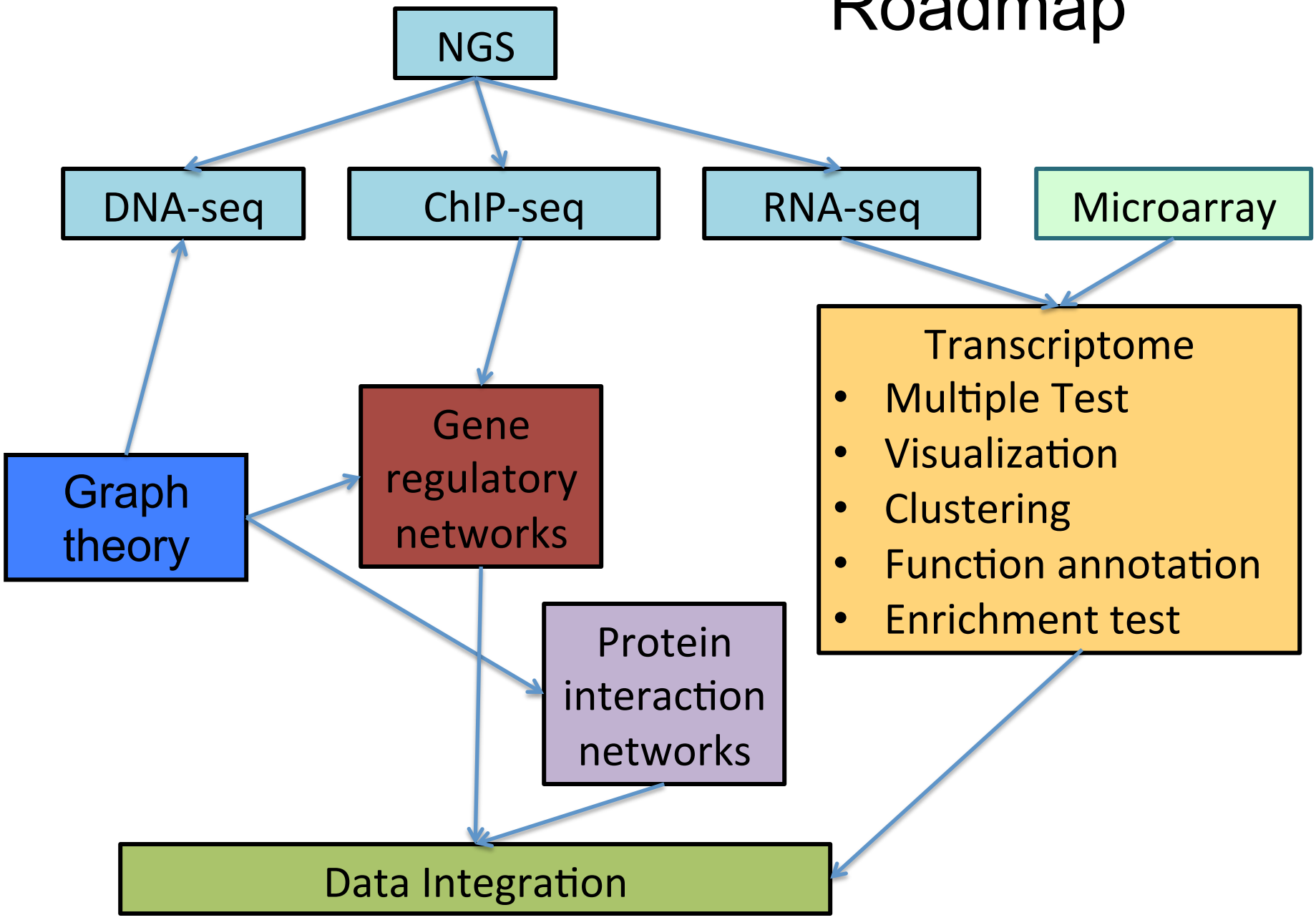
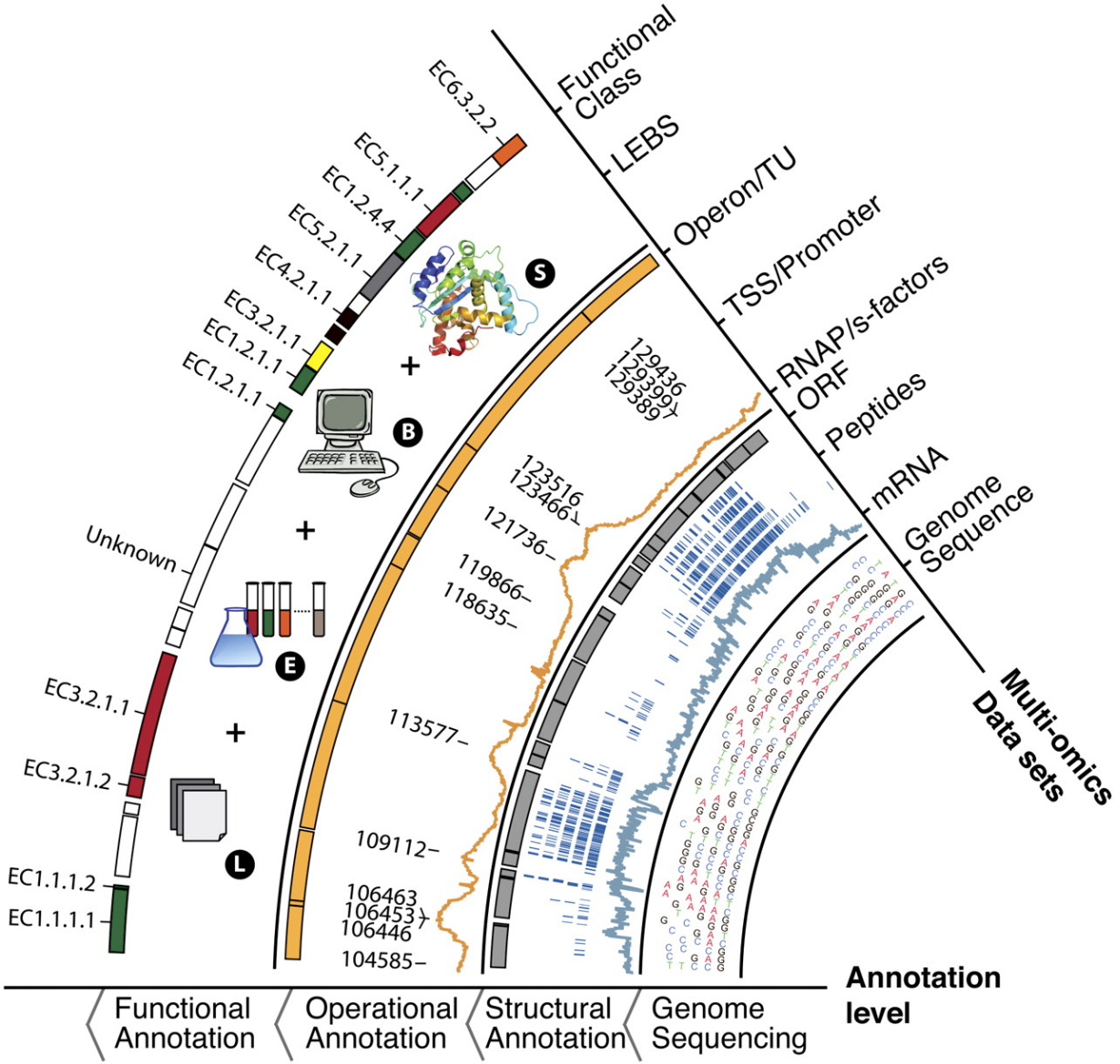
Data Integration

Roadmap


NGS
DNA-seq
ChIP-seq
RNA-seq
Microarray
Transcriptome
• Multiple Test
• Visualization
• Clustering
• Function annotation
• Enrichment test
Graph theory
Gene regulatory networks
Protein interaction networks
Data Integration

Metastructures: systems approach to determine genome annotation.



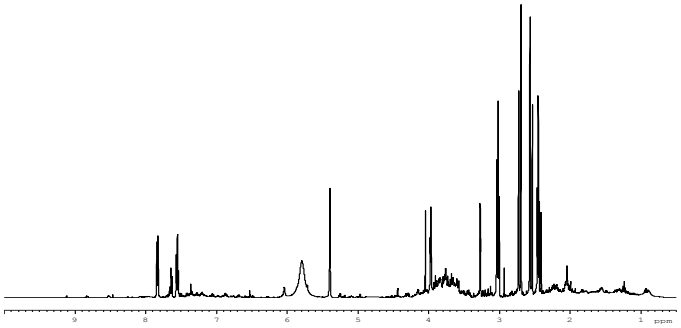
Qiu Y et al. Genome Res. 2010;20:1304-1311

Why do we need to integrate various types of omic data?

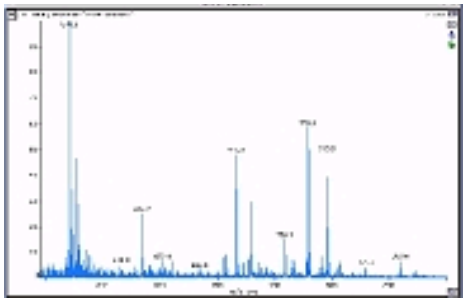
- Get a consensus results (reduce false positive rate)
- Focusing on one type of data may miss an obvious signal

Experimental Platforms

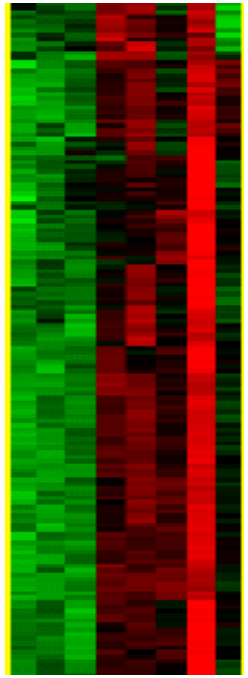
Non-omics and Omics, what are they?



NMR protein structures and proteomics

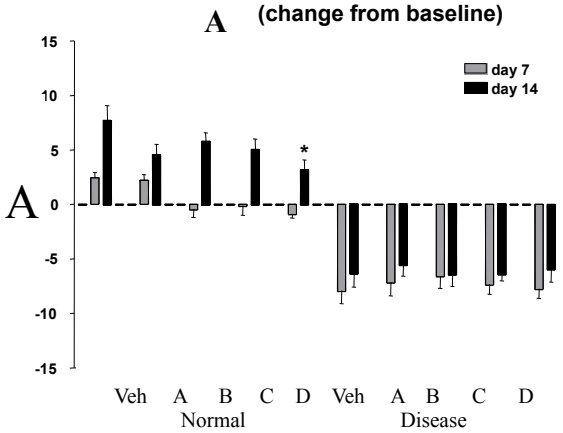


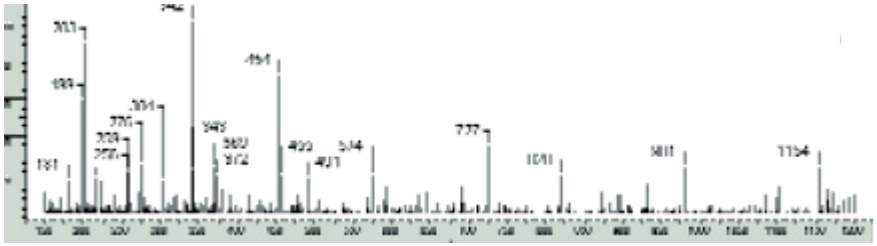
DNA sequences



Transcriptome

"Non-omic" markers





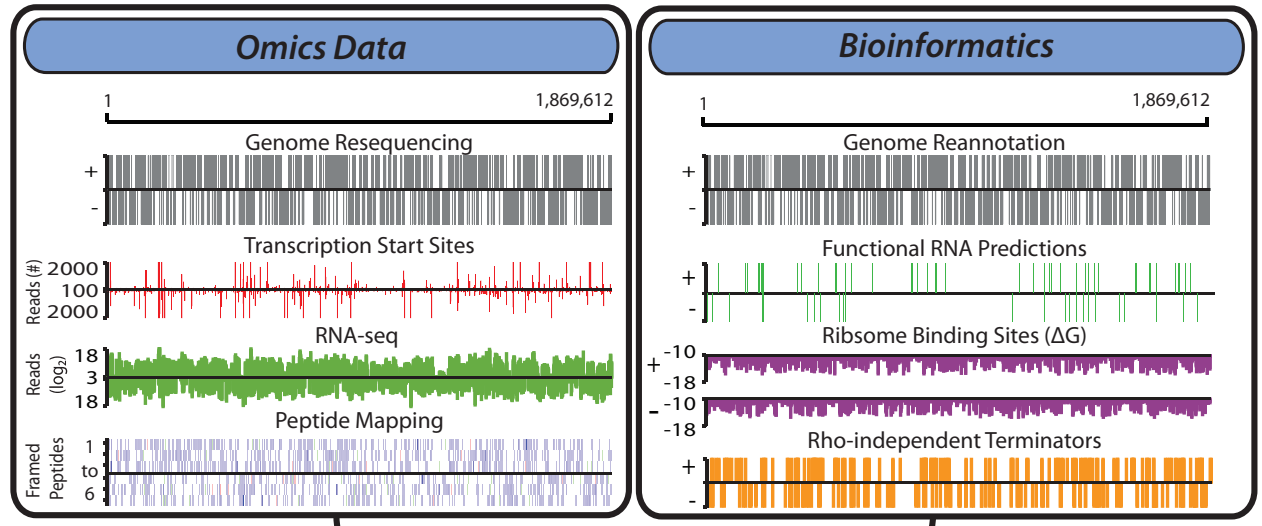
LC-MS protein sequences and interactions

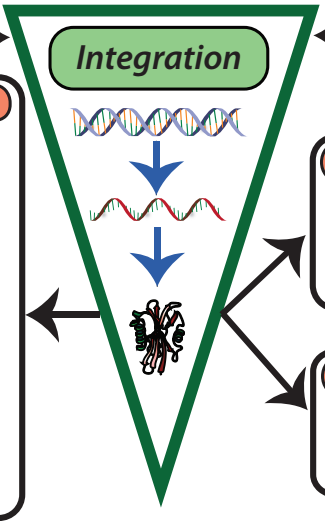
Integration of omics data sets

Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA interactions	Protein–protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification[74]	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification[75]	• Functional annotation[79]	• Functional annotation	• Functional annotation[71,103] • Biomarkers[125]
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation[20]	• Enzyme annotation[109]	• Gene-regulatory networks[76]	• Functional annotation[89] • Protein complex identification[82]		• Functional annotation[102]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[99]	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering[109]
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades[89,102]		• Dynamic network responses[84]
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity[89]
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

Nature Reviews Molecular Cell Biology, **7**:198–210, 2006.

Multi-omic data integration





Gene Annotation Improvement

Protein Coding Genes	**1896**
Comparison with NCBI	
NCBI Total # of Genes	1858
Shared Genes	1830
Genes Dropped	28
Newly Annotated Genes (%)	66 (3.5)
Gene Length Corrections	369
<30 bp Change	252
≥30 bp Change	117
Functional RNAs	**67**
tRNAs/rRNAs	46/3
CRISPRs	8
Others	10

Transcript Characterization

Transcription Units	**748**
Unique TU Starts	676
TUs with TSS Mapped	550
Genes per TU (average)	3.27

Genetic Element Identification

Promoters RBSs

UTRs Rho-independent Terminators

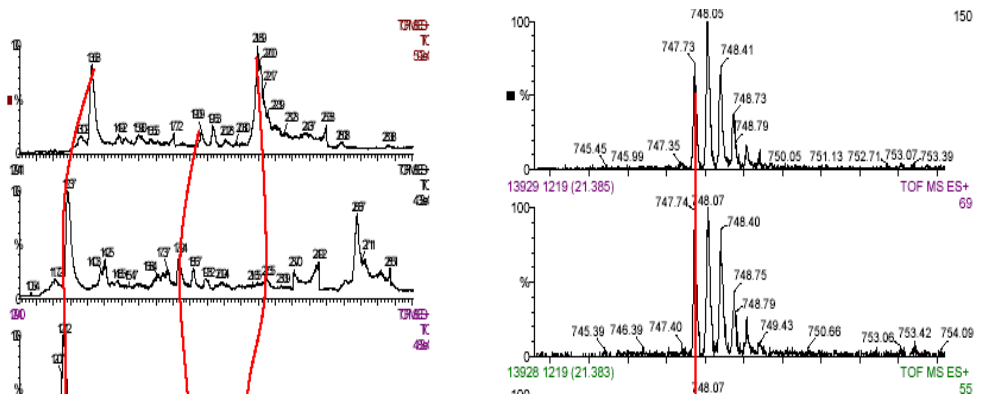
H. Latif et al. PLoS Genetics 2013

Challenges

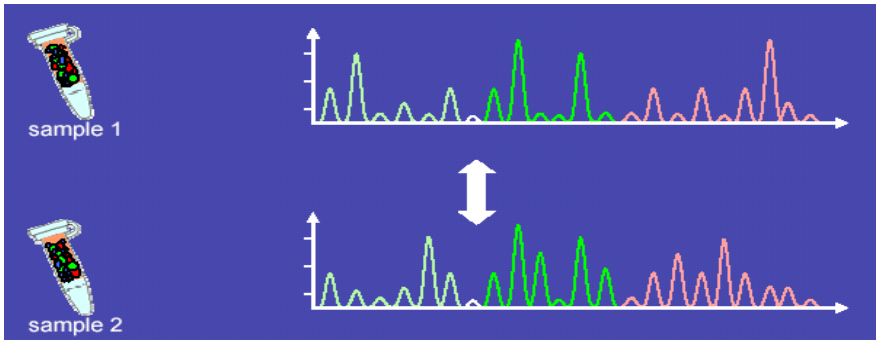
1. Data Pre-processing
2. High Dimensionality
3. Multiple Testing for Marker Selection
4. Data Integration
5. Validation of the Prediction Model

Challenge #1: Data Pre-processing

- Peak Alignment for different platforms

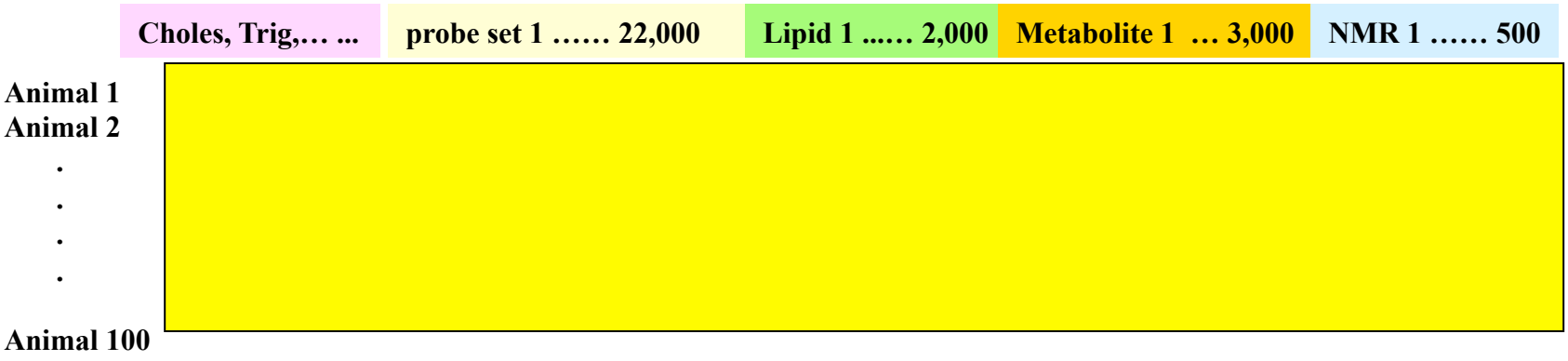


- Normalization among various types of data
  - Why? Remove systematic bias in the data
  - Normalization within the platform makes data comparable across samples



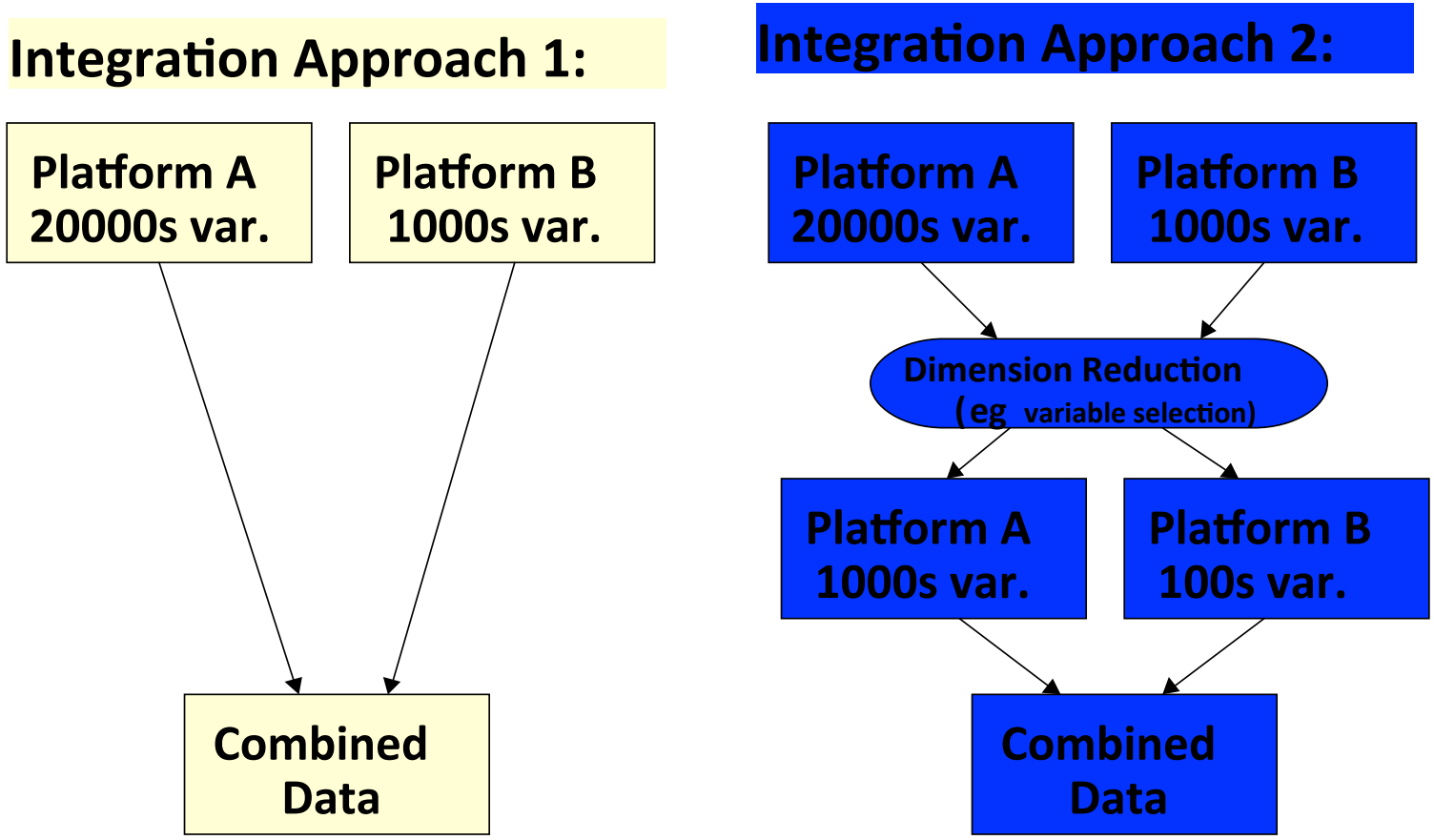
Challenge # 2: High Dimensionality

of subjects << # of variables

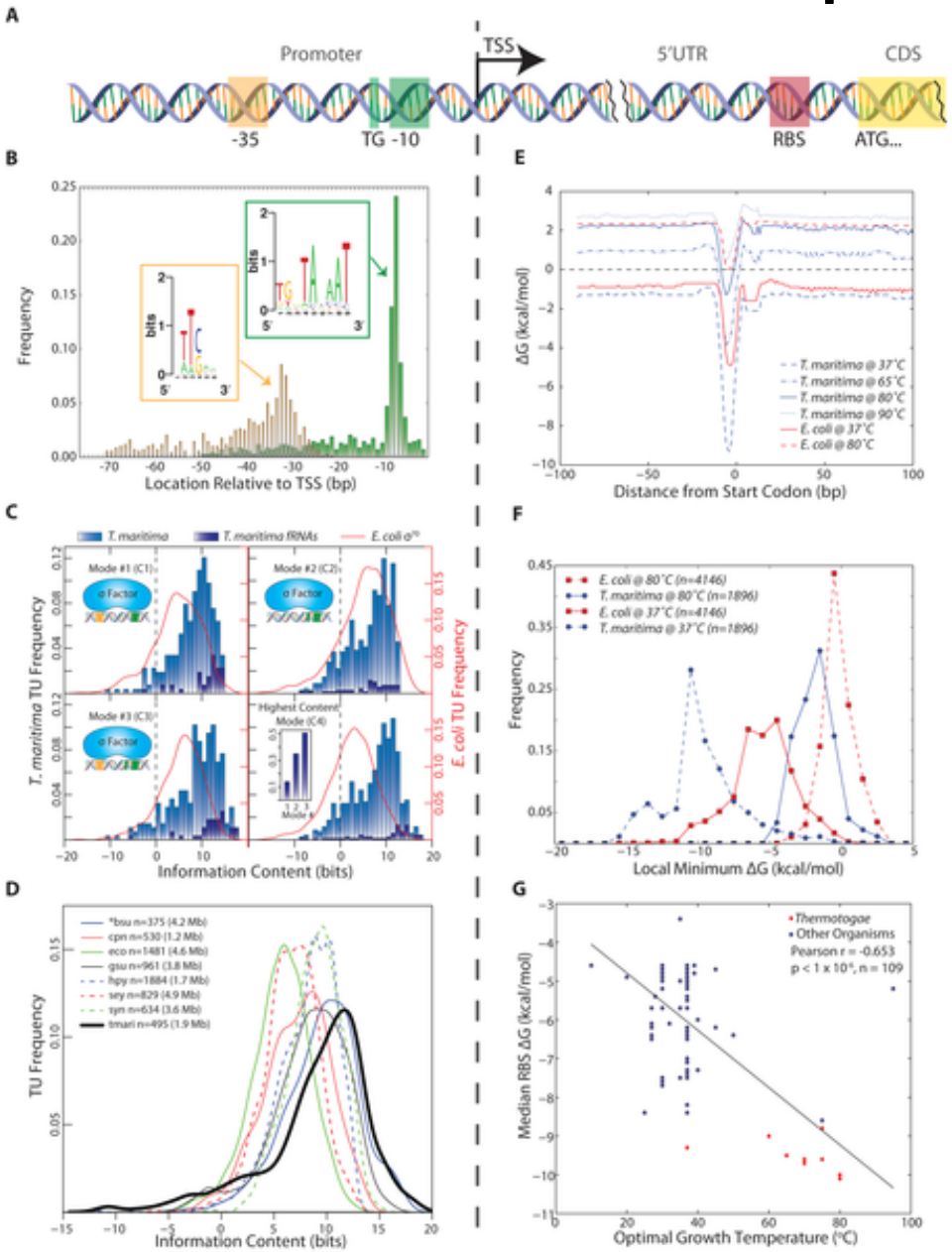


- Genes variants: 5000 peaks
- Gene Expression (RNA-seq): 22,000 probe sets
- Protein-DNA interactions (ChIP-seq): 2, 000 peaks
- Protein interactions: 100,000,000

Challenge #4: Data integration (How?)



Example 1

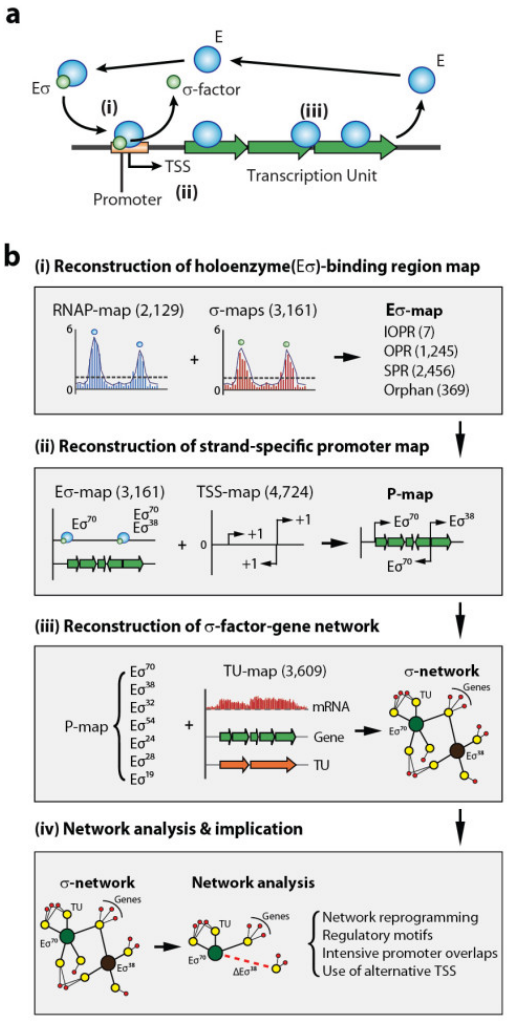


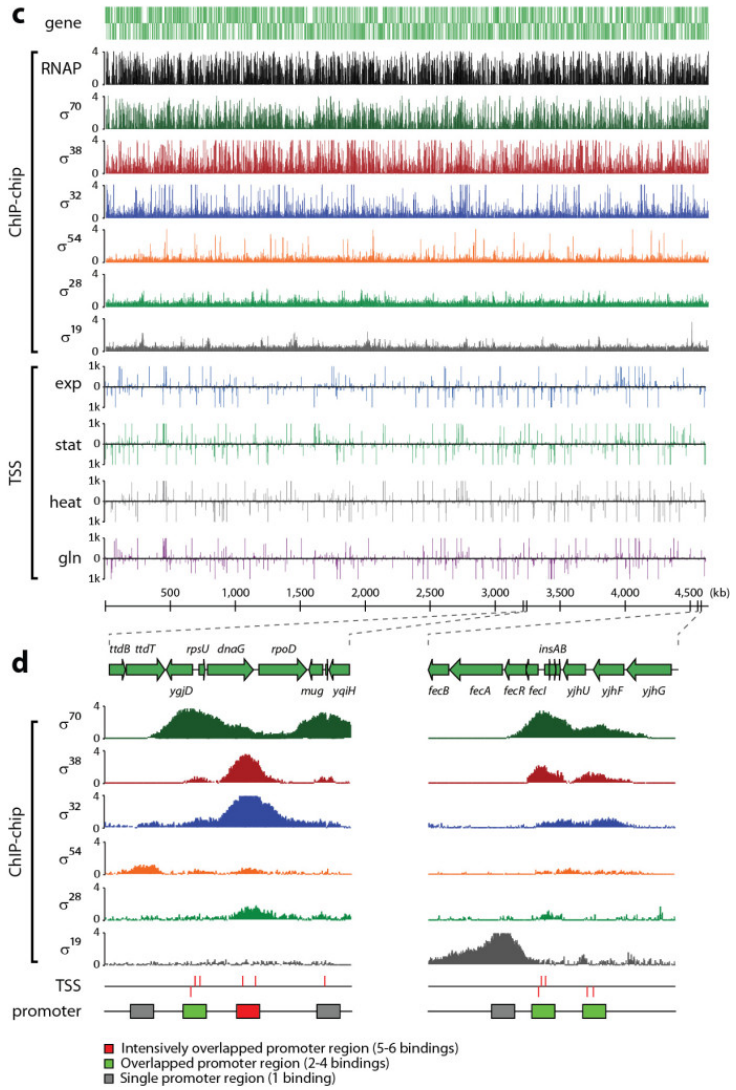
Identification and quantitative comparison of genetic elements for transcription and translation initiation.

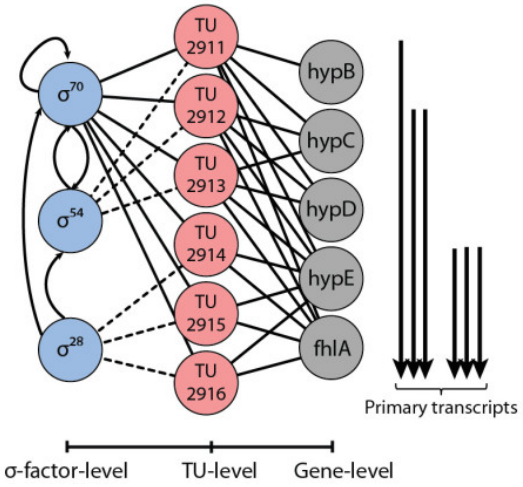
H. Latif et al. PLoS Genetics 2013

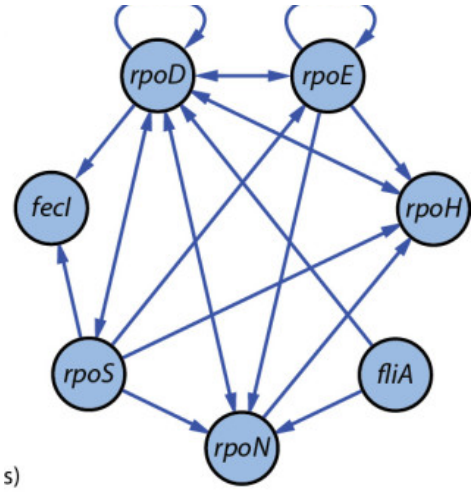
Example 2

σ-factor network in E. coli.









B. Cho et al. BMC Biology 2014

Example 3

Integrative Modeling Defines the Nova Splicing-Regulatory Network and Its Combinatorial Controls

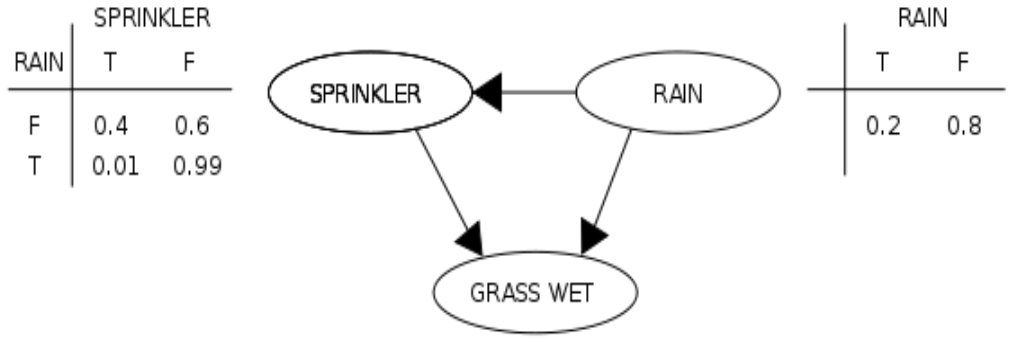
Chaolin Zhang et al. Science 2010

Data Integration

- More and more diverse "omics" data exist
- "It is essential to integrate various kinds of biological information and large-scale omics data sets through systematic analysis" with statistically rigorous and physically sound models
- Bayesian Network
- CLIP-seq (sequencing) + TF binding site (bioinformatics) + expression profile (microarry) +evolutionary signature.

Bayesian Network

- A probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph.
- used to represent causal relationships.



RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

RAIN T	RAIN F
0.2	0.8

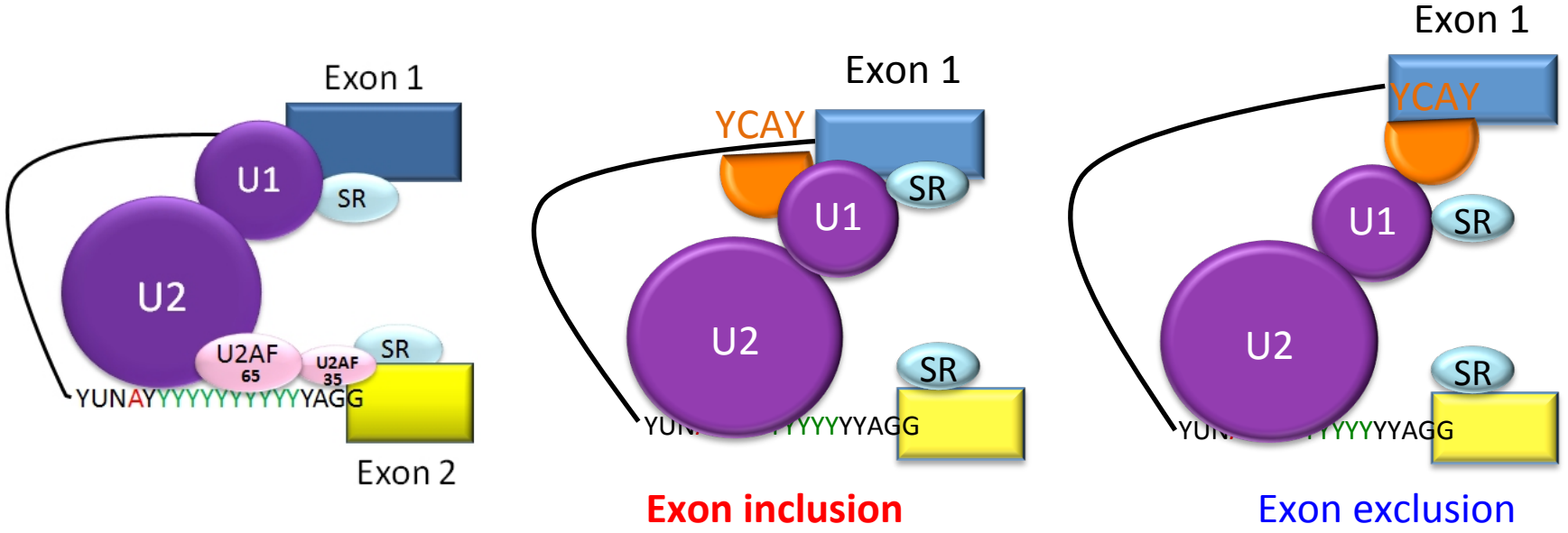
Applications:

1. Inferring unobserved variables
2. Parameter learning

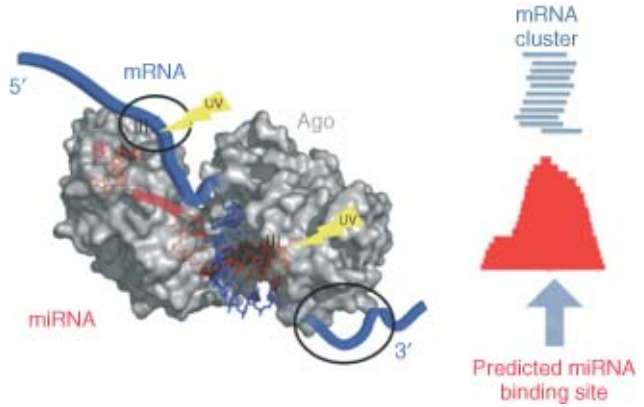
SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Nova-regulated Alternative Splicing

- Nova proteins are a family of neuron-specific alternative splicing factors. (Ule *et al. Nature* 2006)
- The Nova protein binds to pre-message RNA at a binding site of "YCAY" clusters.
- Nova binding to an exonic YCAY cluster changed the protein complexes assembled on pre-mRNA, blocking U1 snRNP binding and exon inclusion.
- Nova binding to an intronic YCAY cluster enhanced spliceosome assembly and exon inclusion.



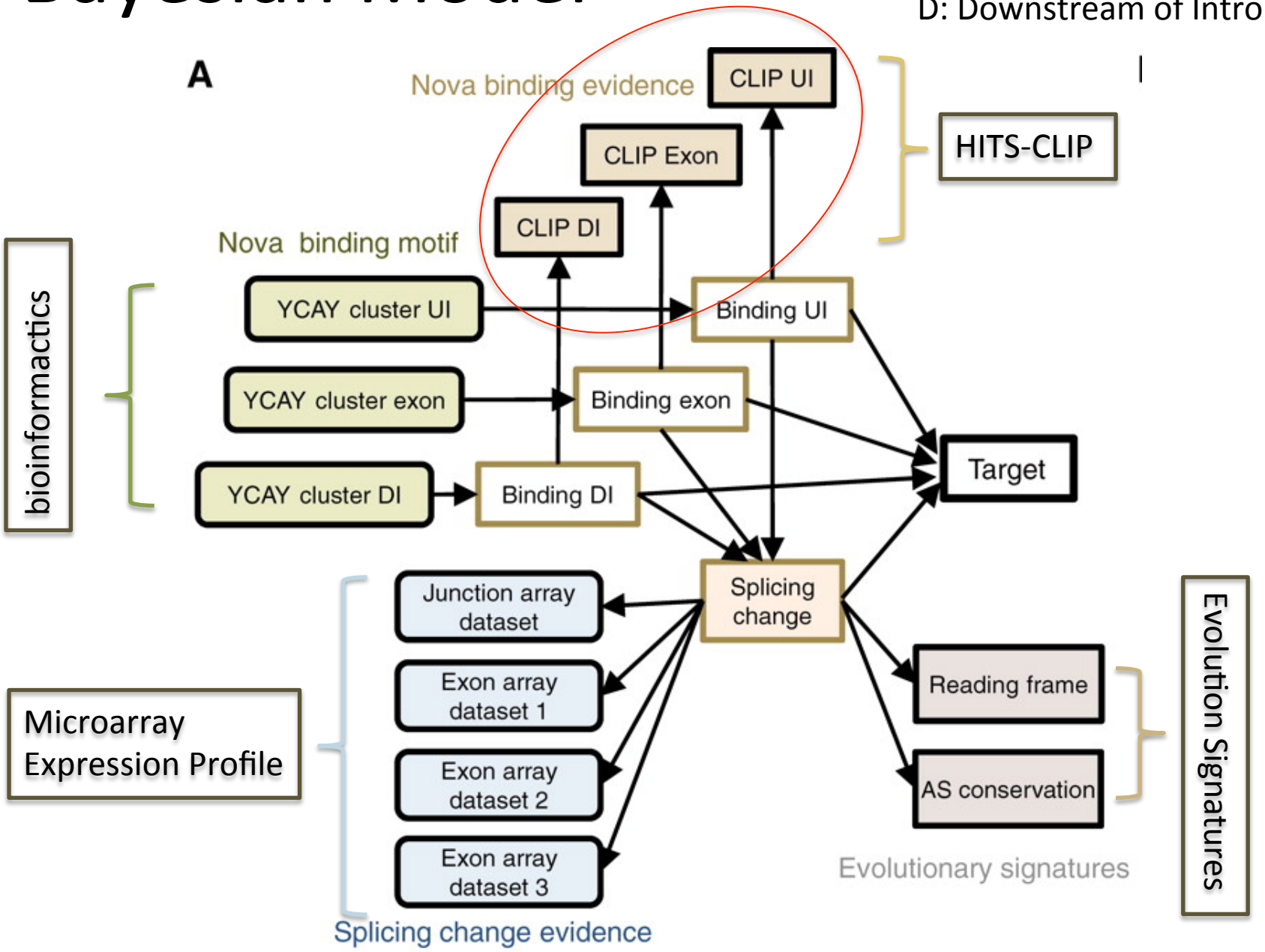
Data sources



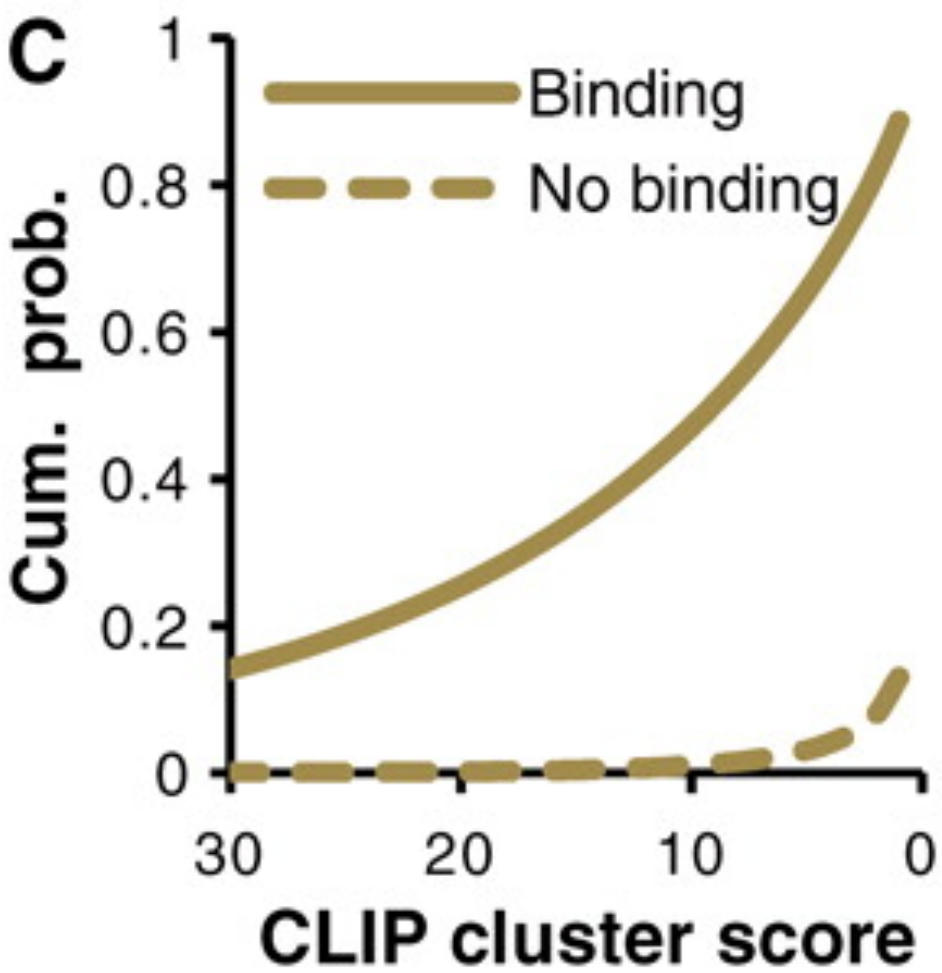
- **HITS-CLIP** (CLIP-seq). Study Protein-RNA binding by crosslinking between RNA and the protein, followed by immunoprecipitation and high-throughput sequencing.
- **Genome-wide searching of YCAY motif**. Bioinformatics approach.
- **Microarray data compared WT and Nova knockout.**
- **Evolution signature.** Conserved Alternating Splicing between human and rats.

Bayesian Model

E: Exon
U: Upstream of Intron
D: Downstream of Intron

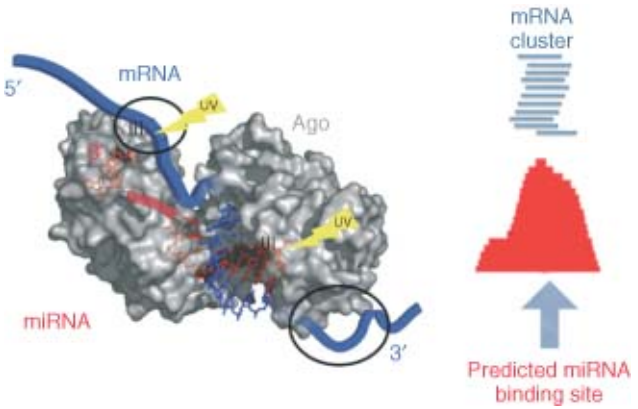


Estimated Conditional prob.

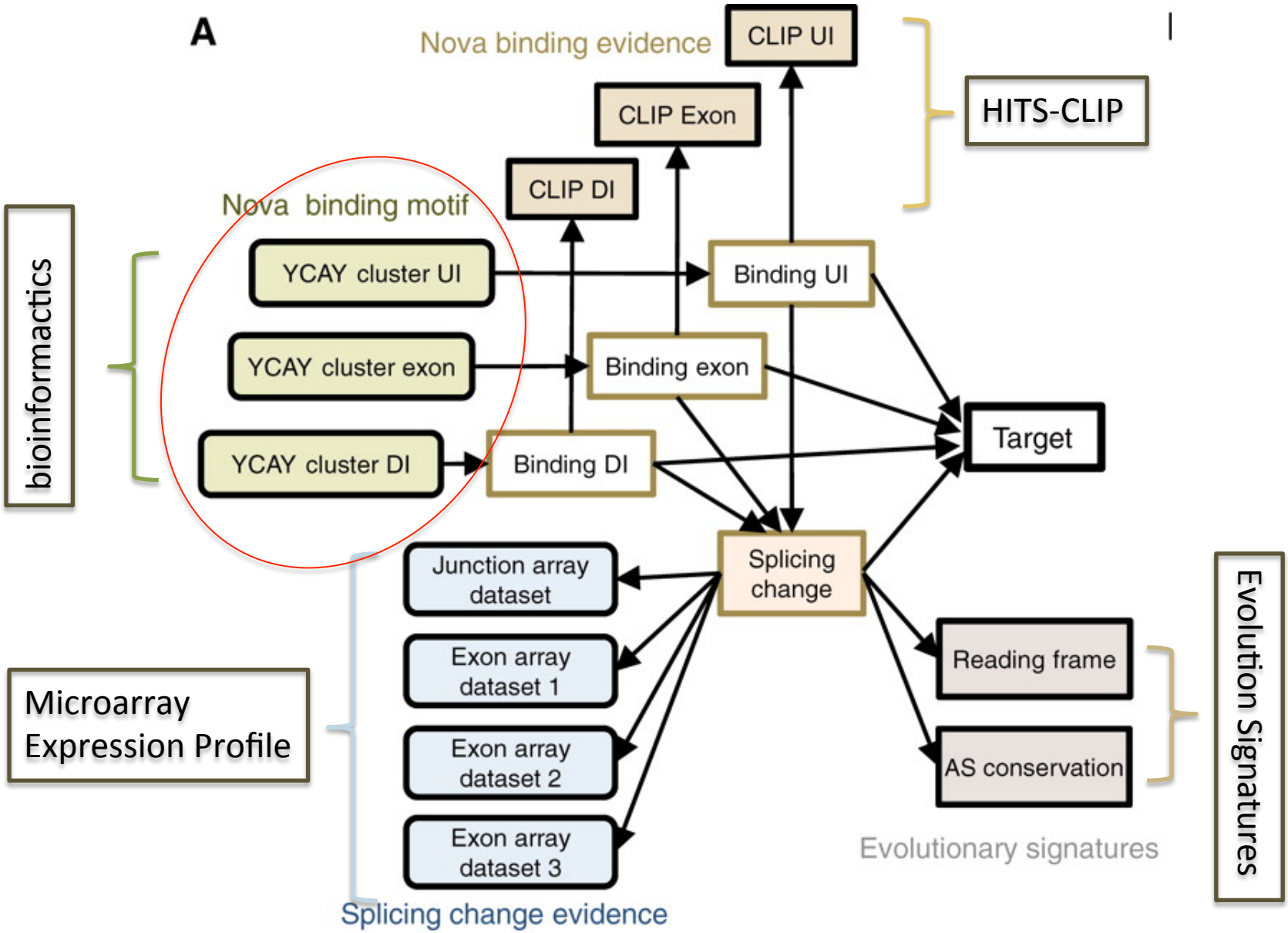


With inferred Nova binding

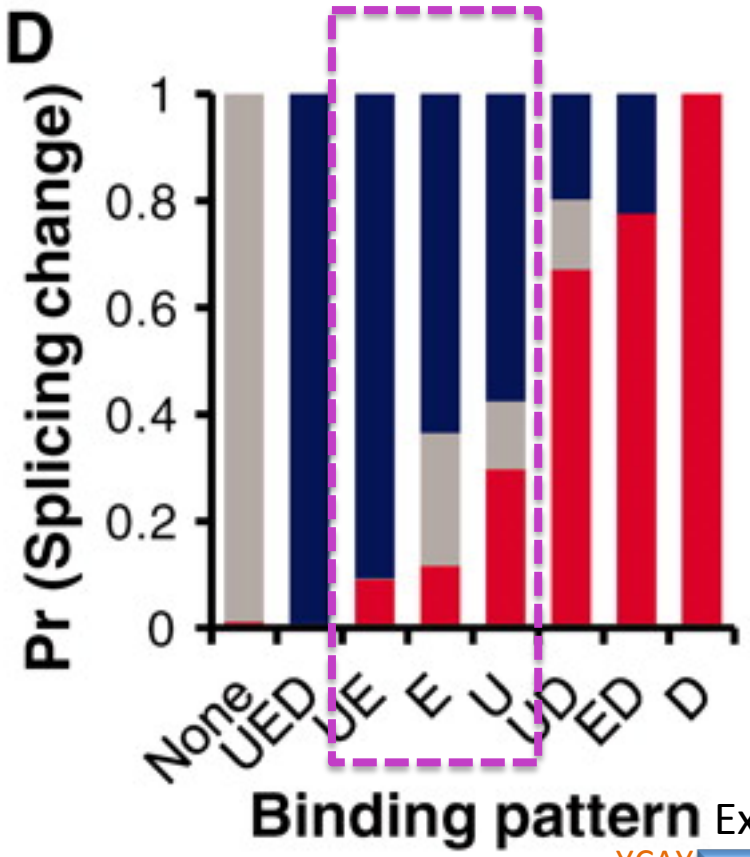
Without inferred Nova binding



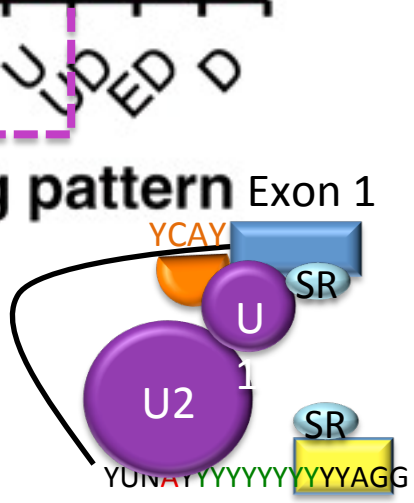
Bayesian Model



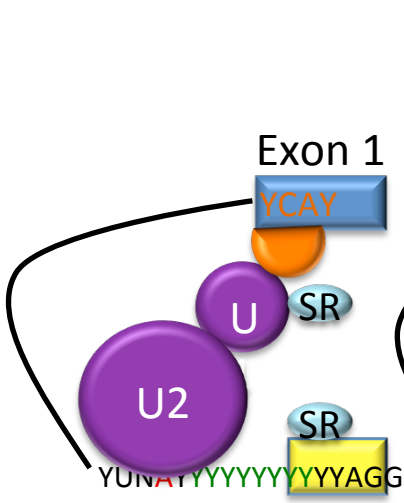
Regulation of Nova binding position

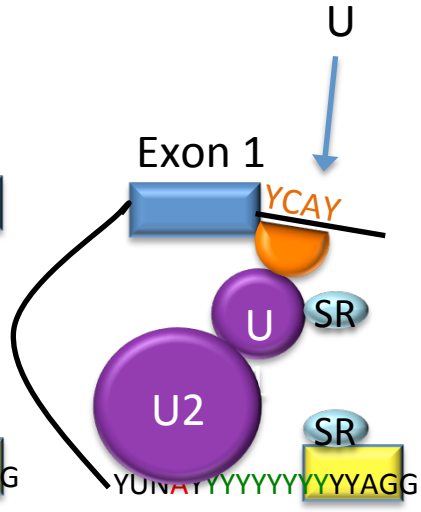


E: Exon
U: Upstream of Intron
D: Downstream of Intron



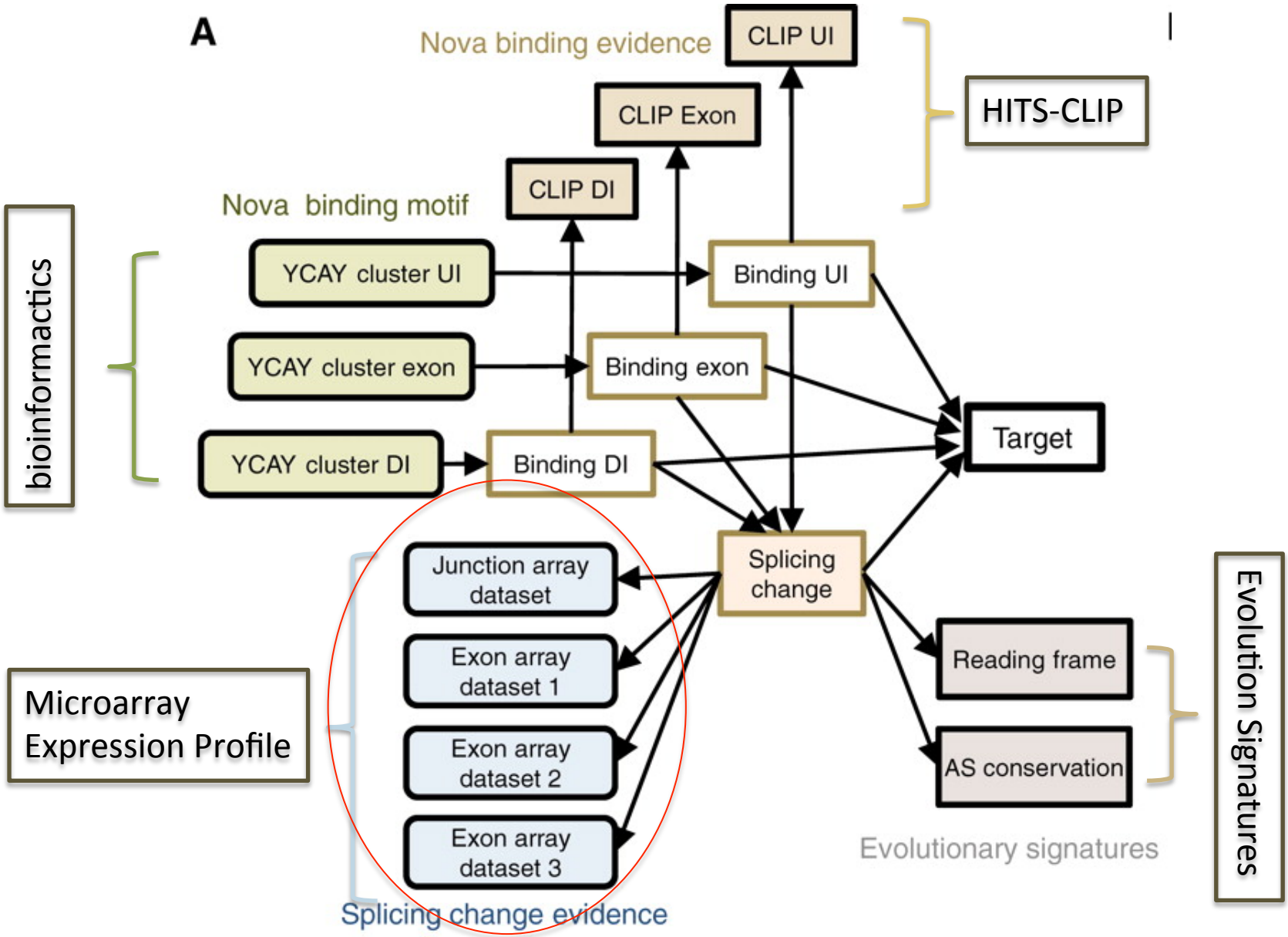
Exon inclusion



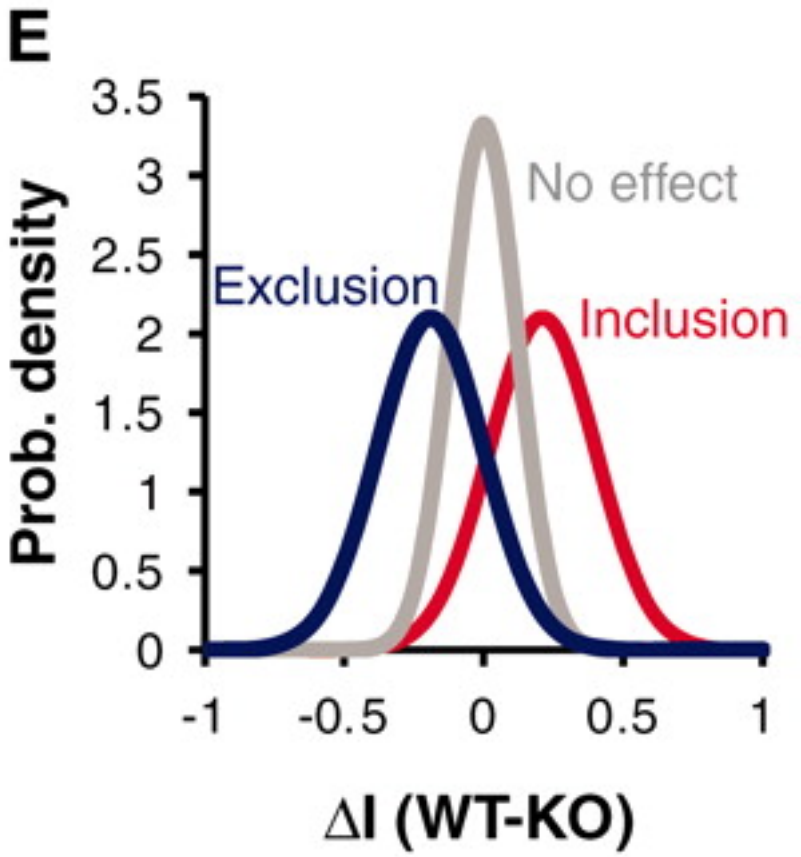


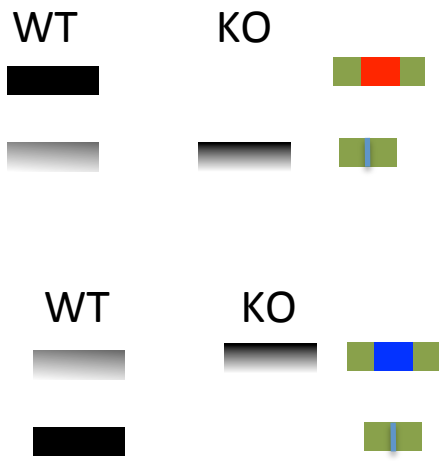
Exon exclusion

Bayesian Model

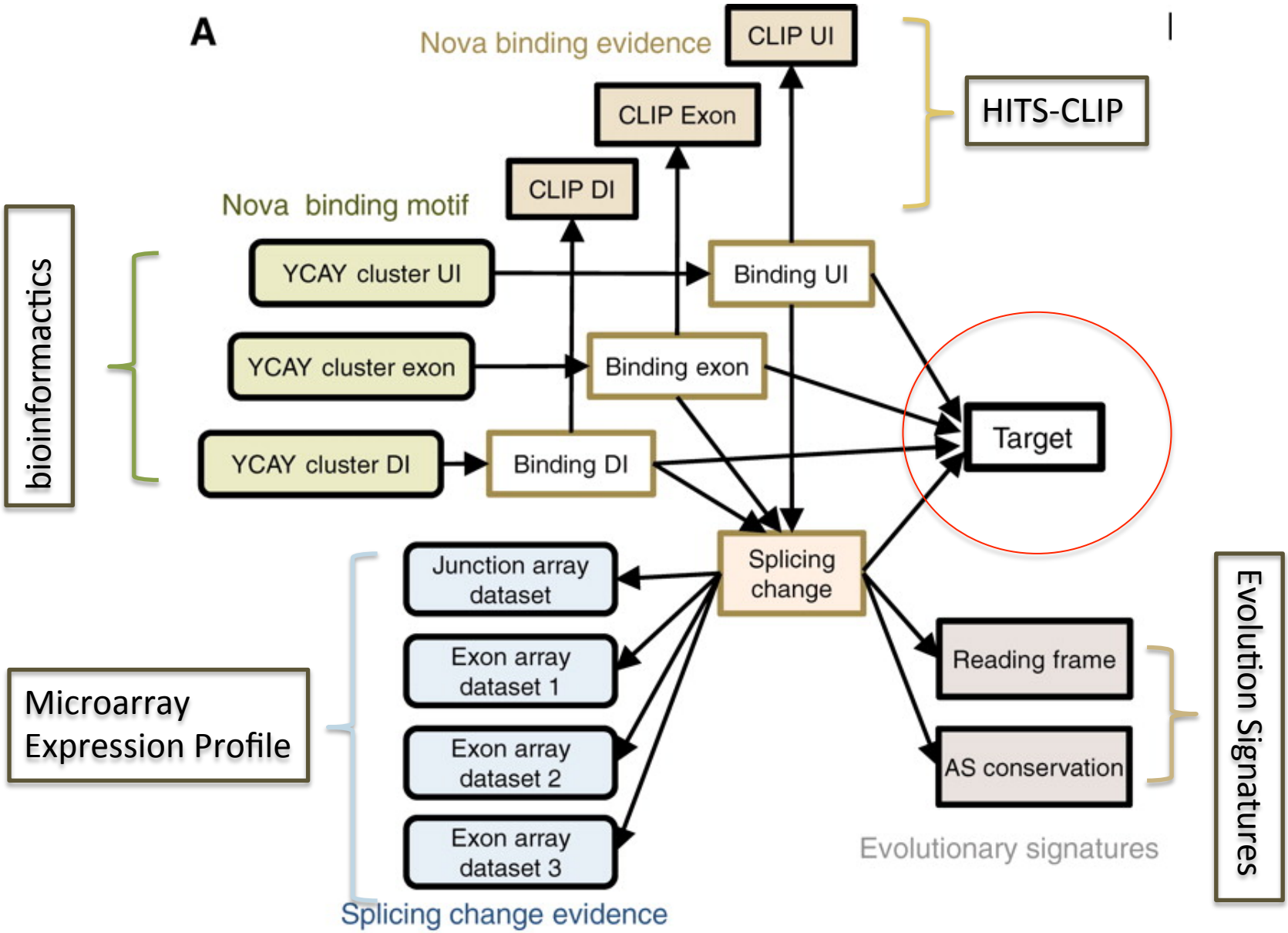


Splicing changes



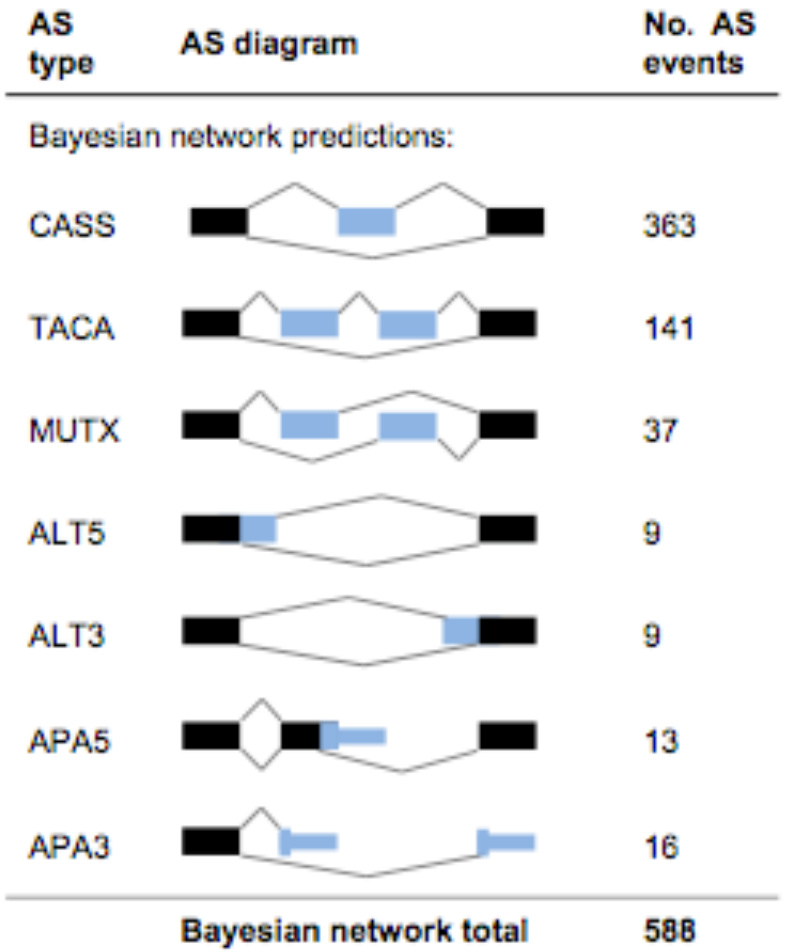


Bayesian Model



Predicted Nova-regulated targets

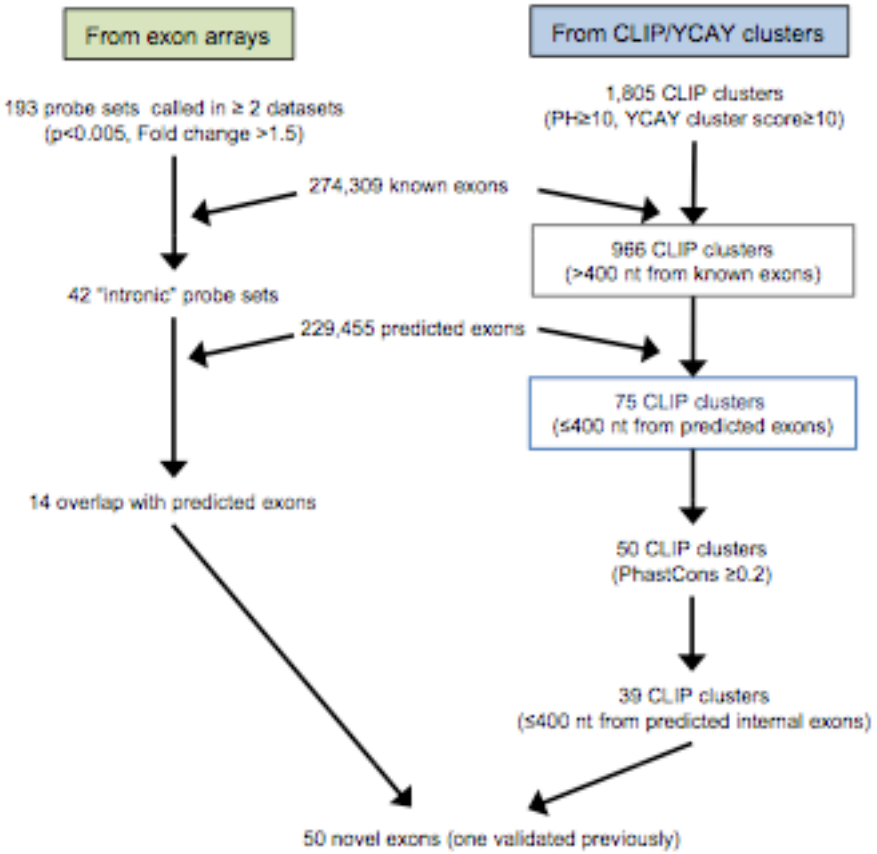
- 13,357 annotated cassette exons



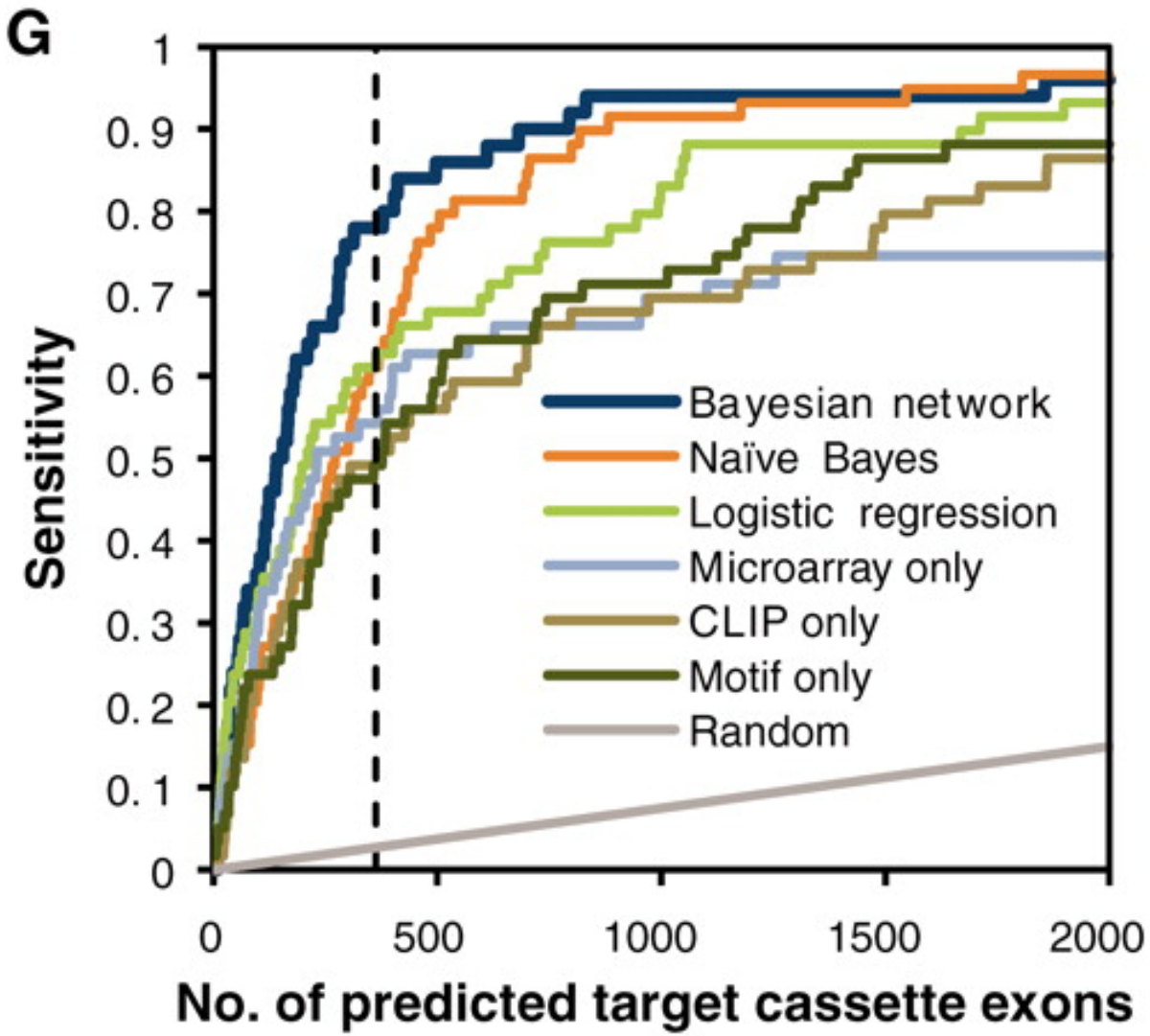
AS type	AS diagram	No. AS events
Bayesian network predictions:		
CASS		363
TACA		141
MUTX		37
ALT5		9
ALT3		9
APA5		13
APA3		16
	Bayesian network total	**588**

Novel Nova-regulated targets

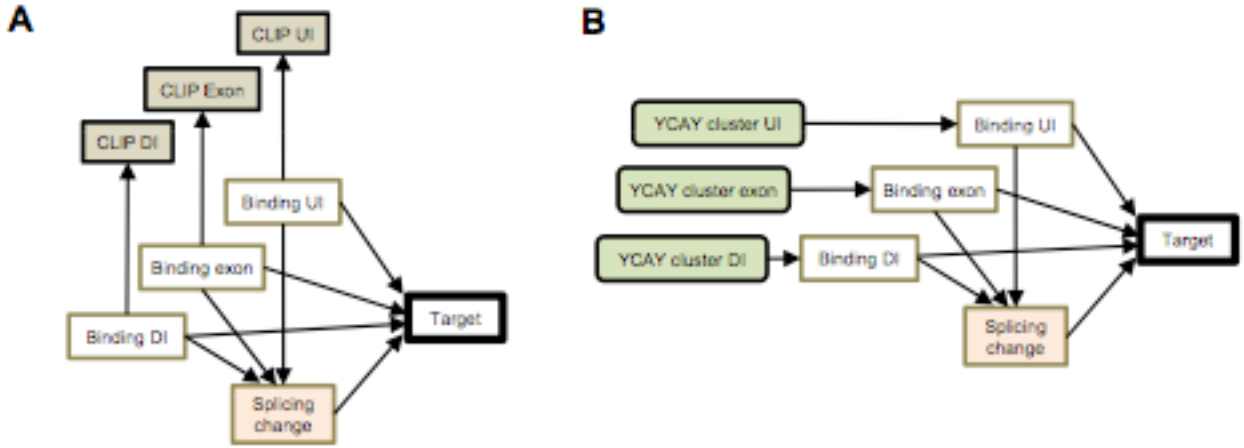
- Besides AS from database, searched novel exons with high sequence conservations.
- Additional 76 novel exons as Nova targets



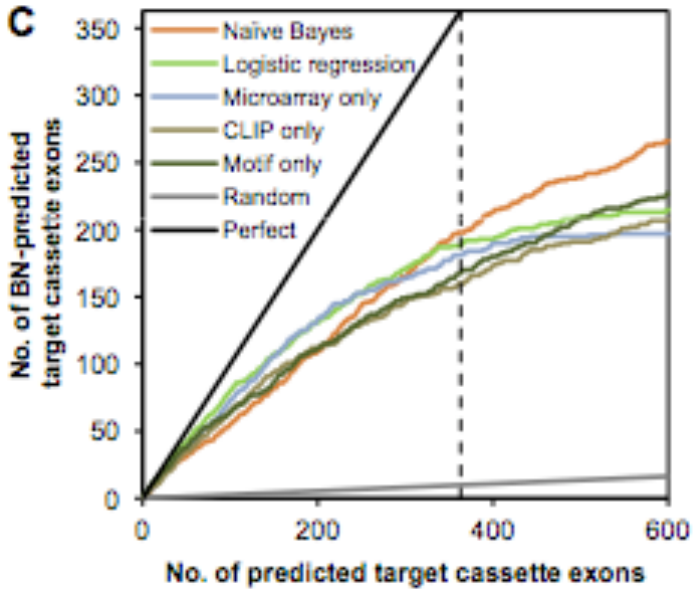
Prediction Performance



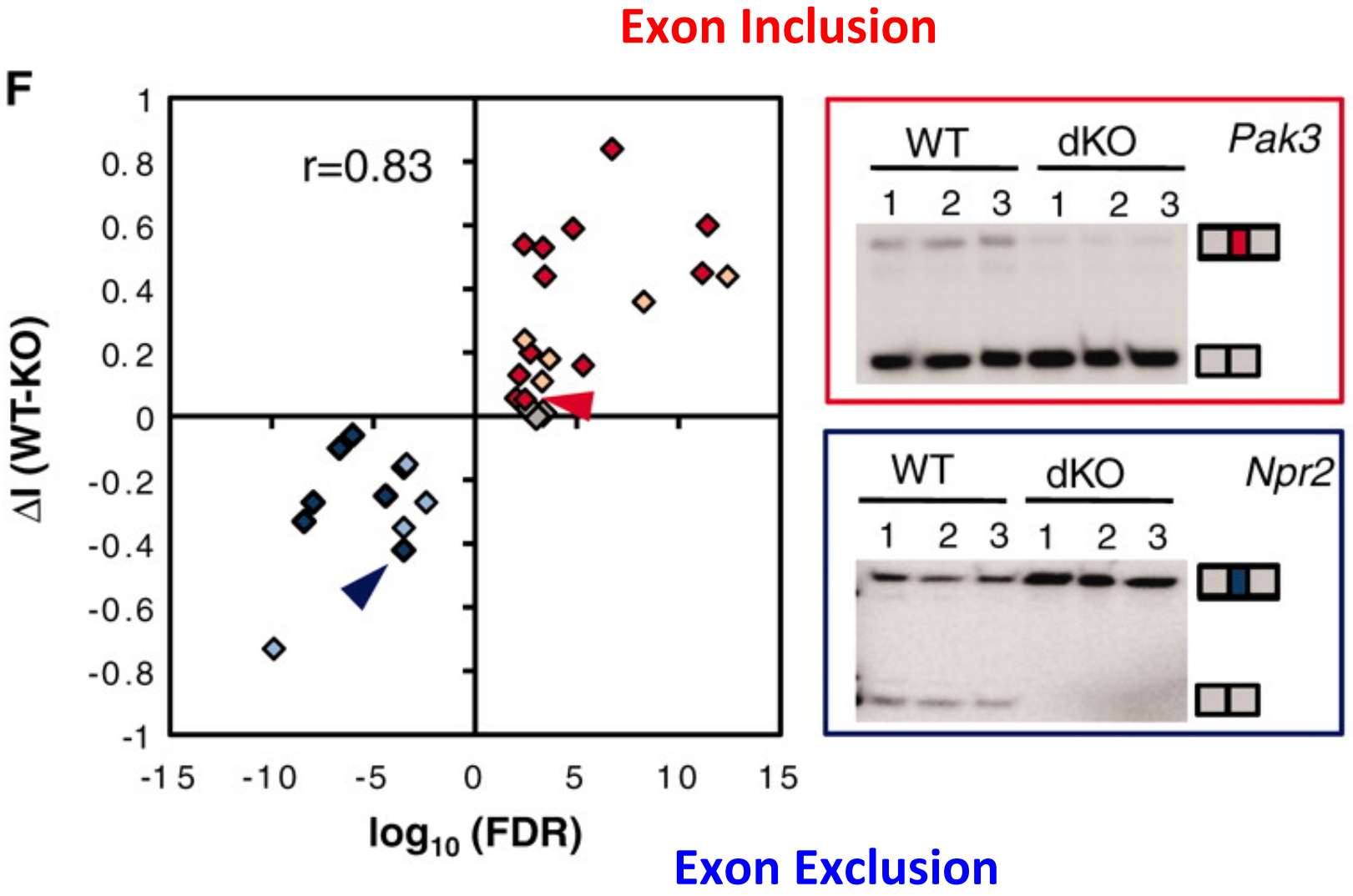
Reduced Bayesian Network



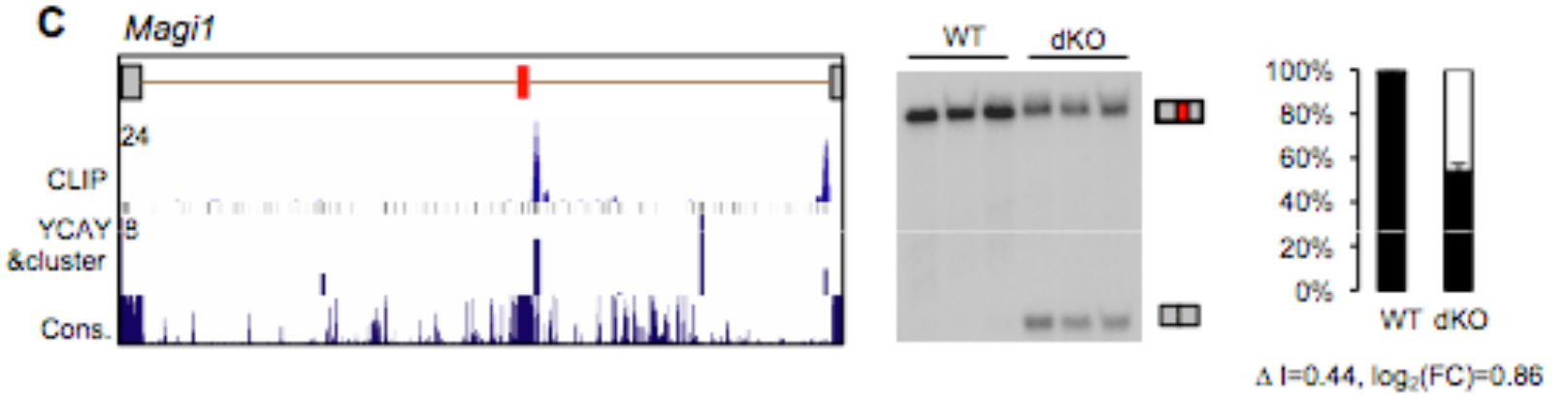
- Clip Data Only
- Motif Data Only



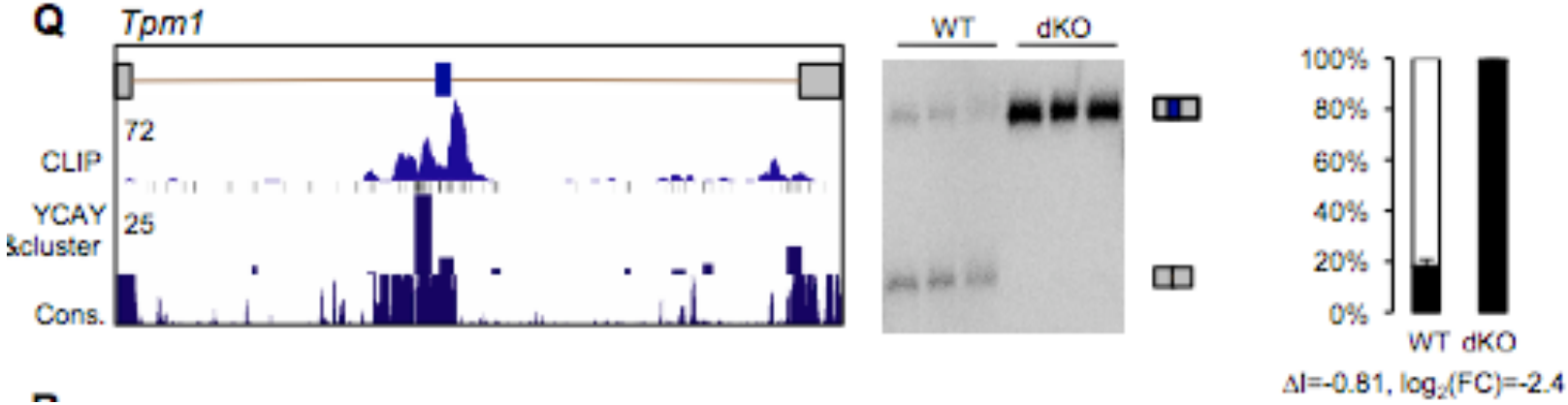
Experimental Validation



Two more Casset Exon Cases

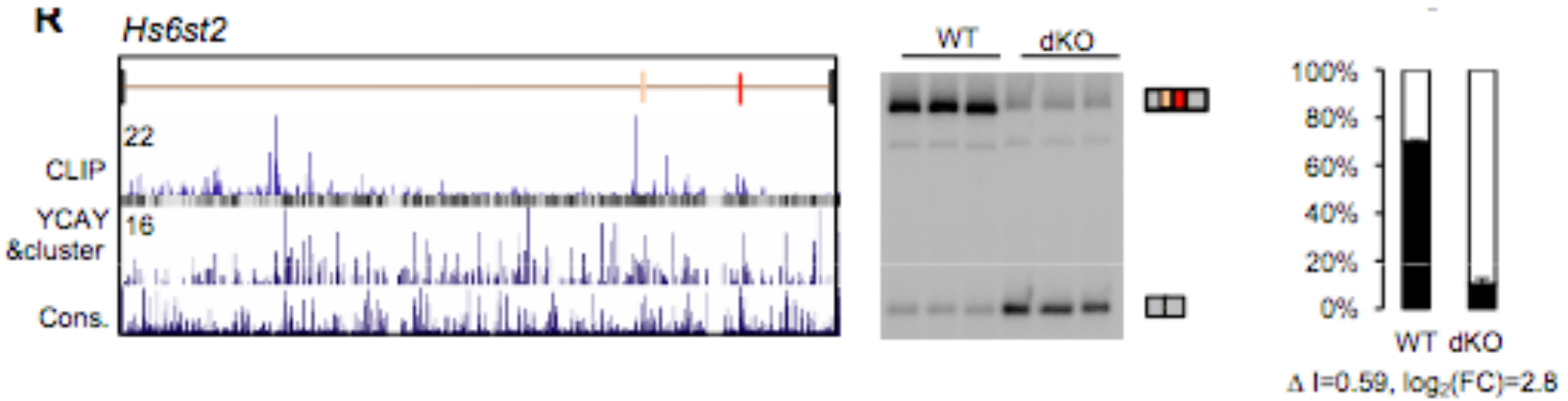


Exon Inclusion

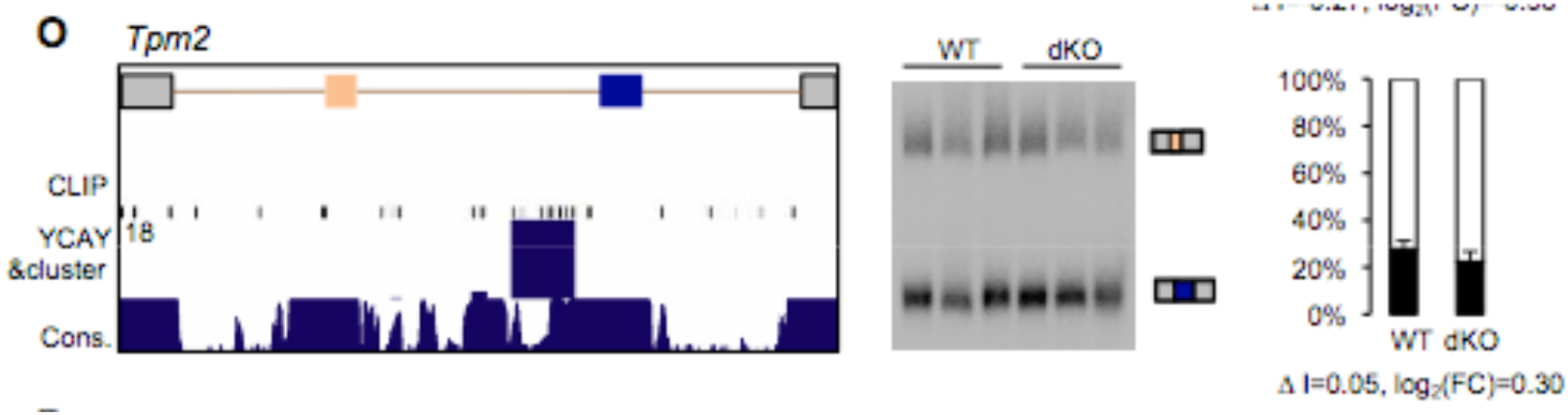


Exon Exclusion

Other examples



TACA **Exon Inclusion**



MUTX **Exon Exclusion**

Conservation regions

Intron Downstream 3’

Intron Downstream 5’

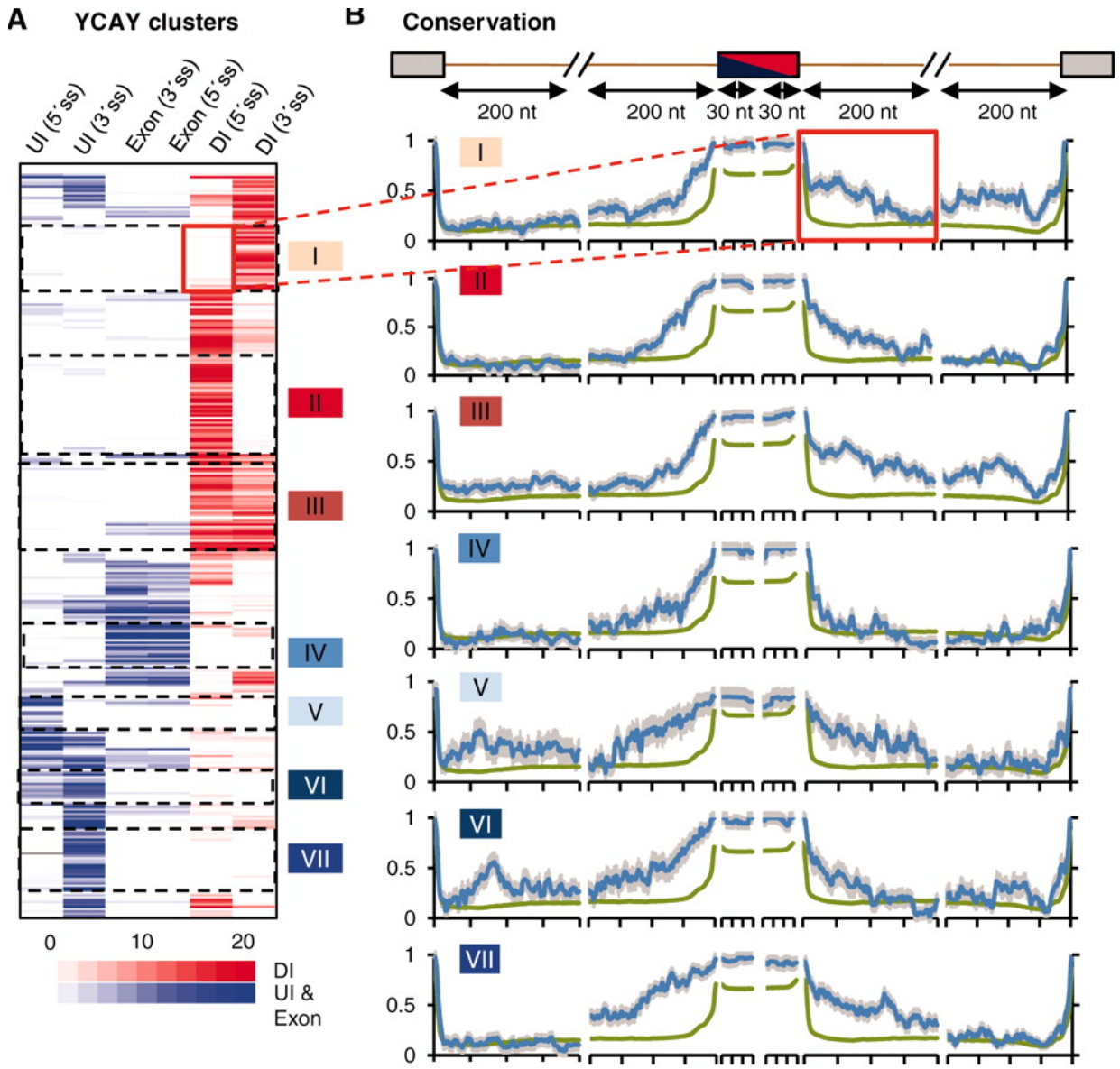
Intron Downstream 5’ and 3’

Exon

Intron upnstream 5’

Intron upnstream 5’ and 3’

Intron upnstream 3’



Functions of Nova targets

- Nova regulates alternative splicing of transcripts encoding synaptic proteins.
- Go-term analysis and KEGG metabolic pathway analysis confirmed this.
- It is unclear how Nova-regulated AS might effect the interactions between those synaptic proteins.
- Protein annotations revealed that about half Nova target transcripts encoded phosphoproteins.

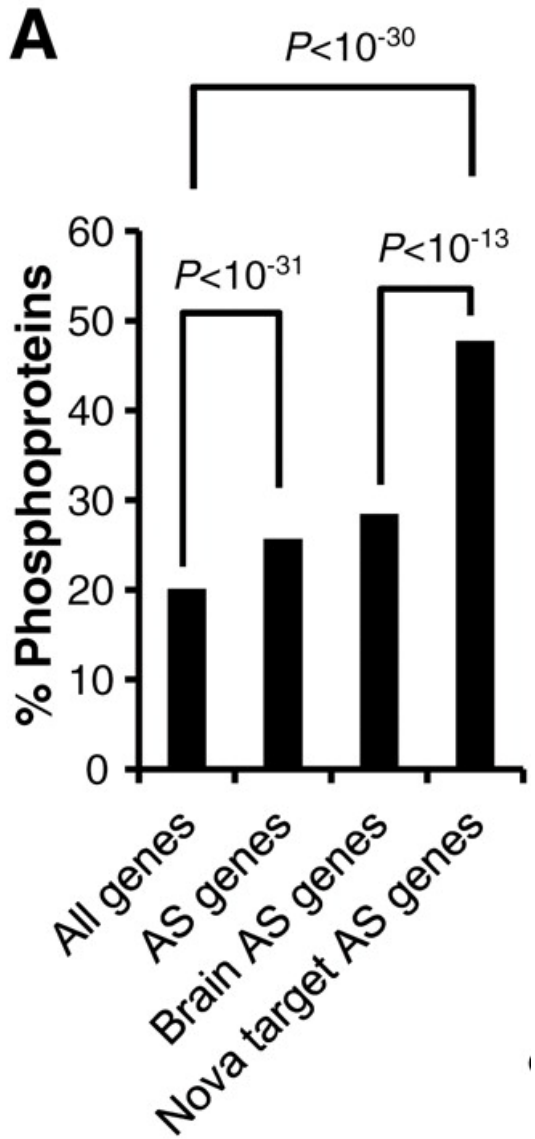
GO-term enrichment of Nova targets

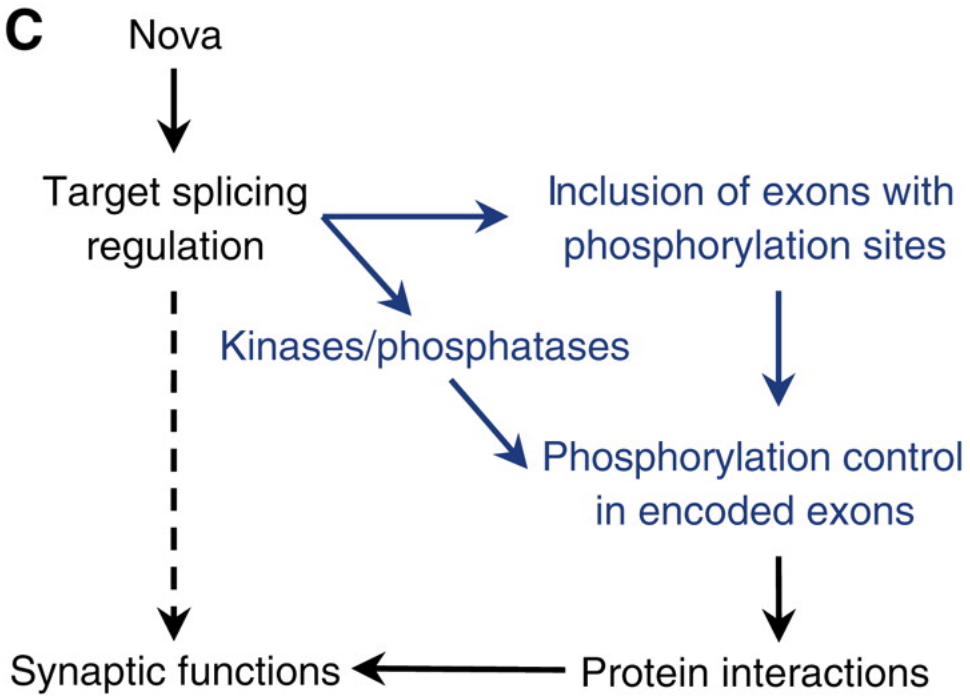
GO Term	Gene count	%	P-Value	Fold Enrichment	Benjamini FDR
Biological process					
GO:0016043~cellular component organization	93	26.05	4.02E-12	2.02	6.93E-09
GO:0007399~nervous system development	53	14.85	1.01E-09	2.48	8.72E-07
GO:0032989~cellular component morphogenesis	31	8.68	2.32E-09	3.56	1.33E-06
GO:0030030~cell projection organization	30	8.40	3.55E-09	3.60	1.53E-06
GO:0007268~synaptic transmission	22	6.16	8.39E-09	4.64	2.90E-06
GO:0048667~cell morphogenesis involved in neuron differentiation	22	6.16	1.18E-08	4.55	3.40E-06
GO:0051179~localization	104	29.13	1.63E-08	1.66	4.03E-06
GO:0000902~cell morphogenesis	28	7.84	1.64E-08	3.57	3.53E-06
GO:0019226~transmission of nerve impulse	24	6.72	2.47E-08	4.01	4.73E-06
GO:0007154~cell communication	34	9.52	4.41E-08	2.93	7.60E-06
Cellular component					
GO:0045202~synapse	38	10.64	1.64E-15	4.85	5.00E-13
GO:0044459~plasma membrane part	82	22.97	2.06E-15	2.51	3.16E-13
GO:0042995~cell projection	47	13.17	3.24E-14	3.62	3.24E-12
GO:0030054~cell junction	42	11.76	3.53E-14	4.00	2.65E-12
GO:0005856~cytoskeleton	61	17.09	6.36E-12	2.60	3.81E-10
GO:0005886~plasma membrane	104	29.13	6.82E-11	1.84	3.41E-09
GO:0042734~presynaptic membrane	11	3.08	1.07E-09	14.80	4.61E-08
GO:0016323~basolateral plasma membrane	19	5.32	3.11E-09	5.83	1.17E-07
GO:0044456~synapse part	23	6.44	5.65E-09	4.55	1.88E-07
GO:0043005~neuron projection	25	7.00	8.93E-09	4.09	2.68E-07
Molecular function					
GO:0005515~protein binding	198	55.46	9.43E-15	1.49	4.50E-12
GO:0008092~cytoskeletal protein binding	35	9.80	2.57E-10	3.51	6.14E-08
GO:0003779~actin binding	23	6.44	9.83E-07	3.40	1.56E-04
GO:0030695~GTPase regulator activity	26	7.28	1.17E-06	3.06	1.39E-04
GO:0060589~nucleoside-triphosphatase regulator activity	26	7.28	1.61E-06	3.01	1.54E-04

Pathway enrichment of Nova targets

KEGG pathway	Gene count	Fold Enrichment	Benjamini FDR	Genes
mmu04020 Calcium signaling pathway	17	3.5	0.001	*Atp2b1, Atp2b2, Cacna1c, Cacna1d, Cacna1b, Cacna1g, Camk2a, Camk2g, Camk2b, Grin1, Gnas, Plcb4, Ppp3cb, Ppp3cc, Ryr2, Slc8a1, Erbb4*
mmu04720 Long-term potentiation	10	5.3	0.003	*Cacna1c, Camk2a, Camk2g, Camk2b, Gria2, Grin1, Plcb4, Ppp1r12a, Ppp3cb, Ppp3cc*
mmu04514 Cell adhesion molecules (CAMs)	12	4.1	0.003	*Alcam, Cadm1, Cadm3, Mpzl1, Neo1, Nrxn3, Nfasc, Nfasc, Ptprf, Ptprm, Nlgn1, Nrcam,Nrxn1*
mmu04520 Adherens junction	10	4.4	0.006	*Actn4, Baiap2, Ctnna2, Ctnnd1, Pard3, Smad2, Smad4, Ptprf, Ptprm,Sorbs1*
mmu04360 Axon guidance	13	3.3	0.006	*Ablim1, Cxcl12, Dcc, Epha5, Efna5, Ablim2, Ntng1, Pak3, Ppp3cb, Ppp3cc, Arhgef12, Robo2, Unc5c*
mmu04912 GnRH signaling pathway	10	3.6	0.017	*Cacna1c, Cacna1d, Camk2a, Camk2g, Camk2b, Gnas, Mapk8, Mapk9, Map2k4, Plcb4*
mmu04310 Wnt signaling pathway	12	2.8	0.032	*Apc, Camk2a, Camk2g, Camk2b, Smad2, Smad4, Mapk8, Mapk9, Plcb4, Porcn, Ppp3cb, Ppp3cc*
mmu04930 Type II diabetes mellitus	6	5.2	0.048	*Cacna1c, Cacna1d, Cacna1b, Cacna1g, Mapk8, Mapk9*
mmu04260 Cardiac muscle contraction	7	4.2	0.049	*Tpm2, Cacna1d, Cacna1c, Tpm1, Ryr2, Slc8a1, Tpm3*
mmu04012 ErbB signaling pathway	8	3.3	0.074	*Camk2a, Camk2g, Camk2b, Mapk8, Mapk9, Map2k4, Pak3, Erbb4*
mmu04530 Tight junction	9	2.7	0.01	*Actn4, Cask, Ctnna2, Pard3, Epb4.1, Epb4.1l1, Epb4.1l2, Epb4.1l3, Magi1*

Nova targets - phosphoproteins





Applications of Bayesian Network

Can we apply Bayesian Network into our research?

- Next generation sequencing data, such as RNA-seq, Chip-seq etc.
- Microarray data
- Motif data, for example, TF binding sites, miRNA sequences etc.
- Genome sequence data, Ath, Maize, Rice, Soybean etc.

Summary

- Recent technological advances present challenging and interesting biological data at molecular level.
- Statistics and multivariate analysis play an important role in understanding and extracting knowledge from these type of data.
- Integrative analysis is even more challenging and we presented some solutions to these challenges. There is plenty of room for improvement.