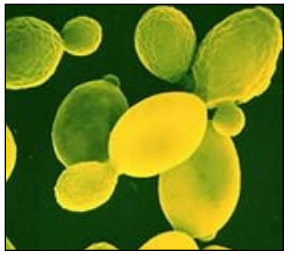
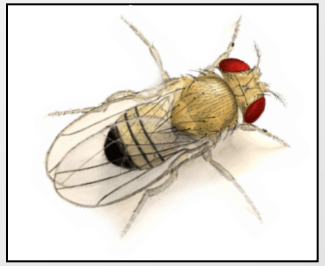
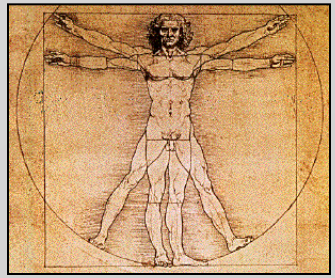




Drosophila modENCODE integrative analysis and insights into human disease

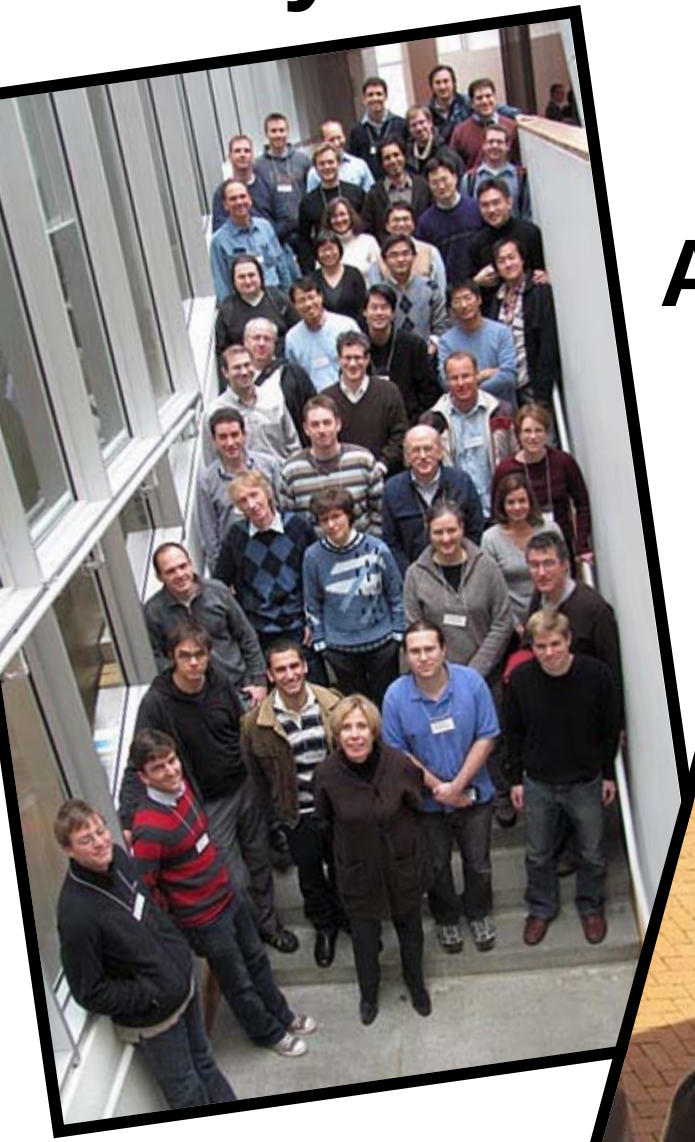
Manolis Kellis



Broad Institute of MIT and Harvard

MIT Computer Science & Artificial Intelligence Laboratory

Analysis Working Group



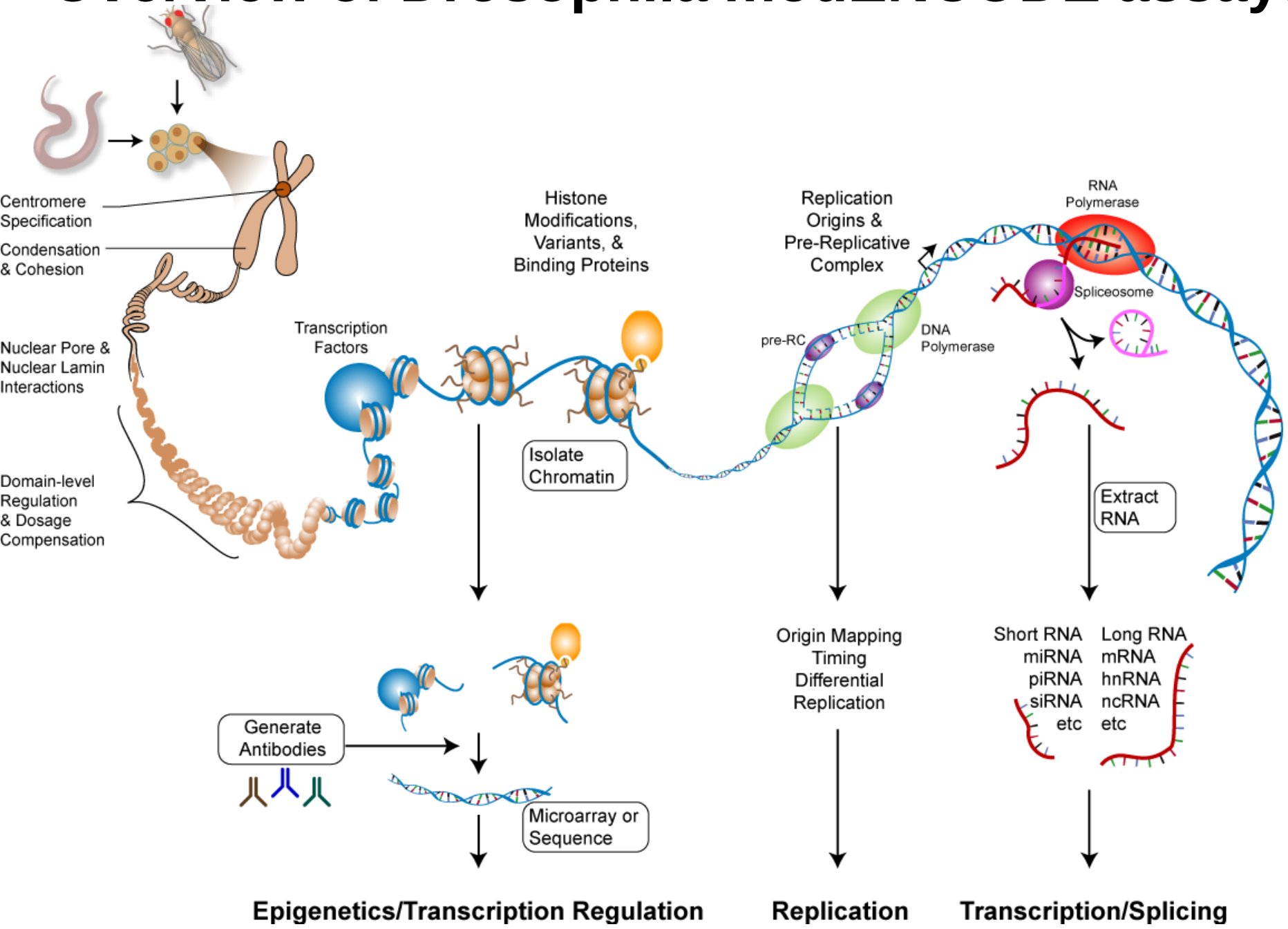
AWG
DAC

Data Analysis Center

Acknowledgements

	Fly	Worm
Transcripts (Celniker/Waterston)	Joe Carlson Jane Landolin Ben Booth Brenton Graveley Ben Brown	Mark Gerstein LaDeana Hillier Kevin Yip, Ashish Agarwal Lukas Habegger
Chromatin (Karpen/Lieb)	Peter Park, Peter Kharchenko, Jason Ernst, Matthew Eaton	Shirley Liu Hyunjin (Gene) Shin
TFs (White/Snyder)	Casey Brown Nicolas Negre	Mark Gerstein Lucas Lochovsky Kevin Yip
Nucleosomes (Henikoff)	Steve Henikoff	Steve Henikoff
smRNAs/3'UTRs (Lai/Piano)	Eric Lai (smRNAs) Nicolas Robine	Kris Gunsalus (3'UTRs) Arun Manoharan Marco Mangone
Origins (MacAlpine)	MacAlpine Mathew Eaton	N/A
Statistics/Infrastructure (Bickel/Gerstein)	Ben Brown and Kevin Yip and Nathan Boley	
Conservation/Submissions	Lincoln Stein, Gos Micklem, DCC	

Overview of Drosophila modENCODE assays



Drosophila modENCODE datasets

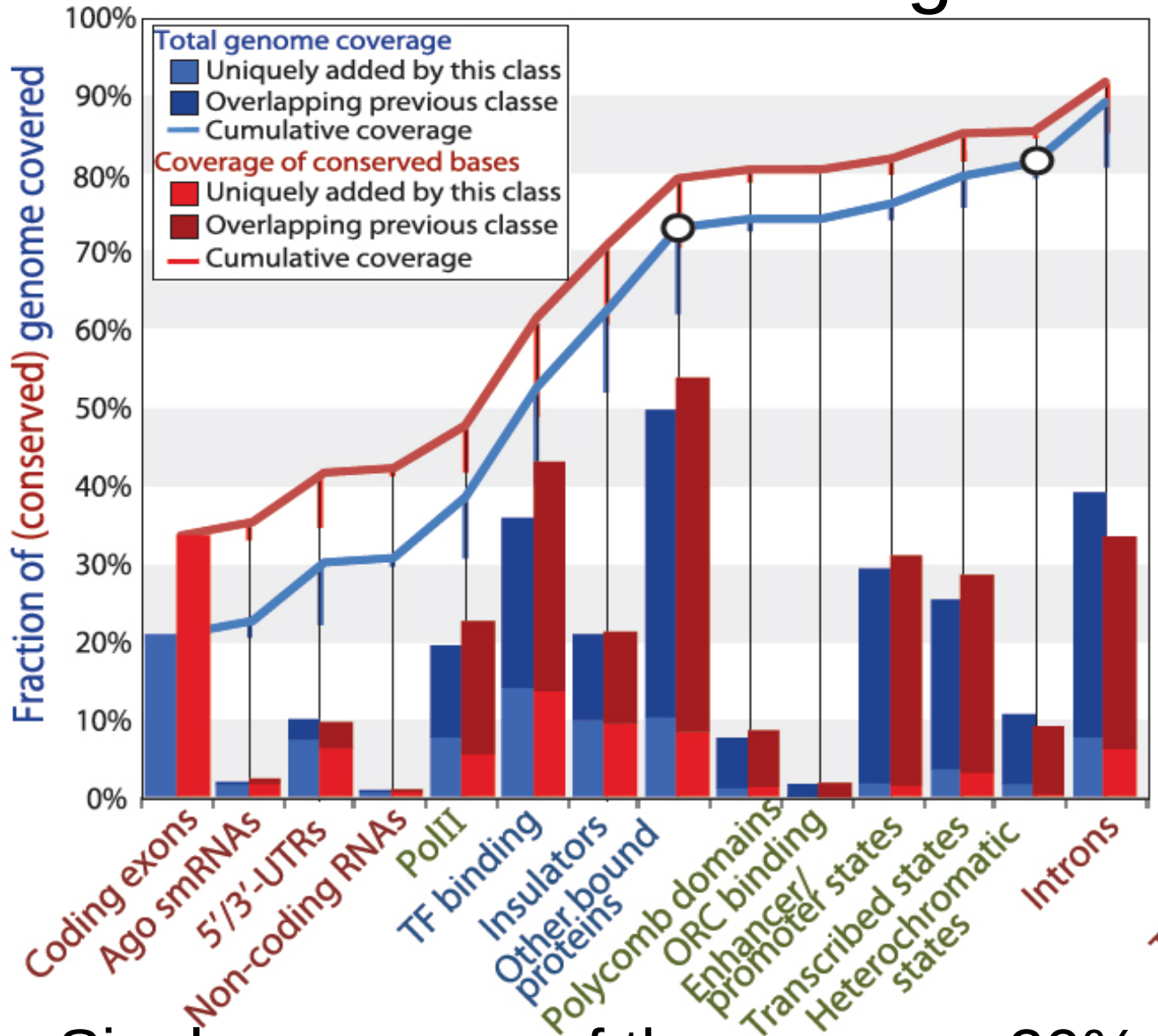
Category	Annotation Types	Assay	Num. Expts	Cell lines					Dev Stages[2]					Tissue	Factors[3,7]
				BG3	CI.8+	KC167	S2-DRSC	Others[1]	Embryo	Larva	Pupa	AdultM	AdultF		
Transcription/Splicing															
Gene Expression	gene transcripts, cell type- and tissue-specific expression	RNA tiling array (polyA+, total RNA)	74	+	+	+	+	22	12	6	6	3	3		
	coding and non-coding genes, transcripts, splice junctions	Total RNA-seq	12	+	+	+	+		12						
		RNA-seq (polyA+)	36	+	+	+	+		12	9	3	3	3		
		cDNA sequencing	3204[4]				+								
	splicing regulatory targets	RNAi/RNA-seq	27				+								**26 splicing factors:** B52, BL, CG17838, CG30122, CG7878, CG7971, CG9373, CG9983, GLO, HEPH, HRB87F, MSI, MUB, QKR58E-2, RBP1, RBP1-like, RSF1, SC35, SF1, SQD, SRP54, TRA2, UPF1, XL6, YTR, YU
	TSS	CAGE (polyA+)	1						1						
		RACE-seq (total RNA)	8727[4]						1			1	1		
Short ncRNAs	ncRNAs	small RNA-seq	76	+	+	+	+	17	9	4	4	4	3	21[5]	**2 processing factors:** AGO1, AGO2
Replication															
Pre-RC Factors	replication origins	ChIP-chip	2			+									**2 replication factors:** ORC2, MCM2-7
		ChIP-seq	5	+		+	+								**2 replication factors:** ORC2, MCM2-7
Repl. origins		BrdU-chip	3	+		+	+								
Timing	early/late domains	BrdU-chip	3	+		+	+								
Copy Number Variation	differential replication	CGH	5				+		1					2[6]	
		CNV-seq	3	+		+	+								
Epigenetics / Transcriptional Regulation															
Transcription Factors	TF binding sites, conserved binding	ChIP-chip	66		+	+	+		53	2			1		**37 TFs & co-factors:** BAB1, BRE1, CAD, CG8478, CHINMO, CNC, CTBP, D, DFD, Dll, DISCO, EN, EVE, EXD, FTZ-F1, GRO, GATAE, GSB-N, H, HKB, INV, KNI, KR, JUMU, MBD-R2, PIWI, RUN, SENS, SBB, SIN3A, STAT92E, TRL, TTK, TLL, UBX, WDS, ZFH1
		ChIP-seq	5						2	1	2				**2 TFs:** CAD, ECR
Chromosomal Proteins	chromatin and chromosomal functions	ChIP-chip	115	+	+	+	+		38	8	3	1	2		**11 histone modifying enzymes:** ASH1, HDAC3, HDAC4, HDAC6, HDACX, E(Z), JIL-1, NEJ, RPD3, SU(VAR)3-9, TRX **2 histone modification binding proteins:** HP1A, PC **7 nucleosome remodeling:** BRM, ISWI, Mi-2, NURF301, MRG15, SNR1, SPT16 **5 insulators:** BEAF-32, CP190, CTCF, SU(HW), MOD(MDG4) **7 others:** CHRO(CHRIZ), HP1c, HP2, PCL, PSC, RNA Polymerase II, SCE/DRING
		ChIP-seq	28						19	5	2	1	1		**5 histone modifying enzymes:** HDAC4, HDACX, HDAC6, HDAC3, NEJ **1 other:** RNA Polymerase II
Histone Modification	active chromatin, enhancing regions, promoters	ChIP-chip	161	+	+	+	+		58	23	5	6	9	2	**21 histone marks:** H1, H2BUbiq, H3K18ac, H3K23ac, H3K27ac, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me1, H3K79me2, H3K9ac, H3K9me2, H3K9me3, H4, H4AcTetra, H4K16ac, H4K5ac, H4K8ac
		ChIP-seq	79						46	18	6	4	5		**6 histone marks:** H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3
Nucleosome Solubility and Turnover	nucleosome occupancy	DNA tiling array	35			+	+	4	6						**6 NaCl fractions (mM):** 80, 80-150, 150, 150-600, 600, 600 mM (pellet), mononucleosomes **6 CATCH-IT conditions:** MET, 20 min, 40 min, 60 min, pulse, AHA 3 hr **2 histone variants:** H3.3, H2Av

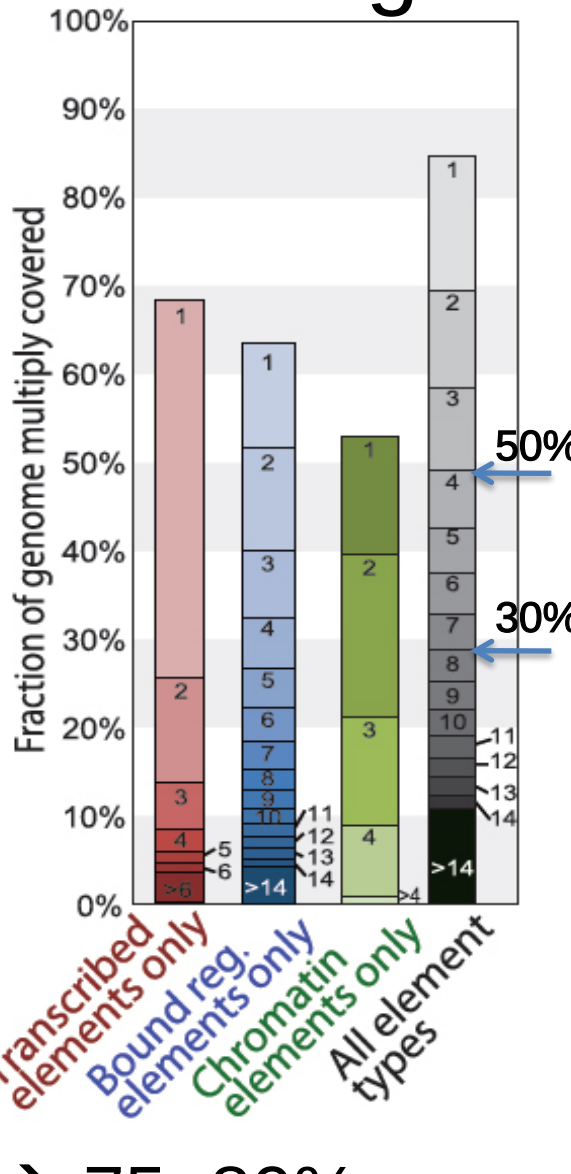
- ~1000 datasets with links to data download (2010)

Drosophila Data Set Submissions (June 2012)

PI	Project	Data Types	Total
S. Celniker	Transcriptome	mRNA, ncRNA, hnRNA; treatments	898
E. Lai	Small RNAs	miRNA, siRNA, piRNA	75
B. Oliver	Comparative Transcriptome	pseudoobscura vs. melanogaster	30
S. Henikoff	Histone Variants	variants, nucleosome turnover	48
G. Karpen	Chromatin	histone modifications, chromosomal proteins	593
D. MacAlpine	Replication	complexes, origins, timing, differential replication	45
K. White	Regulation	transcription factors	474
		TOTAL:	2,163

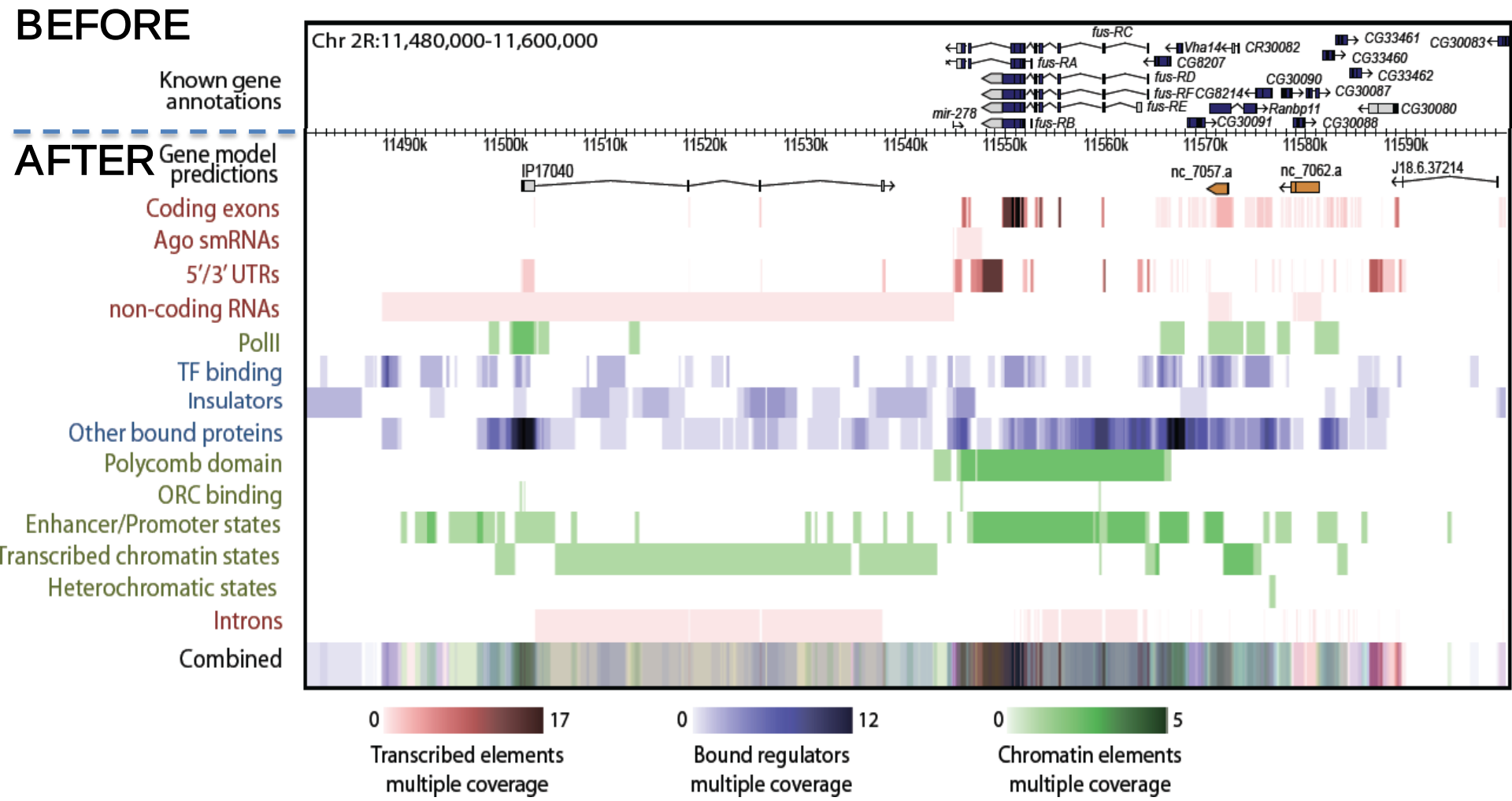
Combined increase in genome coverage





- Single coverage of the genome: 20% → 75~80%
- Multiple coverage: 50% >4 annotations, 30% >8 annot

Annotation: New regions come to life



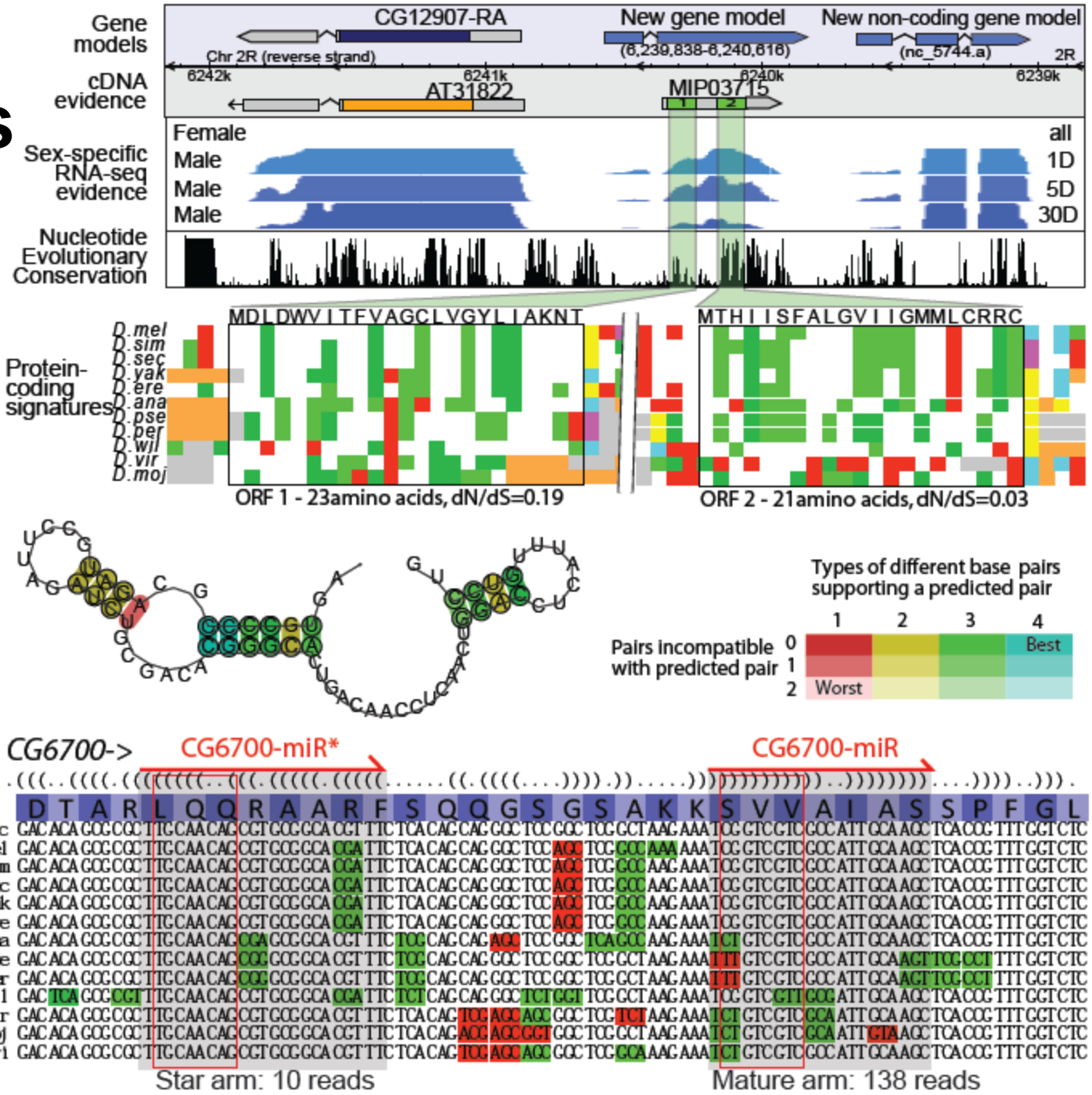
- Goal of modENCODE: Encyclopedia of DNA elements
- Expand annotation of coding, non-coding genome

Insights from integrative analysis

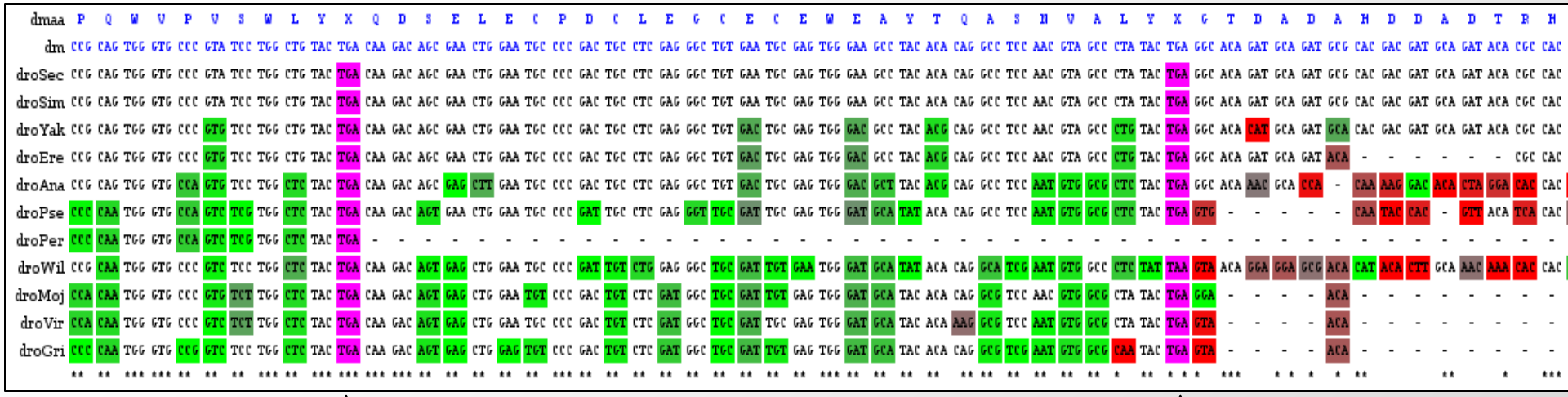
1. **Annotate coding/non-coding genes**
   - Peptides, structures, microRNAs, readthrough
2. **Annotate chromatin regulatory regions**
   - Enhancers, promoters, diversity of functions
3. **Define regulator targets and networks**
   - Hierarchy, TF/miRNA networks, HOT regions
4. **Predictive models of gene regulation**
   - Functional nets → gene function/expression
5. **Implications for human disease**
   - Annotate non-coding SNPs, link to TFs/targets

Genes and Transcripts

- Coding genes
  - ➔ 20AA peptides
- Structured & non-coding RNAs
  - ➔ roX2, HSRω
- microRNAs
  - ➔ New/hybrid mirtrons
  - ➔ within coding exons



Evidence of translational read-through in fly/human



- **New mechanism of post-transcriptional control.**
  - Hundreds of fly genes, handful of human genes.
  - Enriched in brain proteins, ion channels.
  - Initial experiments show potential ADAR role (Reenan Lab).
- **Many questions remain**
  - A-to-I editing of stop codon TAG|TGA|TAA → TGG
  - Cryptic splice sites? RNA secondary structure?

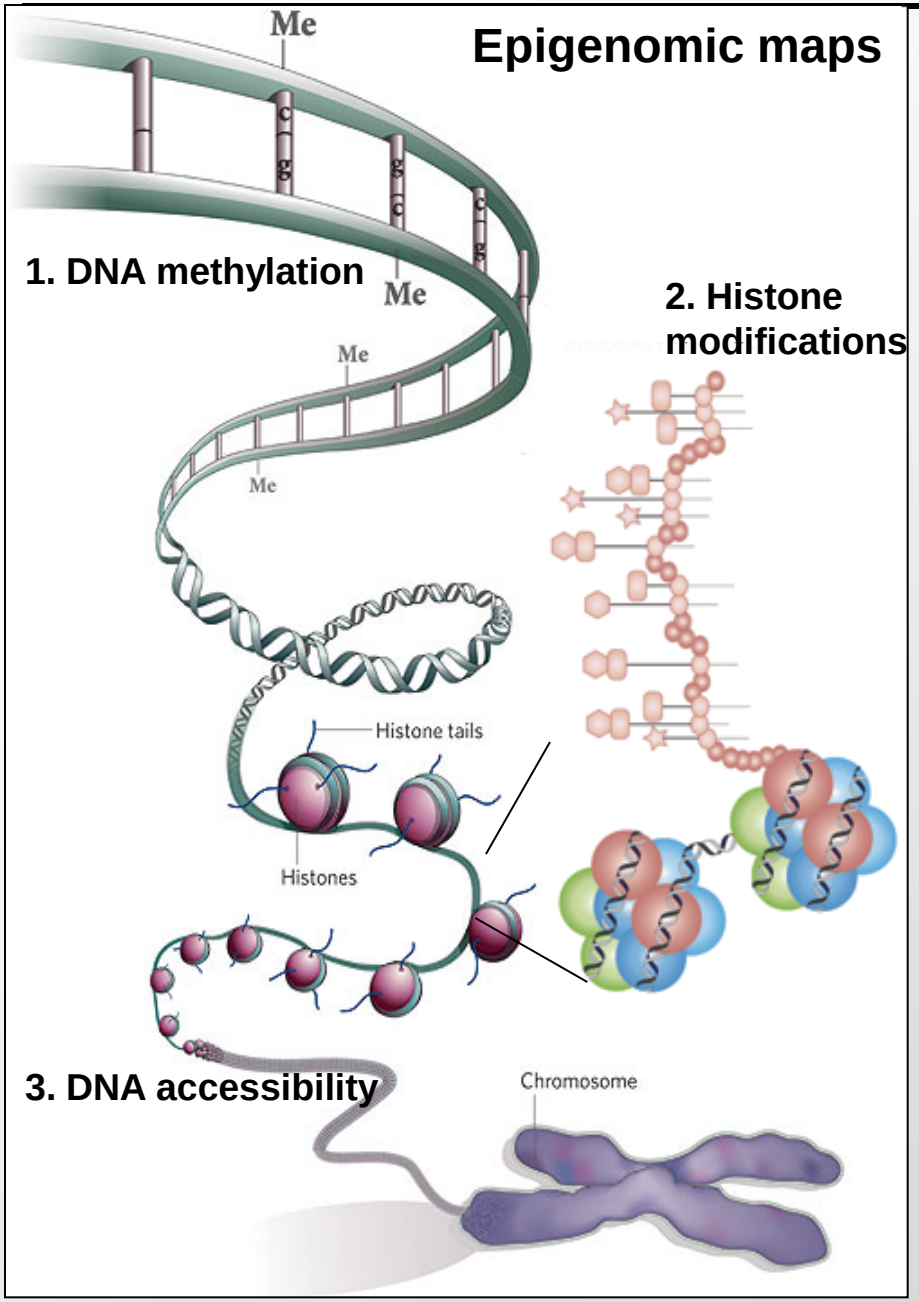
Jungreis, Lin, *et al*, Genome Research 2011

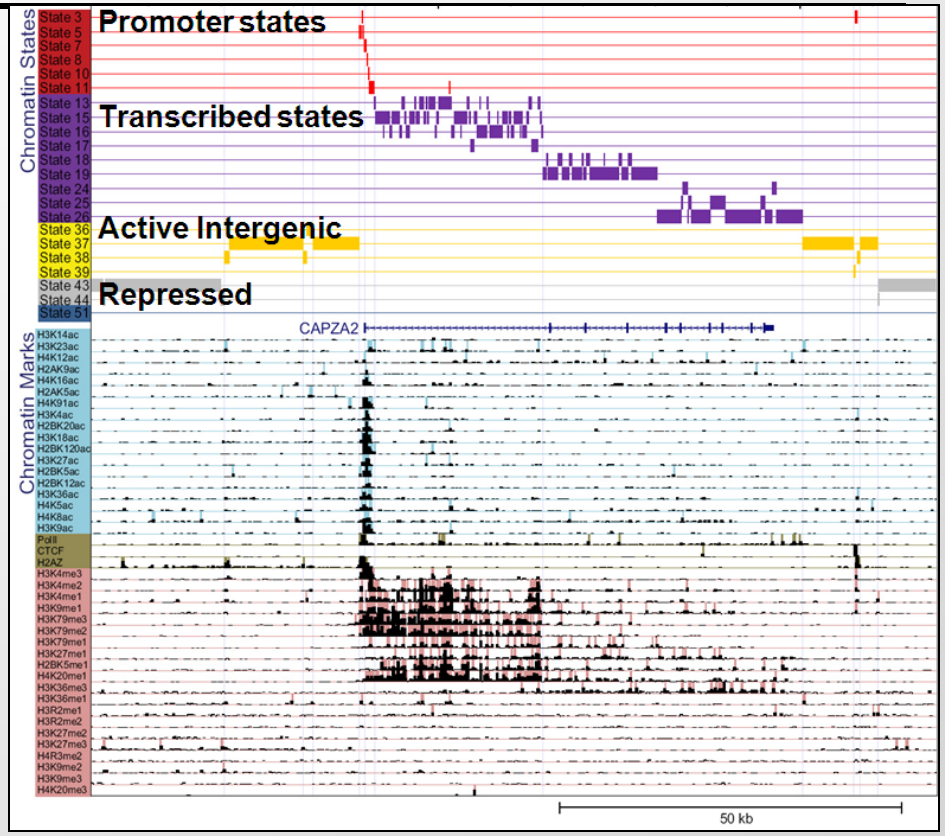
Insights from integrative analysis

1. **Annotate coding/non-coding genes**
   - Peptides, structures, microRNAs, readthrough
2. **Annotate chromatin regulatory regions**
   - Enhancers, promoters, diversity of functions
3. **Define regulator targets and networks**
   - Hierarchy, TF/miRNA networks, HOT regions
4. **Predictive models of gene regulation**
   - Functional nets → gene function/expression
5. **Implications for human disease**
   - Annotate non-coding SNPs, link to TFs/targets

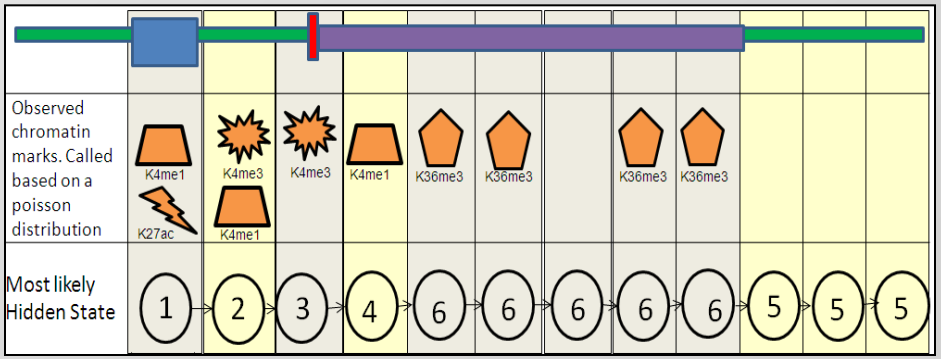
Chromatin states for systematic genome annotation

Epigenomic maps



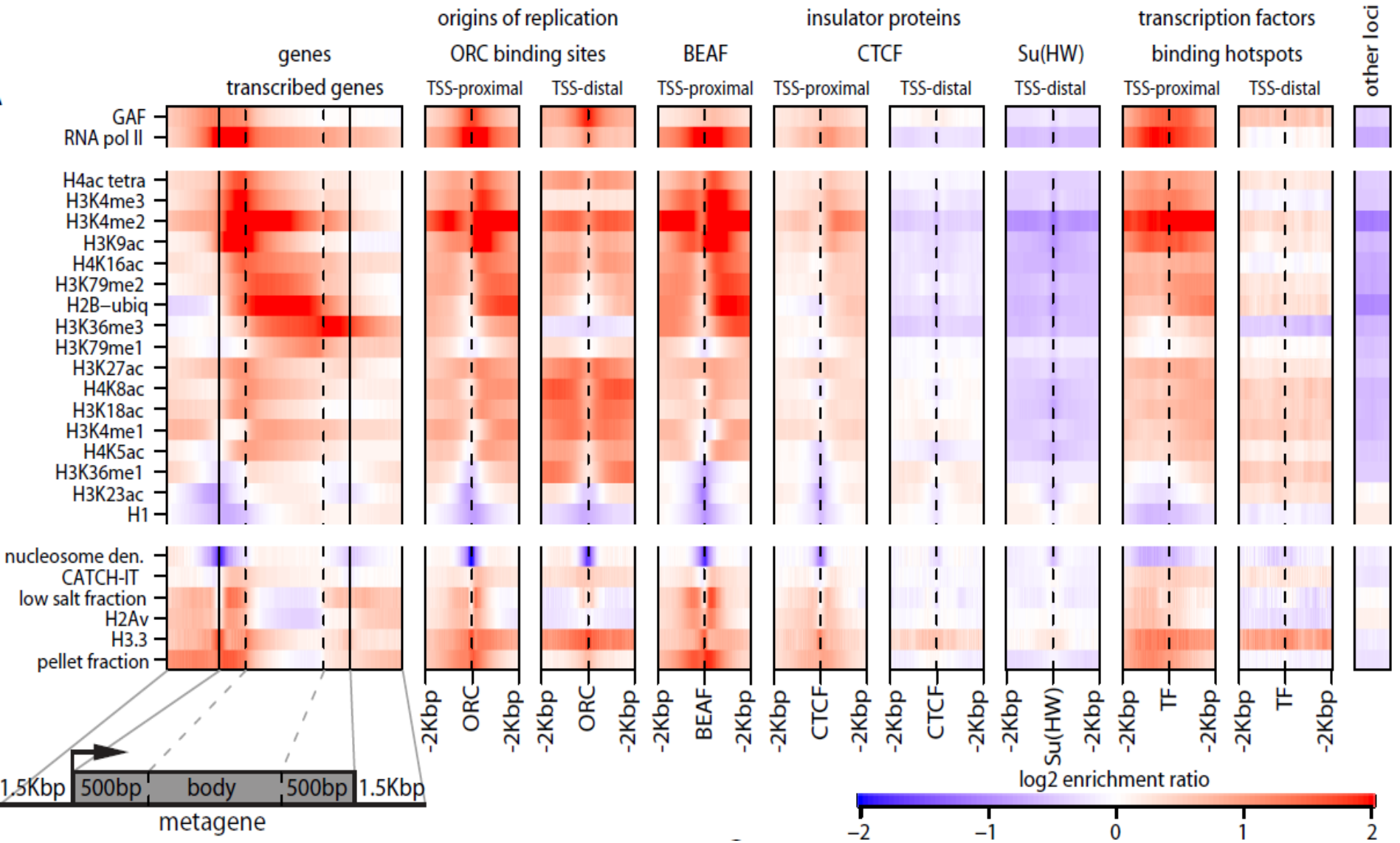


Ernst et al Nature Biotech 2010



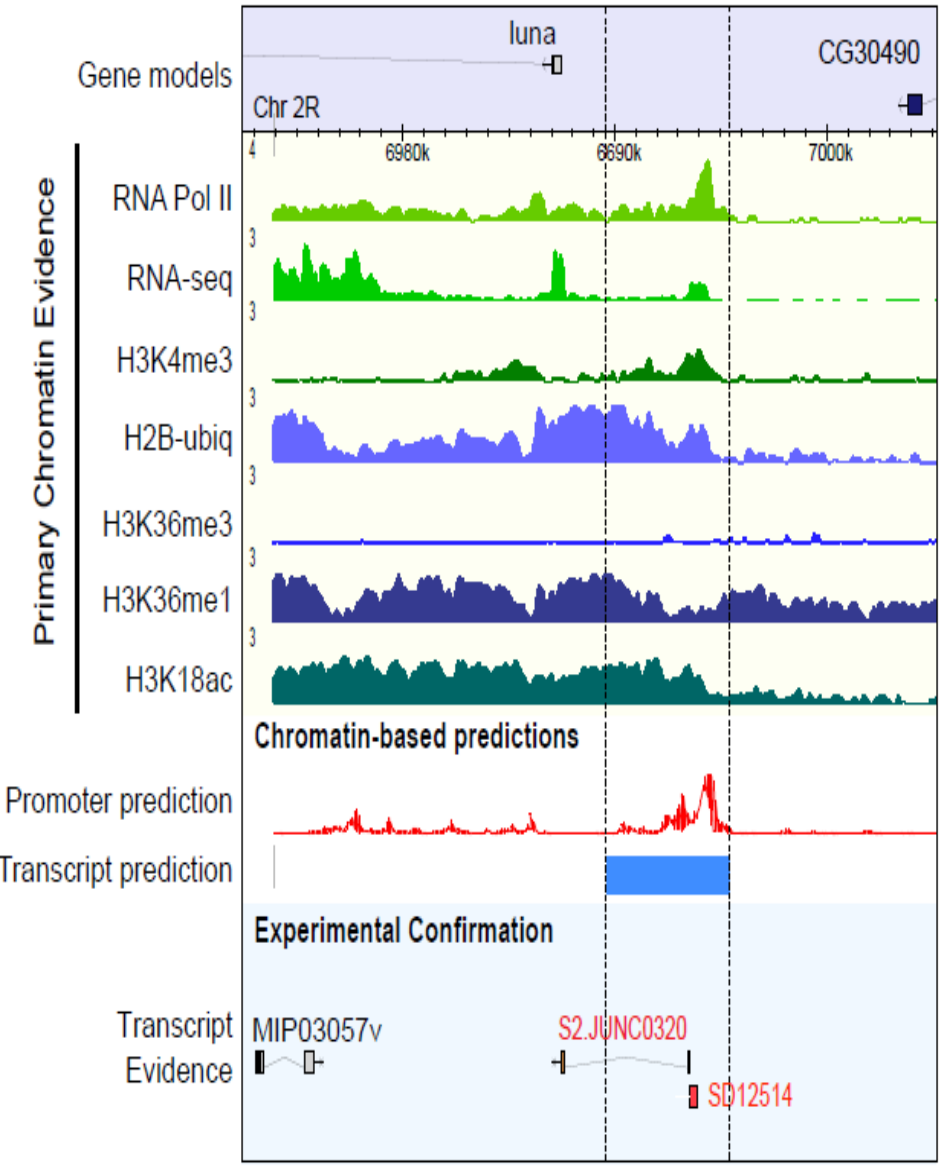
Also: Peter Park, Amos Tanay, Bill Noble.

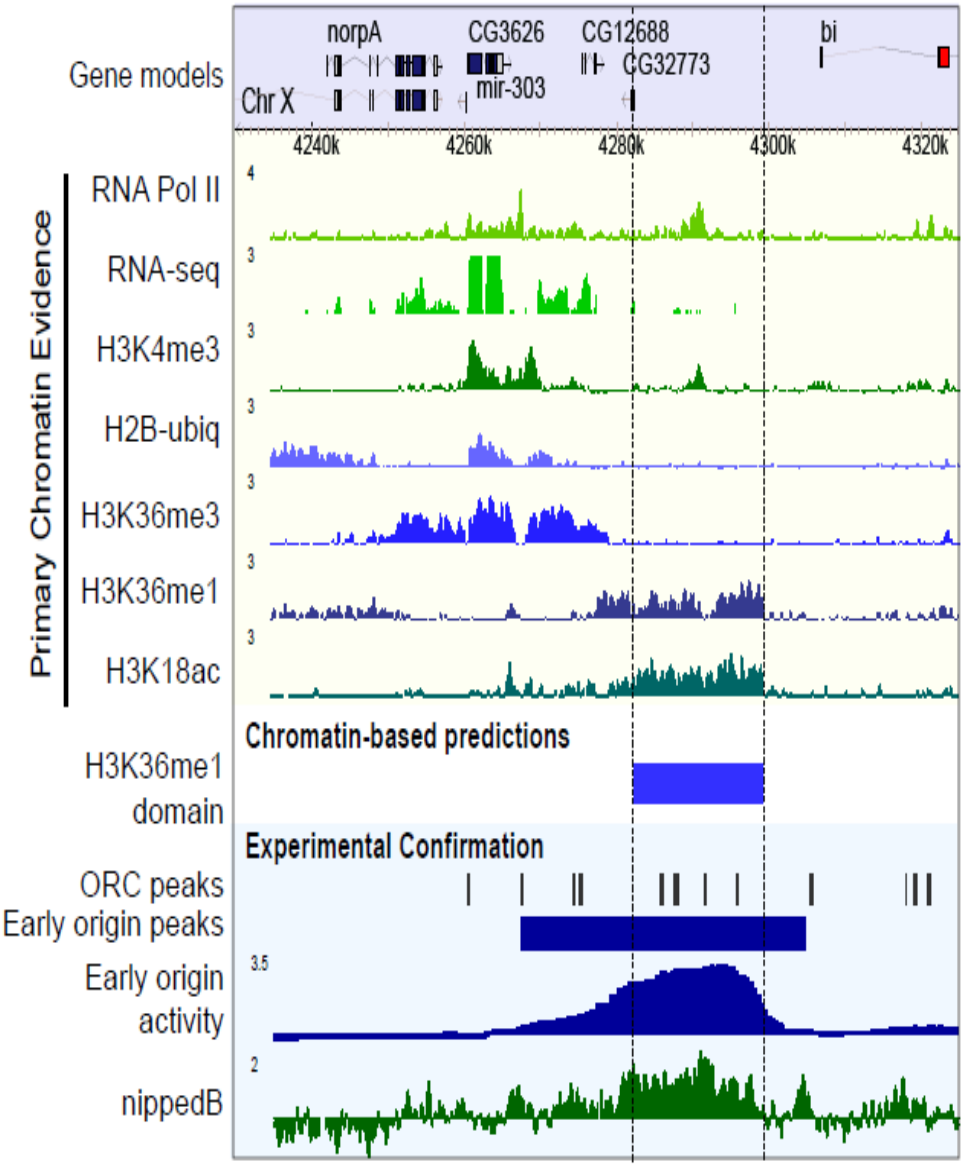
Characteristic chromatin marks / domains



Chromatin signatures predictive of different classes of elements

Examples of new / surprising elements





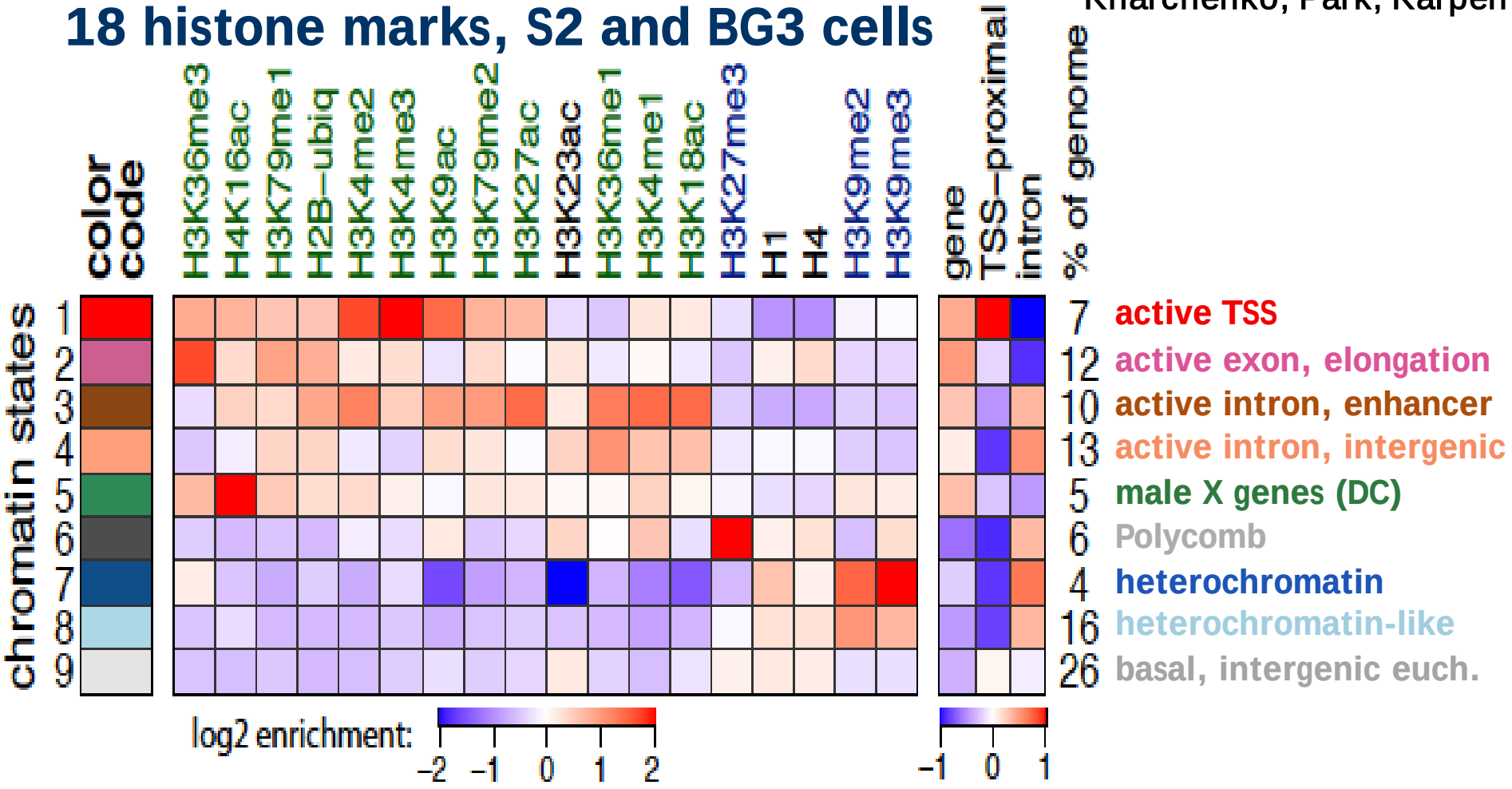
Promoter signatures

H3K36me1 repl. origins

Nine intensity-based chromatin states

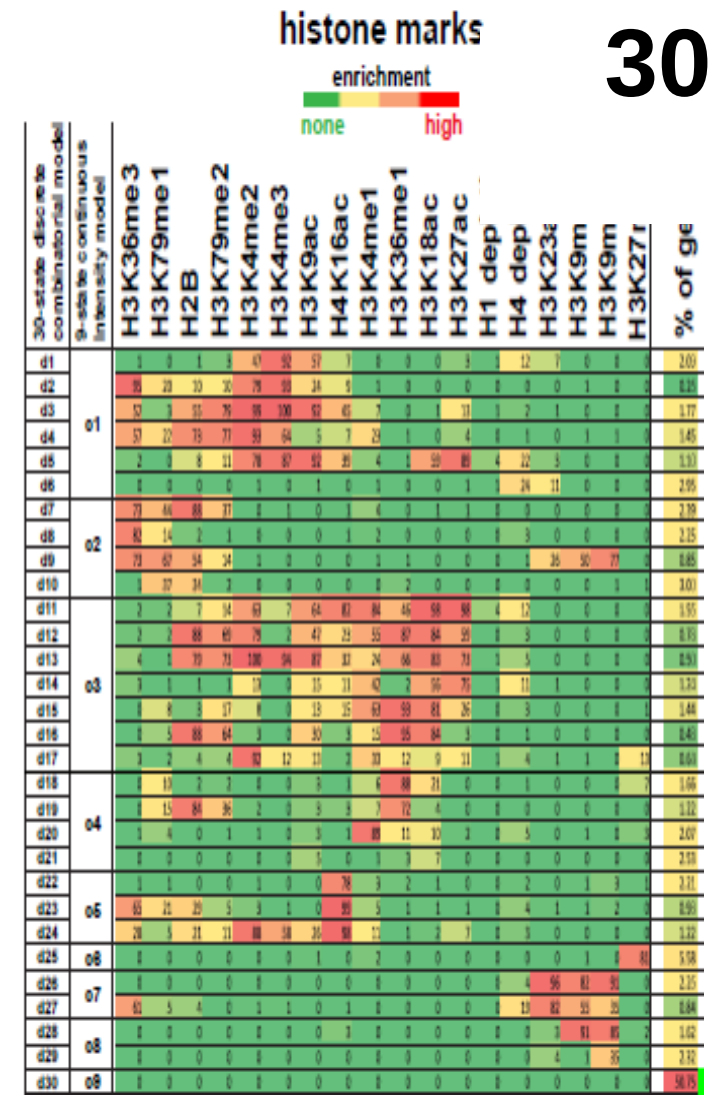
Kharchenko, Park, Karpen

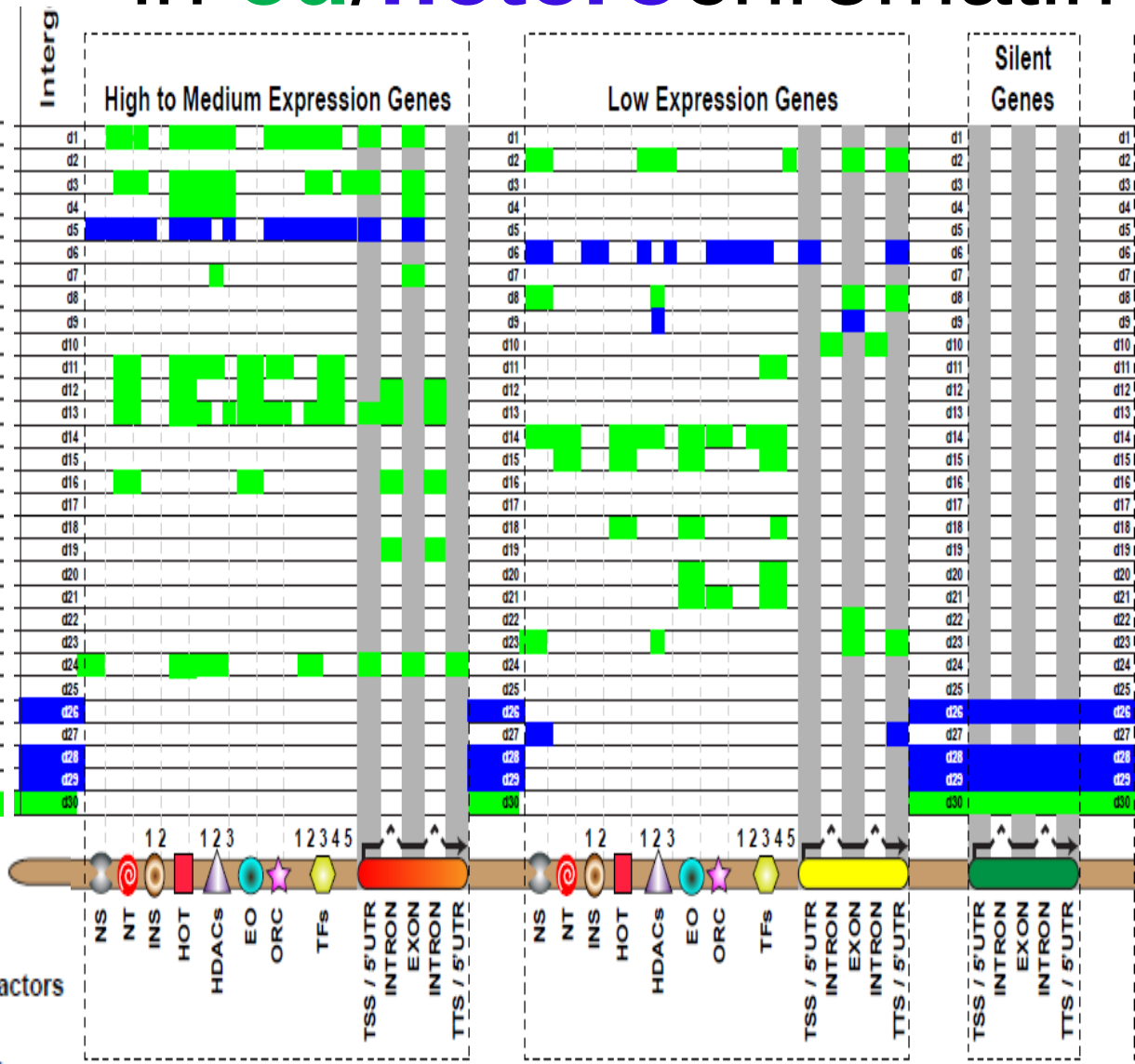
18 histone marks, S2 and BG3 cells



- 9 states captures most major chromatin types
- Summarize combinations and intensity of marks

30 discrete chromatin states in eu/heterochromatin





NS = nucleosome solubility

NT = nucleosome turnover

INS = insulators
- 1 = CTCF/BEAF/CP190
- 2 = Su(Hw)/Mod(mdg4)

HDACs = histone deacetylases
- 1 = TSS +PRE
- 2 = exon
- 3 = TSS

HOT = HOT region

EO = early origin

ORC = ORC binding

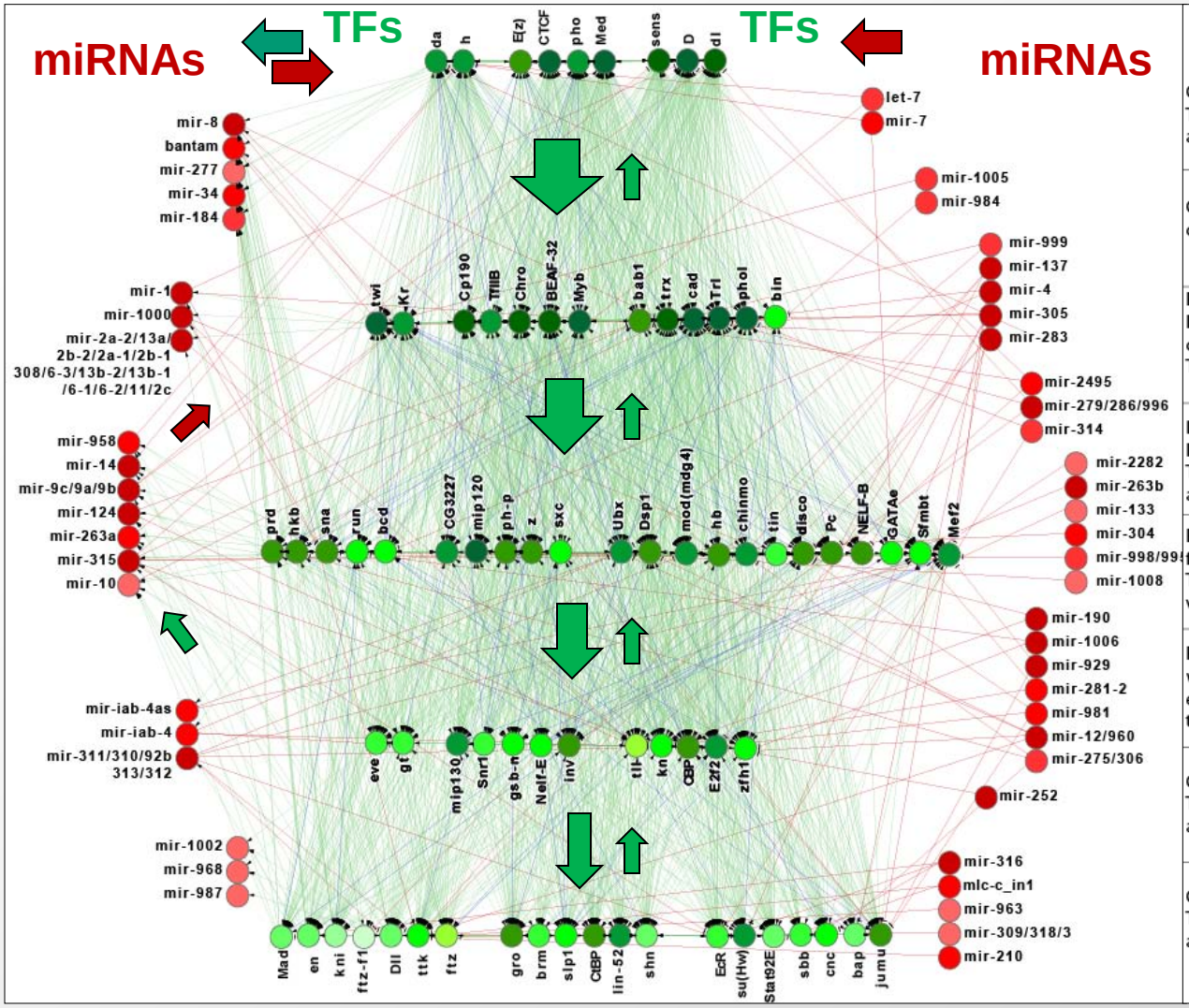
TFs = transcription factors
- 1 = segmentation
- 2 = neural development
- 3 = Polycomb / Trithorax
- 4 = signalling
- 5 = cofactors

- **Diversity of elements captured**

Insights from integrative analysis

1. **Annotate coding/non-coding genes**
   - Peptides, structures, microRNAs, readthrough
2. **Annotate chromatin regulatory regions**
   - Enhancers, promoters, diversity of functions
3. **Define regulator targets and networks**
   - Hierarchy, TF/miRNA networks, HOT regions
4. **Predictive models of gene regulation**
   - Functional nets → gene function/expression
5. **Implications for human disease**
   - Annotate non-coding SNPs, link to TFs/targets

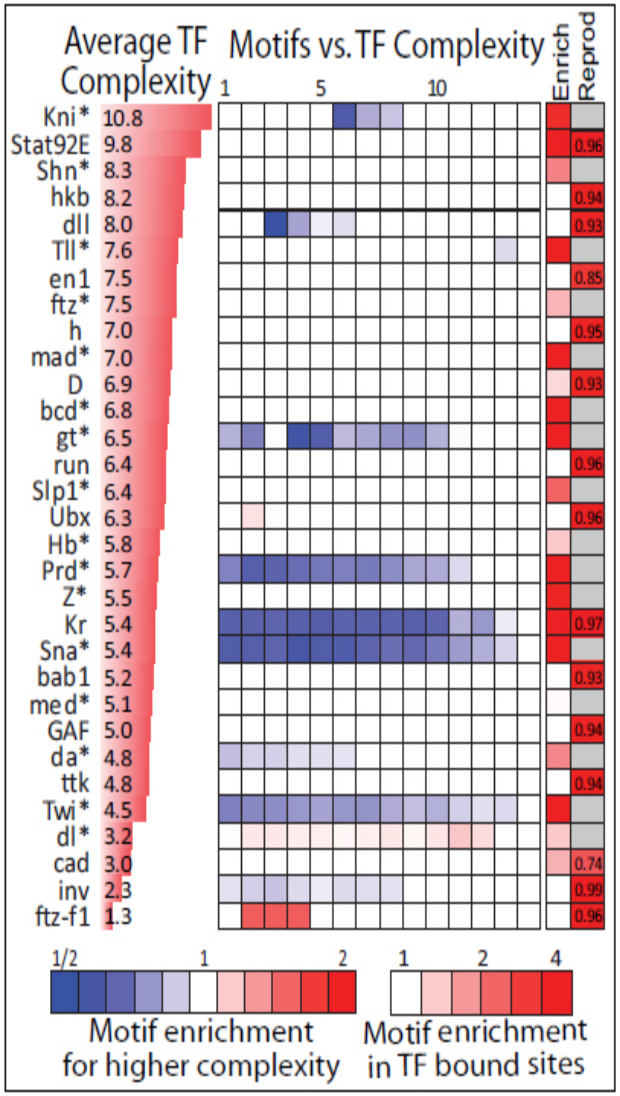
Binding, motifs reveal physical regulatory network

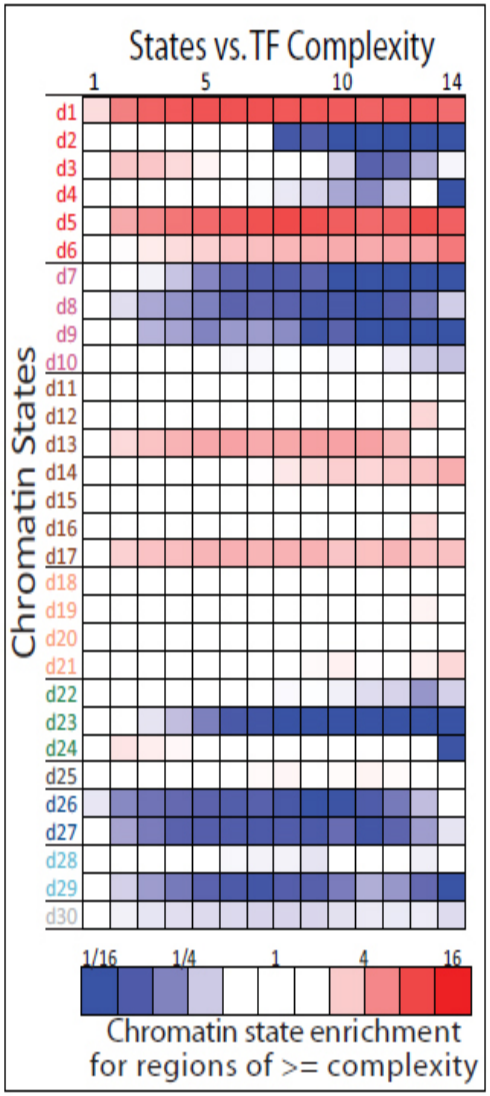


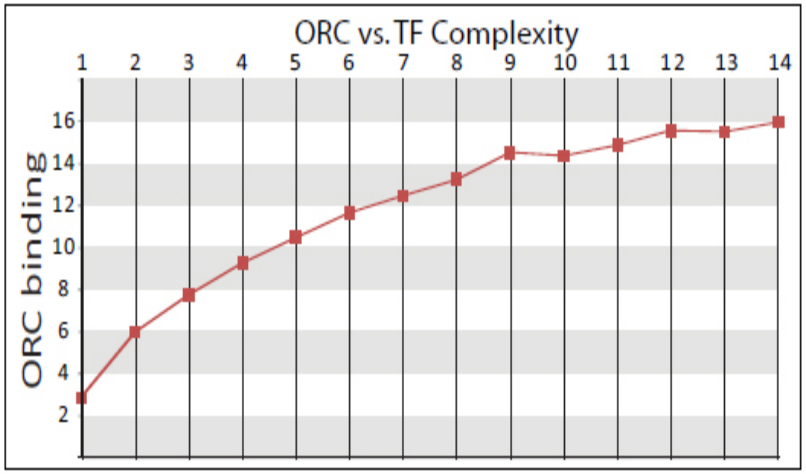
Network Motif	Example genes A	B	C
Cross-regulating TFs co-targeting a miRNA (F=2.996, Z=5.637)	twi	h	bantam
	sna	prd	mir-8
	eve	gt	mir-10
	run	prd	mir-14
	kni	gt	mir-277
Cross-regulatory clique of TFs (2.063, 3.453)	bcd	cad	ph-p
	mod(mdg4)	Kr	Dsp1
	BEAF-32	Myb	Med
	Cp190	mip120	phol
	cad	Chro	dl
Feed-forward loop with cross-regulating TFs and a miRNA (1.969, 2.311)	mir-1	twi	sna
	mir-315	gt	Kr
	mir-14	run	h
	mir-263a	prd	Kr
	mir-8	h	hb
Double feed-forward loop: cross-regulating TFs co-targeted by another TF (1.294, 4.507)	prd	gt	shn
	bab1	trx	disco
	TfIIB	mip130	Myb
	da	Mef2	lin-52
	GATAe	Mef2	z
Feedback loop from downstream TF to upstream TF via a microRNA (1.273, 1.295)	tin	mir-1000	Kr
	sna	mir-1	C15
	Kr	mir-315	sna
	hb	mir-8	nub
	sens	mir-9abc	eve
Feed-forward loop with a miRNA ending at a target gene (1.259, 1.797)	mir-958	hkb	Csk
	bantam	twi	dap
	mir-8	sna	crb
	mir-124	sna	Gli
	mir-263a	run	Mes2
Cross-regulating TFs co-targeting a target gene (1.256, 12.625)	pho	kn	Keap1
	D	da	dnk
	dl	lin-52	px
	phol	Med	tna
	mip120	Myb	Moe
Cross-regulating TFs co-targeting another TF (1.158 Fold Enr., 7.117 Z-score)	z	Med	run
	Dsp1	phol	lin-52
	gt	shn	cic
	trx	disco	CBP
	prd	ph-p	Antp

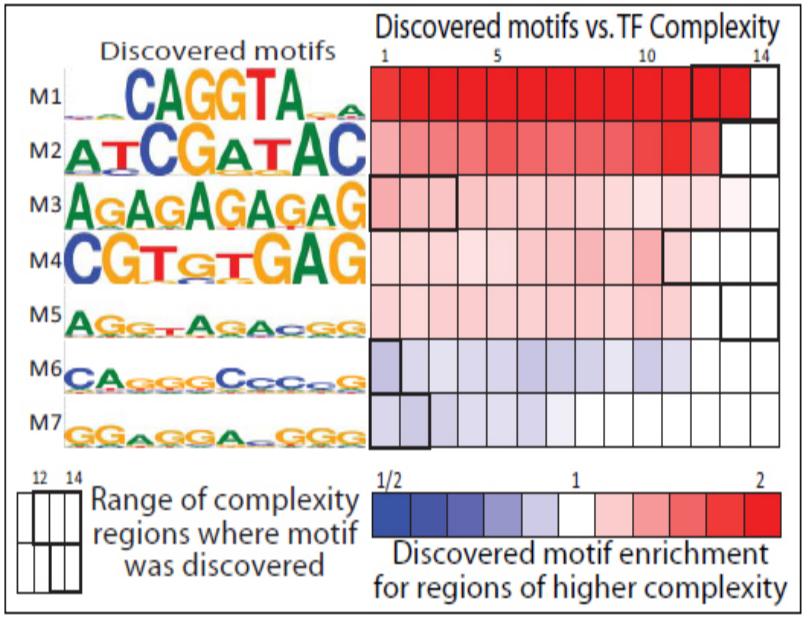
- Hierarchical network: master regulators, 94% down
- Feed-forward, cooperation, feedback through miRs

Hotspots, motifs, states, and origins





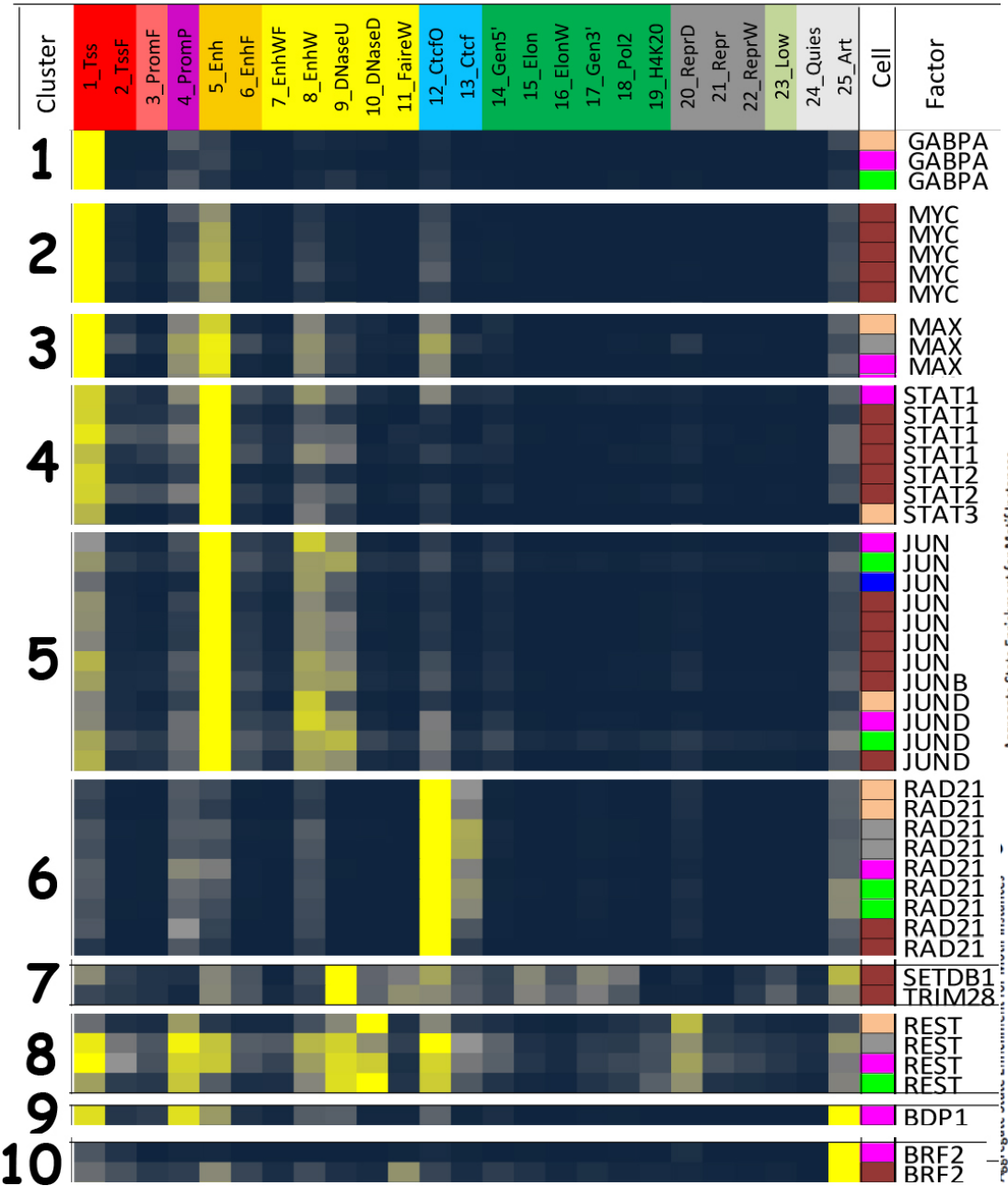




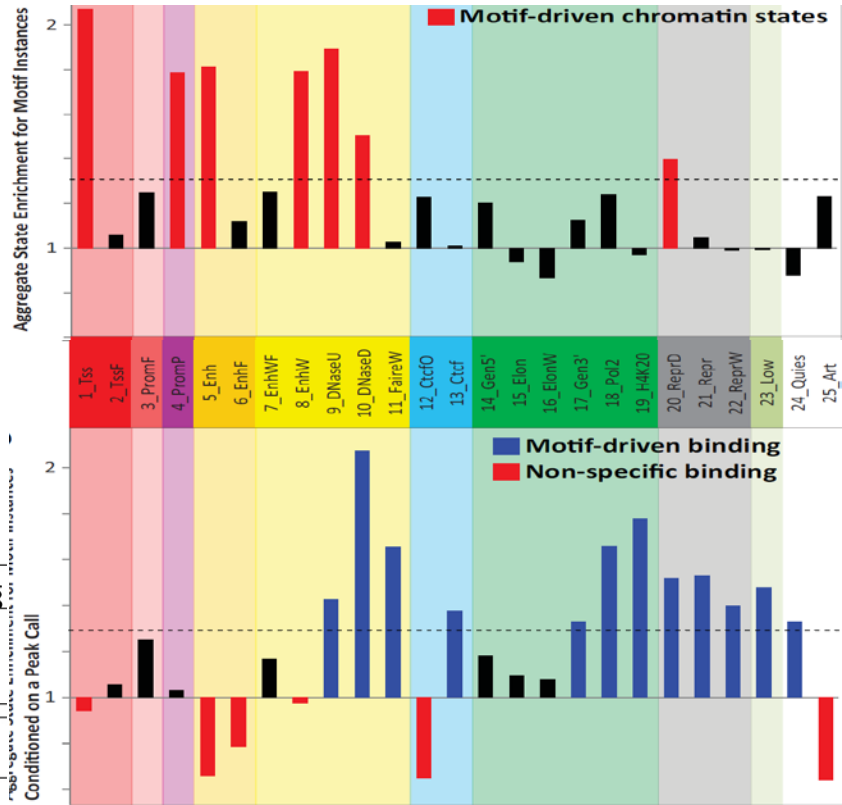
HOT regions **depleted** in known sequence motifs

Enriched in specific states, ORC, new motifs

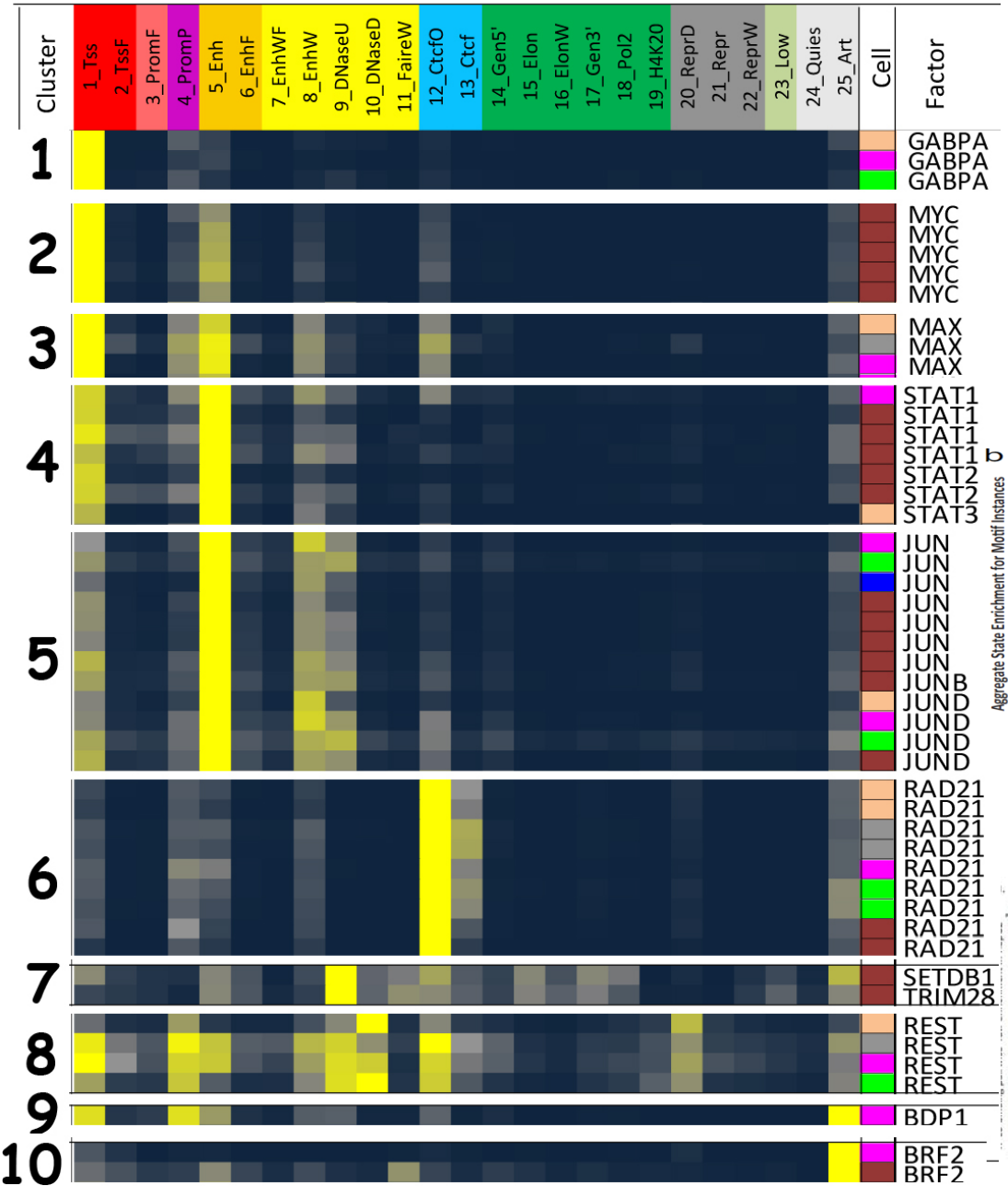
Interplay of TFs, motifs, and chromatin in human



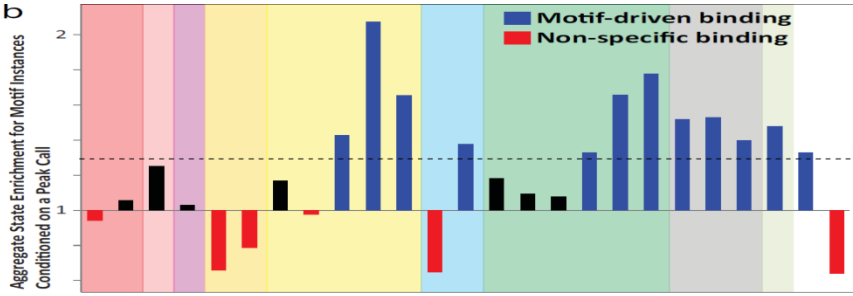
- TFs show distinct chromatin preferences
- Regulatory motifs underlie preferences
- Additional non-specific binding beyond motifs



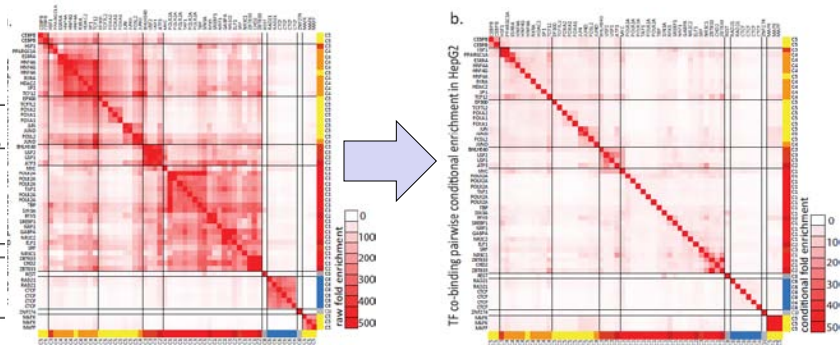
Interplay of TFs, motifs, and chromatin in human



- TFs show distinct chromatin preferences
- Regulatory motifs underlie preferences
- Additional non-specific binding beyond motifs



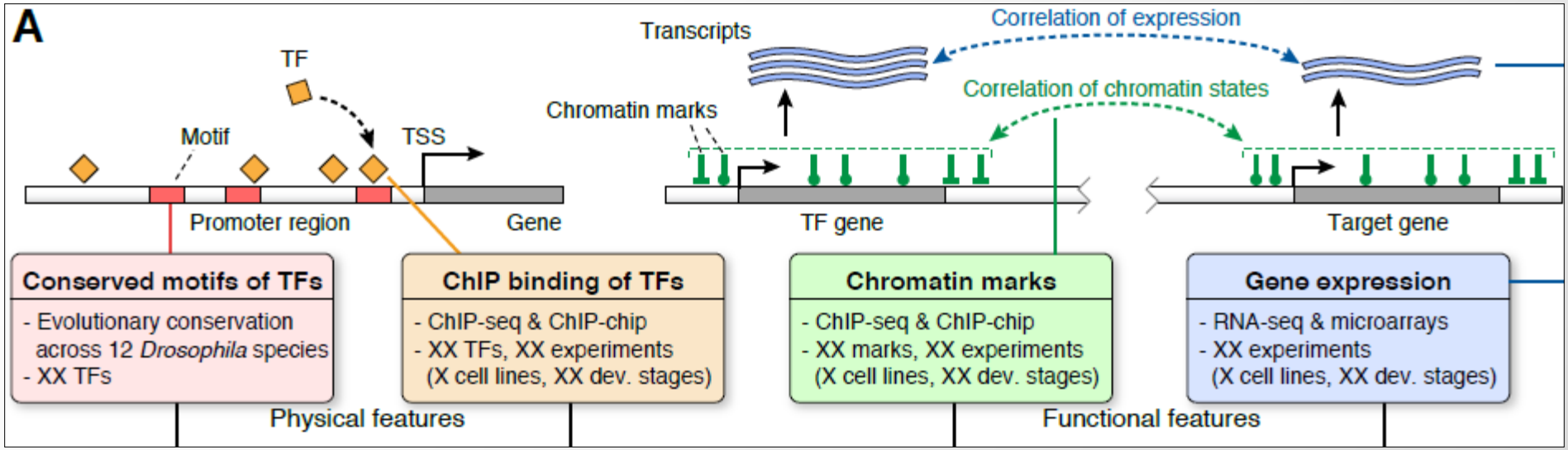
- State prefs predict TF pairwise co-occurrence

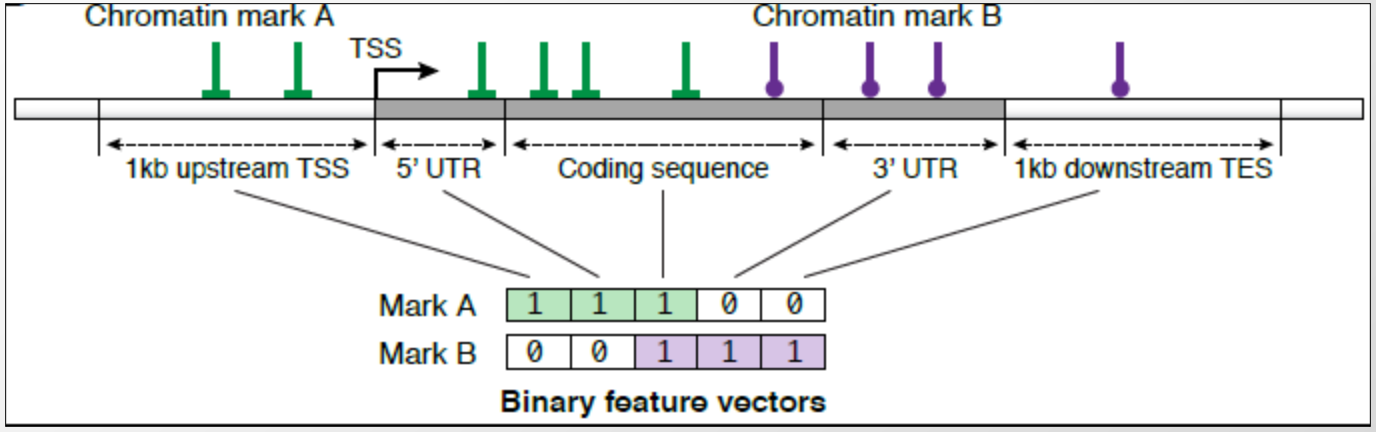


Insights from integrative analysis

1. **Annotate coding/non-coding genes**
   - Peptides, structures, microRNAs, readthrough
2. **Annotate chromatin regulatory regions**
   - Enhancers, promoters, diversity of functions
3. **Define regulator targets and networks**
   - Hierarchy, TF/miRNA networks, HOT regions
4. **Predictive models of gene regulation**
   - Functional nets ➔ gene function/expression
5. **Implications for human disease**
   - Annotate non-coding SNPs, link to TFs/targets

Combined datasets derive functional regulatory net

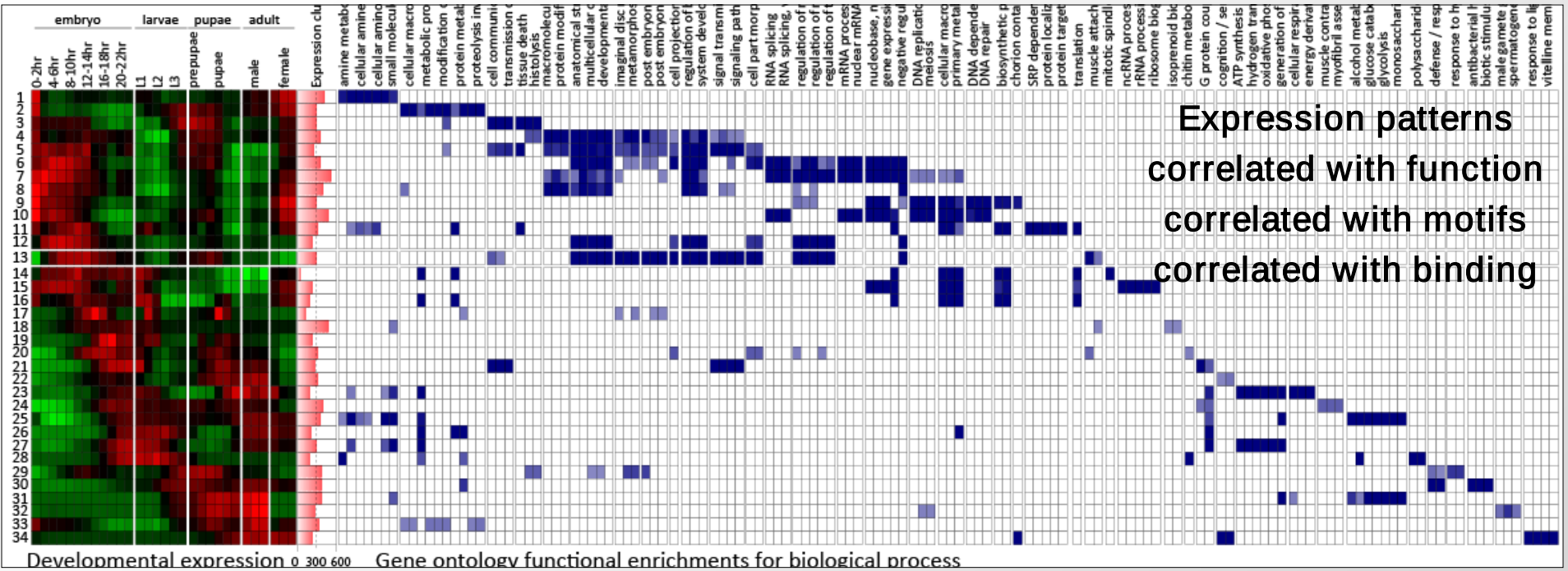




➔ Capture functional consequences of targeting

- **Not all binding is functional, not all targets are direct**
- **Motifs, ChIP, marks, expr as input feature to learning**
- **Unsupervised sum-rule & supervised using REDfly**

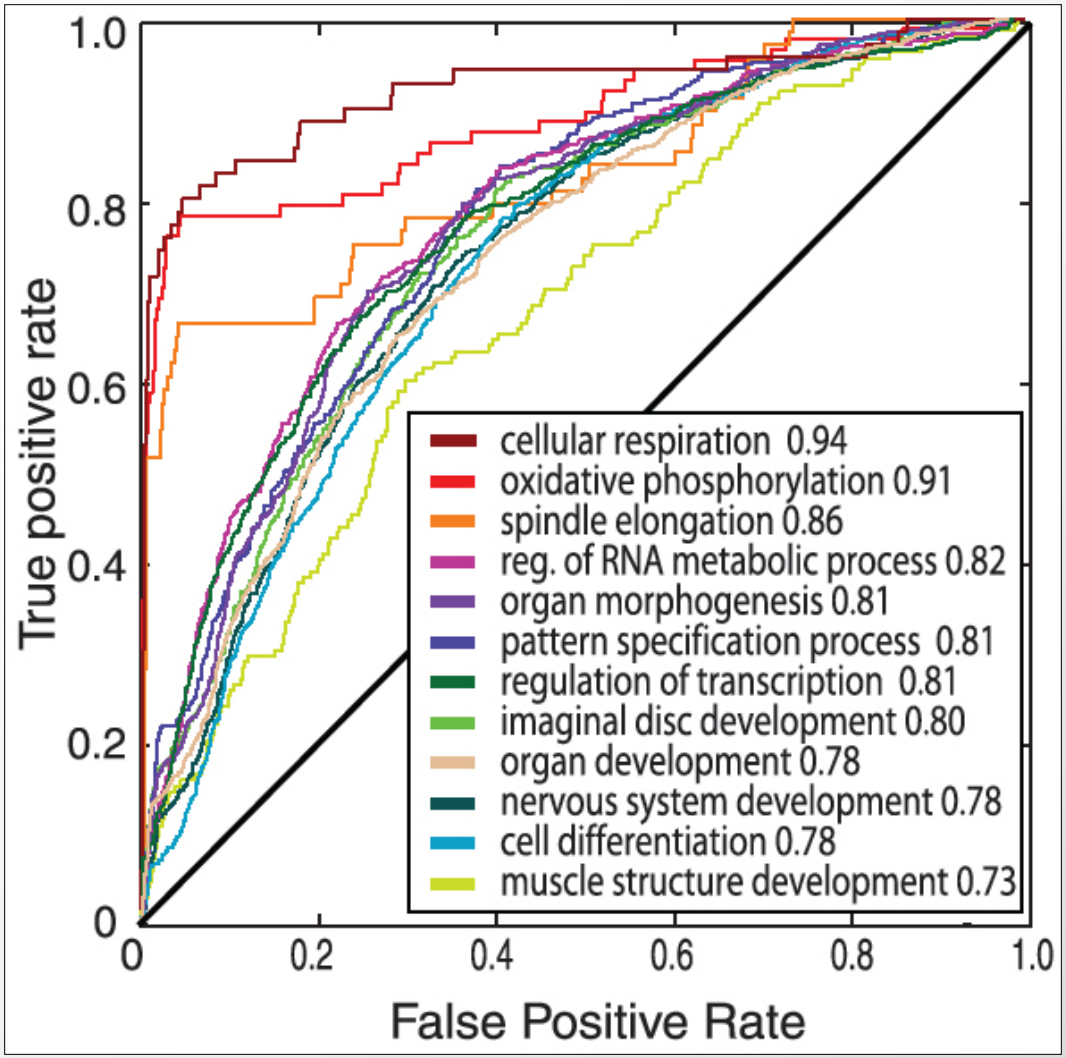
Functional enrichments in integrative network

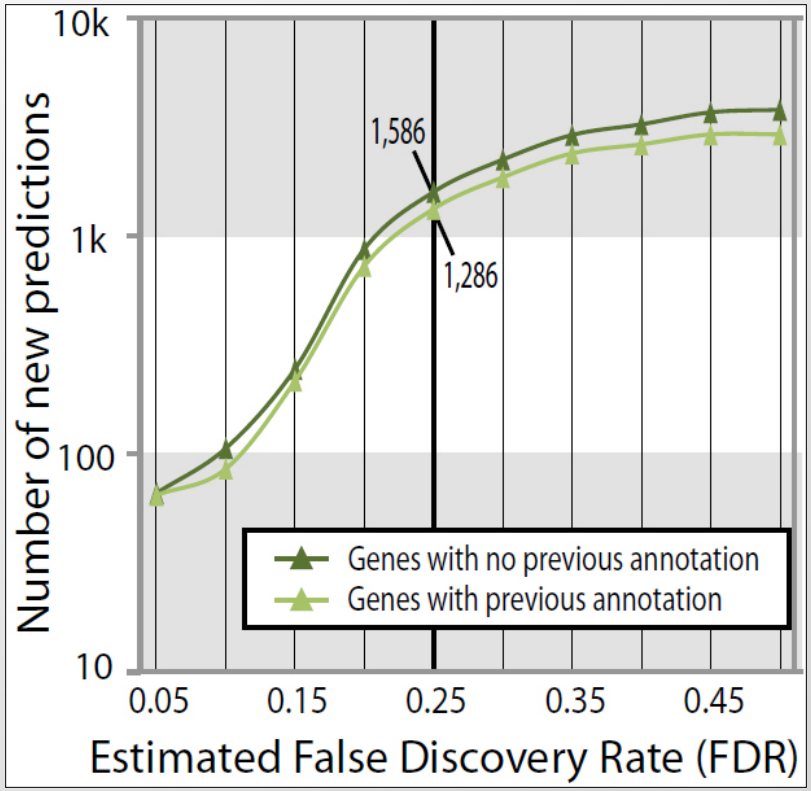


Network name	Network description	Network size: NumTFs	NTargets	Edges/TG	True positive rate for predictions: Function	DevExpr	Tissues	PPI
motifNet	Conserved regulatory motifs in promoters	104	11,090	7	-11%	-4%	-18%	35%
boundNet	ChIP-inferred experimental TF binding sites	76	12,482	13	7%	12%	16%	16%
REDfly	Literature-based known regulatory network	82	88	3	37%	64%	43%	61%
FuncNet	Functional network with chrom/expr activity	576	9,436	24	19%	46%	19%	60%

- **Combine TF binding, motifs, correlated TF/TG activity**
- **Reveal ‘functional’ edges, response determinants**
- **Functional net shows increased predictive value**

Predicting new GO functional annotations for genes

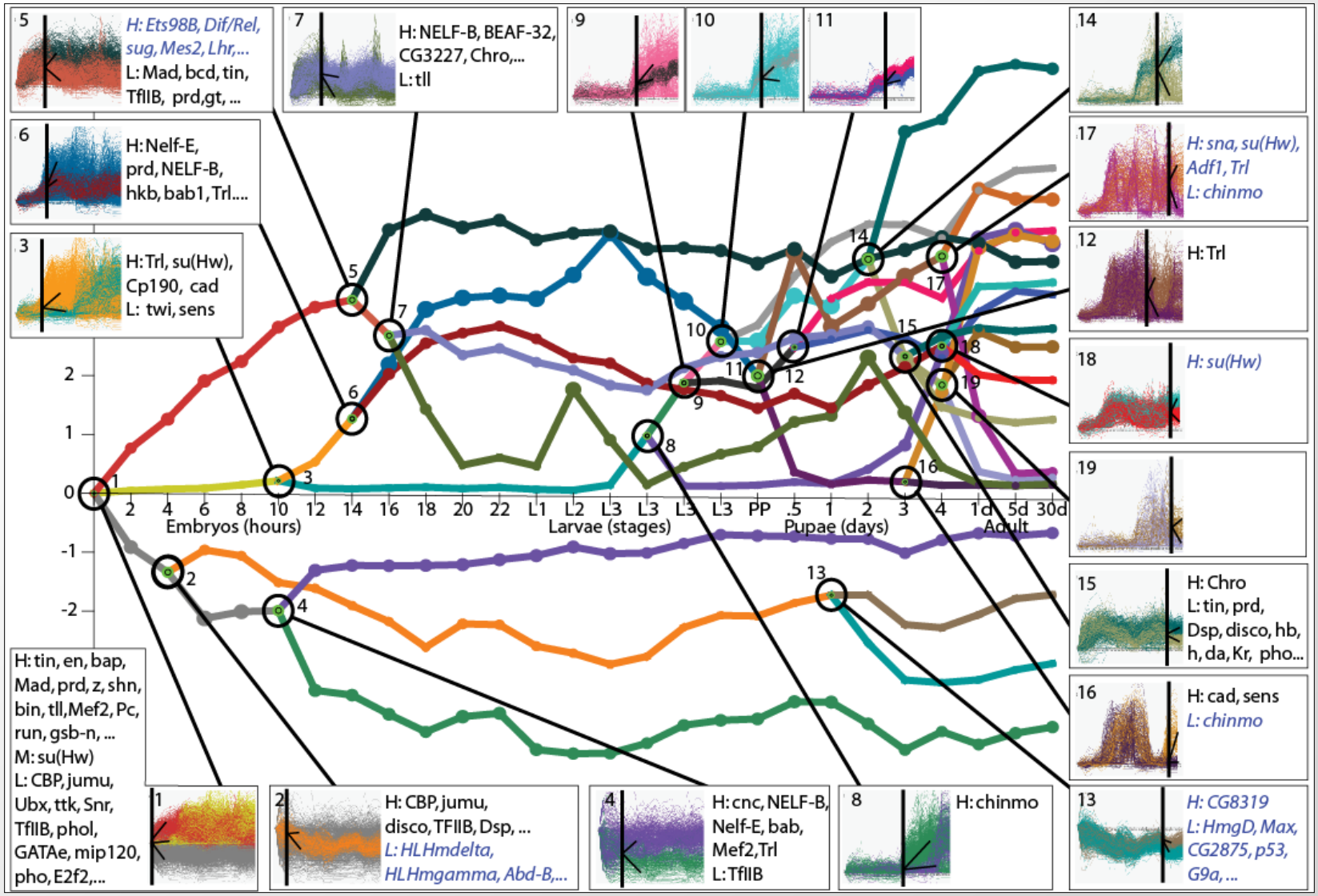




>1000 new functional predictions

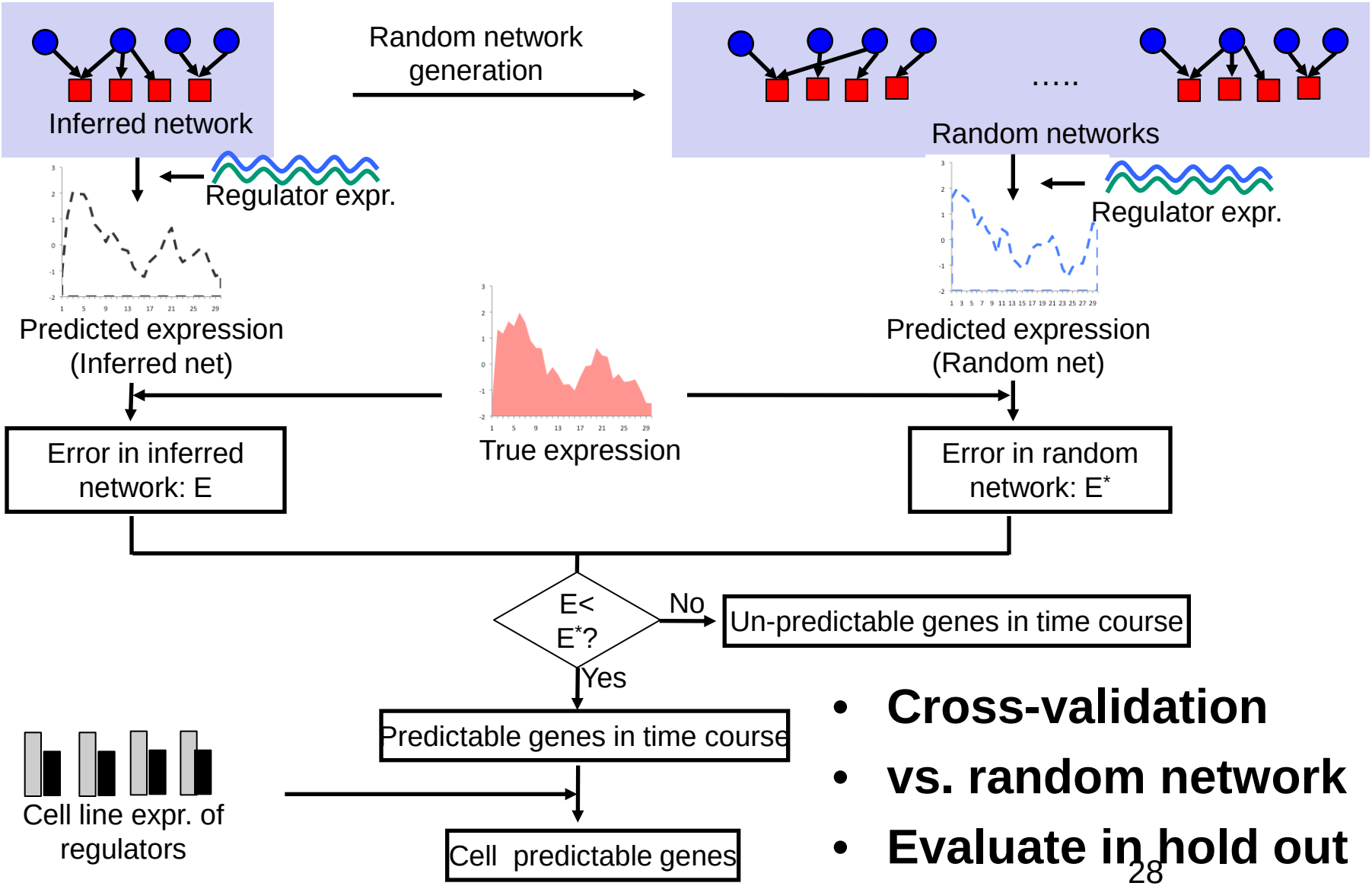
- **Shared activity and regulation ➔ shared function**
- **Tissue expression confirms functional predictions**

Predicting stage-specific regulators of expression



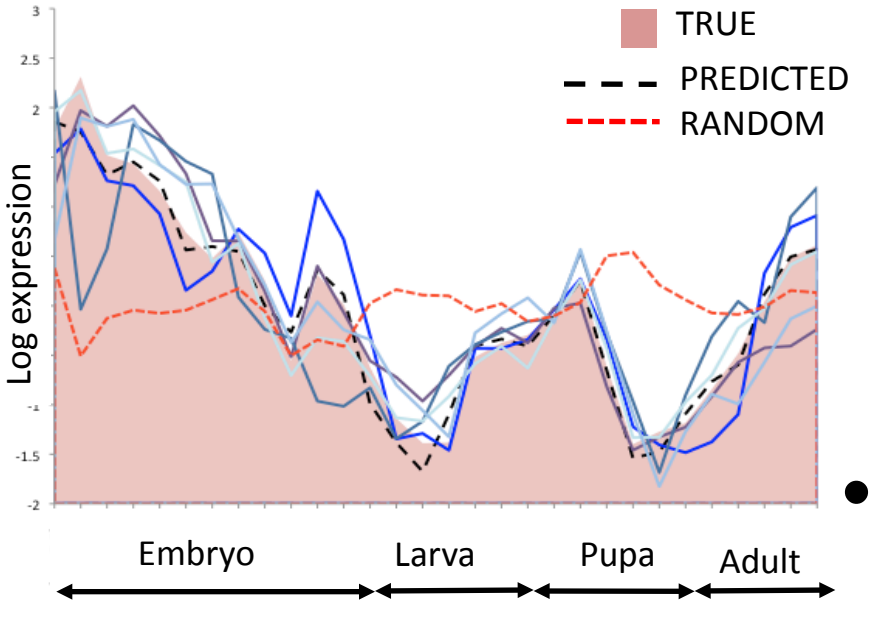
- **Coordinated changes between TFs and their targets**

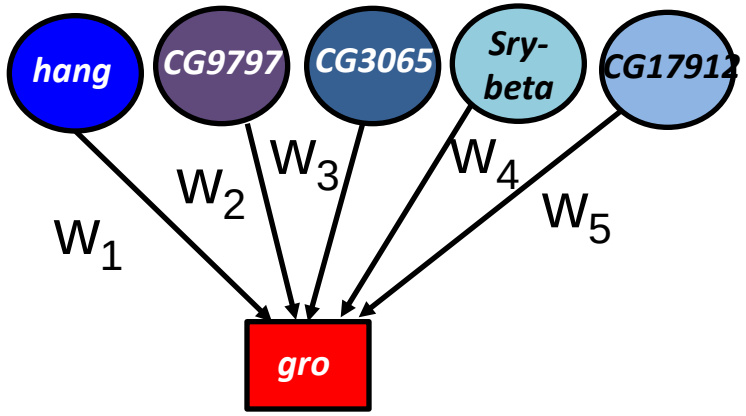
Predictive power for gene expression levels

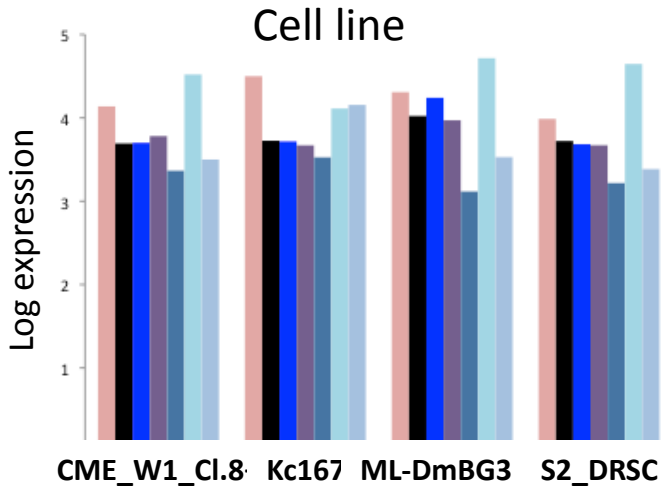


- **Cross-validation**
- **vs. random network**
- **Evaluate in hold out**

Example: predicting *gro* expression





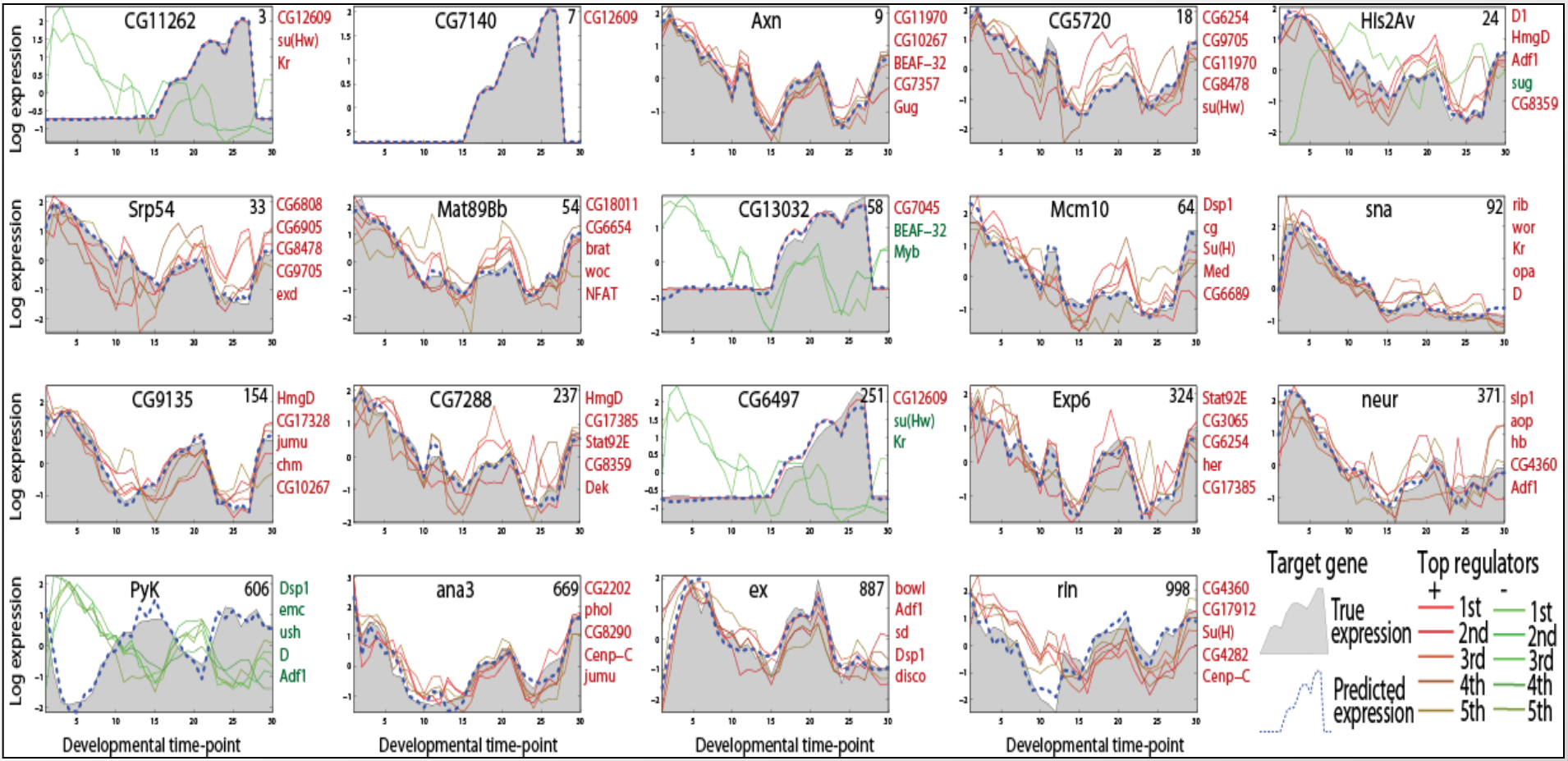


- Predict target expression as a function of TF levels

$$gro = f(TF_1, TF_2, \ldots, TF_n) = w_1TF_1 + w_2TF_2 + \ldots + w_nTF_n$$

- vs. true, random net, TFs
- Predictive in new cell types

Gene expression prediction for 1,500 genes!



- **Linear regression model: Target_expr=F([TF_1_expr,…])**
- **Learn coefficients in 27 time-points, predict in other 3**
- **‘Unpredictable’ genes are also less reproducible**
- **‘Predictable’ genes: learned weights work in cell lines**

Insights from integrative analysis

1. **Annotate coding/non-coding genes**
   - Peptides, structures, microRNAs, readthrough
2. **Annotate chromatin regulatory regions**
   - Enhancers, promoters, diversity of functions
3. **Define regulator targets and networks**
   - Hierarchy, TF/miRNA networks, HOT regions
4. **Predictive models of gene regulation**
   - Functional nets → gene function/expression
5. **Implications for human disease**
   - Annotate non-coding SNPs, link to TFs/targets

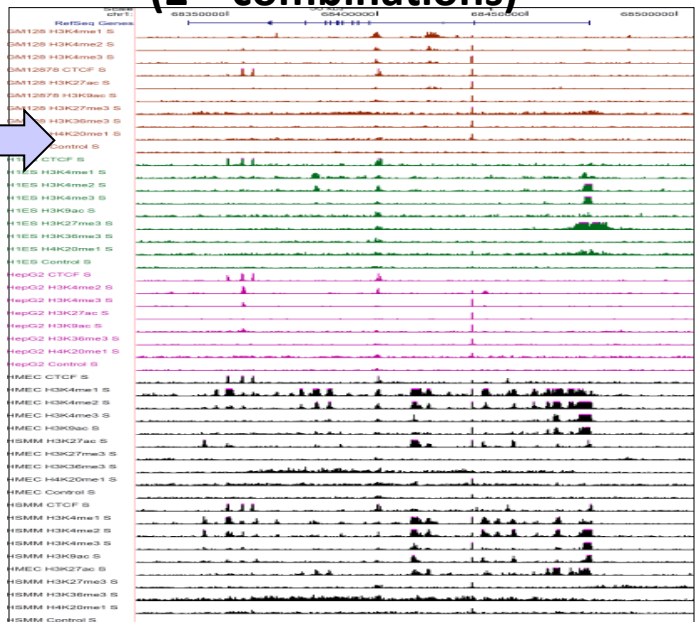
ENCODE: Study nine marks in nine human cell lines

9 marks
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H4K20me1
H3K36me3
CTCF
+WCE
+RNA

X

9 human cell types	
HUVEC	Umbilical vein endothelial
NHEK	Keratinocytes
GM12878	Lymphoblastoid
K562	Myelogenous leukemia
HepG2	Liver carcinoma
NHLF	Normal human lung fibroblast
HMEC	Mammary epithelial cell
HSMM	Skeletal muscle myoblasts
H1	Embryonic

81 Chromatin Mark Tracks (2^{81} combinations)



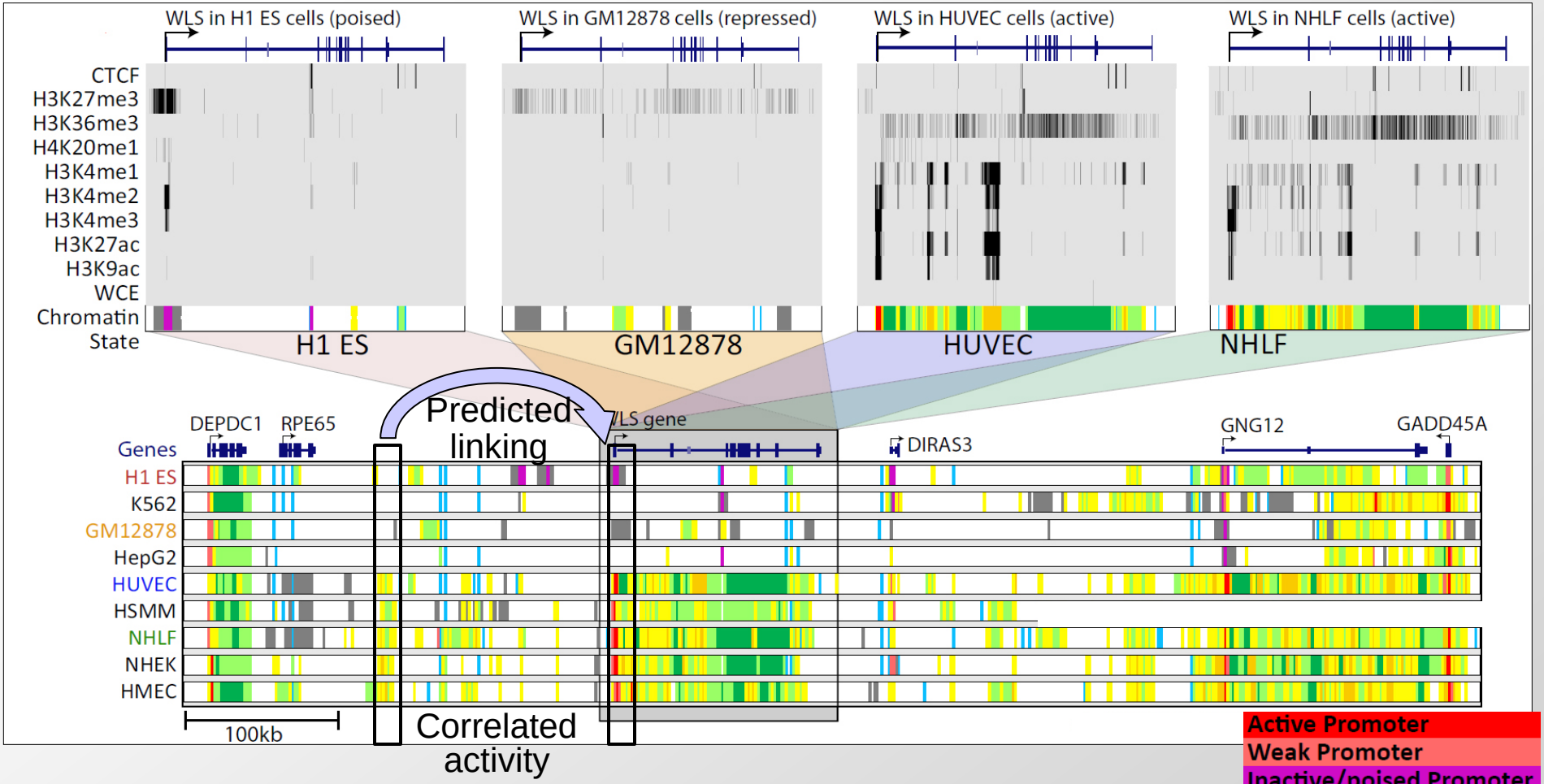
Brad Bernstein ENCODE Chromatin Group

b.

Chromatin States	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage: Median	Coverage: H1 ES	Coverage: GM	Median Length	+/-2kb TSS	Conserved non-exon	DNase (K562)	C-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear Lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active Promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak Promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised Promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb-repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV

Chromatin Mark Observation Frequency (%) | (%) | (fold) | (kb) | (%) | Functional enrichments (fold)

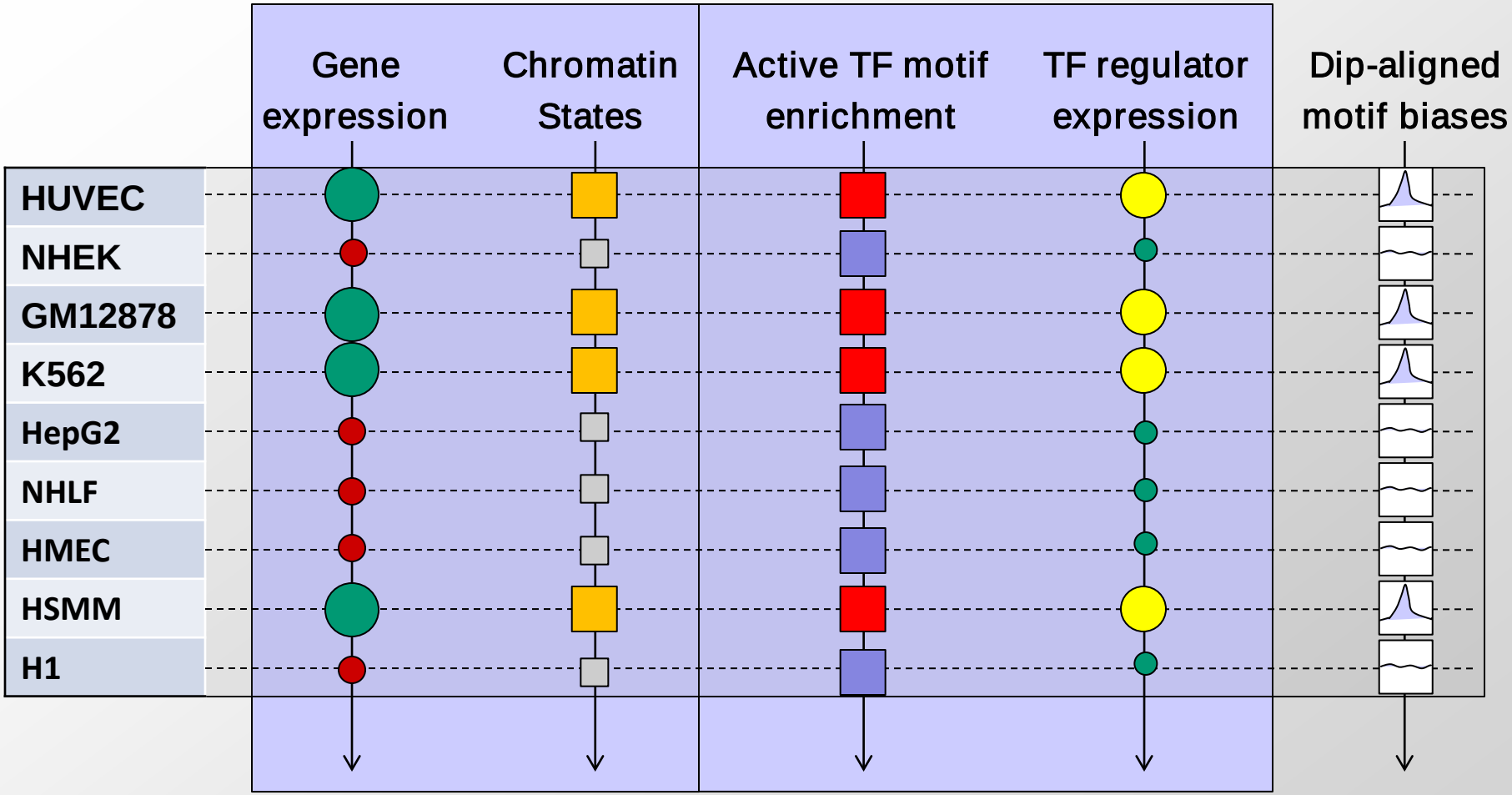
Chromatin states dynamics across nine cell types



- Single annotation track for each cell type
- Summarize cell-type activity at a glance
- Can study 9-cell activity pattern across ↓

Active Promoter
Weak Promoter
Inactive/poised Promoter
Strong enhancer
Strong enhancer
Weak/poised enhancer
Weak/poised enhancer
Insulator
Transcriptional transition
Transcriptional elongation
Weak transcribed
Polycomb-repressed
Heterochrom; low signal

Introducing multi-cell activity profiles



Link enhancers to target genes

Link TFs to target enhancers

Predict activators vs. repressors

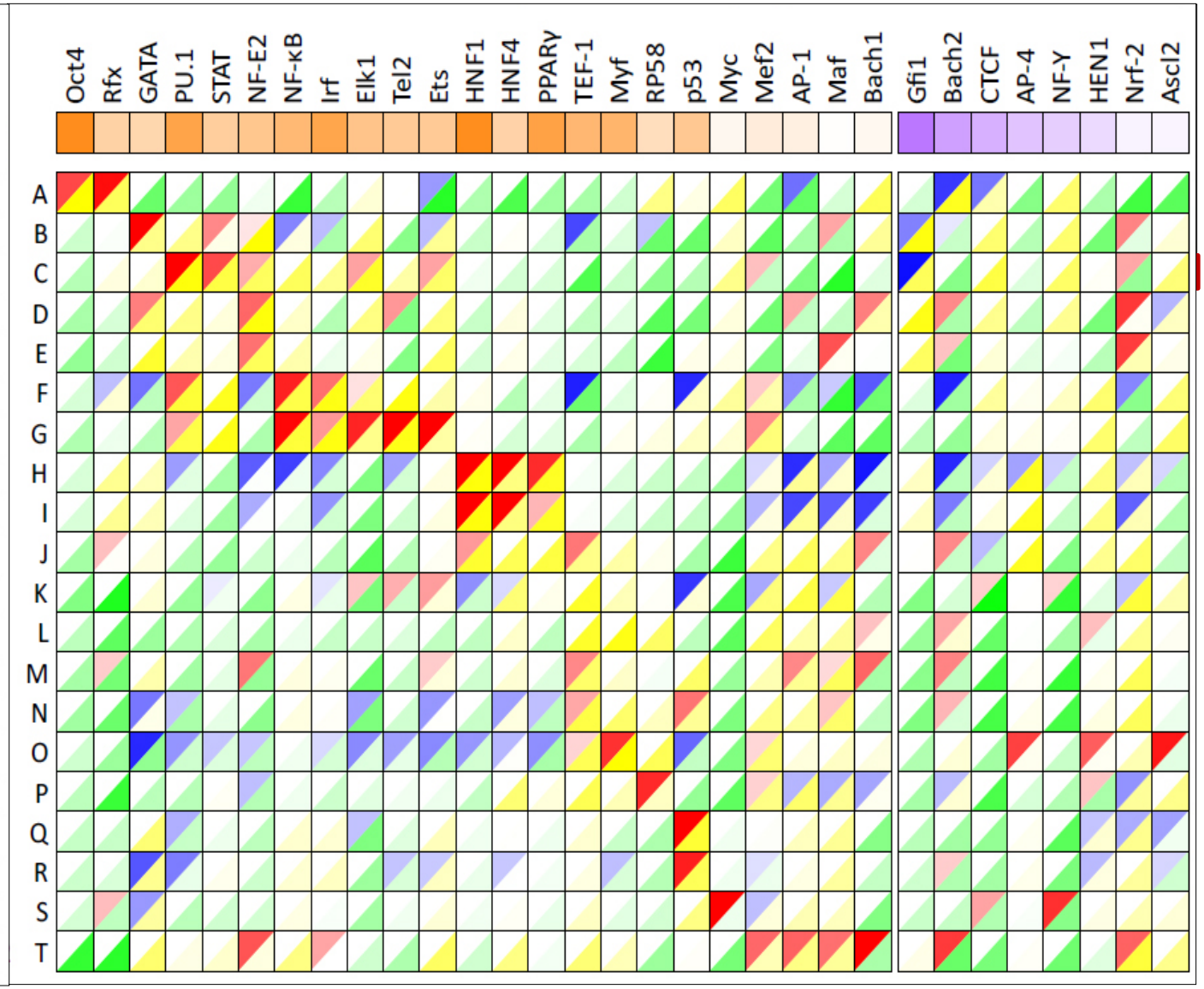
ON
OFF
Active enhancer
Repressed
Motif enrichment
Motif depletion
TF On
TF Off
Motif aligned
Flat profile

Coordinated activity reveals activators/repressors

Enhancer activity

	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
A	44	3	2	3	2	2	2	2	2
B	1	48	2	4	2	2	2	2	2
C	2	48	47	8	5	5	4	5	3
D	4	46	9	10	45	10	11	14	10
E	4	48	12	14	10	10	10	44	26
F	0	1	48	1	1	1	1	1	1
G	2	5	48	6	46	16	13	12	7
H	1	2	2	51	1	1	1	2	1
I	1	1	1	36	1	1	1	1	1
J	3	7	6	49	31	30	18	12	10
K	1	2	2	2	46	3	4	4	3
L	2	3	2	2	47	45	20	7	9
M	3	3	5	6	49	26	19	47	34
N	2	2	6	5	5	47	18	46	31
O	1	1	2	1	2	45	5	2	3
P	1	3	4	2	4	21	45	4	4
Q	2	1	5	5	3	3	4	49	45
R	1	1	3	2	3	3	3	45	8
S	2	4	4	4	8	6	5	8	43
T	11	40	27	25	45	41	39	45	41

Activity signatures for each TF

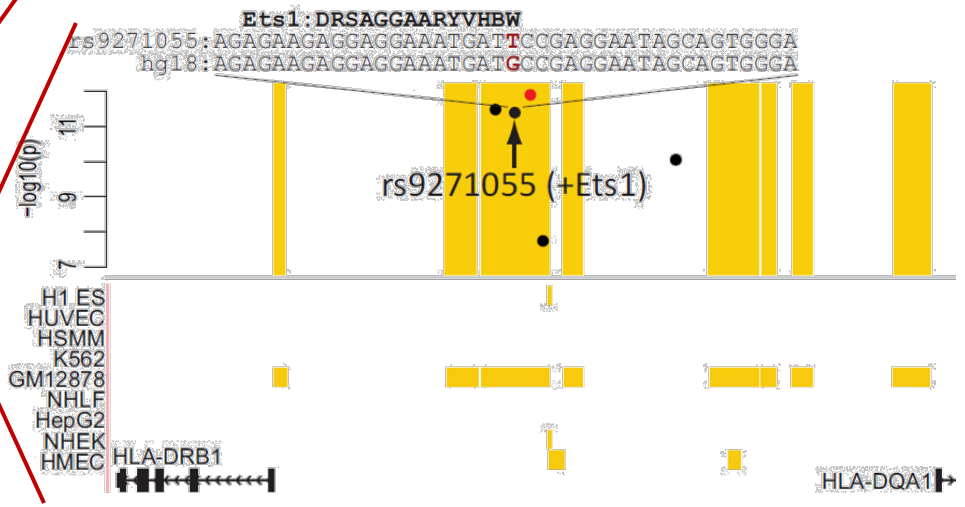


- Enhancer networks: Regulator → enhancer → target gene

Revisiting disease-associated variants

Phenotype	Top Cell Type	Total #SNPs from Study	#SNPs in enh. States 4 and 5	p-value	FDR	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
Erythrocyte phenotypes (Ref. 38)	K562	35	9	$<10^{-7}$	0.02	9	17	4	0	0	1	2	1	1
Blood lipids (Ref. 39)	HepG2	101	13	$<10^{-7}$	0.02	3	5	0	11	2	3	3	4	3
Rheumatoid arthritis (Ref. 40)	GM12878	29	7	2.0×10^{-7}	0.03	0	0	15	0	2	0	0	2	3
Primary billary cirrhosis (Ref. 41)	GM12878	6	4	6.0×10^{-7}	0.03	0	11	41	0	0	0	0	8	8
Systemic lupus erythromatosus (Ref. 42)	GM12878	18	6	9.0×10^{-7}	0.03	0	4	21	0	5	8	0	3	5
Lipoprotein cholesterol/triglycerides (Ref. 43)	HepG2	18	5	1.2×10^{-6}	0.03	17	8	0	24	3	6	4	3	3
Hematological traits (Ref. 44)	K562	39	7	1.7×10^{-6}	0.03	0	12	10	2	1	0	0	1	0
Hematological parameters (Ref. 45)	K562	28	6	2.2×10^{-6}	0.03	0	15	7	0	5	7	7	3	2
Colorectal cancer (Ref. 46)	HepG2	4	3	3.8×10^{-6}	0.03	0	0	0	66	0	12	0	12	12
Blood pressure (Ref. 47)	K562	9	4	5.0×10^{-6}	0.04	0	30	14	0	10	6	7	5	11

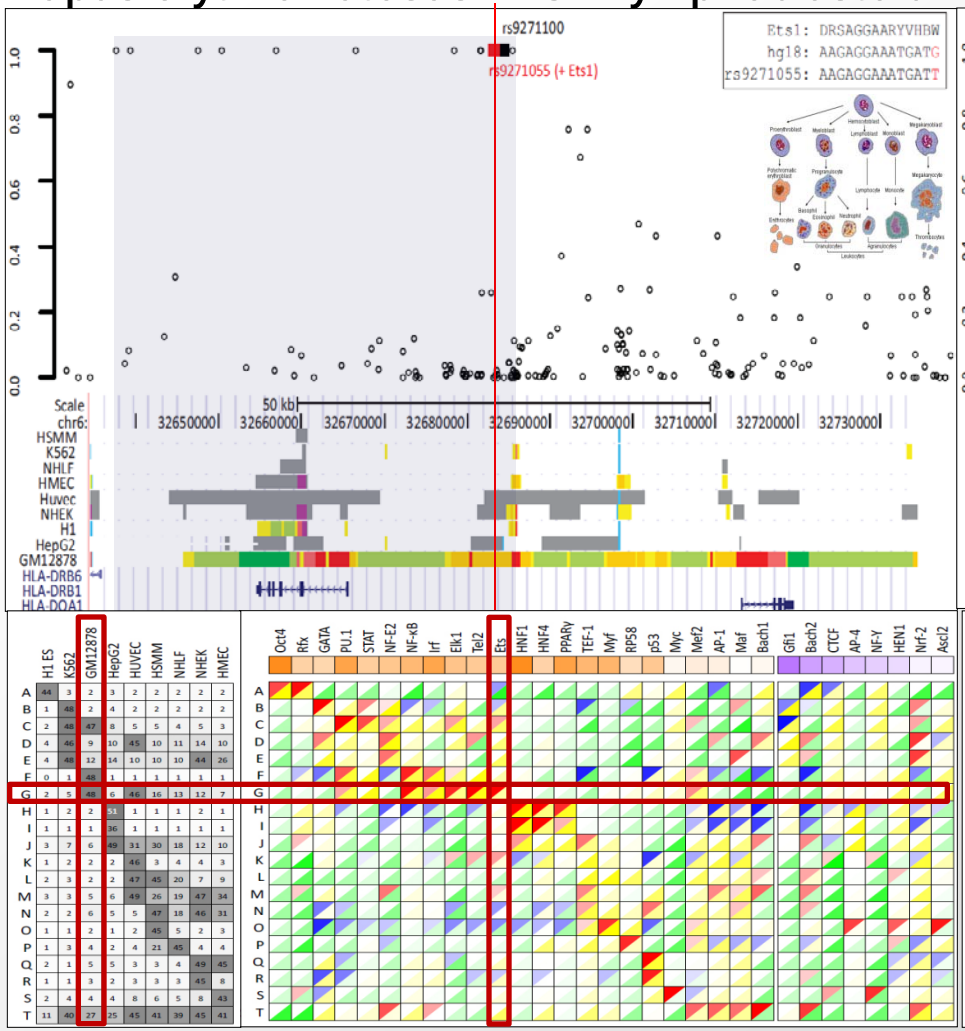
SNP	H1 ES	K562	GM	HepG2	Huvec	HSMM	NHLF	NHEK	HMEC	Chrom. Band	Gene	Link Scc	Distanc
rs13385731										2p22			
rs10036748										5q33			
rs1385374										12q24	MGC16384	-	1
rs2230926										6q23	TNFAIP3	3.7	7
rs4728142										7q32	IRF5	-	4
rs9271100										6p21	HLA-DRB1	4.5	19
rs4917014										7p12	IKZF1	2.2	38
rs7812879										8p23	BLK	2.9	11
rs2205960										1q25			
rs548234										6q21			



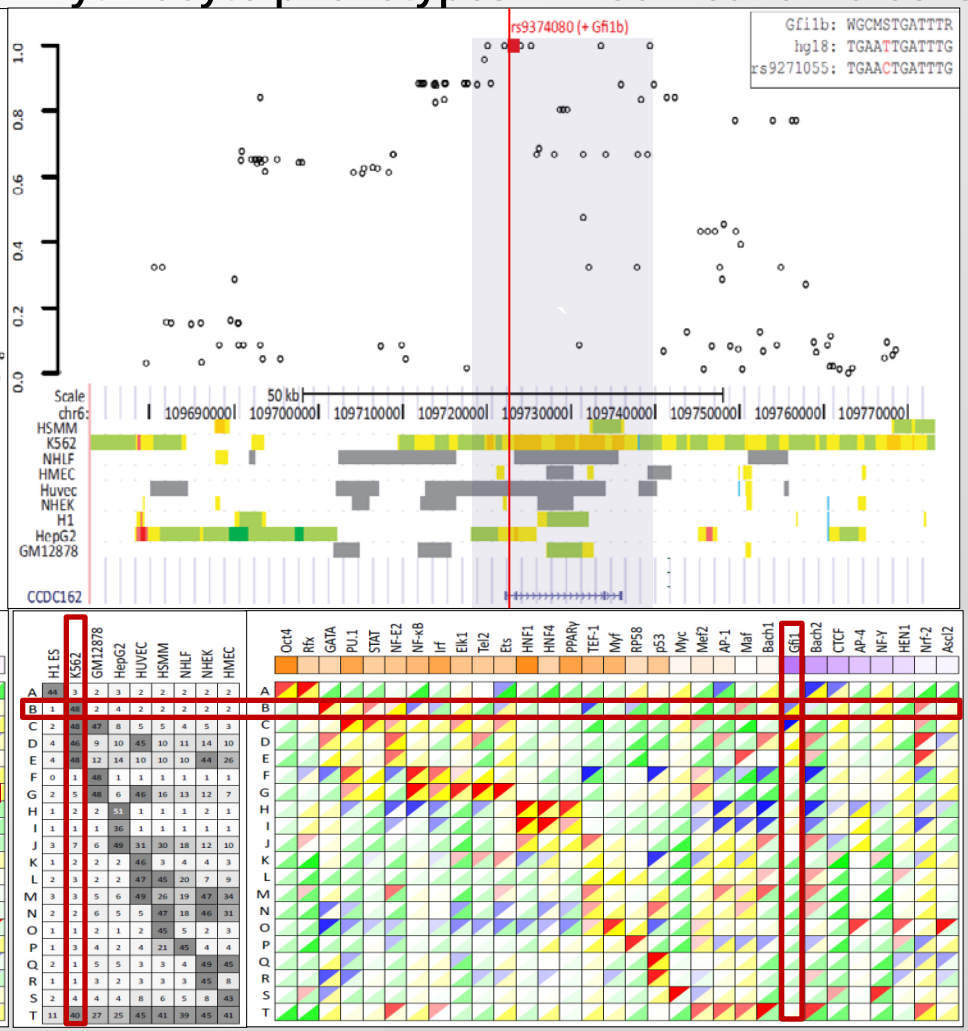
- **Disease-associated SNPs enriched for enhancers in relevant cell types**
- **E.g. lupus SNP in GM enhancer disrupts Ets1 predicted activator**

Mechanistic predictions for top disease-associated SNPs

Lupus erythromatosus in GM lymphoblastoid



Erythrocyte phenotypes in K562 leukemia cells



Disrupt activator Ets-1 motif

➔ **Loss of GM-specific activation**

➔ **Loss of enhancer function**

➔ **Loss of HLA-DRB1 expression**

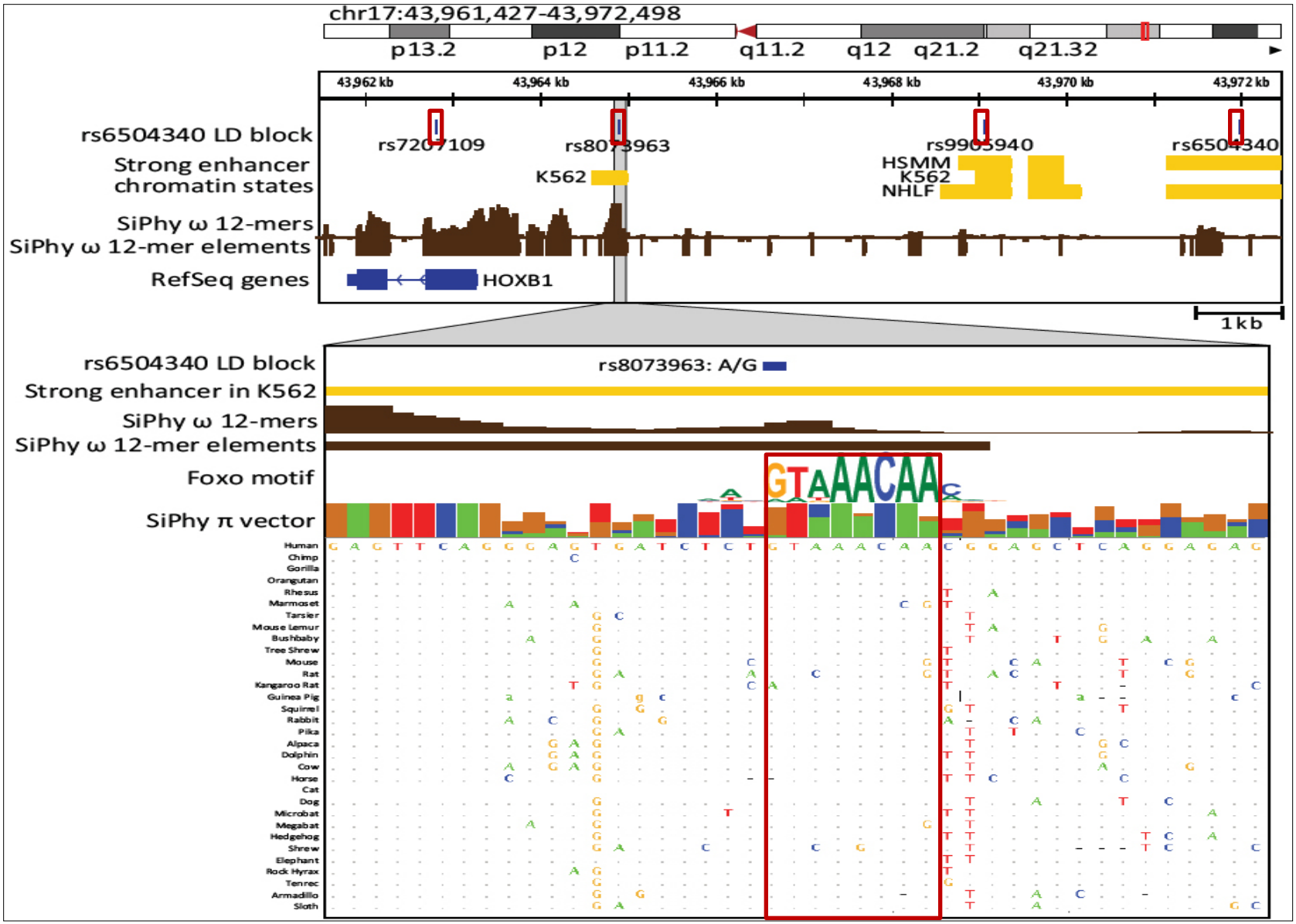
Creation of repressor Gfi1 motif

➔ **Gain K562-specific repression**

➔ **Loss of enhancer function**

➔ **Loss of CCDC162 expression**

Detect SNPs that disrupt conserved regulatory motifs



- Functionally-associated SNPs enriched in states, constraint
- Prioritize candidates, increase resolution, disrupted motifs

Automating prediction of likely causal variants in LD
➔ HaploReg (compbio.mit.edu/HaploReg)

Query SNP: rs17145713 and variants with r^2 >= 0.95

chr	pos (hg19)	LD	variant	Ref	Alt	ASN freq	CEU freq	YRI freq	GERP cons	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	GENCODE genes	RefSeq genes	dbSNP func annot
7	72842724	1	7:72480660	GAC	G	0	0.15	0					PANC-1			5.4kb 5' of FZD9	5.4kb 5' of FZD9	
7	72856430	1	rs1178979	T	C	0.13	0.18	0.3					CLL		GATA	BAZ1B	BAZ1B	intronic
7	72857049	1	rs1178977	A	G	0.14	0.18	0.3							AREB6,DEC	BAZ1B	BAZ1B	intronic
7	72857713	1	rs34604283	CA	C	0.13	0.1	0.2					8 cell types		Sox	BAZ1B	BAZ1B	intronic
7	72868522	1	rs1306476	A	G	0.12	0.18	0.36								BAZ1B	BAZ1B	intronic
7	72883106	1	rs62465144	T	C	0.14	0.18	0.29								BAZ1B	BAZ1B	intronic
7	72885810	1	rs6976930	G	A	0.14	0.18	0.39								BAZ1B	BAZ1B	intronic
7	72904810	1	rs17145713	C	T	0.14	0.18	0.3							ATF3	BAZ1B	BAZ1B	intronic
7	72939244	1	rs11983997	G	C	0.13	0.18	0.26				GM12878, K562	GM12864,GM12878,K562			2.6kb 5' of BAZ1B	2.6kb 3' of BAZ1B	
7	72977249	1	rs34594435	C	T	0.12	0.18	0.03				K562	CMK	KAP1,SETDB1		4.9kb 5' of BCL7B	5.2kb 3' of BCL7B	
7	72988069	1	rs35659126	C	T	0.13	0.18	0.08								TBL2	TBL2	intronic
7	72989141	1	rs34550818	C	CA	0.11	0.12	0						POL2		TBL2	TBL2	intronic
7	72989390	1	rs11974409	A	G	0.13	0.18	0.14								TBL2	TBL2	intronic
7	72998952	1	rs9638180	A	G	0.12	0.18	0.08							Zbtb3	5.8kb 5' of TBL2	5.9kb 3' of TBL2	
7	72999105	1	rs9638182	T	G	0.12	0.18	0.14								6kb 5' of TBL2	6.1kb 3' of TBL2	
7	73007943	1	rs1051921	G	A	0.12	0.18	0.08					4 cell types	POL2		MLXIPL	MLXIPL	3'-UTR

HMEC,iPS,A549,H1-hESC,H7-hESC,HA-h,HNPCEpiC,HRE

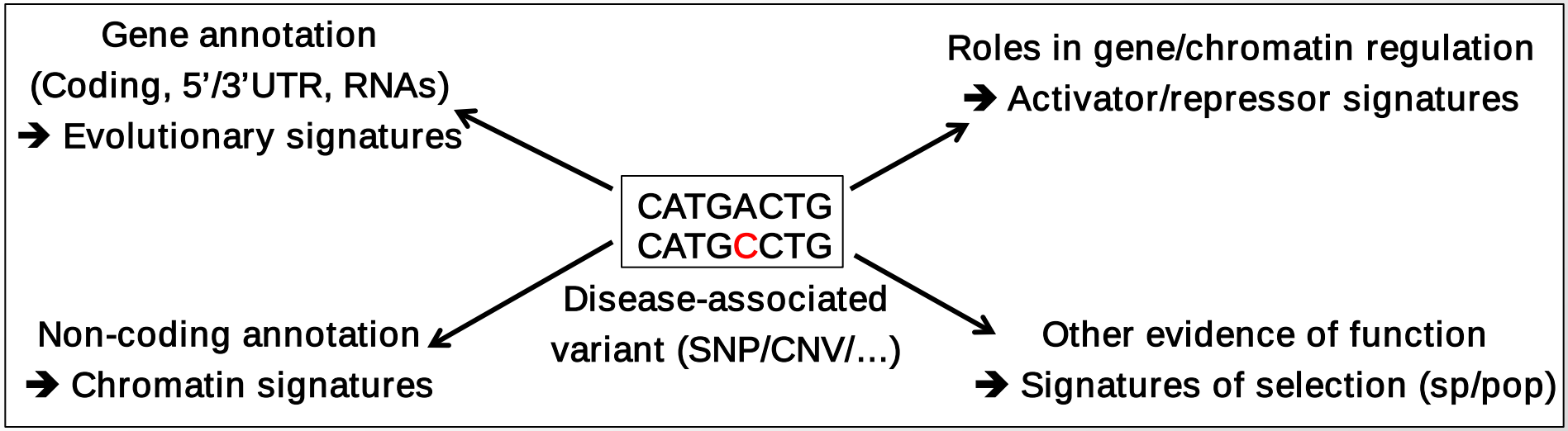
- **Start with any list of SNPs or select a GWA study**
  - Mine publically available ENCODE data for significant hits
  - Hundreds of assays, dozens of cells, conservation, motifs
  - Report significant overlaps and link to info/browser

Ward and Kellis, NAR 2011

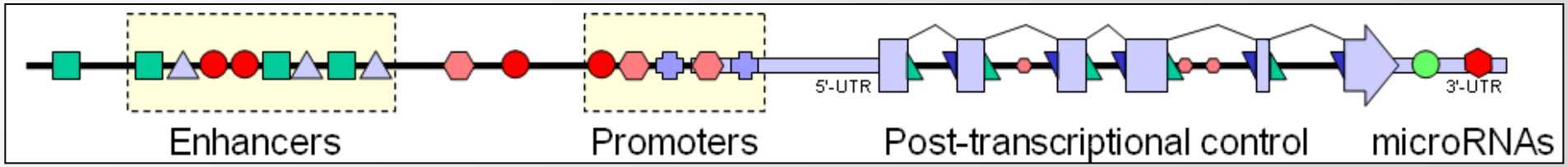
Insights from integrative analysis

1. **Annotate coding/non-coding genes**
   - Peptides, structures, microRNAs, readthrough
2. **Annotate chromatin regulatory regions**
   - Enhancers, promoters, diversity of functions
3. **Define regulator targets and networks**
   - Hierarchy, TF/miRNA networks, HOT regions
4. **Predictive models of gene regulation**
   - Functional nets → gene function/expression
5. **Implications for human disease**
   - Annotate non-coding SNPs, link to TFs/targets

Interpreting complex disease: from regions to models



- **Challenge: from loci to mechanism, pathways, drug targets**



Need: A systems-level understanding of genomes and gene regulation

- The regulators: Transcription factors, microRNAs, sequence specificities
- The regions: enhancers, promoters, and their tissue-specificity
- The targets: TFs→targets, regulators→enhancers, enhancers→genes
- The grammars: Interplay of multiple TFs → prediction of gene expression

➔ The parts list = Building blocks of genome/disease regulatory networks

Integrative Analysis of the *Caenorhabditis elegans* Genome by the modENCODE Project

Mark B. Gerstein,[1,2,3*†] Zhi John Lu,[1,2*] Eric L. Van Nostrand,[4*] Chao Cheng,[1,2*]
Bradley I. Arshinoff,[5,6*] Tao Liu,[7,8*] Kevin Y. Yip,[1,2*] Rebecca Robilotto,[1*] Andreas Rechtsteiner,[9*]
Kohta Ikegami,[10*] Pedro Alves,[1*] Aurelien Chateigner,[11*] Marc Perry,[5*] Mitzi Morris,[12*]
Raymond K. Auerbach,[1*] Xin Feng,[5,22*] Jing Leng,[1*] Anne Vielle,[13*] Wei Niu,[14,15*]
Kahn Rhrissorrakrai,[12*] Ashish Agarwal,[2,3] Roger P. Alexander,[1,2] Galt Barber,[16] Cath[illegible] Brdlik,[4]
Jennifer Brennan,[10] Jeremy Jean
Sergio Contrino,[11] Luke O. Danne
Andréa C. Dosé,[18] Jiang Du,[3] The
Elise A. Feingold,[21] Reto Gassma
Michelle Gutwein,[12] Mark S. Guy
Stefan R. Henz,[29] Angie Hinrichs
Judith Janette,[15] Morten Jensen,
Vishal Khivansara,[23] Ekta Khura
Isabel Latorre,[13] Amber Leahey,
Rebecca F. Lowdon,[21] Yaniv Lublin
Marco Mangone,[12] Sheldon McKa
David M. Miller III,[27] Andrew Mu
Taryn Phippen,[9] Elicia A. Preston
Joel Rozowsky,[1,2] Kim Rutherfor
Andrea Sboner,[1,2] Paul Scheid,[12]
Cindie Slightam,[35] Richard Smith
Teruaki Takasaki,[9] Dionne Vafeac
Christina M. Whittle,[10] Beijing W
Xingliang Zhou,[10] modENCODE
Kristin C. Gunsalus,[12,37†] Gos Mi
LaDeana W. Hillier,[20†] Steven He
Lincoln Stein,[5,6,34†] Jason D. Lieb

We systematically generated larg
Caenorhabditis elegans, a key m
across a developmental time cou
sites, and maps of chromatin org
gene models, including alternati
hierarchical networks
chromosomal location
patterns of chromatin
arms and centers, with
Integrating data types,
gene expression. Overa

Identification of Functional Elements and Regulatory Circuits by *Drosophila* modENCODE

The modENCODE Consortium,* Sushmita Roy,[1,2†] Jason Ernst,[1,2†] Peter V. Kharchenko,[3†]
Pouya Kheradpour,[1,2†] Nicolas Negre,[4†] Matthew L. Eaton,[5†] Jane M. Landolin,[6†]
Christopher A. Bristow,[1,2†] Lijia Ma,[4†] Michael F. Lin,[1,2†] Stefan Washietl,[1†]
Bradley I. Arshinoff,[7,18†] Ferhat Ay,[1,33†] Patrick E. Meyer,[1,30†] Nicolas Robine,[8†]
Nicole [illegible]ashington,[9†] Luisa Di Stefano,[1,31†] Eugene Berezikov,[23‡] Christopher D. Brown,[4‡]
[illegible]rr,[10‡] Irwin Jungreis,[1,2‡]
[illegible]olstorukov,[3‡] Sebastian Will,[1‡]
W. Booth,[6] Angela N. Brooks,[28] Qi Dai,[8]
[illegible][5] Andrey A. Gorchakov,[11] Tingting Gu,[15]
Heather K. MacAlpine,[5] John Malone,[12]
[illegible],[8] Marc Perry,[18] Sara K. Powell,[5]
[illegible]a,[19] Jeremy E. Sandler,[6] Yuri B. Schwartz,[3]
[illegible]rijke van Baren,[20] Kenneth H. Wan,[6]
[illegible],[17] Mark Guyer,[17] Rebecca Lowdon,[17]
[illegible]teven E. Brenner,[28,32] Michael R. Brent,[20]
[illegible]ngeras,[13,16] Robert Grossman,[4]
[illegible]t,[34] Mitzi I. Kuroda,[11] Terry Orr-Weaver,[22]
[illegible]sakony,[26] Bing Ren,[26] Steven Russell,[10]
[illegible]ewis,[9] Gos Micklem,[10] Brian Oliver,[12]
[illegible]§‖ Gary H. Karpen,[6,28§‖] Eric C. Lai,[8§‖]
White,[4§‖] Manolis Kellis[1,2‖]

[illegible]ated into cellular and developmental
[illegible]a of DNA Elements (modENCODE) project
[illegible]tions, chromosomal proteins, transcription
[illegible]leosome properties across a developmental
[illegible]ed more than 700 data sets and discovered
[illegible], and chromatin elements, more than
[illegible]ne. Correlated activity patterns of these
predicts putative new functions for genes,
[illegible]es gene-expression prediction. Our results
[illegible]mputational studies in *Drosophila* and
[illegible]tegration toward comprehensive genomic

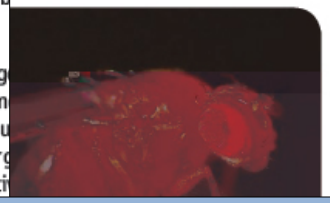
Science Ex[illegible]: Published [illegible]e 12/22/2010

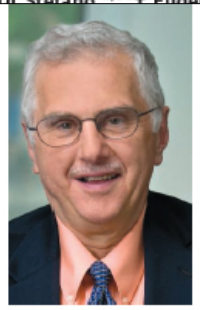
EDITORIAL

Model Organisms and Human Health

IN THIS ISSUE OF *SCIENCE*, WE HIGHLIGHT THE IMPRESSIVE EFFORTS TO DESCRIBE AND ANALYZE the genomes of the two organisms—the fly *Drosophila melanogaster* and the nematode worm *Caenorhabditis elegans*—that serve as the best models for understanding the biology of all animals, including humans. Hundreds of scientists have collaborated in these two major studies, which have moved us far beyond the complete descriptions of the DNA molecules that make up the fly and worm genomes published a little more than a decade ago, an accomplishment that seemed amazing then. As summarized in the Perspective on p. 1758, the new findings reveal essentially all of the tens of thousands of RNA and protein molecules that each of these organisms produces, as well as how their genetic information is packaged. Extensive Web-based databases built on these data are freely available to everyone, greatly accelerating future discoveries. Strange as it may seem, this research, aimed at reaching a deep molecular understanding of how the bodies of a fly and a worm are formed and maintained, will be critical for improving human health.

Most of the government funding for biomedical research in the United States is distributed through the National Institutes of Health. Its budget of $31 billion in 2010 reflects a widespread public appreciation that biomedical research will lead to great improvements in human health. Despite the many advances in our understanding of



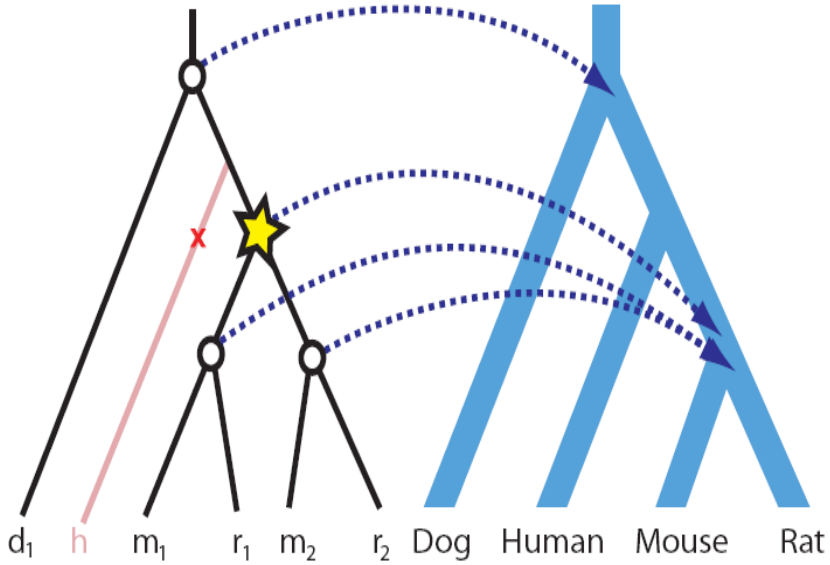


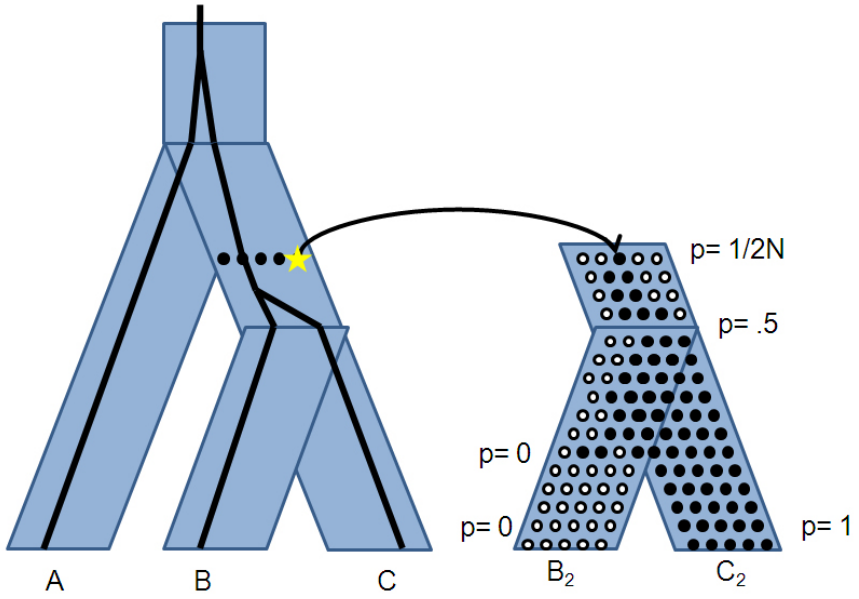
Bruce Alberts is Editor-in-Chief of *Science*.

Next steps: Fly vs. Worm vs. Human

Human-mouse-fly-worm orthologs

- Phylogenomics approaches
  - Species-specific gene-specific rates
  - Incomplete lineage sorting/deep coalescence
  - Unified models of Dup-Loss-Coalescense

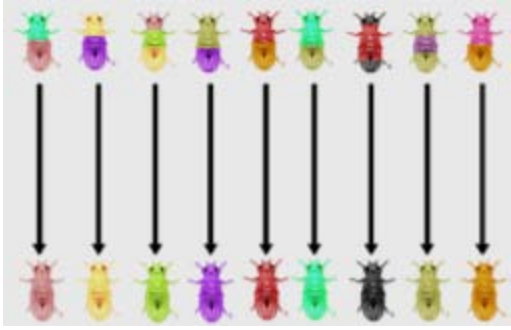




Gene-tree species-tree reconciliation **Deep coalescence of duplicates**

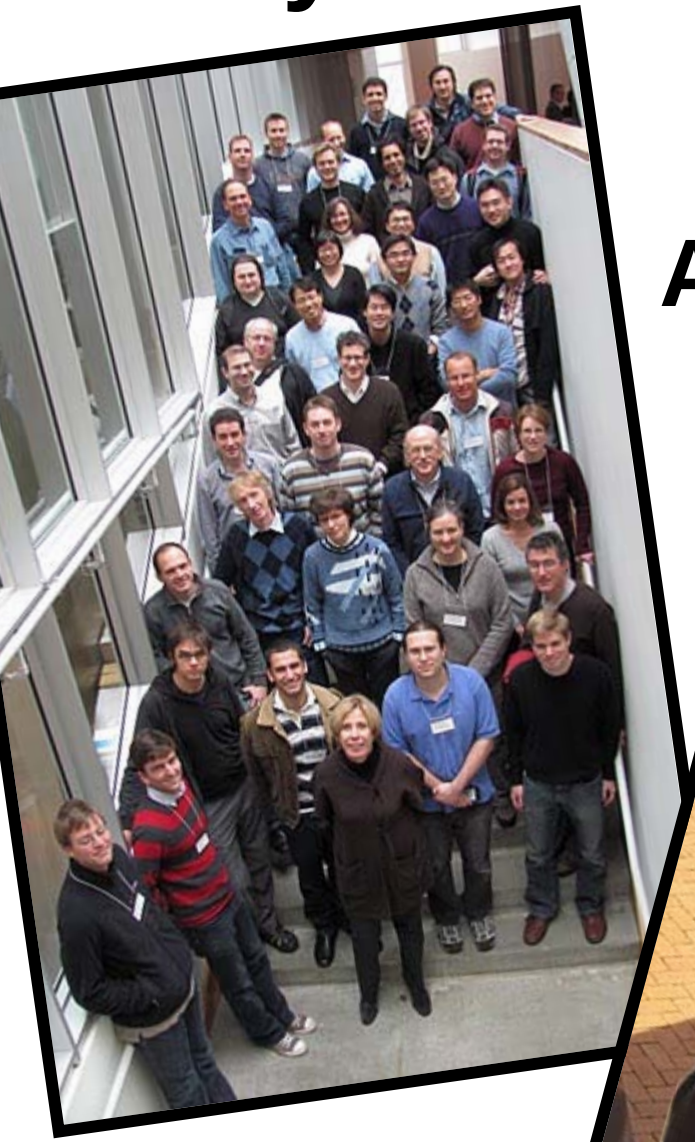
Javier Herrero, Jessica Wu, Matt Rasmussen, Mukul Bansal

Genotype-phenotype



- Population genomics of Drosophila
  - Trudy MacKay, Charles Langley et al
- Selective pressures ⇔ modENCODE
  - Purifying selection, positive selection, recombination hotspots vs. annotations
  - Understand deleterious mutations
- Trait-associated regions ⇔ modENCODE
  - Help annotate trait-associated variants
  - Role of motifs, networks in observed phenotypes
- Systematic mutations and drug screening

Analysis Working Group



AWG
DAC

Data Analysis Center

Acknowledgements

	Fly	Worm
Transcripts (Celniker/Waterston)	Joe Carlson Jane Landolin Ben Booth Brenton Graveley Ben Brown	Mark Gerstein LaDeana Hillier Kevin Yip, Ashish Agarwal Lukas Habegger
Chromatin (Karpen/Lieb)	Peter Park, Peter Kharchenko, Jason Ernst, Matthew Eaton	Shirley Liu Hyunjin (Gene) Shin
TFs (White/Snyder)	Casey Brown Nicolas Negre	Mark Gerstein Lucas Lochovsky Kevin Yip
Nucleosomes (Henikoff)	Steve Henikoff	Steve Henikoff
smRNAs/3'UTRs (Lai/Piano)	Eric Lai (smRNAs) Nicolas Robine	Kris Gunsalus (3'UTRs) Arun Manoharan Marco Mangone
Origins (MacAlpine)	MacAlpine Mathew Eaton	N/A
Statistics/Infrastructure (Bickel/Gerstein)	Ben Brown and Kevin Yip and Nathan Boley	
Conservation/Submissions	Lincoln Stein, Gos Micklem, DCC	

Identification of Functional Elements and Regulatory Circuits by *Drosophila* modENCODE

The modENCODE Consortium,* Sushmita Roy,[1,2]† Jason Ernst,[1,2]† Peter V. Kharchenko,[3]† Pouya Kheradpour,[1,2]† Nicolas Negre,[4]† Matthew L. Eaton,[5]† Jane M. Landolin,[6]† Christopher A. Bristow,[1,2]† Lijia Ma,[4]† Michael F. Lin,[1,2]† Stefan Washietl,[1]† Bradley I. Arshinoff,[7,18]† Ferhat Ay,[1,33]† Patrick E. Meyer,[1,30]† Nicolas Robine,[8]† Nicole L. Washington,[9]† Luisa Di Stefano,[1,31]† Eugene Berezikov,[23]‡ Christopher D. Brown,[4]‡ Rogerio Candeias,[1]‡ Joseph W. Carlson,[6]‡ Adrian Carr,[10]‡ Irwin Jungreis,[1,2]‡ Daniel Marbach,[1,2]‡ Rachel Sealfon,[1,2]‡ Michael Y. Tolstorukov,[3]‡ Sebastian Will,[1]‡ Artyom A. Alekseyenko,[11] Carlo Artieri,[12] Benjamin W. Booth,[6] Angela N. Brooks,[28] Qi Dai,[8] Carrie A. Davis,[13] Michael O. Duff,[14] Xin Feng,[13,18,35] Andrey A. Gorchakov,[11] Tingting Gu,[15] Jorja G. Henikoff,[8] Philipp Kapranov,[16] Renhua Li,[17] Heather K. MacAlpine,[5] John Malone,[12] Aki Minoda,[6] Jared Nordman,[22] Katsutomo Okamura,[8] Marc Perry,[18] Sara K. Powell,[5] Nicole C. Riddle,[15] Akiko Sakai,[29] Anastasia Samsonova,[19] Jeremy E. Sandler,[6] Yuri B. Schwartz,[3] Noa Sher,[22] Rebecca Spokony,[4] David Sturgill,[12] Marijke van Baren,[20] Kenneth H. Wan,[6] Li Yang,[14] Charles Yu,[6] Elise Feingold,[17] Peter Good,[17] Mark Guyer,[17] Rebecca Lowdon,[17] Kami Ahmad,[29] Justen Andrews,[21] Bonnie Berger,[1,2] Steven E. Brenner,[28,32] Michael R. Brent,[20] Lucy Cherbas,[21,24] Sarah C. R. Elgin,[15] Thomas R. Gingeras,[13,16] Robert Grossman,[4] Roger A. Hoskins,[6] Thomas C. Kaufman,[21] William Kent,[34] Mitzi I. Kuroda,[11] Terry Orr-Weaver,[22] Norbert Perrimon,[19] Vincenzo Pirrotta,[27] James W. Posakony,[26] Bing Ren,[26] Steven Russell,[10] Peter Cherbas,[21,24] Brenton R. Graveley,[14] Suzanna Lewis,[9] Gos Micklem,[10] Brian Oliver,[12] Peter J. Park,[3] Susan E. Celniker,[6]§|| Steven Henikoff,[25]§|| Gary H. Karpen,[6,28]§|| Eric C. Lai,[8]§|| David M. MacAlpine,[5]§|| Lincoln D. Stein,[18]§|| Kevin P. White,[4]§|| Manolis Kellis[1,2]||