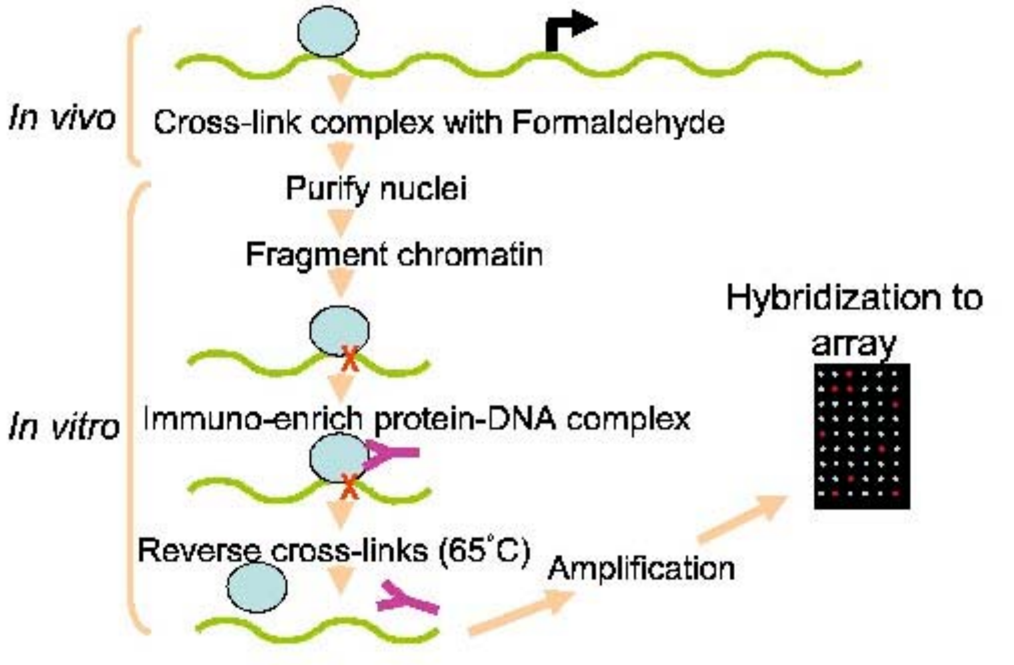
State Research Center of Genetics and Selection of Industrial Microorganisms,
GosNIIGenetika, Moscow, Russia

Some questions of interpretation of results for DNA-protein binding on tiling arrays

October 9, 2008

3rd workshop on algorithms in Molecular Biology, Moscow, 2008

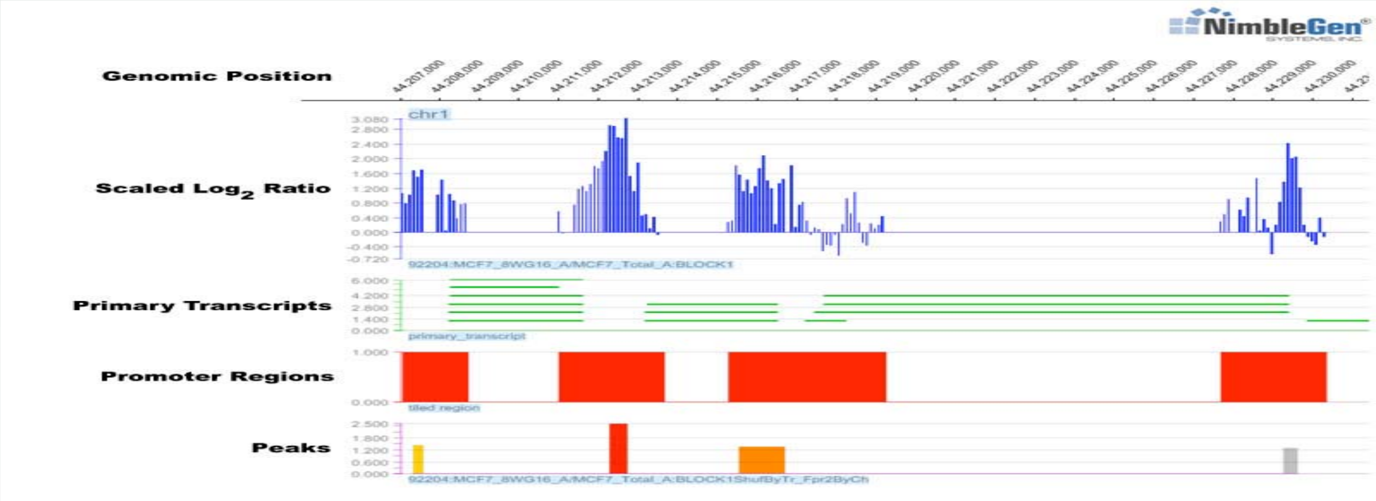
ChIP-chip technology


In vivo
Cross-link complex with Formaldehyde
Purify nuclei
Fragment chromatin
In vitro
Immuno-enrich protein-DNA complex
Reverse cross-links (65˚C)
Amplification
Hybridization to array

Isolation and immunoprecipitation of raw chromatin bound with transcription factors. Non-immunoprecipitated chromatin is also hybridized to the array and signals given by IP and non-IP samples are compared.

From: http://www.tigr.org/

Genome-wide location analysis at tiling arrays



From: http://www.nimblegen.com/

NimbleGen	385,000 50- to 75-mer	RNA polymerase Nature 436: 876-880 (2005)
Affymetrix	$6 *10^6$ 25mer	Estrogen receptor Nat Genet 38: 1289-1297 (2006)
Agilent	244,000 60-mer	Polycomb Cell 125: 301-313 (2006)

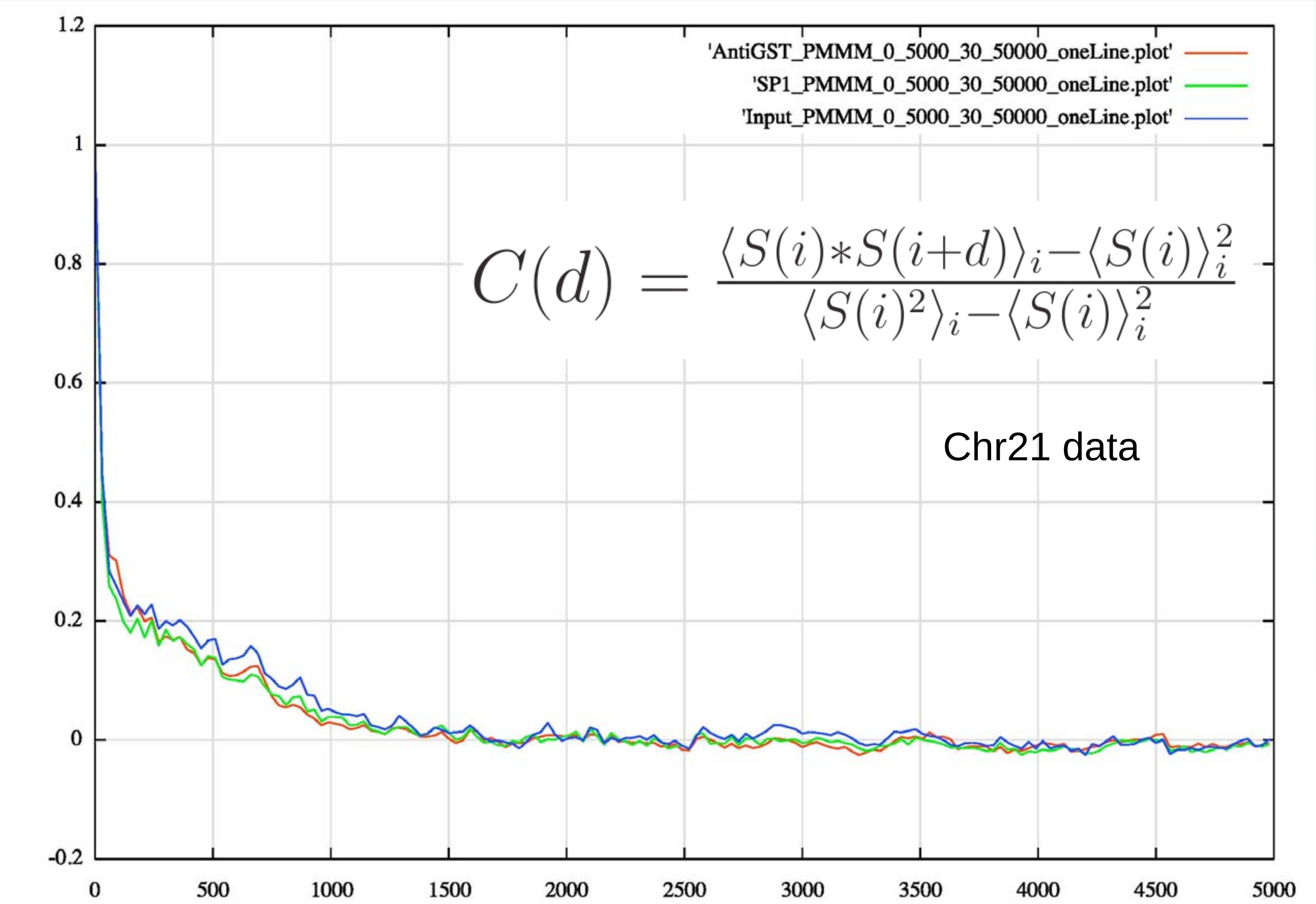
Problem of data quality

- Mishybridization with mismatches -> “genome-wide”
- Hybridization signal depends on the CG content of a probe…

  … and of the test DNA fragment
- Length distribution of DNA fragments after sonication

Correlation in binding to probes neighboring in the genome

C(d)



$$C(d) = \frac{\langle S(i) * S(i+d) \rangle_i - \langle S(i) \rangle_i^2}{\langle S(i)^2 \rangle_i - \langle S(i) \rangle_i^2}$$

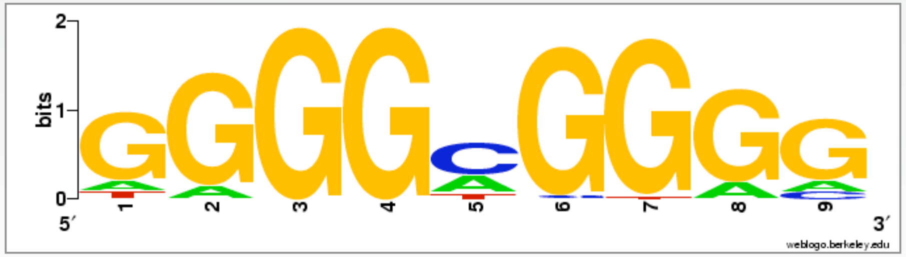
Chr21 data

Distance, b.p

d

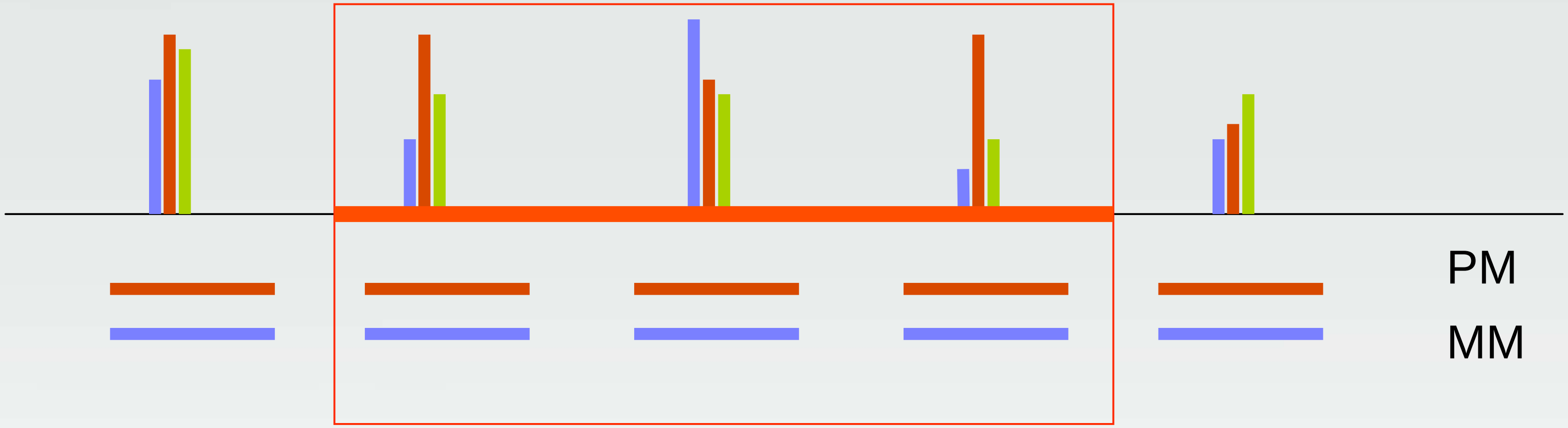
Comparison with bioinformatics

- Sp1 ChIP at Affimetrix
  - human chromosomes 21, 22; 25+5 chip, PM, MM, probes, with two control hybridizations (input DNA and anti-GST)
- TRANSFAC contains many Sp1 binding sites



- Compare ChIP-chip with bioinformatics Sp1 transcription factor binding site predictions

Regions predicted by ChIP-chip



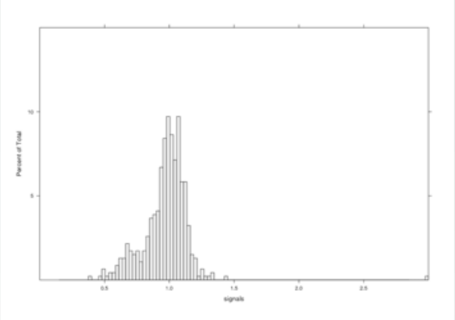
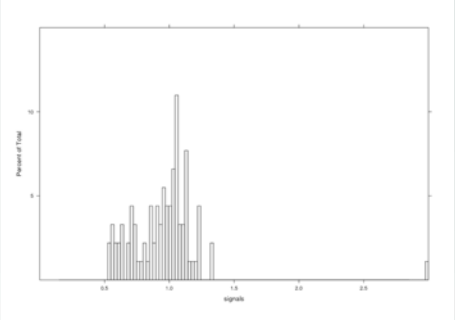
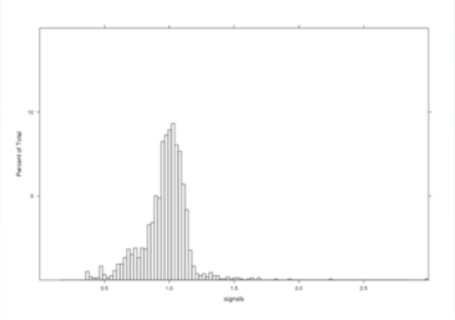
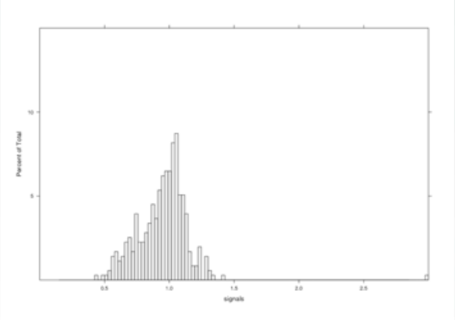
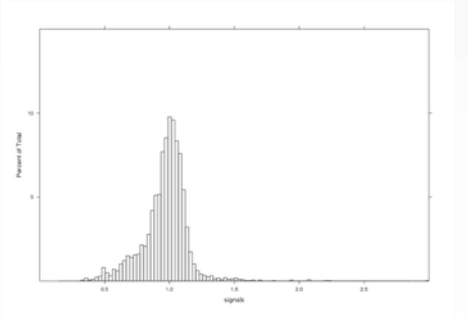
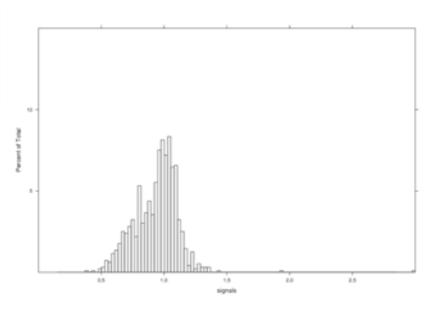
MM – mismatch probe – mishybridisation from other DNA segments

Input – DNA without antibody extraction step

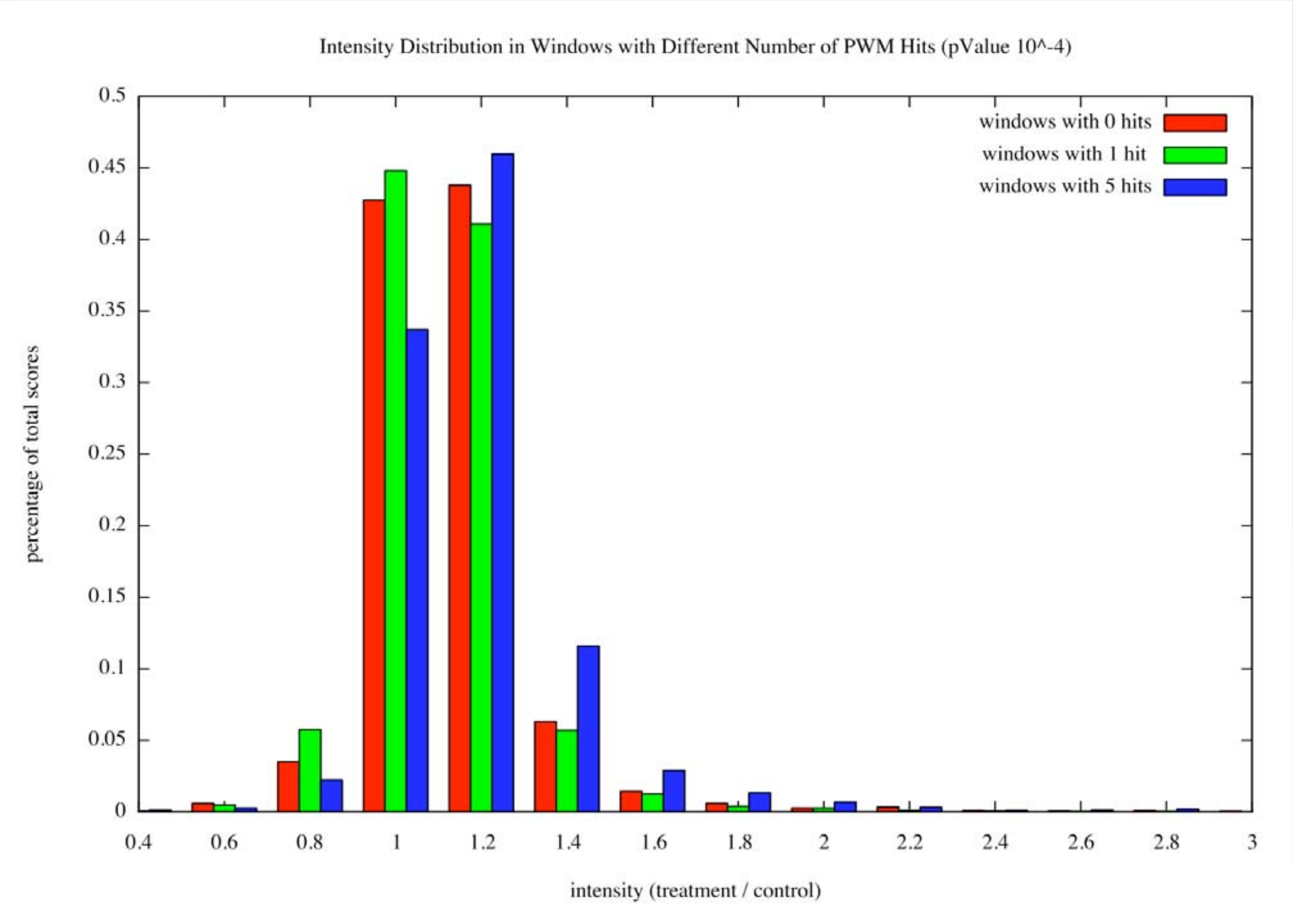
Window – with statistically prevalent PM – usually ~ 1000 bp

Experiments with isolated Sp1 computational hits

Probes Number
Histograms
S/N ChIP

Window	1200 bp. no hits	1200 bp isolated hits
50 bp		
200 bp		
500 bp		

ChIP-chip signal indicate not individual sites but site clusters!



Distribution of intensities in 500 bp window is almost identical for no-PWM-hits, and one-PWM-hit windows, but it is visibly shifted to the left for 5-PWM-hits window.

Conclusions I

- ChIP-chip is a weak filter, concentrating binding regions (up to 30 folds by our evaluation)
- The noise of ChIP-chip is very high
- If one takes 1000 bp windows only about 5% of high-scoring computational Sp1 sites in chromosomes 21 and 22 is covered
    - (Cawley etc. Cell, 2004)
- 50% of ChIP-chip binding regions published by Affimetrix do not contain any signal recognizable with bioinformatics
- Regions identified as ChIP-chip are more likely not individual binding sites but clusters of binding sites.

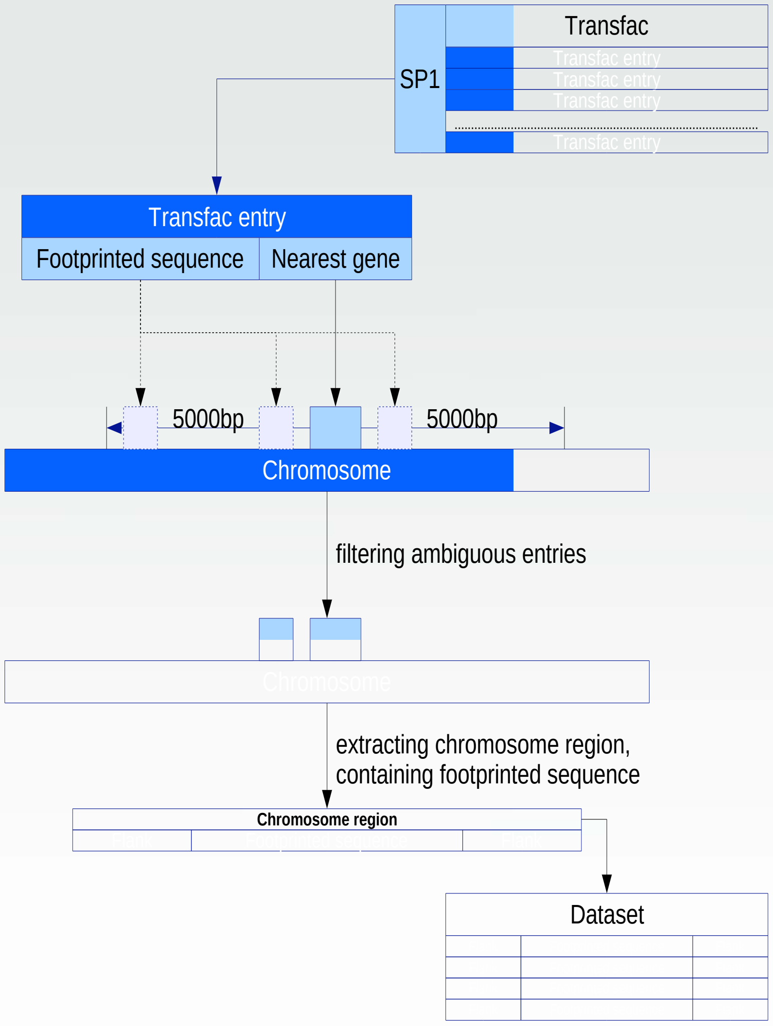
Testground: identification of Sp1 binding motif

Key points: ChIP-chip regions are long – and contain binding sites for many different proteins -> direct identification by bioinformatics is impossible

SELEX – give some idea of binding motif, usually distorted. But it is shows binding to the test protein

Footprint – also can contain mistakes, but can be used as a control, being independent from ChIP-chip and SELEX

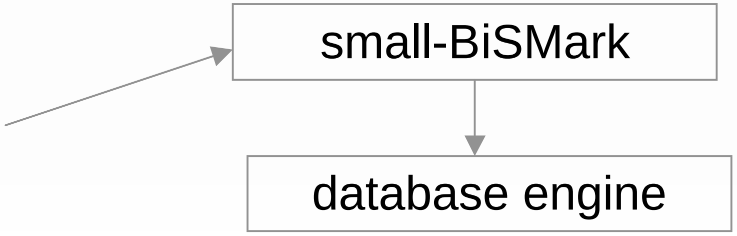
Test set Sp1: obtaining clean data



629 sites total
sequences lengths from 5 to 98
(22 average)

233 sites total
sequences lengths from 9 to 60
(25 average)

Using TRANSFAC as base data source for binding sites of a selected factor



Acknowledgments

- Vsevolod Makeev
- Andreas Heinzel <- From technical university Hagenberg, Austria
- Alexander Favorov
- Valentina Boeva -> Now at Universite Polytechniques, Palaiso, France
- Ivan Kulakovsky
- Dmitry Malko

Financial support

Russian Federation State Innovation Project, Russian Foundation of Basic Research, INTAS, Program in Molecular and Cellular biology, Russian Academy of Sciences

Special thanks to

BioBase GmBH