Freie Universität Berlin



Computational approaches to uncover causes of cancer and hereditary disorders

Rosario M. Piro

Freie Universität Berlin
Charité Universitätsmedizin Berlin
German Cancer Consortium (DKTK)

Politecnico di Milano, DEIB, June 26th, 2018

Outline

Some Molecular Biology

- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data

- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors

- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)

- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways

- PathWave

Outline

Some Molecular Biology

- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data

- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors

- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)

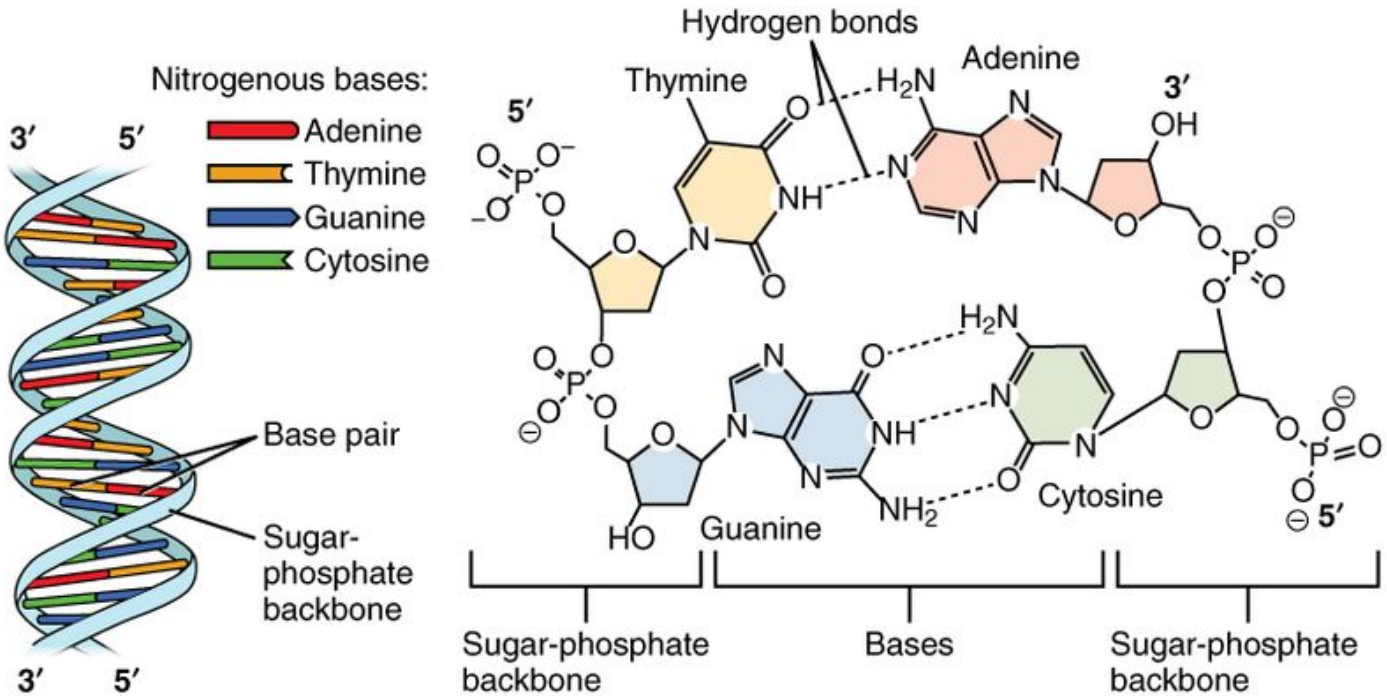
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways

- PathWave

Everything starts from nucleotides

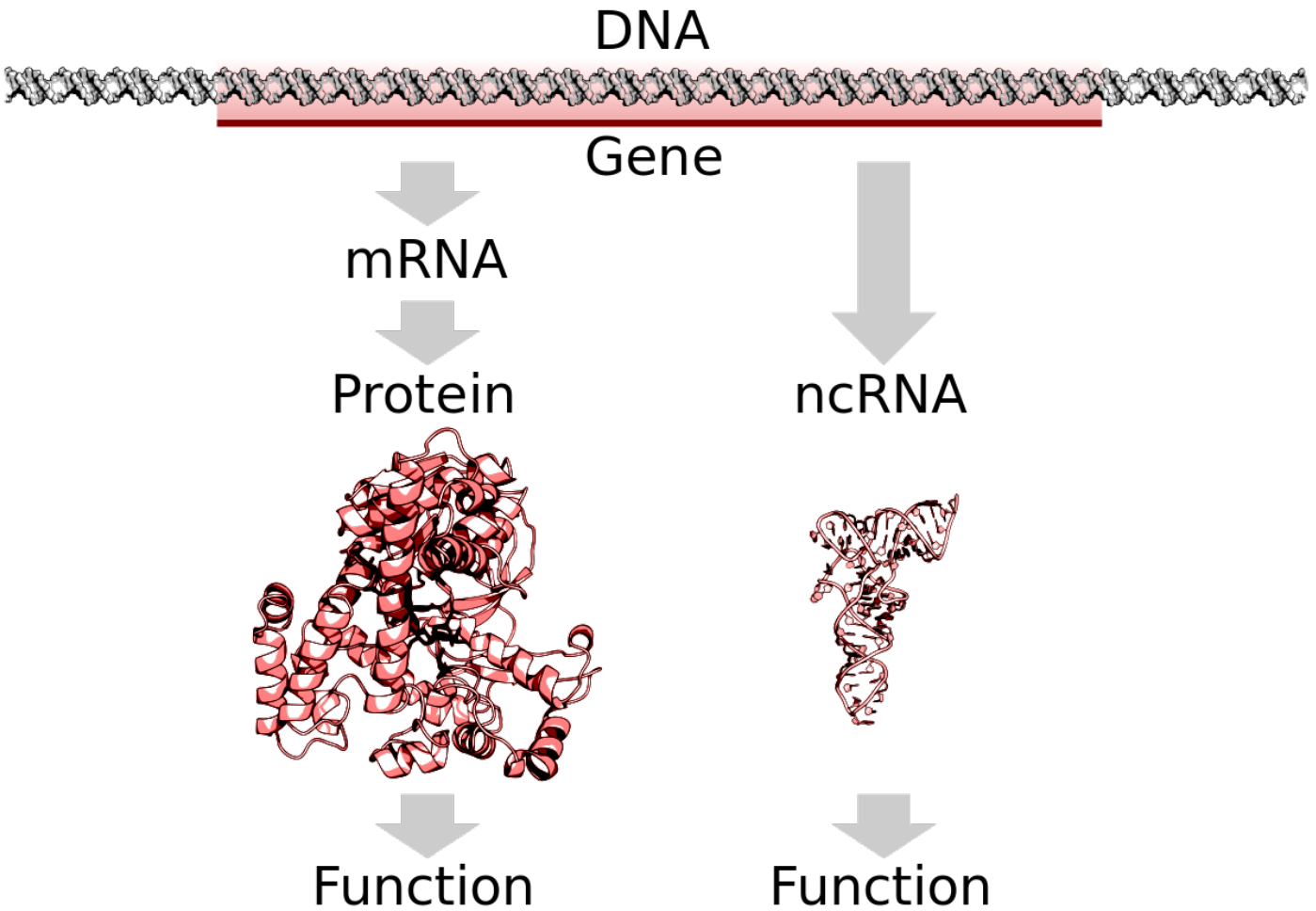
- DNA is composed of adenine (A), cytosine (C), guanine (G) and thymine (T)



from: Wikipedia

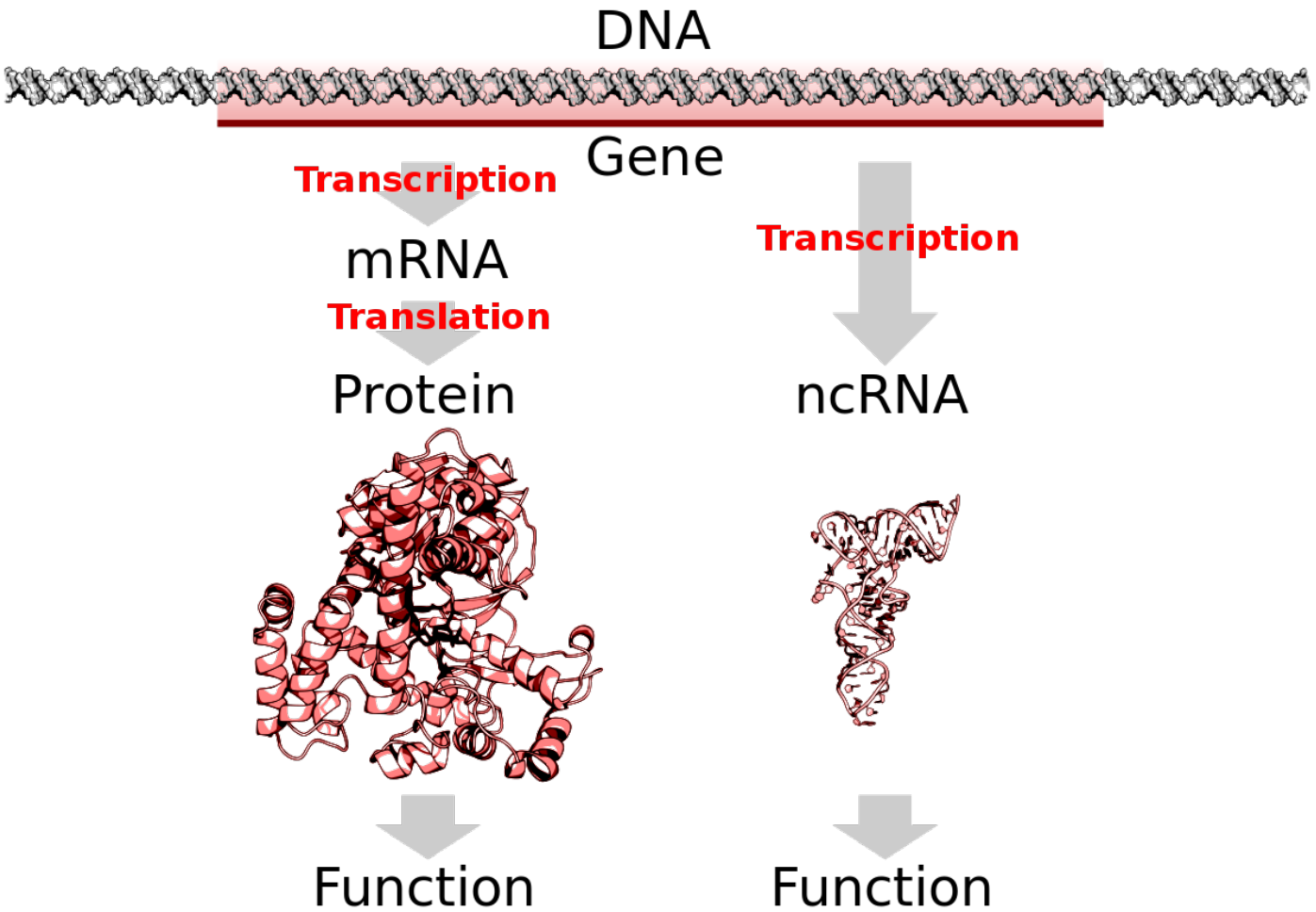
- We often use the term “base” instead of “nucleotide”
- In RNA molecules thymine (T) is substituted by uracil (U)

From DNA to Protein (or non-coding RNAs)



adopted from: Wikipedia

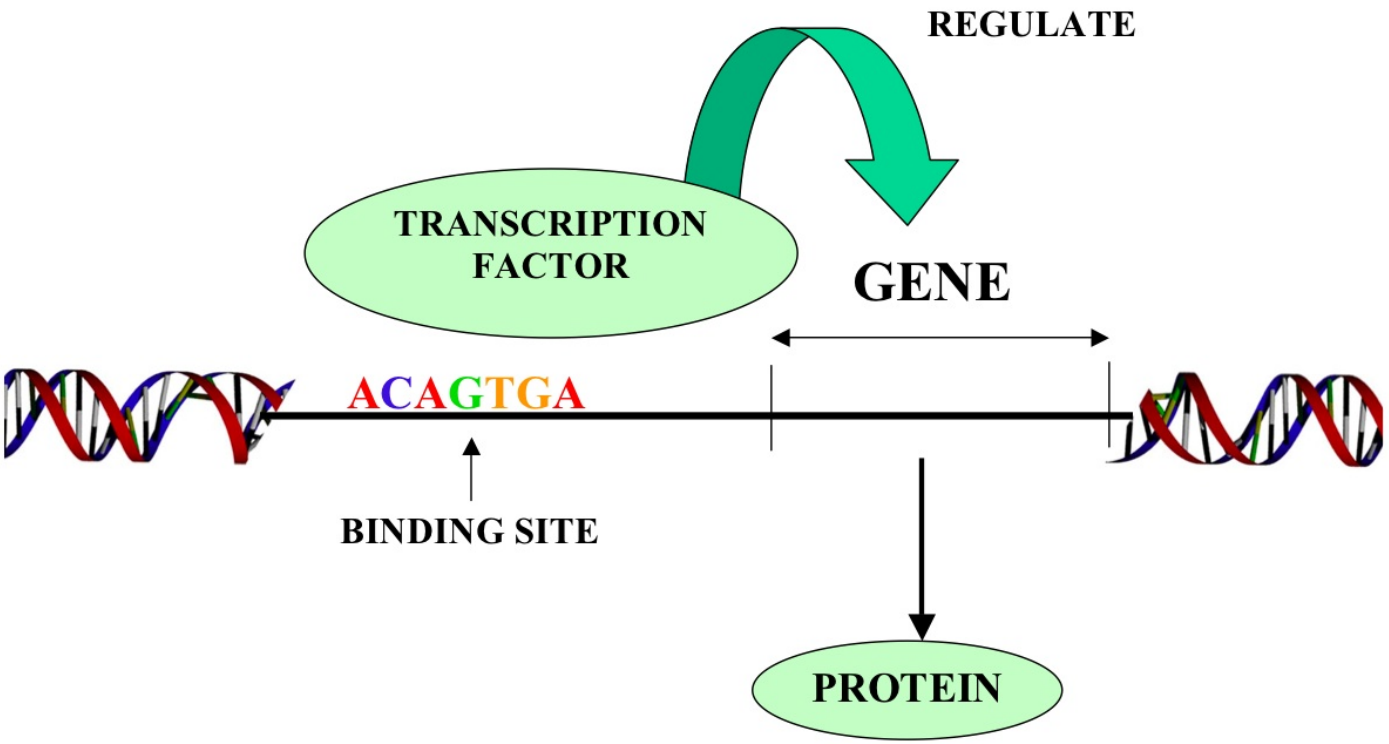
From DNA to Protein (or non-coding RNAs)



adopted from: Wikipedia

Gene regulation

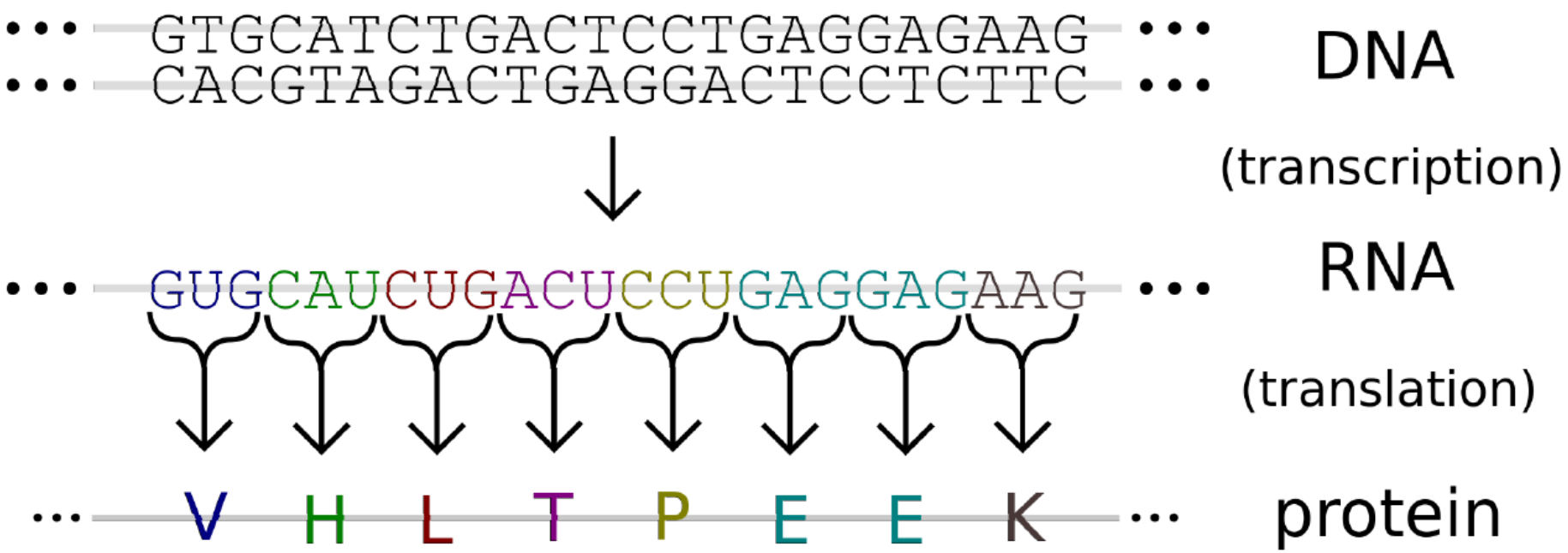
Legend: A transcription factor molecule binds to the DNA at its binding site, and thereby regulates the production of a protein from a gene.

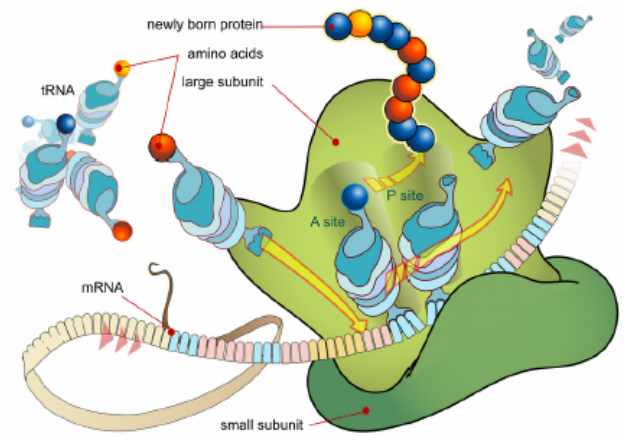


from: Wikipedia

- Also: post-transcriptional regulation, regulation of translation, . . .

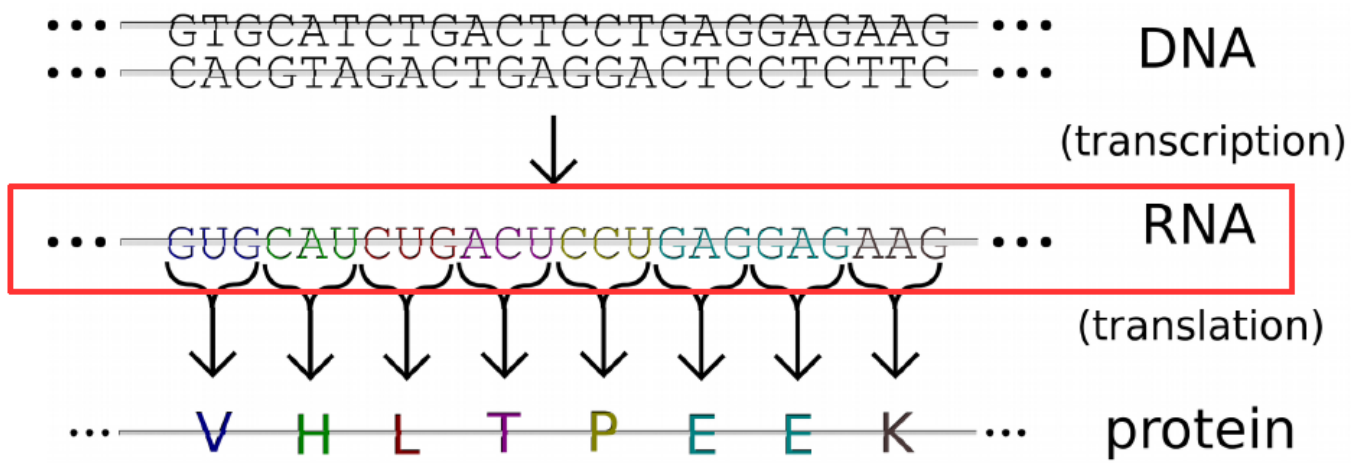
From gene to protein





from: Wikipedia / Wikimedia Commons

Gene expression matrix

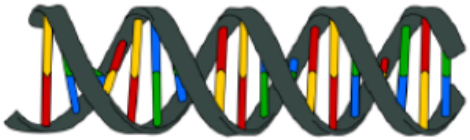


	S1	S2	S3	S4	...									S*n*
G1	1.5	4.7	7.2	1.9	0.4	...								
G2	2.0	6.8	7.1	1.5	8.1	...								
G3	7.8	3.1	0.5	0.8	6.3	...								
G4	...													
...														
G*m*														

m x *n* gene expression matrix

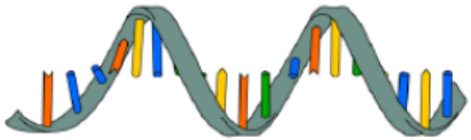
"-omics"

DNA



Genomics-25,000 genes

RNA



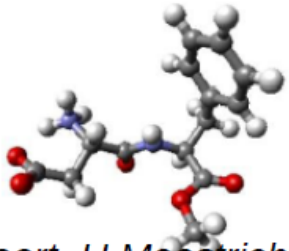
Transcriptomics – 100,000 Transcripts

PROTEIN



Proteomics – 1,000,000 Proteins

BIOCHEMICALS (METABOLITES)



Metabolomics – 1,800 Biochemicals

[slide by *Susan Steinbusch-Coort, U Maastricht*]

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

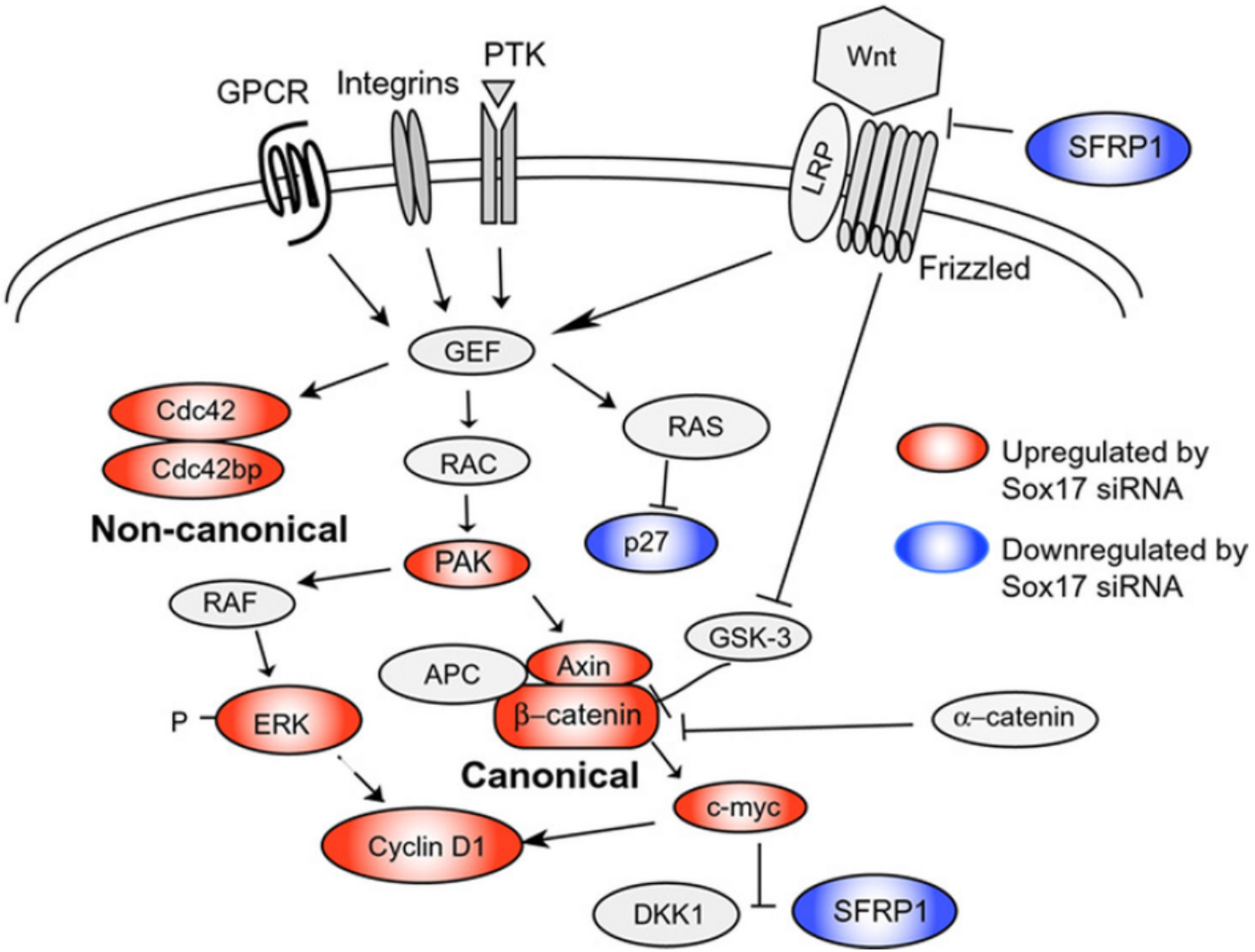
Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways
- PathWave

Example of a signaling pathway



from: Chew et al., *J Neurosci* **31**:13921-13935, 2011

Outline

Some Molecular Biology

- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data

- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors

- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

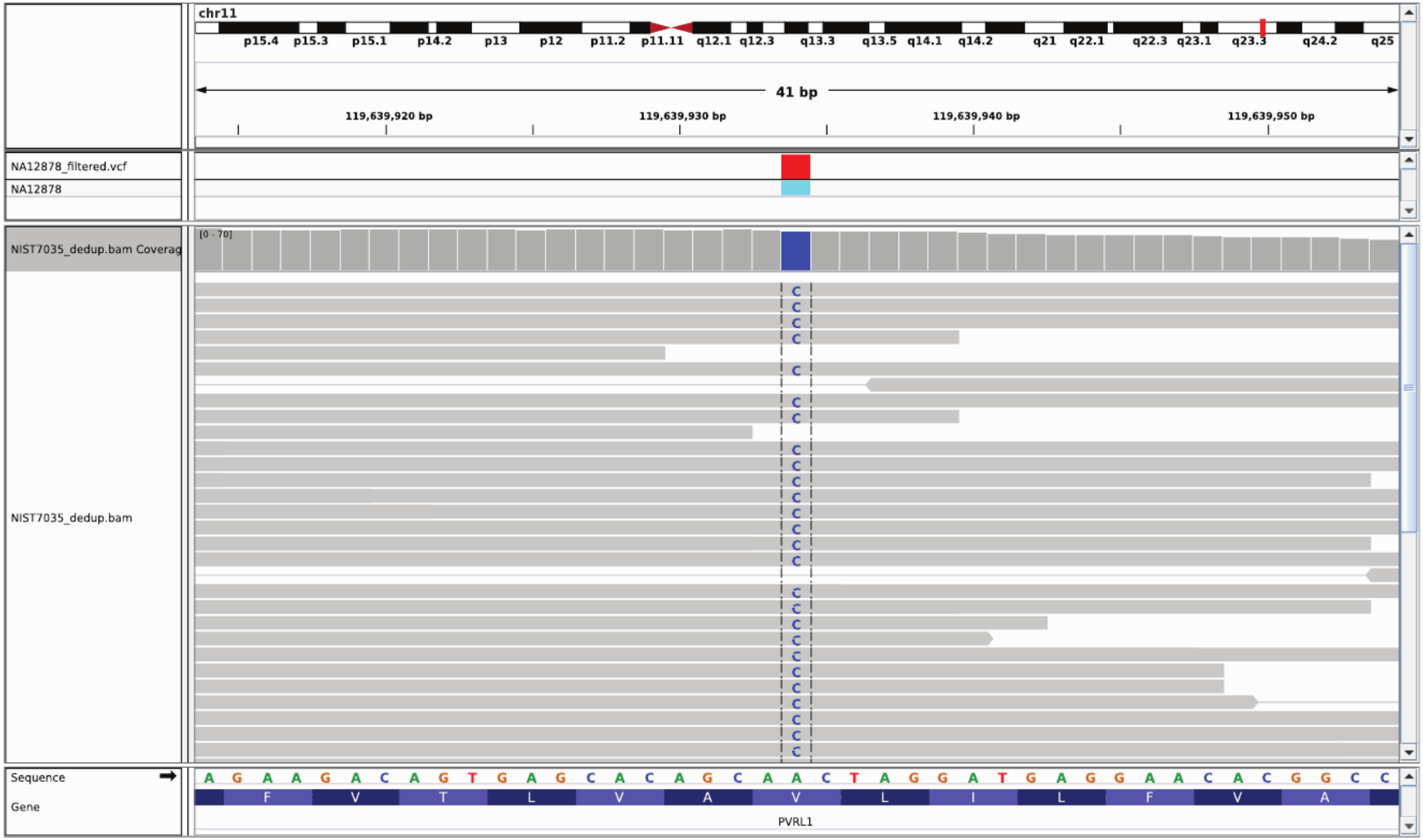
Disease-Gene Prediction (DGP)

- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways

- PathWave

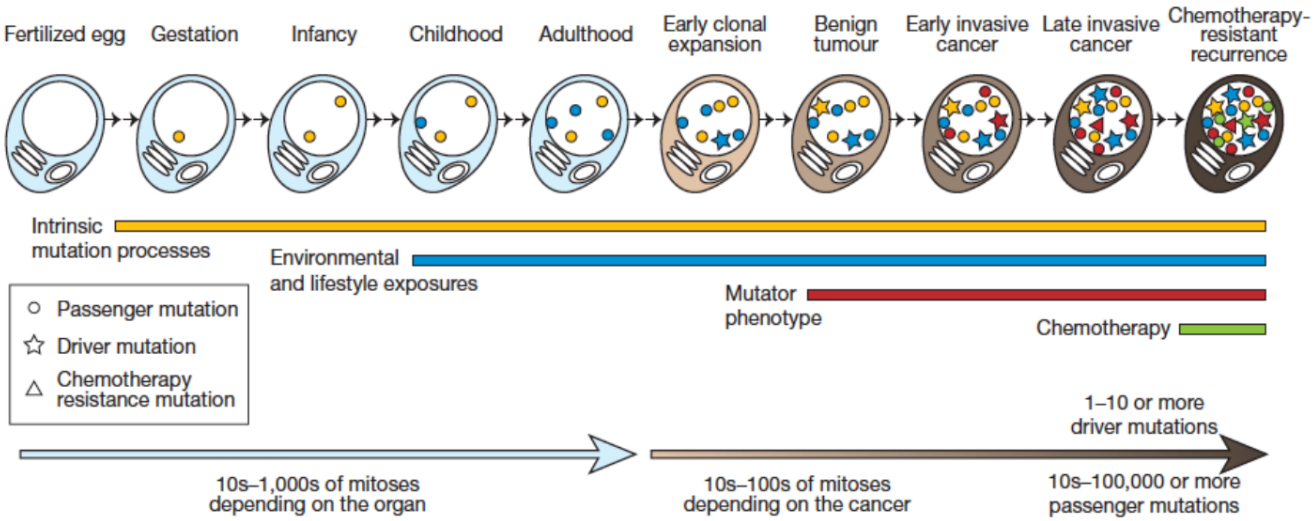
Germline single nucleotide variant (SNV)



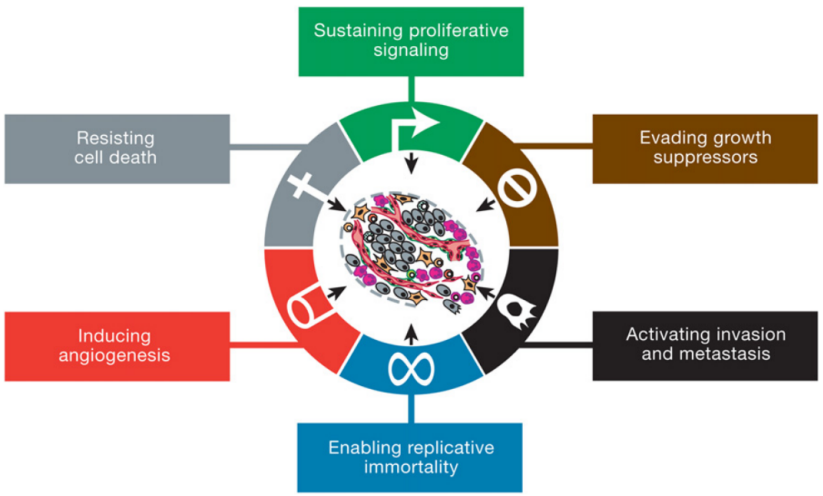
Homozygous germline mutation in the *PVRL1* gene (c.1082T>G, p.Val361Gly).
Note that the *PVRL1* gene lies on the reverse strand of the genome. This IGV figure shows the forward strand (A>C).

from: Robinson, Piro, Jäger. Computational genome and exome analysis. CRC Press, 2017, page 237

Cancer and somatic mutations

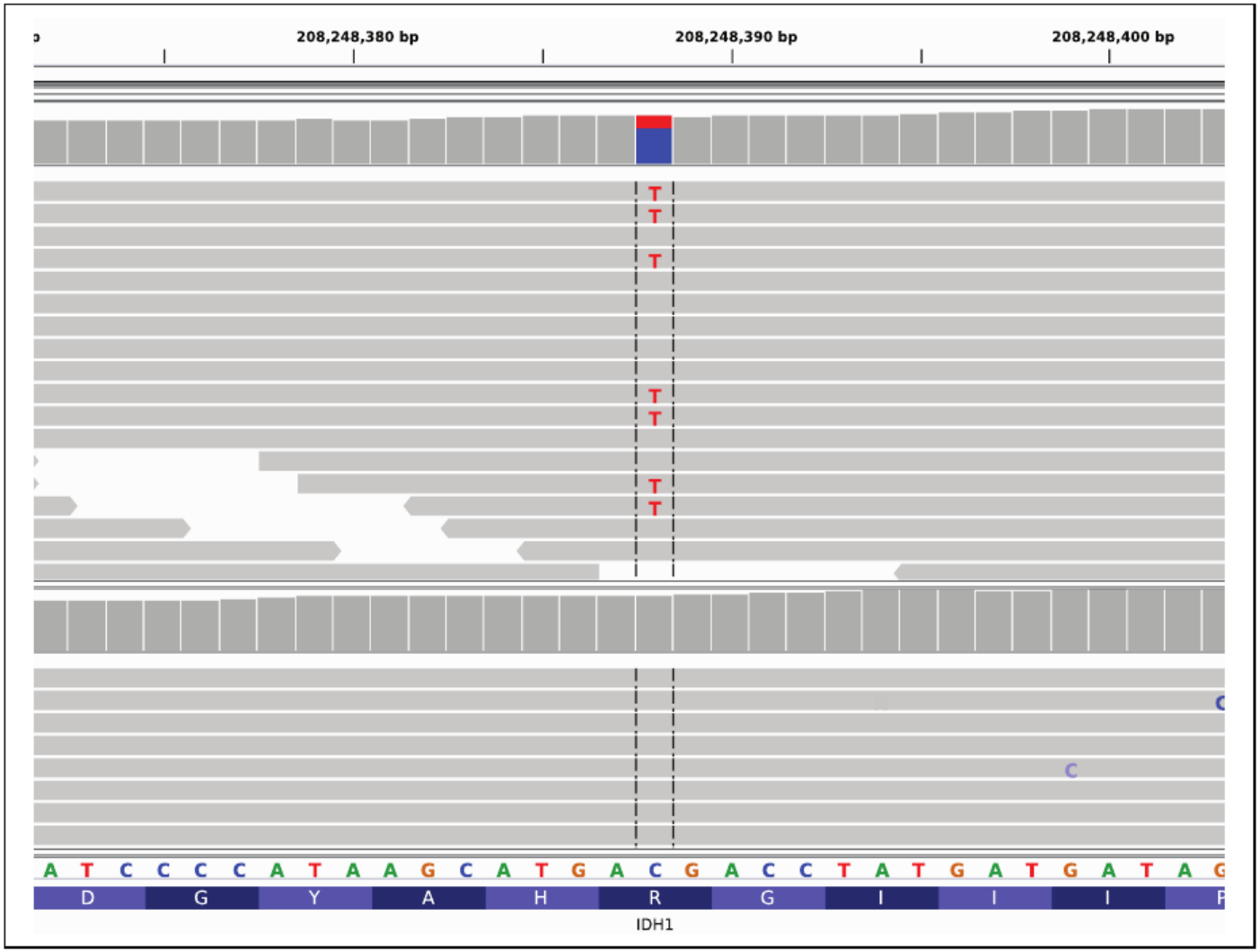


Evolutionary progression of cancer; from: Stratton et al., 2009



“The hallmarks of cancer”; from: Hanahan and Weinberg, 2011

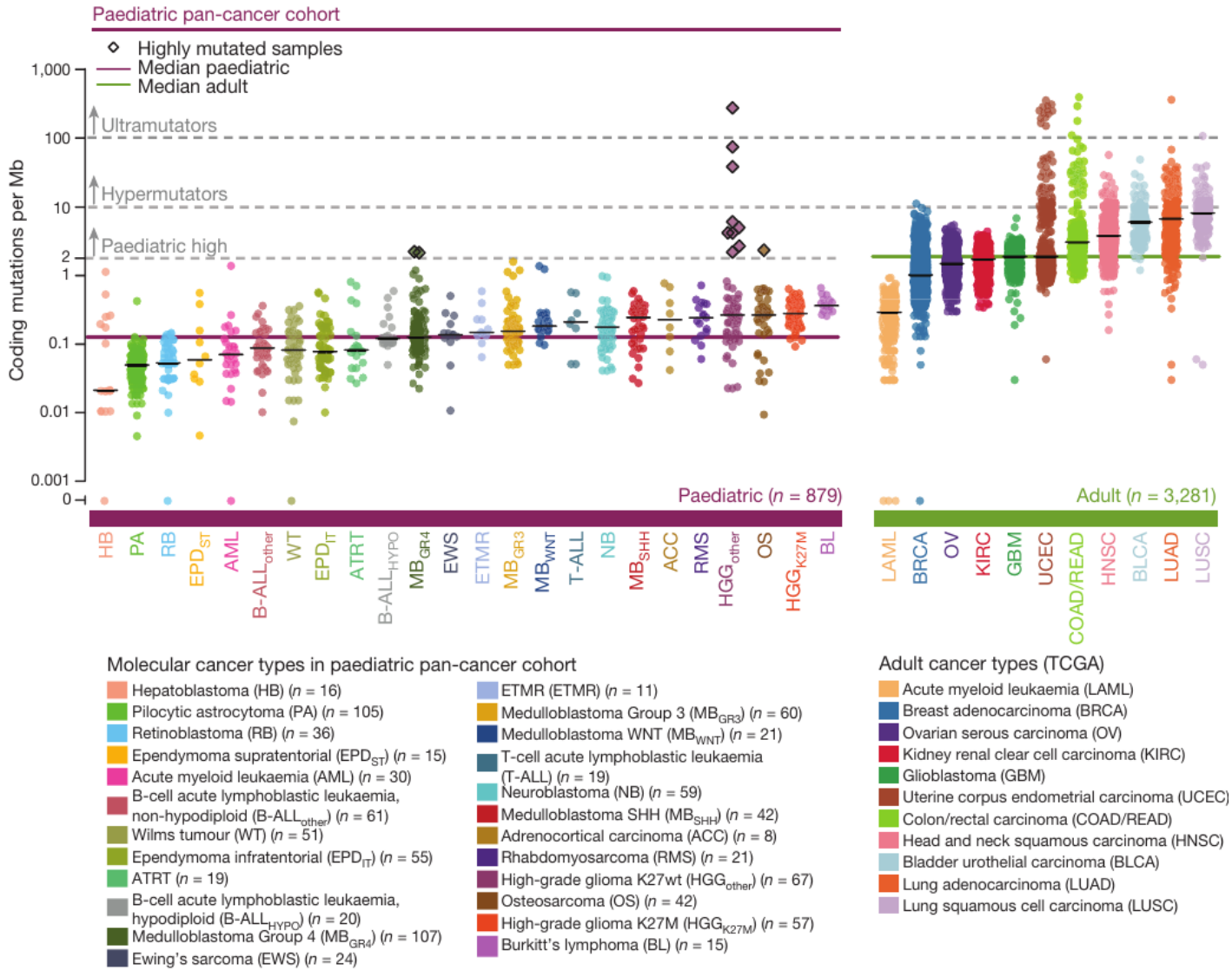
Somatic single nucleotide variant (SNV)



Somatic *IDH1* mutation (c.395G>A, p.R132H) on chromosome 2 in a glioblastoma patient.
Top: aligned sequencing reads from the tumor; bottom: aligned reads from the control sample.

from: Robinson, Piro, Jäger. Computational genome and exome analysis. CRC Press, 2017, page 449

Somatic mutations in protein-coding regions



from: Gröbner et al. The landscape of genomic alterations across childhood cancers. *Nature* **555**:321–327, 2018

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

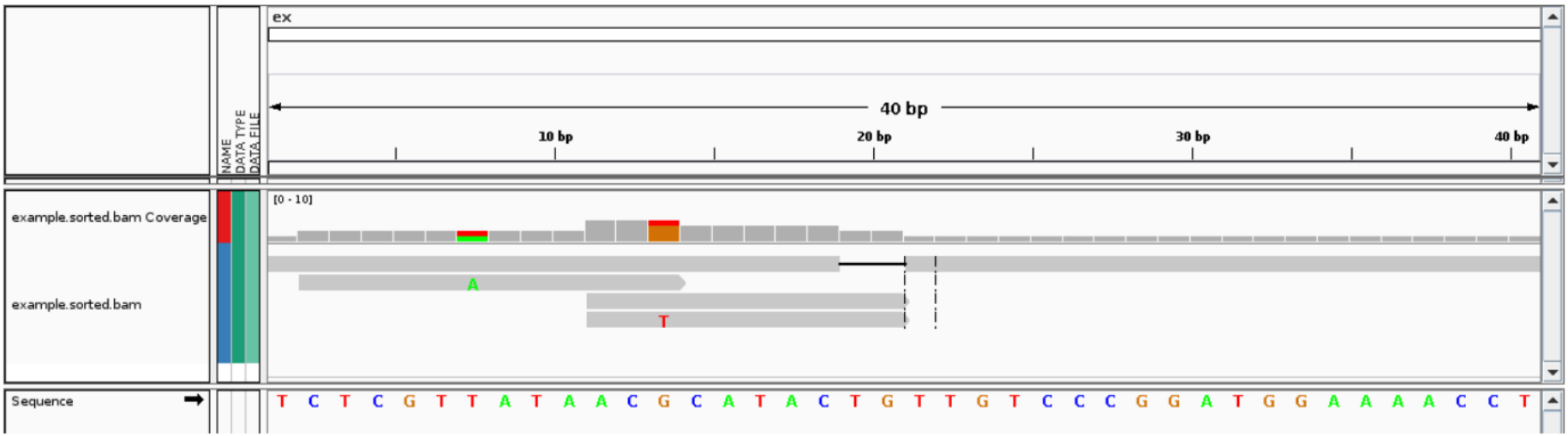
Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways
- PathWave

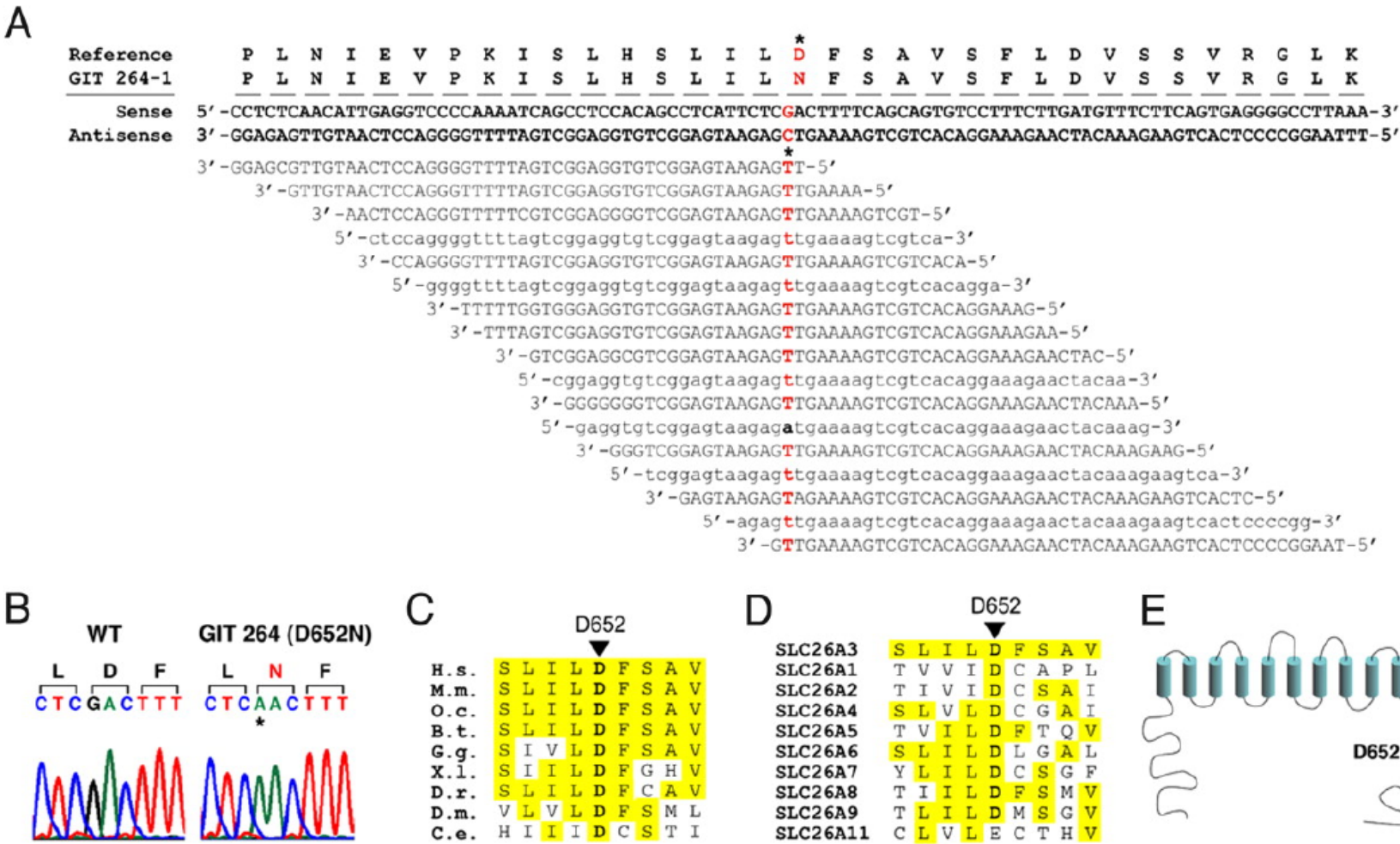
Read Alignment

- Next-generation sequencing (NGS) of DNA (DNA-seq) produces *millions* of short reads
- Reads need to be “aligned” to the reference genome



Variant calling and annotation

- "Variant calling" is the identification of differences with respect to the reference genome; "annotation": how do mutations change genes/proteins?

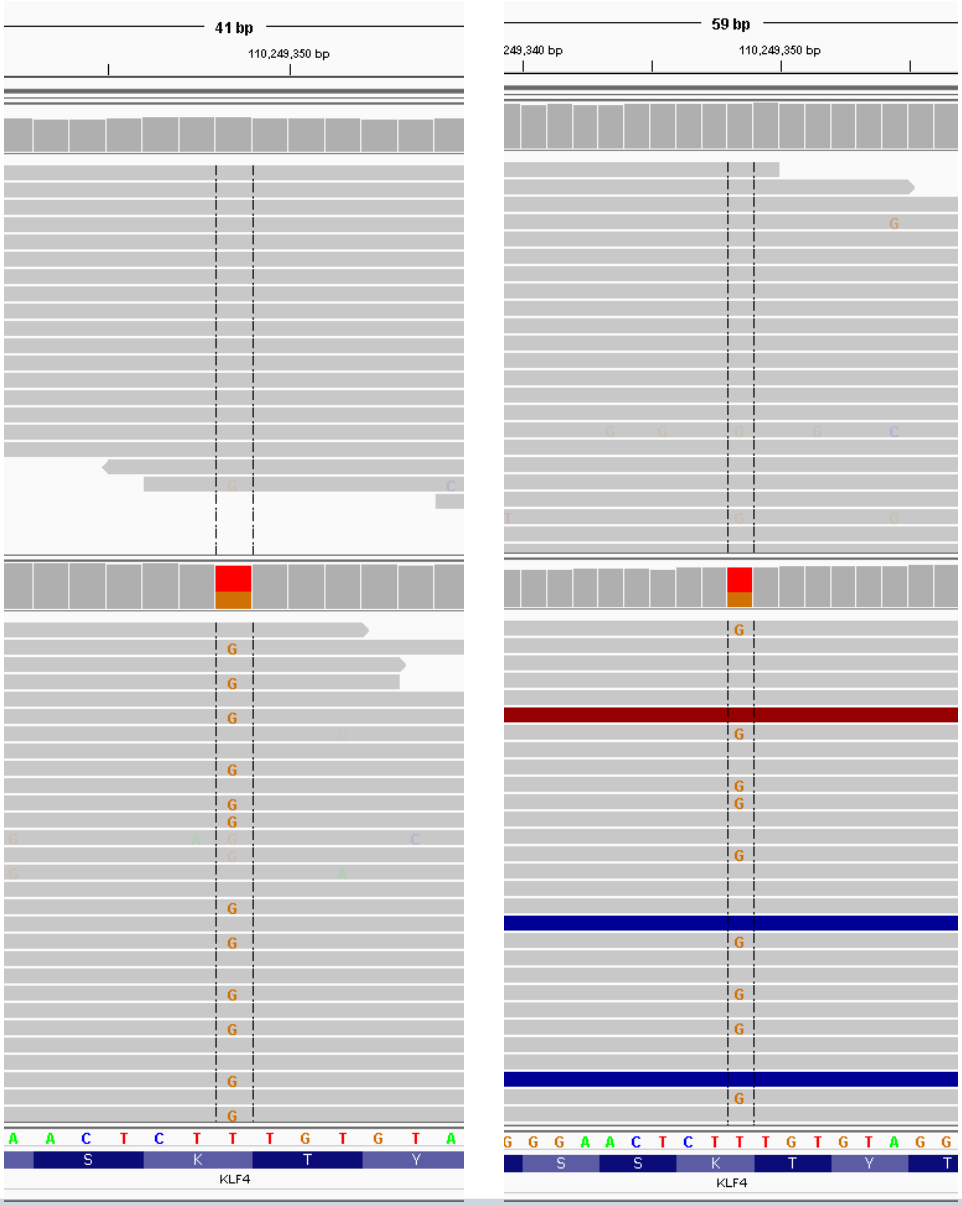


Somatic mutations

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

control

tumor



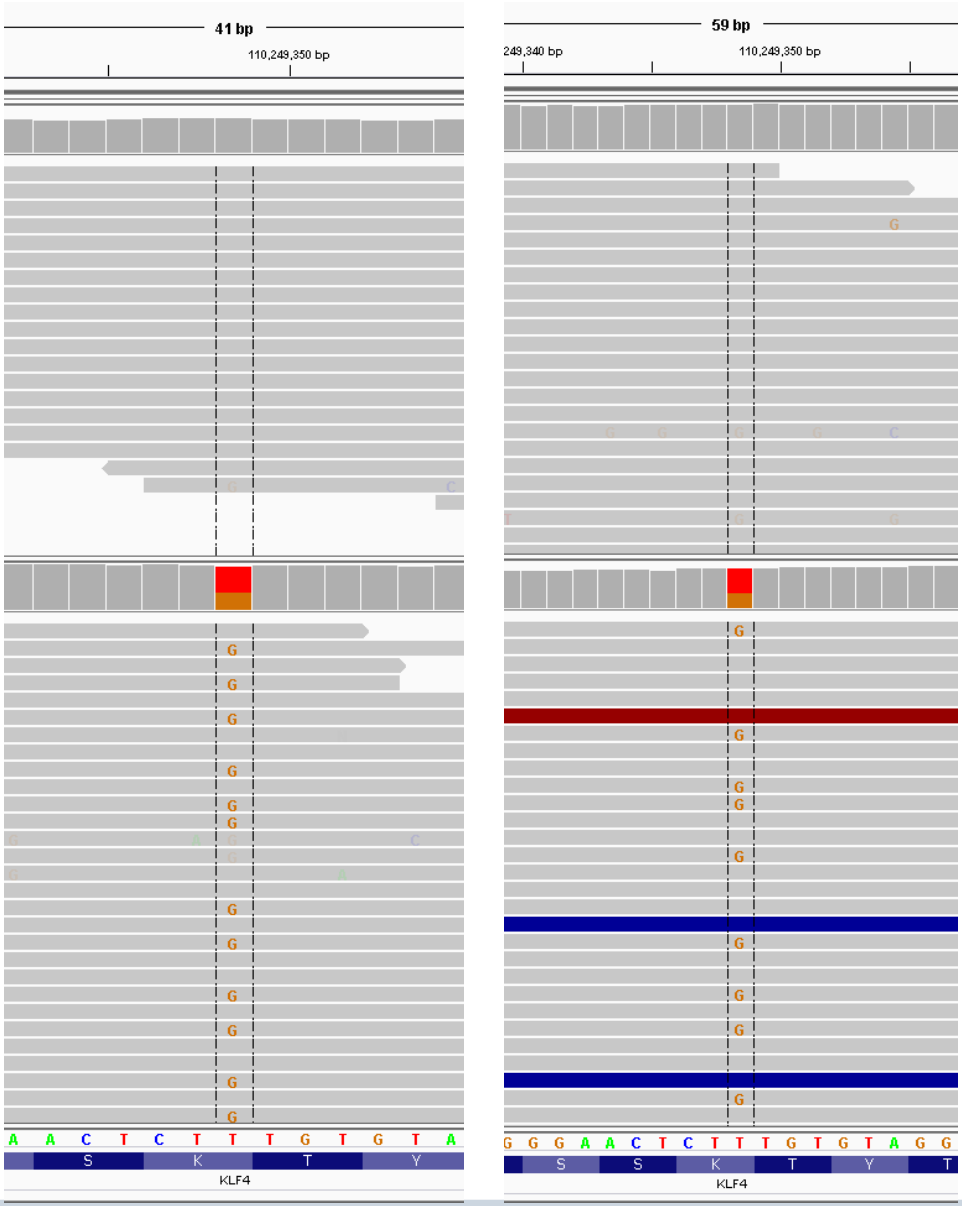
Somatic mutations

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

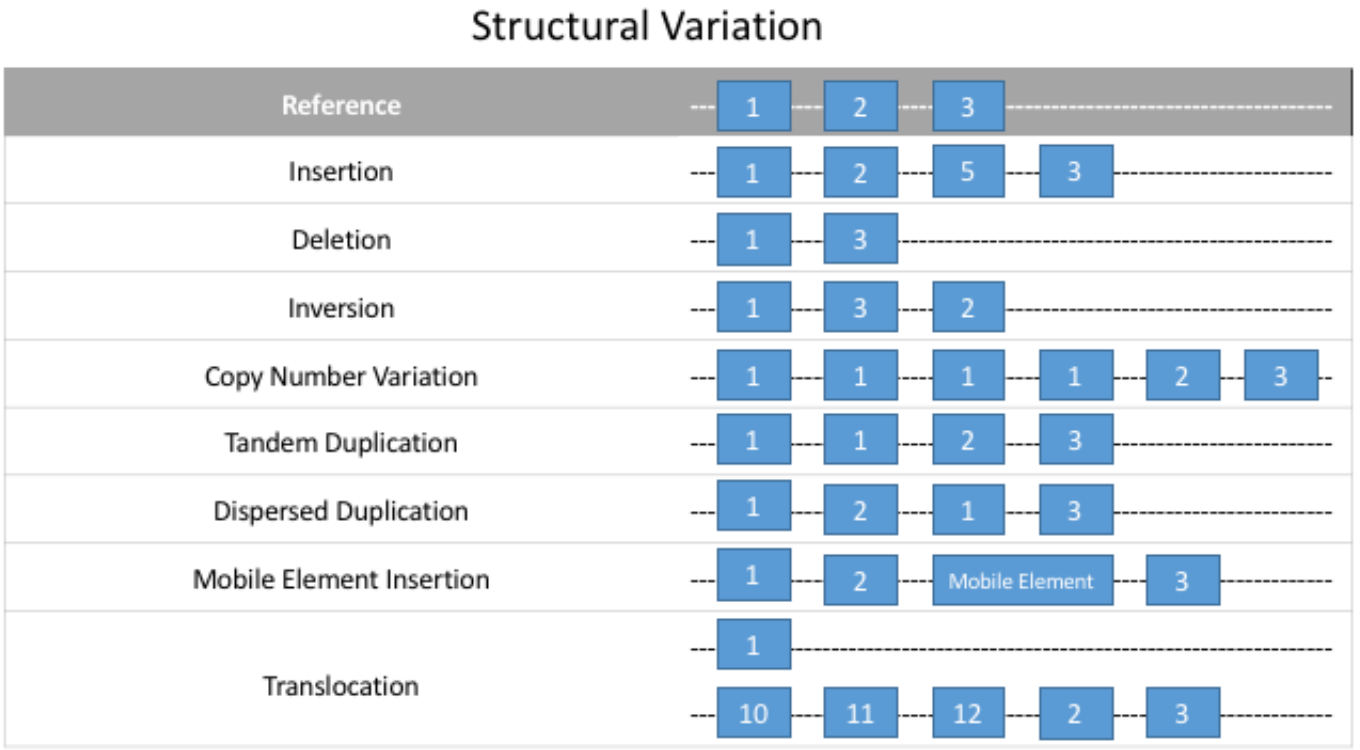


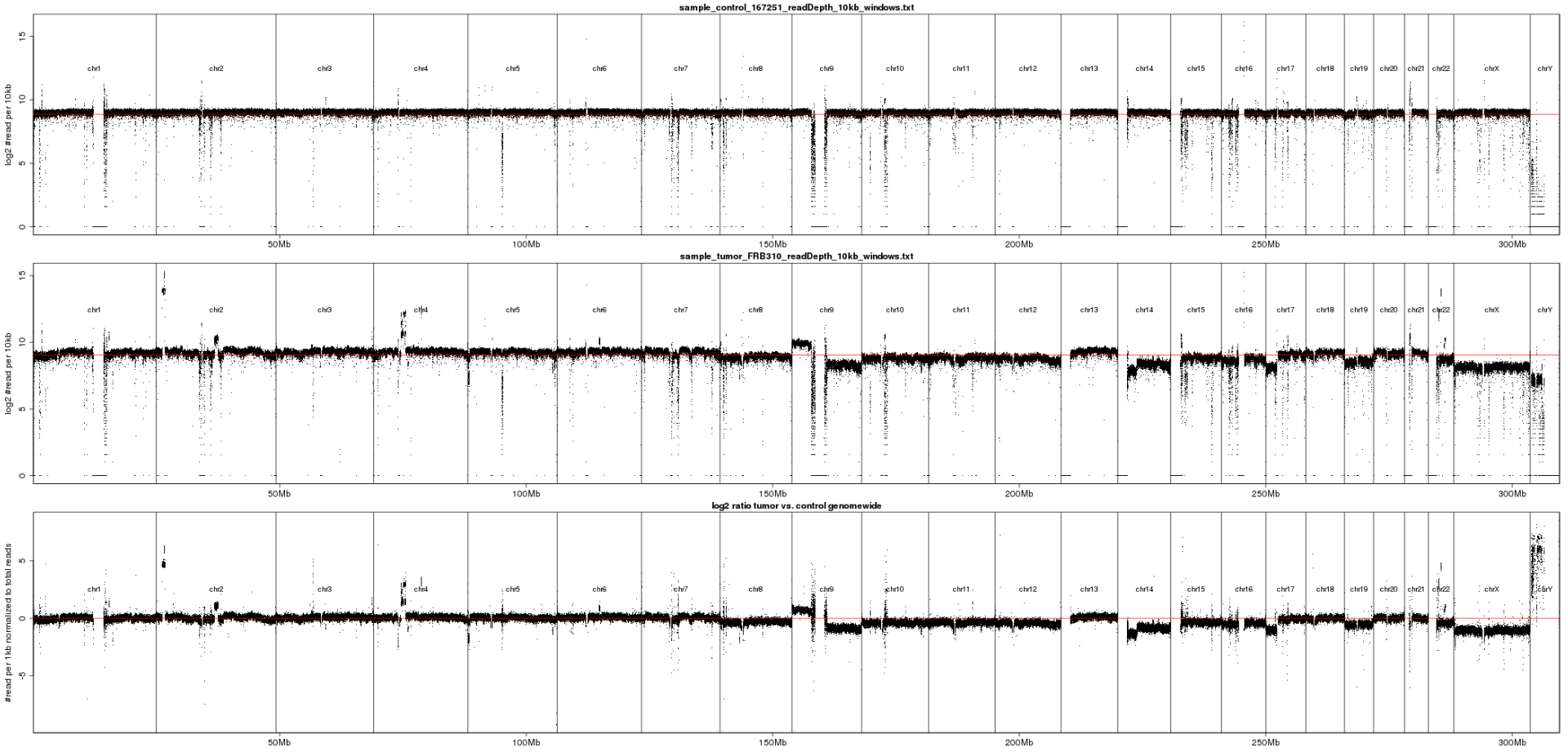
Figure 1: Depiction of different types of structural variants compared to the reference genome. Each different number represents a different gene.

from: Wikipedia / Wikimedia Commons

Copy number variants

Example: identification of CNVs from whole genome sequencing

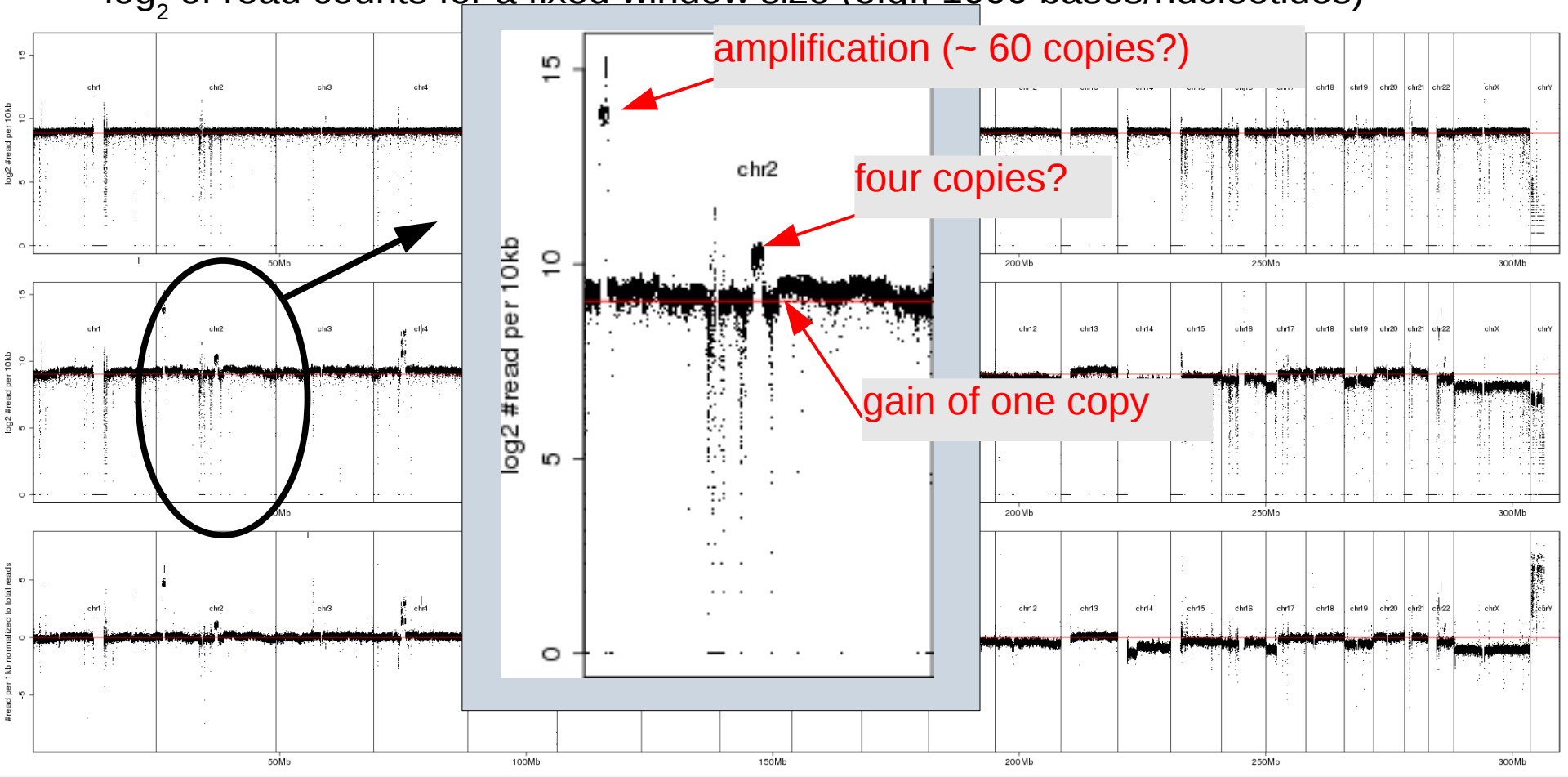
- $\log_2$ of read counts for a fixed window size (e.g., 1000 bases/nucleotides)



Copy number variants

Example: identification of CNVs from whole genome sequencing

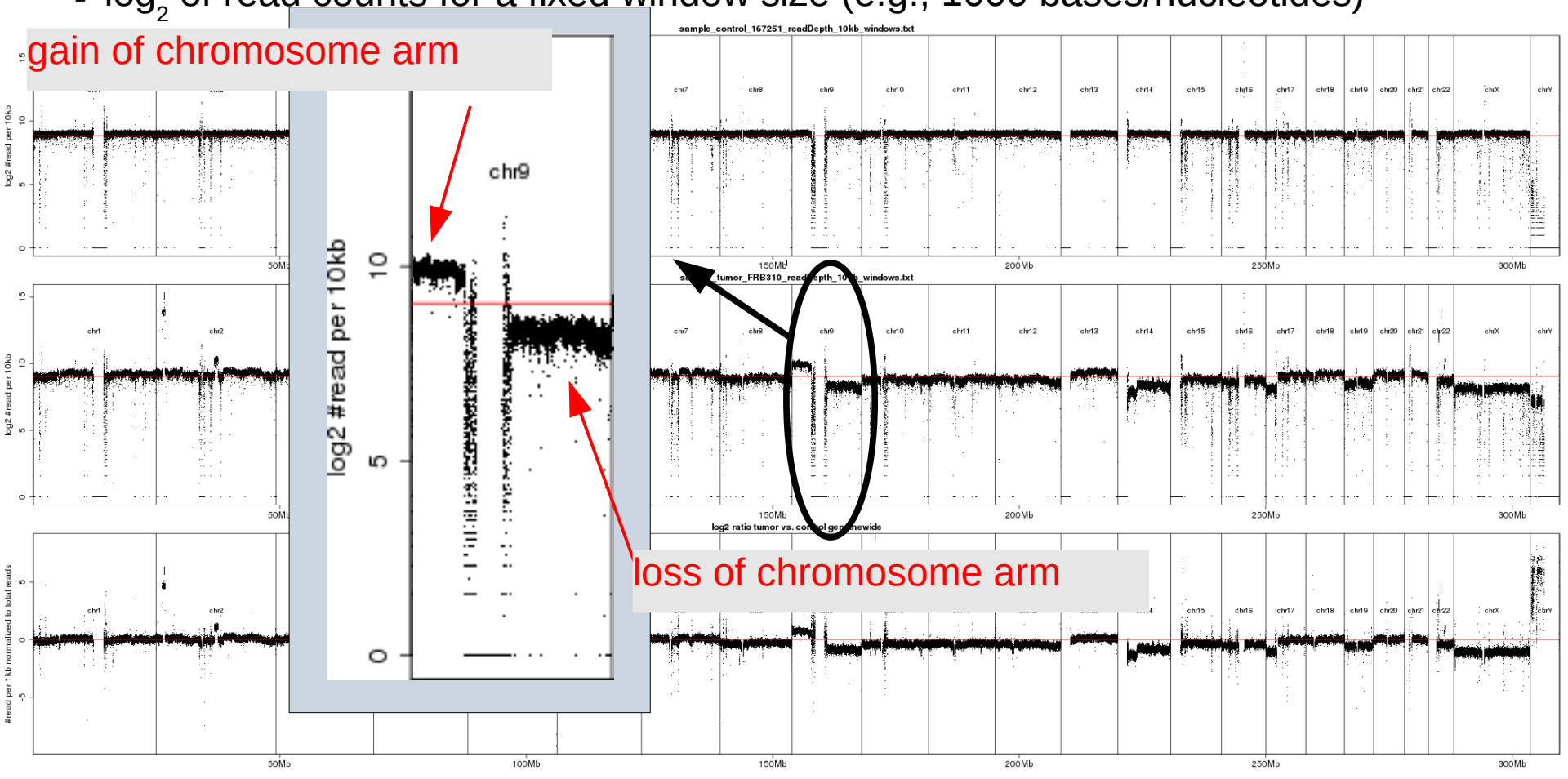
- $\log_2$ of read counts for a fixed window size (e.g., 1000 bases/nucleotides)



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)



Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

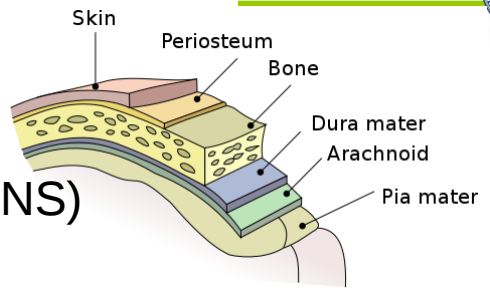
Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

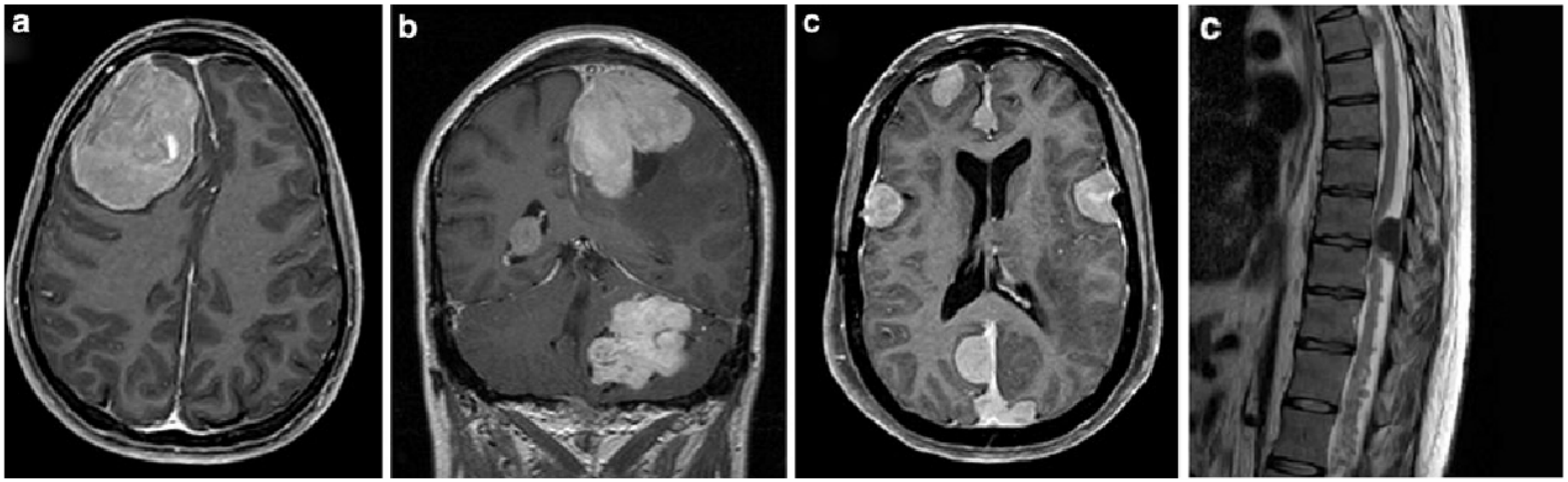
Identifying Regulatory Shifts in Biological Pathways
- PathWave

Meningioma: Introduction



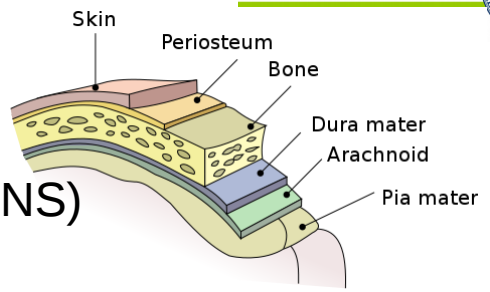
Most common primary brain tumor (~25-30%)

- arise from the meninges (membranes around CNS)



[Goutagny and Kalamarides, J Neurooncol, 2010]

Meningioma: Introduction



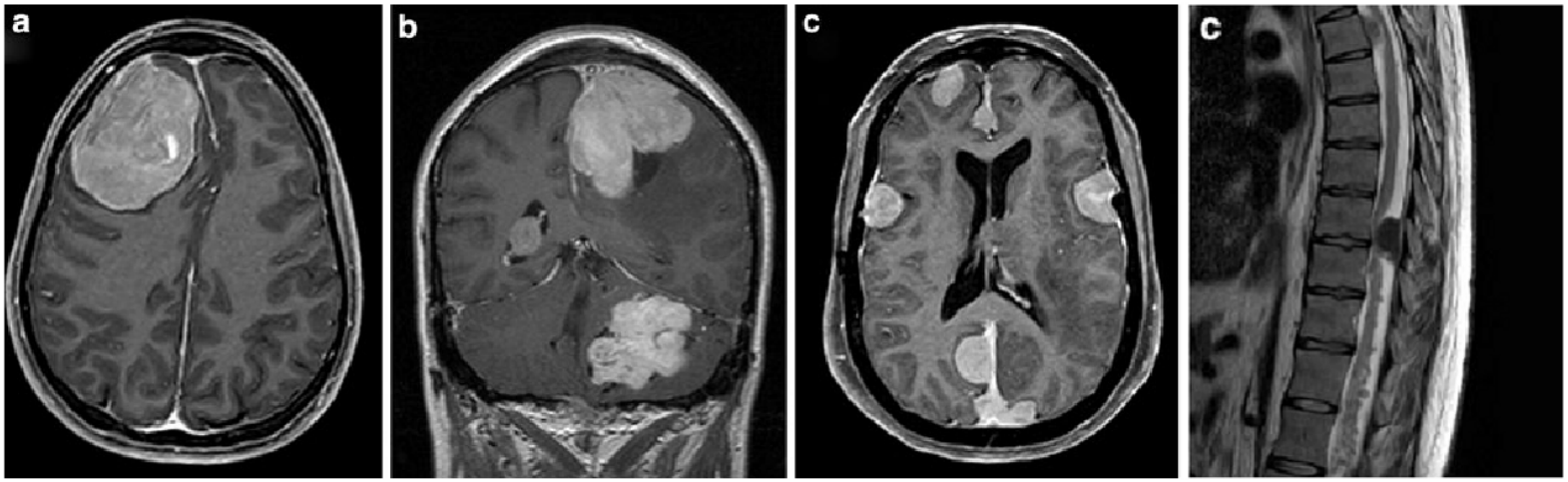
Most common primary brain tumor (~25-30%)

- 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)

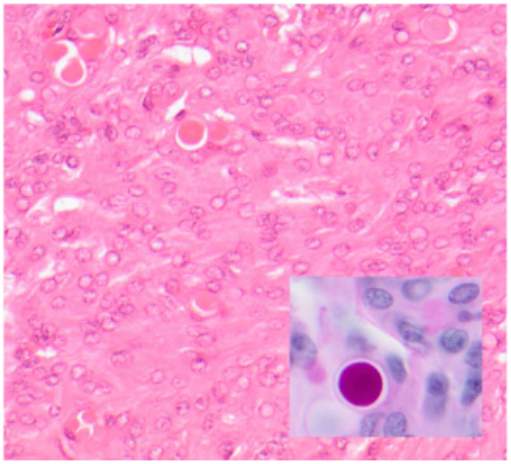


[Goutagny and Kalamarides, J Neurooncol, 2010]

Secretory Meningioma

Subgroup of benign meningiomas (grade I)

- rare occurrence (~3% of meningiomas)



[Reuss et al., Acta Neuropathol, 2013]

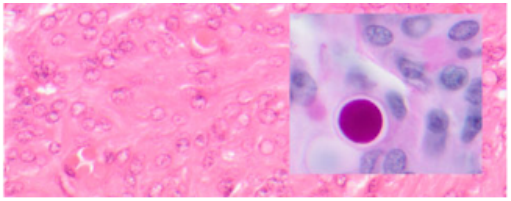
Secretory Meningioma

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]

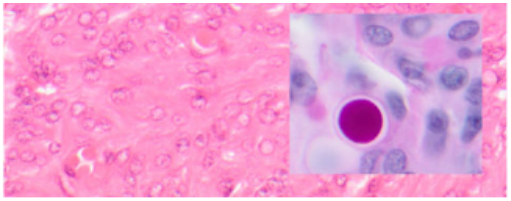
Secretory Meningioma

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
                       .. * *                                        .

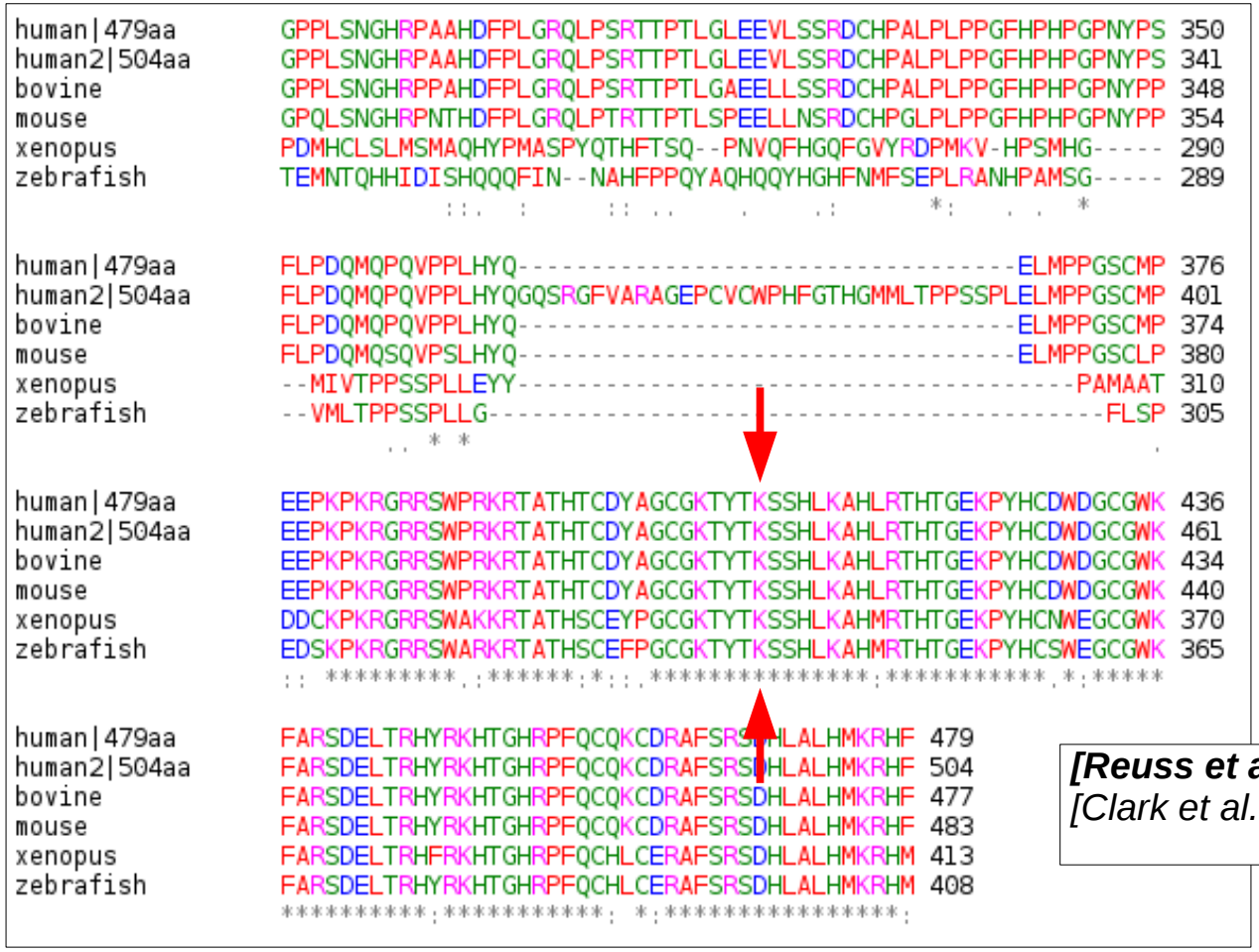
human|479aa      EEPKPKRGRRSWPRKRTATHTCDYAGCGKTYTKSSHLKAHLRTHTGEKPYHCDWDGCGWK 436
human2|504aa     EEPKPKRGRRSWPRKRTATHTCDYAGCGKTYTKSSHLKAHLRTHTGEKPYHCDWDGCGWK 461
bovine           EEPKPKRGRRSWPRKRTATHTCDYAGCGKTYTKSSHLKAHLRTHTGEKPYHCDWDGCGWK 434
mouse            EEPKPKRGRRSWPRKRTATHTCDYAGCGKTYTKSSHLKAHLRTHTGEKPYHCDWDGCGWK 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]

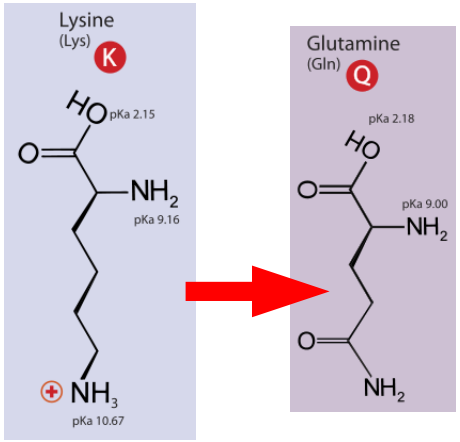
KLF4 K409Q

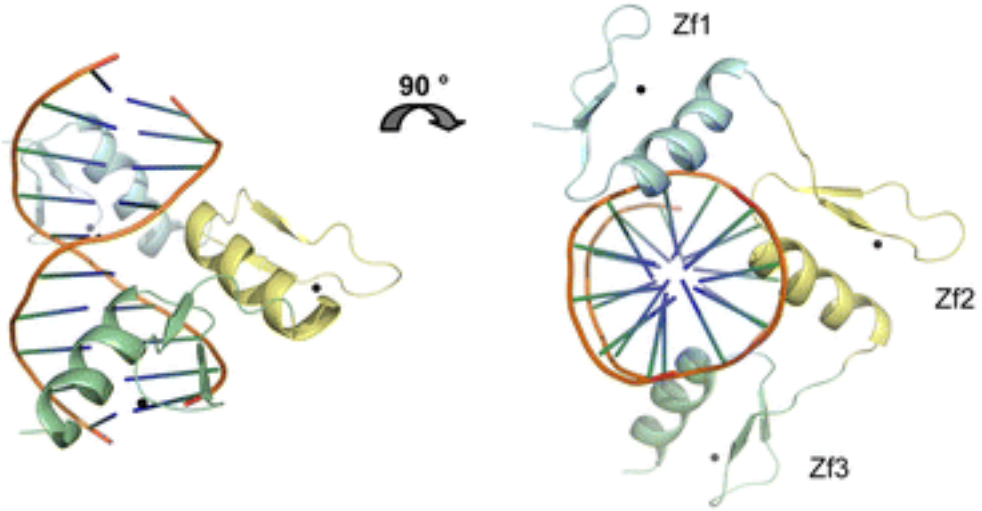
recurrent KLF4 mutation in ALL secretory meningiomas:
chr9:110,249,348 T→G (K409Q; lysine→glutamine**)**

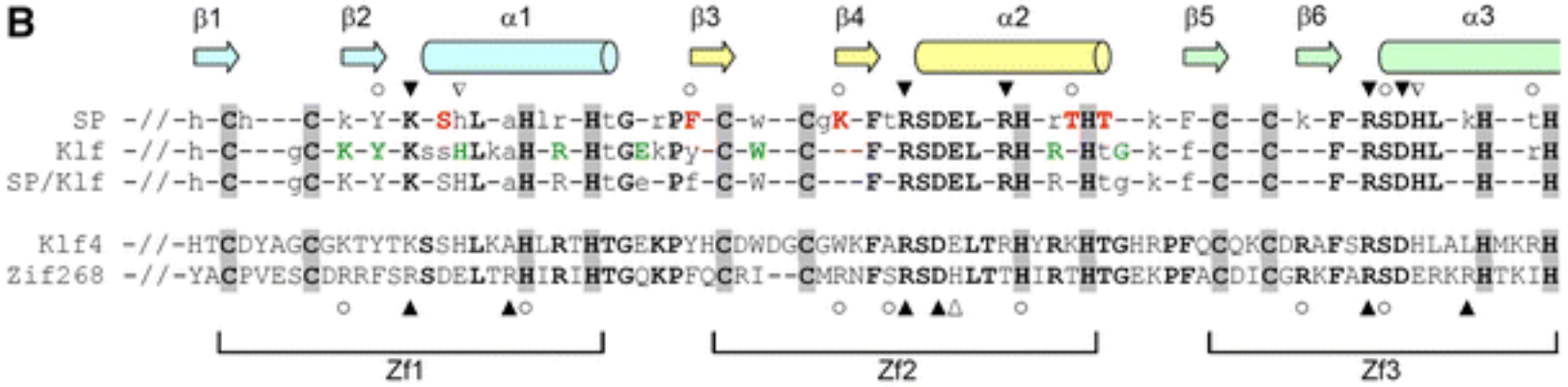


[Reuss et al., Acta Neuropathol, 2013]
[Clark et al., Science, 2013]

KLF4 K409Q: Effect On Function?

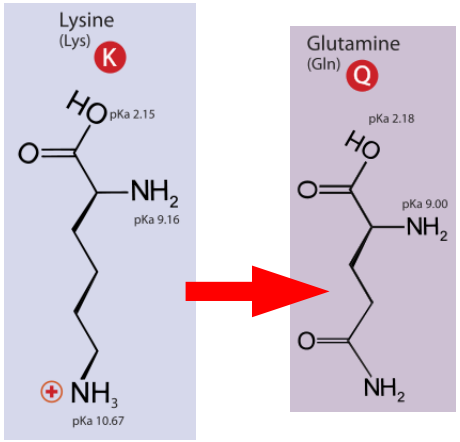


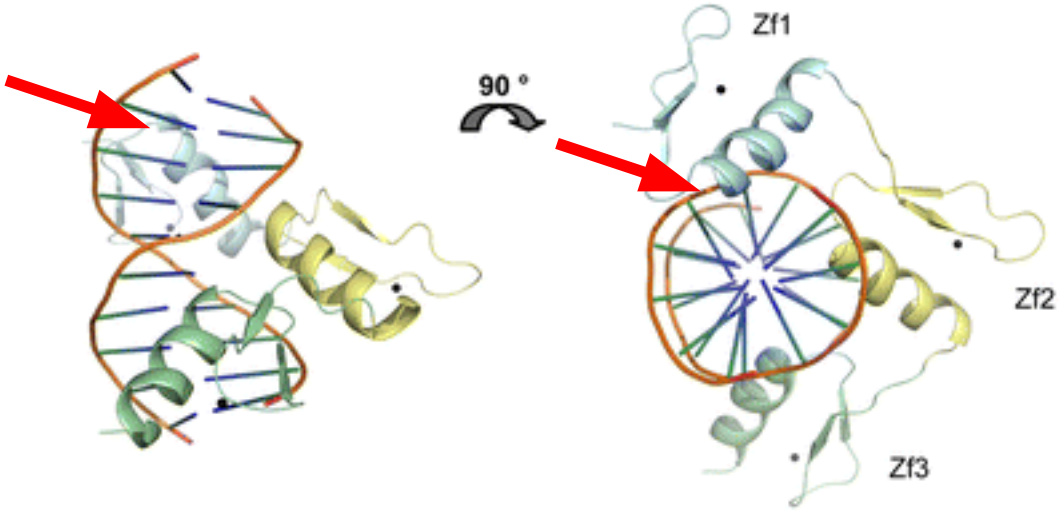


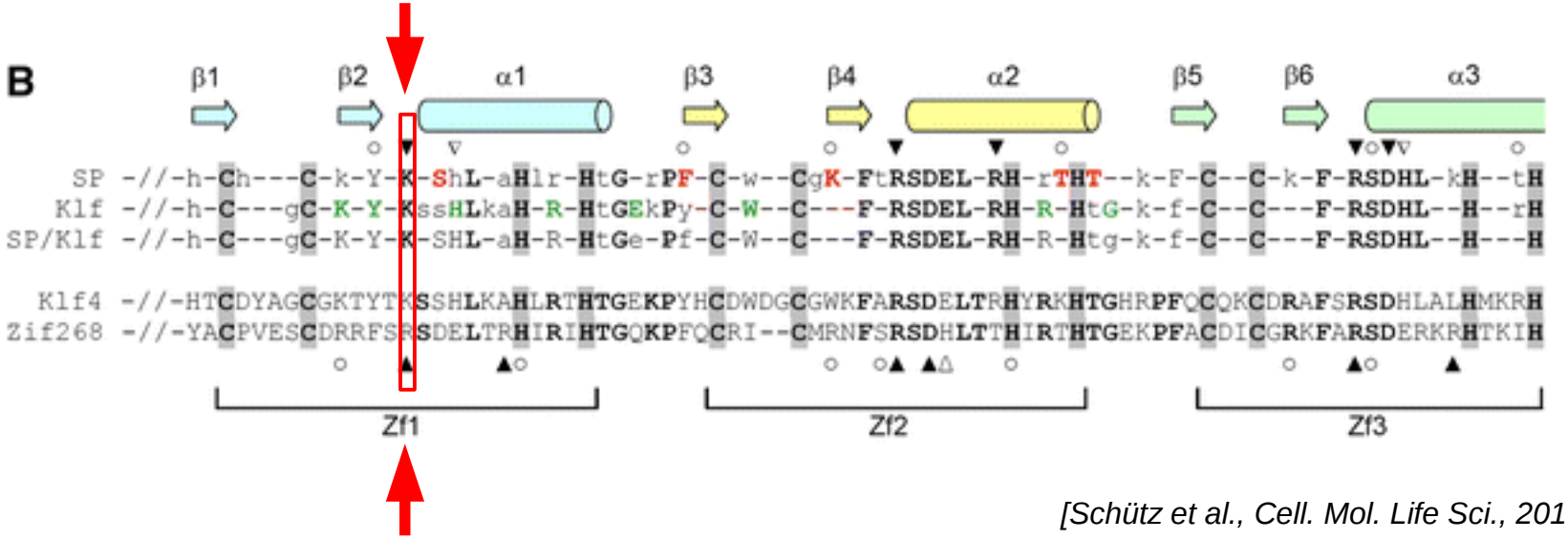


[Schütz et al., Cell. Mol. Life Sci., 2011]

KLF4 K409Q: Effect On Function?

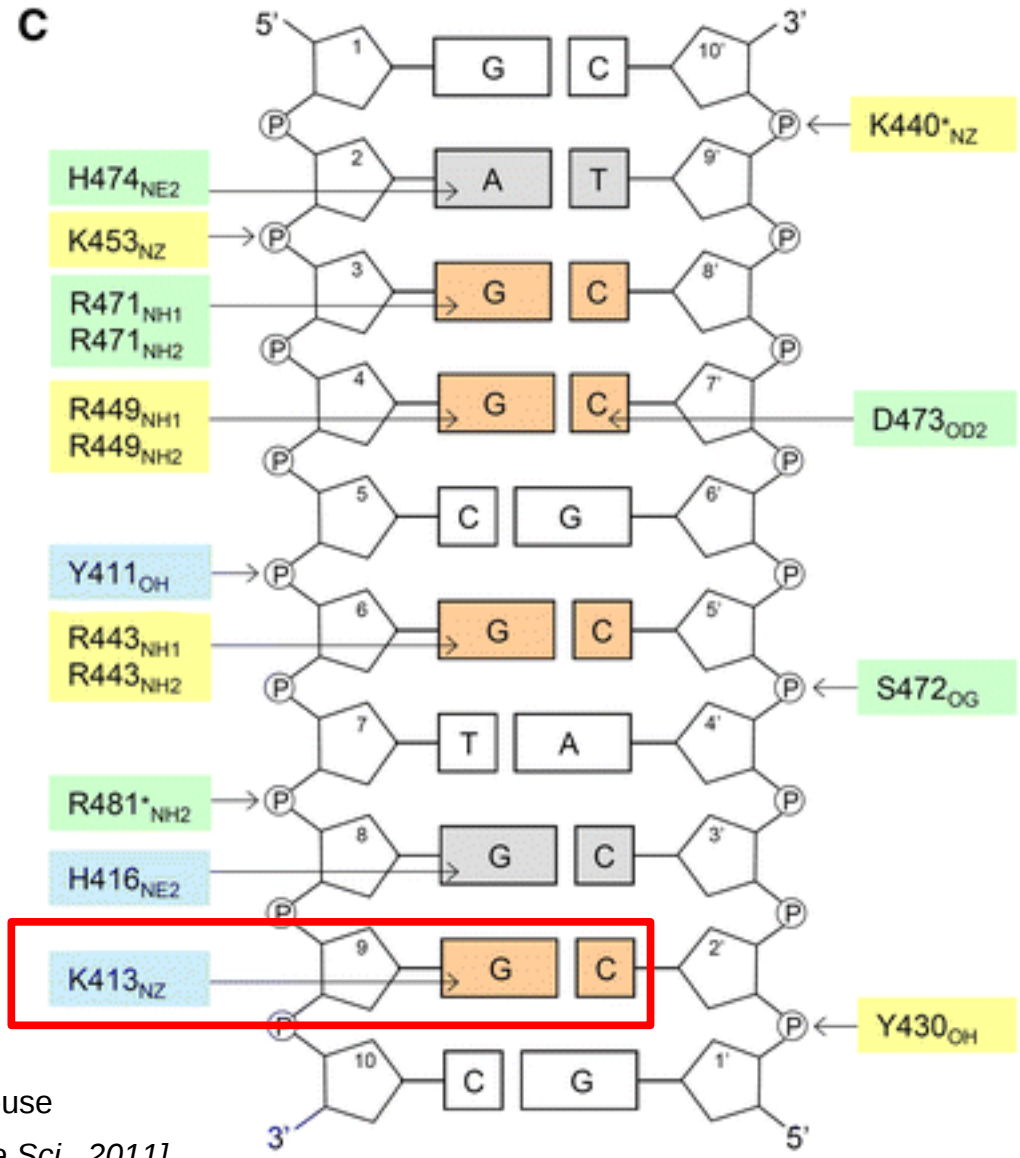






[Schütz et al., Cell. Mol. Life Sci., 2011]

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)

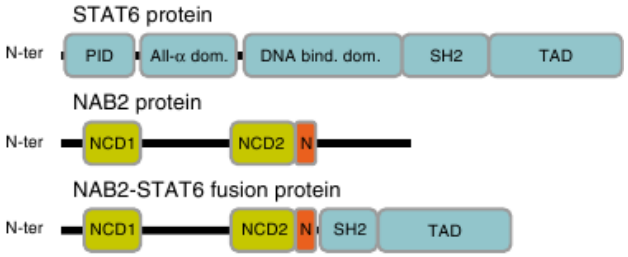
- non-CNS tumors considered as two variants of a single entity (WHO classification of soft tissue and bone tumors)
  - recurrent *NAB2-STAT6* fusion gene identified recently *[Chmielecki et al, 2013; Robinson et al, 2013]*
    - inversion and juxtaposition on Chr. 12q13
    - genes on opposing DNA strands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

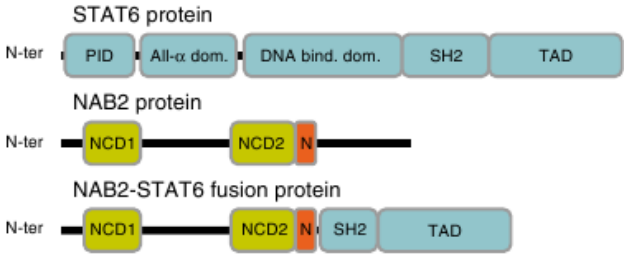
- non-CNS tumors considered as two variants of a single entity (WHO classification of soft tissue and bone tumors)
  - recurrent *NAB2-STAT6* fusion gene identified recently *[Chmielecki et al, 2013; Robinson et al, 2013]*
    - inversion and juxtaposition on Chr. 12q13
    - genes on opposing DNA strands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

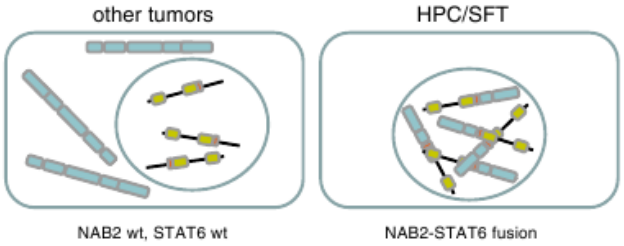
STAT6 IHC Staining In The Nucleus



[Schweizer et al., Acta Neuropathol, 2013]

STAT6 IHC Staining In The Nucleus

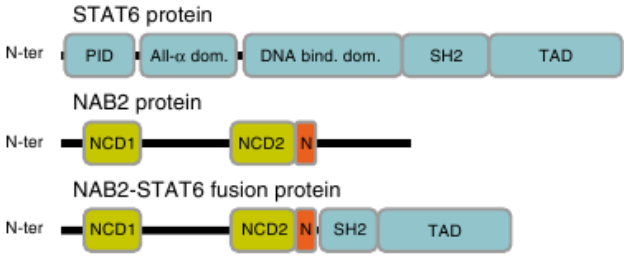


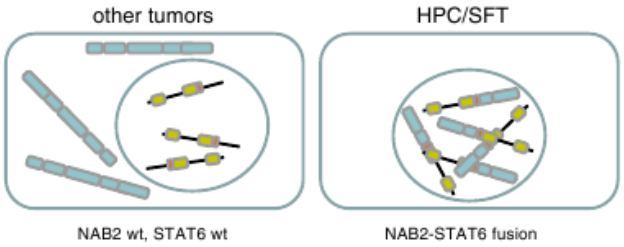


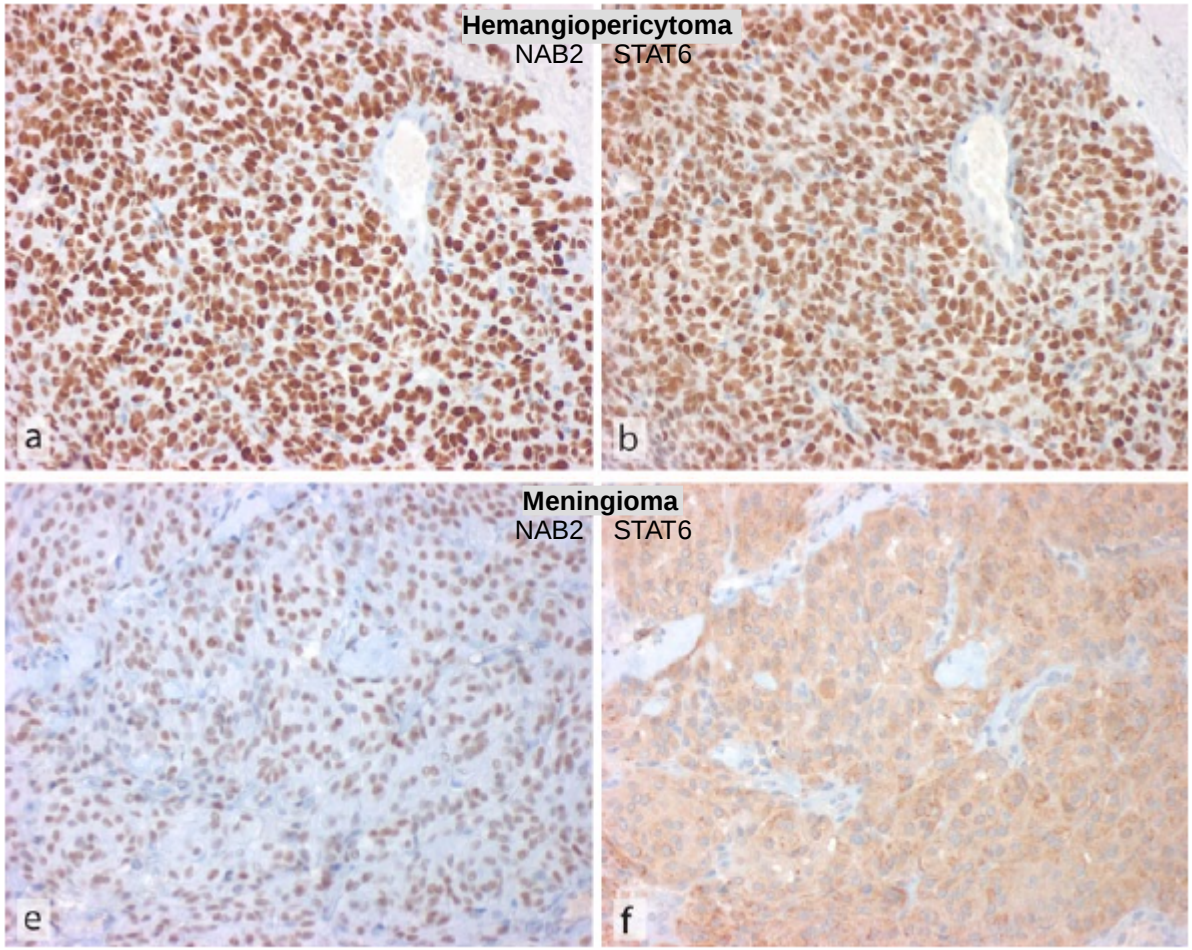
[Schweizer et al., Acta Neuropathol, 2013]

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

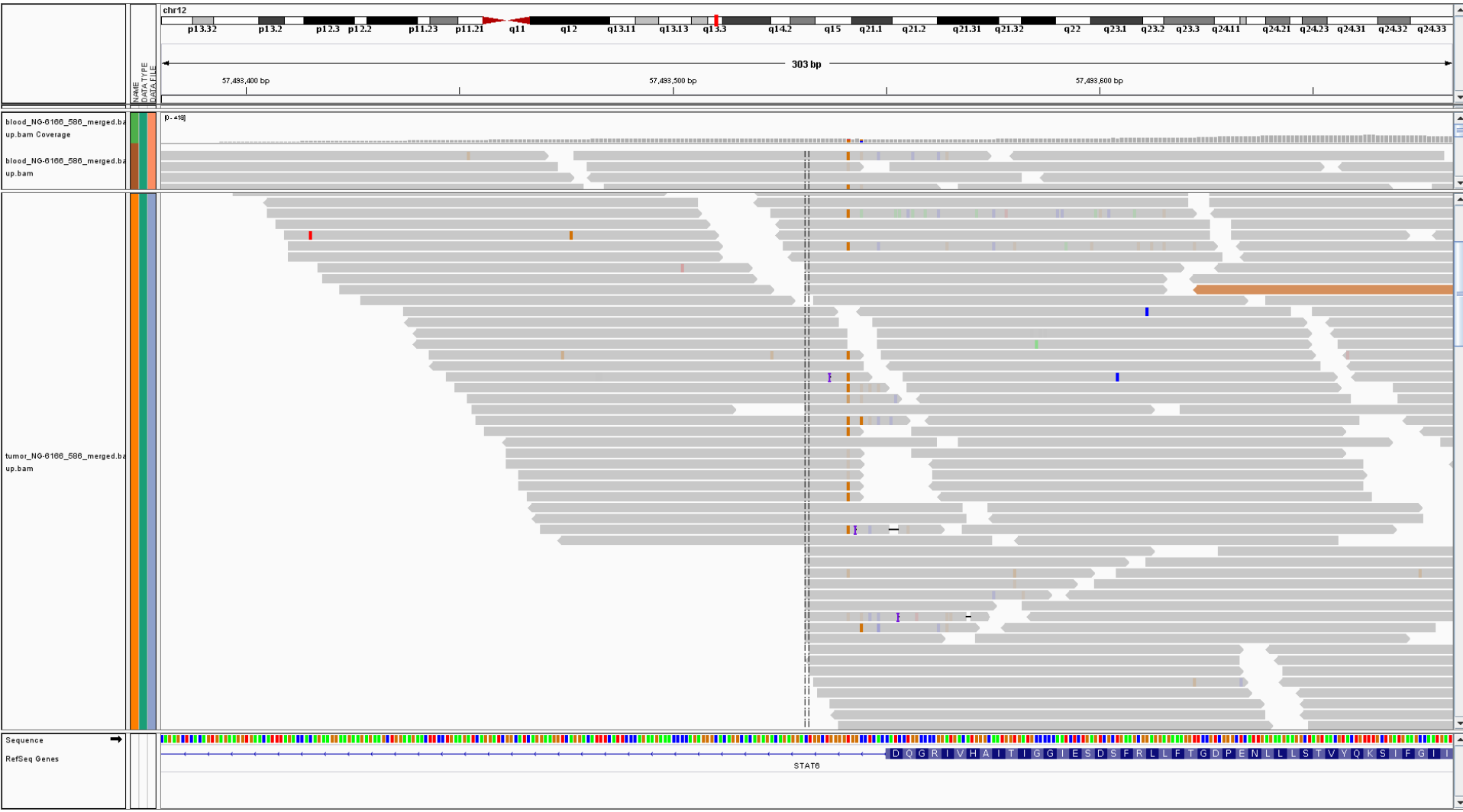




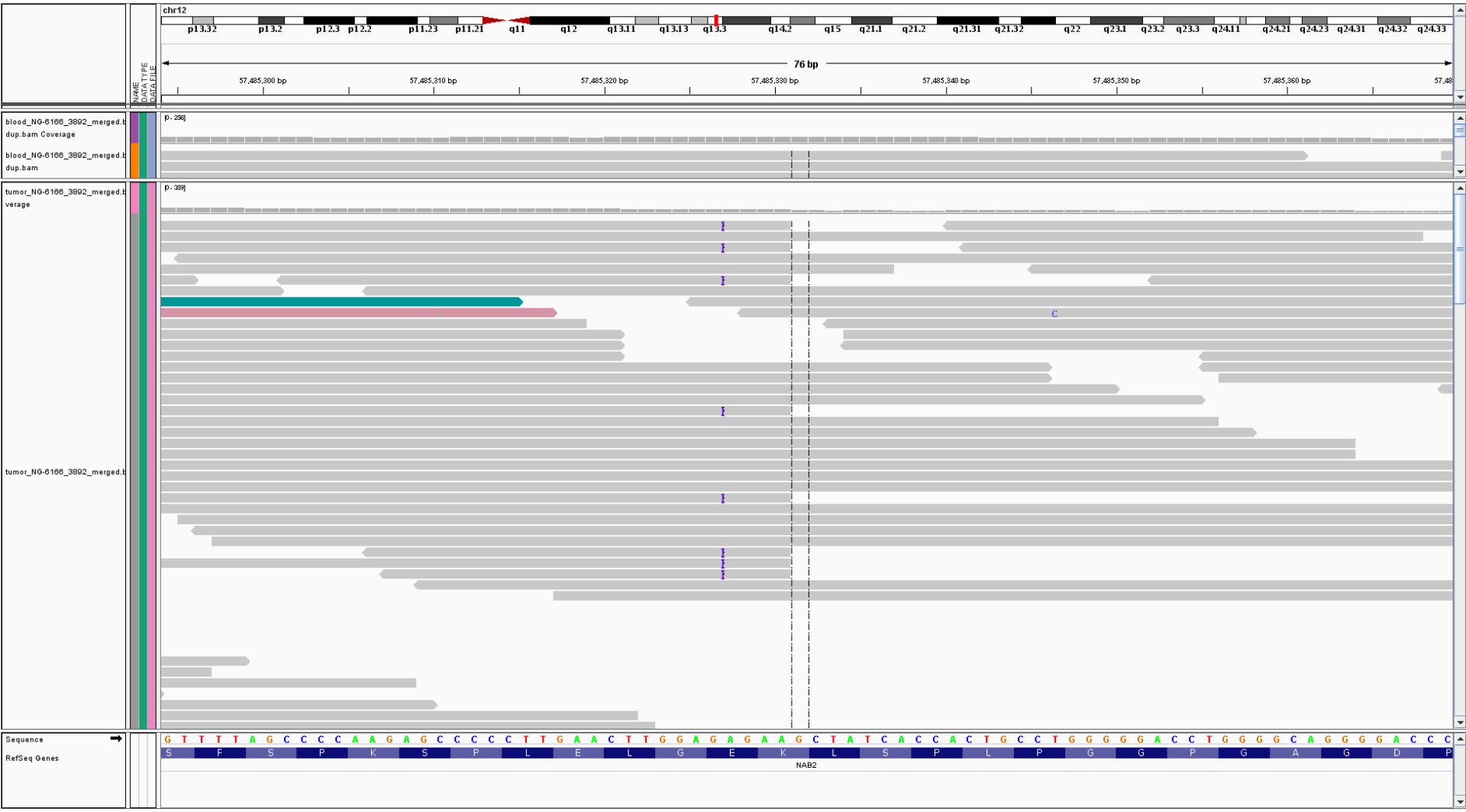


[Schweizer et al., Acta Neuropathol, 2013]

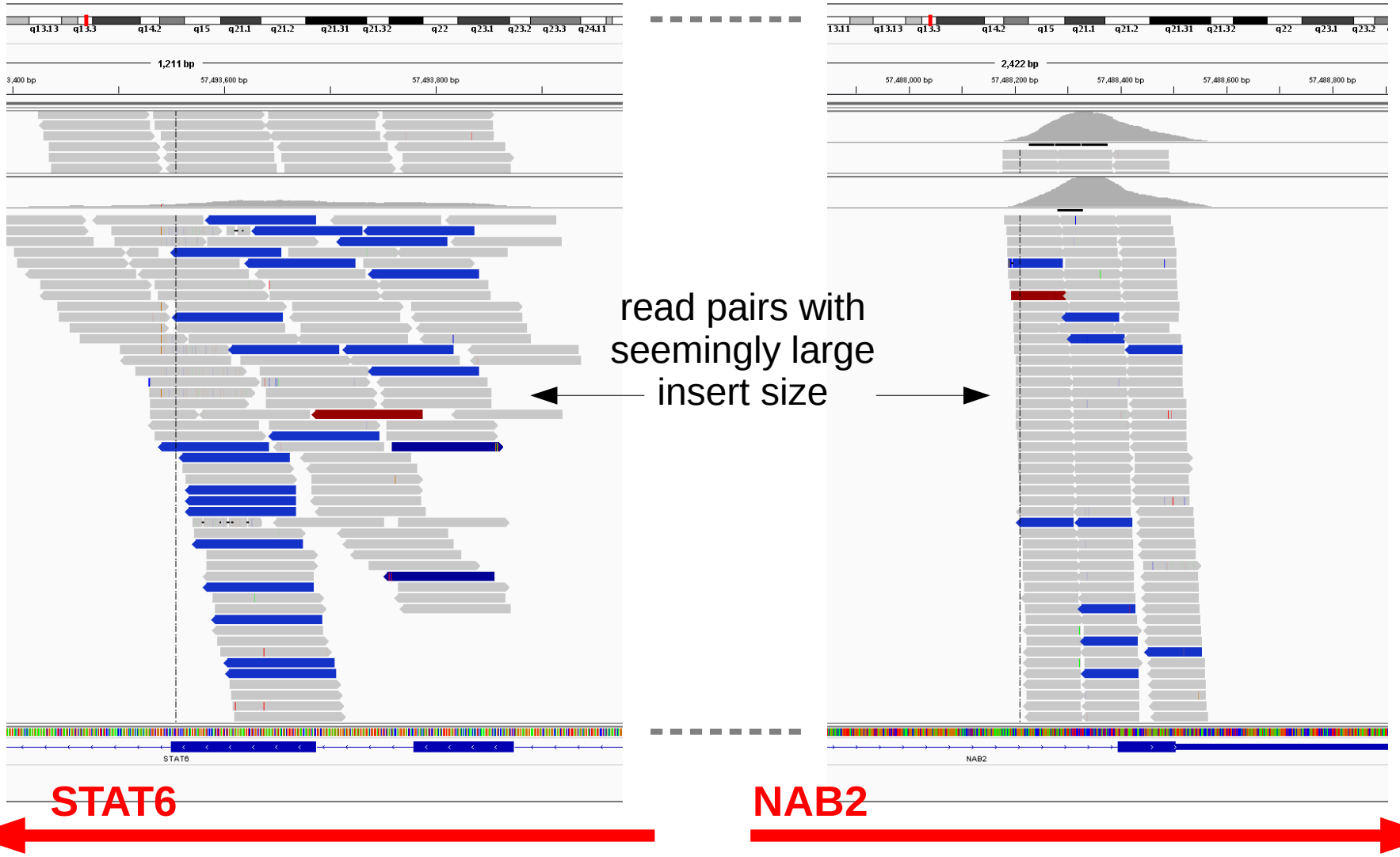
NAB2-STAT6 Fusion: Example 1



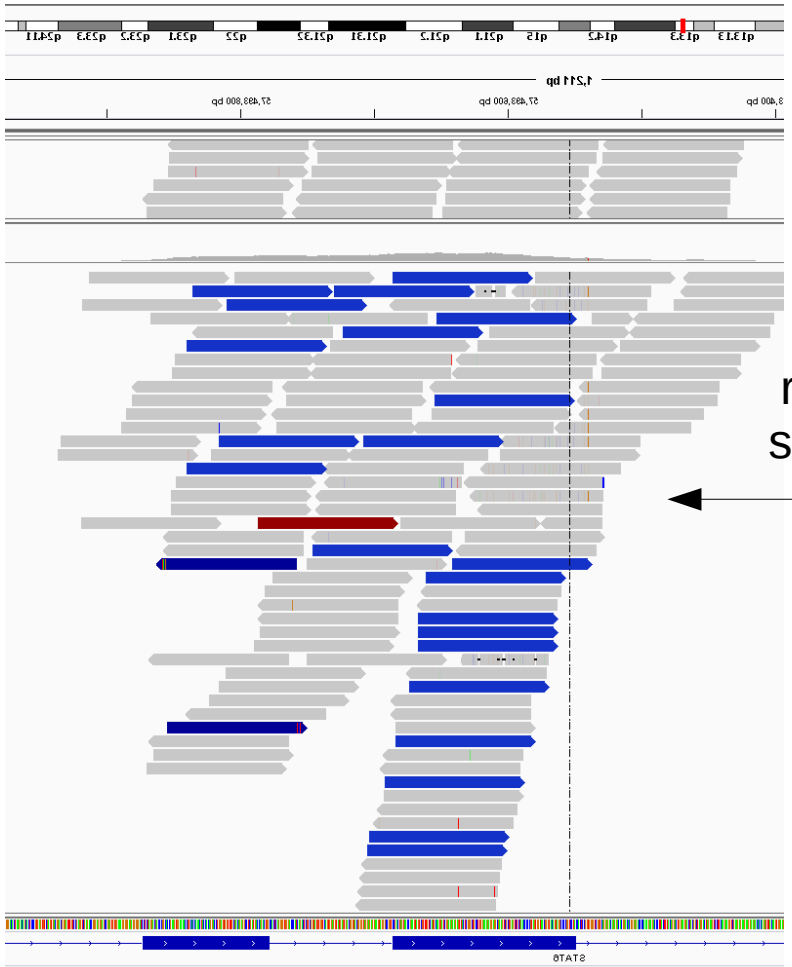
NAB2-STAT6 Fusion: Example 1



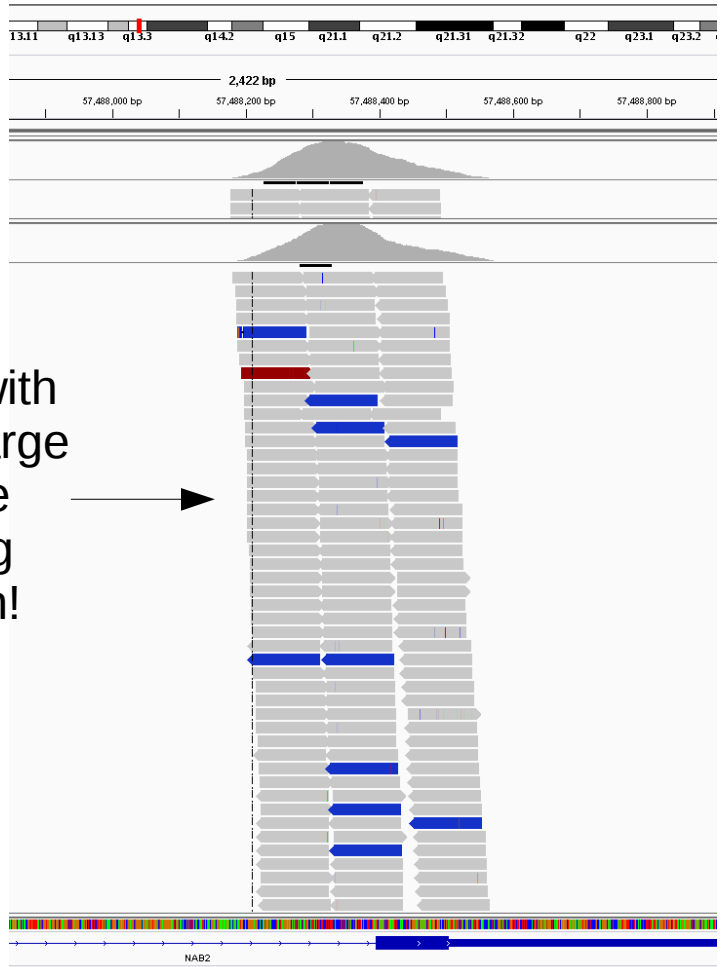
NAB2-STAT6 Fusion: Example 2



NAB2-STAT6 Fusion: Example 2

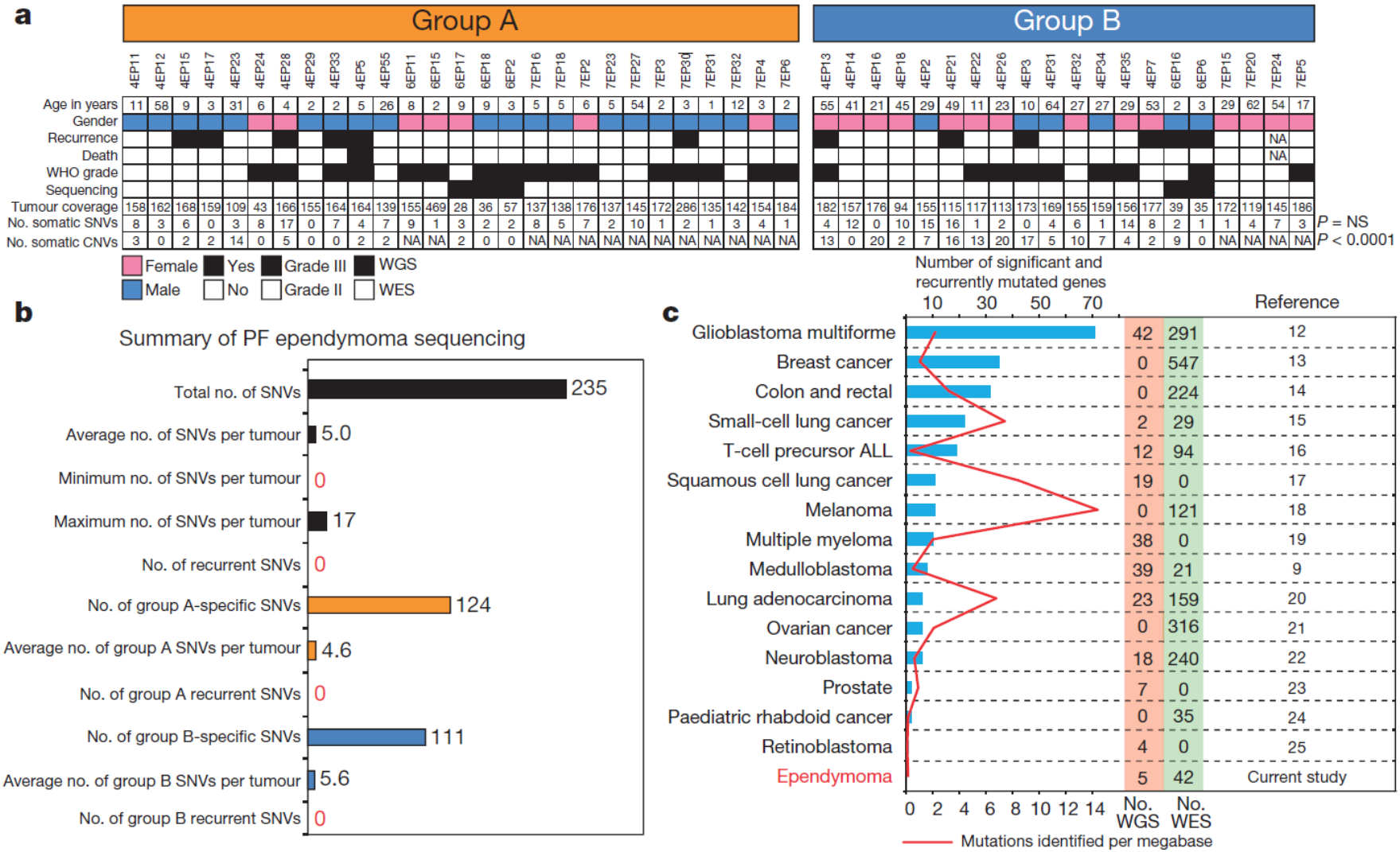


read pairs with seemingly large insert size and wrong orientation!



STAT6 → **NAB2**

Lack of mutations in some tumors



Somatic SNVs in coding regions are rare in posterior fossa ependymoma.

from: Mack, Witt, Piro et al. Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature* **506**:445–450, 2014

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways
- PathWave

Mutational Processes And Somatic Mutations

- Somatic mutations of individual tumors are caused by different mutational processes

Mutational Processes And Somatic Mutations

- Somatic mutations of individual tumors are caused by different mutational processes

- Mutational processes can significantly vary between tumors
  - between different cancer types
  - between individual tumors of the same cancer type

Mutational Processes And Somatic Mutations

- Somatic mutations of individual tumors are caused by different mutational processes
- Mutational processes can significantly vary between tumors
  - between different cancer types
  - between individual tumors of the same cancer type
- The sequence context of mutated bases is important!

Mutational Processes And Somatic Mutations

- Somatic mutations of individual tumors are caused by different mutational processes
- Mutational processes can significantly vary between tumors
  - between different cancer types
  - between individual tumors of the same cancer type
- The sequence context of mutated bases is important!
- Examples
  - Lung cancers of tobacco smokers have a highly increased number of cytosine>adenine (C>A) transversions
  - Spontaneous deamination of 5-methylcytosine (age-related) causes cytosine>thymine (C>T) transitions in the context of CpGs
  - See, for example, Alexandrov and Stratton, Curr Opin Genet Dev 24:52–60, 2014

Mutational Processes And Somatic Mutations

- Somatic mutations of individual tumors are caused by different mutational processes
- Mutational processes can significantly vary between tumors
  - between different cancer types
  - between individual tumors of the same cancer type
- The sequence context of mutated bases is important!
- Examples
  - Lung cancers of tobacco smokers have a highly increased number of cytosine>adenine (C>A) transversions
  - Spontaneous deamination of 5-methylcytosine (age-related) causes cytosine>thymine (C>T) transitions in the context of CpGs
  - See, for example, Alexandrov and Stratton, Curr Opin Genet Dev 24:52–60, 2014
- Mutational processes can be represented by means of “mutational signatures”
  - Reflect the frequencies of base changes within their sequence context

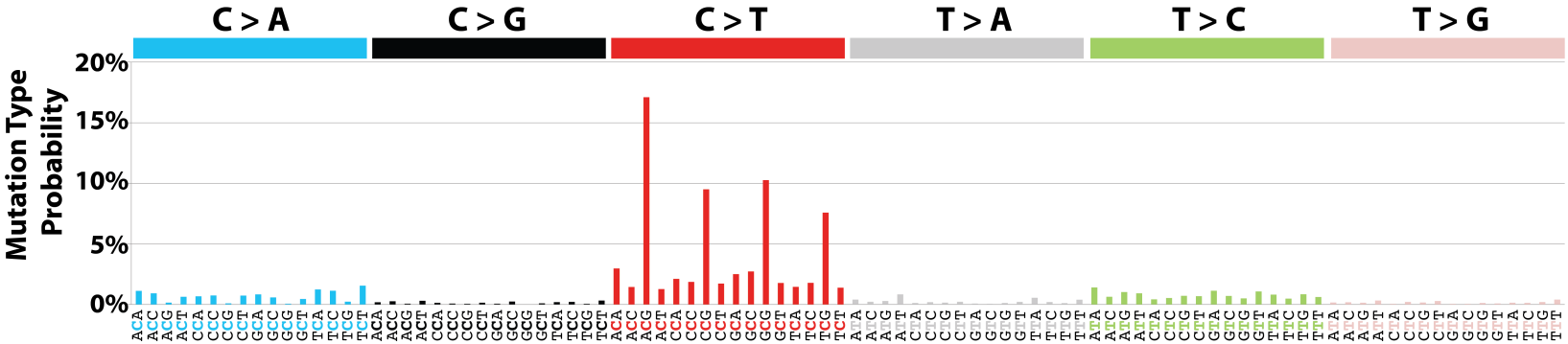
“Alexandrov” Signatures

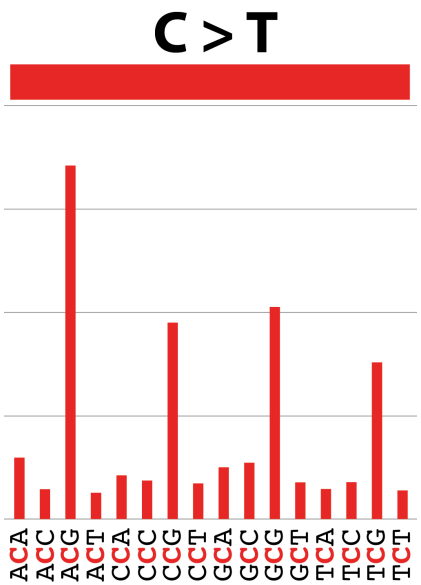
First published concept/notion of mutational signatures:

“Alexandrov” Signatures

First published concept/notion of mutational signatures:

“Full” model (Alexandrov et al, Nature 500:415–421, 2013)

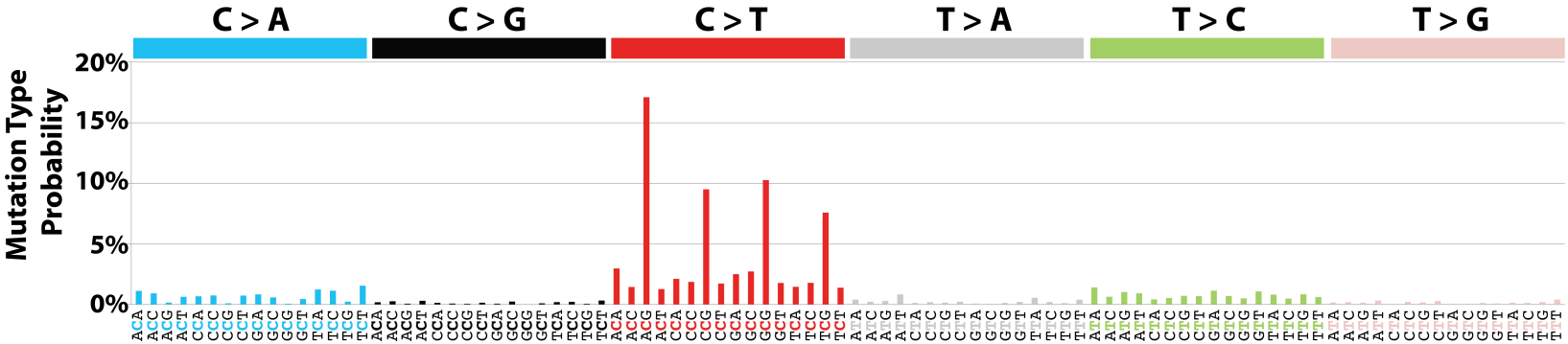


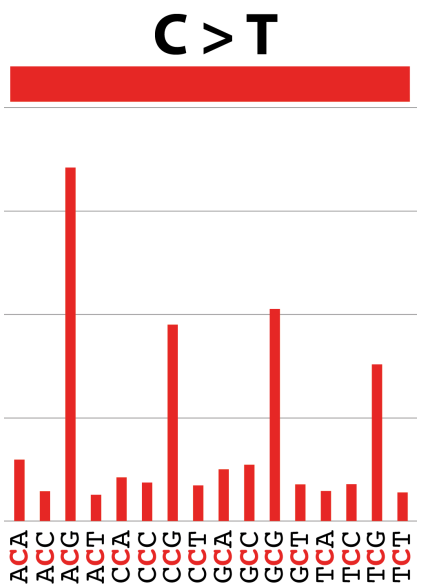


“Alexandrov” Signatures

First published concept/notion of mutational signatures:

“Full” model (Alexandrov et al, Nature 500:415–421, 2013)





- Full dependency between mutated and adjacent bases
- For 3-base sequence contexts: $6 \times 4 \times 4 = 96$ parameters
- Can be described by a vector of 96 probabilities (i.e., sum is 1)

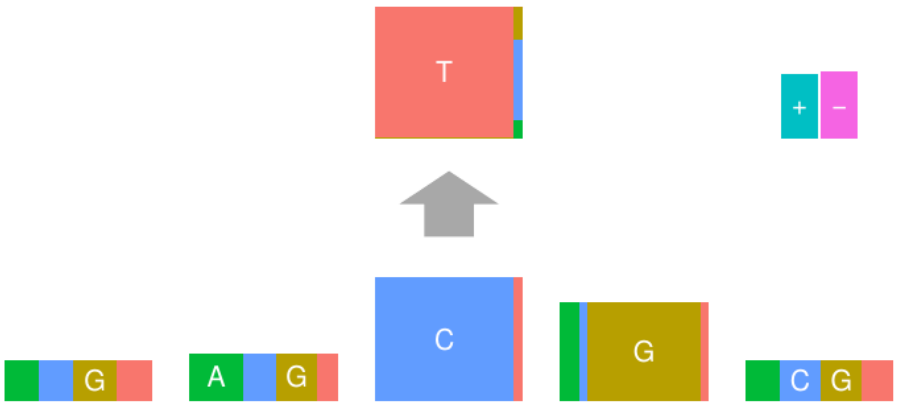
“Shiraishi” Signatures

Alternative concept/notion of mutational signatures:

“Shiraishi” Signatures

Alternative concept/notion of mutational signatures:

“Independent” model (Shiraishi et al, PLoS Genet 11:e1005657, 2015)



Nucleotide change (central base)

C>A	C>G	C>T	T>A	T>C	T>G
0.004	0.006	0.928	0.009	0.038	0.015

Flanking bases

Position	A	C	G	T
-2	0.237	0.228	0.293	0.242
-1	0.362	0.220	0.279	0.139
+1	0.131	0.053	0.764	0.052
+2	0.232	0.277	0.277	0.214

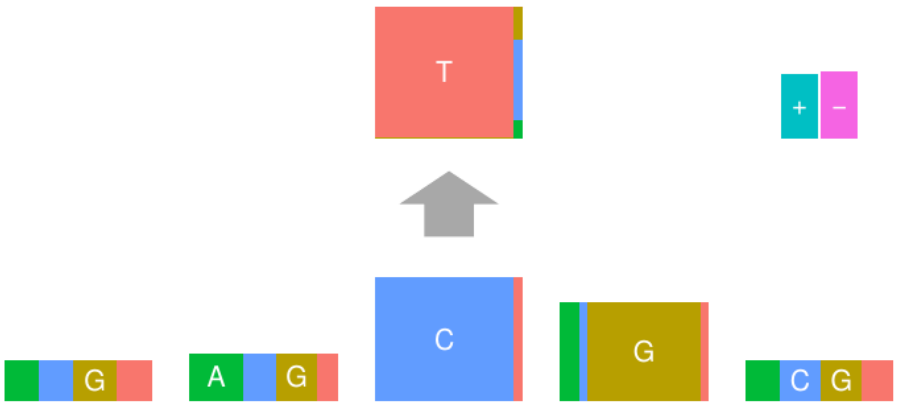
Transcription strand

plus strand	minus strand
0.493	0.507

“Shiraishi” Signatures

Alternative concept/notion of mutational signatures:

“Independent” model (Shiraishi et al, PLoS Genet 11:e1005657, 2015)



Nucleotide change (central base)

C>A	C>G	C>T	T>A	T>C	T>G
0.004	0.006	0.928	0.009	0.038	0.015

Flanking bases

Position	A	C	G	T
-2	0.237	0.228	0.293	0.242
-1	0.362	0.220	0.279	0.139
+1	0.131	0.053	0.764	0.052
+2	0.232	0.277	0.277	0.214

Transcription strand

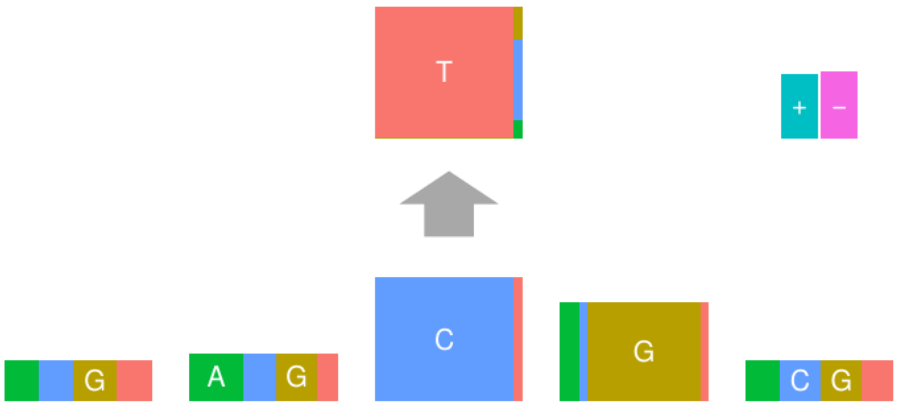
plus strand	minus strand
0.493	0.507

- Mutated base and adjacent bases as independent features

“Shiraishi” Signatures

Alternative concept/notion of mutational signatures:

“Independent” model (Shiraishi et al, PLoS Genet 11:e1005657, 2015)



Nucleotide change (central base)

C>A	C>G	C>T	T>A	T>C	T>G
0.004	0.006	0.928	0.009	0.038	0.015

Flanking bases

Position	A	C	G	T
-2	0.237	0.228	0.293	0.242
-1	0.362	0.220	0.279	0.139
+1	0.131	0.053	0.764	0.052
+2	0.232	0.277	0.277	0.214

Transcription strand

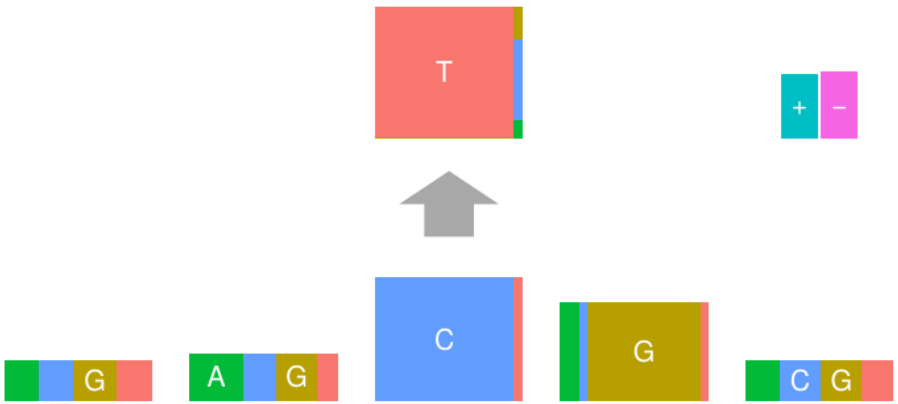
plus strand	minus strand
0.493	0.507

- Mutated base and adjacent bases as independent features
- For 5-base sequence contexts + transcriptional direction:
  $6 + 4 \times 4 + 2 = 24$ parameters

“Shiraishi” Signatures

Alternative concept/notion of mutational signatures:

“Independent” model (Shiraishi et al, PLoS Genet 11:e1005657, 2015)



Nucleotide change (central base)

C>A	C>G	C>T	T>A	T>C	T>G
0.004	0.006	0.928	0.009	0.038	0.015

Flanking bases

Position	A	C	G	T
-2	0.237	0.228	0.293	0.242
-1	0.362	0.220	0.279	0.139
+1	0.131	0.053	0.764	0.052
+2	0.232	0.277	0.277	0.214

Transcription strand

plus strand	minus strand
0.493	0.507

- Mutated base and adjacent bases as independent features
- For 5-base sequence contexts + transcriptional direction:
  $6 + 4 \times 4 + 2 = 24$ parameters
- Can be described by a table

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways
- PathWave

How Are Mutational Signatures Derived?

(The following describes Alexandrov signatures; same for Shiraishi)

- The somatic mutations of a tumor are caused by multiple mutational processed. → We observe an overlap of multiple mutational signatures!

How Are Mutational Signatures Derived?

(The following describes Alexandrov signatures; same for Shiraishi)

- The somatic mutations of a tumor are caused by multiple mutational processed. → We observe an overlap of multiple mutational signatures!

- Basic idea: the 96 mutation frequencies observed in tumor genome $\vec{g}$ can be described as the weighted sum of N signature vectors $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} w_k \vec{s}_k \qquad \text{with} \qquad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

How Are Mutational Signatures Derived?

(The following describes Alexandrov signatures; same for Shiraishi)

- The somatic mutations of a tumor are caused by multiple mutational processed. → We observe an overlap of multiple mutational signatures!

- Basic idea: the 96 mutation frequencies observed in tumor genome $\vec{g}$ can be described as the weighted sum of N signature vectors $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} w_k \vec{s}_k \qquad \text{with} \qquad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

- For a *set of G tumor genomes* we have:

$$\mathbf{G} = \mathbf{S} \times \mathbf{W}$$

  - **G** is the $96 \times G$-matrix of observed mutation frequencies in the tumors;
  - **S** is the $96 \times N$-matrix containing all signatures (one per column); and
  - **W** is the $N \times G$-matrix of weights (also called “exposures” or “contributions”) of the signatures in the single tumors

How Are Mutational Signatures Derived?

- For a *set of G tumor genomes* we have:

$$\mathbf{G} = \mathbf{S} \times \mathbf{W}$$

  - **G** is the $96 \times G$-matrix of observed mutation frequencies in the tumors;
  - **S** is the $96 \times N$-matrix containing all signatures (one per column); and
  - **W** is the $N \times G$-matrix of weights (also called “exposures” or “contributions”) of the signatures in the single tumors

How Are Mutational Signatures Derived?

- For a *set of G tumor genomes* we have:

$$\mathbf{G} = \mathbf{S} \times \mathbf{W}$$

  - **G** is the $96 \times G$-matrix of observed mutation frequencies in the tumors;
  - **S** is the $96 \times N$-matrix containing all signatures (one per column); and
  - **W** is the $N \times G$-matrix of weights (also called “exposures” or “contributions”) of the signatures in the single tumors

- Derive **S** and **W** at the same time!

  - Non-negative matrix factorization (Alexandrov et al, Cell Reports 3:246–259, 2013)
  - Principal component analysis (Gehring et al, Bioinformatics 31:3673–3675, 2015)

How Are Mutational Signatures Derived?

- For a *set of G tumor genomes* we have:

$$\mathbf{G} = \mathbf{S} \times \mathbf{W}$$

  - **G** is the 96 × *G*-matrix of observed mutation frequencies in the tumors;
  - **S** is the 96 × *N*-matrix containing all signatures (one per column); and
  - **W** is the *N* × *G*-matrix of weights (also called “exposures” or “contributions”) of the signatures in the single tumors

- Derive **S** and **W** at the same time!

  - Non-negative matrix factorization (Alexandrov et al, Cell Reports 3:246–259, 2013)
  - Principal component analysis (Gehring et al, Bioinformatics 31:3673–3675, 2015)

- Requires a large set of tumors!

How Are Mutational Signatures Derived?

- For a *set of G tumor genomes* we have:

$$\mathbf{G} = \mathbf{S} \times \mathbf{W}$$

  - **G** is the 96 × *G*-matrix of observed mutation frequencies in the tumors;
  - **S** is the 96 × *N*-matrix containing all signatures (one per column); and
  - **W** is the *N* × *G*-matrix of weights (also called “exposures” or “contributions”) of the signatures in the single tumors

- Derive **S** and **W** at the same time!
  - Non-negative matrix factorization (Alexandrov et al, Cell Reports 3:246–259, 2013)
  - Principal component analysis (Gehring et al, Bioinformatics 31:3673–3675, 2015)

- Requires a large set of tumors!

- What if you what to determine which mutational processes contributed to the mutation load of a *single tumor*???
  (E.g., in a clinical setting)

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

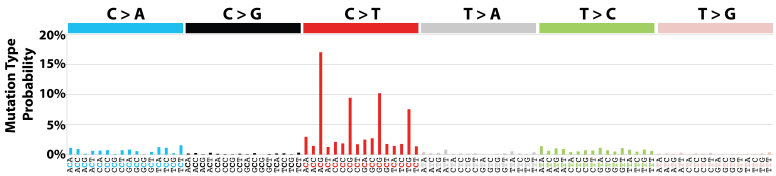
Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways
- PathWave

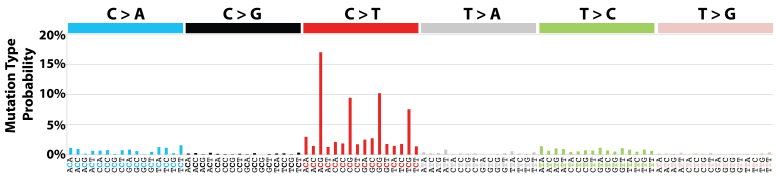
Alexandrov Signatures In A Single Tumor



- Remember: the 96 mutation frequencies observed in *one* tumor genome $\vec{g}$ can be described as weighted sum of N signatures $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} w_k \vec{s}_k \qquad \text{with} \qquad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

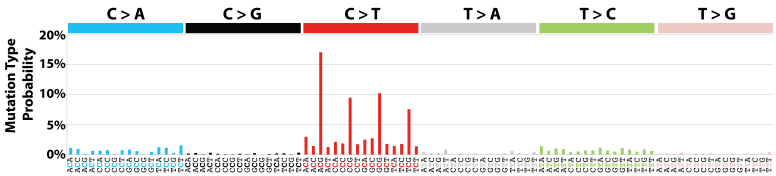
Alexandrov Signatures In A Single Tumor



- Remember: the 96 mutation frequencies observed in *one* tumor genome $\vec{g}$ can be described as weighted sum of N signatures $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} (w_k \vec{s}_k) + \vec{\epsilon} \qquad \text{with} \quad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

Alexandrov Signatures In A Single Tumor

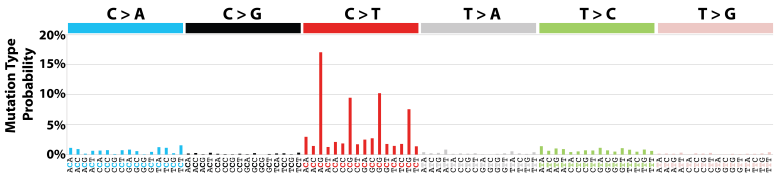


- Remember: the 96 mutation frequencies observed in *one* tumor genome $\vec{g}$ can be described as weighted sum of N signatures $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} (w_k \vec{s}_k) + \vec{\epsilon} \qquad \text{with} \quad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

- Given: $\vec{g}$ (observed mutations), $\vec{s}_k$ (signatures)

Alexandrov Signatures In A Single Tumor

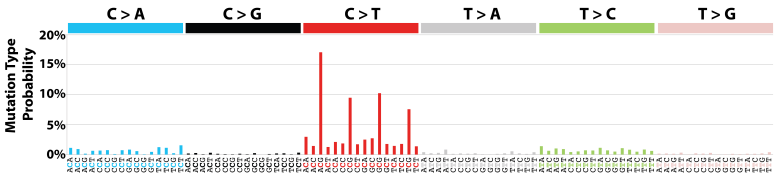


- Remember: the 96 mutation frequencies observed in *one* tumor genome $\vec{g}$ can be described as weighted sum of N signatures $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} (w_k \vec{s}_k) + \vec{\epsilon} \qquad \text{with} \quad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

- Given: $\vec{g}$ (observed mutations), $\vec{s}_k$ (signatures)
- Goal: compute weights w_k and minimize the error terms (ϵ_k)!
  → weights measure the signatures' contributions to the mutational load of the tumor!

Alexandrov Signatures In A Single Tumor

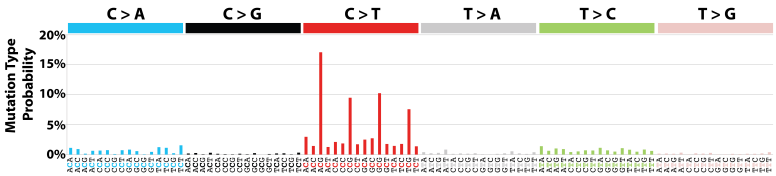


- Remember: the 96 mutation frequencies observed in *one* tumor genome $\vec{g}$ can be described as weighted sum of N signatures $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} (w_k \vec{s}_k) + \vec{\epsilon} \qquad \text{with} \quad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

- Given: $\vec{g}$ (observed mutations), $\vec{s}_k$ (signatures)
- Goal: compute weights w_k and minimize the error terms (ϵ_k)!
  → weights measure the signatures' contributions to the mutational load of the tumor!
- Tool: `deconstructSigs` (Rosenthal et al, Genome Biol 17:31, 2016)
  - R package; constructs a solution by iteratively adding single mutational signatures to minimize the sum-squared error between $\vec{g}$ and $\sum_{k=1}^{N} (w_k \vec{s}_k)$

Alexandrov Signatures In A Single Tumor



- Remember: the 96 mutation frequencies observed in *one* tumor genome $\vec{g}$ can be described as weighted sum of N signatures $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} (w_k \vec{s}_k) + \vec{\epsilon} \qquad \text{with} \qquad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

- Given: $\vec{g}$ (observed mutations), $\vec{s}_k$ (signatures)
- Goal: compute weights w_k and minimize the error terms (ϵ_k)!
  → weights measure the signatures' contributions to the mutational load of the tumor!
- Tool: `deconstructSigs` (Rosenthal et al, Genome Biol 17:31, 2016)
  - R package; constructs a solution by iteratively adding single mutational signatures to minimize the sum-squared error between $\vec{g}$ and $\sum_{k=1}^{N} (w_k \vec{s}_k)$
- For Shiraishi: `decompTumor2Sig`(Krüger & Piro, NETTAB 2017)

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
```

```
### decompTumor2Sig ###
```

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
library(pmsignature)
numsig = 4

# determine signatures from DB of tumor genomes
G <- readMPFile("myMutDB.MPF.txt.gz", numBases = 5, trDir = TRUE)
Param <- getPMSignature(G, K=numsig)

### decompTumor2Sig ###
```

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
library(pmsignature)
numsig = 4

# determine signatures from DB of tumor genomes
G <- readMPFile("myMutDB.MPF.txt.gz", numBases = 5, trDir = TRUE)
Param <- getPMSignature(G, K=numsig)

### decompTumor2Sig ###
```

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
library(pmsignature)
numsig = 4

# determine signatures from DB of tumor genomes
G <- readMPFile("myMutDB.MPF.txt.gz", numBases = 5, trDir = TRUE)
Param <- getPMSignature(G, K=numsig)

### decompTumor2Sig ###
library(decompTumor2Sig)

# read mutations of the individual tumor (function by pmsignature)
T <- readMPFile("queryTumor.MPF.txt.gz", numBases = 5, trDir = TRUE)
```

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
library(pmsignature)
numsig = 4

# determine signatures from DB of tumor genomes
G <- readMPFile("myMutDB.MPF.txt.gz", numBases = 5, trDir = TRUE)
Param <- getPMSignature(G, K=numsig)

### decompTumor2Sig ###
library(decompTumor2Sig)

# read mutations of the individual tumor (function by pmsignature)
T <- readMPFile("queryTumor.MPF.txt.gz", numBases = 5, trDir = TRUE)

# convert mutation data to a matrix (same form as signatures)
genomes <- getGenomesFromMutationFeatureData(T, normalize=TRUE)
```

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
library(pmsignature)
numsig = 4

# determine signatures from DB of tumor genomes
G <- readMPFile("myMutDB.MPF.txt.gz", numBases = 5, trDir = TRUE)
Param <- getPMSignature(G, K=numsig)

### decompTumor2Sig ###
library(decompTumor2Sig)

# read mutations of the individual tumor (function by pmsignature)
T <- readMPFile("queryTumor.MPF.txt.gz", numBases = 5, trDir = TRUE)

# convert mutation data to a matrix (same form as signatures)
genomes <- getGenomesFromMutationFeatureData(T, normalize=TRUE)

# predict exposures for tumor
signatures <- getSignatureListFromEstimatedParameters(Param)
```

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
library(pmsignature)
numsig = 4

# determine signatures from DB of tumor genomes
G <- readMPFile("myMutDB.MPF.txt.gz", numBases = 5, trDir = TRUE)
Param <- getPMSignature(G, K=numsig)

### decompTumor2Sig ###
library(decompTumor2Sig)

# read mutations of the individual tumor (function by pmsignature)
T <- readMPFile("queryTumor.MPF.txt.gz", numBases = 5, trDir = TRUE)

# convert mutation data to a matrix (same form as signatures)
genomes <- getGenomesFromMutationFeatureData(T, normalize=TRUE)

# predict exposures for tumor
signatures <- getSignatureListFromEstimatedParameters(Param)
exposures <- decomposeTumorGenomes(genomes, signatures)
```

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
library(pmsignature)
numsig = 4

# determine signatures from DB of tumor genomes
G <- readMPFile("myMutDB.MPF.txt.gz", numBases = 5, trDir = TRUE)
Param <- getPMSignature(G, K=numsig)

### decompTumor2Sig ###
library(decompTumor2Sig)

# read mutations of the individual tumor (function by pmsignature)
T <- readMPFile("queryTumor.MPF.txt.gz", numBases = 5, trDir = TRUE)

# convert mutation data to a matrix (same form as signatures)
genomes <- getGenomesFromMutationFeatureData(T, normalize=TRUE)

# predict exposures for tumor
signatures <- getSignatureListFromEstimatedParameters(Param)
exposures <- decomposeTumorGenomes(genomes, signatures)

exposures
#$PD4072a
#[1] 0.41464965 0.52231228 0.06303806 0.00000000
```

Quadratic Programming Approach

(Following Lynch, F1000Research 5:1253, 2016)

We want $\mathbf{S}\vec{w} \approx \vec{g}$, so we can solve the following

Problem

$$\begin{aligned} \text{minimize} \quad & (\vec{g} - \mathbf{S}\vec{w})^T(\vec{g} - \mathbf{S}\vec{w}) \\ \text{subject to} \quad & \textstyle\sum_{s=1}^{k} w_s = 1, w_s \geq 0 \end{aligned}$$

Quadratic Programming Approach

(Following Lynch, F1000Research 5:1253, 2016)

We want $\mathbf{S}\vec{w} \approx \vec{g}$, so we can solve the following

Problem

$$\text{minimize} \quad (\vec{g} - \mathbf{S}\vec{w})^T(\vec{g} - \mathbf{S}\vec{w}) = \vec{g}^T\vec{g} - \vec{g}^T\mathbf{S}\vec{w} - (\mathbf{S}\vec{w})^T\vec{g} + (\mathbf{S}\vec{w})^T\mathbf{S}\vec{w}$$
$$\text{subject to} \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

Quadratic Programming Approach

(Following Lynch, F1000Research 5:1253, 2016)

We want $\mathbf{S}\vec{w} \approx \vec{g}$, so we can solve the following

Problem

$$\text{minimize} \quad (\vec{g} - \mathbf{S}\vec{w})^T(\vec{g} - \mathbf{S}\vec{w}) = \vec{g}^T\vec{g} - \vec{g}^T\mathbf{S}\vec{w} - (\mathbf{S}\vec{w})^T\vec{g} + (\mathbf{S}\vec{w})^T\mathbf{S}\vec{w}$$
$$\text{subject to} \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

Since $\vec{g}^T\vec{g}$ is constant and $(\mathbf{S}\vec{w})^T\vec{g} = \vec{g}^T\mathbf{S}\vec{w}$, we can simplify the problem:

$$\text{minimize} \quad -\vec{g}^T\mathbf{S}\vec{w} + \tfrac{1}{2}\vec{w}^T\mathbf{S}^T\mathbf{S}\vec{w} \qquad \text{subject to} \quad \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

Quadratic Programming Approach

(Following Lynch, F1000Research 5:1253, 2016)

We want $\mathbf{S}\vec{w} \approx \vec{g}$, so we can solve the following

Problem

$$\text{minimize} \quad (\vec{g} - \mathbf{S}\vec{w})^T(\vec{g} - \mathbf{S}\vec{w}) = \vec{g}^T\vec{g} - \vec{g}^T\mathbf{S}\vec{w} - (\mathbf{S}\vec{w})^T\vec{g} + (\mathbf{S}\vec{w})^T\mathbf{S}\vec{w}$$
$$\text{subject to} \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

Since $\vec{g}^T\vec{g}$ is constant and $(\mathbf{S}\vec{w})^T\vec{g} = \vec{g}^T\mathbf{S}\vec{w}$, we can simplify the problem:

$$\text{minimize} \quad -\vec{g}^T\mathbf{S}\vec{w} + \tfrac{1}{2}\vec{w}^T\mathbf{S}^T\mathbf{S}\vec{w} \qquad \text{subject to} \quad \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

- Classical quadratic programming problem!

Quadratic Programming Approach

(Following Lynch, F1000Research 5:1253, 2016)

We want $\mathbf{S}\vec{w} \approx \vec{g}$, so we can solve the following

Problem

$$\text{minimize} \quad (\vec{g} - \mathbf{S}\vec{w})^T(\vec{g} - \mathbf{S}\vec{w}) = \vec{g}^T\vec{g} - \vec{g}^T\mathbf{S}\vec{w} - (\mathbf{S}\vec{w})^T\vec{g} + (\mathbf{S}\vec{w})^T\mathbf{S}\vec{w}$$
$$\text{subject to} \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

Since $\vec{g}^T\vec{g}$ is constant and $(\mathbf{S}\vec{w})^T\vec{g} = \vec{g}^T\mathbf{S}\vec{w}$, we can simplify the problem:

$$\text{minimize} \quad -\vec{g}^T\mathbf{S}\vec{w} + \tfrac{1}{2}\vec{w}^T\mathbf{S}^T\mathbf{S}\vec{w} \qquad \text{subject to} \quad \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

- Classical quadratic programming problem!
- Can be easily solved using the R package `quadprog`

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways
- PathWave

Estimation Of Accuracy

Dataset A:

- 21 breast cancer genomes (Nik-Zainal et al, Cell 149:979–993, 2012)
- 4 signatures

Estimation Of Accuracy

Dataset A:

- 21 breast cancer genomes (Nik-Zainal et al, Cell 149:979–993, 2012)
- 4 signatures

Dataset B:

- 435 tumor genomes from 10 tumor types (Alexandrov et al, Nature 500:415–421, 2013)
- 15 signatures

Estimation Of Accuracy

Dataset A:

- 21 breast cancer genomes (Nik-Zainal et al, Cell 149:979–993, 2012)
- 4 signatures

Dataset B:

- 435 tumor genomes from 10 tumor types (Alexandrov et al, Nature 500:415–421, 2013)
- 15 signatures

Procedure (simplified):

Estimation Of Accuracy

Dataset A:

- 21 breast cancer genomes (Nik-Zainal et al, Cell 149:979–993, 2012)
- 4 signatures

Dataset B:

- 435 tumor genomes from 10 tumor types (Alexandrov et al, Nature 500:415–421, 2013)
- 15 signatures

Procedure (simplified):

1. Compute “true” weights w_k using **all** tumors (`pmsignature`).

Estimation Of Accuracy

Dataset A:

- 21 breast cancer genomes (Nik-Zainal et al, Cell 149:979–993, 2012)
- 4 signatures

Dataset B:

- 435 tumor genomes from 10 tumor types (Alexandrov et al, Nature 500:415–421, 2013)
- 15 signatures

Procedure (simplified):

1. Compute “true” weights w_k using **all** tumors (`pmsignature`).
2. Leave out a single tumor (A) or 10% (B), compute signatures based on the remaining tumors (`pmsignature`).

Estimation Of Accuracy

Dataset A:

- 21 breast cancer genomes (Nik-Zainal et al, Cell 149:979–993, 2012)
- 4 signatures

Dataset B:

- 435 tumor genomes from 10 tumor types (Alexandrov et al, Nature 500:415–421, 2013)
- 15 signatures

Procedure (simplified):

1. Compute “true” weights w_k using **all** tumors (`pmsignature`).
2. Leave out a single tumor (A) or 10% (B), compute signatures based on the remaining tumors (`pmsignature`).
3. Predict weights w'_k for each *single* left out tumor (`decompTumor2Sig`) using the signatures from the reduced dataset (step 2).
   - Note: this reflects a realistic case, e.g., in a clinical setting

Estimation Of Accuracy

Dataset A:

- 21 breast cancer genomes (Nik-Zainal et al, Cell 149:979–993, 2012)
- 4 signatures

Dataset B:

- 435 tumor genomes from 10 tumor types (Alexandrov et al, Nature 500:415–421, 2013)
- 15 signatures

Procedure (simplified):

1. Compute “true” weights w_k using **all** tumors (`pmsignature`).
2. Leave out a single tumor (A) or 10% (B), compute signatures based on the remaining tumors (`pmsignature`).
3. Predict weights w'_k for each *single* left out tumor (`decompTumor2Sig`) using the signatures from the reduced dataset (step 2).
   - Note: this reflects a realistic case, e.g., in a clinical setting
4. Compare predicted weights (step 3) to “true” weights (step 1).

Results

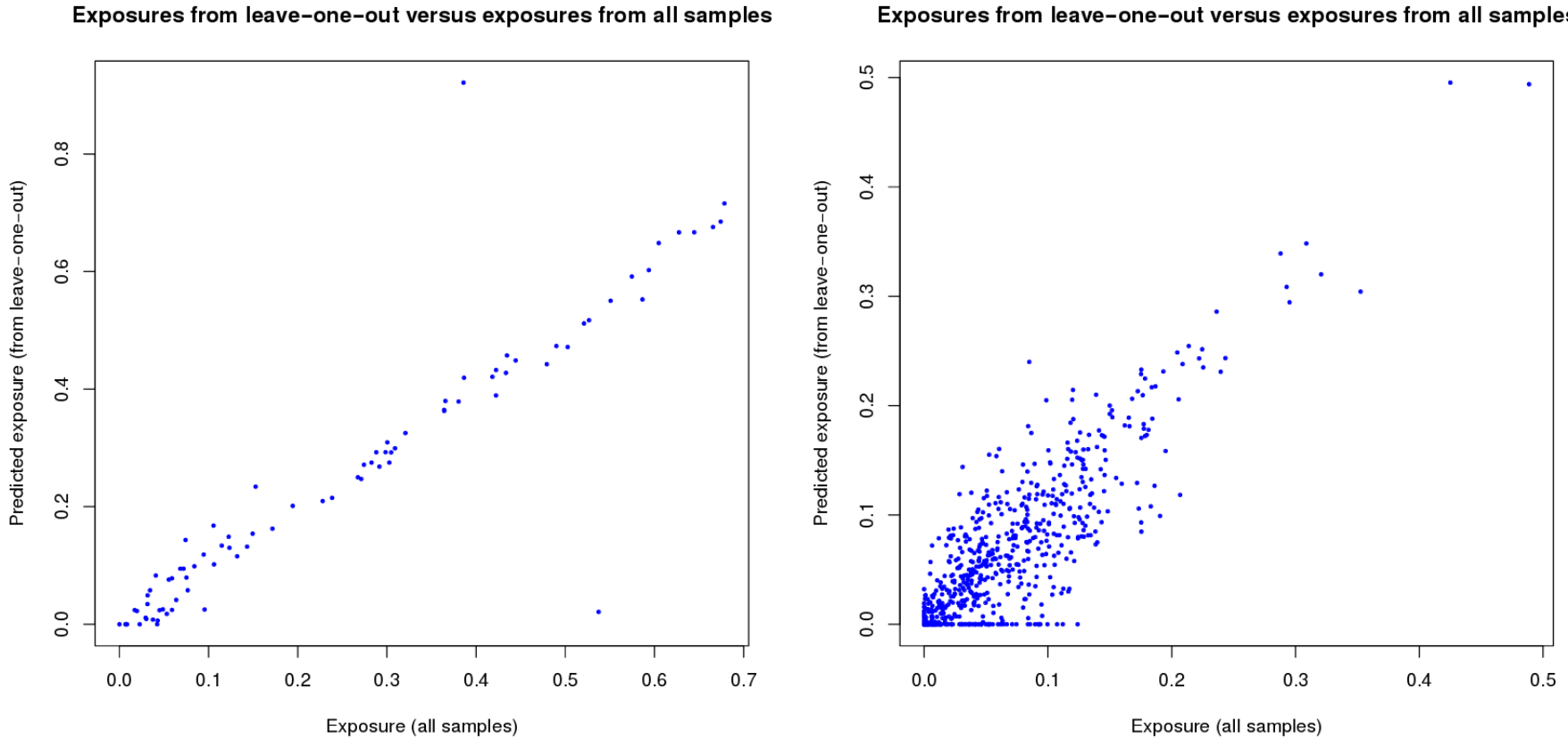


Figure 2: Comparison of contributions/weights ("exposures") predicted for individual tumors (`decompTumor2Sig`; y-axis) and collectively computed (`pmsignature`; x-axis). Left: leave-one-out test on 21 breast cancers ($r = 0.923$). Right: test set of 44 out of 435 tumors ($r = 0.807$).

Median weight differences ($|w_k - w'_k|$): (A) 0.018 and (B) 0.019

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

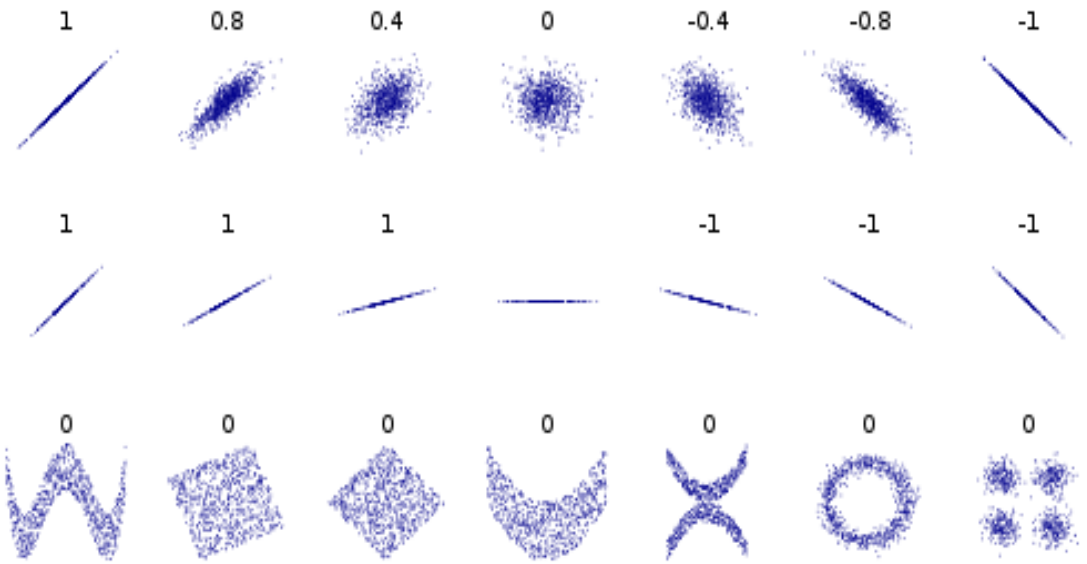
Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

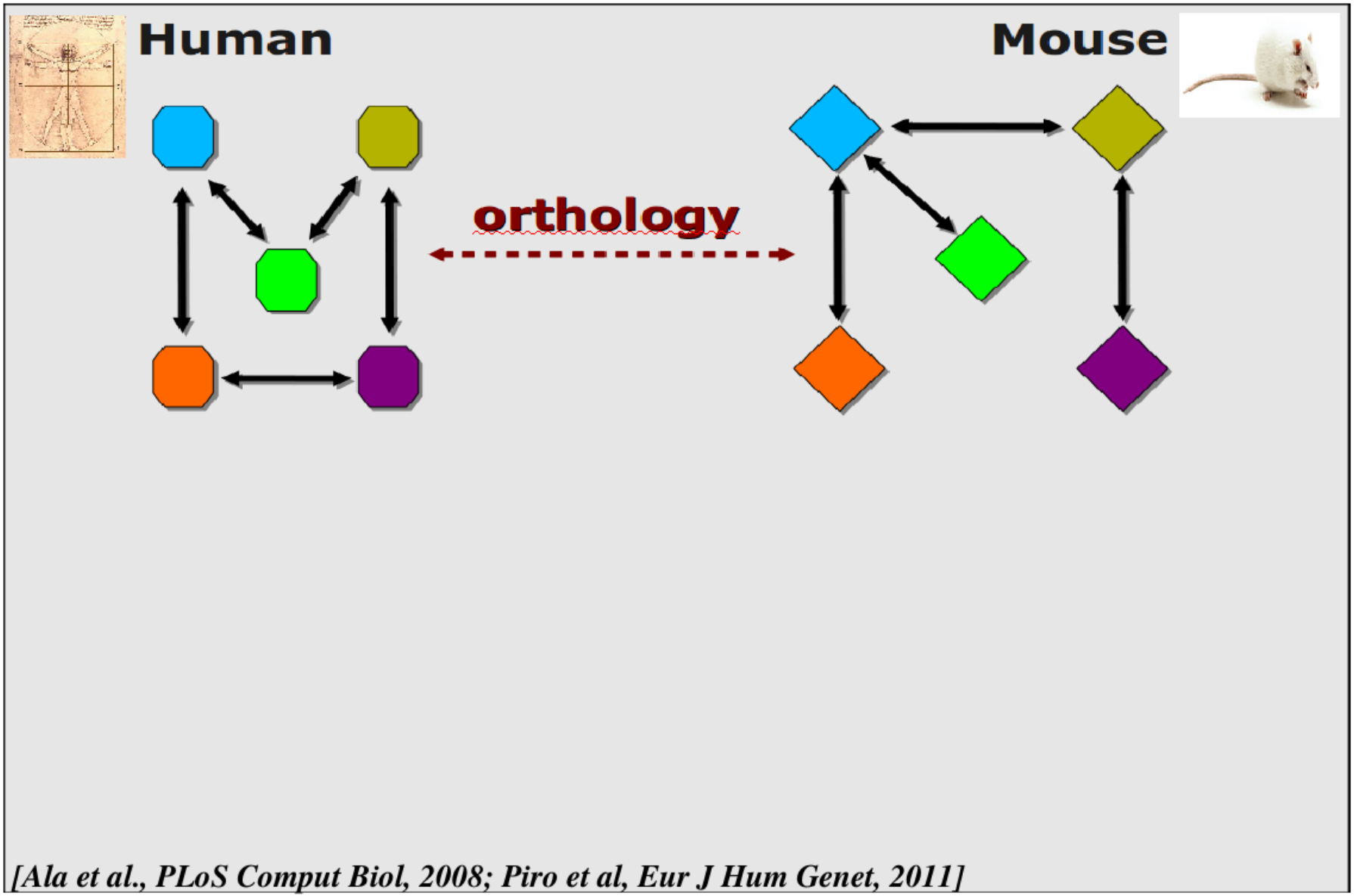
Identifying Regulatory Shifts in Biological Pathways
- PathWave

Pearson correlation coefficient

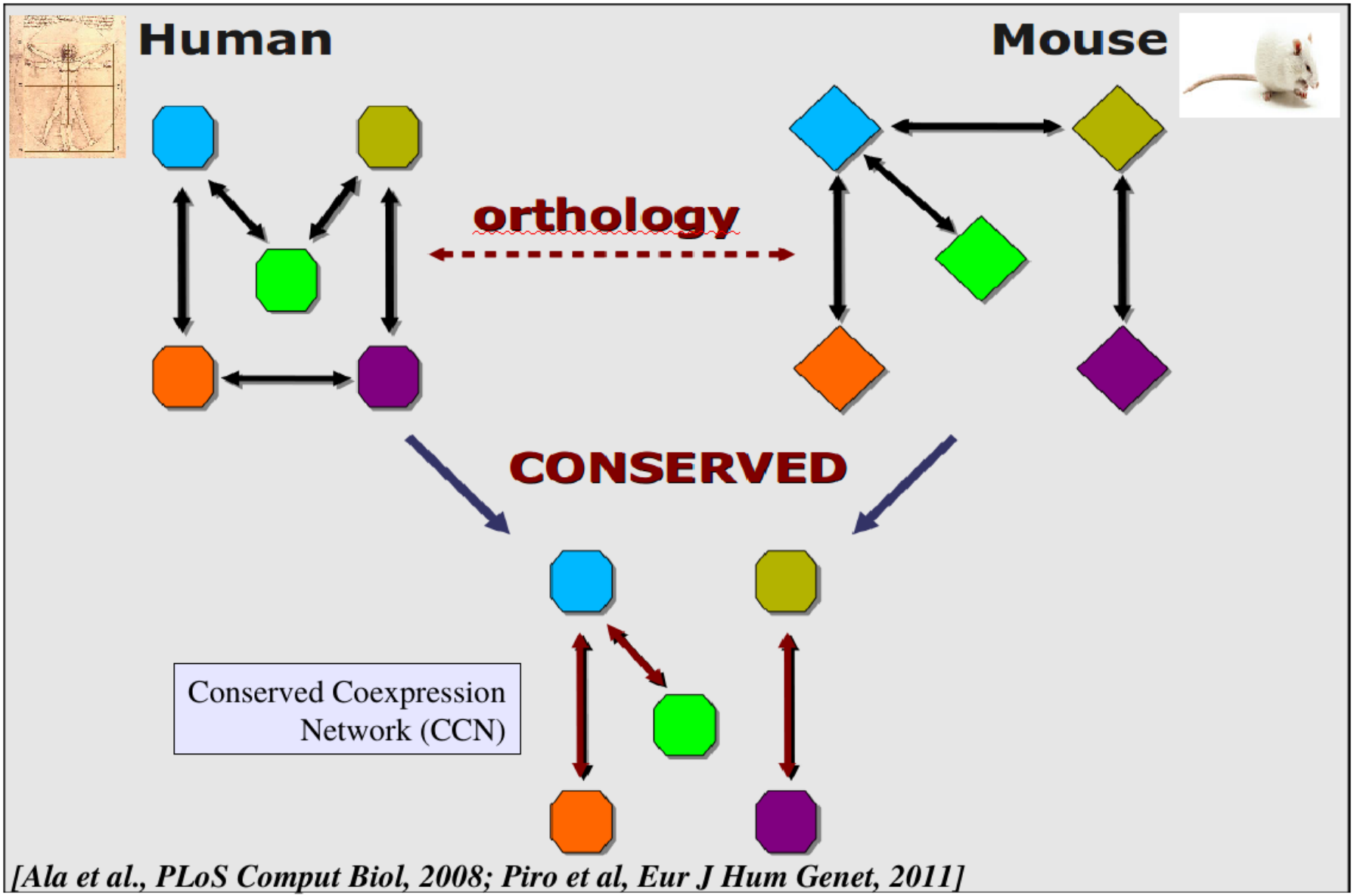
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$



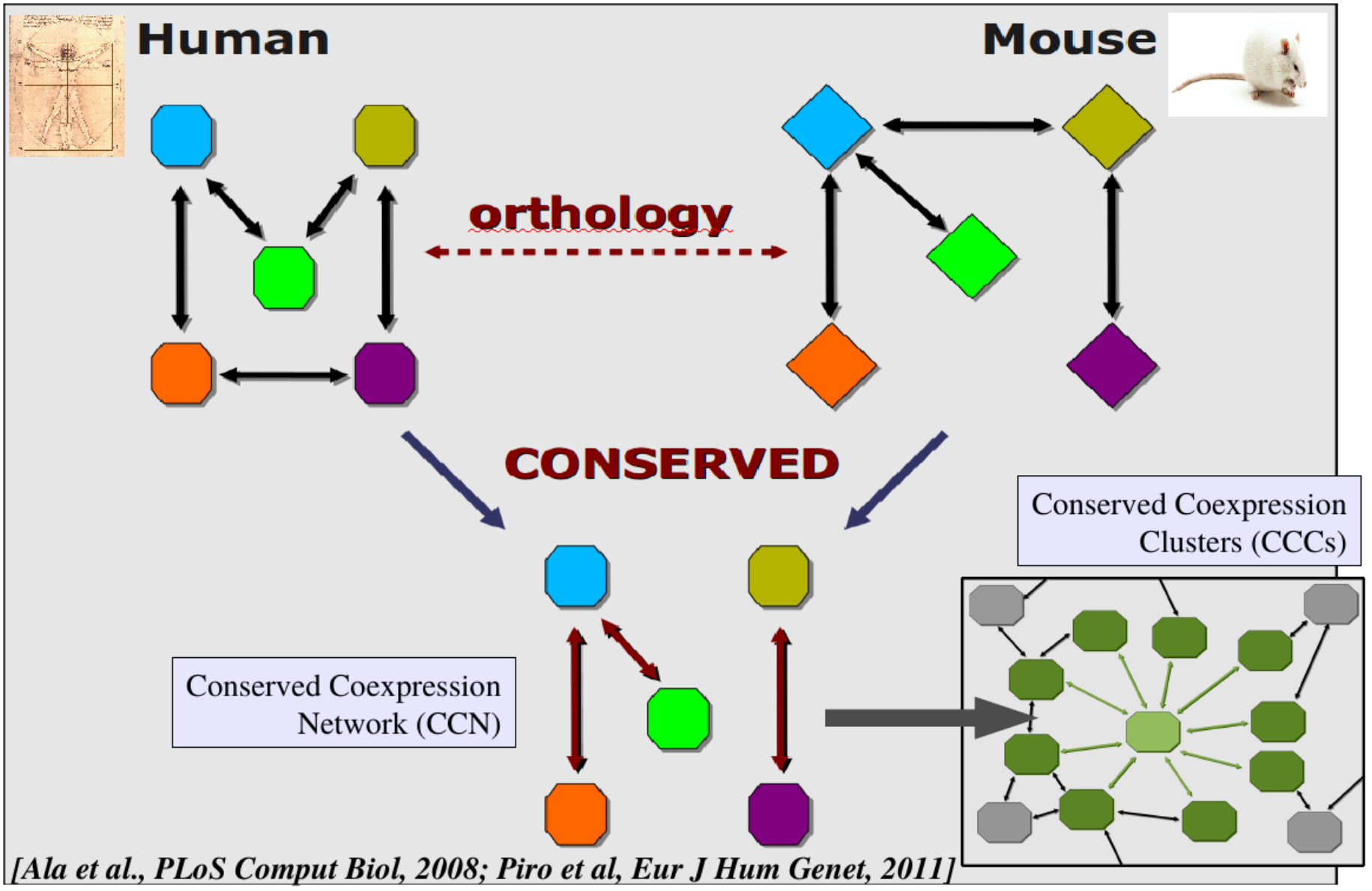
Conserved coexpression-based DGP



Conserved coexpression-based DGP



Conserved coexpression-based DGP



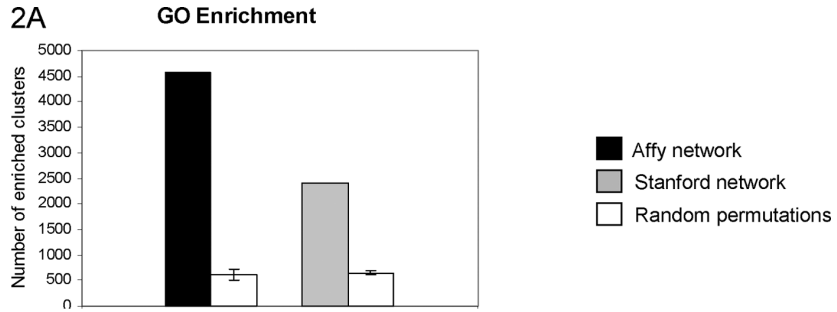
Conserved coexpression-based DGP

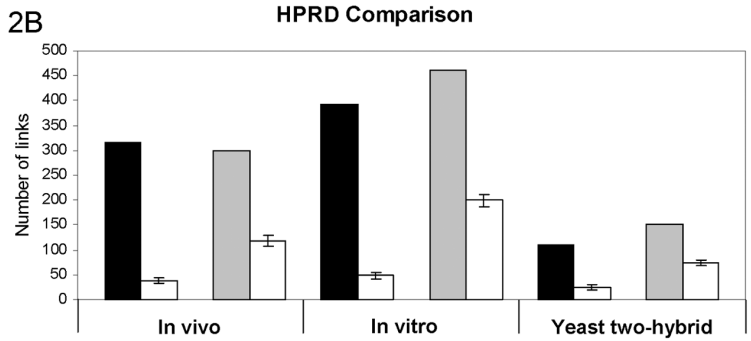
Does conserved coexpression reflect biologically meaningful information?

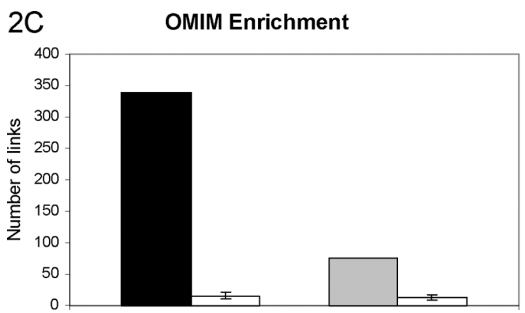
Conserved coexpression-based DGP

Does conserved coexpression reflect biologically meaningful information?

- Clusters are enriched in
  - Gene Ontology annotations
  - Known protein–protein interactions





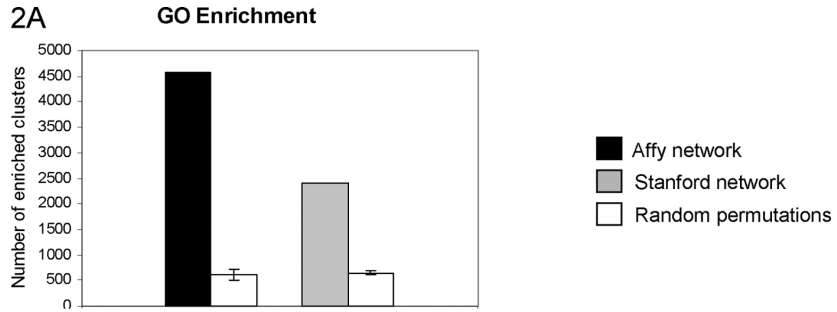


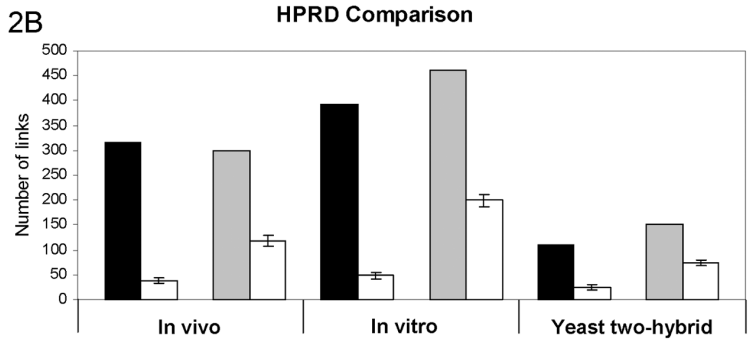
(Ala et al., 2008)

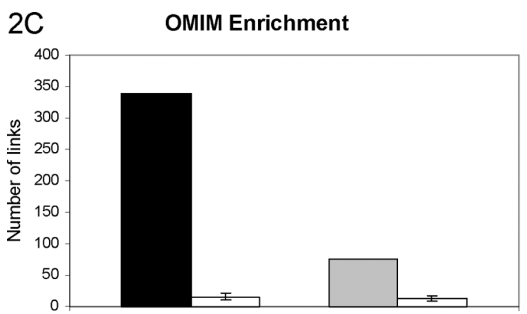
Conserved coexpression-based DGP

Does conserved coexpression reflect biologically meaningful information?

- Clusters are enriched in
  - Gene Ontology annotations
  - Known protein–protein interactions
- Network edges are enriched in
  - Gene pairs involved in similar disease phenotypes (similarity according to MimMiner; van Driel at al., 2006)

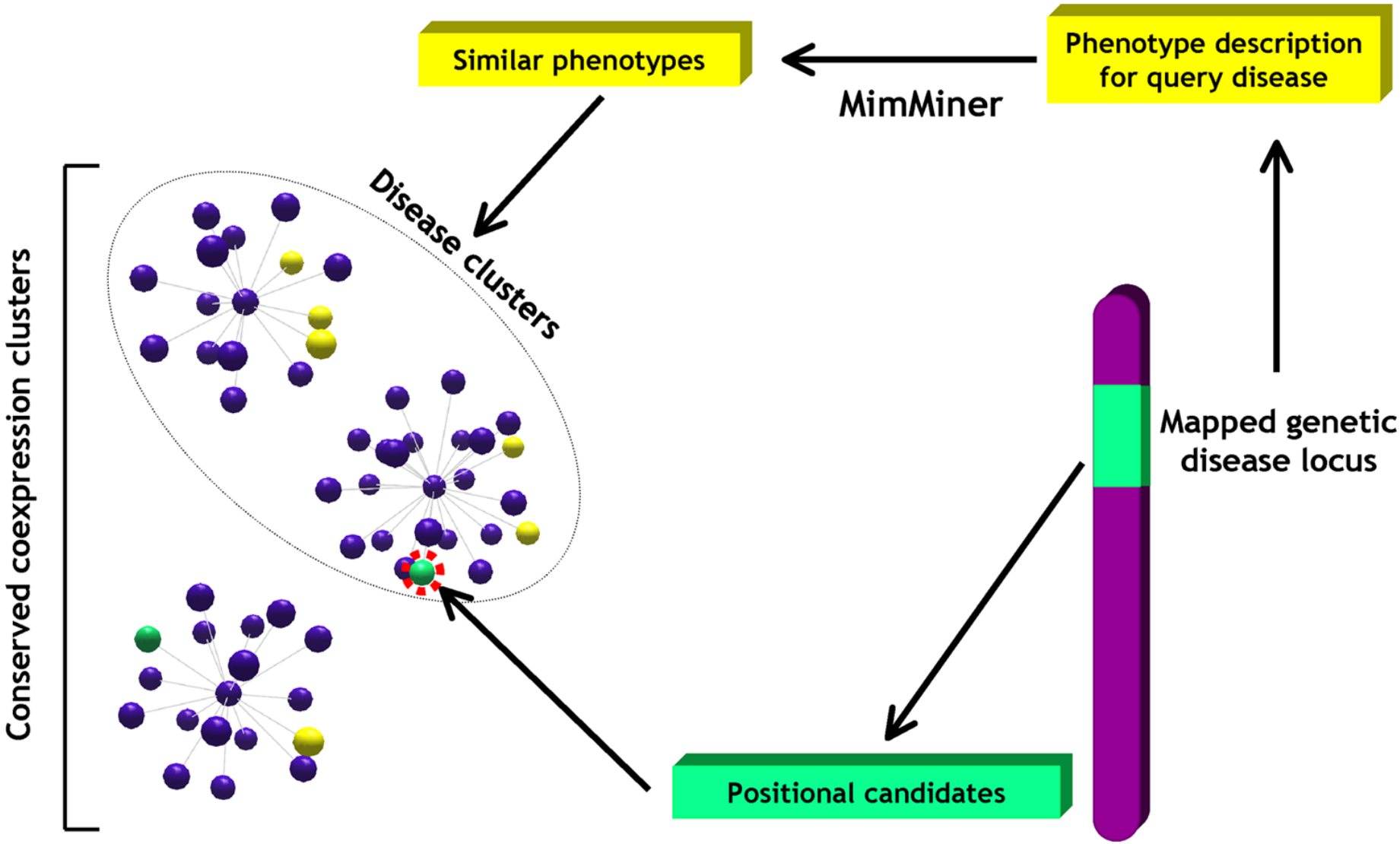






(Ala et al., 2008)

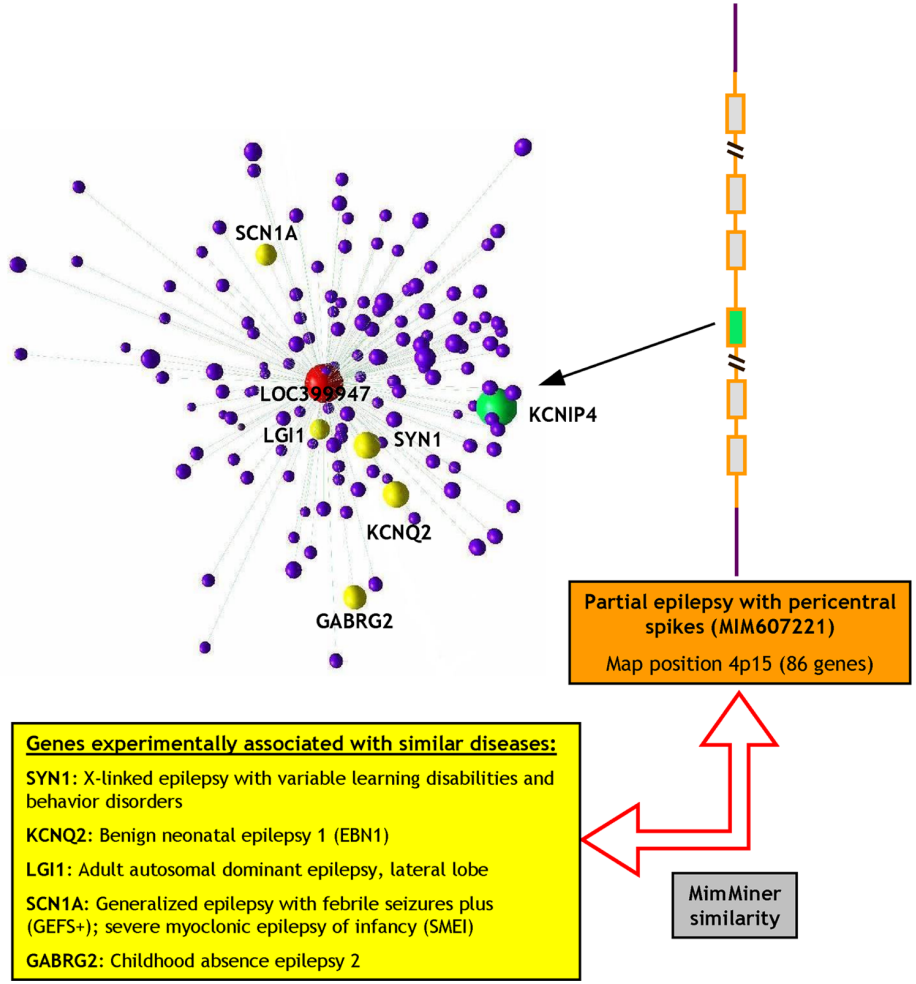
Conserved coexpression-based DGP



(Ala et al., 2008)

Conserved coexpression-based DGP

Example prediction for:
Epilepsy, partial, with pericentral spikes (OMIM %607221)

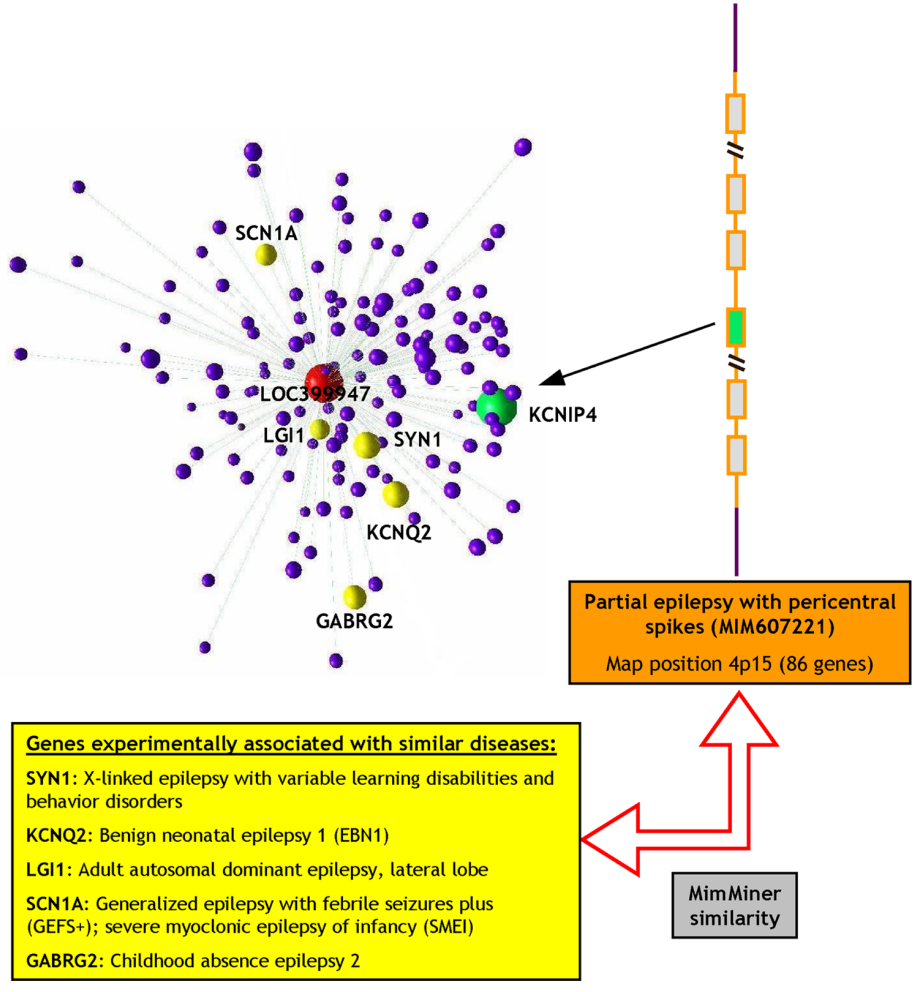


(Ala et al., 2008)

Conserved coexpression-based DGP

Example prediction for:
Epilepsy, partial, with pericentral spikes (OMIM %607221)

- Orphan locus with 86 positional candidates on 4p15

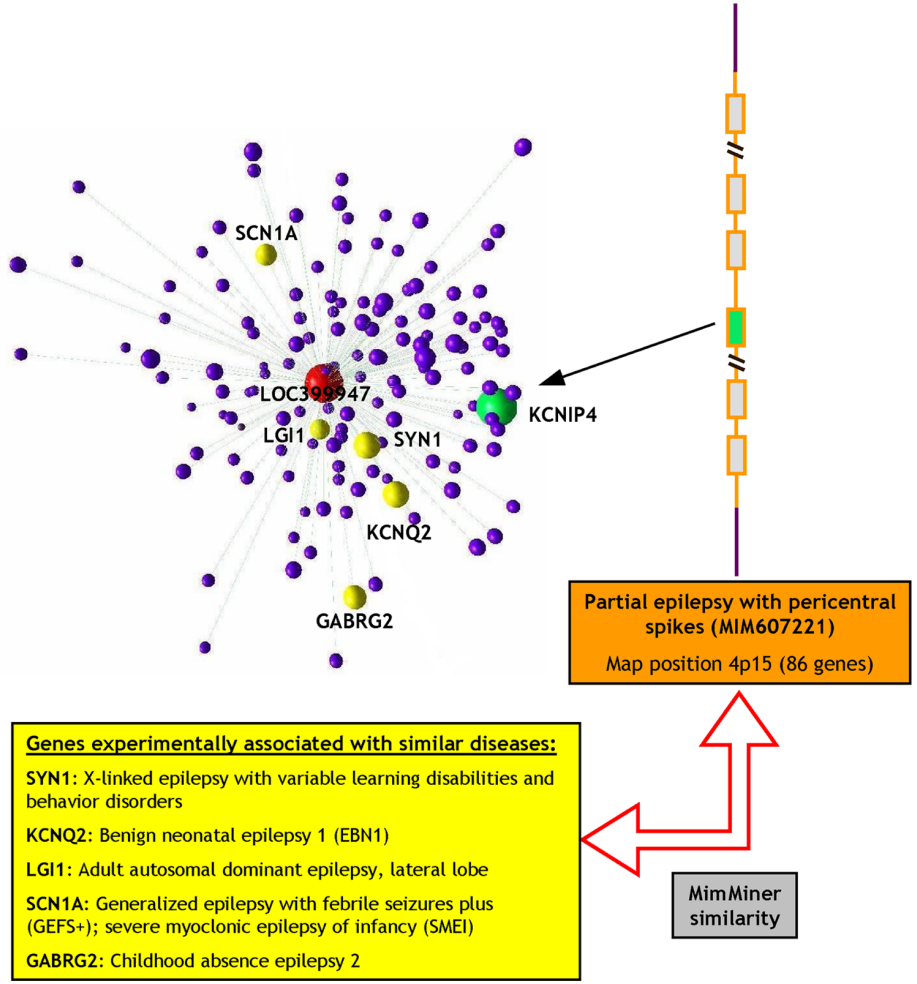


(Ala et al., 2008)

Conserved coexpression-based DGP

Example prediction for:
Epilepsy, partial, with pericentral spikes (OMIM %607221)

- Orphan locus with 86 positional candidates on 4p15
- Candidate (1 of 2): *KCNIP4*; modulates the activity of Kv4 A-type potassium channels that have been implicated in epilepsy



(Ala et al., 2008)

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

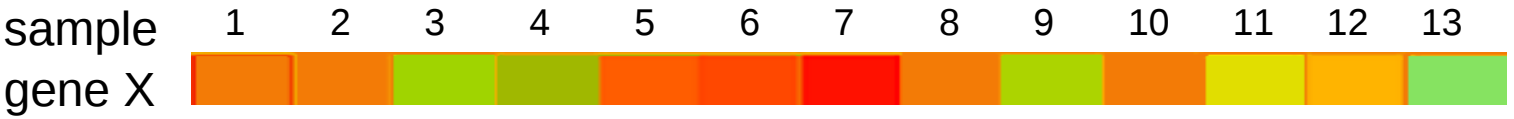
Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways
- PathWave

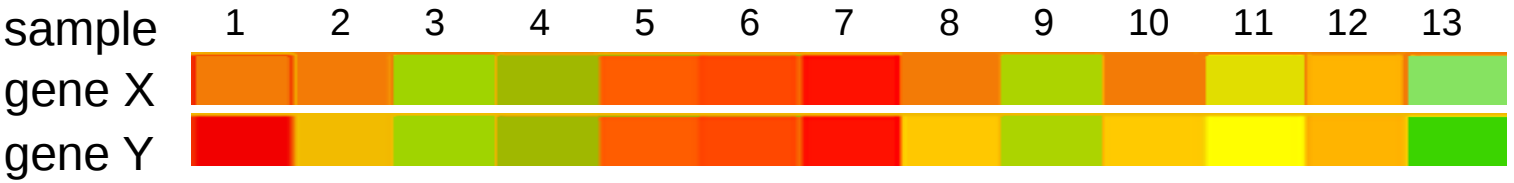
Spatial (3D) gene expression profiles

“Traditional” high-throughput expression data



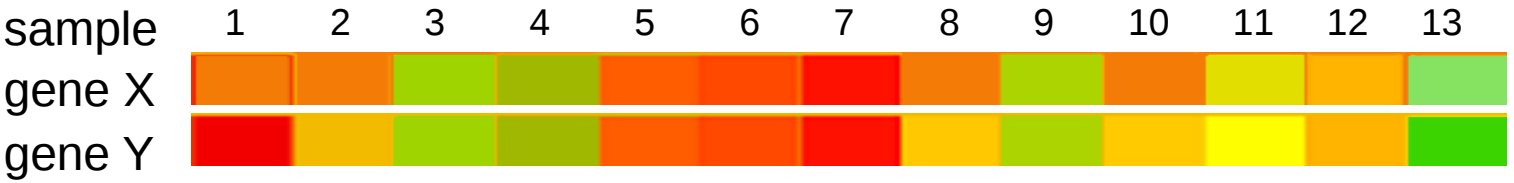
Spatial (3D) gene expression profiles

“Traditional” high-throughput expression data



Spatial (3D) gene expression profiles

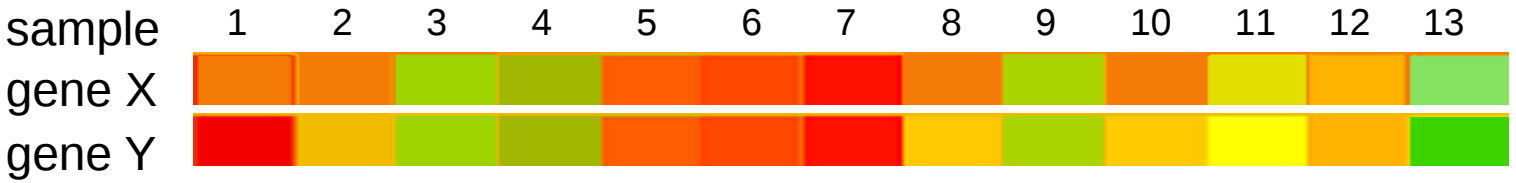
“Traditional” high-throughput expression data



- 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)

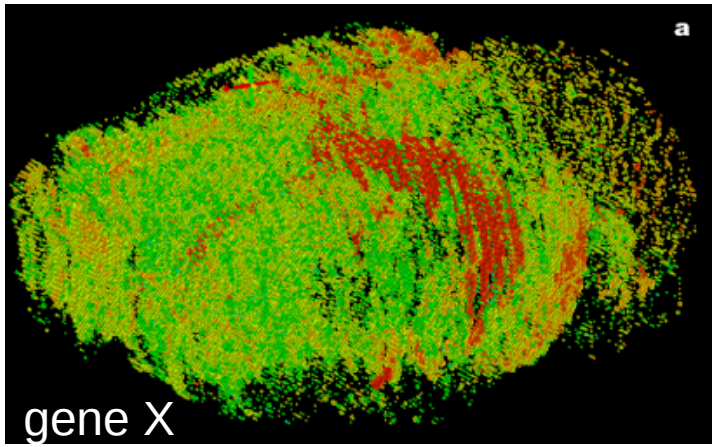
Spatial (3D) gene expression profiles

“Traditional” high-throughput expression data



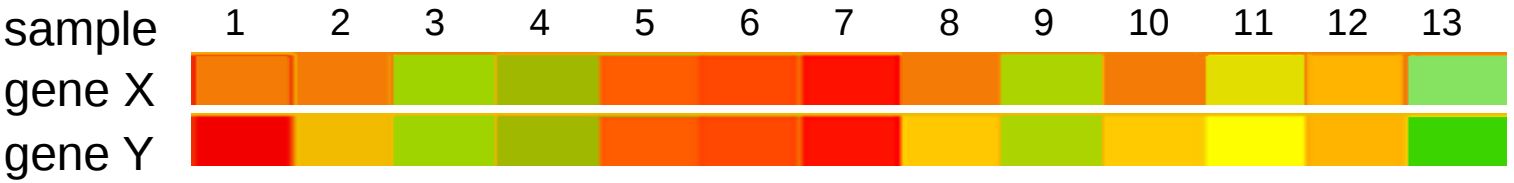
- carry **limited spatial information**, no precise 3D localization
- have a **sparse coverage** (arbitrary, not-consecutive positions)
- problem for tissues/organs with a high degree of spatial organization, e.g. the central nervous system (CNS)

© Allen Institute



Spatial (3D) gene expression profiles

“Traditional” high-throughput expression data

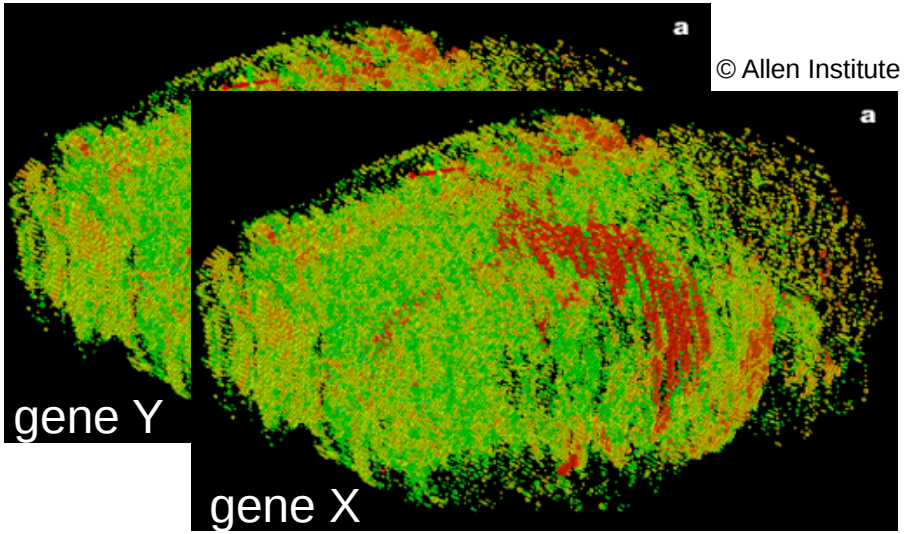


- carry **limited spatial information**, no precise 3D localization
- have a **sparse coverage** (arbitrary, not-consecutive positions)
- problem for tissues/organs with a high degree of spatial organization, e.g. the central nervous system (CNS)

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

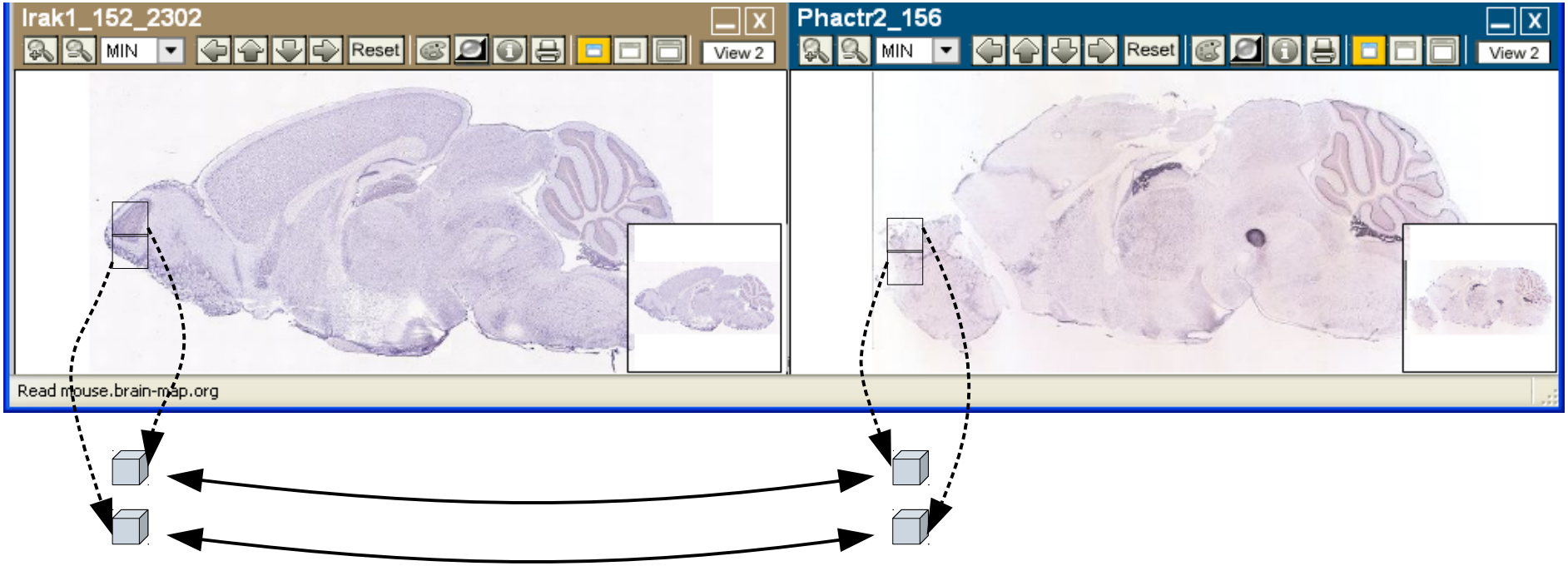


© Allen Institute

Measuring 3D co-expression

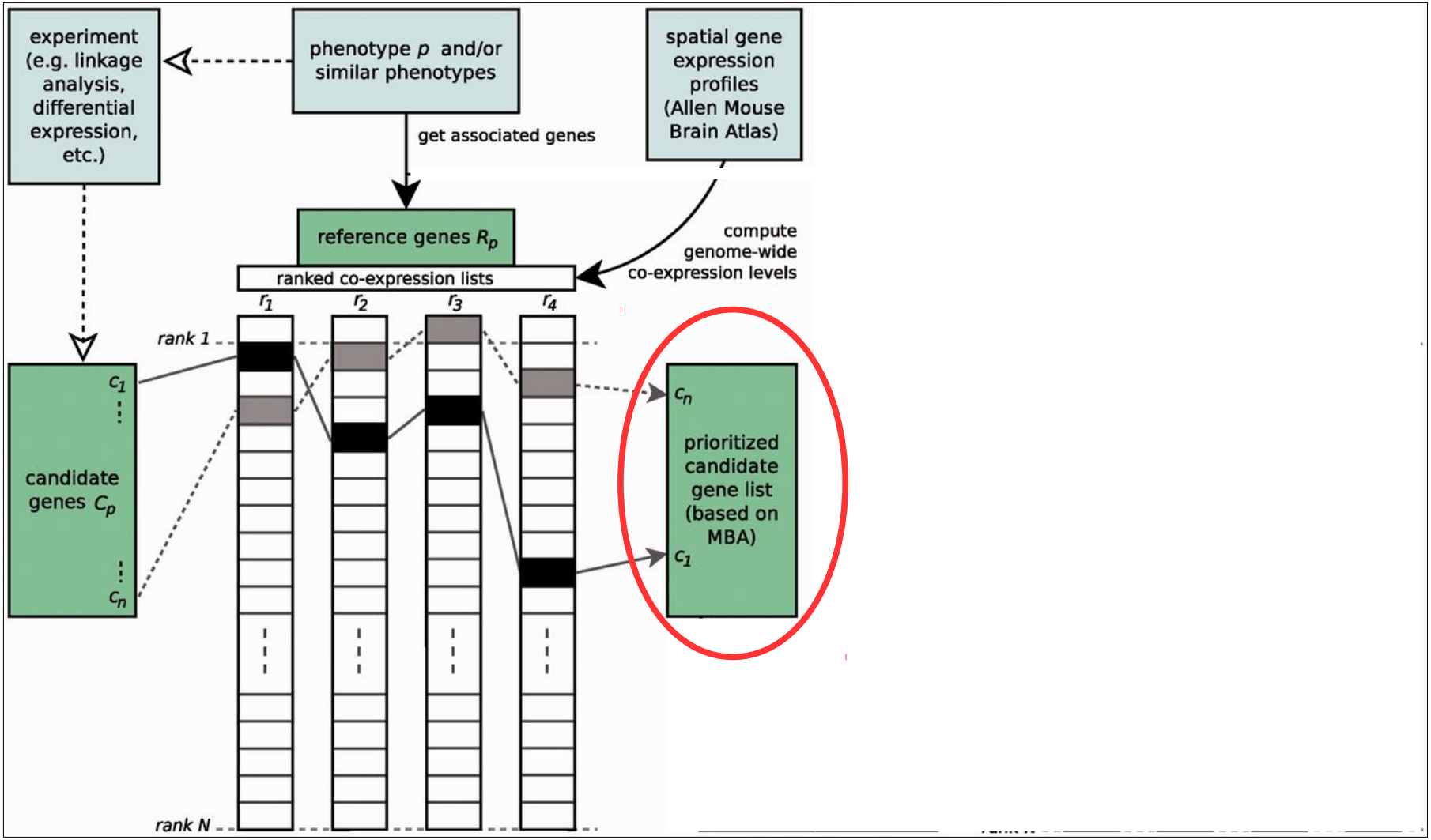
Use of numerical 3D expression patterns

- 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)

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]

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

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]

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]

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**

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

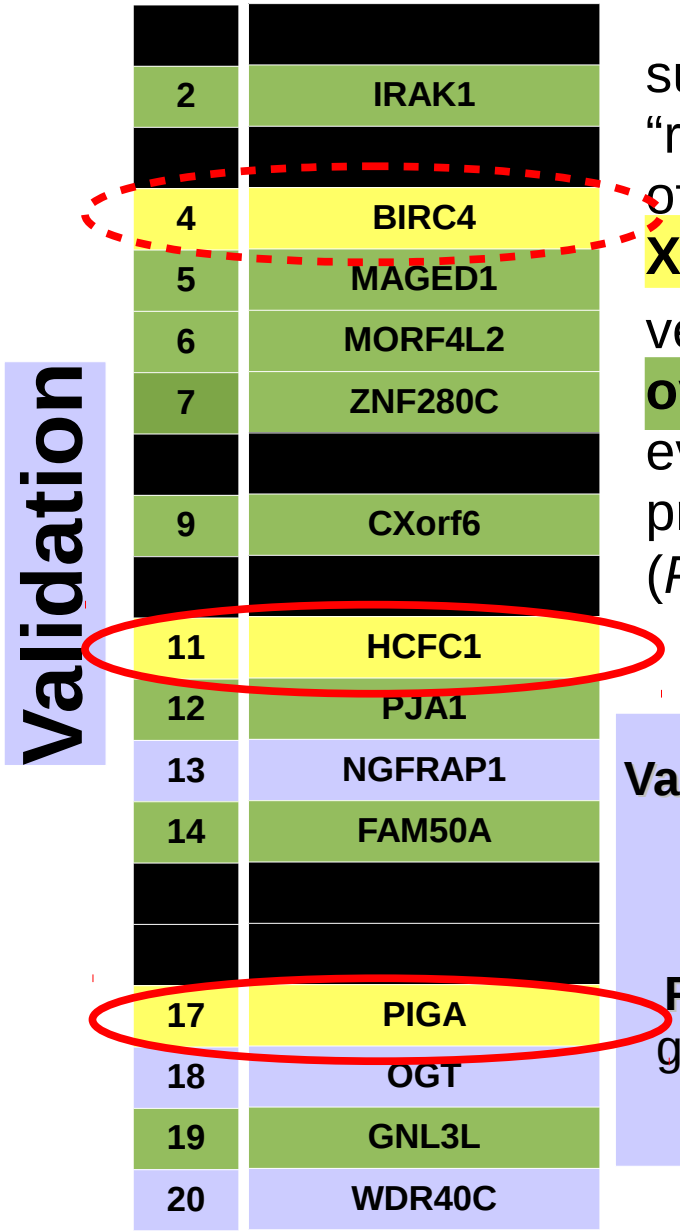
Validation: autosomal MR genes as reference genes

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]

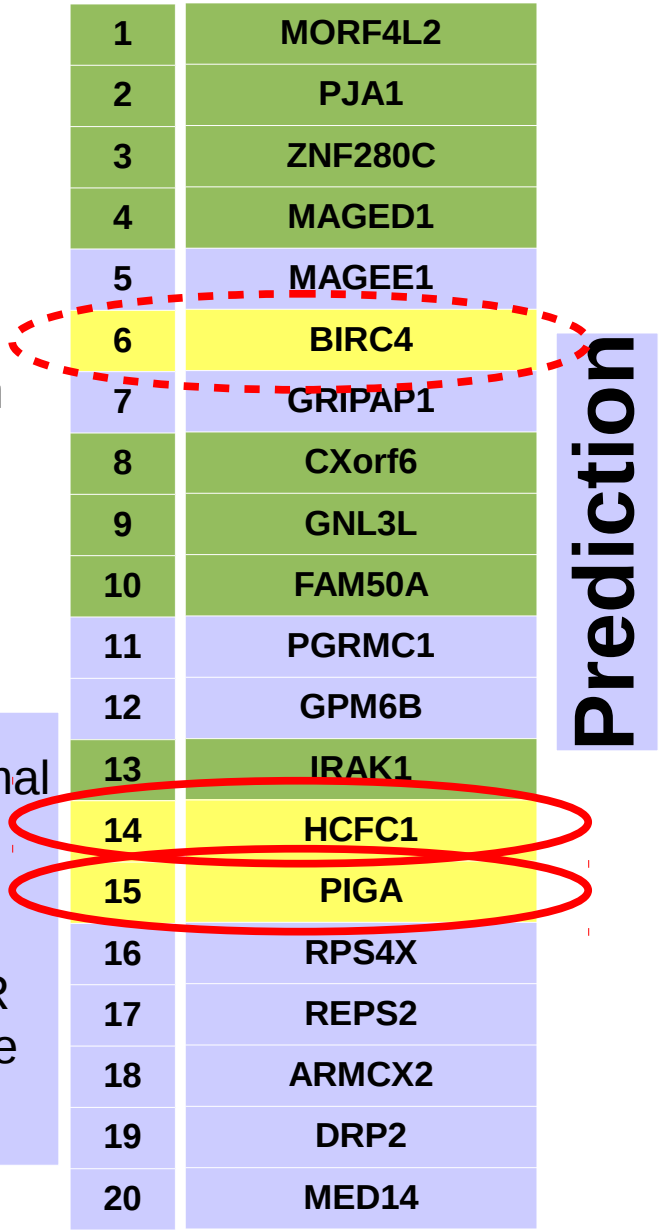


successful "rediscovery" of known **XLMR genes**

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

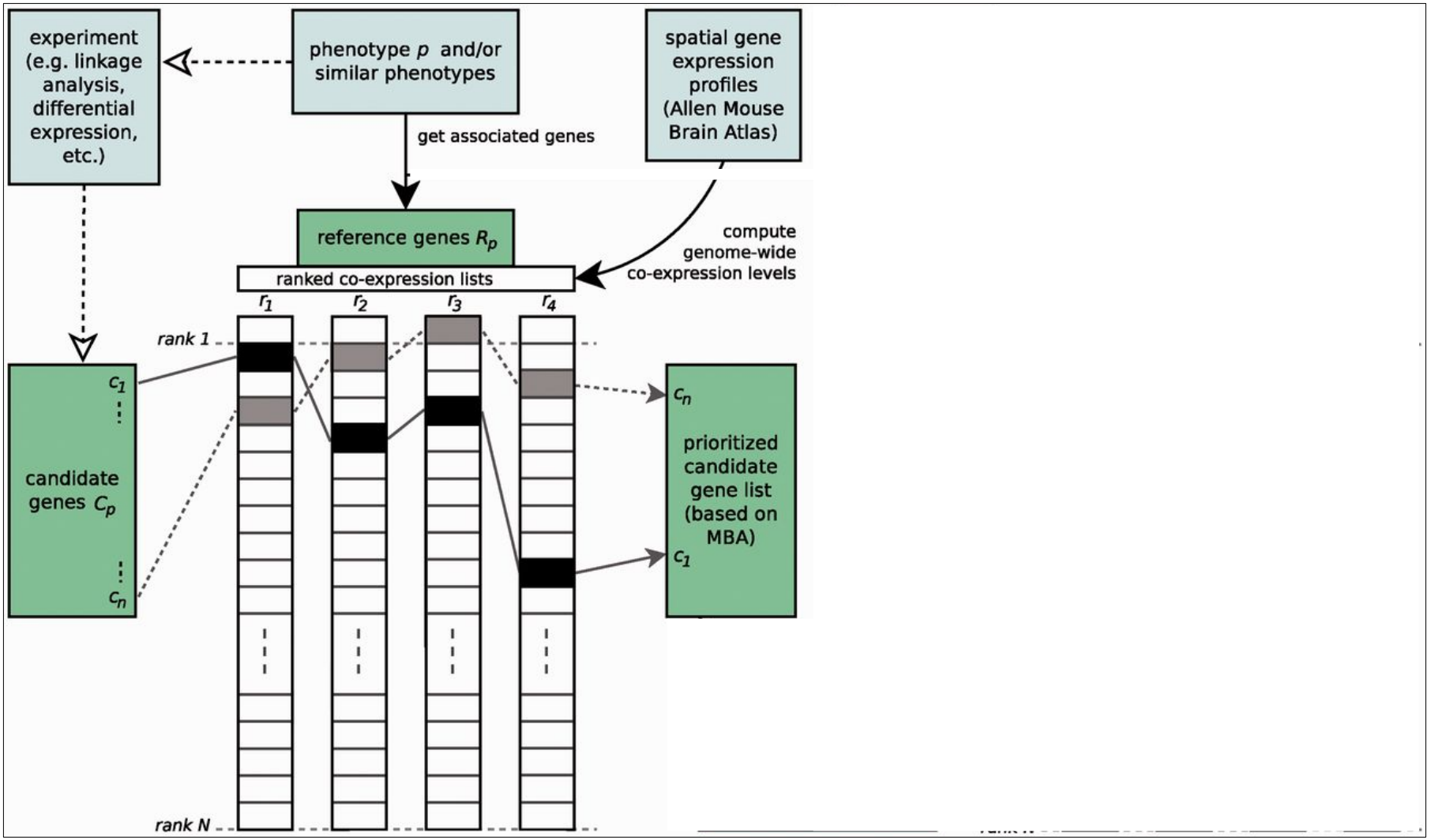
Validation: autosomal MR genes as reference genes

Prediction: XLMR genes as reference genes



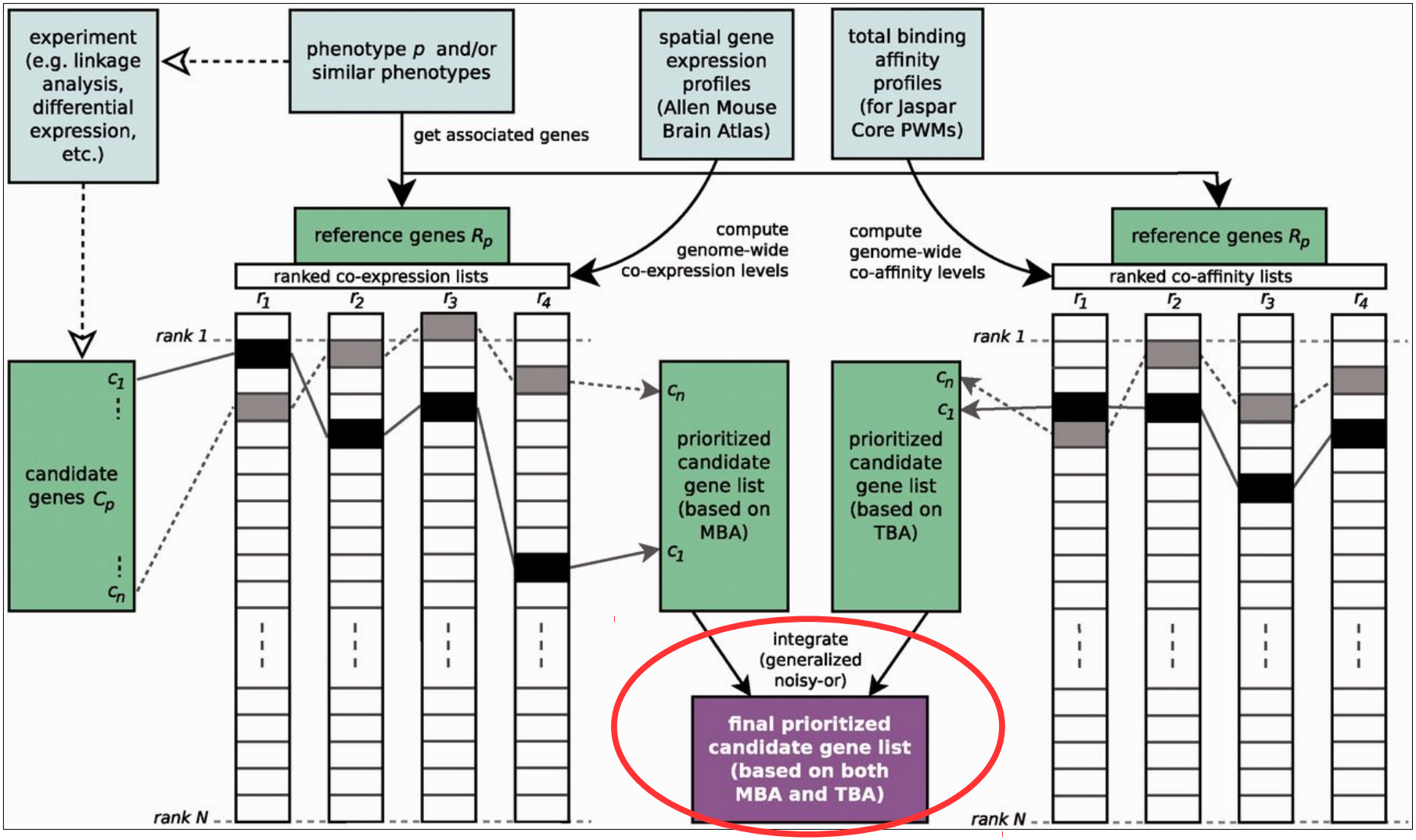
[Piro et al., Bioinformatics, 2010]

Method for disease gene prioritization



[Piro et al., Bioinformatics, 2013]

Method for disease gene prioritization



[Piro et al., Bioinformatics, 2013]

Total Binding Affinity (TBA) – motivation

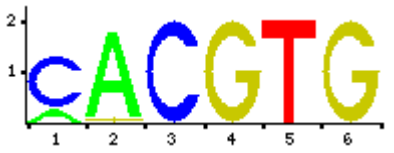
Goal: condition-independent description of regulatory dependencies of promoters to a core set of transcription factors (TFs)

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

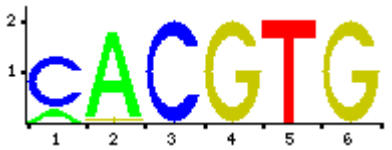


- "Traditional" approach: identification of discrete binding sites for specific TFs within the promoter
  - advantage: allows focused hypotheses and can be experimentally verified

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) 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) – 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:

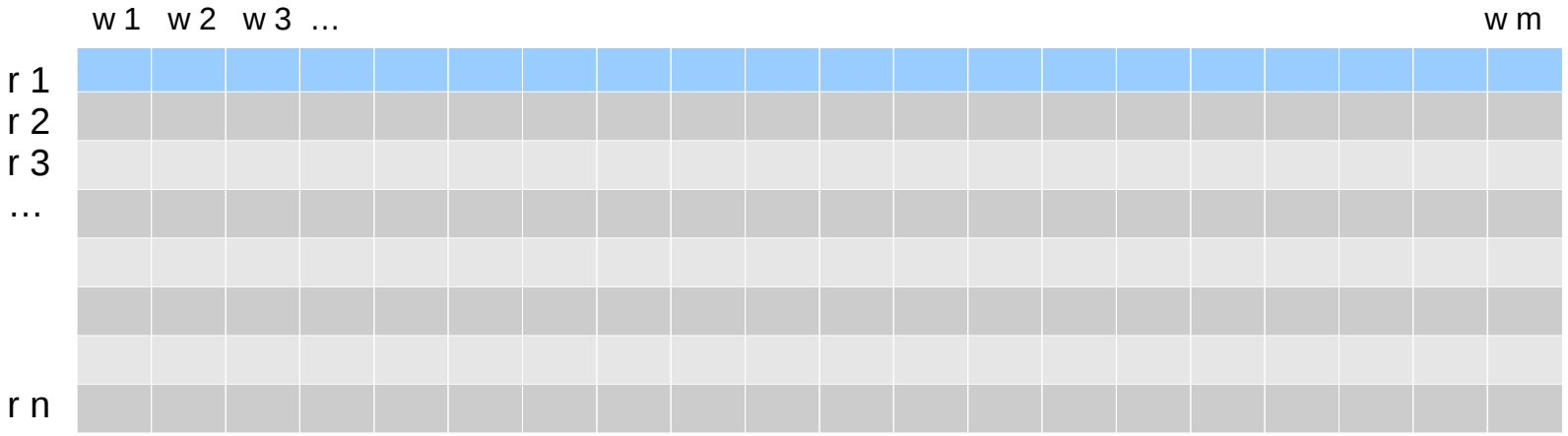
- RefSeq track of the UCSC Genome Browser
- from 1500 bp upstream to 500 bp downstream of the TSS

PFMs / TFs:

- JASPAR Core (130 matrices)

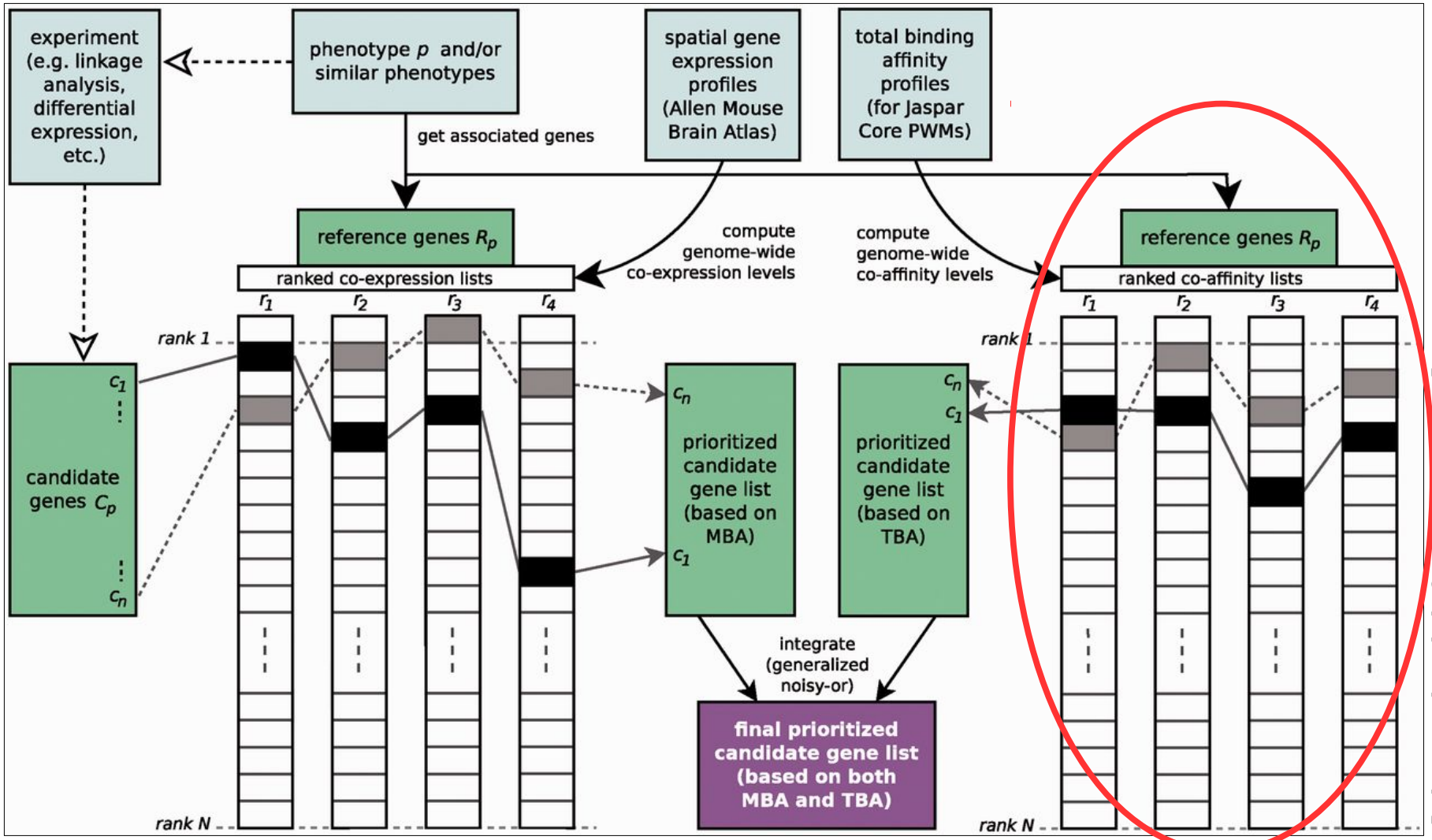
TBAs are computed for all promoters *r* and PFMs *w*

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



[Piro et al., Bioinformatics, 2013]

Method for disease gene prioritization



[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):

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

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

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

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

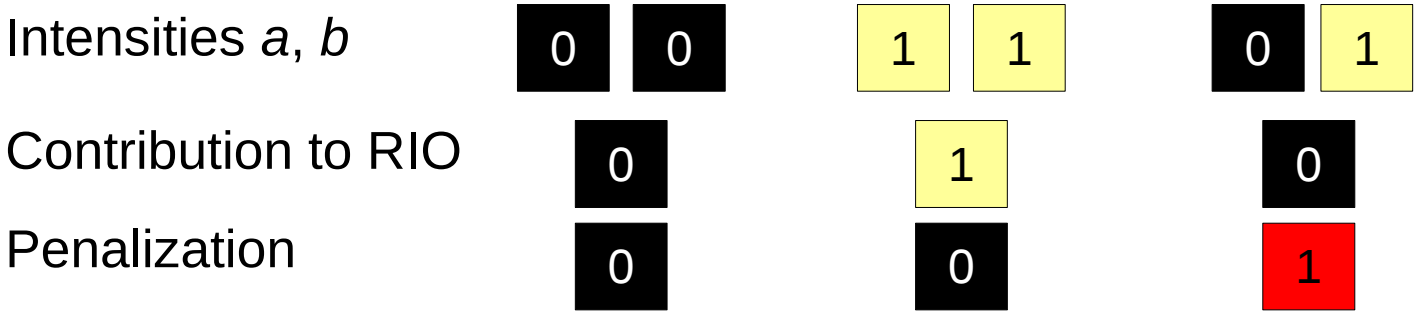
Normalization of expression levels (“intensities”) for each “3D image”

- ranging from 0 to 1

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]

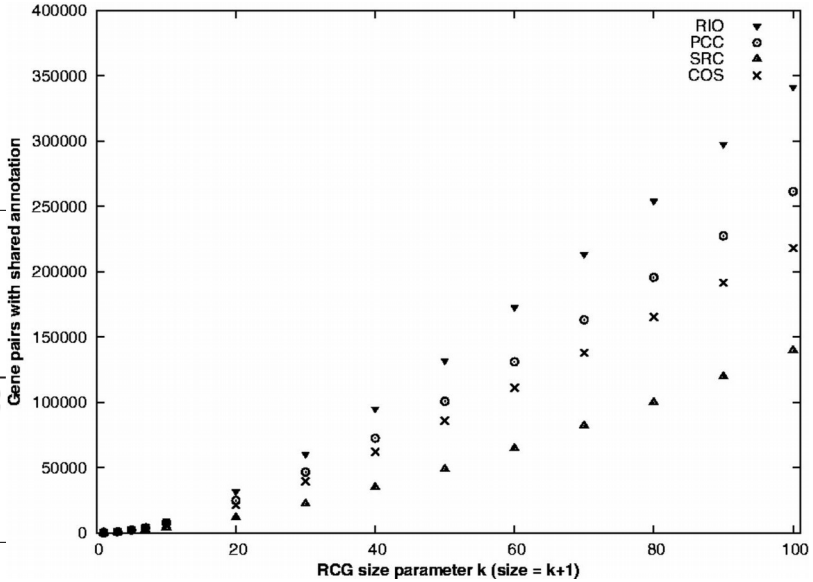
Similarity measure for 3D expression profiles

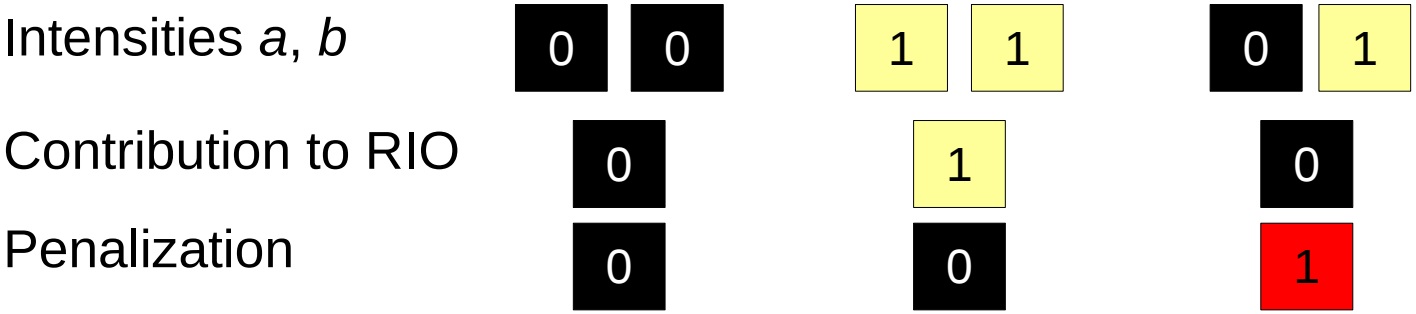
Normalization of expression levels (“intensities”) for each “3D image”

- ranging from 0 to 1

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]

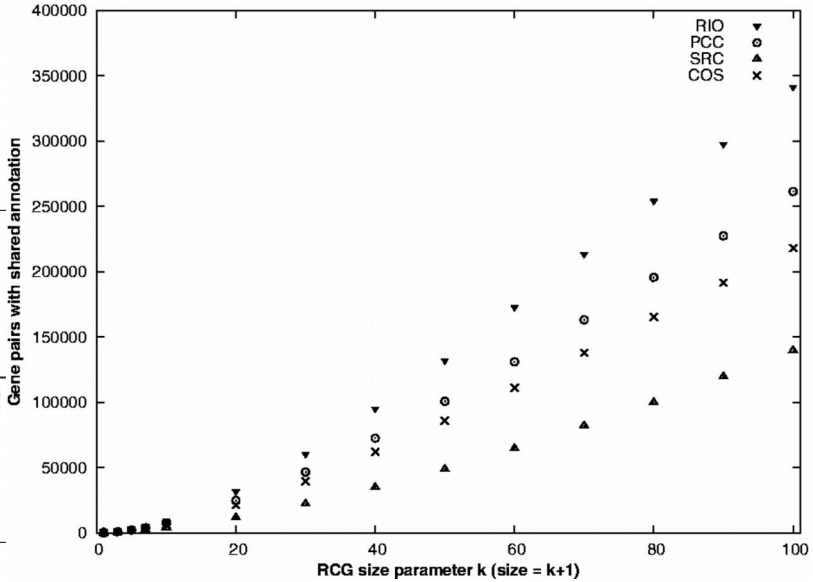
Similarity measure for TBA profiles

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

- ranging from 0 to 1

Relative Intensity Overlap (RIO):

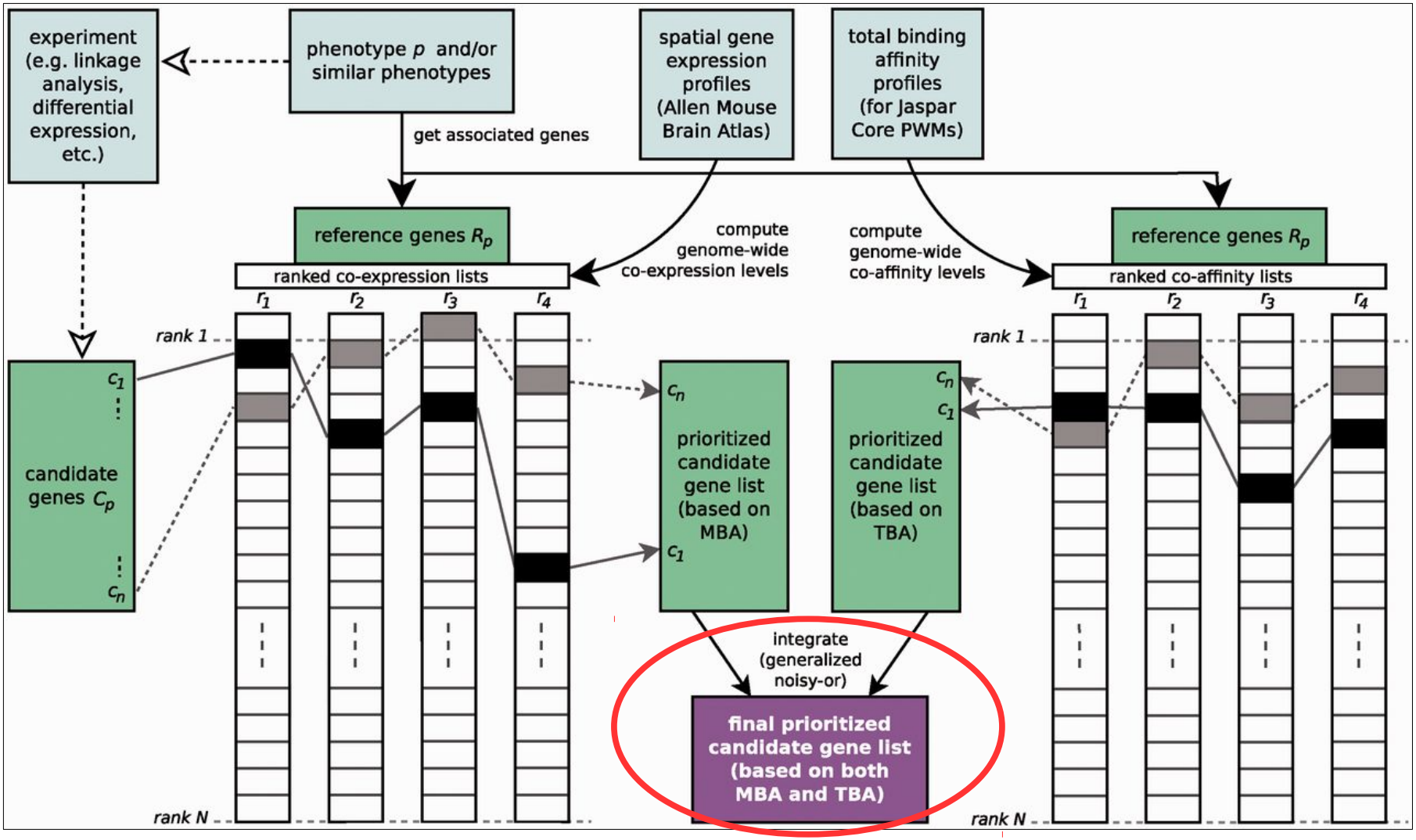
$$\mathrm{RIO}(a,b) = \frac{\sum_{xyz} \left(I^a_{xyz} \times I^b_{xyz}\right)}{\sum_{xyz} \left(max(|I^a_{xyz}|, |I^b_{xyz}|)\right)^2}$$



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

[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*.

Outline

Some Molecular Biology
- DNA → Transcription → RNA (→ Translation → Protein)
- Biological Pathways
- Mutations And Genetic Diseases

Identifying Mutations in DNA-Sequencing Data
- Read Alignment And Variant Calling
- Some results for somatic mutations

Identifying Mutational Signatures In Individual Tumors
- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures
- Mutational Signatures In Individual Tumors (`decompTumor2Sig`)
- Evaluation of `decompTumor2Sig`

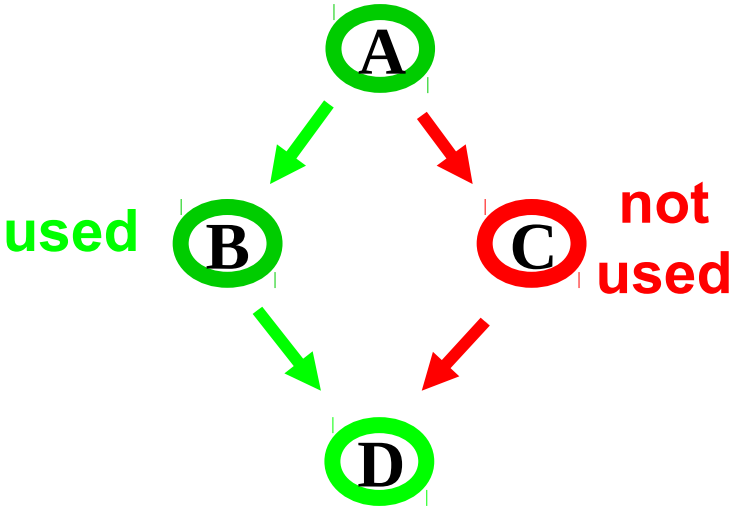
Disease-Gene Prediction (DGP)
- DGP With Conserved Co-Expression Networks
- DGP With 3D Gene Expression Profiles

Identifying Regulatory Shifts in Biological Pathways
- PathWave

PathWave *(Schramm et al., Bioinformatics, 2010)*

Identification of crucial players in cellular networks (metabolic pathways) for specifically targeting pathogenic cells

normal cell

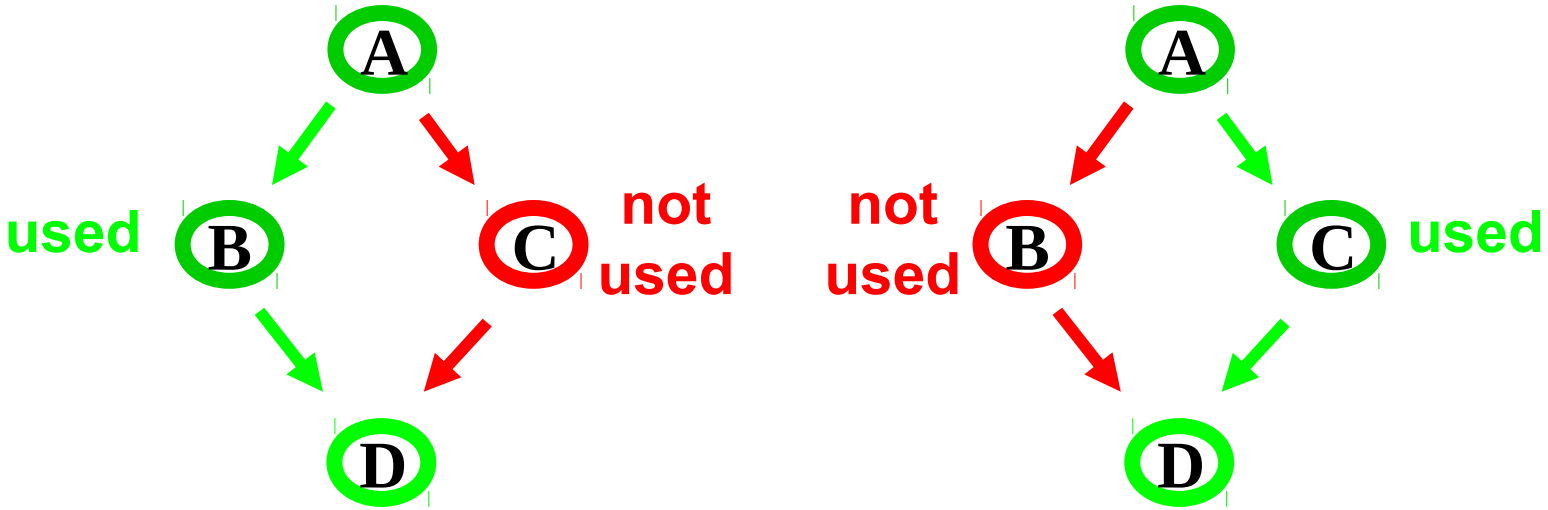


PathWave

(Schramm et al., Bioinformatics, 2010)

Identification of crucial players in cellular networks (metabolic pathways) for specifically targeting pathogenic cells

normal cell	pathogenic cell (tumor)

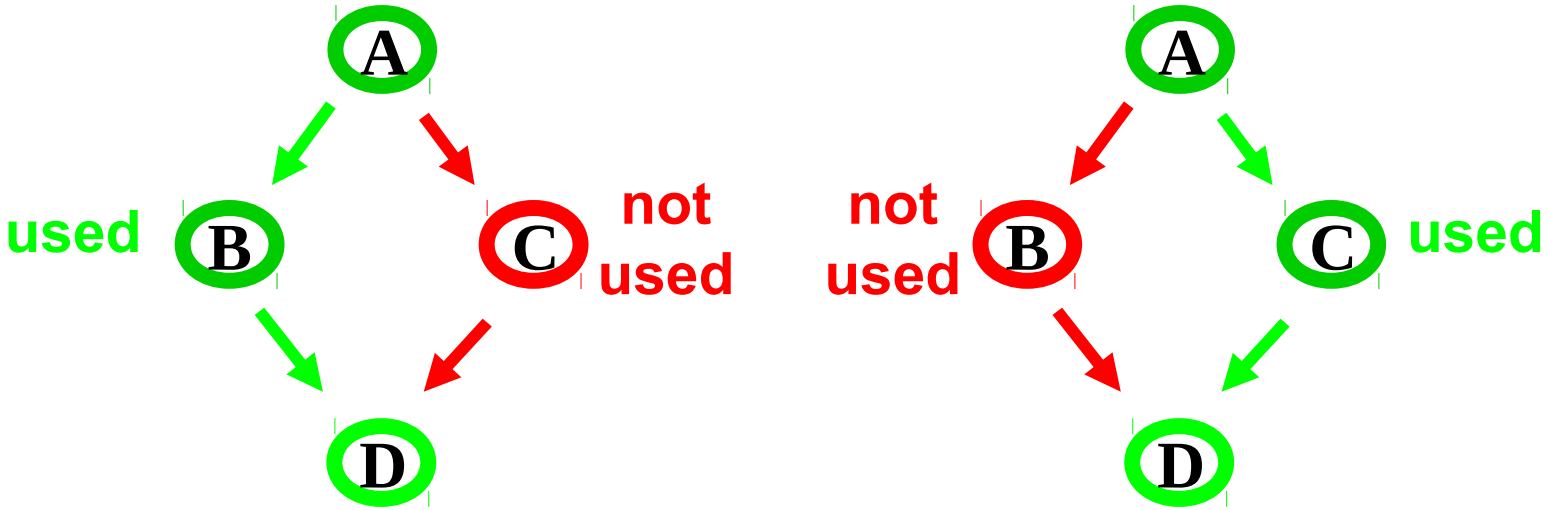


PathWave

(Schramm et al., Bioinformatics, 2010)

Identification of crucial players in cellular networks (metabolic pathways) for specifically targeting pathogenic cells

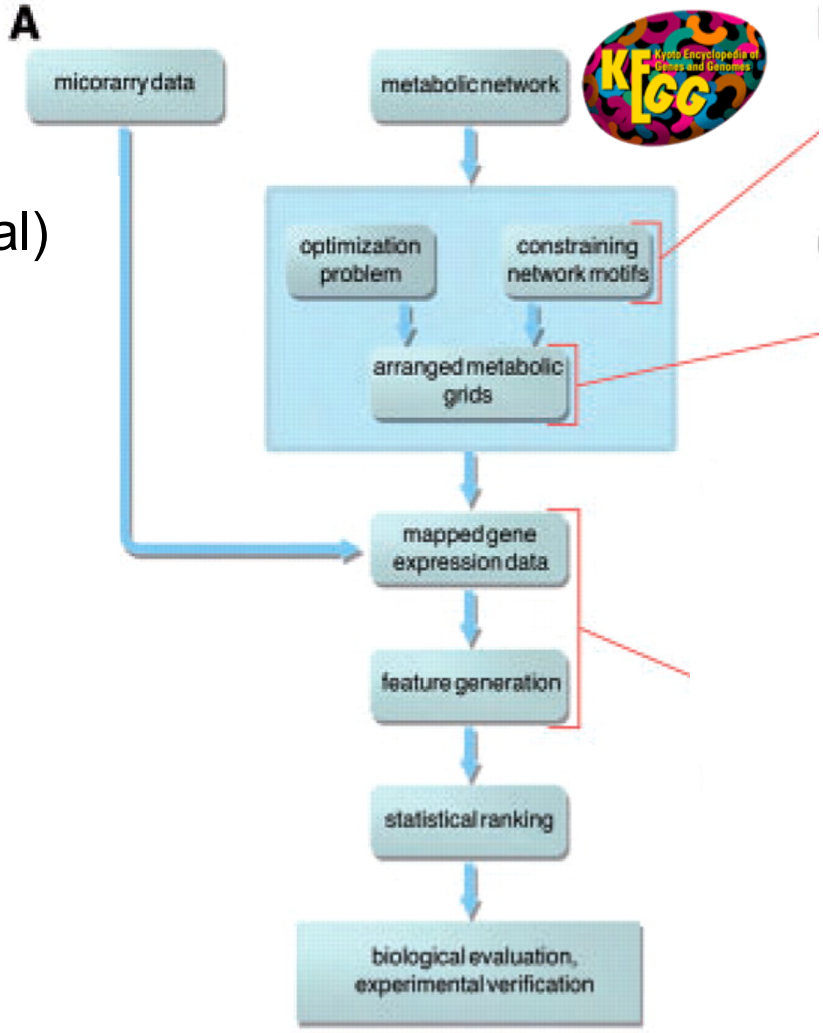
normal cell	pathogenic cell (tumor)



=> targeting C will specifically harm the pathogenic cell

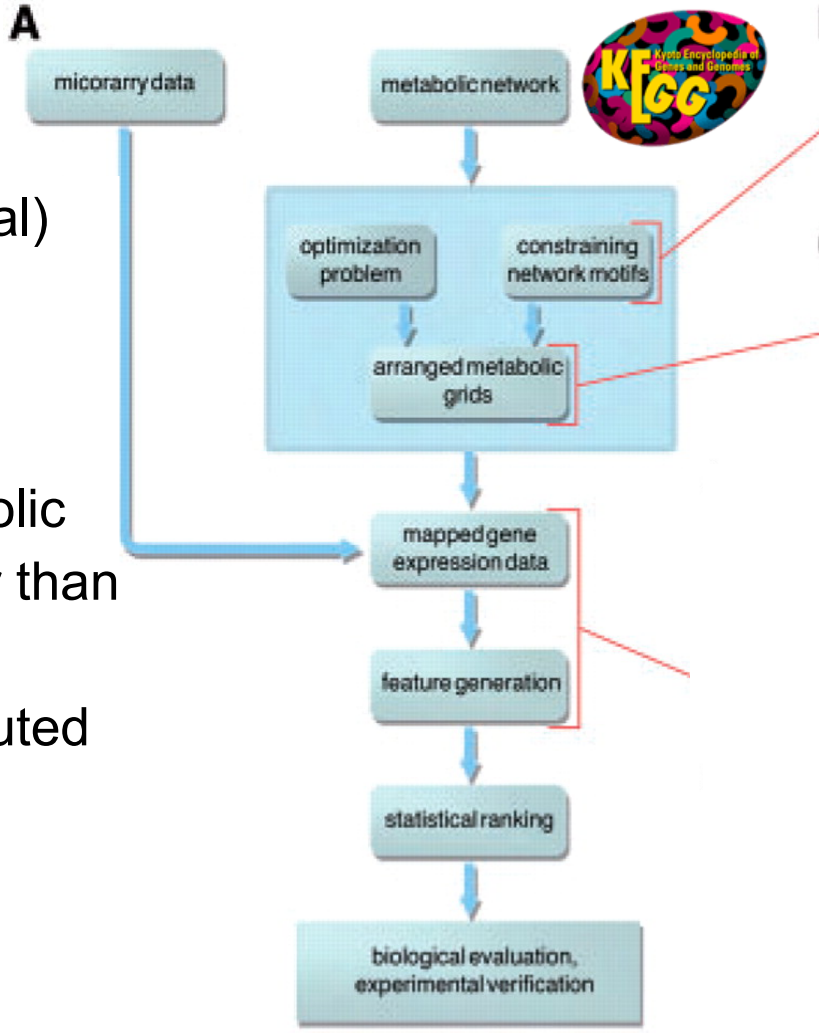
PathWave *(Schramm et al., Bioinformatics, 2010)*

- Identification of reg. patterns: pathways (global) and metabolic switches (local)



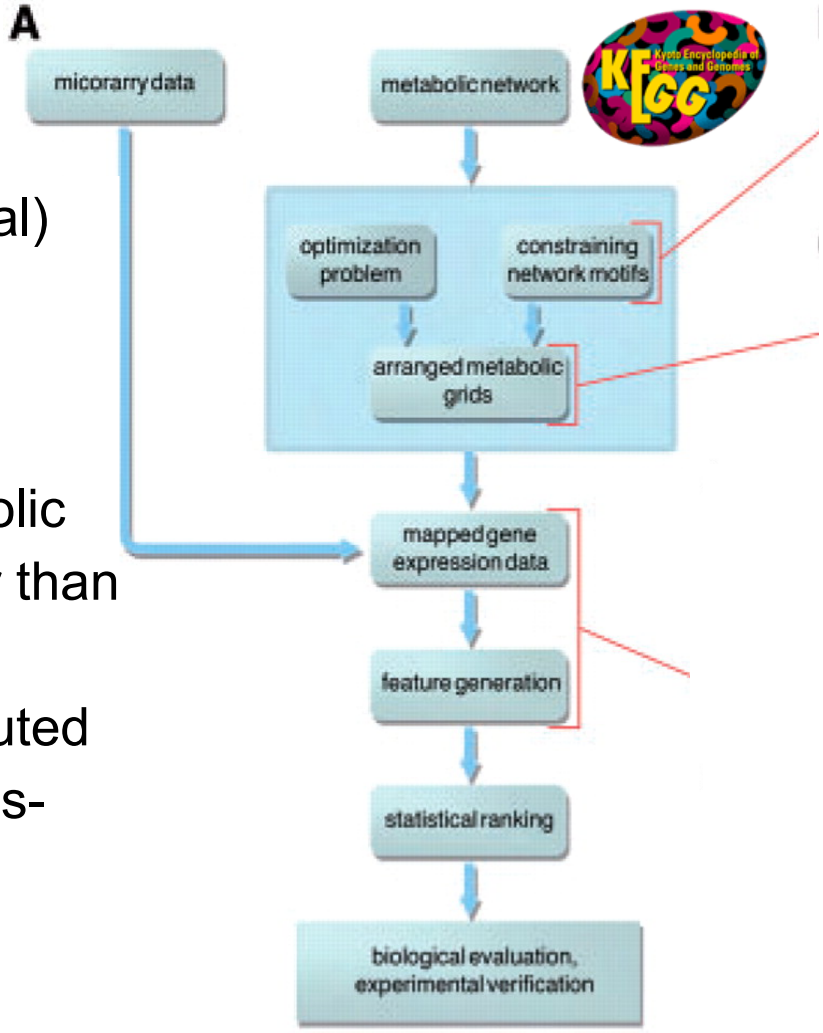
PathWave *(Schramm et al., Bioinformatics, 2010)*

- Identification of reg. patterns: pathways (global) and metabolic switches (local)
- Comparison in terms of metabolic reactions rather than single genes
- Features computed



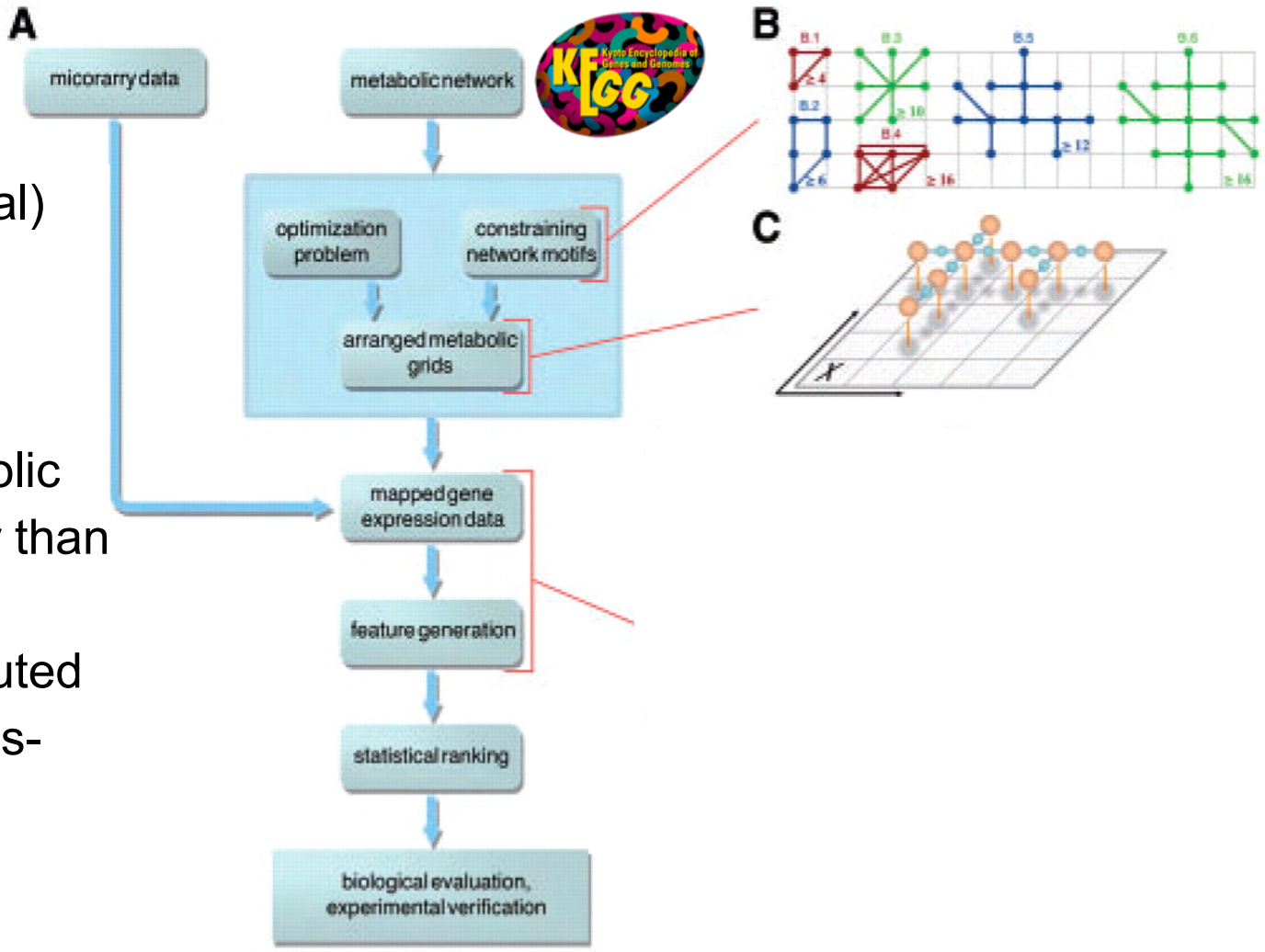
PathWave *(Schramm et al., Bioinformatics, 2010)*

- Identification of reg. patterns: pathways (global) and metabolic switches (local)
- Comparison in terms of metabolic reactions rather than single genes
- Features computed via wavelet trans-forms



PathWave *(Schramm et al., Bioinformatics, 2010)*

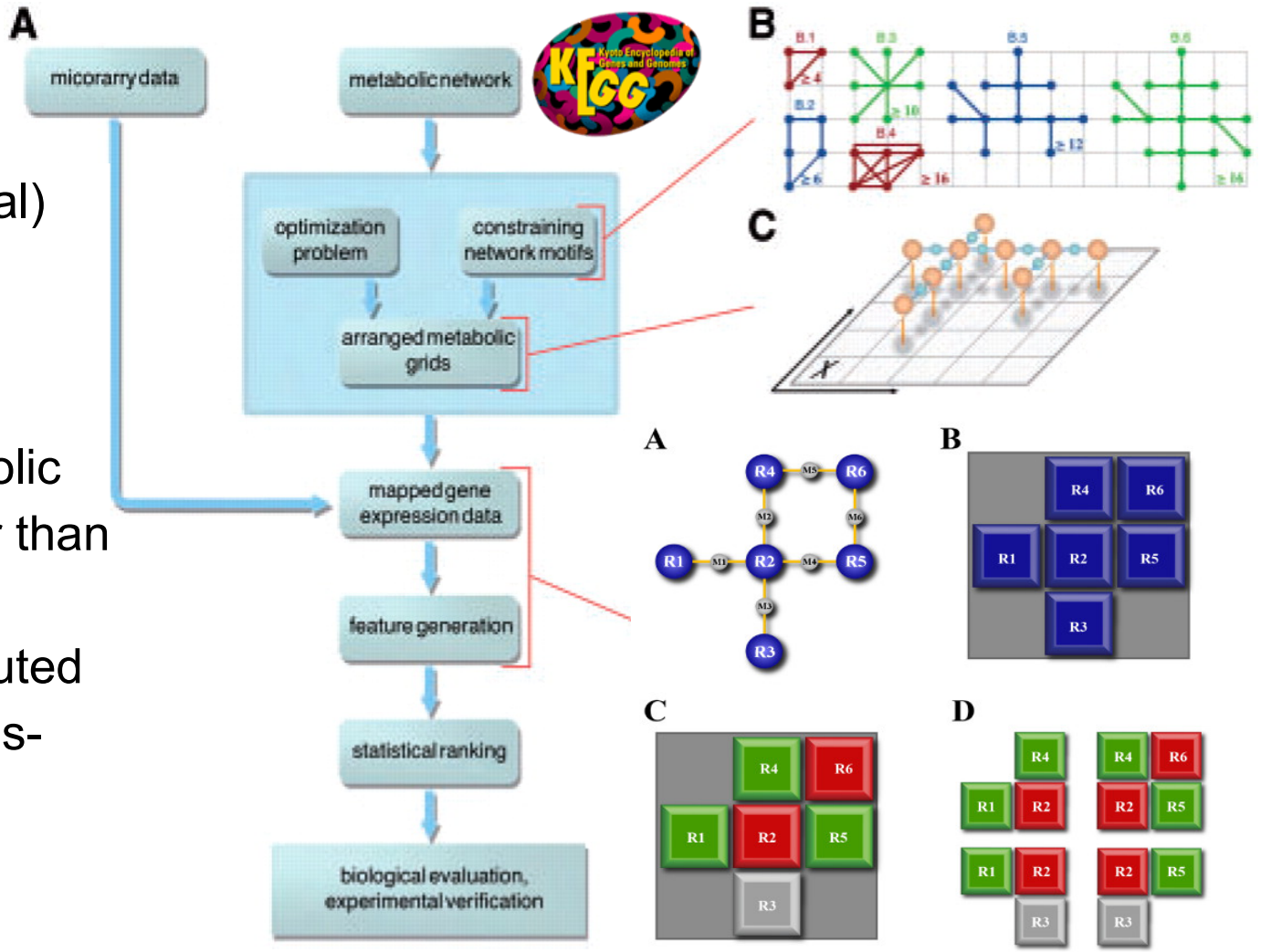
- Identification of reg. patterns: pathways (global) and metabolic switches (local)
- Comparison in terms of metabolic reactions rather than single genes
- Features computed via wavelet trans-forms



PathWave

(Schramm et al., Bioinformatics, 2010)

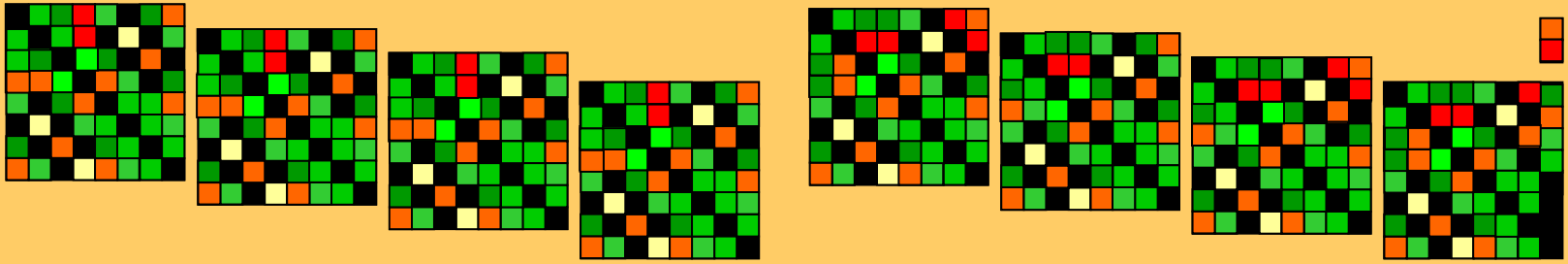
- Identification of reg. patterns: pathways (global) and metabolic switches (local)
- Comparison in terms of metabolic reactions rather than single genes
- Features computed via wavelet trans-forms



integral linear program (CPLEX8.1) to arrange metabolic pathways on 2D grids

PathWave *(Schramm et al., Bioinformatics, 2010)*

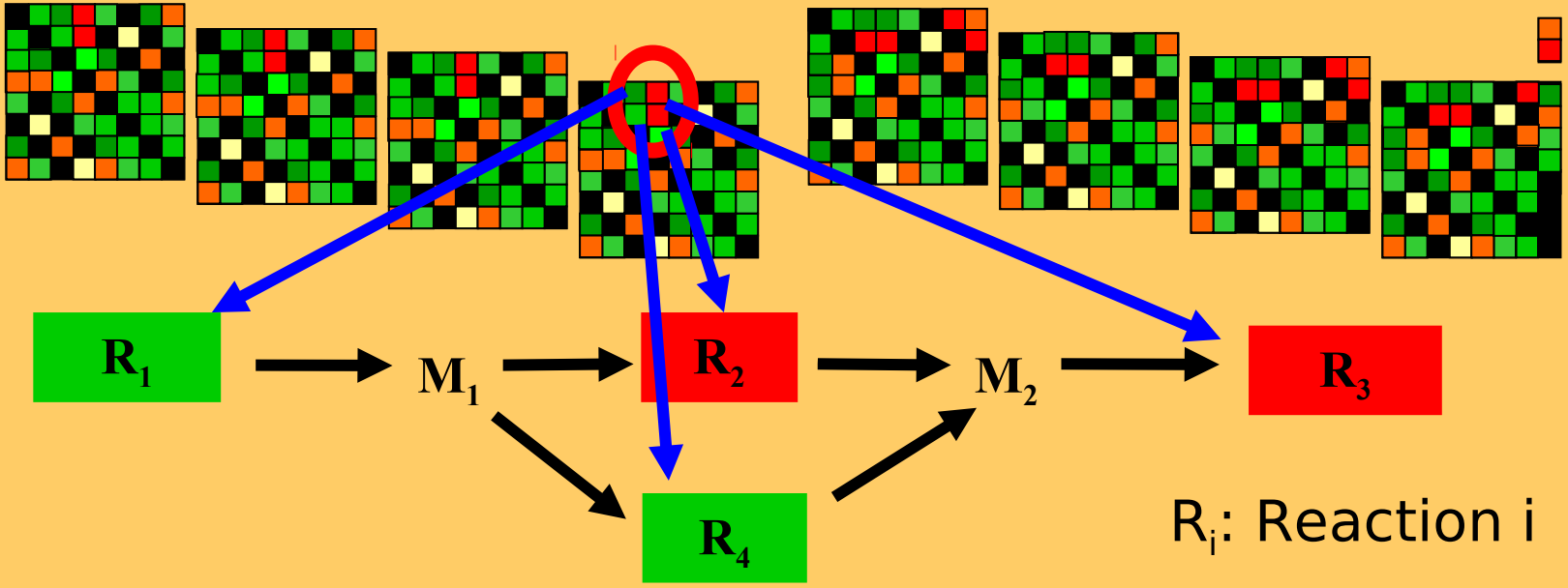
Pattern recognition with Haar wavelet transformations



PathWave

(Schramm et al., Bioinformatics, 2010)

Pattern recognition with Haar wavelet transformations



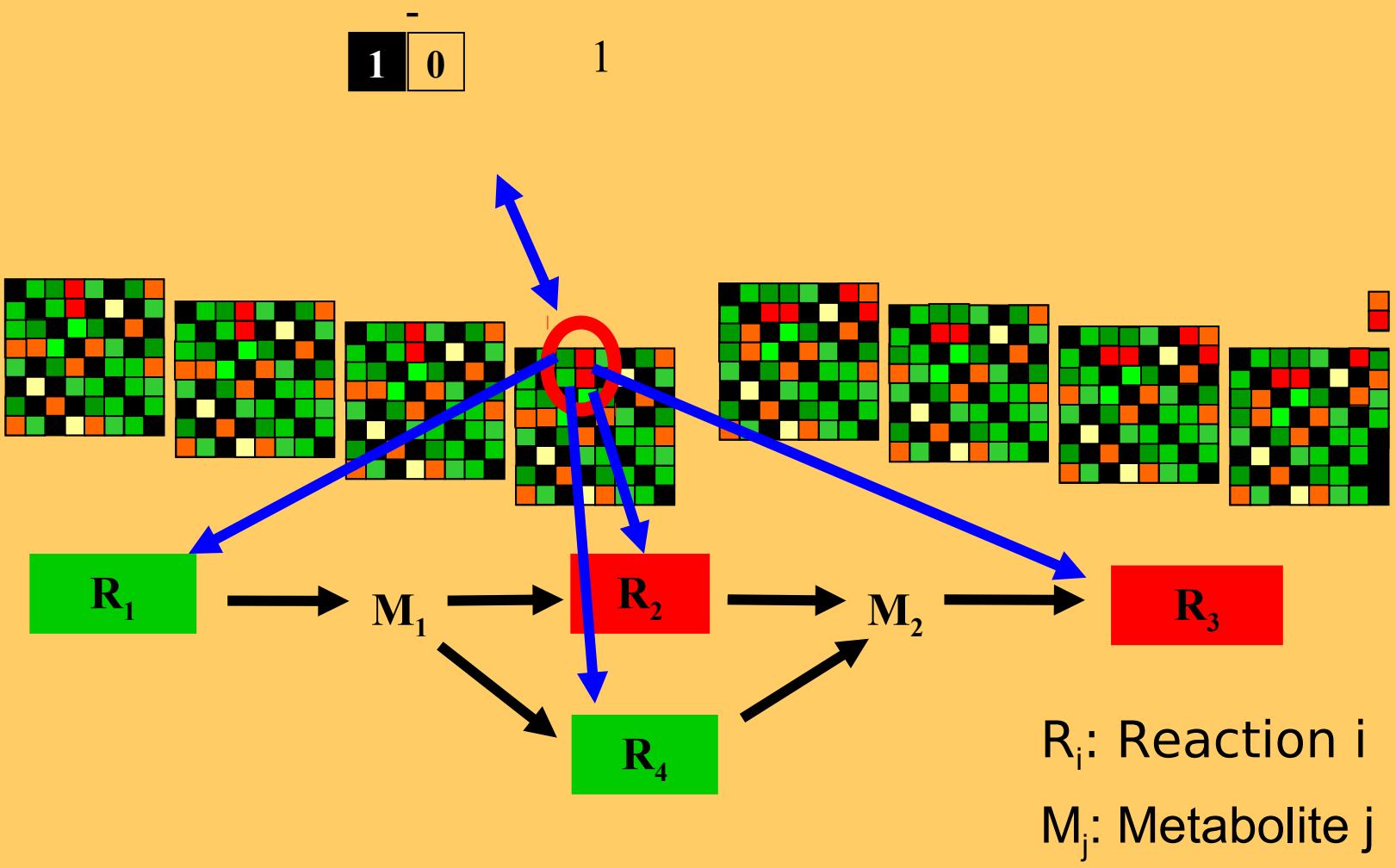
R_i: Reaction i

M_j: Metabolite j

PathWave

(Schramm et al., Bioinformatics, 2010)

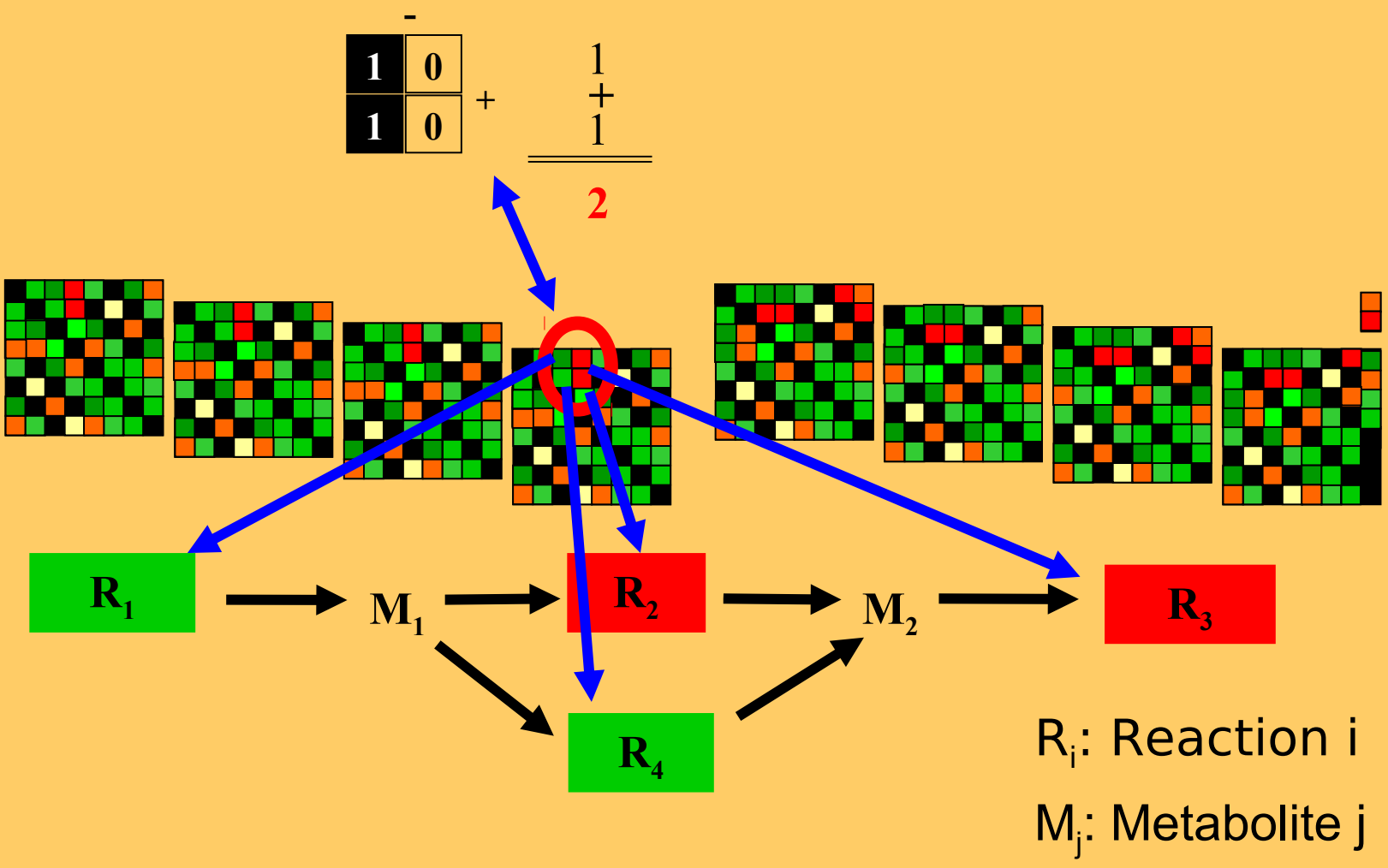
Pattern recognition with Haar wavelet transformations



PathWave

(Schramm et al., Bioinformatics, 2010)

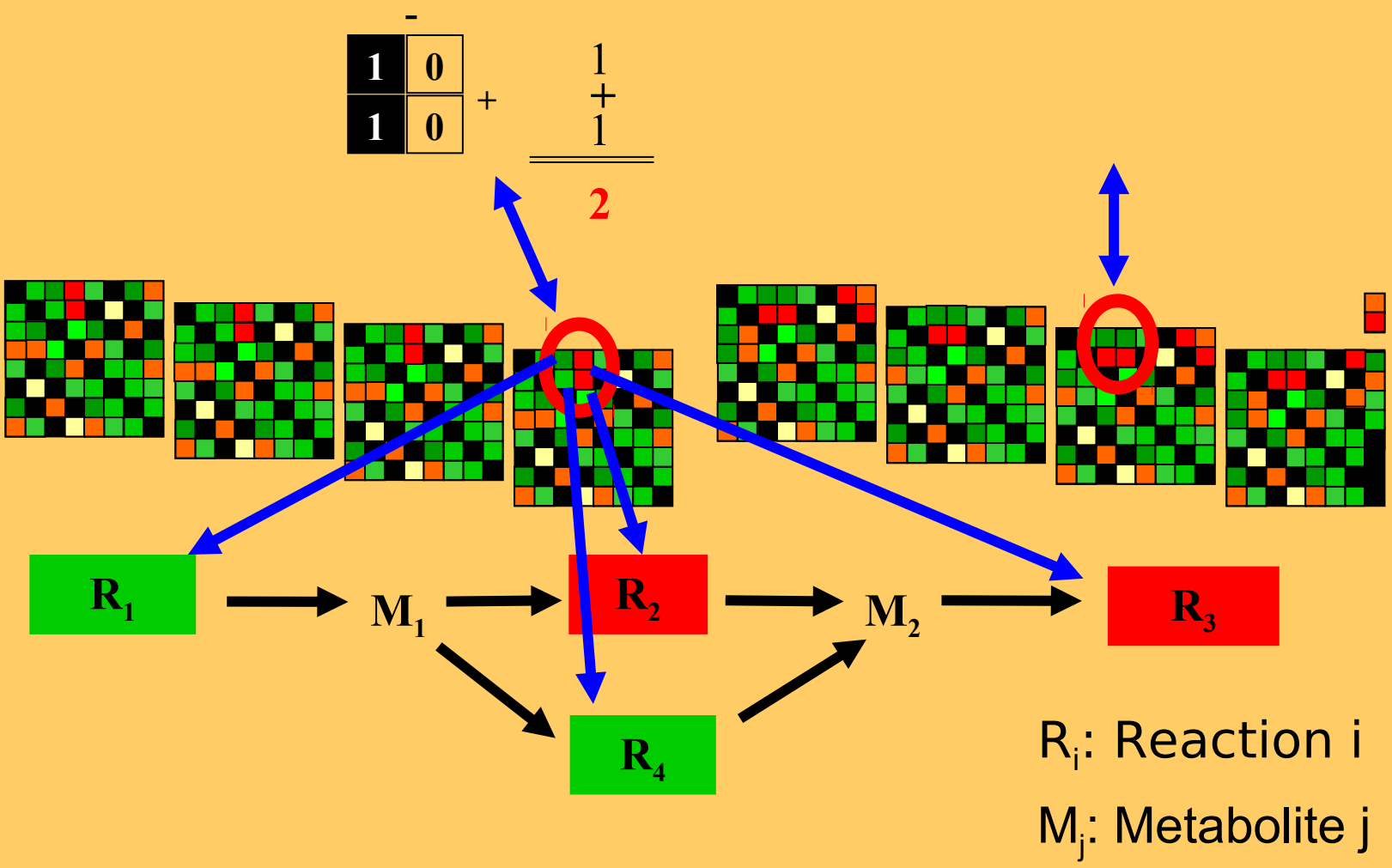
Pattern recognition with Haar wavelet transformations



PathWave

(Schramm et al., Bioinformatics, 2010)

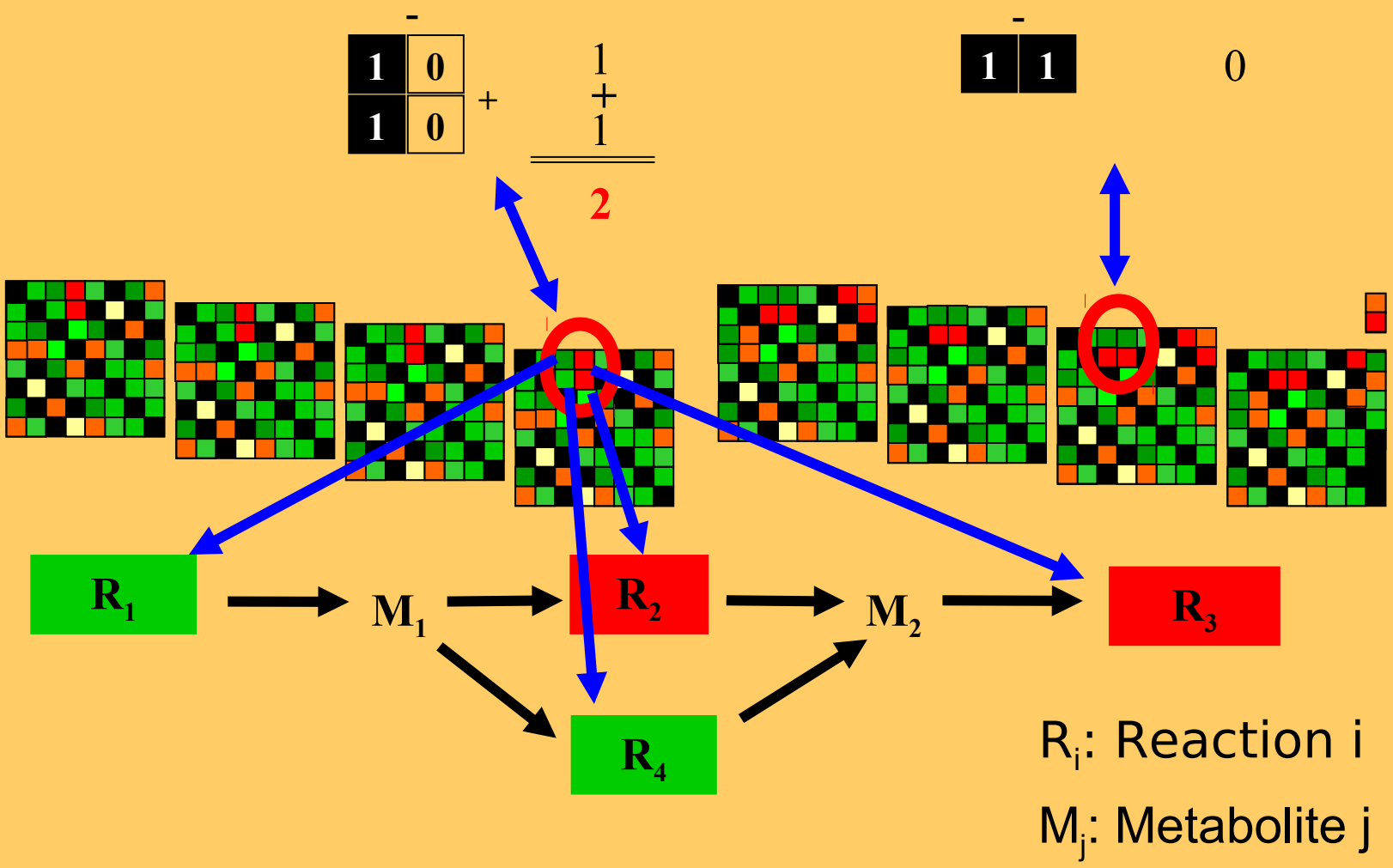
Pattern recognition with Haar wavelet transformations



PathWave

(Schramm et al., Bioinformatics, 2010)

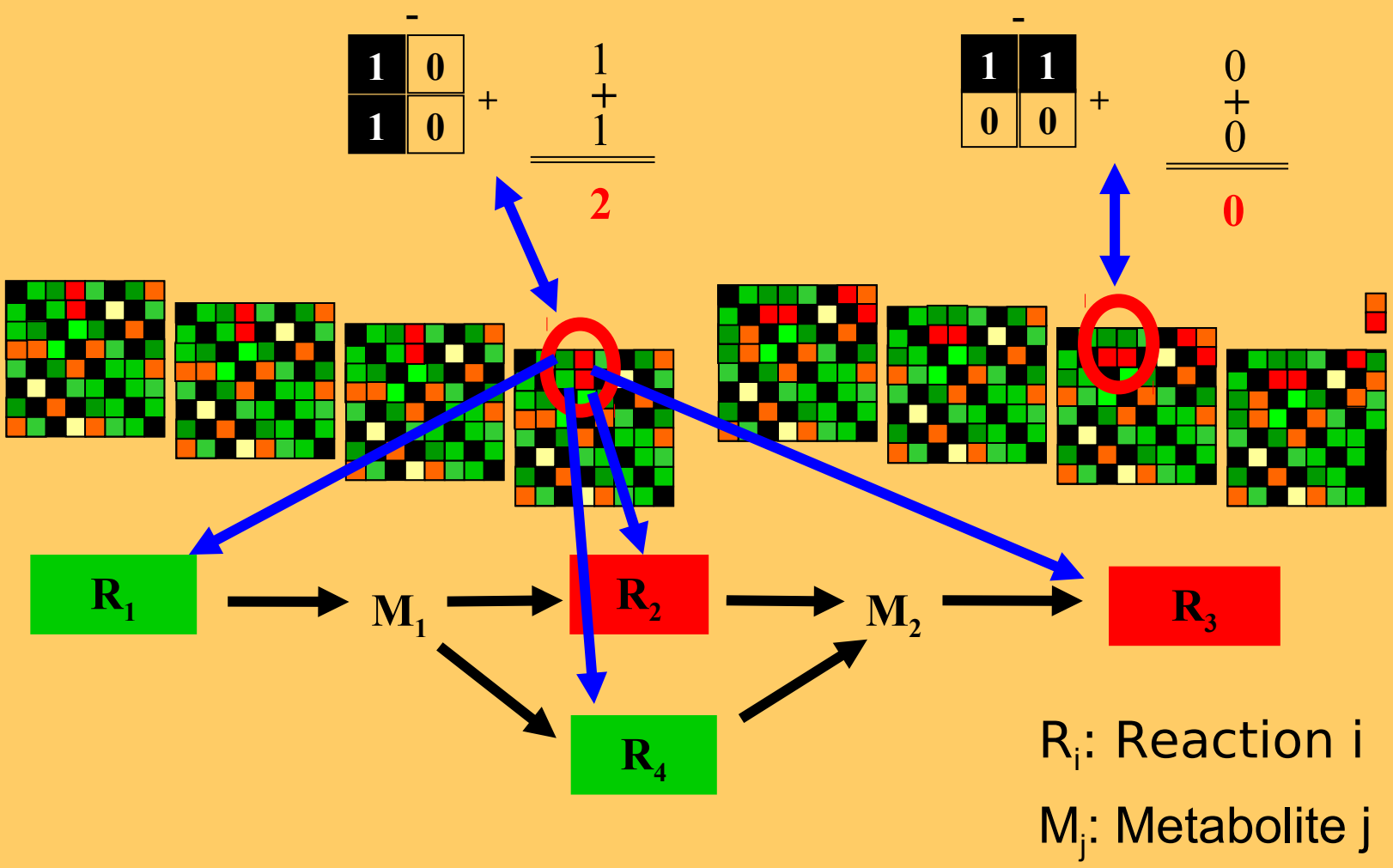
Pattern recognition with Haar wavelet transformations



PathWave

(Schramm et al., Bioinformatics, 2010)

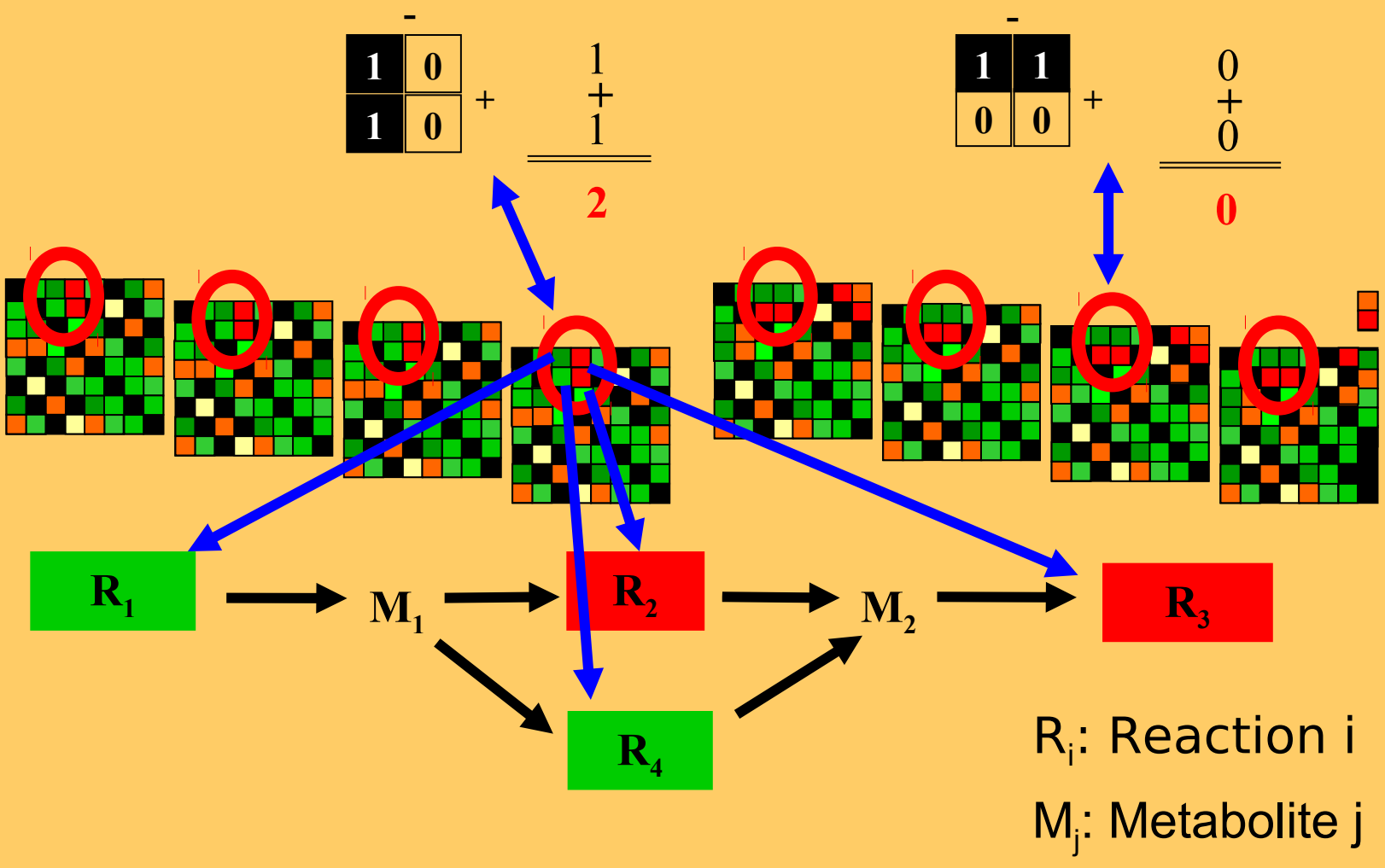
Pattern recognition with Haar wavelet transformations



PathWave

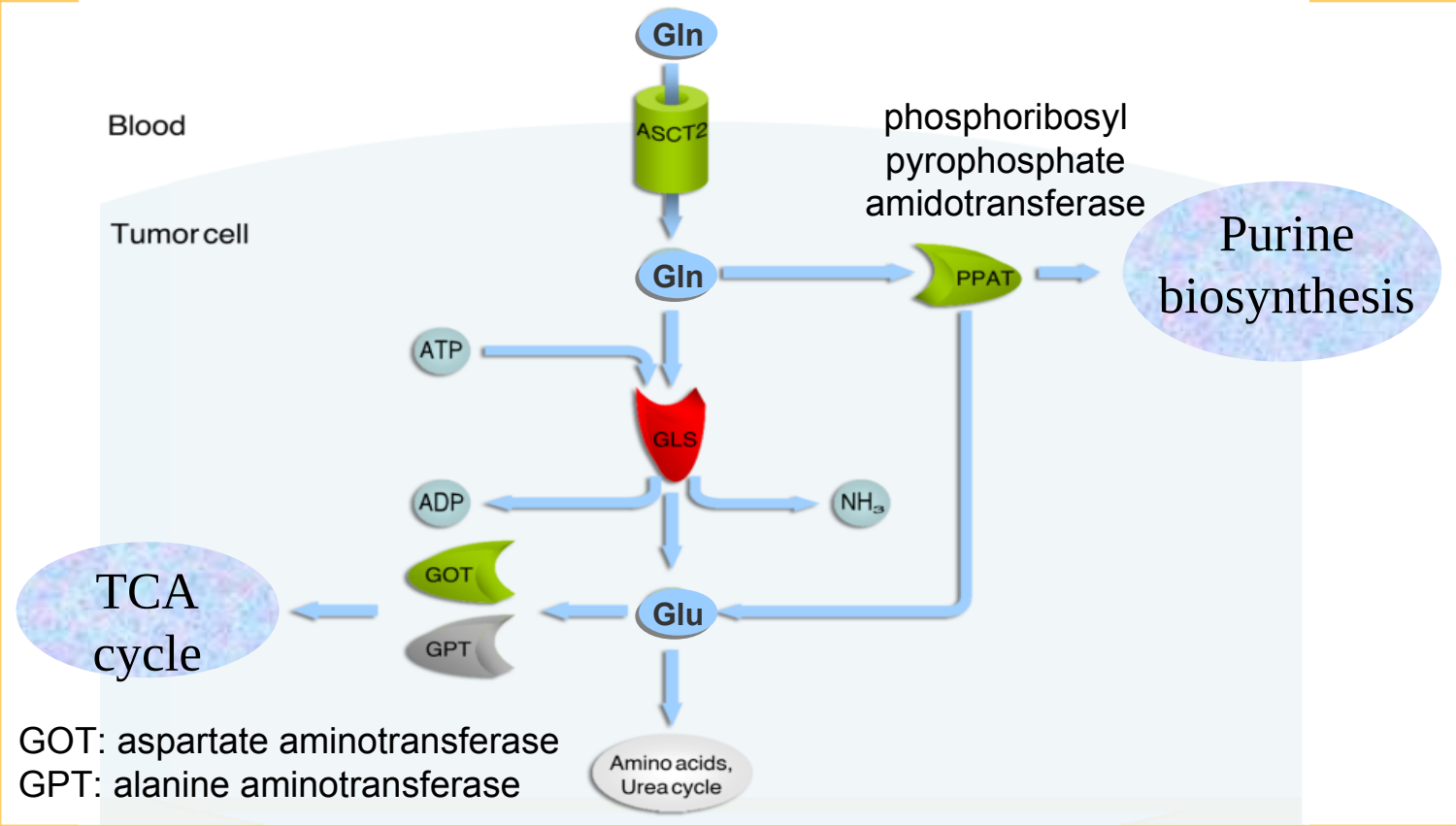
(Schramm et al., Bioinformatics, 2010)

Pattern recognition with Haar wavelet transformations



Neuroblastoma (stage 1 vs. 4 Mycn-amp)

Example: Switch-like regulation pattern in glutamate metabolism in neuroblastoma



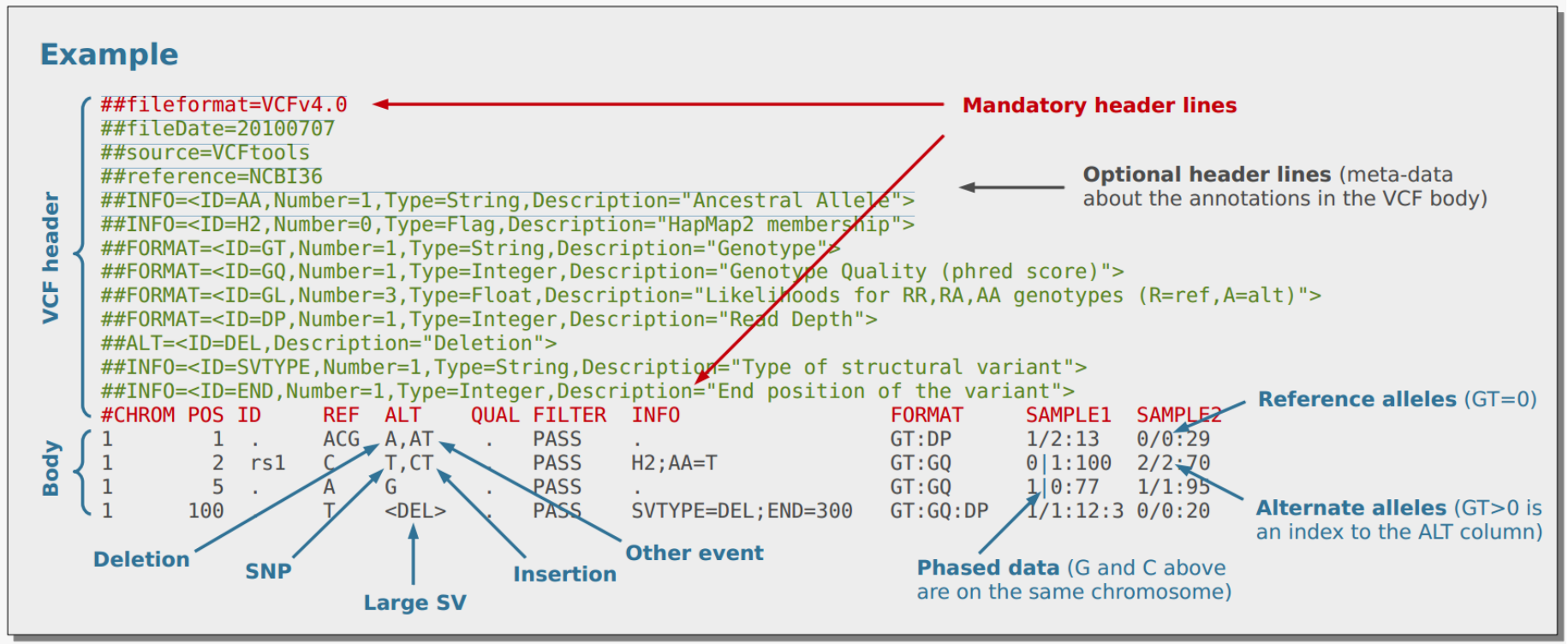
Acknowledgments

- Freie Universtiät Berlin
  - Sandra Krüger, Annalisa Marsico
- Berlin Institute of Health
  - Marten Jäger
- German Cancer Research Center (DKFZ), Heidelberg
  - Susanne Gröbner, Peter Lichter, Stefan Pfister, Marc Zapatka, Hendrik Witt, Natalie Jäger, Barbara Hutter, . . .
- Jackson Laboratory (JAX), Farmington, Connecticut
  - Peter N. Robinson
- Università degli Studi di Torino
  - Paolo Provero, Ferdinando di Cunto, Ugo Ala, Ivan Molineris, . . .
- University of Jena, Germany
  - Rainer König, Marcus Oswald
- The University of Tokyo
  - Yuichi Shiraishi

Questions?



Variant Call Format (VCF)



- Can be used to specify variants/mutations from one or more samples

QP Solving with quadprog

- quadprog implements the method of Goldfarb and Idnani (Mathematical Programming 27:1–33, 1983) for solving quadratic programming problems of the form

$$min(-d^T b + \frac{1}{2} b^T D b)$$

with the constraints $A^T b \geq b_0$.

In our case:

$$min(-\vec{g}^T \mathbf{S} \vec{w} + \frac{1}{2} \vec{w}^T \mathbf{S}^T \mathbf{S} \vec{w}) \qquad \text{subject to } \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

  - D (and hence $\mathbf{S}^T \mathbf{S}$) needs to be positive definite (scalar $b^T D b$ is positive for every non-zero column vector b)
  - We achieve this using function `nearPD` of the `Matrix` package

QP Solving with quadprog

- Differences between approximated, positive definite $\mathbf{S}^T\mathbf{S}$ and original $\mathbf{S}^T\mathbf{S}$ are minimal!
- Example for test of 44 out of 435 tumors with 15 signatures:

