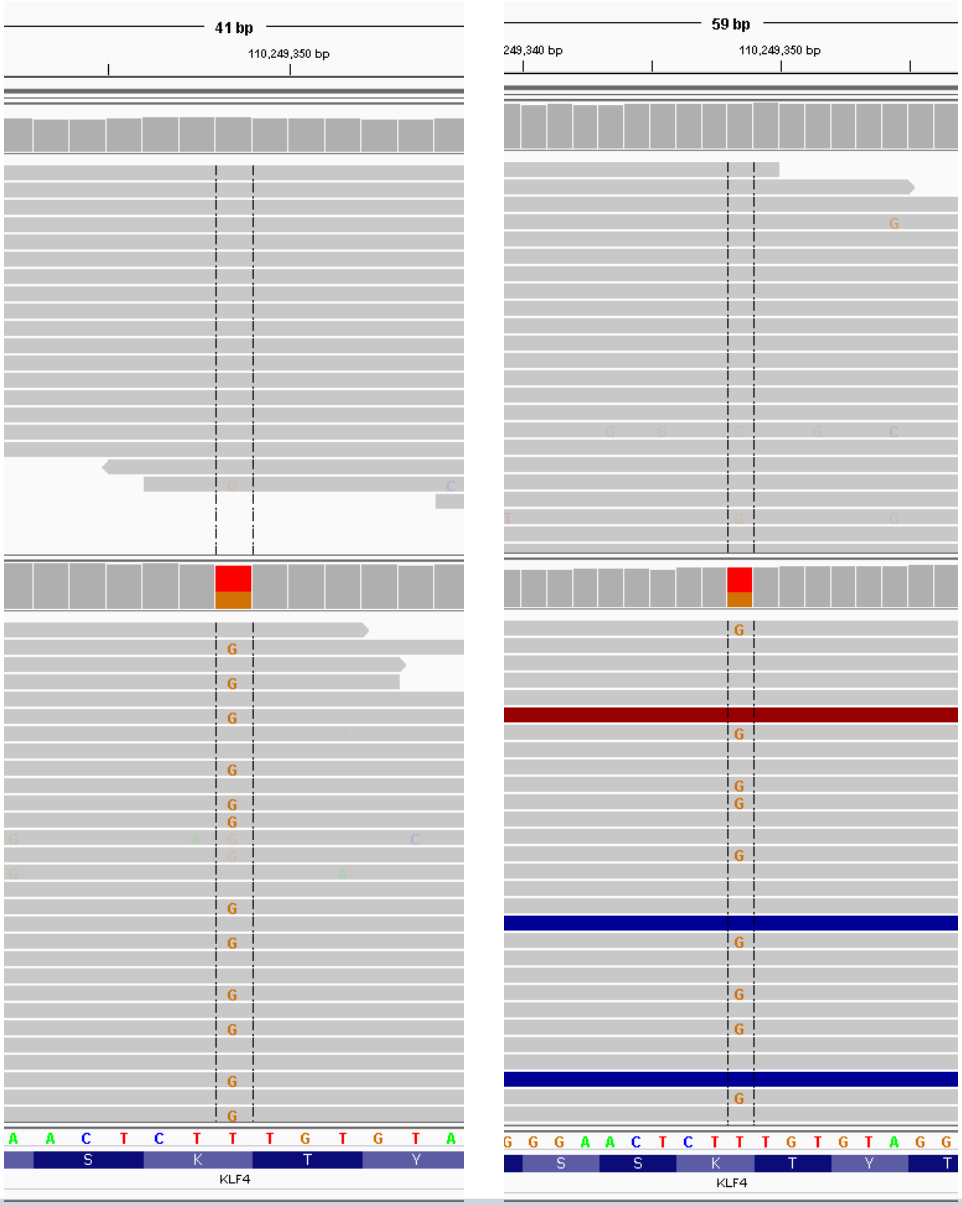


Somatic mutations

Present in a tumor but not in the "germline"/control sample of the same patient

control

tumor



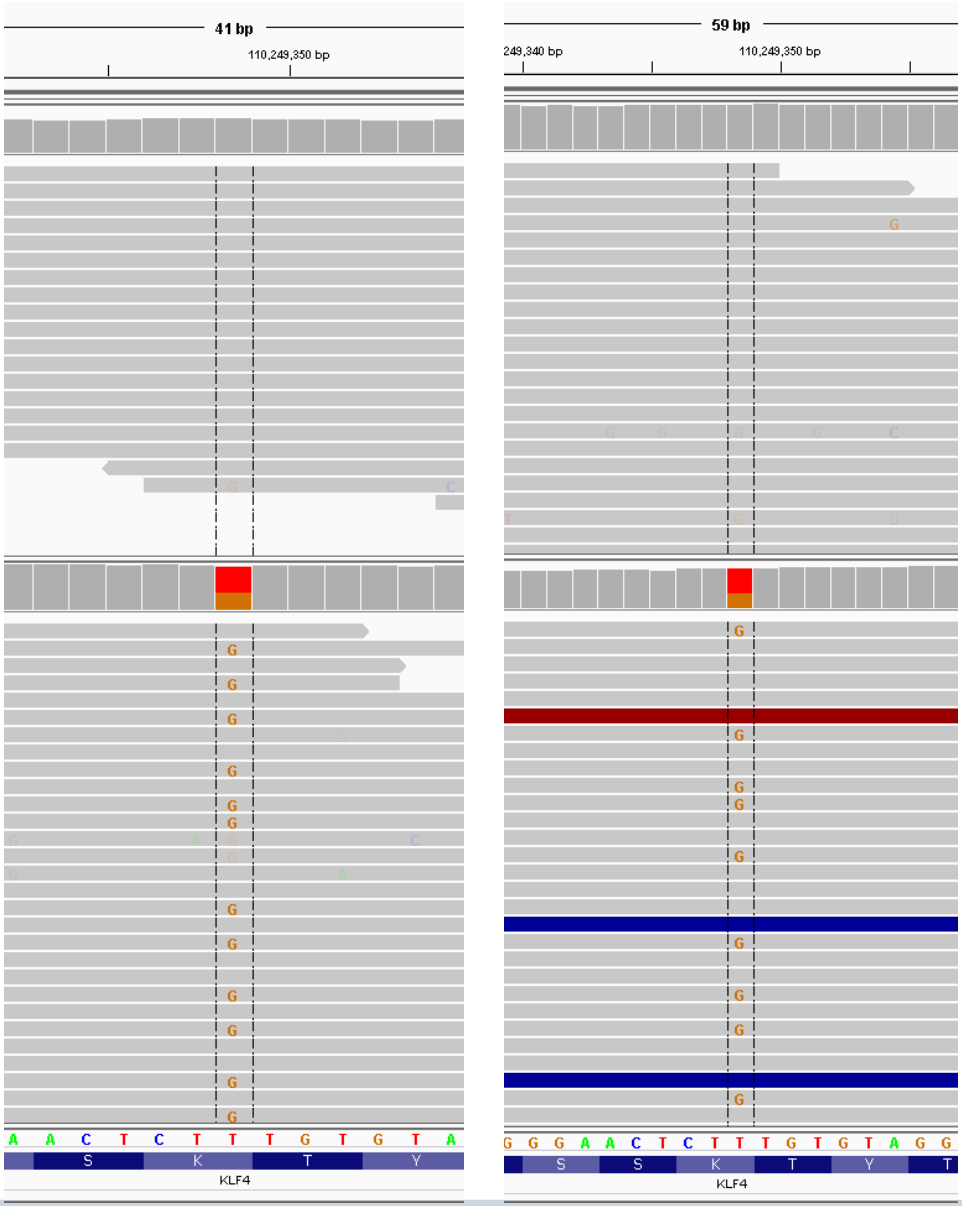
Somatic mutations

Present in a tumor
but not in the
“germline“/control
sample of the
same patient

Can be “driver mutations“
or “passengers“

control

tumor



Structural variants

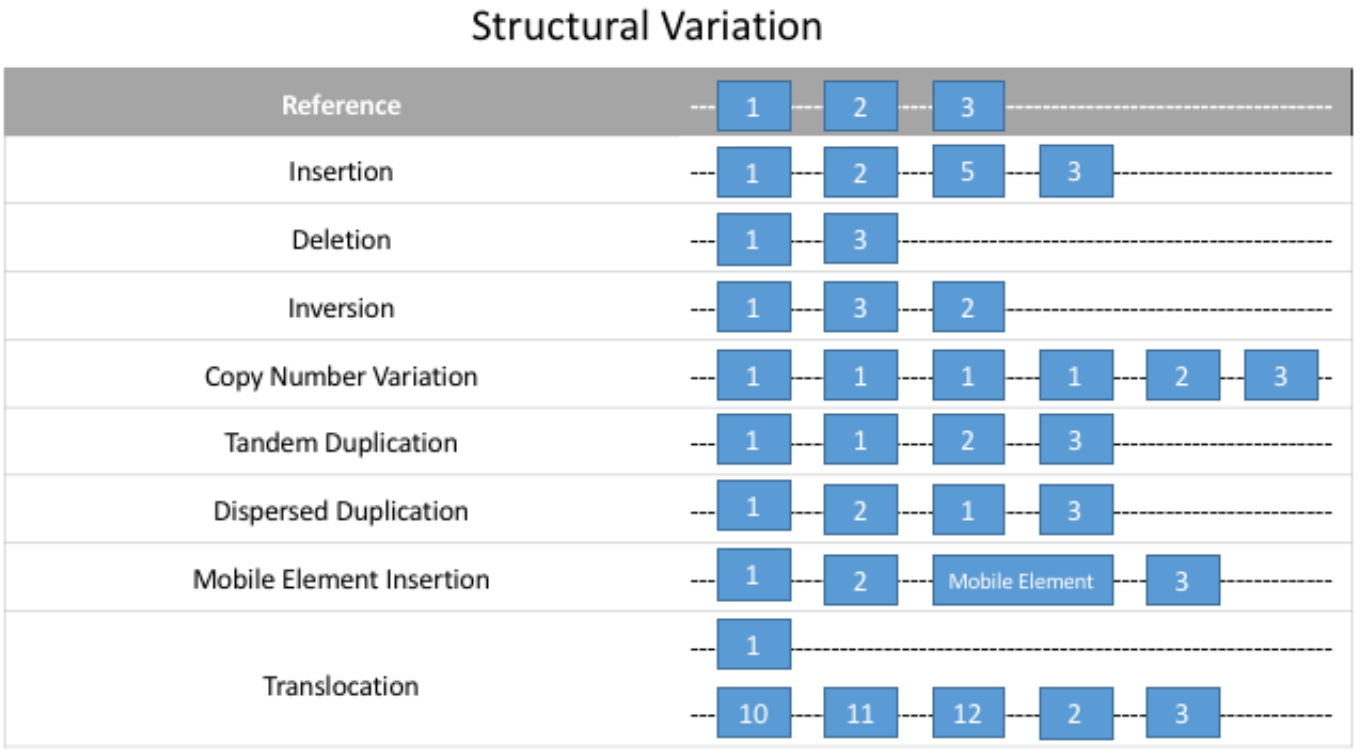


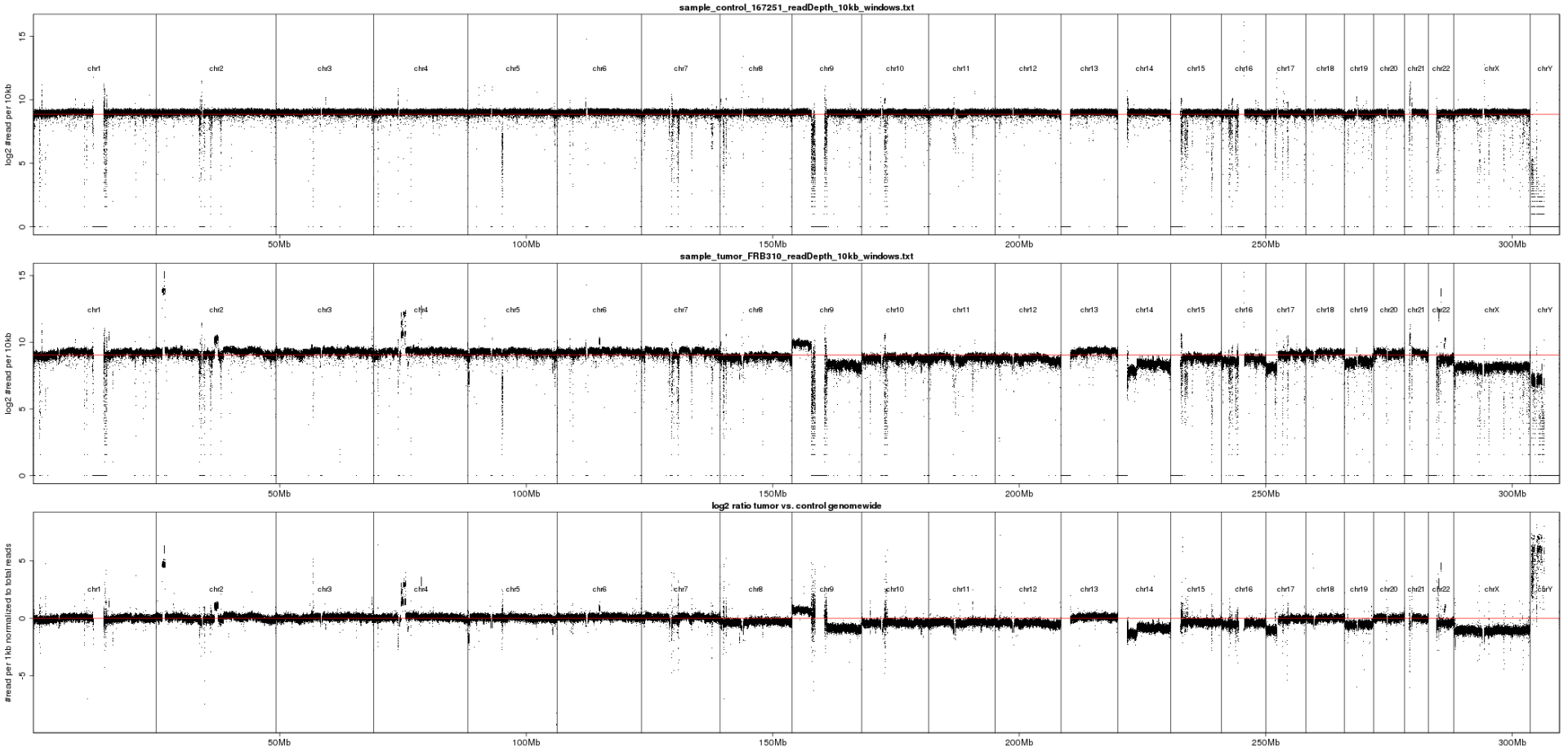
Figure 1: Depiction of different types of structural variants compared to the reference genome. Each different number represents a different gene.

from: Wikipedia / Wikimedia Commons

Copy number variants

Example: identification of CNVs from whole genome sequencing

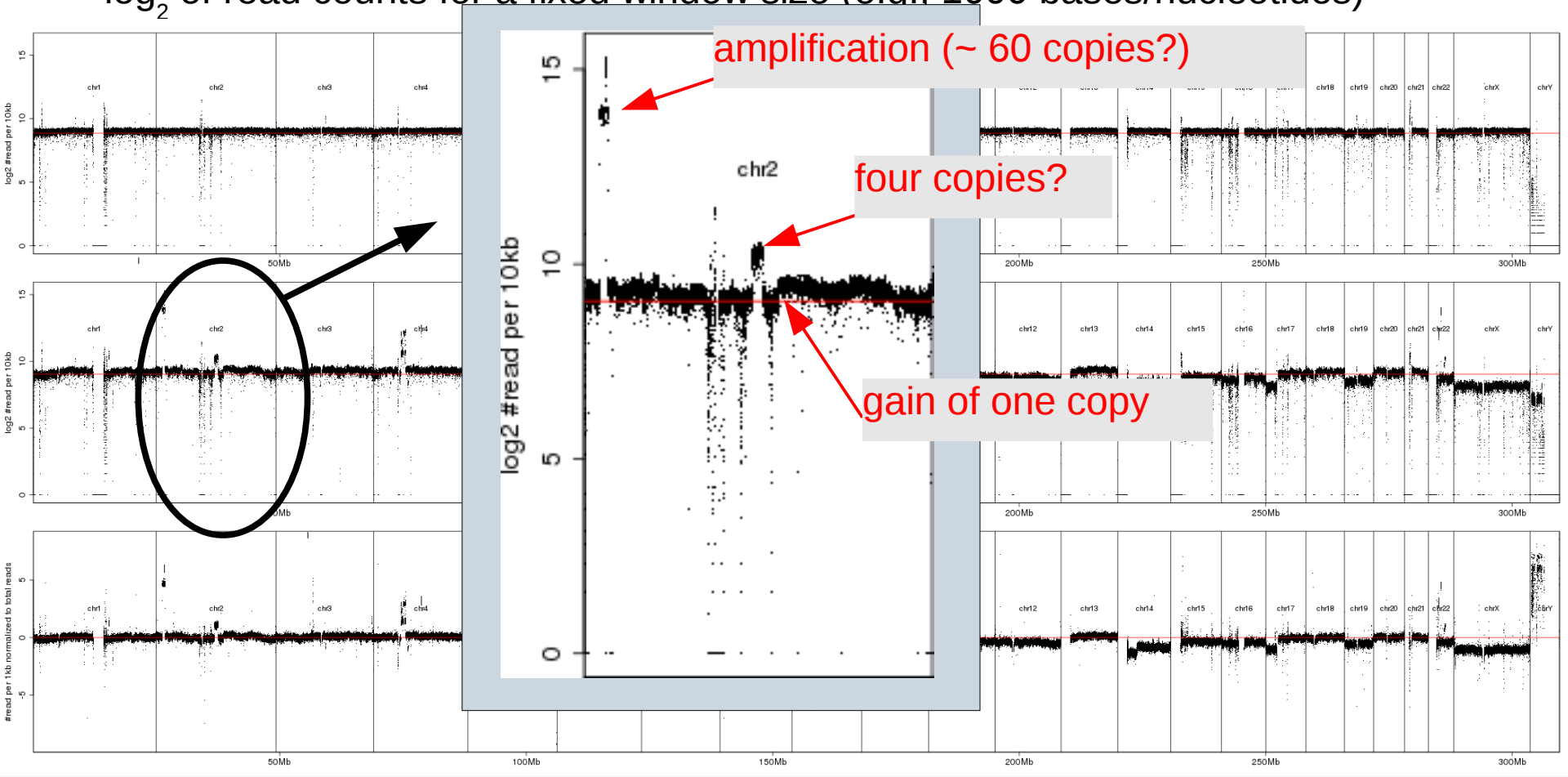
- $\log_2$ of read counts for a fixed window size (e.g., 1000 bases/nucleotides)



Copy number variants

Example: identification of CNVs from whole genome sequencing

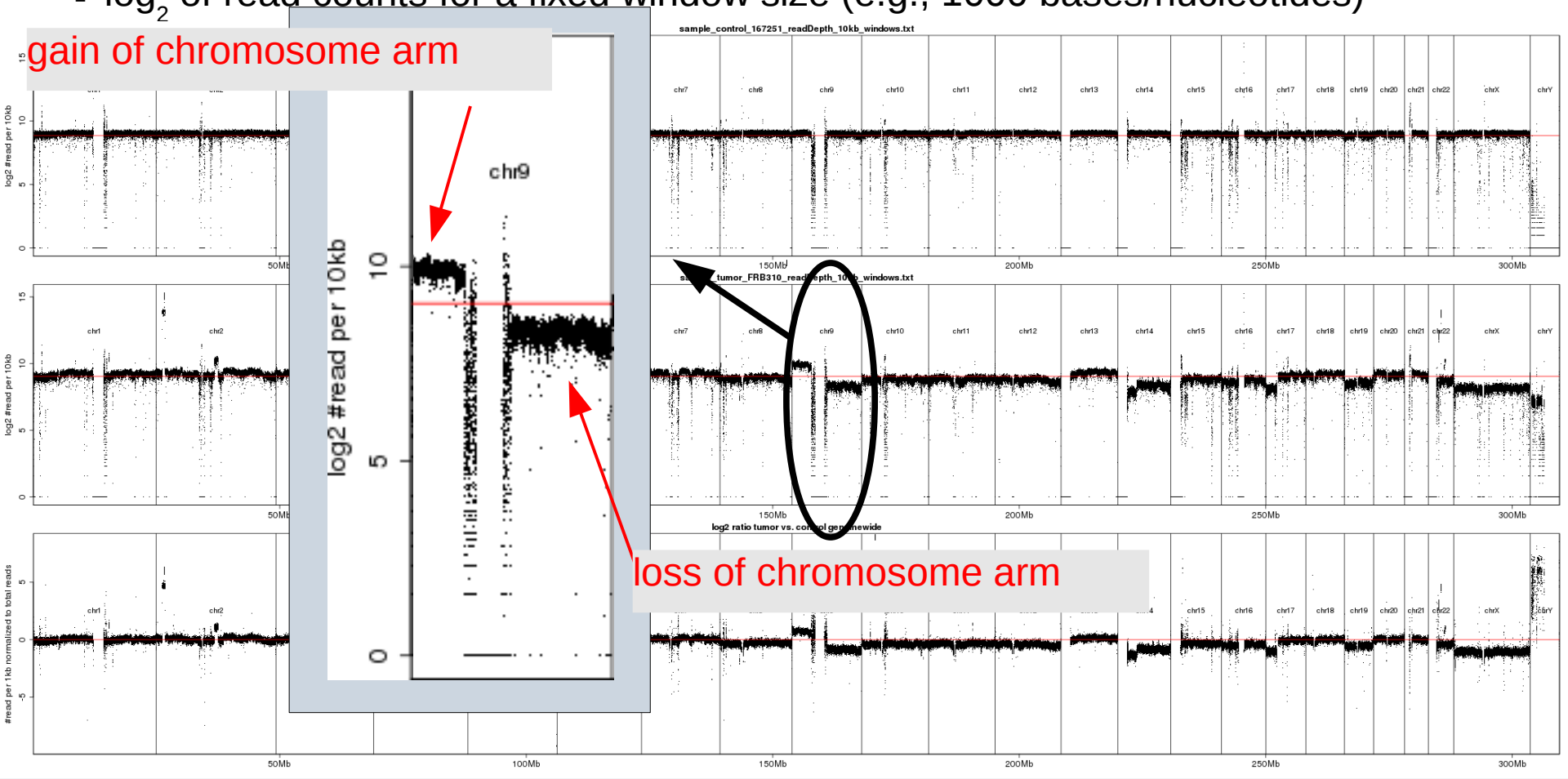
- $\log_2$ of read counts for a fixed window size (e.g., 1000 bases/nucleotides)



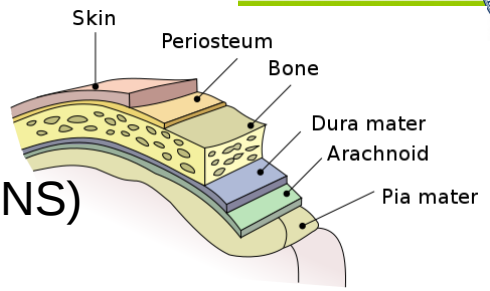
Copy number variants

Example: identification of CNVs from whole genome sequencing

- $\log_2$ of read counts for a fixed window size (e.g., 1000 bases/nucleotides)

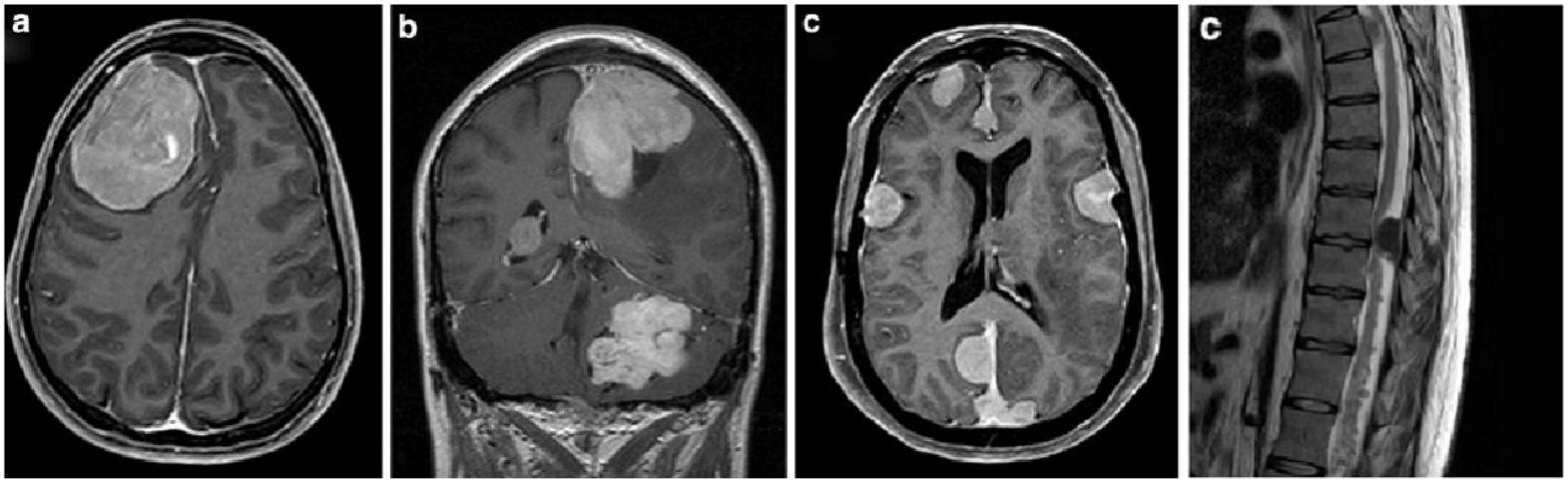


Meningioma: Introduction



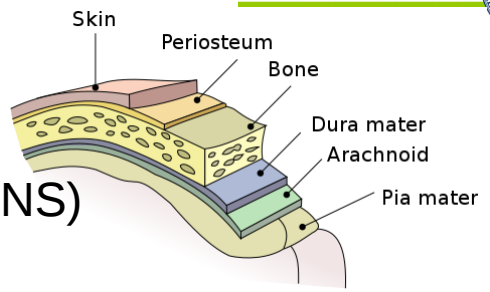
Most common primary brain tumor (~25-30%)

- arise from the meninges (membranes around CNS)



[Goutagny and Kalamarides, J Neurooncol, 2010]

Meningioma: Introduction



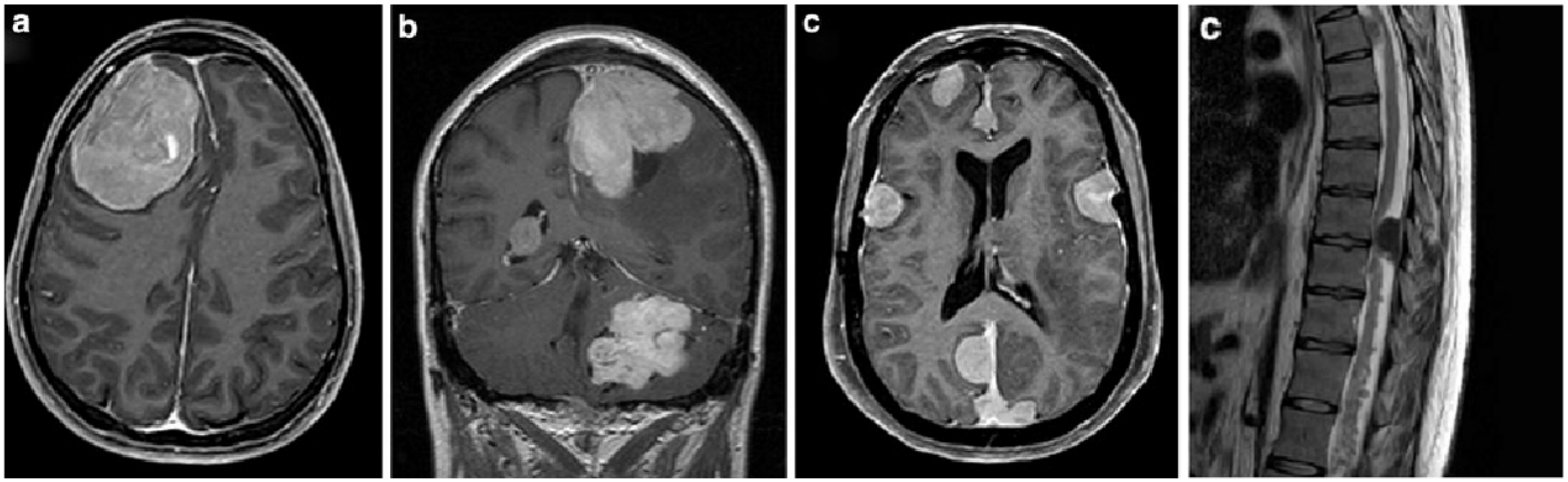
Most common primary brain tumor (~25-30%)

- arise from the meninges (membranes around CNS)

Usually benign (~90% WHO grade I)

- mostly asymptomatic and slow-growing
  - often discovered only after an autopsy (~1-3%; <2.0cm)
  - 10 year recurrence of 9% after complete removal
  - rarely progress to higher grade meningiomas

Some atypical (grade II) and anaplastic/malignant (grade III)

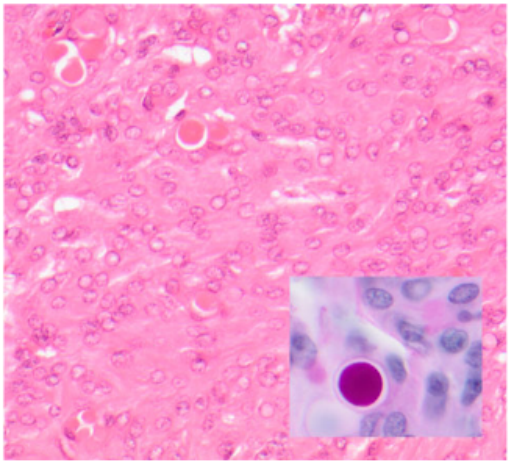


[Goutagny and Kalamarides, J Neurooncol, 2010]

Secretory Meningioma

Subgroup of benign meningiomas (grade I)

- rare occurrence (~3% of meningiomas)



[Reuss et al., Acta Neuropathol, 2013]

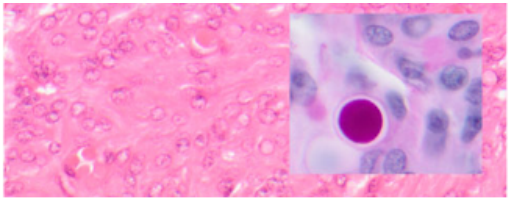
Secretory Meningioma

Subgroup of benign meningiomas (grade I)

- rare occurrence (~3% of meningiomas)
- contains glandular lumina with secretory globules (“pseudopsammomas”)
- often the cause of peritumoral edemas
  - large, nearly hemispheric edemas in ~40-60% of patients
  - can lead to severe, life-threatening complications after surgery

[Regelsberger et al., 2008]





[Reuss et al., Acta Neuropathol, 2013]

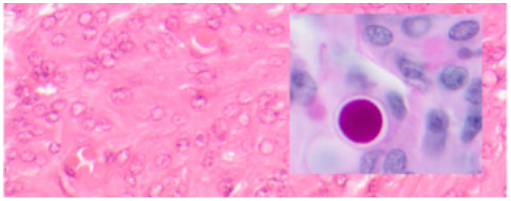
Secretory Meningioma

Subgroup of benign meningiomas (grade I)

- rare occurrence (~3% of meningiomas)
- contains glandular lumina with secretory globules (“pseudopsammomas”)
- often the cause of peritumoral edemas
  - large, nearly hemispheric edemas in ~40-60% of patients
  - can lead to severe, life-threatening complications after surgery
- **<1% possess *NF2* mutations**
  - suggests distinct molecular mechanisms

[Regelsberger et al., 2008]





[Reuss et al., Acta Neuropathol, 2013]

Whole Exome Sequencing

Whole exome sequencing (Illumina HiSeq; Agilent SureSelect) followed by Sanger sequencing for confirmation

Table 1 *KLF4*, *TRAF7* and *NF2* mutations in meningioma

	Meningioma subtype	*n*	*KLF4*	*TRAF7*	*NF2*
Exome series	Secretory WHO grade I	16	16/16 (100 %)	16/16 (100 %)	0/16
Validation series	Secretory WHO grade I	14	14/14 (100 %)	13/14 (93 %)	0/14
	Meningothelial WHO grade I	21	0	5/21 (24 %)	3/18 (17 %)
	Transitional WHO grade I	18	0	0	13/16 (81 %)
	Fibroblastic WHO grade I	16	0	0	8/11 (73 %)
	Atypical WHO grade II	20	0	2/21 (10 %)	9/17 (53 %)
	Anaplastic WHO grade III	14	0	0	9/13 (69 %)

[Reuss et al., Acta Neuropathol, 2013]

Whole Exome Sequencing

Whole exome sequencing (Illumina HiSeq; Agilent SureSelect) followed by Sanger sequencing for confirmation

Table 1 *KLF4*, *TRAF7* and *NF2* mutations in meningioma

	Meningioma subtype	*n*	*KLF4*	*TRAF7*	*NF2*
Exome series	Secretory WHO grade I	16	16/16 (100 %)	16/16 (100 %)	0/16
Validation series	Secretory WHO grade I	14	14/14 (100 %)	13/14 (93 %)	0/14
	Meningothelial WHO grade I	21	0	5/21 (24 %)	3/18 (17 %)
	Transitional WHO grade I	18	0	0	13/16 (81 %)
	Fibroblastic WHO grade I	16	0	0	8/11 (73 %)
	Atypical WHO grade II	20	0	2/21 (10 %)	9/17 (53 %)
	Anaplastic WHO grade III	14	0	0	9/13 (69 %)

[Reuss et al., Acta Neuropathol, 2013]

Whole Exome Sequencing

Whole exome sequencing (Illumina HiSeq; Agilent SureSelect) followed by Sanger sequencing for confirmation

Table 1 *KLF4*, *TRAF7* and *NF2* mutations in meningioma

	Meningioma subtype	*n*	*KLF4*	*TRAF7*	*NF2*
Exome series	Secretory WHO grade I	16	16/16 (100 %)	16/16 (100 %)	0/16
Validation series	Secretory WHO grade I	14	14/14 (100 %)	13/14 (93 %)	0/14
	Meningothelial WHO grade I	21	0	5/21 (24 %)	3/18 (17 %)
	Transitional WHO grade I	18	0	0	13/16 (81 %)
	Fibroblastic WHO grade I	16	0	0	8/11 (73 %)
	Atypical WHO grade II	20	0	2/21 (10 %)	9/17 (53 %)
	Anaplastic WHO grade III	14	0	0	9/13 (69 %)

[Reuss et al., Acta Neuropathol, 2013]

KLF4 K409Q

recurrent KLF4 mutation in ALL secretory meningiomas:
chr9:110,249,348 T→G (K409Q; lysine→glutamine**)**

```
human|479aa      GPPLSNGHRPAAHDFPLGRQLPSRTTPTLGLEEVLSSRDCHPALPLPPGFHPHPGPNYPS 350
human2|504aa     GPPLSNGHRPAAHDFPLGRQLPSRTTPTLGLEEVLSSRDCHPALPLPPGFHPHPGPNYPS 341
bovine           GPPLSNGHRPPAHDFPLGRQLPSRTTPTLGAEELLSSRDCHPALPLPPGFHPHPGPNYPP 348
mouse            GPQLSNGHRPNTHDFPLGRQLPTRTTPTLSPEELLNSRDCHPGLPLPPGFHPHPGPNYPP 354
xenopus          PDMHCLSLMSMAQHYPMASPYQTHFTSQ--PNVQFHGQFGVYRDPMKV-HPSMHG----- 290
zebrafish        TEMNTQHHIDISHQQQFIN--NAHFPPQYAQHQQYHGHFNMFSEPLRANHPAMSG----- 289
                              ::.  :     :: ..    .    .:      *:   . .  *

human|479aa      FLPDQMQPQVPPLHYQ---------------------------------ELMPPGSCMP 376
human2|504aa     FLPDQMQPQVPPLHYQGQSRGFVARAGEPCVCWPHFGTHGMMLTPPSSPLELMPPGSCMP 401
bovine           FLPDQMQPQVPPLHYQ---------------------------------ELMPPGSCMP 374
mouse            FLPDQMQSQVPSLHYQ---------------------------------ELMPPGSCLP 380
xenopus          --MIVTPPSSPLLEYY-----------------------------------PAMAAT 310
zebrafish        --VMLTPPSSPLLG-------------------------------------FLSP 305
                       .. * *                                        .

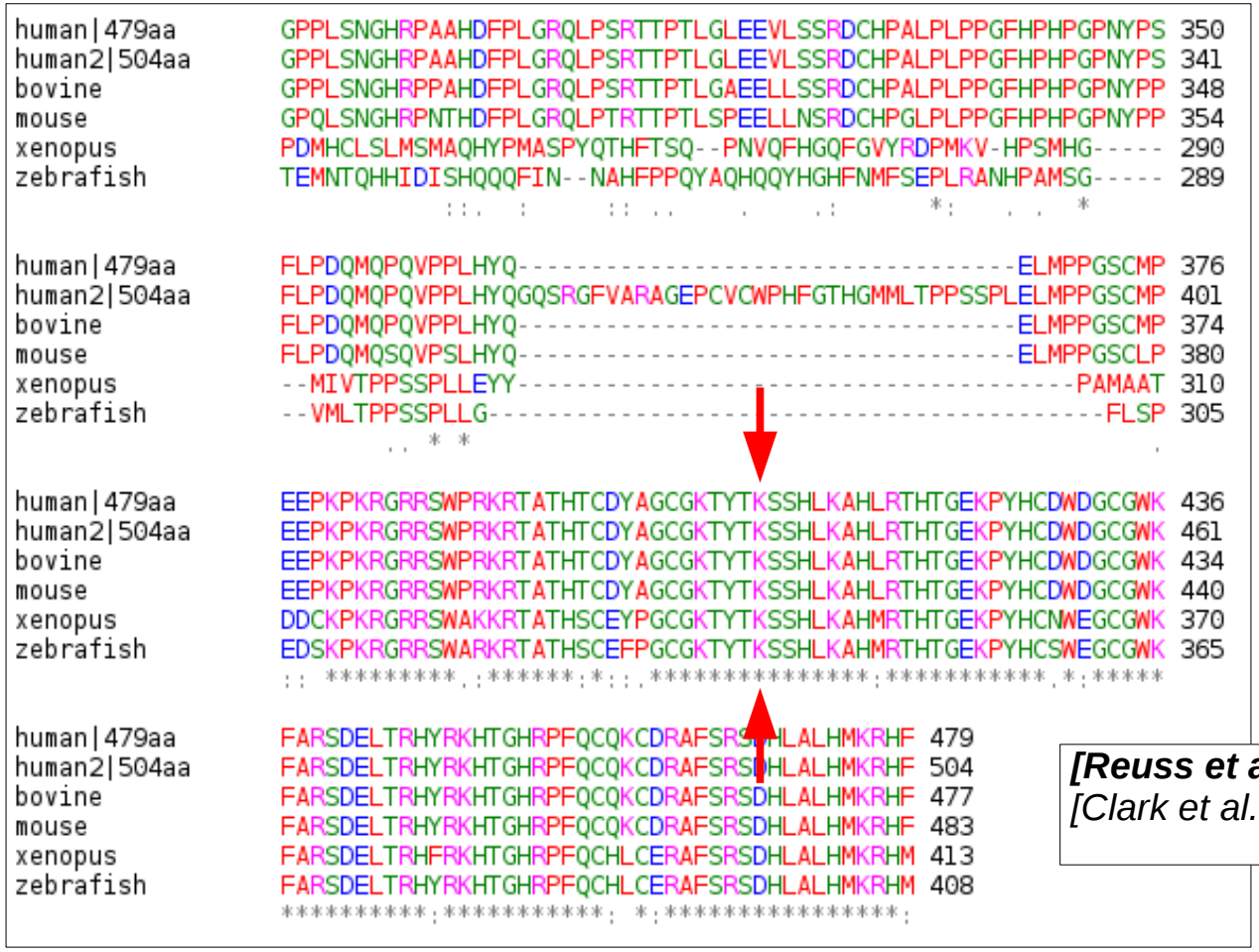
human|479aa      EEPKPKRGRRSWPRKRTATHTCDYAGCGKTYTKSSHLKAHLRTHTGEKPYHCDWDGCGWK 436
human2|504aa     EEPKPKRGRRSWPRKRTATHTCDYAGCGKTYTKSSHLKAHLRTHTGEKPYHCDWDGCGWK 461
bovine           EEPKPKRGRRSWPRKRTATHTCDYAGCGKTYTKSSHLKAHLRTHTGEKPYHCDWDGCGWK 434
mouse            EEPKPKRGRRSWPRKRTATHTCDYAGCGKTYTKSSHLKAHLRTHTGEKPYHCDWDGCGWK 440
xenopus          DDCKPKRGRRSWAKKRTATHSCEYPGCGKTYTKSSHLKAHMRTHTGEKPYHCNWEGCGWK 370
zebrafish        EDSKPKRGRRSWARKRTATHSCEFPGCGKTYTKSSHLKAHMRTHTGEKPYHCSWEGCGWK 365
                 :: *********.:******:*:::**************:***********.*:*****

human|479aa      FARSDELTRHYRKHTGHRPFQCQKCDRAFSRSDHLALHMKRHF 479
human2|504aa     FARSDELTRHYRKHTGHRPFQCQKCDRAFSRSDHLALHMKRHF 504
bovine           FARSDELTRHYRKHTGHRPFQCQKCDRAFSRSDHLALHMKRHF 477
mouse            FARSDELTRHYRKHTGHRPFQCQKCDRAFSRSDHLALHMKRHF 483
xenopus          FARSDELTRHFRKHTGHRPFQCHLCERAFSRSDHLALHMKRHM 413
zebrafish        FARSDELTRHYRKHTGHRPFQCHLCERAFSRSDHLALHMKRHM 408
                 **********:***********: *:*****************:
```

[Reuss et al., Acta Neuropathol, 2013]
[Clark et al., Science, 2013]

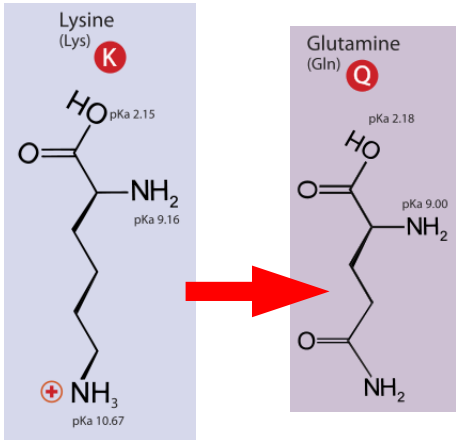
KLF4 K409Q

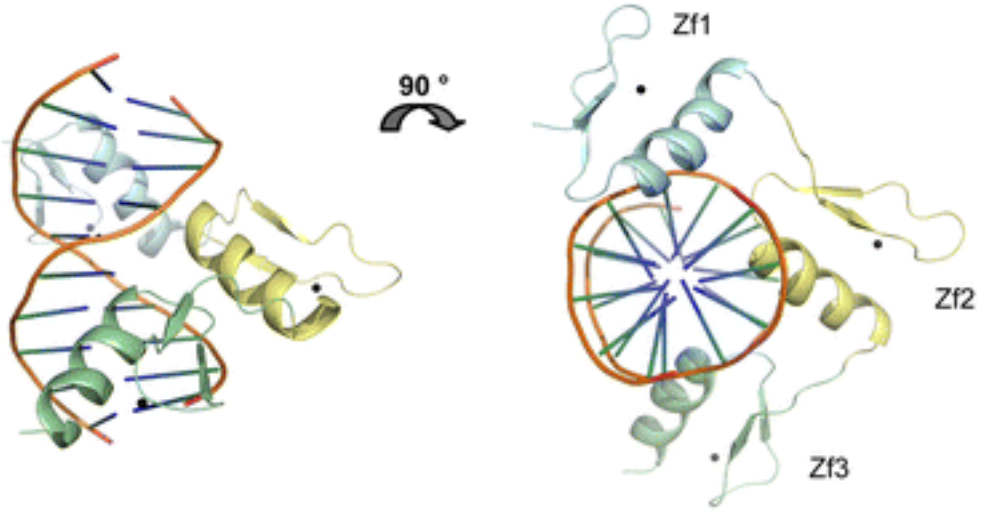
recurrent KLF4 mutation in ALL secretory meningiomas:
chr9:110,249,348 T→G (K409Q; lysine→glutamine**)**

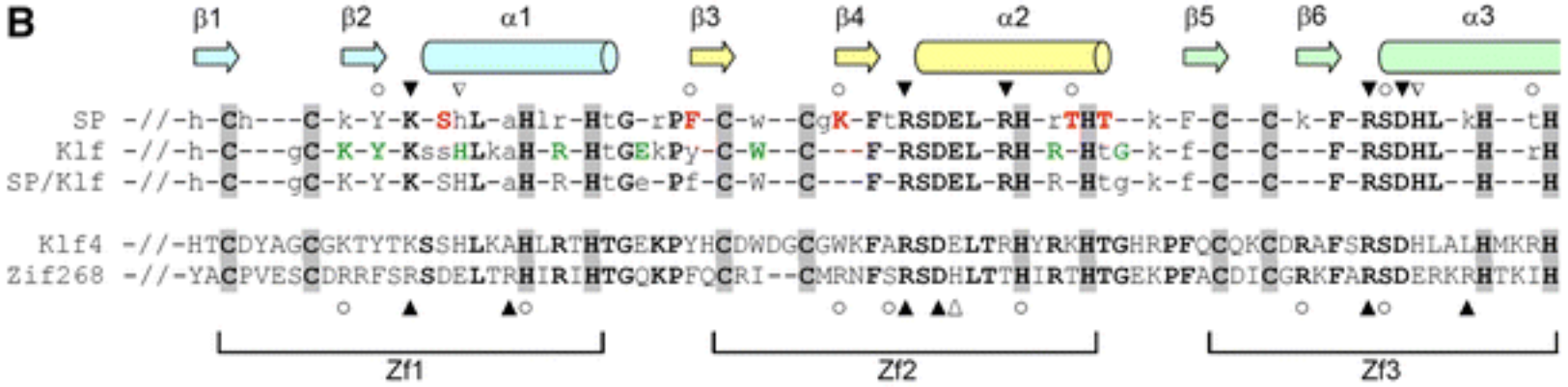


[Reuss et al., Acta Neuropathol, 2013]
[Clark et al., Science, 2013]

KLF4 K409Q: Effect On Function?

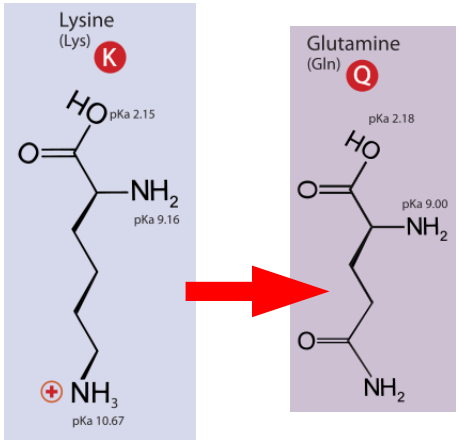


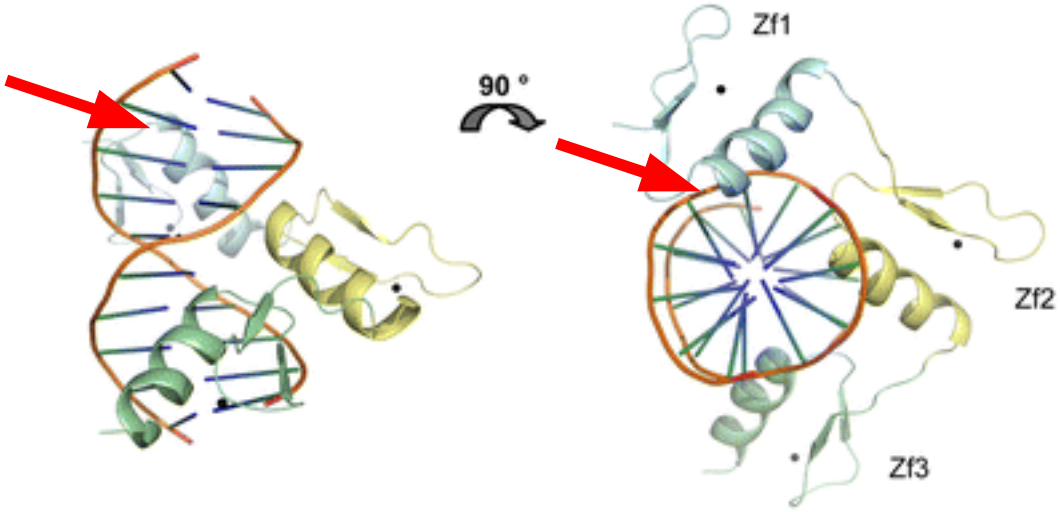


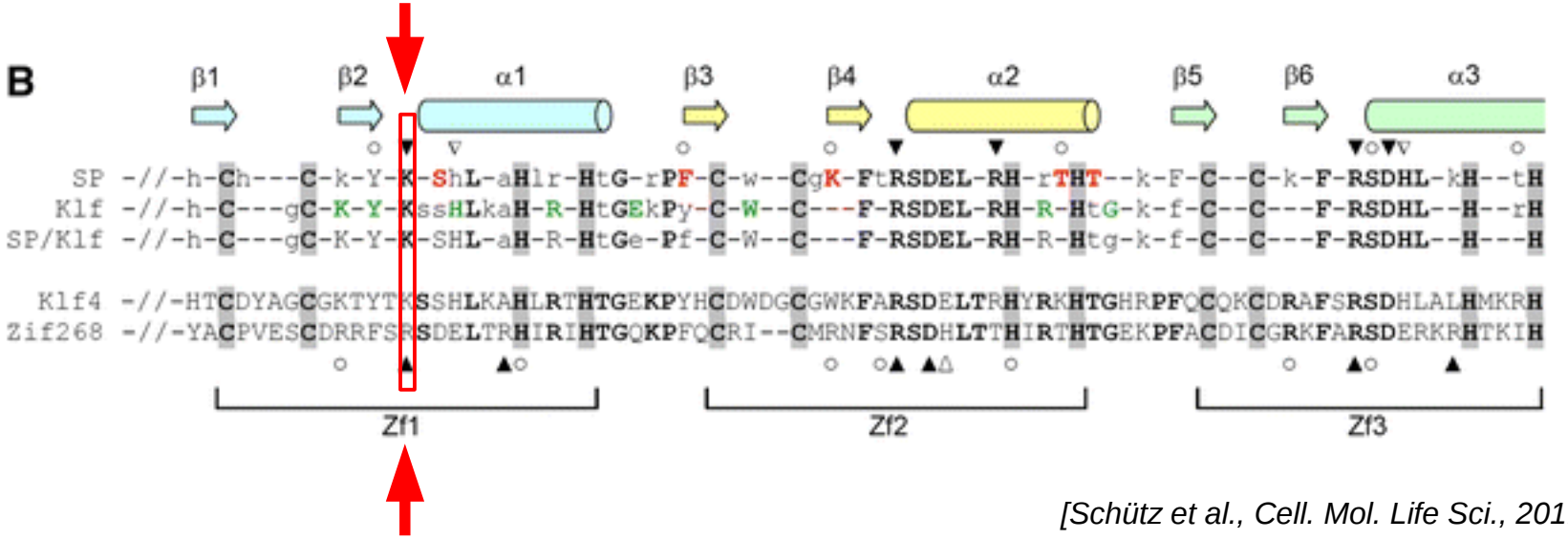


[Schütz et al., Cell. Mol. Life Sci., 2011]

KLF4 K409Q: Effect On Function?

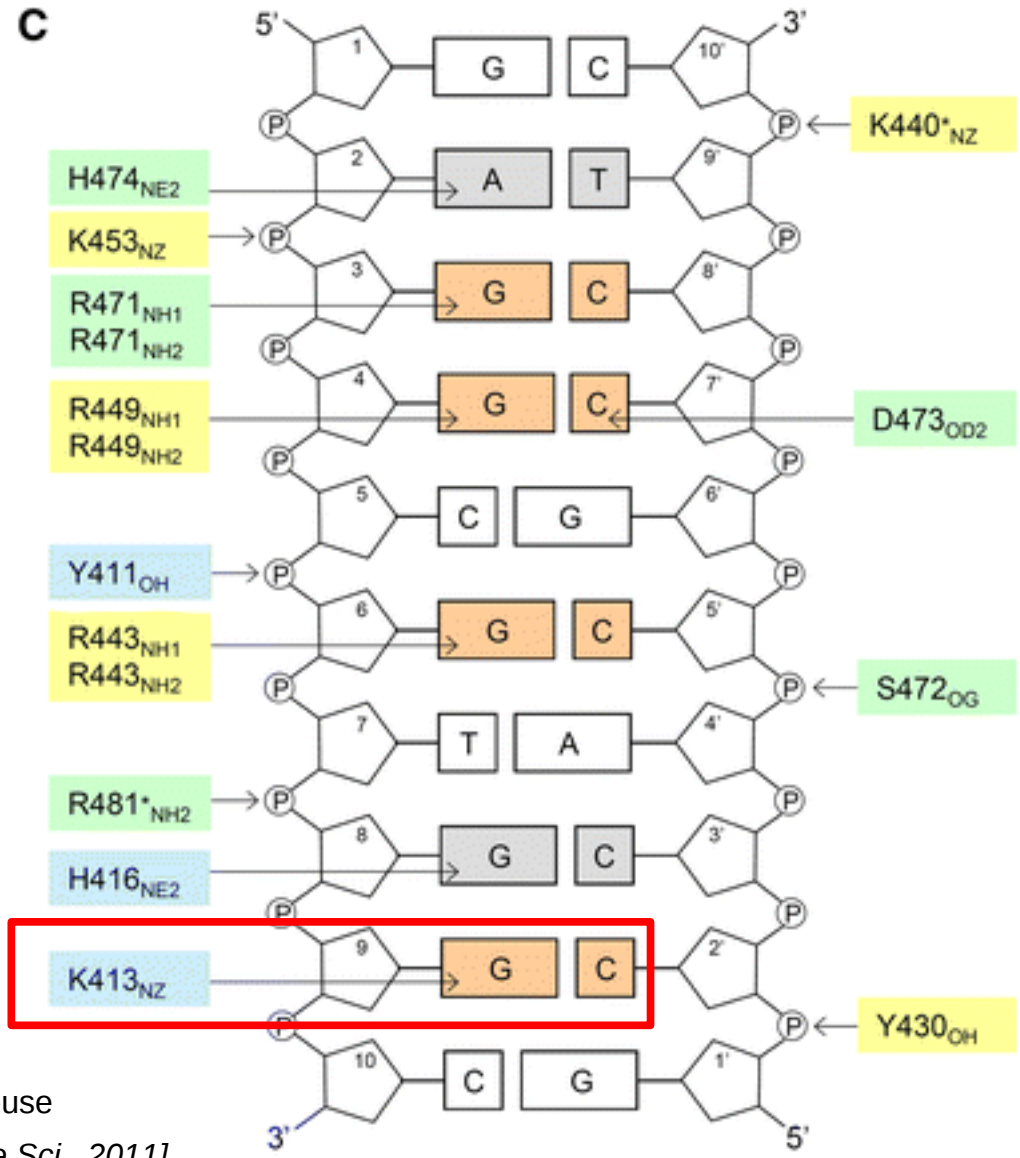






[Schütz et al., Cell. Mol. Life Sci., 2011]

KLF4 K409Q: Effect On Function?



DNA-binding of KLF4 in mouse

[Schütz et al., Cell. Mol. Life Sci., 2011]

HPC & SFT

Hemangiopericytoma (HPC) and solitary fibrous tumor (SFT)

- non-CNS tumors considered as two variants of a single entity (WHO classification of soft tissue and bone tumors)

HPC & SFT

Hemangiopericytoma (HPC) and solitary fibrous tumor (SFT)

- non-CNS tumors considered as two variants of a single entity (WHO classification of soft tissue and bone tumors)
  - recurrent *NAB2-STAT6* fusion gene identified recently *[Chmielecki et al, 2013; Robinson et al, 2013]*
    - inversion and juxtaposition on Chr. 12q13
    - genes on opposing DNA strands with overlapping 3'UTRs

HPC & SFT

Hemangiopericytoma (HPC) and solitary fibrous tumor (SFT)

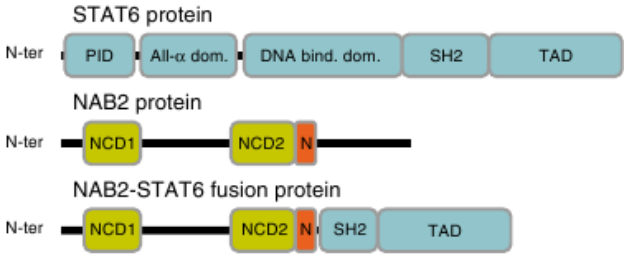
- non-CNS tumors considered as two variants of a single entity (WHO classification of soft tissue and bone tumors)
  - recurrent *NAB2-STAT6* fusion gene identified recently *[Chmielecki et al, 2013; Robinson et al, 2013]*
    - inversion and juxtaposition on Chr. 12q13
    - genes on opposing DNA strands with overlapping 3'UTRs
- CNS tumors were considered as two distinct entities (WHO classification of tumors of the central nervous system)
  - Question: also driven by a *NAB2-STAT6* fusion?

HPC & SFT

Hemangiopericytoma (HPC) and solitary fibrous tumor (SFT)

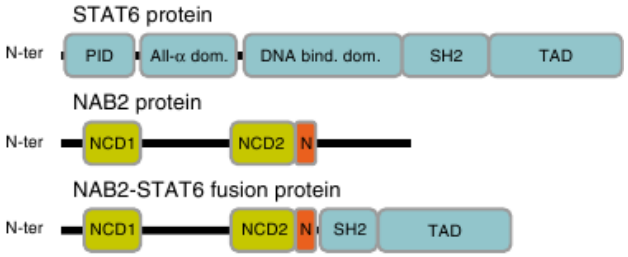
- non-CNS tumors considered as two variants of a single entity (WHO classification of soft tissue and bone tumors)
  - recurrent *NAB2-STAT6* fusion gene identified recently *[Chmielecki et al, 2013; Robinson et al, 2013]*
    - inversion and juxtaposition on Chr. 12q13
    - genes on opposing DNA strands with overlapping 3'UTRs
- CNS tumors were considered as two distinct entities (WHO classification of tumors of the central nervous system)
  - Question: also driven by a *NAB2-STAT6* fusion?
- Whole exome sequencing (*problematic for translocations*)

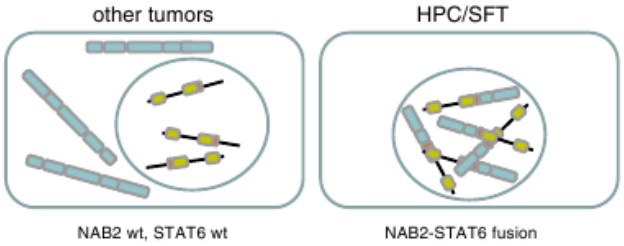
STAT6 IHC Staining In The Nucleus



[Schweizer et al., Acta Neuropathol, 2013]

STAT6 IHC Staining In The Nucleus

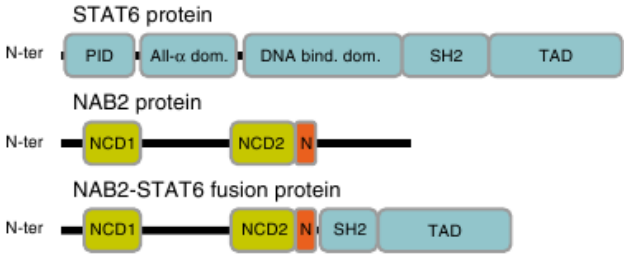


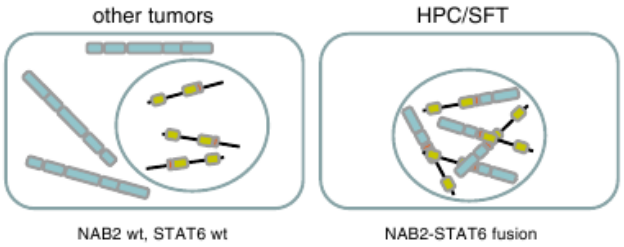


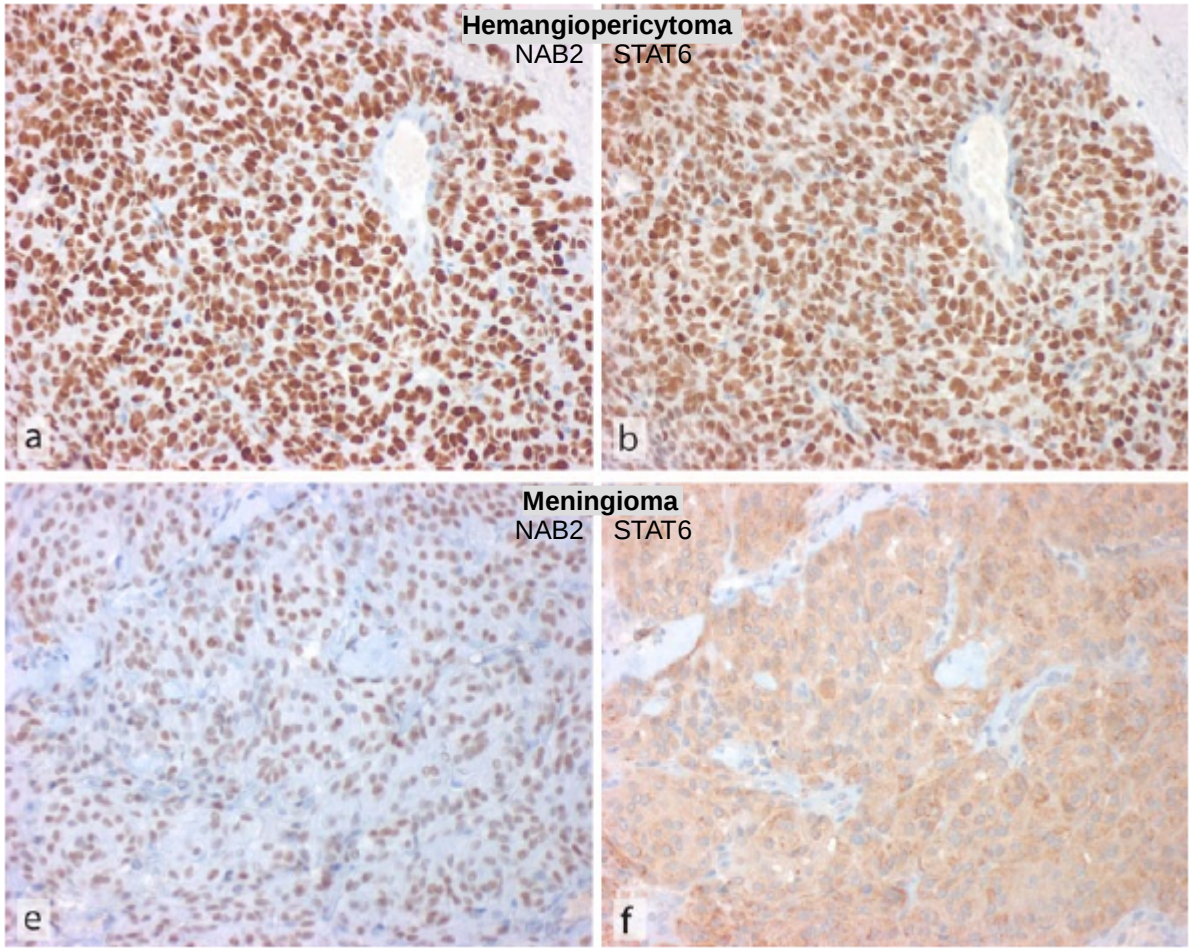
[Schweizer et al., Acta Neuropathol, 2013]

STAT6 IHC Staining In The Nucleus

Diagnosis	N	NAB2 IHC	STAT6 IHC	
		Nucleus	Nucleus	Cytoplasm
Hemangiopericytoma	37	35/37	35/37	0/37
Solitary fibrous tumor	25	25/25	25/25	0/25

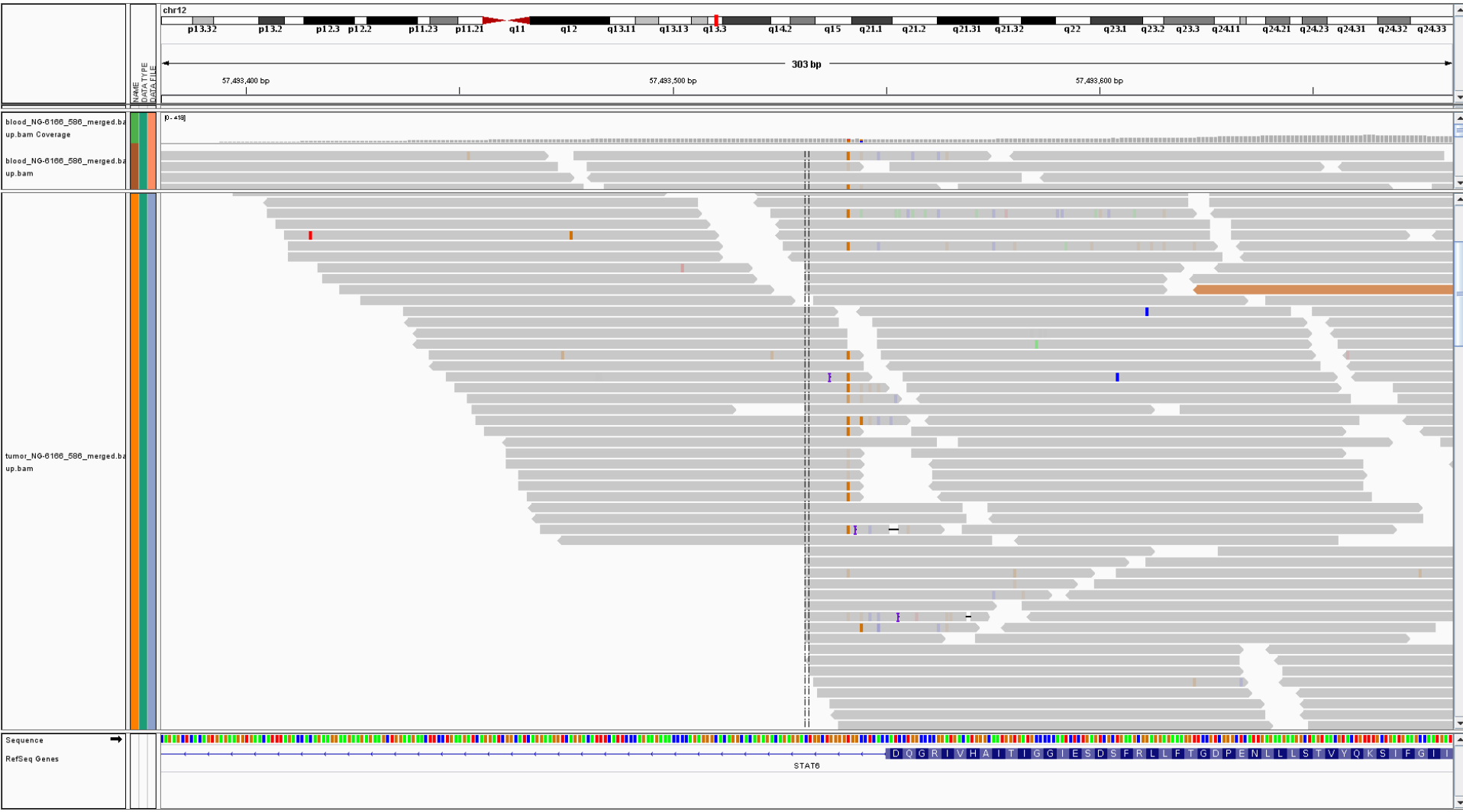




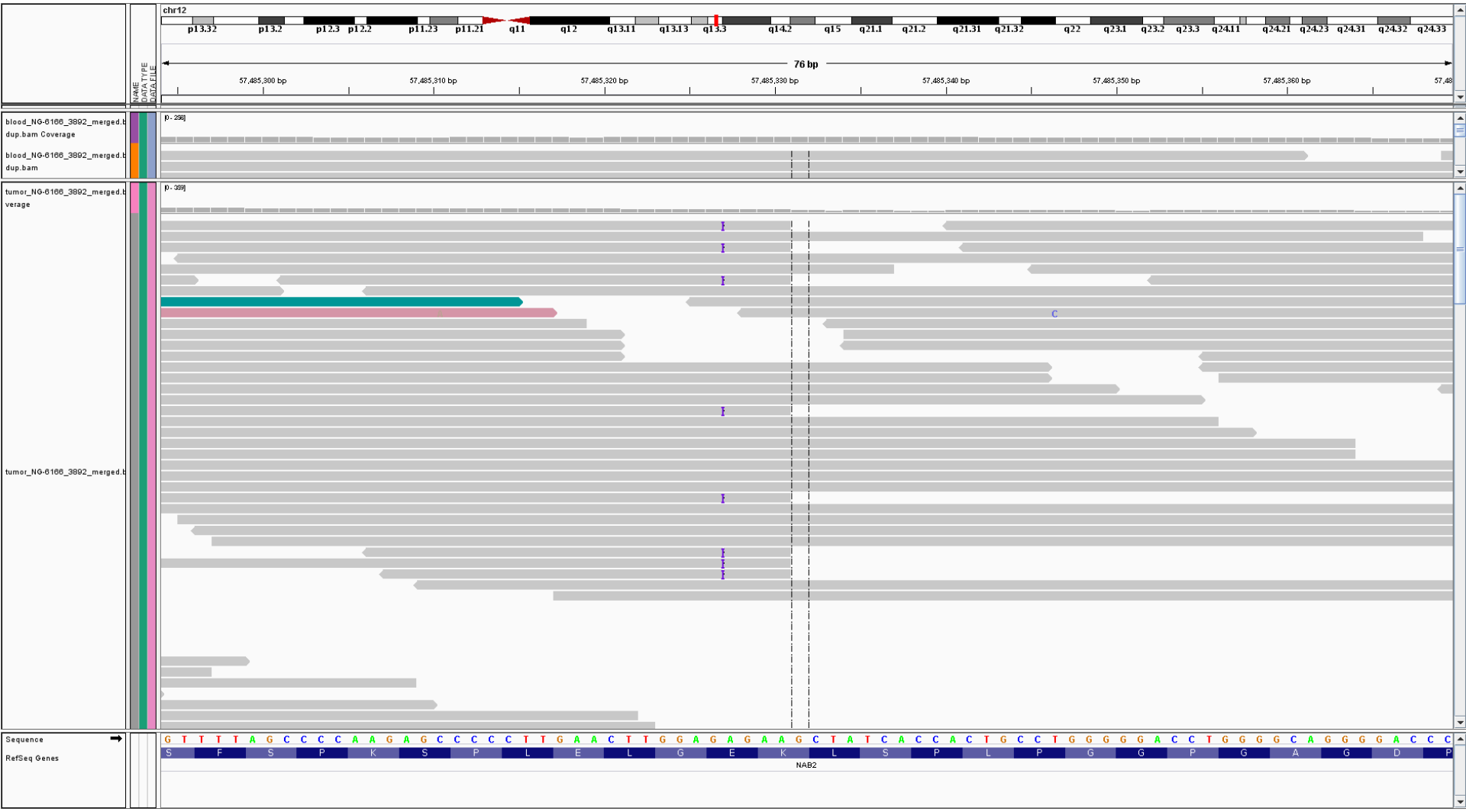


[Schweizer et al., Acta Neuropathol, 2013]

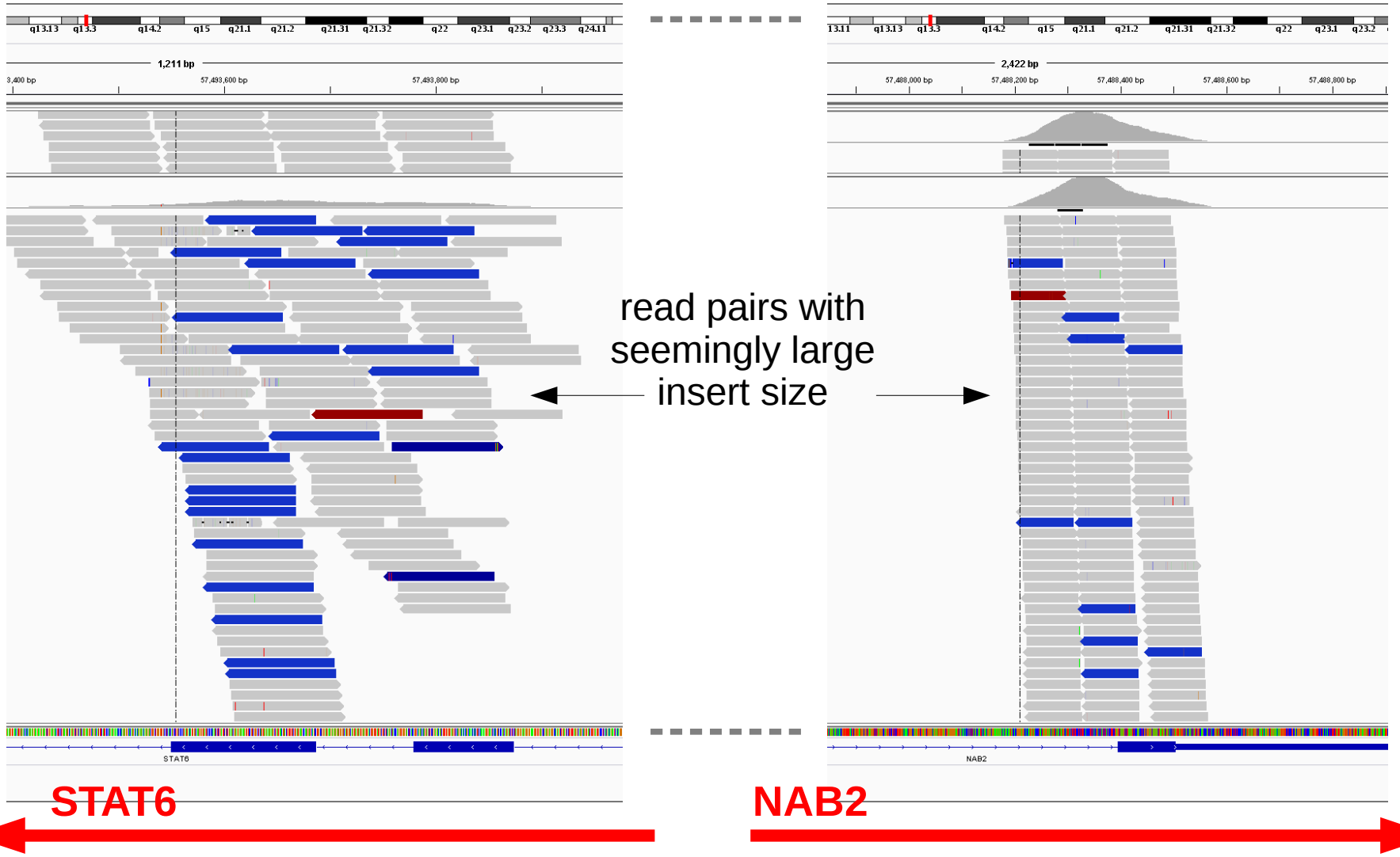
NAB2-STAT6 Fusion: Example 1



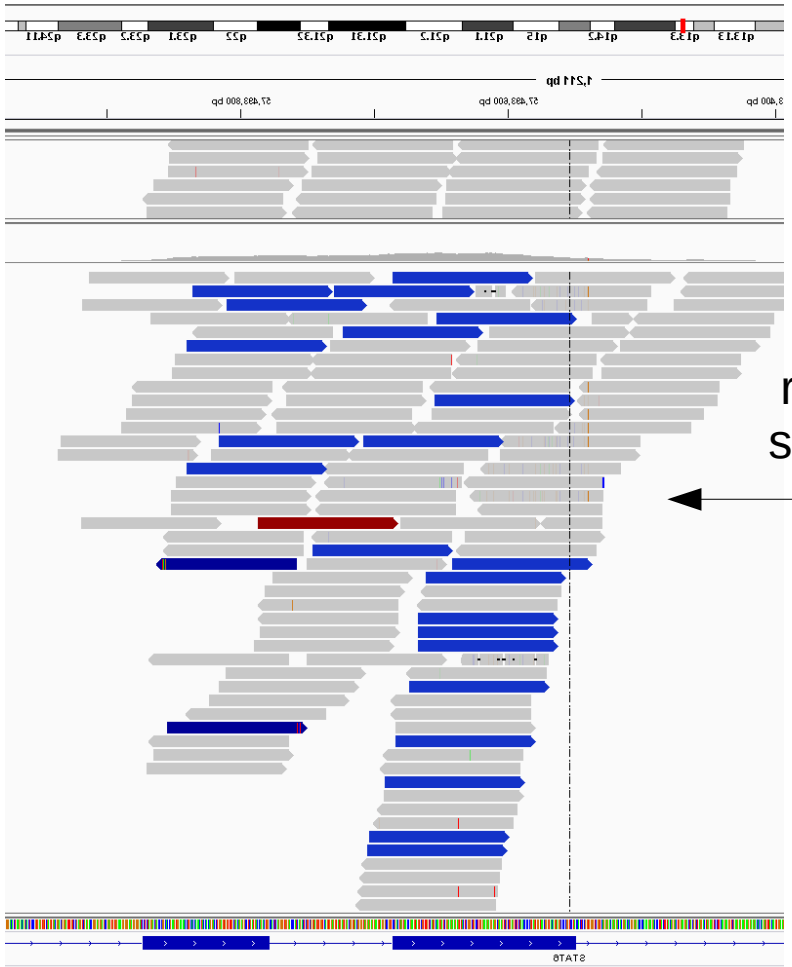
NAB2-STAT6 Fusion: Example 1



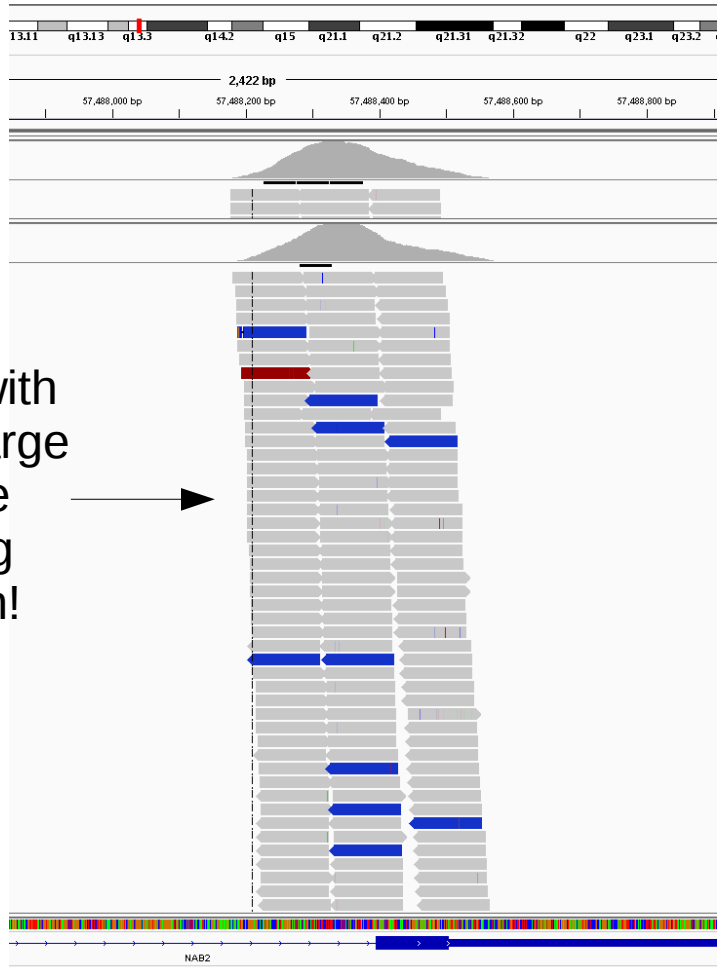
NAB2-STAT6 Fusion: Example 2



NAB2-STAT6 Fusion: Example 2



read pairs with seemingly large insert size and wrong orientation!



STAT6 → **NAB2**

Acknowledgements

Freie Universität Berlin: Computational Neuropathology & Oncology Group

- Anja Seidel
- Veronica Rodrigues de Melo Costa
- Calvinna Caswara
- Marc Horlacher
- Rafael Alberdi Vallejo

University of Torino, Italy

- Ferdinando Di Cunto, Paolo Provero, Ivan Molineris, Ugo Ala

German Cancer Research Center (DKFZ), Heidelberg

- Rainer König (now University of Jena), Barbara Hutter, Natalie Jäger, Peter Lichter, Stefan Pfister, David Jones

University Hospital Heidelberg, Germany

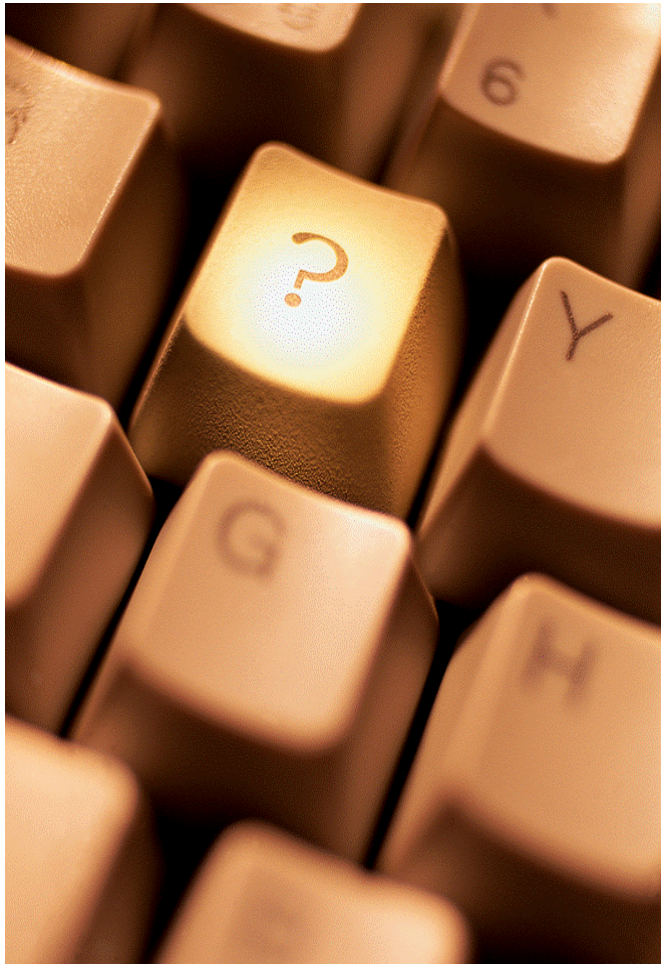
- Andreas von Deimling, David Reuss, David Capper, Leonille Schweizer

Nijmegen Centre for Molecular Life Sciences / Netherlands Bioinformatics Centre

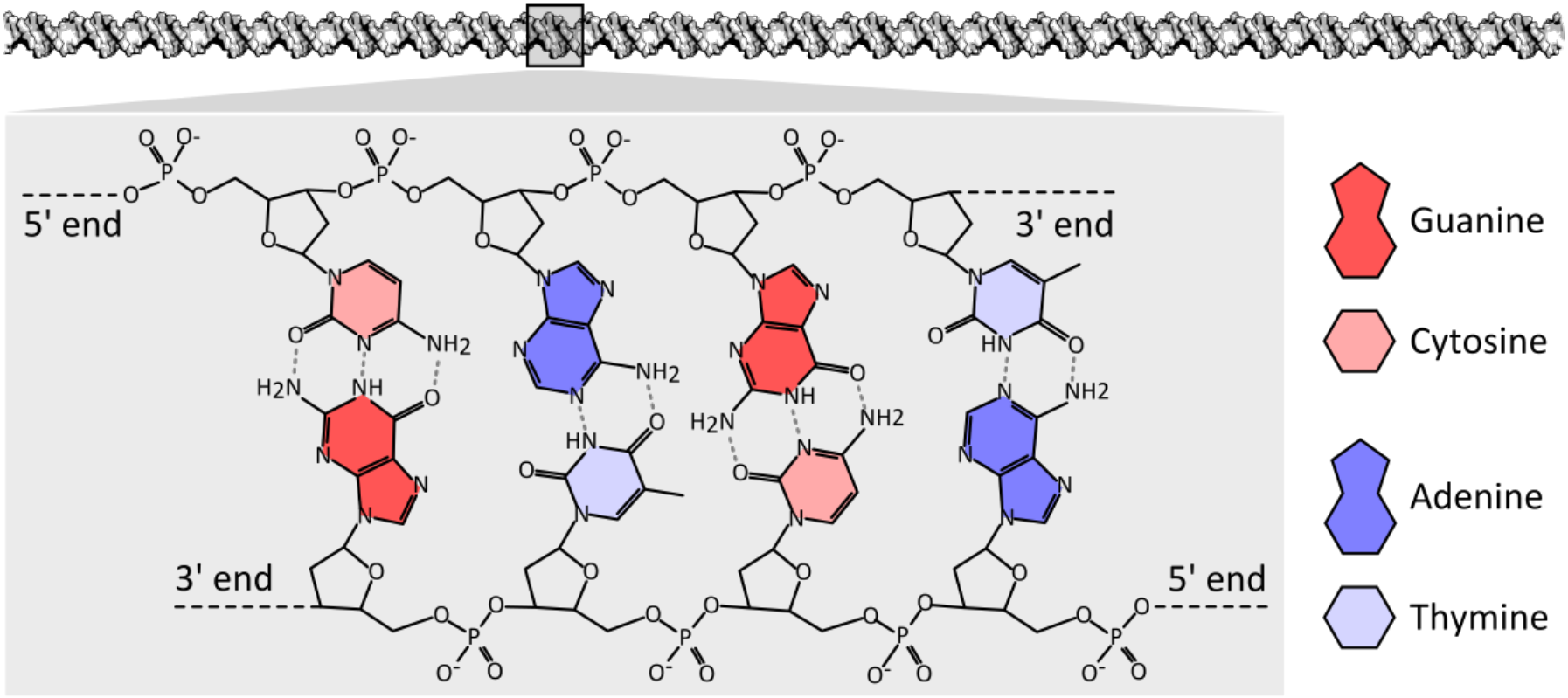
- Han Brunner, Marc van Driel, Martin Oti

Freie Universität Berlin

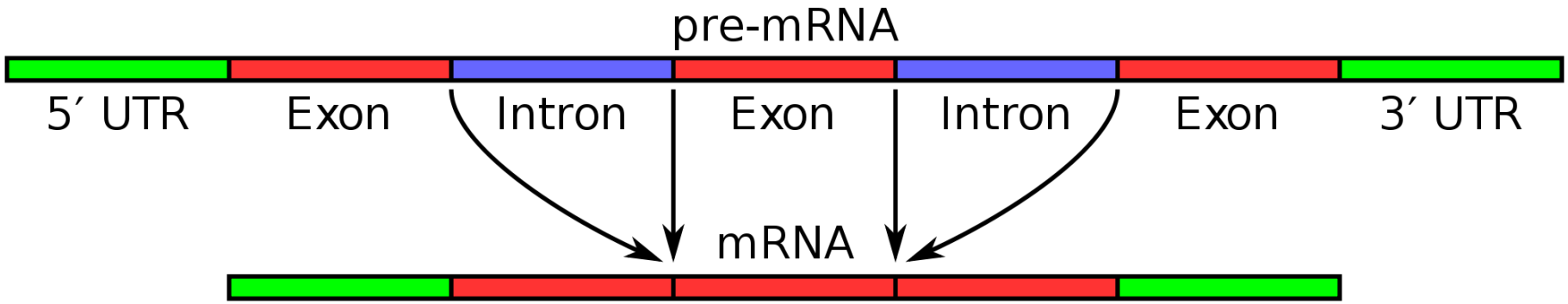
Questions?



DNA



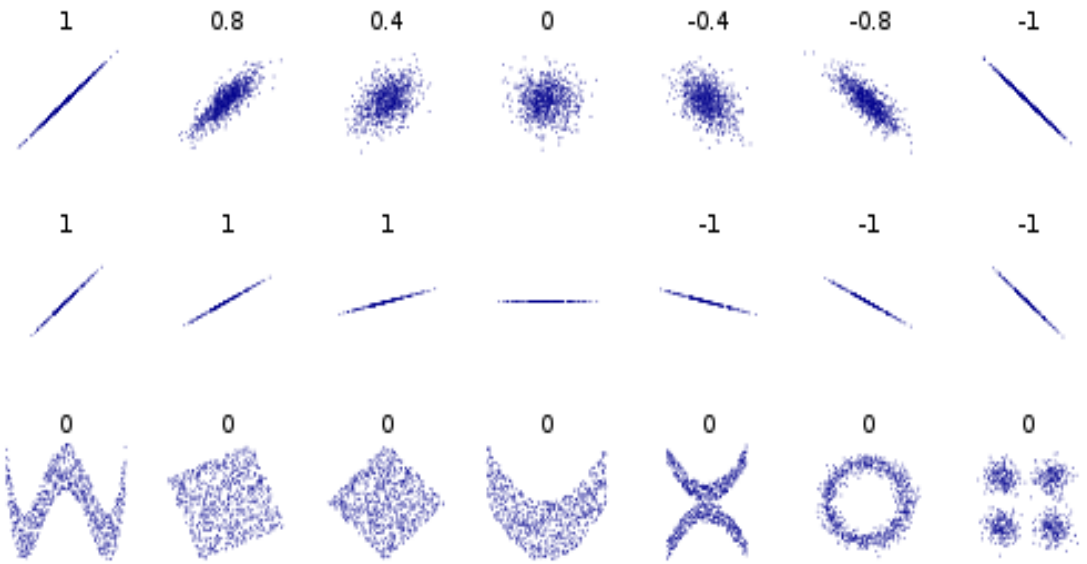
Splicing of messanger RNAs (mRNAs)



from: Wikipedia / Wikimedia Commons

Pearson correlation coefficient

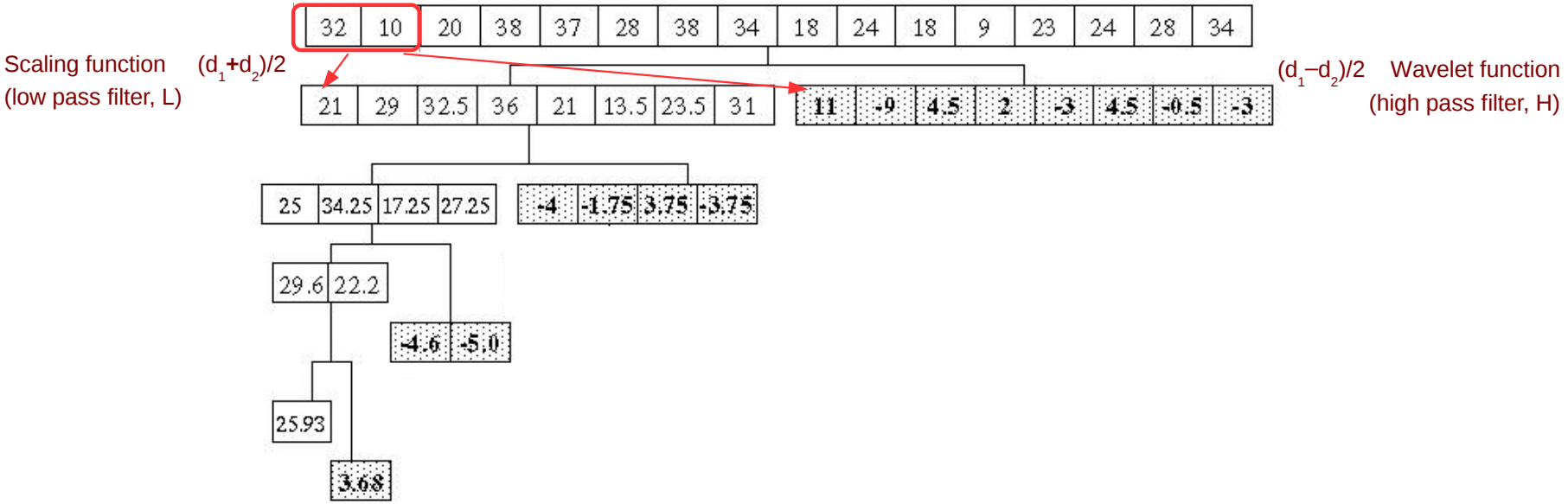
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$



Can we better exploit the spatial information?

Techniques from image processing

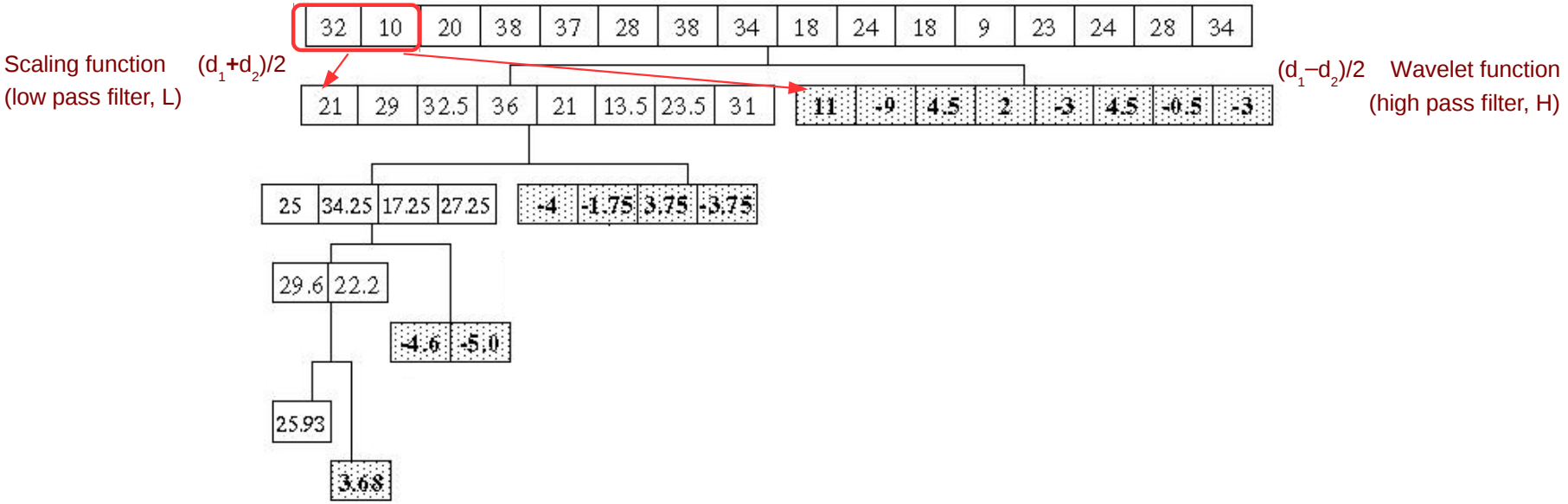
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



Can we better exploit the spatial information?

Techniques from image processing

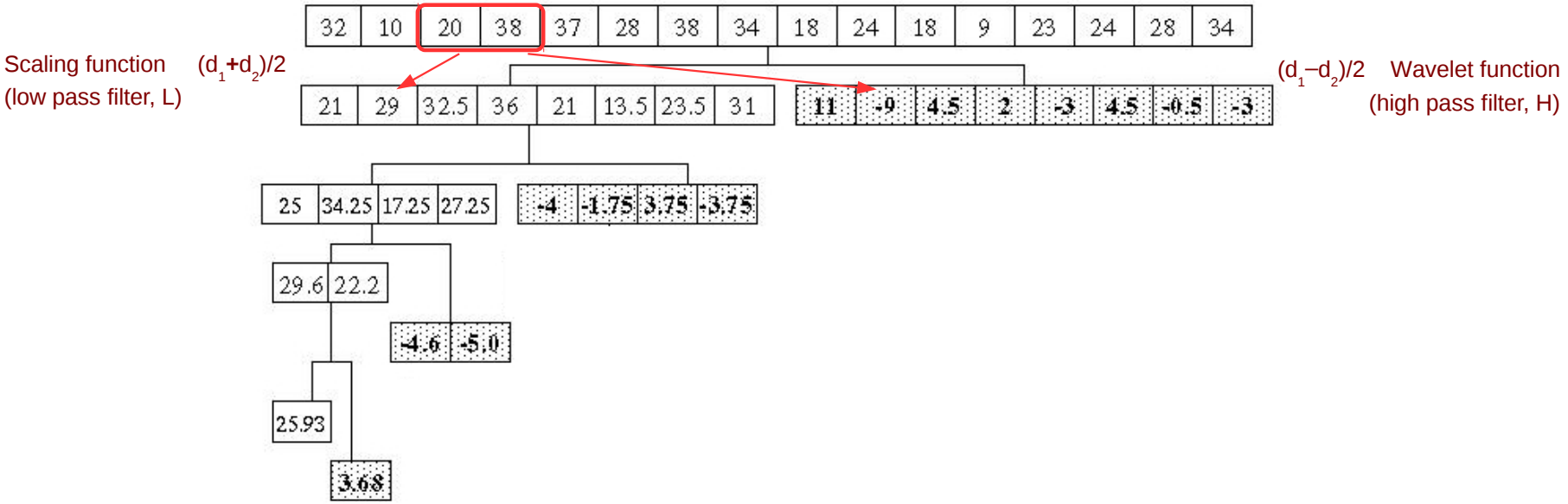
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



Can we better exploit the spatial information?

Techniques from image processing

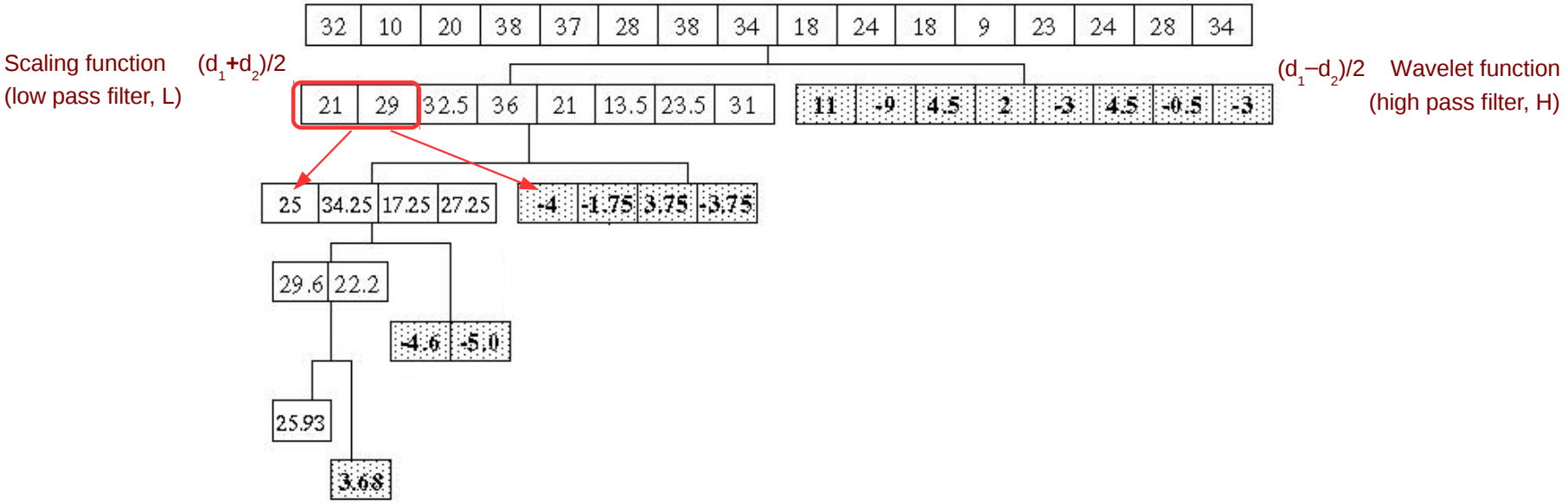
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



Can we better exploit the spatial information?

Techniques from image processing

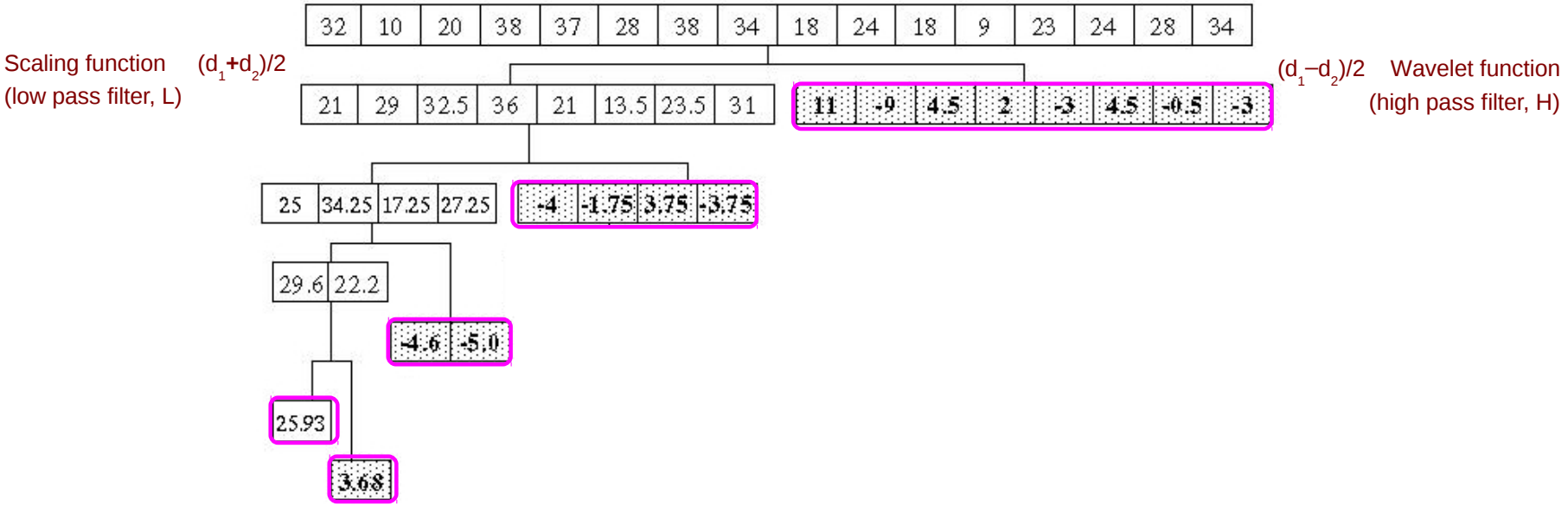
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



Can we better exploit the spatial information?

Techniques from image processing

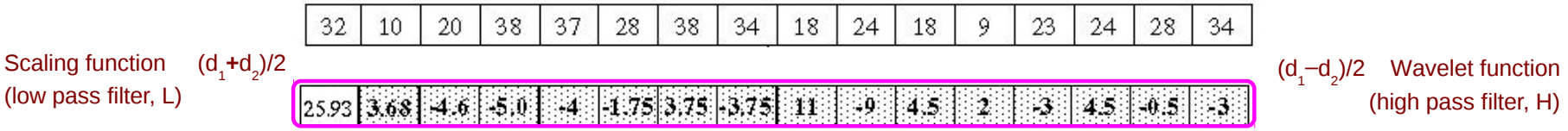
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



Can we better exploit the spatial information?

Techniques from image processing

- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



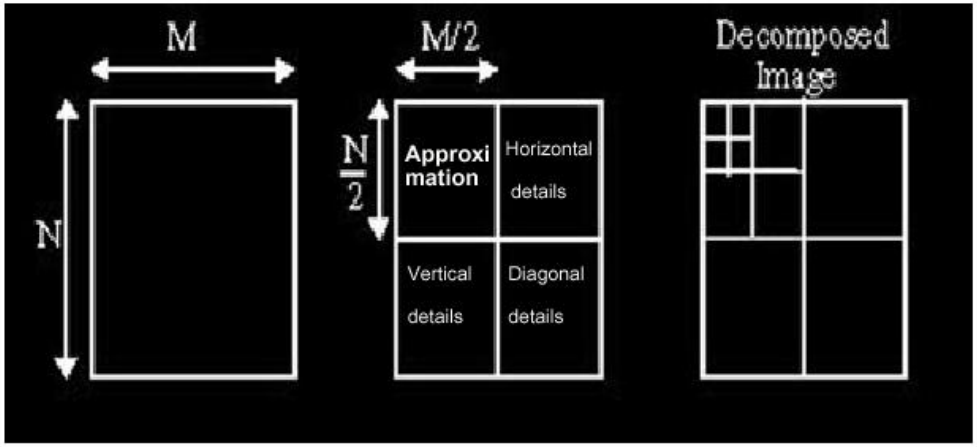
Can we better exploit the spatial information?

Techniques from image processing

- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales

	32	10	20	38	37	28	38	34	18	24	18	9	23	24	28	34
Scaling function (low pass filter, L) $(d_1+d_2)/2$	25.93	3.68	-4.6	-5.0	-4	-1.75	3.75	-3.75	11	-9	4.5	2	-3	4.5	-0.5	-3

$(d_1-d_2)/2$ Wavelet function (high pass filter, H)



Can we better exploit the spatial information?

Techniques from image processing

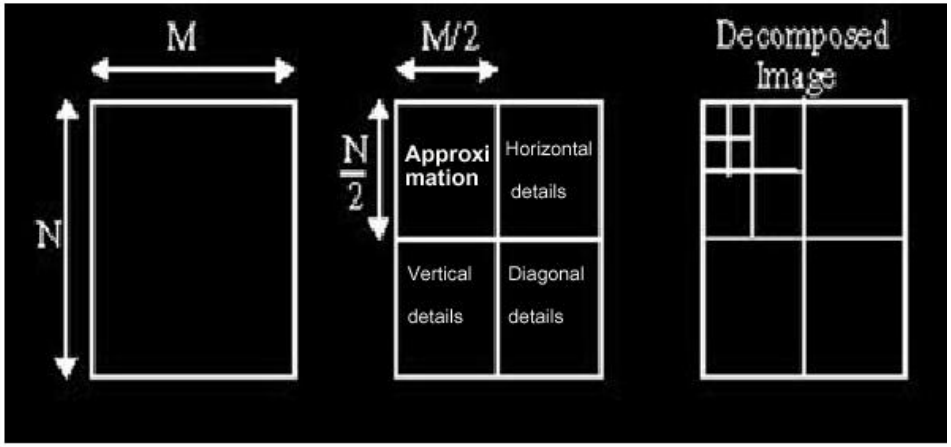
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales

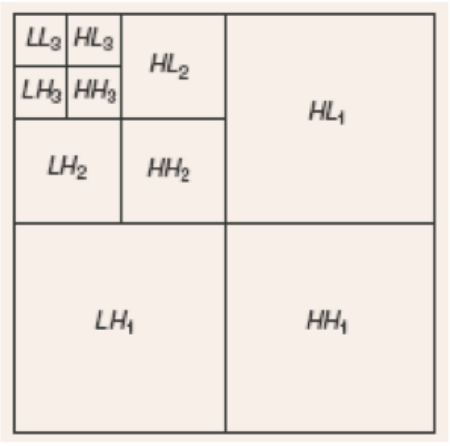
32	10	20	38	37	28	38	34	18	24	18	9	23	24	28	34

Scaling function (low pass filter, L) $(d_1+d_2)/2$

25.93	3.68	-4.6	-5.0	-4	-1.75	3.75	-3.75	11	-9	4.5	2	-3	4.5	-0.5	-3

$(d_1-d_2)/2$ Wavelet function (high pass filter, H)





Can we better exploit the spatial information?

Techniques from image processing

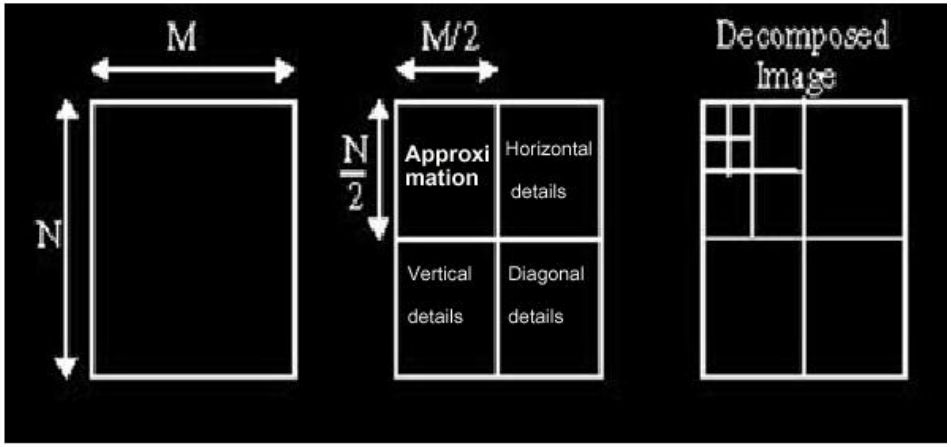
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales

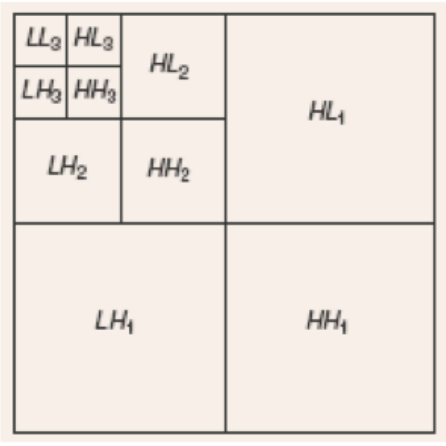
32	10	20	38	37	28	38	34	18	24	18	9	23	24	28	34

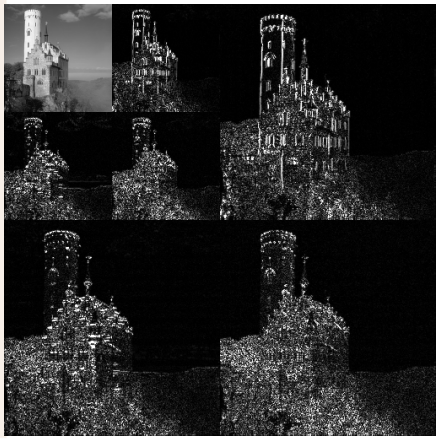
Scaling function (low pass filter, L) $(d_1+d_2)/2$

25.93	3.68	-4.6	-5.0	-4	-1.75	3.75	-3.75	11	-9	4.5	2	-3	4.5	-0.5	-3

$(d_1-d_2)/2$ Wavelet function (high pass filter, H)







Can we better exploit the spatial information?

Techniques from image processing

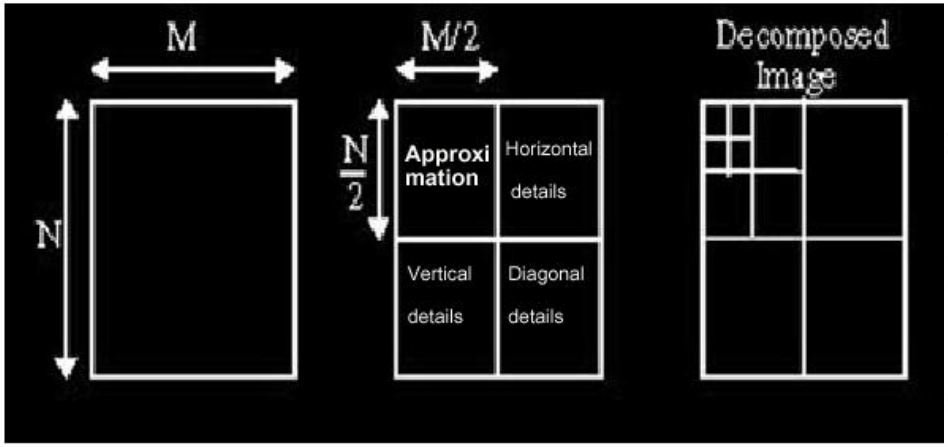
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales

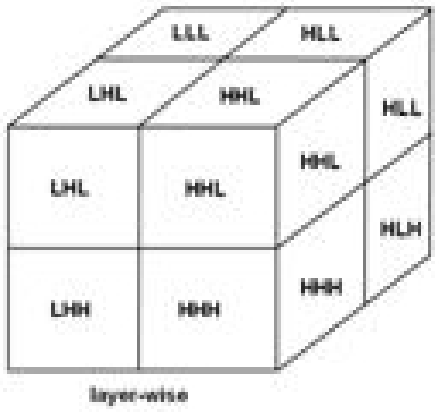
32	10	20	38	37	28	38	34	18	24	18	9	23	24	28	34

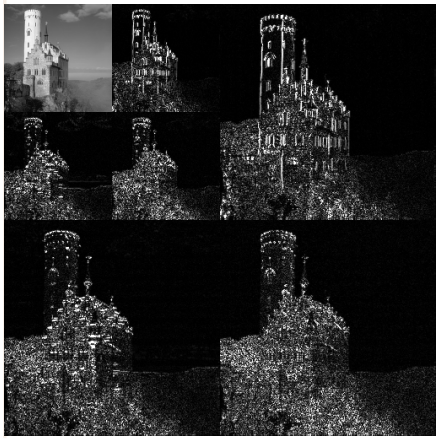
Scaling function (low pass filter, L) $(d_1+d_2)/2$

25.93	3.68	-4.6	-5.0	-4	-1.75	3.75	-3.75	11	-9	4.5	2	-3	4.5	-0.5	-3

$(d_1-d_2)/2$ Wavelet function (high pass filter, H)



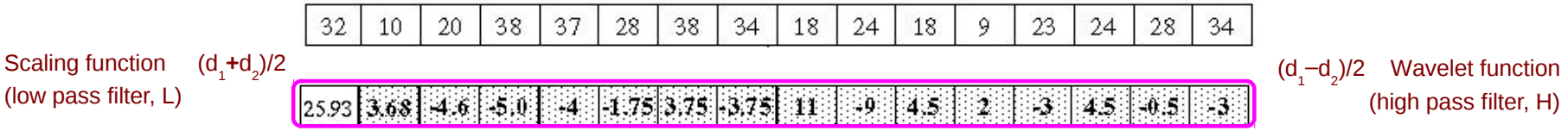




How to compare transformed expr. profiles?

Techniques from image processing

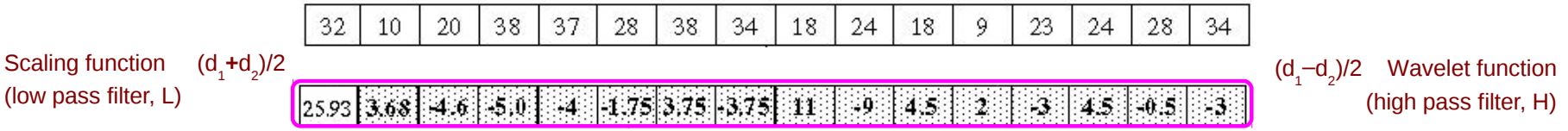
- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



How to compare transformed expr. profiles?

Techniques from image processing

- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



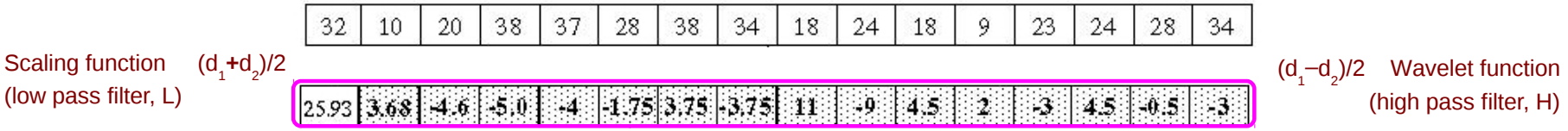
Possibility 1: similarity measures for 1D expression vectors:

- Euklidean distance, Pearson, RIO, …

How to compare transformed expr. profiles?

Techniques from image processing

- Discrete Wavelet Transformation (e.g. Haar WT)
  - applies high and low pass filters on different scales



Possibility 1: similarity measures for 1D expression vectors:

- Euklidean distance, Pearson, RIO, …

Possibility 2: specific approaches for Wavelets:

- Wavelet signature (determine one signature per subband, e.g.LLH2; compare signatures)
- Wavelet coefficient (modified and weighted version of mean squared error)

Preliminary results: leave-one-out

Maß	N	C	g-p P.	Top 1			Top 3			Top 10			Top 10%		
				Tref.	Erw.	p-Wert	Tref.	Erw.	p-Wert	Tref.	Erw.	p-Wert	Tref.	Erw.	p-Wert
a) reduzierte Originalgenexpressionsdaten															
R	50	52	599	23	12	1,49E-03	62	35	6,70E-06	150	116	2,47E-04	117	60	1,08E-12
R	100	111	708	18	7	1,00E-04	44	20	4,39E-07	131	64	2,21E-15	131	71	4,17E-12
R	200	211	714	9	4	7,74E-03	25	11	4,84E-05	73	34	1,04E-09	131	72	8,63E-12
R	400	378	714	7	2	3,24E-03	16	6	2,52E-04	49	19	2,75E-09	124	72	1,11E-09
b) reduzierte transformierte Genexpressionsdaten mit Dekompositionslevel 3															
R	50	52	599	21	12	6,55E-03	58	35	8,66E-05	150	116	2,47E-04	98	60	7,51E-07
R	100	111	708	14	7	5,62E-03	42	20	2,57E-06	114	64	1,04E-09	115	71	1,56E-07
R	200	211	714	10	4	2,48E-03	27	11	6,42E-06	68	34	6,06E-08	116	72	1,47E-07
R	400	378	714	7	2	3,23E-03	15	6	7,45E-04	43	19	8,55E-07	118	72	4,70E-08
c) reduzierte transformierte Genexpressionsdaten mit Dekompositionslevel 6															
R	50	52	599	28	12	1,77E-05	53	35	1,40E-03	145	116	1,39E-03	103	60	3,25E-08
R	100	111	708	23	7	1,96E-07	44	20	4,39E-07	108	64	5,20E-08	117	71	4,80E-08
R	200	211	714	12	4	1,94E-04	31	11	7,35E-08	62	34	4,72E-06	112	72	1,36E-06
R	400	378	714	7	2	3,24E-03	19	6	6,91E-06	45	19	1,38E-07	113	72	8,13E-07

Using RIO: mixed results

Preliminary results: leave-one-out

Maß	N	C	g-p P.	Top 1			Top 3			Top 10			Top 10%		
				Tref.	Erw.	p-Wert	Tref.	Erw.	p-Wert	Tref.	Erw.	p-Wert	Tref.	Erw.	p-Wert
a) reduzierte Originalgenexpressionsdaten															
R	50	52	599	23	12	1,49E-03	62	35	6,70E-06	150	116	2,47E-04	117	60	1,08E-12
R	100	111	708	18	7	1,00E-04	44	20	4,39E-07	131	64	2,21E-15	131	71	4,17E-12
R	200	211	714	9	4	7,74E-03	25	11	4,84E-05	73	34	1,04E-09	131	72	8,63E-12
R	400	378	714	7	2	3,24E-03	16	6	2,52E-04	49	19	2,75E-09	124	72	1,11E-09
b) reduzierte transformierte Genexpressionsdaten mit Dekompositionslevel 3															
R	50	52	599	21	12	6,55E-03	58	35	8,66E-05	150	116	2,47E-04	98	60	7,51E-07
R	100	111	708	14	7	5,62E-03	42	20	2,57E-06	114	64	1,04E-09	115	71	1,56E-07
R	200	211	714	10	4	2,48E-03	27	11	6,42E-06	68	34	6,06E-08	116	72	1,47E-07
R	400	378	714	7	2	3,23E-03	15	6	7,45E-04	43	19	8,55E-07	118	72	4,70E-08
c) reduzierte transformierte Genexpressionsdaten mit Dekompositionslevel 6															
R	50	52	599	28	12	1,77E-05	53	35	1,40E-03	145	116	1,39E-03	103	60	3,25E-08
R	100	111	708	23	7	1,96E-07	44	20	4,39E-07	108	64	5,20E-08	117	71	4,80E-08
R	200	211	714	12	4	1,94E-04	31	11	7,35E-08	62	34	4,72E-06	112	72	1,36E-06
R	400	378	714	7	2	3,24E-03	19	6	6,91E-06	45	19	1,38E-07	113	72	8,13E-07

Using RIO: mixed results

- more disease genes at the top (best or top 3) after wavelet transformation

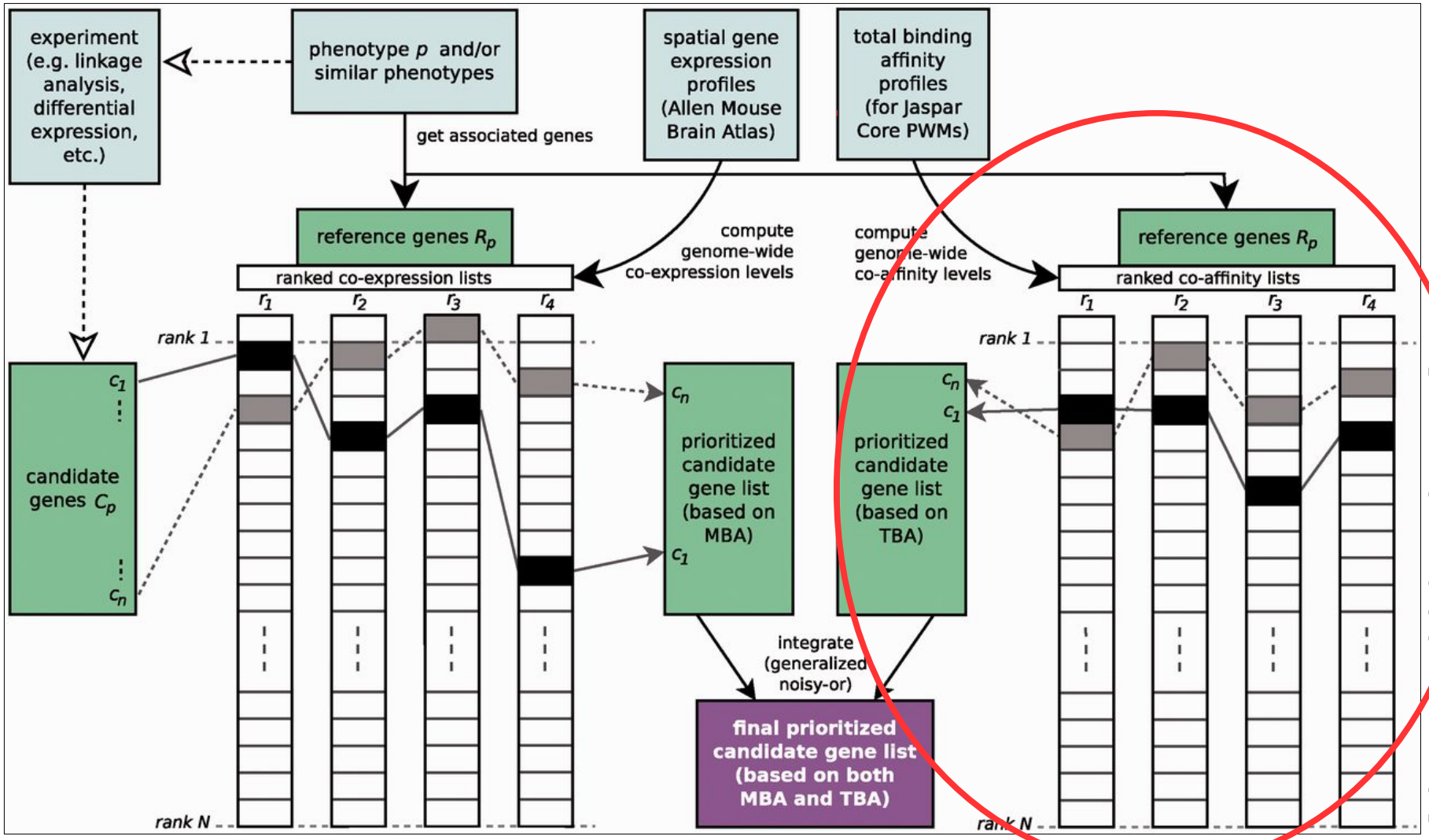
Preliminary results: leave-one-out

Maß	N	C	g-p P.	Top 1 Tref.	Top 1 Erw.	Top 1 p-Wert	Top 3 Tref.	Top 3 Erw.	Top 3 p-Wert	Top 10 Tref.	Top 10 Erw.	Top 10 p-Wert	Top 10% Tref.	Top 10% Erw.	Top 10% p-Wert
a) reduzierte Originalgenexpressionsdaten															
R	50	52	599	23	12	1,49E-03	62	35	6,70E-06	150	116	2,47E-04	117	60	1,08E-12
R	100	111	708	18	7	1,00E-04	44	20	4,39E-07	131	64	2,21E-15	131	71	4,17E-12
R	200	211	714	9	4	7,74E-03	25	11	4,84E-05	73	34	1,04E-09	131	72	8,63E-12
R	400	378	714	7	2	3,24E-03	16	6	2,52E-04	49	19	2,75E-09	124	72	1,11E-09
b) reduzierte transformierte Genexpressionsdaten mit Dekompositionslevel 3															
R	50	52	599	21	12	6,55E-03	58	35	8,66E-05	150	116	2,47E-04	98	60	7,51E-07
R	100	111	708	14	7	5,62E-03	42	20	2,57E-06	114	64	1,04E-09	115	71	1,56E-07
R	200	211	714	10	4	2,48E-03	27	11	6,42E-06	68	34	6,06E-08	116	72	1,47E-07
R	400	378	714	7	2	3,23E-03	15	6	7,45E-04	43	19	8,55E-07	118	72	4,70E-08
c) reduzierte transformierte Genexpressionsdaten mit Dekompositionslevel 6															
R	50	52	599	28	12	1,77E-05	53	35	1,40E-03	145	116	1,39E-03	103	60	3,25E-08
R	100	111	708	23	7	1,96E-07	44	20	4,39E-07	108	64	5,20E-08	117	71	4,80E-08
R	200	211	714	12	4	1,94E-04	31	11	7,35E-08	62	34	4,72E-06	112	72	1,36E-06
R	400	378	714	7	2	3,24E-03	19	6	6,91E-06	45	19	1,38E-07	113	72	8,13E-07

Using RIO: mixed results

- more disease genes at the top (best or top 3) after wavelet transformation
- more disease genes among top 10 or top 10% without wavelet transformation

Method for disease gene prioritization



[Piro et al., Bioinformatics, 2013]

Total Binding Affinity (TBA) – motivation

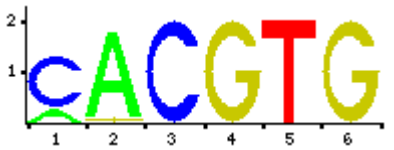
Goal: condition-independent description of regulatory dependencies of promoters to a core set of transcription factors (TFs)

Approach: *in silico* promoter analysis

Total Binding Affinity (TBA) – motivation

Goal: condition-independent description of regulatory dependencies of promoters to a core set of transcription factors (TFs)

Approach: *in silico* promoter analysis

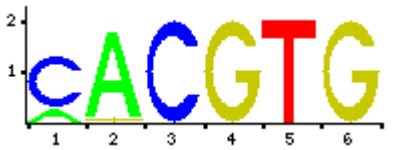


- "Traditional" approach: identification of discrete binding sites for specific TFs within the promoter
  - advantage: allows focused hypotheses and can be experimentally verified

Total Binding Affinity (TBA) – motivation

Goal: condition-independent description of regulatory dependencies of promoters to a core set of transcription factors (TFs)

Approach: *in silico* promoter analysis



- "Traditional" approach: identification of discrete binding sites for specific TFs within the promoter
  - advantage: allows focused hypotheses and can be experimentally verified
  - but: TF binding is a thermodynamic/statistical process → disadvantage: the contribution of weaker but functional binding sites ("low-affinity" binding sites) is often overlooked
- Alternative approach: estimation of the "total binding affinity" (TBA) for specific TFs over the entire promoter region

Total Binding Affinity (TBA) – computation

Total Binding Affinity (TBA) $\boldsymbol{a_{rw}}$ of a promoter ***r*** for a positional frequency matrix (PFM) ***w*** is given by (Foat et al., 2006; Molineris et al., 2011):

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

- ***l*** = length of PWM *w*; ***L*** = length or promoter *r*
- $\boldsymbol{r_i}$ = nucleotide at position *i* of the promoter; $\boldsymbol{r'_i}$ = complement of r_i
- $\boldsymbol{P(w_j,r_i)}$ = probability to have the given r_i at position ***j*** in the PWM *w*
- $\boldsymbol{P(b,r_i)}$ = background probability to observe r_i (= nucleotide frequency in the whole intergenic part of the genome)

Total Binding Affinity (TBA) – computation

Total Binding Affinity (TBA) $\boldsymbol{a_{rw}}$ of a promoter ***r*** for a positional frequency matrix (PFM) ***w*** is given by (Foat et al., 2006; Molineris et al., 2011):

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

- ***l*** = length of PWM *w*; ***L*** = length or promoter *r*
- $\boldsymbol{r_i}$ = nucleotide at position *i* of the promoter; $\boldsymbol{r'_i}$ = complement of r_i
- $\boldsymbol{P(w_j,r_i)}$ = probability to have the given r_i at position ***j*** in the PWM *w*
- $\boldsymbol{P(b,r_i)}$ = background probability to observe r_i (= nucleotide frequency in the whole intergenic part of the genome)

Total Binding Affinity (TBA) – computation

Total Binding Affinity (TBA) a_{rw} of a promoter r for a positional frequency matrix (PFM) w is given by (Foat et al., 2006; Molineris et al., 2011):

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

- l = length of PWM w; L = length or promoter r
- r_i = nucleotide at position i of the promoter; r'_i = complement of r_i
- $P(w_j, r_i)$ = probability to have the given r_i at position j in the PWM w
- $P(b, r_i)$ = background probability to observe r_i (= nucleotide frequency in the whole intergenic part of the genome)

Total Binding Affinity (TBA) – computation

Total Binding Affinity (TBA) a_{rw} of a promoter r for a positional frequency matrix (PFM) w is given by (Foat et al., 2006; Molineris et al., 2011):

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

- l = length of PWM w; L = length or promoter r
- r_i = nucleotide at position i of the promoter; r'_i = complement of r_i
- $P(w_j, r_i)$ = probability to have the given r_i at position j in the PWM w
- $P(b, r_i)$ = background probability to observe r_i (= nucleotide frequency in the whole intergenic part of the genome)

Total Binding Affinity (TBA) – computation

Total Binding Affinity (TBA) a_{rw} of a promoter r for a positional frequency matrix (PFM) w is given by (Foat et al., 2006; Molineris et al., 2011):

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

- l = length of PWM w; L = length or promoter r
- r_i = nucleotide at position i of the promoter; r'_i = complement of r_i
- $P(w_j, r_i)$ = probability to have the given r_i at position j in the PWM w
- $P(b, r_i)$ = background probability to observe r_i (= nucleotide frequency in the whole intergenic part of the genome)

Total Binding Affinity (TBA) – computation

Total Binding Affinity (TBA) a_{rw} of a promoter r for a positional frequency matrix (PFM) w is given by (Foat et al., 2006; Molineris et al., 2011):

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

- l = length of PWM w; L = length or promoter r
- r_i = nucleotide at position i of the promoter; r'_i = complement of r_i
- $P(w_j, r_i)$ = probability to have the given r_i at position j in the PWM w
- $P(b, r_i)$ = background probability to observe r_i (= nucleotide frequency in the whole intergenic part of the genome)

Total Binding Affinity (TBA) – computation

Total Binding Affinity (TBA) a_{rw} of a promoter r for a positional frequency matrix (PFM) w is given by (Foat et al., 2006; Molineris et al., 2011):

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

- l = length of PWM w; L = length or promoter r
- r_i = nucleotide at position i of the promoter; r'_i = complement of r_i
- $P(w_j, r_i)$ = probability to have the given r_i at position j in the PWM w
- $P(b, r_i)$ = background probability to observe r_i (= nucleotide frequency in the whole intergenic part of the genome)

Total Binding Affinity (TBA) – dataset

Promoters:

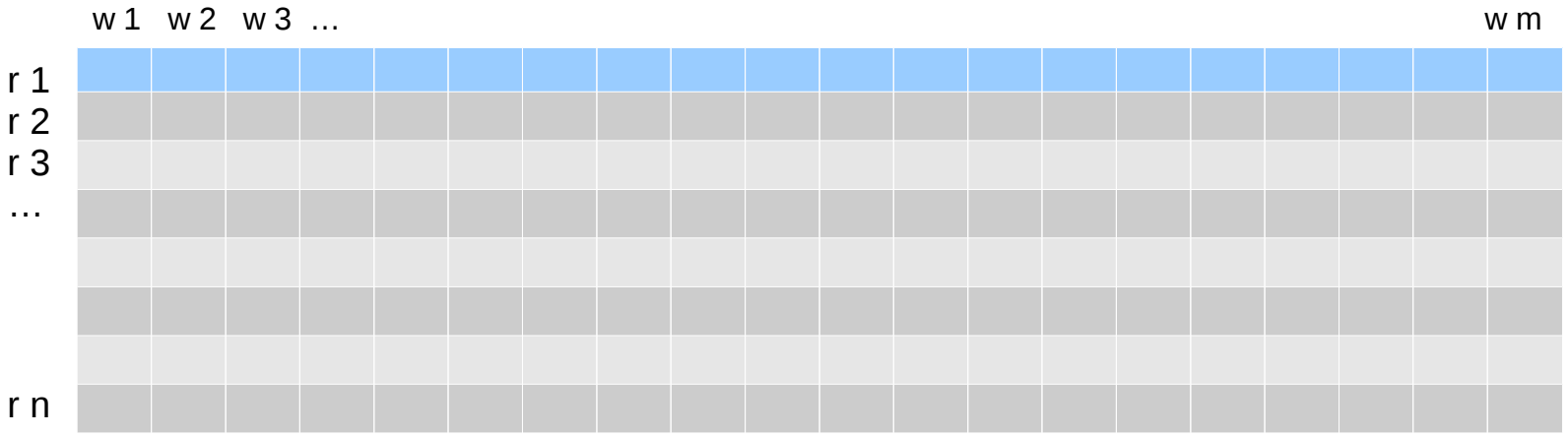
- RefSeq track of the UCSC Genome Browser
- from 1500 bp upstream to 500 bp downstream of the TSS

PFMs / TFs:

- JASPAR Core (130 matrices)

TBAs are computed for all promoters *r* and PFMs *w*

- Result: a binding affinity profile describing regulatory dependencies for each promoter



[Piro et al., Bioinformatics, 2013]

Total Binding Affinity (TBA) – dataset

Promoters:

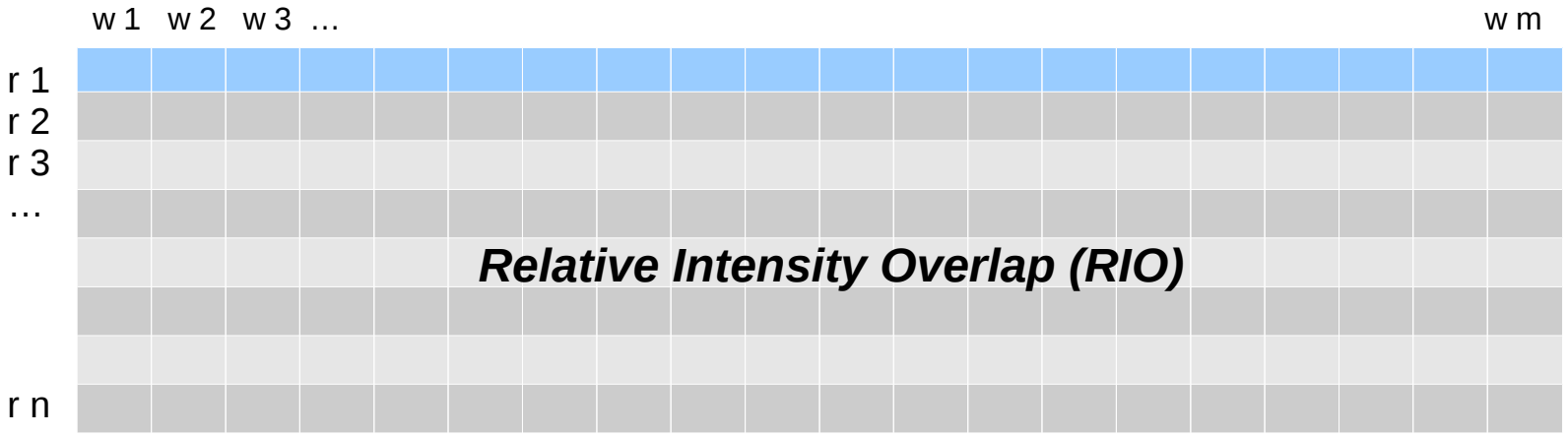
- RefSeq track of the UCSC Genome Browser
- from 1500 bp upstream to 500 bp downstream of the TSS

PFMs / TFs:

- JASPAR Core (130 matrices)

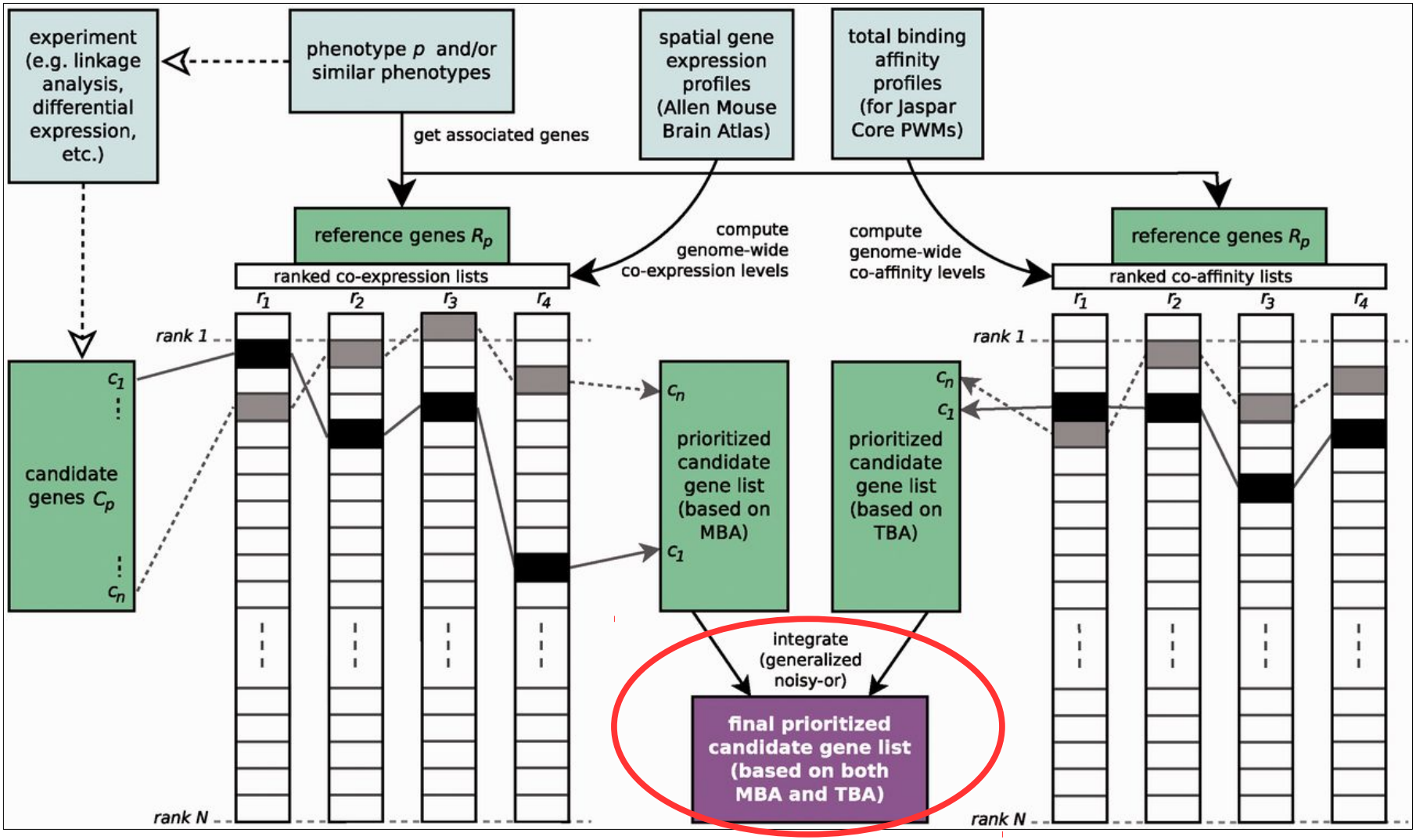
TBAs are computed for all promoters r and PFMs w

- Result: a binding affinity profile describing regulatory dependencies for each promoter



[Piro et al., Bioinformatics, 2013]

Method for disease gene prioritization



[Piro et al., Bioinformatics, 2013]

Integration of 3D expr. and TBA ranking

Integration: (adapted) generalized noisy-OR gate *[Diéz, 1993]*

$$P(\mathcal{R}_c^* = 1 | \mathcal{R}_c^M, \mathcal{R}_c^T) = P(\mathcal{R}_c^* = 1 | \mathcal{R}_c^M) \cdot P(\mathcal{R}_c^* = 1 | \mathcal{R}_c^T)$$ Eq. 1

$$P(\mathcal{R}_c^* = 1 | \mathcal{R}_c^D) = \frac{\left| \{(g,p) | \mathcal{R}_g^D = \mathcal{R}_c^D \} \right| + 1}{\left| \{(g,p)\} \right|}$$ Eq. 2

with:

Datasets *D*: *M*=Mouse Brain Atlas; *T*=TBA.

R_c^D = predicted rank of candidate *c* obtained from dataset *D*;

R_c^* = true rank of candidate *c*.

Probabilities are determined empirically (Eq. 2) from a leave-one-out procedure over all gene-phenotype pairs *(g,p)* with known disease genes *g*.