Multi-dimensional and integrative pipeline for NGS-based datasets to explore cell fate decisions

Mohamed Ashick Mohamed Saleem

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE

ET DE LA SANTE

IGBMC, UM 41/UMR 7104/UMR_S 964

THÈSE présentée par :

Mohamed Ashick MOHAMED SALEEM

soutenue le : **30 Novembre 2015**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Bioinformatique et Biologie des systèmes

Pipeline intégratif multidimensionnel d'analyse de données NGS pour l'étude du devenir cellulaire

THÈSE dirigée par :

GRONEMYER Hinrich — Directeur de recherches, IGBMC, Strasbourg

RAPPORTEURS :

VANDEL Laurence — Directeur de recherches, Université Paul Sabatier, Toulouse

BISCHOF Oliver — Directeur de recherches, Institut Pasteur, Paris

---

AUTRES MEMBRES DU JURY :

KASTNER Philippe — Directeur de recherches, IGBMC, Strasbourg

Acknowledgements

I would like to thank my supervisor Dr. Hinrich Gronemeyer for accepting me as a PhD student and providing me an opportunity to work in his lab. Being a bioinformatician, his encouragement on bridging my bioinformatics skills with broader biological viewpoint has helped me greatly to evolve as a researcher. I am grateful to him for patiently guiding me in all aspects, especially in scientific writing.

I would like to express my special appreciation and thanks to the efforts of Dr. Marco Antonio Mendoza Parra for providing all the necessary knowledge and guidance for proper understanding of epigenetics from a bioinformatics viewpoint when I started to work afresh in the lab. I would like to acknowledge his sincere contributions that have helped me greatly in my research.

I would like to thank Prof. Edith Heard and Dr. Chaligne Ronan for providing me with an excellent and challenging collaboration work. I sincerely thank my examiners Dr. Laurence Vandel, Dr. Oliver Bischof and Dr. Philippe Kastner for their helpful comments on my thesis and being supportive.

Many thanks are due to my current lab-members Valeria, Maxi, Lera, Lisa, Pierre-Etienne, Matthias, Benjamin, Akinchan, Michele, Cathy and Aurelie for providing healthy discussions and suggestions in my research and being very friendly. I would like to thank my former colleagues Irene, Wouter, Gosia, Shankar and Pierre Boris for useful discussions and collaborations. I would also like to thank my friends Ujjwal, Sanjay, Meghna, Nithyha, Tajidh, Kareem, Thanuja, Vivek and Atish in Strasbourg who provided me with an active social life and moral support.

I cannot finish without thanking my family members and friends. I pay sincere and heartfelt admiration to my loving father, mother, brothers and friends back home for their prayers and well wishes. Although, I was far away from them but their constant support and encouragement was a great source of motivation for me.

Table of contents

LIST OF FIGURES AND TABLES **1**
ABBREVIATIONS **3**
OUTLINE OF THE THESIS **6**
ABSTRACT **8**
INTRODUCTION **10**
CHAPTER 1. EPIGENETIC MODIFICATIONS AND ITS ROLE IN CELL FATE DECISIONS **12**
1.1. DNA LEVEL EPIGENETIC MODIFICATIONS FOR GENE REGULATION 12
1.2. HISTONE LEVEL EPIGENETIC MODIFICATIONS FOR GENE REGULATION 13
1.3. EPIGENETIC WRITERS AND ERASERS 17
1.3.1. Histone acetylation *17*
1.3.2. Histone methylation *18*
1.3.3. Other epigenetic modifications in histone tails and core domains *22*
1.4. EPIGENETIC READERS 24
1.5. ROLE OF NON-CODING RNAS IN EPIGENETICS 25
1.6. ROLE OF EPIGENETIC MODIFICATIONS IN CANCER 26
1.7. EPIGENETIC INSTABILITY OF INACTIVE X CHROMOSOME IN BREAST CANCERS 28
1.8. HERITABLE GENE IMPRINTING AND DISORDERS 32
CHAPTER 2. RISE OF NGS DRIVEN STUDIES IN EPIGENETICS **35**
2.1. A BRIEF HISTORY OF NEXT GENERATION SEQUENCING TECHNOLOGY 35
2.2. CHROMATIN IMMUNOPRECIPITATION (CHIP) SEQUENCING FOR EXPLORING GENOME FUNCTION 43
2.3. CAVEATS IN NGS DRIVEN STUDIES 45
CHAPTER 3. A DETAILED BIOINFORMATICS PIPELINE FOR CHIP-SEQ STUDIES **47**
3.1. SEQUENCING QUALITY CONTROL 47
3.1.1. Base quality *48*
3.1.2. Adapter contamination *51*
3.2. MAPPING OF SEQUENCED READS TO A REFERENCE GENOME 52
3.2.1. Unique reads *54*
3.2.2. Uniquely aligned reads *56*
3.3. EXPERIMENTAL QUALITY CONTROL POST ALIGNMENT 59
3.4. PIPELINE USED IN OUR STUDIES 61
3.5. INTERPRETATION OF CUMULATED READ PROFILES 62

3.5.1. Peak detection to identify protein binding regions *62*
3.5.2. Multi-dimensional dataset integration *64*
3.5.3. Integrative and systems biology analysis *65*

CHAPTER 4. SCOPE AND SPECIFIC GOALS OF THIS THESIS **69**

4.1. DEVELOPMENT OF TOOLS TO EVALUATE THE DATA QUALITY AND NORMALIZE TECHNICAL DIFFERENCES IN MULTI-SAMPLE ANALYSIS 69
4.2. INTEGRATIVE ANALYSIS OF EPIGENOMIC AND TRANSCRIPTOMIC STATUS OF THE XI IN BREAST CANCER 70

CHAPTER 5. RESULTS AND DISCUSSIONS **71**

5.1. NGS-QC – A QUALITY CONTROL SYSTEM FOR CHIP SEQUENCING PROFILES 71
5.1.1. NGS-QC generator *72*
5.1.2. NGS-QC Database *74*
5.1.3. Discussion *77*

Manuscript 1

5.2. EPIMETHEUS - A MULTI-PROFILE NORMALIZER FOR EPIGENOME SEQUENCING DATA 78
5.2.1. Methodology *78*
5.2.2. Output *79*
5.2.3. Evaluation of Epimetheus on multiple datasets *82*
5.2.4. Discussion *84*

Manuscript 2

5.3. THE INACTIVE X CHROMOSOME IS EPIGENETICALLY UNSTABLE AND TRANSCRIPTIONALLY LABILE IN BREAST CANCER 86
5.3.1. Allele-specific expression and chromatin state analysis of X chromosome *86*
5.3.2. Discussion *90*

Manuscript 3

CHAPTER 6. CONCLUDING REMARKS AND FUTURE PERSPECTIVES **92**

6.1. UTILITY AND LIMITATIONS OF DEVELOPED TOOLS 92
6.2. DATA MANAGEMENT IN CURRENT BIOINFORMATICS 95

GLOSSARY **97**

SEQUENCING APPLICATIONS AND BIOLOGICAL GLOSSARY: 97
BIOINFORMATIC GLOSSARY: 101

THESIS RÉSUMÉ

Appendices **106**

REFERENCES **111**

List of Figures and tables

Figure 1 Nucleosome core particle

Figure 2 Post-translational modification sites of histone proteins

Figure 3 Epigenetic ‘writers’, ‘erasers’ and ‘readers’ scheme

Figure 4 Acetylation and deacetylation of histone proteins

Figure 5 Methylation of lysine and arginine residues

Figure 6 Major landmarks in random XCI research

Figure 7 Schematic view of the kinetics of X-chromosome inactivation

Figure 8 Genome-wide distribution of identified imprinted genes

Figure 9 Timeline of landmarks in NGS and bioinformatics

Figure 10 An overview of different NGS experiments workflow

Figure 11 Illumina sequencing chemistry

Figure 12 Oxford Nanopore MinION sequencing machine during our testing phase

Figure 13 An overview of ChIP-seq methodology

Figure 14 Boxplot illustrating quality distribution for samples with high and poor quality

Figure 15 Distribution of average quality per read for samples with high and poor quality

Figure 16 Percentage of reads with adapter contamination with its positional distribution

Figure 17 Percentage of unique reads in comparison with total sequences

Figure 18 Comparison of different aligners in single-end and paired-end data illustrating the false positive rate in both types of data

Figure 19 Comparison of ERa peaks across different samples

Figure 20 Illustration of skewed enrichment in RNA degraded transcriptome data

Figure 21 An overall scheme of allele-specific analysis established for X chromosome inactivation analysis perturbation study in breast cancer cells

Figure 22 Dynamic regulatory map of yeast response to amino acid starvation

Figure 23 Display illustrating the database page showing the search panel and violin plots table

Figure 24 Display illustrating the results obtained after performing a query in the NGS-QC database

Figure 25 A scheme of the Epimetheus workflow with illustrative plots

Figure 26 Effects of data normalization.

Figure 27 Allele specific analysis led to the identification of genes escaping XCI specific to cancer cell-lines

Table 1 List of demethylases and their targets.

Table 2 Technical specifications of different platforms.

Table 3 Comparison of different NGS reads aligners.

Abbreviations

5-mC	5-methylcytosine
Ac	Acetylation
ADMA	Asymmetric di-methylarginine
ADP	Adenosine diphosphate
AI	Allelic imbalance
AML	Acute myeloid leukemia
ASCII	American Standard Code for Information Interchange
BET	Bromodomain and extra-terminal
BS	Bisulfite sequencing
CBP	CREB-binding protein
DUB	deubiquitinating enzyme
DNMT	DNA methyltransferases
EC	Embryonic carcinoma
eRNAs	enhancer RNAs
EZH2	Enhancer of zeste homolog 2
FDR	False discovery rate
GNAT	Gcn5-related N-acetyltransferase
GRN	Gene regulatory network
H1/H2A/H2B/H3/H4	Histone proteins
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HKMT	Histone lysine methyltransferase
HMT	Histone methyltransferases
IDR	irreproducible discovery rate
IP	Immuno-precipitated
JHDM	JmjC-domain-containing histone demethylase
JmjC	Jumonji C
lncRNAs	Long non-coding RNAs
LSD1	Lysine specific demethylase 1

MBD	Methyl CpG binding domain
MDS	Myelodysplastic syndrome
Me1,Me2,Me3	Mono-,di-,tri-methylation respectively
MMA	Mono-methylarginine
N-CoR	nuclear receptor corepressor
ncRNAs	Non-coding RNAs
ndsRNAs	Nuclear double stranded RNAs
NGS	Next Generation sequencing
NuA3/NuA4	Nucleosomal acetyltransferases of H3 and H4
PADI4	Petidylargininedeiminase 4
PAR	pseudo autosomal region
PARs	Promoter-associated RNAs
PCAF	P300/CBP-associated factor
PhS	phosphorylated serine
piRNA	Piwi-interacting RNA
PRMT	Protein arginine methyltransferases
PP	Protein phosphatases
PTM	Post translational modification
RCI	Read count intensity
PolII	Polymerase II
RpB	Reads per bin
rRNAs	Ribosomal RNAs
SAGA	Spt/Ada/Gcn5L acetyltransferase
SAM	Sequence alignment/map format
SDMA	symmetric dimethylarginine
siRNAs	Small interfering RNAs
snRNAs	Small nuclear RNAs
snoRNAs	Small nucleolar RNAs
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
TET	Ten-eleven translocation

TFIID	Transcription factor II D
TFTC	TBP-free TAF-containing complex
TMR	Total mapped reads
tRNAs	Transfer RNAs
TSS	Transcription start site
VCF	Variant caller format
WCE	Whole cell extract
XCI	X chromosome inactivation
Xic	X inactivation centre
Xi	Inactive X
Xa	Active X
PHD	Plant homeodomain proteins
BAP1	BRCA1 Associated Protein-1
MBT	Malignant brain tumor
BRCT	BRCA1 C Terminus
PWWP	conserved Proline and Tryptophan
SAND	Sp100, AIRE-1, NucP41/75, DEAF-1
MYND	Myeloid, Nervy, and DEAF-1
SANT	Swi3, Ada2, N-Cor, and TFIIIB

Outline of the thesis

Epigenetics is one of the crucial mechanisms that systematically control the gene regulation in cell fate decisions. Several studies have linked their aberrant behaviour to diseases including cancer. Hence, it is important to understand the molecular mechanisms underpinning epigenetics. In that context, ChIP-seq is widely used for studying epigenetic modifications, especially histone modifications. This biology and informatics blended thesis aims at two aspects (i) development of novel tools to evaluate the quality and correct the sequencing depth variations embedded in NGS driven ChIP-seq assays, and (ii) analysis of the epigenetic status of chromosome X inactivation (XCI) in breast cancer cells.

Following the general introduction, first chapter of this thesis provides a brief literature based description on the biological background of this thesis. I begin by describing a DNA level epigenetic modification called DNA methylation and then proceed to explain histone level modifications categorised as epigenetic 'writers', 'erasers' and 'readers'. Different enzymes involved in each type of modifications and their functional role are discussed. After summarizing about epigenetics, the basic mechanism of one of the exemplary chromosome wide X inactivation will provide the mechanism of X chromosome inactivation (XCI) and the role of epigenetics in it, as one of my studies focus on understanding the deviation of epigenetic status in breast cancer cells. A small summary of imprinted genes have also been discussed, as the comprehensive data is available from XCI study and similar analysis can be used to characterise the epigenetic and allelic status of imprinted genes in breast cancer cells. Imprinted genes analysis is currently ongoing and preliminary results are only available.

Second chapter provides a quick outline on the rise and evolution of next generation sequencing, especially in the context of functional genomics. It also describes several challenges that exist in NGS driven analysis. Third chapter provides a brief literature and experience based description on the bioinformatic background of epigenetic related studies. Best practices to be followed in such analysis are discussed along with the directions and immediate priorities in bioinformatics related challenges in analysis. Fourth chapter provides the broad scope and specific goals of this thesis. Fifth chapter covers the

results and discussions involving the development of two new bioinformatics tools and allele-specific analysis to understand the aberrant behaviours in inactive X chromosome of breast cancer cells. For each manuscript, its corresponding manuscript is attached for the detailed materials and methods, and results, along with a brief overview. Final chapter is intended to provide the future perspectives with concluding remarks. A list of glossary is provided at the end for different NGS applications and bioinformatic terminologies/approaches which are often used in the thesis. I have attached the list of publications that I am part of, including the manuscripts which are submitted.

Abstract

Over the years, various studies have shown that epigenetic modifications have a significant role in gene regulation. Unravelling the mechanisms and functional aspects of such modifications would help us understand why various cells types exhibit different behaviours, though the genomic DNA is same. Since the identification of its crucial role in gene regulation, aberrant changes in such modifications have been observed in several diseases including cancer. As most of these modifications are reversible, recently a large focus has been given on understanding these epigenetic modifications for therapy.

With the rise of next generation sequencing technology, Chromatin ImmunoPrecipitation-Sequencing (ChIP-Seq) has become widely used approach to profile histone modifications. Epigenetic studies may involve sequencing and comparison of multiple factors from different samples. This poses a significant bioinformatic challenges as ChIP-Seq is inherently prone to variabilities embedded in individual assays like antibody efficacy, sequencing depth variation, etc. These underlying technical variabilities and poor enrichment profiles can significantly bias the comparative studies. Hence, there is an imminent need for novel approaches and tools to address these caveats for any such comparative studies. In that context, we have developed NGS-QC, a robust bioinformatics-based quality control system to infer the experimental quality and comparability of the data. This tool and its associated database is publicly available and aids in interpreting the quality of the enrichment datasets and compare them with existing overall quality trend for a given factor from public data. However, even high quality datasets exhibit significant sequencing depth variation and require normalization to correct this variation prior to comparison. Currently existing normalization methods either apply linear scaling corrections and/or are restricted to specific genomic regions. To overcome these limitations, we have developed Epimetheus, a genome-wide quantile-based multi-profile normalization tool for histone modification and related datasets. Comparison with existing methods proves Epimetheus to be more robust, and its outputs are scalable to a variety of downstream analyses.

We employed these newly developed tools in a bioinformatics pipeline to understand the epigenetic status of X chromosome inactivation (XCI) in breast cancer cells. XCI is an epigenetic paradigm and an excellent model to understand the epigenetic system where chromosome-wide repression takes place. Around 50 years ago, disappearance of Barr body (Xi - inactive X chromosome) in breast cancer cells was observed, which was later found to be de-condensation of heterochromatic Xi along with X-linked gene reactivation. An allele specific transcriptomic and epigenetic profiling comparison between normal and breast cancer cells could reveal the regions or genes that are epigenetically disrupted in X chromosome. We established an integrative bioinformatic pipeline to integrate genetic (SNP6 and Exome-seq), epigenetic (ChIP-seq) and transcriptomic (nascent RNA SNP6 and mRNA-seq) data to understand the allelic and epigenetic status of disrupted Barr body in breast cancer cells. Our analysis has revealed perturbation in epigenetic landscape of X-chromosome and aberrant gene reactivation in Xi including the one are associated with cancer promotion.

Introduction

Epigenetics, a term coined by Conrad Waddington, defined as “the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being” (Waddington 1942). This definition was very broad and referred to all molecular pathways modulating the expression of a genotype into a particular phenotype. Rapid growth in the field and technology has resulted in a better understanding of this process and is currently defined as “a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence” (Berger et al., 2009). As a bioinformatic student, it is worth mentioning the analogy given by Prof. Jörn Walter “the hard disk is like DNA, and then the programmes are like the epigenome”.

Though all the cells in the human body carry the same genetic information in its DNA sequence, it is the expression of genes with spatial and temporal specificity that brings about their differentiation into cells and tissues with specialized biological functions. While a complete mammalian genome is composed of approximately 25,000 protein-coding genes, about 30% of the DNA sequence, only a half of them are expressed in any given cell type and most of those expressed are dedicated to cellular homeostasis (Romanoski et al., 2015). The fine control of gene expression is achieved through a complex set of *cis* and *trans* factors both at the 2D and at the 3D level. Genetic elements such as promoters, enhancers, repressors/silencers, insulators, etc., act in *cis* providing binding sites to complex set of factors comprising of transcription factors, co-regulators (activators and repressors), mediators, which act in *trans* for the precise regulation of gene expression. Lately, there is a realization that 3D structure of the chromatin has an important role to play in the organization of these *cis* and *trans* elements facilitating proximity interaction in 3D. DNA in the nucleus is very compactly packed around proteins and condensed into chromatin. Despite such high level compaction, it is accessible to these regulatory effectors and other interactions for gene expression. Recent models have suggested that three-dimensional nuclear organization contributes to genome folding, chromosome compartmentalization and the formation of gene regulatory interactions, ensuring appropriate genome function (Lopes Novo and Rugg-Gunn, 2015). This gives a broader complexity to the regulatory mechanism where the functional activities of the

effectors are spatially facilitated by the chromatin organisation. Remodeling of the chromatin is a dynamic process of chromatin architecture modification, by a variety of factors, to control gene expression. Such remodeling is principally carried out by covalent histone modifications by specific enzymes, and ATP-dependent chromatin remodeling complexes which restructure nucleosomes. The dynamic remodeling of chromatin also imparts an epigenetic regulatory role in several key biological processes, DNA replication and repair; as well as development and pluripotency.

Many of the regulatory factors, though not directly coded the genomic DNA sequence, can also be heritable and these are known as epigenetic factors. These comprise of methylation and other modifications of the DNA nucleotides, chemical modifications and variants of the structural histone proteins constituting the chromatin and a larger variety of non-coding RNAs. Recent studies have shown that many of the epigenetic modifications are influenced by environmental conditions/stresses such as metabolic and biochemical factors and even psychological stresses (Raabe and Spengler, 2013). Thus, epigenetic factors can be hypothesised to provide a way for the organism to pass on the information accumulated through the environmental factors and prepare its progeny. Therefore, it is important to study the epigenetic programming and different machineries involved in gene regulation to decipher their functional role in basic cell processes and their aberrant behavior in diseased cells.

With the advancements in next generation sequencing (NGS) technology and perpetual bioinformatics support, epigenetic modifications can now be studied at a genomic scale. Applications like ChIP-seq and MBD-seq has been widely used for such studies, and FAIRE/ATAC-seq like approaches has been used to identify the open chromatin regions. However, given the influence from multiple factors, the data obtained from these assays is inherently prone to technical variation, which makes the subsequent bioinformatic analysis challenging. Hence, there is an imminent need for novel approaches to evaluate and address these differences to facilitate more accurate analysis.

CHAPTER 1

EPIGENETIC MODIFICATIONS AND ITS ROLE IN CELL FATE DECISIONS

Chapter 1. Epigenetic modifications and its role in cell fate decisions

Epigenetic modifications include DNA methylation, covalent modifications of histone proteins in its tails and core domains and non-coding RNA mediated regulation. In this chapter, each type of modifications and its role in cell fate decisions, especially cancer, are briefly discussed.

1.1. DNA level epigenetic modifications for gene regulation

DNA methylation, an evolutionarily ancient and the only covalent DNA modification known in mammals, occurs at the 5'C of cytosine residues resulting in 5-methylcytosine (5-mC). It occurs predominantly in the symmetric GC context and is estimated to occur at ~70-80% of CG dinucleotides throughout the genome (Ehrlich et al., 1982). The rest of unmethylated CG dinucleotides are mostly found near gene promoters in dense clusters, termed CpG islands (Law and Jacobsen, 2010). The function of DNA methylation seems to vary with the genomic context such as transcriptional start sites with or without CpG islands, in gene bodies, at regulatory elements and at repeat sequences. When a CpG island in the promoter region of a gene is methylated, expression of the gene is typically repressed. Methylated residues of nucleotides serve as sites for the binding of Methyl CpG binding domain (MBD) proteins, which may either directly impede transcription complex binding or recruit histone deacetylates and other chromatin remodeling proteins to form a transcriptionally silent heterochromatin. In the case of cancers, tumor suppressor gene loci, such as retinoblastoma-associated protein 1 (*RB1*), *MLH1*, *p16* and *BRCA1* among others, are known to be frequently hypermethylated and repressed (Jones, 2012). DNA hypomethylating agents such as 5-Azacytidine and 5-Aza 2'-deoxycytidine are used in the treatment of Myelodysplastic Syndrome. They are thought to produce DNA hypomethylation by inhibiting DNA methyltransferases (due to irreversible binding) at low doses, and direct cytotoxicity at higher doses.

The addition of methyl group to DNA backbone is carried out by a family of enzymes called DNA methyltransferases (DNMTs) consisting of five members: DNMT1, DNMT2,

DNMT3a, DNMT3b, DNMT3L (Goll and Bestor, 2005). While DNMT1 is a large protein with 1620 amino acid residues, DNMT2 is a relatively small enzyme and resembles prokaryotic DNA methyltransferases. DNMT1 appears to be responsible for the maintenance of established patterns of DNA methylation, while DNMT3a and 3b seem to mediate establishment of new or *de novo* DNA methylation patterns. Two additional enzymes (DNMT2 and DNMT3L) may also have more specialized but related functions. DNMT3L shares homology with DNMT3a and DNMT3b and was reported to be responsible for establishment of maternal genomic imprinting (Bourc'his et al., 2001).

As opposed to DNA methylation, another important aspect is the removal of a methyl group, termed DNA demethylation. It can either be passive or active, or a combination of both. Passive DNA demethylation refers to loss of 5-mC on newly synthesized DNA strands during successive replication cycles when there is no functional DNA methylation maintenance machinery. Active DNA demethylation is the enzymatic process that removes or modifies methyl group from 5-mC by ten-eleven translocation (TET) enzyme-mediated oxidation. The TET family of 5-mC hydroxylases includes TET1, TET2 and TET3. The broader functions of 5-hmC in epigenetics are still unclear. However, a line of evidence does show that 5-hmC levels are strongly depleted in various tumors (Pfeifer et al., 2013).

1.2. Histone level epigenetic modifications for gene regulation

In the eukaryotic genome, DNA is tightly packed with histone proteins into a protein-DNA complex called chromatin. Chromatin comprises of basic repeating units called nucleosomes, which is an octamer with two copies each of the four core histones H2A, H2B, H3 and H4, and DNA (~146bp) wrapped around the histones. With the help of H1 histone and additional proteins, nucleosomes are further packaged spirally into a 30nm fibre with six nucleosomes per turn (Loyola et al., 2001). This fibre is further looped and coiled to give rise to higher order structures known as chromosomes. Histones have a central globular domain and unstructured N- and C-terminal tails protruding from the central globular domain (Figure 1).

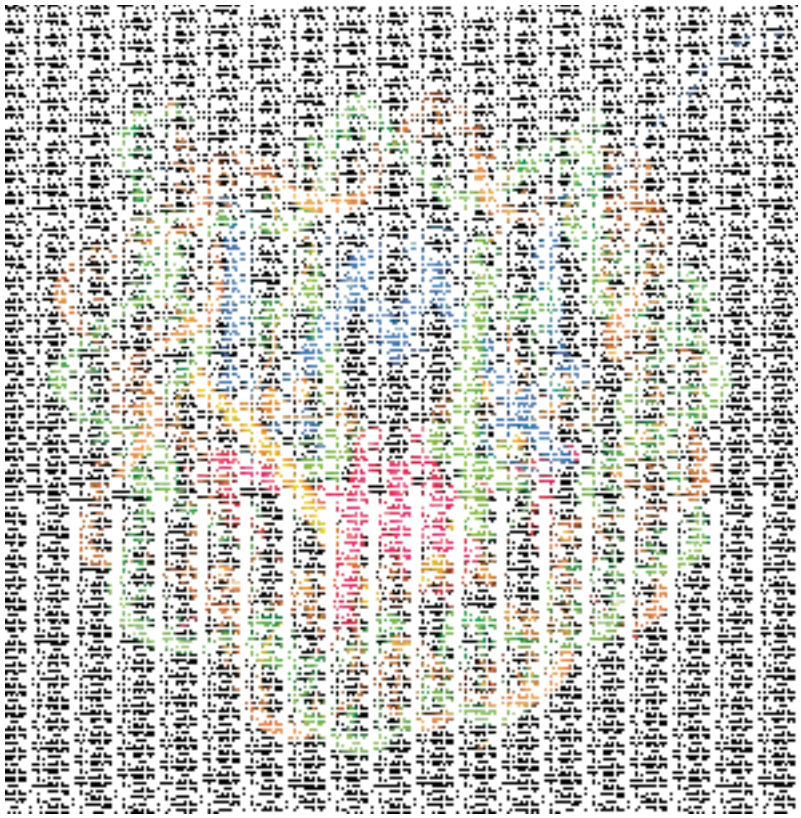


Figure 1. Nucleosome core particle. Bio-molecular structure of octamer histone proteins main chains (blue: H3; green: H4; yellow: H2A; red: H2B) surrounded by 146-bp double stranded DNA phosphodiester backbones (brown and turquoise) with unstructured C- and N- terminal histone tails protruding from the complex. (Taken from Luger et al. 1997).

The N-terminal and C-terminal histone tails along with central globular domain are subjected to post translational modifications (PTMs) such as acetylation, methylation, phosphorylation, ubiquitylation, sumoylation, ADP ribosylation, deimination, biotinylation, butyrylation, N-formylation, and proline isomerization (Cohen et al., 2011). Methylation and acetylation of histone proteins are the most studied histone modifications. Specific enzymes covalently modify the amino acids residues in the histone tails and such that many sites can be potentially modified, resulting in complex patterns of histone modifications (Figure 2). All of these modifications together compartmentalize the chromatin into two states based on their transcriptional status – active ‘euchromatin’ and inactive ‘heterochromatin’.

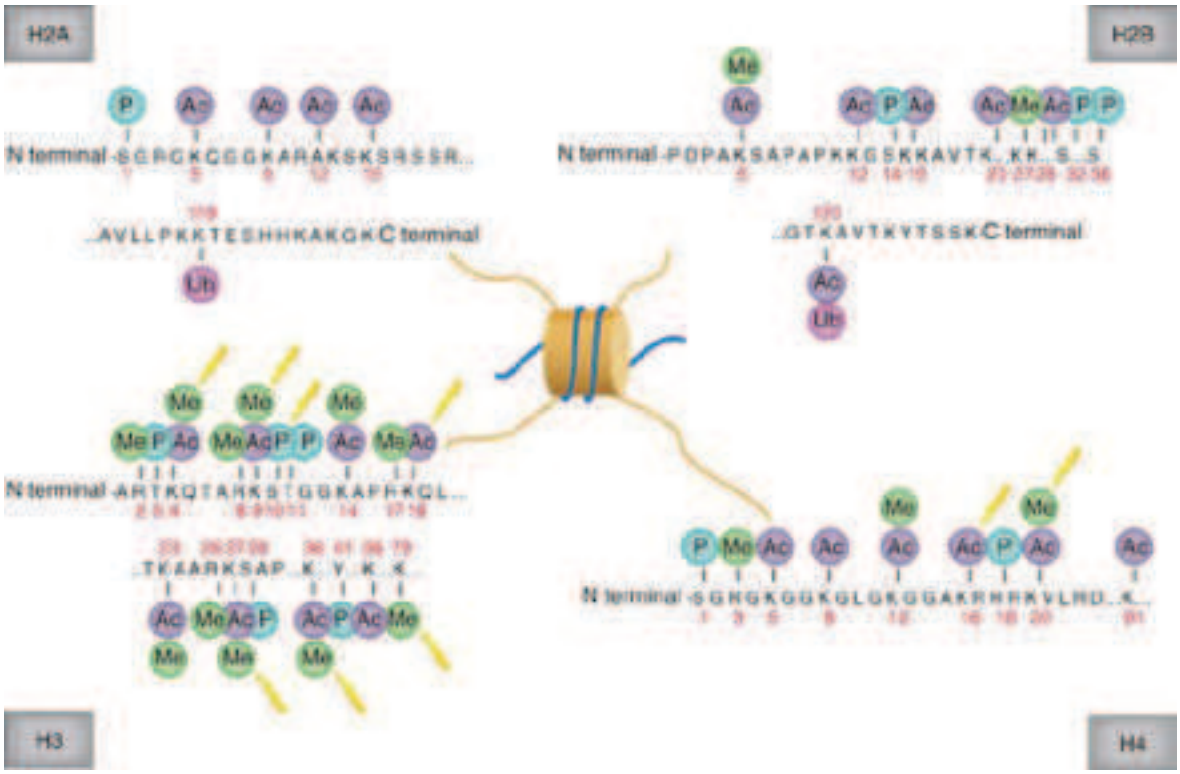


Figure 2. **Post-translational modification sites of histone proteins.** An illustrative view of different histone modification sites along their protruding C- and N-terminal histone tails with type and position in the amino acid sequence. PTMs (Ac-acetylation, Me-Methylation, P-Phosphorylation and Ub-Ubiquitination) that are associated with cancer are highlighted in yellow. (Taken from Rodríguez-Paredes and Esteller 2011).

Similarly, a specific set of enzymes exist that remove these chemical marks (Kouzarides, 2007). Such enzymatic addition and removal of chemical groups is caused by epigenetic modifiers, referred as epigenetic 'writers' and 'erasers' respectively. Interpretation of this epigenetic code is recognised by a set of proteins called epigenetic 'readers' (Falkenberg and Johnstone, 2014) (Figure 3). Such reversible and dynamic epigenetic modifications form a kind of code for the interactions of histones with other proteins, which determines the local chromatin structure and thereby regulating cell specific gene expression (Wu and Grunstein, 2000). Such combinatorial histone modifications may work as a marking system that is recognized/read by regulatory proteins (Quina et al., 2006). Further, these epigenetic modifications have to be replicated along with the DNA during mitosis and to be inherited to the next subsequent cell generations to maintain cell fate (Arzate-Mejía et

al., 2011). The histone code hypothesis predicts that "multiple histone modifications, acting in a combinatorial or sequential fashion on one or multiple histone tails, specify unique downstream functions" (Strahl and Allis, 2000). Signal transduction pathways are responsible for the integration and interpretation of such codes into specific transcriptional states (Schreiber et al., 2002). Such transcriptional states can be maintained through switch-like signalling ('on' or 'off') resulting from feedback loops and these signals converge on chromatin to shape the transcriptional landscape (Bonasio et al., 2010).

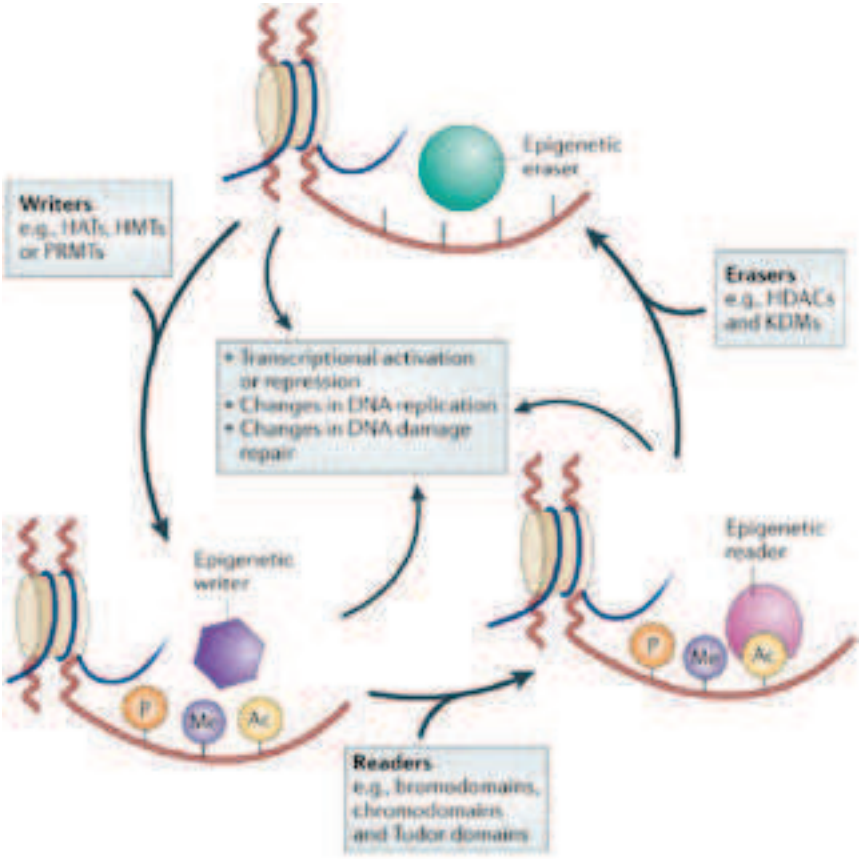


Figure 3. Epigenetic 'writers', 'erasers' and 'readers' scheme. Epigenetic writers (HATs, HMTs and PRMTs) add chemical group on amino acid residues, which are read and interpreted by group of proteins (containing bromodomains, chromodomains, and Tudor domains) called epigenetic readers. Epigenetic erasers catalyse the removal of epigenetic marks. Together, these modifications form a kind of histone code that dynamically regulates gene in precise spatio-temporal manner. (Taken from Falkenberg and Johnstone 2014).

1.3. Epigenetic writers and erasers

1.3.1. Histone acetylation

Histones are covalently modified at the epsilon-amino group of lysines on the N-terminal tail, especially on H3 and H4, by a class of enzymes called histone acetyltransferases (HATs). Acetylation of histones is associated with transcriptionally active euchromatin (Allegra et al., 1987). It neutralizes the positive charge of the target lysine and affects the DNA-histones interaction resulting in an open euchromatin (Shahbazian and Grunstein, 2007). Acetylation of histones is controlled by the opposing action of Histone deacetylases (HDACs) which remove the acetyl group from lysine residues. This interplay between HATs and HDACs activity regulates the level of histone acetylation in the cell (Figure 4). There are three major families of HATs: GNATs, P300/CBP and MYST proteins. Gcn5-related N-acetyltransferase (GNAT) is a well-studied HAT family and has been grouped based on its homology regions and similar acetylation-related motifs. It includes HATs Gcn5, its close relatives and three distantly related Hat1, Elp3, and Hpa2 (Sterner and Berger, 2000). The MYST family includes MOZ, Ybf2/Sas3, Sas2 and Tip60, also has an acetylation-related structural motif. The P300/CBP (CREB-binding protein) family consists of two paralogous proteins, P300 and CBP. These two proteins have interchangeable functions. Members of the P300/CBP family contain many functional domains including a structural motif which is involved in acetyl-CoA binding, three zinc finger regions and a bromo-domain. P300/CBP acts as a co-activators and harbor domains for interaction with many transcription factors (Karmodiya et al., 2014). Similarly, there are four classes of HDACs that have been identified: Class I, II, III, IV. Class I HDACs include 1, 2, 3, and 8, and Class II HDACs includes 4, 5, 6, 7, 9, and 10. Class III includes enzymes called sirtuins. HDAC11 is the only member in Class IV but it has features of both Classes I and II. The first nuclear histone acetyltransferase, Tetrahymena p55 provided the first link between HATs and transcriptional activation (Brownell et al., 1996). Since then, studies have shown that acetylation has an important role in transcription activation, elongation, DNA damage & repair and DNA replication (Bose et al., 2004; Brownell et al., 1996; Lee and Shilatifard, 2007)

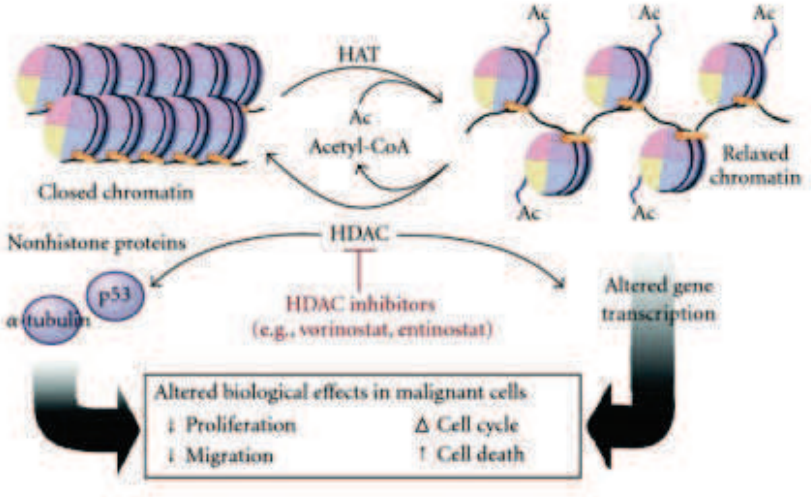


Figure 4. **Acetylation and deacetylation of histone proteins.** Addition of acetyl-CoA via HATs and removal of acetyl-CoA via HDACs resulting in condensed heterochromatin to euchromatin and vis-versa respectively. (Taken from Rodd et al., 2012).

Activation and repression of gene expression is mostly regulated through multi subunit complexes of co-activators and co-repressors. HATs form part of many transcriptional co-activator complexes including SAGA (Spt/Ada/Gcn5L acetyltransferase), PCAF, ADA (transcriptional adaptor), TFIID (transcription factor II D), TFTC (TBP-free TAF-containing complex), and NuA3/NuA4 (nucleosomal acetyltransferases of H3 and H4). Similarly, HDAC containing complexes constitute co-repressors such as SIN3, N-CoR. Genome wide mapping studies have, shown the presence of HDAC complexes at the majority of actively transcribed loci along with repressed ones. HDACs have been shown to prevent cryptic initiation of transcription within coding regions, thus maintaining a precise control of gene expression levels. As genome wide mapping studies accumulate in different cell fate systems, the nature of interaction and role of these co-regulator complexes is starting to become clearer (Perissi et al., 2010; Yang and Seto, 2007).

1.3.2. Histone methylation

Histone methylation occurs on the lysine or arginine residues of histones H3 and H4. Unlike acetylation, methylation has no effect on the charge of the histones (Bannister and Kouzarides, 2011). Histone methylation brings added complexity in histone code as lysine

can be modified into mono-, di-, and tri-methylated states and arginine (Figure 5A) can be modified into mono- and di-methylated (in symmetric or asymmetric configurations) states (Figure 5B) (Zhang and Reinberg, 2001). Histone methylation of lysine is associated with transcriptional activation or repression depending on the site where methylation occurs. For example, di-/tri-methylation of H3K27, tri-methylation of H3K9 and mono-methylation of H4K20 were shown to be involved in transcriptional silencing, whereas di-/tri-methylation of H3K4 and H3K36, and di-methylation of H3K79 were associated with transcriptionally active status (Sims and Reinberg, 2006).

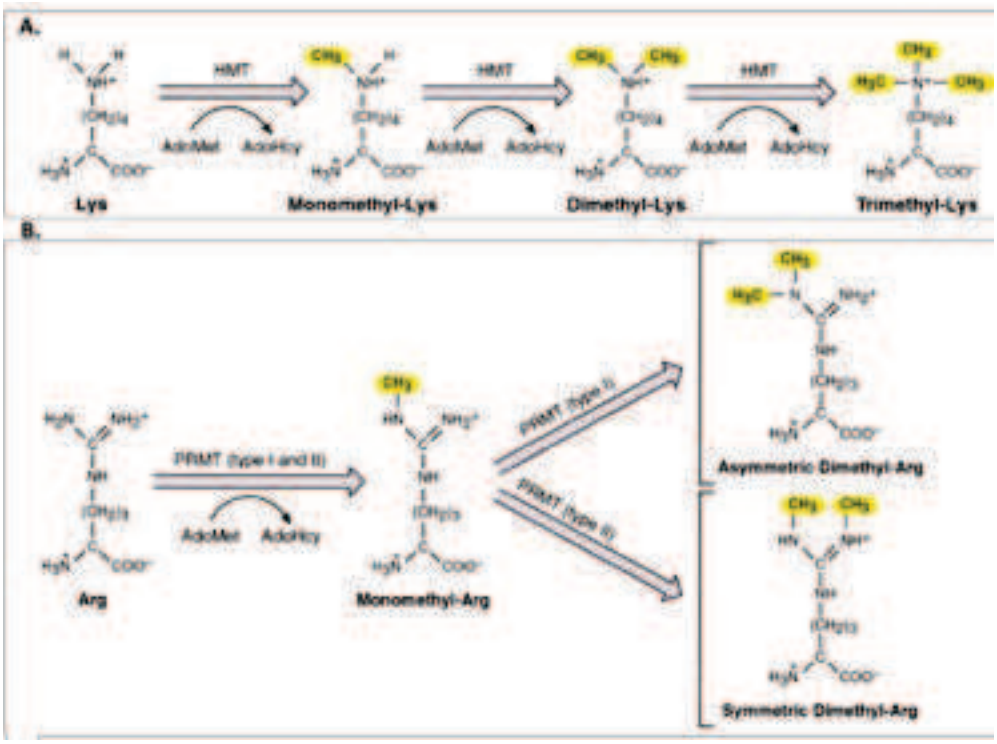


Figure 5. **Methylation of lysine and arginine residues.** (A) Molecular structure of lysine and consequent changes after mono-,di- and tri-methylation of lysine residue (B) Molecular structure of lysine and consequent changes after mono- and di-methylation of arginine residues. Di-methylation of arginine residue can result in either asymmetric or symmetric depending on enzyme that catalyse, type I and II protein arginine methyltransferases respectively. (Adapted from Zhang and Reinberg 2001).

Methylation of lysine and arginine residues is carried out by HKMTs and PRMTs respectively. All of the HKMTs contain SET domain that harbors the enzymatic activity

except Dot1 enzyme. HKMTs tend to be relatively specific enzymes and modify appropriate lysine residues to a specific degree i.e., mono, di, and/or tri-methyl states. X-ray crystallography studies showed that there is a key residue within the enzyme's catalytic activity domain that determines the degree (Bannister and Kouzarides, 2011). PRMTs are classified as either: type I (CARM1, PRMT1, PRMT2, PRMT3, PRMT6, and PRMT8); type II (PRMT5 and PRMT7) or type III. Type I and type II enzymes catalyze the formation of an intermediate mono-methylarginine (MMA), which is further catalyzed into asymmetric di-methylarginine (ADMA) by type I and symmetric dimethylarginine (SDMA) by type II (Di Lorenzo and Bedford, 2011).

In human cells, MLL proteins, SET7/9, and Ash1 are HMTs that catalyze the methylation of H3K4. HMTs like ESET/ SETDB1, G9a, SUV39-h1, SUV39-h2, and Eu-HMTase catalyze the methylation of H3K9. SMYD2 and NSD1 are associated with H3K36 methylation. Enhancer of zeste homolog 2 (EZH2), a polycomb group enzyme is one of the well-studied HMT enzymes involved in oncogenesis, where it is shown to be repressing the expression of several tumor suppressor genes such as p16 INK4a, E-cadherin, DAB2IP, RUNX3, BRCA1, and the adrenergic receptor β2 (Cohen et al., 2011). G9a and EZH2 are HMTs that catalyze methylation of histone H3-K27 (Kouzarides, 2007). As mentioned earlier, both H3K9 and H3K27 methylations mediate heterochromatin formation and also participate in transcriptionally repressing the genes in euchromatin regions.

The discovery of histone demethylases demonstrate that histone methylation is not a permanent modification but rather a more dynamic process (Bannister et al., 2002). PADI4 (Petidylarginine deiminase 4) was the first identified enzyme that functions as a histone deiminase that converts methyl-arginine to citrulline as opposed to directly reversing arginine methylation. However, since PADI4 catalyzes deimination but not demethylation, it cannot strictly be considered a histone demethylase. LSD1 (Lysine specific demethylase 1) was the founding member of demethylase enzymes that directly reverse histone H3K4 or H3K9 modifications by an oxidative demethylation reaction in which flavin is a cofactor. Broadly, two major families of demethylases have been discovered: LSD1 and Jumonji C domain containing (JmjC domain) histone demethylases (JMJD2, JMJD3/UTX

and JARIDs). The specific amino acid residue and degree of methylation determines the demethylation enzyme (Table 1). LSD1 can only remove mono- and dimethyl lysine modifications whereas JmjC-domain-containing histone demethylases (JHDMs) can remove all three histone lysine-methylation states. These demethylases have been found to have potential oncogenic functions and involvement in other pathological processes (Hoffmann et al., 2012).

Name	Synonyms	Targets
KDM1A	LSD1, AOF2	H3K4me2/me1, H3K9me2/me1
KDM1B	LSD2, AOF1	H3K4me2/me1
KDM2A	FBXL11A, JHDM1A	H3K36me2/me1
KDM2B	FBXL10B, JHDM1B	H3K36me2/me1, H3K4me3
KDM3A	JMJD1A, JHDM2A	H3K9me2/me1
KDM3B	JMJD1B, JHDM2B	H3K9me2/me1
KDM4A	JMJD2A, JHDM3A	H3K9me3/me2, H3K36me3/me2
KDM4B	JMJD2B	H3K9me3/me2, H3K36me3/me2
KDM4C	JMJD2C, GASC1	H3K9me3/me2, H3K36me3/me2
KDM4D	JMJD2D	H3K9me3/me2/me1, H3K36me3/me2
KDM4E	JMJD2E	H3K9me3/me2
KDM5A	Jarid1A, RBP2	H3K4me3/me2
KDM5B	Jarid1B, PLU1	H3K4me3/me2
KDM5C	Jarid1C, SMCX	H3K4me3/me2
KDM5D	Jarid1D, SMCY	H3K4me3/me2
KDM6A	UTX, MGC141941	H3K27me3/me2
KDM6B	JMJD3, KIAA0346	H3K27me3/me2
	PHF8, KIAA1111, ZNF422	H3K9me2/me1, H4K20me1
KDM7	KIAA1718	H3K9me2/me1, H3K27me2/me1
KDM8	JMJD5, FLJ13798	H3K36me2

Table 1. **List of demethylases and their targets.** Detailed list of different demethylases with their specific modification sites at different amino acid residue in histone proteins. (Taken from Hoffmann et al. 2012).

1.3.3. Other epigenetic modifications in histone tails and core domains

Histone phosphorylation is the addition of a phosphate group to the histone proteins. Phosphorylation of H2A(X) is an important histone modification that plays a major role in DNA damage response. Phosphorylation of serine 10 in histone H3 (H3S10P) has been shown to correlate with gene activation in mammalian cells and with the induction of transcription during heat-shock response in Drosophila. H2A phosphorylation has also long been correlated with mitotic chromosome condensation, and again serine 10 appears to play a key role. Histone H3 phosphorylation is also known to occur after activation of DNA-damage signalling pathways (Rossetto et al., 2012). Histone dephosphorylation, is the removal of phosphate groups from histone proteins by enzymes called phosphatases. Mammalian serine/threonine-specific protein phosphatases (PPs) are represented by eight

distinct prototypes: PP1, PP2A, PP2B, PP2C, PP4, PP5, PP6 and PP7 (Moorhead et al., 2007; Swingle et al., 2009). Of these, PP1, PP2A and PP4 have all been identified as histone phosphatases: PP1 dephosphorylates H1, which is phosphorylated in a cell-cycle-dependent manner (Paulson et al., 1996). Phospho-H2AX (γ-H2AX) is immediately dephosphorylated after DNA repair by PP2A and PP4 in mammals and yeasts (Chowdhury et al., 2005; Keogh et al., 2006).

Histone ubiquitination is the addition of a small ubiquitin protein (76aa) to the histone proteins. Histone H2A was the first protein identified to be ubiquitinated (Goldknopf et al., 1975). The ubiquitination site has been mapped to the highly conserved residue, Lys 119 (Nickel and Davie, 1989). Around 5-15% of total H2A has been reported to be ubiquitinated in a variety of higher eukaryotic organisms (Robzyk et al., 2000). The majority of ubH2A is in monoubiquitinated form; however, polyubiquitinated H2A has also been detected in many tissues and cell types (Nickel et al., 1989). Deubiquitination is the removal of ubiquitin group from histones by ubiquitin specific peptidases known as deubiquitinating enzymes (DUBs). Several DUBs, including USP16, 2A-DUB, USP21, and BRCA1 associated protein 1 (BAP1) were identified as H2A-specific. Ubp8 and Ubp10 were identified as histone H2B DUBs in yeast (Blankenberg et al., 2001; Henry et al., 2003). In addition to H2A or H2B specific DUBs, several DUBs display dual specificity toward both H2Aub and H2Bub, such as USP3, USP12, and USP46. USP3 is required for cell cycle progression and genome stability, while USP12 and USP46 regulate Xenopus development (Joo et al., 2011; Nicassio et al., 2007). The Ubp8 homolog USP22 is a subunit of coactivator acetyltransferase hSAGA complex. It is recruited to the promoters by activators to deubiquitinate H2A and H2B, and is required for transcription activation (Zhang et al., 2008; Zhao et al., 2008). Multiple histone DUBs were identified, suggesting that they may have redundant functions or act in a context-dependent manner. Although their redundancy was not extensively investigated, current literature supports the notion that these DUBs have context-dependent functions in various processes. Their functions may also be dictated by their expression patterns in different tissues and stages during development.

Histone post-translational modifications occur, not only in the N-terminal tail domains, but also in the core domains (Mersfelder and Parthun, 2006). It has been proposed that the function of PTMs in the globular domain has a direct structural impact on nucleosome dynamics and chromatin regulation whereas the functional importance of PTMs in histone tails is context dependent. For instance, a recent study has demonstrated that the mutation of histone H3K27 in *Drosophila melanogaster* reproduces the effect on gene expression of abolishing H3K27me3 activity, suggesting that it is functionally important (Pengelly et al., 2013). On the contrary, cells with mutated histone H3K4 (a hallmark of active transcription) were viable and still could activate transcription of developmentally regulated genes suggesting limited functional relevance (Hödl and Basler, 2012). However, a recent quantitative modeling study confirmed that the neutralization of positive charges (like lysine acetylation) in the lateral surface of the chromatin could weaken the association of the histone proteins with DNA and thus could directly affect nucleosome dynamics and transcription (Fenley et al., 2010). Several other studies show that the lateral-surface PTMs may directly regulate the nucleosomal DNA accessibility to regulatory factors (e.g., H3K56ac), affect the mobility and stability of nucleosomes and, as a result, functionally contribute to transcription (e.g., H3K122ac) and other chromatin-dependent processes (Tropberger and Schneider, 2013).

1.4. Epigenetic readers

Interpretation of the information conveyed in the epigenetic language or code requires a third class of proteins called epigenetic "readers". Readers typically provide a docking site to accommodate a modified residue, and determine the modification (acetylation/methylation) and degree (such as mono-, di-, or tri-methylation of lysine) (Yun et al., 2011). Various domains such as bromo, chromo, PHD, Tudor, MBT, BRCT, and PWWP that recognize and bind these histone modifications have been identified. These domains recognize and bind to the PTMs produced by the writers and erasers and effect changes in transcription, often through scaffolding the formation of high order transcriptional complexes. Many other chromatin-linked domains are now emerging, including the SAND, PHD, MYND and SANT domains (Bottomley, 2004). BET (bromodomain and extra-terminal) proteins have been shown to regulate the expression of

key oncogenes and anti-apoptotic proteins. The recent discovery of highly specific inhibitors for the BET family has emerged as promising in diverse therapeutic areas like inflammation, viral infection, and especially in oncology (Filippakopoulos and Knapp, 2014). Recent studies have suggested that BET inhibitors may specifically modulate the disease-promoting genes expression without affecting the housekeeping genes. It has been shown that a BET inhibitor (I-BET858) selectively down-regulate genes associated with pathogenesis of Autism spectrum disorders (ASD) (Sullivan et al., 2015).

Bromo and tandem PHD domains target acetylated lysines and thereby regulate transcription, repair, replication and chromosome condensation. PHD, chromo, WD40, Tudor, double/tandem Tudor, MBT, Ankyrin Repeats, zf-CW and PWWP domains target methylated lysines on H3 resulting in either activation or silencing of gene expression. Reader domains of phosphorylation have not been well studied; only two readers BRCT domain of MDC1 and 14-3-3 family have been identified for phosphorylated serine (PhS) in histones (Yun et al., 2011).

1.5. Role of non-coding RNAs in epigenetics

In addition to covalent modifications, several classes of small and long non-coding RNAs (ncRNAs) from intergenic or antisense transcription without protein-coding potential have been identified as key regulators of chromatin remodeling (Pauli et al., 2011). These ncRNAs contribute mechanistically to the establishment of chromatin structure and to the maintenance of epigenetic memory (Malecová and Morris, 2010). The ncRNAs can be broadly classified into two categories: i) infrastructural and ii) regulatory ncRNAs. Infrastructural ncRNAs are constitutively expressed and include rRNAs, tRNAs, snRNAs and snoRNAs (Kaikkonen et al., 2011). Regulatory ncRNAs can be classified into small and long ncRNAs. Small ncRNAs, which include miRNAs, siRNAs and piRNAs, have significant role in RNA degradation and translational repression. Their involvement has been shown in modifying chromatin and target gene expression or guide methylation via RNA interference (RNAi) and other pathways (Collins et al., 2011; Holoch and Moazed, 2015). Long ncRNAs (lncRNAs) are typically polyadenalated and longer than 200nt have been shown to coordinate the access to or dissociation of regulatory proteins from chromatin, recruit chromatin modifiers/remodelers to regulate transcription, and even in

the modulation of the senescent phenotype (Bischof and Martínez-Zamudio, 2015). One of the well-described examples that involve lncRNA is *XIST* RNA mediated X chromosome inactivation which has been elaborately discussed in Chapter 1.6. Recently, a novel class of promoter-associated RNAs (PARs) that help keep PcG complexes tethered to silenced promoters and allow them to be easily released upon gene activation, and enhancer RNAs (eRNAs) that help bring enhancers and promoters together through chromatin looping have been shown (Kaikkonen et al., 2011). Another study has identified the existence of stable nuclear dsRNAs (ndsRNAs) that escape processing and may interact with regulatory engines whose subset has been shown to be interacting with mitotic complex (Portal et al., 2014).

1.6. Role of epigenetic modifications in cancer

As described earlier, regulation of chromatin compaction and DNA accessibility in spatio-temporal manner through epigenetic signals ensures appropriate genomic responses across different developmental stages and tissue types. Given its significance in cell fate decisions, deregulation of epigenetic patterns can lead to propagation of diseased state, especially cancer. Few decades ago, it was suggested that epimutations can act as 1 of Knudson's 2 hits (a hypothesis suggesting that both the copies of the tumor supressing genes must be affected for oncogenesis) required for tumorigenesis (Holliday, 1987). Similar to the frequent occurrence of DNA mutations in specific genes (e.g., *TP53* or *KRAS*), high-frequency epimutations are also observed in specific genes (e.g., *VHL* or *CDKNA*) in several tumor types (Baylin and Jones, 2011). The interplay between genetics and epigenetics is also observed in cancer promotion (Choi and Lee, 2013). DNA methylation can generate mutational hotspots for genetic changes and cancer-specific mutations in genes that are directly involved in epigenome organization are observed in multiple tumor types (Baylin and Jones, 2011).

Involvement of DNA methylation in cancer has been well studied. Cancer cells show genome-wide hypo-methylation and site-specific CpG island promoter hyper-methylation (Esteller, 2008). Furthermore, aberrant DNA hypo-methylation can also account for the activation of some proto-oncogenes and lead to loss of imprinting, as in the case of the *IGF2* gene (encoding insulin-like growth factor-2) in Wilms's tumor (Ogawa et al., 1993).

However, the most recognized epigenetic disruption in human tumors is the CpG island promoter hypermethylation–associated silencing of tumor suppressor genes such as *CDKN2A* (cyclin-dependent kinase inhibitor 2A), *MLH1* (mutL homolog-1), *BRCA1* (breast cancer–associated-1) and *VHL* (von Hippel-Lindau tumor suppressor), an observation that has been expanded through the study of the inactivation of microRNAs with growth-inhibitory features by epigenetic silencing (Lujambio et al., 2007; Saito et al., 2006; Toyota et al., 2008). The disturbance of the DNA methylation landscape in transformed cells has been recently supported by the finding of somatic mutations in *DNMT3A* in acute myeloid leukemia (AML) (Ley et al., 2010). Tumor associated loss of 5hmC in several cancers of lung, brain, breast, liver, kidney, prostate, intestine, uterus and melanoma has been observed. Loss of 5hmC in solid cancers is associated with strong reduction of *Tet1* expression. In breast and liver cancers, significant reduction in the expression of *Tet2* and *Tet3* is also observed along with the reduction in *Tet1* expression. Numerous loss-of-function mutations have been identified in *Tet*, *Dnmt1* and *Dnmt3a* in several cancers (Jin et al., 2011).

Disruption of normal patterns of covalent histone modifications is another hallmark of cancer and is observed during early tumorigenic process. One of the most characteristic examples is the overall reduction of the trimethylated H4K20 and monoacetylated H4K16, along with DNA hypomethylation, at repeat sequences in many primary tumors (Esteller, 2007). Several lines of evidence implicated chromosomal translocations in HATs resulting in fusion proteins in malignancies, like fusions of MLL-CBP, MLL-p300 in MLL (mixed lineage leukemias) and similar fusions of CBP/p300 with MOZ in AML (acute myeloid leukemia. Further, AML1-ETO [t(8;21)(q22;q22)], the most frequent fusion protein in AMLs, requires p300-mediated site-specific acetylation to induce leukemogenesis (Di Cerbo and Schneider, 2013).

While aberrant activity of histone modifying enzymes and histone modifications are implicated in tumorigenesis, the process itself may drive translocations and mutations adversely affecting these epigenetic factors (Sadikovic et al., 2008). Studies have shown selective silencing of tumor suppressor gene, *p16* in a mouse model system developed cancer; thus indicating epigenome change alone can trigger cancer (Yu et al., 2014). Thus,

a complex relationship exists between epigenetic modifications and cancer; however, this also provides an ideal target for chemical intervention in cancers. Several inhibitors and modulators of histone deacetylase and methyl transferases have been successfully tested in cancers (Claude-Taupin et al., 2015; Falkenberg and Johnstone, 2014; Spiegel et al., 2012). Deeper understanding of the global patterns of epigenetic modifications and their corresponding changes in cancer can enable the understanding the role of different epigenetic factors and thus enabling the design of better treatment strategies.

1.7. Epigenetic instability of inactive X chromosome in breast cancers

As described earlier, there is increasing evidence to support the notion that epigenetic modifications accompany tumorigenesis. In theory, epigenetic changes that could lead to aberrant expression of oncogenes or inactivation of tumor suppressor genes can contribute to cancer progression (Sharma et al., 2010). The inactive X chromosome (also known as, Barr body) provides an outstanding example of an epigenetic nuclear landmark where chromosome-wide epigenetic silencing takes place. However, disappearance of the Barr body is frequently observed in cancer cells, particularly in the most aggressive tumors (Chaligné and Heard, 2014). As the X chromosome contains many potential tumour-suppressor or cancer-promoting genes, epigenetic instability in inactive X chromosome has been associated with cancer (Pageau et al., 2007).

X chromosome inactivation (XCI) is a dosage compensation method in mammalian genomes for a genetic imbalance coming from dimorphism between homogametic and heterogametic sexes where one X chromosome is inactivated (females have 2 X chromosomes as compared to 1 in males). In 1949, Barr and Bertram first identified a nuclear body within female cat neurons, but not in the corresponding male cells, subsequently named it as Barr body (Barr and Bertram 1949). This dense Barr body was later identified as X chromosome (Ohno and Hauschka 1960). Shortly thereafter, in 1961, Lyon first proposed ‘X inactivation hypothesis’ that the Barr body X chromosome could be of paternal or maternal origin and that it was genetically inactive (Lyon 1961). This led to further work in the field of X chromosome inactivation research (Figure 6).

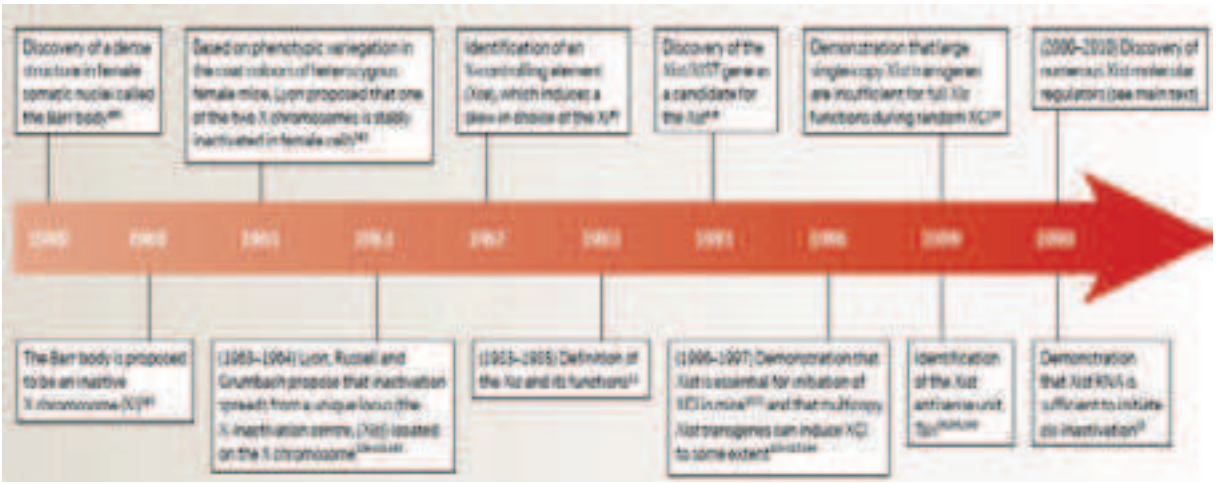


Figure 6. **Major landmarks in random XCI research.** (Taken from Augui et al., 2011).

XIST, a 19-kb long ncRNA in human, transcribes from Xi (inactive X) only (Brown et al., 1992; Hong et al., 2000). XCI proceeds through series of stages namely counting & choice, initiation, propagation and maintenance of silencing (Figure 7). 'Counting' stage is to determine whether XCI is necessary for the cell where the number of X chromosomes and autosomes are counted. 'Choice' stage is when one of the two X chromosomes (either imprinted or random) is chosen for inactivation while the other remains active. The process of counting and choice are overlapping and linked molecularly in the developmental stage by the X inactivation centre (Xic). The Xic contains several non-coding elements, the most important of which are *XIST* and *TSIX. XIST* RNA coats the selected Xi to silence but it alone cannot recapitulate all the roles of Xic. For example, *TSIX*, a *XIST* antisense transcript plays a key role in the choice of which chromosome will be inactivated and is a repressor of *XIST* gene and expressed from Xa. In addition, *trans*-interactions have been proposed to allow the cross-talk between two X chromosomes and likely to be involved in choice. Spreading of silencing is made sure by upregulation of *TSIX* in Xa but downregulation of *TSIX* and upregulation of *XIST* in the future Xi. Chromosome-wide silencing spreads from Xic to both sides of the chromosome. *XIST* is required for the maintenance of stable silencing as well, as it is required for the recruitment of other epigenetic factors related to silencing (Augui et al., 2011).

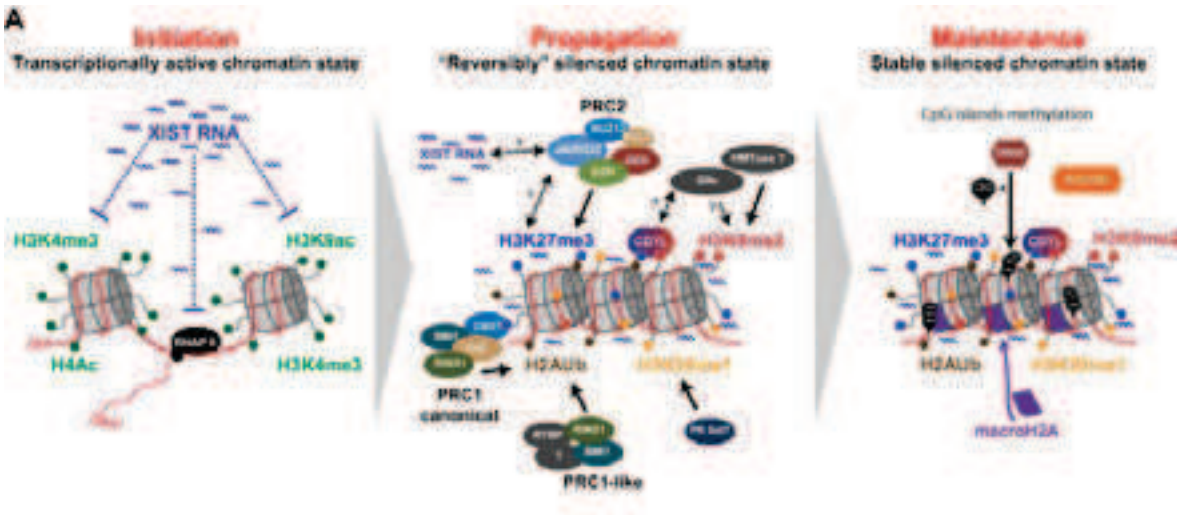


Figure 7. Schematic view of the kinetics of X-chromosome inactivation. Initiation of XCI is associated with expression of *XIST* RNA coating and loss of euchromatin marks (H3K4me2/3, H3K9ac and H4ac), followed by activation of the PRC proteins and the propagation of heterochromatin marks (H3K27me3, H3K9me2, H2Aub1 and H4K20me1). In maintenance phase, promoters of X-linked genes are methylated in DNA with the disappearance of PRC1 and PRC2. (Taken from Chaligné and Heard 2014).

XIST is involved in triggering the inactivation by recruiting the epigenetic marks chromosome-wide like a regular silencing mechanism. X inactivation is an interesting and complex chromosome-wide silencing process which involves co-ordinated epigenetic regulation. Studies focussing on the early changes in chromatin states and structure during inactivation have been reported to assess the role of *XIST* and establishment of silence state in Xi. Loss of euchromatin associated histone modifications like H3K9ac, H3K4me2 and H3K4me3 is the earliest change occurring, followed by global H4 hypoacetylation and passive histone-loss during replication. In addition to these early chromatin changes soon after *XIST* coating, loss of transcription associated factors like RNA polymerase II and nascent transcripts were observed. After one or two cycles, several new histone modifications are recruited on the *XIST*-coated chromosome including H3K27me3, H4K20me1, H3K9me2 and H2Ak119ub1, which are well known repression marks (Chaligné and Heard, 2014).

While the majority of the X-linked genes in Xi are transcriptionally repressed during XCI, a few genes have been shown to escape the inactivation and express from both the chromosomes bi-allelically, termed as 'normal escapees'. There are two types of escapees where one lies within the pseudo autosomal region (PAR) and the other lies outside the PAR. All the escapees from PAR and few from outside PAR have exact homologs on the Y chromosome and show equal expression in both the sexes. However, the expression of the other escapees which does not have any obvious Y-linked homologs can vary considerably depending on the tissue or species. Many of these escapees lie in the short arm of chromosome which gives raise to the hypothesis that the barrier effect of centromeric heterochromatin could be the reason for incomplete silencing. Also these escapees are controlled to not spread as the neighbouring regions are insulated by CTCF (Chaligné and Heard, 2014).

In earlier studies, Barr and Moore described that Barr body is frequently lost in breast cancer which became the evidence of linkage between cancer and Xi reactivation (Barr and Moore 1957). This led to the hypothesis of Xi reactivation being a common event in some cancers. More recent studies suggested the association of Barr body disappearance and over expression of X-linked genes linking the potential role of XCI in tumorigenesis (Ohhata and Wutz, 2013). In some cases, duplication of Xa was also observed in tumors lacking an active chromosome (Chaligné et al., 2015). Two types of mechanisms could explain the loss of Barr body. Epigenetic instability leading to de-condensation of heterochromatin and reactivation of X-linked genes is one possible mechanism but has no evidence to support yet. In another scenario, *XIST* RNA mislocalisation and sporadic Xi reactivation has been observed giving support to random or specific reactivation of certain genes of Xi in cancer cells (Chaligné and Heard, 2014).

There is a growing therapeutic interest in knowing whether epigenetic instability of inactive X chromosome can actually contribute to cancer progression. However, the epigenetic status of inactive X in cancer is less explored (Chaligné et al., 2015). Comparison of allele specific transcriptomic and epigenetic profiling between normal and breast cancer cells could reveal the regions or genes that are epigenetically disrupted from XCI. Novel methods of analyses, such as genetic (SNP6 and Exome-seq), epigenetic

(ChIP-seq) and transcriptomic (nascent RNA SNP6 and mRNA-seq) could be used to understand the allelic and epigenetic status of the disrupted Barr body in breast cancer cells.

1.8. Heritable gene imprinting and disorders

Another mechanism, similar to that of X chromosome inactivation, is known as genomic imprinting, where certain genes are epigenetically marked or imprinted to be silenced in one allele, dependent on the parent-of-origin (Joyce and Schofield, 1998). As opposed to chromosome-wide silencing in X chromosome, imprinted genes are selectively silenced in one allele and they are typically found in clusters of 3-12 genes that are spread over 20-3,700Kb of DNA (Lee and Bartolomei, 2013). The selective silencing of imprinted genes is regulated with the life cycle of the organism (Murphy and Jirtle, 2003). Around ~5-10% of genes expression in the mammalian genome is affected together by XCI and imprinting. Because of parental-origin effects, genetic or epigenetic abnormalities can lead to dosage disequilibrium which in turn can cause human disease syndromes (Lee and Bartolomei, 2013). This dynamic process is complex and it involves various stages namely erasure, establishment, maintenance and implementation of the imprint markings (Murphy and Jirtle, 2003). The process begins with the complete erasing of DNA methylation on starting with the paternal pronucleus within the zygote, while the maternal genome gets demethylated with the subsequent cell divisions. However, imprinted methylation marks present on both the genomes are maintained despite the global demethylation. After complete eradication of methylation, parental-specific methylation is re-established during gametogenesis, in the PGC (primordial germ cells) of the foetus. Remethylation occurs in the sperm postnatally. In the oocytes, the remethylation process is driven by DNMT3 family of protein, DNMT3L and the methyltransferases 3a and 3b. These proteins later also help in recruiting histone deacteylases, altogether these complexes are involved in gene silencing. Parental-specific methylation has to be maintained and carried forward throughout many rounds of DNA replication during growth and development which is carried out by the actions of maintenance methyl-transferases such as DNMT1 (Murphy and Jirtle, 2003).

Imprinting and maintenance is a complex epigenetic mechanism, susceptible to dysregulation at multiple levels. Any dysregulation or altered dosage could result in diverse developmental disorders. As imprinted genes are involved in growth-related pathways, its role has been shown in cancers like Beckwith-Wiedemann syndrome, Wilm's tumor, hepatoblastomas, rhabdomyosarcoma and adrenal carcinoma (Joyce and Schofield, 1998). Similarly, reduced or loss of expression in apoptosis inducing gene *ZAC* has been reported in breast cancer and other primary tumors (Bilanges et al., 1999). A few hundred genes have been identified as imprinted genes (Figure 8). A database (http://www.geneimprint.com/) is also available that provides the list of known and predicted imprinted genes list. Imprinted regions of the genome are associated with several developmental disorders and diseases including cancer due to mutations or impaired regulation leading to alterations in dosage. Similar to XCI, epigenetic and allelic status profiling of known imprinted genes could reveal the role of imprinting loss in breast cancer cells.

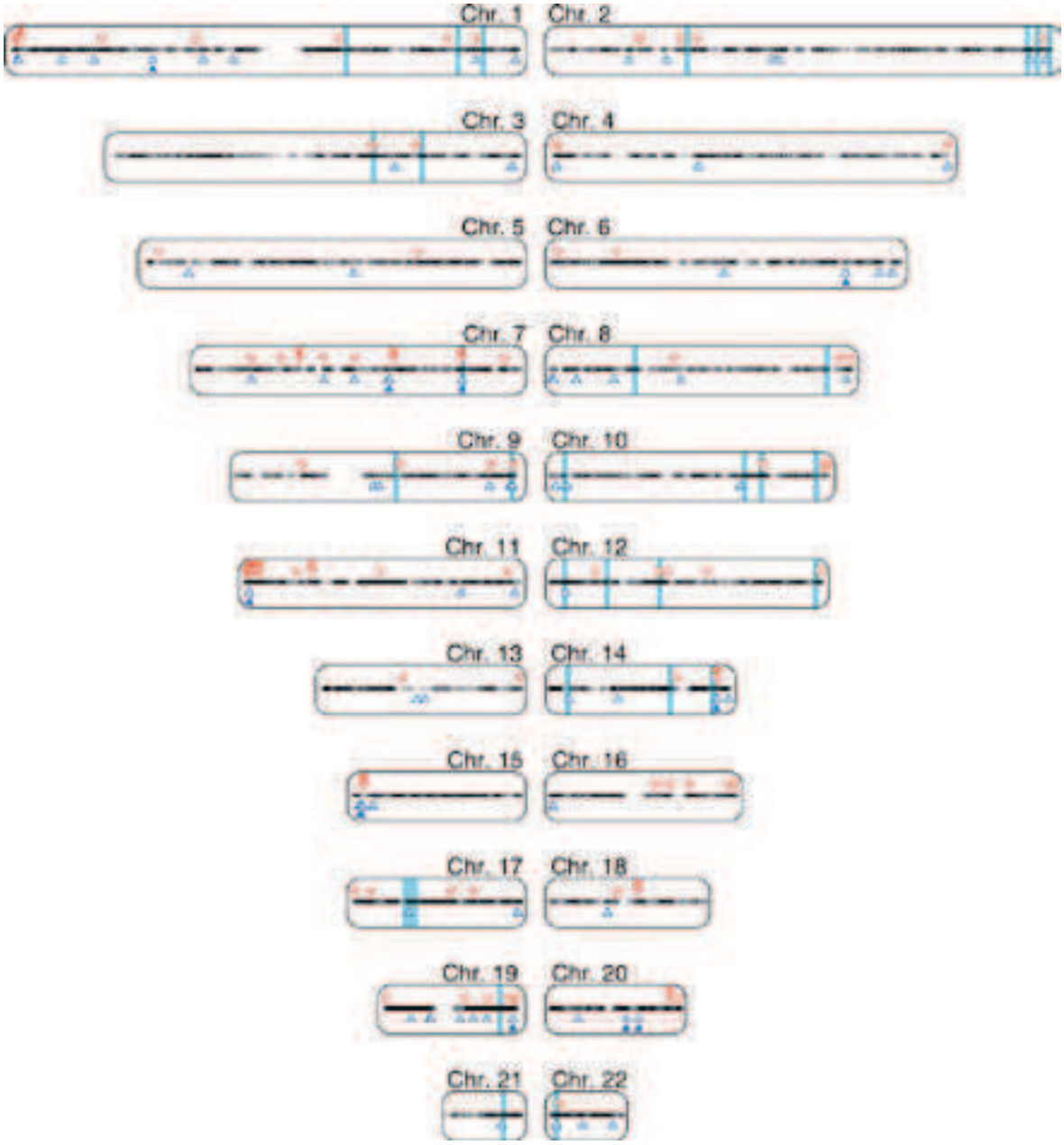


Figure 8. **Genome-wide distribution of identified imprinted genes.** Imprinted genes are highlighted based on their confirmation status and parent-of-origin. On the basis of confirmation status: Filled triangles - proved; Unfilled triangles - predicted to be imprinted with high confidence. On the basis of parent-of-origin: Red downward triangles -Maternally expressed; Blue upward triangles - Paternally expressed; Black dots - b-allelically expressed. Light blue bars highlight a 3-Mb region centered on the linkage regions. (Taken from Luedi et al. 2007).

CHAPTER 2

RISE OF NGS DRIVEN STUDIES IN EPIGENETICS

Chapter 2. Rise of NGS driven studies in epigenetics

Epigenetics is a complex and multi-layered process with potentially profound implications in cell fate decisions including differentiation, cancer, etc. Dissecting how different machineries define functionality of chromatin requires an understanding of their distribution across sequence features such as promoters, gene bodies, intergenic regions, etc. Microarray, a widely used hybridization based technology requires *a priori* knowledge of the genome or the genomic regions to be studied (Hurd and Nelson, 2009). Hence, robust genome-wide studies are required to understand epigenetic mechanism and its role in different cell processes. Next Generation Sequencing (NGS) has become the common medium for global analysis of epigenetic modifications. Large scale NGS based methods are being used to study epigenetic modifications, and changes occurring in different cell types and disease states. This chapter deals with the recent advances in NGS for epigenetic studies.

2.1. A brief history of next generation sequencing technology

The first major foray in DNA sequencing was the Human Genome Project, a 13-year project which was fully completed in 2003 (ConsortiumInternational, 2004). However, even before human genome several other bacterial, viral and fungal genomes were sequenced (Goffeau et al., 1996; Sanger et al., 1977; The Arabidopsis Genome Initiative, 2000; The C. elegans Sequencing Consortium, 1998). Such whole genome *de novo* sequencing assembly of the genome of a particular organism may lead to a better understanding at the genomic level and may assist in predicting genes, protein coding regions, and pathways (Lee et al., 2013). This led to the sequencing of genomes from different organisms and their subsequent characterisation from a functional and evolutionary standpoint. For organisms whose reference genomes are available, resequencing approach is used to better understand its functional aspects. With the basic assumption of existing reference genome as a generic representation of an organism, studies were carried out to understand the changes/differences in an individual genome to identify the inherited deleterious mutations responsible for diseases and disorders. 1000

genome sequencing and several population related studies were executed to provide a comprehensive resource on the human genetic variation (Durbin et al., 2010). Whole genome/exome resequencing studies focussed on genetic level changes such as SNVs, translocations, copy number variations that can potentially influence the gene functionality. However, decades of research on the regulatory mechanisms that control the gene expression expanded the aim of genomic studies to understand these mechanisms as well. This resulted in the use of sequencing in epigenetic modifications (DNA methylation and histone modifications) and transcription factors analysis, thus their effects on expression analysis (RNA) to understand the complex mechanism of cellular processes giving rise to the field of functional genomics. Thus, the sequencing based studies can be differentiated into *de novo* sequencing to assemble genomes and resequencing to study functional genomics. However, sequencing strategies vary between *de novo* and resequencing. For instance, *de novo* sequencing studies require longer reads to resolve assembly related issues in repeat regions. On the other hand, resequencing studies require high throughput to increase the confidence in analysis, but can manage with shorter reads as reference genome is available to map the sequences. Commercial establishments involved in sequencing have also started to focus on developing approaches to address these two distinct requirements separately. One set of platforms (like PacBio and Oxford Nanopore) focus on increasing the length of reads to improve the genome assembly whereas another set of platforms (like Illumina and Ion) focus on increasing the throughput with shorter reads only. An overview of landmarks in NGS driven studies is summarised in Figure 9.

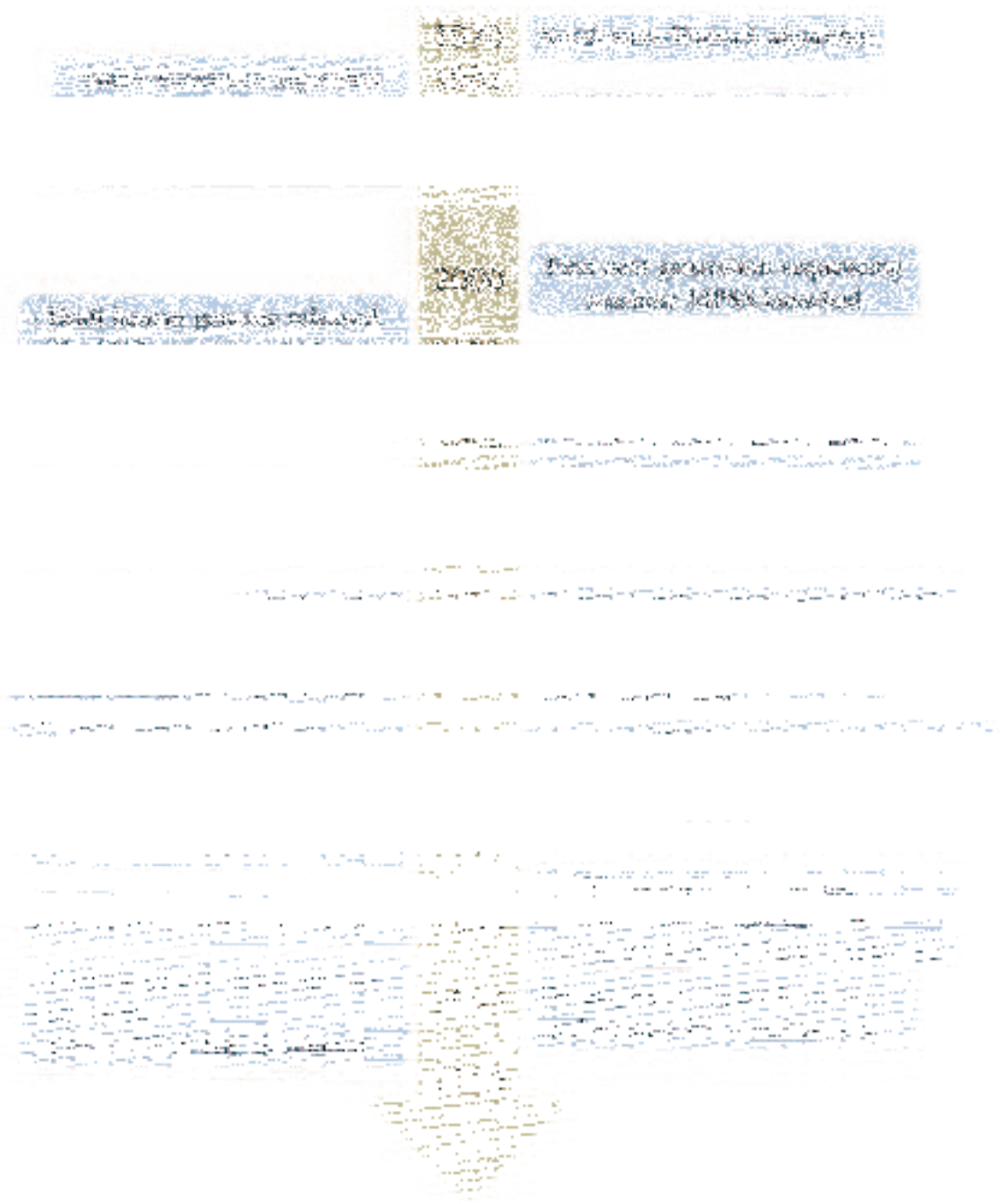


Figure 9. Timeline of landmarks in NGS and bioinformatics.

Advancements in NGS technology and pairing with other technologies led to the development of wide range of applications to target or identify specific regions of interest. For example, coupling microarray as capture technique with sequencing gave rise to exome/target sequencing; thus avoiding the need for sequencing whole genome to identify

mutations, translocations, copy number variations of genes. In 2007, coupling Chromatin ImmunoPrecipitation (ChIP) experiment with sequencing led to the development of ChIP-seq for the identification and characterization of elements in protein-DNA interactions involved in gene regulation (Barski et al., 2007; Johnson et al., 2007). In ChIP-seq assays, specific antibodies are used to target particular DNA-associated proteins (transcription factors, cofactors, histone modifications, etc.) and the pulled down fragments are sequenced. It is followed by an enrichment analysis to identify targeted protein binding regions. Similarly, other approaches like BS-Seq/MeDIP-Seq and ATAC-Seq/FAIRE-Seq were developed to identify genome-wide methylated DNA and open/accessible chromatin regions respectively (Buenrostro et al., 2013; Cokus et al., 2008; Giresi et al., 2007; Jacinto et al., 2008). In parallel, first genome-wide transcriptome profiling using NGS was developed to identify the transcribed regions and its level (Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008). A list of different applications available and their basic overview is summarised below (Figure 10).

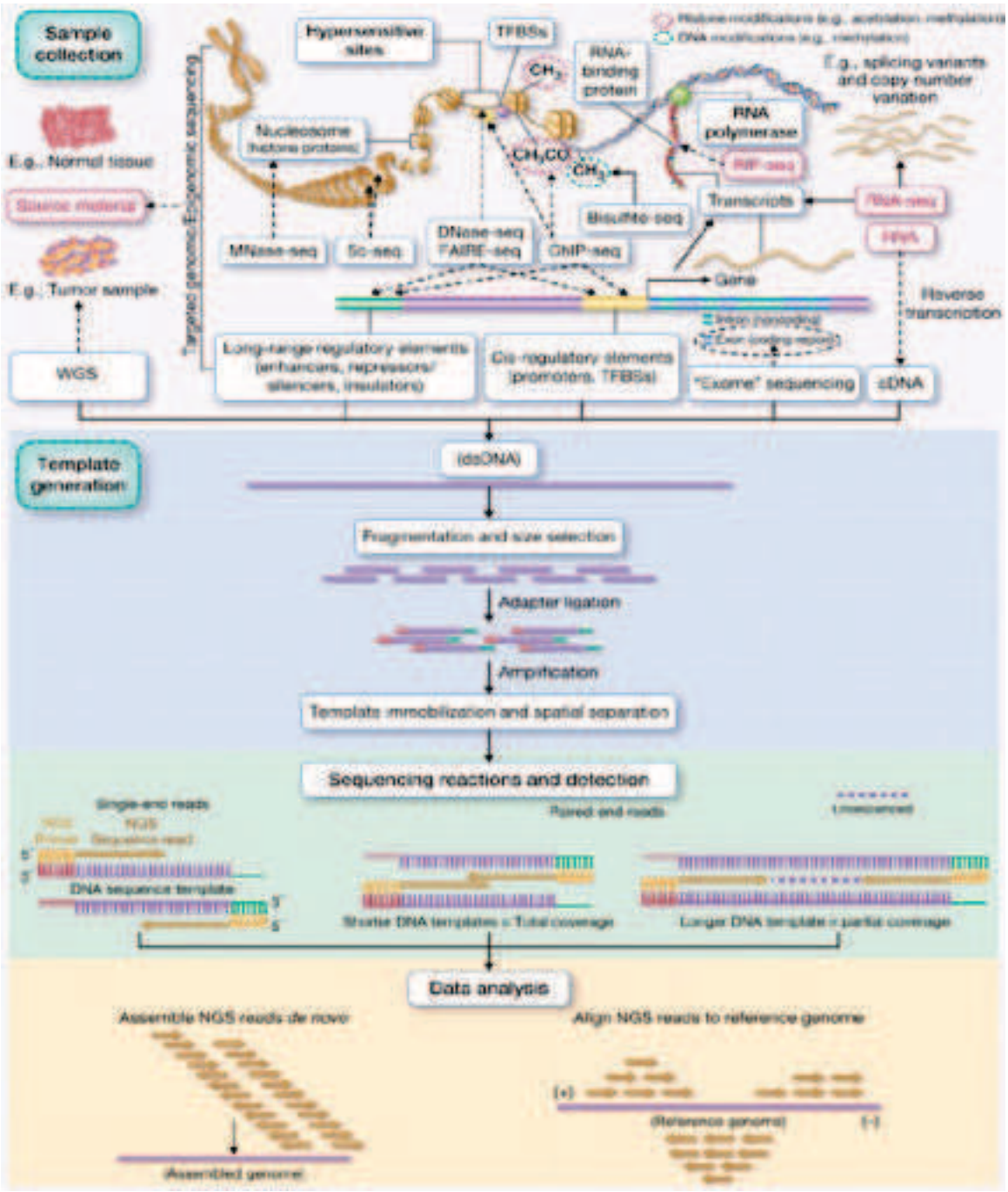


Figure 10. An overview of different NGS experiments workflow. NGS experiments consist of four phases: sample collection (purple), template generation (blue), sequencing reactions and detection (green), and data analysis (orange). Different techniques and methods have been developed to study the various aspects of chromatin architecture and their influence on gene expression. Each technique can have broad applications, depending on the source and nature of the input material, and are described in the glossary. (Taken from Rizzo and Buck 2012).

There has been constant exploration and innovation in sequencing technology to make it more robust and accurate. There are different sequencing platforms such as Illumina, SOLiD, Ion, Helicos, and PacBio are available. Each platform has its own advantages and this depends on the experimental setup and different parameters including but not limited to, need for short or long reads, throughput, error rate and cost effectiveness. Table 2 lists different parameters for existing platforms.

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000	Illumina HiSeq 2500	Ion Proton
Sequence yield per run	0.3-15Gb	20-50 Mb (314); 0.1-0.2 Gb (316); 1Gb (318)	100 Mb	30Gb	600Gb	900Gb-1Tb	10Gb
Run Time	5-55hrs	2 hrs	2 hrs	10 days	11 days	6 days	2-4hrs
Reported Accuracy	> Q30	Q20	< Q10	> Q30	> Q30	> Q30	Q20
Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %	NA	NA
Read length	<300b	~200b	Average 1500b	<150b	<150b	<150b	~200b
Paired reads	Yes	Yes	No	Yes	Yes	Yes	Yes
Insert size	700b	250b	10Kb	700b	700b	700b	NA
Typical DNA requirements	50-1000ng	100-1000ng	~1 µg	50-1000 ng	50-1000 ng	NA	NA
Number of reads	25M	0.6M (314); 3M (316); 5.5M (318)	50K	320-640M	3 billion	4 billion	60-80M

Table 2. Technical specifications of different platforms. Illumina platform tends to outperform the rest of the platforms in terms of sequencing yield and it is widely used in most of the studies. Ion Torrent PGM and PacBio RS generates relatively lesser yield, however the run time is very short, hence it can be used for quick sequencing for finishing (filling gaps and resolving conflicts) genome assembly. While error rate and quality in other platforms are in lower level, PacBio RS is heavily affected by error rates and low sequencing quality. However, PacBio RS II is claimed to have lower error rate with new SMRT technology. (Compiled from the corresponding company website specifications and Quail et al. 2012).

One of the most widely used and cost-effective platforms is Illumina, given its high throughput with relatively low error rates (Quail et al. 2012). A brief summary of Illumina sequencing is described in Figure 11, as the next chapter (chapter 3) mainly focuses on Illumina data. Recently, Oxford technologies developed nanopore based single-molecule sequencing approach where read length could reach up to 30Kbs. However, it is still in the testing phase and has not been made commercially available yet. We have also participated in the testing phase and tried hands-on Oxford Nanopore MinION sequencing (Figure 12). We are unable to share the results due to the existing non-disclosure agreement with Oxford Nanopore technologies.

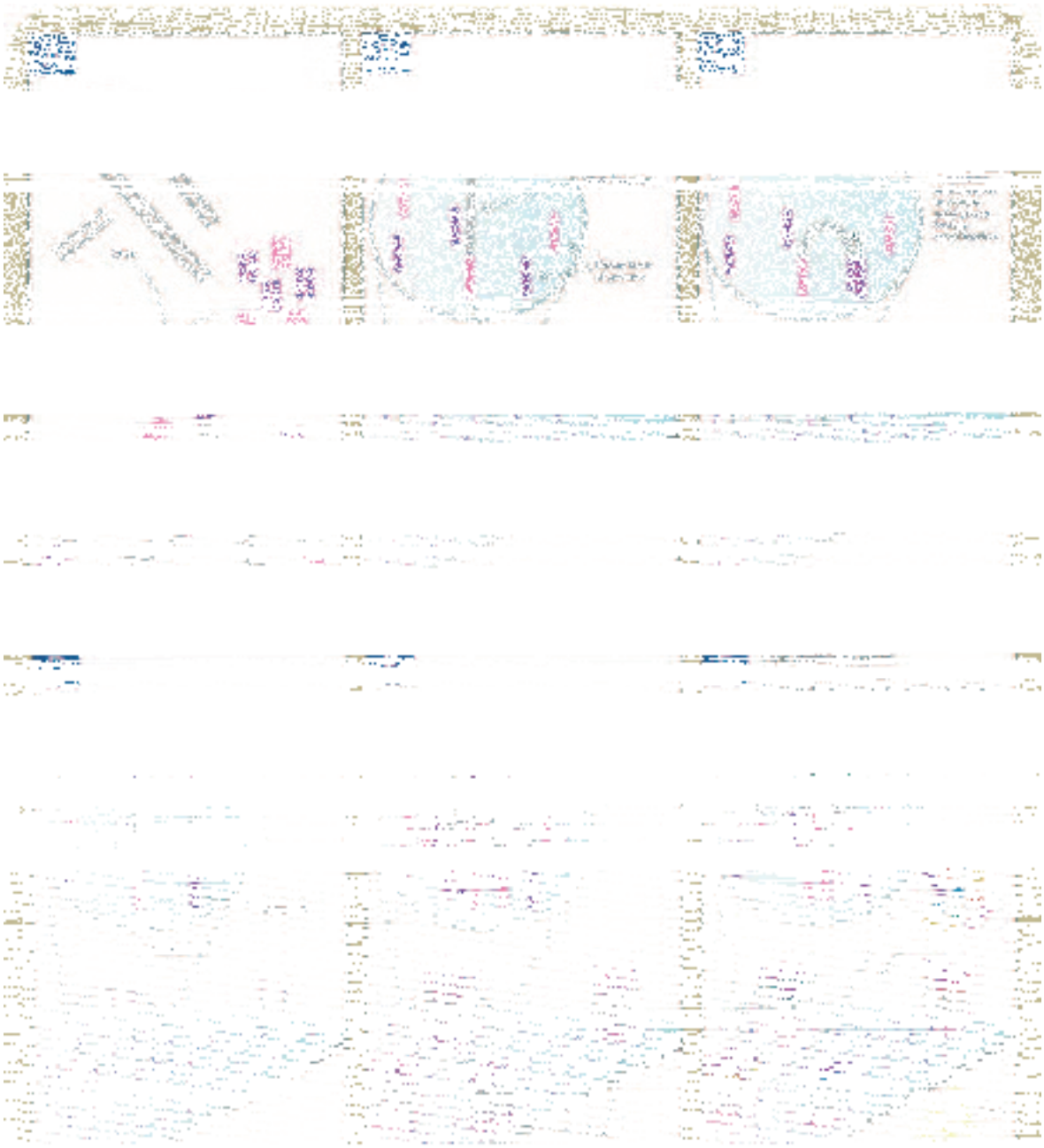


Figure 11. Illumina sequencing chemistry. (A) The DNA sample of interest is sheared to appropriate size (average length 200-700bp) either using sonication or enzyme based digestion depending on the study need. (B) The ends of the fragment are polished, and two Illumina sequencing adapters are ligated to the fragments. Ligated fragments are amplified using specified set of PCR cycles. (C) Illumina uses 'bridge amplification' reaction in the flow cell for polymerase-based extension. (D) Priming occurs as the free hanging end of a ligated fragment "bridges" to a complementary oligo on the surface. The enzyme incorporates nucleotides to build double-stranded

(**Figure. 11 continued**) bridges on the solid-phase substrate. (E) Repeated denaturation and extension results in localized amplification of single molecules in millions of unique locations called “clusters” across the flow cell surface. (F) The first cycle of sequencing consists first of the incorporation of a single fluorescent nucleotide, followed by high resolution imaging of the entire flow cell. These images represent the data collected for the first base. This cycle is repeated, one base at a time to a specified sequencing length, generating a series of images each representing a single base extension at a specific cluster. Base calls are derived with an algorithm that identifies the emission color over time (Compiled from Illumina documentation).

Figure 12. Oxford Nanopore MinION sequencing machine during our testing phase.

2.2. Chromatin immunoprecipitation (ChIP) sequencing for exploring genome function

ChIP-seq is widely used in most of the epigenomic studies. A typical ChIP-seq and related sequencing approach will follow four main steps. (i) cross-linking of the cells using formaldehyde, (ii) shearing of the chromatin using sonication or enzyme based digestion, (iii) pull-down (ChIP) of the DNA fragments that are bound to the protein of interest, and (iv) sequencing the pulled down DNA fragments (Figure 13) (Landt and Marinov, 2012). Sequenced reads are aligned to a reference genome and the identified peaks are annotated in a genomic context.

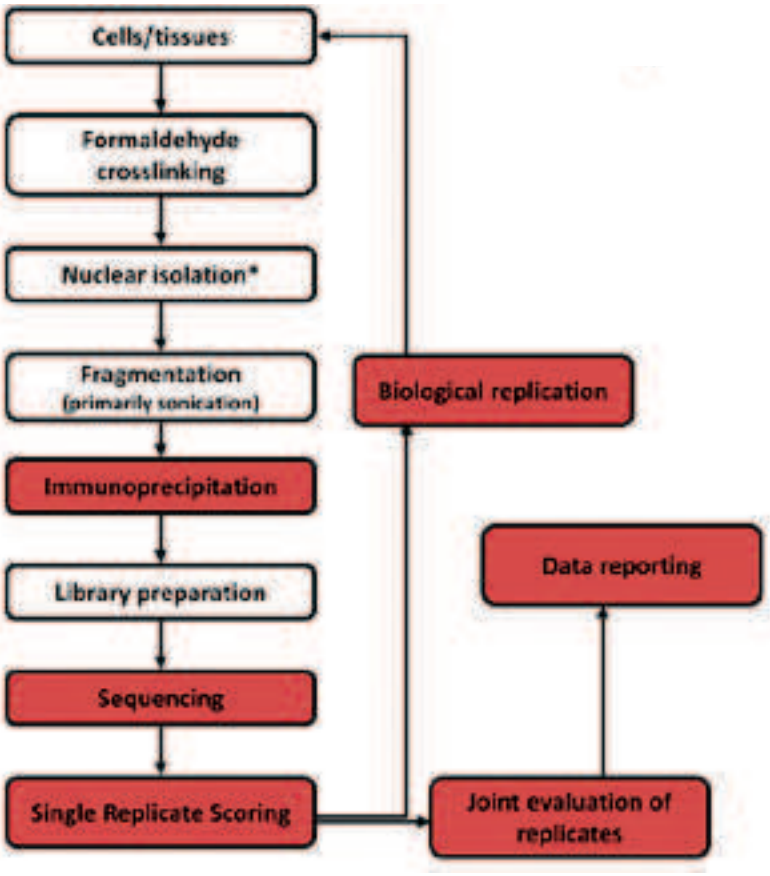


Figure 13. An overview of ChIP-seq methodology. The ChIP process enriches the cross-linked proteins or modified nucleosomes of interest using an antibody specific to the protein or the histone modification. Purified DNA can be sequenced on any of the next-generation platforms. (Adapted from Landt and Marinov 2012).

ChIP-seq studies can result in linear analysis of the associating protein binding regions with nearby genes. The distance (range from 1-20Kb in literature) that is considered for such annotation is very arbitrary and ambiguous. However, given the knowledge of the complex hierarchical organization of chromatin, increasing evidence suggests that distant chromatin regions interact spatially. Chromatin confirmation technologies have identified interactions between gene promoters and distal regulatory elements where chromatin loops bring them together (Göndör and Ohlsson, 2009). Hence, associating peaks in enhancer or promoter regions to nearby genes is not always true. In such case, long range interaction applications such as HiC and ChIA-PET data would help to identify the interaction between enhancer/promoter regions and genes. For instance, a peak in an enhancer region

can be compared with its interactome annotations to determine whether it is interacting with a nearby gene or with a distant region/gene that can potentially be regulated.

2.3. Caveats in NGS driven studies

Though ChIP-seq provides high resolution and sensitivity to results, it also poses significant challenges stemming from both the sequencing technology and the experimental setup. First, sequencing technology related biases are common to all applications. Sequencing errors, GC-bias and PCR related bias are the important biases in sequencing related technologies (Ramachandran et al., 2015). While there are qualities attributed to each sequenced base to evaluate its confidence, sequencing errors *per se* do not have much impact in enrichment analysis as long as it does not affect the alignment accuracy. But to increase the accuracy of the alignment, quality related trimming/filtering is recommended. Data from Illumina has been reported to have sequence specific errors following certain motif regions due to lagging strand dephasing. These sequence-specific errors are consistent in all reads and appear like true variations (Nakamura et al., 2011). Existing variation callers provide strand bias indicator to filter out such sequence specific errors which are represented in reads coming from one of the strand only (DePristo et al., 2011). GC rich regions and PCR related bias result in uneven coverage and over-representation of sequences resulting in false enrichments. GC bias is well documented in Illumina sequencing and GC content normalization is recommended to avoid false positives (Cheung et al., 2011). Over-representation of sequences by PCR due to low library complexity is very crucial in enrichment analysis. For example, when there are accumulation of reads in particular region due GC-bias or clonal reads, a regular peak caller can identify it as true enrichment event. However, this accumulation of reads may not follow a typical peak pattern; some may appear like one resulting in false positive results. Most of the existing tools have a systematic option to exclude such clonal reads (PCR induced over-amplified reads) in ChIP-seq analysis. The rationale behind the need for over amplification is that current ChIP-seq method requires abundant starting material in the range of 1-20 million cells per IP. Studies with less number of cells available invest in more PCR cycles to attain the required amount. To avoid such PCR mediated

amplification bias, several non-PCR amplification methods have been developed like LinDA (Shankaranarayanan et al., 2011).

Secondly, inherent experimental related biases in ChIP due to antibody efficacy pose another challenge in the accuracy of the analysis. Differences in specificity of antibody from different commercial suppliers and batches can bring differences in its performances. Teytelman *et al*., showed that around 238 euchromatic loci (termed as 'hyper-ChIPable') displays high enrichment irrespective of target in *Saccharomyces cerevisiae*. Such enrichments were not a consequence of sequencing related artifacts as confirmed by ChIP-qPCR. This localization of unrelated proteins, including the entire silencing complex to the most highly transcribed genes was attributed to a technical issue with immunoprecipitation (Teytelman et al., 2013). Apart from these technology and experiment related biases, informatics related biases could also bias the results. Effects of such biases and possible solutions are discussed further in the next section.

CHAPTER 3

A DETAILED BIOINFORMATICS PIPELINE FOR CHIP-SEQ STUDIES

Chapter 3. A detailed bioinformatics pipeline for ChIP-seq studies

The rise of next-generation sequencing technologies requires more robust and efficient bioinformatics support. There has been an exponential increase in the data obtained from these high-throughput sequencing technologies to provide higher read coverage in the analysis. Consequently, the need for sophisticated multi-sample analysis to compare different samples like normal vs tumor, different cell-lines has increased to understand the differences in system. Apart from this demand for novel approaches, there is a scope for improving the existing methods by introducing proper quality control and efficient algorithms to handle the variety of biases in the data.

As Illumina is the most widely used sequencing platform, analyses and quality aspects discussed here will be specific to Illumina. The output from the next-generation sequencing technologies depend on the interplay of complex chemistry, optical sensors and hardware. Quality of the data and analysis pipeline complement each other. A robust and efficient pipeline is essential to bring out the best results by avoiding artifacts and biases. This chapter is structured in the order of ChIP-seq workflow as follows,

- Sequencing quality control to identify bias and filter/trim artifactual reads
- Mapping/Aligning reads to the genome to identify the genomic location of each read
- Experimental quality control to evaluate the quality of experiment such as ChIP, capture, RNA extraction.
- Interpretation of results using integrative and comparative analyses

3.1. Sequencing quality control

To identify sequencing related errors and biases, almost all of the sequencing platforms provide quality of confidence for each base. A Phred quality score is used to represent the quality of each base. Phred quality score ‘Q’ is -10log(P) where ‘P’ is probability value of the base called being wrong. *P*-value of 0.01 would result in quality score of 20 which

means that there is a 1 in 100 chance of that base being miscalled. To give simplified representation, quality values are encoded as ASCII values like ‘A’ for 65, ‘B’ for 67, etc., (Cock et al., 2010). A sequencing quality control and statistics check at every analysis step is essential to filter false-positives and bring more accurate results.

There are multiple factors which can influence sequencing quality like library efficacy, fluidics, optics and assay setup (Dai et al., 2010). To avoid bias or artifacts in the analysis, it is important to remove the sequences that contain incorrect bases from the raw data. Using base quality in the aligners has shown to improve the accuracy of the resulting alignment. For example, in one study around 50% of false positive alignments were eliminated upon using base quality in generating alignments (Li and Homer, 2010; Smith et al., 2008). Hence, most of the aligners have incorporated quality score for matches and mismatches.

3.1.1. Base quality

Illumina’s sequencing chemistry, as described in Figure 11, is cycle based where at each cycle, a base of all fragments in the library is sequenced. At the first cycle, the first bases of all fragments are sequenced, second base on second cycle and henceforth. Hence, it is expected to have particular (or set of) cycle to have more low quality bases. However, it has been well documented that Illumina sequences tend to have fall in quality towards the read tails mostly due to reagents scarcity (Yang et al., 2013). These biases can be easily observed in a ‘per base average quality’ plots (Figure 14). Average quality of each read can show the number of reads with more possible erroneous bases (Figure 15). It is important to perform QC and trim or filter the low quality reads to get maximum reliable yield. There are several freely available tools for quality control and processing of raw data. FASTQC and FASTX Toolkit are widely used for quality control and processing respectively.

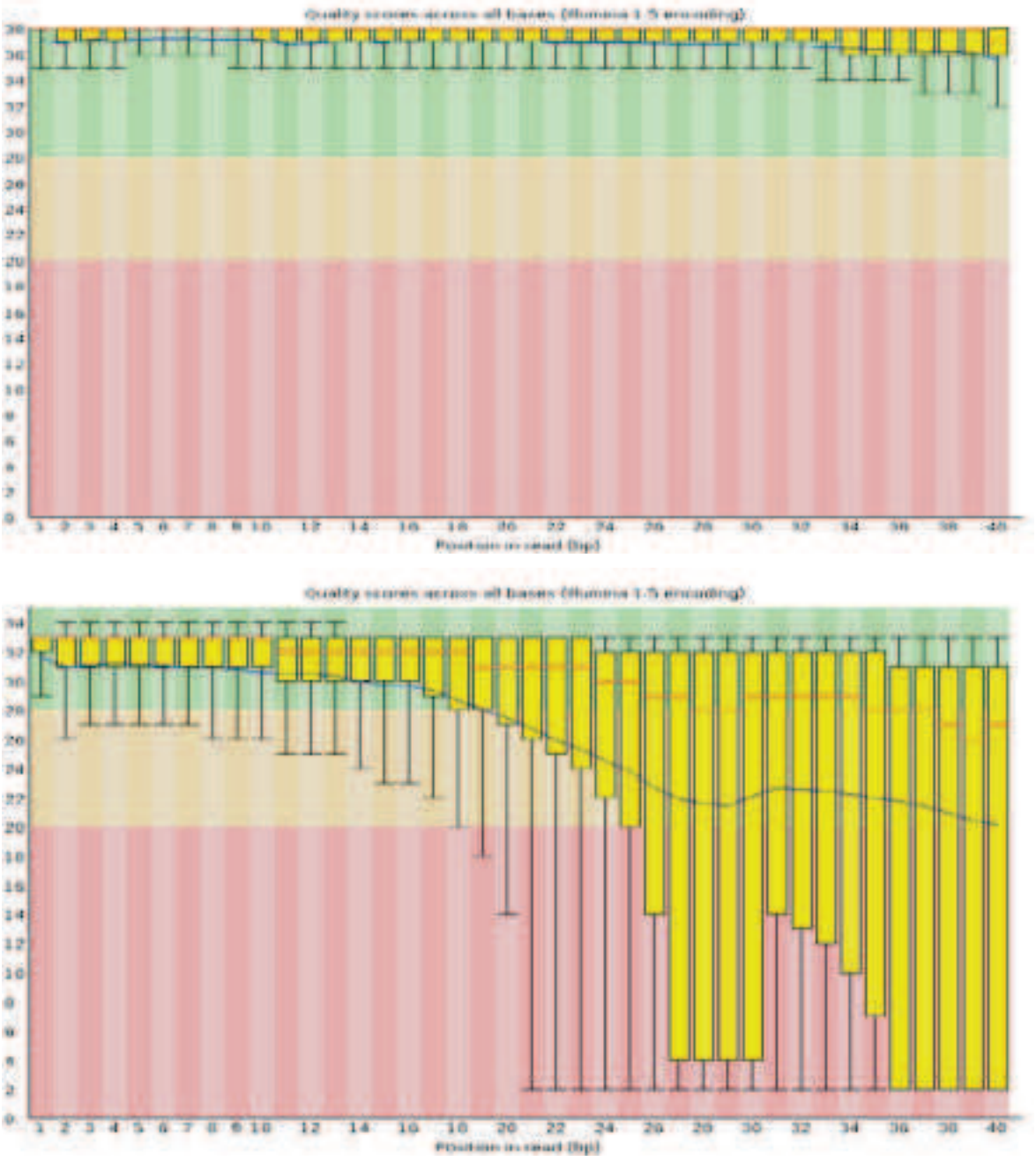


Figure 14. Boxplot illustrating quality distribution for samples with high (top) and poor (bottom) quality. This plot provides overall quality distribution per base from all the reads according to the read length. Background is categorized in three colors: Green (high quality); orange (medium quality); red (poor quality). Blue line in the middle represents the mean quality and red line in each box (yellow) represents median quality. Bottom plot shows data with poor quality where towards end of the reads, there is a rapid fall of quality. (Adapted from FastQC example reports).

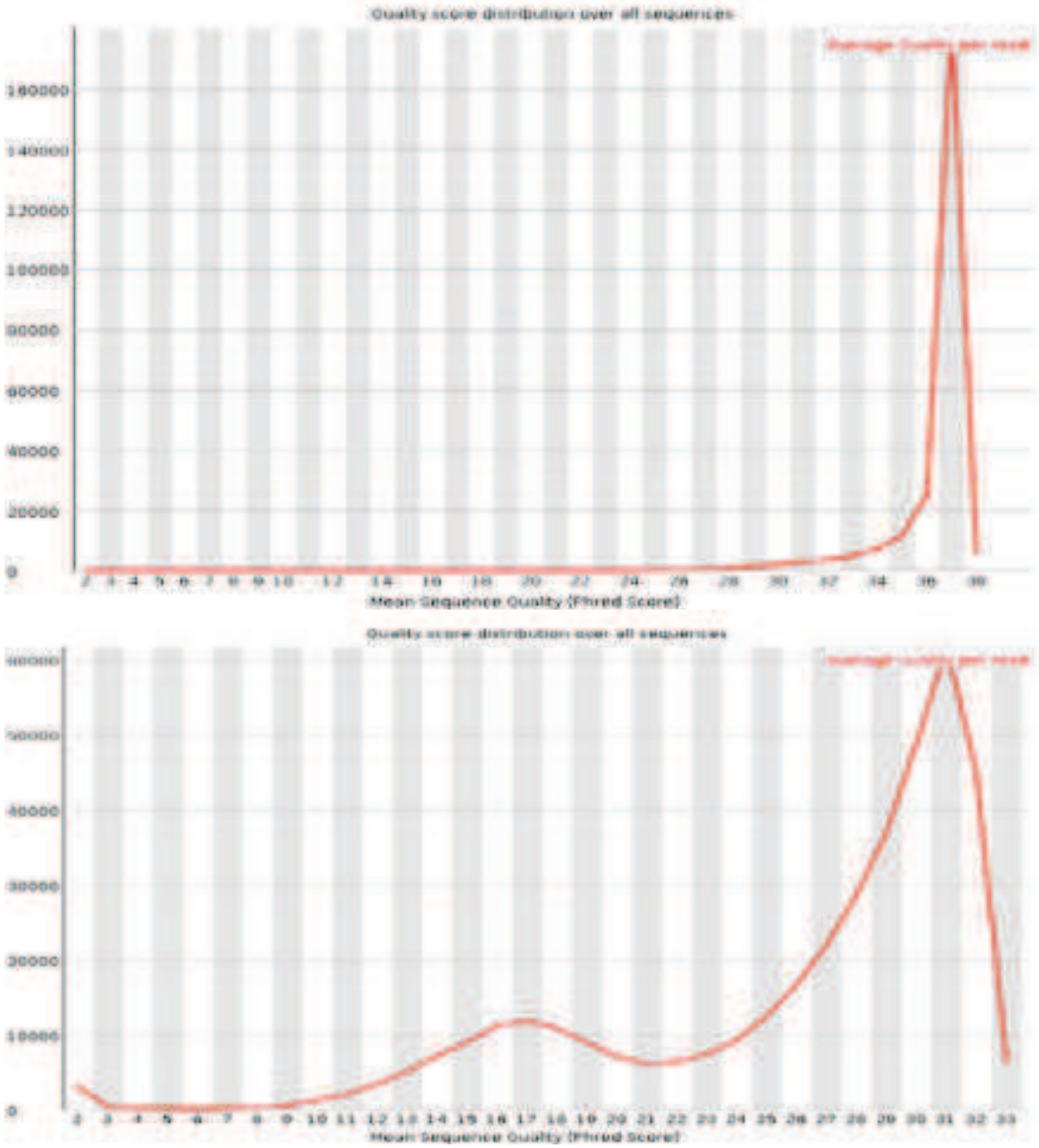


Figure 15. Distribution of average quality per read for samples with high (top) and poor (bottom) quality. This plot provides overall distribution of average quality per read where average of base qualities from each read to illustrate the number of reads with high and poor quality. Bottom plot shows data with poor quality where there is an increase in the middle showing handful of reads with low quality bases in general. (Adapted from FastQC example reports).

3.1.2. Adapter contamination

The length of Illumina sequencing ranges from 36-250bp. When sequencing DNA or RNA fragments that are shorter than the sequencing length, especially in small RNA sequencing, the machines continues to read the 3' adapter sequence. Consequently, the output reads will have adapter sequences at the read tails which will render them unmappable. Though adapter contamination has become less pronounced with recent advancements, it is still observed in datasets especially in smallRNA sequencing approach. It can occur due to over sonication or poor size selection resulting in fragments lesser than sequencing length. There are two possibilities for when such contamination can occur. First, when adapters ligate together to form an adapter dimer. Second, when the fragment length is shorter than sequencing length, sequencing will continue reading the adapter sequences (Patel and Jain, 2012). Adapter dimers are insignificant as they will not be aligned to the genome. However, adapter contamination towards the 3' end of the reads can affects the alignment efficacy in cases where the short sequences at 5' end side of the read matches with the genome. If a high percentage of adapter contamination is observed in QC reports (Figure 16), it is recommended to trim those adapter sequences and recover the maximum usable reads. While alignment or string search based approaches are used for single-end reads, an overlap based approach between pairs have been developed to identify adapter contaminated reads perfectly for paired-end Illumina data (Bolger et al., 2014).

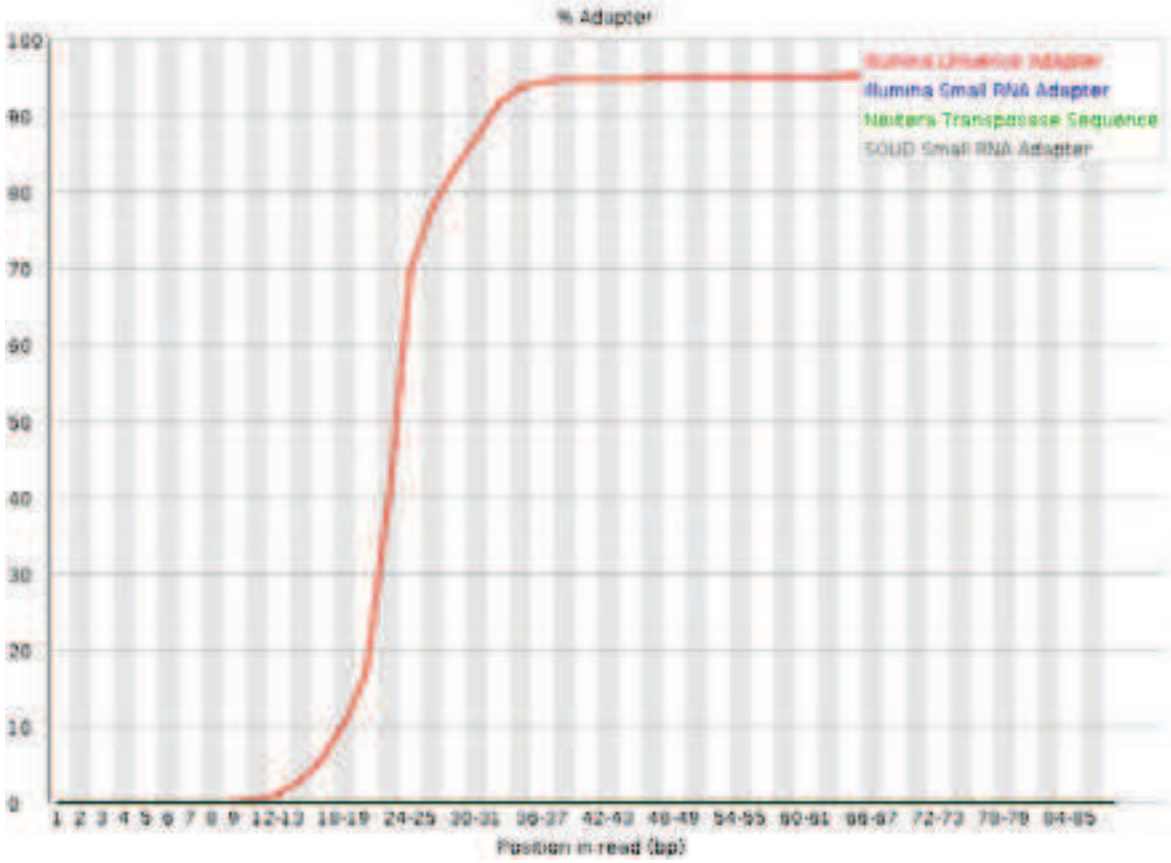


Figure 16. Percentage of reads with adapter contamination with its positional distribution. Reads can have adapter contamination at different lengths depending on the fragment lengths that are sequenced. This plot illustrates the percentage of reads that have adapter contamination and at which position it starts. (Taken from FastQC example reports).

3.2. Mapping of sequenced reads to a reference genome

Mapping (also called aligning) short reads to the reference genome is an essential step in any re-sequencing analysis. Next generation sequencing generates millions of short reads or pairs depending on the application and throughput required. Data generated by these technologies has grown exponentially but this increase has come at the cost of bottlenecks such as chimeric reads that arise as a result of overlapping fluorescent spots. These bottlenecks give rise to inaccurate mappings which in turn can increase false-positives in the final output. So far many algorithms have been developed and application specific aligners also are available for mapping NGS reads (Shang et al., 2014). Given the high throughput and larger genome sizes, fast alignment algorithms build auxiliary data

structures called indices, for the reference genome or read sequences or sometimes both (Li and Homer, 2010). While BWA and Bowtie are the most widely used tools, plenty of improvements have been made in these tools and new tools have also been developed. All of the existing aligners can be classified based on two features: gapped/ungapped, and local/global alignments. The only difference between gapped and ungapped alignment is on allowing gaps for insertion/deletion in alignment. It has been shown that gapped alignment increases the sensitivity by a small percentage but does not show a significant reduction in the false alignments. However, gapped alignment is necessary for identifying indels in data and failure to use gapped alignment may generate false positive SNP calls (Li and Homer, 2010). The local and global alignment approach has major algorithmic difference. With local alignment, when full perfect match is not found, aligners will clip the read's end base by base until a match is found. This enables the tools to exclude the erroneous part of the reads like adapters and low quality ends. On the other hand, global alignment expects to map reads end-to-end with few mismatches allowed in the seed region. When there is a perfect match, both algorithms map them correctly. However, when there are a few errors or contamination in the reads, end-to-end alignment would fail to map them. It has been shown that the local alignment shows less false positive alignments without pre-processing (trim/filter) of the reads. With pre-processing, both approach resulted in similar amount of false positives (Yun and Yun, 2014). Choice of the aligner depends on its feature, capability, analysis requirement, and accuracy; not based on popularity or wide usage (Shang et al., 2014). For instance, read split aligner (e.g., TopHat) has to be used for transcriptome for reads in splicing regions. The following table summarizes the features available in different aligners (Table 3).

Alignment tool	Algorithm	Short or long reads	Gapped alignment	Paired -end	Quality scores used?	Local alignment
Bfast	Hashing ref	Short	Yes	Yes	No	Yes
Bowtie2	FM-Index	Both	Yes	Yes	Yes	Yes
BWA	FM-Index	Both	Yes	Yes	No (soft clip LQ read tails)	Yes (BWA-mem)
Mosaik	Hashing ref	Both	Yes	Yes	No	Partial (only to fix paired-end alignment)
Novoalign	Hashing ref	Short	Yes	Yes	Yes	Yes (read tail)
Shrimp2	Hashing ref	Both	Yes	Yes	Yes	Partial (only soft clip LQ read tails)
SOAP2	FM-Index	Both	No	Yes	Yes	Partial (only soft clip LQ read tails)

Table 3. **Comparison of different NGS reads aligners.** LQ represents low quality. Most of the tools have adapted local alignment option given the reads can have low quality tails or adapter contamination.

3.2.1. Unique reads

As described earlier, clonal reads do not represent the biological reality, but instead an over-representation of same fragments multiple times leading to a bias in the signal counts (Meyer and Liu, 2014). As a PCR step is essential for sequencing adapters ligation, few percentage of clonal reads are expected. However, the percentage of clonal reads can vary drastically among samples and they could bias the analysis. Hence it is highly recommended to remove duplicate reads by keeping only one copy of the duplicate reads using tools like SAMTOOLS (Li et al., 2009) or PICARD (Wysoker et al., 2013)

Existing methods for the identification of clonal reads are less efficient. Clonal reads can be identified by the sequence similarity among reads or post alignment based on the

positional similarity. A summary plot from FASTQC will give an approximate estimation of clonal reads in data (Figure 17). Clonal reads removal approach prior to the alignment depends on the sequence similarity among the reads but sequencing errors can make them appear as unique reads. Hence, identifying clonal reads based on the reads with same alignment position is recommended. Nonetheless, the current approach to identify clonal reads for single-end is less efficient. In single-end (most used method in ChIP-seq), only first few base pairs of the fragment is sequenced. It has been reported that sonication of fragments are biased to sequence specific breaks predominantly in CpG regions (Poptsova et al., 2014). Thus identifying clonal reads based on one end of the fragment could result in false positive results. Though a more robust and sophisticated approach is needed, use of paired-end sequencing where information from both the ends of fragments is available, could make clonal read identification more efficient.

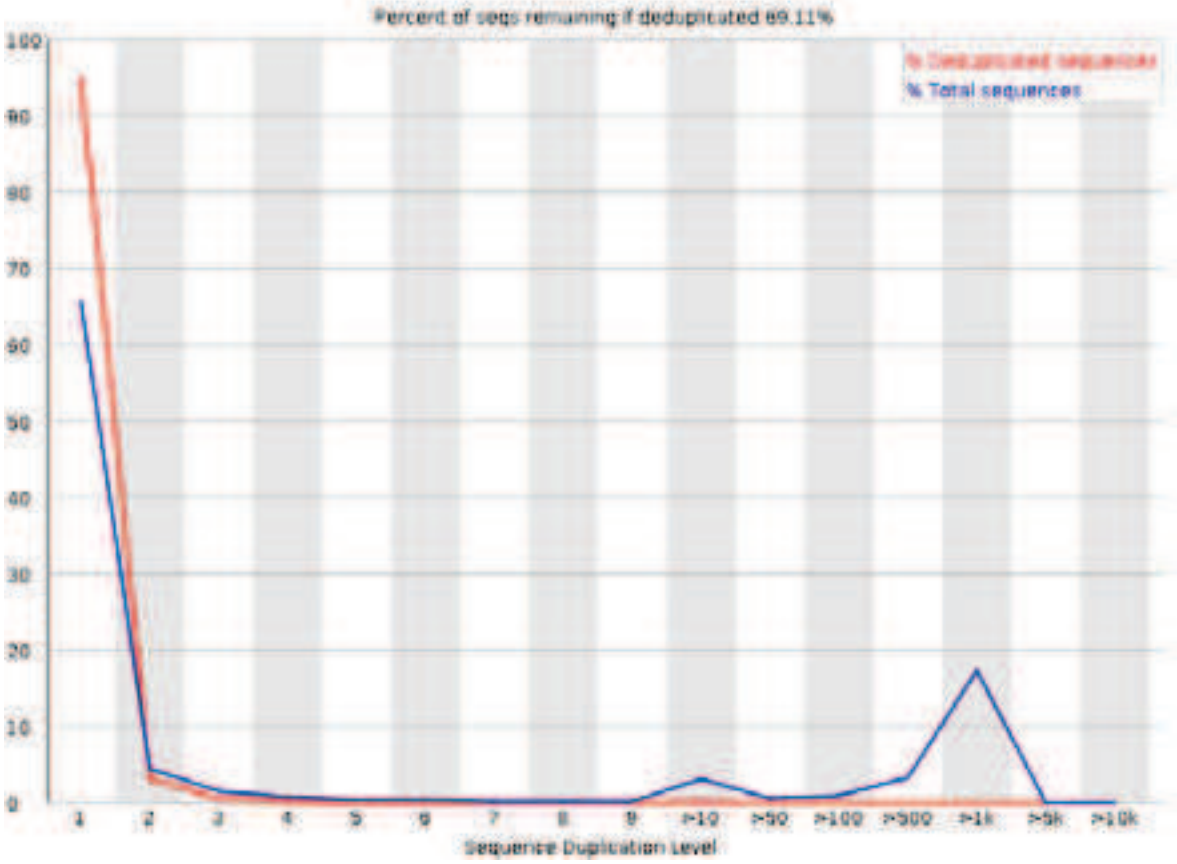


Figure 17. Percentage of unique reads in comparison with total sequences. In this analysis, only first 100,000 sequences are considered to cut down memory requirements. Each sequence is backtracked in the whole file to give a representative count of the overall duplication level. The blue line takes the full sequence set and shows how its duplication levels are distributed. In the red plot the sequences are de-duplicated and the proportions shown are the proportions of the deduplicated set which come from different duplication levels in the original data. (Taken from FastQC example reports).

3.2.2. Uniquely aligned reads

One of the main biases seen in alignment is mapping of reads to multiple regions. As most of the enrichment-based analyses handle shorter reads, uniqueness of each alignment is necessary. A comparison by Heng Li has demonstrated that use of paired-end data has reduced reads aligning to multiple positions as both the reads in a pair has to align in a given fragment size (Figure 18) (Li, 2013). Similarly, longer reads also reduce substantial amount of false positive reads due to increase in the uniqueness of the reads (Derrien et al.,

2012). Given the variable mappability and complexity of targeted genomic regions, reads aligned with low score or to multiple positions should be excluded from the analysis. Most of the existing alignment tools provide a mapping score to evaluate the accuracy of the alignments based on the base quality scores of the read, uniqueness of the match and mismatching bases in the alignment. Mapping quality is the probability value of that particular alignment being wrong. In that context, mapping quality can be used to filter out low quality or ambiguous alignments (Li et al., 2008).

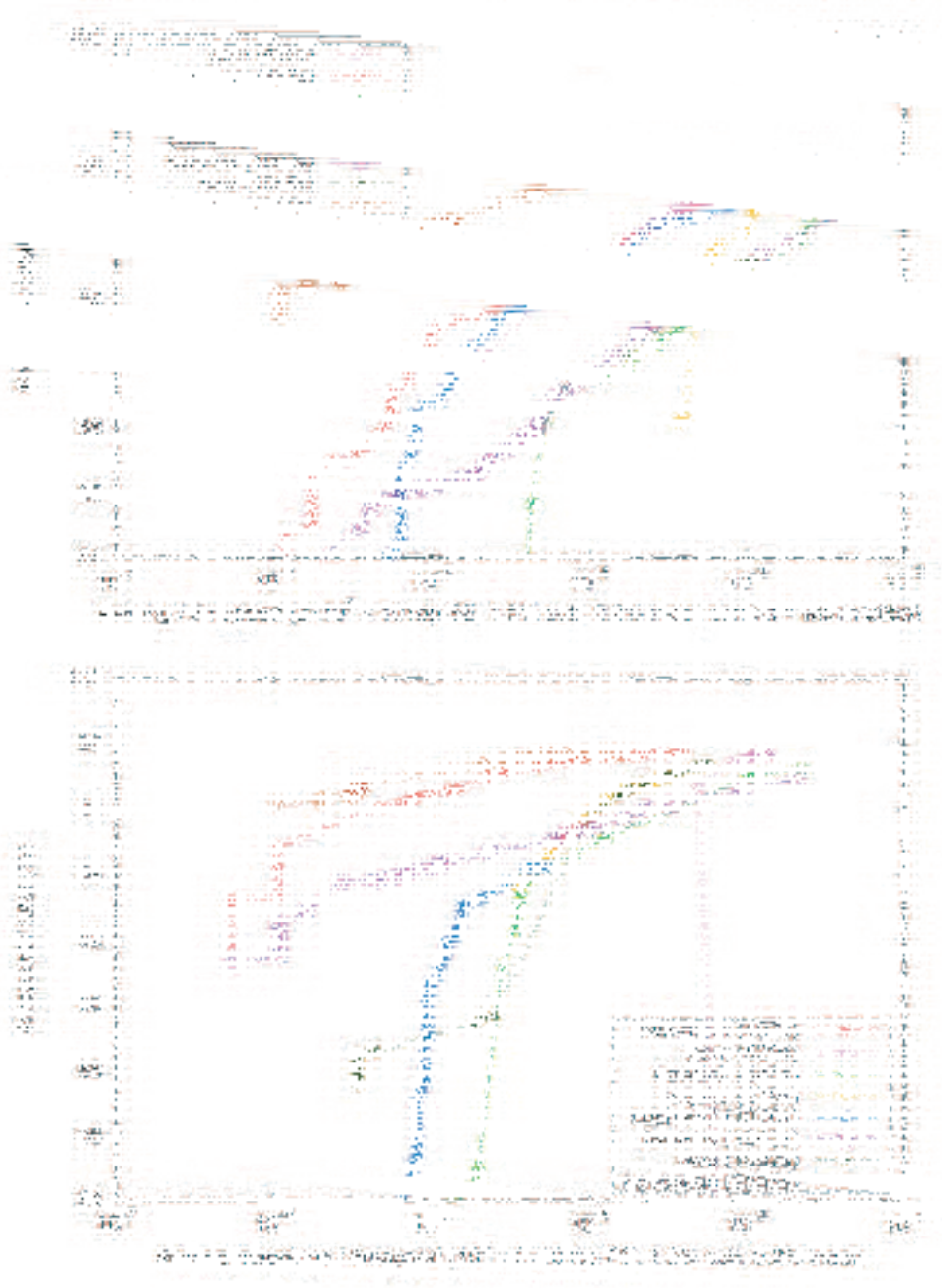


Figure 18. Comparison of different aligners in single-end (top) and paired-end (bottom) data illustrating the false positive rate in both types of data. X-axis represents the ratio of reads that are falsely aligned under different mapping quality cut-off and Y-axis represents the percentage of reads mapped. Bottom panel shows the paired-end data whereas top panel shows the same data aligned as single-end data. It is evident that there is a significant increase in alignment rate under different mapping quality cut-off. (Taken from Li 2013).

3.3. Experimental quality control post alignment

Experimental QC is yet another crucial process to evaluate the quality of the data/experiment altogether. As it is unique to different types of application, different approaches and tools are available to specific analysis. In ChIP-seq, due to the divergence in antibody selectivity/sensitivity and user-chosen sequencing depth, different samples of the same target can exhibit variable distribution of enrichment events (an illustration is shown in Figure 19). For ChIP-seq, ENCODE has recommended fraction of reads in peaks (FRiP) and irreproducible discovery rate (IDR) approach to evaluate the quality of ChIP-seq (Landt and Marinov, 2012). In FRiP, fraction of reads that fall into peak regions identified by a peak-calling tool is calculated, and this metric is used to evaluate the quality of immunoprecipitation. In IDR, the consistency between significant peaks between replicates is used a metric to evaluate the reproducibility of results, which is in turn used to evaluate the quality of data. Though FRiP and IDR are useful metrics, their analysis depends on the annotation from peak callers. Hence, NGS-QC (detailed discussion in Chapter 5.1), a robust sampling based approach has been developed to evaluate the quality of enrichment data (Mendoza-Parra et al. 2013).

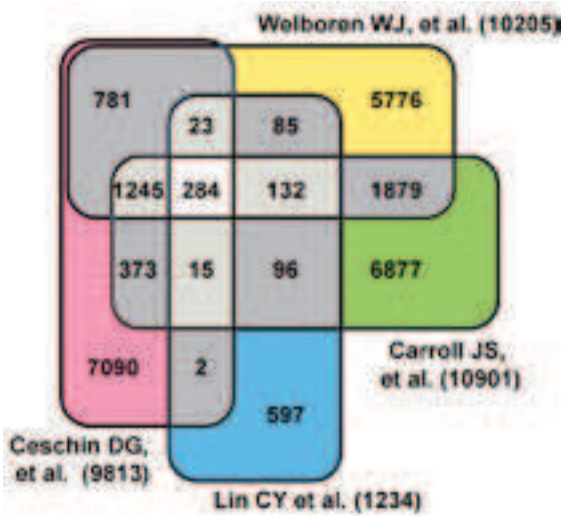


Figure 19. Comparison of ERa peaks across different samples. Diagram shows the number of ERa peaks from three MFC7 datasets (Carroll et al. 2006; Lin et al. 2007; Welboren et al. 2009) and one H3996 dataset (Ceschin et al., 2011) that are common across samples. There was significant difference between three MCF7 samples of same target highlighting the disparities among datasets for the same target. (Taken from Ceschin et al. 2011).

To verify the RNA integrity of mRNA-seq data, 'ngsplot' can be used to verify the enrichment in gene body of all genes. As RNA is sensitive to degradation by post-mortem processes and inadequate sample handling or storage, RNA integrity number (RIN $\geq$ 7) has been suggested as a systematic RNA integrity control prior to sequencing (Pérez-Novo et al., 2005; Thompson et al., 2007). However, it is recommended to verify the enrichment pattern of the RNA to confirm that it is not degraded, as the enrichment would be skewed towards the 3' end (Figure 20).

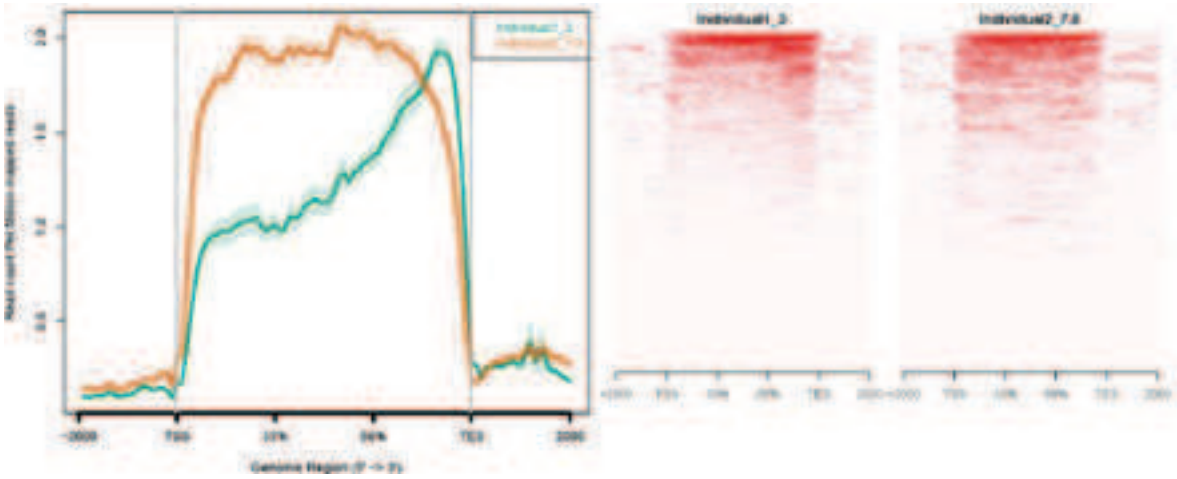


Figure 20. Illustration of skewed enrichment in RNA degraded transcriptome data. An average intensity and heatmap plots over gene-body region highlights a sample's (green color) enrichment bias towards 3' end due to RNA degradation. (Taken from ngsplot example reports).

Most of the recent aligners report the alignments in SAM/BAM format which is widely accepted by many downstream tools (Li et al., 2009). While SAM format provides very detailed alignment information, few of those are important in identifying and filtering faulty or ambiguous alignments. They are,

1. Mapping quality – 5th column, Phred score of alignment being wrong
2. CIGAR value – 6th column, matches/mismatches and gaps/clipping details
3. Paired-end alignment – 8th and 9th column, to identify paired-end alignments not falling within range of given average fragment size
4. First best and second best alignment score tags – extra tags to identify first and second best alignment scores; but these scores are specific to aligner and algorithm

Qualimap is a resourceful tool that provides different plots and criteria to evaluate the quality of alignments (Garcia-Alcalde et al., 2012). BAMTOOLS can be used to filter alignments based on different criteria and tags provided by the aligners (Barnett et al., 2011).

3.4. Pipeline used in our studies

FASTQC provides a detailed report with illustrative plots to assess the sequencing quality in different aspects. All of the data used in our study are subjected to sequencing quality control using FASTQC prior to analysis. The main focus is on base quality, clonal reads and contamination statistics to assess the quality of data. It helps to identify data with poor quality or contamination such that it can be excluded from analysis to avoid bias and can be resequenced. However, with the introduction of incorporating base quality in aligners and addition of local alignment feature, trimming/filtering of the reads are left to the performance of aligners. In that regard, BWA-mem aligner is used for our studies as it can perform local alignment with gapping feature, thus increasing the alignment rate and accuracy. As mentioned earlier in Figure 18, BWA has been shown to have less false positive alignments compared to others. BWA also provides effective mapping qualities for each read that are used in few following tools, especially in variation calling. Further, Picard 'MarkDuplicates' can be used to filter clonal reads effectively than SAMTOOLS rmdup, as it considers clipping and gaps in the alignment. BAMTOOLS is used to filter alignments with low mapping quality (alignments with MQ <10) and to generate alignment statistics. As different aligners use different models to report alignment score, we used our own custom scripts to identify first and second best alignment score tag to filter out reads aligning at multiple positions. If both the first and second best alignment scores are equal for the reads, then those reads are considered as ambiguous alignments and removed from the file. NGS-QC is used to evaluate the ChIP-seq data quality for the reasons discussed in Chapter 3.3. For multi-sample or multi-dimensional based analysis, we evaluate the quality prior to comparative/integrative analysis to avoid biases arising from poor quality datasets.

3.5. Interpretation of cumulated read profiles

3.5.1. Peak detection to identify protein binding regions

Peak calling is the most essential step in ChIP-seq data analysis. It can be defined as the identification of genomic coordinates with accumulation of reads that are indicative of protein binding (Wilbanks and Facciotti, 2010). There are four major steps in the peak calling algorithm namely (i) signal profiling, (ii) background modeling, (iii) peak identification, and (iv) significance analysis.

Signal profiling can be defined as building read count intensities in a sliding window of fixed width across the genome, replacing the tag count at each site with the summed value within the window centered at the site. Consecutive windows exceeding a threshold value are merged. Most of the ChIP-sequencing is single end where only one end of the fragment is sequenced. But in general, the average fragment length after chromatin immuno-precipitation pull down will be around 150-300bp. Most of the ChIP-seq based tools extend the reads to their fragment length such that their combined density will infer a single peak where the summit corresponds closely to the binding site. However, usage of paired-end data can simplify such arbitrary estimation of fragment length and extension thus providing relatively accurate binding sites. Also, alignment accuracy will be increased with paired-end information (Wilbanks and Facciotti, 2010). A new method to identify DNA-protein binding regions at near single nucleotide accuracy has been developed. ChIP-exo, a combination of ChIP-seq and lambda exonuclease digestion (exo) is used to trim the longer DNA fragments on one strand to within a few base pair of the crosslinking point. Thus, it provides higher resolution to identify exact binding regions than regular ChIP-seq (Rhee and Pugh, 2012).

Accurate identification of real peaks by distinguishing the enrichment events from background signals is still challenging. The background model consists of an assumed statistical noise distribution or a set of assumptions that guide the use of control data to filter out certain types of false positives in the treatment data. Poisson distribution based approach is generally to model the background noises. In Poisson distribution, total

number of reads aligned is assumed to be evenly spread in the genome based on which a threshold is set to distinguish peak from background (Pepke et al., 2009).

With the signal profile and background modeling, enrichment events that exceed the predetermined threshold in the sliding window are identified as candidate peaks (Pepke et al., 2009). Subsequently, a control input data (WCE or IgG) is used to exclude technical artifactual enrichment events that are seen in input data as well (Szalkowski and Schmid 2011). However, while control datasets used are generally optimal, a few WCE controls (for example, GSM788366 and GSM768313) exhibit enrichment-like artifactual patterns mostly due to GC or alignment related bias leading to true negative annotation in enrichment sites identification.

Most of the peak callers provide a *P*-value (probability value of peak identified being false) or FDR (false discovery rate) value to describe the quality of identified peaks. Different peak callers follow different statistical models to calculate *P*-value or FDR; hence there is no defined standard cut-off to filter false positive peaks. The number of tags or fold enrichment can also be used to rank peaks for its confidence, though not statistical significance (Pepke et al., 2009).

There is no single generic and universally applicable peak caller for any ChIP-seq data. It has to be chosen based on the nature and pattern of the enrichment. MACS is the most widely used and is very efficient to identify short and sharp peaks with ~80% true positive peaks at ~0.1 FDR (Rye et al., 2011). But for histone modifications with broad peaks or island-like enrichments, HOMER or SICER are more efficient (Zhang et al., 2014). Recently, a pattern/shape learning peak caller has been developed called MeDiChISeq. It is a regression-based approach, which--by following a learning process--defines a representative binding pattern from the investigated ChIP-seq dataset. Using this model MeDiChISeq identifies significant genome-wide patterns of chromatin-bound factors or chromatin modification (Mendoza-Parra et al. 2013). To gain a better result, combination of different peak callers and select peaks based on its consistency has been recommended by a study (Houlès et al., 2015).

Peak calling is followed by association of peaks to a gene by. This is usually done by associating peaks with its nearby genes in a given distance ranging around 5-10Kb depending on the target factor. Tools like HOMER and GREAT are being used for such annotation analysis. But it has been shown that transcription factors can have a long range interaction in both *cis* and *trans* manner (Göndör and Ohlsson, 2009). This raises concern over previously described linear annotation approach where these dynamic non-linear interactions are completely ignored. There is a scope for the development of such non-linear annotation approaches. One solution would be using interactome data like those obtained from HiC or ChIA-PET as a reference for long range interacting factors or regions. Along with transcriptome data and interactome data, a correlative analysis could reveal corresponding genes for enrichment events. Currently, an approach is being developed in Dr. Hinrich Gronemeyer's lab to create a database of long range interacting regions using interactome data such as HiC and ChIA-PET.

3.5.2. Multi-dimensional dataset integration

3.5.2.1. Differential enrichment analysis across samples

Differential analysis of ChIP-seq data involves multiple samples to identify differential regulation of genes and is crucial towards identifying cell-specific differences in regulation. Unfortunately, as mentioned earlier, technical differences in the samples arising mainly due to variation in the sequencing depth makes the data directly incomparable. Several methods such as ChIPnorm, ChIPDiff, MAnorm and diffBind have been proposed for the normalization of multiple samples for comparative studies (Anders and Huber, 2010; Nair et al., 2012; Shao et al., 2012; Xu et al., 2008). A spike-in based experimental quantitative ChIP-seq method has been developed to address sequencing depth variation. An exogenous reference is mixed in the library and used as an internal control, followed by linear scaling based on total number of reads from the external reference control (Bonhoure et al., 2014; Orlando et al., 2014). A detailed comparison and the need for new bioinformatic tools to support such analyses are discussed elaborately in the results section and in the attached Epimetheus manuscript.

3.5.2.2. Genome-wide chromatin state prediction

Recent advancements in NGS have facilitated multi-profile comparisons, in which different ChIP-seqs for histone modifications from one or more samples are compared to assess the different chromatin states within a sample or across several sample. Moreover, chromatin profiling of cells during differentiation and tumorigenesis could give a better understanding of the role of epigenetics in cell fate decisions. ChromHMM is one such tool that focuses on the chromatin state annotation (Ernst and Kellis, 2012). Each chromatin mark's enrichment event is binarized into presence (as '1') or absence (as '0') across non-overlapping windows of a genome. A hidden Markov model is applied to different combinations of chromatin marks to predict the chromatin state for each window. GATE is another tool used for chromatin annotation but in a time-course context to understand the dynamics of chromatin state changes over time (Yu et al. 2013). Similar to ChromHMM, GATE also binarizes enrichment events of each chromatin mark across non-overlapping windows of a genome. Additionally, it clusters the genomic windows, such that each cluster shares a similar combination of epigenetic modifications as well as their temporal changes. While such approaches are complex *per se*, both these approaches binarize local read count intensities to annotate genomic regions as enriched (as '1') or not (as '0') and neglect the intensity differences between different local enrichment events. To further enhance enrichment analysis in a spatio-temporal context the development of a tool, which considers such intensity differences would be very useful.

3.5.3. Integrative and systems biology analysis

While there are several tools and approaches are available for specific analyses, there are only very few tools for the integrative analysis of epigenomic and other 'omics' data, from a single sample, from multiple samples of dynamic epigenome studies or from a set of different epigenome samples which are compared with data sets, such as transcription factor cistromes or chromatin interactomes. Correlating epigenome and the corresponding transcriptome data is important for cross validation. Recently EPITRANS (Cho et al., 2013), a database that integrates epigenomic and transcriptomic data from publicly available datasets has been developed. For transcriptomic data, expression values are calculated by RPKM (Reads Per Kilobase of transcript per Million mapped reads) method

in promoter (2.1Kb) and gene body separately. Similarly, RPKM values are calculated for epigenomic data as well. Both RPKM values are transformed into *Z*-scores, and subsequently correlation coefficient value is determined. This coefficient is used to identify genes that are differentially expressed and epigenetically modified. While this is a greatly useful resource for comparative analysis, it is limited to public datasets and particularly to large consortium datasets. Following this, an integrative analysis tool, Epigenomix was developed to integrate transcriptomic and epigenomic data with quantile normalization approach (Klein et al., 2014). This tool uses expression values obtained from transcript abundance estimation tools and calculates read count intensities (RCI) for the promoter region (3Kb) from epigenomic data. These values are transformed into Z-scores. A positive Z-Score corresponds to equally directed differences and a negative score to unequally directed differences between transcriptomic and epigenomic data. But this approach is limited to promoter regions only and read counts of ChIP-seq data are carried out for the whole promoter (3Kb) region which neglects enrichment pattern and signal-noise ratio differences. For large scale studies involving genomic, epigenomic and transcriptomic data, a robust and multilayer analysis approach is a challenging task from bioinformatics perspective. A pipeline has been developed for epigenetic status of inactive X chromosome study which involves genomic, epigenomic and transcriptomic data (Chaligné et al., 2015). Data from these different applications was integrated and cross-complemented in the analysis (Figure 21).

A similar pipeline is being developed to perform the analysis of imprinted autosomal genes. While the previous study focused only on the X chromosome, imprinted genes analysis focuses on the allele-specificity of the genes which have been annotated as imprinted in literature collected from ‘geneimprint’ database.

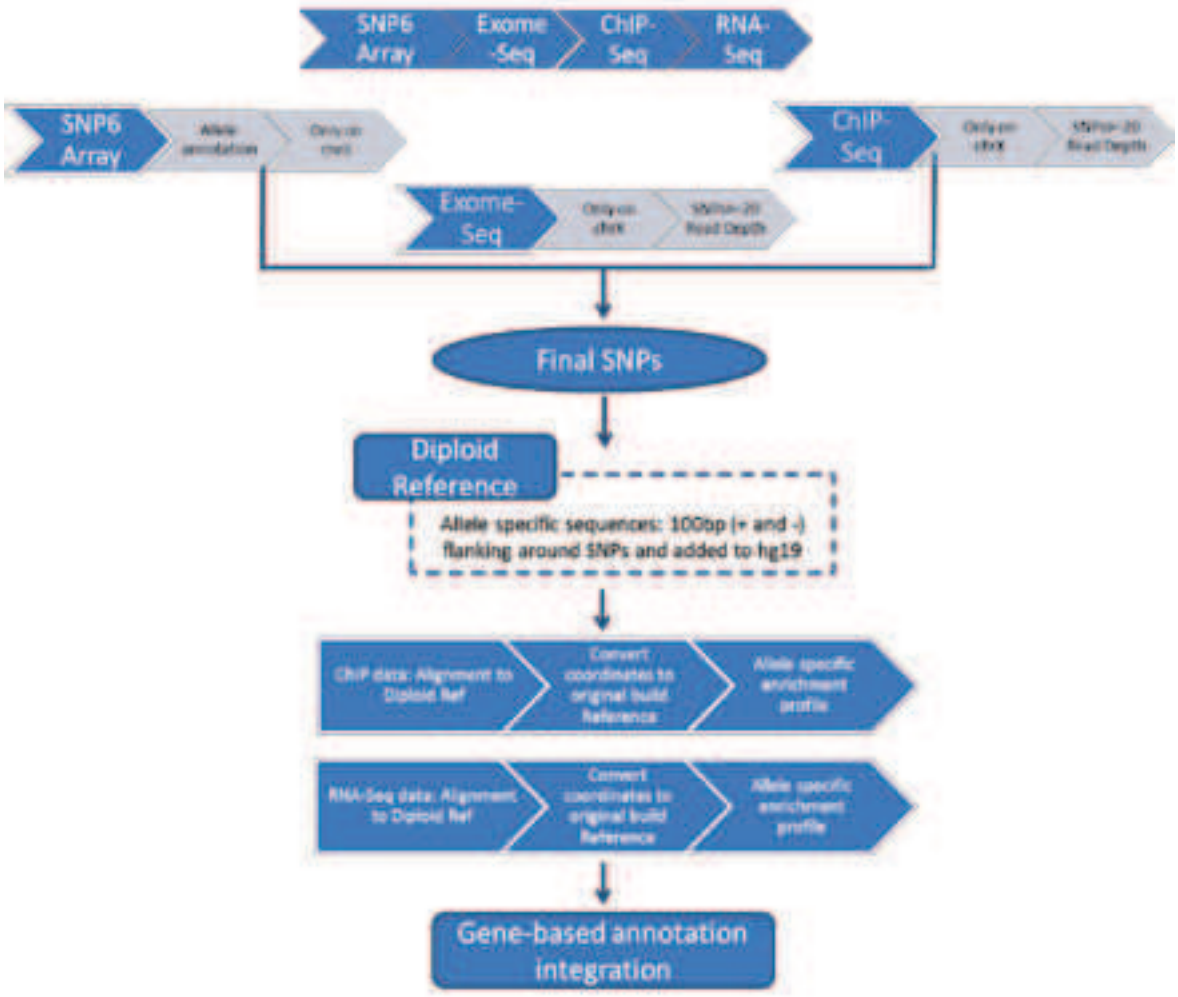


Figure 21. An overall scheme of allele-specific analysis established for X chromosome inactivation analysis perturbation study in breast cancer cells. Cell-line specific variations were identified using epigenomic, transcriptomic and genomic data. Using heterozygous variations, diploid reference genome was built to which reads were re-aligned. Based on number of reads pertaining to each allele, allelic imbalance was calculated for each gene.

System specific dynamic gene regulatory network reconstruction is another integrative approach to map epigenetic and transcriptomic data. While tools like Epigenomix (Klein et al., 2014) are available for such integration of epigenome and transcriptome data, a sophisticated pipeline is required for a dynamic analysis where time-course data is available. In 2007, Ernst et al., have developed the Dynamic Regulatory Events Miner (DREM), a tool that integrates times series and static (or dynamic in DREM 2.0) data

using an Input-Output hidden Markov model. DREM learns to establish a dynamic gene regulatory network by identifying bifurcation points; these are time points at which a group of co-expressed genes diverges (Ernst et al., 2008; Schulz et al., 2012). These bifurcation points are annotated with the transcription factors controlling the split, thus leading to a dynamic model (Figure 22). While DREM provides a collection of curated TF-gene-association database to annotate the graph, another comprehensive collection, called CellNet has been released recently (Cahan et al., 2014; Kim and Schöler, 2014). CellNet has constructed cell-specific gene regulatory networks (GRNs) from 3,419 publicly available gene expression profiles. This annotation base can be used as a reference to build the transcription regulatory database.

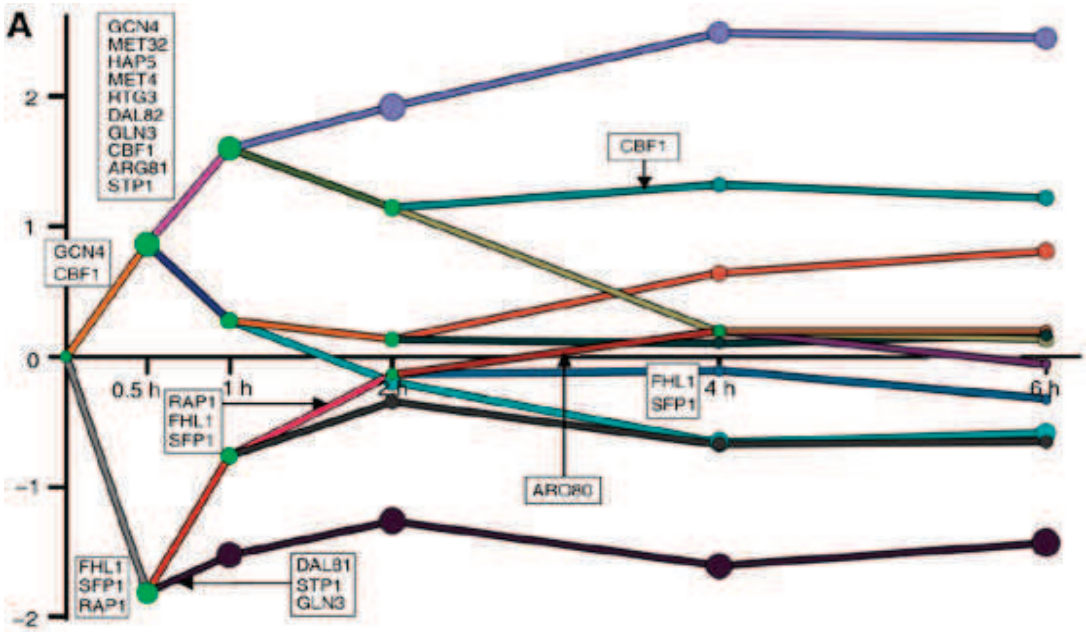


Figure 22. Dynamic regulatory map of yeast response to amino acid starvation. DREM builds dynamic regulatory map from condition-specific binding experiments and time-series expression data. Nodes in the graph represent hidden states and the area of it is proportional to the standard deviation of the genes associated with that node. Significant TFs based on split score is highlighted in ranking order. (Adapted from Ernst et al. 2007).

CHAPTER 4

SCOPE AND SPECIFIC GOALS OF THIS THESIS

Chapter 4. Scope and specific goals of this thesis

Currently, there is a focus on the development of new bioinformatic approaches and tools for different types of epigenomics and ChIP-seq analysis. There is an increase in the interest for multi-dimensional analysis to compare different samples. However, the absence of quality control system for ChIP-seq is a major weakness as the poor quality data can bias such comparative or multi-dimensional analysis. Hence, we wanted to focus on the development of novel tool to evaluate the quality of datasets. Further, we wanted to focus on the addressing the technical differences that are inherent in the ChIP-seq technology by *in silico* tools to normalize the data. This will allow any further downstream multi-profile and integrative analysis to be carried out without such biases. In that context, I aimed to develop novel tools to address such issues, and use those tools to analyse epigenetic status of X chromosome inactivation (refer chapter 1.7).

4.1. Development of tools to evaluate the data quality and normalize technical differences in multi-sample analysis

I focussed on two important technical aspects of ChIP-seq analyses. First, to develop an *in silico* based approach to assess the genome-wide quality of a given enrichment-related dataset. Given the divergence in antibody efficacy and user-chosen sequencing depth, different samples of the same target can exhibit variable distribution of enrichment events (refer Figure 19 for an illustration). We wanted to develop a tool to evaluate the quality of the ChIP-seq and enrichment related datasets. Second, we wanted to develop a normalization tool to correct inherent sequencing depth variation between samples to facilitate fair comparative analysis. As most of the researchers have limited bioinformatics experience and/or access to bioinformatics support, we designed these tools to be highly user-friendly, that also biologists with limited computer knowledge can use them efficiently.

4.2. Integrative analysis for epigenomic and transcriptomic status of the Xi in breast cancer

X inactivation (refer chapter 1.7) is an outstanding example of chromosome-wide epigenetic regulation involving the developmental silencing of approximately one thousand genes. Recent studies have demonstrated the sporadic reactivation of few genes that escape XCI in normal cells (Chaligné and Heard, 2014). Studies have shown the disappearance of Barr body (inactive X) in breast cancer cells suggesting that X chromosome inactivation is compromised in those breast cancer cells (Pageau et al., 2007). However, epigenetic status of inactive X in breast cancers and the extent to which epigenetic instability might account for disappearance of Barr body in some cases is less explored. We wanted to identify the cancer specific escapee genes and study the epigenetic status of chromosome X in breast cancer cell-lines. An integrative analysis of genetic, epigenetic and transcriptomic data is needed to identify the allele-specific expression of the X-linked genes and its corresponding epigenetic status. However, the integration and comprehension of these large scale genomic data need novel bioinformatic tools and approaches, which are developed as a part of my thesis.

CHAPTER 5

RESULTS AND DISCUSSIONS

Chapter 5. Results and Discussions

For each study in this chapter, its corresponding manuscript is attached which describes the methods and results elaborately. A brief summary of methodology and results in the manuscript are discussed for each study below.

5.1. NGS-QC – A quality control system for ChIP sequencing profiles

As described earlier, s comparison or an integration of different datasets requires a specific evaluation of the quality as there can be variability in their profile pattern and technical divergences like the use of different antibodies, sequencing depth and/or immunoprecipitation (IP) efficiency, among many other parameters. For this reason, we have developed NGS-QC Generator, a bioinformatics-based QC system that uses the raw alignment file to (i) infer a set of global QC indicators that reveal the comparability and quality of different NGS data sets; (ii) provide local QC indicators to evaluate the robustness of enrichment events in a given genomic region; (iii) provide guidelines for the optimal sequencing depth for a given target, and (iv) to have quantitative means of comparing different antibodies and antibody batches for ChIP-seq and related antibody-driven studies.

The main rationale behind this method is that beyond a sequencing depth threshold, a ChIP-seq profile changes only in amplitude but not in pattern (Mendoza-Parra et al. 2013). We evaluate this trend by randomly sub-sampling reads (90, 70 and 50% of reads) to see the fluctuations from expected change. Read count intensity (RCI) profiles for original and sub-sampled reads are constructed by counting overlapping reads in continuous non-overlapping windows of the genome. By comparing recovered RCI (recRCI) with original RCI (oRCI), RCI dispersion (δRCI) is calculated for each bin. Then by evaluating the fraction of bins displaying a δRCI within a given interval, a quantitative assessment provides the quality indicators. The detailed statistical model is elaborately discussed in the attached manuscript.

The NGS-QC tool is made publicly available to scientific community in a dedicated galaxy platform (http://galaxy.ngs-qc.org). This galaxy platform has been deployed on our own

powerful servers to accommodate many users simultaneously computationally and in storage. To provide more sophisticated user-friendly experience to users, we have outsourced the design of web portal to a private company called Dreamsoft technologies from India. Since its deployment, a dedicated team has provided automatic pipeline to download the newly released data and process them to keep database updated. Also, several other new modules have been added to the web portal by them.

5.1.1. NGS-QC generator

NGS-QC generator is the dedicated galaxy platform where a user can upload their data to assess the quality and can be compared with the public data in the database. A sample QC report for a public dataset (GSM811204) is shown below for reference. For each sample, NGS-QC Generator generates two output files, (i) a quality control report with global QC indicators and (ii) local QC indicators as BED or WIG files. In the current version, quality control report describes the following information, (i) dataset information namely total mapped reads, the fraction of unique reads (i.e. without the clonal sequences), average read length, genome assembly and target molecule name if provided (refer ‘datasets information’ table in the attached report) (ii) QC parameters section provide details of window size, number of analysis replicates, whether background noise and clonal reads are removed, as these parameters can vary for each report depending on the user’s input (refer ‘QC parameters’ table in the attached report) and (iii) QC results with the global quality indicators and global QC certification (refer ‘results’ table and ‘global QC certification’ stamp in the attached report). The result panel is complemented with a scatter-plot displaying the comparison of original read counts per bin to the recovered counts after multiple random subsampling (90%, 70% and 50% subsampled reads). Furthermore, a global QC certification score (from "AAA" to "DDD" for designating from high to low quality datasets) is provided such that the quality of the analysed dataset is expressed in a rather intuitive manner without the need of getting deep into the assessed quality scores (methodology behind assigning such global QC certification score is elaborately discussed in the attached manuscript). Six illustrative examples of genomic regions display the enrichment patterns complemented by the local QC indicators as heatmap display for the evaluation of robustness of enrichments. When the target molecule identity is provided, a

scatter-plot displays the comparison of quality assessment with other public entries available in the NGS-QC database. For example, in the attached report for H3K4me3 profile, three scatter plots the position of input sample quality among all the publicly available H3K4me3 profiles for each dispersion levels. Such comparison is very useful to verify whether the given profile quality is consistent with the public data.

ATTACHMENT

NGS-QC SAMPLE REPORT

NGS-QC Generator

Data Quality Report

2015-10-05 11:46

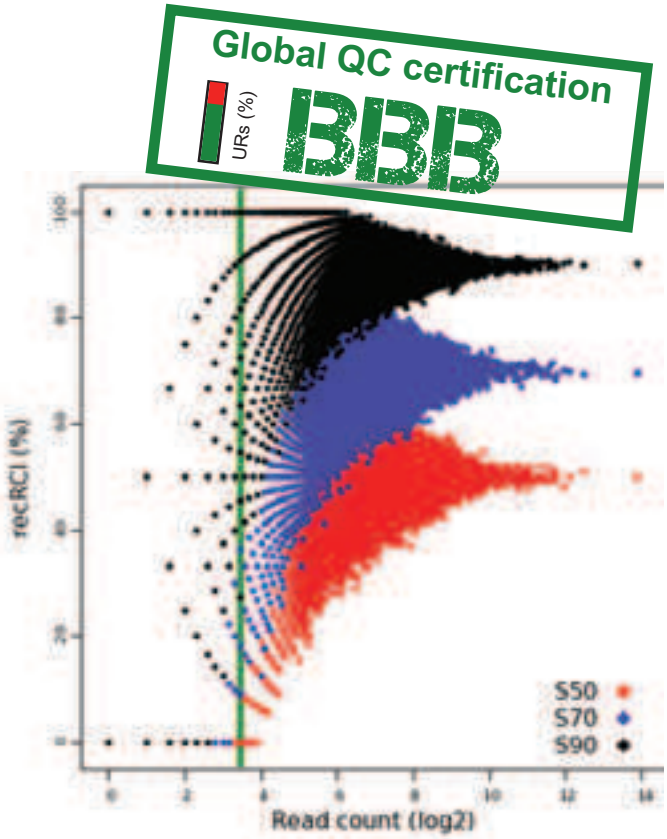
File name: GSM811204_H3K4me3_e2_1M_rep1.bed

Dataset informations	
Total reads	24,800,042
Unique reads (URs)	18,052,251 (72.79%)
Reads's size mean (bp)	49.0
Genome assembly	hg19 (H. sapiens)
Target molecule	H3K4me3

QC parameters	
Sampling percentages	50, 70, 90
Windows size (bp)	500
Replicate number	1/1
Referenced chromosomes	51
Background subtraction	On
Clonal reads removal	Off

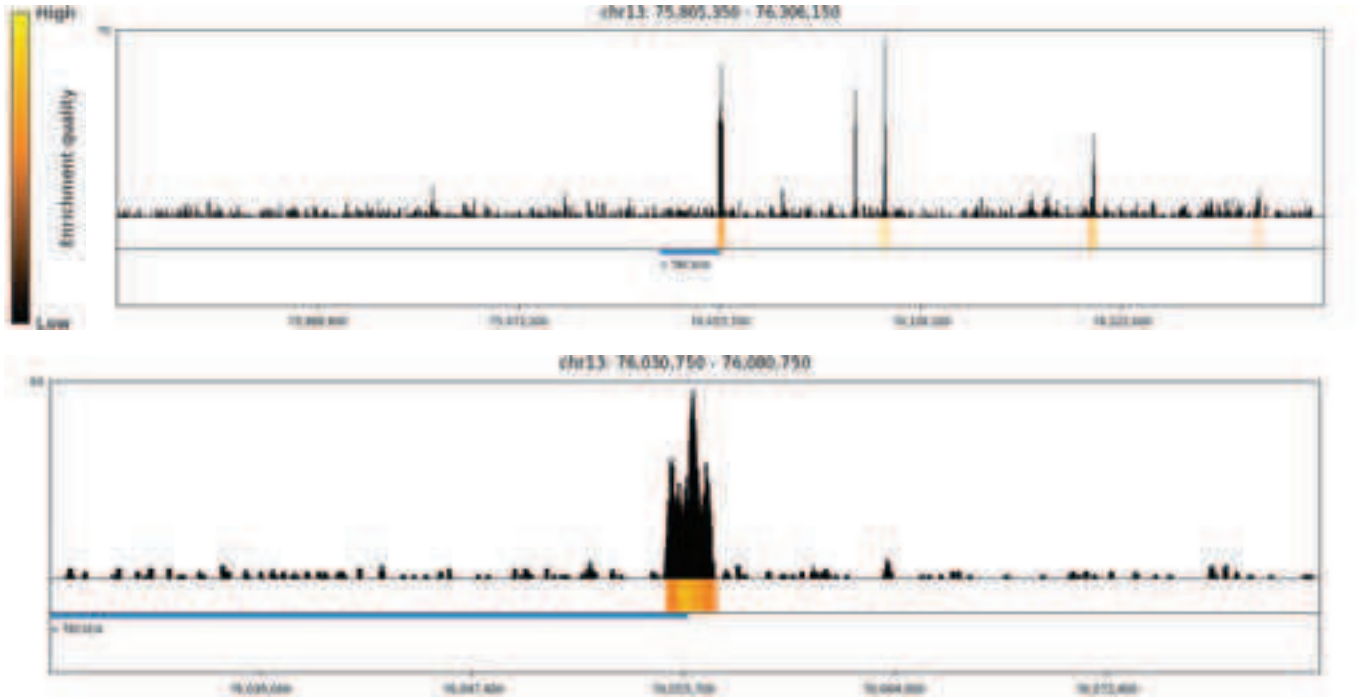
Results		
Considered reads*		24,800,042
QC values denQC (50%) / simQC	2.5%	0.190 / 14.579
	5%	0.999 / 4.789
	10%	3.010 / 2.136

* Reads taken into account to compute the QC indicators.



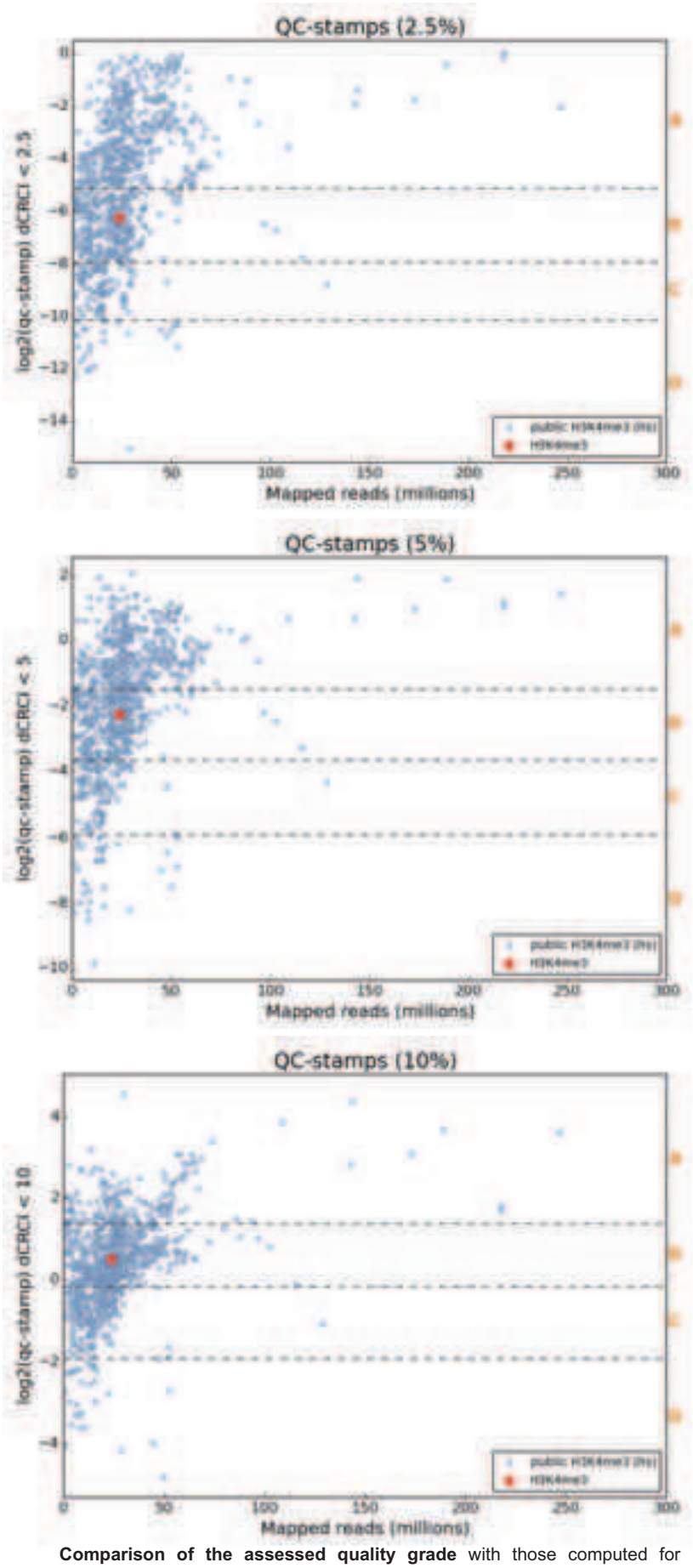
Effect of random sampling on the profile. This figure illustrates the influence of the random sampling subsets (90%: black; 70%: blue; 50%: red) on the recovered read count Intensity (recRCI) per bin. The dark-green vertical line represents the background threshold (11 RCI).

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.



NGS-QC Generator

contact@ngs-qc.org



Comparison of the assessed quality grade with those computed for publicly available datasets currently hosted in the NGS-QC database that correspond to the same antibody target.

contact@ngs-qc.org

5.1.2. NGS-QC Database

In order to facilitate the comparative analysis with public data and to maintain a portal of quality control as reference, we have developed the NGS-QC database by applying NGS-QC approach on a large number of publicly available datasets. Users can retrieve the collection of quality indicators computed for a variety of publicly available datasets in the dedicated website and download QC reports for every dataset. The whole database is built on MySQL and hosted in an independent user-friendly website, which is linked to our laboratory servers (http://www.ngs-qc.org). Regular updates and the availability of additional tools for data analysis are announced on this surface. The query panel (Figure 23A) allows the user to make specific requests through multiple options like the model organism, target molecule, quality grades and also a public identifier from GEO database (GSM or GSE) or from ENCODE consortium (wgEncode). Importantly, the panel is highly user-friendly and multi-modal, such that each of these query options can be used in combinations. The violin plot table below query panel displays the quality scores distributions (QC-Stamp; dRCI<10%) assessed over the whole database content (currently >26,000 datasets), as well as the QC-stamp intervals (from A to D) (Figure 23B). Furthermore, the quality scores distribution per target molecule is displayed such that the users might have a global overview of their associated quality scores.

Figure 23. Display illustrating the database page showing the search panel (top) and violin plots table (bottom). (A) The sophisticated query panel allows users to request through multiple options in a combinatorial manner to yield refined results. For example, one can search for H3K4me3 profiles from different organisms that have specific quality attributes. (B) The violin plot below query panel provides an overall view of quality distribution (QC-Database) and for each target factors individually (e.g., AR – Androgen receptor) that are in the database.

On a given search, results page (Figure 24) provides the following information,

1. **Boxplot table** (Figure 24A) displaying the QC scores distribution for each of the targets included in the request.
2. **Scatter-plot** (Figure 24B) displaying the quality scores (QC-stamp) for each dataset in the context of their total mapped reads (TMRs).
3. **Refinement panel** (Figure 24C left panel) providing further query options to be applied over the initial request to refine the results to specific interest.
4. **Results table** (Figure 24C right panel) displaying a variety of information for each dataset retrieved.

Figure 24. Display illustrating the results obtained after performing a query in the NGS-QC database. A query of H3K27ac profiles from Homo sapiens was made. (A) Scatter-plot displaying the QC-indicators relative to the total mapped reads. (B) Violin plots displaying the different target molecule retrieved on the query. (C) Results table provide several additional information for each dataset (right panel) and the refinement panel (left panel).

5.1.3. Discussion

Over time, there has been a significant evolution in sequencing quality and throughput of data from sequencing machines. However, the experimental quality and throughput is specific to samples and targets. For example, there are around 15 NR2C2 transcription factor profiles in NGS-QC database and all of them exhibit DDD (except two with CCD and CDD) suggesting that either N2C2 enrichments require higher sequencing depth or the antibody specificity for that protein is very poor. Despite the growth in sequencing technology, low quality datasets are still being generated mostly due to multiplexing, sample scarcity, antibody performance, etc. For example, there are 16 H3K4me3 profiles with more than 25 million reads that have CCC quality. This suggests that data with high throughput could also have low quality. We even observed that the input data (control DNA; no antibody use) which exhibits enrichment like patterns could heavily bias the analysis. Hence, assessment to evaluate the quality of the enrichment prior to analysis is imperative to avoid any such biases. NGS-QC tool serves as a robust tool to evaluate the experimental/enrichment quality of datasets.

A large collection of quality assessment for publicly available datasets serves as an excellent repository to compare ones data quality to that of public data. This would help the users to evaluate the performance of a particular antibody by comparing with public data for the same antibody. It also provides a guideline to choose an optimal sequencing depth based on public data. An informative violin plot in the database page provides a detailed summary of quality trend of different targets. NGS-QC database provides an easy way to quickly reuse the vastly available public data. For instance, a comparison of local QC regions (robust enrichments) of a target, across different samples or systems, could reveal samples or systems that exhibit similar enrichment events. More importantly, several studies have reutilised the public data to avoid repeating experiments and further bioinformatic studies to use public data for their application. Hence, NGS-QC database can act as a reference for publicly available enrichment related datasets to select or compare the public data, improving the quality of analyses, reducing bias and to avoid duplication of experiments, thus saving the resources for other goals.

MANUSCRIPT 1

NGS-QC – A QUALITY CONTROL SYSTEM FOR CHIP SEQUENCING PROFILES

Nucleic Acids Research Advance Access published September 14, 2013

Nucleic Acids Research, 2013, 1–13
doi:10.1093/nar/gkt829

A quality control system for profiles obtained by ChIP sequencing

Marco-Antonio Mendoza-Parra*, Wouter Van Gool, Mohamed Ashick Mohamed Saleem, Danilo Guillermo Ceschin and Hinrich Gronemeyer*

Department of Cancer Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)/CNRS/INSERM/Université de Strasbourg, BP 10142, 67404 Illkirch Cedex, France

Received November 3, 2012; Revised August 14, 2013; Accepted August 25, 2013

ABSTRACT

The absence of a quality control (QC) system is a major weakness for the comparative analysis of genome-wide profiles generated by next-generation sequencing (NGS). This concerns particularly genome binding/occupancy profiling assays like chromatin immunoprecipitation (ChIP-seq) but also related enrichment-based studies like methylated DNA immunoprecipitation/methylated DNA binding domain sequencing, global run on sequencing or RNA-seq. Importantly, QC assessment may significantly improve multidimensional comparisons that have great promise for extracting information from combinatorial analyses of the global profiles established for chromatin modifications, the bindings of epigenetic and chromatin-modifying enzymes/machineries, RNA polymerases and transcription factors and total, nascent or ribosome-bound RNAs. Here we present an approach that associates global and local QC indicators to ChIP-seq data sets as well as to a variety of enrichment-based studies by NGS. This QC system was used to certify >5600 publicly available data sets, hosted in a database for data mining and comparative QC analyses.

INTRODUCTION

The recent development of high-throughput sequencing technologies has led to a rapid expansion of studies analyzing the genome-wide patterns of gene regulatory events and features, such as epigenetic DNA and histone modification, and the binding patterns of transcription factors and their co-regulatory complexes, (posttranslationally) modified chromatin-associated factors and chromatin- or transcription-modulatory multi-subunit machineries (1–9). Moreover, the mapping of transcriptomes by RNA-seq (10–13), global nascent RNA sequencing or global run on sequencing (GRO-seq) (14) or ribosome-associated ('ribosome footprinting') RNAs (15), and technologies revealing chromatin conformation are also based on massive parallel sequencing (16–18). A particular challenge is the comparison of multidimensional profiles for several factors, their posttranslational modifications and/or chromatin marks. Indeed, such studies are not easily comparable, as they are performed in different settings by different individuals using different cells and antibodies. Moreover, profiles are established at different platforms with highly variable sequencing depths. As a result, studies performed even with the same cells in different laboratories can differ extensively (3). This presents serious limitations for the interpretation of such global comparative studies and reveals the need for a quantifiable system for assessing the quality and comparability of next-generation sequencing (NGS)-derived profiles and moreover the robustness of local features, such as peaks at particular loci, which are derived from the mapping of read-count intensities (RCIs).

A large number of factors can influence the quality of NGS-based profilings. Particularly in the case of immunoprecipitation-based approaches [e.g. chromatin immunoprecipitation (ChIP-seq), methylated DNA immunoprecipitation (19,20), GRO-seq (21)], experimental parameters like cross-linking efficiencies in different cell types or tissues, shearing or digestion of chromatin or the selectivity and affinity of an antibody (batch) can vary substantially between experiments and different experimenters and will ultimately impact on the overall quality of the final readout. Currently, quality assessment is performed by visual profile inspection of defined chromatin regions and complemented by peak caller predictions. In addition, a number of analytical methods have been described [for a recent summary of the methodologies used by the ENCODE consortium see (22)]. However, none of them has been shown to be applicable to the large variety of ChIP-seq and enrichment-related NGS profiling assays. For instance, methods like fraction of mapped reads

To whom correspondence should be addressed. Tel: +33 3 88 65 34 73; Fax: +33 3 88 65 34 37; Email: hg@igbmc.u-strasbg.fr
Correspondence may also be addressed to Marco-Antonio Mendoza-Parra. Tel: +33 3 88 65 34 19; Fax: +33 3 88 65 34 37; Email: marco@igbmc.fr

retrieved into peak regions (FRiP) (23) or irreproducibility discovery rate (IDR) (24) require prior use of peak calling algorithms for evaluation and are therefore dependent on peak-calling performance of a given tool with the user-defined parameters. Consequently, they cannot be easily used for multi-profile comparisons when different peak callers are required (e.g. transcription factors (TFs) and histone modifications with 'broad' profiles).

In addition to the performance of the immunoprecipitation/enrichment assays, the rapid technological progress provided NGS platforms with largely different sequencing capacities ranging from tens of millions (e.g. Illumina Genome analyzer v1, hereafter referred to as 'GA1') to >3 billion (HiSeq2000) reads per flow cell. As a consequence, the public databases hosting NGS-generated data sets are populated with ChIP-seq profiles presenting a large variety in sequencing depth. Importantly, previous studies have demonstrated that by increasing the sequencing depth, the number of discovered binding sites increases accordingly. Intuitively, it is expected that the number of sequenced reads required to discover all binding events is directly related to their total number and to their binding pattern (i.e. 'broad' regions covering large parts of a genome will require more reads to be properly identified than 'sharp' patterns with few target sites). When evaluating the quality of NGS-based profiling, it is therefore important to assess if a given ChIP-seq profile is performed under optimal sequencing conditions, including the minimal sequencing depth required to discover most of the relevant binding events of a given factor.

For all the above reasons, we have developed a bioinformatics-based quality control (QC) system that uses raw NGS data sets to (i) infer a set of global QC indicators (QCis), which reveal the comparability of different enriched-NGS data sets, (ii) provide local QCis to judge the robustness of cumulative read counts ('peaks or islands') in a particular region, (iii) provide guidelines for the choice of the optimal sequencing depth for a given target and, finally, (iv) to have quantitative means of comparing different antibodies and antibody batches for ChIP-seq and related antibody-driven studies. In addition, we have established a QC indicator database that will be expanded to cover virtually all publicly available enrichment-related NGS profiling assays. Thus, users can compare the quality indicators computed by the NGS-QCi Generator for a given ChIP-seq experiment with the quality indicators for published data sets present in the QC indicator database. This information will guide users toward optimization of the ChIP-seq process, if the QC is lower than that achieved previously by others and/or with other antibodies. Moreover, this QC system will be useful for antibody development and certification. We discuss the simplicity and versatility of the present QC method and database in view of currently existing QC assessment procedures and guidelines. The NGS-QC Database of QC indicators for publicly available profiles and the NGS-QC Generator tool are freely accessible through a customized Galaxy instance at http://igbmc.fr/Gronemeyer_NGS_QC.

MATERIALS AND METHODS

Data sets

Publicly available data sets were downloaded from GEO (25). When available, aligned files (either in BED or BAM format) were used; otherwise sequence data sets, available through the short read archive database, were first aligned to the corresponding reference genome using Bowtie2 under standard alignment options (26).

Assessment of the inherent robustness of ChIP-seq profiles

Based on the rationale that beyond a sequencing depth threshold a ChIP-seq profile changes only in amplitude but not in pattern, we evaluated this property by monitoring the changes of its RCIs after read-subsampling. For this, aligned reads were randomly sampled at three distinct densities (90, 70 and 50%; referred to as s90, s70 and s50 subsets, respectively). To avoid bias, random sampling was performed without replacement; each separately sampled density subset was generated from the original read data set. RCI profiles were constructed by counting the overlaps within a defined window ('bin'). With the aim of having no more than one binding event per bin, it is currently fixed to 500 bp. An empirical evaluation of the influence of this parameter on the assessment of the quality indicators confirmed our initial choice (Supplementary Figure S1d).

Reconstructed profiles from randomly sampled subsets are then compared with that constructed from the initial total mapped reads (TMRs) by computing the recovered RCI (recRCI) per bin after sampling as follows:

$$recRCI = (\frac{samRCI}{oRCI}) * 100$$

Where *samRCI* is the RCI/bin retrieved after sampling and *oRCI* is that found in the original profile. Under the working hypothesis that, as a consequence of random sampling, *recRCI* is directly proportional to the sampling density, the divergence from the expected RCI behavior is measured as follows:

$$\partial RCI = samd - recRCI$$

where *samd* corresponds to the random sampling density; i.e. 90, 70 and 50% for s90, s70 and s50, respectively. Importantly, the RCI dispersion or δRCI is inversely proportional to the original RCI (Supplementary Figure S1C) and it has been empirically observed to present a direct correlation with the quality of ChIP-seq profiles (Supplementary Figure S2). Thus, for providing a quantitative assessment of the changes of RCI dispersion in a given data set, we have evaluated the fraction of bins displaying a δRCI within in a given interval, which has been defined as the global density QC indicator 'denQCi'. This global indicator—evaluated in conditions where only a half of the initial sequenced reads are available (s50)—is systematically used in this study to measure the degree of robustness of the evaluated profile to the read-subsampling treatment (i.e. high denQCi corresponds to low RCI dispersion). In addition, the changes in robustness on subsequent read subsampling has been evaluated

by comparing the denQCi for the sampling closest to the original profile (s90) with that sampling only half of the sequenced reads (s50). This is defined as the similarity QC (simQCi) indicator, computed as ratio between denQCis for the s90 and s50 sampling subsets. The current version of NGS-QCi Generator provides both global quality indicators (denQCi and simQCi) for dispersion intervals of 2.5, 5 and 10%. Further details concerning the assessment of these indicators are provided in the QC report (see Supplementary File S1 and Supplementary Figure S4).

Local QCis

Given that the above analyses were computed for 500-bp bins, the δRCI/bin data can be used as local QCis. The NGS-QCi Generator provides such information in either wiggle or BED formats; the default condition identifies bins with $\delta RCI \leq 10\%$. Local QCis in wiggle file format can be uploaded in the Integrated Genome Browser (IGB) and displayed as a heat-map together with standard RCI wiggle files (as illustrated in Figure 3B). In a similar manner, the corresponding BED file can be uploaded in the UCSC Genome Browser. This display option is useful to visualize predicted δRCIs associated to a given chromatin region of interest. Furthermore, 500-bp chromatin regions with δRCIs thresholds of 2.5, 5 or 10% can be downloaded as a table in BED format. The data sets facilitate comparative analyses of multiple profiles in the context of defined δRCI thresholds.

QC-STAMP and NGS-QCi database

The contribution of the two QCis to the single descriptor QC-STAMP was defined by following equation:

$$\text{QC-STAMP} = \frac{denQCi(s50)}{simQCi}$$

To evaluate the divergence of this global descriptor over all enrichment-related NGS profiles currently compiled in the NGS-QC database, the QC-STAMP distributions assessed for three different RCI dispersion intervals were subdivided in four quantiles to which the following grades have been attributed: 'D', lower quartile (<25%); 'C', interquartile 25–50%; 'B', interquartile 50–75% and 'A' upper quartile (>75%). The NGS-QCi Generator database associates these grades for 2.5, 5 and 10% δRCI to each profile as a three-letter symbol, such that, for example AAA ('triple A') reveals an A grade for all three δRCIs. All available profiles are displayed as a dynamic QC-STAMP versus TMR scatterplot, which allows judging of their QCi similarities in the context of the sequencing depth. Note that the global QC-STAMP descriptor will be dynamically reevaluated when novel entries are provided to the database.

Peak detection approach

In addition to the well-described peak caller MACS (27), peak calling has been performed with MeDiChI, a model-based deconvolution approach originally developed for ChIP-chip assays (28), which we have adapted to ChIP-seq analyses. MeDiChI computes a model from a randomly selected subset of the multiple binding events present in a genome-wide profile. This model is then used as a deconvolution kernel for genome-wide prediction of likely binding events, which are further validated by nonparametric bootstrapping. As we compared ChIP-seq profiles generated at different sequencing depths, we have included a *P*-value/peak intensity product ranking-based approach for defining a common false discovery rate (FDR) during comparison. For this, a ranking coefficient (RC) for the i^{th} peak identified by MeDiChI was calculated by the following equation:

$$RC_i = Int_{Peak\ i} * (-10 * \log_{10}(p - value_i))$$

This RC was sorted from the highest to the lowest value, and the FDR was assessed as follows:

$$FDR_i = -10 * \log_{10}(\frac{i^*}{N} p - value_i)$$

Where i* is the ranking position based on the RC, and N is the total number of peaks. Thus, all ERα ChIP-seq profiles have been compared at a FDR threshold ≥45 or FDR adjusted *P*-value threshold $10^{-4.5}$.

RESULTS

Previous studies described the concept of a 'saturation point' as the sequencing depth after which no new binding sites are identified by a given peak caller with additional sequenced reads (5,29). This concept has been initially evaluated in a retrospective manner by assessing the number of significant binding sites retrieved when only a subset of the original sequenced reads was used for profile reconstruction (random subsampling approach). Intuitively the 'saturation point' concept predicts that beyond such threshold no further binding sites would be discovered and by consequence, the increased sequencing depth should only influence the overall read-count intensity of the corresponding profile.

Following the same concept, the QC system presented here evaluates the stability of the pattern of a given profile beyond the saturation point by measuring the reproducibility of ChIP-seq and enrichment-related NGS profiles under conditions where only a subset of the TMRs are used for reconstruction. In the ideal 'saturation' condition, such a reconstruction will generate a profile with the same read distribution pattern across the genome but with a decrease of the RCIs according to the percentage of TMRs used (Figure 1A). The extent to which this reproducibility is attained is defined as 'robustness' of the original profile and is assessed by the resampling of a given data set at the level of half of the original TMRs (referred to as 's50'). Whereas none of the currently available profiles displays ideal robustness at s50, the evaluation of the deviation from such ideal behavior reflects the degree of robustness and represents a quantitative method for assigning a set of quality descriptors to any NGS-generated profile.

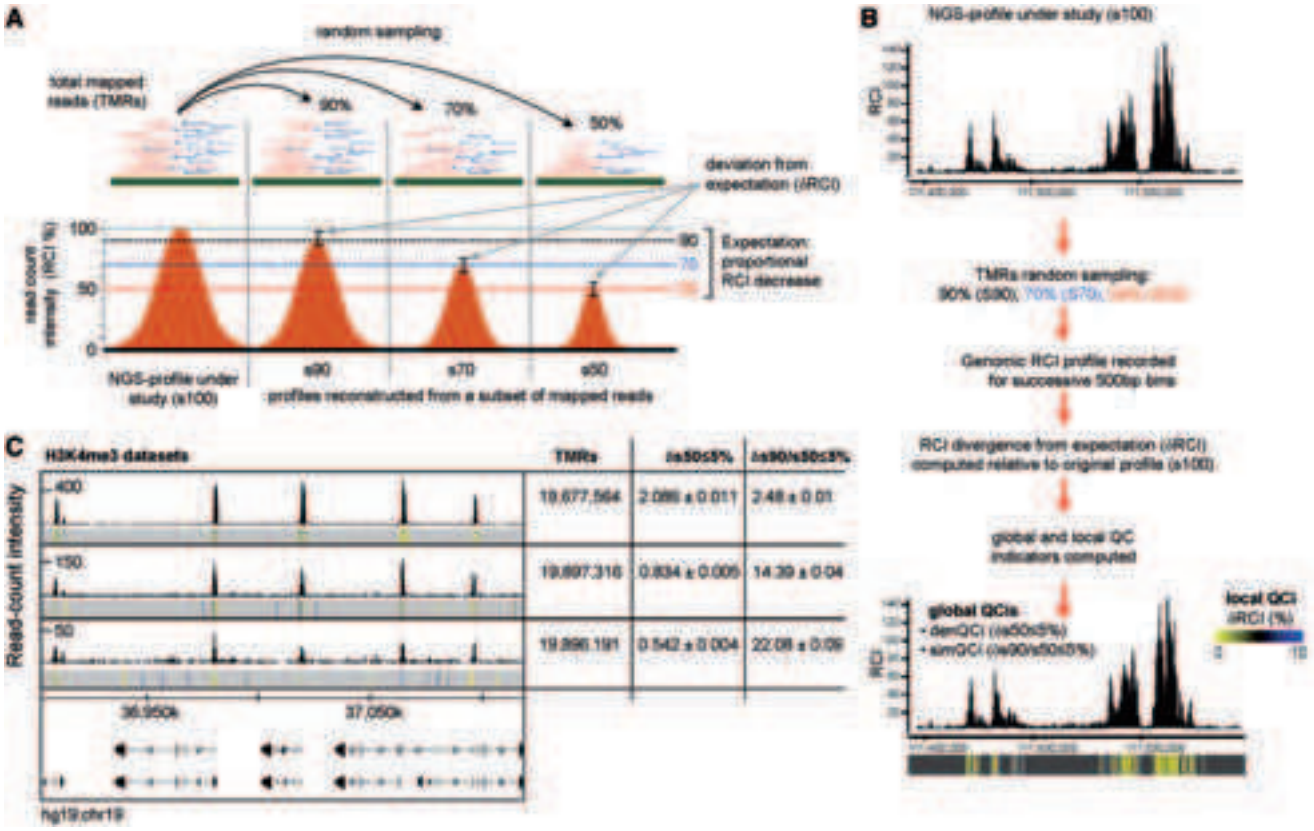


Figure 1. Assessing quality descriptors for ChIP-seq profiles. (**A**) Based on the rationale that a robust profile displays a proportional decrease of its RCIs along the genome when a randomly sampled population of its TMRs is used for profile reconstruction, the present quality assessment method quantifies the deviation from the expected RCI decrease within defined thresholds. (**B**) TMRs are randomly sampled into three distinct populations (90, 70 and 50%), which are used for profile reconstruction by computing the RCIs in 500-bp bins. The RCI divergence from expectation (δRCI) is measured relative to the original profile (s100). This information generates local QCis and is displayed together with the original RCI profile to identify robust chromatin regions (δRCI heat-map below the bottom profile). In addition, two global QCis are calculated, comprising the density QCi [denQCi, defined as the fraction of bins displaying <5% δRCI after 50% TMRs sampling ('δs50/5')] and the similarity QCi (simQCi), defined as ratio of denQCi after 90% sampling over that after 50% sampling ('δs90/s50/5'). (**C**) Genome-browser screenshots of three different H3K4me3 ChIP-seq profiles. In addition, the RCI dispersion per 500-bp bins (local QCi) is illustrated as color-coded heat-map below the corresponding ChIP-seq profiles. Note that while all three profiles present ~19 million TMRs, they differ significantly in their global RCI amplitudes. Furthermore, their corresponding global QCis assessed from 5 random sampling assays are displayed (average ± standard deviation).

ChIP-seq profile's robustness dispersion provides quality descriptors

This QC system evaluates the robustness of RCI dispersion for any given ChIP-seq and enrichment-related NGS profiles by comparing distinct randomly sampled populations derived from the primary data set (Figure 1B). Specifically, TMRs are first resampled at 90, 70 and 50% (referred to as s90, s70 and s50, respectively) of the original data set. The genome-wide read-count distribution within 500 bp bins is then evaluated for the sampled subsets and compared with that observed for the original profile (s100) (for the effect of bin size on measuring profile robustness see Supplementary Figure S1). Under the assumption of a proportional RCI decrease on read subsampling (saturation concept), the bin RCI divergence from expectation is calculated (δRCI or local divergence; defined as the difference between the theoretically expected RCI and that observed after resampling). Furthermore, a global quantitative assessment of the changes in bin RCI dispersion is given by the evaluation of the total bins presenting a defined RCI dispersion. This global indicator, defined as density Quality indicator (denQCi), evaluated in conditions where only a half of the initial sequenced reads are available (s50), is systematically used in this study to illustrate the degree of robustness of the evaluated profile to the reads-subsampling treatment (i.e. $\delta s50 \leq 5\%$ makes reference to the fraction of bins with $\delta RCI \leq 5\%$ when half of the TMRs are used for profile reconstruction).

Furthermore, the changes in robustness on successive read subsampling has been evaluated by comparing the denQCi obtained for the subset closest to the original profile (s90) relative to that assessed from half of all sequenced reads (s50). This second global indicator has been defined as the 'similarity QC indicator' (simQCi) because it reveals the similarity between the robustnesses assessed at s90 and s50. Overall, the higher the denQCi and the lower the simQCi, the more 'robust' is the evaluated profile.

ChIP-seq profiles established from similar TMRs can lead to variable quality patterns as revealed by visual inspection of three ChIP-seq profiles of the tri-methylation of lysine 4 of histone 3 (H3K4me3) generated with antibodies obtained from the same supplier and with similar (~19 millions) TMR levels (Figure 1C). Yet, they present major differences of global RCIs and background levels (note the different scales). Indeed, the computing of the QCis provides quantitative descriptors (denQCi, $\delta s50 \leq 5\%$ and simQCi, $\delta s90/s50 \leq 5\%$) for the relative quality of the three profiles, which fully comply with the visual quality assessment, thus illustrating the usefulness of this approach in providing quantitative QC values for comparing different ChIP-seq data sets. Note that multiple random TMR samplings performed for each of the illustrated profiles revealed a coefficient of variation of <2% for the computed QCis. This demonstrates a high stability of the measurement of global QCis even when derived from a single random drawing (Figure 1C and Supplementary Figure S2).

Sequencing-depth influences the quality of ChIP-seq profiles

ChIP-seq and related assays are in most cases based on reads obtained from a single flow cell channel. Importantly, read densities of flow cells have largely increased over the past few years, ranging from <40 million for the first Genome Analyzer from Illumina (GA1) to >3 billion reads (300 Gb) for the Hiseq2000 platform. Consequently, the TMRs used for profile reconstruction can vary dramatically, inducing questions concerning the comparability of profiles that were constructed with different amounts of TMRs.

To evaluate the direct influence of sequencing depth on NGS-profiling robustness, we performed an analysis of biological replicates for ERα binding in H3396 breast cancer cells (3), which was performed by using one channel of the GA1, GA2X or HiSeq2000 platforms. We also included a comparison with half of a HiSeq channel by using multiplex technology. As expected, the sequencing depth provided by the different sequencing platforms, correlates well with the overall RCIs (Figure 2A). Importantly, TMR sampling analysis revealed a 16.2-fold increase of denQCi and, thus, global profile 'robustness', with increased sequencing depth ('$\delta s50 \leq 5\%$' in Figure 2A).

As expected, the number of TMRs used for ERα profile construction strongly influenced the total number of predicted statistically significant binding sites. In fact, with >50 million reads for the Hiseq2000 profile, 22 150 ERα sites were predicted (FDR adjusted *P*-value threshold $10^{-4.5}$; for peak detection algorithm, see 'Materials and Methods' section). In contrast, only 2038 sites were predicted from ~5 million reads obtained with one GA1 channel (Figure 2B). Albeit the total number of predicted peaks increased strongly with increasing sequencing depths, the number of sites that complied with $\delta s50 \leq 5\%$ shows a much slower increase and entered a plateau phase above 24 million TMRs. This indicates that the 'robust' ERα binding sites approach saturation as defined in previous studies on sequencing depth and *de novo* discovery of transcription factor binding sites (5,29,30).

As we have profiled ERα binding under identical treatment conditions, it was reasonable to assume that the sites identified at low sequencing depth constitute a subpopulation of those identified in the high TMR profiles. In fact, when comparing the ERα binding sites predicted at highest sequencing depth with those derived from the other profiles, not only the number but also the robustness of peaks in the overlapping population increased with increasing sequencing depth. From 1321 ERα sites in the overlap between GA1 and the full channel HiSeq2000 profile, >80% of them (1096 sites) comply with $\delta s50 \leq 5\%$ (Figure 2C). Similarly, the number of ERα binding sites overlapping with the GA2X or half channel HiSeq2000 data sets increased strongly over that obtained with GA1, as did the number of robust peaks.

The above comparison revealed also a significant number of nonoverlapping sites (Figure 2C). While it is reasonable to assume that the outliers of the HiSeq2000 profile (red) result mainly from the incomplete binding site recovery from the other profiles, those outliers that are seen in the low TMR profiles but not in the HiSeq2000 are more likely 'false positives'. Indeed, the number of such sites is variable and does not follow a common trend as the increase of the overlap population with increasing sequencing depth; in this respect, the GA2X data set is suboptimal with 4- to 5-times more outliers (green) than the GA1 (gray) and 1/2Hiseq (blue) ones. Importantly, when considering only the robust peak population, the GA2X outliers were significantly reduced to about the level seen with GA1 and 1/2Hiseq ones. In addition, the nonoverlapping sites, including those of the full channel HiSeq2000, showed consistently lower peak intensities and weaker confidence *P*-values relative to overlapping population (Figure 2D).

Considering the full channel HiSeq data set as 'gold standard', the number of recovered 'true' ERα binding sites increased from <5% for the GA1 data set to ~60% for the half channel HiSeq2000 profile (Figure 2E). Importantly, 80% 'true positive' binding sites were recovered when only robust ERα sites are considered, indicating that the denQCi criterion identifies the highly reliable sites when comparing ChIP-seqs with largely differing sequencing depths.

The QCis are universally applicable to all ChIP-seq and enrichment-related NGS profiling assays

While in previous studies profile saturation has been defined after peak calling (5,29,30), the present QC evaluation system evaluates robustness directly from the raw pattern of genome-aligned reads. Therefore, QCis can be established for any type of enrichment-related NGS profiles, including ChIP-seq, RNA-seq, GRO-seq and others, making this methodology a universal tool for multi-dimensional quality profile comparison. Indeed, we have computed QCis for several types of publicly available NGS-generated profiles and observed a high variability between the corresponding QCis even when data sets with similar TMRs were compared (Figure 3A and Supplementary Figure S3). RNA-seq, which does

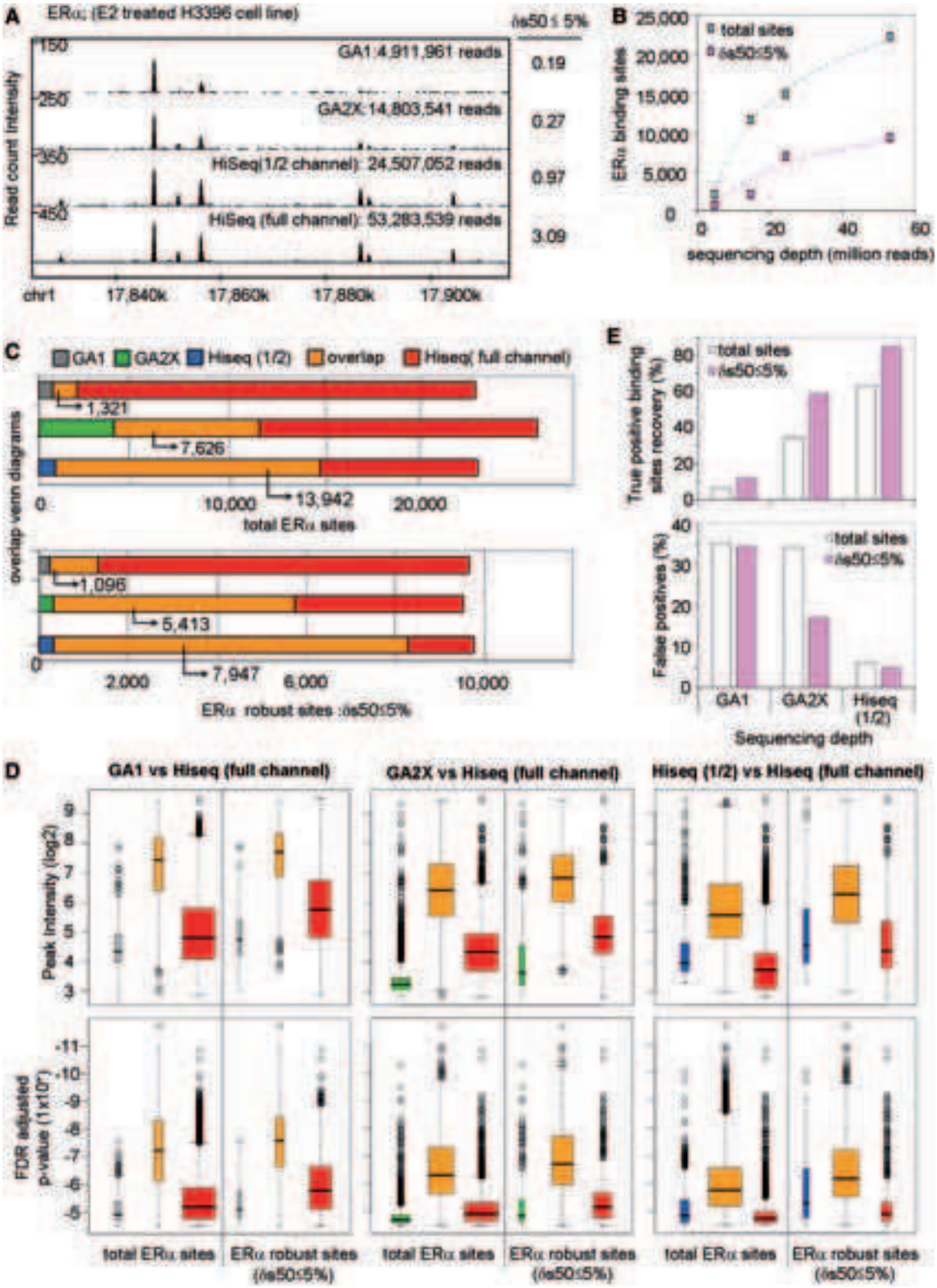


Figure 2. ERα binding sites detection assessed for different sequencing depths. (**A**) ERα RCI profiles obtained from different sequencing platforms [i.e. Genome Analyser 1 (GA1); GA2X and Hiseq2000] are illustrated. Each of the displayed ChIP-seq profiles was obtained by sequencing a single channel of the corresponding platform except for Hiseq2000, where half a channel or a full one was used. The corresponding mapped reads and their associated denQCi ($\delta s50 \leq 5\%$) are displayed. (**B**) Total ERα binding sites identified in ChIP-seq profiles generated at different sequencing depths compared with those that complied with the $\delta s50 \leq 5\%$ criterion. ERα binding sites were predicted with MeDiChI (FDR adjusted *P*-values threshold $10^{-4.5}$; see methods for details). (**C**) Venn diagrams illustrating overlap and outlier populations for ERα binding sites retrieved from sequencing a full HiSeq2000 channel compared with those identified at lower sequencing depths. This analysis was performed for total ERα sites (top panel) and those

(continued)

not involve manipulations like cross-linking and immunoselection, generated the most robust profiles, while a nonenriched input profile (whole-cell extract, WCE) constructed from ~19 million TMRs displayed the worst quality indicators. For nearly identical TMRs, the ChIP-seq profile of H4K20me1 revealed significantly improved QCis, as expected for the immunoselection of specific chromatin regions. Importantly, other histone modification profiles constructed from similar or even lower TMRs displayed better QCis than either H4K20me1 or WCE, thereby revealing that the robustness of a profile depends not only on the sample preparation

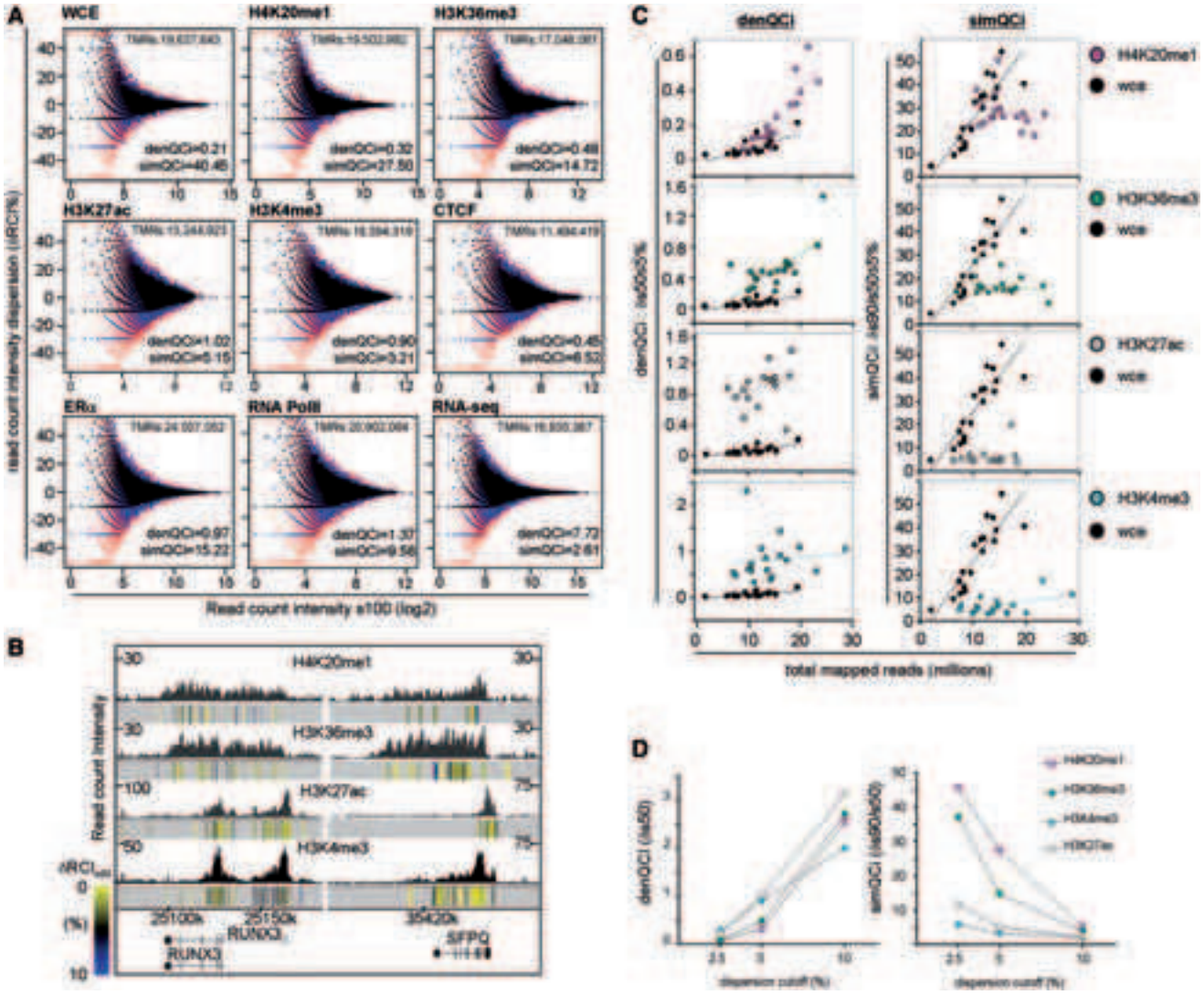


Figure 3. QCis for several types of ChIP-seq and enrichment-related NGS profiles. (**A**) Scatterplots illustrating the RCI dispersion (δRCI%) after sampling for different types of NGS profiles (overlays of s90, black; s70, blue; s50, red). TMR, density (denQCi, $\delta s50 \leq 5\%$) and similarity (simQCi, $\delta s90/s50 \leq 5\%$) QCis are indicated. Note that the input profile has the lowest denQCi and highest simQCi (WCE; top left), whereas the highest denQCi and lowest simQCi were measured for an RNA-seq profile (bottom right). (**B**) RCI dispersion per 500-bp bins is illustrated as color-coded heat-map (indicated at left) below the corresponding ChIP-seq profiles. (**C**) Density and similarity QCis for different profiles of the indicated histone modifications are compared with input WCE profiles. Note the different characteristics of the target profiles on increasing TMRs, which reveals that for H4K20me1 and H3K36me3 profiles presenting TMRs <15 million present QCis similar to the input. (**D**) Density and similarity QCis are displayed at stringent ($\delta s50 \leq 2.5\%$), intermediate ($\delta s50 \leq 5\%$) and relaxed ($\delta s50 \leq 10\%$) dispersion intervals.

Figure 2. Continued
complying with $\delta s50 \leq 5\%$ (bottom panels). (**D**) Boxplots displaying peak intensity and FDR adjusted *P*-value associated to overlap and outlier populations displayed in (C). Note that the ERα sites in the overlaps show systematically higher intensities and confidence than the outliers and that this difference is decreased for the $\delta s50 \leq 5\%$ populations. (**E**) Considering the sites identified with the full HiSeq2000 channel as 'true' sites, the fraction of true sites recovered in the compared profiles (top panel), as well as the false calls, estimated from the outlier population (bottom panel) are illustrated. Note the increase of true sites and a concomitant decrease of false calls in the population that complies with $\delta s50 \leq 5\%$.

and sequencing depth but also on the nature of the immunoprecipitated target. Note that H4K20me1 and H3K36me3 generate rather broad enrichment profiles revealing a spread of the mark over a large chromatin region, while those established for H3K27ac or H3K4me3 exhibit more discrete patterns of locally confined marks (Figure 3B). Our observation that the 500-bp RCI dispersion is generally higher in the H4K20me1 or H3K36me3 profiles compared with those of H3K27ac or H3K4me3 (see heat-map δRCI dispersion in Figure 3B) is likely to originate from the combination of several effects, including (i) the spread, local density and accessibility of the marks and (ii) the quality (i.e. affinity and selectivity) of the antibodies.

In addition to revealing quality differences between data sets for different targets at similar TMRs, the QCi computation also provides important quality information about data sets for the same target at different sequencing depths. Indeed, comparing the QCis for several H4K20me1 data sets generated from largely different TMRs reveals that below 15 million TMRs the QCis become indistinguishable from the WCE profiles, strongly arguing that significantly higher sequencing depths are essential to establish accurate profiles for such targets (Figure 3C). In contrast, H3K4me3 or H3K27ac ChIP-seq profiles have good QCis even for TMRs below 15 million reads.

That we observe major QCi differences between the various data sets reported for similar TMRs indicates that—in addition to the inherent pattern of the evaluated target—other factors, involving most likely all the experimental steps that generate the ultimate DNA library for sequencing, influence the quality of the profile (Figure 3C and Supplementary Figure S3).

Whereas most of the above described QCis have been established for a dispersion interval of 5% ($\delta s50 \leq 5\%$), different dispersion thresholds (e.g. $\delta s50 \leq 2.5\%$ or $\delta s50 \leq 10\%$) may reveal additional characteristics of the studied profiles. Indeed Figure 3D illustrates that the QCis determined for different dispersion intervals do not necessarily show a linear relationship. This information has been used as an additional source for quality evaluation (see below QC-STAMP) and represents a potential method for defining common QCi conditions in the case of multi-profile comparisons by allowing variable robustness dispersion cutoffs (Supplementary File S1).

NGS-QCi Generator: a stand-alone *in silico* platform for computing QCis

The above methodology infers local and global quality indicators for any available NSG-generated profile following a stand-alone approach, as it does not require additional wet-lab efforts. It has been implemented in the NGS-QCi Generator, a computational tool that is accessible at a customized cloud of the web-based platform Galaxy (31–33) (Supplementary File S1). The NGS-QCi Generator provides a comprehensive report summarizing the global QCis (Supplementary Figure S4) and provides access to the computed RCI dispersion per 500-bp bins (wiggle or BED format) defined as local QCis, which can be used to identify the robustness of specific regions of interest (Figure 3B and Supplementary Figure S5). Using the NGS-QCi Generator we have created a QCis database, which comprises at present the QC analysis of >5600 NGS data sets, including ChIP-seq profiles of histone modifications and variants, transcription factors, as well as GRO-seq and RNA-seq profiles (Figure 4A). This QCi database will be expanded to cover virtually all of the publicly available NGS profiles.

To facilitate and simplify the recognition of QCi divergence between profiles we have defined QC-STAMP, a global descriptor that combines the information provided by denQCi and simQCi. The QC-STAMP corresponds to a three-letter code composed of A, B, C and D that is derived from the position of a given profile QCi within the distribution of compiled QCis in the database. The first letter reveals this position for a δRCI dispersion threshold of 2.5%, the second and third letter for 5% and 10% δRCI, respectively. A to D grading was done to specify the following intervals: D, lower quartile (<25%); C, interquartile (25–50%); B, interquartile (50–75%); A, upper quartile (>75%) (Figure 4B). As an example, the H3K4me3 profile derived from 10 007 440 TMRs [arrow (3) in Figure 4A] classified as 'triple A' profile, while nonenriched WCE profiles were, as expected, of the lowest possible quality, 'triple D' (Figure 4C). Similarly expected was the high QC performance of RNA-seq, which does not involve the complex experimentation and immunoprecipitation procedures as ChIP-seq, and consequently received 'triple A' rating [arrow (1) in Figure 4A]. Note that these ratings are meant to provide a simplified view of the evaluated profile's robustness but not to replace the QCis, which provide more specific information.

As the quality of a ChIP-seq profile is the direct consequence of a rather large number of factors (e.g. cross-linking efficiency, chromatin shearing, antibody affinity and selectivity, variability between experiments, experimenters and platforms), the QCis cannot *per se* identify the source for the bad quality of a given profile. However, it does allow identifying data sets of divergent quality, which cannot be compared with each other, even though they might have been generated under similar conditions. Importantly, in contrast to current practice, the sequencing depth applied for generating NGS profiles is a tunable parameter to generate profiles of similar quality. As illustrated in Figures 3 and 5 for similar TMR levels, H4K20me1 or H3K36me3 profiles display in general poorer quality than those of H3K27ac or H3K4me3. However, increasing the sequencing depth will improve their quality descriptors to attain comparable levels, such that, for example, only 'triple A' data sets can be compared (Figure 5). In this respect, we believe that the QCi database will become an important reference to perform *a priori* predictions of the minimal sequencing depth required for a given target to reach a predefined quality.

The NGS-QCis in the context of previously described working standards and guidelines for ChIP-seq assays

Multidimensional comparative analyses of ChIP-seq profiles require prior quality assessment. Currently, this is done by visual inspection of profiles in a genome browser (for instance by evaluating the pattern in

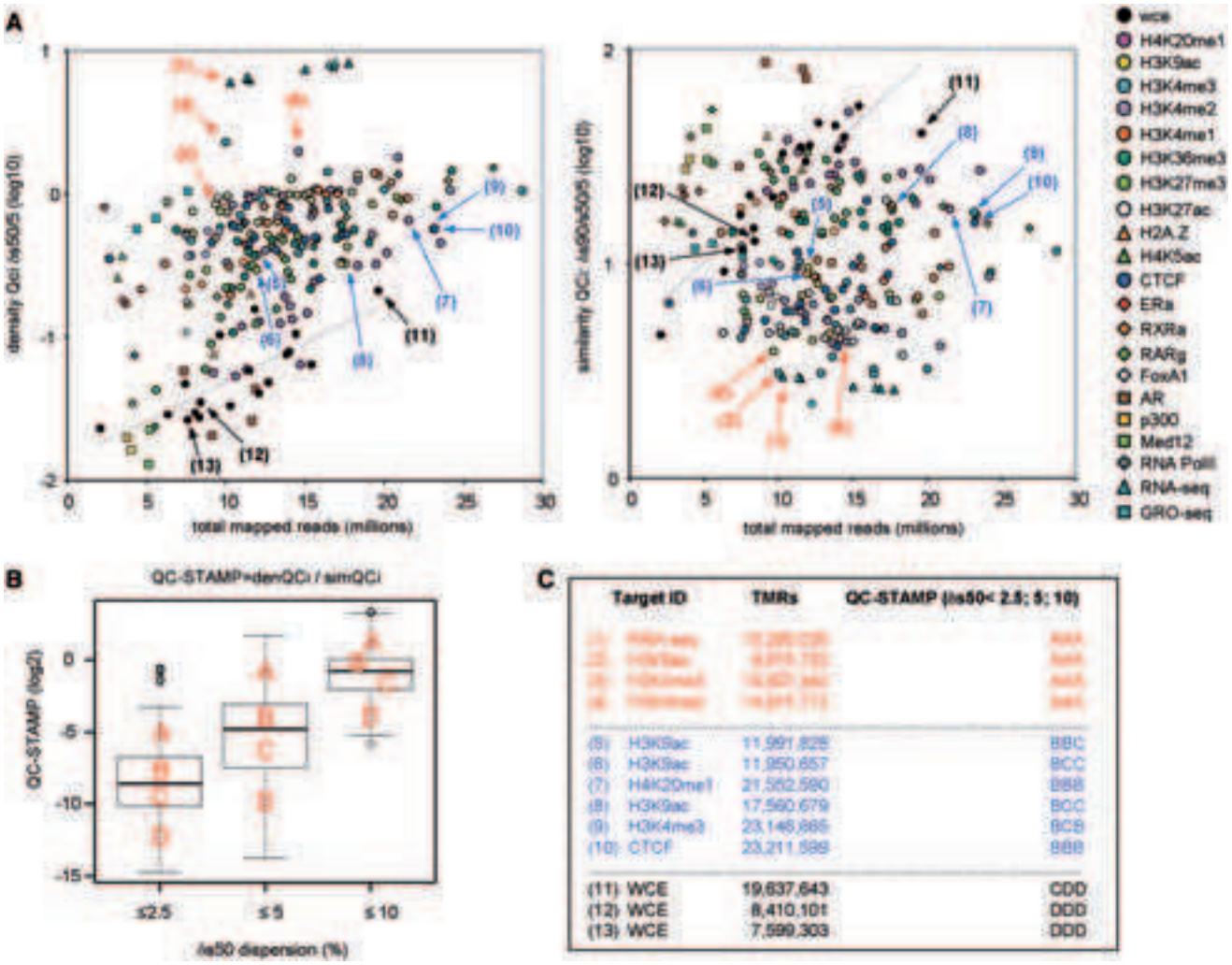


Figure 4. A universal NGS-QCi database for comparative analysis. (**A**) Cloud of NGS-QCis for multiple profiles present in the NGS-QCi database (http://igbmc.fr/Gronemeyer_NGS_QC). Density (left) and similarity (right) QCis are displayed relative to the TMRs; color codes are indicated at the right. QCis of input (WCE) profiles are displayed as black circles; the dashed line is the corresponding fitted curve. Arrows indicate the location of the data sets specified in (C). (**B**) QCis of the evaluated NGS profiles displayed in (A) are expressed in a single term, QC-STAMP, and represented as boxplots for different RCI dispersion intervals (2.5, 5 and 10%). Discrete quality grades 'A' to 'D' were associated with different quantiles (QC-STAMP dist > 75%; >75% QC-STAMP dist > 50%; >50% QC-STAMP dist > 25%; QC-STAMP dist < 25% associated to A, B, C and D qualitative indicators, respectively). (**C**) Examples of NGS profiles associated to different QC-STAMPs.

regions previously described as containing a chromatin enrichment) and complemented peak caller predictions based on (some) user-defined parameters.

In addition to visual inspection, analytical methods have been developed with the aim of providing quantitative quality assessments of NGS-generated profiles [for a recent summary of the methodologies used by the ENCODE consortium see (22)]. Methods like FRiP (23) or IDR (24) require prior use of peak calling algorithms for evaluation and are therefore dependent on peak-calling performance of a given tool with the user-defined parameters. Consequently, they cannot be easily used for multi-profile comparisons when different peak callers are required. This is for example the case when transcription factor profiles are compared with epigenetic profiles that display broad RCI patterns. Note that the IDR approach can only be used when replicate profiles are available, which is strongly suggested but not a routine procedure (see GEO entries). Furthermore, the criteria used for reproducibility by the IDR analysis can be misleading in cases where compared profiles present broad enrichment patterns (Supplementary Figure S6; see also below).

Two other methods; signal distribution skewness (34) and strand cross-correlation analysis (SCC) (22) operate in a peak caller-independent manner. Signal distribution skewness evaluates the asymmetry of genome-wide tag-count distribution, while SCC measures the quality of evaluated ChIP-seq profiles from the sequence tag density on forward and reverse strand reads at target sites. SCC is thus applicable mainly, if not exclusively, to 'sharp' patterns like those observed for transcription factor ChIP-seq data sets. It is rather evident that SCC

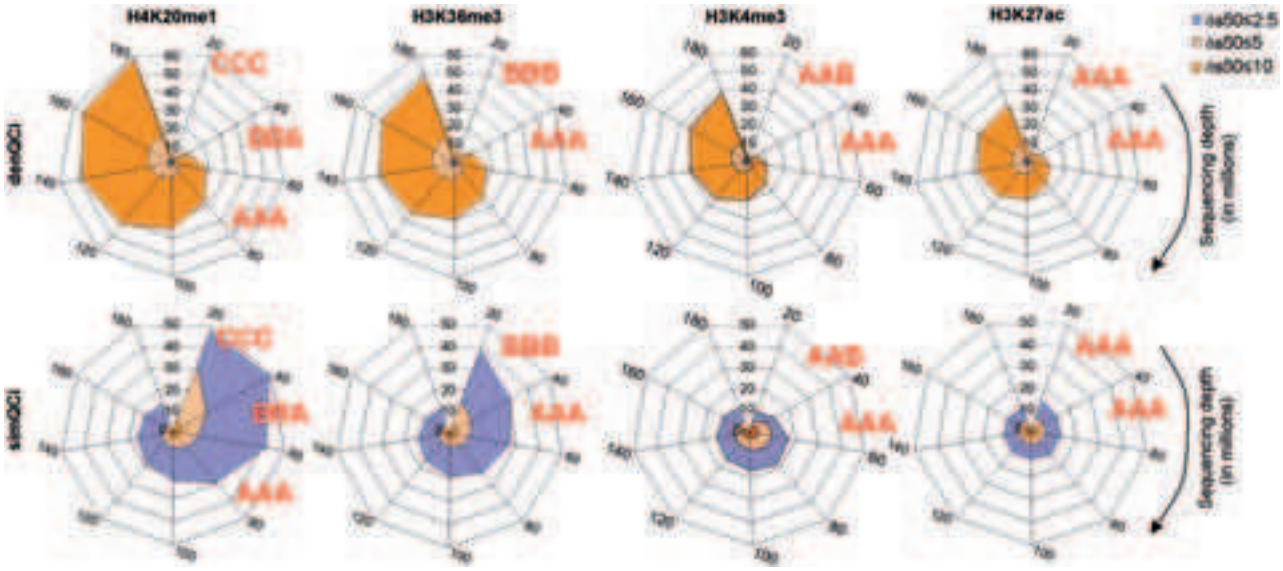


Figure 5. Meta-analysis illustrating the influence of the sequencing depth on the density and similarity QCis. Meta-analysis performed by compiling several profiles and subsequently sampled at defined TMRs ranging from 20 to 180 million. For each resampled subset the corresponding QCis were computed and displayed in spider-web charts, in which denQCi and simQCi are displayed for different δRCI thresholds (color-coded as indicated at the top left). QC-STAMPs have been associated to the evaluated profiles as illustrated. Note that for H4K20me1 sequencing depths of up to 60 million reads are required to obtain a 'triple A' grade, while H3K27ac and H3K4me3 receive this grade with 20 million TMRs.

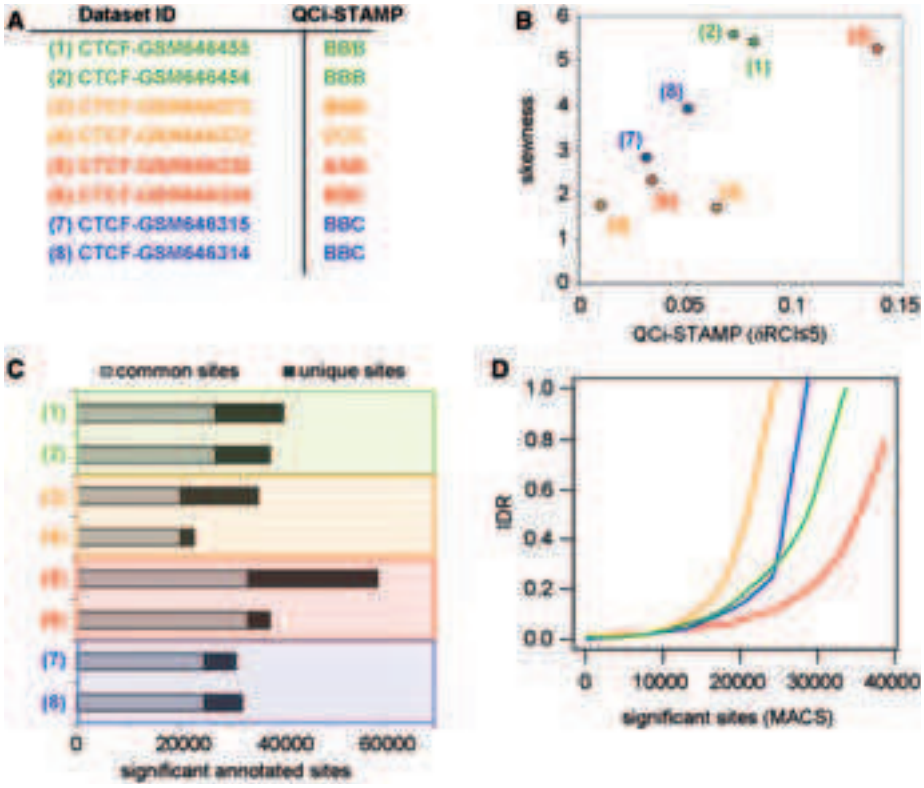


Figure 6. Comparison of QCi-STAMP performance with other analytical methodologies. (**A**) A set of four biological duplicates was selected from publicly available CTCF ChIP-seq profiles (pairs are enhanced by color code) and their corresponding QCi-STAMP descriptors were inferred ('A' for highest and 'D' for lowest quality). (**B**) The skewness of the read-count signal distribution of the biological replicates compared with the predicted QCi-STAMP (dRCI $\leq$ 5%). Note that the QCi-STAMP descriptors discriminate between data set (3) and (4), while their skewness evaluation does not. (**C**) Significant binding sites were predicted by MACS (default *P*-value threshold: 1×10^{-5}) and classified based on their overlap between CTCF replicates (common and unique sites). Common sites were assessed by accepting up to 40-nt distance between MACS-predicted summits. (**D**) 'IDR' among CTCF replicates assessed by sorting significant binding sites according to the corresponding *P*-value. Note that in agreement with the QCi-STAMP descriptors, but differing with the skewness analysis (see panel C), data sets (3) and (4) present the worst IDR, while data sets (5) and (6) present the best IDR pattern.

cannot be used for quality assessment of broad patterns, as significantly enriched reads of such profiles cover large areas. Thus, from the conceptual point of view in addition to the present QCi system, signal distribution skewness appears to constitute the only other universal quality measurement method. To compare signal distribution skewness and our NGS-QC we have evaluated the degree of skewness in four publicly available CTCF ChIP-seq data sets (each of them represented by two biological replicates) and compared it with QCi-STAMP (Figure 6A and B). Both methods provide similar quality predictions, with the important exception that the difference in quality of one pair of the evaluated replicates (GSM646372 and GSM646373 data sets) was predicted by the QCi-STAMP but not by the skewness analysis (Figure 6B). To understand the origin of this discrepancy, we assessed the number of common and unique sites for each pair of replicate data sets [peak caller MACS (27); default *P*-value threshold conditions: 1×10^{-5}), followed by IDR analysis for the predicted binding sites (Figure 6C and D, respectively). Interestingly, this complementary analysis revealed a lower number of significant common sites for replicate GSM646372 ('triple C') and GSM646373 ('triple B') than for the other replicate data sets. This IDR-defined differential quality of the two pairs of replicates was equally well detected by the QCi-STAMP (but not the skewness) approach. Overall, these comparisons show that QCi-STAMP provides a more versatile and reliable quality discrimination of NGS-generated profile than the skewness approach. Moreover, in contrast to IDR, QCi-STAMP reveals which of the replicates should be repeated to increase the overall quality without the necessity of using peak caller approaches.

An additional limitation of the IDR analysis, namely the dependence on peak caller performance, becomes apparent from analysing CTCF (Figure 6; sharp peaks) and H3K4me3 data sets (Supplementary Figure S6; broad peaks). While IDR analysis of CTCF can be done with 40 nt summit distance overlaps (i.e. the maximal distance between predicted summits to consider two binding events as reproduced), such conditions are noninformative for the H3K4me3 data set. To overcome this limitation, larger summit distance thresholds (e.g. 500 nt) have to be used to get informative results (Supplementary Figure S6). It is thus unlikely that comparisons between ChIP-seq profiles presenting different enrichment patterns can be done with IDR. In contrast, the QCi-STAMP reliably predicts the different qualities for the 'triple A' and 'triple B' pair of replicates and the common quality for the two 'triple B' replicates in the case of the evaluated H3K4me3 data sets (Supplementary Figure S6A), as illustrated for the CTCF profiles (Figure 6A).

DISCUSSION

The assessment of the quality of ChIP-seq data sets has been mostly performed by visual inspection in a genome browser and/or by the capacity of peak/island/pattern caller algorithms to predict locally enriched sequence counts. In both cases, it is a rather subjective analysis relying on user-defined criteria, such as the choice of 'representative' regions or thresholds for peak detection, and the statistical models and/or parameters used for assessment of enriched patterns. Only recently, methods are being developed that aim at providing a quantitative measure for the quality of ChIP-seq assays but so far there is no tool that provides a universal quality assessment for past and present NGS-generated profiles.

The present NGS-QC approach provides quantitative QCis generated from the evaluation of a feature common to all NGS-generated profiles, namely the profile construction from sequenced read overlaps. Conceptually, the QC Generator interrogates the robustness of such a profile when fewer sequenced reads are available, irrespective of the underlying experimental approach; simplistically this can be described as a numerical analysis similar to the visual inspection of Figure 2A, which displays RCIs at different TMRs but for the entire genome-aligned profile and not only for a selected region.

This concept has an inherent universal dimension, which is essential for comparative purposes and considering that the public GEO repository represents a powerful source for performing *in silico* data set comparisons, we have established a database of QCis for >5600 profiles. Our ultimate goal is to cover all publicly available ChIP-seq and enrichment-related NGS data sets to provide a comprehensive QCi library to the scientific community. Moreover, we invite all our colleagues to use the QC Generator for evaluation of their own profiles and suggest that all newly reported IP-based NGS profiles (which show the largest variability) are provided with the corresponding global QCis. We also invite the community to import all newly defined QCis into the global QCi database. Collectively, this database will be a highly valuable source of information about the quality that can be achieved, for example, for ChIP-seq of a certain target with a given (batch of) antibodies.

We believe that the universality, together with its simplicity and broad accessibility, makes the present system an attractive tool for QC analysis of profiles before engaging peak detection algorithms. Once a profile has been QCed, the QC descriptors provide objective numerical criteria to any NGS-generated profile that is provided to the community. Thus, existing profiles can be compared with others in multidimensional studies and meta-analyses.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the members of our laboratory and the computational IT support services at the IGBMC, especially Jean-Luc Toussaint, for support in generating the Web site interface. Furthermore, we would like to thank Malgorzata Nowicka for her support on statistical issues related to this study and Pierre-Etienne Cholley for

the current maintenance and content curation of the NGS-QC database, and for the implementation of an executable version of the NGS-QC Generator. M.A.M.P developed the concept of Quality Control indicators and generated all data set comparisons; W.v.G implemented the NGS-QC-Generator, the NGS-QC database and its web interface together with M.A.M.P; D.C provided statistical support during data sets processing. M.A.M.S worked with W.v.G in the implementation of the customized Galaxy instance and is currently responsible for its maintenance. M.A.M.P and H.G. wrote the manuscript and the user tutorial.

FUNDING

Ligue National Contre le Cancer (laboratoire labelisé); the Association pour la Recherche sur le Cancer; the Institut National du Cancer (INCa); the European Community contracts [LSHC-CT-2005-518417 'EPITRON' and HEALTH-F4-2009-221952 'ATLAS']; the Alliance Nationale pour les Sciences de la Vie et de la Santé (AVIESAN)/Institut multi-organismes cancer (ITMO Cancer); D.G.C. was fellow of the Fondation pour la Recherche Médicale (FRM); FRM (aide aux projets innovants) (to W.v.G.). Funding for open access charge: Alliance Nationale pour les Sciences de la Vie et de la Santé (AVIESAN)/Institut Thématique Multi-Organismes Cancer (ITMO Cancer).

Conflict of interest statement. A patent application (EP123406478.4) describing the use of the NGS-QC system has been filed and the software has been deposited at the Agence Pour le Protection des Programmes (Paris).

REFERENCES

1. Harris,R.A., Wang,T., Coarfa,C., Nagarajan,R.P., Hong,C., Downey,S.L., Johnson,B.E., Fouse,S.D., Delaney,A., Zhao,Y. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**, 1097–1105.
2. Laird,P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
3. Ceschin,D.G., Walia,M., Wenk,S.S., Duboe,C., Gaudon,C., Xiao,Y., Fauquier,L., Sankar,M., Vandel,L. and Gronemeyer,H. (2011) Methylation specifies distinct estrogen-induced binding site repertoires of CBP to chromatin. *Genes Dev.*, **25**, 1132–1146.
4. Sims,R.J. III, Rojas,L.A., Beck,D., Bonasio,R., Schuller,R., Drury,W.J. III, Eick,D. and Reinberg,D. (2011) The C-terminal domain of RNA polymerase II is modified by site-specific methylation. *Science*, **332**, 99–103.
5. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
6. Law,J.A. and Jacobsen,S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.*, **11**, 204–220.
7. Margueron,R. and Reinberg,D. (2010) Chromatin structure and the inheritance of epigenetic information. *Nat. Rev. Genet.*, **11**, 285–296.
8. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
9. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
10. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
11. Mamanova,L., Andrews,R.M., James,K.D., Sheridan,E.M., Ellis,P.D., Langford,C.F., Ost,T.W., Collins,J.E. and Turner,D.J. (2010) FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat. Methods*, **7**, 130–132.
12. Wang,Z., Zang,C., Cui,K., Schones,D.E., Barski,A., Peng,W. and Zhao,K. (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, **138**, 1019–1031.
13. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
14. Hah,N., Danko,C.G., Core,L., Waterfall,J.J., Siepel,A., Lis,J.T. and Kraus,W.L. (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, **145**, 622–634.
15. Ingolia,N.T. (2010) Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol.*, **470**, 119–142.
16. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
17. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
18. Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.
19. Down,T.A., Rakyan,V.K., Turner,D.J., Flicek,P., Li,H., Kulesha,E., Graf,S., Johnson,N., Herrero,J., Tomazou,E.M. *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.*, **26**, 779–785.
20. Weber,M., Davies,J.J., Wittig,D., Oakeley,E.J., Haase,M., Lam,W.L. and Schubeler,D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, **37**, 853–862.
21. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
22. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
23. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.
24. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752.
25. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
26. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
27. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
28. Reiss,D.J., Facciotti,M.T. and Baliga,N.S. (2008) Model-based deconvolution of genome-wide DNA binding. *Bioinformatics*, **24**, 396–403.

29. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
30. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
31. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
32. Blankenberg,D., Von Kuster,G., Coraor,N., Ananda,G., Lazarus,R., Mangan,M., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.*, Chapter **19**, Unit 19 10 11–21.
33. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
34. Ho,J.W., Bishop,E., Karchenko,P.V., Negre,N., White,K.P. and Park,P.J. (2011) ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*, **12**, 134.

A quality control system for the analysis and comparison of profiles generated by massive parallel sequencing

Marco-Antonio Mendoza-Parra*, Wouter Van Gool, Mohamed Ashick Mohamed Saleem, Danilo Guillermo Ceschin and Hinrich Gronemeyer*

Department of Cancer Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)/CNRS/INSERM/Université de Strasbourg, BP 10142, 67404 Illkirch Cedex, France

*Corresponding authors:

Marco Antonio Mendoza-Parra
E-mail: marco@igbmc.fr

Hinrich Gronemeyer
E-mail: hg@igbmc.u-strasbg.fr
Phone: +(33) 3 88 65 34 73
Fax: +(33) 3 88 65 34 37

Supplementary Information

Supplementary Figure 1: *Read-count intensity behavior after total mapped reads (TMRs) random sampling and Influence of the window size on the assessment of QC indicators.*

Supplementary Figure 2: *QC indicators reproducibility over TMRs random sampling replicates.*

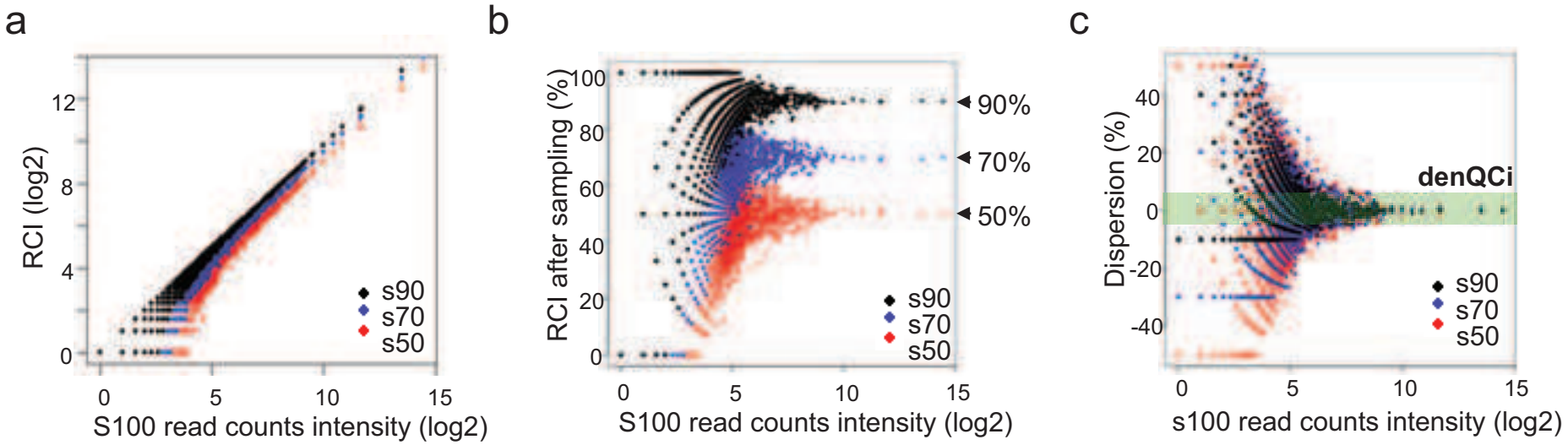
Supplementary Figure 3: QCis assessed for a diversity of NGS-generated profiles.

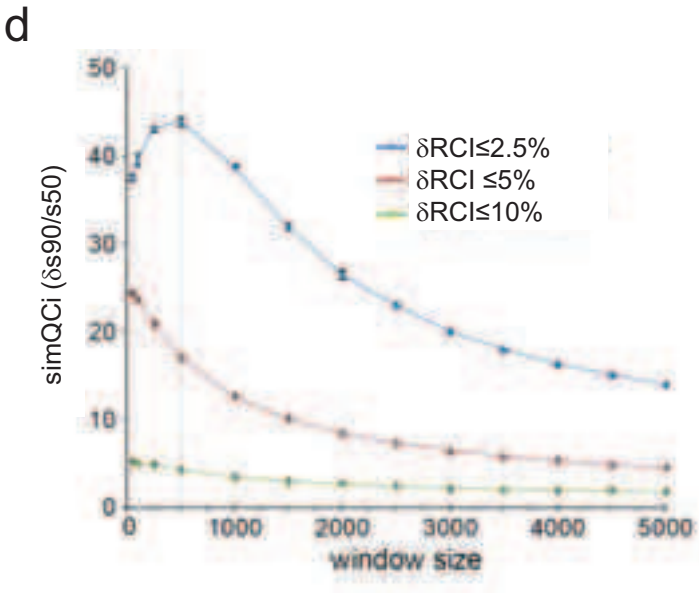
Supplementary Figure 4: *Example of the QCi report generated automatically for each processed NGS-profile by the NGS-QC Generator.*

Supplementary Figure 5: *Local QC indicators inferred from TMR random sampling.*

Supplementary Figure 6*: QCi-STAMP provides the same quality information as skewness and is independent of problems introduced by the use of peak callers.*

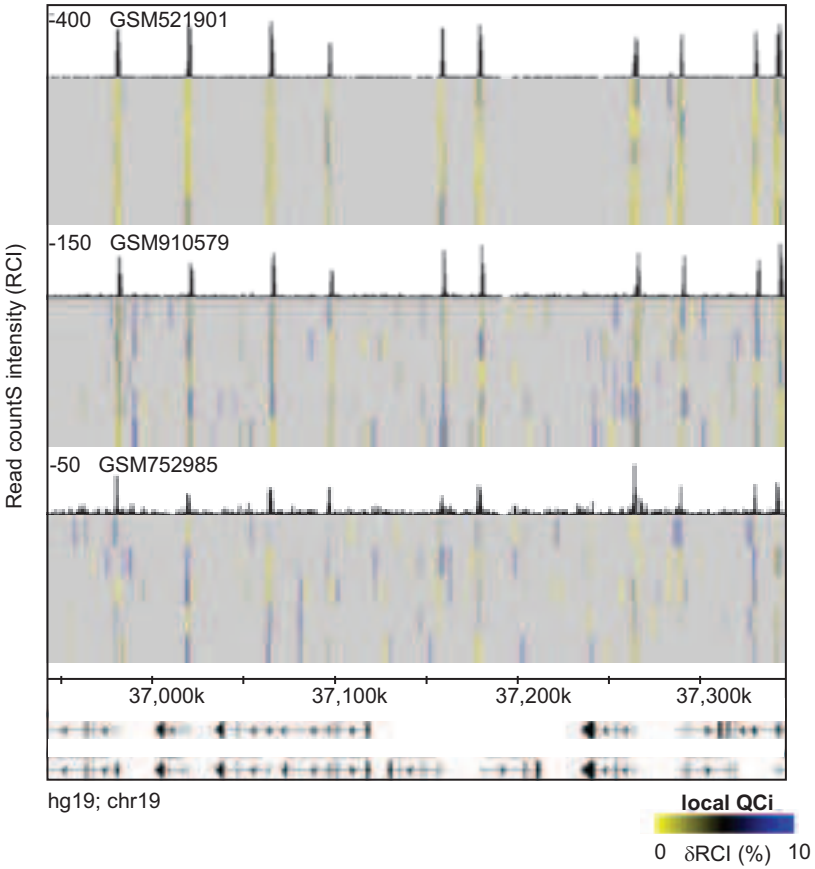
Supplementary File 1: *NGS-QC Generator Tutorial*



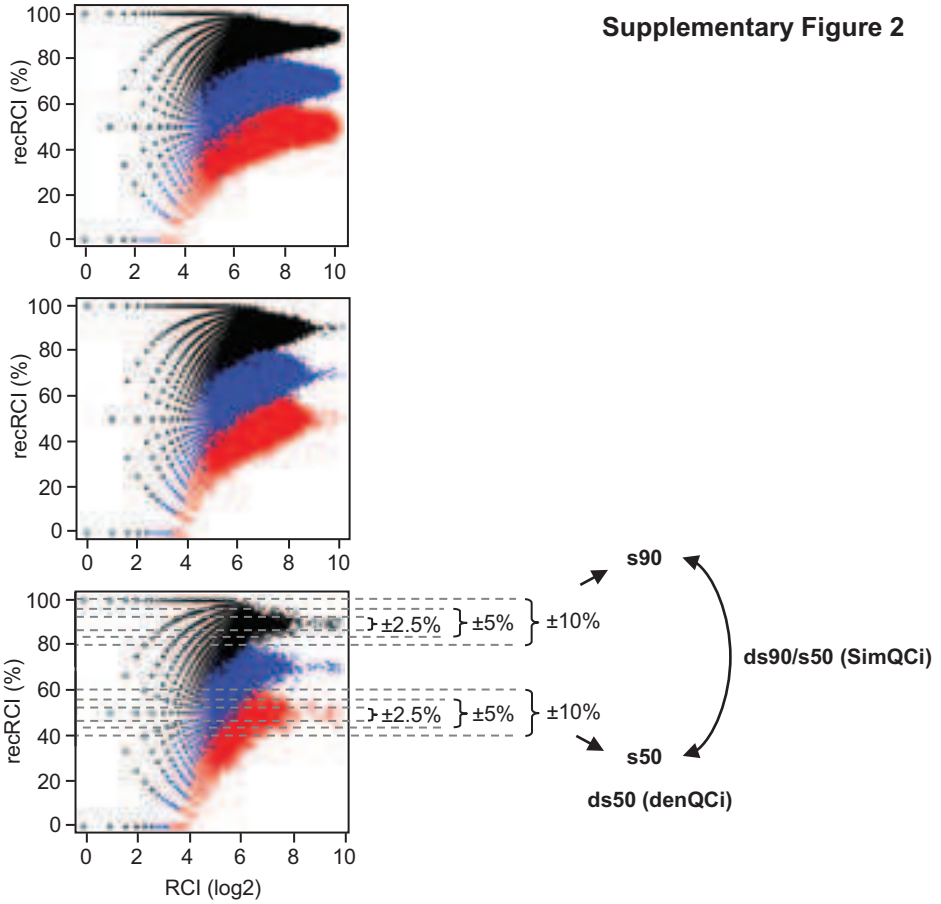


Supplementary Figure 1. (a) Scatterplot illustrating the read-count intensity(RCI) per evaluated bins observed after total mapped read's (TMRs) random sampling (y axis) in comparison to the original RCI (s100; x axis). Each data point corresponds to the RCI within a 500 bp bin. **(b)** To further enhance the influence of random sampling on the read count intensity of a given profile, the RCI per evaluated bin after sampling is represented in percentage relative to the original RCI. Note that for each of the three randomly sampled subsets the recRCI/bin approaches the theoretically expected value with the increase of the RCI/bin in the original profile. **(c)** Scatter plot illustrating the recovered read count intensity dispersion of a given profile. This transformed scatter plot superimposes the three scatter plots obtained after sampling for the same original dataset. The scatter of bins after sampling at s50, s70 or s90 having RCI values that deviate ≤5% from the expected RCI/bin are highlighted (defined as denQCi). **(d)** Influence of the window size on the assessment of QC indicators. Similarity QC indicator (defined as the ratio between the denQCi for s90 relative to that of s50) at different window sizes have been computed for ERα ChIP-seq profile. As highlighted by the vertical gray line, the highest difference for the simQC indicators assessed at three different dispersion intervals (2.5%, 5%, 10%) is retrieved for bins of windows sizes between 250 and 500bp. Note that this value corresponds to the expected chromatin fragmentation size. The ERα ChIP-seq profile used in this study was originally published in (*Ceschin et al. Genes Dev 25, 1132, 2011*).

a



b

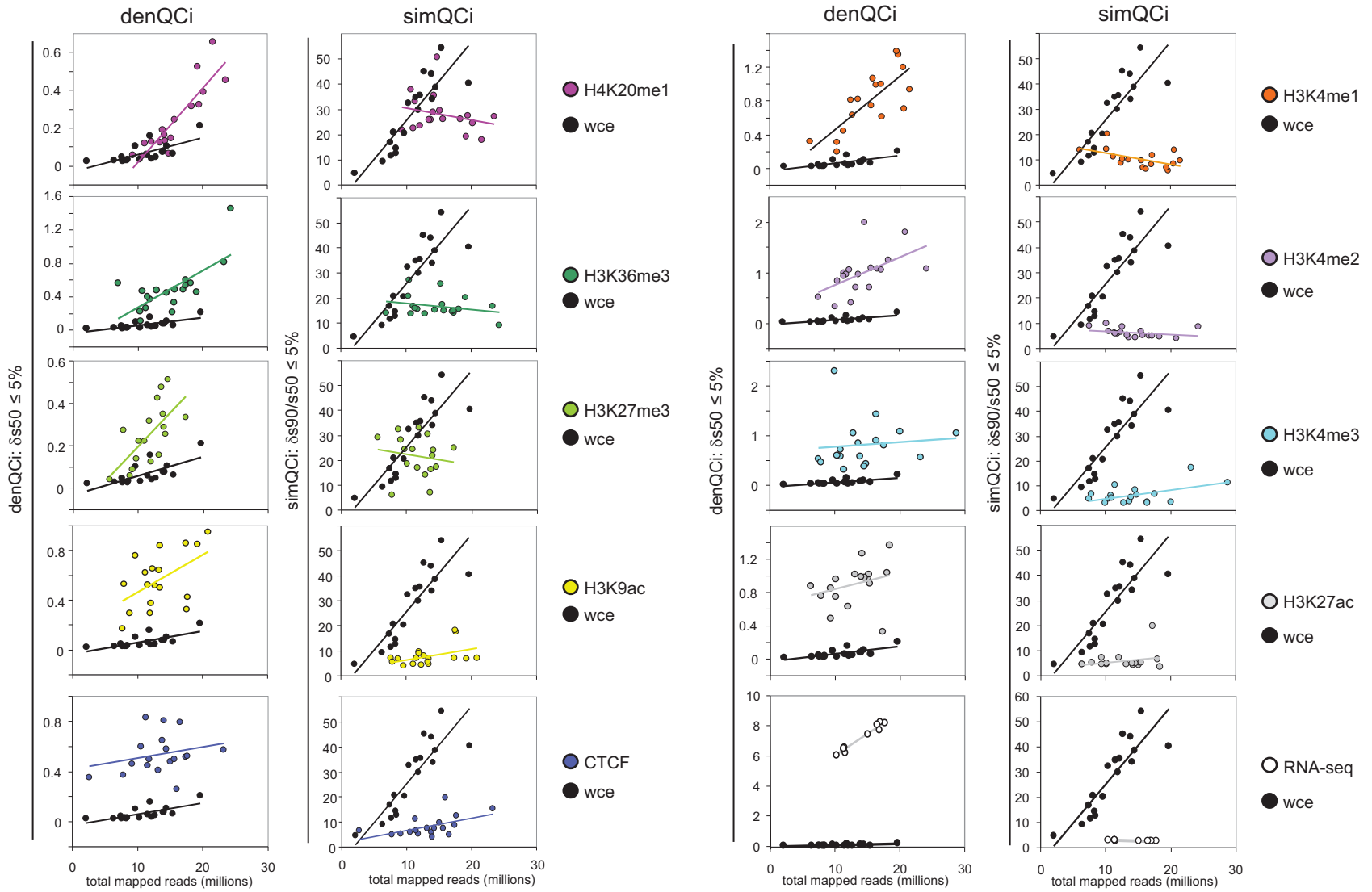


Supplementary Figure 2

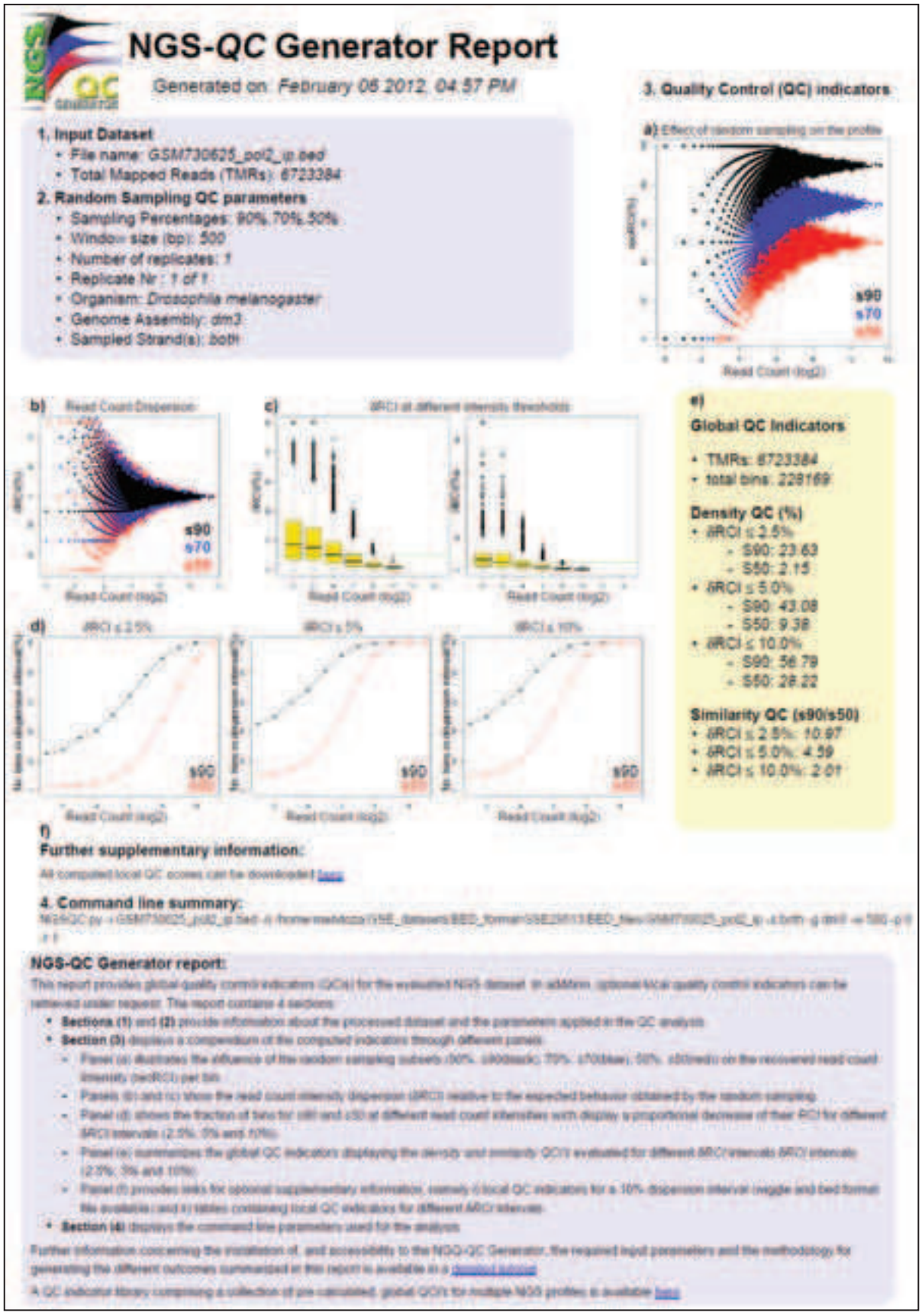
c

	TMRS	ds50≤2.5%	ds50≤5%	ds50≤10%	ds90/s50≤2.5%	ds90/s50≤5%	ds90/s50≤10%
GSM521901	19677564	0.85±0.007	2.086±0.011	3.712±0.004	3.84±0.03	2.48±0.01	1.75±0.0
CV (%)		0.83	0.54	0.12	0.70	0.40	0
GSM910579	19897316	0.18±0.0	0.834±0.005	3.964±0.009	32.39±0.22	14.39±0.04	4.184±0.009
CV (%)		0	0.66	0.22	0.68	0.28	0.21
GSM752985	19896191	0.1±0.0	0.542±0.004	3.544±0.011	53.3±0.6	22.08±0.09	4.74±0.01
CV (%)		0	0.82	0.32	1.09	0.45	0.27

Supplementary Figure 2. QC indicators reproducibility over TMRs random sampling replicates. (**a)** Three different publicly available H3K4me3 ChIP-seq datasets displayed a similar number of TMRs (~19 million reads), nevertheless their associated profiles present important differences in their read-count intensities as well as in their background levels. To assess such differences from a quantitative point of view, their TMRs have been randomly sampled in five replicates at three different sampled subsets (90%, 70% and 50%). RCI dispersion per 500bp bins (local QCi) for each of the sampling replicates are illustrated as color-coded heat map below the corresponding profile. **(b)** Scatter plot illustrating the influence of random sampling on the read count intensity of the evaluated profiles. Note that the dataset displaying the best enrichment pattern (GSM521901) presents a denser scatterplot at higher intensity values than the other compared datasets and that with increasing number of reads the recRCI/bin increasingly approaches the theoretically expected values. As previously indicated, global QCis are assessed by evaluating the fraction of bins presenting a RCI dispersion under a given threshold (i.e. ±2.5%; ±5% and ±10% dispersion from expectation). **(c)** Their corresponding global QCis assessed from all 5 random samplings are displayed (average ±standard deviation) for different denQCi and simQCi threshold conditions. In addition their related coefficient of variation (CV%) has been assessed. Importantly, all global QCis present CVs lower than 2% demonstrating that these global quality descriptors are quite stable even for a single TMRs random sampling assay. Finally, it is worth to mention that the infered QC indicators do correlate with the variable quality of enrichment observed for the compared profiles.



Supplementary Figure 3. QCis assessed for a diversity of NGS-generated profiles. All ChIP-seq and wce displayed datasets were originally published by Ernst J. et al (*Nature, 473, 43-9, 2011*). Furthermore, all displayed RNA-seq datasets were originally published by Trapnell et al. (*Nat. Biotechnol, 28, 511-5, 2010*).



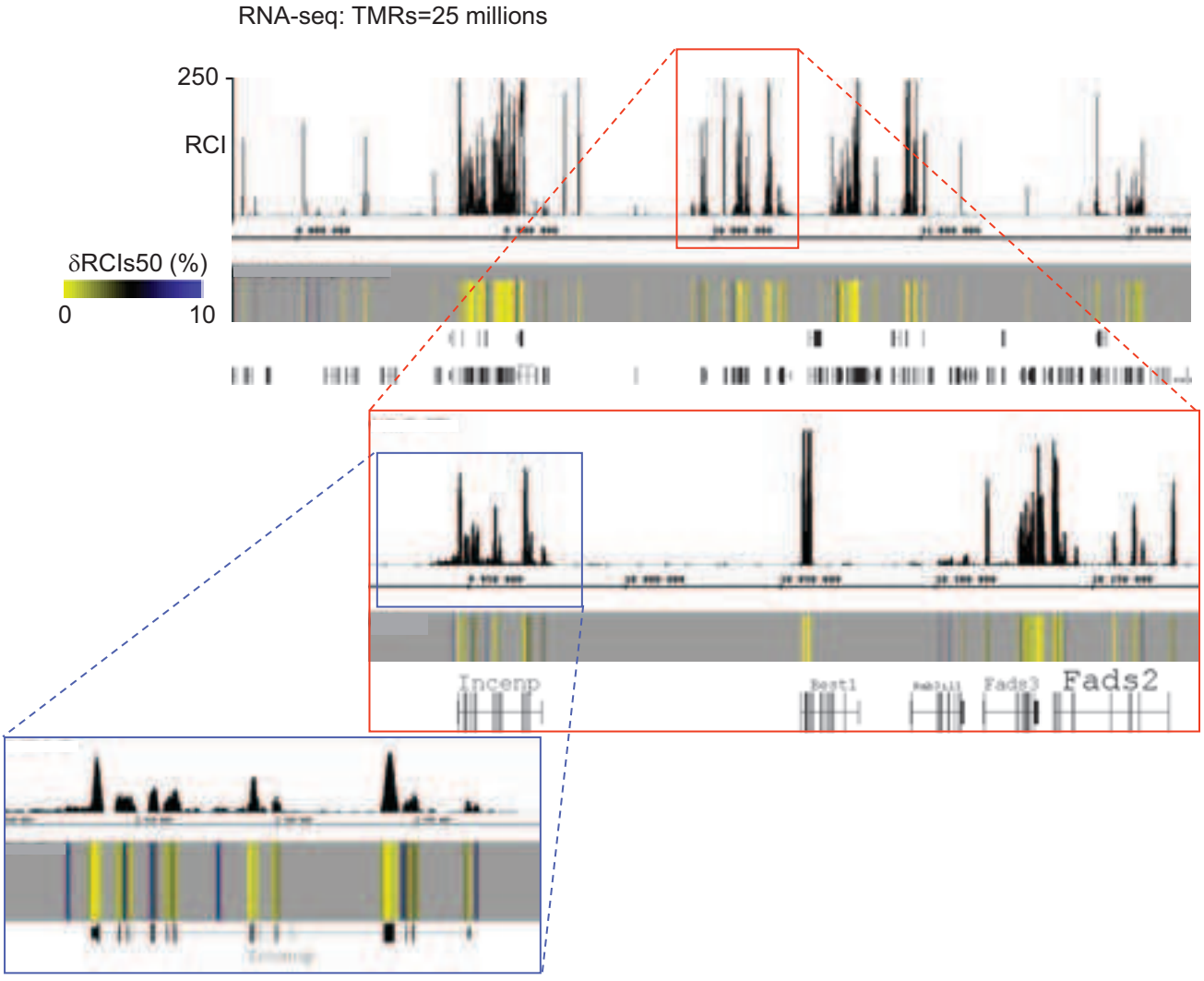
Supplementary Figure 4. Example of the QCi report generated automatically for each processed NGS-profile by the NGS-QC Generator.

a

chrom	start	end	s100	s90	s70	s50	recRCI s90(%)	recRCI s70(%)	recRCI s50(%)	δRCI s90	δRCI s70	δRCI s50
chr1	724001	724500	7	6	5	4	85.71	71.43	57.14	4.29	-1.43	-7.14
chr1	851001	851500	9	8	6	5	88.89	66.67	55.56	1.11	3.33	-5.56
chr1	856501	857000	11	10	8	6	90.91	72.73	54.55	-0.91	-2.73	-4.55
chr1	932501	933000	7	6	5	4	85.71	71.43	57.14	4.29	-1.43	-7.14
chr1	1008501	1009000	17	15	13	8	88.24	76.47	47.06	1.76	-6.47	2.94
chr1	1009001	1009500	268	241	190	137	89.93	70.90	51.12	0.07	-0.90	-1.12
chr1	1014501	1015000	48	42	36	27	87.50	75.00	56.25	2.50	-5.00	-6.25
chr1	1015001	1015500	52	45	41	28	86.54	78.85	53.85	3.46	-8.85	-3.85
chr1	1015501	1016000	33	31	23	19	93.94	69.70	57.58	-3.94	0.30	-7.58
chr1	1121501	1122000	6	5	4	3	83.33	66.67	50.00	6.67	3.33	0.00
chr1	1139501	1140000	7	6	5	3	85.71	71.43	42.86	4.29	-1.43	7.14
chr1	1829501	1830000	12	10	8	6	83.33	66.67	50.00	6.67	3.33	0.00

local QCi unit

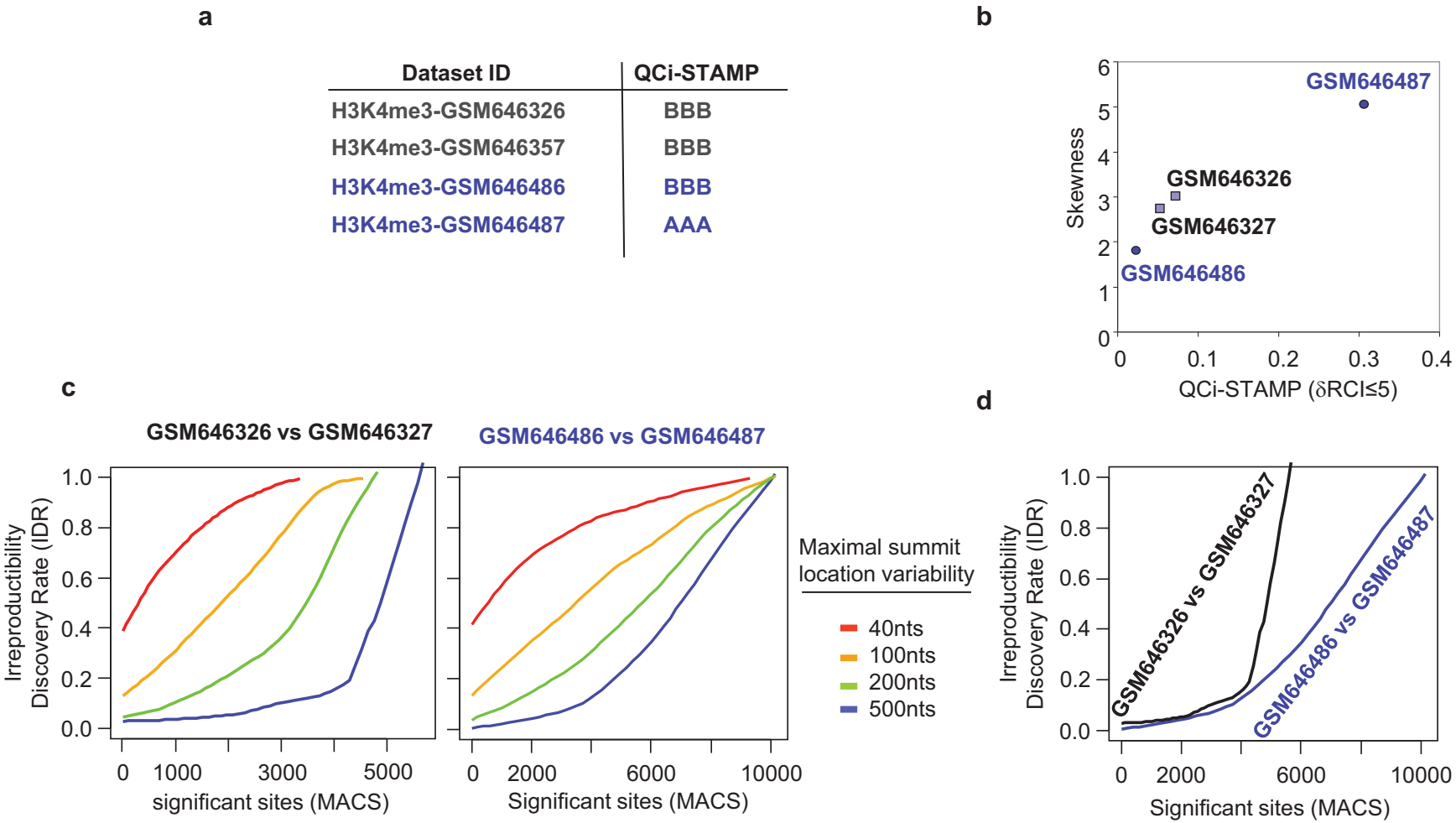
b



Supplementary Figure 5. Local QC indicators inferred from TMR random sampling.
a) Random sampling of total mapped reads (TMRs) at different densities (90%, 70% and 50% described herein as s90, s70 and s50 respectively) followed by read count intensity (RCI) computation per 500bps bins provides local QC indicators. The Local QCi unit correspond to the RCI dispersion assessed per 500bps bin. (b) local QCi displayed for an RNA-seq profile generated from 25 million TMRs. RCI dispersion for a 50% sampling density (δRCIS50) is displayed in a heatmap format and together with the corresponding RCI associated to the profile of interest.

a

Dataset ID	QCi-STAMP
H3K4me3-GSM646326	BBB
H3K4me3-GSM646357	BBB
H3K4me3-GSM646486	BBB
H3K4me3-GSM646487	AAA



Supplementary Figure 6. QCi-STAMP provides the same quality information as skewness and is independent of problems introduced by the use of peak callers. (a) A set of two biological duplicates was selected from publicly available H3K4me3 ChIP-seq profiles and their corresponding QCi-STAMP descriptors were determined. ("A" highest; "D" lowest quality). **(b)** Skewness of the read-counts signal distribution of the biological replicates compared with the predicted QCi-STAMP (δRCI≤5%). **(c)** Irreproducibility Discovery Rate (IDR) among H3K4me3 replicates was assessed by sorting significant binding sites by their p-value and comparing the population of common and unique sites per replicate. Common sites were assessed for different MACS-predicted summit location variability (red: 40nts, orange: 100nts, green: 200nts and blue, 500nts maximal summit location variability). **(d)** Comparison in IDR's performance (500nts summit variability) between H3K4me3 biological replicates. Note that in contrast to "sharp" binding patterns (e.g. CTCF in Fig. 6), IDR analysis requires more relaxed conditions to determine overlapping binding sites.

Supplementary File 1

Next Generation sequencing Quality controls Generator

Version 0.1; June, 2013

1. **Introduction**
2. **Running NGS-QC Generator**
    - *Input Dataset*
    - *Random sampling QC parameters*
    - *QC indicators*
3. **Interpretation of NGS-QC indicators**

    3.1 denQCi and simQCi guide ChIP-seq experiments
4. **A dynamic publicly available database of global and local QC indicators**
5. **Additional applications**
6. **Annexes**

    6.1 NGS-QC Generator availability

    6.2 Command line summary

    6.3 Practical aspects concerning the use of the NGS-QC Generator and database

1. Introduction

Comparative analyses between Next generation sequencing (NGS) generated profiles, such as ChIP-seq, RNA-seq, Gro-seq, or MeDIP-seq require prior characterization of the degree of technical similarity of the various data sets, as individual profiles can vary significantly even between biological replicates, the use of different antibodies and batch-to-batch variations of the same antibody, sequencing depth and immunoprecipitation (IP) quality are only a few of the parameters that impact on the quality of a ChIP-seq profile. The present NGS-QC Generator infers global and local quality indicators based on a stand-alone approach, as it does not require additional wet-lab efforts. This computational approach generates read count intensity profiles from randomly selected subsets of the total originally mapped reads (TMRs) associated to the NGS-profile under study and defines the divergence from the theoretically expected read count intensities (RCIs) recovery after sampling relative to the original profile. For this, TMRs are first randomly sampled at three different densities (90%, 70% and 50%; referred to hereafter as s90, s70 and s50 subsets, respectively); then the genomic RCI profile is recorded for successive 500bp bins and compared to that of the original profile. This comparison is performed to evaluate the divergence from the ideal condition in which the RCI/bin for a s50 subset correspond to 50% of the original RCI/bin value. Importantly, NGS-sampled generated profiles diverge always to different degrees from the hypothesized "ideal behaviour", thereby generating a quantifiable denominator (referred to as profile "robustness"), which is linked to the quality of any NGS-generate profile (Mendoza-Parra et al.; manuscript in preparation).

Below we describe the different steps involved in the NGS-QC Generator's accessibility through the web-based platform GALAXY, the required input parameters and the information provided in the NGS-QC Generator Report. In addition, we provide an interpretation of the different quality control indicators, give examples and discuss additional applications of this methodology.

2. Running NGS-QC Generator

The NGS-QC Generator requires as input a single file containing the genome positions of the uniquely aligned reads (BAM or BED format). Depending on the user-defined analysis, the following additional parameters may be required:

- **Genome**: Currently the following model organism genomes are supported: Homo sapiens (hg19, hg18); Mus musculus (mm9, mm8); Ratus norvegicus (rn4, rn3); Drosophila melanogaster (dm3, dm2); Caenorhabditis elegans (ce6, ce4); Dario rerio (dr6, dr4).

- **Strand specificity**: For ChIPs and related applications the reads/bin are cumulated from both complementary strands during data processing.

! ***NOTE*** The user can define strand-selective analysis for specific applications.

- **Windows size ('bin') for read counts enrichment assessment**: Currently the "default" parameter is set to 500 bp.

! ***NOTE*** For comparative analyses of several profiles, the QC indicators should be calculated using identical bin sizes, thus 500 bp windows should be used to compare QC indicators of a user profile with those displayed in the NGS-QC database available in our website: http://igbmc.fr/Gronemeyer_NGS_QC

We have studied the effects of bin size variation on the NGS-QC generator-calculated indicators and found the highest sensitivity (i.e., highest difference for the QC indicators assessed at different dispersion intervals) at a bin size of 500 bp.

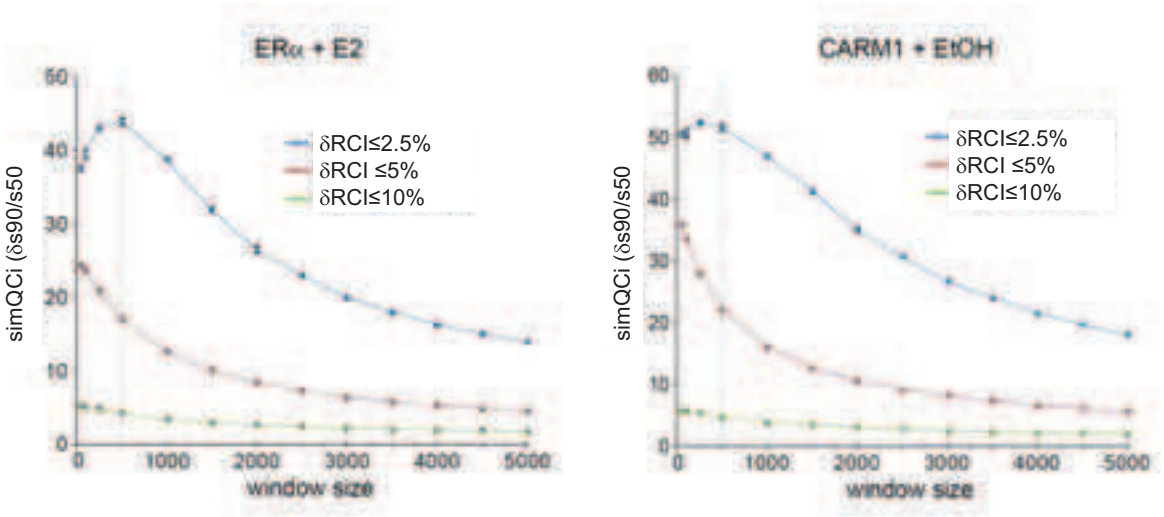
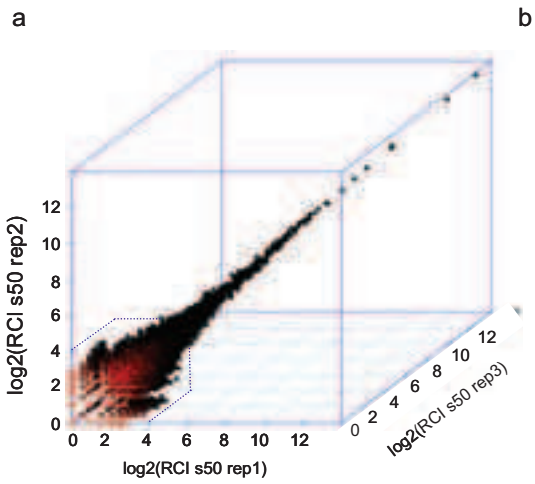


Figure I. Influence of the window size on the assessment of QC indicators for two different ChIP-seq profiles. Similarity QC indicator at different window sizes have been computed for Era and CARM1 ChIP-seq profiles. As highlighted by the vertical gray line, the highest difference for the simQC indicators assessed at three different dispersion intervals (2.5, 5, 10%) is retrieved for bins presenting a windows size between 250 and 500bp. Note that this value is in concordance with the expected chromatin fragmentation size.

- **Number of random sampling replicates**: We have obtained highly reproducible Global QC indicators (less than 2% coefficient of variation among five sampling replicates) when using several sampling replicates; however, users can choose the number of replicate samplings.

! ***NOTE*** This option is supported up to 3 replicates.



b

simQCi (δs90/s50)

dataset 1	rep1	rep2	rep3	rep4	rep5	average	CV%
dRCI≤2.5	8.36	8.33	8.23	8.36	8.24	8.30	0.78
dRCI≤5	6.08	6.08	6.04	6.04	6.04	6.06	0.38
dRCI≤10	3.22	3.23	3.23	3.24	3.22	3.23	0.23

dataset 2	rep1	rep2	rep3	rep4	rep5	average	CV%
dRCI≤2.5	45.75	44.29	45.26	45.80	44.61	45.06	1.37
dRCI≤5	14.39	14.39	14.44	14.32	14.42	14.39	0.33
dRCI≤10	3.76	3.77	3.77	3.78	3.76	3.76	0.14

dataset 3	rep1	rep2	rep3	rep4	rep5	average	CV%
dRCI≤2.5	43.83	44.36	43.51	43.86	44.56	44.04	1.04
dRCI≤5	18.58	18.43	18.44	18.35	18.55	18.47	0.56
dRCI≤10	4.09	4.07	4.06	4.08	4.07	4.08	0.23

dataset 4	rep1	rep2	rep3	rep4	rep5	average	CV%
dRCI≤2.5	50.06	50.43	50.09	51.54	49.20	50.26	1.89
dRCI≤5	20.97	21.22	21.18	21.12	21.08	21.11	0.47
dRCI≤10	4.67	4.88	4.71	4.67	4.66	4.68	0.41

Figure II. QC indicators reproducibility over TMRs random sampling replicates. a) TMRs associated to given ChIP-seq profile have been randomly sampled three times to a 50% density (s50). The read count intensity (RCI) recorded per 500bps bins in all three replicated are compared. Note that for RCI higher than 16 (4 in log2) the RCI correlation is quite high. **b)** In order to quantify replicates sampling robustness, four different ChIP-seq samples were sampled 5 times. After that, the similarity QC indicator (δs90/s50) was compiled for three dispersion intervals (2.5, 5.0 and 10.0%). Chart tables show the different values obtained for each replicate. Average and coefficient of variation (CV%) from all replicates were also computed. Note that in all the cases the CV% is less than 2%.

Several other optional parameters concerning the generation of local QC indicators in wiggle file format, or other complementary data files (see below for details) are implemented. All these parameters can be defined by the predefined entries displayed in the Galaxy platform-based version.
The NGS-QC Generator produces a certain number of output files which are summarized in a report available in PDF format (see **Supplementary Figure 3**). This report is subdivided in 3 sections as described below.

1) ***Input Dataset:*** Information associated with the processed dataset. The filename should contain information specifying target and assay type (targeted factor, epitope, and antibody source and batch specification for ChIP-seq, origin and type of RNA in RNA-seq, GRO-seq, etc), treatment (if any) before ChIP, the model system used and any other information that is considered useful for future *meta* analyses.

2) ***Random sampling QC parameters*:** Specification of the different parameters for data processing, including the percent of sampled reads, the bin window size and the number of replicate samplings. In addition, the genome assembly and strand-specific or global mode of operation are documented.

3) ***QC indicators*:** This section presents a compendium of the computed indicators in visual format and provides the quantitative QC indicators. As is explained below we distinguish as global QC indicators two parameters, the "density QC (denQC)" and "similarity QC (simQC)", and offer the possibility to attach a "local QC" to a given profile. The six panels (**a**) to (**f**) specify the following generated outputs:

a) *Influence of random sampling on the intensity of the evaluated profile*. This scatter plot illustrates the **o**riginal **R**ead **C**ount **I**ntensity per bin (*oRCI*) in the studied profile (x-axis) relative to the **rec**overed **R**ead **C**ount **I**ntensity (*recRCI*) after sampling (y-axis). This relationship is displayed as following:

$$recRCI = \left(\frac{samRCI}{oRCI}\right) * 100$$

Where *samRCI* corresponds to the RCI/bin retrieved after random sampling. As illustrated in **Figure III**, the theoretically expected *recRCI* is directly proportional to the random sampling density, i.e. 90% for s90, 70% for s70 and 50% for s50 respectively.

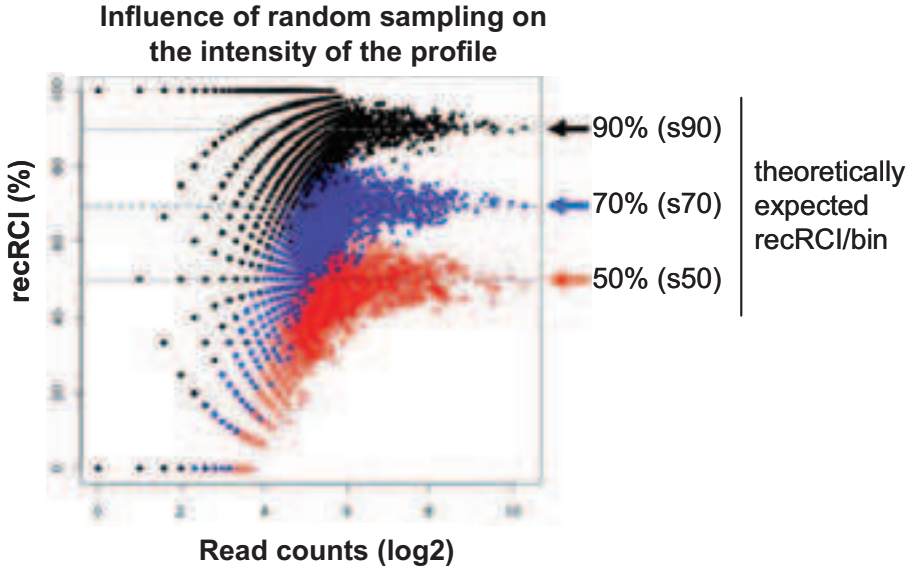


Figure III: Scatter plot illustrating the influence of random sampling on the read count intensity of a given profile. Each data point corresponds to the RCI within a 500 bp bin (x-axis) relative to the fraction of this intensity that is recovered after random sampling (y-axis). Note that for each of the three randomly sampled subsets, a different fraction of bins shows a proportional intensity recovery relative to the original read count measurements, and that with increasing number of reads the recRCI/bin increasingly approaches the theoretically expected value.

This initial analysis provides intuitive information about the quality of the generated profile[1]. In fact, profiles of good quality show high number of bins that display a proportional decrease of RCI/bin in the sampled subsets compared to the original dataset. Thus, the less dispersed the recRCI pattern is, the better is the quality of the associated profile. Note that towards low signal intensities ($<2^4$ read counts/bin) the sampling process inevitably results in increased dispersion.

b) *Read count dispersion.* To compare the dispersion effect relative to the expected proportional decrease in the RCI/bin values induced by random sampling, the

[1] « Quality » is defined here as the degree of dispersion from the theoretically expected recRCI scatter after sampling, which corresponds to a proportional decrease of all RCI/bin values relative to the sampling. With this definition a maximum of the quality indicator is reached when the recRCI/bin values are equal to the *oRCI* multiplied by the sampling percentage (i.e. 50% for s50). Any deviation – for whatever reason - from the expected RCI/bin scatter provides a quantitative indicator of the quality of a given NGS-profile.

scatter plot displayed in Figure 1 has been first centered by the following expression:

$$\partial RCI = samd - recRCI$$

where "*δRCI"* corresponds to the ***read count intensity dispersion*** and "*samd"* to the random sampling density (i.e. 90%, 70% and 50% for s90, s70 and s50 respectively). This transformation facilitates to identify the subset of bins that display a *δRCI within in a given interval*, such as *δRCI≤5* (illustrated in **Figure IV**). This information represents *per se* a quantifiable indicator for the quality of the studied profile. The current version of NGS-QCi Generator provides global quality indicators for dispersion intervals of 2.5%, 5% and 10%. In addition, the quality indicators for each 500bp bin are also generated in a wiggle format (described below as "local QC indicators").

★**Convention:** The measurement of the fraction of bins displaying a δRCI within in a given interval constitutes the global density QC indicator denQCi. The denQCi is described by the term "δs50/5" in which "s50" specifies the sampling in percentage and "5" the δRCI threshold.

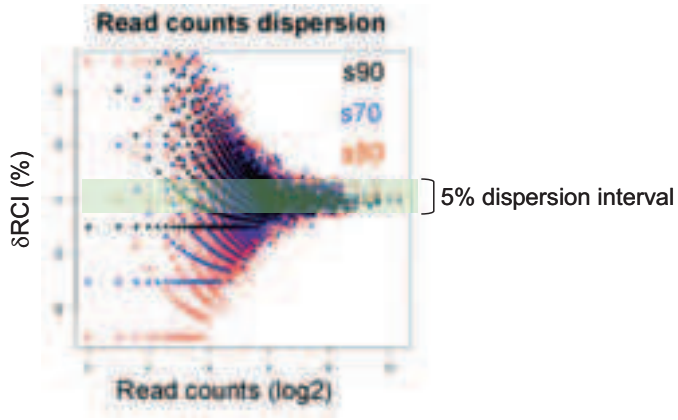


Figure IV: Scatter plot illustrating the recovered read count intensity dispersion of a given profile. This transformed scatter plot superimposes the three scatter plots obtained after sampling for the same original dataset. The scatter of bins after sampling at s50, s70 or s90 having RCI values that deviate ≤5% from the expected RCI/bin are highlighted.

c) *δRCI at different Intensity thresholds. A*s is apparent from **Figure IV** the dispersion of the read counts/bin after sampling and thus, the quality of the profile is inversely proportional to the RCI. In **Figure V** this dispersion (δRCI) is calculated for both the s90 and s50 randomly sampled and reconstructed profiles relative to the original s100 dataset. Note that in the illustrated example, bins with RCIs greater than 16 (4 in log2) present a median δRCI lower than 5% for both the re-sampled data sets. Importantly, for high quality profiles such a 5% threshold extends to lower RCI/bin values than for low quality profiles. Moreover, a similar dispersion pattern in s50 and s90 data sets is a sign for a high quality and "sampling robustness" of the evaluated profile; thus, the degree of similarity of the δRCIs of s90 and s50 data sets is a second quantifiable indicator that is evaluated by the NGS-QCi Generator.

★**Convention:** δRCI(s90/s50) constitutes the global similarity QC indicator simQCi. The simQCI is described as "δs90/s50/5", in which "δs90/s50" corresponds the ratio between the denQCi for s90 and the denQCi for s50 and "5" specifies the δRCI threshold.

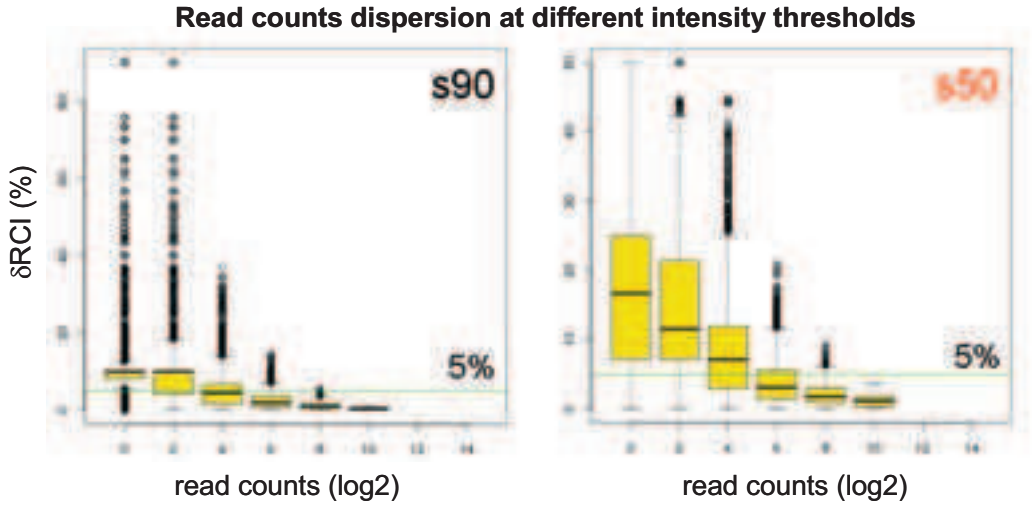


Figure V: δRCI evaluated at different intensity thresholds for both s90 and s50 random sampling subsets. The 5% δRCI threshold is indicated as a green line.

d) *Number of bins at different δRCI intervals.* This analysis computes the fraction of bins in the sampled subset (i.e. s90 or s50) that exhibits a proportional decrease of their RCI for a given RCI threshold. A δRCI of 2.5% defines a very stringent condition, as nearly 50% of bins with RCI values above 256 (2^8) - corresponding to strong signals - are outside this interval (**Figure VI**, left panel). In contrast, more than 80% of such strong signals are within the interval defined by a δRCI threshold of 5% (middle panel); for more relaxed conditions, such as dRCI≤10%, all these signals in within the selected interval (right panel).

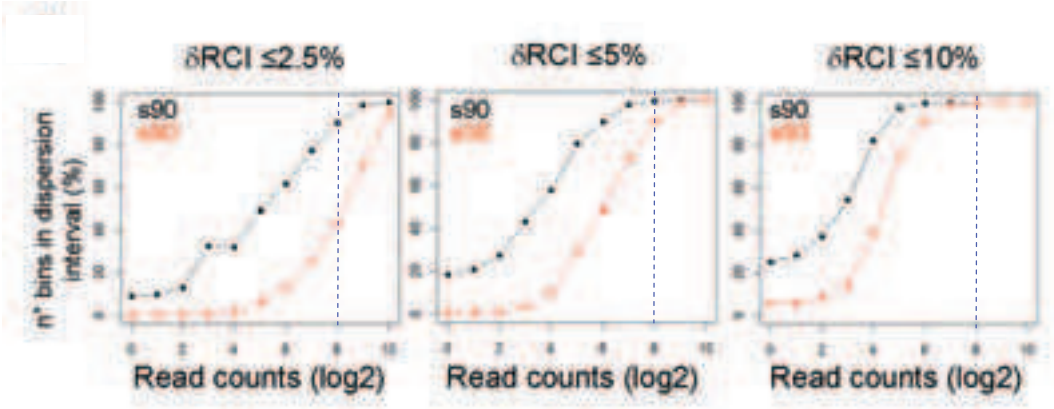


Figure Figure VI: Fraction of bins at different δRCI intervals. For a given RCI threshold the fraction of bins presenting a δRCI equal or lower than the indicated threshold (2.5%, 5% or 10%) is evaluated for s90 and s50 subsets.

e) *Global QC indicators.* All previous Panels (**a**) to (**d**) illustrate several characteristics associated to TMR distribution at different random sampling densities. Such characteristics represent a read-out for the quality of the evaluated

profile, which is the consequence of several factors implicated in its genesis. Below the global QC indicators , which represent "fingerprints" of an evaluated profile, and the corresponding acronyms are summarized:

- **N° of reads**: TMRs used for the profile reconstruction.
- **Total bins**: Number of bins (500bp window size) presenting with at least one read in the original profile.
- **Density QC** (denQCi): The fraction of bins in the s90 or in s50 subsets with a δRCI lower than the default dispersion thresholds (2,5%; 5% and 10%). Note that the higher the density QCi is, the better is the quality of the associated profile.
- **Similarity QC** [simQCi(s90/s50)]: Ratio between the density QCis for s90 and s50 subsets at the different dispersion thresholds. The simQCi reveals the similarity of the s90 and the s50 profiles. As a rule of thumb, the closer this value is to 1, the better is the quality of the studied profile.

Both the density and similarity QCis represent quantifiable NGS-profiles quality indicators, thus they can be used for comparative purposes as described below. Note that QC indicators associated to several publicly available NGS-generated profiles can be retrieved in our website:

http://igbmc.fr/Gronemeyer_NGS_QC

f) *Further supplementary information*. Taken in consideration that the above analyses were computed for 500bp bins, the δRCI/bin data can be used to provide local QC indicators. Such information is provided by the NGS-QC Generator either in a wiggle or in a BED format; the default condition identifies bins with δRCI≤ 10%[2]. **Figure VII** illustrates how the local QCis can be displayed in a heat-map format linked to the original read count intensity profile. This display option is useful to visualize the predicted δRCIs associated to a given chromatin region of interest.

Optionally, 500bp chromatin regions with δRCIs thresholds of 2.5%; 5% or 10% can be downloaded as a table in BED format. All the items in panel (**f**) are user-defined; note that the corresponding files may reach Gb size.

[2] Local QC indicators in wiggle file format can be uploaded in the Integrated Genome Browser (IGB) and displayed in heat-map format; the corresponding BED file can be uploaded in the UCSC browser.

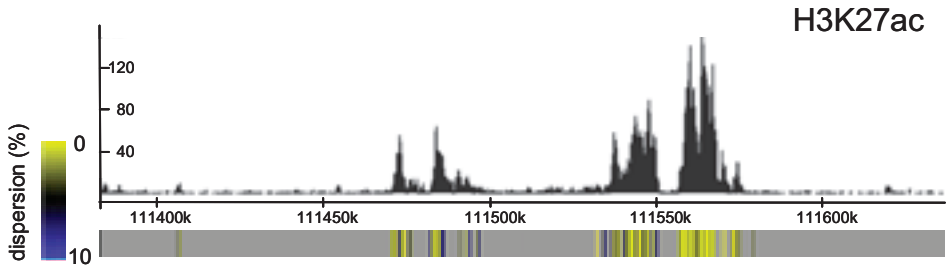


Figure VII: H3K27ac ChIP-seq profile displayed together with the corresponding local QC indicators. Bellow the ChIP-seq profile the corresponding δRCI for each 500bp bin are displayed for a 10% threshold using the heat map illustration indicated on the left. Only bins with δRCI ≤10% are shown.

3. Interpretation of NGS-QC indicators

The quality indicators described by the NGS-QC Generator are derived from the question of how different a given NGS profile would be if only a subset of the total mapped reads were used? The underlying concept is that in the ideal case, the read counts intensities will decrease proportionally to the fraction of sampled reads. From this two quality indicators are derived.

The ***density QC indicator (denQCi)*** makes reference to the fraction of the evaluated chromatin regions (sectioned into 500bp bins) that comply with this proportional within a defined dispersion margin, such as 5% at a sampling ration of 50% (i.e. δs50/5). The maximal theoretical value for denQCi is 100.

The ***similarity QC indicator (simQCi)*** refers to the fraction of chromatin regions which reveal a proportional decrease of RCIs in the subset sampled at 90% relative to that sampled at 50% and is given for a specified dRCI threshold (e.g., ds90/s50/5). The minimal theoretical value for simQCi is 1.

3.1 denQCi and simQCi guide ChIP-seq experiments

Figure VIII illustrates that QC indicators can vary dramatically between experiments; indeed, publicly available ChIP-seq data provide useful information about the range of denQCi and simQCi that have been achieved in previous experiments for a given target and (batch of) antibody, such that a user can judge the QC performance of a ChIP-seq relative to past data sets. Moreover, the library of QC indicators (available at http://igbmc.fr/Gronemeyer_NGS_QC) provides a guide to users about the possible effect of the sequencing depth on ChIP-seq quality. Indeed, the comparison of several H4K20me1 profiles[3] demonstrates that at least 15 million total mapped reads are required to obtain QCis that differentiate between the ChIP-derived and the non-enriched ("input") datasets. In contrast, H3K4me3 ChIP-seq profiles present fairly good QCis even for TMRs lower than 15 million reads.

[3] The compared ChIP-seq profiles were taken from an study performed in nine human cell types following a production pipeline for chromatin immunoprecipitation (Ernst J. et al. 2011 Nature **473**; 43-49)

! ***NOTE*** Importantly, in both profiles individual ChIP-seq profiles can be observed which have been performed at similar sequencing depths but data analysis reveals nevertheless greatly varying global QCi indicators. This underscores the notion that in addition to the sequencing depth (multiple) other factors, whose effects cumulate along the experimental pathwtowards to final data set, influence the quality of the profile.

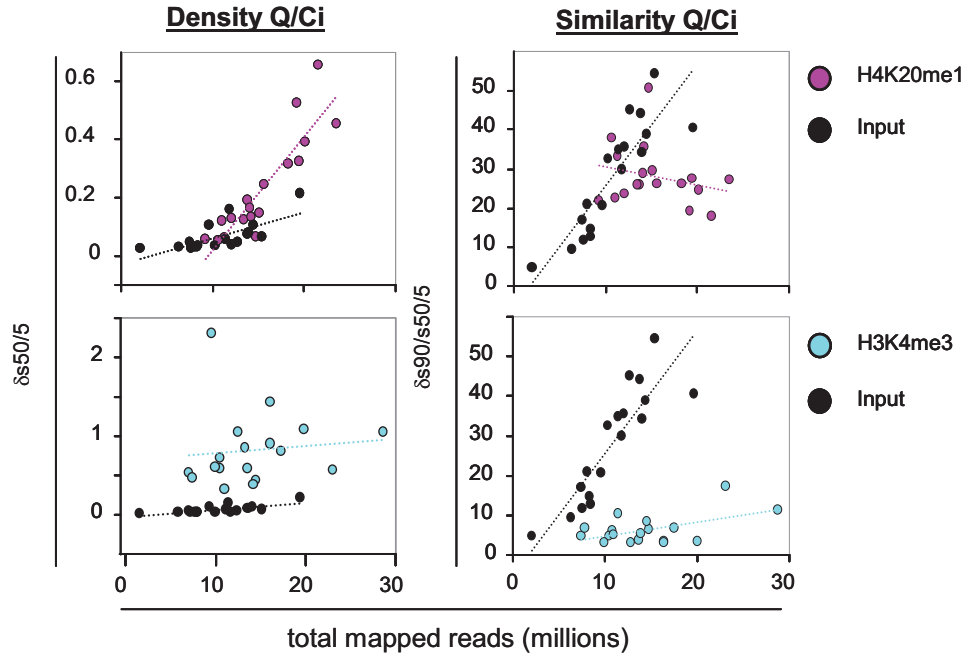


Figure VIII: Density and Similarity QCis for several ChIP-seq profiles in the context of their total mapped reads. Top and bottom panels illustrate QCis for H4K20me1 and H3K4me3 ChIP-seq profiles, respectively. In addition, QCis for the non-enriched input datasets are illustrated for comparative purpose. Notice that in contrast to the H3K4me3 datasets, H4K20me1 profiles reconstructed from up to 15 million reads present QCis similar to those observed for the input datasets. Importantly for such histone modification profiles, increase in the sequencing depth beyond this 15 million reads threshold allows to retrieve QC indicators diverging from the Input datasets behavior.

4. A dynamic publicly available database of global and local QC indicators

With the aim of establishing a dynamic guide for NGS users we have created a QC indicator database comprising a collection of global QCis for multiple NGS profiles. This database, which is available online to the scientific community through our website http://igbmc.fr/Gronemeyer_NGS_QC, will be expanded to include most, if not all, global and local QCis of the NGS profiles currently available from GEO. In addition, future profiles will be integrated and users may evaluate their NGS profiles and compare them with stored QCi.
To facilitate and simplify the recognition of QCi divergence between profiles we have defined QC-STAMP, a global descriptor that combines the information provided by denQCi and simQCi as following:

$$QC\_STAMP = \frac{denQCi}{simQCi}$$

In order to evaluate de divergence of this global descriptor over all enrichment-related NGS profiles currently compiled in the NGS-QC database, the QC-STAMP distributions assessed for three different RCI dispersion intervals was subdivided in four quantiles to which the following grades have been attributed: “D”, lower quartile (<25%); “C”, inter-quartile 25-

50%: “B”, inter-quartile 50-75% and “A” upper quartile (>75%). The NGS-QC Generator database associates these grades for 2.5, 5 and 10% δRCI to each profile as a three letter symbol, such that, for example AAA (“triple A”) reveals an A grade for all three δRCIs. All available profiles are displayed as a dynamic QC-STAMP *vs.* TMR scatterplot, which allows judging of their QCi similarities in the context of the sequencing depth. Note that the global QC-STAMP descriptor will be dynamically re-evaluated when novel entries are provided to the database.

Considering the inherent relationship between the current NGS repositories and our QC database, we aim to integrate in a long term a direct connection between the Galaxy version of the NGS-QCi Generator and the QCi database in order to simplify the repository of this information and to establish links with GEO in order to coordinate the generation of such indicators in a systematic manner[4].

5. **Additional applications**

The presented bioinformatics-based QC system uses the total mapped reads associated to any NGS data sets to infer a set of global QC indicators. In fact, profile’s quality evaluation does not rely in a given Peak calling algorithm, thus it can be directly applied to any type of NGS-generated profile, including RNA-seq, GRO-seq, etc, in addition to the wide variety of ChIP-seq assays (transcription factors, insulators, histone modifications, RNA Polymerase II, etc). For the same reason, the inferred QC indicators are fully comparable, making of this approach a universal tool for multidimensional quality profiles comparison.
We believe that the global QC indicators will be useful for the development, characterization and comparison of antibodies directed towards a particular target. There are considerable variations between different antibodies and different batches of polyclonal antibodies. The certification of antibodies for ChIP-seq using the present QC systems should improve ChIP-seq reproducibility and comparability.

The quality of any NGS profile is the direct consequence of a complex number of factors, including aspects like crosslinking efficiency, chromatin shearing, antibody affinity and selectivity, as well as the variability between experiments and experimenters. While the QC indicators described here cannot *per se* identify the source for quality differences between profiles, they reveal the comparability and non-comparability of different NGS-generated profiles.

! ***NOTE*** The sequencing depth used to generating NGS-profiles can now be used as a tuneable parameter to identify profiles of similar quality. For this, correlative analyses between the inferred QC indicators and the performed sequencing depth will be very useful.

[4] The QCis generated in the current NGS-QCi Generator Galaxy version are not transferred into the QCi database, but in a further version we may establish such link; thus users will be invited to allow such a transfer. In addition, the identity of the sample will not be required, but certain information like the nature of the NGS profile, the antibody source, etc may be requested (without a mandatory condition) in order to associate a comprehensive description of the evaluated samples to the QCi data set.

6. Annexes

6.1 NGS-QC Generator availability

For providing a simple way to access to the community, NGS-QC Generator has been made available through a customized Galaxy cloud instance dedicate to this application (access provided in our website: http://igbmc.fr/Gronemeyer_NGS_QC).

Furthermore, an executable version of the NGS-QC Generator can be downloaded from our above indicated website. Importantly, such stand-alone version requires BEDtools to be installed on the hosting system. It can be retrieved at:

Stable releases: http://code.google.com/p/bedtools
Repository: https://github.com/arq5x/bedtools

A detailed description for the execution of such stand-alone version is available as part of the downloadable file.

6.2 Command line summary

To document and assure accurate reproducibility of the computational treatment we have included in each report the complete command line used for its generation. This information makes reference to the computation core implemented in the heart of the NGS-QC Generator tool:

```
python NGS-QC.pl –i ERa_e2_1h_H3396_sc-543 -o ERa_e2_sampled -s both -g hg19 -w 500 -p 8 -r 1 –-sampleList 90,70,50 –-pcList 2.5,5,10
```

Where:

- "python NGS-QC.py" calls the NGS-QC Generator script
- –i ERa_e2_1h_H3396_sc-543 indicates the dataset to be processed
- -o ERa_e2_sampled refers to the name of the output directory to which all output files will be saved
- -s both can be used to sample both strands. To sample only the forward (reverse) strand, use -s fw (-s rev).
- -z hg19 refers to the processed genome. It requires to be followed by the genome assembly used for TMRs alignment (i.e. mm8, mm9, hg18, hg19, etc).
- -w 500 corresponds to the applied windows size.
- -p 8 corresponds to the number of CPUs used in the parallel processing.
- -r 1 refers to the performed sampling replicas.
- --sampleList 90, 70, 50 corresponds to the random sampled fractions; i.e. 90%, 70%,50% respectively
- –-pcList 2.5,5,10 corresponds to the dispersion percentage thresholds to be used

6.3 Practical aspects concerning the use of the NGS-QC Generator and database

The NGS-QC Generator and the corresponding QC indicator database are accessible from our website (http://igbmc.fr/Gronemeyer_NGS_QC). For assessing the quality of a dataset, users can access the NGS-QC Generator through our customised web-based GALAXY platform. For it users can register by providing an e-mail address as login and a password. This step is mainly required for the use of an FTP server to facilitate the uploading large size data files.

In case a user prefers to remain anonymous five guest accounts are available:

Login account	**password**
guest1@galaxy.igbmc.fr	NTYyM2RiND
guest2@galaxy.igbmc.fr	ZjY4NGFjMz
guest3@galaxy.igbmc.fr	MDBhZTMzM2
guest4@galaxy.igbmc.fr	OTllZWI0Mj
guest5@galaxy.igbmc.fr	YWQ2NDRkM2

Furthermore, due to storage space constraints, uploaded datasets into the Galaxy instance may not be available for more than 24hours, thus we strongly suggest users to download their processed files as early as possible.

When required, some example datasets are available on the “shared_Data” access as part of the Data libraries, thus users may upload them for having a trial run on the NGS-QC generator tool.

5.2. Epimetheus - A multi-profile normalizer for epigenome sequencing data

The preferred technique for epigenetic analysis Chromatin ImmunoPrecipitation-Sequencing (ChIP-Seq) is inherently prone to significant variabilities embedded in individual assays. Multiple factors like antibody efficacy, sequencing library efficiency and depth have a direct impact on data quality and thus on any downstream analysis. But even high quality datasets generally exhibit significant sequencing depth variation, which requires normalization. Currently, existing normalization tools are limited in different aspects, namely (i) annotation dependency, (ii) restriction to specific regions, (iii) less user-friendly and (iv) scalability to a variety of downstream analyses. Moreover, the existing approaches are mostly intended for particular analysis, thus their normalization outputs are not readily exportable to downstream analysis such as chromatin state prediction involving multiple samples; also most of these tools require specialized programming skills. To overcome these restrictions we have developed Epimetheus, a quantile-based multi-profile normalization tool for histone modification data. Epimetheus is written in combination of Perl, C & R and will be freely available to the academic community.

5.2.1. Methodology

There are four main steps involved in Epimetheus pipeline: (i) processing raw alignment data, (ii) building read count intensity (RCI) matrices, (iii) quantile normalization followed by Z-score normalization and (iv) generating normalized BED file and other outputs including plots (Figure 25). As quantile normalization is an absolute read count based approach, any region or technical specific bias will over/under-represent the read counts and lead to bias in downstream analyses. Hence, clonal reads are systematically removed from the analysis unless otherwise specified by the user, and reads are extended to given average fragment size.

A read counts matrix is built by dividing reference genome ‘G’ into small non-overlapping sequential bins and the RCI for each bin is calculated. Size (‘S’) of the bin can be from 100 $\geq S \leq$ 500bp depending on the enrichment pattern (sharp/broad) of histone mark. Let us denote a target histone mark as ‘X’, and ‘Xa’ & ‘Xb’ will represent two samples of same target. Genomic bins for ‘Xa’ can be represented as Xa_1, Xa_2, $Xa_3 \ldots Xa_n$ whereas Xb_1,

Xb_2, Xb_3…Xb_n will represent ‘Xb’, where ‘n’ depends on the sizes of ‘S’ and ‘G’ (G/S). Reads, which overlap with each bin, are counted to calculate reads per bin (RpB) distribution for each sample. As a result, ‘Xa’ and ‘Xb’ will form two libraries as follows.

$$Xa = \{ Xa_i \mid 1 \geq i \leq n \}$$

$$Xb = \{ Xb_i \mid 1 \geq i \leq n \}$$

Similarly ‘Ya’ and ‘Yb’ will form two different libraries of another target histone mark. Using RCI calculation from each sample, a B×N matrix is built for each target mark individually, where B is the total number of bins (for a given ‘G’ and ‘S’) and N is the number of samples. In case of multiple histone marks in the analysis, similarly B×N1, B×N2, etc., will be generated.

Quantile normalization is a rank based normalization, thus different level of intensities are normalized together giving Quantile normalization corrects the coverage differences by (i) sorting each sample’s RpB in ascending order individually, (ii) ranking the values for each sample individually and (iii) calculating the average of corresponding rank values from different samples and re-sort to its original position. This results in a normalized matrix norm(B×N), where each sample will have normalized-RpB (nRpB). Subsequently, Z-score scaling is applied over the normalized matrix to generate znorm(B×N) which is calculated by the distance of each nRpB from a mean value of total nRpB in the sample, divided by the standard deviation as follows.

$$Z_i = \frac{X_i - \overline{X}}{s}$$

Where, X_i is RCI of a given window from population X, Z_i is Z-score normalized RCI of X_i, $\overline{X}$ and s is mean and standard deviation of the population. Quantile normalized results are only meant to be considered for further analysis whereas Z-score normalization is meant for overall inter-target comparison (as plots) only.

5.2.2. Output

Epimetheus generates three types of outputs: (a) visualization files, (b) plots and (c) normalized BED files. Visualization files are text files (in bedgraph format) generated for raw and normalized RCI which can be used in visualization browsers, and can be used for

other downstream analyses as well. To assess the overall difference among samples distribution and the normalization effect, a MA transformation plot is generated to compare samples pairwise before and after normalization. The tool is also capable of extracting target specific (promoter/gene-body/custom regions) matrix of raw and normalized read counts, and produce average RCI plots for the same (refer Figure 25 for example plots).

One of the salient features in Epimetheus as compared to existing methods is that it can produce normalized BED files, representing the changes (normalization effect) in alignment BED file; thus it can be directly used for further downstream analysis. With respect to the change in read counts after normalization for a given bin, increasing counts post normalization is done by adding new reads aligned randomly to a new position within the bin; similarly, existing reads are removed to decrease read counts. As BED format is the most preferred input format in most of the ChIP-seq analysis, Epimetheus enables the direct use of normalized data for downstream ChIP-seq analysis.

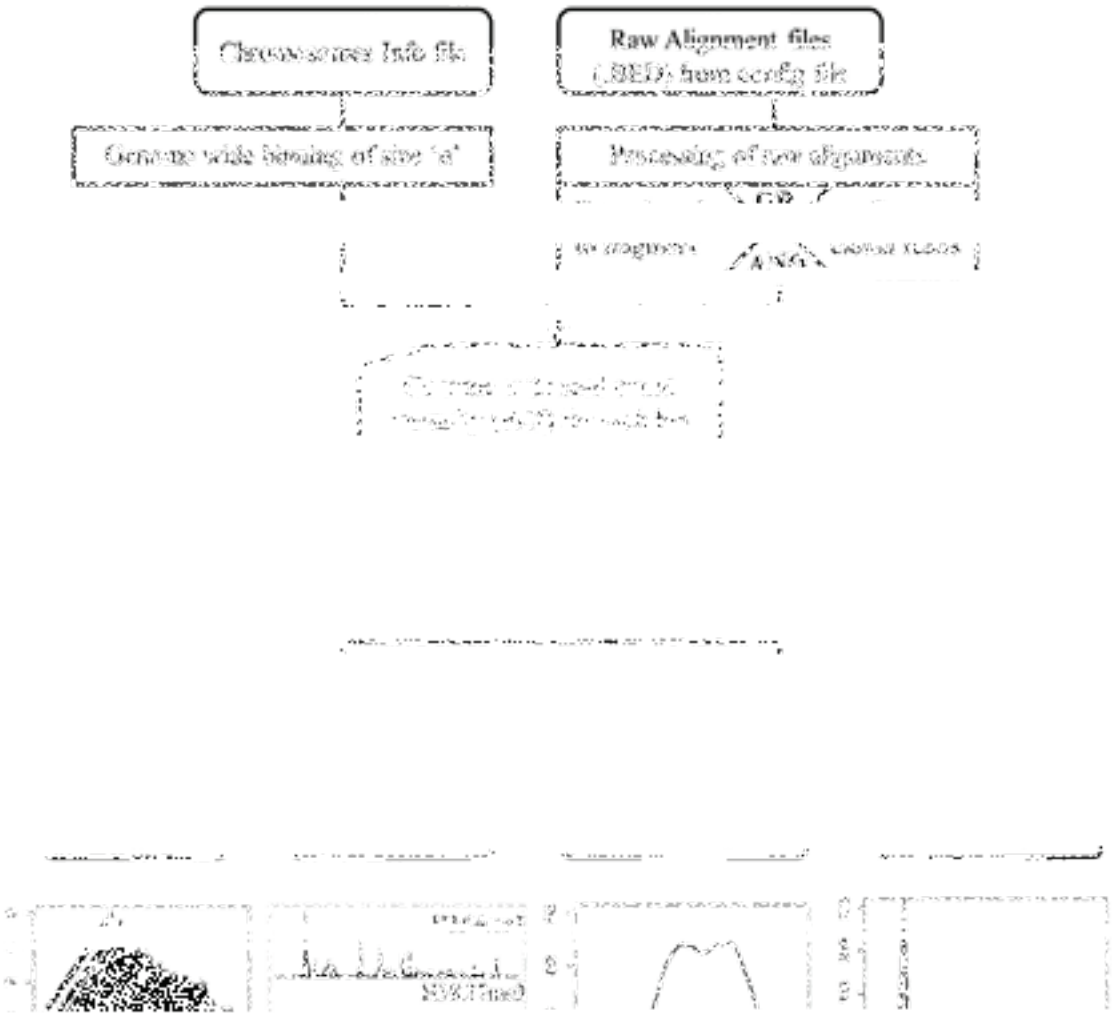


Figure 25. A scheme of the Epimetheus workflow with illustrative plots (refer chapter 5.2.1 for detailed explanation). Apart from normalization, Epimetheus provides illustrative plots to understand the enrichment over promoter region (TSS plot), gene body (gene body plot for RNA PolII) and MA plot to compare the samples before and after normalization. Additionally, it provides BEDGRAPH files which can be loaded into genome browsers to visualize the data.

5.2.3. Evaluation of Epimetheus on multiple datasets

Epimetheus performance has been validated by using biological replicates of H3K4me3 profiles in HepG2 cells (Ernst et al., 2011). Prior to normalization, these replicates exhibited highly similar numbers of promoter-enriched sites as expected. However, most likely due to technical variability they exhibited significant differences in the enrichment level and signal-noise ratios. Epimetheus adjusted such amplitude differences considering the signal-noise ratio disparity among samples. To verify the consequence of normalization on peak calling approach, MACS was used for the HepG2 raw and normalized data sets. Though few differences in peaks counts (due to fluctuation in less enriched sites) were observed, the overall the amplitude differences were corrected (Figure 26A). Using ChromHMM, we compared chromatin state attributions before and after normalization using the previously reported profiles of nine different histone marks in nine cell lines (Ernst et al., 2011). The comparison revealed small but significant changes in chromatin state annotation (2-7%) of genomic bins. Importantly, chromatin state annotations of several genes changed from active to poised and *vice versa*, which generally coincided with their expression levels (Example gene locus: *MYO7A* Figure 26B.

Epimetheus has been applied to evaluate the relative enrichment levels of H3K27me3, H3K4me3 and RNA polymerase II (hereafter termed as PolII) recruitment in temporal analyses of F9 cell differentiation (Mendoza-Parra et al., manuscript submitted). The raw RCIs of ‘repressive’ H3K27me3 marks showed an unexpected, apparently variable enrichment in the *Hoxa* cluster region over time. However, after normalization, the previously described collinear gene activation pattern (Kashyap et al., 2011; Montavon and Duboule, 2013) with progressive loss of ‘repressive’ and gain of ‘active’ histone marks, and PolII recruitment was observed and confirmed by qPCR (Figure 26C).

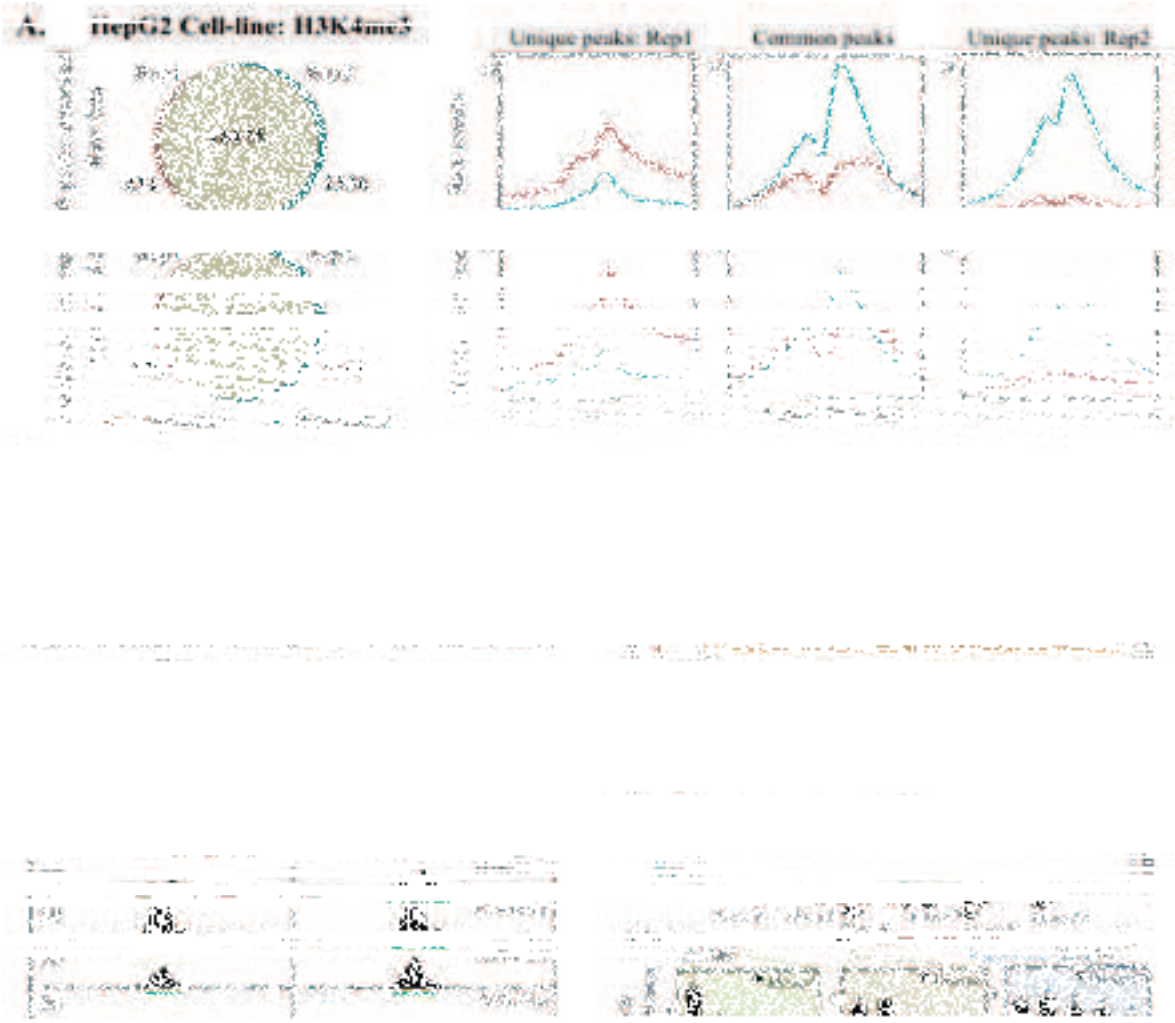


Figure 26. Effects of data normalization. (A) Left panel; Pie charts illustrating common and replicate-specific promoter-associated enrichment events derived from a published dataset before and after normalization. Right panel; Enrichment plots over annotated promoters shows that normalization results in more similar RCI profiles for common peaks and more distinctive profiles for replicate-specific enrichments. (B) Illustration of change in chromatin state annotation for the *MYO7A* locus using the same dataset processed with ChromHMM; note that the *MYO7A* promoter was annotated 'active' from the raw data and changed to 'poised' post normalization, which correlated with the absence of gene expression [Encode data: ENCSR962TBJ]. (C) Signal intensity profile of H3K27me3 enrichment over the *Hoxa* cluster during retinoic acid-induced differentiation of F9 mouse embryo carcinoma cells. Note that in contrast to the raw data normalization results in a gradual decrease of the H3K27me3 profile over time, which correlates with the qPCR data displayed at the bottom.

5.2.4. Discussion

Inherent sequencing depth variation in NGS has made normalization imperative when performing NGS-profiles comparison in the context of their relative signal amplitude levels. Correction by total number of reads (linear normalization) has been widely used in earlier days, especially in RNA-seq. However, one of the main differences between ChIP-seq and other technologies is that specificity and efficacy of the pull-down methods which gives rise to inevitable varying background noise. Such background noise and alignment related bias makes the peak callers unable to differentiate accurately less enriched peaks (small bumps) from background. For the same reason linear normalization cannot be applied to ChIP-seq data. When different antibodies yield different level backgrounds in samples, the normalization by total number of reads would create bias. Hence, to correct different level of intensities among samples, we employed quantile normalization in Epimetheus as it is based on a ranking approach. More importantly, quantile normalization can handle multiple samples as opposed to LOWESS which can perform only pair-wise normalization.

Importantly, as Epimetheus is a genome-wide approach, it is annotation free, thus avoiding bias from external factors. Currently existing tools depend on peak callers' enrichment predictions and/or WCE (whole cell extract, also known as input). Such dependency could lead to potential sources for artifactual normalizations given that diversity in available peak callers' results and the bias introduced by an external dataset like WCE. While most of the control datasets used are generally enrichment-less, few WCE controls can exhibit enrichment-like artifactual patterns (for example, GSM788366 and GSM768313) leading to false negative annotation in enrichment sites identification. This will not only affect the peak calling but will also significantly influence the normalization outcomes given the importance of population/distribution in normalization methods. In that context, we have demonstrated that a selection of population (genome or targeted regions) can influence the normalization (refer Supplementary Figure 3 in the attached manuscript). Hence, a prior quality assessment over control datasets, as for IP assays, is strongly suggested to identify and exclude poor quality datasets from the analysis.

While most of the tools focus on normalization only for differential analysis, the above studies on biological replicates and chromatin state analysis illustrates the need for normalization in any comparative or integrative analysis as well. Though normalization may seem irrelevant in position level comparison of peaks/enrichments, we have observed that amplitude changes influence identification of enrichments. Similarly, changes in amplitude are significant in analyses where enrichment is binarized like that performed by ChromHMM where identification of different patterns and level of enrichments is crucial.

Compared to existing tools, the more robust and sophisticated options in Epimetheus are that it (i) can be customised to variety of requirements, (ii) can be applied genome-wide, or to specific regions (when justified), and (iii) can exclude specific regions, which could be considered to bias the global normalization (e.g. repetitive elements). More importantly, Epimetheus provides analytical outputs which are exportable to variety of downstream analyses.

MANUSCRIPT 2

EPIMETHEUS - A MULTI-PROFILE NORMALIZER FOR EPIGENOME SEQUENCING DATA

Epimetheus - A multi-profile normalizer for epigenomic sequencing data

Mohamed-Ashick M. Saleem, Marco-Antonio Mendoza-Parra, Pierre-Etienne Cholley and Hinrich Gronemeyer*

Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Equipe Labellisée Ligue Contre le Cancer, Centre National de la Recherche Scientifique UMR 7104, InstitutNational de la Santé et de la Recherche Médicale U964, University of Strasbourg, Illkirch, France.

* To whom correspondence should be addressed.

Hinrich Gronemeyer
Tel: +(33) 3 88 65 34 73
Fax: +(33) 3 88 65 34 37
E-mail: hg@igbmc.u-strasbg.fr

ABSTRACT

Exponentially increasing numbers of epigenomic datasets in public repositories like GEO, which harbors presently several thousands, constitutes an enormous source of dramatically expanding information. This fosters and supports a growing interest in integrative and comparative studies to explore the gene regulatory mechanisms to its core. Today's challenge is to define functionally informative local and global patterns of chromatin states for different (patho-) physiological systems in a multi-dimensional perspective. Critically, the most preferred Chromatin ImmunoPrecipitation-Sequencing (ChIP-Seq) is inherently prone to significant variabilities embedded in individual assays, which pose several types of bioinformatic challenges for comparative studies, such as normalization to adjust sequencing depth variation. Currently existing normalization methods either apply linear scaling corrections and/or are restricted to specific genomic regions. To overcome these restrictions we developed Epimetheus, a genome-wide quantile-based multi-profile normalization tool for histone modification data and related datasets.

INTRODUCTION

Epigenetics is a crucial stratum of the complex multi-layered gene regulatory mechanism. With the advancements and cost-reduction in high-throughput sequencing, next generation sequencing (NGS) technology has become a quick and comprehensive medium to explore epigenome and related studies. Studying the epigenome status (and its reorganisation) involves sequencing of several histone modifications with the aim of characterising the state of chromatin in different genomic regions and between different samples. However, its assessment via chromatin immunoprecipitation is inherently prone to significant variabilities embedded in individual assays, posing different bioinformatic challenges for comparative studies - a general caveat in Big Data integrative analysis.

Multiple factors like antibody efficacy, sequencing library accuracy and depth have a direct impact on data quality and thus on any downstream analysis. Therefore, it is imperative to evaluate the quality of data prior to comparative studies (see for example, www.ngs-qc.org)(1). But even high quality datasets generally exhibit significant technology/user-derived signal amplitude differences, which require normalization prior to comparative analysis. Initially linear normalization, where the counts will be represented relative to total number of reads, was widely used. However, inherent differences in signal-noise ratio among samples (for instance generated with different antibodies) proved linear normalization to be unsuitable for ChIP-seq.

While significant computational efforts have been made in the past for single ChIP-seq data analysis, sophisticated computational and experimental methods to correct technical variabilities among multi-sample ChIP-seq analyses is acquiring importance. For example, Taslim *et al.,*(2) proposed a two-step non-linear approach, based on a locally weighted regression (LOESS) method to correct such differences among ChIP-seq data. LOESS's restriction to pairwise normalization led us to develop Polyphemus(3), a multi-profile normalization approach for RNA polymerase II (RNA PolII) datasets based on quantile correction, a widely used method in microarray studies(4). Since then, other quantile based normalization tools have been developed like ChIPnorm(5) and Epigenomix(6) for histone modification data.

Apart from previously described ChIP-seq specific tools, few popular RNA-seq based tools like DESeq(7) and EdgeR(8) are also used for ChIP-seq data. But these tools are limited to linear scaling, unlike RNA-seq, it is problematic considering inherent technical variation (signal-noise ratio) in ChIP-seq. More importantly, all the above-mentioned tools are limited in different aspects, namely (i) annotation dependency, (ii) restriction to specific regions, (iii) less user-friendly and (iv) scalability to a variety of downstream analyses. Moreover, the existing approaches are mostly intended for particular analysis, thus their normalization outputs are not readily exportable to several other basic ChIP-seq analyses involving multiple samples; also most of these tools require some programming skills. To overcome these restrictions we developed Epimetheus, a quantile-based multi-profile normalization tool. The genome-wide normalization procedure applied by Epimetheus enables optimal processing of different enrichment pattern datasets such as broad/sharp histone modification or PolII-seq profiles, chromatin accessibility profiles generated by FAIRE-seq(9) or ATAC-seq(10), or DNase-seq(11), as well as MeDIP-seq (12). Furthermore, users have the possibility to exclude specific genomic regions like, for example, repetitive elements or any other genomic locations for which artifactual enrichments might be expected.

MATERIAL AND METHODS

The basic assumption in quantile normalization is a common read-count distribution of compared datasets. In cases where the enrichment events under comparison comprise factors that are implicated in house-keeping events, it is reasonable to assume that the distribution of the read counts for a given target will be similar across different cell types(5). As for gene expression analysis (RNA-seq and microarrays) or RNA polymerase II enrichment (Polyphemus(3)), where quantile has been widely used, histone modifications are expected to occur at house-keeping as well as at cell/tissue-

specifically expressed/repressed genes. With this assumption, genome-wide quantile normalization is applied separately for each target. Subsequently Z-score scaling is used such that each dataset is represented relative to its mean of distribution, which renders different target histone data comparable. The Epimetheus pipeline involves four main steps: (i) processing raw alignment data, (ii) building read count intensity (RCI) matrices, (iii) two subsequent levels of normalization (quantile and Z-score) and (iv) generating the outputs and plots (detailed scheme - Supplementary Figure 1).

Processing of data

As quantile normalization is an absolute read count based approach, any region or technical specific bias will over/under-represent the read counts and lead to bias in downstream analyses. Clonal reads (i.e., PCR duplicates) constitute such a technical bias. Unfortunately, some level of clonal read contamination is unavoidable in sequencing datasets involving PCR. Epimetheus will remove such clonal reads from the raw alignment data, unless otherwise specified by the user. There are few alignment and platform-specific biases that should be addressed prior to analysis as these are specific to each data and pipeline. Particularly recommended is to remove reads with more than one perfect alignment and those aligned to repeat and centromere regions. Reads are elongated to a specified length to represent the average fragment length (150-300bp) as typically only the first 50-100 base pairs are sequenced in ChIP-seq.

Read count intensities

For quantile normalization, an approach similar to that of Xu *et al*. 2008(13) and Mendoza-Parra *et al*., 2011(3) is followed, where the reference genome 'G' (or custom regions for target-specific normalization) is divided into small non-overlapping sequential bins and the RCI for each bin is calculated. Size ('S') of the bin can be from 100 ≥ S ≤ 500bp depending on the enrichment pattern (sharp/broad) of histone mark. We choose an optimal 100-500bp bin size to preserve the shape of enrichment pattern(1).

Let X be a target histone mark and X_a & X_b be two samples of same target. Genomic bins for X_a will be x_{a1}, x_{a2}, x_{a3}…x_{an} whereas for X_b will be x_{b1}, x_{b2}, x_{b3}…x_{bn}, where 'n' depends on the sizes of 'S' and 'G' (G/S). Reads, which overlap with each bin, are counted to calculate reads per bin (RpB) distribution for each sample. As a result, $X_a = \{x_{ai} \mid 1 \geq i \leq n\}$ and $X_b = \{x_{bi} \mid 1 \geq i \leq n\}$ will be two libraries. Similarly Y_a and Y_b will be two different libraries of another target histone mark.

Normalization

Using RCI calculation results, a B×N matrix is built, where B is the total number of bins (for a given 'G' and 'S') and N is the number of samples. In case of multiple histone marks in the analysis, similarly B×N1, B×N2, etc., will be generated. A quantile based approach cannot be applied to normalize different histone marks with different enrichment patterns as the distribution and amplitude will be highly dissimilar which would nullify the initial assumption. The differences in coverage among samples are adjusted to same level by (i) sorting each sample's RpB in ascending order individually,

(ii) ranking the values for each sample individually and (iii) calculating the average of corresponding rank values and re-sort to its original position.
This results in a normalized matrix norm(B×N), where each sample will have normalized-RpB (nRpB). Subsequently, Z-score scaling is applied over the normalized matrix to generate znorm(B×N) which is calculated by the distance of each nRpB from a mean value of total nRpB in the sample, divided by the standard deviation.

Output

In contrast to previously described methods, Epimetheus produces normalized BED files by adding/removing reads with respect to normalized per-bin RCIs using raw alignment BED files as reference. With respect to the change in read counts after normalization for a given bin, increasing counts post normalization is done by adding new reads aligned randomly to a new position within the bin; similarly, existing reads are removed to decrease read counts. As BED format is the most preferred input format in most of the ChIP-seq analysis, Epimetheus enables the direct use of normalized data for downstream ChIP-seq analysis.
Along with normalized BED output, Epimetheus produces three types of outputs: (a) visualization files, (b) plots and (c) normalized BED files. Visualization files are text files (in bedgraph format) generated for raw and normalized RCI, which can be used for other downstream analyses as well. To assess the difference among samples and the effect of normalization, a MA transformation plot(14) is generated to compare samples pairwise before and after normalization. The tool is also capable of generating target specific (promoter/gene-body/custom regions) matrix with RpB along with corresponding average RCI plots is generated.

RESULTS

Datasets used for the evaluation purpose are subjected to quality control using NGS-QC (www.ngs-qc.org). To avoid biases, clonal reads were excluded from the analysis in all datasets.

Biological replicates

Epimetheus performance has been validated by using biological replicates of H3K4me3 from nine different cell lines (GEO file GSE26320)(15). Biological replicates are a standard procedure to reveal the effect of normalization, as the datasets are expected to be highly similar but may differ in enrichment amplitudes. However, possibly due to technical variability some of these replicates exhibited significant disparities in signal-noise ratios and some differences in the number of enrichment sites. As illustrated in Supplementary Figure 2, GM12878 data exhibit varying signal-noise ratio, whereas HMEC data exhibit similar background and less enriched sites amplitude level but significant differences for the highly enriched sites. In such a situation, linear normalization fails to correct and instead generates artifacts (Supplementary Figure 2A). Epimetheus adjusted such amplitude differences considering the signal-to-noise ratio disparity among samples given its ranking-based approach (Supplementary Figure 2B). To highlight its effect, an average RCI TSS plot displays amplitude differences among different level of enrichment within sample (background, less, medium

and high) before and after normalization (Supplementary Figure 2C). To verify the consequence of normalization on basic ChIP-seq pipeline, peak calling was performed for the HepG2 raw and normalized data sets. Though few differences in peaks count (due to fluctuation in less enriched sites) were observed (Figure 1A), the overall amplitude differences were corrected (Figure 1B). Interestingly, the peaks size was also different between replicates with one being broader than the other. An overall shift in amplitude is evident with LOWESS fit line in MA transformation plot(14) for the raw and normalized data between replicates (Figure 1C).

Chromatin state analysis and Peak calling with normalization

To illustrate performance and scalability of Epimetheus in multi-profile analysis, chromatin state analysis was performed using ChromHMM(16) on nine cell-lines with nine histone marks datasets (previously mentioned GSE26320). ChromHMM identifies enriched regions based on Poisson background distribution, which does not account for differences in background locally. To avoid that, peak calling was carried out on raw and normalized data and peak regions were provided as input to annotate enriched regions for chromatin state analysis. Consistency of peaks before and after normalization for some samples suffered depending on the quality and coverage of the data (Figure 2A). Interestingly, datasets that shows significant disparity between peaks from raw and normalized data are of low quality (using the NGS QC Generator; www.ngs-qc.org) or coverage. For example, the H3K27ac profile of the H1 cell line exhibits higher disparity between peaks from raw and normalized data; as only one replicate data is available, this data set resulted in low coverage and quality of the data - CCD (where AAA is highest and DDD is lowest). Similarly, few other datasets which show high disparity is influenced by either quality or coverage.

Epimetheus normalization had less effect in chromatin states prediction except few enrichment level differences (Figure 2B and 2C). But significant differences were observed in chromatin state annotation of individual genomic bins (2-7%) after normalization (Figure 2E). GM12878, NHEK and NHLF cell lines shows fewer changes in chromatin state annotation but rest of the cell lines show more than 5% change with few of them occurring in promoter regions. Importantly some genes presented their promoter chromatin status changed between active and poised state. For example, chromatin state annotations of *MYO7A* gene changed from active to poised state due to prominent enrichment of H3K27me3 mark post normalization (Figure 2D). On comparison with transcriptome data downloaded from ENCODE(17), no expression signal was found, which correlates well with the chromatin state annotation assessed after, but not with the one before normalization.

Temporal epigenetics dynamics during retinoic acid-induced F9 cell differentiation

We then evaluated Epimetheus performance on time-series data where distinct gradual gain or loss of amplitude is expected. In this perspective, we used the well-characterized F9 mouse embryonal carcinoma (EC) cell model differentiated under retinoic acid treatment(18). In this study, cells after treatment of all-trans retinoic acid (RA) were collected over the first 48 hours (0h, 2h, 6h, 24 and 48h). Each of the collected samples was used for assessing the epigenetic status by profiling the repressive histone modification mark H3K27me3, the transcriptionally active modification mark H3K4me3 and

recruitment of RNA polymerase II (data unpublished). It has been reported that *Hoxa* cluster exhibits collinear gene activation pattern during differentiation where gradual gain of H3K4me3 mark and PolII recruitment but loss of H3K27me3 mark over the time(19)(20). But significant and non-uniform disparity in overall coverage among samples was observed. On inspection over *Hoxa* cluster, all three targets present variable enrichment levels on the *Hoxa* cluster over the assessed time points. Epimetheus corrected those differences as illustrated in Figure 3. After normalization, H3K27me3 mark displays a gradual decrease whereas active mark H3K4me3 and PolII recruitment shows a gradual gain over time presenting the previously described collinear gene activation pattern.

To further support the normalization results, we validated the H3K27me3 enrichment levels on various regions of the *Hoxa* cluster using a quantitative PCR (qPCR) assay. As illustrated in Figure 3, qPCR results are correlating with the same pattern as in normalization results and collinear gene activation pattern.

DISCUSSION

Demonstration of normalization effect on variety of datasets and analyses clearly implicates that normalization is imperative when performing NGS-profiles comparison in the context of their relative signal amplitude levels. Though normalization may seem irrelevant in position level comparison of peaks/enrichments, we have shown in this study (Figure 2 and 3) that amplitude changes influence identification of enrichments. Similarly, changes in amplitude are significant in analyses where enrichment is binarized like that performed by ChromHMM (Figure 2) where identification of different patterns and level of enrichments is crucial.

In contrast to existing normalization tools, Epimetheus provides analytical outputs compatible with variety of downstream analyses. More importantly, the previously described tools depend on peak callers' enrichment predictions and/or control datasets (WCE or input). Given the diversity in available peak callers (and their associated multiple parameters) and the bias introduced by an external dataset like WCE could lead to a potential sources for artifactual normalizations. Specifically, while control datasets used are generally optimal, a few WCE controls exhibits enrichment-like artifactual patterns (for example, GSM788366 and GSM768313) leading to true negative annotation in enrichment sites identification which will significantly influence the normalization outcomes produced by tools like ChIP-norm. For this reason, a prior quality assessment over control datasets, as for IP assays, is strongly suggested (www.ngs-qc.org). Like any other analysis, performance of Epimetheus also depends on quality of the data.

Similarly, we have also demonstrated that selection of population (genome or targeted regions) can influence the normalization. In fact, comparison of different target regions (approach similar to ChIPnorm but different fold values of Input Vs IP were used to identify enriched regions) revealed population related biases in normalization results (Supplementary Figure 3). Compared to existing tools, the more robust and sophisticated options in Epimetheus are that it (i) can be customised to variety of requirements, (ii) can be applied genome-wide, or to specific regions (when justified), and (iii) can exclude specific regions, which could be considered to bias the global normalization (e.g. repetitive elements).

Based on the above results, it is evident that normalization should be made pre-requisite for any comparative analysis on epigenome data. While most of the tools focus on normalization only for differential analysis, above studies on biological replicates and chromatin state analysis are supporting the need of normalization on any comparative, integrative and differential analysis. While linear scaling and RNA-seq based tools alone are to an extent incapable to address the dynamic variations embedded in ChIP-seq. Similarly, quantile and LOWESS based ChIP-seq specific normalization tools are intended for specific analysis and not scalable to other type of analysis. In respect to above issues, Epimetheus is developed to have non-linear normalization with scalability to variety of downstream analysis.

AVAILABILITY

EPIMETHEUS has been written in combination of R, C and Perl and is made freely available at https://github.com/modash/Epimetheus

ACCESSION NUMBER

Details of data and its analysis used are provided in Supplementary data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

Studies in the laboratory of HG were supported by the AVIESAN-ITMO Cancer, the Ligue National Contre le Cancer (HG; Equipe Labellisée); and the Institut National du Cancer (INCa). Support of the Agence Nationale de la Recherche (ANRT-07-PCVI-0031-01, ANR-10-LABX- 0030-INRT and ANR-10-IDEX-0002-02) is acknowledged.

ACKNOWLEDGEMENT

We would like to thank all the members of the IGBMC sequencing platform for sequencing library preparation assays in the context of the F9 cell differentiation project. Furthermore, we thank all members of H. Gronemeyer's laboratory for discussion related to the applications of EPIMETHEUS.

REFERENCES

1. Mendoza-Parra,M.A., Van Gool,W., Saleem,M.A.M., Ceschin,D.G. and Gronemeyer,H. (2013) A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res.*, **41**.

2. Taslim,C., Wu,J., Yan,P., Singer,G., Parvin,J., Huang,T., Lin,S. and Huang,K. (2009) Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics*, **25**, 2334–40.

3. Mendoza-Parra,M. a., Sankar,M., Walia,M. and Gronemeyer,H. (2012) POLYPHEMUS: R package for comparative analysis of RNA polymerase II ChIP-seq profiles by non-linear normalization. *Nucleic Acids Res.*, **40**, e30.

4. Qiu,X., Wu,H. and Hu,R. (2013) The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics*, **14**, 124.

5. Nair,N.U., Das Sahu,A., Bucher,P. and Moret,B.M.E. (2012) Chipnorm: A statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. *PLoS One*, **7**, e39573.

6. Klein,H.U., Schäfer,M., Porse,B.T., Hasemann,M.S., Ickstadt,K. and Dugas,M. (2014) Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. *Bioinformatics*, **30**, 1154–1162.

7. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

8. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–40.

9. Giresi,P.G., Kim,J., Mcdaniell,R.M., Iyer,V.R. and Lieb,J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. 10.1101/gr.5533506.Freely.

10. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

11. Hesselberth,J.R., Chen,X., Zhang,Z., Sabo,P.J., Sandstrom,R., Reynolds,A.P., Thurman,R.E., Neph,S., Kuehn,M.S., Noble,W.S., *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.

12. Jacinto,F. V., Ballestar,E. and Esteller,M. (2008) Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA methylome. *Biotechniques*, **44**, 35–43.

13. Xu,H., Wei,C.-L., Lin,F. and Sung,W.-K. (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, **24**, 2344–9.

14. Shao,Z., Zhang,Y., Yuan,G.-C., Orkin,S.H. and Waxman,D.J. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.*, **13**, R16.

15. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M., *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–9.

16. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

17. Djebali,S., Davis,C. a., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F., *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

18. Alonso,A. and Breuer,B. (1991) The F9-EC cell line as a model for the analysis of differentiation. **397**, 389–397.

19. Kashyap,V., Gudas,L.J., Brenet,F., Funk,P., Viale,A. and Scandura,J.M. (2011) Epigenomic reorganization of the clustered Hox genes in embryonic stem cells induced by retinoic acid. *J. Biol. Chem.*, **286**, 3250–60.

20. Montavon,T. and Duboule,D. (2013) Chromatin organization and global regulation of Hox gene clusters. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **368**, 20120367.

FIGURES LEGENDS

Figure 1. Effects of data normalization. (**A**) Pie charts illustrating common and replicate-specific promoter-associated enrichment events derived from a published dataset2 before and after normalization. (**B**) Enrichment plots over annotated promoters show that normalization results in more similar RCI profiles for common peaks and more distinctive profiles for replicate-specific enrichments. (**C**) MA transformation plot before and after normalization showing the overall effect of normalization between replicates. (**D**) An example region of signal profile displaying the amplitude difference between replicates before and after normalization (note the difference in amplitude range for each track).

Figure 2. Chromatin state analysis using ChromHMM. (**A**) Illustration of peaks' overlap between raw and normalized data for nine different marks on nine different cell-lines distinguished in shape and colour respectively. (**B**) Emission parameters of ChromHMM describing chromatin state differences between raw and normalized peaks. (**C**) An example region illustrating the change after normalization corresponding to the change in chromatin state 14. (**D**) Illustration of change in chromatin state annotation for the *MYO7A* locus using the same dataset processed with ChromHMM; note that the *MYO7A* promoter was annotated 'active' from the raw data and changed to 'poised' post normalization, which correlated with the absence of gene expression [Encode data: ENCSR962TBJ]. (**E**) Stacked bar chart showing the percentage of chromatin state annotation/bin changed after normalization.

Figure 3. Signal intensity profile of H3K4me3, H3K27me3 and RNAPolII enrichment over the *Hoxa* cluster during retinoic acid-induced differentiation of F9 mouse embryo carcinoma cells. Note that in contrast to the raw data normalization results in a gradual decrease of the H3K27me3 profile and

gradual increase of H3K4me3 & RNAPolII profiles over time, which correlates with the qPCR validation of H3K27me3 data displayed at the bottom.

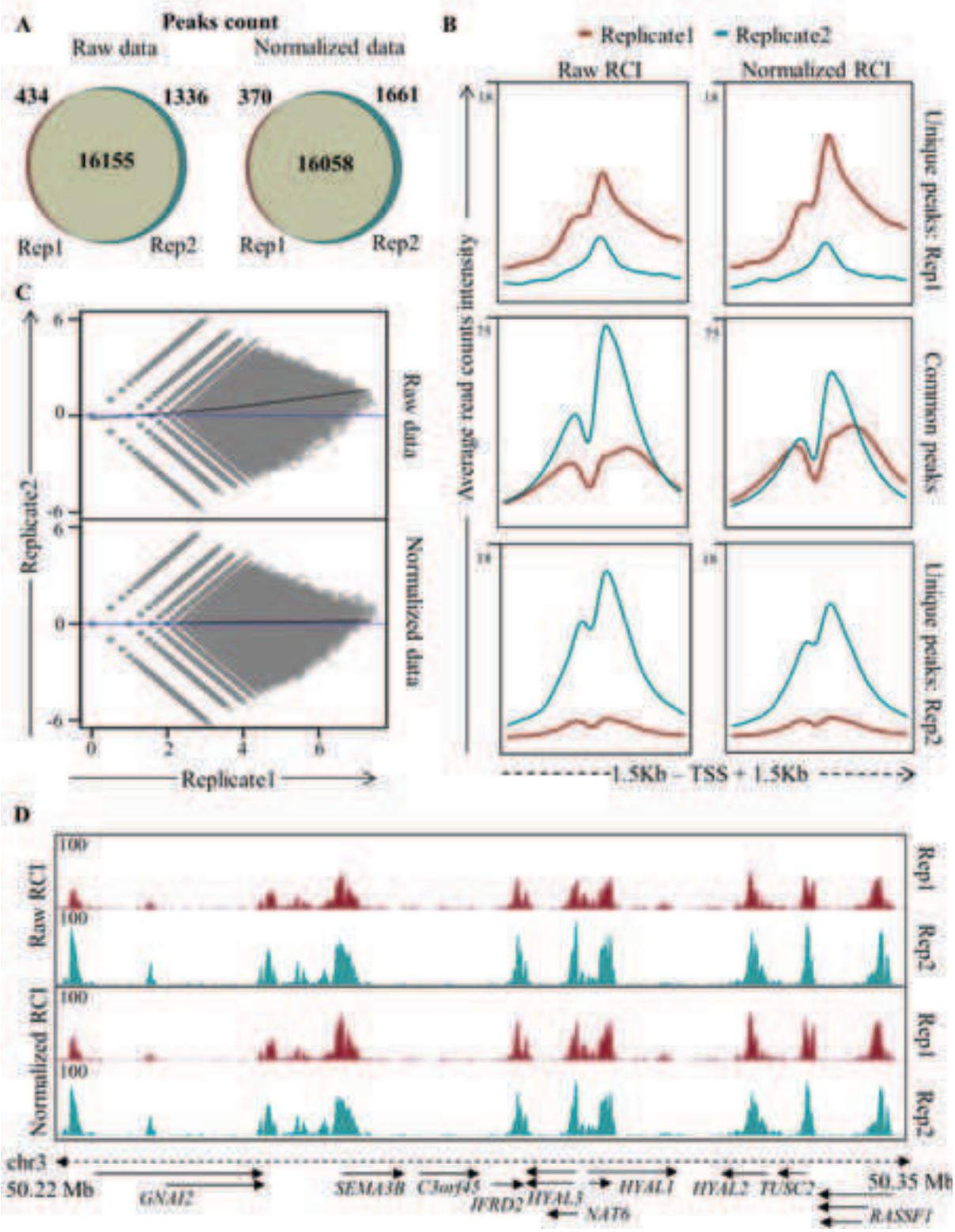

A
Peaks count
Raw data
Normalized data
434
1336
16155
Rep1
Rep2
370
1661
16058
Rep1
Rep2
B
Replicate1
Replicate2
Raw RCI
Normalized RCI
Unique peaks: Rep1
Common peaks
Unique peaks: Rep2
Average read counts intensity
-1.5Kb
TSS
+1.5Kb
C
Raw data
Normalized data
Replicate2
Replicate1
D
Raw RCI
Normalized RCI
Rep1
Rep2
chr3
50.22 Mb
50.35 Mb
GNAI2
SEMA3B
C3orf45
IFRD2
HYAL3
NAT6
HYAL1
HYAL2
TUSC2
RASSF1

Figure 1

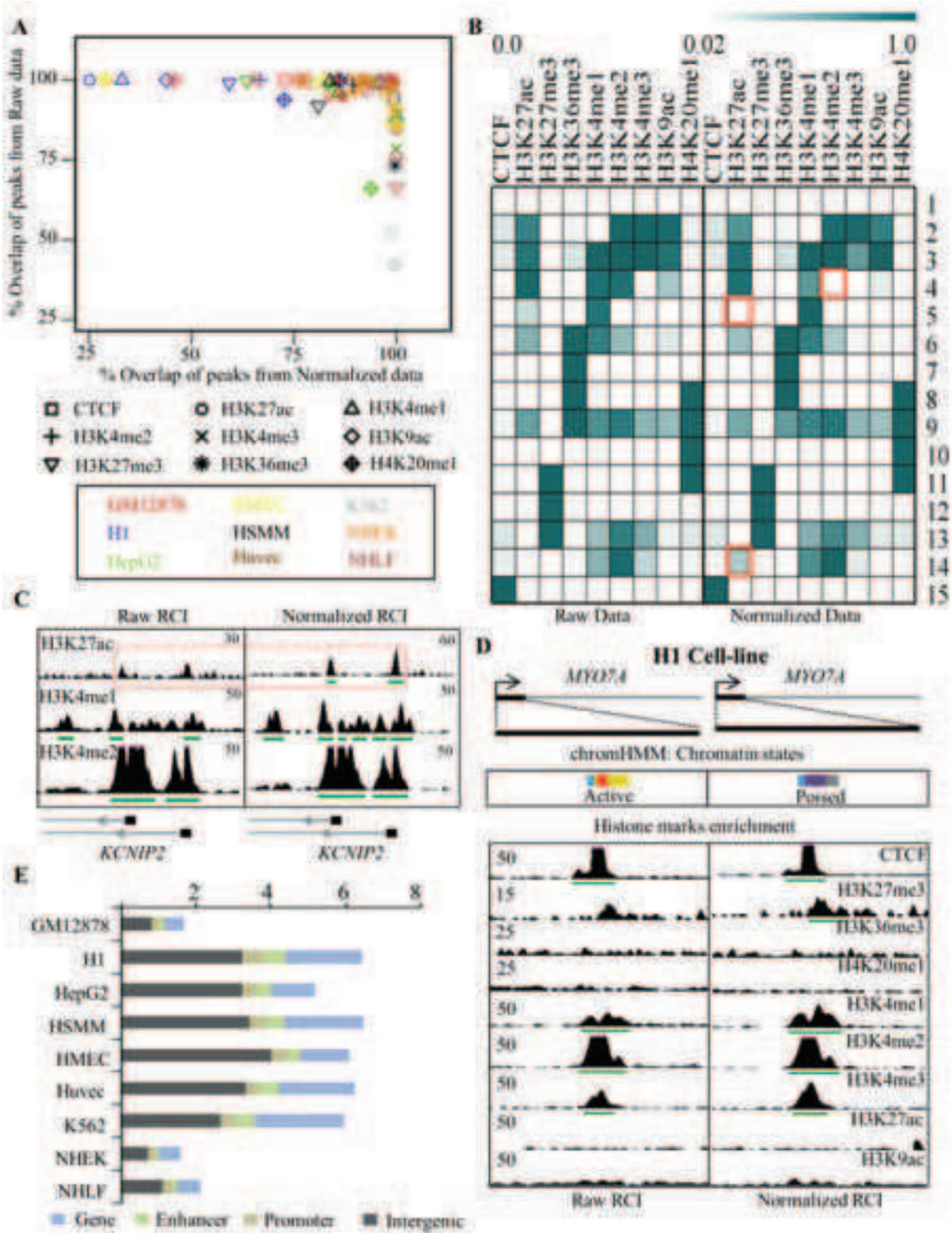

A
% Overlap of peaks from Raw data
% Overlap of peaks from Normalized data
CTCF
H3K27ac
H3K4me1
H3K4me2
H3K4me3
H3K27me3
H3K36me3
H4K20me1
H1
HSMM
B
Raw Data
Normalized Data
C
Raw RCI
Normalized RCI
KCNIP2
D
H1 Cell-line
MYO7A
chromHMM: Chromatin states
Active
Poised
Histone marks enrichment
Raw RCI
Normalized RCI
E
GM12878
HepG2
HSMM
HMEC
Huvec
K562
NHEK
NHLF
Gene
Enhancer
Promoter
Intergenic

Figure 2

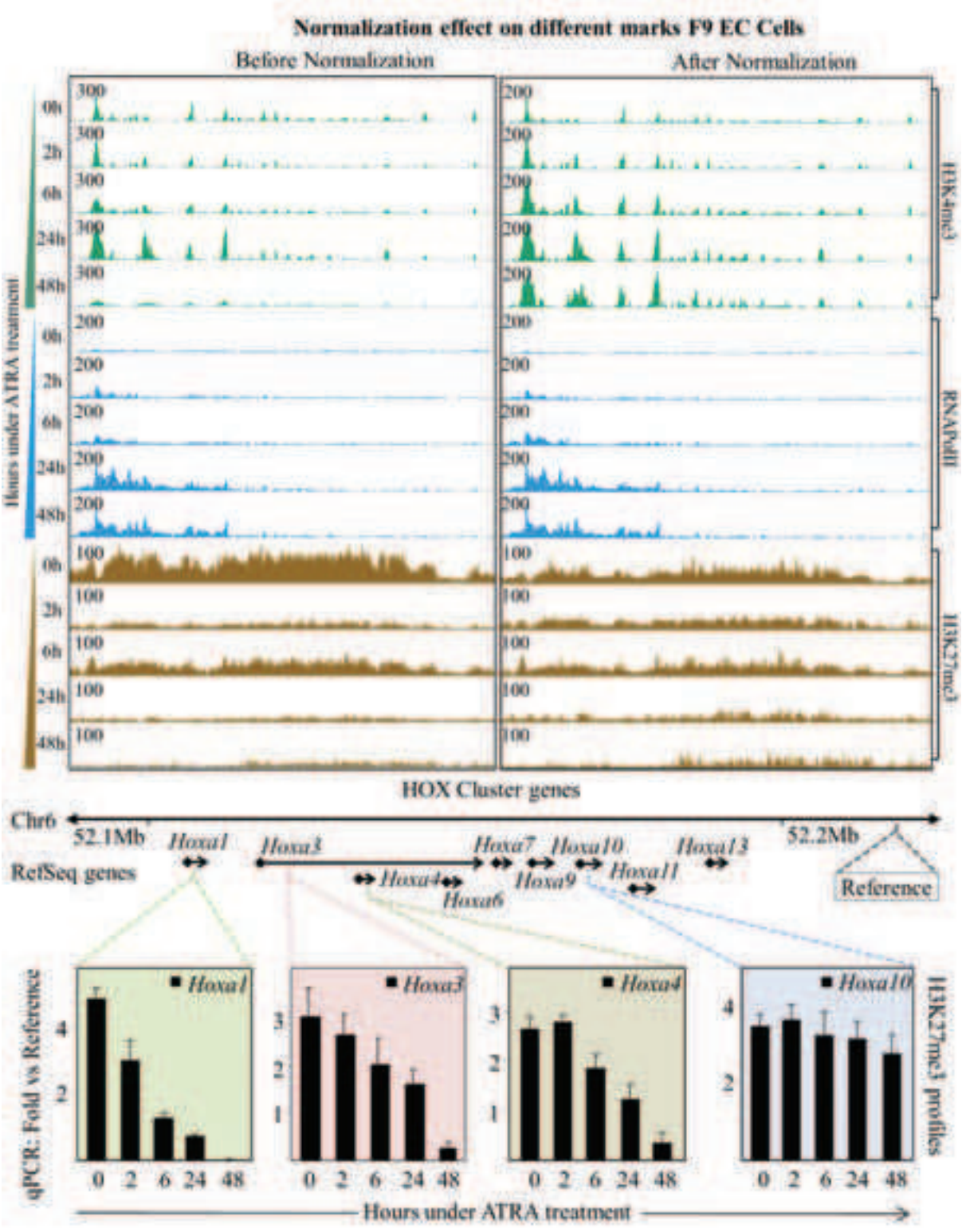
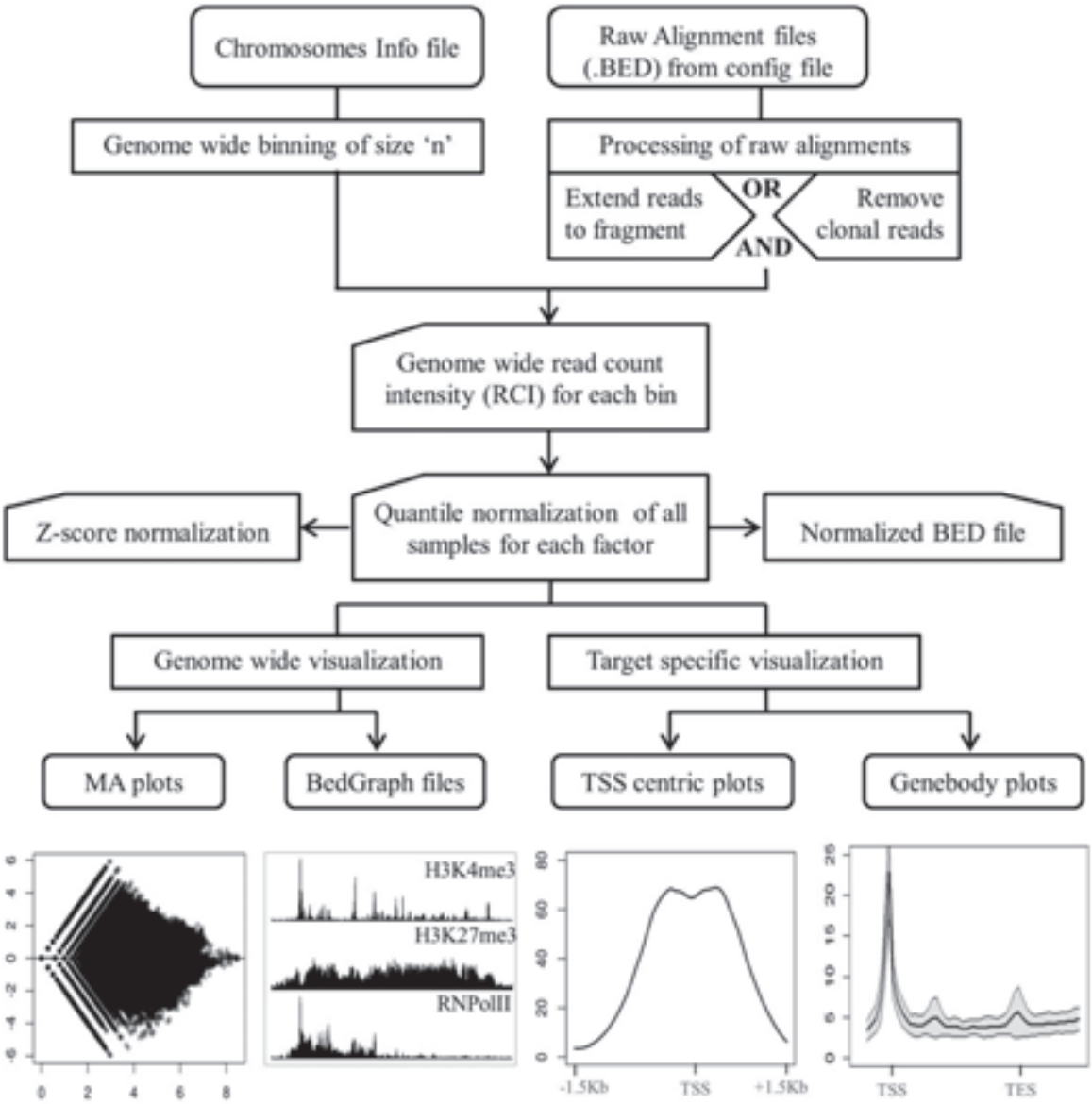

Normalization effect on different marks F9 EC Cells
Before Normalization
After Normalization
Hours under ATRA treatment
0h
2h
6h
24h
48h
H3K4me3
RNAPolII
H3K27me3
HOX Cluster genes
Chr6
52.1Mb
52.2Mb
RefSeq genes
Hoxa1
Hoxa3
Hoxa4
Hoxa6
Hoxa7
Hoxa9
Hoxa10
Hoxa11
Hoxa13
Reference
qPCR: Fold vs Reference
H3K27me3 profiles
Hours under ATRA treatment

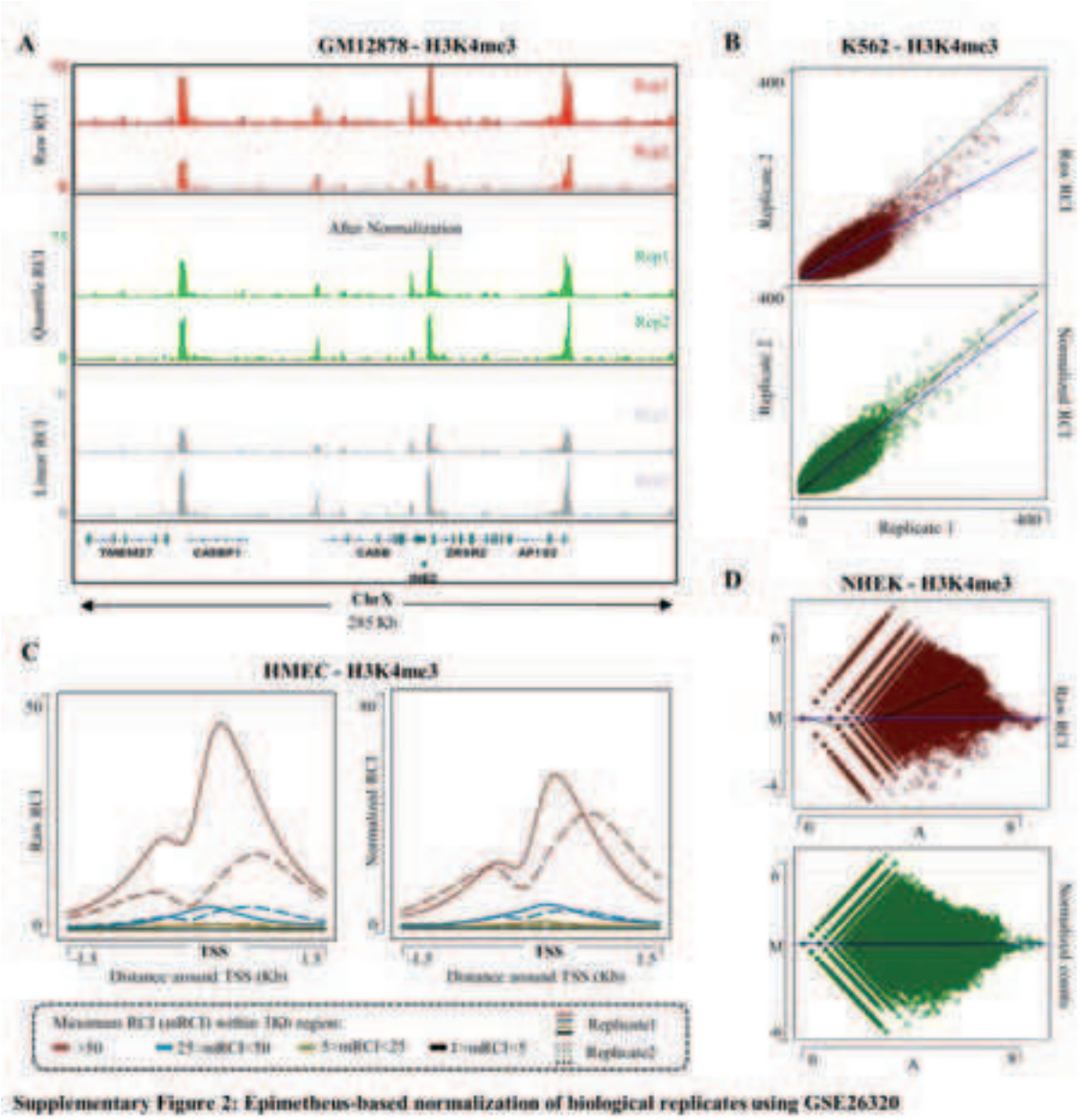
Figure 3

Supplementary

Supplementary Figure 1	A scheme of the workflow of Epimetheus with illustrative plots.
Supplementary Figure 2	Epimetheus-based normalization of biological replicates using GSE26320
Supplementary Note	A detailed summary on methodology of Epimetheus, datasets used and steps followed for different comparisons



Supplementary Figure 1. A scheme of the workflow of Epimetheus with illustrative plots.



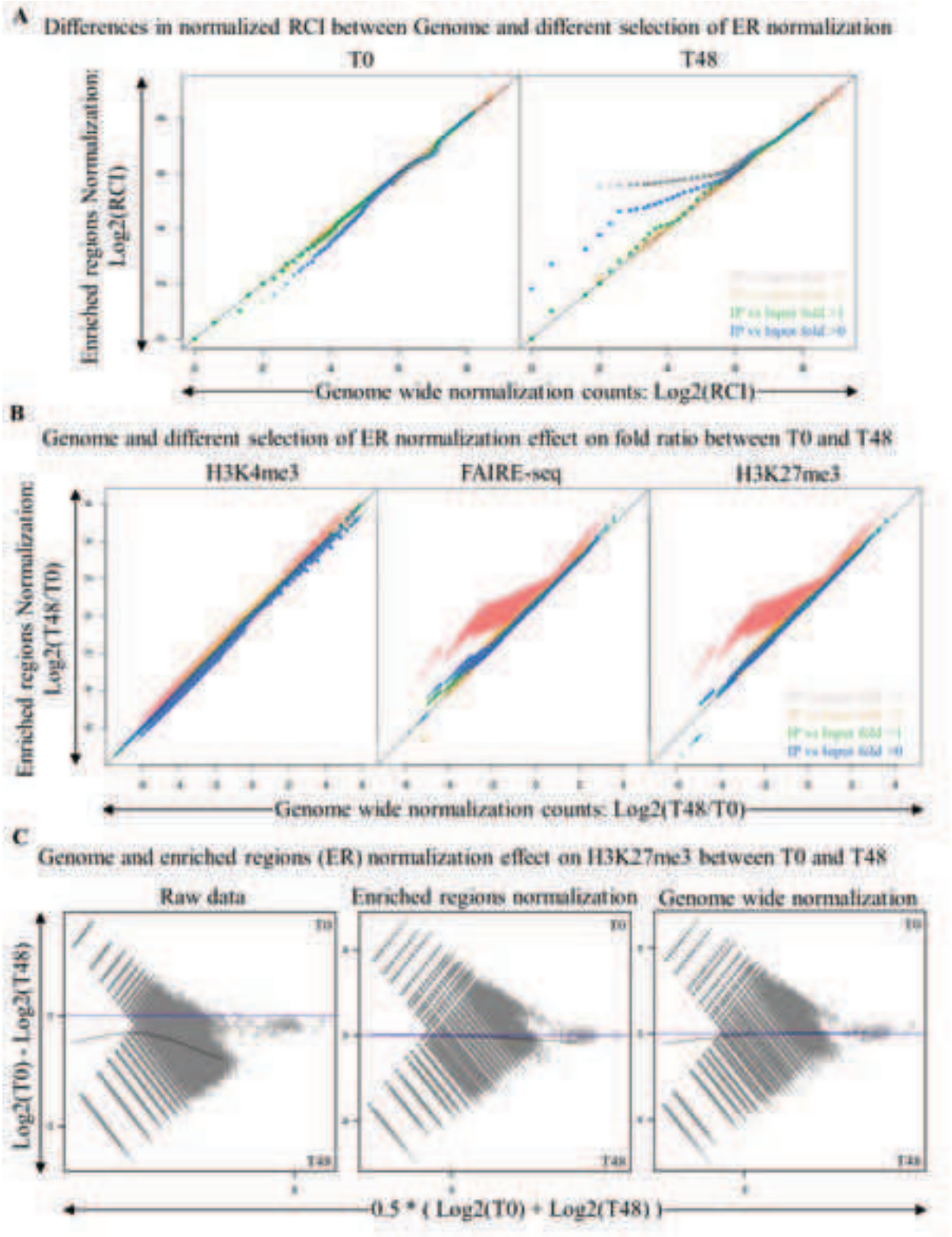
Supplementary Figure 2: Epimetheus-based normalization of biological replicates using GSE26320

Supplementary Figure 2. Epimetheus-based normalization of biological replicates using GSE26320. (**A**) Signal intensity display of H3K4me3 profiles illustrating the effect of quantile normalization compared to linear normalization between replicates exhibiting different background-to-signal enrichment levels. (**B**) Scatter plot of all raw RCIs (red) *versus* normalized RCI (green). (**C**) Comparison of TSS plots of replicates before and after normalization stratified into four intensity levels; color code is given below plot. Note the normalization effect on high intensity (red) enrichments. (**D**) MA plot of raw *versus* normalized RCIs.

Comparison with ChIPnorm-like approach

To illustrate the preference of choosing genome-wide normalization over target specific normalization, we compared both methods using Epimetheus for genome-wide and ChIPnorm(5) like approach for target specific approaches (methodology explained in Supplementary Note). We compared both the approaches on datasets with different enrichment patterns like H3K4me3, H3K27me3 and FAIRE-seq.

ChIPnorm uses input vs IP fold change >1 as a criterion to identify enrichment sites. To verify the effects on extreme cases, we considered fold change of input vs IP difference range from 0 to 4. We observed significant difference from genome-wide approach and also within different fold change datasets in normalization results. Specifically, when fold change criteria of input vs IP is gradually increased for identifying enriched regions, it resulted in gradual increase of discrepancy in differential enrichment (fold change) between samples (Supplementary Figure 3B). Though the more discrepancy is observed on very stringent fold change criteria, it is also evident that it depends on enrichment pattern and its population similarity/diversity between samples. As it is illustrated in Supplementary Figure 3B, sharp enrichment H3K4me3 is relatively less affected than highly diverse FAIRE-seq data and broad enrichment like H3K27me3 where signal intensities are less pronounced.



Supplementary Figure 3. Comparison of genome-wide and enriched regions (ER) only normalization using GSE68291 (unpublished). (**A**) An illustration of change in normalization results with respect to difference in selection of enriched regions on comparison of individual sample's normalization result from genome-wide and ER only normalization approach. Different fold change criteria (Input vs IP) is used to identify enriched regions while its corresponding bins are used from genome-wide

normalization approach (**B**) Effect of selection in target specific normalization in differential analysis; fold change comparison between T0 and T48 sample from H3K4me3, FAIRE-seq and H3K27me3 illustrates different selection in identifying enriched regions has bias differential fold changes between samples. (**C**) MA plot illustration to display the normalization effect on enriched regions (IP vs input fold >1) where LOWESS fit is better in genome-wide normalization.

Supplementary Note

Datasets

Datasets for the comparative and validation analysis on nine cell lines (GSE26320(1)) were downloaded from GEO. Data for F9 cell line data (GSE68291 - unpublished) for temporal analysis was generated in-house.

Processing and alignment of NGS data

For chromatin state analysis on nine cell lines, aligned BED files were directly downloaded and used for the analysis. For F9 cell line data, reads were aligned against mm9 genome using Bowtie (v 1.1.1)(2). Clonal reads were removed before analysis and reads were elongated to 200bp for both the analyses.

Peak Calling

For peak calling, MACS(3) was used with 1e-9 p-value and no-model option. SICER(4) was used to identify broad histone marks islands (H3K27me3, H3K36me3 and H4K20me1) in the nine cell line chromatin state analysis.

ChromHMM

We performed ChromHMM with peaks BED co-ordinate as input. Peaks from both raw and normalized BED files were provided as input separately and ChromHMM was performed with 400 iterations to predict 15 states and annotated with custom scripts for annotation.

Identification of enriched regions (ER) for ER specific normalization

As ChIPnorm is written in MATLAB (commercial software), we couldn't verify it in first hand; instead we wrote scripts following the outline of ChIPnorm workflow. Samples from F9 cell line data were considered for the comparison where T0 and T48 samples are used from three different marks H3K4me3, H3K27me3 and FAIRE-seq. Three main steps in this approach is 1) exclusion of background regions 2) Input and IP are normalized together using quantile and 3) identifying enriched regions based on fold change of Input vs IP. First steps are similar as in Epimetheus and Xu et al where genome is binned into small windows and read counts intensity (RCI) matrix is built for each sample. We then used Poisson distribution to identify number of reads that can be randomly filled in

bins by using total number of reads, effective genome size (with P-value of 0.995). It is followed by applying quantile normalization between input and IP to bring them to same scale to perform fold change analysis to identify enriched bins. In general, fold change >1 is used to identify ER whereas we altered this criterion to different ranges to see the influence of population selection in quantile normalization. We selected fold change greater than 0, 1, 2, 3 and 4. Fold change 0 would include bins in IP which has even one read count more than input where other fold changes consider enrichment based on the ratio. To compare geome-wide and ER only normalization, for each fold change normalization data we considered only its corresponding bins from genome-wide not the whole genome bins. Also, to compare the effect of population selection on differential analysis result between samples, we selected input vs IP fold change >3 range bins as the reference as fold change >4 has very few bins.

Plots

All the plots were generated using custom R scripts; ChromHMM chromatin states heatmap was generated using MeV (Multiple Experiment Viewer) suite and intensity profiles display was generated using UCSC genome browser(6) and IGV(7).

qPCR analysis for F9 data

Details of oligos used for the qPCR validation of the data on *Hoxa* cluster region confirming the normalized results.

Oligo Name	Sequence 5' to 3'	Scale (µmole)	Purification
HoxA_Rctrl_F	GCTGCAGGGGATAAACACAT	0.05	DST
HoxA_Rctrl_R	GCTGGAACATTAAGGCCAAA	0.05	DST
HoxA10_F	ATGAGCGAGTCGACCAAAAA	0.05	DST
HoxA10_R	ATGTCAGCCAGAAAGGGCTA	0.05	DST
HoxaA4_F	TCCTCGAAAGGAGGGAACTT	0.05	DST
HoxaA4_R	CGACACCGCGAGAAAAATTA	0.05	DST
HoxA3_F	GTCTGGAGTTGGGGGATTTT	0.05	DST
HoxA3_R	ACCTAGCCTCCAGACCCTGT	0.05	DST

Supplementary References:

1. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M., *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–9.

2. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

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

4. Zang,C., Schones,D.E., Zeng,C., Cui,K., Zhao,K. and Peng,W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.

5. Nair,N.U., Das Sahu,A., Bucher,P. and Moret,B.M.E. (2012) Chipnorm: A statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. *PLoS One*, **7**, e39573.

6. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The Human Genome Browser at UCSC. 10.1101/gr.229102.

7. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–92.

5.3. The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer

The tools I have developed during my thesis - NGS QC, the QC indicator database and Epimetheus - both enable and simplify the comparative analysis of large numbers of sequencing datasets. These tools were essential in a study, where we addressed an important biological question, namely the aberrations of X chromosome inactivation in breast cancer cells.

X chromosome inactivation is best-studied example of chromosome-wide epigenetic regulation, which involves the silencing of approximately one thousand genes during early embryonic development. The disappearance of the Barr body, the microscopically visible manifestation of the inactivated X chromosome, is considered a hallmark of breast cancer, although it has remained unclear whether this phenomenon corresponds to genetic loss or to epigenetic instability and transcriptional reactivation. X chromosome-wide allele-specific analysis could reveal the genes that are escaping inactivation, and their chromatin status, especially in breast cancer cells. In a collaborative study between the teams of Hinrich Gronemeyer and those of Edith Heard, Marc-Henri Stern and Anne Vincent-Salomon of the Curie Institute, we examined the epigenetic status of inactive X chromosome in normal (HMEC) and breast cancer (ZR-75-1, SK-BR-3 and MDA-MB-436) cells. The main focus of the study was on the integrated analysis of gene expression, chromatin status and nuclear organization of the inactivate X chromosome in breast cancer, using allele-specific and single-cell approaches. My contribution to the study was in the specialized bioinformatic aspects of allele-specific chromatin status and integrated gene expression analyses. Allele-specific regulation and expression analysis is a challenging integrative effect, as three different datasets intersect - Exome-seq/SNP6, RNA-seq and ChIP-seq. Below I will briefly summarize the methodology that has been used to identify allelic and epigenetic status of X-linked genes and X chromosome respectively.

5.3.1. Allele-specific expression and chromatin state analysis of X chromosome

Cell line specific heterozygous variations are the key factors in the identification of allele-specific expression or regulation. A heterozygous locus contains two different alleles, thus

two different bases will be observed. These heterozygous loci can help us to understand the transcriptomic and epigenomic status of each allele. In order to identify cell line-specific heterozygous variations, we have used SNP6 Affymetrix microarrays, which cover 906,600 SNPs and additional 424,000 SNPs in sex and mitochondrial chromosomes. First, to identify the cell line-specific heterozygous variations, SNP6 experiment was performed using genomic DNA (termed hereafter "gDNA SNP6"). Second, to identify the allelic status in RNA expression, SNP6 experiment was performed using nascent RNA (termed hereafter "cDNA SNP6"). Using gDNA and cDNA SNP6 information, allelic expression score was calculated for each SNP. Based on allelic expression score, each SNP was classified into five categories (i) bi-allelic expression, (ii) mono-allelic expression, (iii) marginal call (in-between mono and bi-allelic expression), (iv) contradictory call (disagreement between gDNA and cDNA SNP6 data), and (v) no expression or non-informative locus. Such SNP classification was summarized at the gene level to provide gene-based allelic status (refer Supplementary material of attached manuscript for the detailed discussion).

In order to identify epigenetic status of chromosome X, and allelic status of genes for which SNP6 has less SNP coverage, we performed transcriptome and epigenome sequencing. For epigenetic profiling, we performed ChIP-seq targeting histone modifications H3K4me3 (active) and H3K27me3 (repressive). Further, to trace the transcriptional activity, we performed ChIP-seq targeting RNA PolII and we performed mRNA sequencing for transcriptome profiling. As SNP6 can identify the known SNPs only and that it is mostly focussed on exonic regions, we performed variation calling on three different types of datasets - exome, ChIP-seq and transcriptome data - to collect more informative variations that are wide spread in the chromosome including regulatory regions. For ChIP-seq, to increase the coverage and confidence of variation calling, we merged different ChIP-seq datasets (input + all IPs) for each sample. Combining variations from each type of datasets provided a comprehensive collection of variations to verify the allelic status of epigenome and transcriptome. With the help of these comprehensive variations, we calculated allelic expression ratio (refer glossary) for each X-linked genes and analysed chromatin status of cancer-specific escapees (refer Figure 21 for the workflow scheme). We then calculated allelic imbalance based on number of heterozygous

variations and their read count share between alleles for both epigenome and transcriptome data. Genes that have bi-allelic expression ratio in SNP6, and RNA-seq analysis are identified as ‘escapees’. A comparison of normal and breast cancer cells revealed list of genes which are specific to each cancer cell-line, termed as cancer-specific ‘escapees’.

In general, perturbation (divergence from normal cells) of H3K27me3 was observed in Xi in breast cancer cells (Figure 27A). Other than overall perturbation in H3K27me3 in breast cancer cells, an abnormal presence of RNA PolII and H3K4me3 was observed at cancer-specific escapees, thus complementing the transcriptome results (Figure 27B). Most of the cancer-specific escapees displayed bivalent chromatin which was characterized by both active (H3K4me3 and RNA PolII) and repressive (H3K27me3) marks. Several cancer-specific escapees identified haven previously been shown to be involved in cancer, such as HDAC8 which is shown to be involved in neuroblastoma pathogenesis (Oehme et al., 2009). Similarly, several known normal escapees from XCI such as RAB9A, BCOR, RPL39 and PNPLA4 were repressed in cancer cells due to aberrant epimutations. BCOR gene has been shown to have recurrent mutations that resulted in truncation of encoded proteins in retinoblastoma (Zhang et al. 2012).

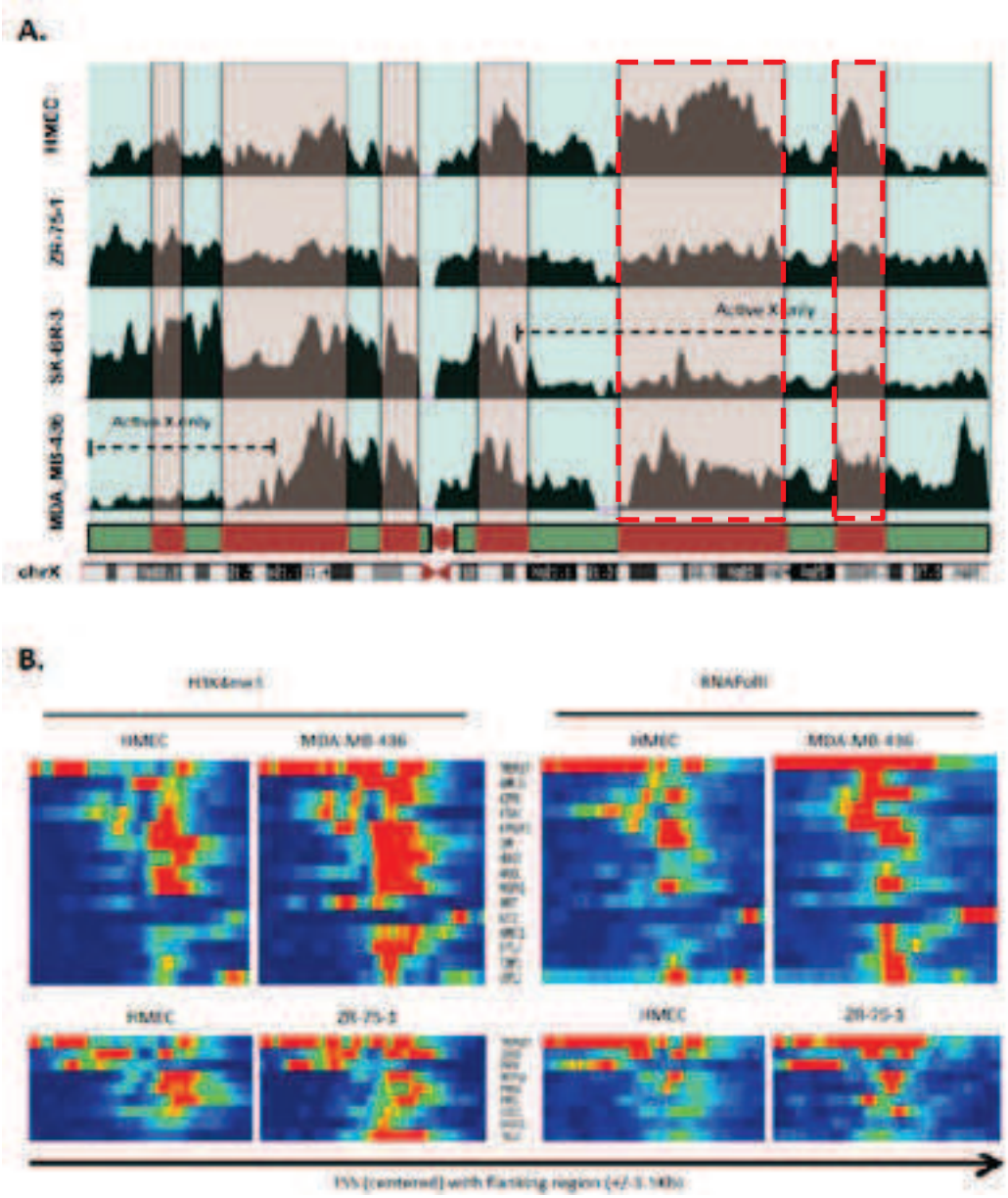


Figure 27. Allele specific analysis led to the identification of genes escaping XCI specific to cancer cell-lines. (A) A scheme of H3K27me3 enrichment across the entire X chromosome shows a regional loss of inactive X. Regional loss of inactive X is highlighted 'active X only' and the two main H3K27me3 enrichment loss in ZR-75-1 and MDA-MB-436 is highlighted in red box. Red and green domains represent H3K27me3 and H3K9me3 enriched regions, respectively, as identified in normal human cells (Chadwick, 2007) (B) Increased abundance of H3K4me3 marks and RNAPolII recruitment are displayed as heatmap plots, as escapee genes are active in both alleles unlike what is observed in a normal cell-line (HMEC).

5.3.2. Discussion

X chromosome inactivation study is a paradigm for epigenetics study experimentally and very challenging bioinformatically. With the currently available technologies, heterozygous variations are the only means to differentiate mono and bi-allelic gene expression globally. Hence, the whole analysis is restricted to the genes with handful number of informative heterozygous variations, where even homozygous variations are not useful. Even more so both gDNA and cDNA SNP6 technologies contain SNP regions widespread with mean spacing between probes around 3Kb, concentrated mostly on coding regions. To collect more informative variations, we have used ChIP-seq data to call variations by merging all different marks as they come from same cell-line. Epimetheus have been used to normalize the profiles among samples to identify the difference in enrichment overall in H3K27me3 and for comparing normal and cancer specific escapees.

In general, the study concluded that a frequent cause of Barr body disappearance is due to the global perturbation of its nuclear organization and disruption of its heterochromatic structure. Though the enrichment level of H3K27me3 is lower and perturbed in Xi of breast cancer cells, it is relatively higher than average enrichment over rest of the genome. Several aberrantly reactivated genes identified have been associated with cancer previously. Hence, such aberrant reactivation of X-linked genes in Xi might contribute to a selective advantage of cancer cells. In conclusion, the perturbed transcriptional and chromatin status of the inactive X chromosome that we have identified in the context of breast cancer, opens up several important clinical perspectives.

Similarly, imprinted genes have also been associated with breast cancer. When dosage disequilibrium of imprinted genes occurs due to aberrant genetic or epigenetic mutation, it can lead to severe disorders including cancer. While the XCI study focused only on X chromosome, a genome wide allele-specific analysis could reveal any aberrant changes of imprinted genes in breast cancer cells. A detailed collection of imprinted genes have been made available would be helpful to contain the analysis to these imprinted genes. Our initial genome-wide analysis has shown aberrant b-allelic expression and epigenome status on few imprinted genes. Among them, three genes namely *ZFP36L2*, *CYP1B1* and *TIGD1* have been observed in two of the cancer cell-lines. CYP1B1 has been shown to have

important role in tumor development as they can metabolize many potential carcinogens and mutagens. It has been shown that CY1PB1 mRNA is the most frequently expressed in breast cancer (Murray et al., 1997). Further investigation with support from experimental data could give more insights on those aberrant imprinted genes and their association with breast cancer.

MANUSCRIPT 3

THE INACTIVE X CHROMOSOME IS EPIGENETICALLY UNSTABLE AND TRANSCRIPTIONALLY LABILE IN BREAST CANCER

Research

The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer

Ronan Chaligné,[1,2,3,4] Tatiana Popova,[1,5] Marco-Antonio Mendoza-Parra,[6] Mohamed-Ashick M. Saleem,[6] David Gentien,[1,7] Kristen Ban,[1,2,3,4] Tristan Piolot,[1,8] Olivier Leroy,[1,8] Odette Mariani,[7] Hinrich Gronemeyer,[6] Anne Vincent-Salomon,[1,4,5,7] Marc-Henri Stern,[1,5,7] and Edith Heard[1,2,3,4]

[1]Centre de Recherche, Institut Curie, 75248 Paris Cedex 05, France; [2]Centre National de la Recherche Scientifique, Unité Mixte de Recherche 3215, Institut Curie, 75248 Paris Cedex 05, France; [3]Institut National de la Santé et de la Recherche Médicale U934, Institut Curie, 75248 Paris Cedex 05, France; [4]Equipe Labellisée Ligue Contre le Cancer, UMR3215, 75248 Paris Cedex 05, France; [5]Institut National de la Santé et de la Recherche Médicale U830, Institut Curie, 75248 Paris Cedex 05, France; [6]Institut de Génétique et de Biologie Moléculaire et Cellulaire, Equipe Labellisée Ligue Contre le Cancer, Centre National de la Recherche Scientifique UMR 7104, Institut National de la Santé et de la Recherche Médicale U964, University of Strasbourg, 67404 Illkirch Cedex, France; [7]Department of Tumor Biology, Institut Curie, 75248 Paris Cedex 05, France; [8]Plate-forme d'Imagerie Cellulaire et Tissulaire at BDD (Pict@BDD), Institut Curie, 75248 Paris Cedex 05, France

Disappearance of the Barr body is considered a hallmark of cancer, although whether this corresponds to genetic loss or to epigenetic instability and transcriptional reactivation is unclear. Here we show that breast tumors and cell lines frequently display major epigenetic instability of the inactive X chromosome, with highly abnormal 3D nuclear organization and global perturbations of heterochromatin, including gain of euchromatic marks and aberrant distributions of repressive marks such as H3K27me3 and promoter DNA methylation. Genome-wide profiling of chromatin and transcription reveal modified epigenomic landscapes in cancer cells and a significant degree of aberrant gene activity from the inactive X chromosome, including several genes involved in cancer promotion. We demonstrate that many of these genes are aberrantly reactivated in primary breast tumors, and we further demonstrate that epigenetic instability of the inactive X can lead to perturbed dosage of X-linked factors. Taken together, our study provides the first integrated analysis of the inactive X chromosome in the context of breast cancer and establishes that epigenetic erosion of the inactive X can lead to the disappearance of the Barr body in breast cancer cells. This work offers new insights and opens up the possibility of exploiting the inactive X chromosome as an epigenetic biomarker at the molecular and cytological levels in cancer.

[Supplemental material is available for this article.]

There is increasing evidence that epigenetic modifications, such as changes in DNA methylation, chromatin structure, noncoding RNAs, and nuclear organization, accompany tumorigenesis (De Carvalho et al. 2012; for review, see Shen and Laird 2013). Even tumors with relatively normal karyotypes can show dramatically perturbed nuclear structures (Huang et al. 1997; for review, see Zink et al. 2004). In theory, epigenetic changes could lead to inactivation of tumor suppressor genes, aberrant expression or function of oncogenes, or more global gene expression changes that perturb genome function, thereby contributing to cancer progression. However, despite the possible use of epigenetic changes as prognostic markers (Elsheikh et al. 2009) or even as therapeutic targets (e.g., Schenk et al. 2012; Zhang et al. 2012), the full extent of epigenetic changes in cancer remains poorly explored.

The inactive X chromosome (Xi), also known as the Barr body, provides an outstanding example of an epigenetic nuclear landmark that is disrupted in cancer. The disappearance of the Barr body in breast tumors was noted many decades ago (Barr and Moore 1957; Perry 1972; Smethurst et al. 1981). To date, only genetic instability had been clearly demonstrated as a cause for Barr body loss (Ganesan et al. 2002; Sirchia et al. 2005; Vincent-Salomon et al. 2007; Xiao et al. 2007; and for review, see Pageau et al. 2007). Past work had implicated *BRCA1*, a major hereditary factor predisposing to breast and ovarian cancer development and a key player in the maintenance of genome integrity (for review, see O'Donovan and Livingston 2010), in promoting *XIST* RNA coating of the Xi and its epigenetic stability (Ganesan et al. 2002; Silver et al. 2007). However, subsequent work in *BRCA1*-deficient tumors indicated that Barr body loss was usually due to genetic loss of the Xi and duplication of the Xa rather than to Xi reactivation and epigenetic instability (Sirchia et al. 2005; Vincent-Salomon et al. 2007; Xiao et al. 2007). *BRCA1*-deficient cancers are usually of the basal-like carcinoma (BLC) subtype, a high-grade, genetically unstable, invasive ductal carcinoma. Indeed, when the genetic status of the X chromosome was explored in BLCs (Richardson et al. 2006), genetic instability/loss of the Xi was found to be a frequent event in both sporadic and $BRCA1^{-/-}$ associated BLCs. Luminal (A and B, expressing

Corresponding authors: Edith.Heard@curie.fr, Anne.Salomon@curie.fr, marc-henri.stern@curie.fr, hg@igbmc.fr
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.185926.114. Freely available online through the *Genome Research* Open Access option.

25:488–503; Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/15; www.genome.org

hormonal receptors) and HER2 (encoded by *ERBB2*) amplified molecular subtypes of invasive ductal carcinoma are more genetically stable and show less frequent loss of the inactive X chromosome (Perou et al. 2000; Turner and Reis-Filho 2006). However, little is known about the epigenetic status of the inactive X in breast cancers and the extent to which epigenetic instability might account for Barr body disappearance in some cases.

X-chromosome inactivation (XCI) ensures dosage compensation for X-linked gene products between XX females and XY males (Lyon 1961). It is a developmentally regulated process that depends on the action of a noncoding RNA, *Xist* (X-inactive-specific-transcript), which becomes up-regulated on one of the two X chromosomes, coating it in *cis* and inducing gene silencing. *Xist* RNA accumulation on the future inactive X rapidly creates a silent nuclear compartment that is depleted of RNA Polymerase II (RNA Pol II), transcription factors, and transcription (as detected by Cot-1 RNA). X-linked genes become repressed during the early stages of XCI (Chaumeil et al. 2006; Clemson et al. 2006; Chow et al. 2010). *Xist* RNA also induces a cascade of chromatin changes, involving Polycomb group proteins and other complexes, and results in various histone modifications, such as the hypoacetylation of histones 3 and 4, trimethylation of histone 3 lysine 27 (H3K27me3), and the loss of di- and trimethylation at histone 3 lysine 4 (H3K4me2/3) (Csankovszki et al. 1999; Heard et al. 2001; Boggs et al. 2002). Promoter DNA methylation of X-linked genes occurs downstream from *Xist* RNA coating, with gene-specific timing of promoter methylation (Gendrel et al. 2013). The Xi adopts a unique three-dimensional (3D) chromosome organization that is dependent on *Xist* RNA (Splinter et al. 2011; for review, see Chow and Heard 2010). Furthermore, the chromatin landscape of the inactive X has been investigated in adult human cells and seems to be divided into large blocks of H3K9me3 or H3K27me3 (Chadwick 2007; Chadwick and Willard 2004). In somatic cells, the majority of X-linked genes are stably repressed on the Xi, with spontaneous reactivation of single genes being observed at a frequency of $<10^{-8}$, presumably due to synergistic epigenetic mechanisms (Csankovszki et al. 2001). However, a subset of genes can escape XCI in somatic cells (Carrel and Willard 2005; Kucera et al. 2011; Cotton et al. 2013). In cancer, aberrant escape from XCI has previously been speculated to occur (Pageau et al. 2007; Agrelo and Wutz 2010; Carone and Lawrence 2013; Yildirim et al. 2013). However, the extent to which the normally stable epigenetic state of the Xi is perturbed in cancer has never been systematically explored.

The X chromosome is of interest from a cancer perspective. First, several of the approximately 1000 genes located on the X have been implicated in cancer, including the cancer/testis (C/T) genes (Grigoriadis et al. 2009); tumor suppressors such as *AMER1* (also known as *WTX*), *FOXP3* (Bennett et al. 2001; Rivera et al. 2007); chromatin remodelers related to disease, e.g., *ATRX*; or chromatin modifying factors, e.g., *KDM6A* (also known as *UTX*), *PHF8, HDAC8* (Nakagawa et al. 2007; for reviews, see Agrelo and Wutz 2010; Portela and Esteller 2010). A few of these genes are known to escape X inactivation in normal cells (e.g., *KDM6A*), but most are normally stably repressed on the inactive X. In the cases of *AMER1* and *FOXP3*, tumorigenesis has been linked to clonal expansion of cells in which the wild-type copy is on the inactive X in female patients heterozygous for a mutation (Bennett et al. 2001; Rivera et al. 2007).

Although reactivation of X-linked genes has been previously hypothesized to occur in a cancer context (Spatz et al. 2004), few actual examples have been reported, presumably due to the technical challenges in specifically detecting the Xi. For example, deletion of *Xist* was reported to lead to hematological dysplasia and leukemia in mice; however, the allele-specific transcriptional activity of the inactive X chromosome and its heterochromatin structure were not examined (Yildirim et al. 2013). In another study, reactivation of the X-linked *MPP1* gene and disrupted *XIST* expression were reported in an ovarian cancer cell line (Kawakami et al. 2004). In breast tumors, DNA hypomethylation and abnormal expression of a single X-linked gene analyzed, *VBP1,* was detected on the Xi (Richardson et al. 2006). A systematic analysis of the transcriptional and epigenetic status of the Xi in breast tumors has been lacking however. Here we perform an integrated analysis of gene expression, chromatin status, and nuclear organization of the inactive X chromosome in breast cancer, using allele-specific and single-cell approaches.

Results

Aberrant nuclear organization of the inactive X chromosome in breast cancer cells

To evaluate the status of the inactive X chromosome in different types of breast cancer, we selected three cell lines that represent the main breast cancer molecular subtypes: ZR-75-1 (luminal), SK-BR-3 (HER2+), and MDA-MB-436 (Basal-Like Carcinoma [BLC], *BRCA1* null). WI-38 (embryonic lung fibroblasts) and Human Mammary Epithelial Cells (HMECs) were analyzed in parallel as nonmalignant ("normal") female primary cells. Using RNA FISH, we found that ZR-75-1 and MDA-MB-436 cell lines possess one *XIST* RNA domain, whereas SK-BR-3 cells have two domains. X-chromosome paint DNA FISH combined with *XIST* RNA FISH, and 3D microscopy revealed that *XIST* RNA signals overlapped to a great extent with the X chromosome DNA in both normal and tumor cell lines. However, punctate *XIST* RNA signals beyond the X-chromosome territory could be detected in the tumor cell lines, particularly in ZR-75-1 and MDA-MB-436 (Fig. 1A; Supplemental Fig. S1A). RT-qPCR revealed that *XIST* was expressed at slightly lower levels in the tumor cell lines, and the associated RNA FISH signal was slightly weaker and was more dispersed in the breast cancer cell lines (Supplemental Fig. S1B,C,E). Importantly, all of the tumor cell lines revealed a markedly weaker DNA enrichment of the Barr body (Supplemental Fig. S1D,E).

Given the complex genomes of breast cancer cells, we investigated the precise genetic constitution of the active and inactive X chromosomes using single nucleotide polymorphism array (Human SNP Array 6.0) analysis and DNA FISH (Fig. 1B; Supplemental Fig. S1F). ZR-75-1 contains three X-chromosome segments, each carrying an XIC/*XIST* locus, but *XIST* RNA coated only one of them, suggesting the presence of two Xa chromosomes and one Xi (in agreement with allelic imbalance of the XIC locus). SK-BR-3 possesses four X-chromosome fragments, each with an XIC locus, but only two are associated with *XIST* RNA. MDA-MB-436 displayed the most complex situation, with six X-chromosome fragments visible by DNA FISH on metaphase spreads, but with only two XIC loci and one *XIST* RNA domain (Fig. 1A,B; Supplemental Fig. S1F). We also evaluated X-chromosome constitution in these cell lines through the expression of two X-linked genes: *KDM5C*, known to escape from XCI, and *HUWE1*, subject to XCI (Cotton et al. 2013). Our observations concur with the expected expression profiles in the two normal and three cancer cell lines, i.e., *KDM5C* is expressed from all X chromosome fragments that carried the gene, and *HUWE1* is

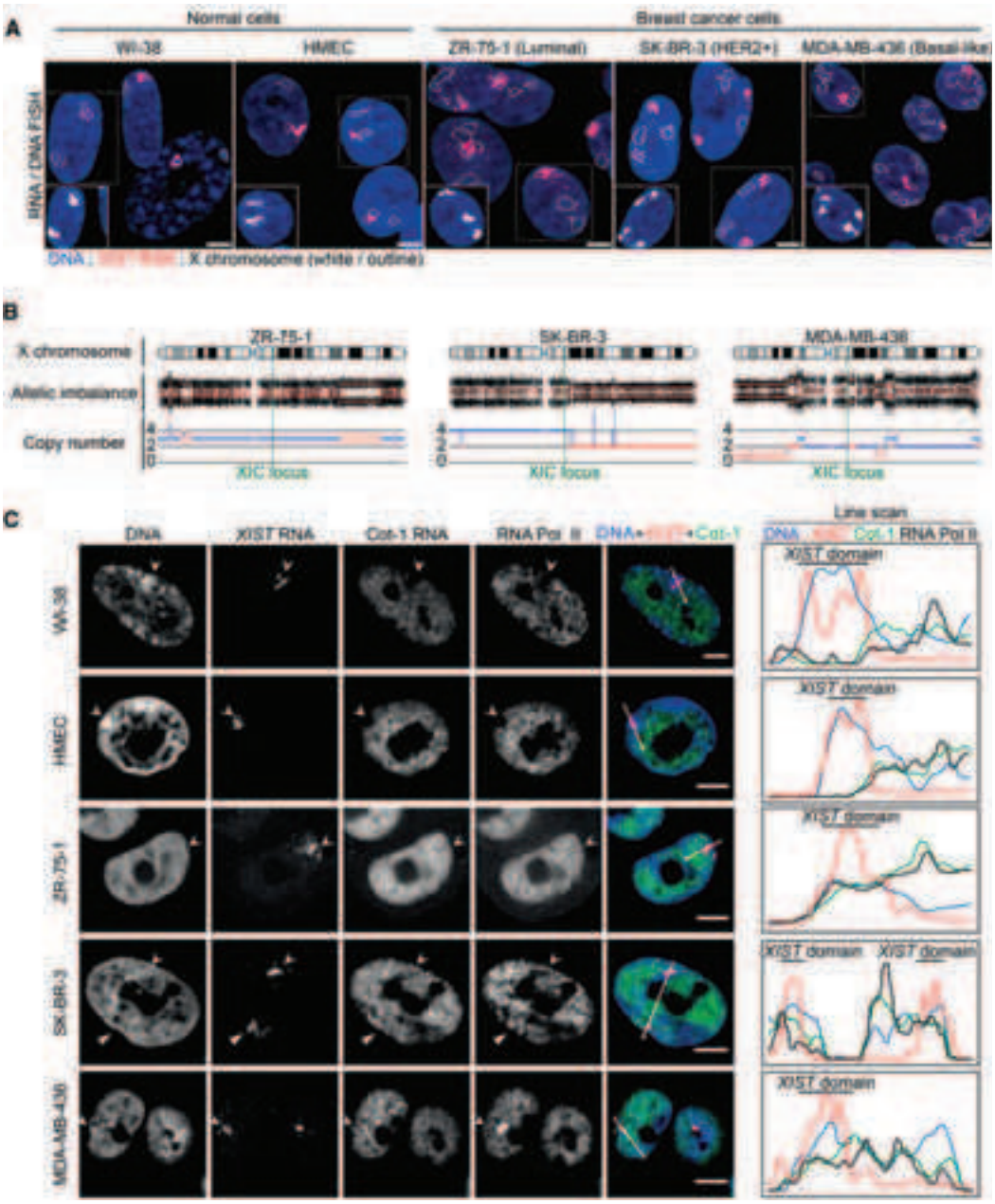


Figure 1. The *XIST*-coated X-chromosome silent compartment is severely disrupted in breast cancer cell lines. (*A*) Z-projections of sequential 3D RNA/DNA FISH show examples of *XIST* RNA coating (red) and X-chromosome territories (white or outlined) in normal (WI-38 and HMEC) and breast cancer cell lines (ZR-75-1, SK-BR-3, and MDA-MB-436). Scale bar, 5 µm. (*B*) Human SNP Array 6.0 (Affymetrix) genomic analysis (Popova et al. 2009) shows the copy number and allelic imbalance of X-chromosome fragments in breast cancer cell lines. The XIC locus is indicated with a green dotted line. (*C*) Immuno-RNA FISH using anti-RNA Pol II antibody, *XIST*/Cot-1 RNA FISH, and DAPI staining show the level of exclusion of RNA Pol II and Cot-1 RNA, as well as the level of chromatin compaction (i.e., Barr body) on *XIST* RNA domains (arrowheads) in normal and breast cancer cell lines. On the *right*, line scans (white arrows) show the relative levels of Cot-1 RNA (green), RNA Pol II (black), and DNA density (blue) at the *XIST* domain (black bar). Scale bar, 5 µm.

expressed only from the non-*XIST* RNA-coated X fragments that carried it (Supplemental Fig. S1G). Thus, all three tumor cell lines contain at least one fragment of an Xi chromosome.

We then investigated whether *XIST* RNA-coated Xi fragments were depleted for RNA Pol II and Cot-1 RNA as previously described for the Xi in female somatic cells (Chaumeil et al. 2006; Clemson et al. 2006; Chow et al. 2010). In WI-38 and HMEC cells, both Cot-1 RNA and RNA Pol II were excluded from the *XIST* domain, which was associated with a DAPI-dense, heterochromatic Barr body. However, all tumor cells showed a

frequent absence of a DAPI-dense Barr body and a defective depletion of Cot-1 RNA and RNA Pol II within the *XIST* domain (Fig. 1C; Supplemental Figs. S1H–K, S2A,C). Together, these results reveal major aberrations of nuclear organization and chromosome condensation of the *XIST* RNA-coated X chromosome in breast cancer cells.

Aberrant chromatin hallmarks of the inactive X chromosome in breast cancer cell lines

We next investigated whether heterochromatic hallmarks of the Xi were preserved. Detection of H3K27me3 by IF combined with *XIST* RNA FISH revealed a marked lack of H3K27me3 enrichment at the *XIST*-coated chromosome in all three tumor cell lines (Fig. 2A). In HMECs, H3K27me3 is about twofold more enriched on the Xi than on the non-*XIST*-coated genome (Fig. 2B; Supplemental Fig. S2A,B). In tumor cells, the lowest enrichment was found in ZR-75-1 and MDA-MB-436 with a median of 1.25 and 1.37-fold, respectively, whereas for SK-BR-3 it is 1.68 (Fig. 2A,B; Supplemental Fig. S2I). Decreased H3K27me3 enrichment at the *XIST* domain was further supported by super resolution structured illumination microscopy (SIM) (Fig. 2C). Indeed, ZR-75-1 and MDA-MB-436 showed the lowest degree of *XIST* and H3K27me3 colocalization with a Pearson colocalization coefficient of 0.15, whereas SK-BR-3 had a coefficient at 0.35. HMEC and WI-38 displayed colocalization coefficients of 0.44 and 0.45, respectively (Supplemental Fig. S2D).

Depletion of euchromatic histone modifications is another hallmark of the Xi. Using IF combined with *XIST* RNA FISH, we found that H3K9 and H4 acetylation were present within the *XIST* RNA domain in tumor cells in contrast to normal cells (Fig. 2D,E; Supplemental Fig. S2E,F,I). The H3K4me2 mark was less perturbed, being globally absent from the Xi, except in ZR-75-1 cells (Supplemental Fig. S2G–I). Similar results were obtained for H3K4me3 staining (with, for example, median at 0.69 and 0.71, respectively, for HMEC and MDA-MB-436) (data not shown). Closer examination by SIM revealed that H3K9ac and *XIST* RNA signals were intermingled in the majority of breast cancer nuclei (Fig. 2F), whereas H3K4me2 was largely but not completely depleted within the *XIST* RNA compartment (Supplemental Fig. S2J). SIM of RNA Pol II also revealed substantial intermingled overlap with *XIST* RNA domains (Supplemental Fig. S2K). Thus, there is a major disruption of chromatin hallmarks over the *XIST* RNA-coated chromosome, most strikingly in ZR-75-1 and MDA-MB-436 cell lines. We confirmed that *XIST* is always expressed from only one allele, excluding the possibility of aberrant *XIST* expression and coating of the Xa instead of the Xi (Supplemental Fig. S3A). In summary, the heterochromatic structure of the Xi is disrupted in the three tumor cell lines to variable extents. The variability in Xi perturbation between cells was not found to be linked to a specific stage of the cell cycle (Supplemental Fig. S3B–D). Furthermore the global levels of histone modifications in the different cell lines did not correlate with the aberrant chromatin status of the Xi (Supplemental Fig. S3E).

To specifically compare the chromatin status of the Xi and Xa in the tumor cell lines, we used metaphase spreads to monitor chromatin marks by IF followed by X-chromosome paint DNA FISH as described (Fig. 3; Keohane et al. 1996; Chaumeil et al. 2002). In all tumor cell lines, we could readily distinguish Xa from Xi fragments using H3K27me3, H4ac, and H3K4me2 (Fig. 3A–C). The only exception was MDA-MB-436, where from the two main Xi fragments, only the XIC-linked (and *XIST*-coated) fragment is enriched for H3K27me3 (Fig. 3C,D; Supplemental Fig. S3F), whereas the other (non-XIC-linked) X fragment lacked H3K27me3 enrichment, although it was still depleted for H4ac and H3K4me2. Thus, the XIC is required for H3K27me3 enrichment but is dispensable for depletion of euchromatin marks on the Xi in these cancer cells (Fig. 3A,B). We also noted from the analysis of metaphase spreads that in MDA-MB-436 and SK-BR-3 cells, where the Xi is translocated to an autosomal region, H3K27me3 enrichment was seen beyond the X chromosome paint signal, implying that it can spread aberrantly into autosomal regions (Fig. 3C).

Reactivation of X-linked genes on the inactive X chromosome in breast cancer cell lines

We next assessed whether the heterochromatic disruption of the Xi observed in breast tumor cell lines corresponded to aberrant abnormal transcriptional activity from the Xi. To take advantage of SNPs that lie within introns of genes, we used an allele-specific transcriptional analysis based on nascent RNA hybridization to Human SNP Array 6.0 (henceforth called RNA SNP6) (Fig. 4A,B; Supplemental Fig. S4A; Gimelbrant et al. 2007). Due to the randomness of the XCI, clonal populations of cells are required to investigate Xi status. This was the case for all three tumor cell lines and for subclones derived from primary WI-38 cells (Supplemental Fig. S4B–E). In both WI-38 clones and the tumor cell lines, we saw the expected overall biallelic expression from autosomal regions (Chromosome 2 is shown as an example in Fig. 4A; Supplemental Fig. S4F). On the other hand, the X chromosome showed a globally monoallelic expression pattern in WI-38 clones, with the exception of genes in the pseudoautosomal regions that are known to behave as autosomes and to escape fully from XCI (Fig. 4B; Supplemental Fig. S4G). In tumor cells, we observed a generally monoallelic expression pattern from the X chromosome, although several regions showed biallelic expression, particularly in MDA-MB-436 cells (Fig. 4B). A gene-based analysis detected several previously described X-linked escapees (including *DHRSX, TRAPPC2*, *CD99*, or *KDM6A*) (Carrel and Willard 2005; Kucera et al. 2011; Cotton et al. 2013), confirming the efficiency of this approach. We used known escapees and genes subject to XCI (Carrel and Willard 2005; Cotton et al. 2013) to define a threshold to consider that a given X-linked gene is expressed from inactive and active alleles. Thus, we defined "cancer-specific" escapees as genes reactivated in at least one of the three cancer cell lines, but strictly expressed from the Xa in WI-38 clones and/or identified previously as subject to XCI (Fig. 4C). With these stringent criteria, we identified five, five, and nine "cancer-specific" escapees in the ZR-75-1, SK-BR-3, and MDA-MB-436 cells, respectively. To increase the number of informative X-linked genes evaluated, we also performed an RNA-seq analysis on mRNA from two additional WI-38 clones and the three tumor cell lines. We identified six, one, and 15 "cancer-specific" escapees in the ZR-75-1, SK-BR-3, and MDA-MB-436 lines, respectively (Fig. 4D). We validated Xi-linked reactivation for several of these genes (Supplemental Figs. S4H–J, S5A). In conclusion, although RNA SNP6 and RNA-seq analyses do not necessarily reveal exactly the same "cancer-specific" escapees (15%–23% overlap was found, depending on cell line) due to the different SNPs assessed by the two methods (mainly intronic and mainly exonic, respectively), the combination of both techniques allowed us to identify 10 (9% of informative X-linked genes), five

(8%), and 20 (13%) Xi-linked genes as being abnormally reactivated in ZR-75-1, SK-BR-3, and MDA-MB-436 cell lines, respectively (Fig. 4E; Supplemental Table S1).

The preceding allele-specific analysis could not identify genes that are fully silenced in somatic cells and reactivated from only one allele in cancer cells, such as members of the C/T antigen

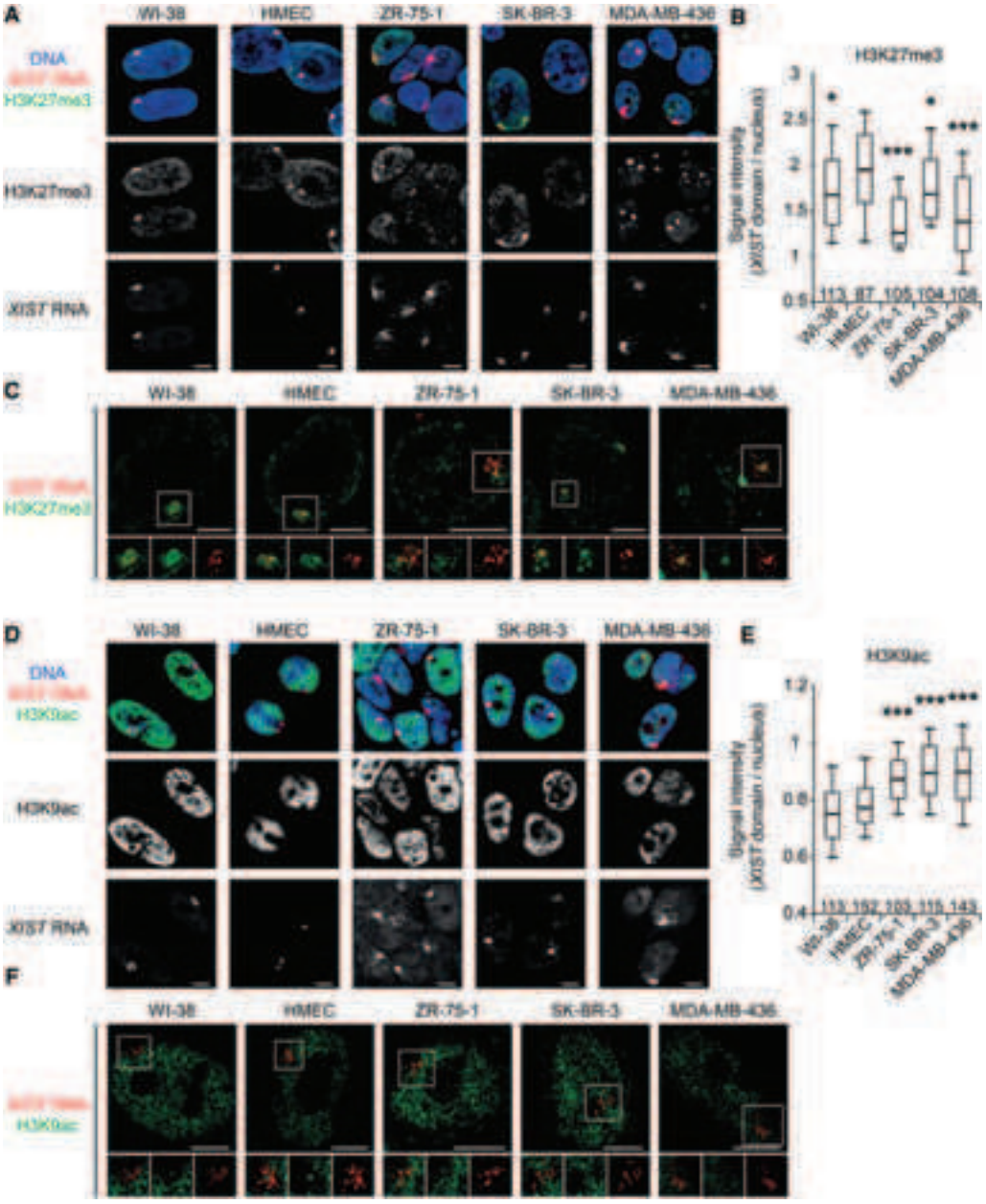


Figure 2. H3K27me3 and H3K9ac profiles associated with *XIST*-coated X chromosomes are impaired in breast cancer cell lines. (*A*) Z-projections of 3D immuno-RNA FISH show representative examples of the level of H3K27me3 enrichment (green) on *XIST* RNA domains (red) in normal (WI-38 and HMEC) and breast cancer cell lines (ZR-75-1, SK-BR-3, and MDA-MB-436). NB: In MDA-MB-436, the highly H3K27me3 enriched bodies visible in each nucleus do not belong to the X chromosome (nor in metaphase [Fig. 3C] or in interphase [Supplemental Fig. S3F]). (*B*) Boxplot shows the levels of H3K27me3 enrichment on *XIST* domains relative to the rest of the nucleus. Numbers of analyzed nuclei are shown *above* the *x*-axis. For details on quantification method see Supplemental Figure S2A,B. (*C*) High-resolution immuno-RNA FISH shows representative examples of H3K27me3 enrichment (green) on *XIST* RNA domains (red) in normal and breast cancer cell lines. *Insets* for H3K27me3, *XIST* RNA, and merge are shown *below* each cell line. (*D*) Single section of 3D immuno-RNA FISH shows representative examples of the level of H3K9ac depletion (green) on *XIST* RNA domains (red) in normal and breast cancer cell lines. (*E*) Boxplot shows the levels of H3K9ac depletion on *XIST* domains relative to the rest of the nucleus. The numbers of analyzed nuclei are shown *above* the *x*-axis. For details on the quantification method, see Supplemental Figure S2A,C. (*F*) High-resolution immuno-RNA FISH shows representative examples of H3K9ac depletion (green) on *XIST* RNA domains (red) in normal and breast cancer cell lines. *Insets* for H3K9ac, *XIST* RNA, and merge are shown *below* each cell line. (Boxplots) Upper whisker represents 90%, upper quartile 75%, median 50%, lower quartile 25%, and lower whisker 10% of the data set for each cell line. (***) $P < 0.001$; (**) $P < 0.01$; (*) $P < 0.05$ using the Student's *t*-test. All data sets are compared with HMEC data set. Scale bar, 5 µm.

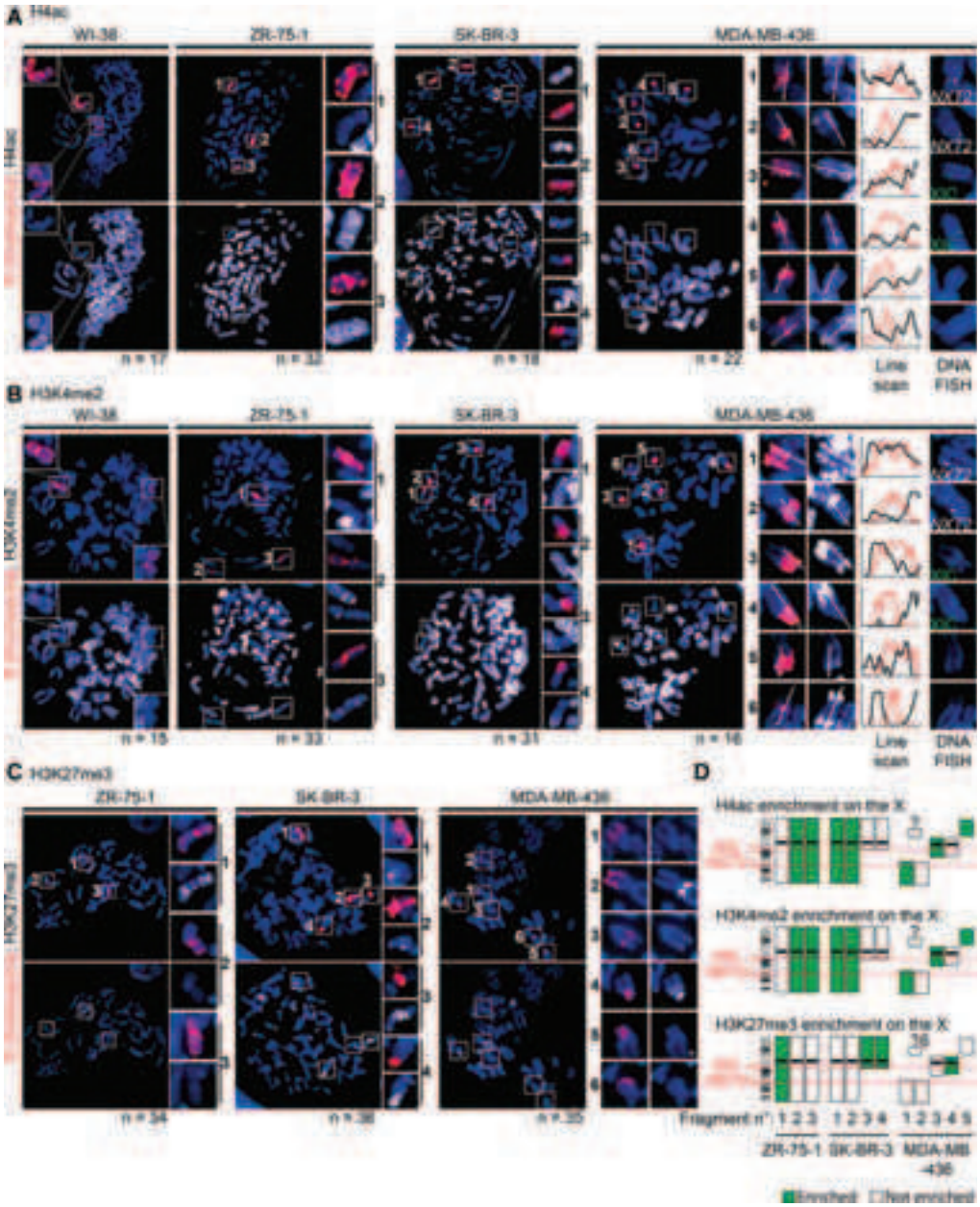
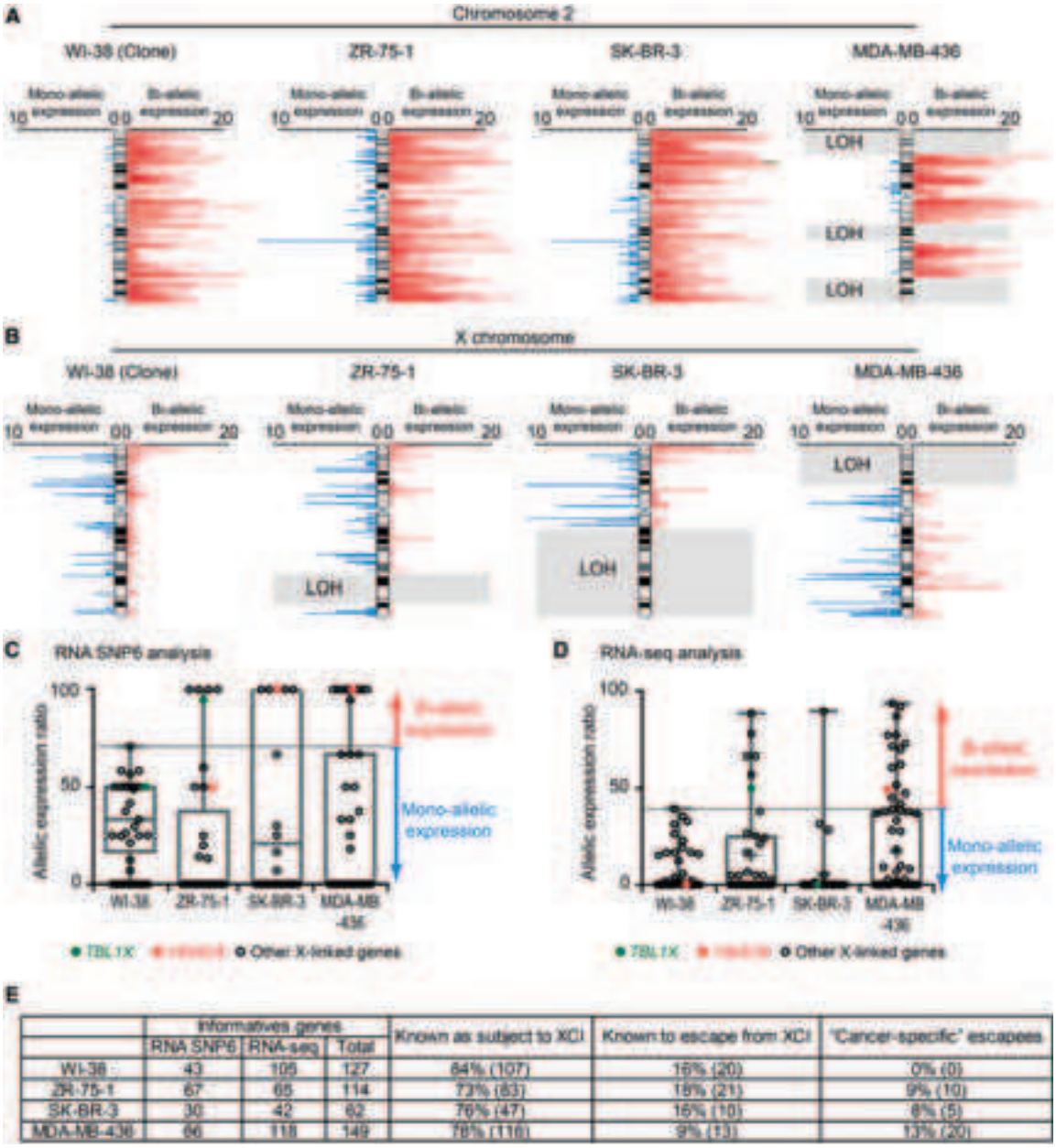


Figure 3. The inactive X chromosome is still epigenetically distinguishable from its active counterpart. (*A*) Representative examples of immunofluorescence show the status of H4ac (white) depletion/enrichment on X chromosomes (X-paint DNA FISH, red) on metaphase spreads from normal (WI-38) and breast cancer cell lines (ZR-75-1, SK-BR-3, and MDA-MB-436). On the *right*, MDA-MB-436 cells carry six X-chromosome fragments with a "2-by-2" homology, as assessed by the presence or absence of the *NXT2* (white) or XIC loci (green), and line scans show H4ac enrichment variation between these X-fragments and the neighboring autosomal regions. As expected, one X chromosome (Xi) lacks H4ac staining in normal WI-38 cells (and HMEC, not shown). ZR-75-1 and SK-BR-3 cell lines harbor a reduced H4ac staining on one and two X chromosomes, respectively, in agreement with the number of *XIST*-coated X chromosomes shown in Figure 1A. In MDA-MB-436 cells, homologous X-chromosome fragments (two containing the XIC locus, two containing the *NXT2* locus, and two with none of them) display opposite H4ac staining, suggesting that there is still one inactive and one active X chromosome linked to those loci, although fragmented. (*B*) Representative examples of immunofluorescence show the status of H3K4me2 (white) depletion/enrichment on X chromosomes (X-paint DNA FISH, red) on metaphase spreads from normal and breast cancer cell lines. On the *right*, line scans show H3K4me2 enrichment variation between the six X-fragments (for details, see *A*) and the neighboring autosomal regions in MDA-MB-436 cells. In each tumoral cell line, H3K4me2 depletion patterns follow the H4ac profiles found in *A*. (*C*) Representative examples of immunofluorescence show the status of H3K27me3 (white) enrichment on X chromosomes (X-paint DNA FISH, red) in metaphase spreads from breast cancer cell lines. ZR-75-1 and SK-BR-3 cell lines harbor an accumulation of H3K27me3 on one and two X chromosomes, respectively, in agreement with the number of *XIST*-coated X chromosomes shown in *A*. In MDA-MB-436 cells, H3K27me3 staining was only enriched on the X-chromosome fragment, where the XIC region lies. Indeed, RNA/DNA FISH analysis showed that this X fragment corresponds to the one coated by *XIST* RNA in interphase cells, which is not the case for the other fragments (Supplemental Fig. S3F). In SK-BR-3 and MDA-MB-436 cell lines, H3K27me3 spreads into the autosomal fragments translocated to the XIC-containing fragment. (*D*) Schematic view of H4ac, H3K4me2, and H3K27me3 patterns on X-chromosomes in the three tumor cell lines.



	Informatives genes			Known as subject to XCI	Known to escape from XCI	"Cancer-specific" escapees
	RNA SNP6	RNA-seq	Total			
WI-38	43	105	127	84% (107)	16% (20)	0% (0)
ZR-75-1	67	65	114	73% (83)	18% (21)	9% (10)
SK-BR-3	30	42	62	76% (47)	16% (10)	8% (5)
MDA-MB-436	66	118	149	78% (116)	9% (13)	13% (20)

Figure 4. Abnormal reactivation of the inactive X chromosome in breast cancer cell lines. (*A*,*B*) RNA SNP6 analysis shows the expression status of an autosomal chromosome, as example Chromosome 2 (*A*), and the X chromosome (*B*) in normal (WI-38) and breast cancer cell lines (ZR-75-1, SK-BR-3, and MDA-MB-436). Red bars indicate biallelic expression, and blue bars indicate monoallelic expression. The bar length represents the number of expressed informative SNPs on a 50-SNP sliding window. Gray rectangles correspond to noninformative regions due to loss of heterozygosity (LOH). Two WI-38 subclones (#1 and #28), carrying an inactive X chromosome of opposite parental origin, show clear monoallelic expression from either the maternal or paternal X chromosome confirming the clonality of the subclones (see Supplemental Fig. S4B). Allele-specific PCR analysis also confirmed the clonality of the three breast tumor cell lines (see Supplemental Fig. S4C–E). (*C*) RNA SNP6 analysis shows levels of X-linked gene allelic expression. X-linked genes known as subject to XCI (Carrel and Willard 2005; Cotton et al. 2013) and/or considered as monoallelically expressed in WI-38 clones (i.e., for each informative gene, <2/3 of the SNPs were observed as biallelically expressed) are shown on the boxplots. (*D*) RNA-seq analysis shows levels of X-linked gene allelic expression. X-linked shown on the boxplots are known to be subject to XCI (Carrel and Willard 2005; Cotton et al. 2013) and/or are considered as monoallelically expressed in WI-38 clones (i.e., for each informative gene, the allelic expression ratio is <40, i.e., expressed <20% on one of the two alleles). (*E*) Summary of the informative genes identified by the RNA SNP6 and RNA-seq approaches. Genes "known as subject to XCI" or "known to escape from XCI" refer to previous studies (Carrel and Willard 2005; Cotton et al. 2013). WI-38 data correspond to the two clones.

family that show aberrant expression in cancer cells (Grigoriadis et al. 2009). By assessing the overall expression of C/T members, we found increased expression of several C/T antigens in the cancer cell lines but not in normal cells (Supplemental Fig. S5B). For one C/T antigen gene (*MAGEA6*), we used RNA FISH to show that this aberrant expression usually originated from the

active rather than the inactive X in tumor cells (Supplemental Fig. S5C).

In order to assess allelic expression of specific genes at the single-cell level, we developed RNA FISH probes for several X-linked escapee genes, bypassing the issue of uninformative SNPs. We confirmed that *HDAC8* is expressed from the *XIST* RNA-associated Xi chromosome only in MDA-MB-436 and SK-BR-3 cells (Supplemental Fig. S5D), whereas *TBL1X* was expressed from Xi only in ZR-75-1 (Fig. 5A). We also confirmed that *APOOL* and *SYTL4* are only escaping from XCI in MDA-MB-436 by RNA FISH (data not shown). *ATRX* was used as a control gene that is subject to XCI in all five cell lines (Supplemental Fig. S5E).

We then investigated the degree to which reactivation could impact on gene dosage for *TBL1X*, one of the "cancer-specific" escapees in ZR-75-1 cells. Using IF against TBL1X combined with RNA FISH, we correlated the protein levels of TBL1X to its expression from the Xi (Fig. 5B). On average, in ZR-75-1, the IF signals appear highly heterogeneous but also stronger than the four other cell lines (in agreement with the RNA level) (Fig. 5C; Supplemental Fig. S5F). We noted that MDA-MB-436 cells also showed slightly increased protein levels, consistent with overall *TBL1X* expression levels in this cell line, which must be due to higher expression of the single active allele (on the Xa) in this cell line. To determine whether the higher protein levels in ZR-75-1 are due to reactivation of *TBL1X* on the Xi or to overexpression of the active alleles on the Xa (as in MDA-MB-436), we quantitated the IF signal in cells that do, or do not, show *TBL1X* transcriptional reactivation on the Xi (Fig. 5D). Significantly more TBL1X staining was seen in ZR-75-1 nuclei that displayed *TBL1X* reactivation (Fig. 5E). We also sorted ZR-75-1 cells by FACS based on TBL1X staining intensity (Fig. 5F) and observed significantly more biallelic expression of *TBL1X* in cells with the highest levels of TBL1X protein staining (Fig. 5G).

Local epigenetic erosion affects genes that escape XCI in cancer

To investigate further the underlying causes of Xi gene reactivation in cancer cells, we investigated the chromatin status of "cancer-specific" escapees at the molecular level. First, the DNA methylation status of multiple X-linked gene promoters was investigated using EpiTYPER analysis (Sequenom) (Supplemental Fig. S6A). All escapees (normal or "cancer-specific") showed low levels of DNA methylation at their promoters (e.g., *KDM5C*, *HDAC8*). However, we noted that some genes subject to XCI (i.e., only expressed from the Xa) in cancer cell lines, nevertheless showed low promoter methylation (e.g., *TBL1X* in SK-BR-3 cells or *HDAC8* in ZR-75-1). This suggests that they might be more prone to reactivation in a cancer context, with outright reexpression from the Xi in only some cell lines.

We also performed chromatin immunoprecipitation and sequencing (ChIP-seq) on normal and cancer cell lines to assess Xi chromatin status. We investigated H3K27me3 (associated with the inactive state of the Xi), H3K4me3 (enriched at transcriptional start sites [TSSs] of active genes), and RNA Pol II. The comparison of quantile-normalized H3K27me3 profiles revealed major changes for the X chromosomes between normal and tumor cells (Fig. 6A). In HMECs, low-resolution chromosome-wide profiles exhibited a pattern of domains that is highly reminiscent of the distinct nonoverlapping regions of the human Xi previously reported for H3K9me3 and H3K27me3 (Chadwick 2007; Chadwick and Willard 2004). Indeed, comparing the H3K9me3 and H3K27me3 data from the ENCODE Project Consortium (2012) with our H3K27me3 ChIP-seq data sets, these different types of heterochromatin domains are readily detectable in normal HMEC and WI-38 cells (Supplemental Fig. S6C; data not shown). In contrast, the organization of these H3K27me3-enriched domains was found to be heavily perturbed in ZR-75-1 and MDA-MB-436. In ZR-75-1 cells, the X chromosome displays a global, nearly uniform pattern of H3K27me3, with no discernable enriched domains (Fig. 6A). The analyzable parts of the X in SK-BR-3 cells (where an Xi is retained) are much less perturbed, apparently respecting the H3K27me3 domains. These results are in line with the reorganization of the Xi in interphase cells by IF/FISH (Fig. 2A–C). The X chromosome in MDA-MB-436 shows a heavily segmented H3K27 methylation profile, as (1) the beginning of the short arm shows no H3K27me3 marks (evident consequence of the loss of the Xi fragment); (2) the rest of the short arm displays significant H3K27me3 enrichment, although the profile is rather different from that seen for HMEC; (3) the region surrounding the XIC shows a profile similar to that seen in normal cells; and (4) the region spanning Xq21.33 to the end of the long arm, which is no longer linked to the XIC (see Fig. 3), does not display discernable H3K27me3 domains; in particular, the two highly enriched domains visible in normal cells are lacking (Fig. 6A, red dotted rectangles). To further consolidate these observations, we compared the variation of H3K27me3 signals along the X chromosome between HMEC and the other four cell lines (WI-38 and the three tumor cell lines). Highly variable H3K27me3 patterns across the X chromosome were observed in the tumor cell lines, and several regions for which an Xi copy was still present showed a drastic decrease in H3K27me3 levels (e.g., the Xq21.33-Xq24 region in ZR-75-1 and MDA-MB-436) (Fig. 6C). On the other hand, much less pronounced variation in H3K27me3 distributions on the Xi was observed when HMEC and WI-38 cells were compared, despite their divergent tissue origins (lung fibroblasts versus mammary epithelial cells) (Fig. 6C). Importantly, in the breast cancer lines, the perturbations were not unique to the Xi, as we also noted aberrant H3K27me3 landscapes across autosomal regions of cancer cells (e.g., Chromosome 17 on Supplemental Fig. S6B), indicating that this is a genome-wide characteristic of tumor cells. Thus, we conclude that both genome-wide and Xi-specific distributions of H3K27me3 are severely disrupted in breast tumor cell lines. Although this is partly due to genetic changes (Xi translocations and regional losses), the Xi epigenomic landscape is clearly disorganized, consistent with our aforementioned observations using IF.

Next, we assessed patterns of H3K4me3 and RNA Pol II around the TSS of X-linked genes, and noted that the escapees identified in each cell line displayed a generally higher enrichment of RNA Pol II and H3K4me3 than X-linked genes that were expressed only from the Xa (Supplemental Fig. S6D,E). Similarly, "cancer-specific" escapees generally exhibited higher enrichment at their TSS in the cell lines where they escaped compared to HMECs (Fig. 6D; Supplemental Fig. S7A) with a few exceptions (e.g., *CFP*, *FLNA*, and *MOSPD1* in MDA-MB-436 cells displayed no obvious differences in TSS profiles) (Supplemental Fig. S7B). We also noted that "cancer-specific" escapees, such as *HDAC8* or *NXT2*, exhibit additional and/or enlarged H3K4me3 sites in tumor cells when compared to HMEC (Supplemental Fig. S7B).

As H3K27me3 is normally rather broadly distributed on the Xi, rather than being TSS centered (Marks et al. 2009; Simon et al. 2013), we examined the local environment of genes that normally escape XCI (e.g., *KDM5C* and *SMC1A*) or are silenced on the Xi (e.g., *HUWE1*) and found them to display the expected low and high enrichments, respectively (Fig. 6B, center panel).

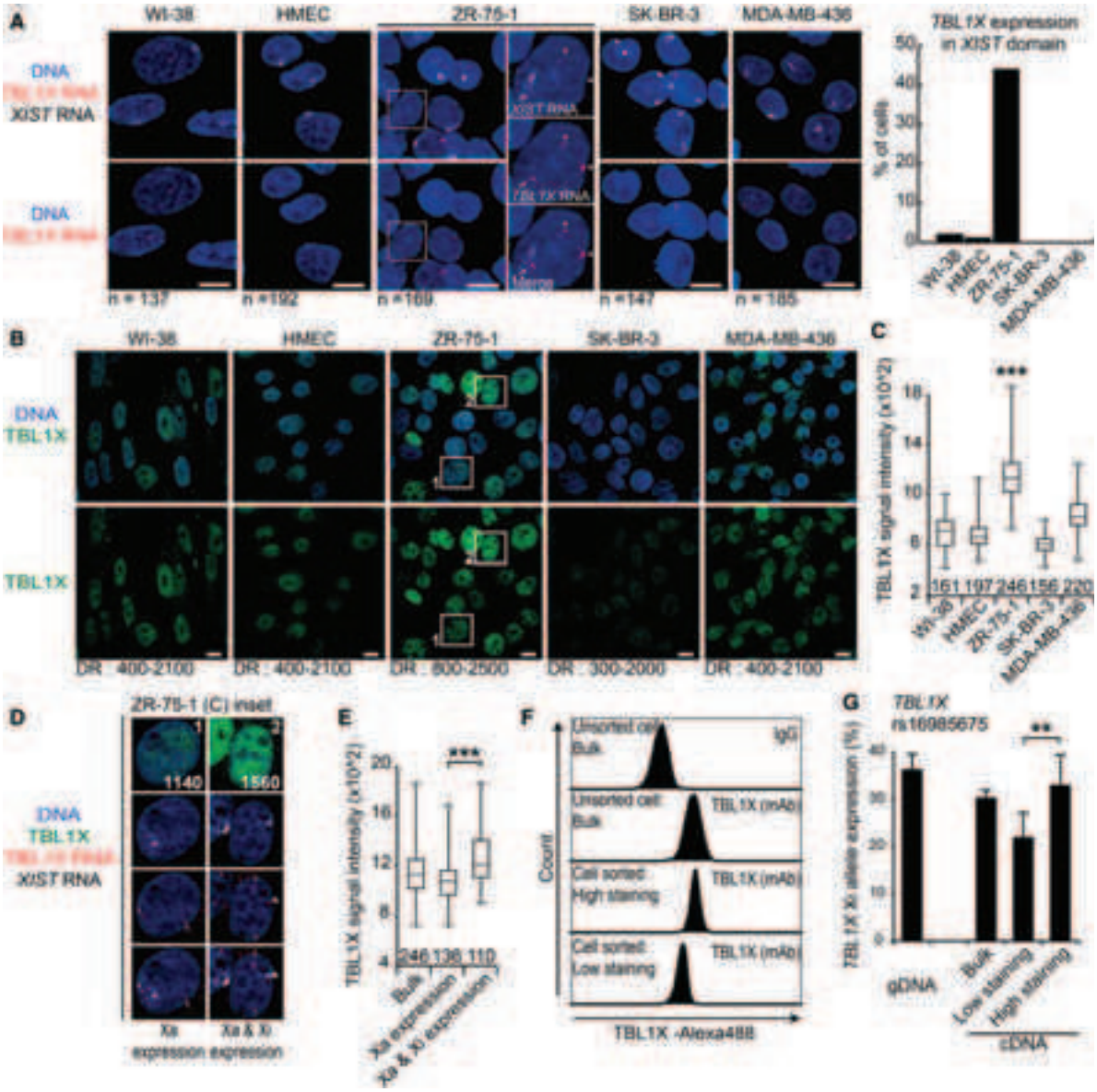


Figure 5. Reactivation of X-linked genes in breast cancer cell lines can lead to an increase of protein amount. (*A*) Z-projections of 3D RNA FISH show representative examples of *TBL1X* expression (red) at *XIST* domains (white) in normal (WI-38 and HMEC) and breast cancer cell lines (ZR-75-1, SK-BR-3, and MDA-MB-436). In ZR-75-1 cells, arrowheads indicate active X chromosomes and the arrow the *XIST*-coated chromosome. On the *right*, bar graph shows levels of *TBL1X* expression from *XIST* domains, with reactivation in ZR-75-1 cells. (*B*) Immunostaining shows TBL1X protein (green). The dynamic range (DR) of the brightness and contrast of each image (ImageJ) is indicated *below*. (*C*) Boxplot shows the intensity of TBL1X immunostaining for each cell line. The upper whisker represents the maximum value, upper quartile 75%, median 50%, lower quartile 25%, and lower whisker the minimum value of the data set. The number of nuclei analyzed is indicated *above* the *x*-axis. (***) $P < 0.001$ using the Student's *t*-test. WI-38, ZR-75-1, SK-BR-3, and MDA-MB-436 are compared with HMEC. (*D*) The *inset* of two ZR-75-1 nuclei from *C* shows a combination of TBL1X protein immunofluorescence staining (green) and RNA FISH for *TBL1X* (red) and *XIST* (gray). In the *left* nucleus, where *TBL1X* is expressed only from the active X chromosome, the IF signal intensity is 1140 a.u., whereas in the *right* nucleus, where both Xa and Xi *TBL1X* alleles are expressed, the intensity is as high as 1560 a.u. (*E*) Boxplot shows the levels of TBL1X signal intensity either in the whole cell population (bulk; *left* box) or in cells in which *TBL1X* is expressed only from the active X chromosome (*middle* box) or when *TBL1X* is expressed from all X chromosomes (*right* box). The upper whisker represents the maximum value, upper quartile 75%, median 50%, lower quartile 25%, and lower whisker the minimum value of the data set. Nuclei number analyzed is indicated *above* the *x*-axis. (*F*) Cell sorting of ZR-75-1 cells based on TBL1X signal intensity. An IgG antibody has been used as negative control. (*G*) Bar graph shows the level of *TBL1X* expression from the *XIST*-coated X chromosome by pyrosequencing at SNP rs16985675. *Left* bar represents the gDNA control, which is in agreement with the allelic imbalance (i.e., one Xi allele and two Xa alleles). Data represent the mean values ±SEM. (***) $P < 0.001$; (**) $P < 0.01$; (*) $P < 0.05$ using the Student's *t*-test.

For "cancer-specific" escapees (*TBL1X* in ZR-75-1, *HDAC8* in MDA-MB-436 and SK-BR-3), no obvious systematic correlation between local H3K27me3 levels and escape/silencing could be seen (Fig. 6B, left and right panels). Although the global disorganization of H3K27me3 domains in tumor cell lines is not necessarily reflected locally at the level of genes, H3K27me3

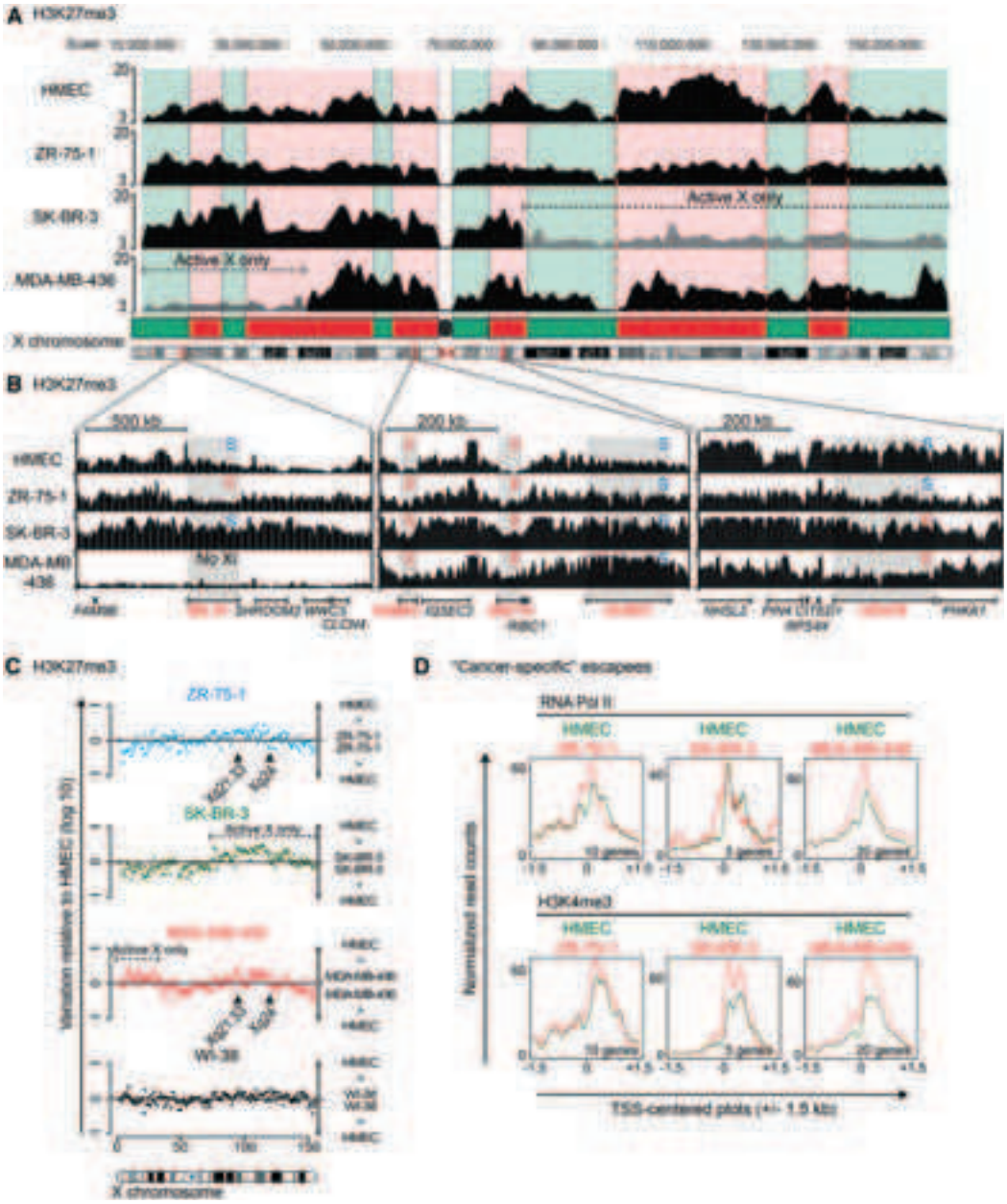


Figure 6. Chromatin landscape of the inactive X chromosome is disrupted in breast cancer cell lines. (*A*) Scheme of H3K27me3 enrichment (ChIP-seq) across the whole X chromosome. Red and green domains represent H3K27me3 and H3K9me3 enriched regions, respectively, as identified in normal human cells (Chadwick 2007). Regional loss of inactive X is indicated (and depicted by gray region). The two main enriched H3K27me3 domains' loss in ZR-75-1 and MDA-MB-436 are depicted by the two red dotted rectangles. (*B*) H3K27me3 enrichment is detailed for three regions of the X chromosome carrying genes subjected (S) or escaping XCI (E). (*C*) Dot plots show variation of H3K27me3 enrichment along the X chromosome (1-Mb bins) of the three tumoral cell lines and WI-38 relative to HMEC. (*D*) TSS-centered plots (±1.5 kb) show RNA Pol II and H3K4me3 enrichment for the "cancer-specific" escapees (cf. Supplemental Table S1) of each tumoral cell line (red line) and HMEC (green line). The number of genes analyzed is indicated *below* each plot.

disorganization may nevertheless affect long-range regulatory landscapes, creating a context favoring escape in concert with additional events.

Finally, we monitored allele-specific enrichment of H3K4me3, RNA Pol II peaks, and H3K27me3 enrichment across genes with informative SNPs. *HUWE1* revealed exclusively monoallelic enrichment for all three marks, consistent with its silence on the Xi in all lines (Supplemental Fig. S7C), whereas escapees *SMC1A* and *DDX3X* and several "cancer-specific" escapees displayed biallelic H3K4me3 and RNA Pol II, with monoallelic H3K27me3 (Supplemental Fig. S7C,D). Thus, for informative escapees in the three cancer cell lines, H3K27me3 is observed on one allele, whereas both alleles show signs of active transcription (H3K4me3 and RNA Pol II occupancy).

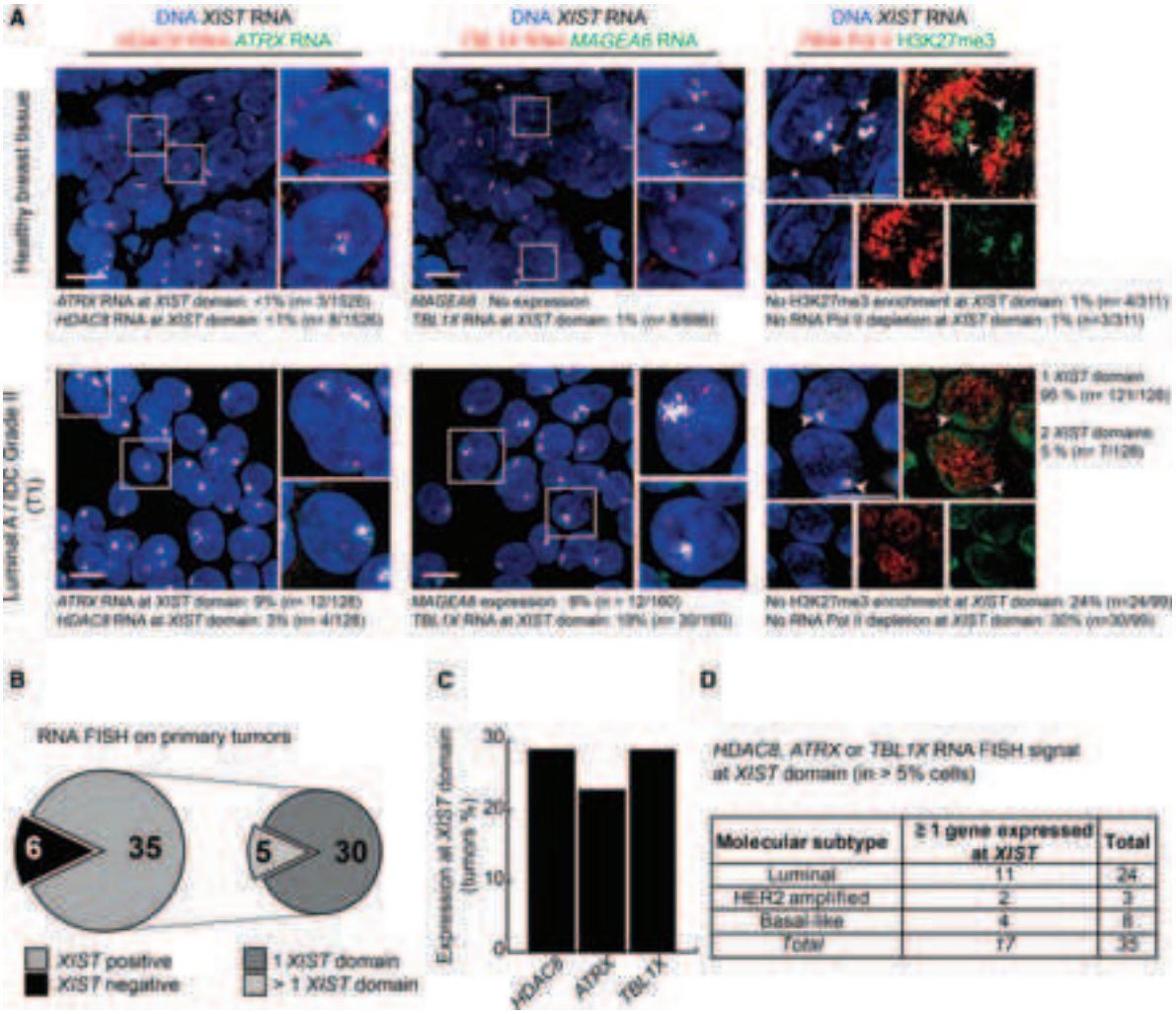


Figure 7. The inactive X chromosome is reactivated in primary breast tumors. (*A*) Z-projections of 3D RNA FISH show representative examples of expression of *HDAC8* (red) and *ATRX* (green) (*left*) or *TBL1X* (red) and *MAGEA6* (green) (*middle*) at *XIST* domains (gray) in healthy breast tissue and invasive ductal carcinoma (IDC; Luminal A Grade III tumor). On the *right*, Z-projections of super-resolutive 3D immuno-RNA FISH show representative examples of the level of H3K27me3 enrichment (green) and RNA Pol II depletion (red) on *XIST* RNA domains (gray) in healthy and tumoral breast tissues. Arrowheads indicate the *XIST* domains. Quantification of RNA Pol II exclusion and H3K27me3 enrichment at *XIST* domains have been carried out on images acquired with a confocal spinning-disk microscope. Scale bar, 10 µm. (*B*) Summary of the *XIST* domain positive (domains in >10% of the nuclei) and negative tumors among the 41 primary breast tumors studied. (*C*) Summary of the number of tumors harboring *HDAC8*, *ATRX*, or *TBL1X* expression at *XIST* domain (assessed by RNA FISH). A gene showing expression within the *XIST* domain in >5% of the nuclei is considered as reactivated in this tumor. (*D*) The table recapitulates the number of *XIST* positive tumors with Xi-linked gene reactivation according to their molecular subtypes: Luminal, HER2 amplified, or Basal-like (BCL).

Perturbation of the inactive X chromosome is also found in primary breast tumors

We next assessed whether epigenetic disruption of the Xi also occurs in primary breast tumors. Due to the cellular heterogeneity in such samples, as well as the variable presence of normal stromal cells, we focused on single-cell techniques (IF and RNA FISH) to investigate the Xi. We analyzed seven tumors using a tumor stamp technique with fresh samples (see Methods) to evaluate the degree of enrichment of H3K27me3 at sites of *XIST* RNA accumulation (Fig. 7A; Supplemental Fig. S8). H3K27me3 enrichment on the Xi in tumors was highly variable, showing almost no enrichment in four of the seven tumors analyzed: T1, T2, T4, and T4meta. This confirmed our observations from cell lines that Xi chromatin status is frequently disrupted in breast cancer. We also noted that H3K27me3 enrichment within a *XIST* RNA domain was not necessarily accompanied by a depletion of RNA Pol II (e.g., tumors T1 and T2) (Fig. 7A; Supplemental Fig. S8A). We also noted a significant decrease in DNA enrichment at the level of the *XIST* RNA domain in primary tumors (Supplemental Fig. S8E,F). Taken together, these results demonstrate that the Xi shows significant chromosome disorganization and chromatin disruption in primary breast tumors, similarly to the tumor cell lines described above and that suggesting that disappearance of the Barr body in certain breast cancers is indeed due to epigenetic instability.

We next assessed whether the aberrant chromatin status of the Xi also translated into X-linked gene reactivation by assessing

XIST together with *HDAC8*, *ATRX*, *MAGEA6*, and *TBL1X* expression on fresh tumor stamps (including those analyzed above) or tumor-tissue cryosections. These genes were chosen because (1) they are robustly detected by RNA FISH; (2) they are "cancer-specific" escapees in some tumor cell lines (except *ATRX* and *MAGEA6*); and (3) *HDAC8* and *ATRX* lie in proximity to each other and to *XIST* (within a few megabases), thus minimizing their chances of being separated by translocations and facilitating RNA FISH analysis in tumors. We analyzed 41 primary breast tumors with corresponding normal tissue for 15 of them (examples shown in Fig. 7A; Supplemental Fig. S8). Thirty-five tumors were *XIST*-positive, with at least one *XIST* RNA domain in ≥10% of nuclei (Fig. 7B). The number, organization, and intensity of *XIST* RNA domains varied substantially between tumors and even among cells of the same tumor (Supplemental Fig. S8). For X-linked genes, aberrant reactivation from the Xi was considered to occur if ≥5% of nuclei harbored a nascent RNA FISH signal at or within a *XIST* RNA domain of a given sample. With these criteria, we found 28%, 20%, and 29% of tumors displayed aberrant *HDAC8*, *ATRX*, and *TBL1X* expression from the inactive X, respectively (Fig. 7C). Note that in healthy breast tissue, we never observed >1% of nuclei showing X-linked gene RNA FISH signal within the *XIST* RNA domain. Furthermore, we did not observe higher degrees of reactivation for any of these three X-linked genes in particular cancer subtypes, although only a limited number of HER2+ and basal-like tumors were analyzed (Fig. 7D). We also analyzed the cancer/testis antigen family 1, member 6 *MAGEA6* gene, which is normally silent on both Xa and Xi. None of the primary tumors showed reactivation from the Xi, although in some tumors, *MAGEA6* expression was detected from the presumed Xa (Fig. 7A; Supplemental Fig. S8B–D), similarly to our data in breast cancer cell lines. In summary, RNA FISH analysis of 35 *XIST*-positive primary breast tumors of the luminal, HER2+, and basal-like subtypes, revealed that all three X-linked genes tested, *HDAC8*, *TBL1X*, and even *ATRX*, show Xi reactivation in a significant proportion of tumor cells in stark contrast to the situation in healthy breast tissue from the same patient.

To extend our findings, we analyzed publicly available data for biallelic expression of X-linked genes, using a data set for which both RNA-seq and DNA SNP6 data were available (Shah et al. 2012). After we filtered out tumors of "poor" quality (see Supplemental Methods) and those contaminated by normal cells (Popova et al. 2009), we identified 25 BLC tumor samples with a heterozygous X chromosome, suggesting they likely retained an inactive X or at least some region of the Xi (Supplemental Fig. S9A; Supplemental Table S2). Among these tumors, we identified 183 informative genes, of which 78 were expressed biallelically and 105 monoallelically. Almost half of these biallelically expressed genes are subject to XCI in healthy human cells (Supplemental Fig. S9B; Cotton et al. 2013). Furthermore, in agreement with our findings in the three tumor cell lines, *TBL1X*, *NXT2*, and *DOCK11* were among the 14 genes that were biallelically expressed in at least two primary breast tumors (Supplemental Fig. S9C). We identified no obvious correlation between the degree of "cancer-specific" escape from XCI and the BRCAness of the tumor (as defined in Popova et al. 2012; Supplemental Table S2).

In summary, our analysis of Xi transcriptional status in a total of about 140 primary breast tumors of the luminal, HER2+, and basal-like subtypes, using both RNA FISH and RNA-seq analyses, revealed that multiple X-linked genes are reactivated on the inactive X chromosome.

Discussion

We have conducted an in-depth investigation of the nuclear organization, chromatin status, and chromosome-wide transcriptional activity of the inactive X chromosome in breast cancer cell lines and primary tumor samples. We can conclude that a frequent cause of Barr body loss in breast cancer is due to the global perturbation of its nuclear organization and disruption of its heterochromatic structure. Furthermore, the aberrant epigenomic landscapes we have uncovered for the Xi in breast cancer cells are accompanied by a significant degree of sporadic gene reactivation, which in some cases can lead to aberrant dosage at the protein level (Supplemental Fig. S10A).

Epigenetic erosion of the Barr body in breast cancer

Epigenetic perturbations of the inactive X chromosome were found at multiple levels in breast cancer. Based on microscopy, *XIST* RNA coating was often found to be highly dispersed, with variable H3K27me3 enrichment, and a marked absence of an RNA Pol II-depleted nuclear compartment. Based on ChIP-seq, abnormal presence of both RNA Pol II and H3K4me3 was observed at "cancer-specific" escapees, reminiscent of the chromatin organization of the normally escapees from XCI in noncancer cells (Kucera et al. 2011; Cotton et al. 2013). Importantly, however, virtually all informative "cancer-specific" escapees displayed simultaneously repressive (H3K27me3) and active (H3K4me3, RNA Pol II recruitment) chromatin marks (see Supplemental Fig. S7D), suggestive of bivalent chromatin, as observed in ES cells (Bernstein et al. 2006), which may reflect, or even underlie, metastable states of gene expression from the Xi in a cancer context. The Xi was also severely perturbed at a more global chromatin level, with aberrant distributions of H3K27me3 and acetylation of H3 and H4 present in interphase breast cancer cells. The disruption of H3K27me3 domains that we observed based on ChIP-seq in breast cancer cell lines may reflect the nuclear disorganization of the Xi, as it has been shown that H3K27me3 enriched domains in normal cells tend to be clustered together in interphase and most likely participate in the specific chromosomal and nuclear organization of the Barr body (Chadwick and Willard 2004). Nevertheless, despite these global and local epigenetic perturbations in all the breast cancer cell types examined, the Xi could still be distinguished from the Xa. For example, although the degree of enrichment for H3K27me3 on the Xi is lower in cancer cells when compared to HMEC and WI-38 cells, it is still higher than the mean enrichment found over the rest of the genome (Fig. 2B; Supplemental Fig. S10B). Similarly, although exclusion of Cot-1 RNA, RNA Pol II, and euchromatic marks is not complete on the Xi in cancer samples, some degree of exclusion is nevertheless detectable in a subset of cells. Furthermore, "cancer-specific" escapees (like normal escapees) were never expressed to the same levels as their counterparts on the active X.

Possible causes of the epigenetic instability of the inactive X chromosome in breast cancer

Epigenetic instability of the Xi appears to occur across a broad spectrum of breast cancer types with no obvious specificities for particular molecular subclasses. For example, elevated genetic instability, such as in *BRCA1* null and basal-like breast tumors (Richardson et al. 2006; Vincent-Salomon et al. 2007) cannot explain the marked epigenetic instability that we found in all subtypes. We believe that the underlying causes of the structural

and transcriptional lability of the Xi in cancer are probably a result of both genetic and epigenetic defects. For example, the slightly lower levels of *XIST* expression that we observed in most cases might lead to less efficient chromosome coating and contribute to the disruption of the silent nuclear compartment normally present in somatic cells (Chaumeil et al. 2006; Clemson et al. 2006), as well as to the aberrant distribution of H3K27me3 and other chromatin marks. Furthermore, the precise combination of epigenetic factors that ensure the inactive state of different genes on the inactive X chromosome in somatic cells is still very much an open question. Indeed, our study revealed that the rather global epigenetic misregulation in tumor cells results in rather sporadic X-linked gene reactivation, and escape from silencing may be dependent on a gene's local environment, as neighboring genes can behave very differently in a cancer context. For example, the *NXT2* gene was found to show aberrant transcription, whereas its close neighbor, *NUP62CL*, remained silent in MDA-MB-436 cells, although both lie in a non-*XIST*-coated/H3K27me3 depleted region of the Xi (Supplemental Table S1).

Consequences of Xi erosion in breast cancer cells

The epigenetic instability of the Xi in breast cancer, which can result in aberrant X-linked gene expression, might in some cases contribute to a selective advantage for cancer cells. Indeed, several "cancer-specific" escapees identified here have previously been shown to be involved in cancer, such as *HDAC8*, which is implicated in cellular transformation (Oehme et al. 2009) and metastasis formation (Park et al. 2011). *TBL1X*, for which we demonstrated increased protein dosage in the context of its aberrant reactivation from the Xi, belongs to a complex with *HDAC3* that is directly linked to several forms of cancer (Spurling et al. 2008; López-Soto et al. 2009; Kim et al. 2010; Müller et al. 2013; Miao et al. 2014). Aberrant dosage of such X-linked chromatin-associated factors could easily be imagined to lead to pleiotropic effects in a cancer context, promoting or enhancing more genome-wide misregulation. Further studies will be required to explore the extent to which X-linked gene reactivation might contribute to cancer progression.

Importantly, in addition to the aberrant reactivation of genes on the inactive X, aberrant silencing of several genes that normally escape XCI, such as *RAB9A*, *BCOR*, *RPL39*, or *PNPLA4* was also observed in tumor cell lines. *BCOR* mutations have already been implicated in some cancers (Zhang et al. 2012). Aberrant repression of such genes in a cancer context might be due to sporadic epimutation or to impaired protection from XCI through perturbation of boundary elements (Filippova et al. 2005). Finally, we also showed that abnormal activation of cancer/testis Antigen genes, which are known to be aberrantly expressed in cancer, was from the active rather than the inactive X chromosome in one case (*MAGEA6*), pointing to differences in the stability of silent genes on the active versus the inactive X chromosomes in cancer.

Consequences of genetic instability on the epigenetic status of the Xi in cancer cells

Our study also reveals how chromosomal rearrangements, such as deletions or translocations can have an impact on the epigenetic status of a chromosome through loss of the XIC from an inactive X fragment and/or juxtaposition of the XIC to an autosome. We found such a scenario in the MDA-MB-436 cell line, where loss of the XIC from an Xi fragment resulted in reduced H3K27me3 enrichment on the Xi, as expected from previous reports demonstrating that PRC2 is recruited (directly or indirectly) to the Xi via *XIST* RNA (Wutz et al. 2002; Plath et al. 2004; Maenner et al. 2010). However, the H3K27me3 profile on this Xi fragment is not equivalent to a euchromatin region, indicating that other mechanisms may act to maintain an intermediate heterochromatic organization. Furthermore, loss of *XIST* RNA coating and reduced H3K27me3 was not sufficient to result in notably higher rates of sporadic gene reactivation of the inactive X-chromosome fragment when compared to Xi fragments carrying an XIC and expressing *XIST* (Supplemental Fig. S10A). This is presumably because other marks, such as hypoacetylation of H4, hypomethylation of H3K4, and promoter DNA methylation, are not fully perturbed and can propagate the inactive state. Thus, although *XIST* RNA and PRC2-associated chromatin changes may participate in maintaining the inactive state, they do not appear to be essential in the context of this particular cell line. We also made the intriguing observation that in X:autosome translocations involving an Xi fragment still carrying an XIC and expressing *XIST* RNA, H3K27me3 enrichment could be found to spread into the autosomal sequences adjacent to the XIC (for example in the SK-BR-3 and MDA-MB-436 cell lines) (Fig. 3C). Although we were not able to evaluate whether this results in aberrant gene silencing, such a spread of heterochromatin into autosomal regions as previously shown (Cotton et al. 2014) could clearly have important implications in a cancer context by inducing functional LOH for critical genes such as tumor suppressors.

In conclusion, the perturbed transcriptional and chromatin status of the inactive X chromosome that we have identified in the context of breast cancer opens up several important clinical perspectives. Today, there is still no rapid and efficient way to evaluate the epigenetic instability of tumor cells in a clinical context (Portela and Esteller 2010). In theory, detection of X-linked gene reactivation and aberrant chromatin status using IF and RNA FISH in breast tumors could provide valuable biomarkers to assess epigenetic status and/or to evaluate responsiveness of tumors to drug treatments (Huang et al. 2002). Whether the same degree of Xi epigenetic instability will be found in other types of cancer remains an interesting question for the future.

Methods

RNA, DNA FISH, and immunofluorescence

For *XIST* RNA FISH, a combination of two probes covering 16 kb of *XIST* mRNA was used (Okamoto et al. 2011). For nascent transcript detection by RNA FISH, the following BAC (CHORI) probes were used: *HDAC8* (RP11-1021B19), *TBL1X* (RP11-451G24), *ATRX* (RP11-42M11), *HUWE1* (RP11-155O24), and *KDM5C* (RP11-258C19). The correct chromosomal location of BACs was first verified using DNA FISH on metaphase spreads. A FISH probe for *MAGEA6* was generated by cloning the genomic sequence in pCR-XL-TOPO vector. Human Cot-1 DNA (Invitrogen) was used for Cot-1 RNA FISH. Probes were labeled by nick translation (Vysis) with Spectrum Red-dUTP, Spectrum Green-dUTP, or Cy5-dUTP following the manufacturer's instructions. RNA and DNA FISH were performed as described previously (Chaumeil et al. 2008). For more details see Supplemental Methods.

Microscopy

Images were generated using a Nikon confocal spinning disk microscope fitted with a 60×/1.4 OIL DIC N2 PL APO VC objective.

For super resolution imaging, structured illumination (3D-SIM) was performed using a DeltaVision OMX microscope (GE Healthcare).

Human SNP Array 6.0 DNA and nascent RNA experiments

DNA copy number profiles

Genomic profiling was performed at Institut Curie using Affymetrix Human SNP Array 6.0; cell files were processed by Genotyping Console 3.0.2 (Affymetrix, reference model file HapMap270, version 29). Human SNP Array 6.0 data were mined using the previously described and validated GAP method (Popova et al. 2009). Segmental absolute copy numbers and allelic contents (major allele counts) were detected. R scripts and full details of the application are available at http://bioinfo-out.curie.fr/projects/snp_gap/ and have been previously reported (Popova et al. 2009). For more details see Supplemental Methods.

Nascent RNA allelic expression

Preparation of samples and analysis of nascent RNA were performed as described previously (Gimelbrant et al. 2007). Briefly, we purified nuclei of assessed cell lines (Nuclei Pure Isolation Kit, Sigma) and subsequently purified nuclear RNA (by classical phenol:chlorophorm extraction). Then, we hybridized cDNA obtained by reverse transcription of nuclear RNA of each sample onto Affymetrix Human SNP Array 6.0. Data was normalized by Genotyping console, and raw single-SNP intensities were taken as allelic expression of corresponding genes. Each SNP was characterized by (1) global expression level score; (2) allelic expression ratio score; and (3) genomic status (loss or retention of heterozygosity score), which were summarized into a biallelic and monoallelic expression status. Genome-wide biallelic and monoallelic expression profiles were obtained by cumulating SNP status in a 50-SNP window and at gene level.

Primary tumors

A hematein-eosin-safran (HES)–stained tissue section was made in each primary tumor to evaluate tumor cellularity and diagnosis. Characterization of the tumor samples was completed by the determination of estrogen receptor, progesterone receptor, ERBB2, cytokeratin 5/6, and epidermal growth factor receptor (EGFR) status determined by immunohistochemistry done according to previously published protocols (Azoulay et al. 2005). All experiments were performed in accordance with the French Bioethics Law 2004-800, the French National Institute of Cancer (INCa) Ethics Charter, and after approval by the Institut Curie review board and the ethics committees of our institution ("Comité de Pilotage of the Groupe Sein"). In the French ethics law, patients gave their approval for the use of their surgical tumor specimens for research. Data were analyzed anonymously.

For details on experimental procedures used for cell culture, DNA methylation analysis, Sanger sequencing, real-time PCR, allele-specific PCR, pyro-sequencing, RNA sequencing analysis, and chromatin immunoprecipitation analysis, see Supplemental Methods.

Data access

All high-throughput data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE62907.

Acknowledgments

We would like to thank all members of the Heard team for helpful discussions and in particular Julie Chaumeil and Simao da Rocha for critical reading of the manuscript. We would also like to thank Michel Wassef and Raphael Margueron for discussions and feedback on this work. Support is gratefully acknowledged from l'Association pour la Recherche sur le Cancer (ARC) (post-doctoral fellowship to R.C.); FRM (Equipe FRM to E.H. from 2006–2009); Equipe labellisée "La Ligue Contre Le Cancer" (Equipe Labellisée to E.H. since 2011); EU FP7 MODHEP EU Grant no. 259743 (to E.H.); Labex DEEP (ANR-11-LBX-0044) part of the IDEX Idex PSL (ANR-10-IDEX-0001-02 PSL); and ERC Advanced Investigator Award no. 250367. This work was also supported by the AVIESAN-ITMO Cancer-INCa grant "Epigenomics of breast cancer" (E.H., H.G., and M.-H.S.). Work in the laboratory of H.G. was supported by the Agence Nationale de la Recherche (ANR-10-LABX-0030-INRT and ANR-10-IDEX-0002-02), the Ligue Contre le Cancer (H.G.; Equipe Labellisée), SATT Conectus Alsace, and the Institut Nationale du Cancer. The authors greatly acknowledge the PICT-IBiSA@BDD (UMR3215/U934) Imaging facility of the Institut Curie, the cytometry platform of the Institut Curie, and the "Centre de Ressources Biologiques" for access to the tumor materials.

References

Agrelo R, Wutz A. 2010. ConteXt of change—X inactivation and disease. *EMBO Mol Med* **2:** 6–15.

Azoulay S, Laé M, Fréneaux P, Merle S, Al Ghuzlan A, Chnecker C, Rosty C, Klijanienko J, Sigal-Zafrani B, Salmon R, et al. 2005. KIT is highly expressed in adenoid cystic carcinoma of the breast, a basal-like carcinoma associated with a favorable outcome. *Mod Pathol* **18:** 1623–1631.

Barr ML, Moore KL. 1957. Chromosomes, sex chromatin, and cancer. *Proc Can Cancer Conf* **2:** 3–16.

Bennett CL, Christie J, Ramsdell F, Brunkow ME, Ferguson PJ, Whitesell L, Kelly TE, Saulsbury FT, Chance PF, Ochs HD. 2001. The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of *FOXP3*. *Nat Genet* **27:** 20–21.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

Boggs BA, Cheung P, Heard E, Spector DL, Chinault AC, Allis CD. 2002. Differentially methylated forms of histone H3 show unique association patterns with inactive human X chromosomes. *Nat Genet* **30:** 73–76.

Carone DM, Lawrence JB. 2013. Heterochromatin instability in cancer: from the Barr body to satellites and the nuclear periphery. *Semin Cancer Biol* **23:** 99–108.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434:** 400–404.

Chadwick BP. 2007. Variation in Xi chromatin organization and correlation of the H3K27me3 chromatin territories to transcribed sequences by microarray analysis. *Chromosoma* **116:** 147–157.

Chadwick BP, Willard HF. 2004. Multiple spatially distinct types of facultative heterochromatin on the human inactive X chromosome. *Proc Natl Acad Sci* **101:** 17450–17455.

Chaumeil J, Okamoto I, Guggiari M, Heard E. 2002. Integrated kinetics of X chromosome inactivation in differentiating embryonic stem cells. *Cytogenet Genome Res* **99:** 75–84.

Chaumeil J, Le Baccon P, Wutz A, Heard E. 2006. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev* **20:** 2223–2237.

Chaumeil J, Augui S, Chow JC, Heard E. 2008. Combined immunofluorescence, RNA fluorescent in situ hybridization, and DNA fluorescent in situ hybridization to study chromatin changes, transcriptional activity, nuclear organization, and X-chromosome inactivation. *Methods Mol Biol* **463:** 297–308.

Chow JC, Heard E. 2010. Nuclear organization and dosage compensation. *Cold Spring Harb Perspect Biol* **2:** a000604.

Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, Attreed M, Avner P, Wutz A, Barillot E, et al. 2010. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141:** 956–969.

Clemson CM, Hall LL, Byron M, McNeil J, Lawrence JB. 2006. The X chromosome is organized into a gene-rich outer rim and an internal core containing silenced nongenic sequences. *Proc Natl Acad Sci* **103:** 7688–7693.
Cotton AM, Ge B, Light N, Adoue V, Pastinen T, Brown CJ. 2013. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol* **14:** R122.
Cotton AM, Chen CY, Lam LL, Wasserman WW, Kobor MS, Brown CJ. 2014. Spread of X-chromosome inactivation into autosomal sequences: role for DNA elements, chromatin features and chromosomal domains. *Hum Mol Genet* **23:** 1211–1223.
Csankovszki G, Panning B, Bates B, Pehrson JR, Jaenisch R. 1999. Conditional deletion of *Xist* disrupts histone macroH2A localization but not maintenance of X inactivation. *Nat Genet* **22:** 323–324.
Csankovszki G, Nagy A, Jaenisch R. 2001. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J Cell Biol* **153:** 773–784.
De Carvalho DD, Sharma S, You JS, Su SF, Taberlay PC, Kelly TK, Yang X, Liang G, Jones PA. 2012. DNA methylation screening identifies driver epigenetic events of cancer cell survival. *Cancer Cell* **21:** 655–667.
Elsheikh SE, Green AR, Rakha EA, Powe DG, Ahmed RA, Collins HM, Soria D, Garibaldi JM, Paish CE, Ammar AA, et al. 2009. Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. *Cancer Res* **69:** 3802–3809.
The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.
Filippova GN, Cheng MK, Moore JM, Truong JP, Hu YJ, Nguyen DK, Tsuchiya KD, Disteche CM. 2005. Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell* **8:** 31–42.
Ganesan S, Silver DP, Greenberg RA, Avni D, Drapkin R, Miron A, Mok SC, Randrianarison V, Brodie S, Salstrom J, et al. 2002. BRCA1 supports XIST RNA concentration on the inactive X chromosome. *Cell* **111:** 393–405.
Gendrel AV, Tang YA, Suzuki M, Godwin J, Nesterova TB, Greally JM, Heard E, Brockdorff N. 2013. Epigenetic functions of Smchd1 repress gene clusters on the inactive X chromosome and on autosomes. *Mol Cell Biol* **33:** 3150–3165.
Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science* **318:** 1136–1140.
Grigoriadis A, Caballero OL, Hoek KS, da Silva L, Chen YT, Shin SJ, Jungbluth AA, Miller LD, Clouston D, Cebon J, et al. 2009. CT-X antigen expression in human breast cancer. *Proc Natl Acad Sci* **106:** 13493–13498.
Heard E, Rougeulle C, Arnaud D, Avner P, Allis CD, Spector DL. 2001. Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. *Cell* **107:** 727–738.
Huang S, Deerinck TJ, Ellisman MH, Spector DL. 1997. The dynamic organization of the perinucleolar compartment in the cell nucleus. *J Cell Biol* **137:** 965–974.
Huang KC, Rao PH, Lau CC, Heard E, Ng SK, Brown C, Mok SC, Berkowitz RS, Ng SW. 2002. Relationship of *XIST* expression and responses of ovarian cancer to chemotherapy. *Mol Cancer Ther* **1:** 769–776.
Kawakami T, Zhang C, Taniguchi T, Kim CJ, Okada Y, Sugihara H, Hattori T, Reeve AE, Ogawa O, Okamoto K. 2004. Characterization of loss-of-inactive X in Klinefelter syndrome and female-derived cancer cells. *Oncogene* **23:** 6163–6169.
Keohane AM, O'Neill LP, Belyaev ND, Lavender JS, Turner BM. 1996. X-Inactivation and histone H4 acetylation in embryonic stem cells. *Dev Biol* **180:** 618–630.
Kim HC, Choi KC, Choi HK, Kang HB, Kim MJ, Lee YH, Lee OH, Lee J, Kim YJ, Jun W, et al. 2010. HDAC3 selectively represses CREB3-mediated transcription and migration of metastatic breast cancer cells. *Cell Mol Life Sci* **67:** 3499–3510.
Kucera KS, Reddy TE, Pauli F, Gertz J, Logan JE, Myers RM, Willard HF. 2011. Allele-specific distribution of RNA polymerase II on female X chromosomes. *Hum Mol Genet* **20:** 3964–3973.
López-Soto A, Folgueras AR, Seto E, Gonzalez S. 2009. HDAC3 represses the expression of NKG2D ligands ULBPs in epithelial tumour cells: potential implications for the immunosurveillance of cancer. *Oncogene* **28:** 2370–2382.
Lyon MF. 1961. Gene action in the *X*-chromosome of the mouse (*Mus musculus* L.). *Nature* **190:** 372–373.
Maenner S, Blaud M, Fouillen L, Savoye A, Marchand V, Dubois A, Sanglier-Cianférani S, Van Dorsselaer A, Clerc P, Avner P, et al. 2010. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol* **8:** e1000276.
Marks H, Chow JC, Denissov S, Françoijs KJ, Brockdorff N, Heard E, Stunnenberg HG. 2009. High-resolution analysis of epigenetic changes associated with X inactivation. *Genome Res* **19:** 1361–1373.
Miao LJ, Huang FX, Sun ZT, Zhang RX, Huang SF, Wang J. 2014. Stat3 inhibits Beclin 1 expression through recruitment of HDAC3 in nonsmall cell lung cancer cells. *Tumour Biol* **35:** 7097–7103.
Müller BM, Jana L, Kasajima A, Lehmann A, Prinzler J, Budczies J, Winzer KJ, Dietel M, Weichert W, Denkert C. 2013. Differential expression of histone deacetylases HDAC1, 2 and 3 in human breast cancer—overexpression of HDAC2 and HDAC3 is associated with clinicopathological indicators of disease progression. *BMC Cancer* **13:** 215.
Nakagawa M, Oda Y, Eguchi T, Aishima S, Yao T, Hosoi F, Basaki Y, Ono M, Kuwano M, Tanaka M, et al. 2007. Expression profile of class I histone deacetylases in human cancer tissues. *Oncol Rep* **18:** 769–774.
O'Donovan PJ, Livingston DM. 2010. BRCA1 and BRCA2: breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair. *Carcinogenesis* **31:** 961–967.
Oehme I, Deubzer HE, Wegener D, Pickert D, Linke JP, Hero B, Kopp-Schneider A, Westermann F, Ulrich SM, von Deimling A, et al. 2009. Histone deacetylase 8 in neuroblastoma tumorigenesis. *Clin Cancer Res* **15:** 91–99.
Okamoto I, Patrat C, Thépot D, Peynot N, Fauque P, Daniel N, Diabangouaya P, Wolf JP, Renard JP, Duranthon V, et al. 2011. Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. *Nature* **472:** 370–374.
Pageau GJ, Hall LL, Ganesan S, Livingston DM, Lawrence JB. 2007. The disappearing Barr body in breast and ovarian cancers. *Nat Rev Cancer* **7:** 628–633.
Park SY, Jun JA, Jeong KJ, Heo HJ, Sohn JS, Lee HY, Park CG, Kang J. 2011. Histone deacetylases 1, 6 and 8 are critical for invasion in breast cancer. *Oncol Rep* **25:** 1677–1681.
Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. 2000. Molecular portraits of human breast tumours. *Nature* **406:** 747–752.
Perry M. 1972. Evaluation of breast tumour sex chromatin (Barr body) as an index of survival and response to pituitary ablation. *Br J Surg* **59:** 731–734.
Plath K, Talbot D, Hamer KM, Otte AP, Yang TP, Jaenisch R, Panning B. 2004. Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. *J Cell Biol* **167:** 1025–1035.
Popova T, Manie E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH. 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol* **10:** R128.
Popova T, Manié E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, Delattre O, Sigal-Zafrani B, Bollet M, Longy M, et al. 2012. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with *BRCA1/2* inactivation. *Cancer Res* **72:** 5454–5462.
Portela A, Esteller M. 2010. Epigenetic modifications and human disease. *Nat Biotechnol* **28:** 1057–1068.
Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S. 2006. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* **9:** 121–132.
Rivera MN, Kim WJ, Wells J, Driscoll DR, Brannigan BW, Han M, Kim JC, Feinberg AP, Gerald WL, Vargas SO, et al. 2007. An X chromosome gene, *WTX*, is commonly inactivated in Wilms tumor. *Science* **315:** 642–645.
Schenk T, Chen WC, Göllner S, Howell L, Jin L, Hebestreit K, Klein HU, Popescu AC, Burnett A, Mills K, et al. 2012. Inhibition of the LSD1 (KDM1A) demethylase reactivates the all-*trans*-retinoic acid differentiation pathway in acute myeloid leukemia. *Nat Med* **18:** 605–611.
Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, et al. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486:** 395–399.
Shen H, Laird PW. 2013. Interplay between the cancer genome and epigenome. *Cell* **153:** 38–55.
Silver DP, Dimitrov SD, Feunteun J, Gelman R, Drapkin R, Lu SD, Shestakova E, Velmurugan S, Denunzio N, Dragomir S, et al. 2007. Further evidence for BRCA1 communication with the inactive X chromosome. *Cell* **128:** 991–1002.
Simon MD, Pinter SF, Fang R, Sarma K, Rutenberg-Schoenberg M, Bowman SK, Kesner BA, Maier VK, Kingston RE, Lee JT. 2013. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* **504:** 465–469.
Sirchia SM, Ramoscelli L, Grati FR, Barbera F, Coradini D, Rossella F, Porta G, Lesma E, Ruggeri A, Radice P, et al. 2005. Loss of the inactive X chromosome and replication of the active X in BRCA1-defective and wild-type breast cancer cells. *Cancer Res* **65:** 2139–2146.
Smethurst M, Bishun NP, Fernandez D, Allen J, Burn JI, Alaghband-Zadeh J, Williams DC. 1981. Steroid hormone receptors and sex chromatin frequency in breast cancer. *J Endocrinol Invest* **4:** 455–457.
Spatz A, Borg C, Feunteun J. 2004. X-chromosome genetics and human cancer. *Nat Rev Cancer* **4:** 617–629.
Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJ, Zhu Y, Kaaij LJ, van Ijcken W, Gribnau J, Heard E, et al. 2011. The inactive X chromosome

adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev* **25:** 1371–1383.
Spurling CC, Godman CA, Noonan EJ, Rasmussen TP, Rosenberg DW, Giardina C. 2008. HDAC3 overexpression and colon cancer cell proliferation and differentiation. *Mol Carcinog* **47:** 137–147.
Turner NC, Reis-Filho JS. 2006. Basal-like breast cancer and the BRCA1 phenotype. *Oncogene* **25:** 5846–5853.
Vincent-Salomon A, Ganem-Elbaz C, Manié E, Raynal V, Sastre-Garau X, Stoppa-Lyonnet D, Stern MH, Heard E. 2007. X inactive–specific transcript RNA coating and genetic instability of the X chromosome in *BRCA1* breast tumors. *Cancer Res* **67:** 5134–5140.
Wutz A, Rasmussen TP, Jaenisch R. 2002. Chromosomal silencing and localization are mediated by different domains of *Xist* RNA. *Nat Genet* **30:** 167–174.
Xiao C, Sharp JA, Kawahara M, Davalos AR, Difilippantonio MJ, Hu Y, Li W, Cao L, Buetow K, Ried T, et al. 2007. The *XIST* noncoding RNA functions independently of BRCA1 in X inactivation. *Cell* **128:** 977–989.
Yildirim E, Kirby JE, Brown DE, Mercier FE, Sadreyev RI, Scadden DT, Lee JT. 2013. Xist RNA is a potent suppressor of hematologic cancer in mice. *Cell* **152:** 727–742.
Zhang J, Benavente CA, McEvoy J, Flores-Otero J, Ding L, Chen X, Ulyanov A, Wu G, Wilson M, Wang J, et al. 2012. A novel retinoblastoma therapy from genomic and epigenetic analyses. *Nature* **481:** 329–334.
Zink D, Fischer AH, Nickerson JA. 2004. Nuclear structure in cancer cells. *Nat Rev Cancer* **4:** 677–687.

Received October 28, 2014; accepted in revised form January 28, 2015.

The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer

Ronan Chaligné, Tatiana Popova, Marco-Antonio Mendoza-Parra, et al.

Genome Res. 2015 25: 488-503 originally published online February 4, 2015
Access the most recent version at doi:10.1101/gr.185926.114

Supplemental Information

The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer

Ronan Chaligné[1,2,3,8], Tatiana Popova[1,4], Marco-Antonio Mendoza-Parra[5], Mohamed-Ashick M. Saleem[5] , David Gentien[1,6], Kristen Ban[1,2,3,8], Tristan Piolot[1,7], Olivier Leroy[1,7], Odette Mariani[6], Hinrich Gronemeyer*[5], Anne Vincent-Salomon*[1,4,6,8], Marc-Henri Stern*[1,4,6] and Edith Heard*[1,2,3,8]

Extended Experimental Procedures

Cell Culture

Human Mammary Epithelial Cells (HMEC, Invitrogen) were grown in serum-free medium (HuMEC, Invitrogen). WI-38, ZR-75-1, SK-BR-3 and MDA-MB-436 cells were grown in Dulbecco's modified Eagle's medium (DMEM; Invitrogen) containing 10% fetal bovine serum (FBS).

DNA Methylation analysis.

We bisulfite-treated 2 µg of genomic DNA using Epitect bisulfite kit (Qiagen). Bisulfite converted DNA was amplified with bisulfite primers listed in Table S3. All primers incorporated a T7 promoter tag, and PCR conditions are available upon request. We analyzed PCR products by MALDI-TOF mass spectrometry after in vitro transcription and specific cleavage (EpiTYPER by Sequenom®). For each amplicon, we analyzed two independent DNA samples and several CG sites in the CpG Island. Design of primers and selection of best promoter region to assess (approx. 500 bp) were done by a combination of UCSC Genome Browser (http://genome.ucsc.edu) and MethPrimer (http://www.urogene.org). All the primers used are listed (Table S3). NB: MAGEC2 CpG analysis have been done with a combination of two CpG island identified in the gene core.

Analysis of RNA allelic expression profiles (based on Human SNP Array 6.0)

DNA and RNA hybridizations were normalized by Genotyping console. Based on Log2ratios and Allelic Differences of DNA profile absolute segmental copy numbers were inferred. Single SNP raw intensities were considered from RNA profiles. Allelic expression scores were calculated based on both DNA and RNA profiles, as follows:

1. RNA expression of SNP (log(Signal A +Signal B)) was smoothed using a 5 SNPs sliding window and excluding low expressed SNPs (log(Signal A +Signal B)<6.5). Each SNP expression score was calculated by subtracting median and normalizing by standard deviation of expression level shown by exonic and intronic SNPs. Parameters were chosen based on comparisons to mRNA expression profile measured by Affymetrix 133plus2 array.
2. Each SNP homozygosity score was calculated based on the inferred segmental copy number and major allele counts. If the segment was annotated to have a homozygous allelic status, all SNPs from the segment were annotated to be homozygous. If the segment was annotated to have a heterozygous allelic status, Allelic Difference was centered to the median of heterozygous band and normalized by standard deviation of heterozygous band. Homozygosity score of each SNP was set to centered and normalized Allelic Difference.
3. Each SNP allelic expression ratio score was calculated as 2tangent(α) where α corresponded to the angle defined by (Signal A; Signal B) vector (α = 2*arctan(Signal A/Signal B)-pi/2). Balanced allelic expression corresponded to 0.
4. Based on three thresholds: (1) Total expression score, (2) Heterozygous DNA call, (3) Allelic expression ratio score, all SNPs were classified into 6 groups:
    1) No expression or non-informative: designated by 0
    2) Contradictory 1 (homozygous SNP and bi-allelic expression): designated by -1
    3) Contradictory 2 (homozygous AA SNP and mono-allelic BB expression or vice versa): designated by -2
    4) Mono-allelic expression: designated by 1
    5) Bi-allelic expression: designated by 2
    6) Marginal call (in-between mono-allelic and bi-allelic): designated by 1.5

    Further analysis and quality controls showed good correspondence between attributions and low number of contradictory calls.
5. Profile of bi-allelic expression was obtained by summarizing bi-allelic expression calls in a sliding window of 50 SNPs.

5[bis]. Single SNP classification was summarized on the gene level, and each gene allelic expression has been score based on:

1) On the number of informative SNPs in the gene core (belonging to group 4, 5 or 6, see above)

2) On confidence attributed to the SNPs (depending on the distance from the threshold)

3) Consistence between SNPs within the same gene

Sanger sequencing, real-time PCR, allele-specific PCR and pyro-sequencing

Total RNA was isolated from cells using Trizol Reagent (Invitrogen) and purified on columns combined with DNase-treated (Qiagen) to remove contaminating DNA. First-strand cDNA was prepared from 5 µg of RNA and random hexamers using Superscript III (Invitrogen) at 50°C for 1 hr. gDNA was isolated using DNazol reagent (Invitrogen). cDNA and gDNA genotyping status (i.e. single variable position) was determined: either by a real-time PCR single nucleotide polymorphism (SNP) detection system with fluorescent competitive probes using an ViiA7 analyzer (Applied Biosystems), by Sanger sequencing (3130xl Genetic Analyzer, Applied Biosystems) of purified PCR product using BigDye V3.1 kit as recommended by the provider (Applied Biosystems) or by pyrosequencing as recommended by the manufacturer (Qiagen, Pyromark Q24). All real-time PCR reactions used SybrGreen Master Mix (Applied Biosystems) to a final volume of 10 µl. Each sample was analyzed at least in triplicate. All the primers used are listed (Table S3).

RNA sequencing analysis

We performed RNA-sequencing and DNA exome-sequencing on ZR-75-1, SK-BR-3 and MDA-MB-436 cell lines. The RNA-sequencing correspond of paired-reads lane 2x100 bp sequencing on poly-A RNA purified. The DNA exome-sequencing has been done by paired-reads lane 2x100 bp after a SureSelect® array-capture. Both sequencing has been performed on high-throughput Illumina HiSeq sequencer. Burrows-Wheeler Aligner (BWA) was used for the mapping. Briefly, SNPs were called from DNA exome-seq data using The Genome Analysis Toolkit (GATK, Broad Institute) and dbSNP database. For allelic expression analysis, we then only kept SNPs supported by >= 10 RNA-seq reads. For each gene, when several SNPs were informative, we assessed the allelic expression from the most informative SNP. We next calculated the allelic expression ratio as (100-(absolute(Allele A%-Allele B%))). Genes above 40 are considered as bi-allelically expressed and genes below 40 are categorized as mono-allelically expressed. We noted that remarkably fewer X-linked genes were retained in SK-BR-3, compared to the other

samples. This is likely due to the large region of LOH on the long arm of the X-chromosome in this cell line. NB: We chose to keep the same threshold (of 40) for all cell lines, knowing that we would presumably underestimate the degree of Xi-gene reactivation for the ZR-75-1 line which is trisomic for the X-chromosome. Indeed, when the threshold is change for ZR-75-1 to account for ploidy, this would have led to just one more gene being included as "cancer-specific" escapee: *ARHGEF9*. We therefore chose to keep the same threshold of 40 for all lines, in order to simplified data presentation without impacting on the general conclusions.

Chromatin Immunoprecipitation analysis

Chromatin immunoprecipitation assays and massive parallel sequencing

Cancer and normal cells were fixed with 1% para-formaldehyde during 30 minutes at room temperature. Chromatin from fixed cells was fragmented by sonication and immunoprecipitated in lysis buffer (50mM TrisHCl pH=8, 1mM EDTA, 140mM NaCl, 1% Triton, 0.1% Na-deoxycholate) complemented with protease inhibitor cocktail (Roche cat# 11873580001). After overnight immunoprecipitation in presence of the corresponding antibodies, 2 washes with lysis buffer, 2 washes with lysis buffer containing 360mM NaCl, 2 washes with washing buffer (10mM TrisHCl pH=8, 250mM LiCl, 0.5% NP-40, 1mM EDTA, 0.5% Na-deoxycholate) and 2 washes with 1xTE were performed before chromatin elution at 65°C (15 min in elution buffer: 50mM TrisHCl pH=8, 10mM EDTA, 1% SDS). The immunoprecipitated chromatin was decrosslinked overnight (65°C in presence of 1%SDS; 1xTE solution), the remaining proteins were removed by proteinase K treatment (Roche; cat# 03115852001) and phenol-chloroform extraction. The purified DNA was validated by quantitative real-time PCR (qPCR, Roche LC480 light cycler device; Qiagen Quantitect PCR reagents) and libraries for massive parallel sequencing were prepared following standard procedures (NEXTflexTM ChIP-seq library kit; cat# 514120). Chromatin immunprecipitation assays were performed with antbodies directed against RNA Polymerase II (Santa Cruz; sc-9001; H-224), H3K4me3 (Abcam; ab8580) and H3K27me3 (Millipore; ab07449).

ChIP-seq libraries were prepared according to the standard Illumina protocol and sequenced with the HiSeq2500 system. Single end sequencing was carried out to obtain around 25-150 million (M), 100bp long reads per sample.

Alignment and Quality control

Datasets were subjected to two types of quality control. FASTQC-0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to assess the quality of sequencing and potential adapter or cross contaminants. Average sequencing quality (phred score) per base was above 30 ($Q \geq 30$) for all datasets. In addition, aligned datasets were then subjected to NGS-QC ((Mendoza-Parra et al. 2013); www.ngs-

qc.org) to assess the robustness of enrichment. The majority of data sets were of «triple A» quality, no data set was below «triple B».

For exome-seq and ChIP-seq, alignment was performed using BWA-MEM-0.7.7 (Li and Durbin 2009) with default parameters, which simultaneously checks for both global and local alignment for reads. Alignment was followed by three sets of filters to prevent bias in the analysis. 1) duplicate reads (PCR clonal reads) were filtered out using Picard tools-1.86 (http://picard.sourceforge.net) 2) reads with mapping quality less than 10 were filtered out using Bamtools-2.2.3 (https://github.com/pezmaster31/bamtools) and 3) reads with more than one alignment reported were filtered out using in-house scripts. Further analysis was carried out on processed alignment file which is around 10-100M reads after these filters.

ChIP Allele specific analysis

To prepare the allele information for each cell-line for allele-specific analysis down the line, SNP analysis was carried out along with Human SNP Array 6.0 data. To identify novel variation (apart from known SNPs from Human SNP Array 6.0 data), SNP analysis was carried out on all three ChIP-seq, Exome-seq and RNA-seq data individually. ChIP-seq data of different marks (H3K4me3, H3K27me3, RNA Pol II and Input) were merged for each cell line to increase the depth and confidence for variation calling. Variation calling was performed for each cell line separately for ChIP-seq and Exome-seq following the best practice GATK-2.6.5 pipeline by filtering reads with Mapping quality ≥ 1 (Van der Auwera GA et al., 2013; http://www.broadinstitute.org/gatk/guide/best-practices?bpm=DNAseq). Variation calling for RNA-seq was carried out following the methods of Piskol, Robert et al., 2013 to avoid artifacts specific to RNA-seq data (Piskol et al. 2013). A final list of allele information was generated by combining the SNP information from the different data sets for each cell line. To increase the allele-specific sensitivity for the alignment, reads were additionally realigned in an allele-specific manner following the method of Satya et al. (Satya et al. 2012). Read counts for each allele and SNP position were extracted for each mark using in-house scripts. SNP positions with at least three reads from both alleles were considered as heterozygous positions.

Peak calling and annotation

Peak calling was performed using HOMER ((Heinz et al. 2010); http://homer.salk.edu/homer/index.html) with default parameters. For H3K27me3 and RNA Pol II, the 'style' parameter was chosen as 'histone' due to the broad patterns for this mark, whereas for H3K4me3, which generally give sharp peaks, the parameter 'factor' was chosen. Genomic context annotation over identified peaks were carried out using the HOMER annotation module but with basic annotation by excluding references other than coding genes and non-coding RNA.

Integration of annotations

A gene-based analysis of annotation integration was carried out using in-house scripts to integrate all annotation from peak and variation calling (informative SNP counts, read depths, homo/heterozygous SNP count and weighted allelic imbalance). To include annotations from regulatory regions 1Kb sequences upstream from the TSS and downstream of the TES were considered. A weighted arithmetic average was calculated for each gene by calculating average Allelic Imbalance (AI) where each SNP's AI was weighted by its read depth.

Normalization of ChIP-seq data

To illustrate the comparisons across cell lines, ChIP-seq data were normalized using an in-house developed tool called 'Epimetheus', which is based on quantile normalization (manuscript under preparation). Read Count Intensity (RCI) was calculated for a window of 100bp bin size across chromosomes and then these intensities were normalized using quantile normalization from the limma package. The impact of normalization was assessed using MA plots before and after normalization. Specified genomic feature based normalized RCI was constructed, which are illustrated in Figure 6. For TSS-centered plots and heatmaps, a separate TSS-based normalization was carried out with 20bp bin size to obtain higher resolution.

RNA, DNA FISH and immunofluorescence.

For DNA FISH, cells were denatured in 50% formamide/2X SSC at 80°C for 30 min and rinsed several times in cold 2X SSC prior to overnight hybridization at 42°C. Labeled BAC probes were denatured and competed with Cot-1 DNA (3 µg/coverslip) for 15 min at 37°C. Preparation of the X chromosome paint probe was performed according to the supplier's instructions (CytoCell). After hybridization, coverslips were washed three times in 50%formamide/ 2X SSC and three times in 2X SSC at 42°C for RNA-FISH and DNA-FISH, and then stained with DAPI (0.2 mg/ml). Immunofluorescence RNA-FISH was performed as described previously (Chaumeil et al. 2008). For immunofluorescence, the following antibodies were used: RNA polymerase II (clone CTD 4H8; Millipore cat# 05-623), H3K9Ac (Millipore cat# 06-942), H4Ac (Millipore cat# 06-946), H3K27me3 on interphase (clone 7B11), H3K27me3 on metaphase (ActiveMotif cat# 39155), H3K4me2 (Millipore cat# 07-030) at a 1/200 dilution. For RNA FISH on tumors, either tumor stamps were generated from fresh tissue samples and immediately frozen at -80°C; or else cryosections (10µm thick) were generated. Just prior to IF / FISH, these stamps/sections were fixed in 3% paraformaldehyde/PBS for 10 min, then permeabilized in 1X PBS/0.5% Triton X-100/2 mmol/L vanadyl ribonucleoside complex (New England Biolabs) on ice for 4 min. After three washes in PBS, the sample was

dehydrated through an ethanol series of washes prior to RNA FISH or DNA FISH, which was then performed as described previously (Vincent-Salomon et al. 2007). For immunofluorescence alone, samples were used directly, without prior ethanol treatment and dehydration.

Sequential Immunofluorescence and RNA FISH analysis with super resolution OMX® microscopy

Sequential Immuno-RNA FISH was performed as previously described (Chaumeil et al. 2008). Antibodies used for immuno-staining were: anti-H3K9Ac (Millipore cat# 06-942), anti-H3K27me3 (clone 7B11), anti-H3K4me2 (Millipore cat# 07-030) and anti-RNA polymerase II (clone CTD 4H8; Millipore cat# 05-623). Structured illumination image acquisition was carried out using a DeltaVision OMX version 3 system (Applied Precision, Issaquah, WA) coupled to three EMMCD Evolve cameras (Photometrics, Tucson, AZ). Multi-channel image alignment was performed using ImageJ (Rasband, W.S., ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/, 1997-2012) and UnwarpJ plugin (Sorzano et al. 2005). At least, thirteen nuclei were analyzed for each experiment.

Data processing of 104 Basal-like breast carcinomas (BLC)

We obtained controlled access to the EGA datasets from the study EGAS00001000132 including Human SNP Array 6.0-arrays, RNA exome and whole genome sequencing data (Shah et al. 2012). Human SNP Array 6.0-arrays were processed using the GAP method to obtain absolute copy number and allelic content profiles (Popova et al. 2009). Samples that were classified as bad, average quality or contaminated by normal tissue were discarded. In details, we obtained 81 primary BLC tumors with RNA-seq and SNP-array data available. 39 samples classified as “bad” quality were removed from the analysis: 2 cases were identified as “normal”; 7 cases were identified with bad quality hybridization to SNP-arrays and 30 cases were identified with more than 50% contamination by the normal tissue, which showed low signal to noise ratio in SNP-array copy number and allelic imbalance profiles. Finally, 42 BLCs with good SNP-array quality (i.e. <50% contamination by the normal tissue and high signal to noise ratio) and RNA-seq data available were analyzed (cf Table S2). After evaluation of allelic status of X chromosome, we ended up with 25 samples exhibiting heterozygosity, of at least some region of the X chromosome for further evaluation. Allelic

expression was obtained as the number of reads and corresponding allelic frequency covering known SNP positions. SNPs coverage was obtained based on the SAMtools pileup processing of RNA-seq data (Li et al. 2009).

Supplemental Figure Legends

Figure S1. Xi characterization in normal and cancer cells

(A) Whole X chromosome DNA FISH (X paint in white). Continue of the Figure 1A.

(B) *XIST* expression level assessed by real-time PCR. Normalization was performed using *TBP* expression levels (Kwon et al. 2009). Data represent the mean values +/- SEM.

(C) *XIST* RNA FISH signal intensity. Quantification was performed using ImageJ software (NIH, Betesda) by quantifying FISH signal at *XIST* RNA domain (we normalized *XIST* intensity signal to the general RNA FISH background that we measured in proximity to the Xa identify by using HDAC8 RNA FISH). The number of nuclei analyzed is indicated at the bottom of the box plot.

(D) DAPI signal intensity. Quantification was performed using ImageJ software by comparing the DAPI signal at the *XIST* RNA domain versus DAPI signal associated with the *HDAC8* RNA FISH signal at the Xa. The number of nuclei analyzed is indicated at the bottom of the box plot.

(E) Example of DAPI signal intensity quantification on HMECs and ZR-75-1 cells.

(F) X-paint DNA FISH on metaphase spreads (grey) shows the number of X chromosomes (red) in normal and breast cancer cell lines.

(G) *KDM5C* and *HUWE1* expression at *XIST* domain assessed by nascent transcript RNA FISH (*KDM5C*, red; *HUWE1*, green; *XIST, grey)*

Left Bar chart: Quantification (in %) of nuclei showing mono- or bi-allelic expression of *HUWE1* associated with *XIST* RNA domain.

Right Bar chart: Quantification (in %) of nuclei showing mono- or bi-allelic expression of *KDM5C* associated with *XIST* RNA domain.

(H) Boxplot of the relative levels of *XIST* RNA coating and Cot-1 RNA exclusion. The number of nuclei analyzed is indicated at the bottom of the box plot. Examples of nuclei used for this analysis are shown in figures 1C. The quantification has been done by ImageJ software on images acquired on a Nikon confocal spinning disk microscope. For details on quantification method see Figure S2A and S2C.

(I) Visual quantification of presence or exclusion of Cot-1 RNA at *XIST* RNA domain. The number of nuclei analyzed is indicated at the bottom of the bar chart. Quantification has been done on images acquired on a Nikon confocal spinning disk microscope.

(J) Boxplot of the relative levels of *XIST* RNA coating and RNA Pol II exclusion. The number of nuclei analyzed is indicated at the bottom of the box plot. Examples of nuclei used for this analysis are shown in figures 1C. The quantification has been done by ImageJ software on images acquired on a Nikon confocal spinning disk microscope. For details on quantification method see figure S2A and S2C.

(K) Visual quantification of RNA Pol II presence or exclusion at *XIST* RNA domain. The number of nuclei analyzed is indicated at the bottom of the bar chart. Quantification has been done on images acquired on a Nikon confocal spinning disk microscope.

Box plot on this figure: Upper whisker represents 90%, upper quartile 75%, median 50%, lower quartile 25% and lower whisker 10% of the dataset for each cell line.

(***) $p< 0.001$, (**) $p<0.01$, (*) $p<0.05$ using the Student's *t*-test. All the dataset are compared with HMEC dataset.

Scale bar: 5µm

Figure S2. Histone post-translational modifications associated with the X chromosome

(A) Schematic view of the quantification done by ImageJ home-made macro to evaluate the enrichment / depletion of immuno-staining (or RNA FISH) signal at *XIST* domain compare to the non-*XIST* coated DNA. This quantification has been carried-out on immuno-RNA FISH images (see Figure 1C for example). Nucleoli have not been considered in the analysis. For MDA-MB-436 cells, to avoid bias in the quantification, we also excluded from the analysis the highly H3K27me3 enriched bodies which do not belong to X chromosome (nor in metaphase (Figure 3C) or in interphase (Figure S3F)). NB: Evaluation of the H3K27me3 enrichment on the inactive X chromosome was also performed using an X-chromosome DNA FISH paint (Figure S3F) to identify the Xi territory. This revealed the same changes in H3K27me3 enrichment in cancer cell lines as when *XIST* was used as a read out for the Xi.

(B) Example of results obtained with ImageJ macro in the evaluation of H3K27me3 enrichment at *XIST* domain.

(C) Example of results obtained with ImageJ macro in the evaluation of Cot-1 RNA depletion at *XIST* domain.

(D) Pearson's co-localization coefficients have been evaluated for *XIST* RNA coating and H3K27me3 association on DeltaVision OMX microscope. Examples of nuclei used for the analysis are shown in Figures 2C. Data represent the mean values +/- SEM. The number of nuclei analyzed is indicated at the bottom of the box plot.

(E) Immuno-RNA FISH revealing the degree of H4ac (green) depletion at *XIST* RNA domains (red).

(F) Boxplot of the relative levels of *XIST* RNA coating and H4ac association. The number of nuclei analyzed is indicated at the bottom of the box plot. The quantification has been done by ImageJ software on images acquired on a Nikon confocal spinning disk microscope. For details on quantification method see Figure S2A and S2C.

(G) Immuno-RNA FISH revealing the degree of H3K4me2 (green) depletion at *XIST* RNA domains (red).

(H) Boxplot of the relative levels of *XIST* RNA coating and H3K4me2 association. The number of nuclei analyzed is indicated at the bottom of the box plot. The quantification has been done by ImageJ software on images acquired on a Nikon confocal spinning disk microscope. For details on quantification method see Figure S2A and S2C. Similar results have been obtained by immuno-RNA FISH for H3K4me3 and *XIST* RNA (Data not shown).

(I) Visual quantification at *XIST* RNA coating of H3K27me3 enrichment; H3K9ac exclusion; H3K4me2 exclusion and H4ac exclusion. The number of nuclei analyzed is indicated at the bottom of the bar chart. Examples of nuclei used for this analysis are shown in figures 2A, 2D, S2E and S2G. Quantification has been done on images acquired with Nikon confocal spinning disk microscope.

(J) Immuno-RNA-FISH for *XIST* (red) and H3K4me2 (green). Acquisition has been carried out by super-resolution structured illumination on DeltaVision OMX microscope. Inset for H3K4me2, *XIST* RNA and merge is shown below each cell lines.

(K) Immuno-RNA FISH for *XIST* (red) and RNA Pol II (green). Acquisition has been carried out by super-resolution structured illumination on DeltaVision OMX microscope. Inset for RNA Pol II, *XIST* RNA and merge is shown below each cell lines.

(***) $p< 0.001$, (**) $p<0.01$, (*) $p<0.05$ using the Student's *t*-test. All the dataset are compared with HMEC dataset.

Box plot on this figure: Upper whisker represents 90%, upper quartile 75%, median 50%, lower quartile 25% and lower whisker 10% of the dataset for each cell line.

Figure S3. Nuclear organization of the *XIST* RNA coated chromosome

(A) Genomic DNA Sanger sequencing of *XIST* provides genotype information for at least one SNP in each tumoral cell lines. cDNA sequencing reveals that *XIST* is mono-allelically expressed in the three tumor cell lines.

(B) Bar chart of the percentage of nuclei observed as EdU positive or EdU negative. Briefly, we performed a 30min EdU pulse on cultured cells and then detected incorporated Edu using the Click-It assay (Invitrogen). At least 100 nuclei were analyzed for each cell line. Indeed, we wondered whether the perturbed state of the Xi in some cells might be dependent on cell cycle, for example due to S phase when chromatin must be replicated. Thus, we further examined MDA-MB-436 cells, which showed the highest proliferation rate of the three cancer cell lines.

(C) The MDA-MB-436 cell line was used for a sequential IF / RNA FISH (EdU pulse and detection / H3K27me3 immuno-staining / *XIST* RNA FISH). Using ImageJ macro, we quantified the degree of H3K27me3 enrichment in EdU positive and EdU negative cells to explore the impact of cell cycle on Xi epigenetic chromatin mark instability in tumoral cell lines. The level of H3K27me3 enrichment is slightly lower in this experiment compare to Figure 2B, presumably due to a slight decrease in immuno-staining quality following EdU “Click-It” detection. No particular correlation could be seen between EdU positive (S phase) or EdU negative (G1 or G2 phase) cells and disrupted H3K27me3 enrichment on the Xi , indicating that the disrupted chromatin patterns observed are not necessarily linked to a specific stage of the cell cycle such as S phase.

(D) Example of H3K27me3 signal intensity quantification on EdU negative or EdU positive MDA-MB-436 cells (Immuno-RNA FISH: DNA, blue; *XIST* RNA, red; EdU, green and H3K27me3, white).

(E) Immuno-blotting was performed on protein nuclear extracts prepared as follows. After washing with PBS, cells were incubated on ice for 10 minutes in buffer A (10 mM HEPES pH 7.8, 10 mM KCl, 2 mM MgCl2, 0.1 mM EDTA) with protease inhibitor cocktail, added 10% NP40, and centrifuged at 14,000 rpm for 20 seconds. The supernatant was removed. The pellet was suspended in buffer B (50 mM HEPES pH 7.8, 50 mM KCl, 300 mM NaCl, 0.1 mM EDTA, protease inhibitor cocktail) and incubated on ice for 30 minutes. Nuclear debris was pelleted by centrifugation at 14,000 rpm for 10 minutes, and the supernatants were used as nuclear extracts. Nuclear proteins (20µg) were separated on SDS–polyacrylamide gel electrophoresis and transferred to polyvinylidene difluoride (PVDF) membranes by a standard procedure. Antibodies used for immunoblotting were: H3K27me3 (ActiveMotif cat# 39155), H3K4me2 (Millipore cat# 07-030), H4ac (Millipore cat# 06-946), H3ac (Millipore, cat# 06-599), H3 (Abcam, cat# 1791), Lamin A/C (Millipore, cat# 05-714). Immunoblots were revealed using enhanced chemiluminescence (ECL+, Amersham). Histone 3 and Lamin A/C are used as loading normalization.

(F) Sequential immuno-RNA/DNA FISH was performed as described (Chaumeil et al. 2008). Briefly, staining and images acquisition were first carried out and positions saved. Slides were then treated with RNase A and RNase H for 1h at 37°C. After several washes, slides were used for X chromosome paint DNA FISH. At least, 80 nuclei were analyzed for each cell lines. Grey and red drawing outlines of *XIST* RNA, H3K27me3 and H3K4me2 panels represent the X chromosome territories (from the X chromosome panel). For each cell lines either one or two planes are shown.

Figure S4. Validation of X-linked genes allelic expression status

(A) Schematic outline of our allele-specific transcriptional analysis of X-chromosome transcriptional activity.

(B) Dilution- limited cultures enabled us to derive 22 independent clones from the primary WI-38 cell line. Analysis of each clone by allele-specific PCR reveals a clear mono-allelic expression of *CLCN4* from either the maternal or paternal X chromosome (eleven from each origin were obtained). Amongst this 22 clones, we then chose two clones, displaying inactivation of one or the other X, for RNA-seq and two further clones (again with inactivation of opposite alleles) for the RNA SNP6 approach. Similarly, we also derived clones from HMEC cells by dilution-limit culture. However the cloning efficiency of HMEC cells was far lower than for WI-38 cells. Only two HMEC clones were obtained, both with very limited cell proliferation capacity and were therefore not used for the RNA SNP6 array or RNA-seq analysis but only for allele-specific PCR. Each clone shows a clear mono-allelic expression of *NXT2* from maternal or paternal chromosome (data not shown).

(C-E) We also derived single cell clones from the ZR-75-1, SK-BR-3 and MDA-MB-436 cells. Each tumor cell line revealed the same Xi/Xa allelic profile in all clones analyzed and the parental bulk cell line, as expected if these tumors (and the cell lines derived from them) were originally clonal, unlike WI-38 and HMEC primary cells which are polyclonal, with a mixed population of cells harboring a maternal or paternal inactivated X chromosome. For example, *NXT2* allelic expression is strictly the same between the bulk population and the clones for ZR-75-1 (mono-allelic expression) and MDA-MB-436 (bi-allelic expression).

(F) For the allele-specific transcriptome experiment we used two different WI-38 clones (with alternative paternal Xi / maternal Xi profiles). The allele-specific expression profile obtained on autosomal genes are almost the same in both clones. This demonstrates the robustness and accuracy of the approach and the fact that both clones present a similar allelic expression pattern on autosomes.

(G) WI-38 clone #1 harbors an inactive X chromosome of different parental origin compared to clone #28, however the allelic expression patterns observed were very similar between the two clones revealing that at this level of resolution

there are no striking differences in X-chromosome inactivation status between the maternal and paternal X chromosomes.

(H-I) Allele-specific PCR using TaqMan® probes for the analysis of *HDAC8* rs5912136 (H) and *APOOL* rs4828121 (I). The x axis show expression from allele A and the y axis expression from allele B. RT minus sample was used as negative control (i.e. no amplification of alleles A and B). To determine threshold for 100% of allele A or 100% allele B, we used pure gDNA material. For example, qRT-PCR with TaqMan® probes demonstrated that *APOOL,* is mono-allelically expressed in WI-38 cells, but is bi-allelically expressed in tumor cell lines with similar expression levels from inactive and active alleles in MDA-MB-436 cells, and lower levels (about 30%) from the inactive allele in ZR-75-1 cells.

(J) Genomic DNA Sanger sequencing of several genes provides genotype information for at least one SNP. cDNA sequencing reveals whether the gene expression is mono- or bi-allelic. For example, cDNA and gDNA Sanger sequencing on *SYTL4* reveal an mono-allelic expression in normal WI-38 cells and in the ZR-75-1 cell line, but an bi-allelic expression in MDA-MB-436 cells, (and uninformative in SK-BR-3 cells due to LOH).

Figure S5. Transcriptional activity of the X chromosome

(A) Allele specific PCR based on TaqMan® probes for analysis of *NXT2* rs3204027. The x axis shows expression from allele A and y axis expression from allele B. RT minus sample was used as negative control (i.e. no amplification of alleles A and B). To determine threshold for 100% of allele A or 100% allele B, we used pure gDNA material.

(B) Cancer Testis (C/T) antigen mRNA expression analysis. We investigated expression of several X-linked members of the C/T antigen family. This was performed by normalizing data to *TBP* expression for each sample and then reported to HMEC expression to evaluate expression increased in breast cancer cell lines. These genes are normally only expressed in the testis and are silent in somatic tissues, on both the active and inactive X chromosomes, but have been reported to be over-expressed in breast tumors. We found that some of these genes showed no expression at all (*MAGEA12, SAGE1, XAGE3*, data not shown) in all cell lines examined, while others (*MAGEA4, MAGEA6* and *MAGEC2*) showed increase expression in the cancer cell lines but not in normal cells. The aberrant expression of X-linked C/T antigens could either be due to reactivation on the Xi or the Xa. In the case of SK-BR-3, only the active X chromosome alleles, of those three genes, are present (due to LOH) meaning that the over-expression we observed must be due to the re-activation of the alleles on the active X chromosome. To detect *MAGEA6* expression in the other cell lines, RNA FISH was used. Data represent the mean values +/- SEM.

(C) *MAGEA6* expression analysis by RNA FISH on normal and tumoral cell lines (green). HDAC8 (red) and *XIST* (grey) RNA FISH has been used as control to localize Xi and Xa region within the nucleus. In normal cells (HMEC and

WI-38), no expression of *MAGEA6* was found. *MAGEA6* expression could be detected from the active X in a significant proportion of SK-BR-3 (47%) and ZR-75-1 (32%) cells, but never from the Xi. In MDA-MB-436 cells, as the *MAGEA6* loci are associated with an X chromosome fragment that is no longer linked to the XIC, it was not possible to determine from which allele the gene was expressed.

(D) Z-projections of 3D RNA FISH show representative examples of HDAC8 expression (green) at XIST domains (grey) in normal (WI-38 and HMEC) and breast cancer cell lines (ZR-75-1, SK-BR-3, and MDA-MB-436). In SK-BR-3 cells, arrowheads indicate active X chromosomes and arrows XIST-coated chromosomes. On the right, bar graph shows levels of HDAC8 expression from XIST domains, with reactivation in SK-BR-3 and MDA-MB-436 cells.

(E) *ATRX* expression assessed by nascent transcript RNA FISH on DAPI-stained nuclei (ATRX, green; *XIST*, grey; DNA, blue).

(F) *TBL1X* expression level assessed by real-time PCR. Normalization was performed using *TBP* expression level. Data represent the mean values +/- SEM.

(***) p< 0.001, (**) p<0.01, (*) p<0.05 using the Student's *t*-test. All the dataset are compared with HMEC dataset.

Figure S6. Perturbation of the Xi chromatin landscape in breast cancer cells

(A) Gene promoter DNA methylation analysis. Each histogram indicates the ratio of promoter methylation (0 to 1) according to gene and to cell line. The position of the gene is indicated on the X chromosome. Color code indicates the known allelic expression status on the Xi for each gene in different cell lines (subject to XCI, blue; escape from XCI, red; LOH i.e. no locus on Xi, brown; unknown; black). Data represent the mean values +/- SEM. Primers used for analysis by EpiTYPER are available in Table S3. DNA methylation levels of X-linked promoters examined were consistent with the Xa:Xi chromosome ratios in different cell lines. For example, in ZR-75-1 a general reduction in DNA methylation of X-linked gene promoters was found compared to HMECs, consistent with the presence of two Xa versus one Xi. This was most pronounced for regions presenting LOH, where only the Xa allele is present.

(B) Variation of H3K27me3 signals of 1Mb bins along the chromosome 17 between HMEC and either Wi-38 or the three tumor cell lines. Above 0 mean more enrichment in HMEC cells, and below 0 more enrichment in the cell line used in the comparison. The profile appears much more variable in the three tumor cell lines than by comparing HMEC to WI-38.

(C) UCSC Genome Browser (Kent et al. 2002) whole X chromosome view of H3K4me3 and H3K27me3 ChIP-seq data. Our data have been normalized (see Materials and Methods for more details). The quality of ChIP-seq data sets was validated with the NGS-QC Generator and received QC Stamps between “triple A” and “BAA” ((Mendoza-Parra et al. 2013); www.ngs-qc.org). In addition to our own HMEC data, HMEC profiles for H3K4me3, H3K27me3 and H3K9me3 were obtained from ENCyclopedia Of DNA Elements (ENCODE) project and were used to : 1- compare the ChIP-seq quality of our dataset; 2- refine the position of the two distinct chromatin type identified on the Xi in normal human cells : H3K9me3 (green) or H3K27me3 (red) enriched. At the bottom, the X chromosome schematic view highlights the H3K9me3 (green) or H3K27me3 (red) enriched domain. The percentages correspond to the frequency of detection of those particular regions in human cells by immuno-staining on metaphase by Chadwick, B (Chadwick 2007). Asterisk indicates a preferentially H3K9me3 enriched region which has not been observed by immuno-staining on metaphase by Chadwick, B, but which is clearly visible by ChIP-seq analysis on HMEC likely due to the higher resolution.

(D) TSS-centered plots for RNA Pol II and H3K4me3 enrichment of X-linked subject to XCI or escaping from the XCI. For each cell lines, X-linked subject to the XCI or escaping from the XCI have been choose based on RNA-seq and RNA SNP6 analyzed done previously (cf Table S1) (but for *HMEC, as we do not have allelic expression analysis available, we used X-linked genes list obtained from analyzing WI-38 clones). Genes escaping the XCI show higher enrichment of RNA Pol II and H3K4me3 at the TSS region, indicating that expression of an additional copy is sufficient to observe enrichment increase. Furthermore, in ZR-75-1 cells, there two active X and one inactive X chromosome meaning that we are still detecting expression of one additional copy out of three (even though the difference is reduced compare to the other cell lines).

(E) Heatmap for enrichment of RNA Pol II and H3K4me3 at the TSS region. We represented heights X-linked: four are escaping XCI in all the cell lines (*KDM6A*, *KDM5A*, *RPS4X* and *SMC1A*) and four are silenced on the Xi in all the cell lines (*RBM41*, *DLG3*, *HUWE1* and *RRAGB*).

Figure S7. Local perturbation of the Xi chromatin in breast cancer cells

(A) TSS-centered plots (+/- 1.5kb) show RNA Pol II and H3K4me3 enrichment for all the X-linked genes (except regional Xi loss) of each tumoral cell lines and HMEC. The number of genes analyzed is indicated below each plot.

(B) Heatmap for RNA Pol II and H3K4me3 enrichment at TSS +/- 1,5 kb. Genes listed are “cancer-specific” escapees (in respect to each of the three tumoral cell lines; see figure 6D to have the averaging of the TSS region enrichment).

(C) Example of H3K27me3, RNA Pol II and H3K4me3 Allelic Imbalance (AI) enrichment for known escaping or silenced X-linked genes in the three tumoral cell lines. AI has been calculated, for a given gene, by the weighted arithmetic mean of all the informative SNPs lying within the gene body (+1kb before TSS and +1kb after the TES). As for RNA-seq analysis, AI < 40 is considered as a mono-allelic enrichment (i.e. < 20% enrichment for the lowest enriched allele) and above AI > 40 as bi-allelic enrichment. X-linked genes escaping the XCI show an enrichment of H3K4me3 and RNA Pol II on the active and inactive allele. At the contrary, H3K27me3 is only detected enriched on one allele. X-linked subject to the XCI show mono-allelic enrichment of H3K27me3, RNA Pol II and H3K4me3.

(D) Allelic imbalance (AI) of H3K27me3, RNA Pol II or H3K4me3 enrichment is shown for “cancer-specific” escapees in MDA-MB-436 cells. AI for a given gene represents the weighted mean of informative SNPs lying within the gene body +/- 1kb. As for RNA-seq analysis, AI < 40 is considered as a mono-allelic enrichment.

Figure S8. Assessment of the epigenetic status of the inactive X chromosome in primary breast tumors

(A-D) Examples of RNA FISH for HDAC8, ATRX, TBL1X nascent transcripts and *XIST* RNA; or immuno-RNA FISH for RNA Pol II and H3K27me3 combined with *XIST* RNA FISH on primary breast tumor samples: Luminal A sub-type IDC (Invasive Ductal Carcinoma) of grade II (T2) (A); HER2 amplified sub-type IDC of grade II (T3) (B); Basal-like sub-type IDC of grade III (T4) (C) and lymph node metastasis coming from patient on panel C (T4meta) (D).

Left panel: *HDAC8*, *ATRX* and *XIST* expression was assessed by RNA FISH on breast tumor stamps.

Middle panel: TBL1X, *MAGEA6* and *XIST* expression was assessed by RNA FISH on stamps of breast tumor stamps.

Right panel: Immuno-RNA FISH reveals the degree of H3K27me3 enrichment (green) and of RNA Pol II (red) depletion on the *XIST* coated chromosome (gray). Acquisition was carried out by super-resolution structured illumination on DeltaVision OMX microscope for the images displayed on the panel. Quantification of RNA Pol II exclusion and H3K27me3 enrichment at *XIST* domain have been carried-out on images acquired with a Nikon confocal spinning-disk microscope and with the DeltaVision OMX microscope.

(E) DAPI signal intensity of the Barr body. Quantification was performed on primary breast tumor samples and healthy breast tissues using ImageJ software by comparing the DAPI signal at the *XIST* RNA domain versus DAPI signal associated with the ATRX RNA FISH signal at the Xa. The number of nuclei analyzed is indicated at the bottom of the box plot. (***) $p < 0.001$, (*) $p<0.05$ using the Student's *t*-test. All the dataset are compared with healthy breast tissue #1.

(F) Example of DAPI signal intensity quantification on one healthy breast tissue and one luminal A sub-type IDC of grade II (T1).

Scale bar : 10µM

Figure S9. Assessment of the inactive X chromosome transcriptional reactivation in primary breast tumors

(A) Description of allelic expression analysis performed with dataset from 104 Basal-like breast tumors (Shah et al. 2012).
(B) The left chart shows allelic expression status of the 183 informative X-linked genes expressed in the 25 selected tumors (see Figure S9A). The right chart indicates allelic expression status of the 78 bi-allelically expressed genes in normal non-tumor cells (Cotton et al. 2013).
(C) List of genes that escape in at least two primary breast tumors in a “cancer-specific” manner. The three genes in red were already identified as escapees in the three tumor cell lines (see Table S1).

Figure S10. Summary of the inactive X epigenetic erosion in breast cancer cells

(A) Summary table of the overall inactive X-chromosome status in normal and breast cancer cells.

(B) Schematic view of the inactive X chromosome erosion in breast cancer cells.

Table S1. List of the genes identified as subject to XCI or to escape from XCI in the WI-38 clones, ZR-75-1, SK-BR-3 and MDA-MB-436.

Table S2. List of the “good quality” primary breast tumors analyzed based on Shah et al dataset (Shah et al. 2012).

Table S3. List of primers used.

Supplementary references

Chadwick BP. 2007. Variation in Xi chromatin organization and correlation of the H3K27me3 chromatin territories to transcribed sequences by microarray analysis. *Chromosoma* **116**(2): 147-157.

Chaumeil J, Augui S, Chow JC, Heard E. 2008. Combined immunofluorescence, RNA fluorescent in situ hybridization, and DNA fluorescent in situ hybridization to study chromatin changes, transcriptional activity, nuclear organization, and X-chromosome inactivation. *Methods Mol Biol* **463**: 297-308.

Cotton AM, Ge B, Light N, Adoue V, Pastinen T, Brown CJ. 2013. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome biology* **14**(11): R122.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**(4): 576-589.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome research* **12**(6): 996-1006.

Kwon MJ, Oh E, Lee S, Roh MR, Kim SE, Lee Y, Choi YL, In YH, Park T, Koh SS et al. 2009. Identification of novel reference genes using multiplatform expression data and their validation for quantitative gene expression analysis. *PloS one* **4**(7): e6162.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Mendoza-Parra MA, Van Gool W, Mohamed Saleem MA, Ceschin DG, Gronemeyer H. 2013. A quality control system for profiles obtained by ChIP sequencing. *Nucleic acids research* **41**(21): e196.

Piskol R, Ramaswami G, Li JB. 2013. Reliable identification of genomic variants from RNA-seq data. *American journal of human genetics* **93**(4): 641-651.

Popova T, Manie E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH. 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome biology* **10**(11): R128.

Satya RV, Zavaljevski N, Reifman J. 2012. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic acids research* **40**(16): e127.

Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G et al. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**(7403): 395-399.

Sorzano CO, Thevenaz P, Unser M. 2005. Elastic registration of biological images using vector-spline regularization. *IEEE transactions on bio-medical engineering* **52**(4): 652-663.

Vincent-Salomon A, Ganem-Elbaz C, Manie E, Raynal V, Sastre-Garau X, Stoppa-Lyonnet D, Stern MH, Heard E. 2007. X inactive-specific transcript RNA coating and genetic instability of the X chromosome in BRCA1 breast tumors. *Cancer Res* **67**(11): 5134-5140.

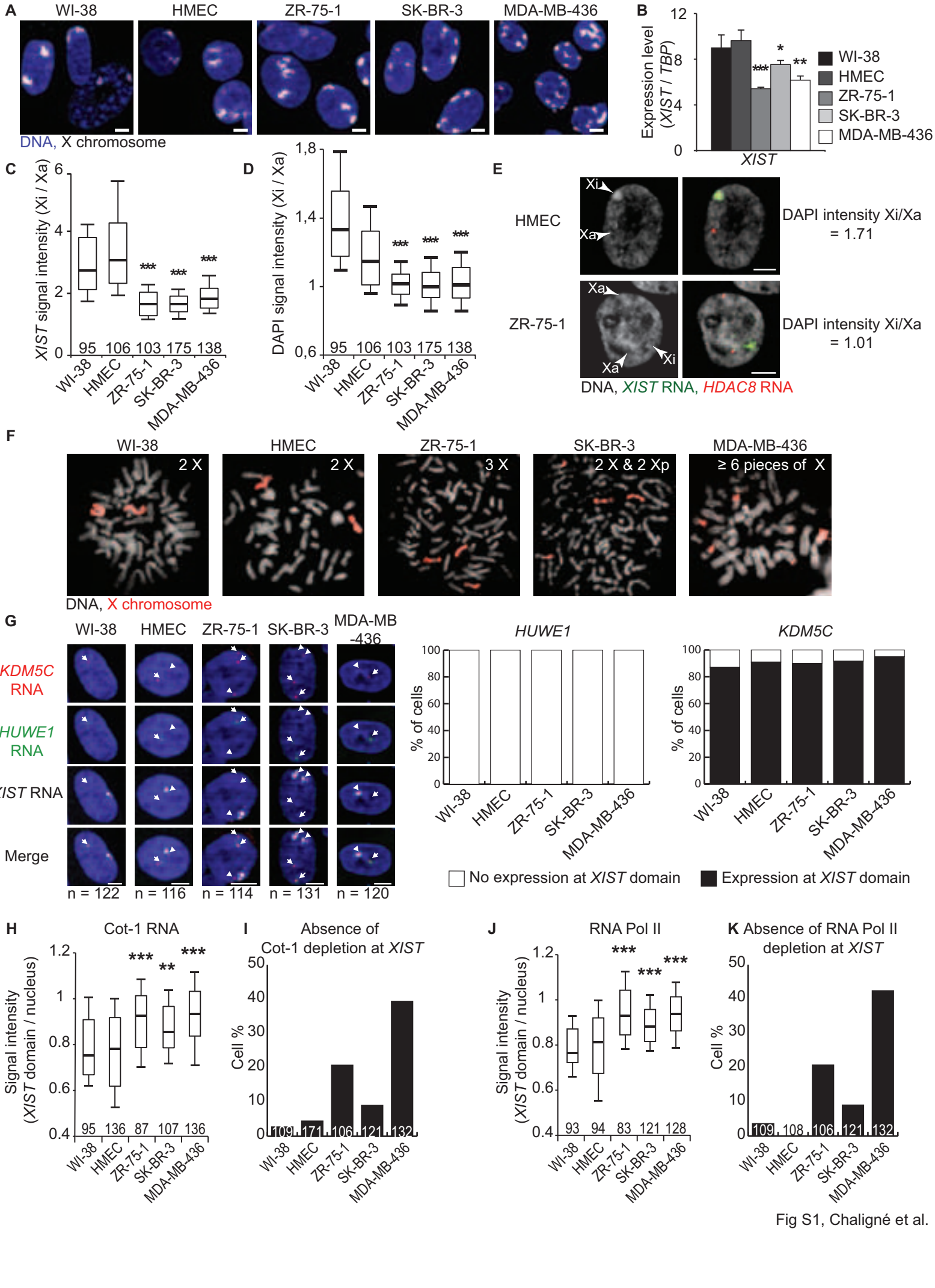


Fig S1, Chaligné et al.

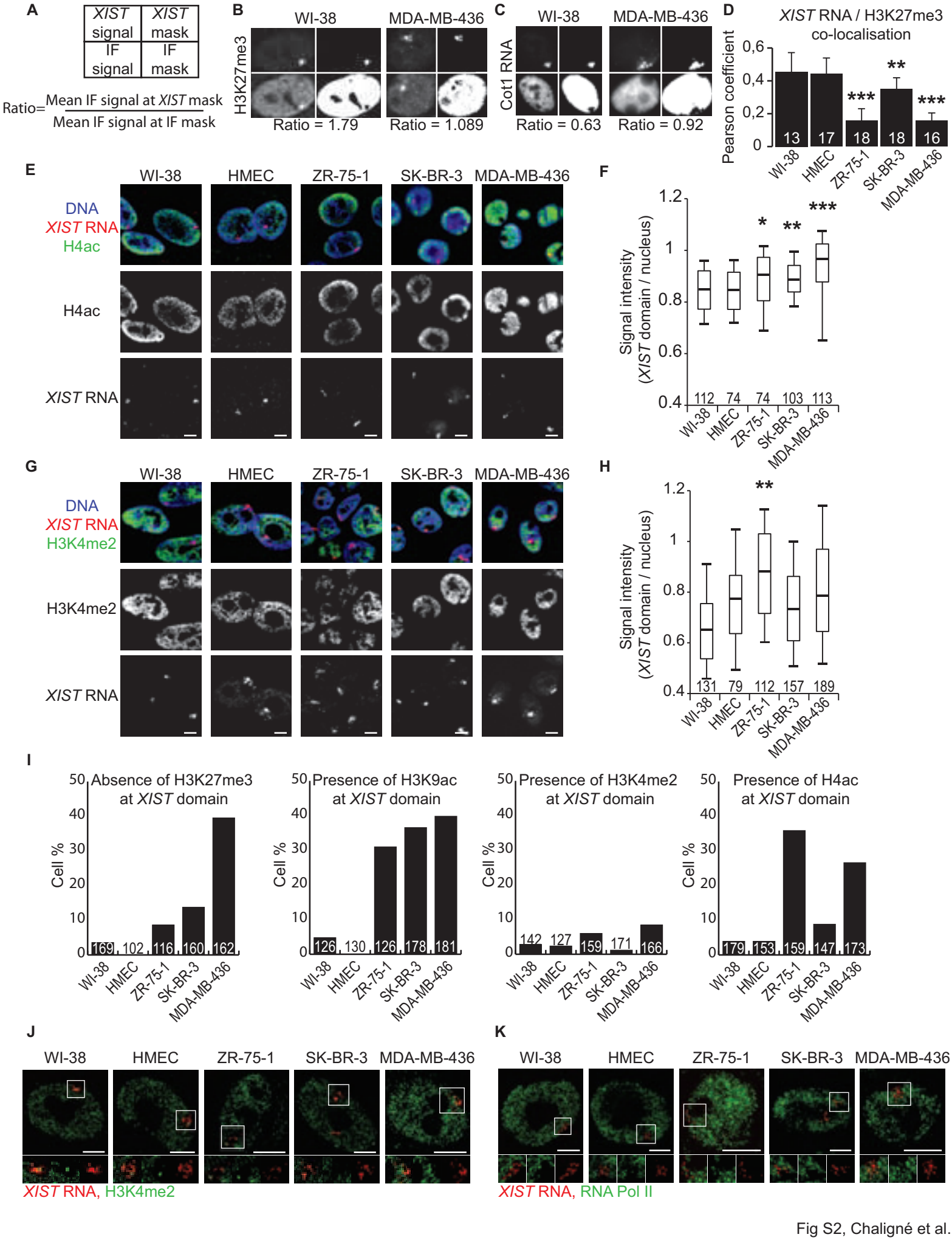


Fig S2, Chaligné et al.

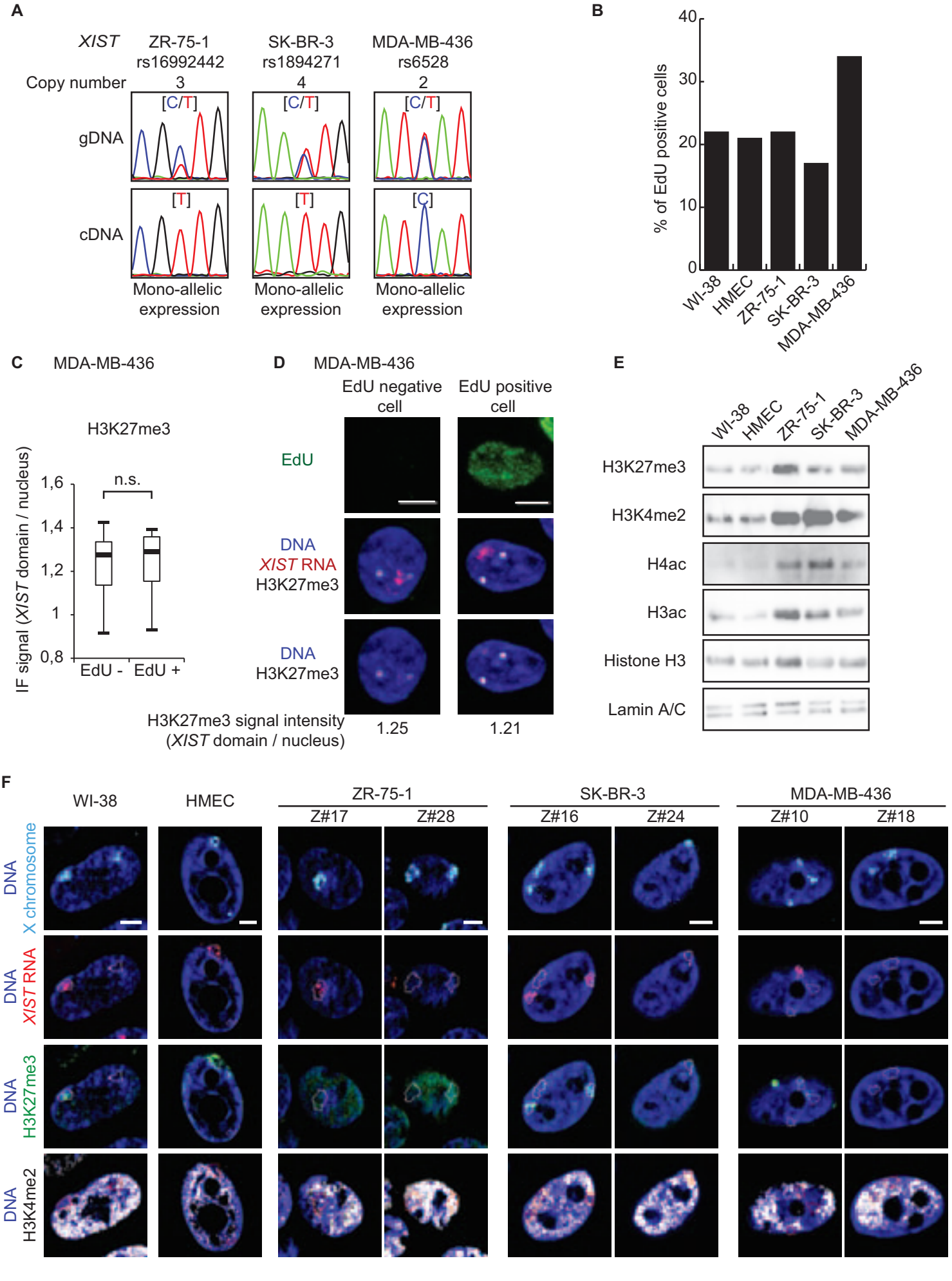


Fig S3, Chaligné et al.

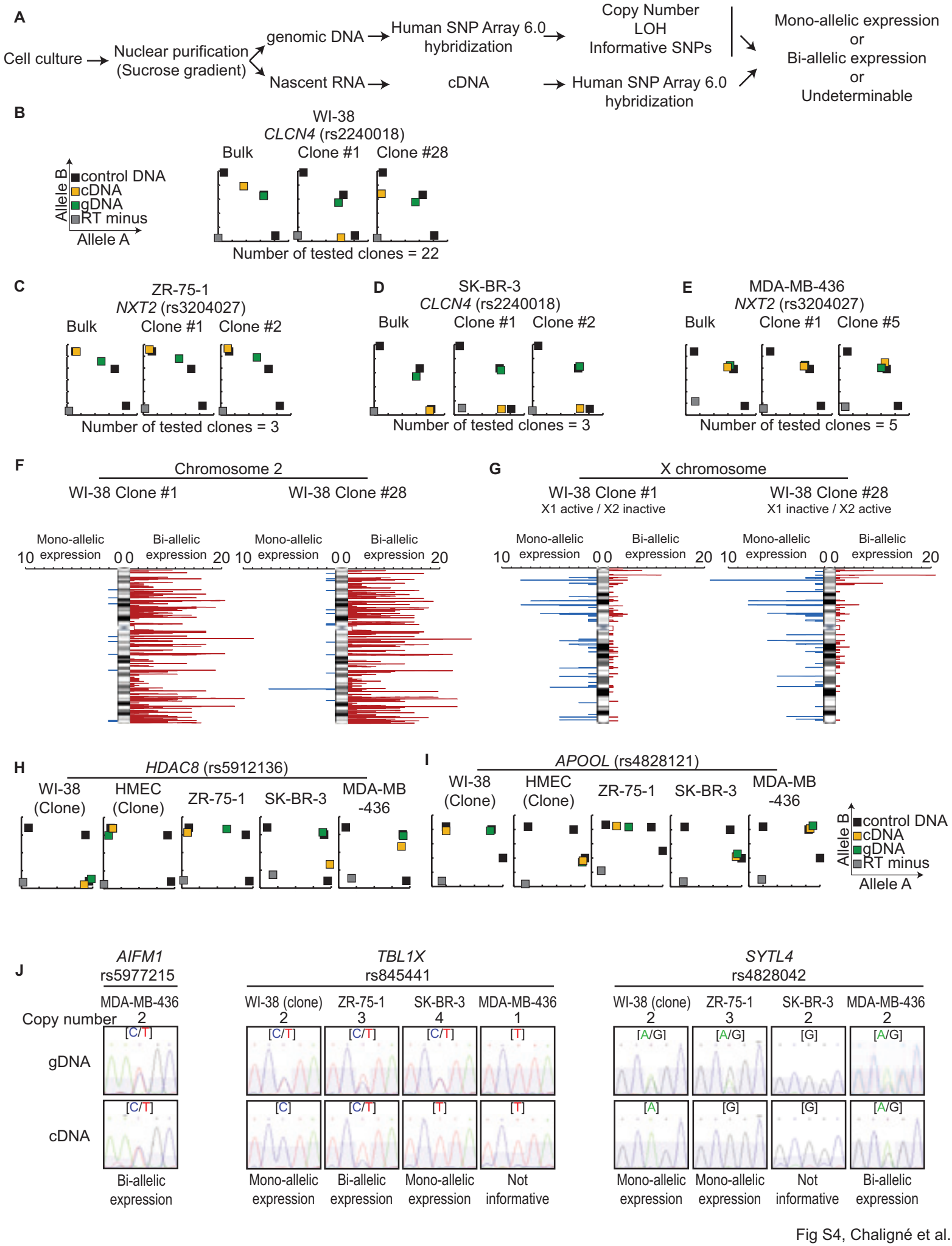


Fig S4, Chaligné et al.

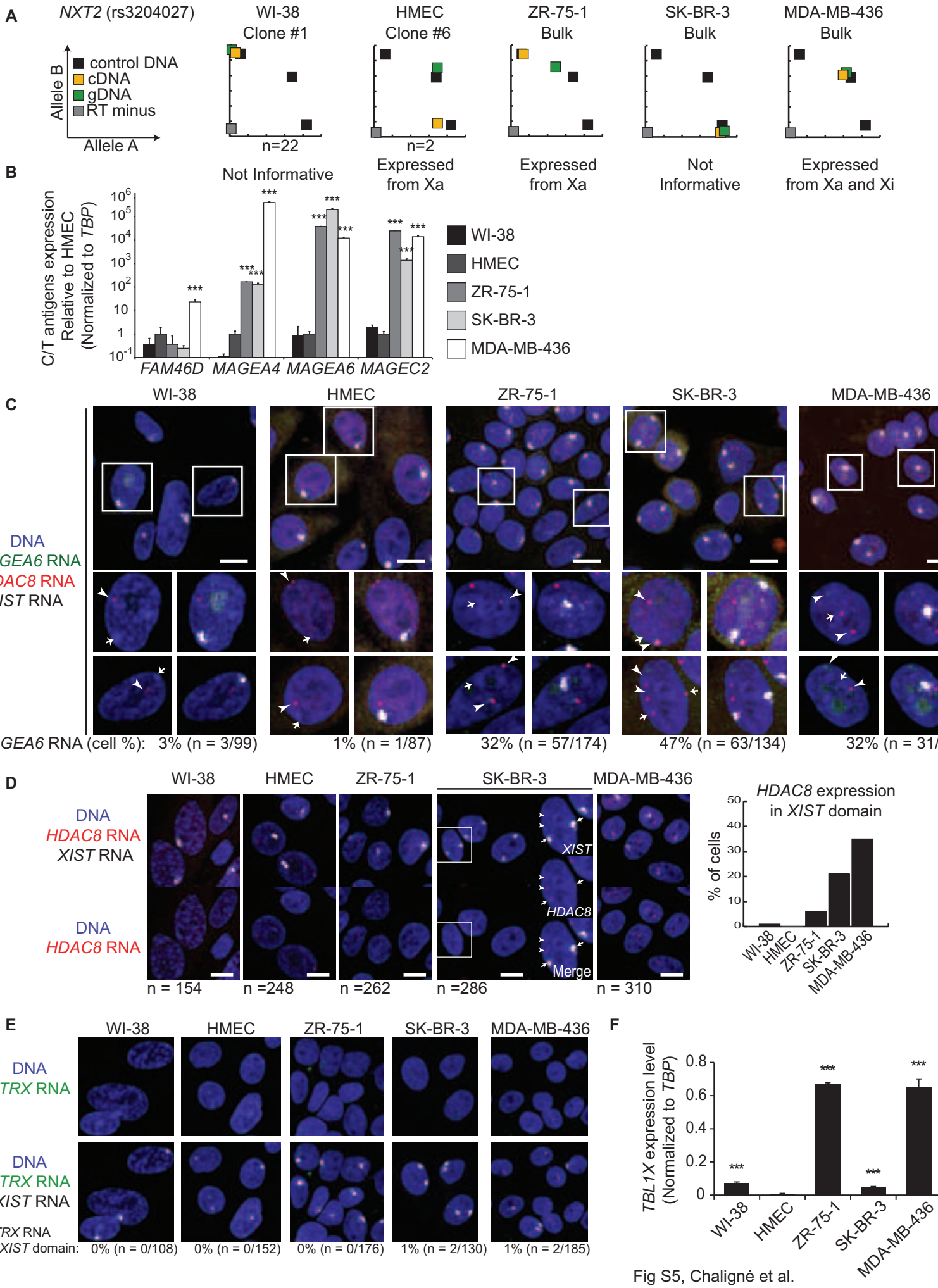


Fig S5, Chaligné et al.

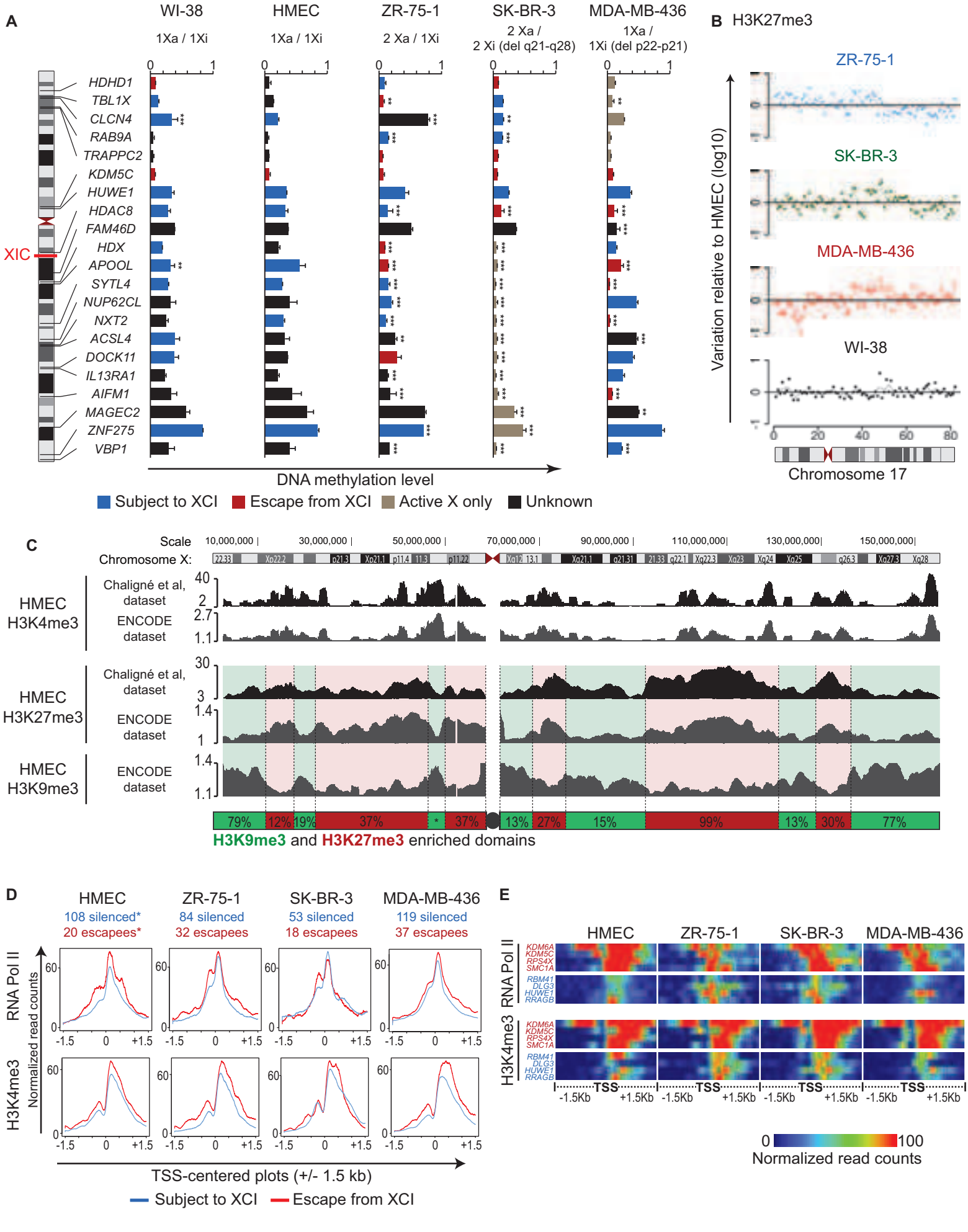


Fig S6, Chaligné et al.

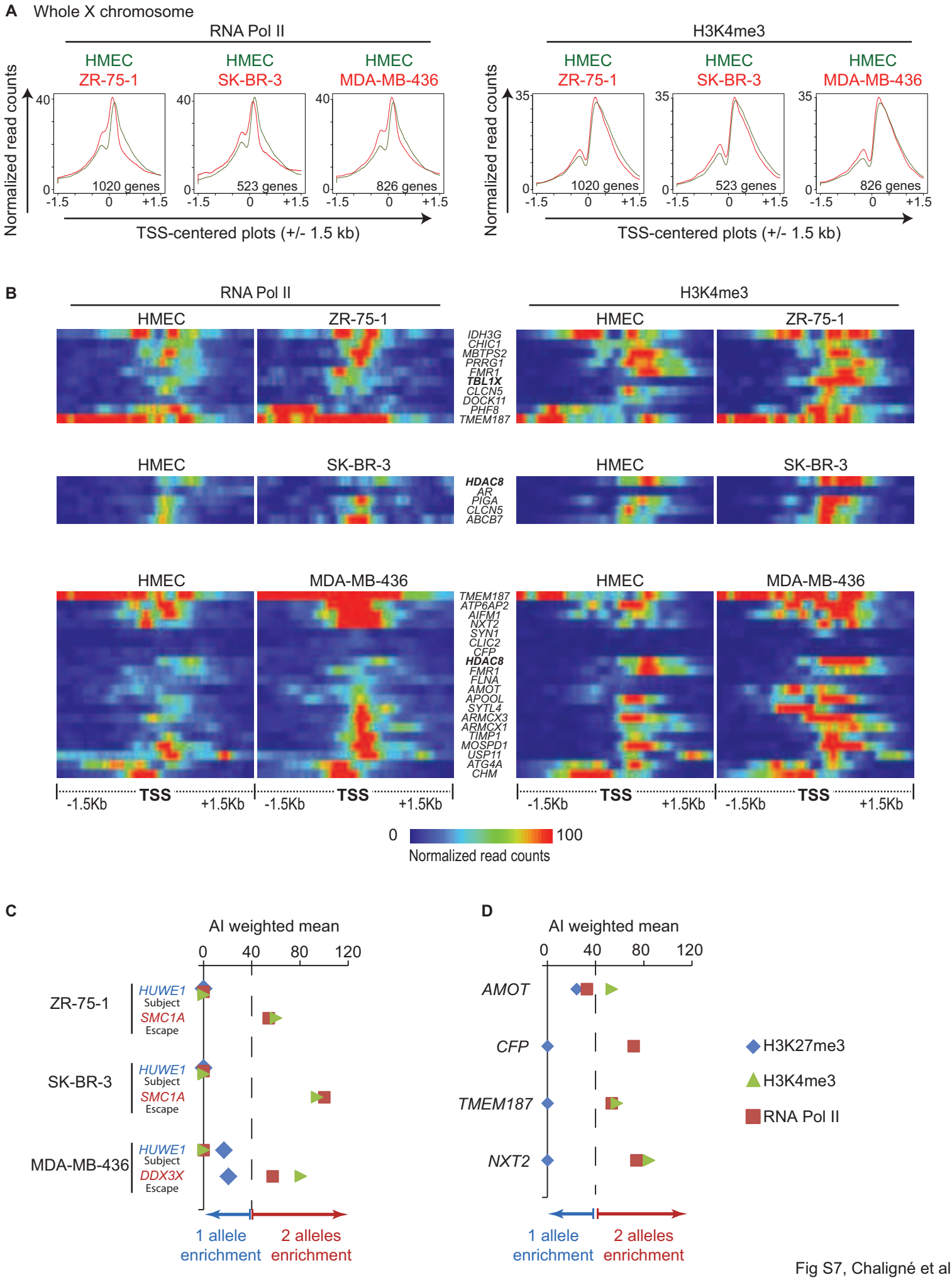


Fig S7, Chaligné et al.

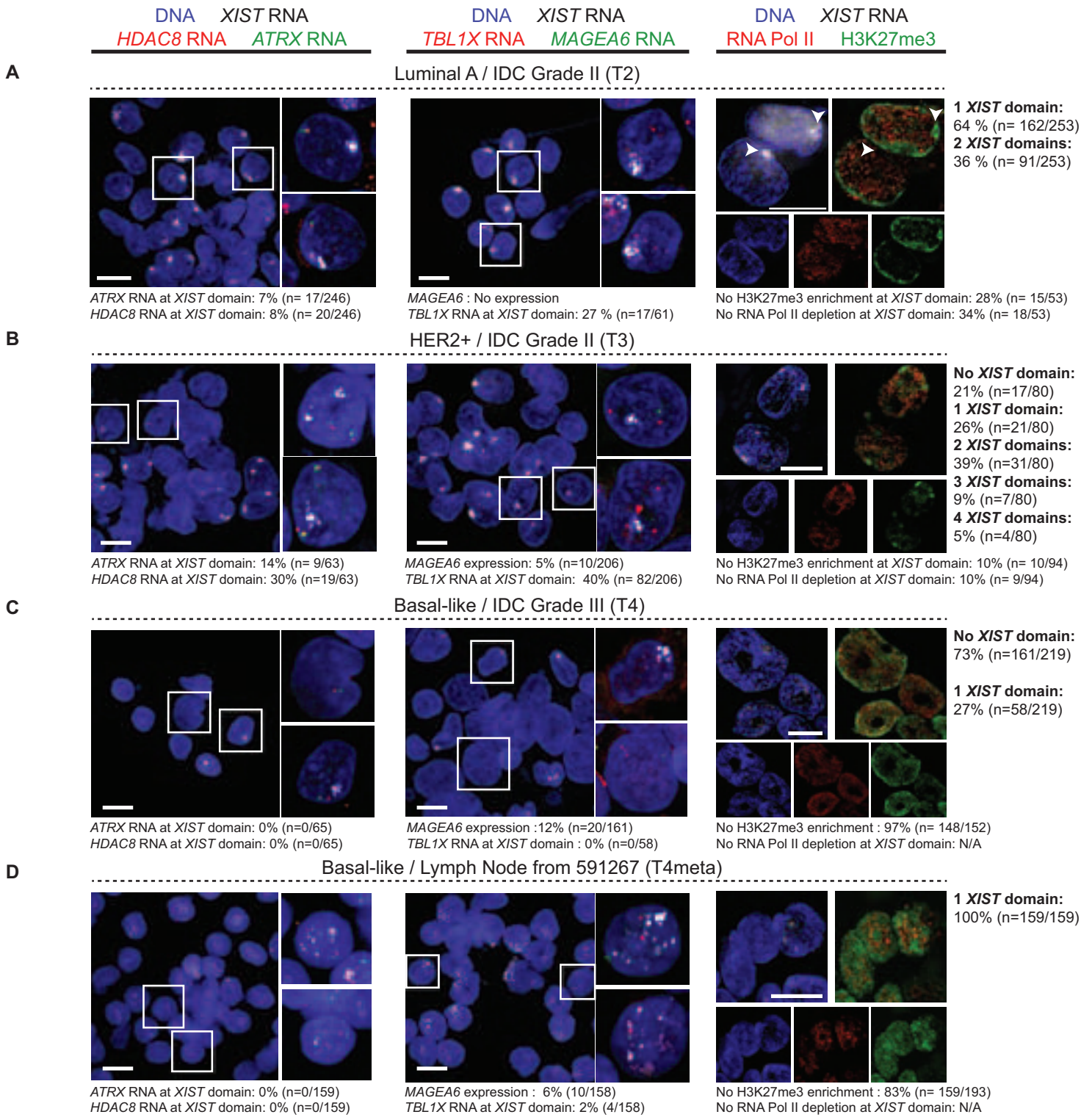


Fig S8, Chaligné et al.

A Data from Shah et al. Nature, 2012

104 primary breast tumors (BLC)

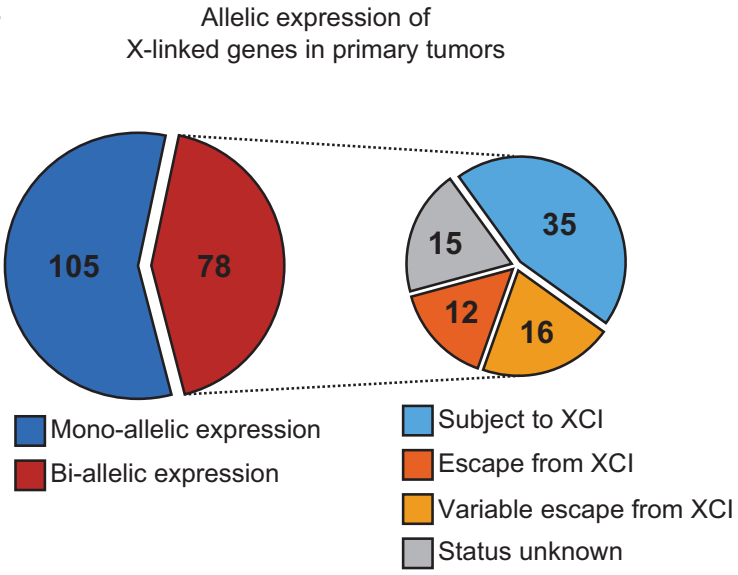
SNP6 array dataset:
X chromosome region with no LOH
High purity and high quality tumors

25 primary breast tumors

RNA-seq dataset:
> 20 reads per SNP
0,3 < RNA Allelic imbalance > 0,7

Allelic expression status
of X-linked genes

B



C

Gene	Tumor number: Bi-allelic expression	Tumor number: Mono-allelic expression
GNL3L	14	2
TMEM164	4	1
XIAP	4	9
CYBB	3	1
DOCK11	3	7
NXT2	3	3
TBL1X	3	5
CLCN4	2	6
PJA1	2	8
SASH3	2	8
SH3BGRL	2	3
SLC25A43	2	5
TCEAL4	2	4
ZNF275	2	9

Fig S9, Chaligné et al.

A

	Xi genetic pattern	Barr body formation	Silencing compartment	Dense *XIST* coating	H3K27me3 enrichment	H3K4me2 paucity	H3/H4 acetylation paucity	"Cancer-specific" escapees	Local heterochromatin formation
WI-38 / HMEC Non tumoral cells		++	++	++	++	++	++	0%	++
ZR-75-1 Luminal breast cancer		+/-	+/-	+/-	+/-	+/-	+/-	9% 10/114	+/-
SK-BR-3 HER2+ breast cancer		+/-	+/-	+	+	+	+/-	8% 5/62	+/-
MDA-MB-436 Basal-like / BRCA1 null		+/-	+/-	+/-	+/-	+	+/-	14% 11/77	+/-
		-	?	-	-	+	+/-	13% 9/72	+/-

B

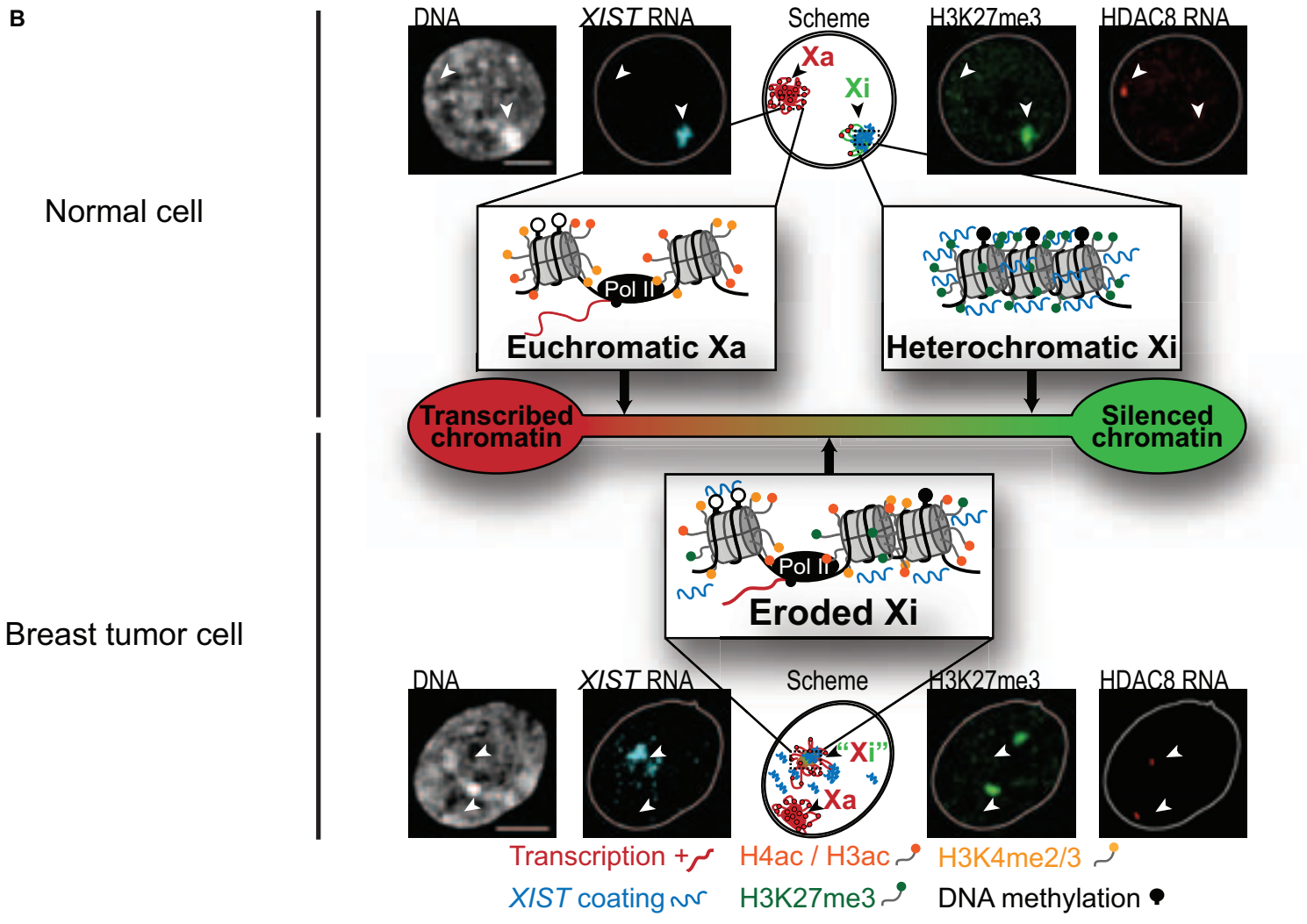


Figure S10, Chaligné et al

CHAPTER 6

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Chapter 6. Concluding remarks and future perspectives

6.1. Utility and limitations of developed tools

Despite carrying basically the same DNA in each cell, instead of becoming an ever-expanding mass of identical cells, we are made up of a wide variety of specialized tissues. The human genome project has identified that only ~1.5% of the genome codes for proteins (Lander et al., 2001). Somehow, each of the 200 different kinds of cells in the human body must be reading off a different set of the hereditary instructions written into the DNA. It has become evident that the hereditary instructions are mediated by a complex regulatory mechanism, termed as 'epigenetics'. These comprise a variety of molecular and structural modifications to DNA and the histone proteins to which it is compactly bound, without changing the underlying sequence but ensure that the right genes are expressed at the right time (Martens et al., 2011).

The past decade has seen a tremendous growth in the field of epigenomics, largely facilitated by the aid of massive parallel sequencing. Different specialized techniques have been developed to understand the epigenetics at various levels. Currently, ChIP-seq is one of the important techniques that has been widely used to study epigenetic modifications. Though it is a mostly a robust technique, it is inherently prone to significant variabilities embedded in individual assays including, but not limited to antibody efficacy and sequencing depth variation. One of the important aspects that directly influence ChIP-seq data is the quality of the antibody used. Egelhofer *et al*., have shown that 25% (out of 200 antibodies tested) of commercially available antibodies are unsuitable for ChIP-seq experiments with either experimental validation or reanalysis of the data (Egelhofer et al. 2011). A significant number of antibodies for histone modifications failed the western blot or dot blot specificity test. Hence, there is a growing concern about the need for extensive quality controls of antibodies. In this regard, NGS-QC can provide an independent global assessment of ChIP-seq data quality. This would be a significant step towards improving the quality of the data produced, leading to more reliable results and efficient use of resources in terms of man hours and money. Publishing journals and data repositories such

as GEO can include NGS-QC quality certification in their pipeline through which researchers will be informed about the reliability of the data.

Local QCi annotated regions produced from NGS-QC are helpful to understand the robustness of each enriched regions and comparability of different datasets. Local QC regions of NGS-QC can provide robust enriched regions for a given dataset. In future, such annotation can be used to quickly compare different datasets to identify the similarity among them. For instance, when there are multiple samples available for a target, methods that perform overlap along population analysis can reveal the highly similar and dissimilar datasets based on the robust enrichments identified in local QC file. Thus, other than global quality indicator to evaluate the overall quality of a sample, NGS-QC provides several other additional information to further analyse/annotate the enrichment events.

ENCODE consortium has recommended two different approaches to assess the ChIP-seq data quality, FRiP, IDR and strand cross-correlation methods (Landt and Marinov, 2012). Though FRiP and IDR (described in chapter 3.3) are useful metrics, their analysis depends on the annotation from peak callers. Such peak identification can vary across different tools and within a tool depending on user-controlled parameters. IDR can assess the samples which have replicates and can provide information about reproducible fraction of peaks only. Strand-cross correlation approach computes the characteristic asymmetric pattern between forward and reverse strand reads in enrichment events. Importantly, while this approach is applicable to profiles with sharp peaks, such as those seen for transcription factors, it cannot be used for the broad profiles often seen for histone marks (Mendoza-Parra and Gronemeyer, 2014). Hence, NGS-QC, a robust annotation-free approach can provide genome-wide quality assessment for massive parallel DNA sequencing from ChIP-seq and other DNA/RNA enrichment-based technologies. Though NGS-QC provides genome-wide robust analysis, we have observed that, some of the less enriched regions may escape from random sampling by chance, thus appearing to be a robust enriched region. NGS-QC tackles this issue by providing an option to repeat sampling in a single run, along with the three levels of sampling (90%, 70% and 50%). Though such scenario with multiple levels sampling is less expected, still it is probabilistic.

Next to antibody efficacy related issue, sequencing depth variation among samples is another technical issue that is observed in ChIP-seq. However, epigenetic studies involve multiple samples to compare and identify differential regulation of genes. It is crucial towards identifying cell-specific differences in regulation, especially in cancer cells to identify the role of epigenetics in it. Earlier, linear normalization, where read counts are scaled with respect to total number of reads, is commonly used in ChIP-seq data (Bailey et al., 2013). However, this approach does not account for signal-noise ratio difference among difference samples in ChIP-seq (Aleksic et al., 2014). Hence, we developed Epimetheus, a genome-wide, annotation free normalization approach for epigenome ChIP-seq data. Our comparison of normalization effect on integrative analysis tools like ChromHMM has shown that prior normalization improves the results in which multiple samples ChIP-seq data are involved. The main advantage of Epimetheus is that the normalization results can be easily used in any downstream analyses. However, like any other analysis, the performance of Epimetheus also depends on quality and comparability of the data. Hence, a prior quality assessment is essential to avoid poor quality datasets affecting the normalization approach, as normalization techniques cannot inherently improve data quality. For example, when lower-quality datasets are used for integrative analyses that are sensitive to false-negative rates, incorrect inferences and conclusions become likely due to high disparity in distribution (Marinov et al., 2014). The basic assumption in quantile normalization is that the read count distributions of the samples to be similar. It is reasonable to assume that their probability distribution of the read counts over the whole genome is similar across different cell types, and in cases where the enrichment events under comparison comprise factors that are implicated in house-keeping events (histone modifications datasets) (Nair et al., 2012). However, an extensive study is needed to verify the suitability of quantile normalization in transcription factor data.

Recently, an experimental spike-in based normalization approach to provide quantitative ChIP-seq data has been developed (Bonhoure et al. 2014; Orlando et al. 2014). In this approach, an exogenous reference is mixed in the library and used as an internal control. A linear scaling based on the total number of reads of exogenous reference is used in analysis to correct the sequencing depth differences, and claimed to provide quantitative and directly comparable results. But this approach is not addressing the technical differences

coming from antibody performances resulting in signal-noise ratio differences. In such cases, we propose to use the quantile normalization prior to linear scaling from spike-in approach. We believe Epimetheus pipeline is imperative even with experimental based normalization approaches. Epimetheus is intended for use with histone modification profiles as the enrichment pattern comparability is higher in it.

6.2. Data management in current bioinformatics

One of the most important challenges that the biological research community is currently facing is in regard to data management. A single run of Illumina HiSeq 2500 machine alone can generate a terabyte of data. Storage of the data produced by modern DNA sequencing instruments has become a major concern. While the capacity of computing hardware doubles every 18 months, new biological data is doubling every 9 months (Bao et al., 2014). For example, as of 2013, European Bioinformatics Institute (EBI) has stored 2 petabyte of genomics related data (Marx, 2013). With this trend, data management has become one of the major challenges in bioinformatics. Though all the files are compressed to reduce the storage occupancy, there has to be novel approaches to compress the data more efficiently without relying on generic approaches. For instance, a data-specific compression tools like Fqzcomp can compress the FASTQ files more efficiently than generic ones (Bonfield and Mahoney, 2013; Nicolae et al., 2015). Fqzcomp, a FASTQ file specific compression tool, has been shown to compress the files approximately to one tenth of its original size, which is two folds lesser than a regular compressing approach 'gzip' (Bonfield and Mahoney, 2013). In such data specific tools, the known formats are encoded into numeric representation where only changes are stored. For example, in a FASTQ file, when there are two read identifiers as follows: @ SRR062634.2724180 and @SRR062634.2724181, the compressor stores the second ID as (18)(1), which means an increment from previous ID in 18^{th} position. Similarly, sequences are packed as k-mers for which numbers are encoded. For example, sequences can be split at 4 base interval and each 4 bases is encoded into bytes. In this way, repetition of same identifiers or sequences can be reduced drastically during compression (Bonfield and Mahoney, 2013). In another aspect, storing FASTQ file as a default option can be replaced by storing alignment (BAM) file alone, where one can extract the sequence and quality from it if needed. Thus, it can

avoid the double storage of same information. A similar reference based compression approach has been proposed as an alternative to SAM/BAM (a standard alignment format), where sequences of aligned reads are not stored but only variations are stored (Jones et al. 2012). In this way, sequences can be extracted from the reference genome and noted variations can be incorporated.

Glossary

Sequencing applications and biological glossary:

ATAC-seq:

Assay for transposase-accessible chromatin (ATAC) sequencing captures open chromatin sites using a simple two-step protocol with 500–50,000 cells and reveals the interplay between genomic locations of open chromatin, DNA-binding proteins, individual nucleosomes and chromatin compaction at nucleotide resolution. Hyperactive Tn5 transposase loaded in vitro with adaptors for high-throughput DNA sequencing can simultaneously fragment and tag a genome with sequencing adaptors. Avoids potentially loss-prone steps like such as adaptor ligation, gel purification and cross-link reversal as ATAC is minimal (Buenrostro et al., 2013).

Barr body:

Barr, named after Murray Barr, is the inactive X chromosome in female where one of the two X chromosomes as a dosage compensation method. In 1949, Barr and Bertram first identified a nuclear body within female cat neurons, but not in the corresponding male cells (Barr and Bertram 1949).

BS-seq:

This is a method which identifies methylation of cytosine (5mC) genome-wide. Cytosine methylation plays important role in gene regulation and chromatin remodelling. In this method, sodium bisulphite chemistry is used to convert non-methylated cytosines to uracil which is converted to thymine in the sequence reads or data output. After bisulphite conversion the DNA is sheared and sequenced using next generation sequencing technologies (Fraga and Esteller, 2002).

ChIA-PET:

Chromatin Interaction analysis using paired end tags (ChIA-PET) is used identify functional targets (Chromatin interactions) of DNA bound protein. ChIA-PET provides genome-wide high-resolution data for interactions that involve a given DNA-binding

protein. This method involves use of formaldehyde for cross linking the interactions, after this chromatin is immune-precipitated using antibody against protein of interest. DNA ligase is used to create chimeric DNA fragments. This is followed by restriction digestion and sequencing by next generation sequencing technologies (Wit and Laat, 2012).

ChIP-exo:

It is more precise method to probe exact binding of protein in a protein-DNA interaction genome wide. It uses lambda exonuclease to digest the DNA which is not bound to proteins. The exonuclease also removes contaminating DNA from the reaction. After this normal immunoprecipitation is performed using specific antibody. The isolated DNA is subjected to next generation sequencing technologies to get precise binding sites (Rhee and Pugh, 2012).

ChIP-seq:

Chromatin Immunoprecipitation (ChIP) coupled to sequencing is a method to probe sites of protein-DNA interaction. This method is widely used to map global binding of sites of transcription factors in a genome. In this method, protein-DNA interaction is cross linked using formaldehyde. Further, chromatin is sheared and immune-precipitated using specific antibody against a protein associated with the DNA. This purified ChIP DNA is sequenced using next generation sequencing technologies (Meyer and Liu, 2014).

Exome-seq:

Exome refers to the protein coding regions of the genome. This method comprises of selective capture of exome coupled to next generation sequencing methods. This is most widely used method of targeted sequencing. This is a cost effective method to genome sequencing if the interest lies in coding regions of the genome. It is used for identification of structural variants and SNPs. Its application lies in population and disease genetics (Ng et al., 2010).

FAIRE-seq:

Formaldehyde-assisted isolation of regulatory elements (FAIRE) sequencing is used to identify the open/accessible chromatin regions. Formaldehyde is used to crosslink

chromatin, and phenol–chloroform is used to isolate sheared DNA which then will be sequenced (Giresi et al., 2007).

GRO-seq:

Global run-on sequencing assay to quantify transcriptionally engaged polymerase density genome-wide. It is used to profile the activity of engaged PolII along transcribed regions providing real-time transcriptional behaviour. Capture of nascent transcripts helps to identify variety of RNA species beyond the regular genes encoding proteins (Allison et al., 2014; Core et al., 2008).

HiC:

HiC is a method that assesses the three-dimensional architecture of whole genomes by coupling proximity-based ligation with high-throughput sequencing. In HiC, chromatin is cross-linked with formaldehyde, then digested using restriction enzymes, and re-ligated in such a way that only DNA fragments that are covalently linked together form ligation products. A biotin-labelled nucleotide is incorporated at the ligation junction to enrich chimeric DNA ligation junction for sequencing (Belton et al., 2012).

MeDIP-seq:

DNA methylation plays key role in gene expression and chromatin organisation. Methylated DNA immunoprecipitation or MeDIP is a method which employs antibodies against 5m Cytosine to immuno-precipitate all the methylated DNA in the genome. This method aids in analysis of methylome. The immune-precipitated DNA is sequenced using next generation sequencing technologies (Weber et al., 2005).

MNase-seq:

This method is used to probe the nucleosome positioning and density in the genome. It can also be used to find the nucleosome free regions in the genome. This method uses micrococcal nuclease (MNase) digestion to locate nucleosomes and after this isolated DNA is sequenced by next generation sequencing technologies (Meyer and Liu, 2014).

RIN number:

The RNA integrity number (RIN) is a software tool designed to help scientists estimate the integrity of total RNA samples. A RIN number is computed for each RNA resulting in the classification of RNA samples in 10 numerically predefined categories of integrity. The output RIN is a decimal or integer number in the range of 1–10: a RIN of 1 is returned for a completely degraded RNA samples whereas a RIN of 10 is achieved for intact RNA sample. In general, RIN number greater than 7 is recommended for experiments.

RNA-seq:

This method comprises of sequencing RNA (whole transcriptome, mRNA or small RNA) using next generation sequencing methods. Sequencing transcriptomes is a major advance in the field of gene expression studies as it allows visualisation of whole transcriptome rather than subset of predefined genes in expression microarray studies. It provides comprehensive view of cellular transcriptomic profile and it also aids in identification of novel transcripts genome-wide (Morin et al., 2008).

Single-end and Paired-end sequencing:

In single-end sequencing, the sequencer reads a fragment from only one end to its specified sequencing length. In paired-end sequencing, the sequencer reads at one end of a fragment to its specified read length, and then starts another round of reading from the opposite end of the fragment. The main advantage of paired-end sequencing is that the information from both the ends of a fragment provides higher alignment accuracy.

WGS:

Whole genome sequencing (WGS) is the complete genome sequencing starting from genomic DNA without any prior capture or pull-down to attain reads covering the whole genome. WGS reveals the complete DNA make-up of an organism, enabling us to better understand variations both within and between species. It can be used to identify an individual's complete genome sequence (coding and noncoding regions); including copy number variation (e.g., repeats, indels) and structural rearrangements (e.g., translocations) (Rizzo and Buck, 2012).

Bioinformatic glossary:

Background modeling:

The background modeling is performed in peak calling to exclude background noises. It can be defined as an assumed statistical noise distribution or a set of assumptions that guide the use of control data to filter out certain types of false positives in the treatment data.

Base quality:

Base quality is the value provided by the sequencing machines to represent the confidence of a base called being correct. To give simplified representation, quality values are encoded as ASCII values like ‘A’ for 65, ‘B’ for 67, etc. Higher the value represents higher the confidence of base called being correct.

Clonal reads/PCR duplicates:

Clonal reads are the over-representation of the same fragment multiple times which are induced due to PCR step involved in sequencing. Clonal reads can arise from various reasons such as differences in GC content, whereby a higher GC content can lead to an increased PCR amplification. The resulting clonal reads can contribute disproportionally to read coverage data. Hence, it is recommended to remove the clonal reads to avoid bias in the analysis.

False discovery rate (FDR):

False discovery rate is similar to P-value, where it is used to represent the significance of results. In ChIP-seq peak calling, it is used to represent the chance of a result being wrong by performing peak calling at each p-value to find ChIP peaks over control and control peaks over ChIP. For example, if there are 1,000 peaks whose p-value $\leqq$ 0.00018, MACS uses the input sample as the IP and IP as the control to identify peaks again. If totally there are 48 peaks in the input sample over the IP whose p-value $\leqq$ 0.00018. The FDR for the peak X = 48 / 1000 = 0.048.

FASTQ

FASTQ files are the raw sequence files that are generated from the sequencing machine. FASTQ format is represented in four parts at consecutive lines: (i) sequence/read ID starting with '@' symbol (ii) sequence (iii) quality ID starting with '+', and (iv) quality values encoded in ASCII format. Symbols '@' and '+' are used to distinguish the sequence and quality values, as quality values can also have ATGC/atgc characters.

Gapped alignment and ungapped alignment:

When aligners search for a sequence match in the reference genome, gapped alignment allows gaps along with mismatches in the string match. As a given sample can have insertions or deletions, allowing gaps in the alignment can facilitate alignment of reads that carry an insertion or deletion. Subsequently, variation callers use this information to identify insertions or deletions in the sample. Ungapped alignment allows does not allow gaps in the string match; hence it cannot be used to identify insertions or deletions in the data.

Hidden Markov model:

Hidden Markov Model (HMM) is a full probabilistic model i.e., scores and the parameters used are all probability values that can be manipulated and optimized in a variety of ways. Hence, it has found wide applications in computational biology and it is used to solve a variety of problems, including gene finding, profile searches, multiple sequence alignment and regulatory site identification. In biological sequence analysis it relies on the fact that different parts of a sequence have different statistical properties. For example, GENSCAN – a gene finder that employs HMMs internally – assigns different probabilistic values for different states and transition probabilities for the transition between these states. Once a DNA sequence is given as an input, a HMM parses the sequence generating state paths and an observed path. Finally, the most probable state path is given as our gene model. HMMs have proven to be very successful in such applications.

Irreproducible discovery rate (IDR):

IDR is recommended as a quality standard to evaluate the reproducibility information from the replicates by ENCODE consortium. The basic idea is that between two biological replicates, the most significant peaks are expected to have high consistency. However, the peaks with low significance, which are more likely to be false-positive, are expected to have low consistency. If the consistency between a pair of rank lists (peaks) that contains both significant and insignificant findings is plotted, a transition in consistency is expected (Fig. 1C). This consistency transition provides an internal indicator of the change from signal to noise and suggests how many peaks have been reliably detected.

Local alignment and global alignment:

The very basic difference between a local and a global alignment is that the local alignment tries to match a substring (part of a sequence) of the read with the reference whereas the global alignment performs an end to end alignment with the reference. With the use of local alignment, aligners try to improve the alignment rate and accuracy by clipping the low quality or contaminated part of the read.

Local QC indicators (local QCi):

Local QCis are the wiggle file output from NGS-QC, where δRCI (dispersion of RCI after sampling per bin) information is presented. Such local QCi regions can be used to judge the robustness of read accumulation in a given bin.

Mapping quality:

A mapping quality is basically the probability that a read is aligned in the wrong place. Mapping quality is represented in phred-scaled probability value. Different aligners use different approach to calculate mapping quality, but mostly it estimated based on the base quality scores of the read, uniqueness of the match and mismatching bases in the alignment.

P-value:

P-value is the probability value of particular results being wrong, thus it helps to determine the significance of results. In data analysis, several tools provide P-value to represent the confidence value for a particular analysis. For example, in ChIP-seq peak calling analysis, P-value is used to determine the probability value of particular peak called being wrong.

Phred score:

A Phred quality score is used to represent the quality of each base. Phred quality score 'Q' is -10log(P), where 'P' is probability value of the base called being wrong. *P*-value of 0.01 would result in quality score of 20 which means that there is a 1 in 100 chance of that base being miscalled.

Poisson distribution:

The Poisson distribution can be used to calculate the probabilities of various numbers of "true positive" identification based on the mean number of successes.

Read count intensity or Reads per base:

Read count intensity or reads per bin is the cumulative count of reads within a specified region or window.

Read shifting:

Read shifting is in silico approach of extending reads to its specified average fragment length.

SAM/BAM:

SAM (Sequence Alignment/Map format) is a standard alignment format. A SAM file is a TAB-delimited text format consisting of a header containing reference sequence information, and a line per read alignment containing alignment related information. BAM is the compressed binary format of SAM.

Soft clip/trim:

In aligner's perspective, clipping is that when there is no full match for a read alignment, aligners may clip the reads tail if they contain low quality bases to improve the alignment rate.

Z-score:

Z-score is a statistical measurement of a value from population indicating how many standard deviations from the mean of the population. It is calculated by the distance of an element from a mean value of population, and divided by the standard deviation. A positive Z-score represents a value greater than the mean and negative score represents a value less than the mean. Z-score of 0 represents an element equal to the mean.

THESIS RÉSUMÉ

Analyse intégrative de données issues de séquençage à haut débit de cellules cancéreuses du sein

L'expression génique peut être affecté ou régulée génétiquement ou épigénétiquement. Des études ont montré, au fil des ans, que les modifications épigénétiques ont un rôle significant dans la régulation génique. La découverte des méchanismes et aspects fonctionnels de ces modifications nous aiderait à mieux comprendre pourquoi différents types cellulaires possèdent de multiples comportements à partir du même ADN. Depuis la découverte du rôle essentiel de ces modifications, des changements aberrants ont été observés dans plusieurs maladies, dont le cancer. Puisque la majorité de ces modifications sont réversibles, un réel effort a été fait afin de les utiliser au cours de thérapies. Les avancées technologiques ainsi que la diminution des coûts ont fait des méthodes de séquençage à haut débit un moyen rapide et exhaustif d'explorer les effets de l'expression des gènes. Parmi ces méthodes, le séquençage de fragments immuno-précipités par immuno-précipitation de chromatine (ChIP-Seq) est couramment utilisé pour détecter des interactions protéines-ADN et établir des profils épigénomiques de cellules afin de comprendre la différentiation des cellules souches, la cancérogenèse, etc. L'étude de l'épigénome nécessite le séquençage de plusieurs modifications d'histones pour comprendre l'état de la chromatine dans différentes régions entre différents échantillons (e.g. traitement/contrôle, sain/malade) au cours du temps. De même, les techniques « Exome-Seq » et « RNA-Seq » ont été largement utilisées pour comprendre les variations génétiques et leurs effets au niveau transcriptionel. Cela s'est traduit en une importante accumulation de données à intégrer pour pouvoir avancer des conclusions. Cela pose un réel défi bioinformatique puisque la technique du ChIP-Seq est, par nature, sujette à des variations entre échantillons dues à l'efficacité de l'anticorps, la profondeur de séquençage, etc. Ces variabilités couplés à des profils peu enrichis peuvent considérablement biaiser les études comparatives, d'où un besoin de nouvelles approches et de nouveaux outils pour répondre à ces limites. Étant données les limites des méthodes de séquençages et l'apparition de variations techniques entre échantillons (bruit), certaines données ne peuvent pas être directement comparées. Une approche bioinformatique est nécessaire afin de corriger de manière robuste ces différences les rendant comparable, puis de réaliser une analyse intégrative multidimensionnelle dans le but de comprendre le mécanisme de régulation génique.

La première partie de ma thèse décrira le développement d'outils novateurs et leur importance vis-à-vis des problèmes préalablement décrits. La seconde partie détaillera l'intégration, grâce

aux outils développés, de différents types de données dans le but de comprendre le rôle épigénétique de l'inactivation et de la réactivation du chromosome X, ainsi que des gènes soumis à empreinte dans les cellules cancéreuses du sein.

A. Développement d'outils d'analyse de données de séquençage à haut débit

L'accroissement du nombre d'expériences épigénomiques disponibles dans des bases de données publiques comme GEO, qui possède actuellement plusieurs milliers de profils, contribue fortement à l'expansion de l'information. Cela encourage l'intérêt porté aux études intégratives et comparatives visant à explorer les mécanismes de régulation génique. Les défis d'aujourd'hui consistent à définir de manière fonctionnelle les motifs locaux et globaux des états de la chromatine pour différents systèmes physiologiques dans une perspective multidimensionnelle. La technique « ChIP-Seq » si largement utilisée est intrinsèquement encline à générer des variations entre expériences, ce qui pose des problèmes d'ordre bioinformatique lors d'analyses comparatives, un problème récurrent dans les analyses « big data ». Plusieurs facteurs, tels que l'efficacité de l'anticorps ou de la librairie de séquençage, ont un impact direct sur la qualité des données et donc sur toutes les analyses ultérieurs. Il est, par conséquent, impératif d'évaluer la qualité des données avant toute étude comparative. Cependant, l'absence de systèmes de contrôle de qualité (QC) représente un frein majeur aux analyses comparatives de données séquencées. Cela concerne d'autant plus les expériences portant sur l'étude des interactions protéines-ADN (ChIP-seq), mais aussi celles basées sur un enrichissement (MeDIP-Seq, GRO-Seq, RNA-Seq).

Comme décrit précédemment, la comparaison ou l'intégration de différents profils requiert une évaluation spécifique de la qualité puisque les motifs peuvent être variables et qu'il peut exister des divergences techniques entre profils dues à l'utilisation de différents anticorps, un séquençage plus ou moins profond ou une immuno-précipitation (IP) plus ou moins efficace, etc. Pour pallier à ces problèmes, nous avons développé NGS-QC Generator, un outil bioinformatique de contrôle qualité qui utilise les données de séquençage pour (i) attribuer des indicateurs de qualité globaux indiquant le degré de comparabilité des plusieurs profils NGS ; (ii) proposer des indicateurs de qualité locaux afin d'évaluer la robustesse des enrichissements dans une région génomique précise ; (iii) recommander une profondeur de séquençage optimale pour une cible donnée ; et (iv) donner des moyens de comparer différents anticorps et lots d'anticorps lors d'expériences ChIP-Seq ou toute autre expérience usant d'anticorps. C'est pourquoi nous avons développé une approche associant des indicateurs de qualité

locaux et globaux à des profils ChIP-Seq ainsi qu'à d'autres types de profils issus de séquençage à haut débit. Cette approche a été utilisée pour certifier plus de 20,000 profils disponibles dans des bases de données publiques. Les résultats ont été compilés dans une base de données afin de permettre la comparaison des rapports de qualité (www.ngs-qc.org).

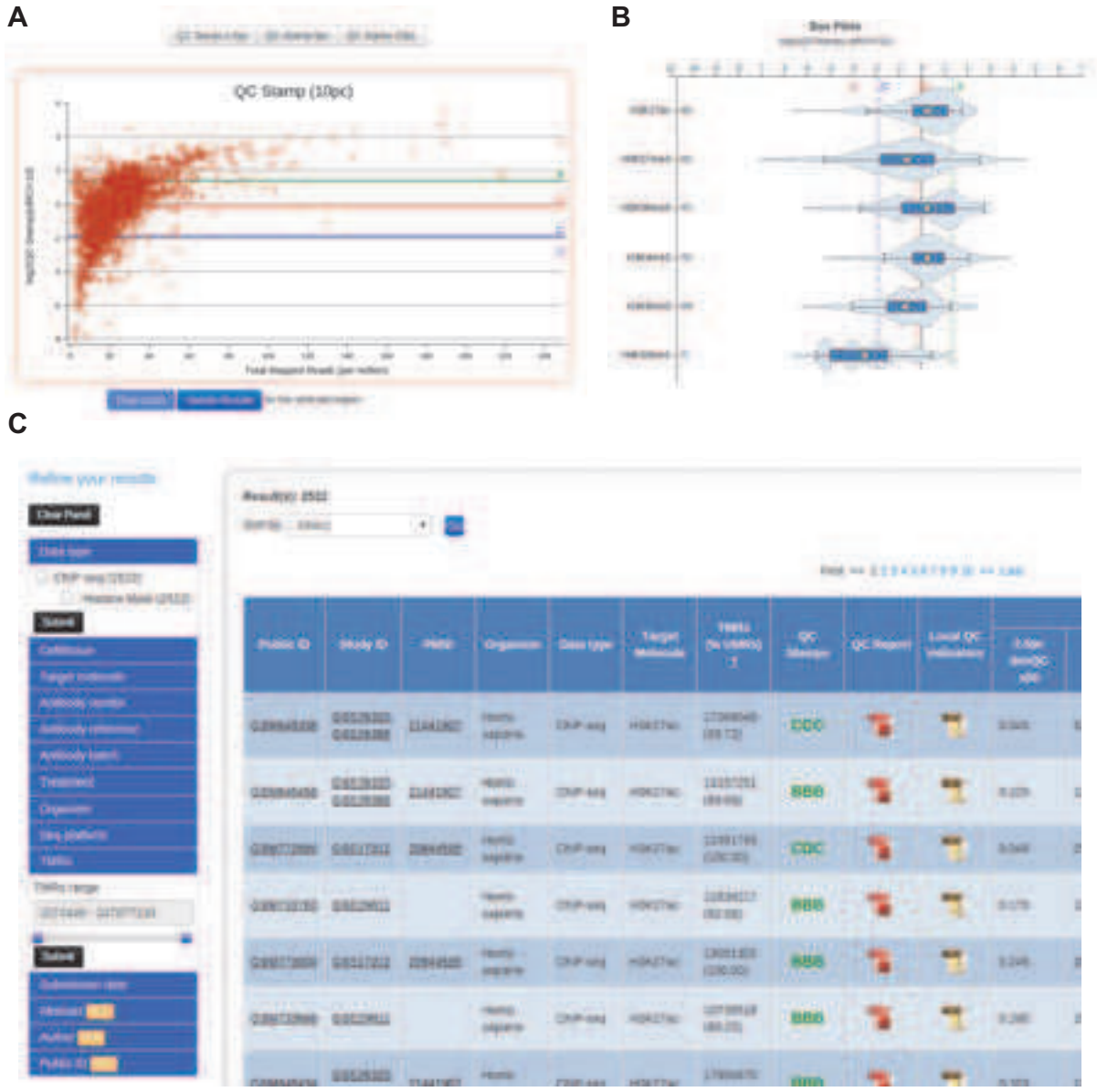


Figure1. Capture d'écran des résultats d'une requête depuis la base de données NGS-QC sur des profils H3K27ac de l'Homo sapiens. **(A)** Nuage de points représentants les indicateurs de qualité par rapport au nombre total de *reads* alignés. **(B)** Diagrammes en violon représentant les différentes cibles retournées par la requête. **(C)** Tableau contenant des informations complémentaires pour chaque expérience (à droite) et outil de filtrage (gauche).

La logique de cette méthode est qu'au-delà d'une certaine profondeur de séquençage, un profil ChIP-Seq change d'amplitude mais pas de motif (Mendoza-Parra et al., 2013). Nous évaluons cette tendance par sélectionnant aléatoirement des *reads* (90 %, 70 % et 50 % du total) pour observer les divergences par rapport aux changements attendus. Des motifs d'intensité de comptage des *reads* (RCI) sont construits pour le profil original et les profils échantillonnés aléatoirement par comptage des *reads* chevauchant des fenêtres génomiques non-chevauchantes de taille fixe. En comparant le RCI observé (recRCI) et le RCI original (oRCI), nous calculons une dispersion RCI (δRCI) pour chaque fenêtre génomique. Puis, en mesurant la fraction de fenêtre possédant un δRCI dans un intervalle donné, une évaluation quantitative détermine les indicateurs de qualité. Afin de faciliter l'analyse comparative de données publiques et de maintenir un portail de contrôle qualité de référence, nous avons développé la base de données NGS-QC en appliquant la méthode à un important nombre de profils disponibles dans des bases de données publiques. Sur un site Internet dédié, les utilisateurs peuvent accéder à une collection d'indicateurs de qualité calculés sur de nombreux profils et, pour chaque profil, télécharger un rapport de contrôle qualité (Figure 1).

Cependant, même les profils de haute qualité présentent des variations dans la profondeur du séquençage, impliquant la nécessité de normaliser les données avant toute étude comparative. Les méthodes existantes de normalisation se basent sur une correction linéaire, et/ou sont limitées à certaines régions génomiques. Pour pallier ces limites, nous avons développé Epimetheus, un outil de normalisation multi-profils, basé sur les quartiles, pour données de modifications des histones.

Epimetheus procède en deux étapes de normalisation : une correction du classement des quartiles, puis une méthode du Z-score pour corriger les divergences inter-profil et inter-cible. Différents graphiques et fichiers de visualisation (données brutes et normalisées) sont, en outre, produits. Enfin, contrairement aux autres outils, Epimetheus produit des fichiers BED normalisés (le format BED est un format standard utilisé par de nombreux outils) par ajout ou suppression de *reads* tout en respectant la divergence entre l'intensité des données brutes et normalisées grâce au fichier d'alignement, ce qui peut être utilisé pour des analyses ultérieures (Figure 2).

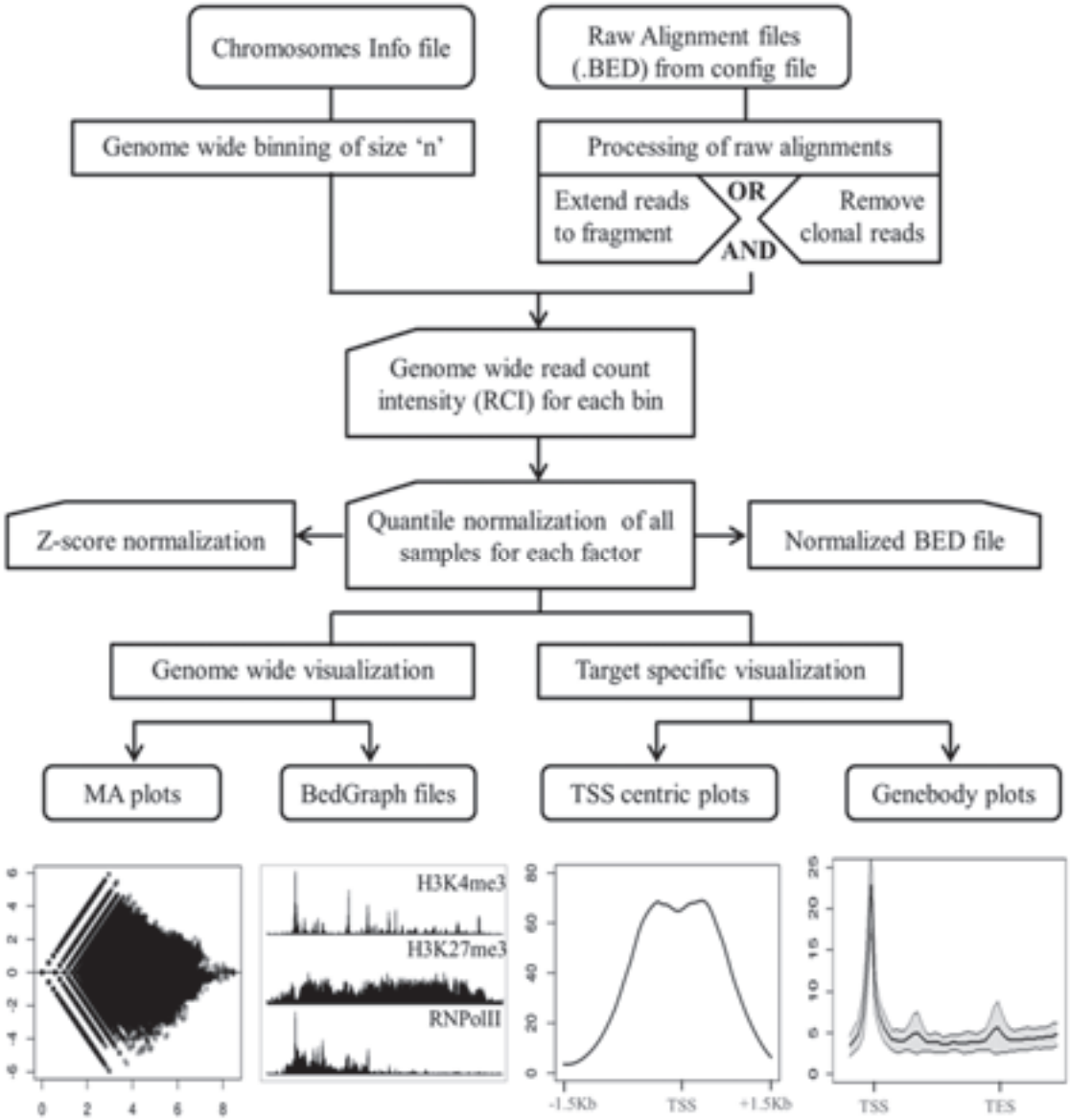


Figure 2. En plus de la normaliser, Epimetheus propose des graphiques pour comprendre l'enrichissement au niveau du promoteur (graphique TSS) et du corps de gène (graphique pour l'ARN PolII) ainsi qu'un graphique MA afin de comparer les expériences avant et après normalisation. Enfin, des fichiers BedGraph sont générés et peuvent être chargés dans un *genome browser* pour visualiser les données.

Epimetheus a été validé en utilisant des réplicas biologiques de profils H3K4me3 dans des cellules HepG2 (Ernst et al., 2011). Avant normalisation, ces réplicas présentaient des nombres similaires de sites promoteurs enrichis, comme prévu. Cependant, du fait de variabilités techniques, ils présentaient également des différences significatives au niveau de l'enrichissement et du ratio signal/bruit. Epimetheus a permis d'ajuster ces différences grâce à la disparité des ratios signal/bruit entre les expériences. Pour vérifier les conséquences de la normalisation sur la détection des pics, nous avons utilisé MACS sur les données HepG2 brutes et normalisées. Si quelques différences ont été observées en comptant les pics (ceci étant dû aux fluctuations des sites moins enrichis), les différences globales d'amplitude ont été corrigées (Figure 3A). Nous avons utilisé chromHMM (Ernst et Kellis, 2012) pour comparer les attributions aux états de la chromatine avant et après normalisation sur des profils de neuf différents marqueurs d'histones dans neuf lignées cellulaires (Ernst et al., 2011). Cette comparaison a révélé des différences, petites mais néanmoins significatives, dans les annotations de l'état de la chromatine (2-7 %) de fenêtres génomiques. Il est à noter que les annotations de l'état de la chromatine de plusieurs gènes sont passées d'actives à suspendues et vice-versa ce qui, de manière générale, coïncidait avec leurs niveaux d'expression (exemple avec le gène MYO7A, Figure 3B).

Epimetheus a été utilisé pour évaluer les niveaux d'enrichissement relatifs du recrutement de H3K27me3, H3K4me3 et ARN polymérase II (PolII) dans des analyses de la différenciation des cellules F9 (Mendoza-Parra et al., manuscrit soumis). Les RCIs des données brutes des marqueurs H3K27me3 répressifs ont montré un enrichissement variable non-attendu au niveau de la région Hoxa au cours du temps. Cependant nous avons observé, après normalisation, le motif d'activation génique colinéaire, précédemment décrit (Kashyap et al., 2011 ; Montavon et Duboule, 2013) avec une perte progressive de marqueurs d'histone répressifs et un gain de marqueurs d'histone actifs, ainsi que le recrutement de PolII. Ces observations ont été confirmées par qPCR (Figure 3C). Epimetheus, par son approche sophistiquée et sa facilité d'utilisation, est un outil universel et flexible de normalisation de données épigénomique ou issues d'enrichissement (FAIRE/ATAC-Seq, PolII-Seq, MeDIP-Seq, etc.). Il combine plusieurs langages de programmation tels que Perl, C, et R, et son manuscrit a été soumis.

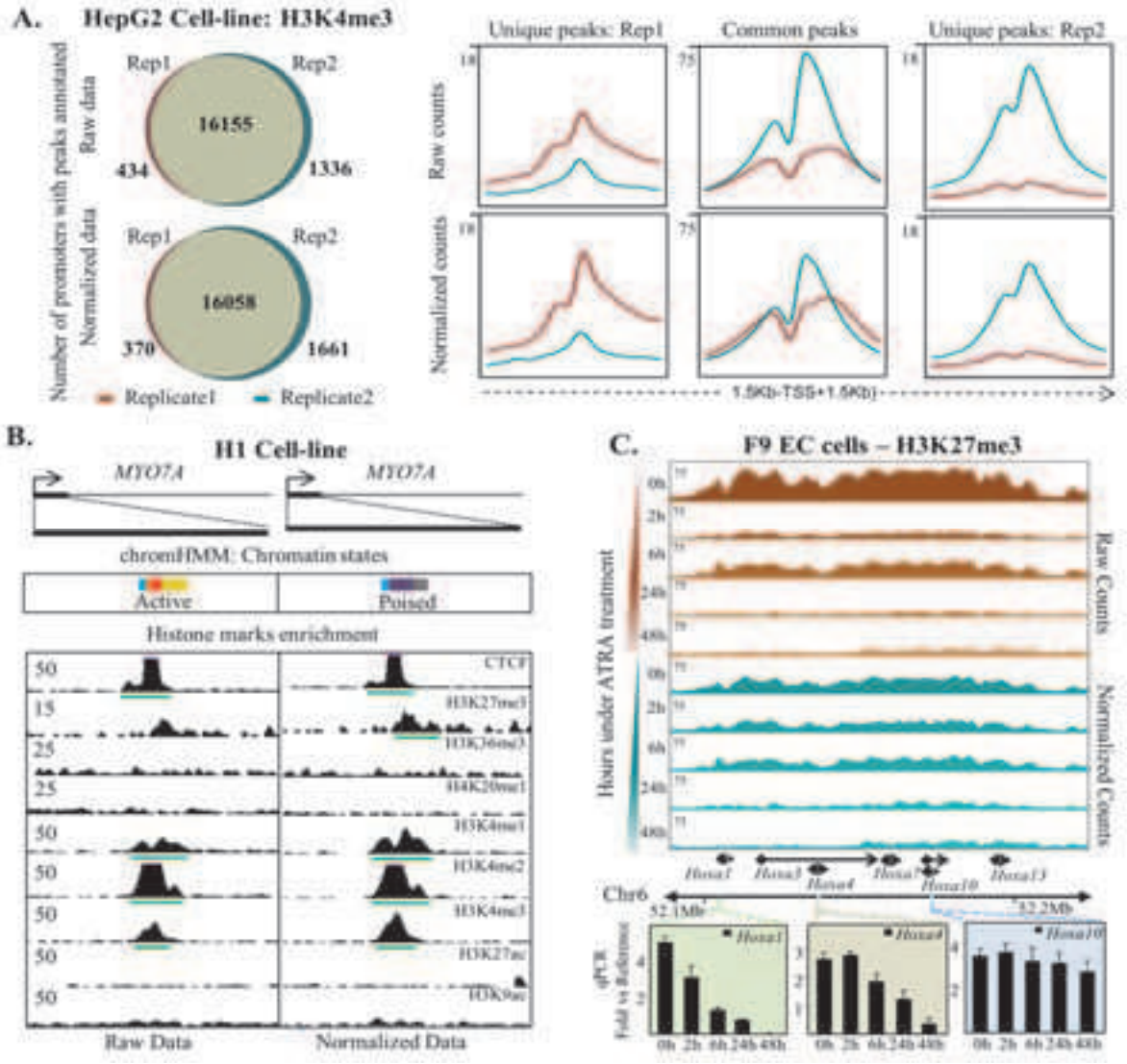


Figure 3. Effets de la normalisation. (A). De gauche à droite : diagrammes circulaires illustrant les évènements d'enrichissement au niveau du promoteur (par réplica ou en commun) avant et après normalisation. Les enrichissements de promoteurs annotés montrent que la normalisation donne des RCI plus similaires pour les pics communs et plus distinctes pour les enrichissements spécifiques du réplica. **(B)** Illustration du changement d'annotation de l'état de la chromatine pour *MYO7A* en utilisant la même expérience analysée ave ChromHMM. À noter que le promoteur de *MYO7A* a été annoté « actif » avec les données brutes puis « suspendu » avec les données normalisées, ce qui concorde avec l'absence d'expression génique [données ENCODE : ENCSR962TBJ]. **(C)** Profils d'intensité de l'enrichissement de H3K27me3 dans la région Hoxa pendant la différenciation par l'acide rétinoïque de cellules F9 de carcinome embryonnaire chez la souris. Contrairement aux données brutes, les données normalisées présentent une diminution graduelle du profil H3K27me3, ce qui concorde avec les résultats qPCR placés sous les profils.

B. Développement de pipelines et exploration de données sur le cancer

Il devient de plus en plus évident que les modifications épigénétiques telles que les changements dans la méthylation de l'ADN, la structure de la chromatine, les ARN non-codant, et l'organisation nucléaire, accompagnent la cancérogenèse lorsque ces modifications sont interrompues de façon aberrante (Berdasco et Esteller, 2010). Le chromosome X inactivé (Xi), ou corpuscule de Barr, est un très bon exemple d'un évènement épigénétique interrompu par le cancer. Bien que la disparition des corpuscules de Barr soit considérée comme un signe de cancer, la raison de cette disparition reste incertaine : cela peut être dû à une perte de matériel génétique ou à une instabilité épigénétique suivie d'une réactivation de la transcription. Les études de la chromatine et de la transcription dévoilent, au sein des cellules cancéreuses, des éléments épigénomiques modifiés, ainsi qu'une aberrante expression des gènes au niveau du chromosome X inactivé, dont plusieurs gènes impliqués dans le développement de cancer. Nous avons observé que les tumeurs et lignes cellulaires du sein présentent souvent une importante instabilité épigénétique du chromosome X inactivé, accompagné d'une organisation tridimensionnelle du noyau anormale et des perturbations de l'hétérochromatine, comme une augmentation en marqueurs euchromatiques et des distributions aberrantes de marqueurs répressifs comme H3K27me3 et la méthylation d'ADN promoteur.

Nous avons démontré que nombre de ces gènes sont réactivés de façon aberrante dans les tumeurs primaires du sein (MDA-MB-436, SK-BR-3 & ZR-75-1), puis nous avons démontré que l'instabilité épigénétique du chromosome X inactivé peut entraîner un mauvaise concentration en facteurs *X-linked.* Ainsi, notre étude propose la première analyse intégrée de chromosome X inactivé dans le contexte du cancer du sein et établit que son érosion épigénétique peut entraîner la disparition du corpuscule de Barr dans les cellules cancéreuses du sein. Ce travail offre de nouvelles idées et donne la possibilité d'utiliser le chromosome X inactivé comme une bio-marqueur épigénétique au niveau moléculaire et cytologique pour le cancer. Nous avons conduit une étude approfondie de l'organisation nucléaire, de l'état de la chromatine et de l'activité de la transcription de chromosome X inactivé dans des lignes cellulaires du cancer du sein et des échantillons de tumeurs primaires. Nous avons conclu qu'une cause fréquente de la disparition du corpuscule de Barr dans le cas du cancer du sein est la perturbation globale de son organisation nucléaire et de sa structure hétérochromatique. Enfin, les aberrations épigénomiques découvertes dans le Xi présent les cellules cancéreuses

du sein sont accompagnées par une degré significatif de réactivations de gènes sporadiques ce qui, dans certains cas, peut entraîner des concentrations aberrantes au niveau protéique.

Un pipeline a été développé pour intégrer et réaliser des analyses transcriptomiques et épigénomiques allèle-spécifiques. Les résultats du NGS-QC ont été utilisés pour déterminer la qualité des données avant analyse. Pour les données épigénomiques et transcriptomiques, une analyse allèle-spécifique a été menée en créant une référence de diploïdes grâce aux informations sur les SNP collectées de données SNP6, éxomiques et ChIP-Seq. Avant analyse, les données ChIP-Seq ont été normalisées avec Epimetheus pour ajuster les divergences entre échantillons. Une analyse génique d'annotations a été menée pour intégrer les annotations de piques et d'expressions géniques (nombre de SNP informatifs, profondeur de *read*, nombres de SNP homozygotes et hétérozygotes). Une moyenne arithmétique pondérée a été calculée pour chaque gène par calcul du déséquilibre allélique (DA) où le DA de chaque SNP a été pondéré par sa profondeur de *read*. Les gènes ont ainsi été classés par expressions mono-alléliques ou bi-alléliques. Cette étude a été publiée dans le journal *Genome Research* (http://genome.cshlp.org/content/early/2015/02/04/gr.185926.114.full.pdf+html).

De même, les gènes imprimés sont un autre exemple où les gènes sont marqués épigénétiquement ou éteints dans un allèle dépendant du parent d'origine. Il a été remarqué que les modifications de ces gènes imprimés, quand les gènes exprimés mono-alléliquement deviennent bi-alléliques ou totalement inactifs, peuvent entraîner des effets négatifs comme des tumeurs ou maladies (e.g. tumeur de Wilms, rhabdomyosarcome embryonnaire, etc.). Plusieurs études ont été publiées, indiquant qu'une perte d'empreinte ou une méthylation différentielle des gènes imprimés est lié au cancer du sein. Une analyse intégrative est en cours afin de déterminer le rôle des gènes imprimés dans le développement du cancer du sein.

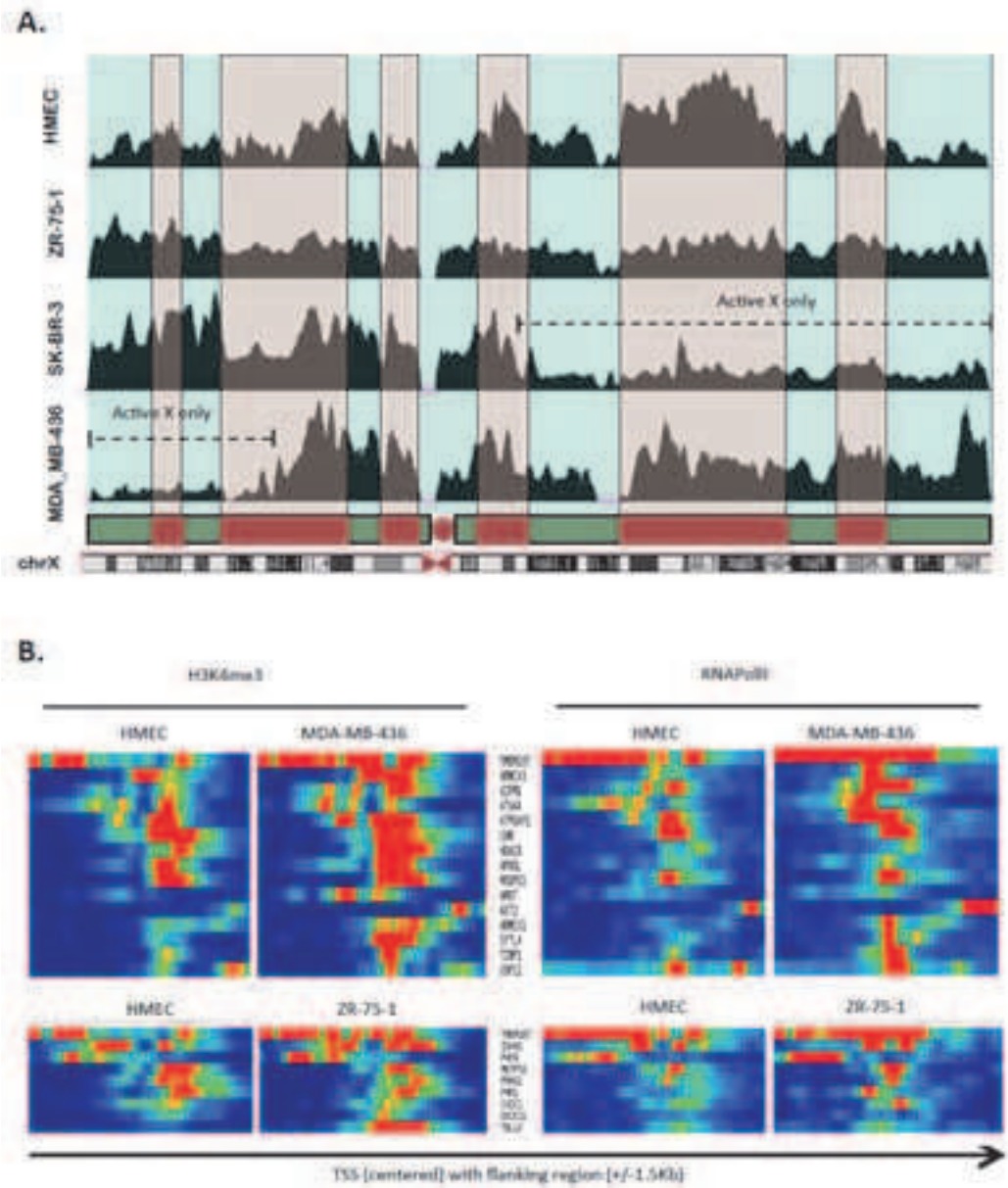


Figure 4. Analyse allèle-spécifique conduisant à l'identification de gènes échappant à l'inactivation du chromosome X spécifique de lignées cellulaires cancéreuses. (A) Le schéma de l'enrichissement en H3K27me3 sur l'ensemble du chromosome X montre une perte localisée de l'inactivation du chromosome X. Ces régions sont annotées « active X only » (chromosome X activé uniquement) et les deux principales pertes d'enrichissements H3K27me3 dans les lignes cellulaires ZR-75-1 et MDA-MB-436 sont en rouge. Les domaines en rouge et vert représentent, respectivement, les régions enrichies en H3K27me3 et H3K9me3, telles qu'identifies dans les cellules humaines normales (Chadwick, 2007). **(B)** Les augmentation de l'abondance en marqueurs d'histone H3K4me3 et recrutement d'ARN Pol II sont affichées sous forme d'*heat maps*, où les gènes échappés sont actifs dans les deux allèles, contrairement à ce qui est observé dans une ligne cellulaire normale (HMEC).

REFERENCES

Berdasco, M., and Esteller, M. (2010). Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry. Dev. Cell *19*, 698–711.

Bilanges, B., Varrault, A., Basyuk, E., Rodriguez, C., Mazumdar, A., Pantaloni, C., Bockaert, J., Theillet, C., Spengler, D., and Journot, L. (1999). Loss of expression of the candidate tumor suppressor gene ZAC in breast cancer cell lines and primary tumors. Oncogene *18*, 3979–3988.

Chaligné, R., and Heard, E. (2014). X-chromosome inactivation in development and cancer. FEBS Lett. *588*, 2514–2522.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Joyce, J.A., and Schofield, P.N. (1998). Genomic imprinting and cancer. Mol. Pathol. *51*, 185–190.

Kashyap, V., Gudas, L.J., Brenet, F., Funk, P., Viale, A., and Scandura, J.M. (2011). Epigenomic reorganization of the clustered Hox genes in embryonic stem cells induced by retinoic acid. J. Biol. Chem. *286*, 3250–3260.

Mendoza-Parra, M. a., Sankar, M., Walia, M., and Gronemeyer, H. (2012). POLYPHEMUS: R package for comparative analysis of RNA polymerase II ChIP-seq profiles by non-linear normalization. Nucleic Acids Res. *40*, e30.

Mendoza-Parra, M.A., Van Gool, W., Saleem, M.A.M., Ceschin, D.G., and Gronemeyer, H. (2013). A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Res. *41*.

Montavon, T., and Duboule, D. (2013). Chromatin organization and global regulation of Hox gene clusters. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *368*, 20120367.

Nair, N.U., Das Sahu, A., Bucher, P., and Moret, B.M.E. (2012). Chipnorm: A statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. PLoS One *7*, e39573.

Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S.H., and Waxman, D.J. (2012). MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol. *13*, R16.

COMMUNICATIONS SCIENTIFIQUES

Contribution à un chapitre de livre :

Marco A. Mendoza-Parra, Mohamed Saleem MA, Matthias Blum, Pierre-Etienne Cholley and Hinrich Gronemeyer; *NGS-QC Generator: A Quality control system for ChIP-seq and related deep sequencing-generated datasets*; **Statistical Genomics : Methods and Protocols; Springer Book Series**; (à paraître).

Articles:

Mohamed Ashick Mohamed Saleem, Marco Antonio Mendoza-Parra, Pierre-Etienne Cholley and Hinrich Gronemeyer; *Epimetheus - A multi-profile normalizer for epigenomic sequencing data*; (soumis).

Marco Antonio Mendoza-Parra, Wouter Van Gool, Mohamed Saleem MA, Danilo Guillermo Ceschin and Hinrich Gronemeyer; *A quality control system for profiles obtained by massive parallel sequencing*; **Nucleic Acid Research**; 2013 Nov 1; 41(21):e196. doi: 10.1093/nar/gkt829. [IF 8.808] (BREVETE).

Ronan Chaligné, Tatiana Popova, Marco Antonio Mendoza Parra, Mohamed Ashick Mohamed Saleem, David Gentien, Kristen Ban, Tristan Piolot, Olivier Leroy, Odette Mariani, Hinrich Gronemeyer, Anne Vincent-Salomon, Marc-Henri Stern and Edith Heard; *The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer*; **Genome Research**; 2015 Apr, 25(4): 488-503. [IF 13.852].

Marco Antonio Mendoza-Parra, Valeriya Malysheva, Mohamed Ashick Mohamed Saleem and Hinrich Gronemeyer; *Reconstructing divergent retinoid-induced cell fate-regulatory programs in stem cells*; **Molecular Systems Biology** (en révision).

Valeriya Malysheva, Marco Antonio Mendoza-Parra, Mohamed Ashick Mohamed Saleem and Hinrich Gronemeyer; *Reconstructing gene regulatory networks of tumorigenesis*; **Cancer research** (en révision).

Appendices

Contributed to a chapter in book

Marco A. Mendoza-Parra, Mohamed Saleem MA, Matthias Blum, Pierre-Etienne Cholley and Hinrich Gronemeyer; *NGS-QC Generator: A Quality control system for ChIP-seq and related deep sequencing-generated datasets;* **Statistical Genomics : Methods and Protocols; Springer Book Series;** (in press).

Publications

Marco Antonio Mendoza-Parra, Wouter Van Gool, Mohamed Saleem MA, Danilo Guillermo Ceschin and Hinrich Gronemeyer; *A quality control system for profiles obtained by massive parallel sequencing*; **Nucleic Acid Research**; 2013 Nov 1; 41(21):e196. doi: 10.1093/nar/gkt829. **[IF 8.808] (PATENTED).**

Mohamed Ashick Mohamed Saleem, Marco Antonio Mendoza-Parra, Pierre-Etienne Cholley and Hinrich Gronemeyer; *Epimetheus - A multi-profile normalizer for epigenomic sequencing data*; (in submission).

Ronan Chaligné, Tatiana Popova, Marco Antonio Mendoza Parra, Mohamed Ashick Mohamed Saleem ,David Gentien, Kristen Ban, Tristan Piolot, Olivier Leroy, Odette Mariani, Hinrich Gronemeyer, Anne Vincent-Salomon, Marc-Henri Stern and Edith Heard ; *The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer***; Genome Research**; 2015 Apr, 25(4): 488-503. **[IF 13.852].**

Marco Antonio Mendoza-Parra, Valeriya Malysheva, Mohamed Ashick Mohamed Saleem and Hinrich Gronemeyer; *Reconstructing divergent retinoid-induced cell fate-regulatory programs in stem cells*; **Molecular Systems Biology** (in review).

Valeriya Malysheva, Marco Antonio Mendoza-Parra, Mohamed Ashick Mohamed Saleem and Hinrich Gronemeyer; *Reconstructing gene regulatory networks of tumorigenesis*; **Cancer research** (in review).

Reconstructing divergent retinoid-induced cell fate-regulatory programs in stem cells

Marco-Antonio Mendoza-Parra*, Valeriya Malysheva, Mohamed Ashick Mohamed Saleem and Hinrich Gronemeyer*

Department of Functional Genomics and Cancer, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Equipe Labellisée Ligue Contre le Cancer, Centre National de la Recherche Scientifique UMR 7104, Institut National de la Santé et de la Recherche Médicale U964, University of Strasbourg, Illkirch, France.

*Corresponding authors:

Marco Antonio Mendoza-Parra
E-mail: marco@igbmc.fr
Hinrich Gronemeyer
E-mail: hg@igbmc.u-strasbg.fr
Phone: +(33) 3 88 65 34 73
Fax: +(33) 3 88 65 34 37

Running title: gene regulatory programs in RA-induced Cell fate transitions

Key words: RA-induced cell differentiation / Gene regulatory programs / functional Genomics / ChIP-seq / cell-fate transition programs

ABSTRACT (175 words)

Cell lineages, which shape body architecture and specify cell functions, derive from the integration of a plethora of cell intrinsic and extrinsic signals. These signals trigger a multiplicity of decisions at several levels to modulate the activity of dynamic gene regulatory networks (GRNs), which ensure both general and cell-specific functions within a given lineage, thereby establishing cell fates. Cellular 'differentiation' models conserved certain sequences of events within a cell fate acquisition process. These models are important homogenous experimental systems to study the complex interplay between extrinsic signals and alterations at different levels in the gene regulatory hierarchies from a systems biology perspective. Here we have dissected the GRNs involved in the neuronal or endodermal cell-fate specification responses to retinoic acid (RA) in two stem cell models by integrating dynamic RXRa binding, chromatin accessibility and promoter epigenetic status with the transcriptional activity inferred from RNA polymerase II mapping and

transcription profiling. Our data reveals how RA induces a network of transcription factors which direct the temporal organization of cognate GRNs, thereby driving neuronal/endodermal cell-fate specification. By applying CRISPR/Cas9 editing approaches, we have first verified the relevance of early induced neuronal-specific factors, but in addition we have redirected cell-fate specification from endodermal to neuronal commitment, demonstrating that a systems view of cell fate specification provides the necessary insight for directional intervention. These results are encouraging in view of cell/tissue engineering for regenerative medicine.

Reconstructing gene regulatory networks of tumorigenesis

Valeriya Malysheva, Marco-Antonio Mendoza-Parra, Mohamed-Ashick M. Saleem and Hinrich Gronemeyer*

Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Equipe Labellisée Ligue Contre le Cancer, Centre National de la Recherche Scientifique UMR 7104, InstitutNational de la Santé et de la Recherche Médicale U964, University of Strasbourg, Illkirch, France.

* To whom correspondence should be addressed.

Hinrich Gronemeyer
Tel: +(33) 3 88 65 34 73
Fax: +(33) 3 88 65 34 37
E-mail: hg@igbmc.u-strasbg.fr

Abstract

The mechanistic links between transcription factors and the epigenetic landscape, which coordinate the deregulation of gene networks during cell transformation are largely unknown. We used an isogenic model of stepwise tumorigenic transformation of human primary cells to monitor the progressive deregulation of gene networks upon immortalization and oncogene-induced transformation. By combining transcriptome and epigenome data for each step during transformation and by integrating transcription factor (TF) - target gene associations, we identified 142 TFs and 24 chromatin remodelers/modifiers (CRMs), which are preferentially associated with specific co-expression paths that originate from deregulated gene programming during tumorigenesis. These TFs are involved in the regulation of divers processes, including cell differentiation, immune response and establishment/modification of the epigenome. Unexpectedly, the analysis of chromatin state dynamics revealed patterns that distinguish groups of genes, which are not only co-regulated but also functionally related. Further decortication of TF targets enabled us to define potential key regulators of cell transformation, which are engaged in RNA metabolism and chromatin remodeling. Our study suggests a direct implication of CRMs in oncogene-induced tumorigenesis and identifies new CRMs

involved in this process. This is the first comprehensive view of gene regulatory networks that are altered during the process of stepwise human cellular tumorigenesis in a virtually isogenic system.

References

Aleksic, J., Carl, S.H., and Frye, M. (2014). Beyond library size: a field guide to NGS normalization. bioRxiv 006403.

Allegra, P., Sterner, R., Clayton, D.F., and Allfrey, V.G. (1987). Affinity chromatographic purification of nucleosomes containing transcriptionally active DNA sequences. J. Mol. Biol. *196*, 379–388.

Allison, K. a., Kaikkonen, M.U., Gaasterland, T., and Glass, C.K. (2014). Vespucci: A system for building annotated databases of nascent transcripts. Nucleic Acids Res. *42*, 2433–2447.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Arzate-Mejía, R.G., Valle-García, D., and Recillas-Targa, F. (2011). Signaling epigenetics: Novel insights on cell signaling and epigenetic regulation. IUBMB Life *63*, 907–921.

Augui, S., Nora, E.P., and Heard, E. (2011). Regulation of X-chromosome inactivation by the X-inactivation centre. Nat. Rev. Genet. *12*, 429–442.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput. Biol. *9*, e1003326.

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. *21*, 381–395.

Bannister, A.J., Schneider, R., and Kouzarides, T. (2002). Histone Methylation: Dynamic or Static? Cell *109*, 801–806.

Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. a, Jiang, H., and Feng, G. (2014). Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. Lib. Acad. *13*, 67–82.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stŕmberg, M.P., and Marth, G.T. (2011). Bamtools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691–1692.

BARR, M.L., and BERTRAM, E.G. (1949). A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. Nature *163*, 676.

BARR, M.L., and MOORE, K.L. (1957). Chromosomes, sex chromatin, and cancer. Proc. Can. Cancer Conf. *2*, 3–16.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Baylin, S.B., and Jones, P. a. (2011). A decade of exploring the cancer epigenome — biological and translational implications. Nat. Rev. Cancer *11*, 726–734.

Belton, J.-M., McCord, R.P., Gibcus, J., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. Methods *58*, 10.1016/j.ymeth.2012.05.001.

Berger, S.L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics An operational definition of epigenetics. 781–783.

Bilanges, B., Varrault, A., Basyuk, E., Rodriguez, C., Mazumdar, A., Pantaloni, C., Bockaert, J., Theillet, C., Spengler, D., and Journot, L. (1999). Loss of expression of the candidate tumor suppressor gene ZAC in breast cancer cell lines and primary tumors. Oncogene *18*, 3979–3988.

Bischof, O., and Martínez-Zamudio, R.I. (2015). MicroRNAs and lncRNAs in senescence: A re-view. IUBMB Life *67*, 255–267.

Blankenberg, D., Kuster, G. Von, Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2001). Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. In Current Protocols in Molecular Biology, (John Wiley & Sons, Inc.),.

Bolger, a. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bonasio, R., Tu, S., and Reinberg, D. (2010). Molecular signals of epigenetic states. Science *330*, 612–616.

Bonfield, J.K., and Mahoney, M. V. (2013). Compression of FASTQ and SAM Format Sequencing Data. PLoS One *8*, e59190.

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., and Delorenzi, M. (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. *24*, 1157–1168.

Bose, M.E., McConnell, K.H., Gardner-Aukema, K.A., Müller, U., Weinreich, M., Keck, J.L., and Fox, C.A. (2004). The Origin Recognition Complex and Sir4 Protein Recruit Sir1p to Yeast Silent Chromatin through Independent Interactions Requiring a Common Sir1p Domain. Mol. Cell. Biol. *24*, 774–786.

Bottomley, M.J. (2004). Structures of protein domains that create or recognize histone modifications. EMBO Rep. *5*, 464–469.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536–2539.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527–542.

Brownell, J.E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D.G., Roth, S.Y., and Allis, C.D. (1996). Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell *84*, 843–851.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: Network Biology Applied to Stem Cell Engineering. Cell *158*, 903–915.

Di Cerbo, V., and Schneider, R. (2013). Cancers with wrong HATs: the impact of acetylation. Brief. Funct. Genomics *12*, 231–243.

Ceschin, D.G., Walia, M., Wenk, S.S., Duboé, C., Gaudon, C., Xiao, Y., Fauquier, L., Sankar, M., Vande, L., and Gronemeyer, H. (2011). Methylation specifies distinct estrogen-induced binding site repertoires of CBP to chromatin. Genes Dev. *25*, 1132–1146.

Chadwick, B. (2007). Variation in Xi chromatin organization and correlation of the H3K27me3 chromatin territories to transcribed sequences by microarray analysis. Chromosoma *116*, 147–157.

Chaligné, R., and Heard, E. (2014). X-chromosome inactivation in development and cancer. FEBS Lett. *588*, 2514–2522.

Chaligné, R., Popova, T., Mendoza-Parra, M.-A., Saleem, M.-A.M., Gentien, D., Ban, K., Piolot, T., Leroy, O., Mariani, O., Gronemeyer, H., et al. (2015). The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer. Genome Res. gr.185926.114 – .

Cheung, M.-S., Down, T. a., Latorre, I., and Ahringer, J. (2011). Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Res. *39*, e103–e103.

Cho, S.Y., Chai, J.C., Park, S.J., Seo, H., Sohn, C.-B., and Lee, Y.S. (2013). EPITRANS: A database that integrates epigenome and transcriptome data. Mol. Cells *36*, 472–475.

Choi, J.D., and Lee, J.-S. (2013). Interplay between Epigenetics and Genetics in Cancer. Genomics Inform. *11*, 164–173.

Chowdhury, D., Keogh, M.-C., Ishii, H., Peterson, C.L., Buratowski, S., and Lieberman, J. (2005). γ-H2AX Dephosphorylation by Protein Phosphatase 2A Facilitates DNA Double-Strand Break Repair. Mol. Cell *20*, 801–809.

Claude-Taupin, A., Boyer-Guittaut, M., Delage-Mourroux, R., and Hervouet, E. (2015). Use of epigenetic modulators as a powerful adjuvant for breast cancer therapies. Methods Mol. Biol. *1238*, 487–509.

Cock, P.J. a., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. *38*, 1767–1771.

Cohen, I., Poreba, E., Kamieniarz, K., and Schneider, R. (2011). Histone Modifiers in Cancer: Friends or Foes? Genes Cancer *2*, 631–647.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Collins, L.J., Schönfeld, B., and Chen, X.S. (2011). The Epigenetics of Non-coding RNA. Handb. Epigenetics 49–61.

ConsortiumInternational, H.G.S. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Dai, M., Thompson, R.C., Maher, C., Contreras-Galindo, R., Kaplan, M.H., Markovitz, D.M., Omenn, G., and Meng, F. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics *11*, S7.

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A. a, del Angel, G., Rivas, M. a, Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast Computation and Applications of Genome Mappability. PLoS One *7*, e30377.

Durbin, R.M., Altshuler, D.L., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., Donnelly, P., et al. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Ehrlich, M., Gama-Sosa, M. a., Huang, L.H., Midgett, R.M., Kuo, K.C., Mccune, R. a., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res. *10*, 2709–2721.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Ernst, J., Vainas, O., Harbison, C.T., Simon, I., and Bar-Joseph, Z. (2007). Reconstructing dynamic regulatory maps. Mol. Syst. Biol. *3*.

Ernst, J., Beg, Q.K., Kay, K. a., Balázsi, G., Oltvai, Z.N., and Bar-Joseph, Z. (2008). A Semi-Supervised Method for Predicting Transcription Factor–Gene Interactions in Escherichia coli. PLoS Comput. Biol. *4*, e1000044.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nat. Rev. Genet. *8*, 286–298.

Esteller, M. (2008). Epigenetics in Cancer. N. Engl. J. Med. *358*, 1148–1159.

Falkenberg, K.J., and Johnstone, R.W. (2014). Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. Nat Rev Drug Discov *13*, 673–691.

Fenley, A.T., Adams, D. a., and Onufriev, A. V. (2010). Charge state of the globular histone core controls stability of the nucleosome. Biophys. J. *99*, 1577–1585.

Filippakopoulos, P., and Knapp, S. (2014). Targeting bromodomains: epigenetic readers of lysine acetylation. Nat. Rev. Drug Discov. *13*, 337–356.

Fraga, M.F., and Esteller, M. (2002). DNA methylation: a profile of methods and applications. Biotechniques *33*, 632,634,636–649.

Garcia-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Gotz, S., Tarazona, S., Dopazo, J., Meyer, T.F., and Conesa, a. (2012). Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics *28*, 2678–2679.

Giresi, P.G., Kim, J., Mcdaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. 877–885.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546,563–567.

Goldknopf, I.L., Taylor, C.W., Baum, R.M., Yeoman, L.C., Olson, M.O., Prestayko, A.W., and Busch, H. (1975). Isolation and characterization of protein A24, a “histone-like” non-histone chromosomal protein. J. Biol. Chem. *250* , 7182–7187.

Goll, M.G., and Bestor, T.H. (2005). EUKARYOTIC CYTOSINE METHYLTRANSFERASES. Annu. Rev. Biochem. *74*, 481–514.

Göndör, A., and Ohlsson, R. (2009). Chromosome crosstalk in three dimensions. Nature *461*, 212–217.

Henry, K.W., Wyce, A., Lo, W.S., Duggan, L.J., Emre, N.C.T., Kao, C.F., Pillus, L., Shilatifard, A., Osley, M.A., and Berger, S.L. (2003). Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-associated Ubp8. Genes Dev. *17*, 2648–2663.

Hödl, M., and Basler, K. (2012). Transcription in the absence of histone H3.2 and H3K4 methylation. Curr. Biol. *22*, 2253–2257.

Hoffmann, I., Roatsch, M., Schmitt, M.L., Carlino, L., Pippel, M., Sippl, W., and Jung, M. (2012). The role of histone demethylases in cancer therapy. Mol. Oncol. *6*, 683–703.

Holliday, R. (1987). Epigenetic Defects. Science (80-.). *238*, 163–170.

Holoch, D., and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. Nat. Rev. Genet. *16*, 71–84.

Hong, Y.K., Ontiveros, S.D., and Strauss, W.M. (2000). A revision of the human XIST gene organization and structural comparison with mouse Xist. Mamm. Genome *11*, 220–224.

Houlès, T., Rodier, G., Le Cam, L., Sardet, C., and Kirsh, O. (2015). Description of an optimized ChIP-seq analysis pipeline dedicated to genome wide identification of E4F1 binding sites in primary and transformed MEFs. Genomics Data *5*, 368–370.

Hurd, P.J., and Nelson, C.J. (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. Briefings Funct. Genomics Proteomics *8*, 174–183.

Jacinto, F. V., Ballestar, E., and Esteller, M. (2008). Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA methylome. Biotechniques *44*, 35–43.

Jin, S.G., Jiang, Y., Qiu, R., Rauch, T. a., Wang, Y., Schackert, G., Krex, D., Lu, Q., and Pfeifer, G.P. (2011). 5-hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations. Cancer Res. *71*, 7360–7365.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.

Jones, P. a (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Jones, D.C., Ruzzo, W.L., Peng, X., and Katze, M.G. (2012). Compression of next-generation sequencing reads aided by highly efficient de novo assembly. Nucleic Acids Res. *40*, e171–e171.

Joo, H.Y., Jones, A., Yang, C., Zhai, L., Smith IV, A.D., Zhang, Z., Chandrasekharan, M.B., Sun, Z.W., Renfrow, M.B., Wang, Y., et al. (2011). Regulation of histone H2A and H2B deubiquitination and xenopus development by USP12 and USP46. J. Biol. Chem. *286*, 7190–7201.

Joyce, J.A., and Schofield, P.N. (1998). Genomic imprinting and cancer. Mol. Pathol. *51*, 185–190.

Kaikkonen, M.U., Lam, M.T.Y., and Glass, C.K. (2011). Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc. Res. *90*, 430–440.

Karmodiya, K., Anamika, K., Muley, V., Pradhan, S.J., Bhide, Y., and Galande, S. (2014). Camello , a novel family of Histone Acetyltransferases that acetylate histone H4 and is essential for zebrafish. 1–9.

Kashyap, V., Gudas, L.J., Brenet, F., Funk, P., Viale, A., and Scandura, J.M. (2011). Epigenomic reorganization of the clustered Hox genes in embryonic stem cells induced by retinoic acid. J. Biol. Chem. *286*, 3250–3260.

Keogh, M.-C., Kim, J.-A., Downey, M., Fillingham, J., Chowdhury, D., Harrison, J.C., Onishi, M., Datta, N., Galicia, S., Emili, A., et al. (2006). A phosphatase complex that dephosphorylates γH2AX regulates DNA damage checkpoint recovery. Nature *439*, 497–501.

Kim, K.-P., and Schöler, H.R. (2014). CellNet—Where Your Cells Are Standing. Cell *158*, 699–701.

Klein, H.U., Schäfer, M., Porse, B.T., Hasemann, M.S., Ickstadt, K., and Dugas, M. (2014). Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. Bioinformatics *30*, 1154–1162.

Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*, 693–705.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Landt, S., and Marinov, G. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome … 1813–1831.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. *11*, 204–220.

Lee, J.-S., and Shilatifard, A. (2007). A site to remember: H3K36 methylation a mark for histone deacetylation. Mutat. Res. Mol. Mech. Mutagen. *618*, 130–134.

Lee, J.T., and Bartolomei, M.S. (2013). X-Inactivation, Imprinting, and Long Noncoding RNAs in Health and Disease. Cell *152*, 1308–1323.

Lee, C., Chiu, Y., Wang, L., Kuo, Y., Chuang, E.Y., Lai, L.-C., and Tsai, M. (2013). Common applications of next-generation sequencing technologies in genomic research. Transl. Cancer Res. *2*, 33–45.

Ley, T.J., Ding, L., Walter, M.J., McLellan, M.D., Lamprecht, T., Larson, D.E., Kandoth, C., Payton, J.E., Baty, J., Welch, J., et al. (2010). DNMT3A Mutations in Acute Myeloid Leukemia. N. Engl. J. Med. *363*, 2424–2433.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. arXiv *00*, 3.

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. *11*, 473–483.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell *133*, 523–536.

Lopes Novo, C., and Rugg-Gunn, P.J. (2015). Chromatin organization in pluripotent cells: emerging approaches to study and disrupt function. Brief. Funct. Genomics elv029.

Di Lorenzo, A., and Bedford, M.T. (2011). Histone arginine methylation. FEBS Lett. *585*, 2024–2031.

Loyola, A., LeRoy, G., Wang, Y.H., and Reinberg, D. (2001). Reconstitution of recombinant chromatin establishes a requirement for histone-tail modifications during chromatin assembly and transcription. Genes Dev. *15*, 2837–2851.

Luedi, P.P., Dietrich, F.S., Weidman, J.R., Bosko, J.M., Jirtle, R.L., and Hartemink, A.J. (2007). Computational and experimental identification of novel human imprinted genes. Genome Res. *17*, 1723–1730.

Luger, K., Mäder, a W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251–260.

Lujambio, A., Ropero, S., Ballestar, E., Fraga, M.F., Cerrato, C., Setién, F., Casado, S., Suarez-Gauthier, A., Sanchez-Cespedes, M., Gitt, A., et al. (2007). Genetic unmasking of an epigenetically silenced microRNA in human cancer cells. Cancer Res. *67*, 1424–1429.

LYON, M.F. (1961). Gene Action in the X-chromosome of the Mouse (Mus musculus L.). Nature *190*, 372–373.

Malecová, B., and Morris, K. V (2010). Transcriptional gene silencing through epigenetic changes mediated by non-coding RNAs. Curr. Opin. Mol. Ther. *12*, 214–222.

Marinov, G.K., Kundaje, a., Park, P.J., and Wold, B.J. (2014). Large-Scale Quality Analysis of Published ChIP-seq Data. G3: Genes|Genomes|Genetics *4*, 209–223.

Martens, J.H. a., Stunnenberg, H.G., and Logie, C. (2011). The Decade of the Epigenomes? Genes Cancer *2*, 680–687.

Marx, V. (2013). Biology: The big challenges of big data. Nature *498*, 255–260.

Mendoza-Parra, M.A., and Gronemeyer, H. (2014). Assessing quality standards for ChIP-seq and related massive parallel sequencing-generated datasets: When rating goes beyond avoiding the crisis. Genomics Data *2*, 268–273.

Mendoza-Parra, M.A., Van Gool, W., Saleem, M.A.M., Ceschin, D.G., and Gronemeyer, H. (2013a). A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Res. *41*.

Mendoza-Parra, M.-A., Nowicka, M., Van Gool, W., and Gronemeyer, H. (2013b). Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. BMC Genomics *14*, 834.

Mendoza-Parra, M.-A.A., Van Gool, W., Mohamed Saleem, M.A., Ceschin, D.G., Gronemeyer, H., Saleem, M.A.M., Ceschin, D.G., and Gronemeyer, H. (2013c). A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Res. *41*, e196.

Mersfelder, E.L., and Parthun, M.R. (2006). The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. Nucleic Acids Res. *34*, 2653–2662.

Meyer, C. a., and Liu, X.S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat. Rev. Genet. *15*, 709–721.

Montavon, T., and Duboule, D. (2013). Chromatin organization and global regulation of Hox gene clusters. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *368*, 20120367.

Moorhead, G.B.G., Trinkle-Mulcahy, L., and Ulke-Lemee, a (2007). Emerging roles of nuclear protein phosphatases. Nat. Rev. Mol. Cell Biol. *8*, 234–244.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques *45*, 81–94.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth *5*, 621–628.

Murphy, S.K., and Jirtle, R.L. (2003). Imprinting evolution and the price of silence. BioEssays *25*, 577–588.

Murray, G.I., Taylor, M.C., McFadyen, M.C.E., McKay, J.A., Greenlee, W.F., Burke, M.D., and Melvin, W.T. (1997). Tumor-specific Expression of Cytochrome P450 CYP1B1. Cancer Res. *57* , 3026–3031.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Sci. *320* , 1344–1349.

Nair, N.U., Das Sahu, A., Bucher, P., and Moret, B.M.E. (2012). Chipnorm: A statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. PLoS One *7*, e39573.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, a., Takahashi, H., et al. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. *39*, e90–e90.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Abigail, W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Evan, E., et al. (2010). Targeted Capture and Massicely Parallel Sequencing of twelve human exomes. Nature *461*, 272–276.

Nicassio, F., Corrado, N., Vissers, J.H. a., Areces, L.B., Bergink, S., Marteijn, J. a., Geverts, B., Houtsmuller, A.B., Vermeulen, W., Di Fiore, P.P., et al. (2007). Human USP3 Is a Chromatin Modifier Required for S Phase Progression and Genome Stability. Curr. Biol. *17*, 1972–1977.

Nickel, B.E., and Davie, J.R. (1989). Structure of polyubiquitinated histone H2A. Biochemistry *28*, 964–968.

Nickel, B.E., Allis, C.D., and Davie, J.R. (1989). Ubiquitinated histone H2B is preferentially located in transcriptionally active chromatin. Biochemistry *28*, 958–963.

Nicolae, M., Pathak, S., and Rajasekaran, S. (2015). LFQC : A lossless compression algorithm for FASTQ files. Bioinformatics 1–8.

Oehme, I., Deubzer, H.E., Wegener, D., Pickert, D., Linke, J.P., Hero, B., Kopp-Schneider, A., Westermann, F., Ulrich, S.M., Von Deimling, A., et al. (2009). Histone deacetylase 8 in neuroblastoma tumorigenesis. Clin. Cancer Res. *15*, 91–99.

Ogawa, O., Eccles, M.R., Szeto, J., McNoe, L.A., Yun, K., Maw, M.A., Smith, P.J., and Reeve, A.E. (1993). Relaxation of insulin-like growth factor II gene imprinting implicated in Wilms' tumour. Nature *362*, 749–751.

Ohhata, T., and Wutz, A. (2013). Reactivation of the inactive X chromosome in development and reprogramming. Cell. Mol. Life Sci. *70*, 2443–2461.

OHNO, S., and HAUSCHKA, T.S. (1960). Allocycly of the X-chromosome in tumors and normal tissues. Cancer Res. *20*, 541–545.

Orlando, D.A., Chen, M.W., Brown, V.E., Solanki, S., Choi, Y.J., Olson, E.R., Fritz, C.C., Bradner, J.E., and Guenther, M.G. (2014). Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome. Cell Rep. *9*, 1163–1170.

Pageau, G.J., Hall, L.L., Ganesan, S., Livingston, D.M., and Lawrence, J.B. (2007). The disappearing Barr body in breast and ovarian cancers. Nat. Rev. Cancer *7*, 628–633.

Patel, R.K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One *7*, e30619.

Pauli, A., Rinn, J.L., and Schier, A.F. (2011). Non-coding RNAs as regulators of embryogenesis. Nat. Rev. Genet. *12*, 136–149.

Paulson, J.R., Patzlaff, J.S., and Vallis, a J. (1996). Evidence that the endogenous histone H1 phosphatase in HeLa mitotic chromosomes is protein phosphatase 1, not protein phosphatase 2A. J. Cell Sci. *109 (Pt 6*, 1437–1447.

Pengelly, A.R., Copur, Ö., Jäckle, H., Herzig, A., and Müller, J. (2013). A Histone Mutant Reproduces the Phenotype Caused by Loss of Histone-Modifying Factor Polycomb. Sci. *339* , 698–699.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nat. Methods *6*, S22–S32.

Pérez-Novo, C.A., Claeys, C., Speleman, F., Van Cauwenberge, P., Bachert, C., and Vandesompele, J. (2005). Impact of RNA quality on reference gene expression stability. Biotechniques *39*, 52–56.

Perissi, V., Jepsen, K., Glass, C.K., and Rosenfeld, M.G. (2010). Deconstructing repression: evolving models of co-repressor action. Nat Rev Genet *11*, 109–123.

Pfeifer, G.P., Kadam, S., and Jin, S.-G. (2013). 5-Hydroxymethylcytosine and Its Potential Roles in Development and Cancer. Epigenetics Chromatin *6*, 10.

Poptsova, M.S., Il'icheva, I. a., Nechipurenko, D.Y., Panchenko, L. a., Khodikov, M. V., Oparina, N.Y., Polozov, R. V., Nechipurenko, Y.D., and Grokhovsky, S.L. (2014). Non-random DNA fragmentation in next-generation sequencing. Sci. Rep. *4*, 1–6.

Portal, M.M., Pavet, V., Erb, C., and Gronemeyer, H. (2014). Human cells contain natural double-stranded RNAs with potential regulatory functions. Nat. Struct. Mol. Biol. *22*, 89–97.

Quail, M., Smith, M.E., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics *13*, 1.

Quina, a. S., Buschbeck, M., and Di Croce, L. (2006). Chromatin structure and epigenetics. Biochem. Pharmacol. *72*, 1563–1569.

Raabe, F.J., and Spengler, D. (2013). Epigenetic Risk Factors in PTSD and Depression. Front. Psychiatry *4*, 80.

Ramachandran, P., Palidwor, G.A., and Perkins, T.J. (2015). BIDCHIPS : bias decomposition and removal from ChIP - seq data clarifies true binding signal and its functional correlates. Epigenetics Chromatin 1–16.

Rhee, H.S., and Pugh, B.F. (2012). ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. Curr. Protoc. Mol. Biol. *0 21*, 10.1002/0471142727.mb2124s100.

Rizzo, J.M., and Buck, M.J. (2012). Key Principles and Clinical Applications of "Next-Generation" DNA Sequencing. Cancer Prev. Res. *5*, 887–900.

Robzyk, K., Recht, J., and Osley, M.A. (2000). Rad6-dependent ubiquitination of histone H2B in yeast. Science *287*, 501–504.

Rodd, A.L., Ververis, K., and Karagiannis, T.C. (2012). Current and Emerging Therapeutics for Cutaneous T-Cell Lymphoma: Histone Deacetylase Inhibitors. Lymphoma *2012*, 1–10.

Romanoski, C.E., Glass, C.K., Stunnenberg, H.G., Wilson, L., and Almouzni, G. (2015). Epigenomics: Roadmap for regulation. Nature *518*, 314–316.

Rossetto, D., Avvakumov, N., and Côté, J. (2012). Histone phosphorylation: a chromatin modification involved in diverse nuclear events. Epigenetics *7*, 1098–1108.

Rye, M.B., Saetrom, P., and Drablos, F. (2011). A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. Nucleic Acids Res. *39*, e25–e25.

Sadikovic, B., Al-Romaih, K., Squire, J. a, and Zielenska, M. (2008). Cause and consequences of genetic and epigenetic alterations in human cancer. Curr. Genomics *9*, 394–408.

Saito, Y., Liang, G., Egger, G., Friedman, J.M., Chuang, J.C., Coetzee, G. a., and Jones, P. a. (2006). Specific activation of microRNA-127 with downregulation of the proto-oncogene BCL6 by chromatin-modifying drugs in human cancer cells. Cancer Cell *9*, 435–443.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. Nature *265*, 687–695.

Schreiber, S.L., Schreiber, S.L., Bernstein, B.E., and Bernstein, B.E. (2002). Signaling network model of chromatin. Cell *111*, 771–778.

Schulz, M.H., Devanny, W.E., Gitter, A., Zhong, S., Ernst, J., and Bar-Joseph, Z. (2012). DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. BMC Syst. Biol. *6*, 104.

Shahbazian, M.D., and Grunstein, M. (2007). Functions of Site-Specific Histone Acetylation and Deacetylation. Annu. Rev. Biochem. *76*, 75–100.

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res. Int. *2014*, 309650.

Shankaranarayanan, P., Mendoza-Parra, M.-A., Walia, M., Wang, L., Li, N., Trindade, L.M., and Gronemeyer, H. (2011). Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. Nat. Methods *8*, 565–567.

Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S.H., and Waxman, D.J. (2012). MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol. *13*, R16.

Sharma, S., Kelly, T.K., and Jones, P. a. (2010). Epigenetics in cancer. Carcinogenesis *31*, 27–36.

Sims, R.J., and Reinberg, D. (2006). Histone H3 Lys 4 methylation: Caught in a bind? Genes Dev. *20*, 2779–2786.

Smith, A.D., Xuan, Z., and Zhang, M.Q. (2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics *9*, 128.

Spiegel, S., Milstien, S., and Grant, S. (2012). Endogenous Modulators and Pharmacological Inhibitors of Histone Deacetylases in Cancer Therapy. Oncogene *31*, 537–551.

Sterner, D.E., and Berger, S.L. (2000). Acetylation of Histones and Transcription-Related Factors. Microbiol. Mol. Biol. Rev. *64*, 435–459.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41–45.

Sullivan, J.M., Badimon, A., Schaefer, U., Ayata, P., Gray, J., Chung, C., Schimmelmann, M. Von, Zhang, F., Garton, N., Smithers, N., et al. (2015). Autism-like syndrome is induced by pharmacological suppression of BET proteins in young mice. J. Exp. Med. *212*.

Swingle, M.R., Amable, L., Lawhorn, B.G., Buck, S.B., Burke, C.P., Ratti, P., Fischer, K.L., Boger, D.L., and Honkanen, R.E. (2009). Structure-activity relationship studies of fostriecin, cytostatin, and key analogs, with PP1, PP2A, PP5, and(beta12-beta13)-chimeras (PP1/PP2A and PP5/PP2A), provide further insight into the inhibitory actions of fostriecin family inhibitors. J. Pharmacol. Exp. Ther. *331*, 45–53.

Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc. Natl. Acad. Sci. U. S. A. *110*, 18602–18607.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796–815.

The C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012–2018.

Thompson, K.L., Pine, P.S., Rosenzweig, B. a, Turpaz, Y., and Retief, J. (2007). Characterization of the effect of sample quality on high density oligonucleotide microarray data using progressively degraded rat liver RNA. BMC Biotechnol. *7*, 57.

Toyota, M., Suzuki, H., Sasaki, Y., Maruyama, R., Imai, K., Shinomura, Y., and Tokino, T. (2008). Epigenetic silencing of microRNA-34b/c and B-cell translocation gene 4 is associated with CpG island methylation in colorectal cancer. Cancer Res. *68*, 4123–4132.

Tropberger, P., and Schneider, R. (2013). Scratching the (lateral) surface of chromatin regulation by histone modifications. Nat. Struct. Mol. Biol. *20*, 657–661.

Waddington, C.H. (2012). The epigenotype. 1942. Int. J. Epidemiol. *41*, 10–13.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schübeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. *37*, 853–862.

Wilbanks, E.G., and Facciotti, M.T. (2010). Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. PLoS One *5*, e11471.

Wit, E. De, and Laat, W. De (2012). A decade of 3C technologies : insights into nuclear organization. 11–24.

Wu, J., and Grunstein, M. (2000). 25 Years after the nucleosome model: Chromatin modifications. Trends Biochem. Sci. *25*, 619–623.

Wysoker, A., Tibbetts, K., and Fennell, T. (2013). Picard tools version 1.90. Http://picard.sourceforge.net.

Xu, H., Wei, C.-L., Lin, F., and Sung, W.-K. (2008). An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics *24*, 2344–2349.

Yang, X.-J., and Seto, E. (2007). HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. Oncogene *26*, 5310–5318.

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., Zhao, F., and Zhu, B. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. BMC Bioinformatics *14*, 33.

Yu, D., Waterland, R. a, Zhang, P., Schady, D., Chen, M., Guan, Y., and Gadkari, M. (2014). and Reduces Survival in Mice. *124*, 3708–3712.

Yu, P., Xiao, S., Xin, X., Song, C.-X., Huang, W., McDee, D., Tanaka, T., Wang, T., He, C., and Zhong, S. (2013). Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. Genome Res. *23*, 352–364.

Yun, S., and Yun, S. (2014). Masking as an effective quality control method for next-generation sequencing data analysis. BMC Bioinformatics *15*, 382.

Yun, M., Wu, J., Workman, J.L., and Li, B. (2011). Readers of histone modifications. Cell Res. *21*, 564–578.

Zhang, Y., and Reinberg, D. (2001). Transcription regulation by histone methylation: Interplay between different covalent modifications of the core histone tails. Genes Dev. *15*, 2343–2360.

Zhang, J., Benavente, C. a., McEvoy, J., Flores-Otero, J., Ding, L., Chen, X., Ulyanov, A., Wu, G., Wilson, M., Wang, J., et al. (2012). A novel retinoblastoma therapy from genomic and epigenetic analyses. Nature *481*, 329–334.

Zhang, X., Pfeiffer, H.K., Thorne, A.W., and McMahon, S.B. (2008). USP22, an hSAGA subunit and potential cancer stem cell marker, reverses the polycomb-catalyzed ubiquitylation of histone H2A. Cell Cycle *7*, 1522–1524.

Zhang, Y., Lin, Y.-H., Johnson, T.D., Rozek, L.S., and Sartor, M. a. (2014). PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. Bioinformatics *30*, 2568–2575.

Zhao, Y., Lang, G., Ito, S., Bonnet, J., Metzger, E., Sawatsubashi, S., Suzuki, E., Le Guezennec, X., Stunnenberg, H.G., Krasnov, A., et al. (2008). A TFTC/STAGA module mediates histone H2A and H2B deubiquitination, coactivates nuclear receptors, and counteracts heterochromatin silencing. Mol. Cell *29*, 92–101.

Prénom NOM

TITRE de la thèse

Résumé : **Pipeline intégratif multidimensionnel d'analyse de données NGS pour l'étude du devenir cellulaire**

L'épigénomique pourrait nous aider à mieux comprendre pourquoi différents types cellulaires montrent différents comportements. Puisque, dans le cadre d'études épigénétiques, il peut être nécessaire de comparer plusieurs profils de séquençage, il y a un besoin urgent en nouvelles approches et nouveaux outils pour pallier aux variabilités techniques sous-jacentes. Nous avons développé NGS-QC, un système de contrôle qualité qui détermine la qualité de données et Epimetheus, un outil de normalisation d'expériences de modifications d'histones basé sur les quartiles afin de corriger les variations techniques entre les expériences. Enfin, nous avons intégré ces outils dans un pipeline d'analyse allèle-spécifique afin de comprendre le statut épigénétique de XCI dans le cancer du sein où la perte du Xi est fréquent. Notre analyse a dévoilé des perturbations dans le paysage épigénétique du X et des réactivations géniques aberrantes dans le Xi, dont celles associées au développement du cancer.

Résumé en anglais : **Multi-dimensional and integrative pipeline for NGS-based datasets to explore cell fate decisions**

Epigenomics would help us understand why various cells types exhibit different behaviours. Aberrant changes in reversible epigenetic modifications observed in cancer raised focus towards epigenetic targeted therapy. As epigenetic studies may involve comparing multi-profile sequencing data, there is an imminent need for novel approaches and tools to address underlying technical variabilities. We have developed NGS-QC, a QC system to infer the experimental quality of the data and Epimetheus, a quantile-based multi-profile normalization tool for histone modification datasets to correct technical variation among samples. Further, we have employed these developed tools in an allele-specific analysis to understand the epigenetic status of X chromosome inactivation in breast cancer cells where disappearance of Xi is frequent. Our analysis has revealed perturbation in epigenetic landscape of X and aberrant gene reactivation in Xi including the ones that are associated with cancer promotion.