# Linking changes in DNA methylation and chromatin structure during cell differentiation

Sergei Lebedev

Advisors: Oleg Shpynov, Roman Chernyatchik

JetBrains

December 13, 2012

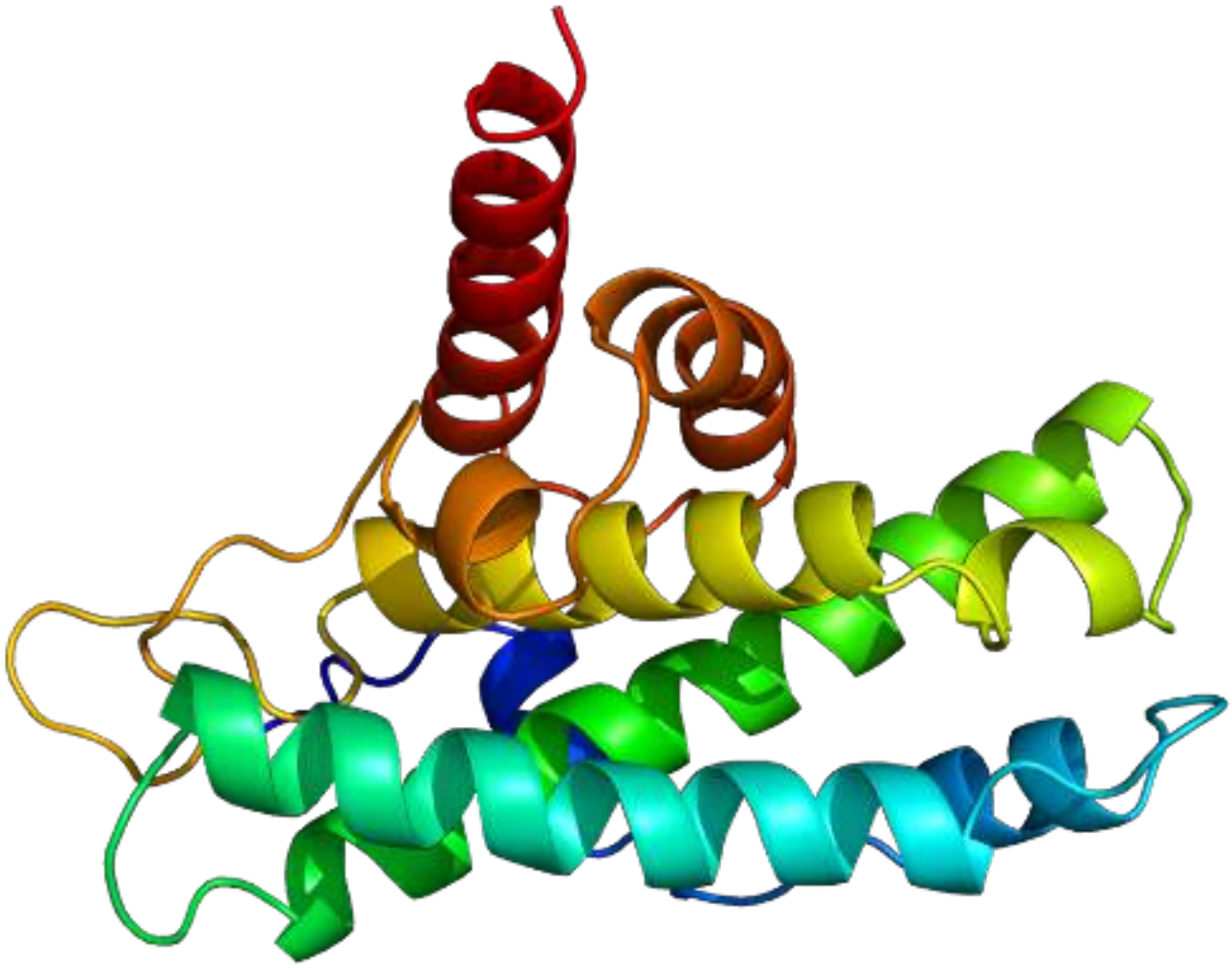


Figure : **R**etino**bl**astoma protein, RB1

# ~~Linking changes in DNA methylation and chromatin structure during cell differentiation~~ ChIP-seq motif finding *or* learn bioinformatics the hard way!

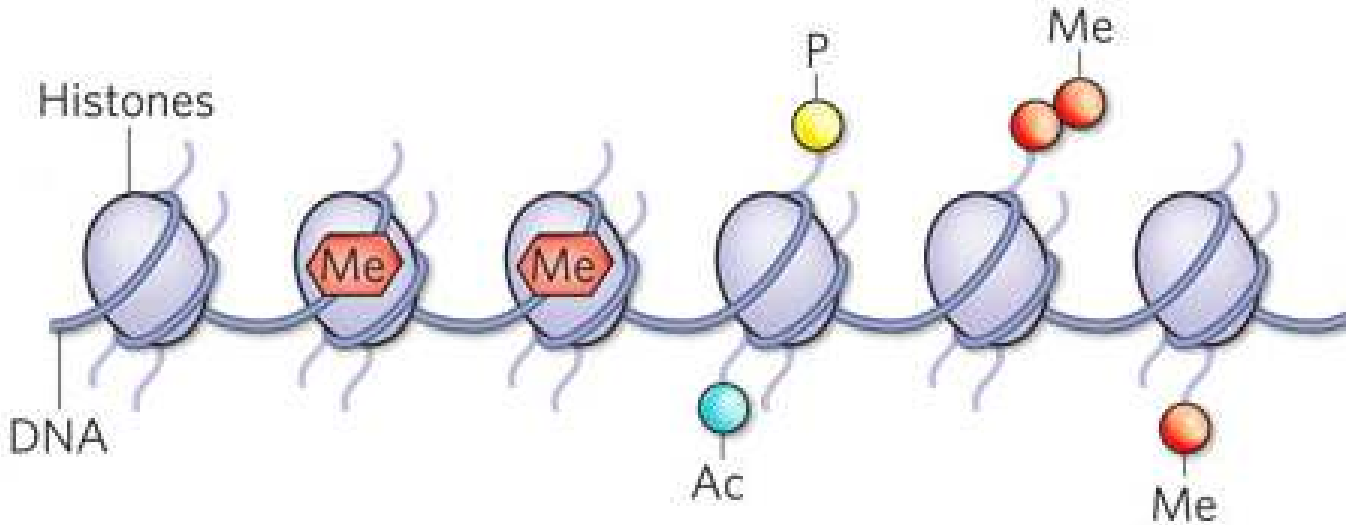
Sergei Lebedev

Advisors: Oleg Shpynov, Roman Chernyatchik

JetBrains

December 13, 2012

# Motivation or Epigenetics in a single slide



- Given binned ChIP-Seq and BS-Seq reads for multiple variously differentiated cell lines, find genomic regions, similarly and differentially enriched between cell lines.

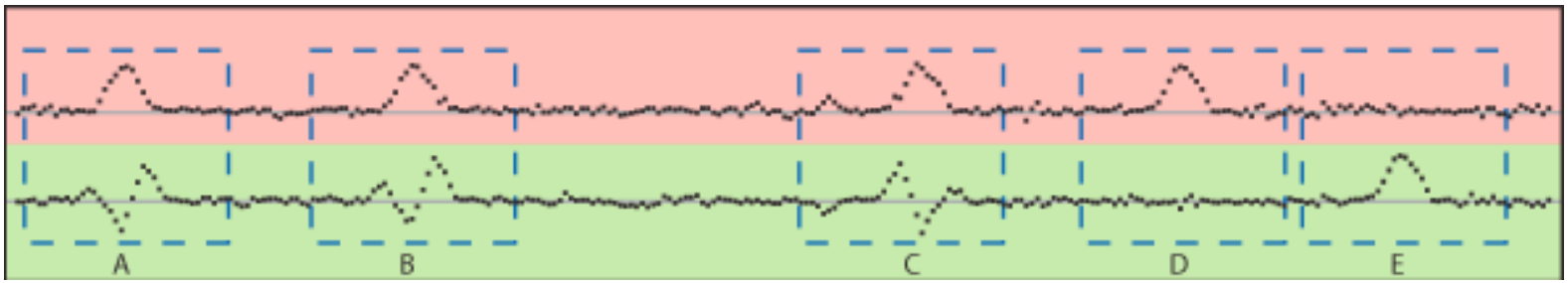



Figure : Schematic overview of the problem[1]

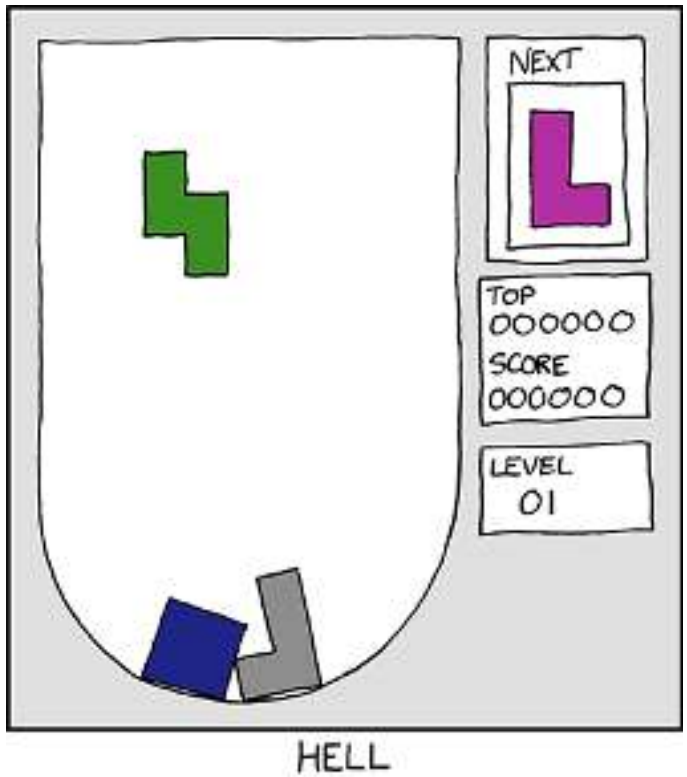
- Already solved for ChIP-Seq and BS-Seq data separately, ex: edgeR [2], MACS [3], ChromaSig [1].
- But looking for differential enrichment in histone modification **and** DNA methylation simultaneously is still challenging.

[1] doi:10.1371/journal.pcbi.1000201.g001

# The Plan

- Pick a tool, which solves part of the problem; we've chosen ChromaSig [1] – a motif finder for ChIP-Seq data.
- See if we can change it to operate on multiple cell lines. Turns out we **can**, but it's harder then we thought initially.
- See if it yields any interesting results, when we add tracks with BS-Seq data.
- A lot to experiment with, the data is very noisy, try existing filtration and normalization schemes or come up with a new one.

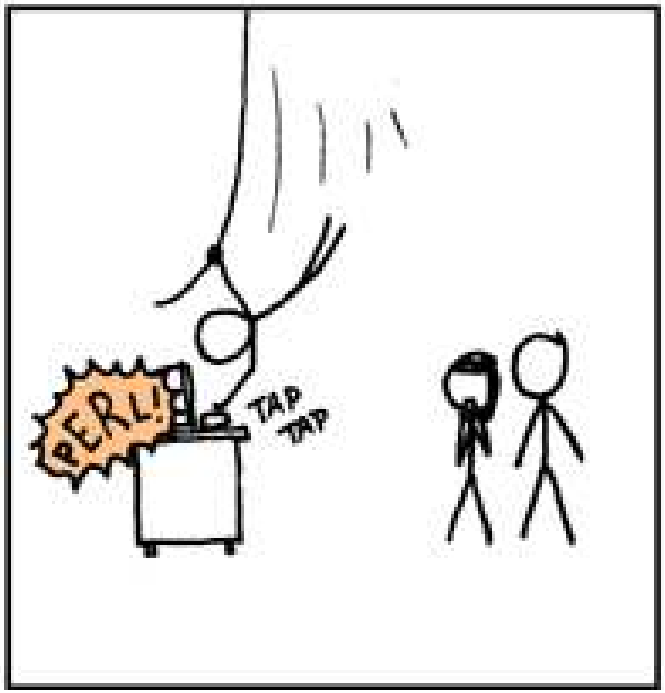
# Learn the hard way: Academia[2]




## Rule #2

If you think of re-implementing an algorithm from a paper – think **again**.

[2] http://xkcd.com/724




## Rule #17

If you think of re-implementing an algorithm in Perl – see Rule #2.

[3] http://xkcd.com/208

```
for(my $index = 0 ; $index < scalar(@sorted) ; $index++) {
    my $i = $sorted[$index];
    my $is_local_max = 1;
    my @identical = ();
    for (my $j = $i - $overlap_half_window_size ;
         $j <= $i + $overlap_half_window_size ;
         $j++) {

        if (($j == $i) || (not defined $sig_locs->{$j})) {
            next;
        }
        # check for identity
        if ($sig_locs->{$j}->{sum} == $sig_locs->{$i}->{sum}) {
            push(@identical, $j);
        }
        if ($sig_locs->{$j}->{sum} > $sig_locs->{$i}->{sum}) {
            $is_local_max = 0;
            last;
        }
    }
}
```

## Rule #8

If you think that algorithm implementation **matches** the description in the paper — see Rule #2.

- You can only ~~pray~~**hope** that the implementation is *slightly* similar.
- And that you have enough time and patience to get through all the glory detail details, encrypted in the academia-style code.

- ChromaSig re-implemented from scratch in Java, lots of fun!
- An attempt to evaluate the implementation on real ChIP-seq data failed; we search for motifs *de-novo* – how to do the evaluation?
- Simulations to the rescue!

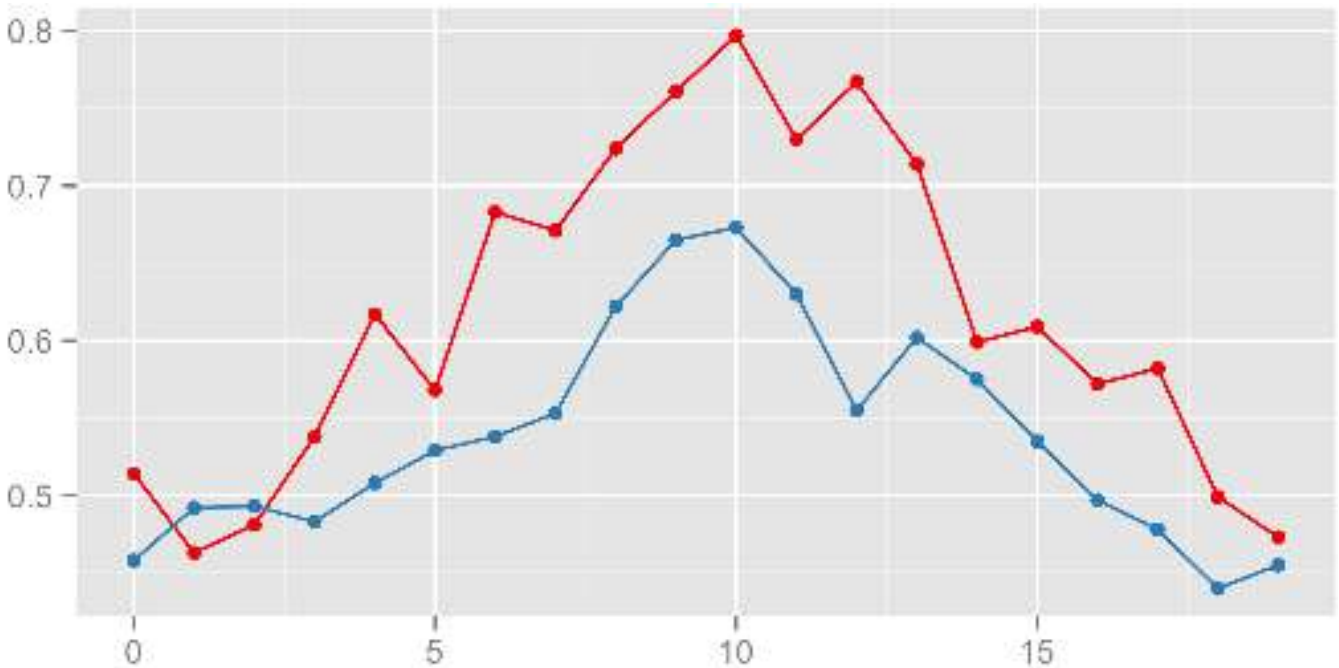



Figure : Simulated ChIP-seq enrichment patterns

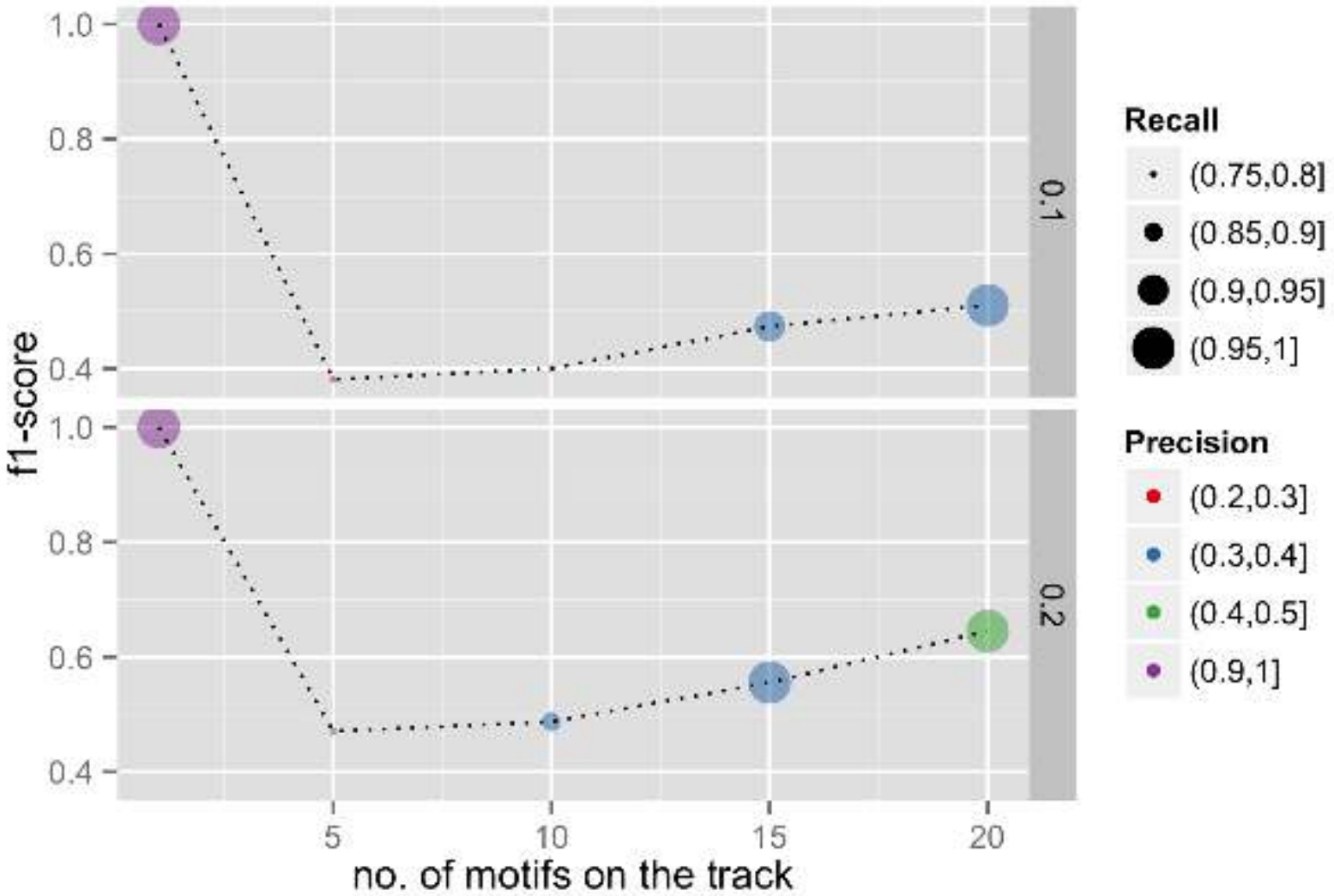



Figure : Median F1-score for different distance cutoff values

Still lots of *rules* to be learned.

Thank you!


G. Hon, B. Ren, and W. Wang.
ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome.
*PLoS Comput. Biol.*, 4(10):e1000201, Oct 2008.


M. D. Robinson, D. J. McCarthy, and G. K. Smyth.
edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.
*Bioinformatics*, 26(1):139–140, Jan 2010.


Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu.
Model-based analysis of ChIP-Seq (MACS).
*Genome Biol.*, 9(9):R137, 2008.