The ChIP-Seq project

Giovanna Ambrosini, Philipp Bucher

EPFL-SV Bucher Group

April 19, 2010 Lausanne

Overview

- Focus on technical aspects

  Description of applications (C programs)

  Where to find binaries, data files and src packages

  Release on SourceForge

  The Web Interface

- Large data sets have become available starting from the year 2007

  Barski et al. (2007): human CD4+ cell lines
  Histone modifications, POL II, CTCF (~2 millions tags per experiment)

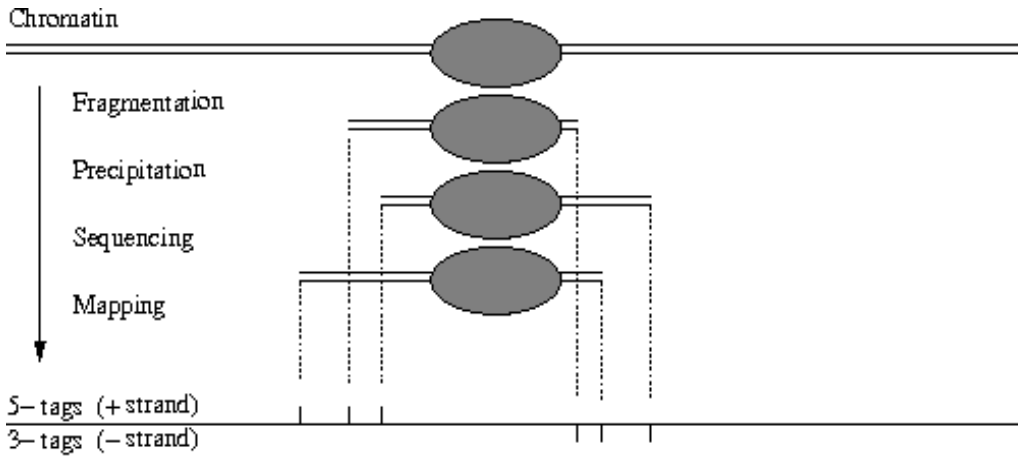
  Mikkelsen et al. (2007): four mouse cell lines
  Histone modifications (~2 millions tags per experiment)

  Robertson et al. (2007). INF-gamma stimulated HeLa cells
  STAT1 (>20 million tags per experiments)

- Very good data quality and reproducibility

- Choose to develop simple algorithms to achieve good results

ChIP-Seq Technique an Data



Our representation: SGA (Simple Genome Annotation) format

```
NC_000001.9     stim    559139  -       1
NC_000001.9     stim    559333  +       1
NC_000001.9     stim    559356  -       1
NC_000001.9     stim    559765  -       1
NC_000001.9     stim    559766  +       3
NC_000001.9     stim    559767  +       1
NC_000001.9     stim    559768  +       1
NC_000001.9     stim    559777  +       3
NC_000001.9     stim    559778  +       2
...
```

Fields of SGA format

1. Sequence ID
2. Feature
3. Position
4. Strand
5. Counts
6. Description

ChIP-Seq programs require SGA files to be sorted by chromosome name and position!

setenv LANG C; sort –s –k1,1 –k3,3 –k4,4

ChIP-seq data analysis: Biological questions

- ChIP-seq data for specific transcription factors
  - *Average length of an immuno-precipitated fragment (protected DNA regions)*
  - *The location of in vivo occupied sites*
  - *The strength of in vivo occupied sites*
  - *The in vivo binding specificity (consensus sequence, matrix)*
  - *Contextual features of in vivo occupied binding sites*
- ChIP-seq data for histone variants
  - *Which regions of the genome are enriched in a particular variants*
  - *Nucleosome phasing, position of individual nucleosomes*
  - *Epigenetic genome organization – definition of chromatin domains*
- Combined analysis
  - *Position of TF binding sites relative to nucleosomes*

ChipSeq Tools: Design principles and available tools

- Design principles

  - *Simple tools (easy to understand to non-specialists)*
  - *Fast algorithms*
  - *Generic methods if possible (not restricted to ChIP-seq data)*
  - *C-programs for basic programming tools*
  - *Auxiliary Perl tools to perform format conversion tasks and other useful tasks such as repeat masking and SGA fetching*

- Web interface (*http://ccg.vital-it.ch/chipseq*)

  - *Access and analysis of selected public data*
  - *Upload and analysis of private (user-owned) data*
  - *Combined analysis of private and public data*
  - *Interoperability with sequence analysis program (e.g. motif discovery)*
  - *Link to genome browser (preparation of customized WIG and BED files)*

ccg.vital-it.ch/chipseq

SIB Swiss Institute of Bioinformatics

ChIP-Seq

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Computational Cancer Genomics I ExPASy I EPFL

ChIP-Seq Analysis Server

The ChIP-Seq Server aims at providing tools for the analysis of ChIP-seq data and other types of MGA (mass genome annotation data).
The programs offered by this web server are listed below.

Access to ChIP-Seq tools	
ChIP-cor	Feature Correlation Tool
ChIP-center	Tag Centering Tool
ChIP-peak	Signal Peaks Detection Tool
ChIP-part	Partitioning Tool
WIG Files	WIG Files for public chIP-seq data

ChIP-Cor Application

❑ Input

- *Genomic tag count distributions for two features (reference, target)*
- *Features may be + and – strand tags from same experiments*
- *Applicable to other types of features, e.g. TSS positions*
- *Relative correlation distance, histogram step size and normalization*

❑ Output

- *A count correlation histogram (text file indicating the frequency of the target feature as a function of the relative distance to the reference feature)*

❑ Method

- *Consider reference positions which have at least one tag count.*
- *For each position, computes number of tag pairs that fall into a distance range.*
- *Different normalization options:*
  - *count density of target feature*
  - *global → relative target feature frequency (over-representation)*

❑ Purpose

- *Identification of average fragment size*
- *Reveals length distribution of enriched domains*
- *Provides clues for choosing parameters for peak and partitioning algorithms*
- *Positional relationship to other genomic features, e.g. transcription start sites*

❖ Output options on the Web

- *Histogram graph, feature extraction option*

Correlation plot: Example

Input data:

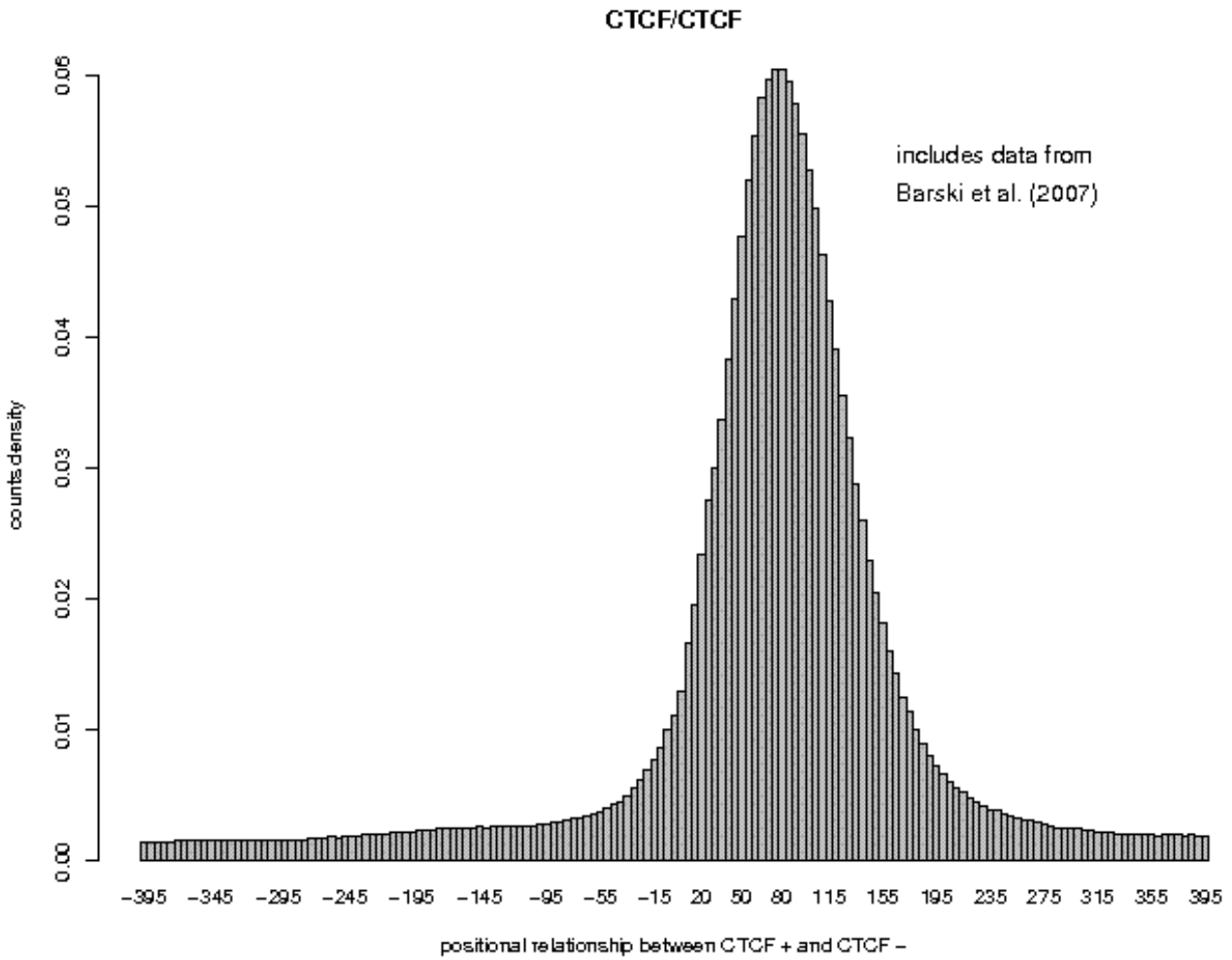
Ref: CTCF 5’ tags
Target: CTCF 3’ tags

Analysis parameters:

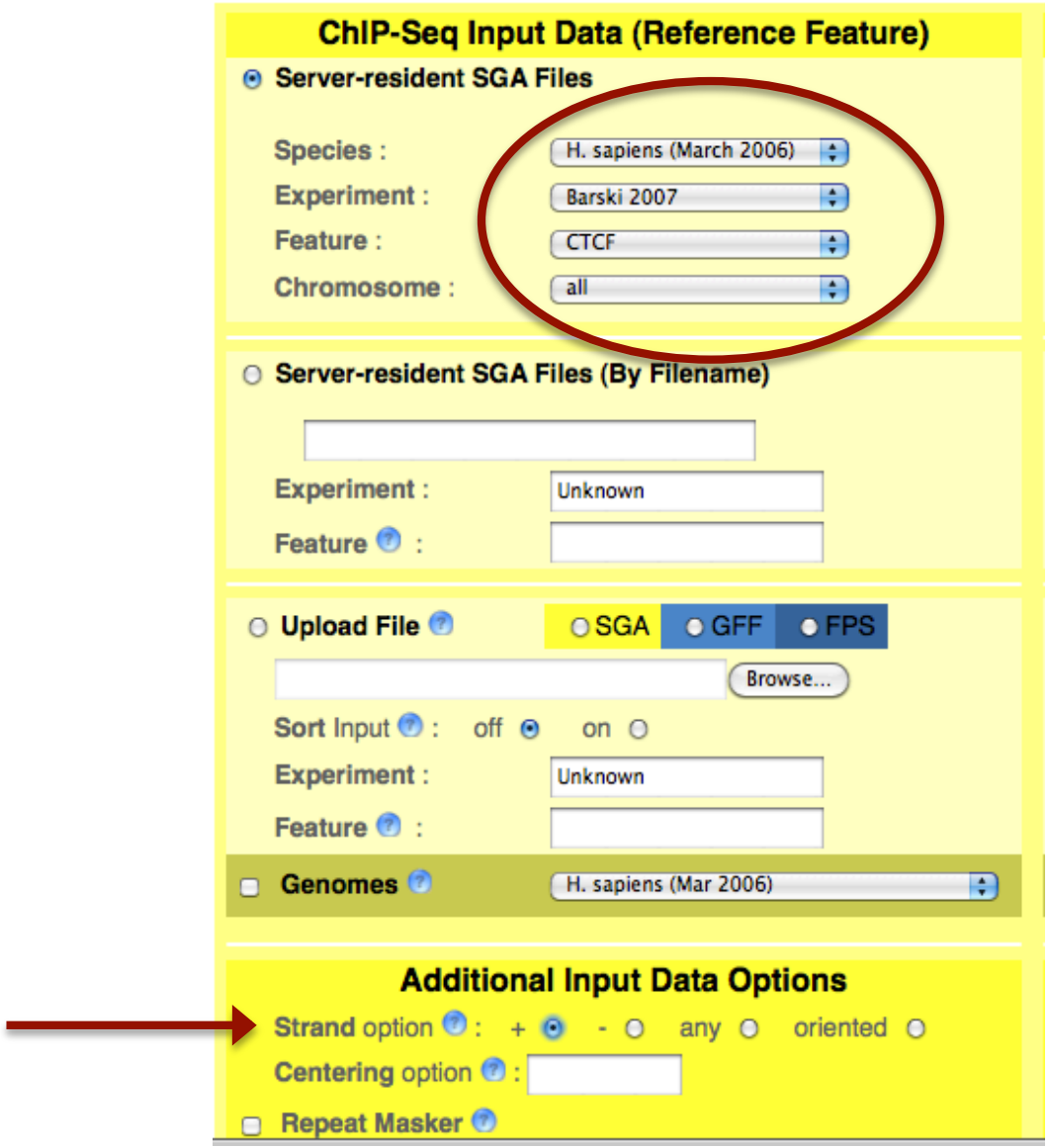
Range: –400,+400
Window width: 5
Count cut-off: 3
Y-axis: count-density

➢ Observations

➢Peak center: ~75
➢Peak count density: 0.06
➢Background: < 0.002



Correlation plot: input Data



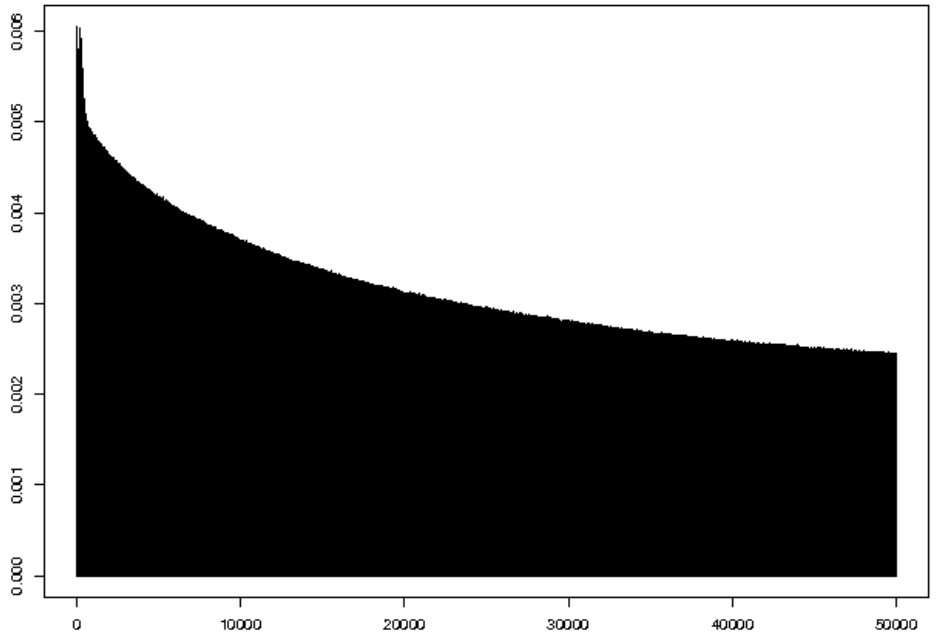
Auto-correlation plots for different histone modifications

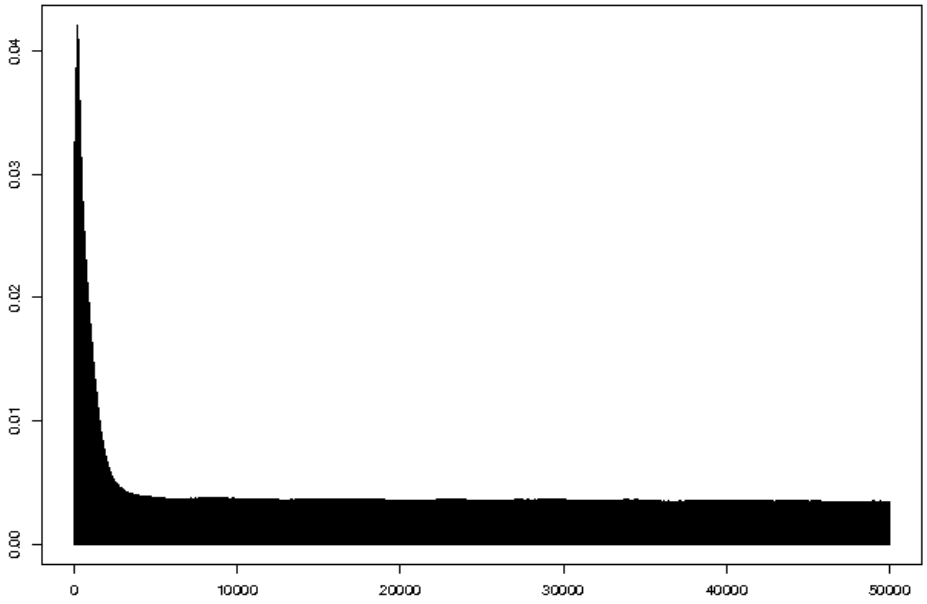
Top:

Auto-correlation plot of H3K36me3 in mouse ES cells

Bottom:

Auto-correlation plot of H3K4me3 in mouse ES cells





- Observations
  - H3K36me3 → long range correlation
  - H3K4me3 → short range correlation

ChIP-Score Application

❑ Input

- *Genomic tag count distributions for two features (reference, target)*
- *Count output threshold, distance range*

❑ Output

- *All reference sites that are enriched or depleted in target feature sites.*

❑ Method

- *Consider reference positions which have at least one tag count.*
- *For each position, computes cumulative target tag counts that fall into a distance range.*
- *Select those reference positions, which:*
  - *Have cumulative target tag count ≥ threshold (enriched feature)*
  - *Or have cumulative target tag count < threshold (depleted feature)*

❑ Purpose

- *Identification of enriched/depleted domains*
- *Further correlation to other genomic features, e.g. transcription start sites*

❖ Special ChIP-Cor Web server option

ChIP-Center Application

❑ Input

- *Oriented tag counts for a Chip-Seq feature*
- *Shift amount for centering tag positions*

❑ Output

- *Centered, un-oriented tag counts*

❑ Method

- *Moves by a given shift value tag positions to estimated center positions of DNA fragments*

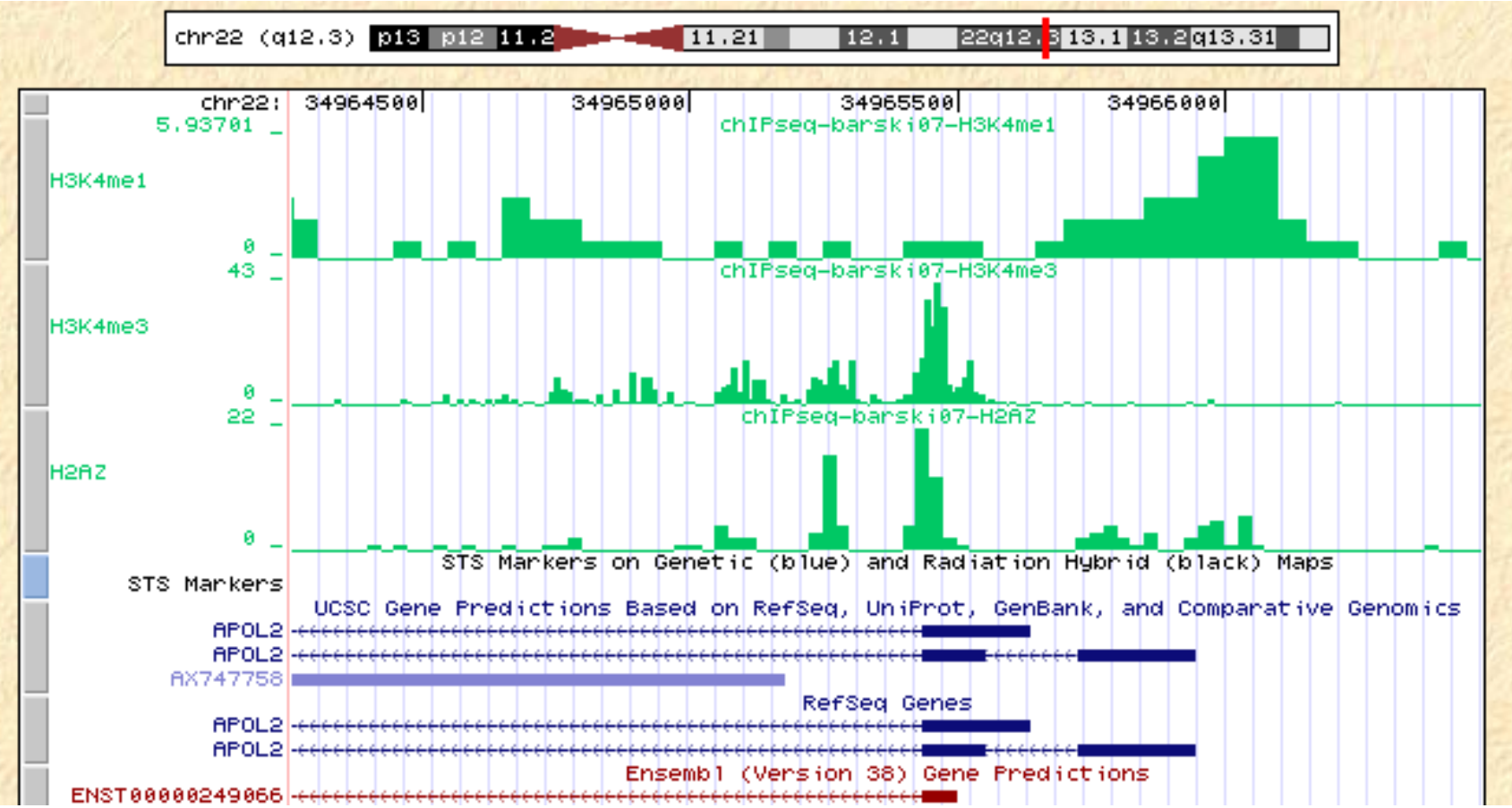
❑ Motivation

- *5' and 3' tag positions show relative displacement to each other*
- *best estimates for protein-binding site position:*
  5' end position + ½ fragment length
  or 3' end position – ½ fragment length
- *centered tag count distribution more useful as input for peak recognition and partitioning algorithm*

❖ Output options on the Web

- *WIG files for viewing data in a genome browser environment*

Example ChIP-center: Viewing customized WIG files in a UCSC browser environment



Based on data: from Barski et al. 2007, Cell 129, 823-837).
ChIP-Seq tags from both strands centered by 70 bp .
WIG file resolution: H3K4me1 50bp, H3K4me3 10 bp, H2A.Z 25 bp.

ChIP-peak Application

❑ Input

- *Centered tag counts*
- *Peak threshold (t), integration range of tag counts (w), Vicinity range (r)*

❑ Output

- *List of peak center positions (SGA or FPS format)*

❑ Method

- *Consider only positions which have at least one tag count.*
- *For each position, determines cumulative tag counts in windows of width w.*
- *select as peaks those positions, which*
  - *have cumulative tag count ≥ threshold t.*
  - *are local maximum with range ± r.*

❖ Special server options

- Download of sequences around peak center positions
- Provide several output formats: WIG, GFF, FPS

Example ChIP-Peak: Locating *in vivo* STAT1-binding sites

Input data

Robertson et al. (2007) Nature Methods 4, 651-657.
Cell material: Interferon γ-stimulated HeLa S3 cells.
About 15 million tags in total

Analysis parameters

Centering: 70bp, window 200bp, exclusion range 200 bp, threshold 100 counts

➢ Result: 4446 peaks

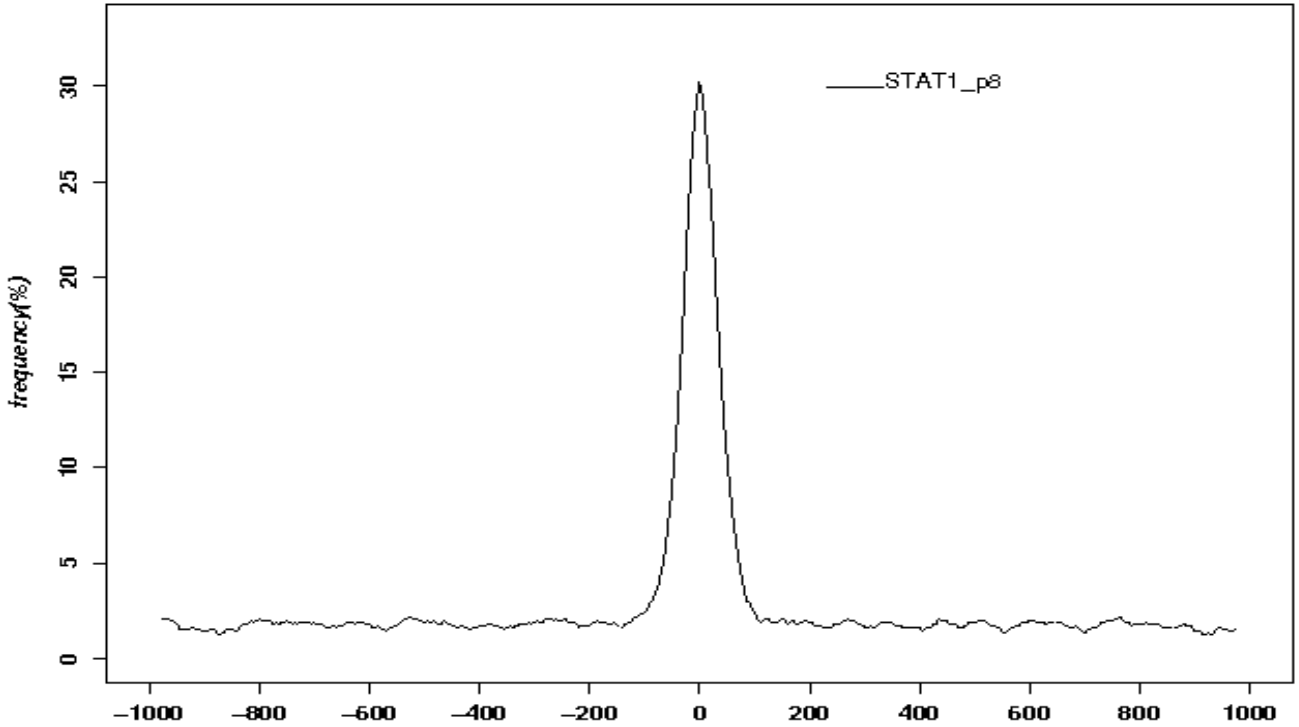
Sequence extraction range for downstream analysis: –1000, 1000

Donwstream sequence analysis

✧Distribution of TTCNNNGAA around STAT1 peak

✧Sliding window size 50

✧Figure produced with OPROF (SSA server)



STAT1 Sequence Motif Defined by ChIP-Seq data

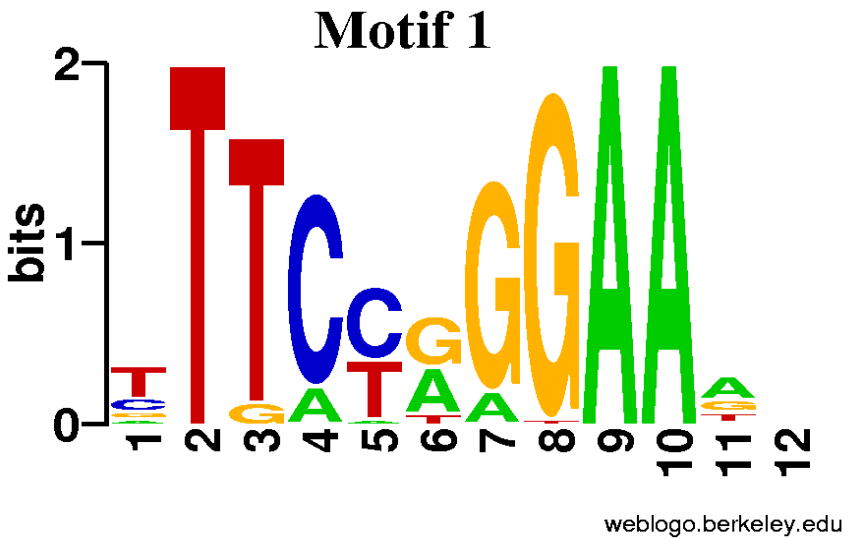
Input

4446 ChIP peak regions

200 bp

ab initio motif discovery

MEME (zoops)



Matrix from experimental *in vivo* sites

	-7	-6	-5	-4	-3	-2	-1	0
A	23	38	15	0	2	29	8	19
C	33	17	13	0	0	67	56	31
G	35	17	12	4	2	4	15	31
T	10	27	60	96	96	0	21	19

Matrix from SELEX

	-7	-6	-5	-4	-3	-2	-1	0
A	6	62	26	2	2	5	2	2
C	57	13	27	2	3	89	95	48
G	23	14	10	2	2	2	2	49
T	14	11	36	95	93	4	2	2

ChIP-partition Application

❑ Input

- Centered tag counts
- Count density threshold, transition penalty

❑ Output

- List of signal-enriched regions (beginning, end)

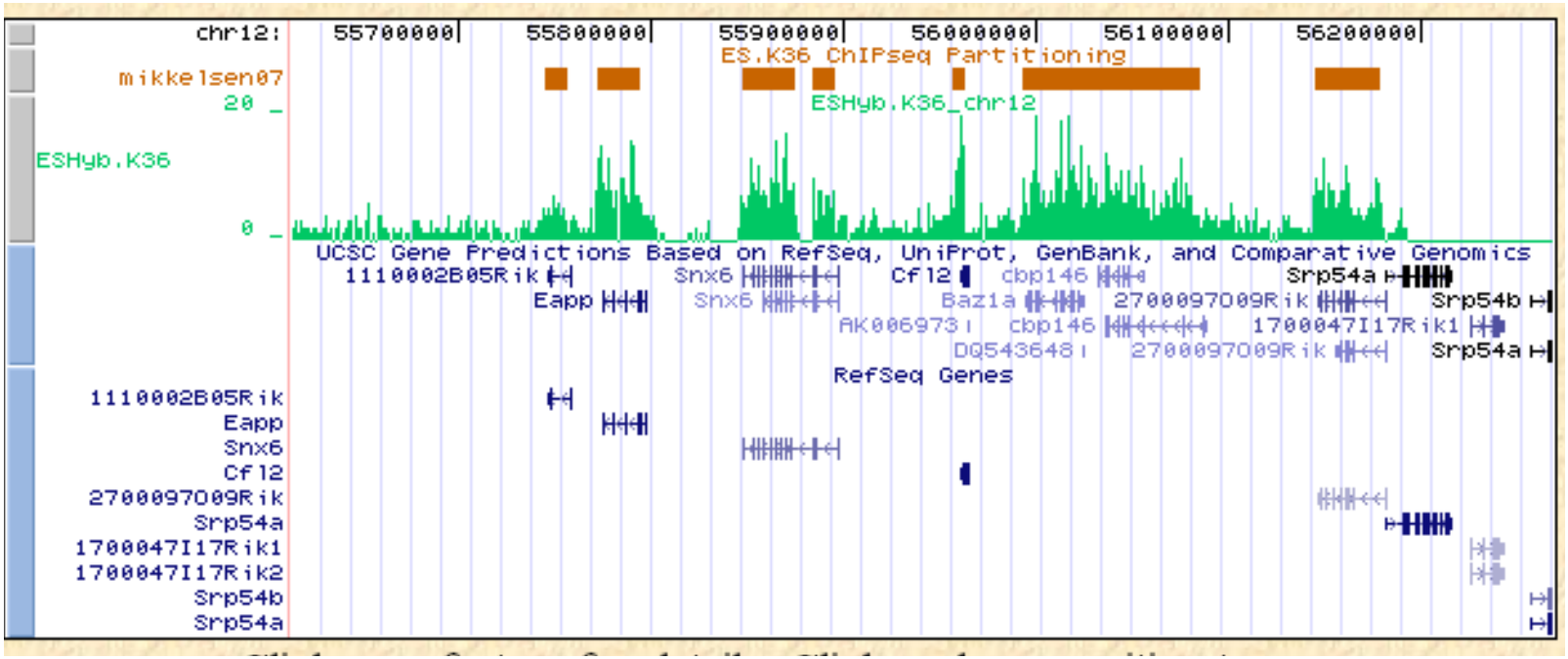
❑ Method

- *Optimization of a partition scoring function by a fast dynamic programming algorithm*
- *Two parameters: count density threshold, transition penalty*
- *Scoring functions for global results: sum of scores of*
  - *Transitions (penalty)*
  - *Signal-rich region: length × (local count-density – threshold)*
  - *Score for signal-poor region: length × (threshold – local count-density)*

❖ Output options on the Web

- *GFF, BED file for genome browser*

Viewing the results of the partitioning program in the genome browser



Custom tracks:

Mikkelsen07: results of ChIP-partition program (BED file)

ESHyb.K36: from: http://www.isrec.isb-sib.ch/WIG/HSM07_ESHyb.K36_m_chr12.wig

Where to find source and binary files

- On Vital-IT (at SIB – http://www.isb-sib.ch)

  /mnt/common/share/chip-seq-1.1.0

  /bin.em64t -> /mnt/local/bin

  chipcor chipscore chipcenter chippeak chippart

  /src

  chipcor.c chipscore.c chipcenter.c chippeak.c chippart.c

  /perl

  eland2sga.pl bed2sga.pl gff2sga.pl sga2.bed.pl sga2fps.pl counts_filters.pl fetch_sga.pl

- On SourceForge

  http://sourceforge.net/projects/chip-seq

  tarball file: chip-seq.1.1-0.tar.gz

  www : http://chip-seq.sourceforge.net

- Documentation

  Man pages, README files, C programs

  On going projects: ChIP-seq Web Tutorial, PDF User Manual, Reference Technical Manual

Where to find data files

- On all platforms (within the Vital-IT environment)

/db/chipseq

Experiments

Barski et al. (2007): human CD4+ cell lines
Histone modifications, POL II, CTCF (~2 millions tags per experiment).

Mikkelsen et al. (2007): four mouse cell lines
Histone modifications (~2 millions tags per experiment).

Robertson et al. (2007). INF-gamma stimulated HeLa cells
STAT1 (>20 million tags per experiments).

Boyle et al. (2008): CD4+ cell lines
Open chromatin studies (~10 millions tags per experiment).

Wang et al. (2008): human CD4+ cells lines
Histone acetylations and methylations (3,4 millions tags per experiments).

Schones et al. (2008): human CD4+ cells lines
Regulation of nucleosome positioning (100 millions tags per experiments).

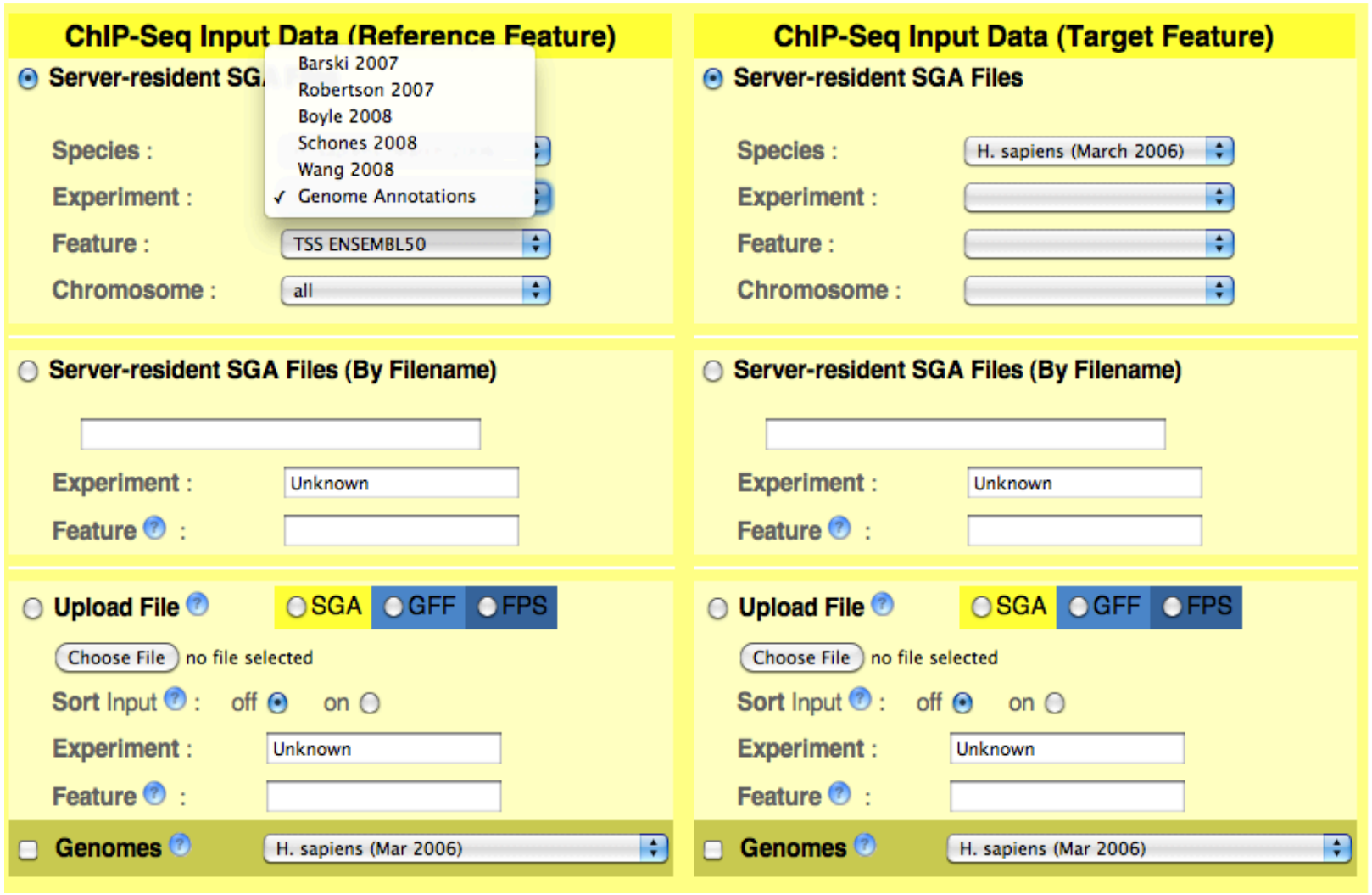
....

Genome Annotations

CAGE, ENSEMBL, DBTSS7 and EPD TSS
Repeat masks
Phastcons tracks from UCSC
ENSEMBL POLYA

Where to find data files (cont.)

- On the Web Server



Web Access Statistics

- Chip-Seq Web Server on ccg.vital-it.ch *Week: Jan 10 2010 - Jan 17 2010*

  - ✓ *Tot Number of Accesses : 828*
  - ✓ *Tot Number of IPs : 95*

  - ✓ *Number of Accesses from Switzerland: 316*
  - ✓ *Number of IPs from Switzerland: 10*

  - ✓ *Number of Accesses from Abroad: 512*
  - ✓ *Number of IPs from Abroad: 85*

  - ✓ *Number of Accesses from inside the UNIL: 4*
  - ✓ *Number of IPs from the UNIL: 3*
  - ✓ *Number of Accesses from the EPFL: 293*
  - ✓ *Number of IPs from the EPFL: 4*