ATAC-Seq pipelines

Oleg Shpynov

JetBrains Biolabs

July 7, 2016

Agenda

- ATAC
- Who is who?
- ATAC vs FAIRE
- Pipelines

ATAC

ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding

Tamaki Suganuma[1], José L Gutiérrez[1,2], Bing Li[1], Laurence Florens[1], Selene K Swanson[1], Michael P Washburn[1], Susan M Abmayr[1] & Jerry L Workman[1]

The Ada2a-containing (ATAC) complex is an essential *Drosophila melanogaster* histone acetyltransferase (HAT) complex that contains the transcriptional cofactors Gcn5 (KAT2), Ada3, Ada2a, Atac1 and Hcf. We have analyzed the complex by MudPIT (multidimensional protein identification technology) and found eight previously unidentified subunits. These include the WD40 repeat protein WDS, the PHD and HAT domain protein CG10414 (herein renamed Atac2/KAT14), the YEATS family member D12, the histone fold proteins CHRAC14 and NC2β, CG30390, CG32343 (Atac3) and CG10238. The presence of CG10414 (Atac2) suggests that it acts as a second acetyltransferase enzyme in ATAC in addition to Gcn5. Indeed, recombinant Atac2 displays HAT activity *in vitro* with a preference for acetylating histone H4, and mutation of *Atac2* abrogated H4 lysine 16 acetylation in *D. melanogaster* embryos. Furthermore, although ATAC does not show nucleosome-remodeling activity itself, it stimulates nucleosome sliding by the ISWI, SWI–SNF and RSC complexes.

Be careful!

ATAC-Seq

Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) is a method for mapping chromatin accessibility genome-wide.

High resolution mapping

High-Resolution Mapping and Characterization of Open Chromatin across the Genome

Alan P. Boyle,[1] Sean Davis,[3] Hennady P. Shulha,[2] Paul Meltzer,[3] Elliott H. Margulies,[4] Zhiping Weng,[2] Terrence S. Furey,[1,*] and Gregory E. Crawford[1,*]
[1]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708, USA
[2]Biomedical Engineering Department, Boston University, Boston, MA 02215, USA
[3]Center for Cancer Research, National Cancer Institute
[4]National Human Genome Research Institute
National Institutes of Health, Bethesda, MD 20892, USA
*Correspondence: terry.furey@duke.edu (T.S.F.), greg.crawford@duke.edu (G.E.C.)
DOI 10.1016/j.cell.2007.12.014

1

In addition, and unexpectedly, our analyses have uncovered detailed features of nucleosome structure.

[1] 650+ cited

DNAse I

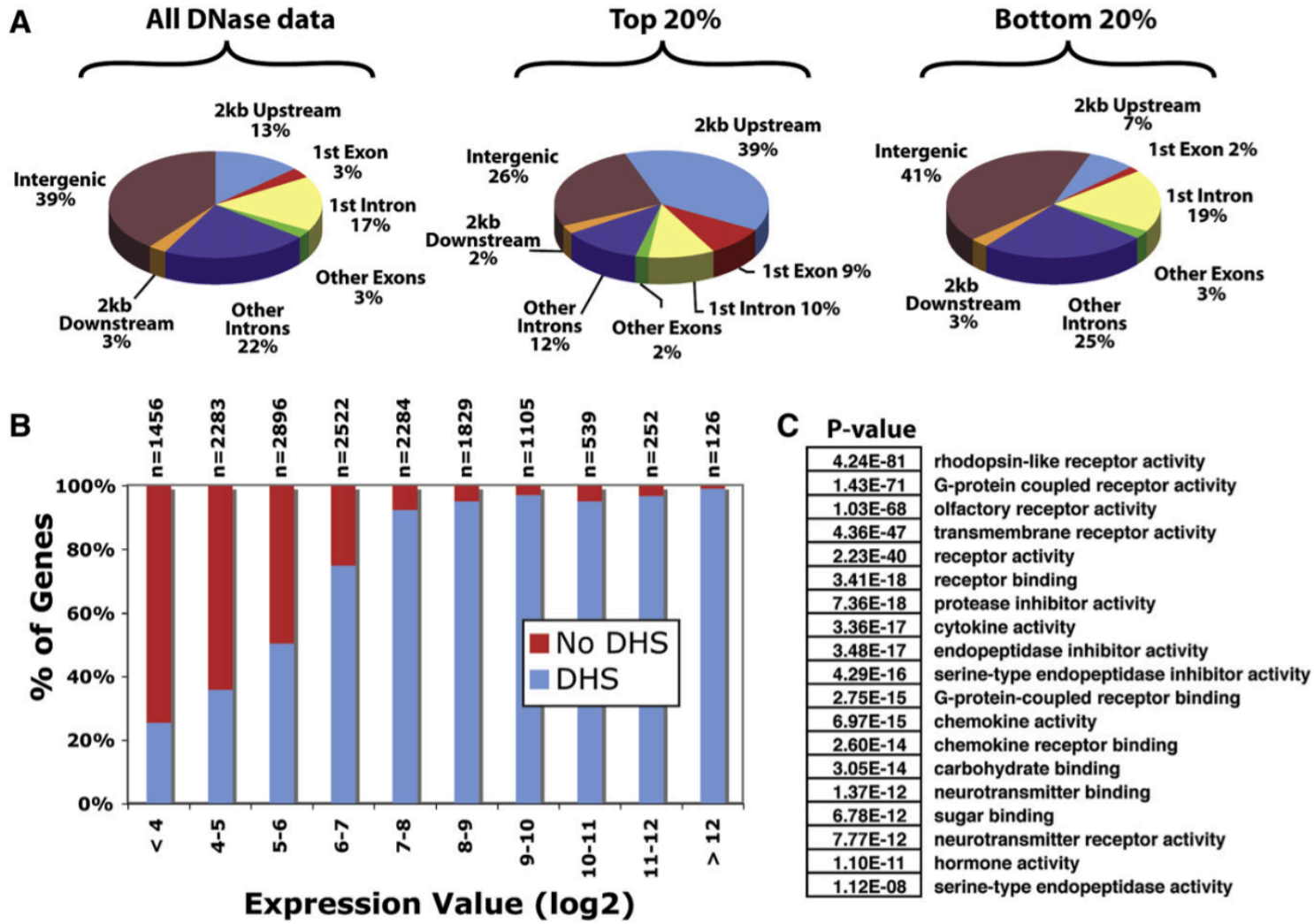


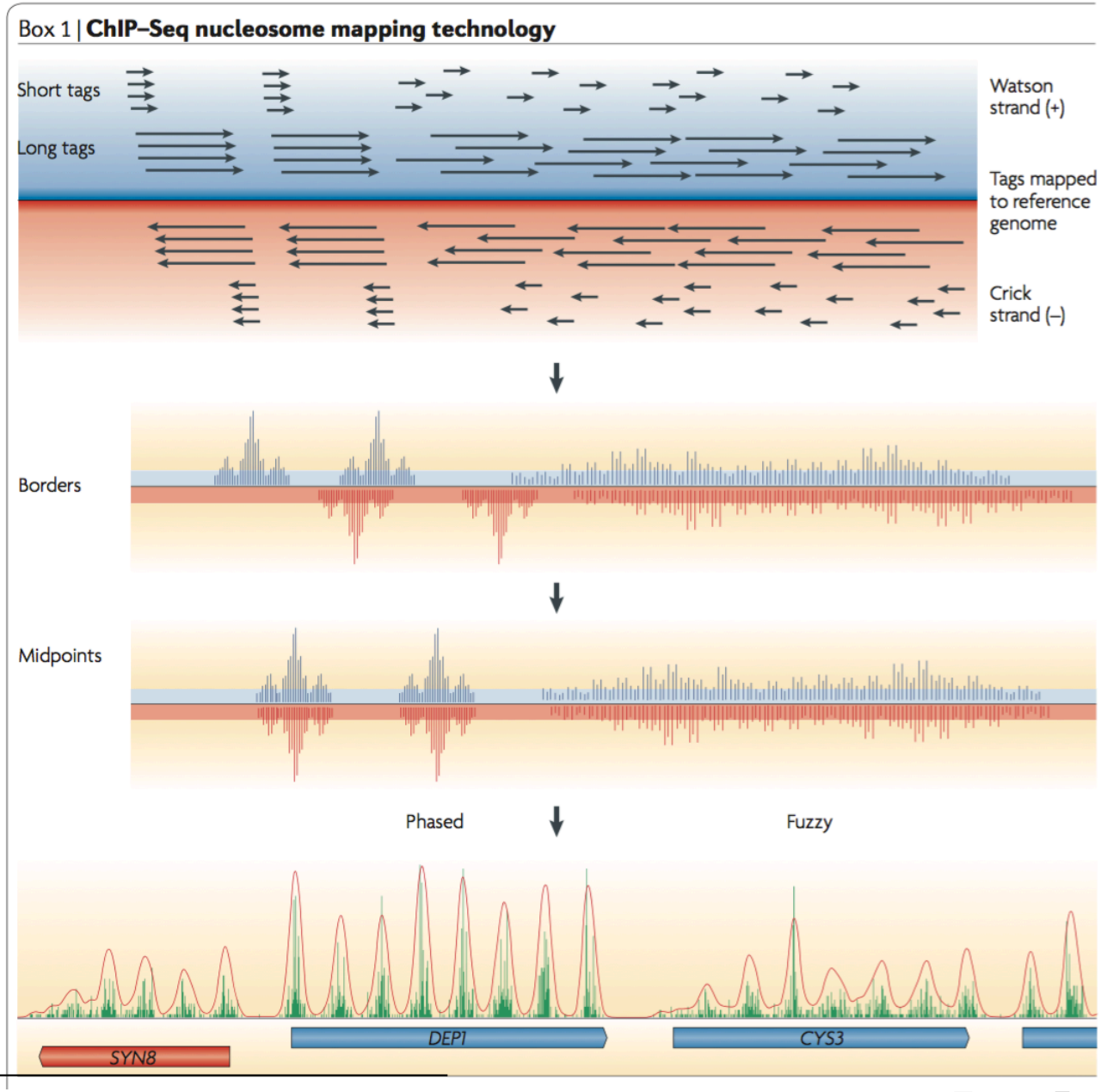
Figure 3. Location of DNase I Hypersensitive Sites Relative to Annotated Genes

(A) The locations of DNase I hypersensitive (DHS) sites relative to gene annotations. Shown are the locations of all DNase I HS sites, the strongest scoring DNase I HS sites (top 20%), and the weakest scoring DNase I HS sites (bottom 20%).

(B) Genes that have high expression (>9) are likely to have a DNase I HS site at the 5′ end, while genes lacking a 5′ DNase I HS site are more likely to have low expression.

(C) GO categories and probabilities related to genes that are lacking 5′ DNase I HS sites.

MNAse



[2]

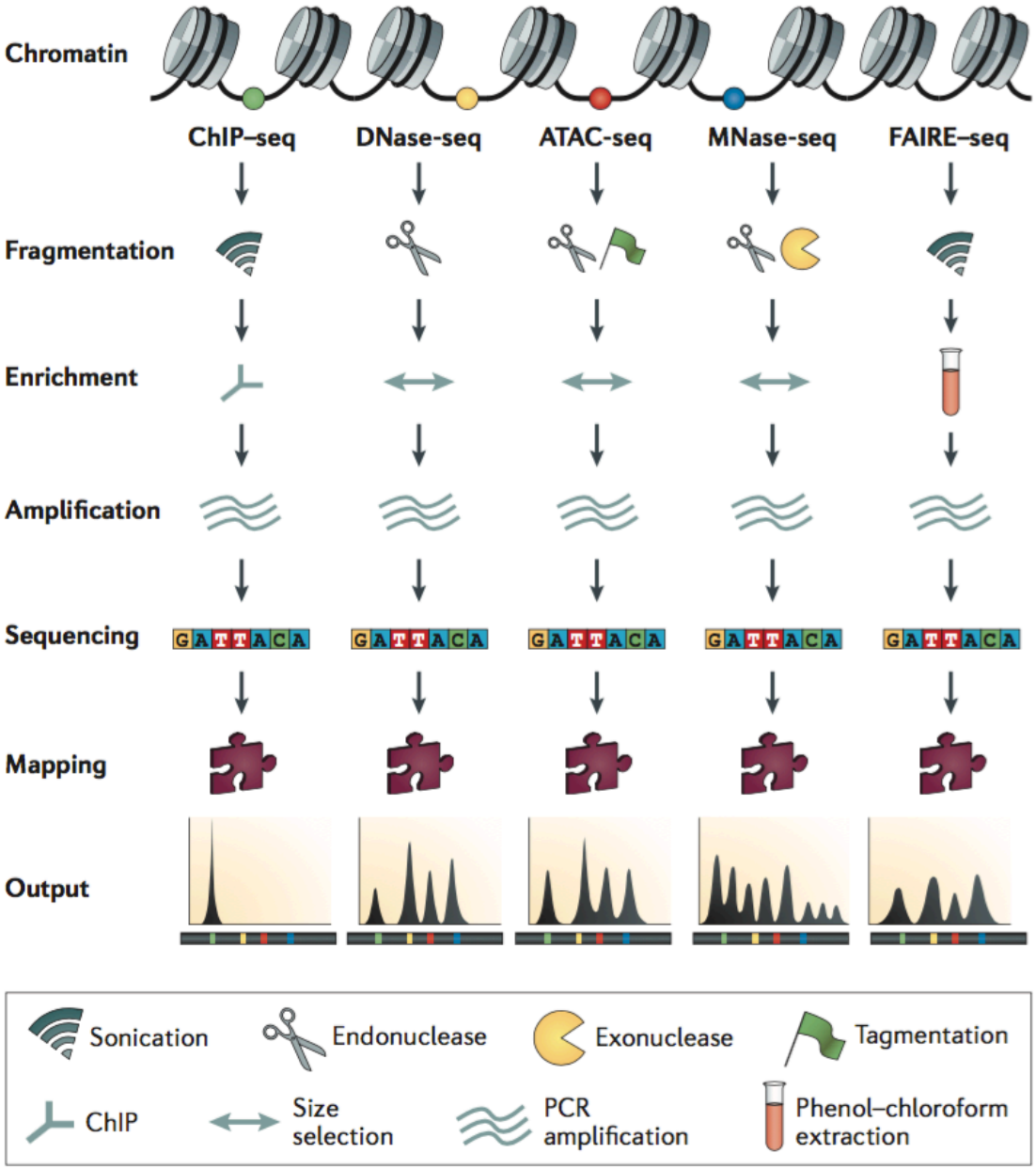
[2] http://www.nature.com/nrg/journal/v10/n3/full/nrg2522.html

Who is who?

Identifying and mitigating bias in next-generation sequencing methods for chromatin biology

Clifford A. Meyer and X. Shirley Liu

Abstract | Next-generation sequencing (NGS) technologies have been used in diverse ways to investigate various aspects of chromatin biology by identifying genomic loci that are bound by transcription factors, occupied by nucleosomes or accessible to nuclease cleavage, or loci that physically interact with remote genomic loci. However, reaching sound biological conclusions from such NGS enrichment profiles requires many potential biases to be taken into account. In this Review, we discuss common ways in which biases may be introduced into NGS chromatin profiling data, approaches to diagnose these biases and analytical techniques to mitigate their effect.


Chromatin
ChIP–seq
DNase-seq
ATAC-seq
MNase-seq
FAIRE–seq
Fragmentation
Enrichment
Amplification
Sequencing
GATTACA
Mapping
Output
Sonication
Endonuclease
Exonuclease
Tagmentation
ChIP
Size selection
PCR amplification
Phenol–chloroform extraction

Bias, Bias, Bias, ...

- Chromatin fragmentation and size selection: sonication, enzymatic cleavage.
- Controls for enzymatic cleavage assays.
  Genomic assays that are based on the selection of fragments produced from enzymatic DNA cleavage including ATAC- seq, DNase-seq and MNase-seq may be influenced by the tendency of the enzyme to cleave some DNA sequences more efficiently than others.
- Identifying Bias.
  The **ChiLin** quality control pipeline is a good starting point for understanding the quality and bias characteristics of ChIPseq, DNase-seq and ATAC-seq samples.

Summary: ATAC-seq requires fewer cells and less experimental calibration, but bias characteristics are not as well understood as those of DNase-seq.[3]

[3] http://www.ncbi.nlm.nih.gov/pubmed/24317252

ChiLin

[ЦИТИРОВАНИЕ] **ChiLin**: ChIP-seq data analysis and **quality control Pipeline**
Q Qin, H Sun, S Mei, L Taing, M Brown, F Li, HW Long… - 2014
Похожие статьи Цитировать Сохранить

FURTHER INFORMATION

ChiLin: http://liulab.dfci.harvard.edu/software

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

4

Welcome to Cistrome

The cistrome refers to "the set of cis-acting targets of a trans-acting factor on a genome-wide scale, also known as the in vivo genome-wide location of transcription factor binding-sites or histone modifications". Here we build integrative analysis pipelines (Cistrome) to help experimental biologists, and conduct efficient data integration to better mine the hidden biological insights from publicly available high throughput data.

5

[4] Correct url: `http://cistrome.org/chilin/`

[5] `http://cistrome.org/Cistrome/Cistrome_Project.html`

ATAC-Seq

Paper 2014:

Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position

Jason D Buenrostro[1–3], Paul G Giresi[2,3], Lisa C Zaba[2,3], Howard Y Chang[2,3] & William J Greenleaf[1]

Protocol for humans 2015:

ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide

Jason D. Buenrostro,[1,2] Beijing Wu,[1] Howard Y. Chang,[2] and William J. Greenleaf[1]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California
[2]Program in Epithelial Biology and the Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California

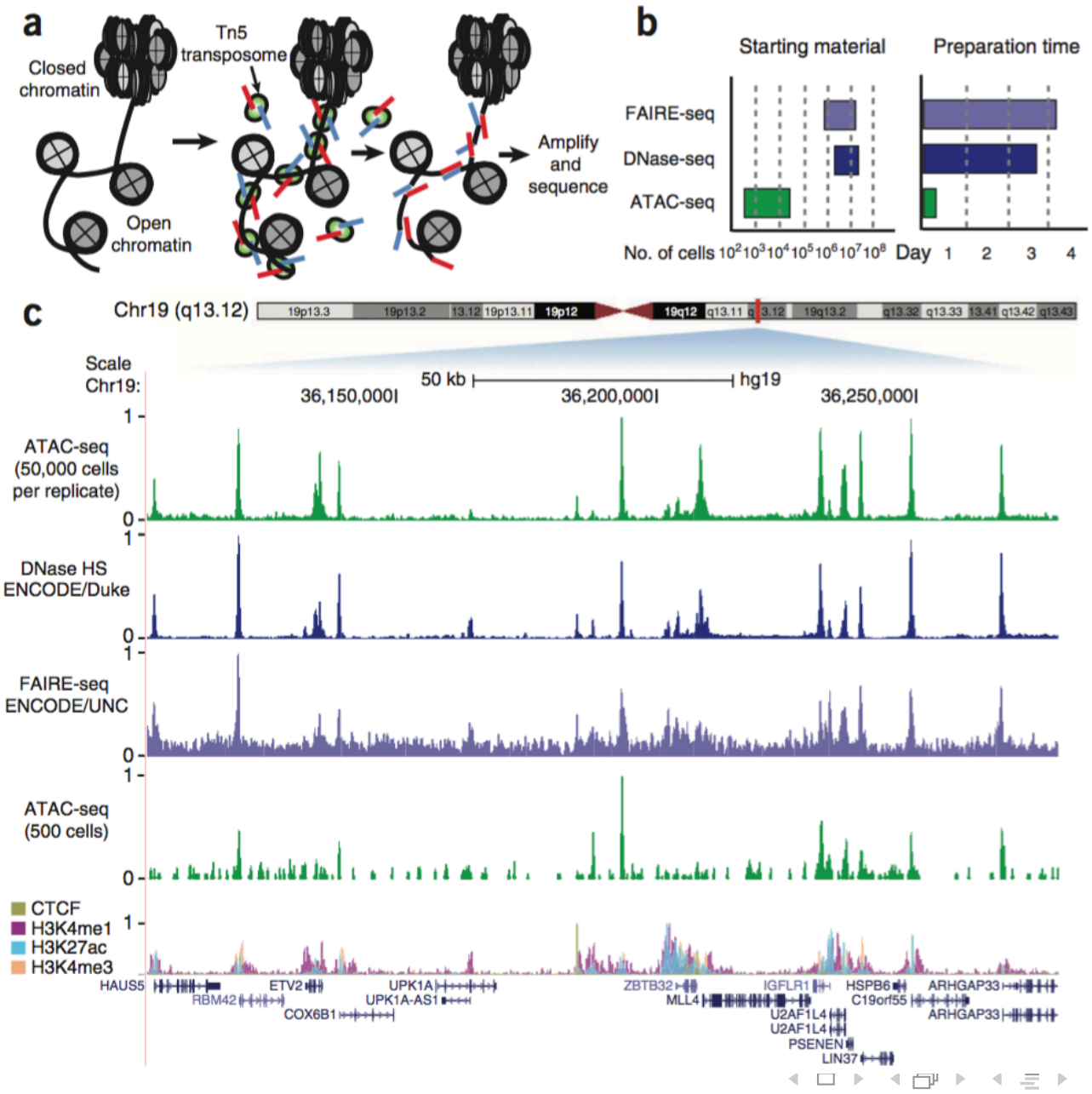
Motivation

MNase-seq, ChIP-seq, and DNase-seq in particular have proven to be information-rich, genome-wide analysis methods for understanding this epigenetic structure, providing information on transcription factor binding, the positions of modified and canonical nucleosomes, and chromatin accessibility...
However, **tens to hundreds of millions** of cells as input material are necessary.[6]

- Average over and drown out heterogeneity in cellular populations
- Cells must often be grown ex vivo to obtain sufficient material
- No personal epigenomes on diagnostic timescales

[6] 1-50mln for GM12878 in paper 2014

Tagmentation, material, time



ATAC-Seq length distribution

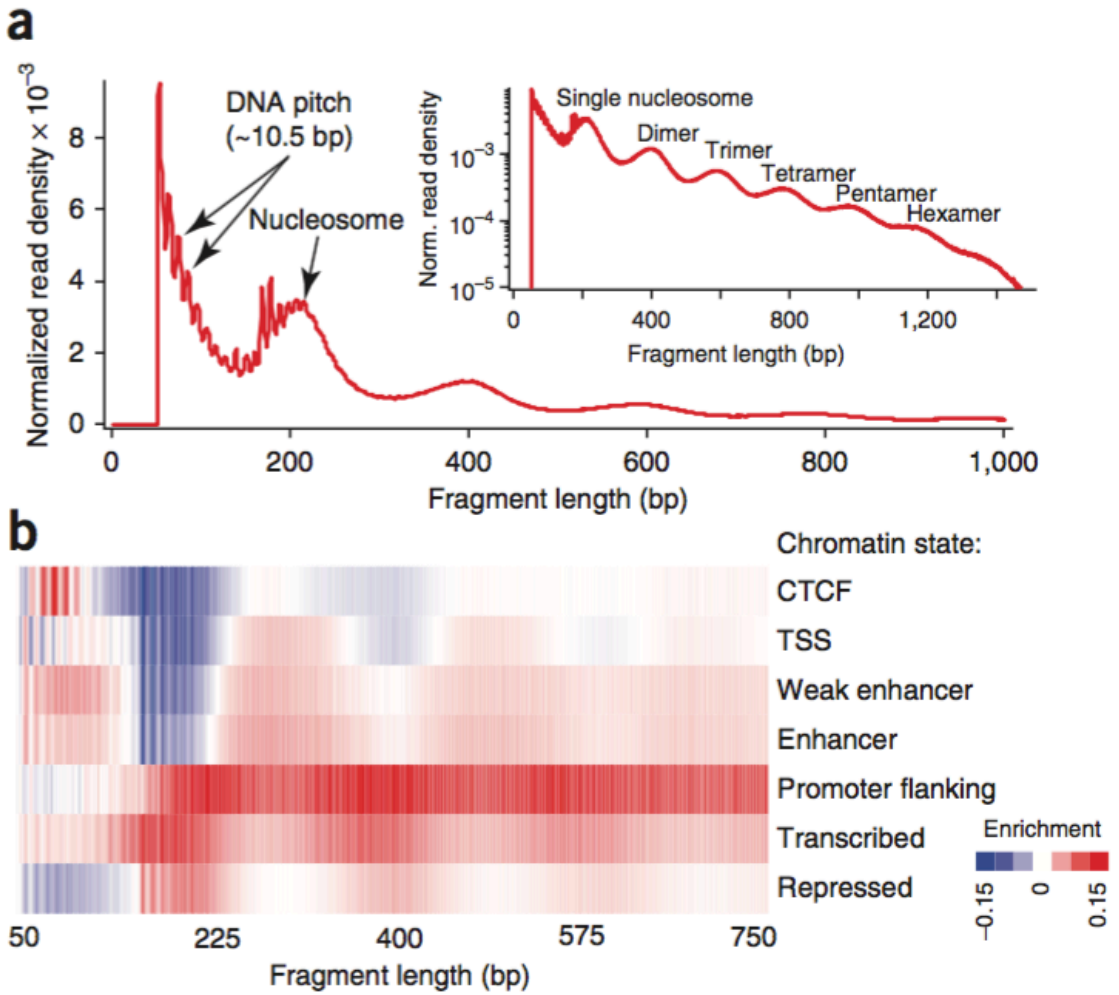


Figure 2 | ATAC-seq provides genome-wide information on chromatin compaction. (**a**) ATAC-seq fragment sizes generated from GM12878 nuclei. Inset, log-transformed histogram shows clear periodicity persists to six nucleosomes. (**b**) Normalized read enrichments for seven classes of chromatin state previously defined[17].

Summary

Experimental:

- Peak intensities highly reproducible between technical replicates (R = 0.98) and DNase-seq data (R = 0.79 and R = 0.83)
- Sensitivity was **diminished** for small $\leq$ 500 numbers of input material.

Protocol:

- Relatively simple protocol that can be carried out in hours for a standard sample size of 50000 cells.
- The **biggest** source of failure comes from variations in cell number.[7]
- For nucleosome mapping, paired-end sequencing is preferred.
- Difference in human open chromatin human $\geq$ 50mln mapped reads
- TF foot-printing $\geq$200mln mapped reads.
- Half of reads are approximately of subnucleosomal length (150 bp) and half of the reads are longer.

[7] Too few cells causes over-digestion of chromatin - reads that map to inaccessible regions (noise); too many cells causes under-digestion and creates high-molecular-weight fragments.

VS others

Discovery of Transcription Factors and Regulatory Regions Driving *In Vivo* Tumor Development by ATAC-seq and FAIRE-seq Open Chromatin Profiling

Kristofer Davie[1‡], Jelle Jacobs[1‡], Mardelle Atkins[2,3], Delphine Potier[1], Valerie Christiaens[1], Georg Halder[2,3], Stein Aerts[1*]

1 Laboratory of Computational Biology, Center for Human Genetics, KU Leuven, Leuven, Belgium, **2** Laboratory of Growth Control and Cancer Research, Center for Human Genetics, KU Leuven, Leuven, Belgium, **3** VIB Center for the Biology of Disease, Laboratory for Molecular Cancer Biology, Leuven, Belgium

‡ These authors contributed equally to this work.
* stein.aerts@med.kuleuven.be

Here we ask whether open chromatin profiling can be used to identify the entire repertoire of active promoters and enhancers underlying tissue-specific gene expression during normal development and onco- genesis in vivo.
In conclusion, we show that FAIRE-seq and ATAC-seq based open chromatin profiling, combined with motif discovery, is a straightforward approach to identify functional genomic regulatory regions, master regulators, and gene regulatory networks controlling complex in vivo processes.

Conclusions

- Both methods are robust in identifying accessible or open regions.
- ATAC-seq shows slightly lower background levels.
- ATAC-seq shows a higher recall of true enhancers than FAIRE-seq.
- ATAC can detect both smaller and greater significant differences between normal and tumor states than FAIRE.

NucleoATAC[8]

Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions

Alicia N. Schep,[1] Jason D. Buenrostro,[1] Sarah K. Denny,[2] Katja Schwartz,[1] Gavin Sherlock,[1] and William J. Greenleaf[1,3]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; [2]Biophysics Program, Stanford University School of Medicine, Stanford, California 94305, USA; [3]Department of Applied Physics, Stanford University, Stanford, California 94305, USA

[8] https://github.com/GreenleafLab/NucleoATAC

Existing tools

DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing
K Chen, Y Xi, X Pan, Z Li, K Kaestner, J Tyler… - Genome …, 2013 - genome.cshlp.org
Abstract Recent developments in next-generation sequencing have enabled whole-genome profiling of nucleosome organizations. Although several algorithms for inferring nucleosome position from a single experimental condition have been available, it remains a challenge ...
Цитируется: 56 Похожие статьи Все версии статьи (15) Цитировать Сохранить

NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model
A Polishko, N Ponts, KG Le Roch, S Lonardi - Bioinformatics, 2012 - Oxford Univ Press
Abstract Motivation: Nucleosomes are the basic elements of chromatin structure. They control the packaging of DNA and play a critical role in gene regulation by allowing physical access to transcription factors. The advent of second-generation sequencing has enabled ...
Цитируется: 21 Похожие статьи Все версии статьи (12) Цитировать Сохранить

Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions
AN Schep, JD Buenrostro, SK Denny… - Genome …, 2015 - genome.cshlp.org
... green) or DANPOS2 (orange). **NucleoATAC** enables high-resolution nucleosome calling in S. cerevisiae. **NucleoATAC** identified the ... time (Supplemental Fig. 12). **NucleoATAC** can be applied across species. Because histones are among ...
Цитируется: 4 Похожие статьи Все версии статьи (11) Цитировать Сохранить

9

[9] Based on gaussian mixtures.

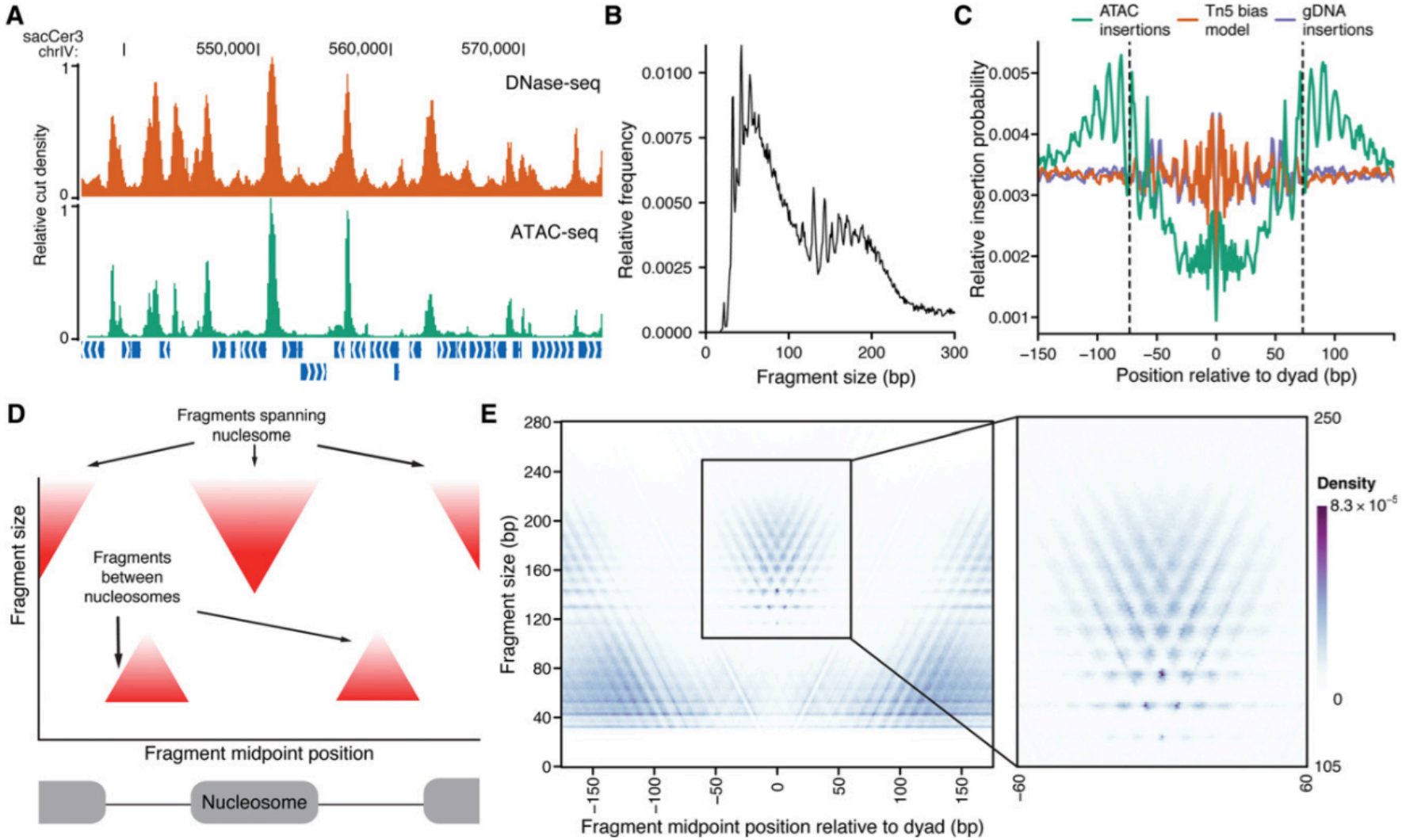


Figure 1. ATAC-seq signal is highly structured around nucleosomes. (*A*) ATAC-seq (green) insertion track for *S. cerevisiae* shows enrichment of insertions at accessible chromatin regions, similar to DNase-seq cut density (orange). Both tracks were smoothed by 150 bp and scaled so that the maximum density in the region is 1. (*B*) Fragment-size distribution for *S. cerevisiae* ATAC-seq samples. (*C*) Insertion probabilities for ATAC-seq (teal), genomic DNA (purple), and predicted by sequence bias (orange) (see Methods) around nucleosomes defined by chemical mapping. (*D*) Schematic illustration of expected V-plot pattern around a well-positioned nucleosome. (*E*) V-plot (fragment size versus fragment center position) of ATAC-seq fragments around well-positioned nucleosomes called by chemical mapping, with *inset* showing region with nucleosome-spanning fragments.

V-plot

Density of fragment sizes versus fragment center locations relative to a genomic feature of interest.[10]

- Apex of the V represents the smallest possible fragment that spans the DNA protected by a nucleosome (117bp).
- Horizontal and vertical periodicity likely reflects both the steric hindrance of the transposase (vertical and horizontal periodicity) and 10-bp rotational positioning of nucleosomes in *yeast*.
- Stochastic breathing of DNA - most abundant position in the V-plot represents fragments of 143 bp centered at the dyad.
- NucleoATAC identified the positions of 13344 nucleosomes across broad open chromatin regions in the yeast[11] out of 17015 chemical mapping.

[10] Introduced in`http://www.cs.duke.edu/courses/compsci662/spring15/pdf/henikoff.2011.pdf`

[11] Z-score$\geq$3, log-likelihood ratio$\geq$ 0

NucleoATAC vs DANPOS2

Table 1. Positional concordance metrics for nucleosome calls in *S. cerevisiae*

Assay	Inference method	Number of calls	Distance AUC	Sensitivity	Specificity	Rotational specificity
ATAC	NucleoATAC	13,344	0.721	0.479	0.611	0.340
ATAC	DANPOS2	14,261	0.685	0.436	0.521	0.149
MNase	NucleoATAC	15,725	0.764	0.604	0.653	0.371
MNase	DANPOS2	18,519	0.719	0.600	0.551	0.157
MNase	PuFFIN	17,452	0.750	0.629	0.613	0.188
None	Random tiling	19,185	0.512	0.273	0.242	0.061

Summary

- Reads QC - ChiLin
- Nucleosome positioning - NucleoATAC

Fin