Analyzing ChIP-seq Data

Robert Gentleman + many others

FRED HUTCHINSON
CANCER RESEARCH CENTER
A LIFE OF SCIENCE

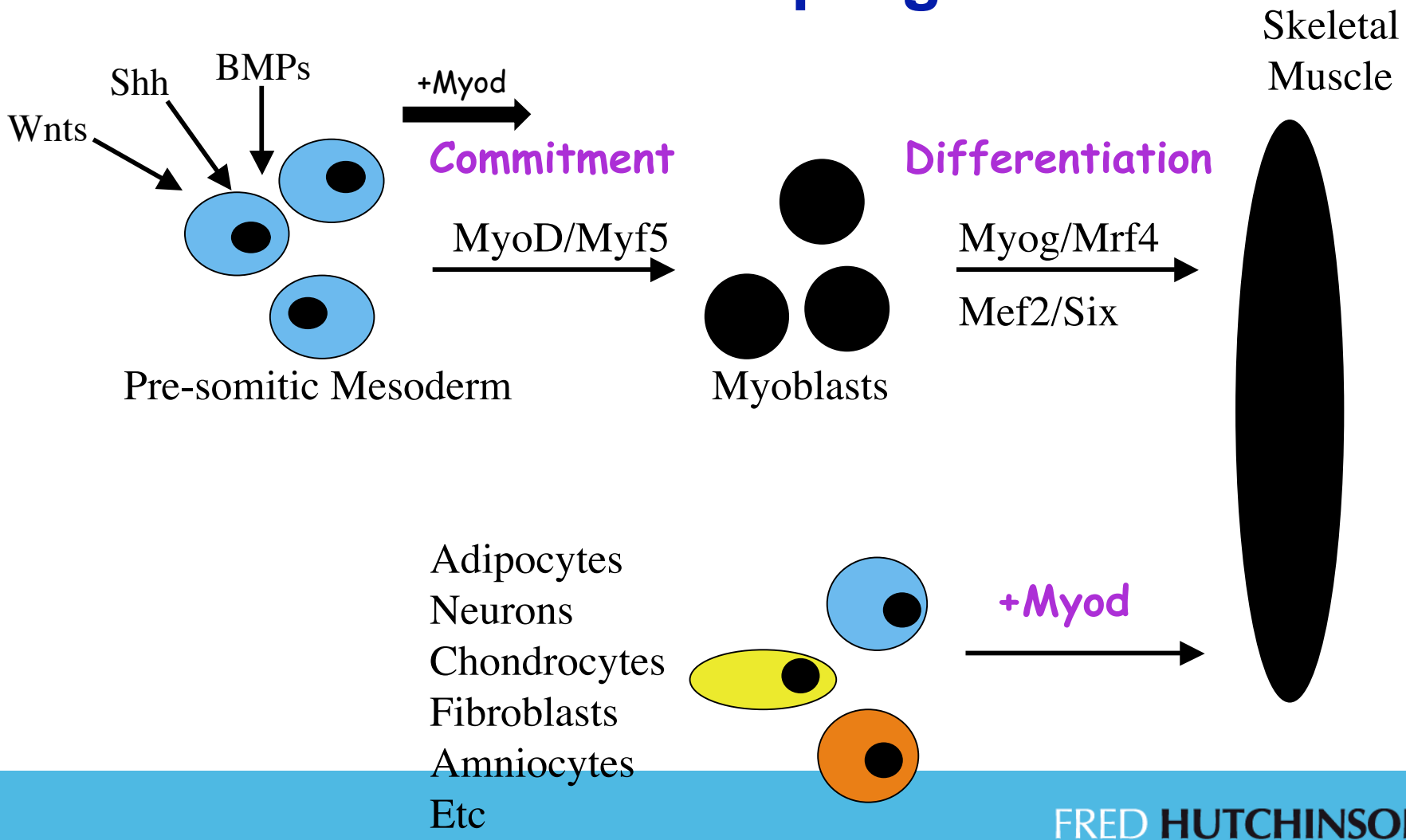
Outline

- discuss our experiment in some detail (this is more of a progress report)
- some results concerning the TF binding sites (eboxes)
- some of the many QA methods we are working on for short reads
- some of the data
- indications of what part of the pipeline can be handled by Bioconductor packages

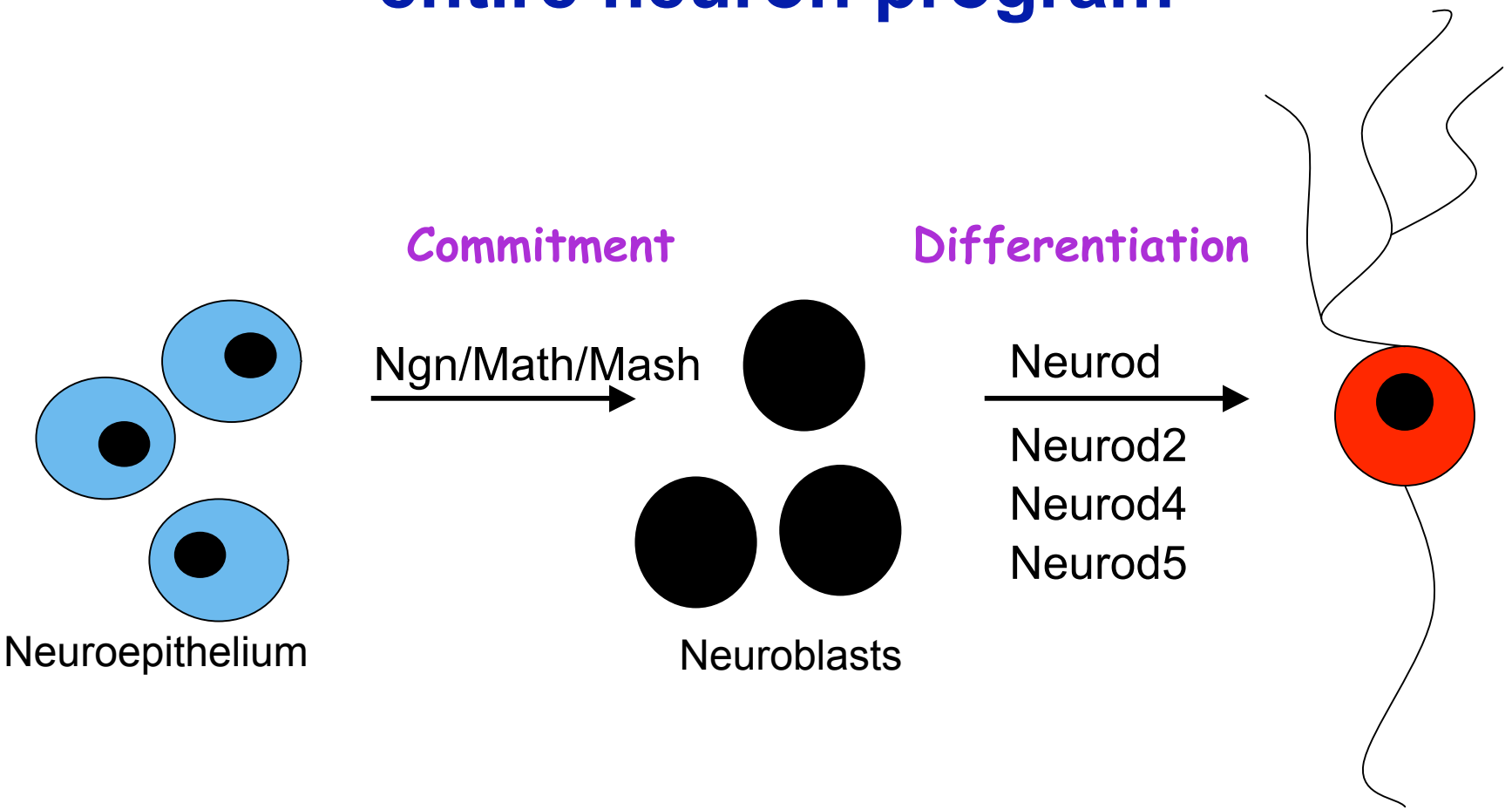
The clean experiment

- if you want to see how simple, and clean things can be, have a look at
  - Genome-Wide Mapping of in Vivo Protein-DNA Interactions, Johnson et al, Science, 2007, 316, p 1497-1502
  - they had a mono-clonal antibody and a consensus binding sequence that was 31nt long
- by contrast, we have polyclonal antibodies and a consensus sequence that is closer to 4nt long

Myogenic bHLH factors regulate the entire muscle program



Neurogenic bHLH factors regulate the entire neuron program

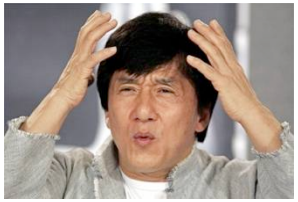


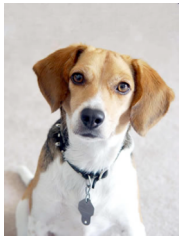
Myod & NeuroD2

- Belong to the same family of bHLH protein
- Both dimerize with E-protein
- Both bind to the same consensus sequence CANNTG (ebox)

Questions: How could these two factors maintain a common core program but modulate cell-type specific genes expression at the same time

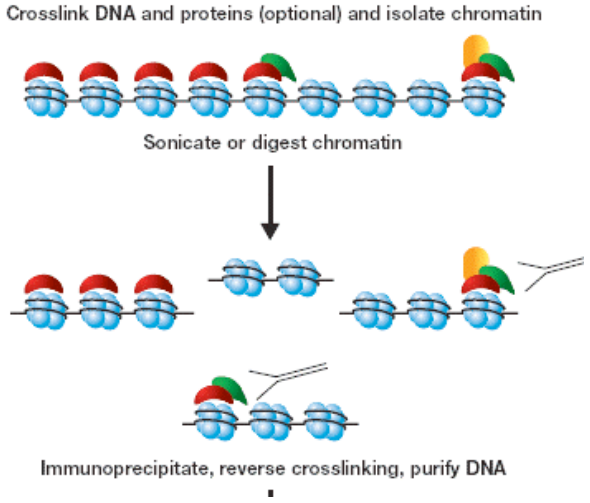
Experimental Design

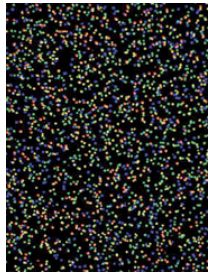




Isolate fibroblasts

+Myod





Solexa Sequencing

Chromatin Immunoprecipitation

Isolate ES cells

+NeuroD2

Data analysis

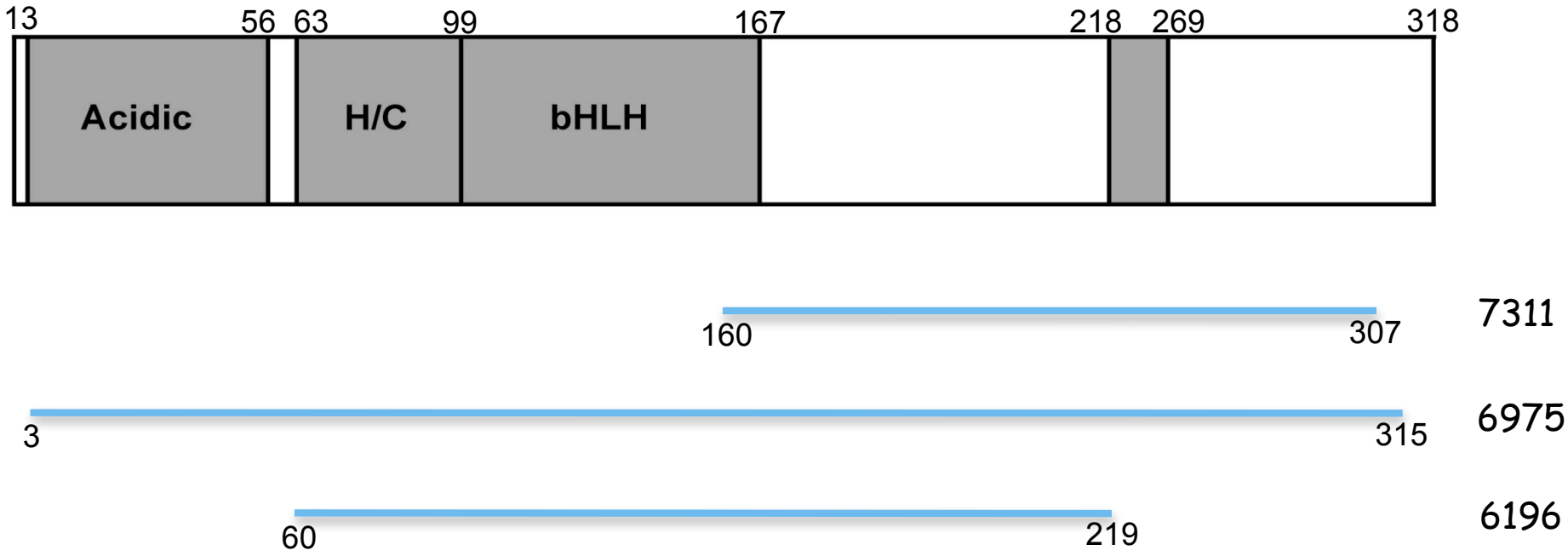
- Solexa:
  Myod binding sites in three species (three antibodies)
  Neurod2 binding sites
  conserved regions adjacent to the binding sites

- Compare to mRNA expression profiling

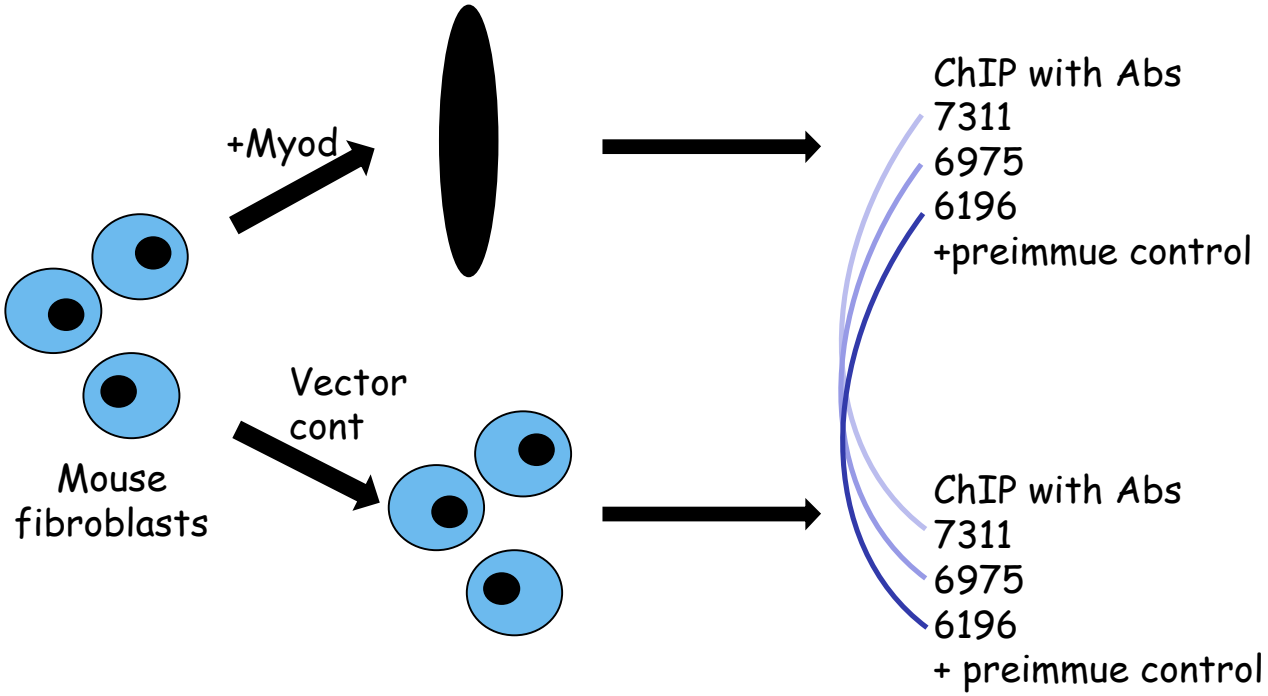
- Compare to microRNA expression profiling

- Modeling protein-DNA interactions

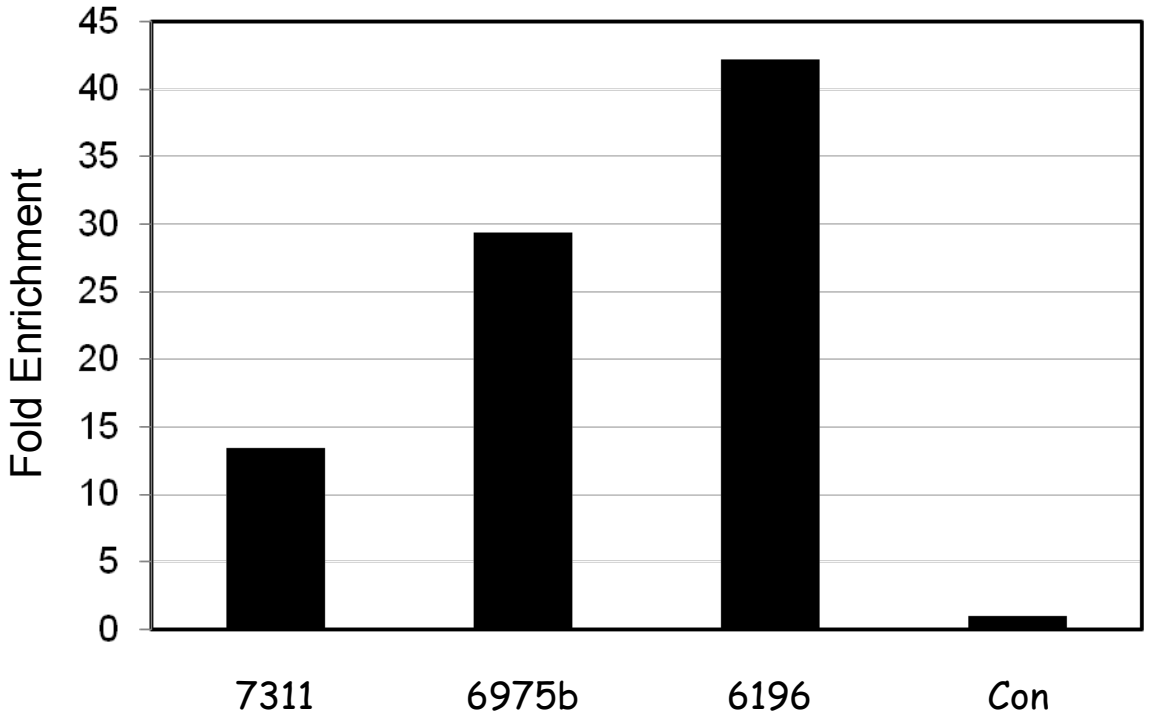
MyoD: The blue lines indicate regions used to raise antibodies



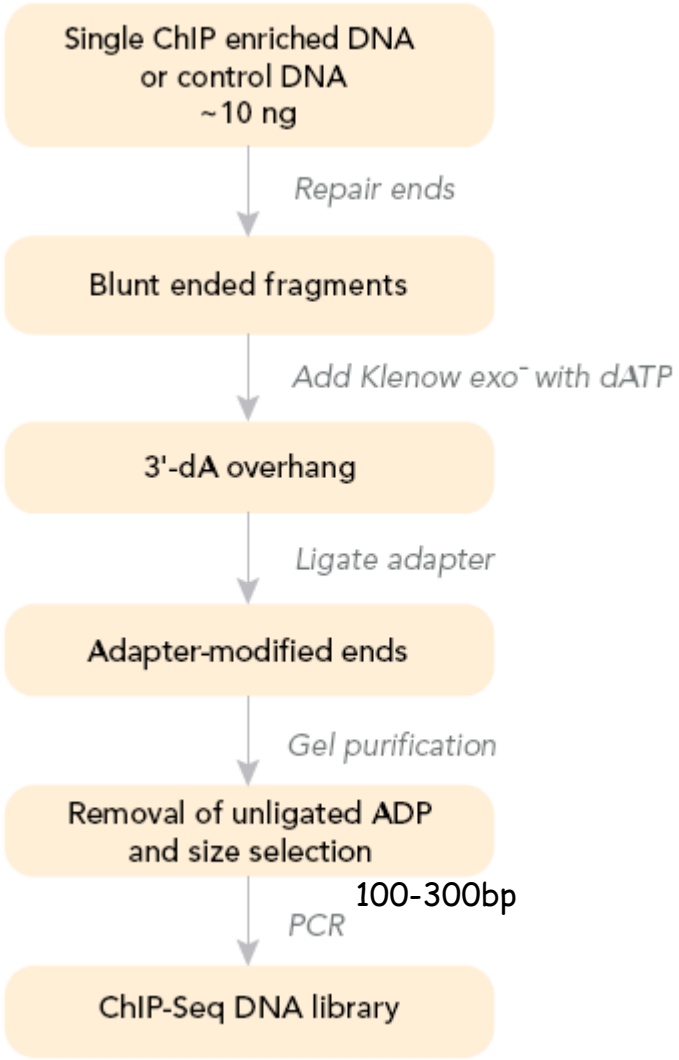
ChIP with MM cells



Enrichment on Myog promoter

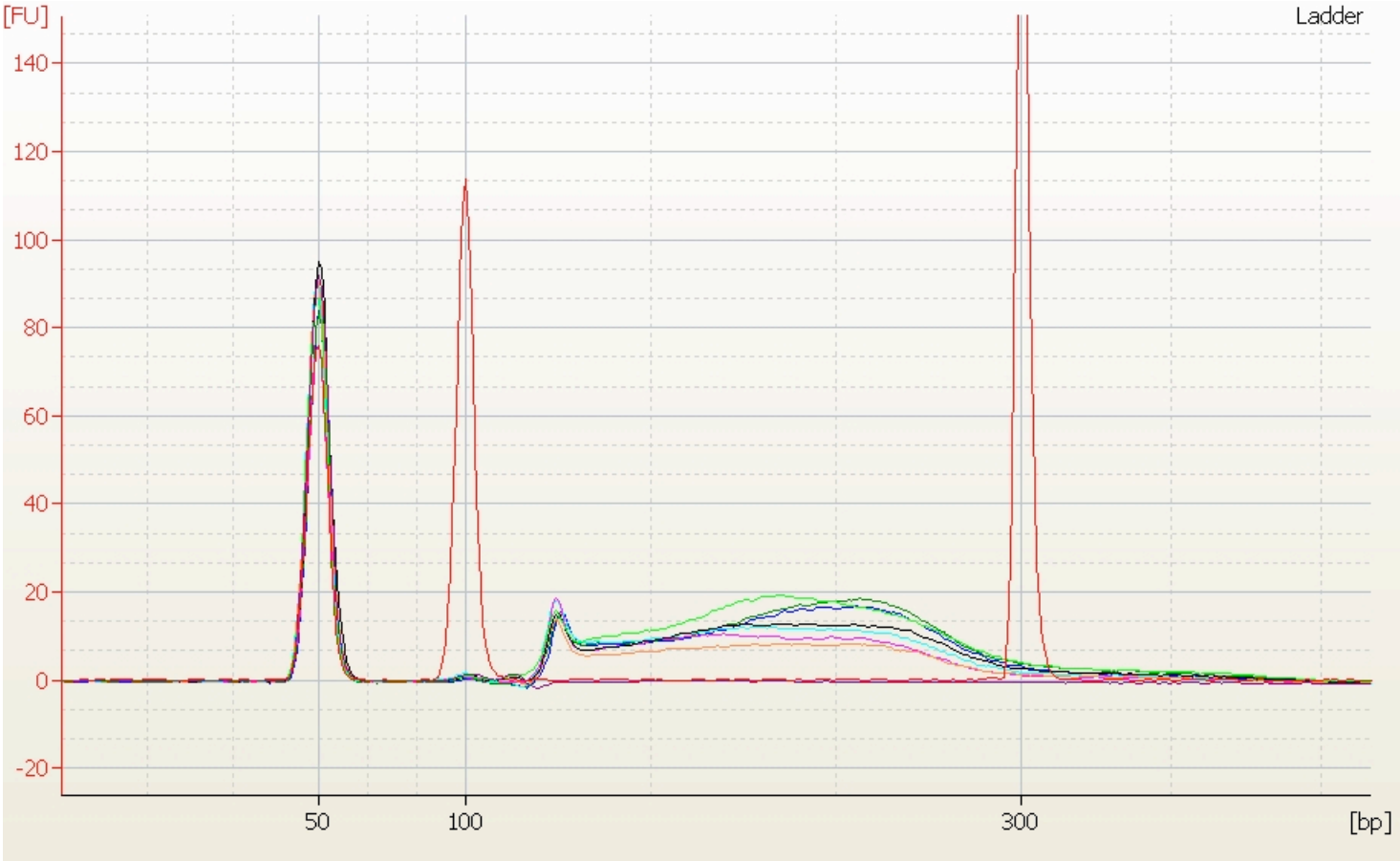


ChIP-seq Sample prep



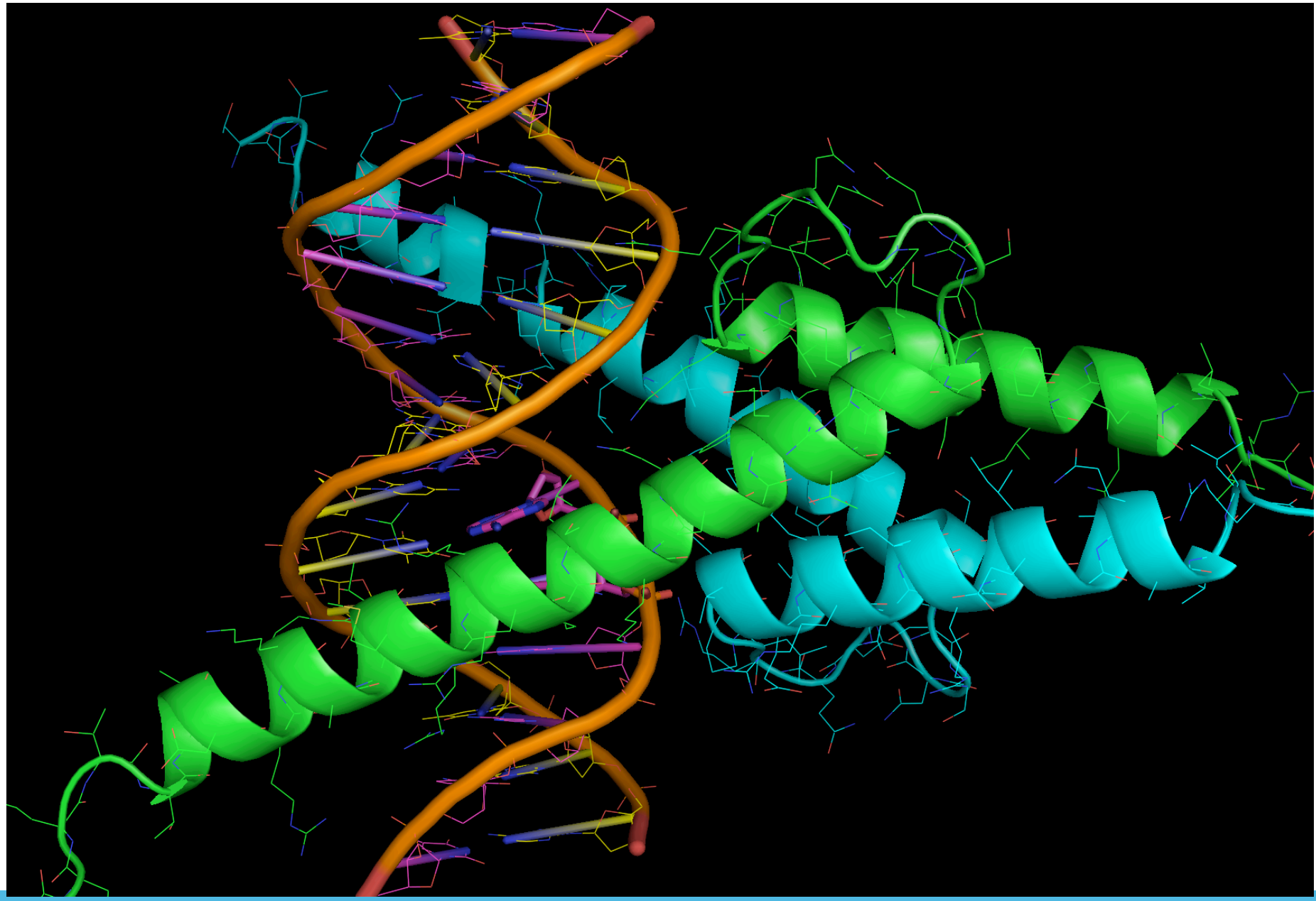
Load **2 picomoles** on the machine

Bioanalyzer analysis



bHLH Transcription Factors

- the basic-Helix-Loop-Helix family of transcription factors is known to form dimers (hetero and homo) that typically (but not always) bind eboxes
- The ebox sequence is CANNTG, which is quite common
  - 15.1 million (+ strand) Human
  - 14.2 million (+ strand) Mouse
  - 12.7 million (+ strand) Dog



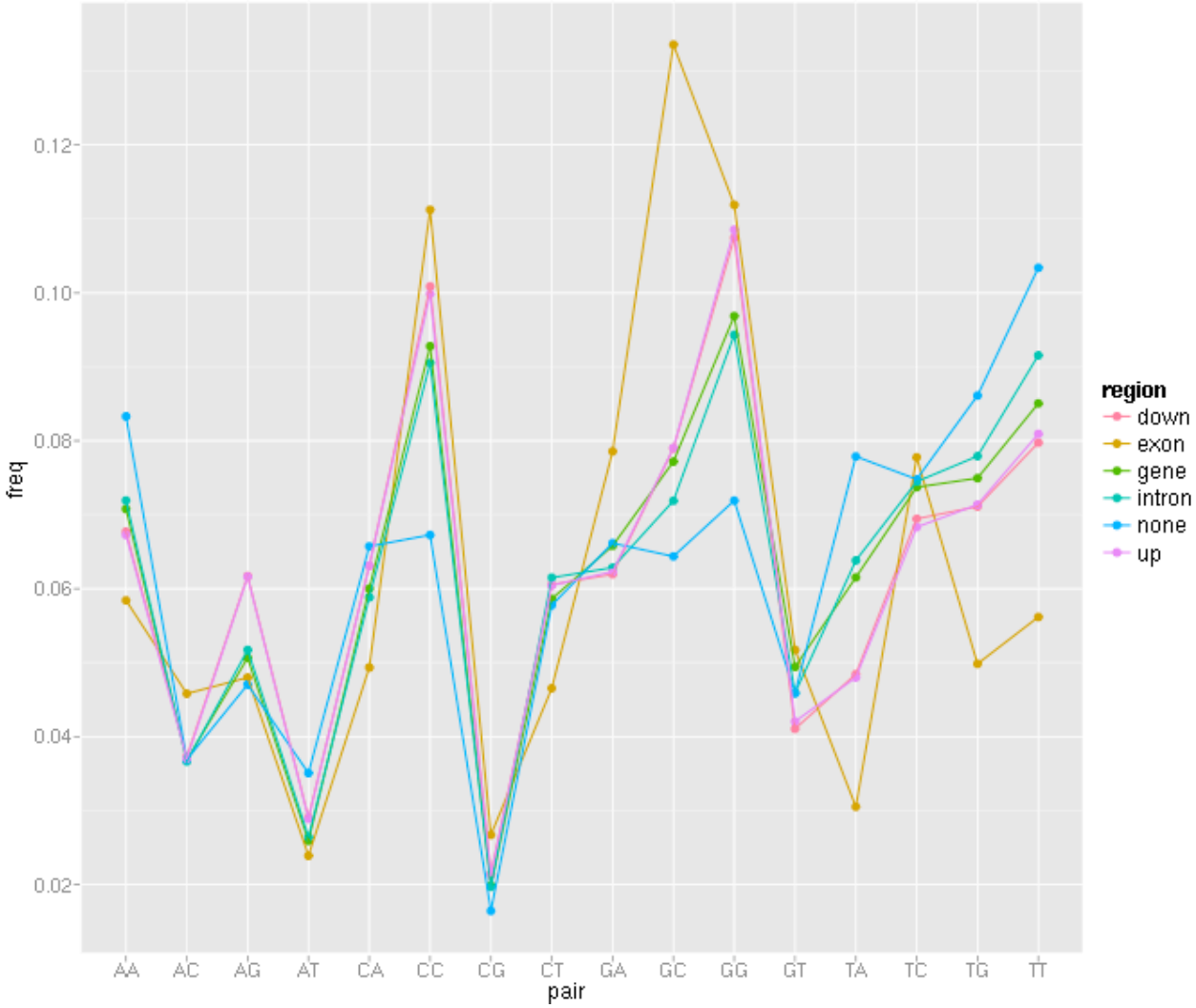
FRED HUTCHINSON
CANCER RESEARCH CENTER

A LIFE OF SCIENCE

EBOXES

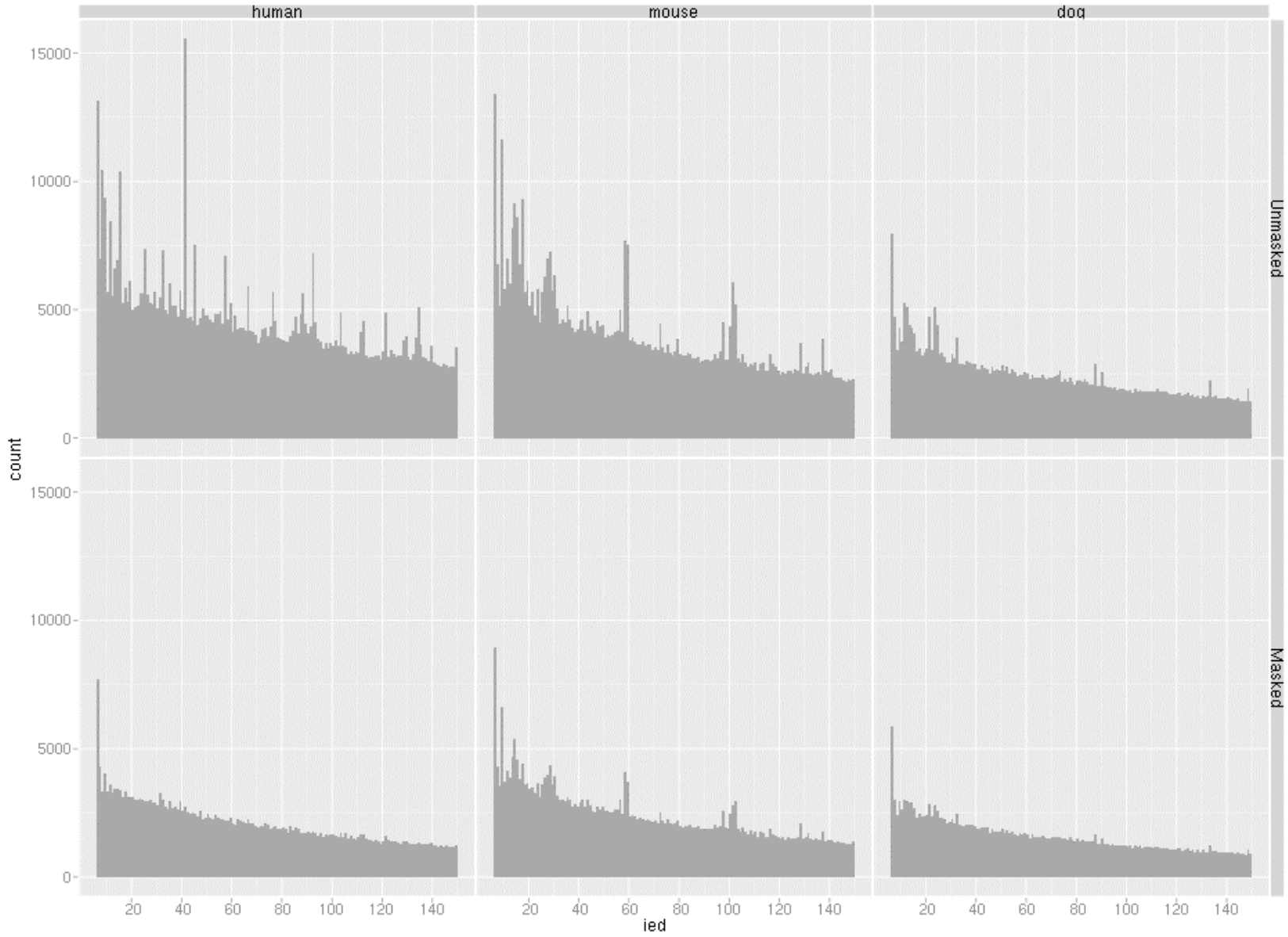
- there are 16 variants, some are reverse-complement palindromes
- it is of some interest to develop algorithms that can characterize the behaviors, ideally identifying EBOXES that are likely to be used by specific transcription factors etc.
- we divided the (mouse) genome into regions: upstream, downstream, intron, exon, none and counted frequencies of the NN nucleotides

Di-nucleotide Frequencies for eboxes in the repeat masked genome



EBOXES

- we computed the distance between sequential pairs of eboxes, separately for each chromosome (and each organism)
- the distances show some interesting characteristics (typically different ones for different species) that indicate that some distances are preferred.
- but most of it disappears when we use a repeat masked genome instead



EBOXES

- next steps:
  - look at eboxes that are conserved
  - eboxes by region (as done for the nucleotide frequency)
  - eboxes that are occupied in our ChIP-seq experiments

The Data

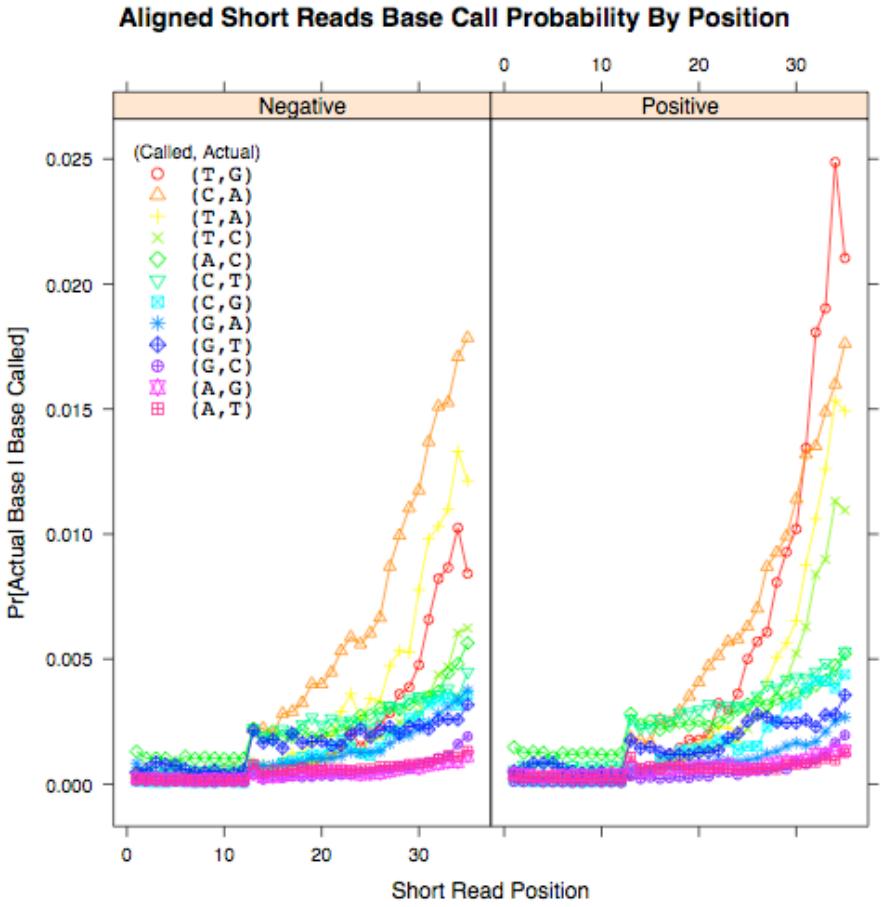
- we used Solexa to do the sequencing:
  - 8 lanes of data, one used for QA (a phage genome is sequenced)
  - expt1: 1 lane for each antibody, for both MyoD stimulated fibroblasts and unstimilated
  - something in the neighborhood of 4 million reads/lane (fewer if we use unique reads)
- one thing many others have done, is to not use unique reads, but to restrict to reads that map once to the genome
  - that seems sort of backwards to me

Data Quality

- we have been working on a number of tools to help assess quality
  - four different tutorials you can attend
    - Martin Morgan (ShortRead)
    - Patrick Aboyoun (Alignments)
    - Herve Pages (Biostrings - matching)
    - Michael Lawrence (rtracklayer - genome browser)
- most (but not all) of what we are doing is in the current development versions of these packages
- it does seem prudent to try aligning a few tens of thousands of unmatched reads

Data Quality

- these are conditional probabilities
- they suggest that the matching algorithms could be improved by accounting for base and position



Lane 1 Expt 1 Mouse

Unique reads: 2745164, unique reads close to linker: 10778

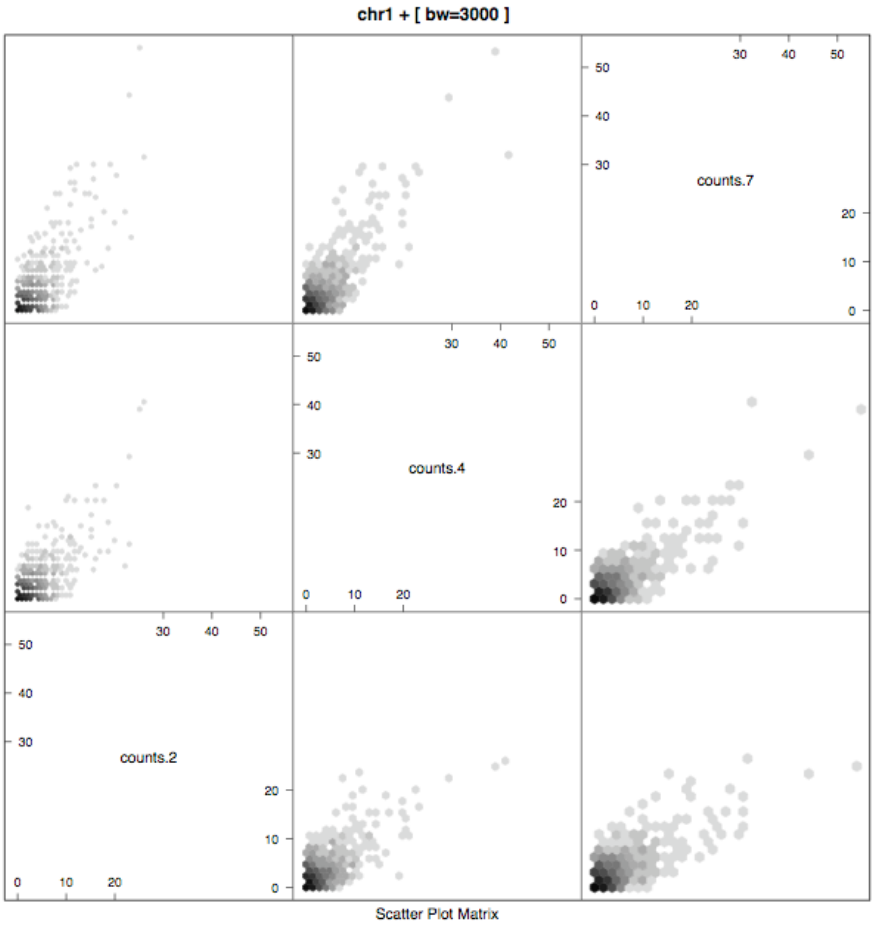
	NM=0	NM<=1	NM<=2
NH=0	1896925 (69.1%)	1713351 (62.4%)	1641265 (59.8%)
NH=1	702248 (25.6%)	801725 (29.2%)	801453 (29.2%)
2 <= NH <= 10	55890 (2.0%)	74783 (2.7%)	91811 (3.3%)
11 <= NH <= 100	41523 (1.5%)	56077 (2.0%)	65946 (2.4%)
101 <= NH <= 1000	25732 (0.9%)	43303 (1.6%)	59836 (2.2%)
1001 <= NH <= 10000	17162 (0.6%)	33321 (1.2%)	41913 (1.5%)
10001 <= NH	5684 (0.2%)	22604 (0.8%)	42940 (1.6%)

Data

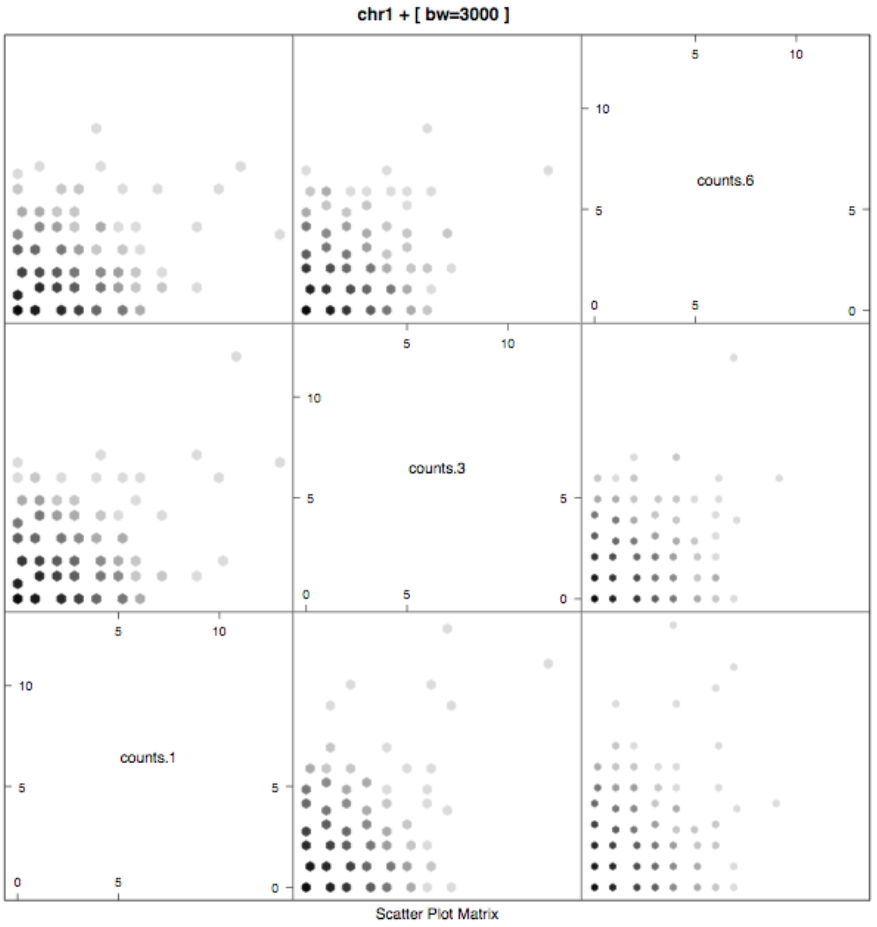
- Some obvious questions
  - why do so few match (MAQ gets essentially the same number)?
  - I had hoped that by taking those that match to two or more places we would gain a lot (we don't seem to)
  - those that match a lot, are quite common and will slow down any matching algorithm
  - repeat masking helps (but could be having other effects)

Is there signal?

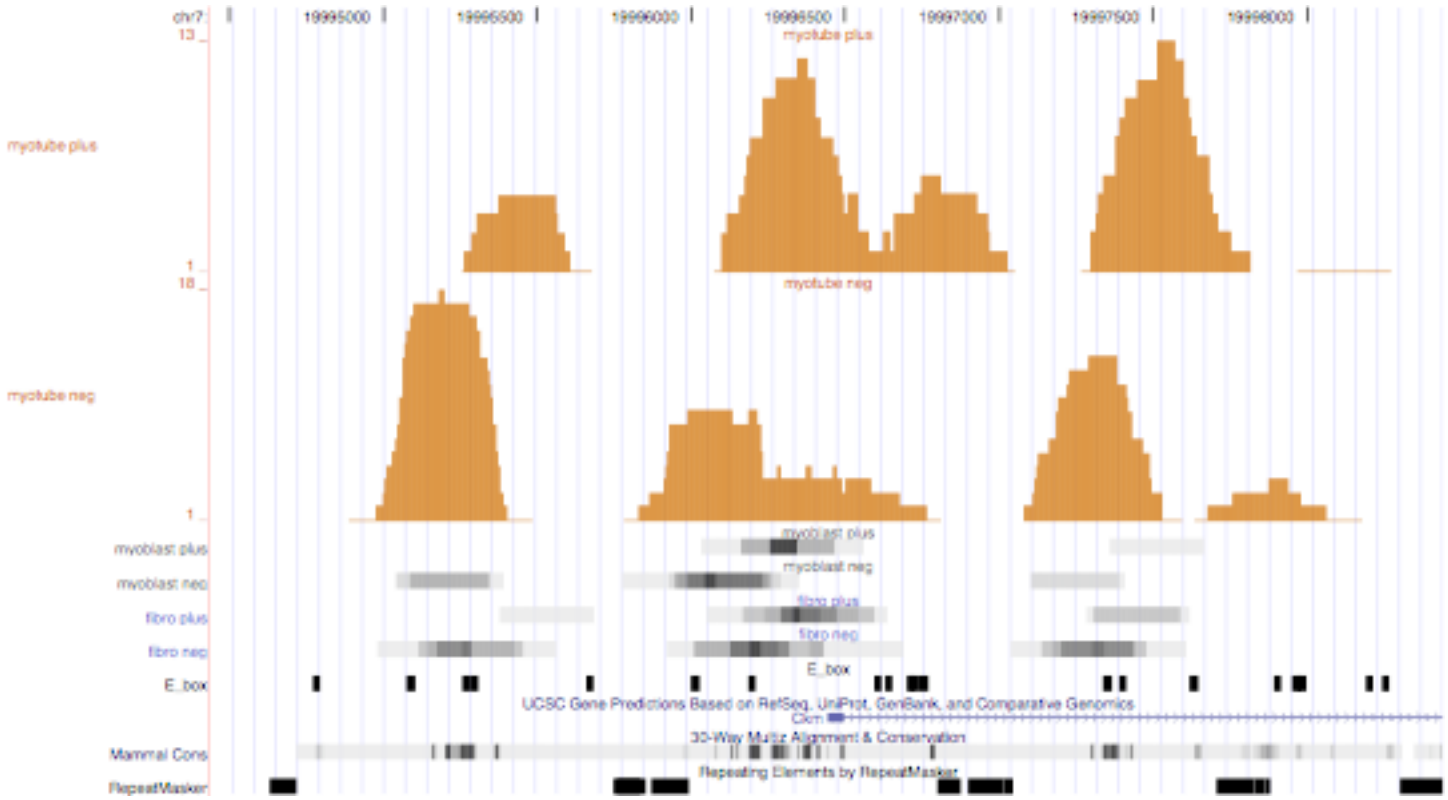
MyoD expressed



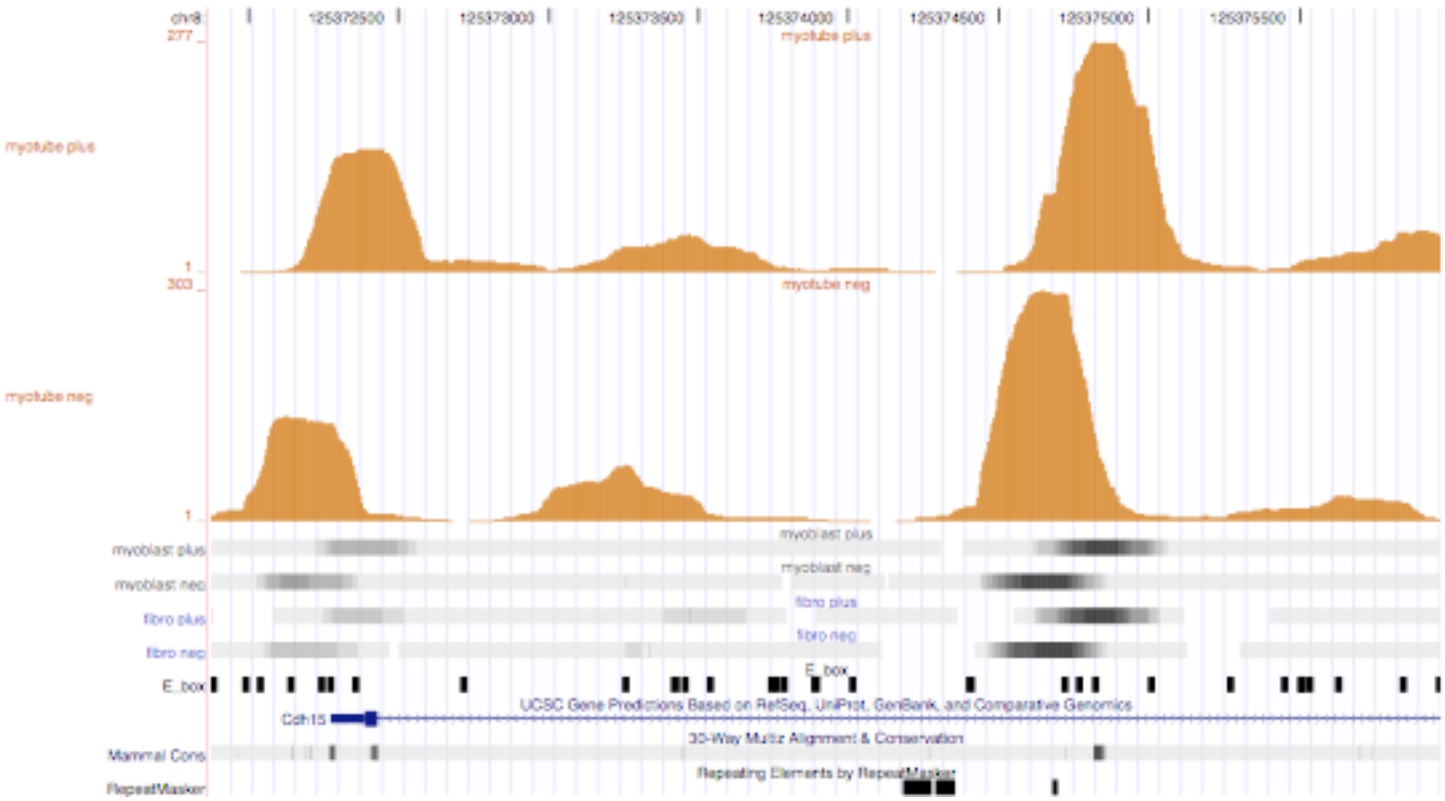
Control Lanes



Binding to CKM



Cdh15


chr8
125372500
125373000
125373500
125374000
125374500
125375000
125375500
277
1
303
1
myotube plus
myotube neg
myoblast plus
myoblast neg
fibro plus
fibro neg
E_box
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
Cdh15
30-Way Multiz Alignment & Conservation
Mammal Cons
Repeating Elements by RepeatMasker
RepeatMasker

Things we can do in BioC

- matching to genome
- alignment to genome
- finding TF binding sites
- nucleotide frequencies
- depth of coverage
- peak finding
- read and write Genome browser tracks
- working on relating two sets of intervals

Contributions

- Yi Cao
- Stephen Tapscott
- Phil Bradley
- Deepayan Sarkar
- Herve Pages
- Patrick Aboyoun
- Zizhen Yao
- Larry Ruzzo
- Michael Lawrence
- Marc Carlson
- Martin Morgan