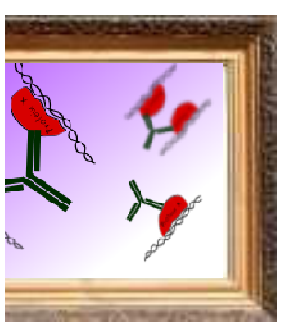
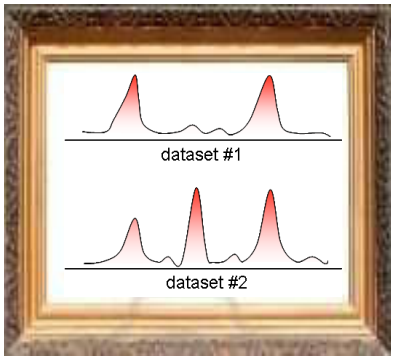
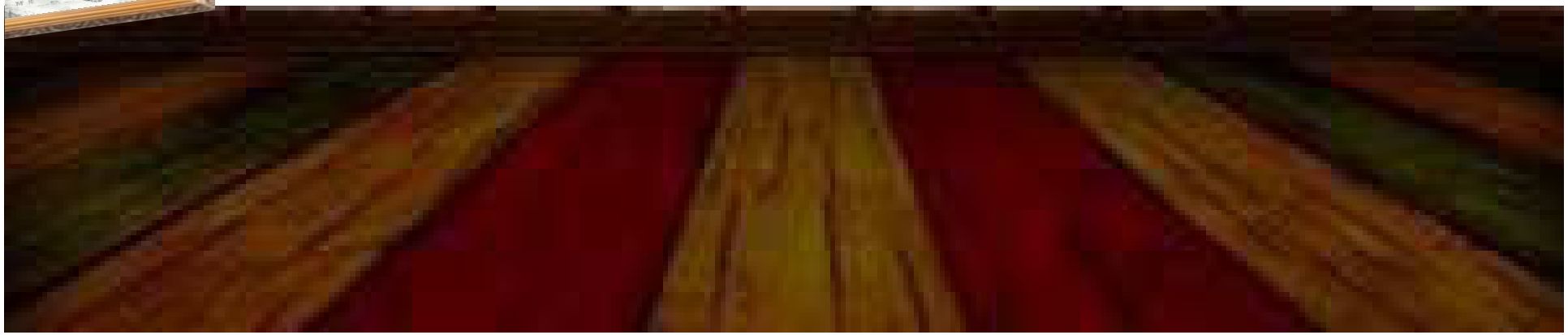
A quality control system for the certification of ChIP-seq and enrichment-related NGS-generated datasets



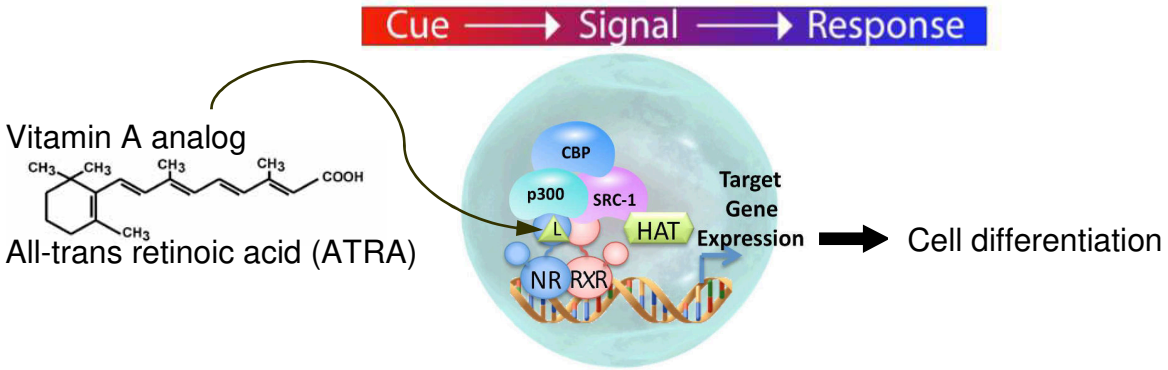


Marco Antonio Mendoza Parra

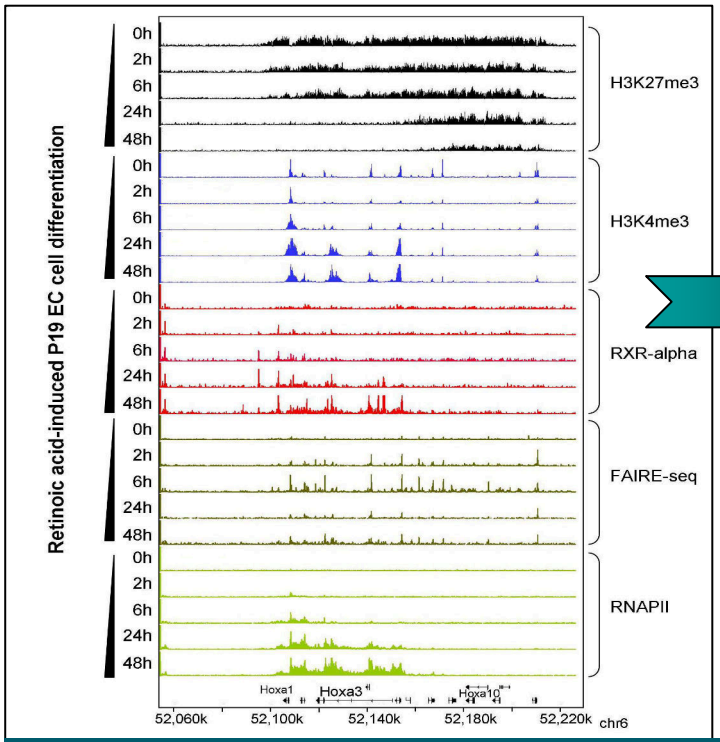
Dept. Functional Genomics and Cancer
IGBMC; Strasbourg / France



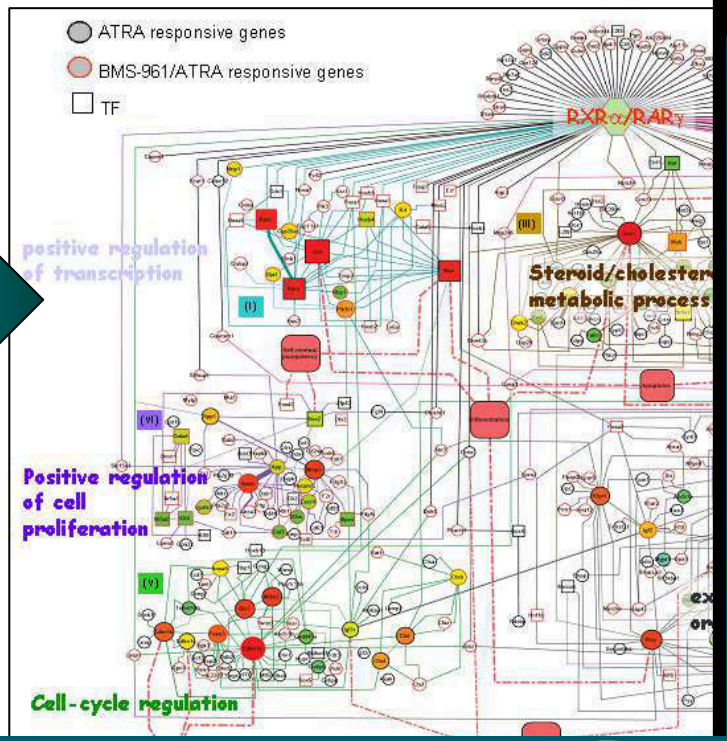
Systems biology of retinoic acid-induced cell fate transitions



Functional Genomics



Systems Biology



igbmc
institut de génétique et de biologie moléculaire et cellulaire

Team leader

Hinrich GRONEMEYER
Tel. : +33 (0)3 88 65 34 73
hinrich.gronemeyer@igbmc.fr

The Team

Researchers
Hinrich GRONEMEYER
Marco Antonio MENDOZA PARRA

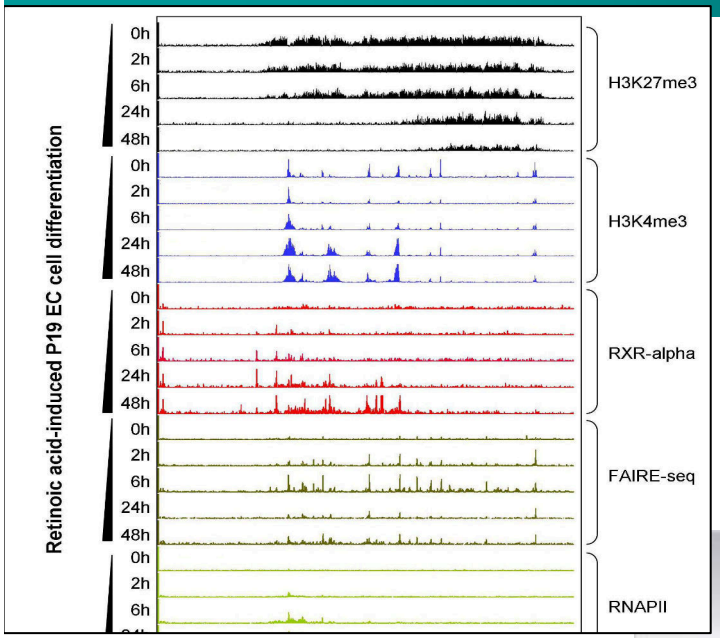
Post-Doctoral Fellowships
Valeria PAVET-PORTAL
Maximiliano PORTAL

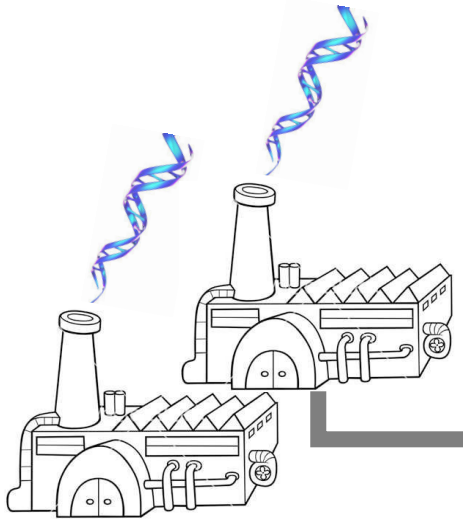
Phd Students
Pierre BORIES
Valeriya MALYSHEVA
Mohamed Ashick MOHAMED SALEEM
Yelyzaveta SHLYAKHTINA

Engineers & Technicians
Benjamin BILLORE
Matthias BLUM
Pierre-Etienne CHOLLEY
Cathie ERB
Michèle LIEB

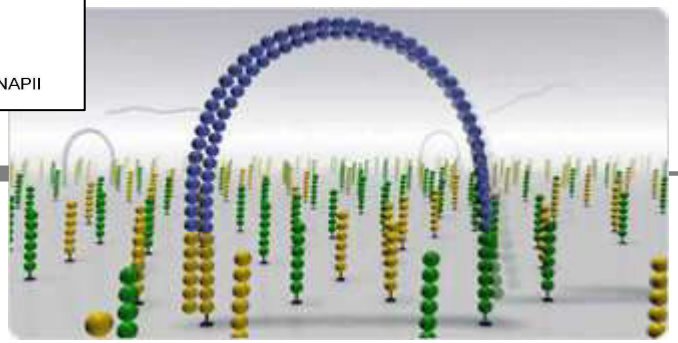
Administration
Irène YUJNOVSKY

Massive parallel DNA sequencing…





Genomics centers



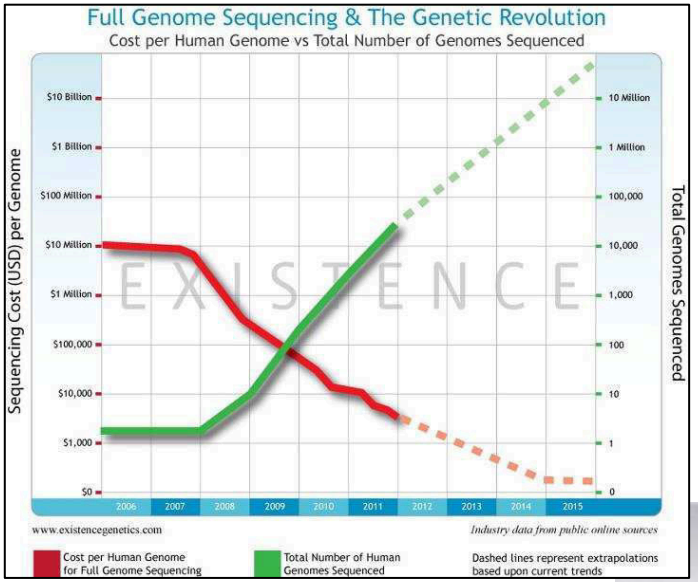
Massive parallel DNA sequencing

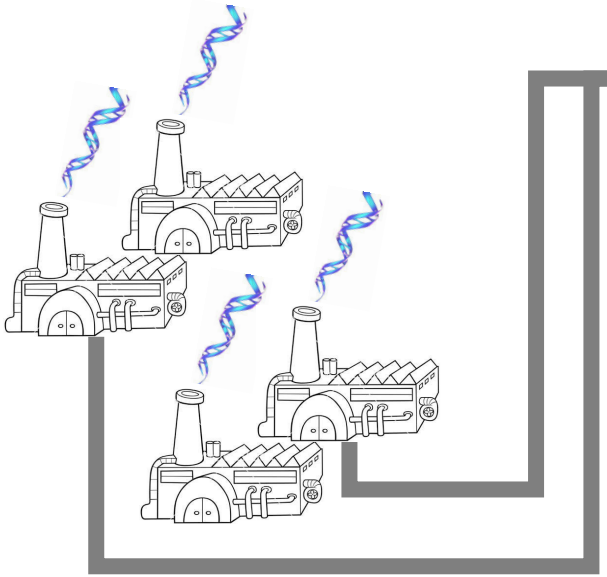
Genomic labs

Genomic labs

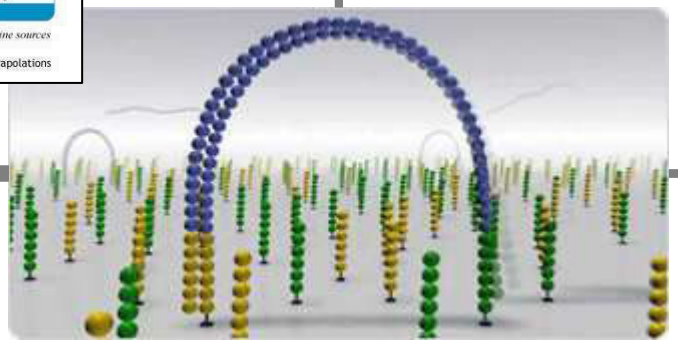
Genomic labs

Massive parallel DNA sequencing…Revolution: The democratisation





Genomics centers



Massive parallel DNA sequencing

Genomic labs

Genomic labs

Genomic labs

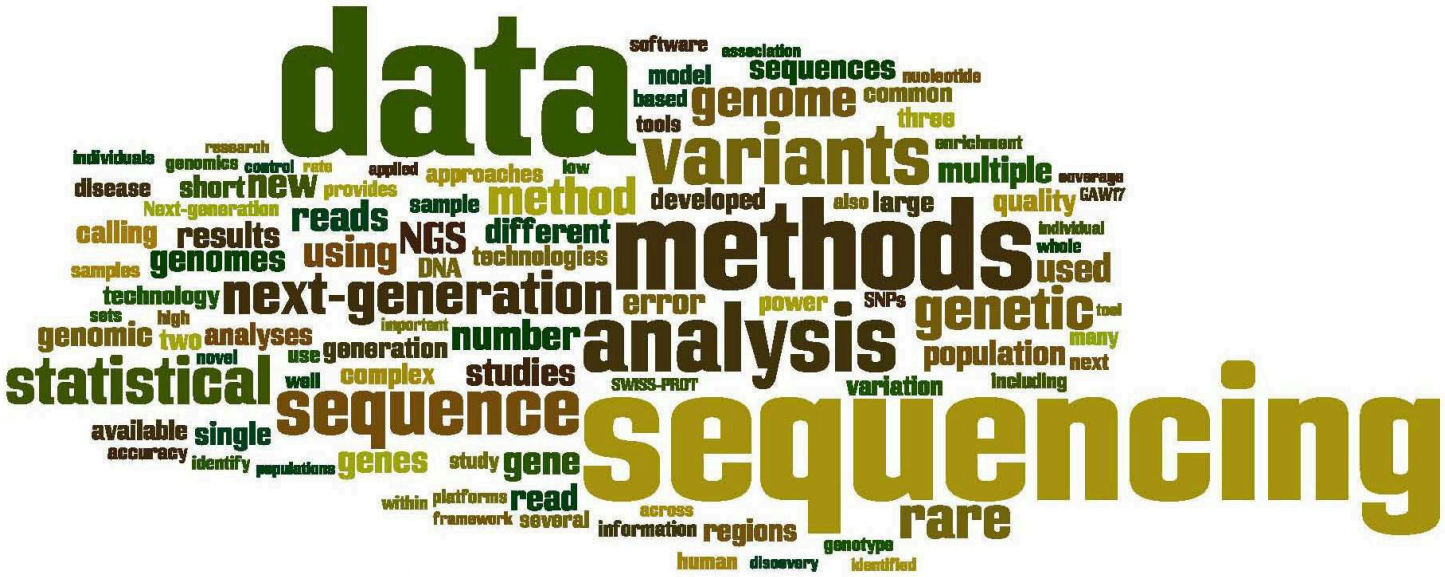
Genomic labs

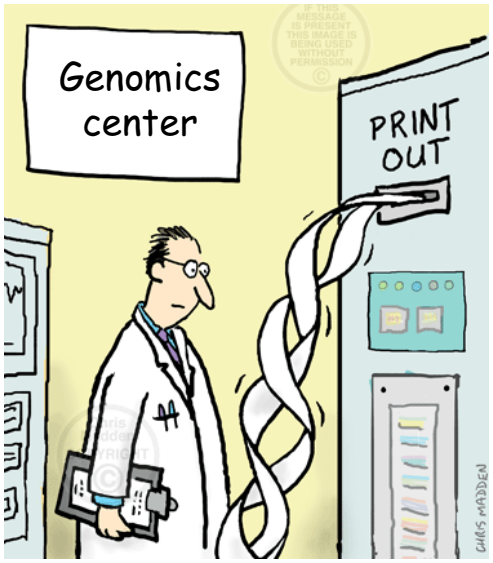
Genomic labs

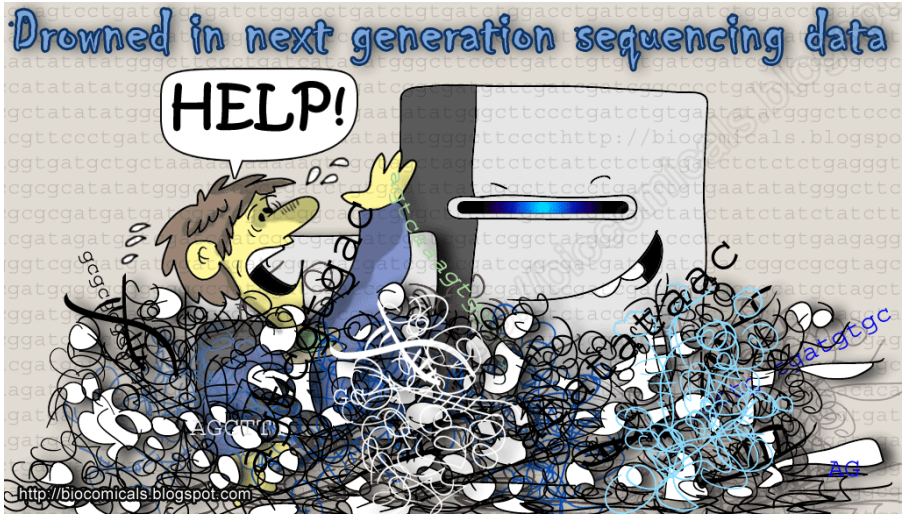
Genomic labs

Genomic labs

Massive parallel DNA sequencing…Revolution: The democratisation

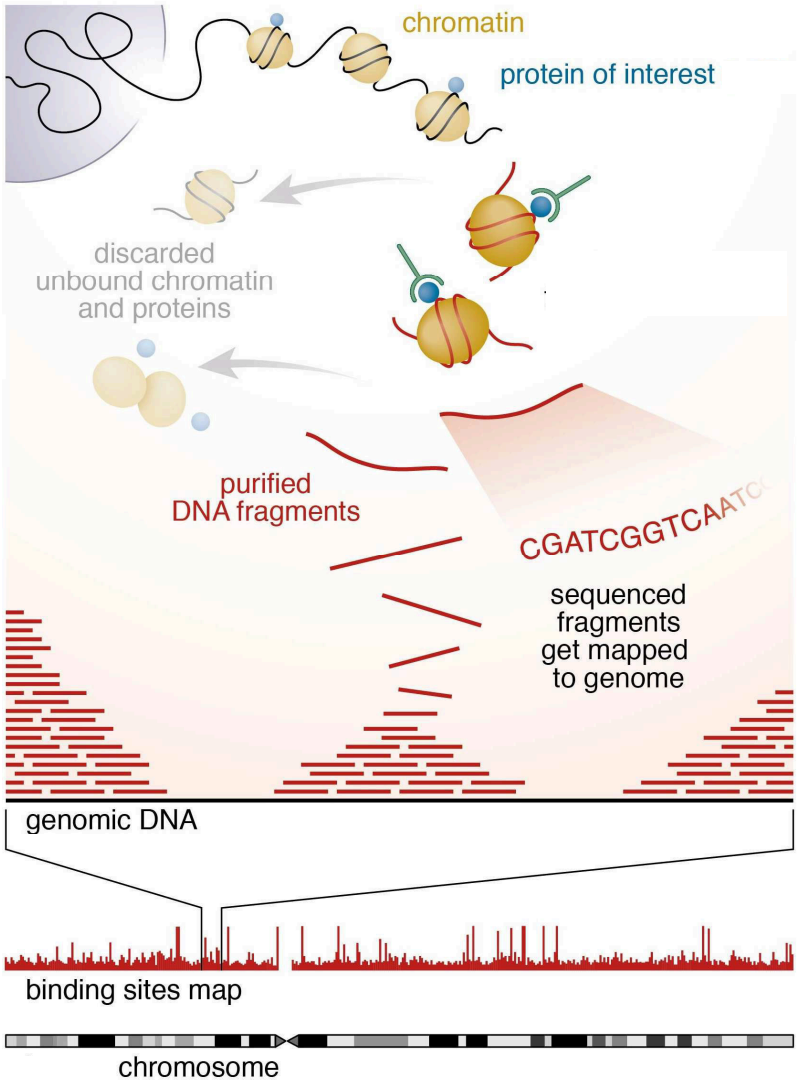






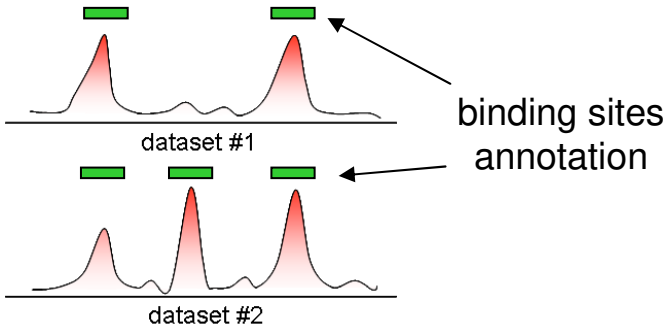
Omics datasets: The analytical path

ChIP-seq: Chromatin immunoprecipitation combined with massive parallel sequencing



ChIP-seq datasets analysis

Extract and compare meaningful biological information

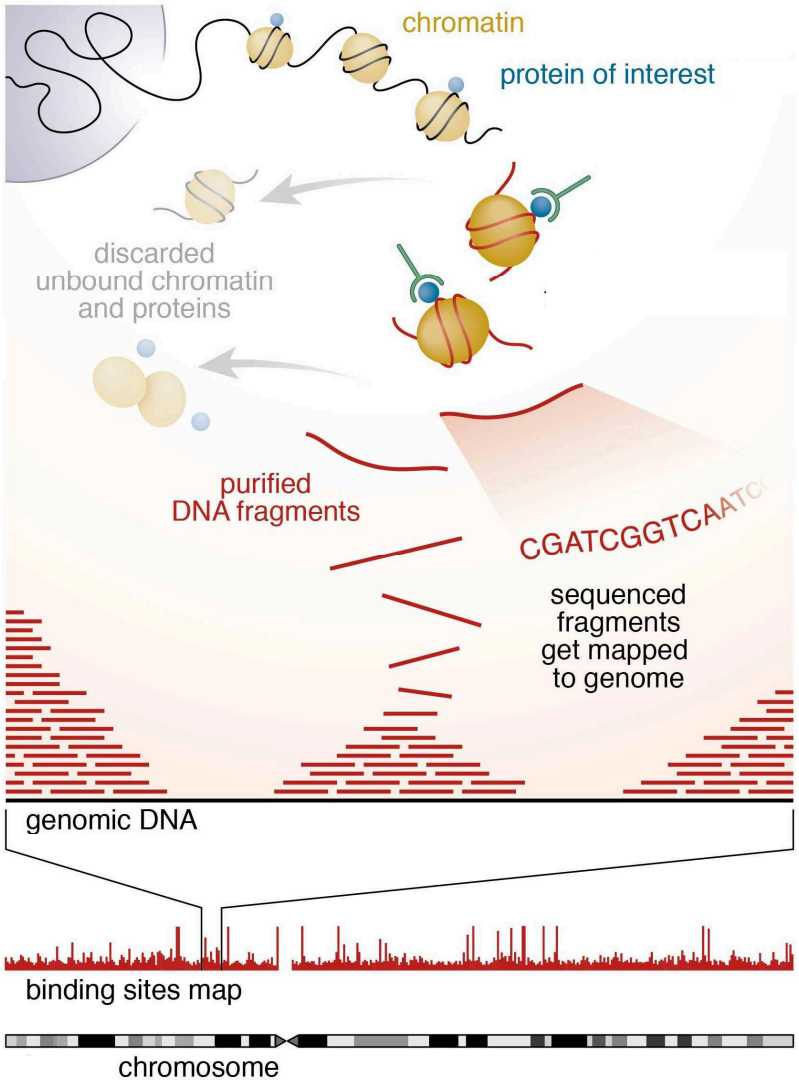


A regression-based method for peak detection in protein-chromatin interaction profiles (ChIP-seq).

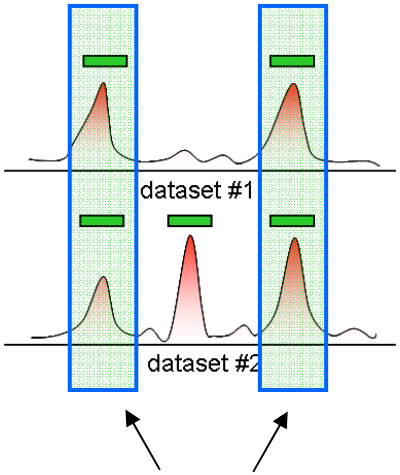
M. A. Mendoza-Parra et al.
BMC Genomics; Nov. 2013

Omics datasets: The analytical path

ChIP-seq: Chromatin immunoprecipitation combined with massive parallel sequencing

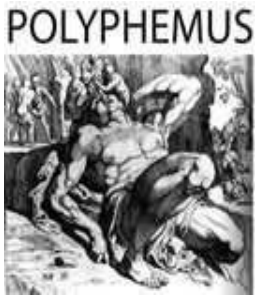


ChIP-seq datasets analysis



Extract and compare meaningful biological information

Compare binding
genomic position b
read-counts intens

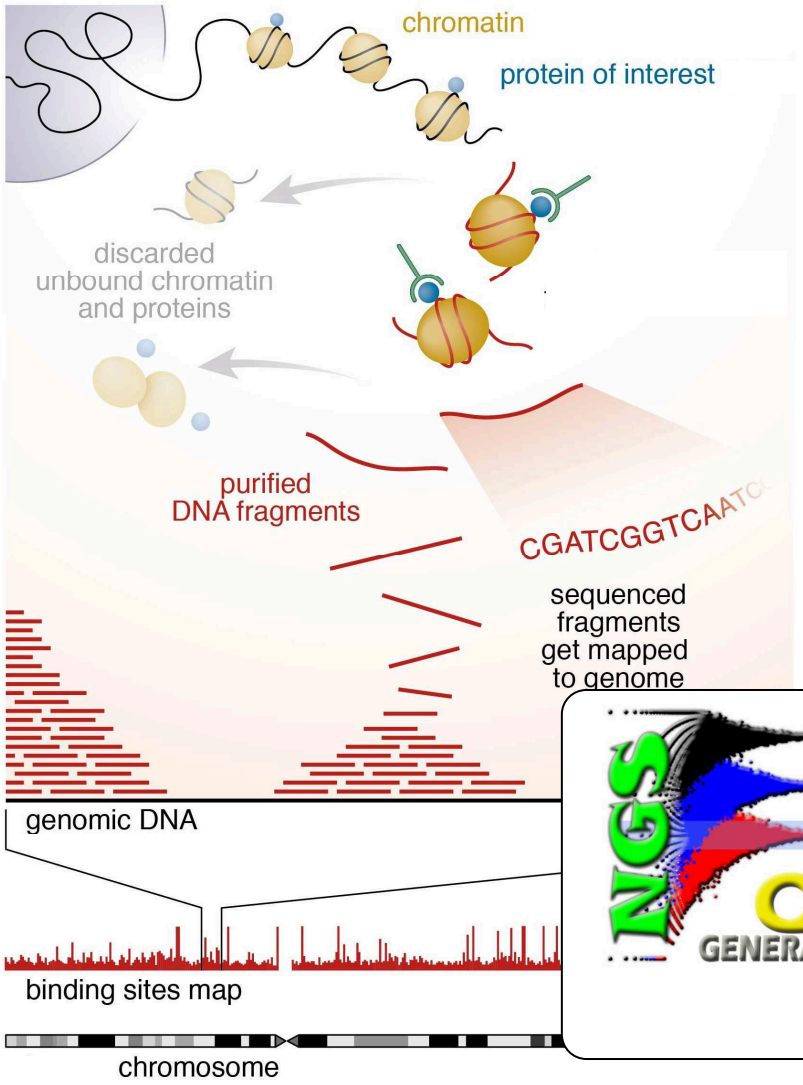


Comparative analysis of RNA Polymerase II ChIP-seq profiles by non-linear normalisation approaches

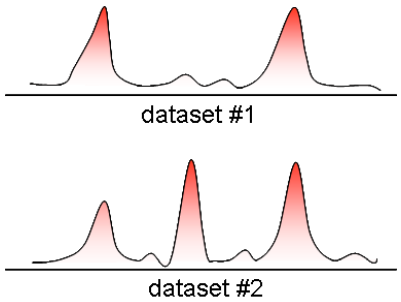
M. A. Mendoza-Parra et al.
Nucl. Acids Res; Nov. 2011

Omics datasets: The analytical path

ChIP-seq: Chromatin immunoprecipitation combined with massive parallel sequencing



ChIP-seq datasets analysis



Extract and compare meaningful biological information

evaluate the "quality" of compared datasets which may impact the assessment of the relevant biological information

QC?

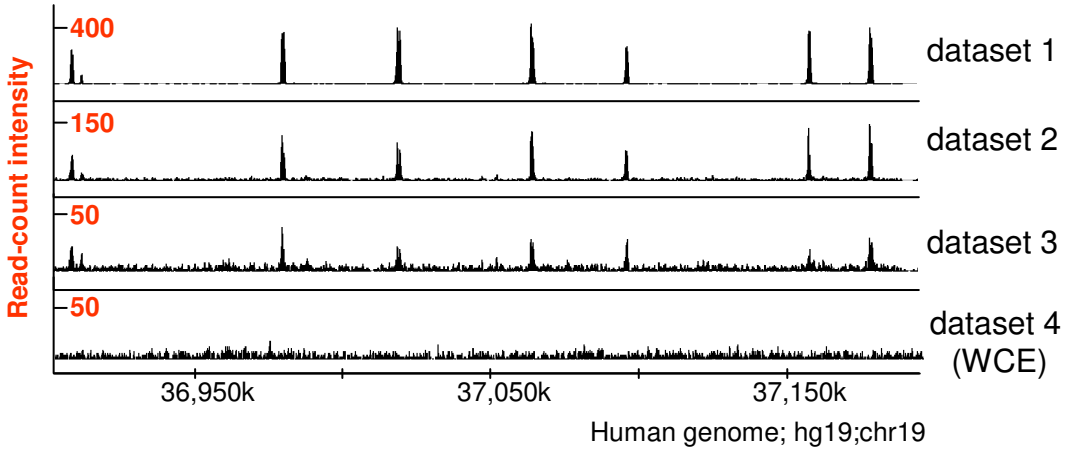
a bioinformatics-based quality control (QC) system that uses raw NGS dataset information to have quantitative means for comparing different ChIP-seq profiles.

A quality control system for profiles obtained by ChIP sequencing

M. A. Mendoza-Parra et al. ***Nucl. Acids Res; Nov. 2013***

Are you concerned about these differences in ChIP-seq assays performance?

Profiles of three publicly available H3K4me3 ChIP-seq datasets.

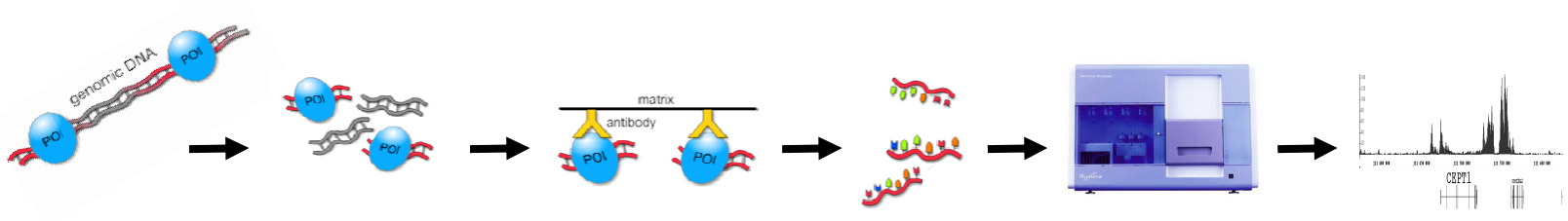


- Common binding events are retrieved,
- However there are significant differences :
  - Global read-count intensity
  - Observed background levels

- With the **same commercial antibody** source (including batch reference).
- Produced from a **similar number of mapped reads** (~19 million).
- Performed in **different laboratories**.

Source of technical variability in immuno-selection-NSG based approaches

ChIP-seq workflow

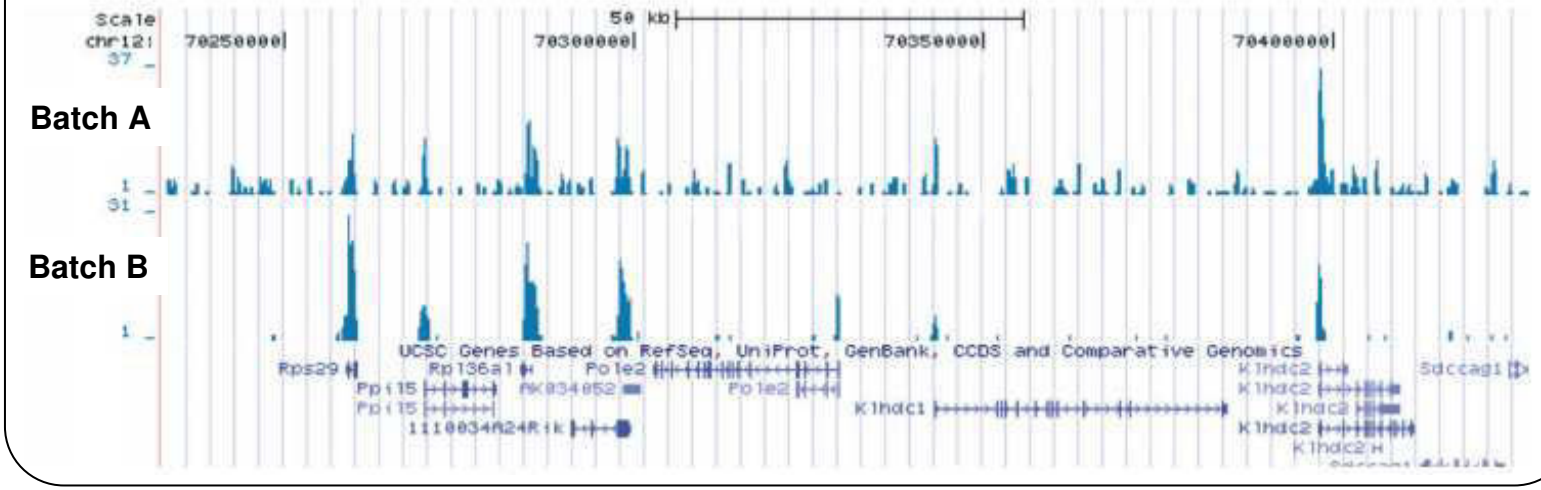


Protein-chromatin crosslinking → *Chromatin fragmentation* → *immunoprecipitation* → *DNA purification* → *Parallel sequencing* → *Computational analysis*

ChIP-seq grade Abs?

Quality ChIP-seq profile=f(chrom shearing + cell fixation + user skills + Ab efficacy + Seq depth + …)

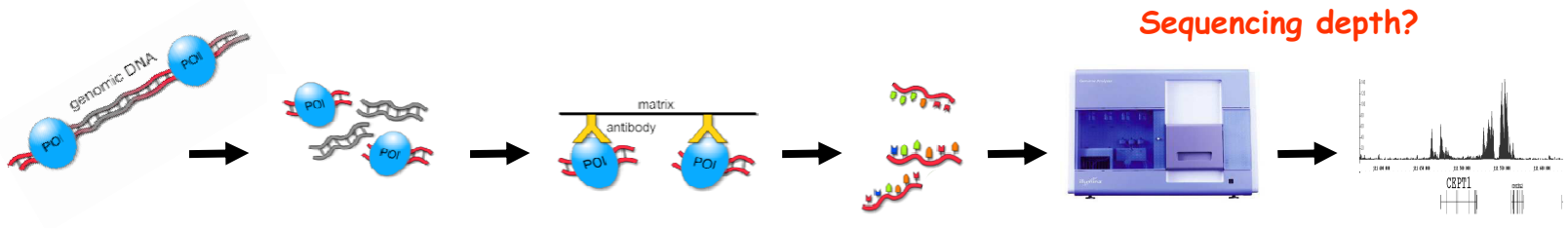
Ab against H3K4me3



Source: Diagenode H3K4me3 antibody; technical data sheet

Source of technical variability in immuno-selection-NSG based approaches

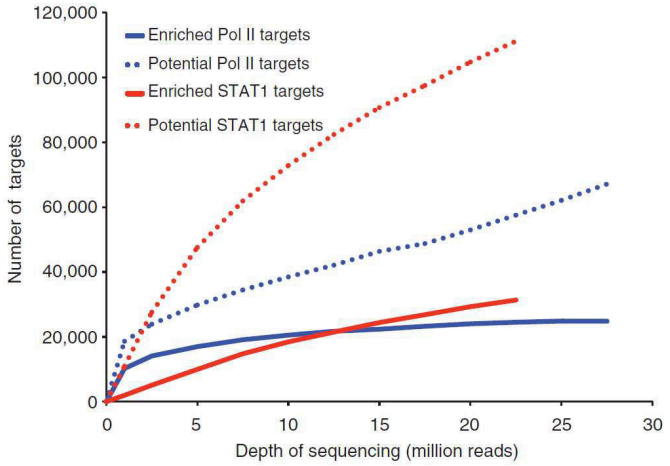
ChIP-seq workflow



Protein-chromatin crosslinking — *Chromatin fragmentation* — *immunoprecipitation* — *DNA purification* — *Parallel sequencing* — *Computational analysis*

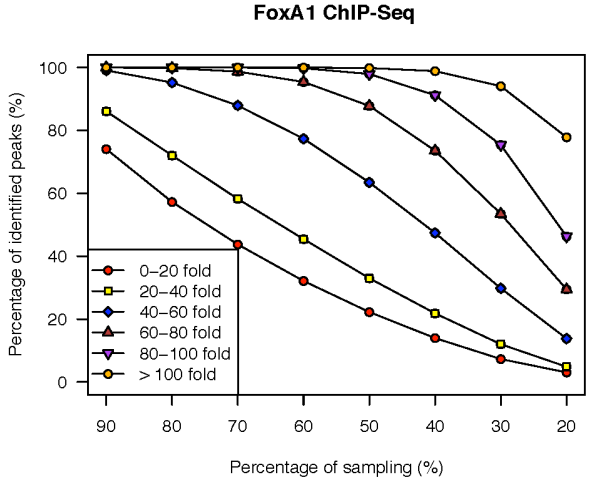
Quality ChIP-seq profile=f(chrom shearing + cell fixation + user skills + Ab efficacy + Seq depth + …)

number of identified peaks increases with the sequencing depth



Rozowsky J. et al. Nature Biotechnol 2009

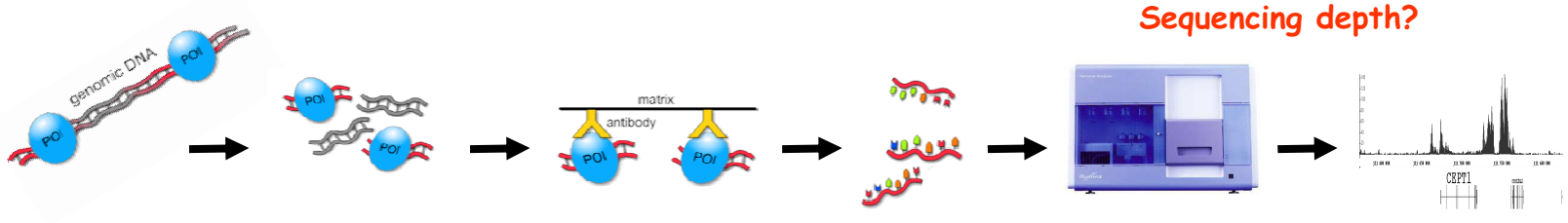
Peaks confidence is proportional to the sequencing depth



Zhang Y. et al. Genome Biol 2008

Source of technical variability in immuno-selection-NSG based approaches

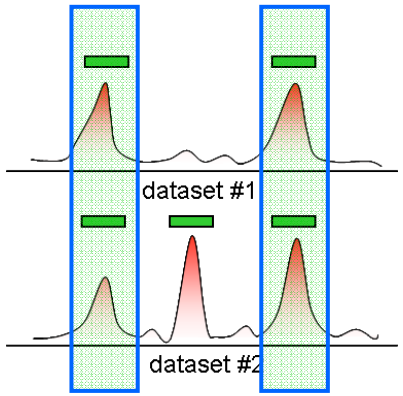
ChIP-seq workflow



Protein-chromatin crosslinking — *Chromatin fragmentation* — *immunoprecipitation* — *DNA purification* — *Parallel sequencing* — *Computational analysis*

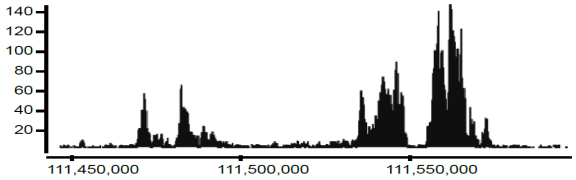
ChIP-seq grade Abs?

Quality ChIP-seq profile=f(chrom shearing + cell fixation + user skills + Ab efficacy + Seq depth + …)



as individual profiles could present significant technical variability (due to the use of different antibodies, sequencing depth or other factors able to influence the immunoprecipitation (IP) quality), comparative analyses between Next generation sequencing (NGS) generated profiles may require prior characterization of the degree of technical similarity of the various data sets

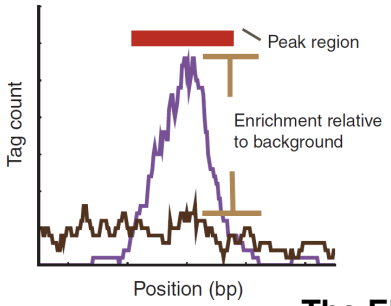
Current practices for the evaluation of the quality of ChIP-seq datasets:



Visual inspection

Drawback:
- only in a local context
- not quantitative

Peak calling-based approaches



global ChIP enrichment evaluated by the **Fraction of mapped reads in identified Peaks (FRiP)**

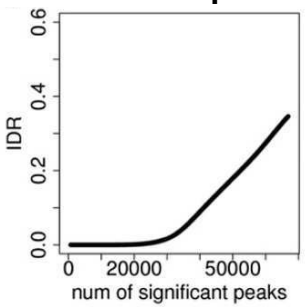
The ENCODE consortium recommends a 1% FRiP threshold

Quality Control systems

Evaluate the **binding sites reproductibility between biological replicate** datasets (IDR)

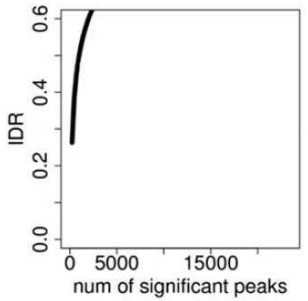
IDR: Irreproductibility Discovery Rate

RAD21 replicates



high reproductibility

SPT20 replicates



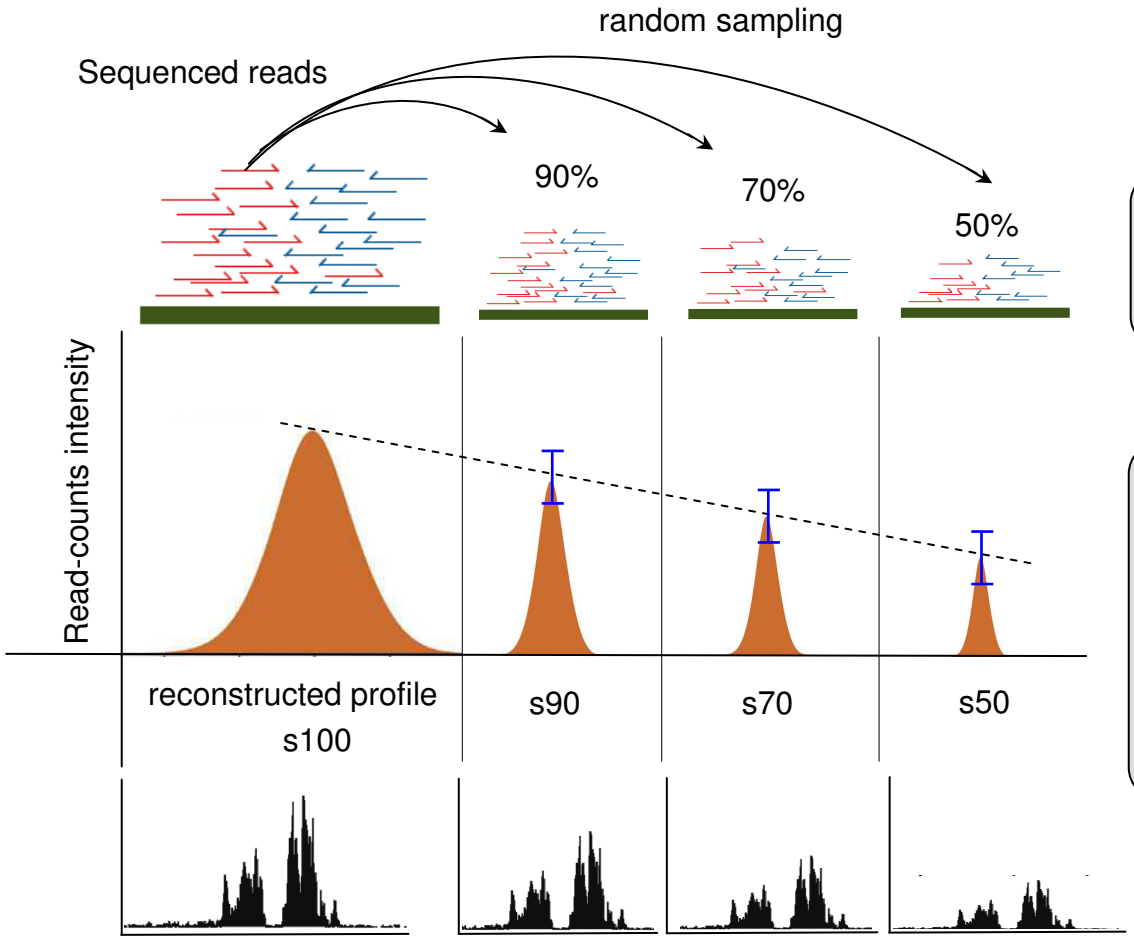
low reproductibility

Drawback:
they **depend on the peak calling performance** thus their quality assessement may **not** represent **a universal approach** for Comparing any type of ChIP-seq datasets

Landt S. et al. Genome Research 2012

A quality control (QC) system for the analysis and comparison of NGS-generated profiles

Aim of the methodology:
Evaluate changes in enrichment patterns when a fraction of the total mapped reads (TMRs) are used for its reconstruction.



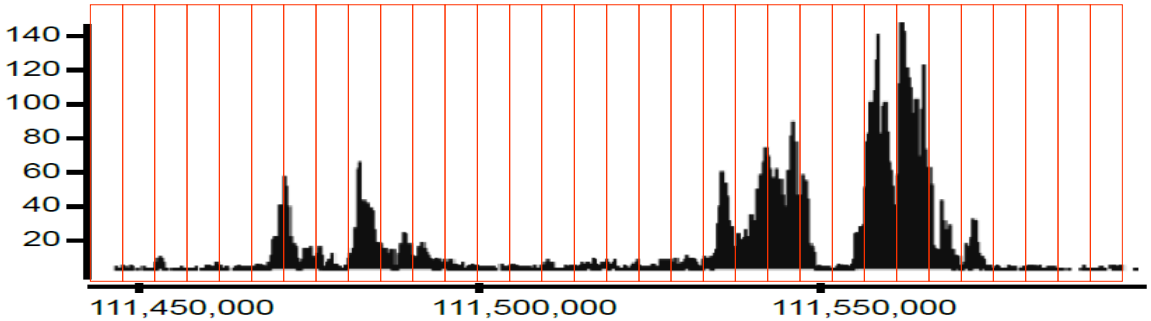
In an ideal case the reconstructed profile will be reproduced after random sampling with a proportional decrease in the read-count intensity

- the deviation from the ideal behavior reflects the degree of reproducibility/robustness of the profile under study
- It provides a quantitative method for assigning quality descriptors

A quality control (QC) system for the analysis and comparison of NGS-generated profiles

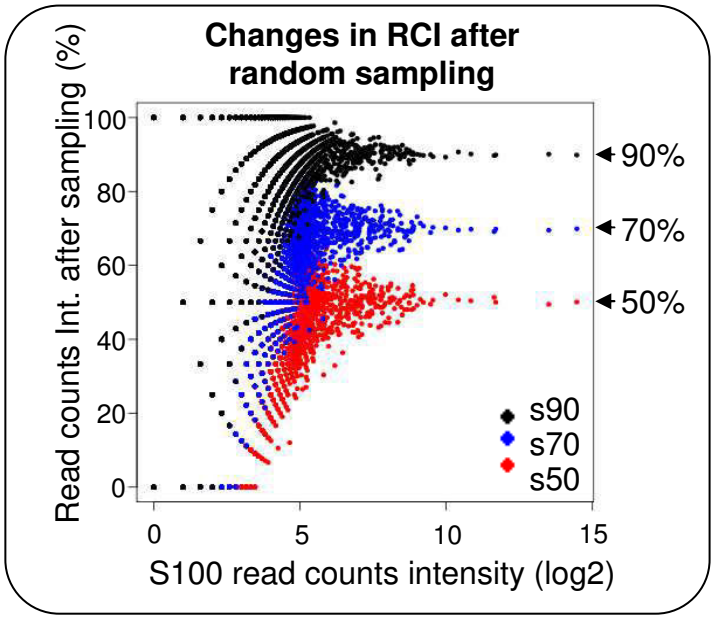
Computational treatment:

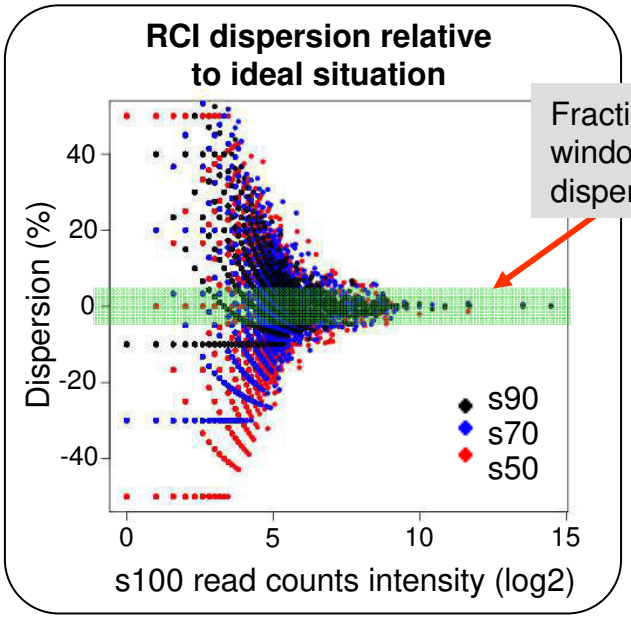
Segmented read-count intensity profile (500nts window size)



RCI: read count intensity
S100: Original dataset
S90: 90% reads sampled
S70: 70% reads sampled
S50: 50% reads sampled

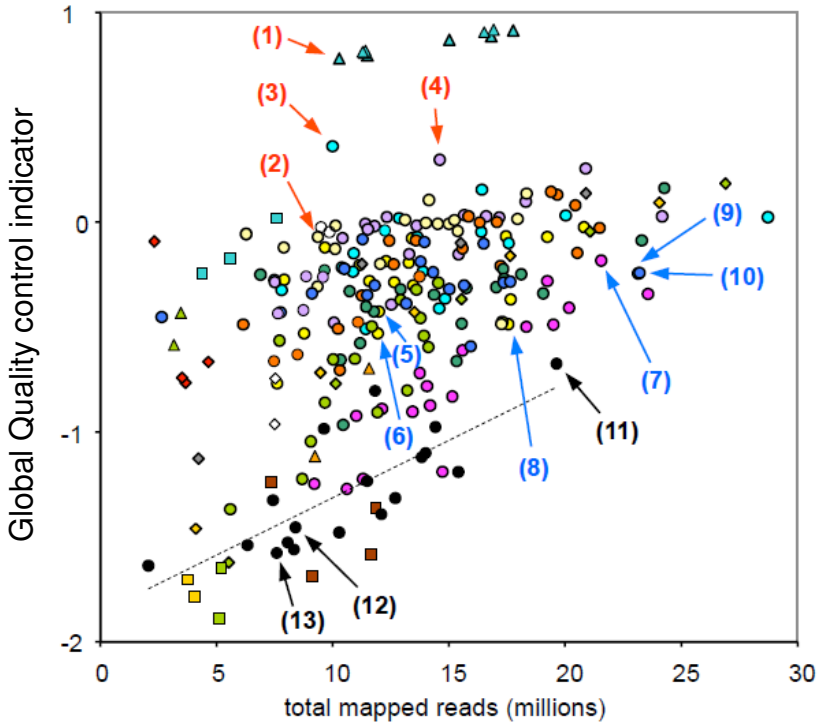
Analysis on ChIP-seq segments (500bp bins)

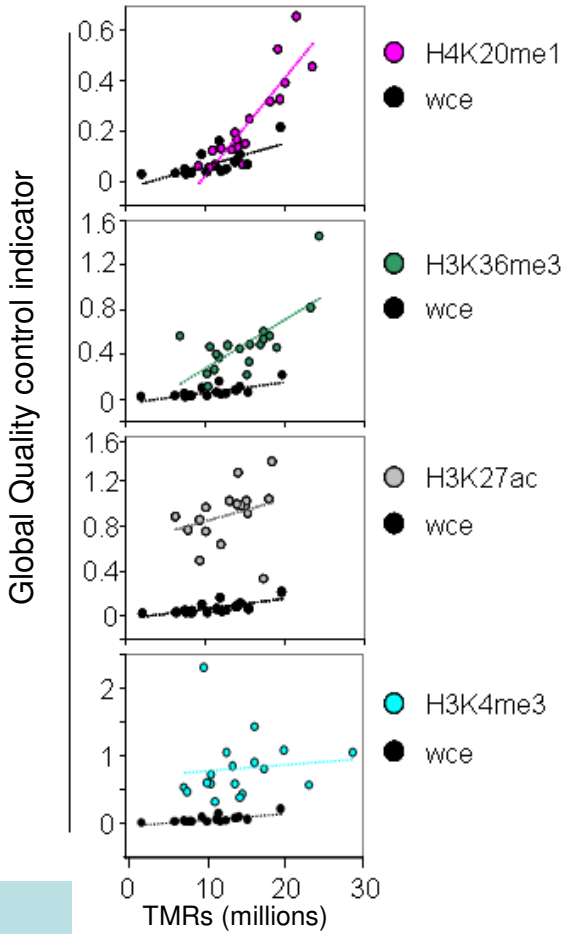




Fraction of genomic windows in a given dispersion interval

A universal NGS-QCi system for the stratification of current and past generated profiles

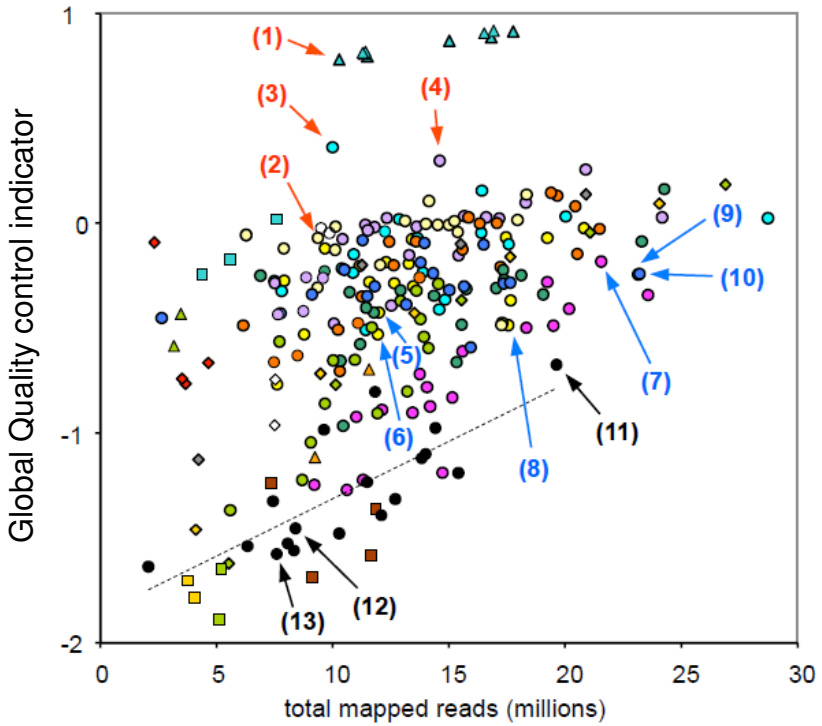


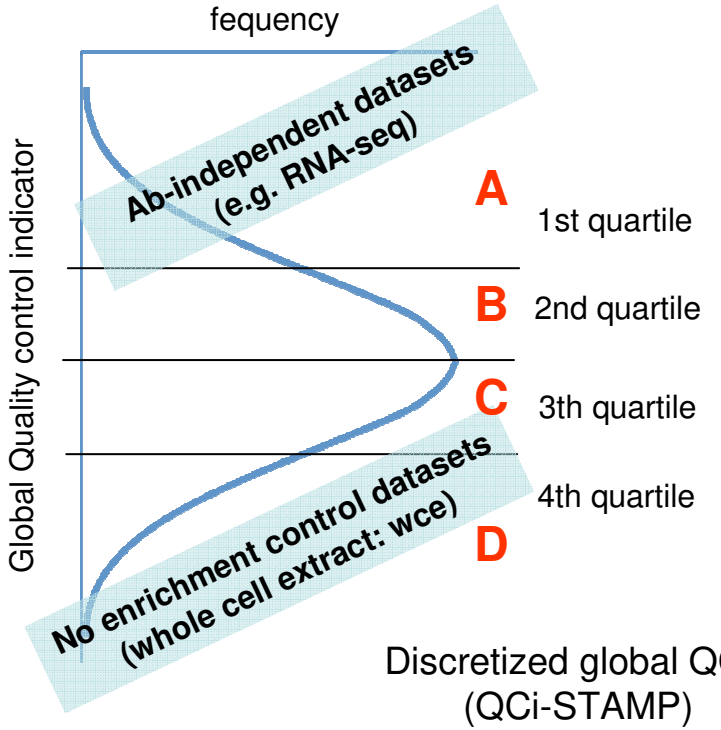


the QCi analysis provides important quality information about datasets for the same target at different sequencing depths

A universal NGS-QCi system for the stratification of current and past generated profiles

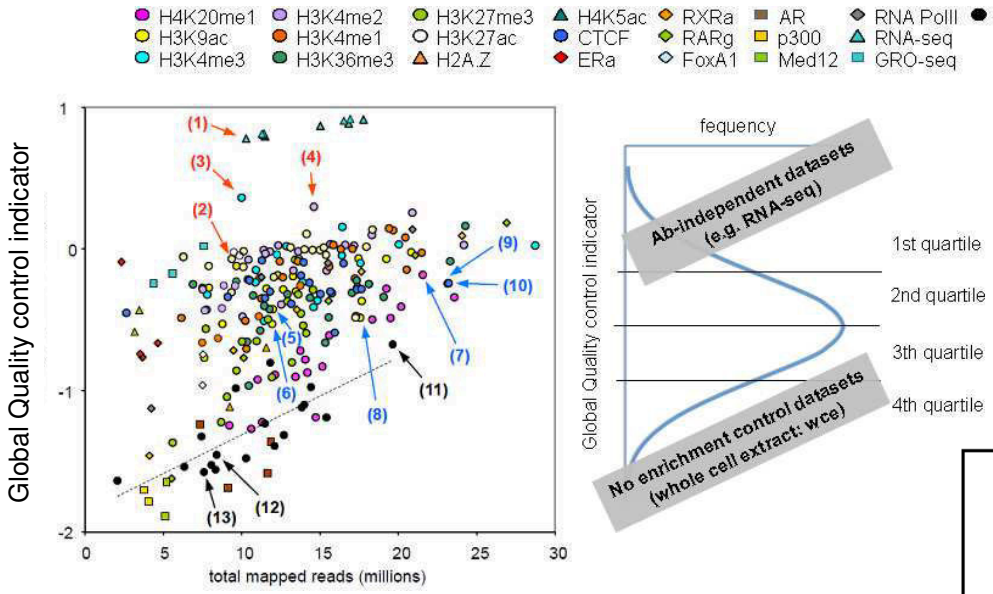
H4K20me1 H3K4me2 H3K27me3 H4K5ac RXRa AR RNA PolII wce
H3K9ac H3K4me1 H3K27ac CTCF RARg p300 RNA-seq
H3K4me3 H3K36me3 H2A.Z ERa FoxA1 Med12 GRO-seq

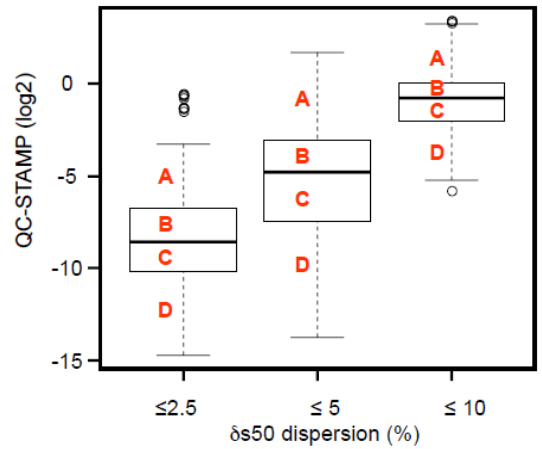




Discretized global QCis
(QCi-STAMP)

A universal NGS-QCi system for the stratification of current and past generated profiles



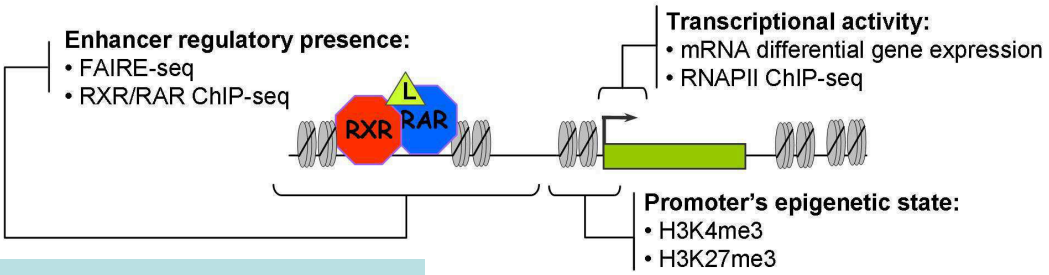


QCis assessed at three different confidence threshold conditions

Target ID	TMRs	QC-STAMP (δs50< 2.5; 5; 10)
(1) RNA-seq	10,285,026	AAA
(2) H3K9ac	9,675,720	AAA
(3) H3K4me3	10,007,440	AAA
(4) H3K4me2	14,615,772	AAA
(5) H3K9ac	11,991,828	BBC
(6) H3K9ac	11,950,657	BCC
(7) H4K20me1	21,552,590	BBB
(8) H3K9ac	17,560,679	BCC
(9) H3K4me3	23,146,665	BCB
(10) CTCF	23,211,599	BBB
(11) WCE	19,637,643	CDD
(12) WCE	8,410,101	DDD
(13) WCE	7,599,303	DDD

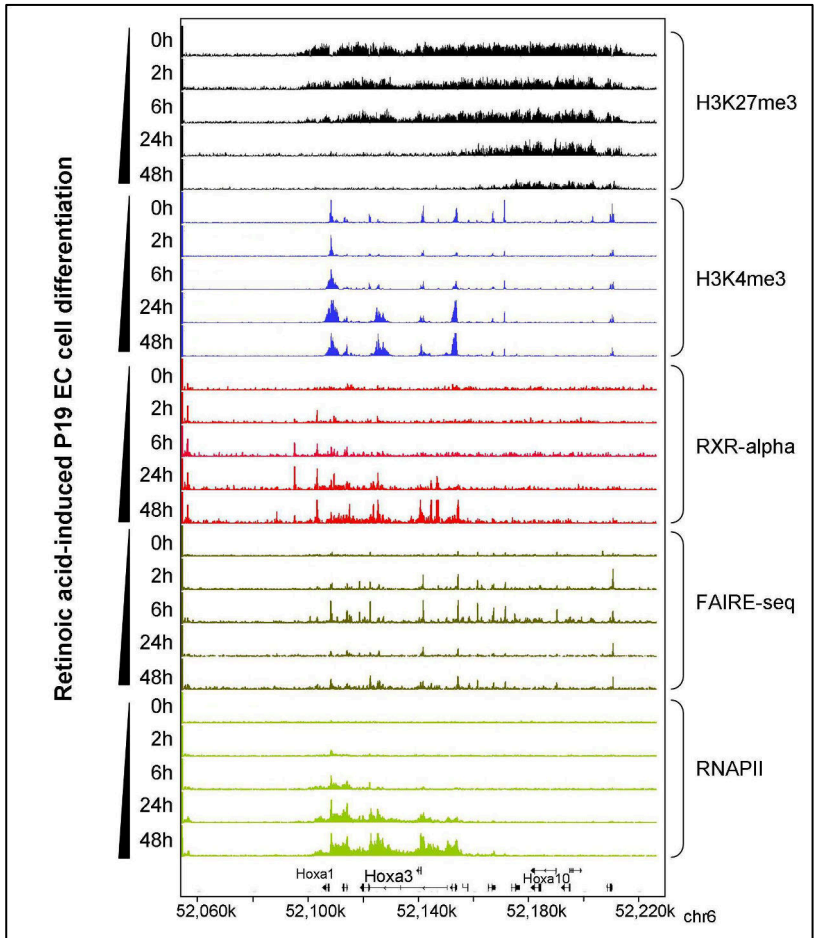
Using the NGS-QC certification system for validating ChIP-seq profiles prior further analysis

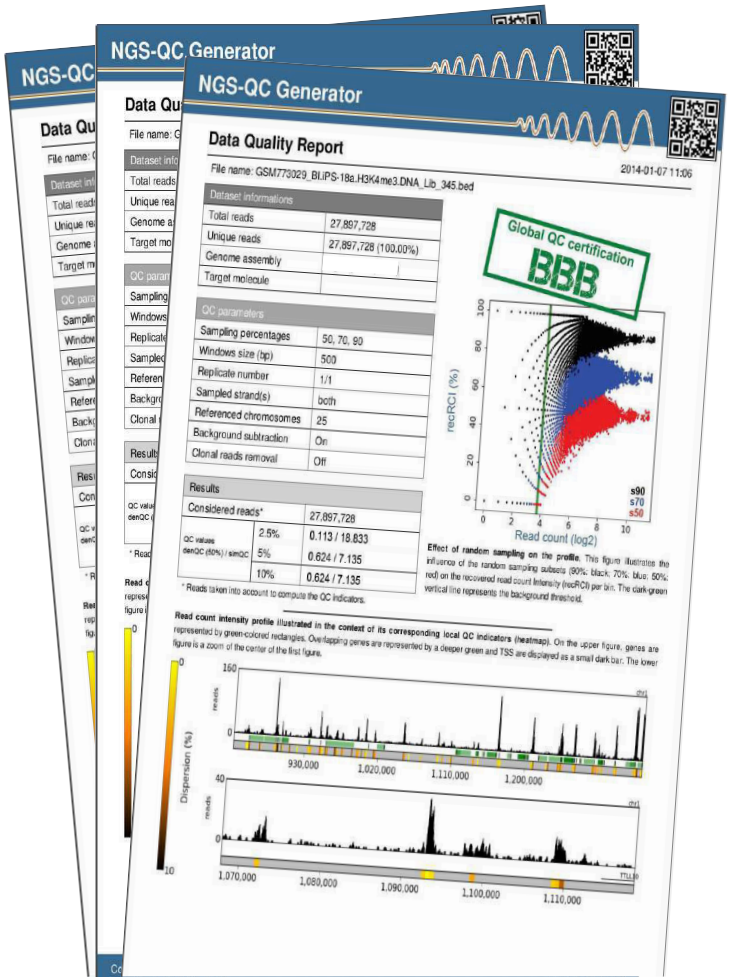
A



Omics datasets integration

Compared profiles present at least a « BBB » certification !!!!



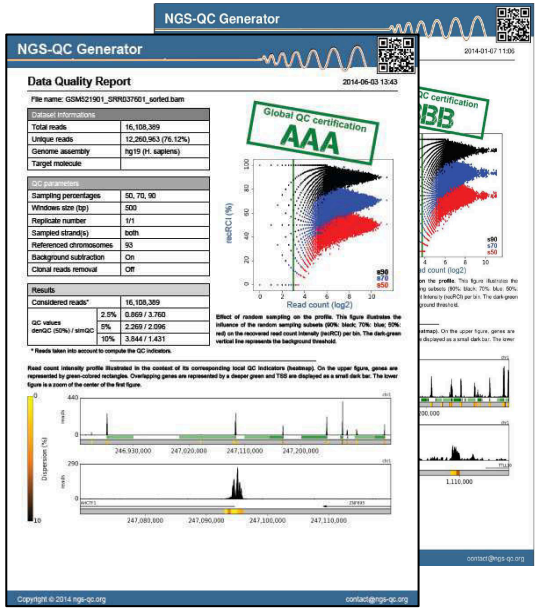


A universal NGS-QCi database for comparing current and past generated profile's indicators

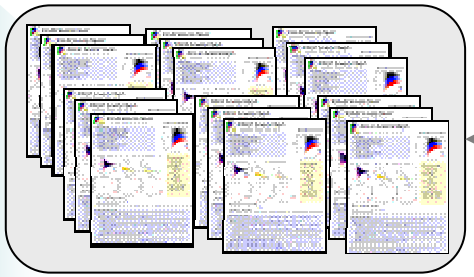
NGS-datasets repositories

Automated certification pipeline

Datasets Quality?

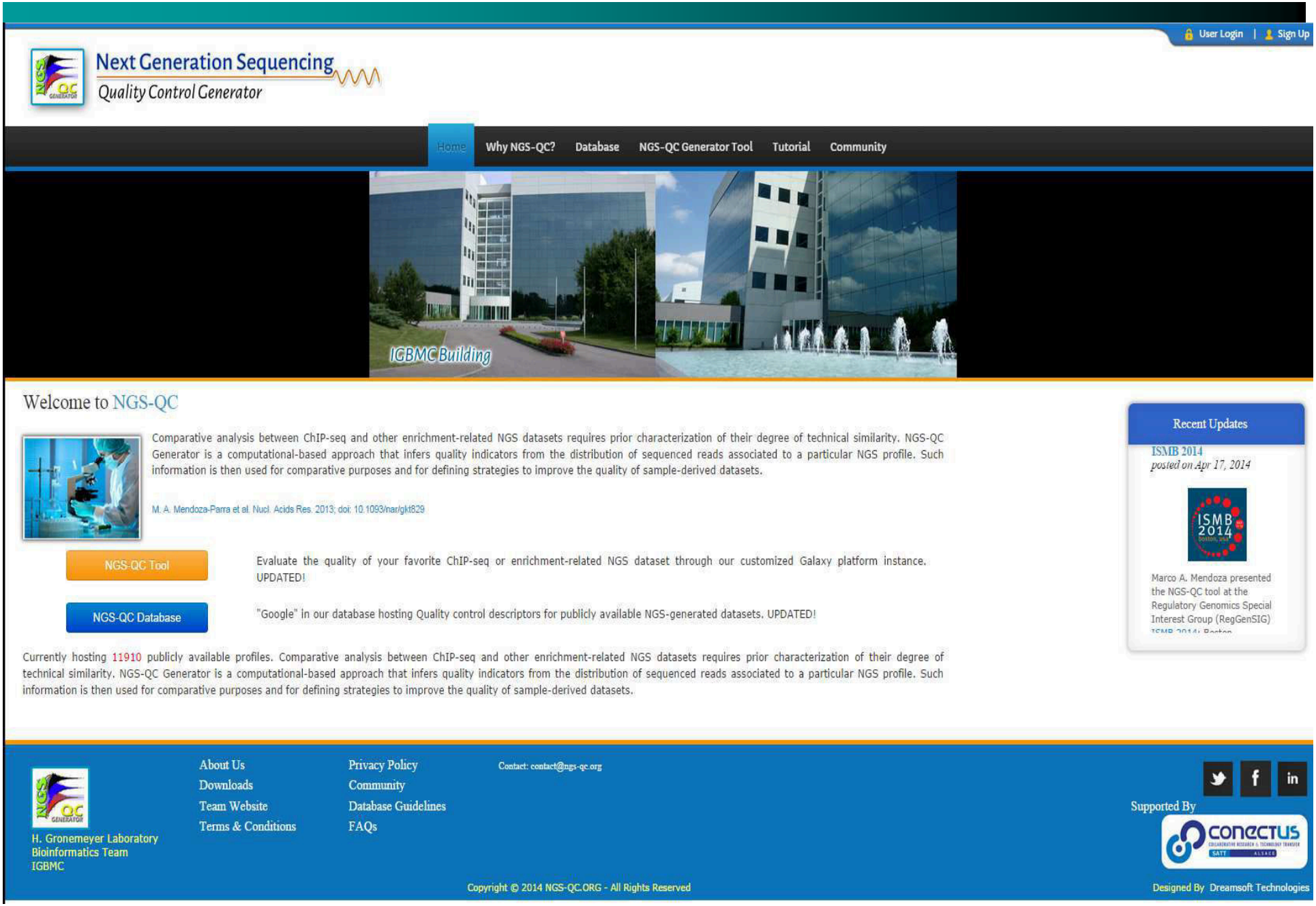


NGS-QCi database



www.ngs-qc.org

(currently hosting > 11000 entries)

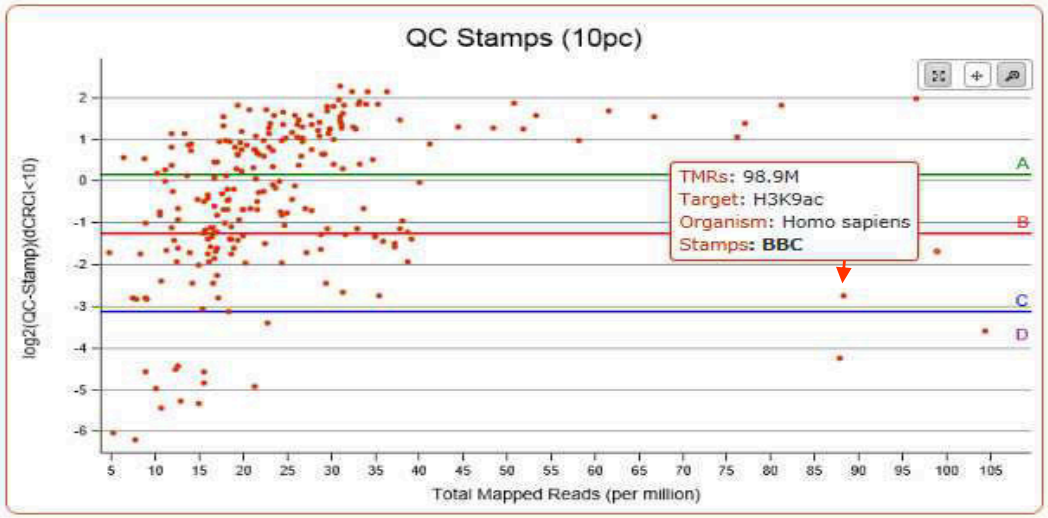


www.ngs-qc.org

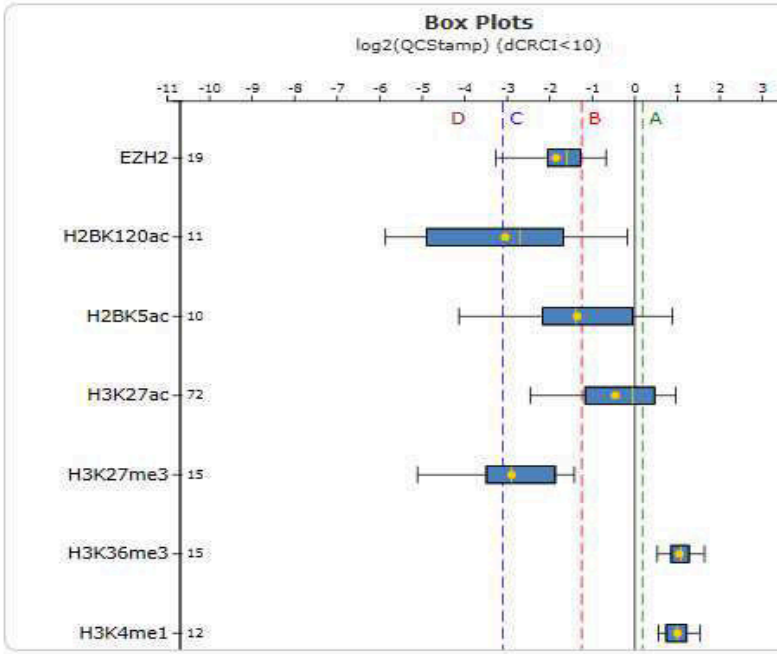
Next Generation Sequencing
Quality Control Generator

Home | Why NGS-QC? | Database | NGS-QC Generator Tool | Tutorial | Community

QC Stamp 2.5pc | QC Stamp 5pc | QC Stamp 10pc



Clear Zoom | Update Results for the selected region



Hide Panel

Refine your results:

Clear Panel

DataType
Cell/Tissue
Target Molecule
Antibody Vendor
Antibody Reference
Antibody Lot
Treatment

Public ID	Study ID	PMID	Organism	Data type	Target Molecule	TMRs	QC Stamps	QC Report	Local QC indicators	Global QC indicators					
										2.5pc denQC s50	2.5pc simQC s50/s90	5pc denQC s50	5pc simQC s50/s90	10pc denQC s50	10pc simQC s50/s90
GSM802399	GSE32340	22499665	Homo sapiens	ChIP-seq	H3K18ac	46270774	BBB			0.108	28.238	0.568	10.977	2.604	3.277
GSM802400	GSE32340	22499665	Homo sapiens	ChIP-seq	H3K18ac	40619778	AAA			0.700	16.321	4.018	6.022	15.315	2.168
GSM956007	GSE16256	19829295, 20452322, 20944595, 21289626	Homo sapiens	ChIP-seq	H3K18ac	38846564	AAA			0.358	19.252	2.119	7.457	8.795	2.472
GSM956006	GSE16256	19829295, 20452322,	Homo	ChIP-seq	H3K18ac	21874494	CBB			0.092	63.247	0.777	12.336	3.322	3.871

www.ngs-qc.org

A web-browser access for running the NGS-QC Generator on your favourite NGS dataset!!!

www.ngs-qc.org

Next Generation Sequencing
Quality Control Generator

Home | Why NGS-QC? | Database | NGS-QC Gen

Galaxy / NGS-QC — Analyze Data | Workf

Tools — Load Data

search tools

Get Data

NGS-QC Generator

NGS-QC Generator - A quality control system for profiles obtained by ChIP sequencing

Local QCis Generator displays regions from an existing dataset

Replicates Generator allows to general several random replicates of a BED file.

Text Manipulation

Operate on Genomic Intervals

Workflows

- All workflows

NGS-QC Generator (version 1.5.7)

Input file:

151: (as bed) Join on data 150 and data 149

Genome:

Human Feb. 2009 (GRCh37/hg19)

Number of replicates:

(•) One
( ) Two
( ) Three

Exclude background (noise) reads:

[x] Uncheck this option if you want to disable the background subtraction.

Exclude clonal (PCR Duplicates) reads:

[ ] Use this option if you want to remove clonal reads.

Resolution (bp):

500,000

Used for local regions displayed within the report

Execute

What it does

Comparative analysis between Next Generation Sequencing (NGS) generated
technical similarity.

NGS-Q/C Generator infers global and local quality indicators based on a stand
evaluates the distribution of the sequencing reads associated to a particular N

Specifically, the total and uniquely aligned mapped reads (TMRs) are randoml
respectively). Then, the genome-wide distribution of the sampled reads popul

Further information and a tutorial how to analyze sequencing profiles with NG

Running time:

It takes NGS-QC Generator around 15 minutes to analyze profiles smaller tha

NGS-QC Generator

Data Quality Report

2014-01-07 11:06

File name: GSM773029_BI.iPS-18a.H3K4me3.DNA_Lib_345.bed

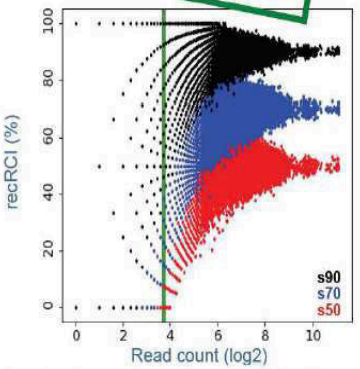
Dataset informations	
Total reads	27,897,728
Unique reads	27,897,728 (100.00%)
Genome assembly	hg19 (H. sapiens)
Target molecule	

QC parameters	
Sampling percentages	50, 70, 90
Windows size (bp)	500
Replicate number	1/1
Sampled strand(s)	both
Referenced chromosomes	25
Background subtraction	On
Clonal reads removal	Off

Results		
Considered reads*		27,897,728
QC values denQC (50%) / simQC	2.5%	0.113 / 18.833
	5%	0.624 / 7.135
	10%	0.624 / 7.135

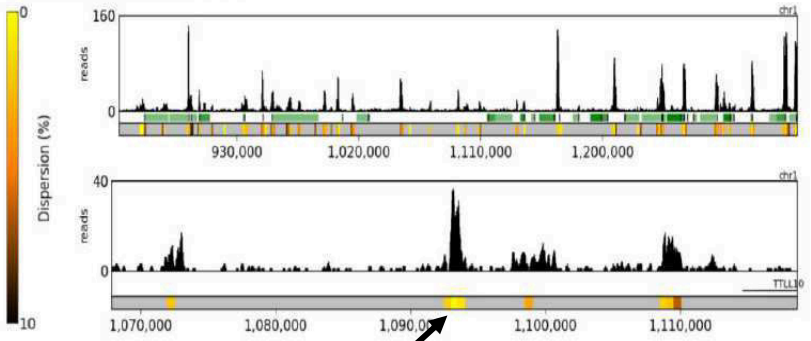
* Reads taken into account to compute the QC indicators.

Global QC certification
BBB



Effect of random sampling on the profile. This figure illustrates the influence of the random sampling subsets (90%: black; 70%: blue; 50%: red) on the recovered read count Intensity (recRCI) per bin. The dark-green vertical line represents the background threshold.

Read count intensity profile illustrated in the context of its corresponding local QC indicators (heatmap). On the upper figure, genes are represented by green-colored rectangles. Overlapping genes are represented by a deeper green and TSS are displayed as a small dark bar. The lower figure is a zoom of the center of the first figure.



Local QC indicator

contact@ngs-qc.org

www.ngs-qc.org

Next Generation Sequencing
Quality Control Generator

Home | Why NGS-QC? | Database | NGS-

Galaxy / NGS-QC

Tools | Load Data

search tools

Get Data

NGS-QC Generator

NGS-QC Generator - A quality control system for profiles obtained by ChIP sequencing

Local QCis Generator displays regions from an existing dataset

Replicates Generator allows to general several random replicates of a BED file.

Text Manipulation

Operate on Genomic Intervals

Workflows

- All workflows

Local QCis Generator (version 1.0)

Dataset: 139: P19_H3K4me3_T48.bed

Local QC file: 169: Loc. densities (rep. 1) for data 139

Genome: Human Feb. 2009 (GRCh37/hg19)

Resolution (bp): 500,000

Number of local QCis: 10

Change this value if you want to get more figures

Exclude clonal (PCR Duplicates) reads:

Use this option if you want to remove clonal reads.

Execute

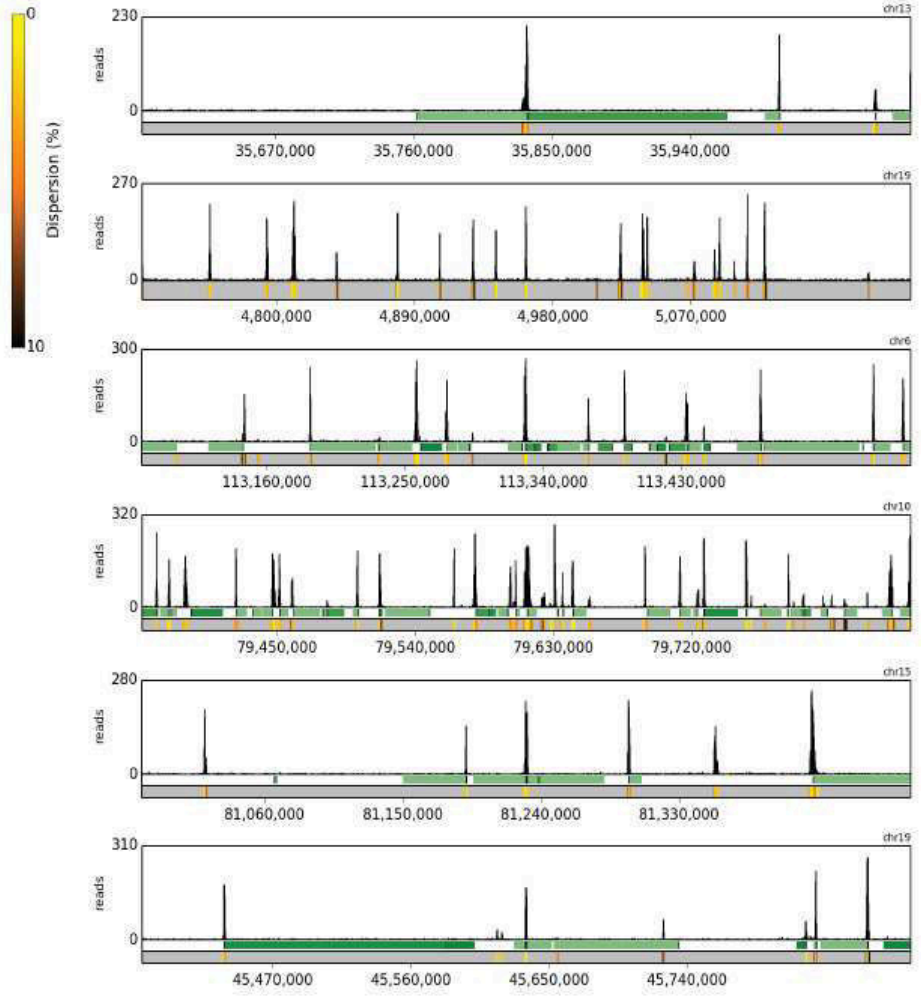
What it does

This tool generates local QC indicators from a bed file

Options

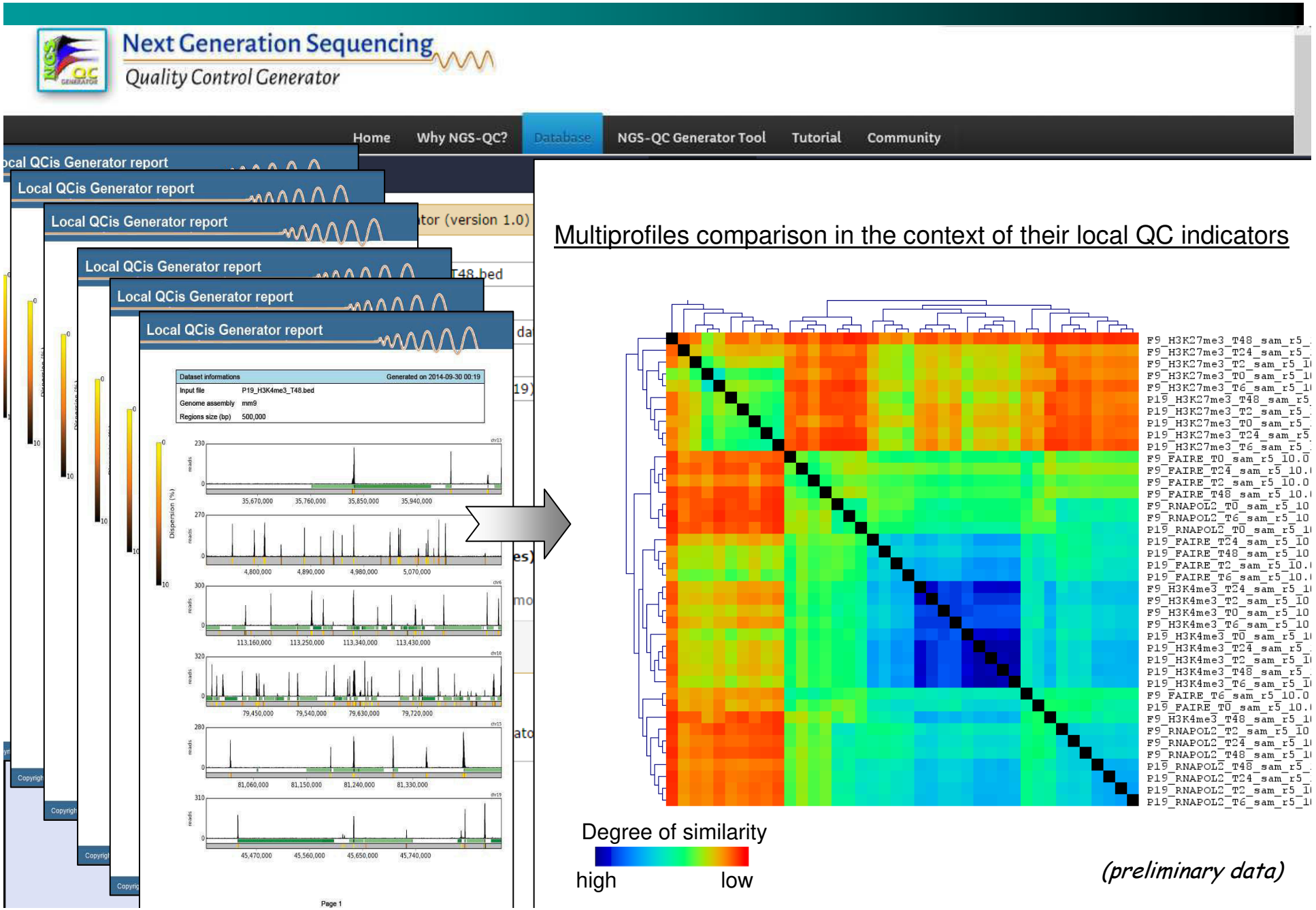
Local QCis Generator report

Dataset informations	Generated on 2014-09-30 00:19
Input file	P19_H3K4me3_T48.bed
Genome assembly	mm9
Regions size (bp)	500,000

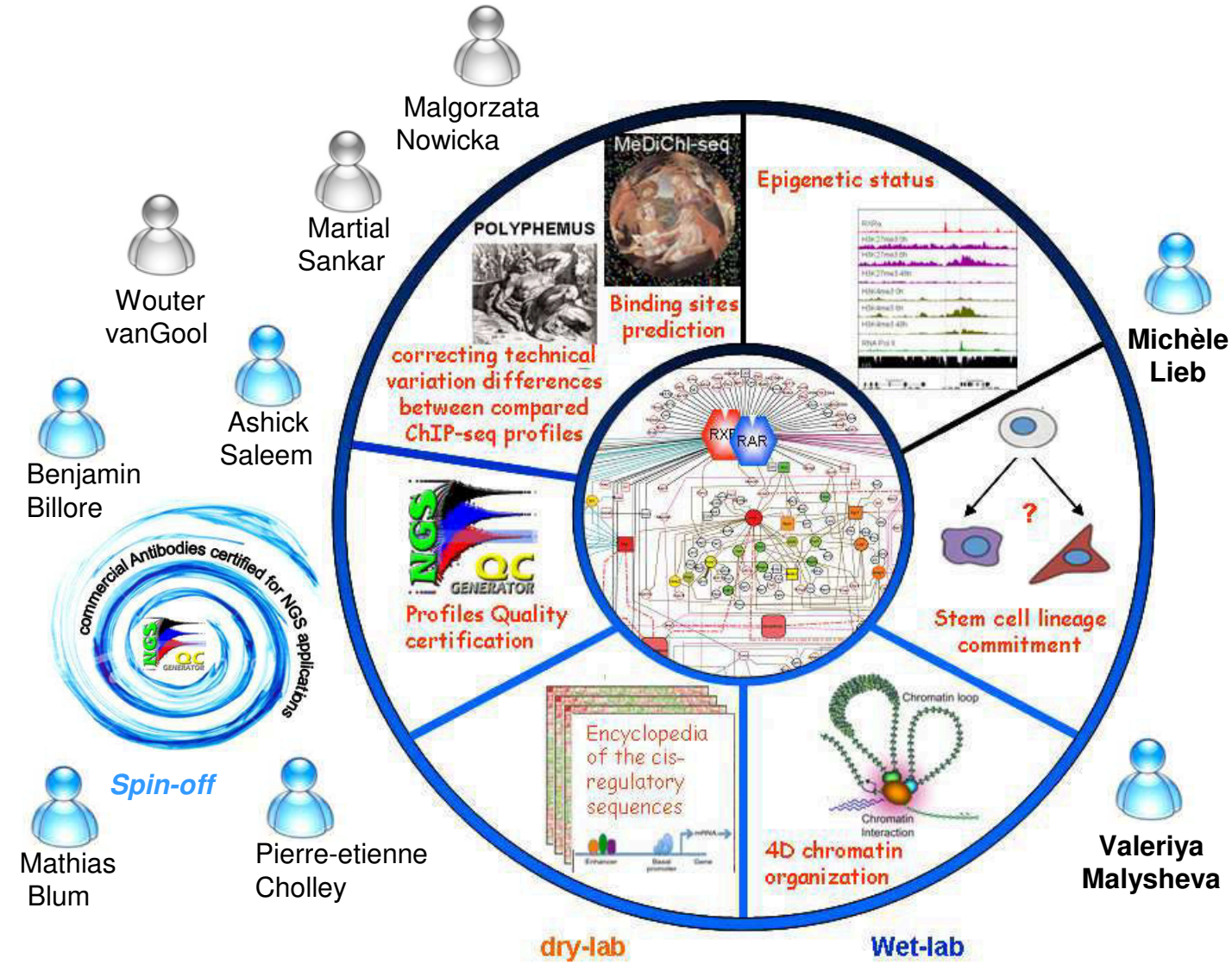


Page 1

www.ngs-qc.org



acknowledgements



igbmc
institut de génétique et de biologie moléculaire et cellulaire

Functional Genomics & Cancer

The Team

Researchers
Hinrich GRONEMEYER
Marco Antonio MENDOZA PARRA

Post-Doctoral Fellowships
Valeria PAVET-PORTAL
Maximiliano PORTAL

Phd Students
Pierre BORIES
Valeriya MALYSHEVA
Mohamed Ashick MOHAMED SALEEM
Yelyzaveta SHLYAKHTINA

Engineers & Technicians
Benjamin BILLORE
Matthias BLUM
Pierre-Etienne CHOLLEY
Cathie ERB
Michèle LIEB

Administration
Irène YUJNOVSKY

cnrs Inserm UNIVERSITÉ DE STRASBOURG