CSEP 590B "Summary"

Below, as a somewhat unusual "course summary," I have decided to give the bulk of a research talk I presented in our CompBio seminar last spring, partly because I think the content is interesting, but more to show how deeply "computation" is embedded in modern "bio" research, and to show that many of the themes of the course are directly relevant.

Asides emphasizing these connections are highlighted in a sprinkling of boxes like this.

Also note that the last ~40 slides of Lecture 9 "CMs" were actually presented in Lecture 10, but conceptually and logistically it was easier to split the slides this way...

Please cite this article in press as: Cao et al., Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming, Developmental Cell (2010), doi:10.1016/j.devcel.2010.02.014

Developmental Cell
Resource

Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming

Yi Cao,[1,7] Zizhen Yao,[2,7] Deepayan Sarkar,[2] Michael Lawrence,[2] Gilson J. Sanchez,[1,4] Maura H. Parker,[3] Kyle L. MacQuarrie,[1,4] Jerry Davison,[2] Martin T. Morgan,[2] Walter L. Ruzzo,[2,5] Robert C. Gentleman,[2,*] and Stephen J. Tapscott[1,3,6,*]

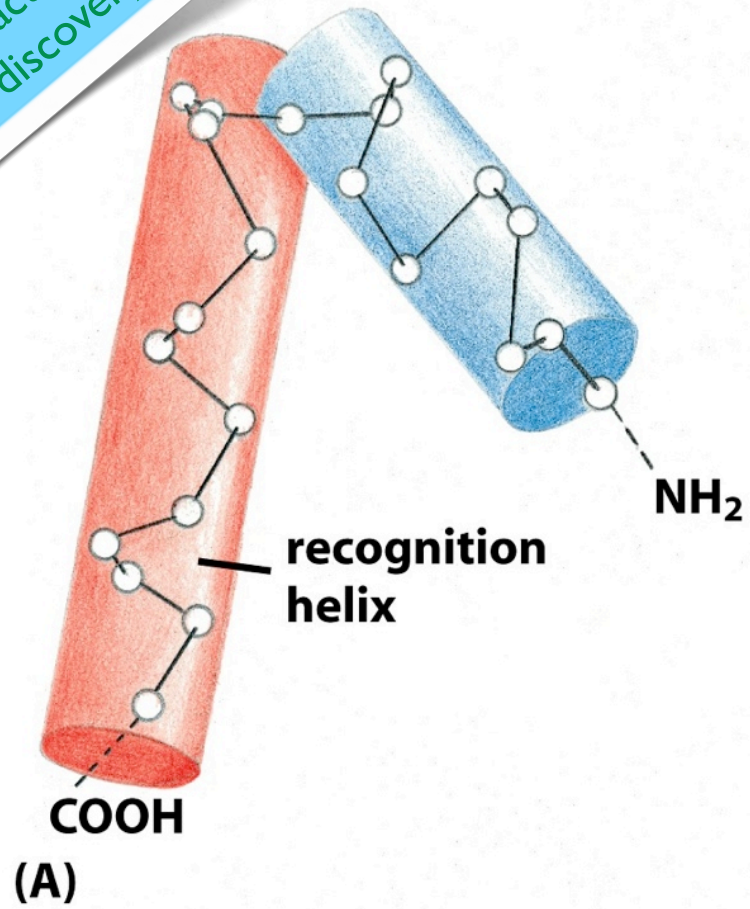
April 2010

Goal: To give you a sense of where the "comp" fits in a modern "bio" paper.

Outline

Transcription factors & MyoD
Chromatin Immunoprecipitation (ChIP)
ChIP-seq
Computational Methods
Results

Transcription factors & TFBS motif discovery



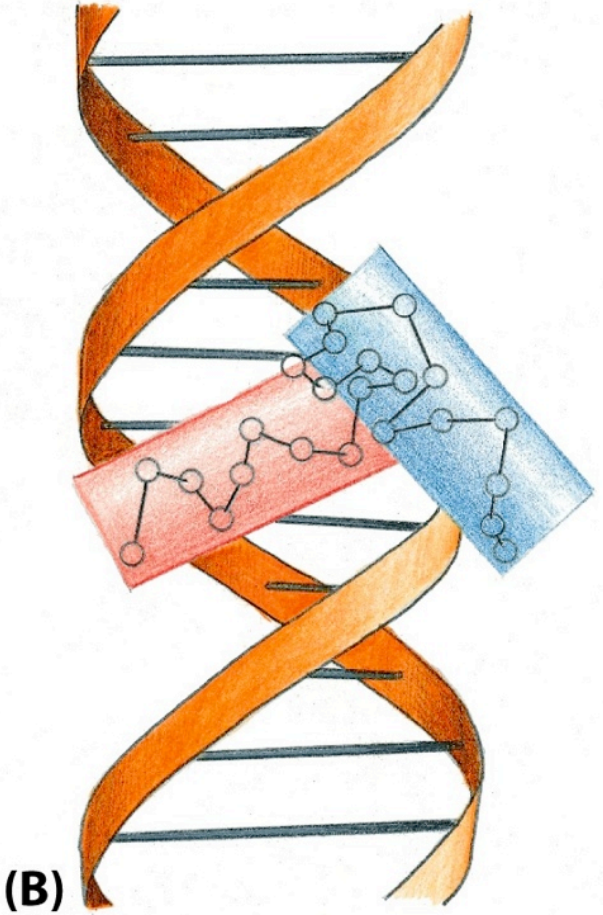
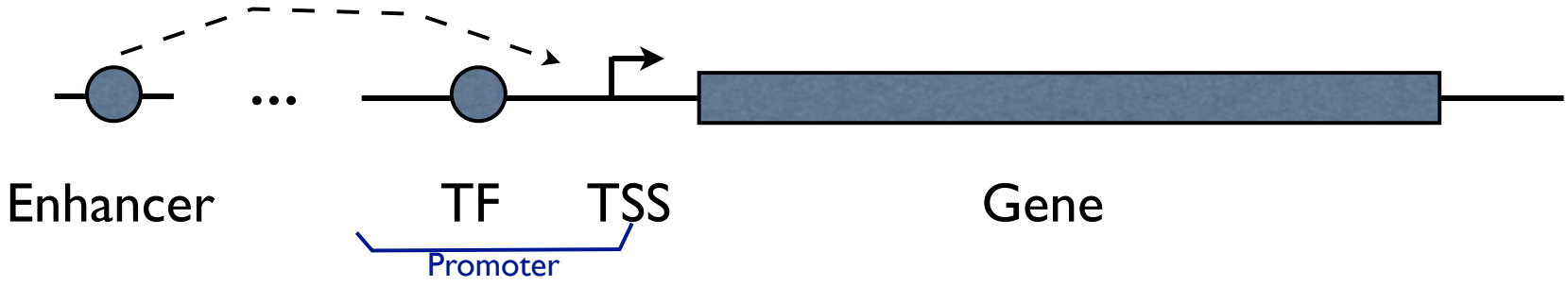


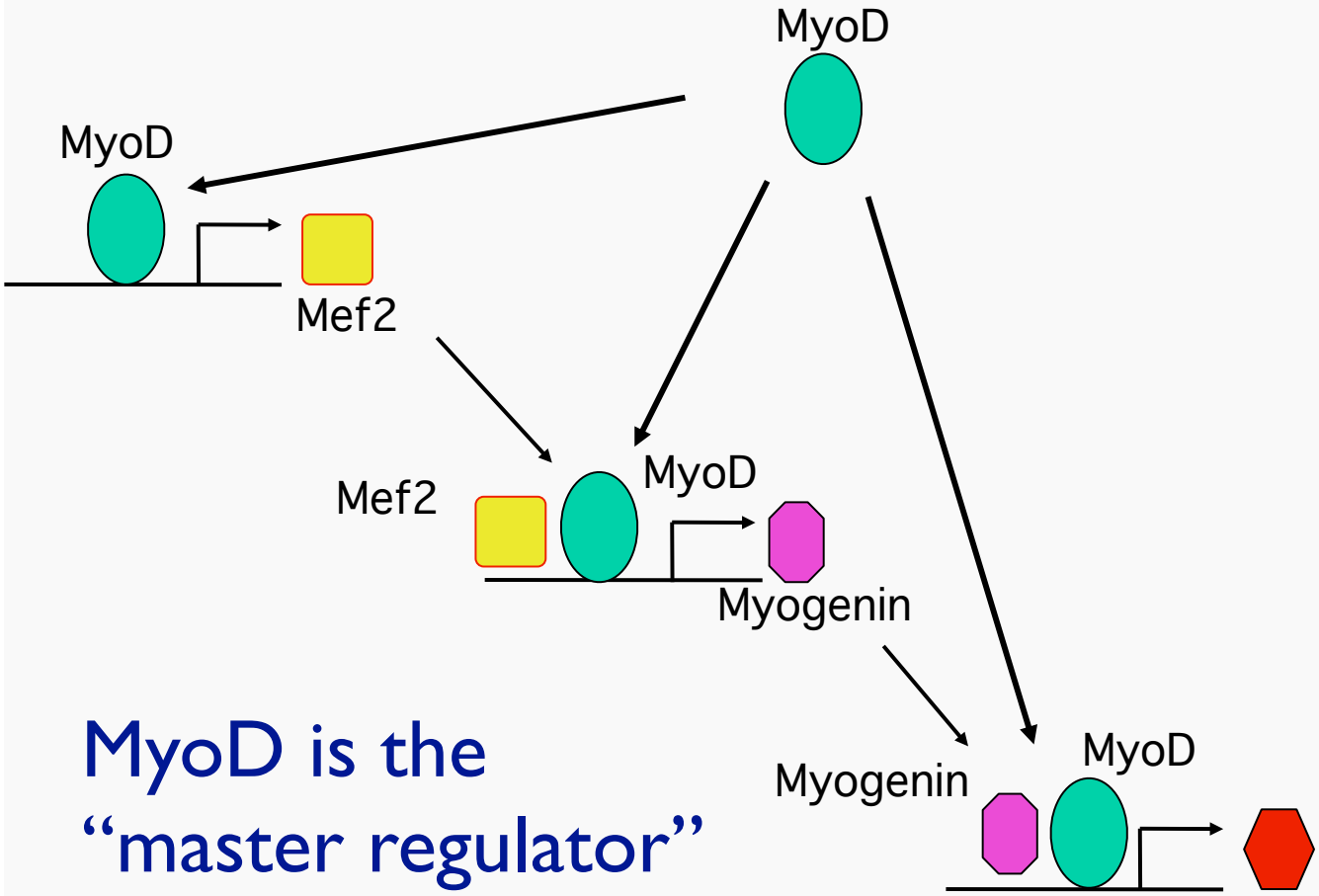
Figure 7-10 Molecular Biology of the Cell 5/e (© Garland Science 2008)



Myogenesis:

Myoblast – a muscle precursor

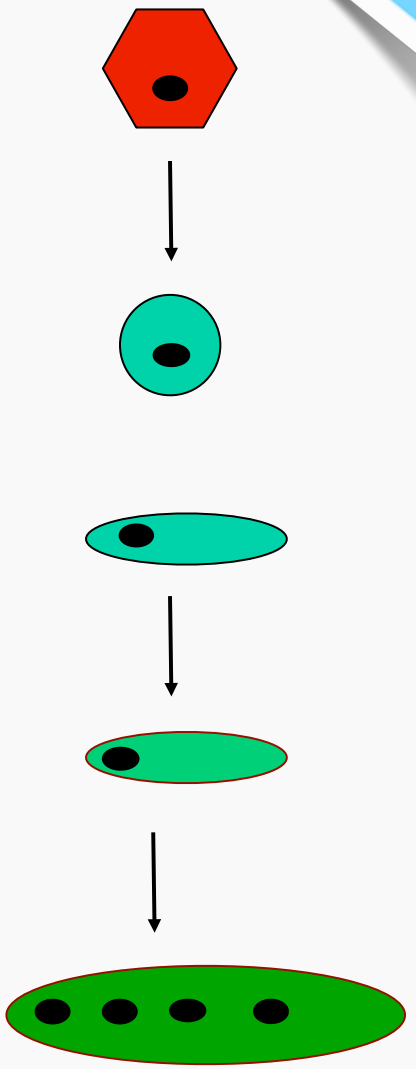
Myotube – differentiated skeletal muscle cell



MyoD is the "master regulator"

Other players: Mef2, MyoG, ...

Network motifs



"Standard Model"

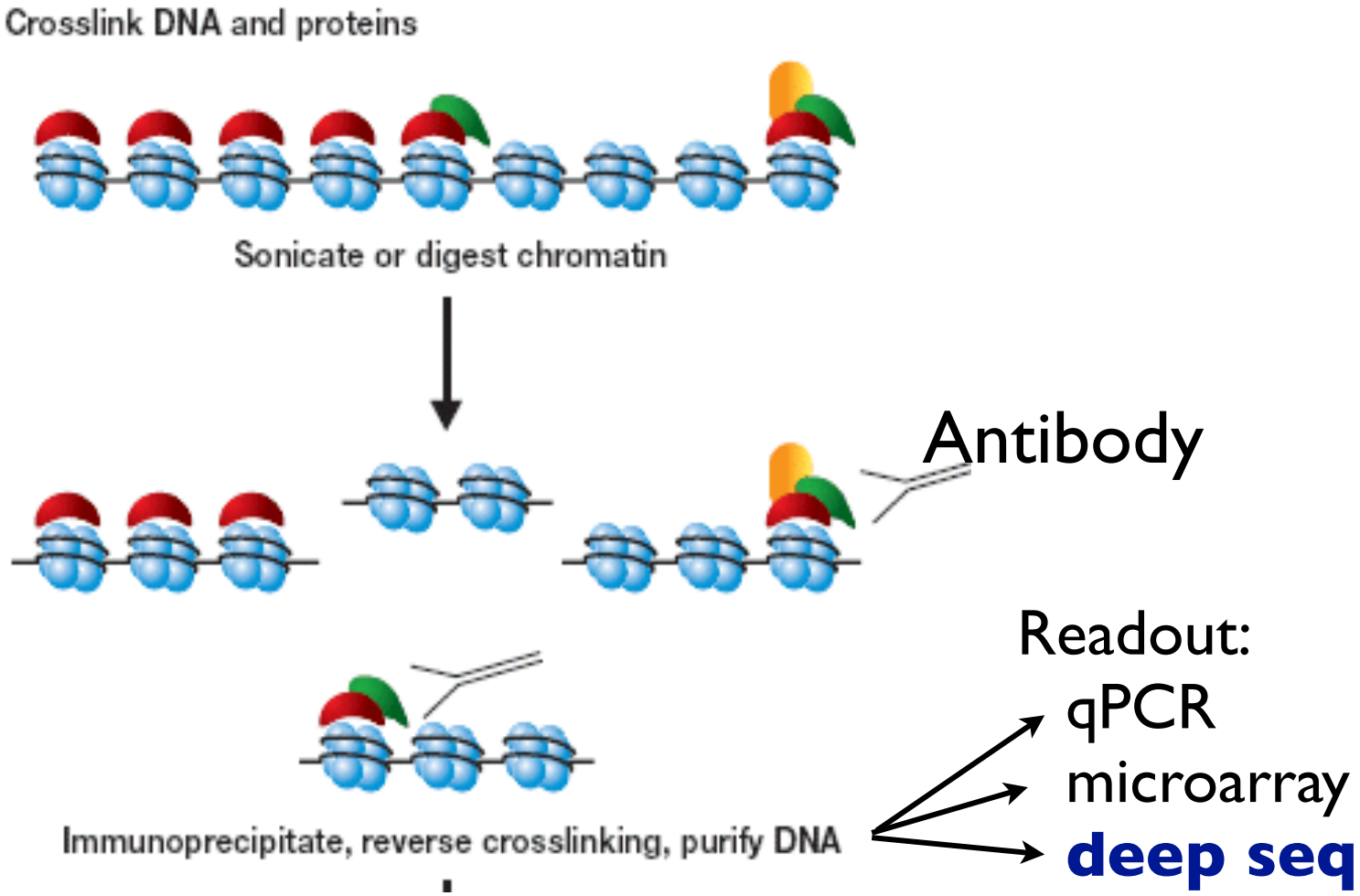
MyoD absent or low in myoblasts

Triggering it in myoblasts (or many other cell types) starts a cascade leading to myotubes

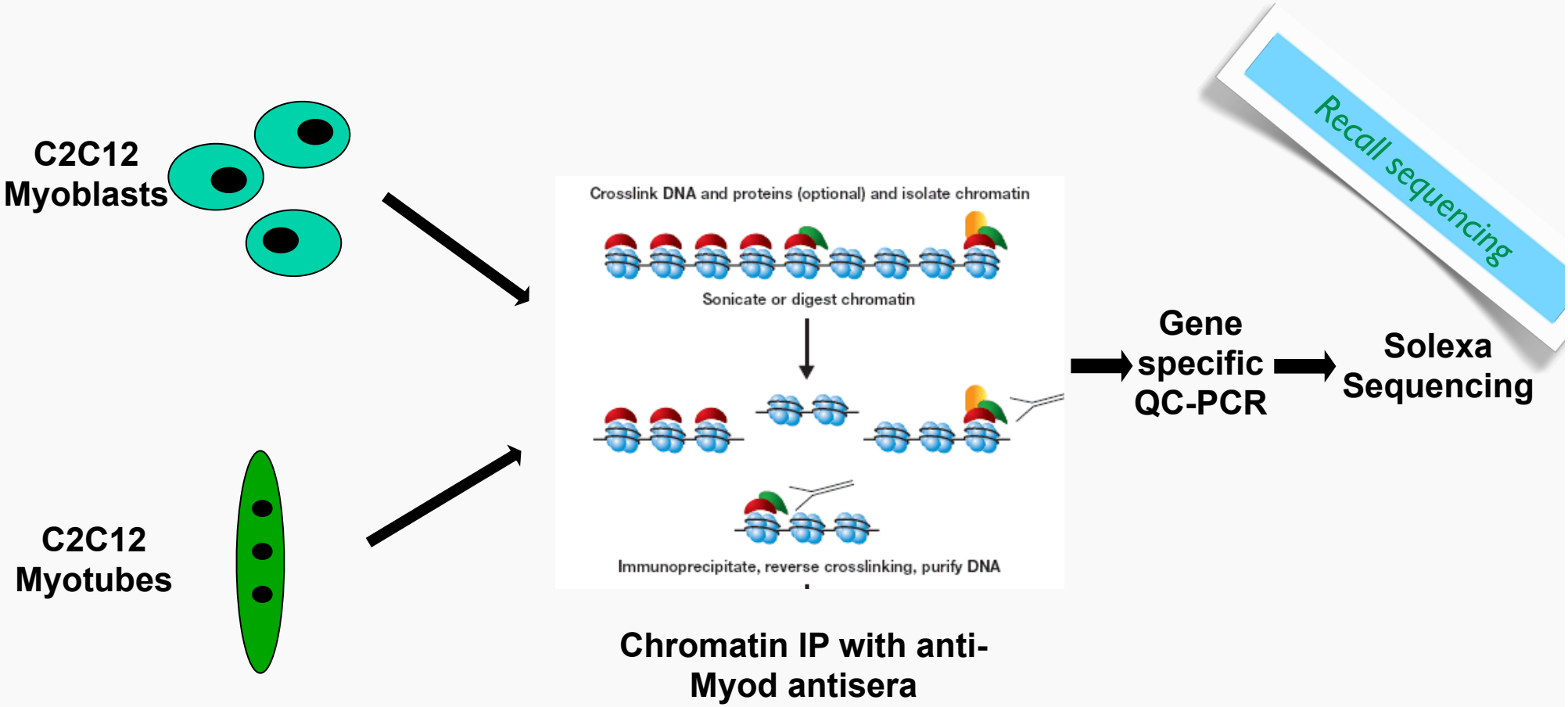
500-1500 genes show differential expression between myoblasts & myotubes

Expectation: *MyoD drives those changes, by binding their promoters, plus a few enhancer sites*

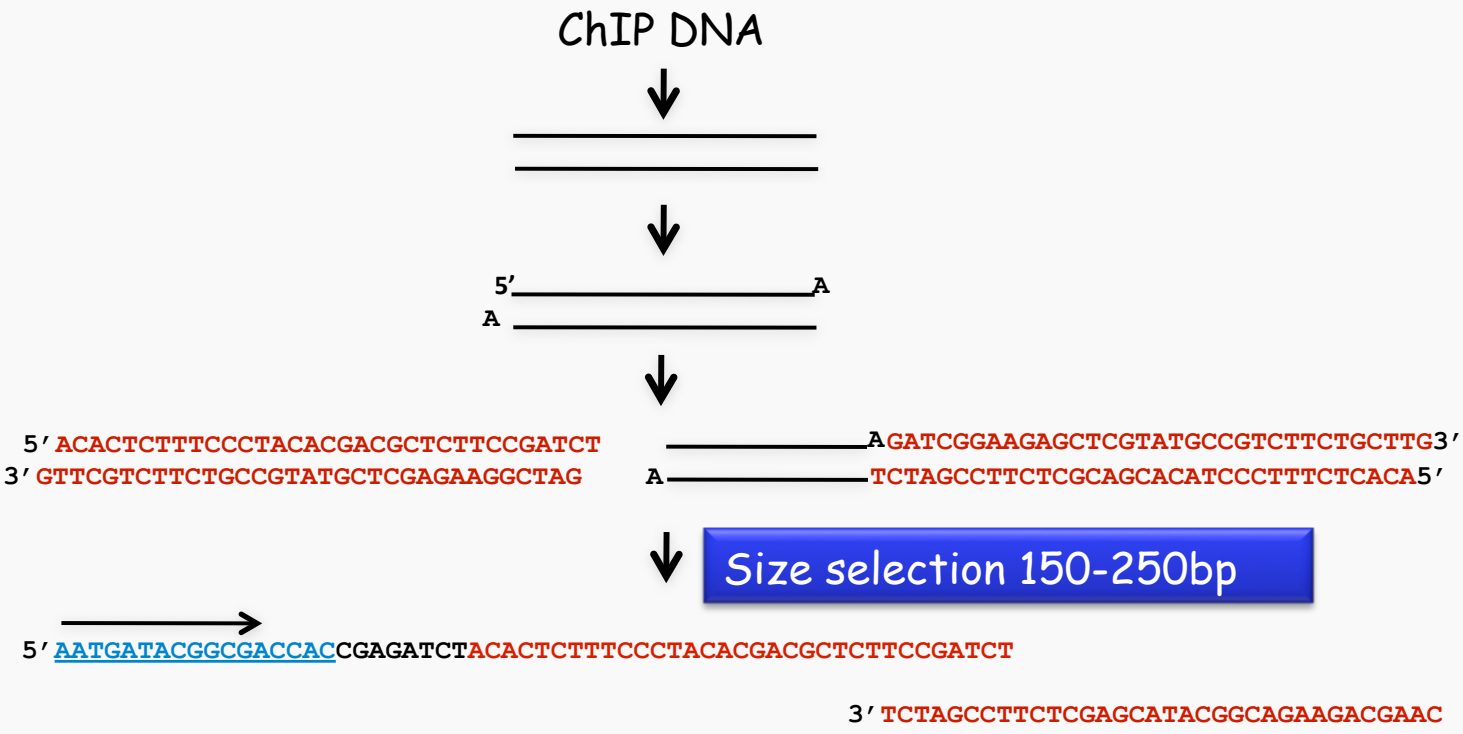
Chromatin Immunoprecipitation



MyoD Experimental Design



ChIP-seq Sample Prep



End repair

3'-dA overhang

Adapter ligation

PCR amplification

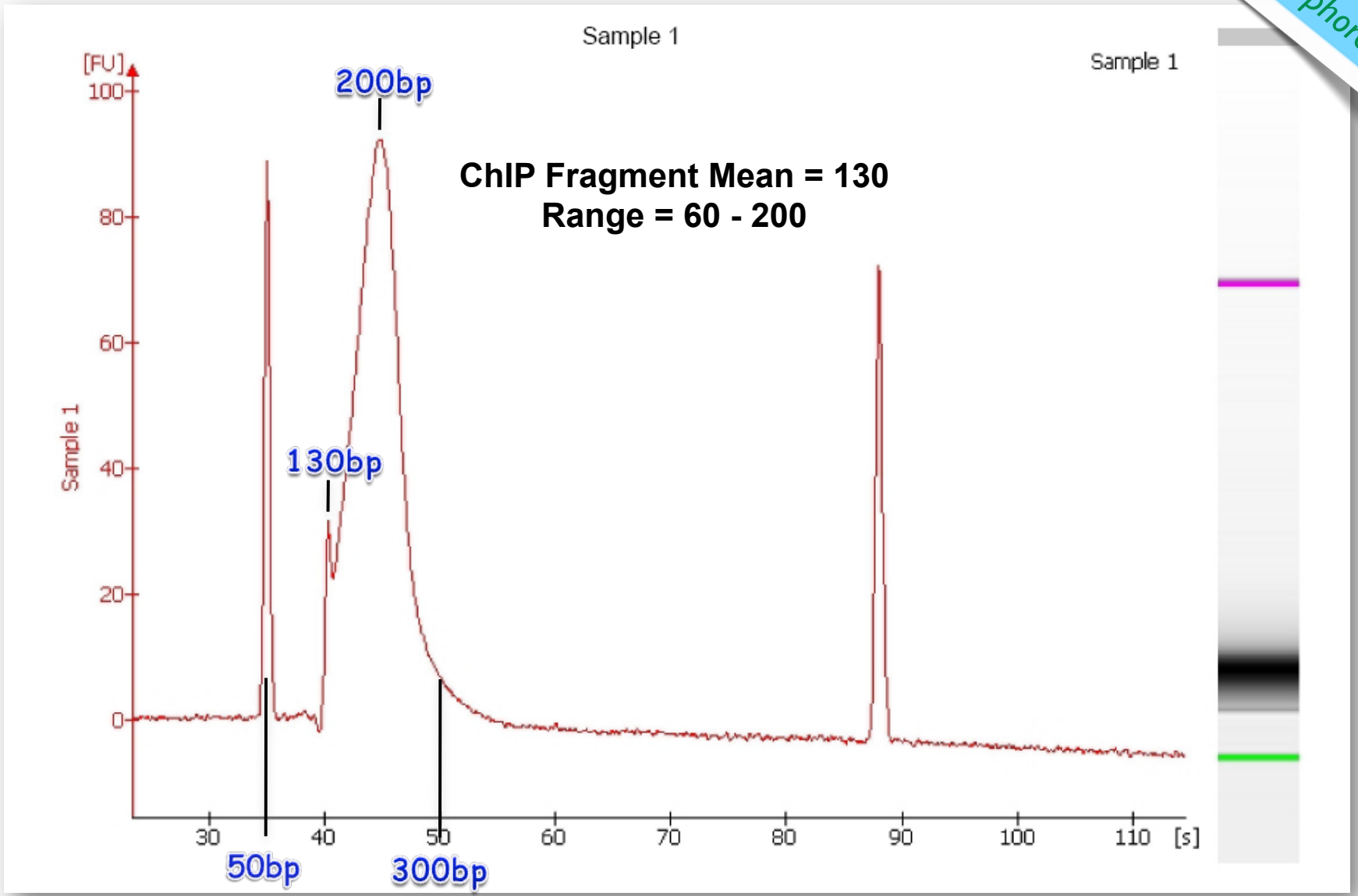
Recall PCR

Adaptor total length: 67bp

Load 2 **picomoles** on the machine

Bioanalyzer Analysis

Recall gel electrophoresis



Analysis & Methods

ChIP-seq Analysis

Yields 5-20M "reads" per lane (8 lanes per run, usually 8 different samples)

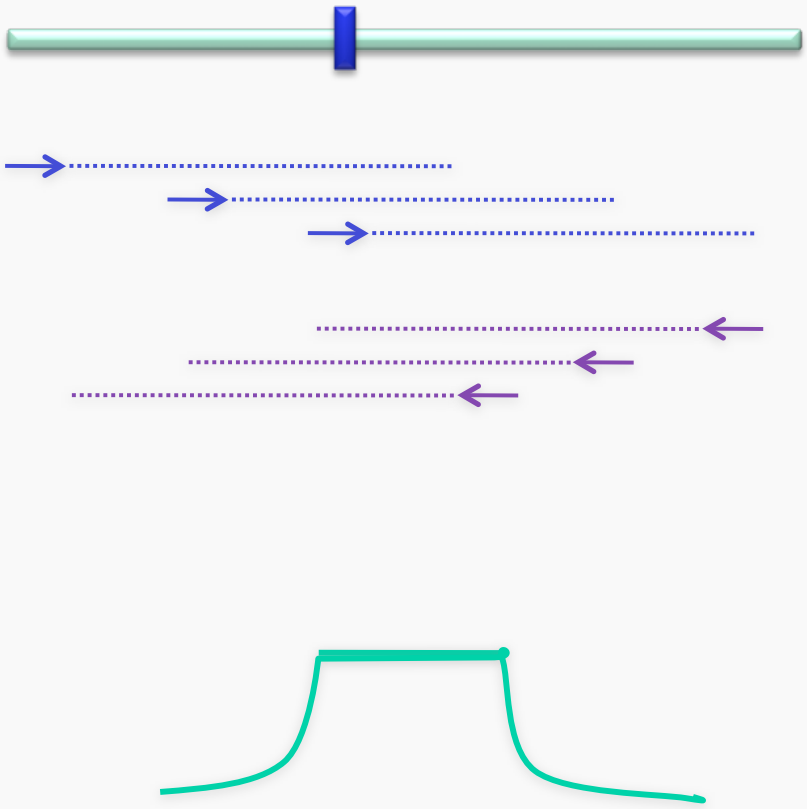
Reads (35-55 bp, depending on run) are mapped back to the mouse reference genome.

- only one copy of dup reads retained (PCR artifacts?)
- tolerate 2 bp mismatch among 1st 28 bp
- reads not mapping uniquely are discarded

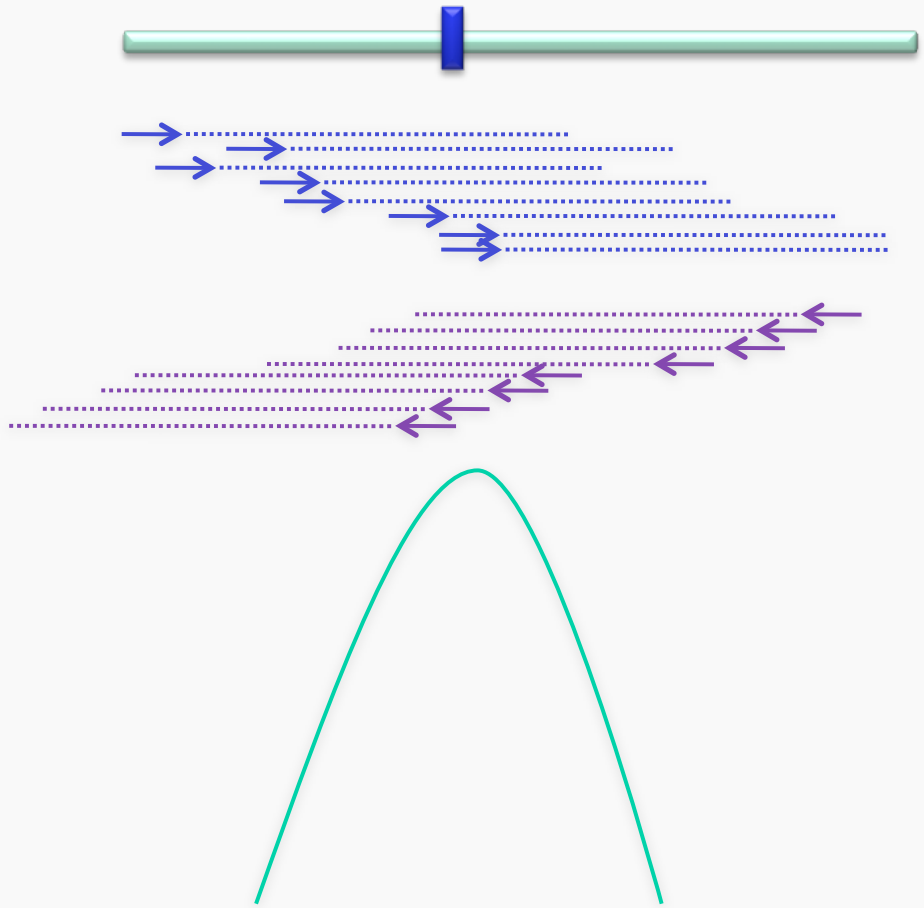
"Extended read" – pretend each is 200 bp

Overlapping extended reads presumably mark binding sites

Identification of Binding Regions



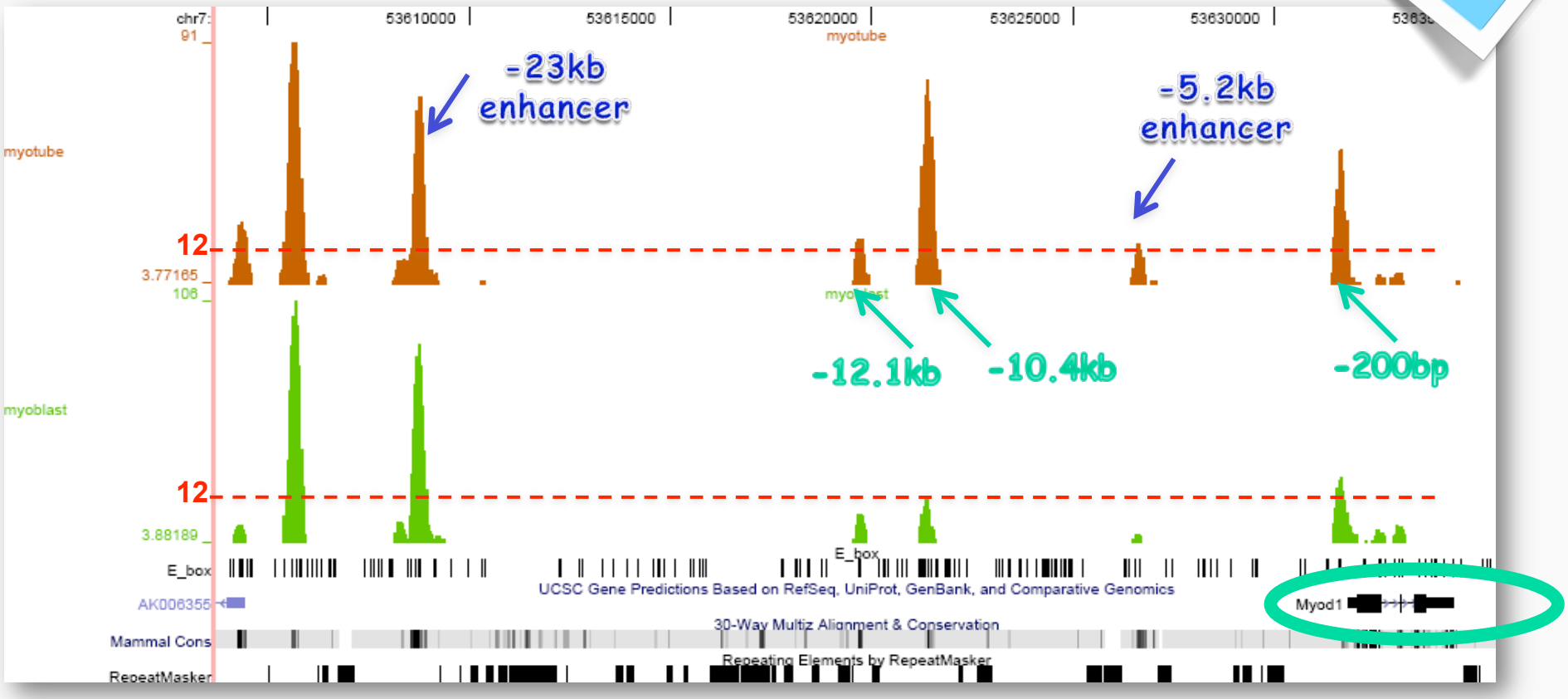
Few reads = short flat peaks



Many reads = High sharp peaks

Myod Locus

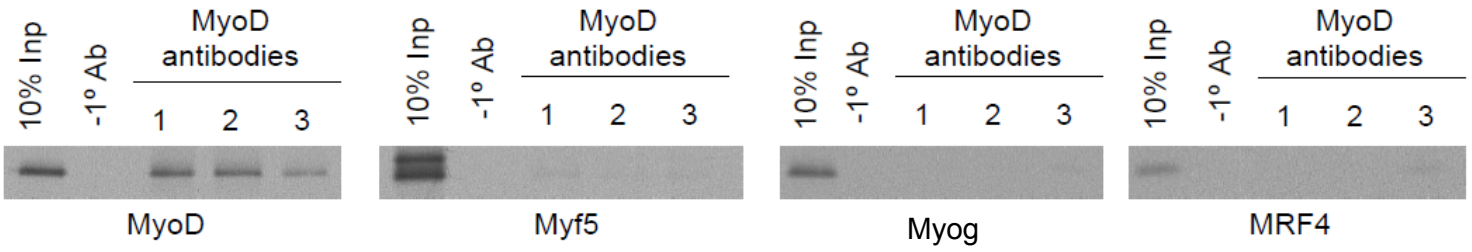
Recall gene, TFBS prediction. Also, multi-way alignment, repeat discovery,...



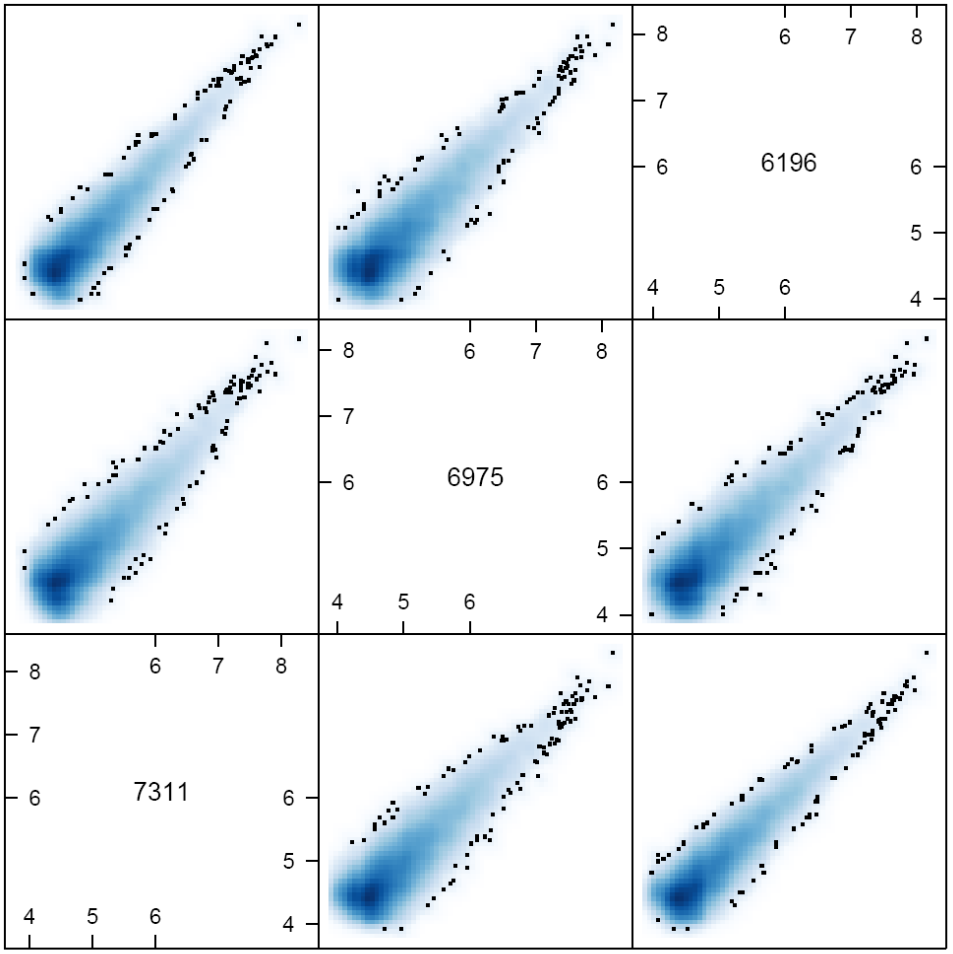
MyoD

Are the antibodies any good?

A

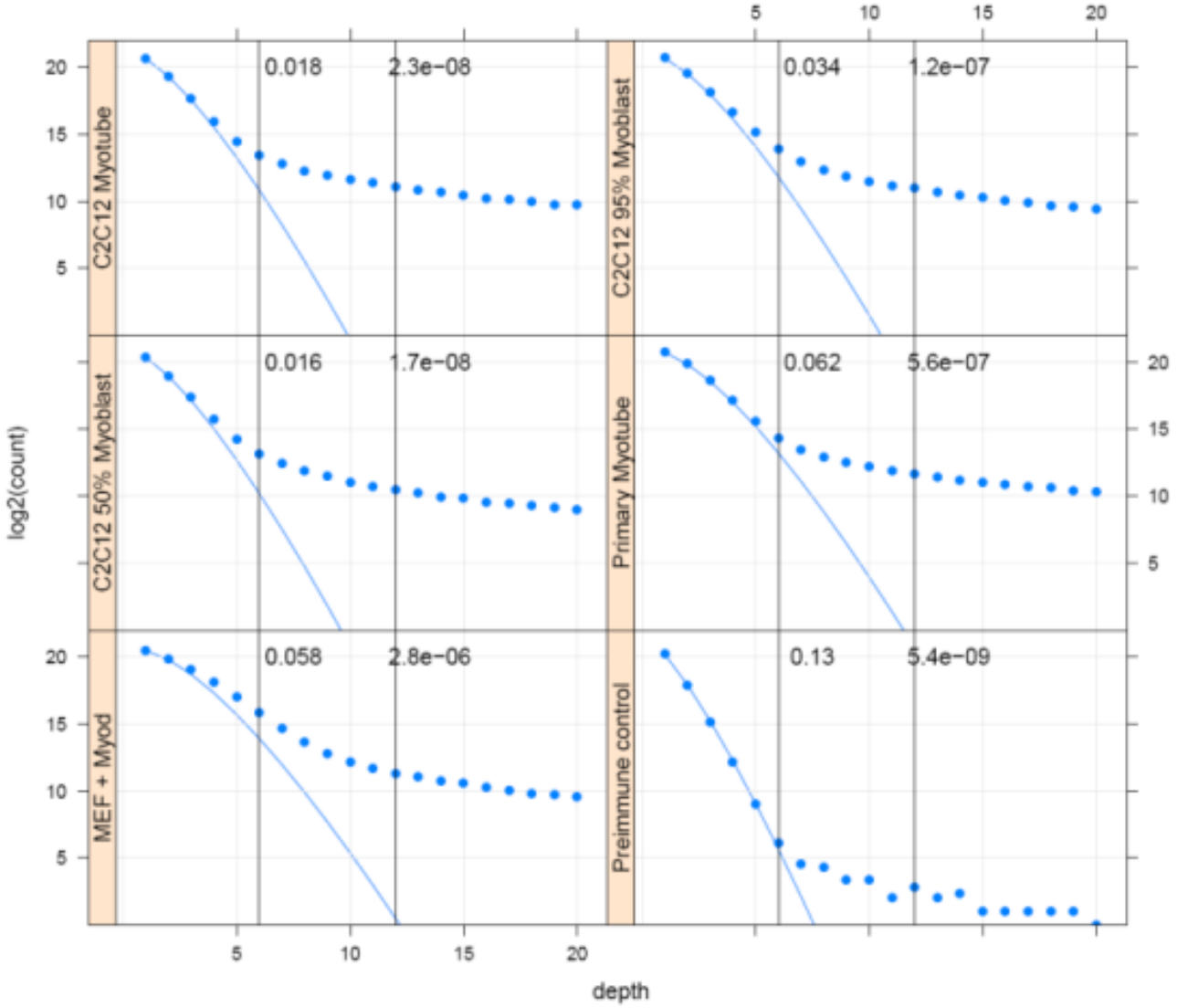


B



Analysis questions: How tall must a peak be?

Recall likelihood ratio tests, etc.



Estimate Poisson null model from "islands" of height 1, 2.

How likely is height 6? 12?

Results

“Standard Model”

MyoD binding absent or rare in myoblasts

Triggering it in myoblasts (or many other cell types) starts a cascade leading to myotubes

500-1500 genes show differential expression between blasts & tubes

Expectation: MyoD binds their promoters & drives those changes

How Many Peaks?

As opposed to the 500-1500 genes changed, we find MyoD bound to 25,956 loci in myotubes (at 12-read cutoff; FDR < 10^{-6})

In myotubes and myoblasts both

> 60,000 at ~.01 FDR

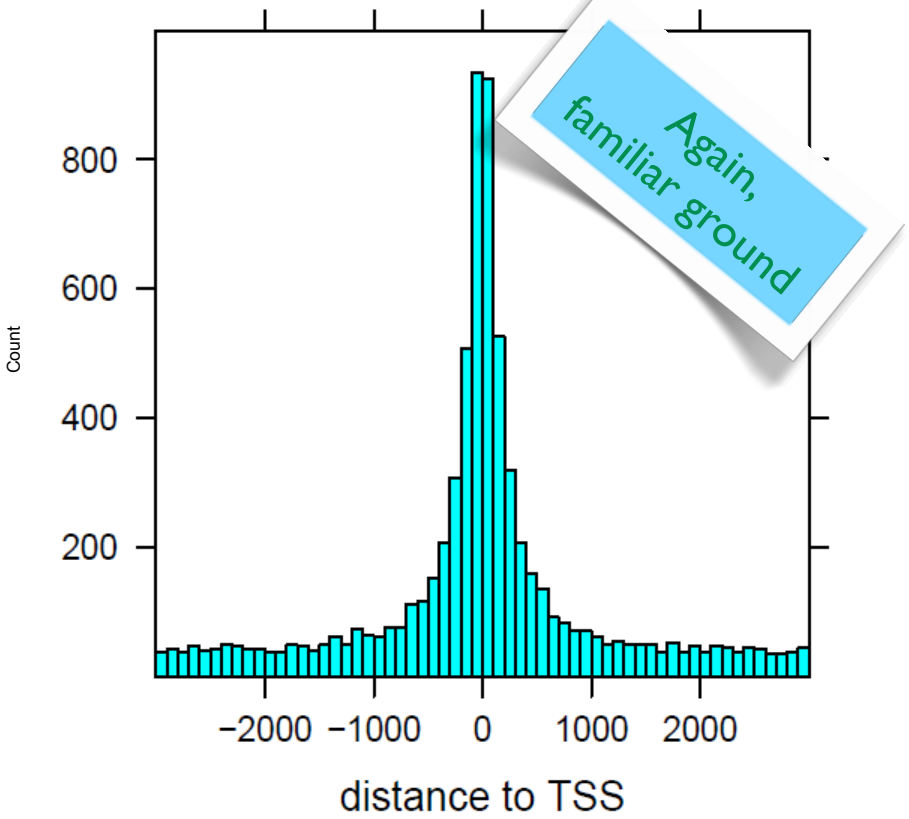
(Excludes X, Y, repetitive regions)

much computational analysis

Where are peaks?

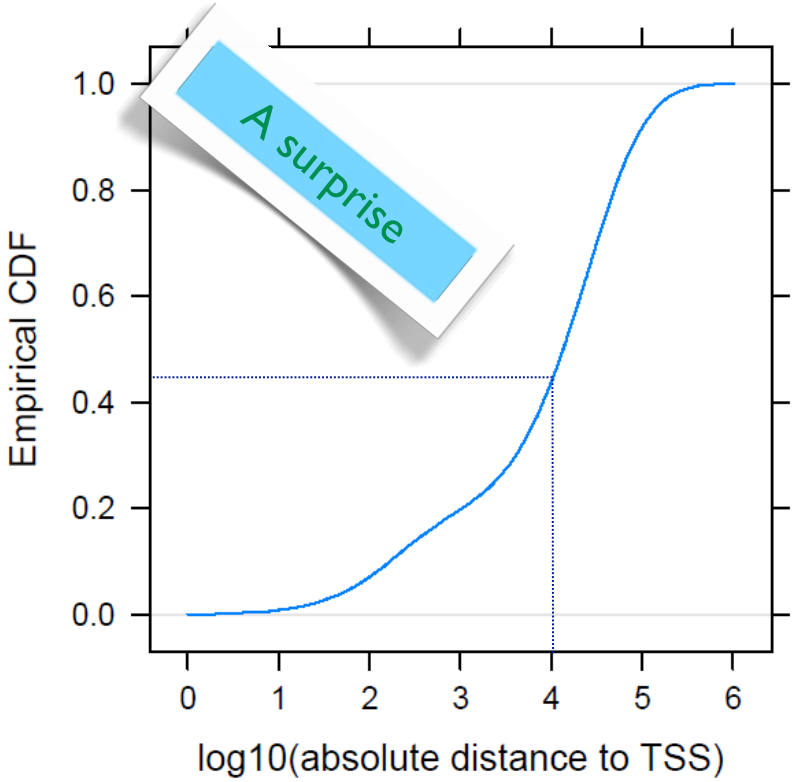
Concentrated at promoters

Again, familiar ground



But 50% are >10k from any TSS

A surprise



Binds 41% of genes.

much computational analysis

Where are peaks?

Table 1. Gene Context Analysis of Myod-Binding Sites

	Number of Peaks[a]		Number of Genes[b]		Number of Peaks/kb[c]	
	Myotube	Myoblast	Myotube	Myoblast	Myotube	Myoblast
Promoter[d]	4772	4982	4256	4502	0.153	0.160
Promoter proximal[e]	6349	6417	5085	5313	0.055	0.056
3′[f]	694	517	579	433	0.022	0.017
Exon	2031	1615	1554	1283	0.032	0.026
Intron	8780	7239	5443	4957	0.011	0.010
Upstream[g]	3739	3124	2600	2272	0.015	0.012
Downstream[h]	3985	3237	2901	2447	0.015	0.013
Intergenic[i]	6254	5776	0	0	0.005	0.004
Total	25956	23271				

[a] Number of Peaks: the number of peaks in each category.
[b] Number of Genes: the number of genes that peaks are associated with in each category, measured by unique Entrez IDs. If one or more peaks are located in multiple alternative splice variants of one gene, only one gene is counted.
[c] Number of Peaks/kb: number of peaks divided by the total size of the corresponding genomic region in kilobases.
[d] Promoter: ±500 bp from the transcription start site (TSS).
[e] Promoter proximal: ±2 kb from the TSS.
[f] 3′ end: ±2 kb from the end of the transcript.
[g] Upstream: −2 kb to −10 kb upstream of the TSS.
[h] Downstream: +2 kb to +10 kb from the end of the transcript.
[i] Intergenic: >10 kb from the annotated gene.

much computational analysis

What's it doing?

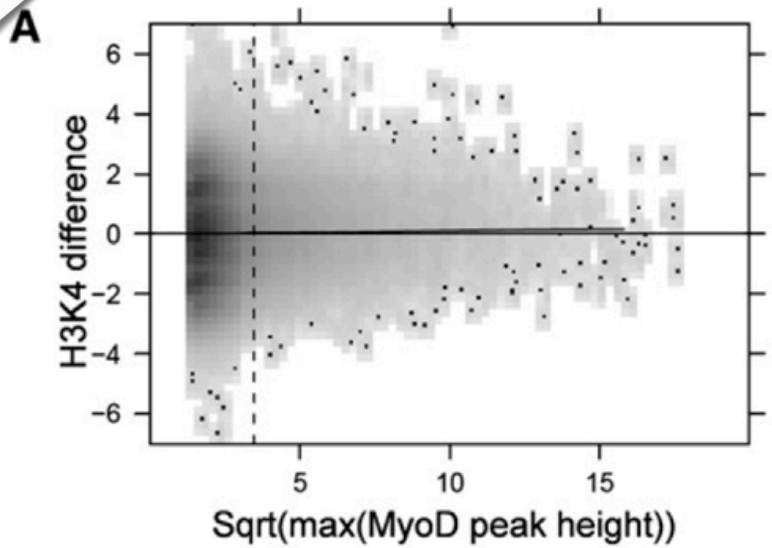
2. Gene Ontology Analysis on Differentially Bound Peaks in Myoblasts versus Myotubes

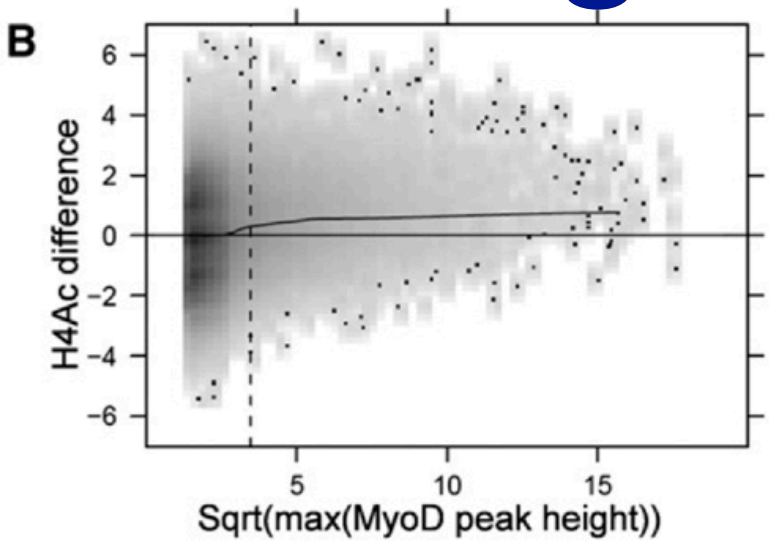
O Categories Enriched in Genes Associated with Myotube-Increased Peaks

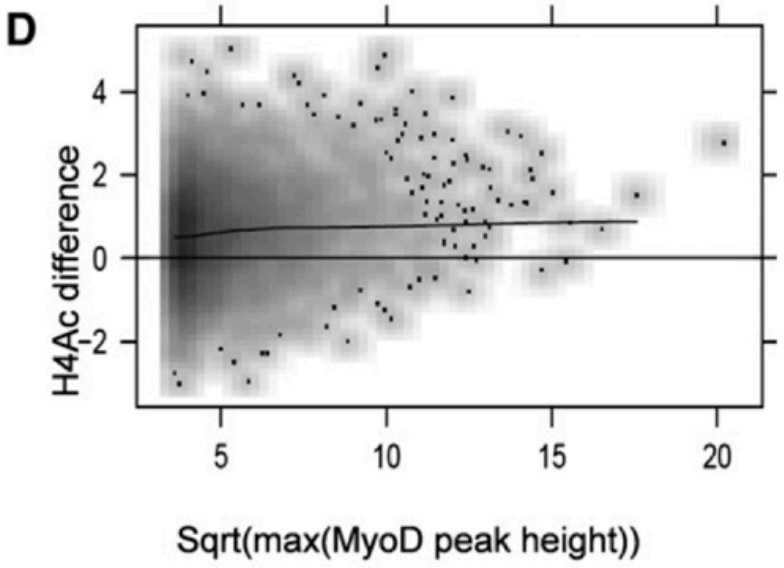
GOID	Term	P Value	OR[a]	Count[b]	Size[c]	Ont[d]
GO:0005856	cytoskeleton	2.05E-11	2.40	94	490	CC
GO:0043292	contractile fiber	6.98E-09	5.85	22	58	CC
GO:0030016	myofibril	1.96E-08	5.74	21	56	CC
GO:0044449	contractile fiber part	2.58E-08	5.97	20	52	CC
GO:0030017	sarcomere	4.95E-08	6.04	19	49	CC
GO:0008092	cytoskeletal protein binding	3.69E-07	2.52	47	227	MF
GO:0007519	skeletal muscle development	2.50E-06	4.13	20	65	BP
GO:0015629	actin cytoskeleton	4.73E-06	3.08	27	111	CC
GO:0003779	actin binding	7.52E-06	2.59	34	159	MF
GO:0006936	muscle contraction	1.93E-05	4.22	16	51	BP
GO:0044430	cytoskeletal part	2.23E-05	2.03	51	294	CC
GO:0031674	I band	2.27E-05	5.67	12	32	CC
GO:0003012	muscle system process	2.54E-05	4.11	16	52	BP
GO:0030029	actin filament-based process	2.89E-05	2.73	27	119	BP
GO:0007517	muscle development	5.06E-05	2.69	26	116	BP
GO Categories Enriched in Genes Associated with Myotube-Decreased Peaks						
GO:0044421	extracellular region part	4.59E-09	3.43	37	229	CC
GO:0005576	extracellular region	1.88E-08	2.54	56	457	CC
GO:0007167	enzyme linked receptor protein signaling pathway	7.88E-07	3.17	29	188	BP
GO:0005615	extracellular space	1.70E-06	3.83	21	116	CC

much computational analysis

What else is it doing?



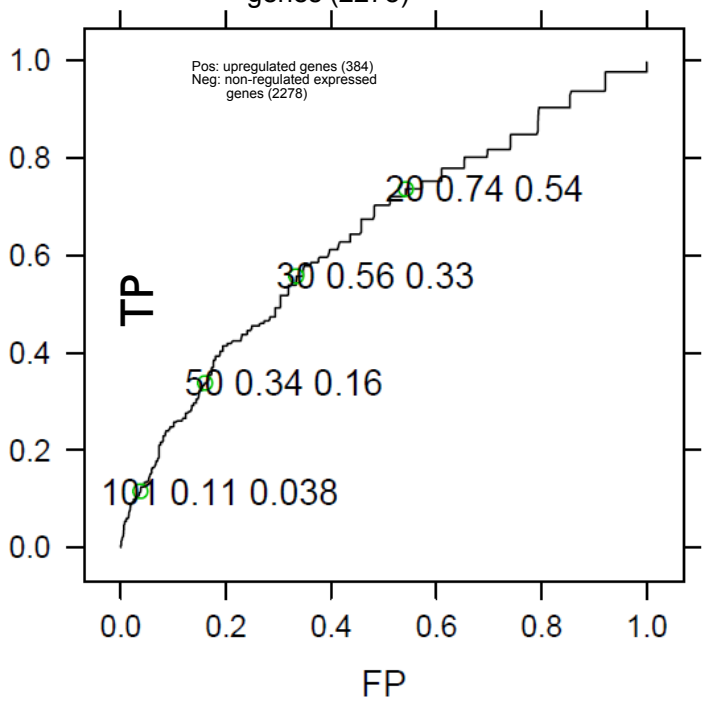




much computational analysis

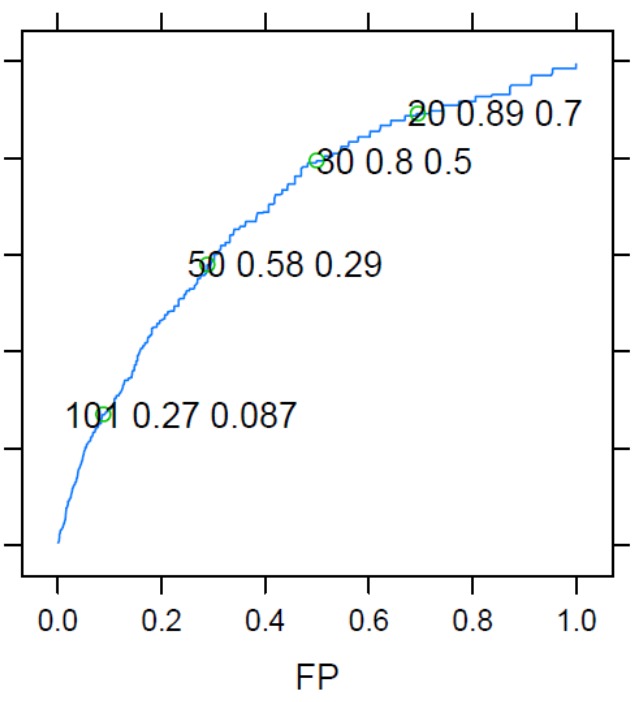
Promoter regions
AUC=0.64

Pos: upregulated genes (384)
Neg: non-regulated expressed genes (2278)



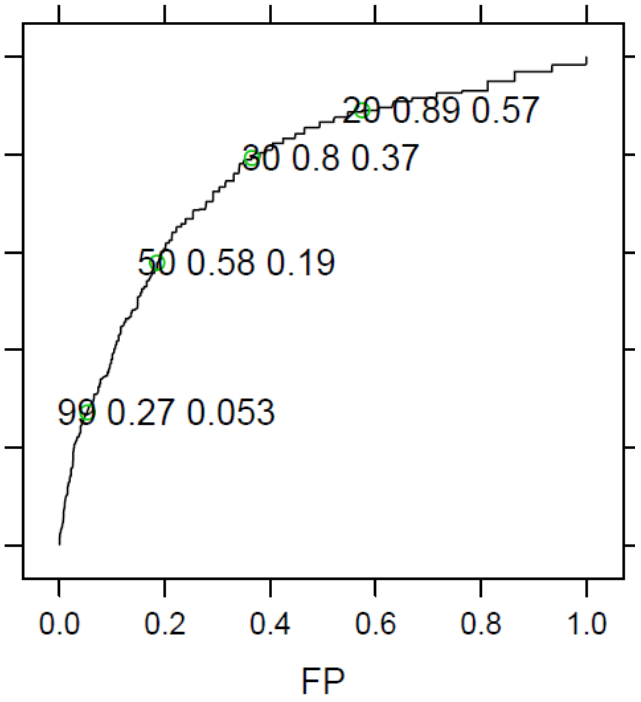
CTCF domains
AUC=0.70

Pos: upregulated genes (504)
Neg: non-regulated expressed genes (3789)



CTCF domains
AUC=0.77

Pos: upregulated genes (504)
Neg: intergenic (1294)



Binding Site/Cofactor Motifs (See paper)

TF	consensus	Myotube	Control	Myotube/Control
MyoD	CACCTGNY	1670	77	21.7
AP-4	CWCAGCTGG	1859	100	18.6
AP-2	MKCCCSCNGGCG	16	1	16.0
E47	VSNGCAGGTGKNC	2983	242	12.3
Sp1	GGGGCGGGGY	334	30	11.1
ITF-2	AACAGATGKT	789	71	11.1
FosB	TGACTCANNSK	567	52	10.9
Lmo2	CNNCAGGTGB	1004	95	10.6
USF2	CACGTG	254	25	10.2
TGIF	AGCTGTCANNA	531	57	9.3
NF-1/L	TGGNNNNNNGCCAA	482	55	8.8
Egr-1	GTGGGSGCRRS	244	28	8.7
c-Ets-1				

Discriminative motif discovery, on very large scale
E.g., 3 papers with related approaches appeared in *Bioinformatics* today

Summary

MyoD present (& bound) in both myoblasts & myotubes

Binds most genes, not just differentially expressed ones
Significant genome-wide binding

Although differentially bound peaks are associated with changed expression, peak height is a weak predictor of function

Implicated in broad chromatin modifications (histone H4 acetylation)

Motif discovery possible (but of limited predictive value in isolation)

Summary

MyoD present (& bound) in both myotubes

Binds most genes, not just ed ones
Significant genome-wide

Although different associated with changed expre weak predictor of function

Implica tin modifications (histone H4

Motif c sible (but of limited predictive value in isolation)

And math, stat, computational analysis is deeply interwoven with everything here...

Thanks for a fun quarter!