UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE – ED 414 - Science de la vie et de la Santé

THÈSE présenté par:

Valeriya MALYSHEVA

Soutenue le: 10 Novembre 2016

Pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité: Aspects Moléculaires et Cellulaires de la Biologie

RECONSTRUCTION OF GENE REGULATORY NETWORKS DEFINING THE CELL FATE TRANSITION PROCESSES

THÈSE dirigée par:

M. GRONEMEYER Hinrich	Dr., IGBMC

RAPPORTEURS:

M. BISCHOF Oliver	Dr., Institut Pasteur
M. SPICUGLIA Salvatore	Dr., TAGC
Mme. HIBNER Urszula	Dr., IGMM

EXAMINATEURS:

M. BARILLOT Emmanuel	Dr., Institut Curie
M. FRASER Peter	Dr., Babraham Institute
M. SEXTON Thomas	Dr., IGBMC
M. SPITZ François	Dr., Institut Pasteur

ACKNOWLEDGEMENTS

First of all, I would like to thank my PhD supervisor Dr. Hinrich Gronemeyer for giving me this outstanding opportunity to work in his team on this extremely interesting and challenging project, for his courage to accept me as a PhD student, while I am coming from a different scientific field without the necessary experience, for constant enormous support and advice during my PhD and for his always positive attitude. This was THE place to learn and grow. This was my… fate.

Next, I would like to thank all my Jury members Dr. Emmanuel Barillot, Dr. Oliver Bischof, Dr. Peter Fraser, Dr. Urszula Hibner, Dr. Thomas Sexton, Dr. Salvatore Spicuglia, and Dr. Francois Spitz for kindly accepting to read and evaluate my PhD work.

I would like to thank Marco, my supervisor who was leading me through all these years and formed me as a scientist, for his help, guidance and training. Marco, you not only taught me new molecular biology techniques but also showed how to decorticate the problems when experiments do not work, you taught me how to be stable to failures, you made me strong. Without you this work would not be possible and I am infinitely grateful for everything you have done for me.

I would like to thank Matthias Blum, for his enormous every day support and patience while teaching me programming and correcting my Python scripts, when I was lost in manuals and documentation.

I thank all the other current members of the Gronemeyer team and those who left: Valeria, Maxi, Akin, Pierre-Etienne, Michele, Cathie, Aurelie, Lisa, Ashick, Ben and Gosia - thank you for all the help and support throughout these years.

I am also grateful for all the help provided by the IGBMC facilities and administration, Valérie and Violaine for their kind and friendly nature, support and help with translations.

Last but not the least, I would especially like to thank my family that always believed in me, supported me every day and gave me energy, courage and love. I dedicate this work to you.

CONTENTS

TABLE OF FIGURES

LIST OF ABBREVIATIONS

3C	Chromosome Conformation Capture
4C	Chromosome Conformation Capture on Chip
5C	Carbon-Copy Chromosome Conformation Capture
ATRA (RA)	All-Trans Retinoic Acid
ChIA-PET	Chromatin Interaction Analysis with Paired-End Tag Sequencing
ChIP	Chromatin Immunoprecipitation
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRMs	Chromatin Remodelers/Modifiers
CT	Chromosome territory
EC	Embryo Carcinoma
ESCs	Embryonic Stem Cells
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
FISH	Fluorescence In-situ Hybridization
GEO	Gene Expression Omnibus
GRN	Gene Regulatory Network
GWAS	Genome-Wide Association Studies
HOT	High-Occupancy Target
HTS	High-throughput Sequencing
iPSC (iPS)	induced Pluripotent Stem Cell
lncRNA	long non-coding RNA
LOGIQA	Long-range Genome Interactions Quality Assessment
LOH	Loss of Heterozygosity
mRNA	messenger RNA
PRC	Polycomb Repressive Complex
TAD	Topologically Associating Domains
TF	Transcription Factor
TF-TG	Transcription Factor - Target Gene

FOREWORD

"The eventual goal of science is to provide a single theory that describes the whole universe. However, the approach most scientists actually follow is to separate the problem into two parts. First, there are the laws that tell us how the universe changes with time. (If we know what the universe is like at any one time, these physical laws tell us how it will look at any later time.) Second, there is the question of the initial state of the universe. (…) it seems ... reasonable to suppose that there are also laws governing the initial state."

Stephen Hawking "A Brief History of Time"

Though the idea of the eventual goal of the science is debatable and every scientist has different objectives, we are all working on the creation of the grand unifying theory of the universe by fulfilling the gaps of unknown. Starting from different points of cosmology, physics or biology, integrating the knowledge of on-edge sciences, we reconstruct the multi-dimensional puzzle of our universe. Keeping this idea in my mind and in my heart I worked on my PhD projects, amazed by the complexity of our universe and hoping that it will help to put at least one small piece of the puzzle in place.

INTRODUCTION

1. A (VERY) BRIEF HISTORY OF TIME: THE CELL FATE.

1.1. THE UNCERTAINTY PRINCIPLE. THE FATE OF A CELL AND ITS PLASTICITY.

Every organism can be seen as a (complex) system that functions according to a biological/chemical program that is specified by genetically encoded information whose storage, maintenance and reading is based on the laws of physics. The fate of each cell is defined by this program and adapted to the developmental history and environmental context in which the cell is placed. However, the algorithms of this program are not yet fully understood nor the limits of the cell fate potential specified in this program.

Initially, cell fate acquisition has been viewed as an irreversible unidirectional path from pluripotent to the differentiated state; Waddington depicted it as a path of a ball down the hill of a landscape[1]. According to this model the destiny of line-committed cells was pre-defined, unidirectional and irreversible. However, experiments involving the transfer of somatic nuclei into an enucleated egg or fusion of a somatic cell with a pluripotent stem cell provided the proofs of cellular fate plasticity and demonstrated moreover that somatic cell memory can be erased and the cell can be reprogrammed to the pluripotent state [2,3].

Decades later it became clear that differentiated cells can be not only rejuvenated but also directly converted from one cell type to another bypassing the pluripotent state (*trans*-differentiation) by ectopic expression of a single transcription factor [4–7].

The discovery of the induced pluripotent stem cells (iPSCs) became a milestone in the history of reprogramming. Takahashi and Yamanaka demonstrated that pluripotent stem cells can be directly generated from differentiated cells by the addition of only a few defined transcription factors (OSKM factors: OCT3/4, SOX2, KLF4 and MYC), showing the great plasticity and potential of the cell [8]. Thus, the deterministic view of a cell fate is no longer valid and cells may adopt other cell fates if needed. The regenerating lens of the newt is a perfect illustration for such naturally occurring cell fate re-adaptation/ *trans*-differentiation [9]. When the lens is removed, pigmented epithelial cells (PECs) from the dorsal iris dedifferentiate and proliferate to create a new lens vesicle, and then differentiate into the mature cells of the lens. Microarray analyses

have revealed that during this process, PECs upregulate cancer and apoptosis-related genes, along with epigenetic modifiers, such as histone deacetylases and histone demethylases [10].

All these studies supported the idea of the transcription factors acting as master regulators of cell identity and fate. However, albeit key TFs have been identified that are sufficient for cell reprogramming [11–14], our knowledge about the temporal evolution and regulation of those TF-specified gene networks that execute cell fate acquisitions and which are essential to understand cell plasticity, has remained fragmentary.

1.2. Cell transformation – aberrant fate of a cell

Cell fate transitions are at the basis of essentially all biological processes in multicellular organisms and are tightly controlled. However, escape from the control mechanisms can lead to pathophysiological phenomena. Cancer is such a progressive multistep transition process that – due to (a few or a plethora of) mutations that lead to deregulation of control and failsafe mechanisms of the system - ultimately leads to cell transformation, characterized by aberrant proliferation or survival of cells that have escaped the (immune)surveillance mechanism of the host organism, lost their own control mechanisms, and acquired specific features that enable them to develop an "organism inside the organism".

1.2.1. Principles of cell transformation

The complex process of cancer development typically involves multiple genetic, epigenetic and chromatin changes. In a landmark paper Hanahan and Weinberg summarized the common traits for the majority of cancer types in hallmarks of cancer [15]. The defined traits are limitless proliferative potential, self-sufficiency in growth signaling and insensitivity to growth inhibitory signals, resistance to cell death, induction of angiogenesis and the ability to invade tissue and form metastases. These, together with the recent additions of evasion from the immune system and modification to adapt to the altered metabolism of a transformed cell [16] (**Figure 1**), describe a prototypic cancer phenotype.

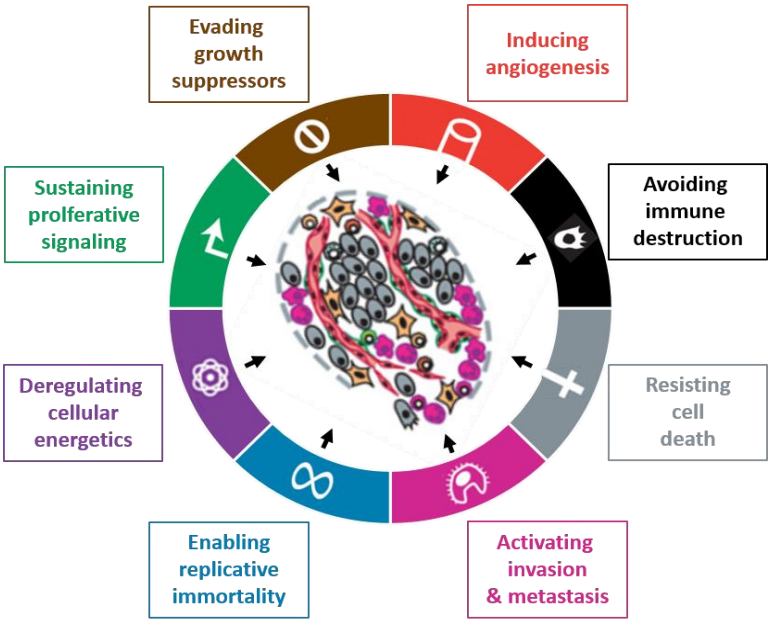


Figure 1. The hallmarks of cancer. The defining characteristics common for all type of cancer. *Adapted from* [16].

In fact, the tumorigenic transformation can be seen as an aberrant cell fate transition that happened due to the abnormal re-wiring of the gene regulatory network underlying the cell state, achieved through gradual accumulation of (epi)genetic changes. A general estimation of the number of these changes suggests that around 2-6 suffice for tumorigenesis [17] and given (1) that the mutation rate in normal human cells is extremely low (100 - 200 mutations per generation [18]) and (2) that the majority of them are in non-coding regions, some of the cancer driving mutations (a mutation that is causally implicated in oncogenesis [19]) may for example target regulators of genome stability, key factors involved in differentiation, factors regulating the cell suicide in case of serious damage or modulators of the immune system. In some cases, such as for colon carcinoma, the order in which the mutations appear has been defined, suggesting that each one of them is necessary for the next step of cell transformation [20]. However, recent reports reveal the existence of an additional rather dramatic mechanism of tumor development that occurs in about 2-3% of all types of cancers, involving massive chromosomal rearrangements in a single step catastrophic event termed chromotripsis [21].

With the improved sequencing technologies, it is now possible to identify the critical (epi)genome changes that are responsible for the development of human tumors, and a concept has been developed which discriminates between the actual "driver" mutations, necessary for

tumor growth and secondary "passengers" events, which are not causally involved in the generation of tumor clones. Several large-scale studies gave an extensive description of this dichotomy [22,23] and identification of novel driver mutations [24,25] can further be applied for the design of novel anti-cancer therapies.

Besides genetic changes, cancer cells are also characterized by epigenetic alterations - heritable gene expression modifications that do not involve changes in the DNA sequence. In general, cancer cells exhibit enhanced global DNA hypomethylation, gene-specific local hypermethylation (e.g. of tumor suppressor genes) and altered functions, expression [26] and/or recruitment of epigenetic modulators. Each of these features contributes to global genome instability, repression of tumor suppressors and other cancer-specific changes [27,28].

The genetic and epigenetic processes can act in concert, such that epigenetic changes influence the genome function and vice versa; indeed, oncogene signaling can reshape the epigenetic landscape [29]. For example, deamination of 5-methylcytosine (5mC) creates a T:G mismatch which is a hotspot for somatic mutations [30]. The inverse also applies when somatic mutations give rise to epigenetic changes, as seen in the case of mutations in genes coding for some of the epigenetic enzymes, such as DNA methytransferase like DNMT3A or the histone methyltransferase KMT6A (also known as EZH2), commonly found deregulated in AML patients [31] and lymphomas [32,33].

Understanding cancer was for a long time limited to purely correlative observations and cancer heterogeneity remained largely unexplored. However, recent technological advances have facilitated insight into cause-consequence relationships and the functional complexity of cancer such that a major focus is now on single-cell cancer genomics and systems biology studies of the epigenome [34], offering a view into the complex molecular architecture of cancer. A large amount of data obtained using these approaches has fostered our understanding of the (molecular and cellular) origins of cancer [35,36] and aided in the design of novel cancer therapies [37,38].

1.2.2. GENETIC REGULATORS OF CELL TRANSFORMATION

Genes that regulate cell transformation are divided into two functional groups: oncogenes and tumor suppressor genes. While tumor suppressors - molecular brakes of tumor development – require generally "loss of function" mutations (or deletion) of both alleles in order to produce an

effect, oncogenes require only one hit, which endows them with a "gain of function" mutation that suffices for tumor development.

Oncogenes. Proto-oncogenes normally exist in the genome and code for proteins that promote cell proliferation and growth, but due to mutations and/or overexpression their function(s) become uncontrolled/corrupted and contribute to cancer development. Based on their functions, they can be divided into several categories: growth factors, growth factor receptors, signal transducers (such as the tyrosine kinase *Src*, the serine/threonine kinase *Raf-1* or the small GTPase *Ras* family), transcription factors (*Fos, Jun, Myc, Myb*) and cell death regulators (like *Bcl-2*).

The first oncogene was *Ras,* identified in 1982 as a transforming agent in NIH-3T3 mouse fibroblasts and cloned from the T24 and EJ bladder carcinoma cell lines [39–41]. In my studies I used a stepwise cellular transformation system where the oncogene c-Myc was used, one of the most prominent oncogenes in humans. *c-Myc*, which together with *N-Myc* and *L-Myc* forms the *Myc* family is a gene coding for a transcription factor that was discovered in patients with Burkitt's lymphoma. These lymphomas originate from characteristic chromosomal translocations of *c-Myc* to distinct loci, such as the immunoglobulin heavy chain in the most frequent t(8;14)(q24;q32) translocation, which puts *c-Myc* under the control of the *IGH* gene [42]. Despite some controversial views on the mechanism by which MYC regulates genes [43], it became increasingly evident that MYC differentially controls discrete sets of genes (up to 15% of the complete genome [44]) affecting global transcript levels and altering diverse cellular processes, including cell growth and cell cycle, by deregulation of other TFs and chromatin remodelers. MYC is also known to block cell adhesion, cell-cell communication and/or terminal differentiation and influences apoptosis [45–47].

Heterocomplexes of MYC and the MYC-associated factor X (MAX) enhance transcription by binding to target sequences ('E boxes') within the promoters/enhancers of cognate genes. They recruit additional transcriptional activators and chromatin remodelers [48] (such as histone acetyl transferases - GCN5, TIP48) that leads to transcriptional upregulation of target genes. This action of MYC is antagonized by formation of a second type of MAX complex [49] (MAD-MAX or MNT-MAX), which also binds E-box elements, but instead recruits co-repressors and leads to

decreased target gene transcription. Thus, depending on the balance of E box occupancy with MYC-MAX or MAD-MAX heterodimers, target genes will be either activated or repressed.

Myc also acts as a transcriptional repressor of multiple target genes (*p15*, p*21, p27)* by blocking the action of the appropriate transcription factors (such as SMAD, YY-1, SP1, MIZ-1) [50]. In the latter case MYC does not bind to target DNA directly, but instead binds to MIZ1 at the site of the core promoter. Gene repression is achieved through competition of MYC and the coactivator p300 for binding to MIZ-1, but also through MYC's ability to recruit the DNA methyltransferase Dnmt3a to MIZ-1 regulated genes [51].

MYC is a short-lived protein (t1/2 ~ 20 min), but controls a significant number of genes; it is sensitive to subtle changes in amounts that are accompanied by changes in co-regulator recruitment. The model of MYC action suggests that it does not bind to all targets at the same time, but that they all ultimately become transiently occupied in a certain short period [45].

MYC is tightly controlled at multiple levels. Various signaling cascades, such as WNT, RAS/RAF/MAPK, JAK/STAT, TGFβ and others, contribute to increases in *MYC* transcription. Additionally MYC is heavily controlled at the posttranscriptional level through phosphorylation, ubiquitinylation or acetylation, which affect its stability and activity [52–56].

Importantly MYC stability is altered by other oncogenes and RAS/RAF/ERK pathway through phosphorylation [57], which suggests oncogenic synergy in signaling, where MYC probably acts as a central regulator of cellular transformation [26,58]. In support of this, the crucial role of MYC signaling is seen *in vivo* in mice models of RAS-induced lung adenocarcinoma and SV40-driven pancreatic tumor model, where systemic MYC inhibition by a dominant negative mutant ('Omomyc'), led to tumor regression [59,60]. These mice also showed profound changes in proliferating tissues, which is in accordance with the well-described central role of c-MYC in cell pluripotency, as it is a part of Takahashi/Yamanaka reprogramming cocktail [8,61]. Indeed, cancer and stem cells have some similarities, e.g. ability to proliferate extensively and in case of cancer stem cells to generate populations of non-tumorigenic cells [62] in the way normal cancer cells give rise to differentiated progeny. Our studies also indicated that during the stepwise tumorigenesis cells gradually acquire embryonic stem cell traits, suggesting that oncogene induces or facilitates the re-wiring of normal cells GRNs to stem cell GRNs [26]. That implies that

deciphering of these GRN transitions will help to understand the principles of key processes of tumorigenic cell transformation.

Tumor suppressors. Based on the role they perform, tumor suppressor genes can be divided into two categories - caretakers and gatekeepers.

Gatekeepers sense stress or damage within a cell that represents a threat to the fidelity of replication and act to halt proliferation. Once gatekeeper pathways are activated, cell can either be physically removed by apoptosis or permanently growth arrested by becoming senescent. Key regulators of these two processes are the same and the two most important ones are TP53 and RB1. The tumor suppressor TP53 is a transcription factor that is stabilized upon DNA damage and other stress, and acts as a transcriptional repressor of anti-apoptotic genes like *BCL-2* and a transcriptional activator of pro-apoptotic genes, therefore leading to apoptosis induction. Conversely, activation of TP53 can also favor senescence via induction of the cyclin-dependent kinase inhibitor (CDKI) CDKN1A, which blocks cell proliferation. The other major tumor suppressor, RB1 is active in its hypo-phosphorylated state and functions by blocking the progression of the cell cycle from G1 to S. In the presence of stress or DNA damage signals, CDKN2A interacts with CDK4 and CKD6, blocking their phosphorylation of RB1, thus keeping it in its active state. Inherited mutations in gatekeeper genes require only one additional mutation in the second allele to produce an effect. Thus, mutations in gatekeepers greatly increase the risk of cancer and these genes are relatively often found in sporadic mutations.

Caretakers have a role in maintaining the genome integrity and preventing the formation of mutations. They are generally involved in DNA repair and can be either the sensors of the DNA lesions (like ATR or BRCA1 or 2) or part of the repair machinery. A single mutation in a caretaker gene needs a mutation in the other allele (or undergo LOH) to become prevalent and yet does not lead to neoplasia but only to higher incidence in the acquisition of other mutations; and thus caretaker-driven tumorigenesis is rarely seen in sporadic cancers [63].

1.2.3. STEPWISE TRANSFORMATION SYSTEMS.

Human cancer cell lines derived from human tumor specimens are extensively used for identification of molecules and pathways involved in malignant transformation as well as for preclinical testing of potential therapeutic anti-cancer compounds. However, these experimental

models suffer from several limitations. As such, human-derived cancer cell lines can bear a high number of genetic mutations that complicates deduction of the cause-consequence relationships and reconstruction of the information flux from the initial signal. For the same reason it is difficult to generate stable cell lines using tumor explants, as they accumulate mutations in prolonged cultures. The functional consequences of these mutations are unknown and introduce bias in experiments performed with non-isogenic cells. Additionally, continuous passaging of cell lines derived from human tumors can lead to the selection of fast growing sub-clones that can progressively dominate the culture and do not represent anymore the original cancer type studied, thus introducing a serious bias in the study results.

Due to the multiple levels of tumor complexity mentioned before, a reductionist approach has been developed to understand the basic principles of cancer development. It consists of identification of the minimal fundamental changes required in different cell types for their transformation. In their landmark paper Hahn and Weinberg described a stepwise tumorigenic model system, in which defined genetic changes had been introduced into several primary normal cell types in order to generate cancer cells. Successful transformation of normal cells was achieved by expression of catalytic subunit of telomerase hTERT (which prevents telomere shortening), the oncoproteins of the Simian virus early region [64] (SV40 ER, expressing small and large T) and an overexpressed oncogene. The genetic elements introduced allow cells to bypass several pre-existing barriers in cancer development. Blocking tumor suppressors by SV40 expression among others blocks TP53, RB1 (by large T) and PP2A (by small t) and prevents cell senescence. The additional expression of hTERT enables cells to surpass cell crisis. The advantage of such cell models for experimental studies is their isogenicity, which we validated in our studies [26], thus enabling solid conclusions about the net effect and the role of the introduced genetic elements in tumorigenesis and to accurately compare the immortalized and tumor stages with their normal counterpart.

Full transformation was achieved in this stepwise model in multiple cell types, confirming that the rules of tumorigenic transformation are somewhat universal and that despite the heterogeneity of cancer, there are basic mechanisms that govern the ontogeny of cancer cells. The fully transformed cells exhibit cancer-specific characteristics, such as anchorage-independent growth (as validated in our study [26]), tumor formation in nude mice [65,66] and they are

sensitive to TRAIL-induced apoptosis [26,46,67,68]. These systems provide a valuable tool in studying the processes of transformation, the transformation-related characteristics and is a perfect platform for prediction of new key regulators of tumorigenesis that could be further studied as a potential target in cancer treatment development.

2. SPACE AND TIME. CHROMATIN STRUCTURE AND GENE REGULATION

2.1. THE THREE-DIMENSIONAL STRUCTURE OF CHROMATIN

The 3D architecture of the genome influences key cellular processes such as gene regulation, replication and differentiation [69]. In order to preserve the integrity and ensure functionality, the DNA in the eukaryotic cell nucleus has to adopt an adjustable and robust non-random dense structure that would at the same time guarantee the accessibility of various DNA binding components of the replication or transcription machineries, epigenome modulators/interpreters and DNA repair enzymes. Thus, tight regulation of chromatin organization in space and time is a key for a proper functioning of the cell fate program. The following section describes different experimental approaches to investigate the chromatin structure itself and the hierarchy of chromatin organization, ranging from the DNA polymer to functional chromatin/nuclear territories.

2.1.1. CHROMATIN FIBER

The DNA consists of two helical chains of 1nm radius centred around the common axis and wrapped around the octamer of histone proteins in 1.65 turns covering 145-147 bp, thereby forming the nucleosomes – repeating building blocks of chromatin separated by linker DNA of 20-50 bp [70]. By coiling chromatin folds into highly ordered structures: first, nucleosomal "beads-on-a string" fibers of 11nm in diameter, which are further condensed into 30nm fibers with the help of linker histones [71]. Despite efforts during the last decades, the exact arrangement of chromatin into these higher-order structures remained largely uncharacterized, proposing still debatable models of arrangement like solenoid and Zig-Zag models [72–74], which may actually coexist depending on the functional context [75]. Similarly, the dynamics and integrity of this fiber during the transcription, cell cycle, differentiation and tumorigenesis, and its topological anchoring and consistence, have remained elusive.

2.1.2. TOPOLOGICALLY ASSOCIATED DOMAINS

The next level of organization of metazoan interphase chromosomes are topologically associating domains (TADs) – self-interacting regions of chromatin at a sub-megabase scale [76–79]. Detected by methods such as microscopy [80] and HiC [76,81], these contiguous regions favor internal contacts and are relatively insulated from the neighboring domains, though contacts between TADs do occur at a relatively lower level [82]. Interestingly, TADs have not (yet) been detected in plants [83] and yeast [84], demonstrating that they possess (an) alternative mode(s) of genome folding. Co-localization between TADs composed of similarly transcriptionally permissive or inert chromatin leads to the establishment of A and B compartments, or open and closed compartments, respectively [81].

An increasing number of studies describe an important functional role of TADs in gene regulation [85–87]. Some TADs have homogeneous interiors, while others have a rather nested structure and are partitioned into smaller sub-TADs that are thought to vary and may facilitate changes in gene expression during cell differentiation [88], DNA replication [89] and development [90]. Upon experimental deletion of a TAD boundary, the TAD spreads to the next boundary, indicating that inter-TAD contacts are not hard-wired and that boundary-associated elements play crucial roles [77].

Formation and maintenance of TADs are mediated by various architectural proteins, including CTCF, and the Cohesin and the Mediator complexes [91], which stabilize these contacts and restrict the distance over which enhancer-promoter interaction can occur [92]. However, despite an enrichment at TAD boundaries, neither the presence of CTCF/Cohesin sites nor CTCF binding is sufficient to establish TAD boundaries; indeed only 15 % of all CTCF binding sites are found at TAD boundaries [76]. Similarly, insulator-binding proteins do not always block inter-TAD chromatin interactions. At the same time, knockdown of CTCF results in less well-defined TAD boundaries, reducing intra-TAD and increasing inter-TAD interactions, which is accompanied by changes in gene expression[93]. Neither disruption of the Cohesin complex nor its deletion destabilizes TAD boundaries [93,94], though its disruption leads to a diminution of intra-TAD interaction. All these studies indicate, that although architectural proteins are required for proper chromatin organization in some cases, they are not necessary for TAD boundary formation, which may depend on contextual factors.

Another open question is mechanism of TAD formation. Based on polymer simulations of the chromatin fiber it was recently proposed that cis-acting loop-extruding factors (potentially Cohesins) form progressively larger loops by extrusion, but are stalled by boundary elements, such as CTCF at TAD boundaries [95]. The proposed mechanism suggests that TADs consist of dynamically forming, growing and dissociating loops. Interestingly, this model stands against the popular view of TADs as stable loops, as the modeling of such scenario provides some of the worst fits to HiC data [95]. Importantly, the loop extrusion mechanism (**Figure 2**) recapitulates the results of TAD boundary deletion experiments [77], further supporting this hypothesis.

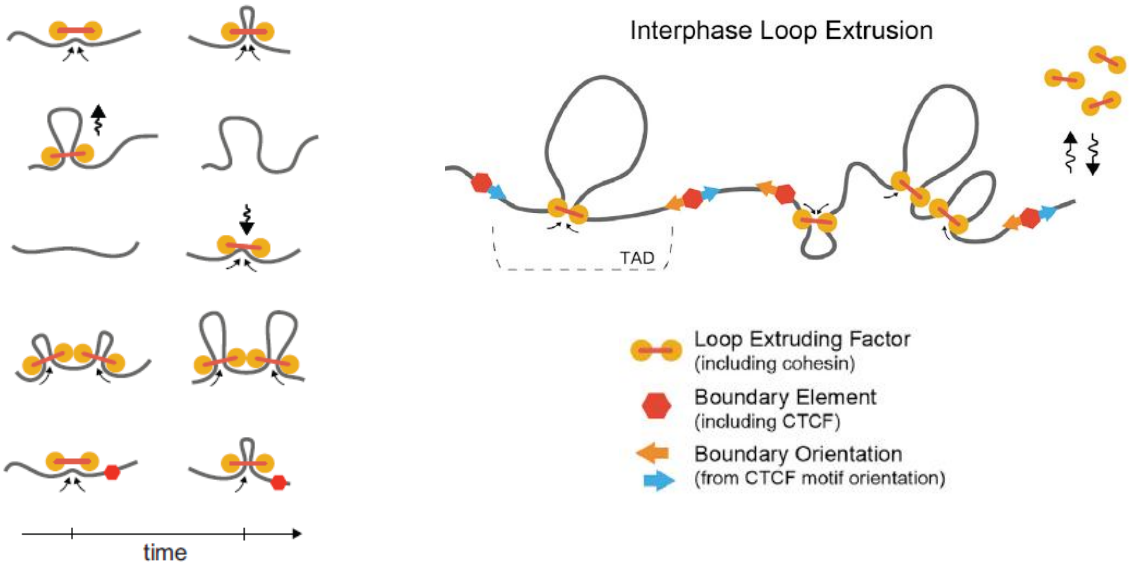


Figure 2. Loop extrusion model proposing that tads consist of dynamically forming, growing and dissociating loops. *Adapted from* [95].

2.1.3. CHROMOSOME TERRITORIES

Each chromosome, subdivided into many TADs, resides within a discrete volume of space known as a chromosome territory (CT), as has been demonstrated by microscopy-based approaches [96]. The potential formation of CTs has been described also by several polymer models (equilibrium model, fractal globule [81]). Transcriptionally repressed genes tend to be positioned at the nuclear periphery and are often attached to the nuclear lamina [97], while

transcriptionally active genes prefer interior nuclear regions [98]. However, some exceptions do exist, like the case of rod photoreceptor cells of nocturnal animals, where the euchromatin is expelled to the nuclear periphery and heterochromatin occupies the central part of the nucleus, thus serving as an optical lens for efficient light detection [99].

2.2. Experimental techniques for chromatin structure investigation

There are two major approaches to investigate chromatin architecture: one is imaging, including microscopy combined with various fluorophores, the other is based on the Chromosome Conformation Capture (3C) assays, giving the read-outs through qPCR or HTS.

2.2.1. Microscopy-based approaches

The first method applied in studies of chromosome shape and size was microscopy, which enabled the establishment of karyotypes of human cells. Giemsa staining further improved the method, resulting in the detection of G-bands of chromosomes and of large-scale genome aberrations, such as chromosomal translocations.

Fluorescence In-situ Hybridization (FISH)[100] uses fluorophore-labelled custom DNA probes, which hybridize with genomic DNA, allowing targeted visualization of loci at a 200nm resolution. To trace DNA in a 3D space, 3D FISH has been developed by taking advantage of multiple fluorophores and guide DNAs. RNA-FISH allows the detection of various RNA species, like messenger RNA (mRNA) and long non-coding RNA (lncRNA). Combining such spatial data from a few hundred cells one can estimate the frequency of co-localization between selected loci. However, this technique is limited in coverage, such that only a few loci can be monitored simultaneously.

2.2.2. Chromosome Conformation Capture and its derivatives

Chromosome Conformation Capture (3C) was the first molecular method in a 3C family [101]; it investigates the genome organization relying on proximity ligation (**Figure 3**). In brief, the chromatin undergoes sequentially through the following steps: crosslink [i], digestion by a

[i] generally, with the zero-length crosslinker formaldehyde

restriction enzyme [ii], ligation, purification and analysis by PCR, qPCR using sequence specific primers or by sequencing.

There are a number of 3C-derived methods, like 4C and 5C, aiming to investigate larger number of interactions, with HiC ultimately monitoring long-range chromatin interactions at genome-wide scale, albeit with still fairly low resolution (**Figure 3**).

In **4C – Chromosome Conformation Capture on Chip** [102] – the 3C library is cut with a second restriction enzyme. The fragments are circularized during the second round of ligation and further amplified by Inverse Polymerase Chain Reaction. The advantage of this additional circularization step is that the amplification reaction can be done using only one end of the fragment of interest - bait. The 4C library is further analyzed using hybridization to DNA microarray or by HTS [103]. Thus this procedure allows the discovery of all the interactions with one site of interest [iii].

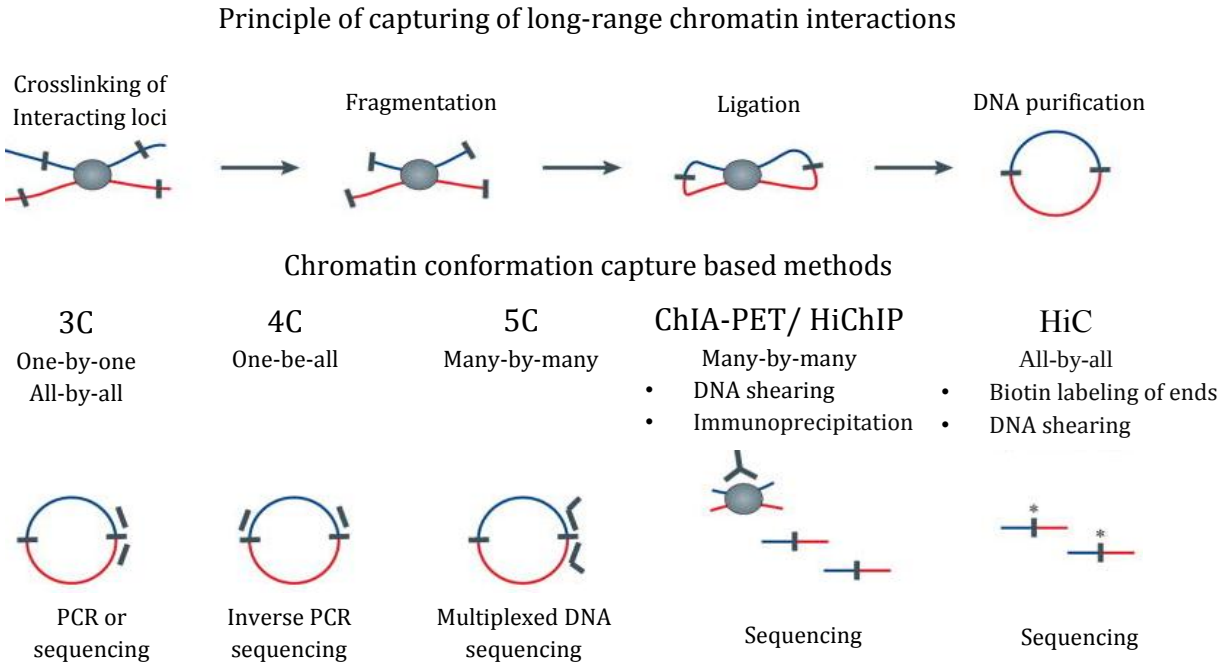


Figure 3. 3C based techniques. Adapted from [82].

[ii] often HindIII is used; sonication/tagmentation approaches are less used but attractive alternatives in view of increased resolution

[iii] the so-called "viewpoint"; this approach is also referred to as "one-by-all"

The **Carbon-Copy Chromosome Conformation Capture (5C)** [104] gave the possibility to inspect the interactions of many different sites at the same time, as the library is amplified with multiple primers through the multiplex ligation-mediated amplification (LMA). This procedure allows the capture of any fragments defined by the primer set [iv]. These primers are custom selected, though in the majority of studies these primers cover a continuous genomic region of several megabase.

The development of the **HiC** method in 2009 [81] revolutionized the world of chromatin organization studies, as this method reports interactions between any pair of loci in the genome. The concept of HiC is similar to 3C with several important modifications (**Figure 3, Figure 4**). After the digestion with the restriction enzyme, the DNA overhangs are filled in with nucleotides, one of which is biotinylated, followed by the blunt-end ligation. After purification of the DNA, it is sheared by sonication and the biotinylated fragments are pulled down with streptavidin-coated magnetic beads to enrich the final library with the ligation products, which are further amplified by PCR and sequenced. While the idea of HiC is very simple, the original protocol contained several weak points and needed optimization, which I performed during the experimental work in the context of my PhD studies (see Materials and Methods of Publication N° 4 Malysheva et al. 2016. 'Chromatin structure dynamics directs cell fate acquisition', manuscript in preparation).

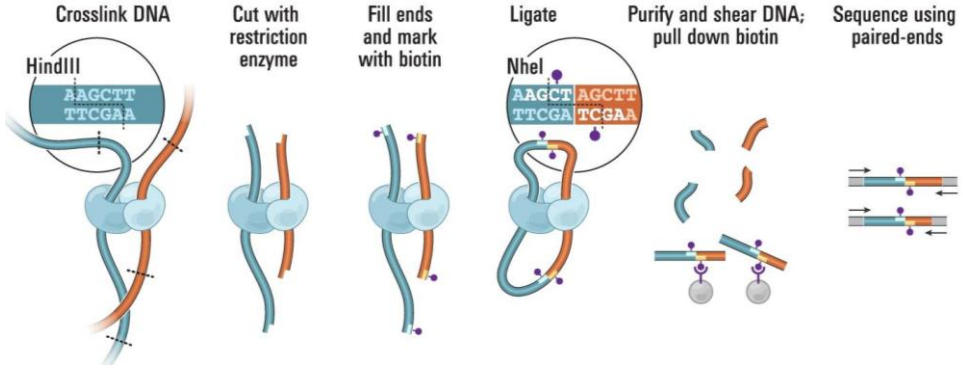


Figure 4. Overview of the HiC procedure. Taken from [81].

Although in theory HiC gives the most comprehensive map of interactions, the complexity of the HiC libraries is very high and to reveal all interactions that took place in an experimental sample needs a very high sequencing depth. Assuming that every restriction fragment can interact with

[iv] this approach is referred to as "many-by-many"

any other fragment, one would expect a theoretical number of 10^{11} possible HindIII restriction fragment pairs from the human genome. Moreover, the more frequent are the restriction sites of the enzyme in use, the higher should be the sequencing depth, as more potential interacting fragments will be produced. For example, the usage of DpnII as a restriction enzyme would theoretically generate 10^{13} possible restriction fragment interaction pairs. Thus, it is difficult to generate a Hi-C library with enough complexity at a sequencing depth that covers all possible restriction fragment interactions. This indicates that the current HiC datasets of large genomes are far from being sequenced at optimal depths, while for small genomes, e.g. *Drosophila*, this is much easier to attain. Our tool for the quality estimation of long-range chromatin interactions (LOGIQA [105]) confirms this notion, as the quality of the HiC datasets are in general increasing with increasing sequencing depth, while for *Drosophila* quality-vs-depth curve reaches a plateau much earlier. Techniques such as ChIA-PET, 4C, Capture HiC and the recently developed HiChIP can help to localize the view [v], thus to reduce the library complexity and the minimum needed coverage depth, while providing more details in the interactions of the regions of interest.

In **Capture-HiC** [106], sequences of interest can be enriched from a Hi-C DNA library to obtain highly multiplexed, targeted interaction profiles (**Figure 5**). This involves the hybridization of biotinylated capture-probes to DNA sequences of interest followed by capturing of this library of probe–target sequence complexes on streptavidin beads.

Another limitation of C-based techniques is that the primary signal is averaged over millions of cells. Though this data may uncover the preferred conformation of loci, it doesn't give an information about cell-to-cell variability in a way that DNA FISH does. To address the heterogeneity of the sample **single-cell HiC** has been developed [107]. As any given site can only be ligated only ones (or maximum n times with copy number n), the amount of signal from a single-cell experiment is much less than in Hi-C. However, pooling maps from single cell experiments results in interaction matrices similar to HiC, showing it to be a faithful average of single-cell data. Comparison of whole chromosome contact maps suggested that domain intactness is generally conserved at the single cell level, with intra-domain structures showing much less variability than inter-domain contacts [107]. This corroborates the previously observed stability of TADs.

[v] to the interactomes mediated by a particular protein or to the interactomes of the regions of interest

The first technique developed to investigate chromatin interactions, mediated by a protein of interest, in a genome wide manner was **Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET)** [108]. In this method the crosslinked chromatin is first immunoprecipitated prior ligation by using the antibody directed against the protein of interest. The ChIA-PET library is further read by HTS. This technique has been used to study interactions involving subsets of functional genomic elements bound by estrogen receptor 1, ESR1, RNA Polymerase II, CTCF, SMC1A and RAD21 as well as various histone marks, such as H3K4me1, H3K4me2, H3K4me3 and H3K27ac [109–115].

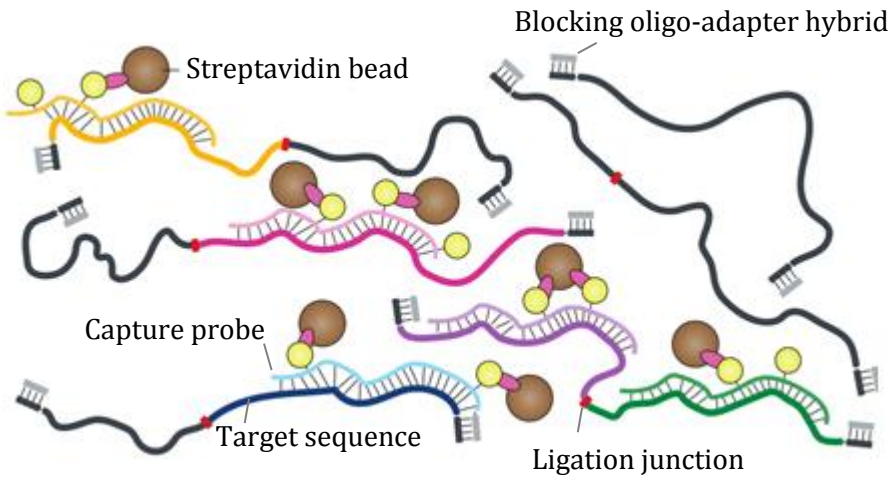


Figure 5. Capture-HiC principle. Adapted from [260].

However, ChIA-PET requires hundreds of millions of cells per experiment and results in a small fraction of informative reads for a given sequencing depth [116]. The recently developed HiChIP technique [117] somewhat reversed the ChIA-PET protocol by performing the ChIP with ligated chromatin (including also some additional technical modifications) that improved the yield of conformation informative reads by over 10-fold and requires over 100-fold less of the input material relative to ChIA-PET [117]. Thus, HiChIP is a new promising method for 3D genome structure studies.

2.3. Transcription regulation by epigenome and chromatin architecture

The epigenetic environment and the chromatin structure into which a gene is embedded have a strong influence on transcription, as many nuclear regulatory mechanisms act locally in a 3D

nuclear space. As such the local concentration of transcription factors, RNA polymerase II and the associated factors/complexes/machineries, as well as the accessibility of local chromatin have a large impact on transcription[118]. Similarly, histone modifications and DNA methylation patterns influence gene transcription as well. Finally, the spatial distribution of small RNA molecules can also affect gene regulation, as in the case of the silent mammalian X-chromosome in females, which is inactivated [vi] by the actions of the *Xist* non-coding RNA and its regulators, such as the repressor *Tsix*, the activator Rnf12 and other putative positive regulators (*Jpx*, *Ftx* and *Xpr*), resulting in specific chromatin modifications, spatial reorganization of the chromosome and its almost complete transcriptional silencing[119–121]. Interestingly, these studies showed a functional connectivity between the chromatin organization, epigenetics and gene expression/silencing. Indeed, *Tsix* transcription levels were correlated with TAD compaction using RNA and DNA FISH in the same cells. It was thus proposed that structural fluctuations within TADs may underlie differential transcriptional status and contribute to generating asymmetries between the two X chromosomes, hence influencing choice during the onset of X chromosome inactivation [122].

2.3.1. EPIGENETIC REGULATION OF GENE EXPRESSION

Ground-breaking studies in the mid-20th century on position-effect variegation and transposable elements [123,124], followed by the discovery of X-chromosome inactivation [125] and imprinting [126,127] led to the concept that identical genetic material can be maintained in different 'on' or 'off' states in the same nucleus affecting the phenotype. These observations supported the idea of the epigenetic changes – initially coined by Waddington as changes of phenotype without changes in genotype – as being an additional regulation mechanism of cell-type identity, transducing the inheritance of gene expression patterns without altering the underlying DNA sequence.

First reported epigenetic modification was the DNA methylation, which connection with gene expression has been established in numerous studies on ovalbumin and globin genes [128,129], showing the anti-correlation between the level of DNA methylation and gene expression levels. Soon thereafter, the implication of global DNA hypomethylation (at CpG dinucleotides) in cancer has been reported [27] and further local DNA hypermethylation of tumour suppressor genes [130]. Examples of genes affected by hypomethylation include oncogene *HRAS* [131], *CCND2* in

[vi] with the exception of a few interesting 'escapee' genes [120]

gastric carcinoma [132], human papillomavirus 16 (HPV16) in cervical cancer [133], etc. Indeed, the frequency of hypomethylated sites appears to be high, as indicated by high throughput genomic-methylation analysis of tumors [134], including cancers of stomach, colon, pancreas, liver, uterus, lung and many others. Moreover, pre-malignant adenomas also had generally altered DNA methylation patterns [135,136].

Discoveries of post-translational modifications of histones and development of modification- or site-specific antibodies implication in gene regulation enabled the identification of the role of these modifications, in addition to DNA methylation, in regulating gene activity [137–139]. Today we possess the knowledge about a large spectrum of histone modifications that led us to distinguish active and repressed genomic regions. The epigenetic landscape of chromatin is not even and there are modification-rich 'islands', which tend to be the regions that regulate transcription or are the sites of active transcription. As such, active transcriptional enhancers are marked with H3K4me1 and H3K27ac [140–142], while promoters of active genes possess a high enrichment of H3K4me3, H3K9ac and in some cases H3K27ac. In addition, H3K36me3 is highly enriched throughout the entire transcribed region. At the same time, trimethylation of lysines 27 and 9 of H3 – are classical markers of repressed transcription.

Furthermore, bivalent domains, defined by the co-existence of a H3K4me3 permissive histone mark and a repressive H3K27me3, are thought to play an important role in pluripotency by keeping the developmental genes in a poised state ready for activation upon differentiation of ESCs [143,144] or of epiblasts [145]. However, the nature of bivalency has been recently questioned. It has been proposed to be an *in vitro* artifact resulting from suboptimal culture conditions [146] or from technical difficulties associated with the low amounts of available material [147,148]. There have been also controversy reports from non-mammalian species, with bivalent domains present in zebrafish [149] but absent in Xenopus or Drosohila embryo [150]. However recent studies in primordial germ cells (PGCs), embryonic precursors of the germline, have shown developmental regulatory genes remaining bivalent and silent *in vivo* [151], but they do not maintain these features in the adjacent somatic cells, which represents a scenario similar to cultured ESCs that differentiate. Potentially, the maintenance of bivalency through the germline could provide the basis for the controversially disputed transgenerational epigenetic inheritance [for a classical example of this hypothesis, see [152]]. Interestingly, loss of the H3K27me3 mark from bivalent

promoters has been reported to lead to activation of cancer-promoting genes in colorectal cancer [153], including stem cell regulators, oncogenes and proliferation-associated genes.

New advances in technology allow now the analysis of single-cell epigenomes with more precision [154,155], indicating that almost the entire genome is transcribed, giving rise to a range of ncRNA with distinct regulatory functions [156] and many others among them remaining under investigation.

2.3.2. GENE REGULATION IN 3D CONTEXT

Metazoan genomes are organized in linear clusters of co-expressed genes that span about 100kb in *Drosophila* melanogaster [157] and 1 Mb in humans [158], this size corresponds to the average TAD in these species. Furthermore, TADs were found to overlap with the chromatin states [76,78,159], thus assigning a certain chromatin type to each TAD.

The position of a TAD in the nucleus relatively to other TADs or nuclear structures, such as nuclear lamina, can change during the development, supporting a role of a TAD localization in cell type specification. For example, entire TADs on the X-chromosome re-localize to the nuclear lamina during the X-chromosome inactivation in the early embryonic development [77]. TADs harbor multiple genes and correlate with active and repressive epigenetic marks. This discovery brought a missing link to chromosome biology, linking thousands of genes and enhancers in a structured way. Genes in the same domain tend to be physically close and have similar epigenetic make-up, such as chromatin marks or DNA methylation patterns.

Chromatin looping can occur between a variety of genetic elements within a given cell type, linking local genome organization to cis-regulation of both, gene expression and alternative splicing. Studies conducted by the ENCODE consortium demonstrated that many promoters in a given cell are contacted by multiple enhancers, and vice versa, and that gene expression driven from a given promoter positively correlates with the number of enhancers contacting it in a cell population[160]. As the one of the main drivers of cell type-specific gene expression, enhancer usage is dynamic during the cell proliferation, differentiation and other cell physiological processes [161]. Correlating chromatin state and RNA polymerase II occupancy at enhancers and promoters enabled the identification of co-regulated elements that tend to co-localize within a same TAD, thus supporting the model that functional promoter-enhancer interactions are

delimited by TAD boundaries [87,92]. At the same time, genes that are located in-between TADs, so-called TAD boundaries, are able to change the direction of their interaction preferences, switching their interactions from one TAD to another, like in the case of Hox cluster genes [85].

2.3.3. TRANSCRIPTION FACTORIES

Previous studies suggested that the transcriptional activity in a cell may take place in a few hundred transcription factories[162–164]. According to this hypothesis RNA polymerase is immobilized, while the DNA is going through it creating nascent RNA[165]. Accordingly, genes move in and out of these factories, creating bursts of transcription. The fact that the foci remain stable even after the inhibition of RNA polymerase II lends some support to the hypothesis of transcription factories; indeed, if transcription factories were just aggregations of active genes one could expect that they would fall apart if transcription stopped. Moreover, it has been demonstrated that co-regulated genes binding the same transcription factors tend to co-localize at common transcription factories[166]. Taking into account that co-regulated genes and enhancers tend to be found in the same TAD it is tempting to hypothesize that there is a cross-talk between TADs and transcription factories.

In addition, recent integrative study of long-range chromatin interactions in K562 cells showed that high-occupancy target (HOT) regions, marking promoters of highly expressed genes [167,168], were enriched at interacting loci and tended to interact with other HOT regions. This finding supports the transcription factory model. The strong enrichment for cohesion, CTCF, and ZNF143 at all interacting loci including HOT regions suggests that these factors are possible regulators or facilitators of transcription factories [115].

However, the hypothesis of transcription factories is still debatable raising the questions of their generality and importance for transcriptional regulation. In particular, it is unknown how such a ‘factory’ is capable of transcribing genes on the (+)-strand and (-)-strand at the same genomic locus at the same time. There is also no convincing experimentally supported model revealing how the polymerase remains immobilized or how and to what structure it is tethered to.

3. GENE REGULATORY NETWORKS

All the aspects discussed above deal with the structural organization of chromatin and its functional modifications as a complex regulatory platform that has one function: to regulate the expression of genes in a dynamic, temporally defined manner, specific for a particular cell type and responding to the cognate signaling inputs. Altogether, this results in cell fate-specific expression of coding and non-coding RNAs, which is instructed by a gene-regulatory program that triggers cell homeostasis and cell fate progression along a physiological or pathological trajectory. Thus, one could look at the organism with its multitude of diverse cell types, each of them with an imprinted/memorized history and (more or less, depending on the cell) specified future, as a temporal program that defines a complex map of cell fate probabilities. To understand this map of cell fates we have to reconstruct the roadmaps of cell fate regulation – i.e., the gene regulatory networks.

3.1. NETWORK INCEPTION

The dramatic progress in molecular biology, biotechnology and bioinformatics over the past years has allowed us to discover plethora of novel molecular interactions giving rise to metabolic circuits, signaling networks and molecular machineries. Each of these circuits doesn't function as an isolated complex but contains up to several thousands of different types of interconnected components engaged in a complex regulatory network. This network, when extended to the cell level, ultimately represents a complex map of cell abilities, a plethora of programs that a cell has to follow or, alternatively, could potentially follow. The interactions of these programs will specify the characteristics of tissues, organs and finally of the whole organism. Thus, understanding of the global topological organization of such complex networks is a crucial step towards elucidating a comprehensive functional map for the entire cell, and is critical for deciphering the acquisition of the diverse cell fates, and the maintenance and dynamics of cell functions. All these features are essential to guarantee the development and proper function of cells and organs in the compartmentalized mammalian body. While some variations are tolerated and rescue/failsafe systems are operative surveillance units, intolerable deviations can occur, particularly at key nodes and lead to pathological malfunction. Therefore, network analysis has emerged also as a powerful approach to elucidate disease processes[169].

Though being a relatively recent problem to solve, the history of network "decryption" goes back to 1736 and the famous 'Euler's problem of Seven Bridges of Königsberg' [170] (**Figure 6**). The initial problem was as simple as finding a way of walking around the city by crossing each of the

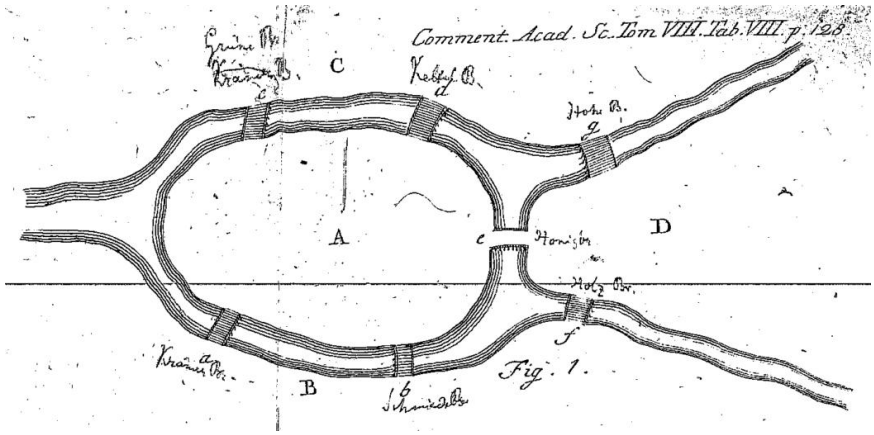


Figure 6. Problem of Seven Bridges of Konigsberg. Adapted from [170].

bridges exactly once. He also addressed the generalized problem: given any division of a river into branches and any arrangement of bridges, is there a general method for determining whether such a route exists. Though in the particular case of the seven Königsberg bridges such a walk was impossible, the reasoning why it is actually impossible led to some of the original concepts of node-edge relationships and the following constraints of a walk through the graph. These ideas initiated the topology and graph theories, the concepts of which have evolved significantly through the past several hundred years and have been applied in studies of diverse networks across multiple disciplines.

Real-world complex systems, abstracted to networks, including biological networks share common global architecture termed the 'small-world'[171] and 'scale-free'. 'Small-world' stands for a network with small characteristic path lengths and a relatively high level of clustering [vii]. 'Scale-free' refers to the to the node connectivity in the real world networks, which have been shown to fit a power-law distribution, with most nodes having few connections and a few nodes being highly connected (scale-free networks) [172,173]. These two key observations initiated a new approach to model biochemical reactions in a cell. Instead of viewing reactions in pathways as interaction of enzyme with a substrate followed by generation of a product or a binding reaction, biochemical interactions were now abstracted to nodes and links ('edges') forming a network[174].

[vii] i.e., groups of nodes have many interactions with one another

3.2. GENE REGULATORY NETWORK RECONSTRUCTION – HOW AND WHY?

There are two fundamental approaches to use the graph theory in the analysis of regulatory biological networks. The first provides an understanding of the global organization of such networks. For this, the properties and attributes computed for individual nodes, links, and/or groups of nodes and links are averaged, or the distribution of such properties is analyzed and compared with the distributions found in randomly reorganized network. The second approach uses the prior knowledge of multivariate experiments (e.g., microarray data sets) in the context of known pathways and networks to infer cause-consequence relationships and regulatory links. Depending on the question of a particular study one can use both approaches or rather try to find key regulators/pathways by analyzing the attributes of nodes or edges.

An important attribute/property of nodes is their ***degree*** - the number of direct neighbors of a node. Different types of biochemical networks across different species were found to have a connectivity degree distribution that fits a power-law function[172,175,176]; this can be explained by the fact the proteins in the cell are pleiotypic, serving many different functions. In real world-networks most nodes have few neighbors but a substantial number of nodes have a high degree, termed ***hubs***. The identification of hubs is often of interest, as they have been shown to be topologically and functionally important: the deletion of genes encoding hub proteins frequently correlates with lethality in yeast (the centrality-lethality rule[175]). Hubs might be master regulators of biological processes[177] and have been found to be preferentially targeted by both bacterial and viral pathogens[178].

In addition, another layer of topological metrics can be analyzed – bottleneck nodes, defined as those interconnecting highly connected nodes or hubs in the system. Previous reports demonstrated that bottleneck nodes might represent highly relevant components of the signal transduction process [179].

One of the approaches to construct a network is to query different interaction databases to identify the 'interactors' of a list of genes or proteins of interest (e.g., differentially expressed genes). The query of protein-protein databases would result in an undirected network where the information of a signal flux direction is not represented, while the use of other databases/tools like CellNet [180] would result in a directed network. In the latter case the database contains the

information of transcription factor- target gene (TF-TG) that enables following the propagation of the signal through the network. Once the network is reconstructed one can identify the hubs and/or bottlenecks of the network. This approach enables the identification of a larger network for analysis than in the case where one restricted the interactions to only those that occur between nodes in the gene/protein list. In addition, this way of network analysis can help to identify sub-networks that are enriched in co-regulated genes, or identify non-differentially expressed nodes that are topologically important in the network, both of which would not be identified otherwise.

However, analysis of complex comprehensive network is a challenge, as the number of nodes and connections can easily extend to tens of thousands. Moreover, biological networks, are not static entities[181], and as the cell undergoes diverse process (e.g., (trans)differentiation), hubs may (dis)appear or the spectrum of hub actions can vary along the temporal dimension[182,183]. In this case integrating contextual information, such as gene expression data, with standard network analysis can provide information about the most relevant key factors and sub-networks in a particular context[184–186]. To address this challenge and identify hubs in networks a number of tools has been developed, including Hubba[187], APID2Net[188], PinnacleZ[189], NetworkAnalyzer[190,191] and CentiScaPe[192]; these tools are based on different node parameters, such as degree, Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC) and other parameters.

Altogether network reconstruction may aid in the identification of potential drug-targets for the development of novel therapies, including not only cancer but also inflammation, degenerative diseases or infectious disease caused by emerging pathogens [193]. Examples of cancer systems biology [194] are the identification of (rare) driver mutations in cancer [195,196] or of pathways associated with survival of cancer patients in view of a personalized therapy [197], or the mode of action of pharmacological compounds [198]. Another area, which gains progressively attention in the scientific community, is the assessment of cells destined for regenerative medicine. Obviously, the demonstration of the ability of such stem/precursor cell to adopt *in vitro* a cell fate and a functionality that is essential for its use in regeneration, as deduced from the reconstructed network and comparison with the (normal) cells to be substituted, will be a milestone achievement towards a successful therapeutic use.

3.3. CHROMATIN STRUCTURE IMPLICATION IN GENE REGULATORY NETWORKS

Systematic mapping of transcription factor binding sites and open chromatin regions have uncovered complex regulatory networks revealing mechanisms of gene regulation [199]. However, in addition to local interactions, 3D contacts between distal regulatory elements play an important role in gene regulation [166,200,201]. Comparison of long-range interactions between cell types revealed that enhancer-promoter interactions are highly cell type-specific [115]. Construction and comparison of distal and proximal regulatory networks revealed a difference in structure and biological functions. Proximal binding events appear to be enriched in genes with housekeeping functions, while many cell-type-specific and dynamic biological processes were regulated by distal binding of TFs [115].

Supporting a causative relationship between cell type-specific GRNs and genome organization, loss of Klf [202], Nanog or Oct4[203] disrupted pluripotency-specific long-range chromatin contacts in pluripotent cells. Furthermore, ectopic recruitment of Nanog to chromatin was sufficient to induce chromatin interaction between the targeted locus and other Nanog-bound regions[203]. These functional studies clearly show that regulatory factors play causal roles in the establishment of chromatin organization

However, despite these advances the mechanism of establishment and regulation of long-range interactions in cell fate acquisition process as well as their role in this process remains elusive. Moreover, how the aberrant re-wiring of chromatin regulatory interactions, their cross-talk with epigenome and their role in GRN establishment of tumorigenesis is a key question to answer in order to understand the cell transformation process. Answering these questions requires (i) a comprehensive map of short-range and long-range interactions between regulatory elements (‘chromatin interactome’), (ii) detailed maps of transcription factor binding (‘cistromes’), (iii) histone modification (‘epigenome’) and (iv) gene expression (‘transcriptome’) analysis in the investigated cells and (v) integration of all these levels along the temporal dimension during the processes of normal and aberrant cell fate acquisition.

To address these questions for two model systems of cell fate transitions, namely the neuronal and endodermal cell differentiation induced by the morphogen retinoic acid and the stepwise

tumorigenesis of primary human cells, which are the topics of my PhD project, we conducted integrative transcriptome, epigenome and chromatin architecture studies. Through extensive integration with thousands of available genomic data sets, we deciphered the gene regulatory networks of these processes and revealed new insights in the molecular circuitry of cell fate acquisition.

THESIS OBJECTIVES:

Study I. Reconstruction of gene regulatory networks of tumorigenesis to define key transcription factors and chromatin remodelers in cell transformation. (Malysheva Valeriya, Marco-Antonio Mendoza-Parra, Mohamed Ashick Mohamed Saleem and Hinrich Gronemeyer. Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis. *Genome Medicine*. (2016) **8,** 1–16 2016)

Alterations in genetic and epigenetic landscapes are known to contribute to the development of different types of cancer. However, the mechanistic links between transcription factors and the epigenome which coordinate the deregulation of gene networks during cell transformation are largely unknown.

To monitor the progressive deregulation of gene networks upon immortalization and oncogene-induced transformation while ensuring cell-to-cell comparability, a stepwise human cellular transformation model was chosen for the current study. In this model primary human cells (BJ) were first immortalized and pre-transformed into BJEL cells by the introduction of hTERT (the catalytic subunit of telomerase) and the large T and small t-antigen of the SV40 early region. The full transformation into *bona fid*e tumor cells was achieved by overexpression of the c-MYC oncogene. The experimental advantage of this system is that normal, immortalized, and tumor cells are near isogenic, as revealed by single-nucleotide polymorphism (SNP) analysis, such that data obtained for the pre-transformed and cancer cell can be accurately compared with the normal counterpart.

We applied a systems biology approach by combining transcriptome and epigenome data for each step during transformation and integrated transcription factor–target gene associations in order to reconstruct the gene regulatory networks that are at the basis of the transformation process. The following questions were addressed:

(i) how are the global patterns of gene expression and chromatin organization changed;

(ii) how are these levels coordinated during tumorigenesis; and

(iii) what is the regulatory role of chromatin remodelers.

We reconstructed gene regulatory networks that revealed the alterations occurring during human cellular tumorigenesis. Using these networks, we predicted and validated several transcription factors as key players for the establishment of tumorigenic traits of transformed cells. Our study suggested a direct implication of chromatin remodelers/modifiers (CRMs) in oncogene-induced tumorigenesis and identified new CRMs involved in this process. This was the first comprehensive view of the gene regulatory network that is altered during the process of stepwise human cellular tumorigenesis in a virtually isogenic system; it generated a working basis for understanding how this interplay is deregulated in a cellular model of human cancer.

Study II. Reconstruction of cell fate-regulatory programs in stem cells in order to reveal pivotal regulatory factors in cell differentiation (Marco-Antonio Mendoza-Parra, Valeriya Malysheva, Mohamed Ashick Mohamed Saleem, Michele Lieb, Aurelie Godel, and Hinrich Gronemeyer. Reconstructed cell fate-regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis. *Genome Research.* (2016). doi:10.1101/GR.208926.116)

Cell lineages, which shape the 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. Even a single chemical trigger, such as the morphogen all-trans retinoic acid (ATRA), can induce the complex network of gene-regulatory decisions that matures a stem/precursor cell to a particular step within a given lineage. The use of RA (rather than complex culture conditions) as defined trigger of regulatory events is essential to elucidate the dynamically regulated "downstream" gene networks.

Embryo carcinoma (EC) cells can differentiate into all three primary germ layers [204]. While F9 EC cells differentiate into primitive endoderm when treated with RA in monolayer, parietal or visceral endodermal differentiation is observed when RA is either complemented with cyclic AMP or when cells are cultured as embryoid bodies in suspension. P19 EC cells differentiate

into either skeletal muscle or neuronal cell types upon treatment with dimethlysulfoxide or RA, respectively. Thus, RA can induce cell fate commitment towards two distinct primary germ layers. However, the temporal evolution of the corresponding gene programs and the regulatory mechanisms remained elusive.

In this study we dissected the GRNs involved in the RA-induced neuronal or endodermal cell fate specification by integrating dynamic RXRA binding, chromatin accessibility, promoter epigenetic status and the transcriptional activity inferred from RNA polymerase II mapping and transcription profiling. Our data revealed how RA induces a network of transcription factors (TFs), which directs the temporal organization of cognate GRNs, thereby driving neuronal/endodermal cell fate specification. Modeling signal transduction propagation using the reconstructed GRNs indicated critical TFs for neuronal cell fate specification, which were confirmed by CRISPR/Cas9-mediated genome editing.

Overall, this study demonstrated that a systems view of cell fate specification combined with computational signal transduction models provides the necessary insight in cellular plasticity for cell fate engineering. This integrated approach can be used to monitor the *in vitro* capacity of (engineered) cells/tissues to establish cell lineages for regenerative medicine.

Study III. Quality assessment of long-range chromatin interaction datasets (Marco-Antonio Mendoza-Parra, Matthias Blum, Valeriya Malysheva, Pierre-Etienne Cholley and Hinrich Gronemeyer. LOGIQA: a database dedicated to long-range genome interactions quality assessment. *BMC Genomics* (2016) **17,** 355)

Massive parallel DNA sequencing in combination with a variety of molecular biology techniques provides functional insights into a plethora of regulatory levels and functions, including epigenomics and protein-genome interactions (e.g., ChIP-seq, MeDIP-seq), global transcriptional activity (e.g., RNA-seq, GRO-seq, Ribo-seq), protein-RNA interactions (e.g., CLIP/RIP-seq), chromatin accessibility (e.g., DNase-seq, FAIRE-seq, ATAC-seq, MNase-seq) and the 3-dimensional chromatin organisation (HiC, ChIA-PET, etc.).

While data acquisition is not anymore an issue, today's challenge is the availability of user-friendly computational solutions to interrogate and integrate - in a comparative manner - billions of data points from different types of functional genomics datasets. In addition, the number of genomics data linked to various cell/(patho)physiological functions increase exponentially in public repositories like the Gene Expression Omnibus (GEO). However, the exploitation of the functional genomics information in these repositories is seriously limited by the lack of information on the quality of these datasets, meanwhile the evaluation of the data sets ensuing their comparability is a crucial step in integrative studies.

In this context, we have developed previously a quality control system dedicated to ChIP-seq and enrichment-related datasets [205] (www.ngs-qc.org). Subsequently, we created LOGIQA (www.ngs-qc.org/logiqa), a database hosting quality scores for long-range genome interaction assays accessible through a user-friendly web-based environment dedicated to quality-scored visualization of long-range interaction maps. Currently, LOGIQA harbors QC scores for >900 datasets, which provides a global view of their relative quality and reveals the impact of genome size, coverage and other technical aspects. LOGIQA provides a user-friendly dataset query panel and a genome viewer to assess local genome-interaction maps at different resolution and quality-assessment conditions.

Study IV. Chromatin structure dynamics in cell fate acquisition (Malysheva Valeriya, Marco-Antonio Mendoza-Parra*, Matthias Blum and Hinrich Gronemeyer*. Chromatin structure dynamics directs cell fate acquisition. *Manuscript in preparation*)

The cell fate acquisition is a highly complex phenomenon that involves a plethora of intrinsic and extrinsic instructive signals that direct the lineage progression of stem cells, the regulatory circuitry to generate, for example, the early basic architecture and functions of an organ acts rather cell autonomously, as cerebral organoids have been generated *in vitro* from ES or iPS cells [206]. The understanding of regulatory mechanisms that underlie the cell fate decision processes not only brings the fundamental understanding of cause-and-consequence relationships inside the cell, but also open the doors to the directed trans-differentiation.

We have previously defined the dynamic gene-regulatory networks underlying endodermal and neuronal differentiation induced by the morphogen all-*trans* retinoic acid (RA) [207]. Here we assessed the contribution of the chromatin interactome to commitment and selective acquisition of these two cell fates and observed a previously unrecognized highly dynamic re-wiring of chromatin domains during cell differentiation.

One of the major challenges in the current functional genomics era resides in the possibility of explaining biological systems behavior from the integration of multiple "omics" readouts. In this context, to give a comprehensive view of regulatory mechanisms we integrated the temporal transcriptional response together with the reorganization of the epigenetic make-up and 3-dimensional chromatin organization.

This systems biology integrative approach indicated key regulatory elements that respond to the initial signal. Overall, our data revealed an enormous capacity of the morphogen to reorganize long-range chromatin interactions as a means to "read" distant epigenetic signals to drive cell fate acquisition and suggest that the differential establishment of chromatin contacts directs the acquisition of the two cell fates

Study V. Chromatin structure dynamics in tumorigenesis (Malysheva Valeriya, Marco-Antonio Mendoza-Parra, Matthias Blum and Hinrich Gronemeyer. Chromatin dynamics during tumorigenic transformation. *Manuscript in preparation*)

The 3-dimensional structure of cancer cell chromatin has become an interest of recent research but the focus has been so far on the effect of frequent chromosomal translocations (e.g., *BCR-ABL*, *MYC-IGH*) or on mutations in key architectural factors, like the subunits of the cohesion complex, which were found in a diverse set of cancers [208]. One of the insights gained from these studies is that the distribution of chromosomal alterations is related to the positioning of these alterations in the 3D chromatin architecture [209]. Only very recently comparative direct global 3D chromatin structure studies between a particular cancer and the normal cells of origin have been reported, as for prostate cancer [210]. Yet, in all these studies normal tissue is compared with very late stages of the tumorigenic evolution, including the development of multiple clonal cancer cell lineages and major chromosomal aberration (i.e., loss/gain of parts of chromosomes/alleles

including LOH, generation double minutes, chromosomal translocations) due to genomic instability.

As a continuation of the Study I, here we have set out to understand the net consequences of cell immortalization and c-Myc protooncogen-induced tumorigenesis of the global chromatin structure of normal primary human cells in a stepwise tumorigenesis model, which does not show any of the major consequences of genome instability. Thus, we describe the very early and nevertheless global alterations in chromatin architecture due to two precisely defined immortalizing and oncogenic insults.

Ananalysis of the dramatic changes of chromatin interactome observed during tumorigenesis in view of the immortalizing actions of genes expressed from the early regin of SV40 and the transforming activity of MYC, including the integration of chromatin structure data with our previously described transcriptome and epigenetic landscape [26] coupled with the analysis of chromatin accessibility (FAIRE-seq) for each step of tumorigenic transformation is ongoing and will help to better understand the impact of tumorigenic elements on the chromatin structure and, in particular, the mechanisms through which MYC is acting as a global chromatin remodeler inducing the acquisition of aberrant (tumorigenic) cell fate.

RESULTS

PUBLICATION N°1

Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis.

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

Genome Medicine 8:57, 2016.

Malysheva et al. Genome Medicine (2016) 8:57
DOI 10.1186/s13073-016-0310-3

Genome Medicine

RESEARCH Open Access

Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis

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

Abstract

Background: Alterations in genetic and epigenetic landscapes are known to contribute to the development of different types of cancer. However, the mechanistic links between transcription factors and the epigenome which coordinate the deregulation of gene networks during cell transformation are largely unknown.

Methods: 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. We applied a systems biology approach by combining transcriptome and epigenome data for each step during transformation and integrated transcription factor–target gene associations in order to reconstruct the gene regulatory networks that are at the basis of the transformation process.

Results: We identified 142 transcription factors and 24 chromatin remodelers/modifiers (CRMs) which are preferentially associated with specific co-expression pathways that originate from deregulated gene programming during tumorigenesis. These transcription factors are involved in the regulation of divers processes, including cell differentiation, the immune response, and the 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. Decortication of transcription factor targets enabled us to define potential key regulators of cell transformation which are engaged in RNA metabolism and chromatin remodeling.

Conclusions: We reconstructed gene regulatory networks that reveal the alterations occurring during human cellular tumorigenesis. Using these networks we predicted and validated several transcription factors as key players for the establishment of tumorigenic traits of transformed cells. 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 the gene regulatory network that is altered during the process of stepwise human cellular tumorigenesis in a virtually isogenic system.

Background

During the past decade great progress has been made in identifying landscapes of genetic alterations which act at different gene regulatory levels and lead to the development of numerous cancer phenotypes. While much is known about altered signaling, recent studies have shown that the epigenomes of cancer cells can also dramatically deviate from those of the corresponding normal cells. However, little is known about the global deregulation of the transcriptome and epigenetic landscapes, as well as their crosstalk during the multistep process of cell transformation.

The deregulatory processes that ultimately turn a normal cell into a tumor cell are conceptually well understood and have been described as "hallmarks of cancer" [1]. At the same time, the sequencing of cancer genomes provided an encyclopedia of somatic mutations, revealing the difficulty of working with primary human cancer cells that carry a small number of "driver" and a high number of variable "passenger" mutations [2]. To reduce this complexity and ensure cell-to-cell comparability, a

* Correspondence: hg@igbmc.u-strasbg.fr
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

stepwise human cellular transformation model [3] was chosen for the current study. In this model primary human cells (BJ) were first immortalized and pre-transformed into BJEL cells by the introduction of hTERT (the catalytic subunit of telomerase) and the large T and small t-antigen of the SV40 early region. The full transformation into bona fide tumor cells was achieved by overexpression of the c-*MYC* oncogene (Fig. 1a). The experimental advantage of this system is that normal, immortalized, and tumor cells are near isogenic, as revealed by single-nucleotide polymorphism (SNP) analysis (Additional file 1: Figure S1), such that data obtained for the pre-transformed and cancer cell can be accurately compared with the normal counterpart.

Epigenetic modifications comprising both DNA methylation and post-translational histone modifications or histone variants have been shown to affect transcription regulation. Different methylation patterns of lysine residues of histone H3 are widely used markers to describe the active and silenced states of transcription at the corresponding chromatin loci [4]. However, we know very little about how this regulation is altered during the process of

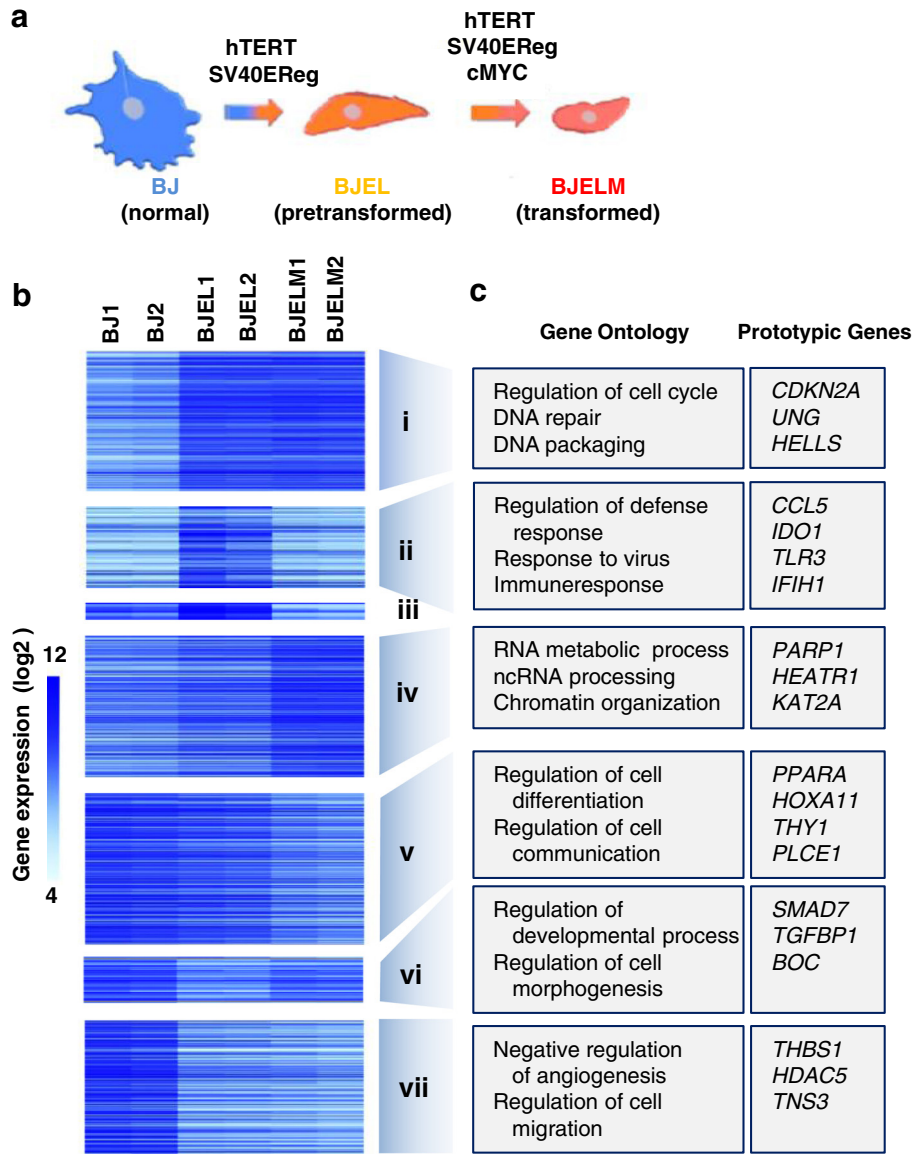


Fig. 1 Transcriptional analysis of the stepwise cell transformation process. **a** BJ stepwise transformation cell model system. **b** Changes in the expression rate of differentially expressed genes (DEGs) in normal, immortalized, and transformed cells. **c** Biological process-based Gene Ontology analysis (performed with DAVID, $p < 0.05$; Additional file 2: Figure S2) for each co-regulated group of genes (co-expression pathways i to vii) and prototypic genes

tumorigenesis. The current study is among the first to reveal the interplay between the epigenome and transcriptome in a stepwise tumorigenesis system; it generates a working basis for understanding how this interplay is deregulated in a cellular model of human cancer. Here we addressed the following questions: (i) how are the global patterns of gene expression and chromatin organization changed; (ii) how are these levels coordinated during tumorigenesis; and (iii) what is the regulatory role of chromatin remodelers.

Methods

Cell culture

Primary human diploid BJ foreskin fibroblasts were obtained from the American Type Culture Collection (ATCC). Genetically defined cells of the BJ stepwise system (BJ and BJEL) were generously provided by Drs Hahn and Weinberg. BJELM cells were produced previously in our laboratory by retroviral transfection of the BJEL cell with pBabe-MYC-ER [5]. Cells were cultured in monolayer conditions in Dulbecco's modified Eagle's medium (DMEM)/M199 (4:1) (with 1 g/l glucose) supplemented with 10 % heat inactivated fetal calf serum (FCS) and gentamicin. The medium for BJEL was supplemented with G-418 (400 μg/μl) and of hygromycin (100 μg/μl). The medium for BJELM was supplemented with G-418 (400 μg/μl), hygromycin (100 μg/μl) and puromycin (0.5 μg/ml) and continuously grown with 10^{-6} M 4-hydroxytamoxyfen (4-OHT).

TRAIL-induced apoptosis measurement

Cells were seeded in 24-well plates and incubated until the subconfluent state and incubated with recombinant human tumor necrosis factor-related apoptosis-inducing ligand (TRAIL; 200 ng/ml) for 8 h to monitor and measure apoptosis. The whole cell content, with floating (apoptotic) and attached cells, was collected for apoptosis measurement. Cell pellets were permeabilized on ice with 100 μg/ml digitonin and stained with APO2.7 (1:5; Beckman Coulter, USA). Apoptosis was measured by fluorescence-activated cell sorting (FACS) and quantified by detection of 7A6 mitochondrial antigen.

Western blotting and antibodies

The whole cell protein extract was prepared using lysis buffer comprising 0.5 M LSBD (0.5 M NaCl, 50 mM Tris–HCl pH 7.9, 20 % glycerol, 1 % NP-40, 1 mM DTT), 0.3 % NP-40, 1× Protease Inhibitor Cocktail (Roche), 1 mM NaF, 10 mM Na_3VO_4, 1 mM PMSF, 0.125 μM okadaic acid. The protein concentration of extracts was measured using a Protein Assay reagent (Bio-Rad Laboratories). Proteins (50 μg) were separated by SDS PAGE, transferred to nitrocellulose membranes, and incubated with indicated antibodies. Antibodies used were as follows: c-MYC (N-262, rabbit; Santa Cruz, sc-764), SV40 T (Pab 108, mouse; Santa Cruz, sc-148), and β-actin (C-11, goat; Santa Cruz, sc-1615).

Double nickase transfection by CRISPR-Cas9

Transfections were performed using double nickase plasmids using the manufacturer's protocol and targeting the following factors: DHX33 (sc-404530-NIC2), CHD7 (sc-404017-NIC2), NOLC1 (sc-402907-NIC2), GTF3C4 (sc-411269-NIC2), PRMT3 (sc-406688-NIC2). Lipofectamin 2000 was used as the transfection reagent at a final concentration of 50 nM.

In brief, cells were seeded in six-well plates and grown for 24 h in antibiotic-free standard growth medium to achieve 80 % confluence. Transfection was performed with 1 μg of CRISPR plasmids followed by 24-h incubation. At the end of the incubation period the medium was replaced with a standard medium with antibiotics. Successfully transfected cells were sorted 24 h later by FACS, using green fluorescent protein as a marker, and used for other experiments.

Test for anchorage-independent colony formation on soft agar

First, six-well plates were covered with "bottom agar" consisting of 4 % FCS, DMEM, and 0.7 % agar. Afterwards, 1000 cells (per one replicate) were mixed with a "top agarose" preparation consisting of DMEM 1×, 10 % FCS, and 0.35 % agar. The final mix was put on the top of the "bottom agar". Cells were cultured with appropriate controls in soft agar medium for 21 days. Cells were fed once or twice per week with cell culture medium. Following this incubation period, formed colonies were stained with 0.5 ml of 0.005 % Crystal Violet for several hours and the number of colonies formed per well was quantified.

Real-time quantitative PCR

Total RNA was extracted from cells using the GenElute™ Mammalian Total RNA Miniprep kit (Sigma). The extracted RNA (2 μg) was used for reverse transcription (AMV-RTase, Roche; Oligo(dT) New England Biolabs; 1 h incubation at 42 °C followed by 10 min at 94 °C). Transcribed cDNA was diluted tenfold and used for real-time quantitative PCR (RT-qPCR; Roche instrument LC480). For confirmation of introduced gene and marker gene expression the following primers were used: *TERT*, forward GCCTTCAAGAGCCACGTC, reverse CCACGAACTGTCGCATGT; *MYC*, forward CACCAG CAGCGACTCTG, reverse GATCCAGACTCTGACCT TTTGC; *CCND2*, forward GGACATCCAACCCTAC ATG, reverse CGCACTTCTGTTCCTCACAG; *THBS1*, forward CAATGCCACAGTTCCTGATG, reverse TGG AGACCAGCCATCGTC; *CHD7*, forward CACCTGAA

GCATCACTGTAACAA, reverse TCACTTCTTGTCTT AGGTAGTACAGCA; *DHX33*, forward TGGTGAAAG CTGCACAGAAG, reverse CCATCGTAGCTGACATC ACAA; *NOLC1*, forward ATAAGTTCGCCAAAGCGA CA; reverse CTAAGAGGGAAGAGGCATTGG; *PRMT3*, forward AGGATGAGGACGATGCAGAT, reverse TTCT TCAGCAGATGTGAATAACCT; *GTF3C4*, forward TTG CTCCATGACAGCATTG; reverse GGGGCTTTGCAG TAACCTCT.

To assess relative gene expression, all qPCR measurements were normalized relative to the constitutively expressed GAPDH mRNA levels assessed with the following primers: *GAPDH*, forward CGACCACTTTGT CAAGCTCA; reverse AGGGGTCTACATGGCAACTG.

Transcriptomics

Transcriptome analysis was performed using an Affymetrix Gene 1.0 ST Array in two biological replicates for each cell line, providing 1 μg of extracted RNA for library production. For comparing BJ, BJEL, and BJELM cells' generated transcriptomes, we normalized all raw CEL files with the Affymetrix software Expression Console version 1.3.1 to calculate probe-set signal intensities using RMA algorithms with default settings. High reproducibility between the corresponding biological replicates was evaluated by calculating the Pearson correlation coefficient and skewness parameter between replicates and between BJEL and BJELM relative to BJ (Additional file 2: Figure S2).

To identify differentially expressed genes (DEGs), we compared BJEL versus BJ and BJELM versus BJ (in biological replicates). Thus, to identify confident DEGs, we used a modified *t*-test [6] for measurements coming from independent normal populations with unequal variances; this method aims to specifically address the question of differential expression in tests involving two samples (BJ versus BJEL or BJ versus BJELM) in which the experiments were performed in repeats. Finally, the probability of having a t-statistic value by chance was calculated and a threshold (significance level of 0.05) was applied.

Inferring transcription factors involved in deregulated gene expression during cell transformation

For the selection of transcription factors (TFs) associated with particular co-expression pathways we used the CellNet database of TF–target gene (TG) associations. We first selected TFs that are associated with >10 % of DEGs that constitute a given co-expression pathway. Then we selected TFs with promoter-associated RNA polymerase II (RNA Pol II), which gave rise to 142 TFs. Finally, we assessed the relevance of these TFs in distinct co-expression pathways using a hypergeometric distribution test with subsequent hierarchical clustering (Fig. 2).

Chromatin immunoprecipitation assays

BJ, BJEL, and BJELM cells were fixed with 1 % paraformaldehyde (Electron Microscopy Sciences) for 10 min at room temperature. Chromatin immunoprecipitation (ChIP) assays were performed according to the following conditions: chromatin sonication and immunoprecipitation in lysis buffer (50 mM Tris–HCl pH 8, 140 mM NaCl, 1 mM EDTA, 1 % Triton, 0.1 % Na-deoxycholate) complemented with protease inhibitor cocktail (Roche 11873580001); two washes with lysis buffer; two washes with lysis buffer containing 500 mM NaCl; two washes with washing buffer (10 mM Tris–HCl pH 8, 250 mM LiCl, 0.5 % NP-40, 1 mM EDTA, 0.5 % Na-deoxycholate); two washes with TE buffer; elution at 65 °C; 15 min at 65 °C in elution buffer (50 mM Tris–HCl pH 8, 10 mM EDTA, 1 % SDS).

RNA Pol II (Santa Cruz sc-9001 H-224), H3K27me3 (Millipore 07-449), H3K4me3 (Abcam 8580), H3K9ac (Abcam 4441), and H3K27ac (Abcam 4729) antibodies were purchased from their corresponding commercial suppliers. RNA Pol II ChIP assays were performed with 3×10^6 cells, while histone modification marks were evaluated with 1×10^6 cells. All ChIP assays were validated using positive and negative controls. Specifically, enrichment performance was assessed at promoter regions of genes *SRSF6* and *NEK1* (for H3K4me3, H3K9ac, H3K27ac, RNA Pol II), *NEUROG1* and *MB* (for H3K27me3 validation), and *DPP10* as a cold region, using the following primers: *SRSF6*, forward CGTTCGACAACCAGCCCTT, reverse GGCCC GACTCACCCATTTT; *NEK1*, forward CGTTACCGC CTCTCCAACTT, reverse CTTACCCTACCCTGGCC TCT; *NEUROG1*, forward ACAGATAGAAAGGCGC TCAGA, reverse CGCAACTGGCACAGAGTAAC; *MB*, forward GGCTCACTGGGTGTCCTG, reverse AAG GTATAAAAACGCCCTTGG; *DPP10*, forward GTTT CCAATTTCATCCATGTCC, reverse CACATCAAAC TGGTGGGTGA.

ChIP validation assays were performed by RT-qPCR (Roche instrument LC480 light cycler) using a QuantiTect kit (Qiagen).

Massive parallel sequencing

qPCR-qualified ChIP assays were quantified (Qubit dsDNA HS assay kit, Invitrogen); 3–10 ng of the ChIP material was used for preparing Illumina sequencing libraries following a multiplexing approach (NEXTflexTM ChIP-seq Bioo Scientific, reference 5143-02). Prepared sequencing libraries were sequenced on an Illumina HiSeq2000 instrument. Regular Illumina pipelines were used for image processing and base calling. Sequence files were then aligned to the human genome assembly using default parameters (hg19; Bowtie).

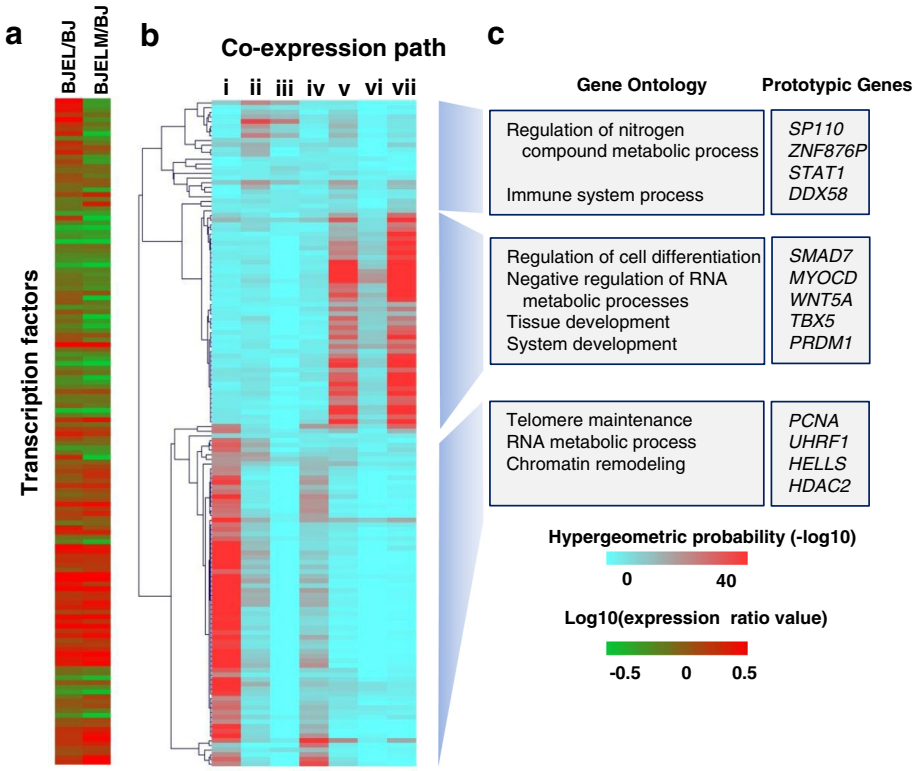


Fig. 2 Association of key transcription factors (TFs) with co-expression pathways. Using the CellNet database of TF–TG associations revealed 142 TFs that were associated with more than 10 % of DEGs. **a** Expression ratios (relative to BJ) of TFs associated with particular co-expression pathways. **b** Heat map of hierarchical clustering illustrates the prevalence of corresponding TFs in the regulation of particular co-expression pathways (the *color bar* corresponds to the – log10(hypergeometric distribution value); *red* corresponds to high-confidence TF–TG associations, *blue* to low-confidence associations). **c** Biological process-based Gene Ontology analysis of clustered groups of TFs associated with particular co-expression pathways ($p < 0.05$) and prototypic genes

Quality control of sequence data

Sequence-aligned files were then qualified for enrichment performance using the NGS-QC Generator tool [7] (http://www.ngs-qc.org/). This methodology provides enrichment quality descriptors in a scale ranging from triple A (best) to triple D (worst). Based on this quantitative method, all ChIP-seq datasets described in this study had at least a triple B quality grade to ensure that only high quality datasets were used for downstream integrative analyses.

Enrichment pattern detection and normalization of ChIP-seq intensity profiles

Relevant binding sites in all ChIP-Seq (except the H3K27me3 dataset, which was analyzed with the SICER tool [8]) datasets were identified with MeDiChI-Seq [9]. To enable a comparison of ChIP-seq profiles of the same target between different cell lines, a normalization procedure over profile global amplitudes prior to further analyses was applied using a Quintile-based approach. Briefly, we calculate read count intensity for a non-overlapping window of 100 bp across the genome and then normalize these intensities using quantile normalization from the "limma" package. Quantile normalization is a ranking-based approach where calculated read count intensities are sorted and ranked for each sample. The corresponding ranked values between samples are adjusted into a mean value. The impact of normalization was assessed using MA plots before and after normalization. First, we normalize all datasets associated with a given target; then normalized target datasets are brought to the same scale via a z-score normalization. A detailed description of this quantile normalization procedure, which is applicable for a variety of ChIP-seq and enrichment-related next-generation sequencing datasets and is available as part of Epimetheus, a user-friendly dedicated tool, is going to be presented in a further publication (in preparation).

Integration of transcriptome and epigenome data

To integrate transcriptome data with chromatin state dynamics, we performed unsupervised clustering of ChIP signals for each target that was assessed in the current study within a 1-kb window of each transcription

start site (TSS) for all DEGs, comprising a total of 7616 transcripts. Histone marks or RNA Pol II binding were tagged as "present" at the TSS of the DEG if it satisfied the following criteria: (i) the peak was detected—the summit of the detected peak (by MeDiChI-Seq [9] or by SICER [8]) was 500 bp up- or downstream of the TSS of the DEG; (ii) the peak was of high intensity after normalization—following quantile and Z normalization the Z-score of a given peak was >1.65 (P_{95}); (iii) the peak was robust—the signal had to be robust with less than 15 % dispersion after the subsampling procedure (NGS-QC tool, http://www.ngs-qc.org/). Afterwards, unsupervised clustering of all the possible combinations of histone marks and RNA Pol II at the TSS of DEGs was performed. A heatmap of chromatin state dynamics represents the median enrichment for each cluster of genes within ±1.5 kb of a TSS of a DEG at each stage of the stepwise transformation model (Fig. 3a, d). At the next step we assessed whether dynamic patterns of chromatin states are associated with particular groups of co-expressed genes (Fig. 3b, d).

Gene regulatory network reconstruction

To provide a comparative view of the signal transduction events governing the cell transformation in the stepwise model system, we reconstructed a gene regulatory network (GRN) by combining several layers of information from three different databases: (i) the MiMI, which contains protein–protein, DNA–protein, and other interaction data, querying the interactions only between DEGs [10]; (ii) CellNet, a collection of directed TF–TG interactions [11, 12], where the TFs listed in the CellNet Gene Regulatory Network (GRN) collection were associated with genes differentially expressed in the BJ/BJEL/BJELM model system; (iii) several publically available MYC-targeted ChIP-seq datasets (see "Methods"). The integration of DEG-related interactions in the Cytoscape platform (version 2.8.3) resulted in a dense cell type-specific GRN composed of 1265 nodes and 5327 edges which were then organized according to the transformation steps and gene co-expression pathways. Furthermore, a two-step GRN reduction process was applied by using a double screening system in the Hubba tool [13] to define the highly connected nodes ("hubs"). In addition, a second layer of topological metrics reduction was applied by scoring for "bottleneck" nodes since previous reports demonstrated that, in addition to highly connected nodes ("hubs"), bottleneck nodes (defined as those interconnecting highly connected nodes or hubs in the system) might represent highly relevant components in the system [14]. In particular, bottleneck nodes in signal transduction systems might correspond to essential entities required for the flow of the signal transduction driving the phenomenon of interest. Definition of hubs and bottlenecks was performed using topological metrics, such as MNC (Maximum Neighborhood Component), DMNC (Density of Maximum Neighborhood Component) and Bottleneck [13]. This reduction process generated GRNs composed of 253 nodes and 2657 edges. The organization of the reduced GRN and its visualization were performed with the Cytoscape package Cerebral [15]. As part of the visualization options in Cytoscape, the differential expression levels in BJEL and BJELM cells per node were presented in a heat map format such that the transcriptome dynamic changes could be visualized. The changes of node color for groups ii, iii, iv, and v in Fig. 4a and 4b indicate the change in expression of co-regulated genes during the transformation process.

Analysis of publically available ChIP-seq datasets used for targeting MYC enrichment

The following ChIP-seq datasets from the Gene Expression Omnibus (GEO) were used to identify MYC enrichment sites: GSM1088663 (HeLa cells), GSM896988 and GSM1000576 (BJ cells), and GSM748557 (NHEK cells). The raw sequencing files were aligned with Bowtie using default parameters and processed with MACS for peak annotation. A threshold of – log10(p value) ≥300 was applied to select peaks with high confidence.

Results

Transcriptome dynamics during the cell transformation process

Following validation of the stepwise tumorigenesis model, which included the determination of TRAIL (tumor necrosis factor-related apoptosis-inducing ligand) sensitivity [16] (Additional file 2: Figure S2), we assessed the global gene expression in all three cell lines and the ratio of expression levels of immortalized to normal cells (BJEL/BJ) and cancer relative to normal cells (BJELM/BJ). Genes exhibiting >2- and <0.5-fold expression changes with a significance level of $p < 0.05$ (modified t-test, "Methods") were considered up- and down-regulated, respectively, and classified as differentially expressed genes (DEGs). The resulting 1700 DEGs were subdivided into seven groups of co-regulated genes according to their expression characteristics during the subsequent steps of transformation (Fig. 1; Additional file 2: Figure S2a, b). Nearly half of the genes (47 %) showed an altered expression level at the pre-transformation step and 65 % of genes changed expression level after full transformation by c-*MYC* expression. Interestingly, about 12 % of these genes changed expression after both immortalization and c-MYC-mediated transformation.

Gene Ontology (GO) analysis revealed that each pathway of DEGs is enriched for distinct GO terms (Fig. 1b; Additional file 3: Figure S3a). Notably, those enriched in

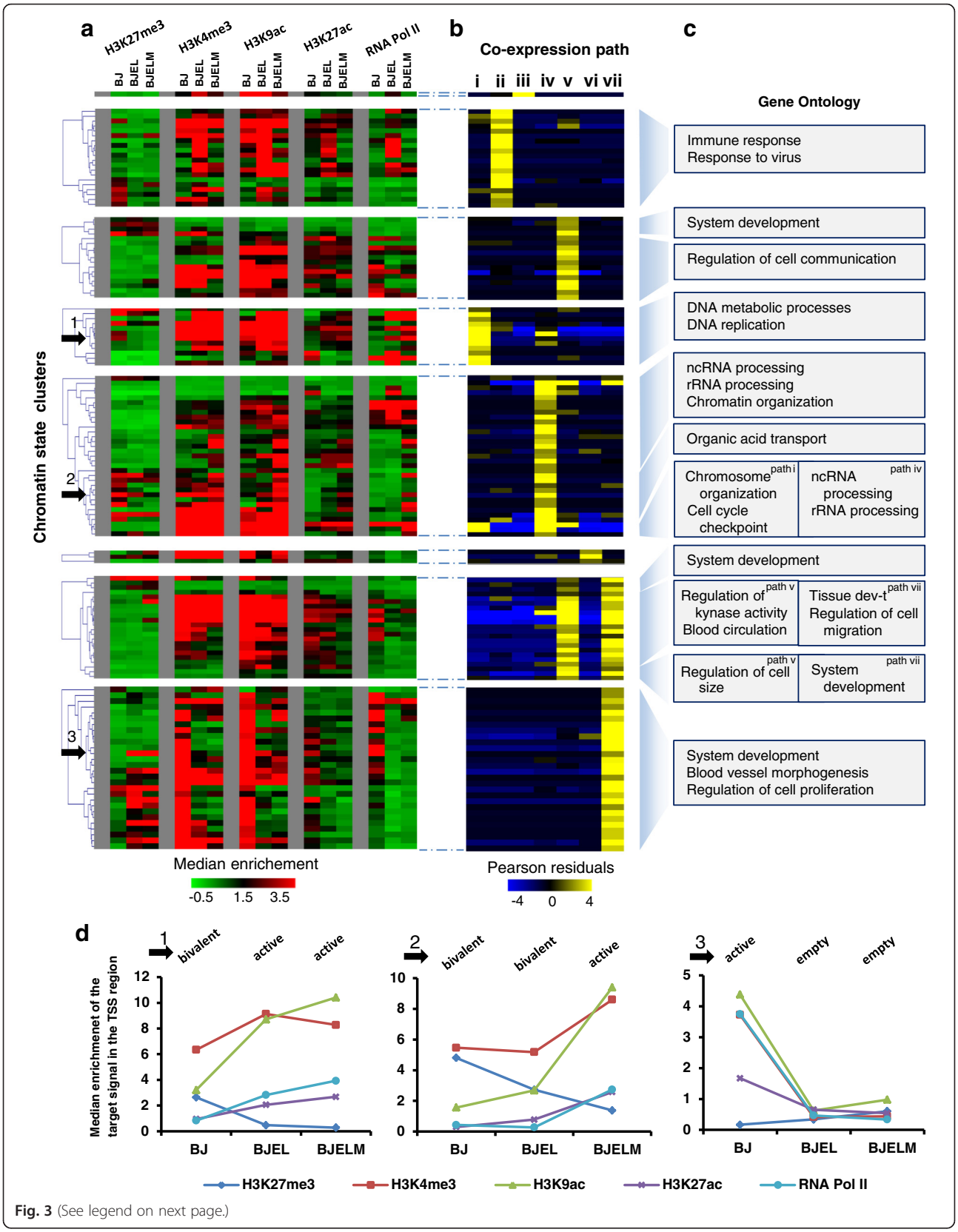


Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Chromatin state transitions in promoters of differentially expressed genes during the cell transformation process and integration of epigenetic data (chromatin state clusters) with transcriptome dynamics (co-expression pathways). **a** Hierarchical clustering of transcripts based on enrichment of histone modifications and RNA Pol II at the promoter of DEGs. The *color* represents the median enrichment for each cluster of genes within ±1.5 kb of a TSS of a DEG. **b** Heat map illustrating the prevalence of chromatin state clusters in particular co-expression paths. The *color* represents Pearson residuals. *Yellow* indicates significant enrichment of transcripts in the corresponding expression pathways with a corresponding chromatin state cluster. **c** Biological process-based Gene Ontology analysis of chromatin state clusters, regrouped by hierarchical clustering (hierarchical tree in **a**), and associated with the same co-expression pathway. **d** Three examples of chromatin state clusters illustrating the evolution of the epigenetic landscape in the stepwise transformation process (*black arrows* in **a**). *Panel 1* correspond to the changes from the bivalent chromatin state in BJ cells to the active state in BJEL and BJELM cells. In the same manner, *panel 2* corresponds to the changes from the bivalent chromatin state in BJ and BJEL cells to the active state in BJELM cells. Finally, *panel 3* corresponds to the chromatin state cluster that characterizes the group of downregulated genes in BJEL and BJELM cells; the promoters of these genes are in the active state in BJ cells but lose all marks in the BJEL and BJELM cells

the group of genes that are down-regulated in transformed cells (pathway v) are associated with regulation of cell motion, cell communication, and regulation of cell differentiation, as well as suppression of angiogenesis, while genes that are progressively induced from the normal to the tumorigenic stage (pathway iv) are significantly enriched for the GO terms ribosome biogenesis and noncoding RNA and rRNA processing. Disease-related GO analysis using DAVID [17, 18] showed significant enrichment of DEGs characteristic for several types of cancers, among them breast, bladder, stomach, and lung cancer (Additional file 3: Figure S3b).

Together these results show that the stepwise transformation model shares multiple similarities with different types of human cancers and is a convenient and reliable cell model for tumorigenesis research. Importantly, several of the deregulated gene expression pathways affect phenomena that are well-described as hallmarks of cancer, such as the activation of angiogenesis, the activation of invasion and metastasis, and regulation of cell cycling [1].

Multiple chromatin remodelers/modulators are dysregulated during cell transformation

To monitor changes of the epigenome during the stepwise BJ transformation process, we assessed first whether chromatin remodelers/modulators (CRMs) were deregulated. Indeed, we detected 24 differentially expressed CRMs, belonging to all three classes of writers, erasers and readers, and other chromatin remodelers (Additional file 4: Figure S4; Additional file 5: Table S1). Fourteen of these changed their expression at the last step of transformation as a consequence of the overexpression of c-*MYC*; interestingly, 12 genes among these were upregulated and are members of such functional groups as methyltransferases, acetyltransferases, demethylases, and related CRMs, indicating that MYC-induced transformation leads to dramatic changes in the epigenome involving CRMs.

The majority of CRMs defined in the current study are deregulated in different types of cancer, such as ovarian, bladder, lung, and many other types (see Additional file 5: Table S1 and Additional file 6 for references). For several CRMs, such as PRMT5 and MINA, the interaction with MYC was reported to be an essential step in cancer development (Additional file 5: Table S1, references 31, 32, 36–39 (listed in Additional file 6)). These observations suggest that CRMs are involved in regulation of tumorigenesis and mediate at least some of the transforming activities of overexpressed *c-MYC*. We would like to emphasize, moreover, that our present approach has identified new candidate CRMs, some of which are putative MYC targets that have not been previously recognized, two "writers" (GTF3C4 and PRMT3), three "readers" (LBR, AKAP1, and MBD5), and the PcG group member MTF2. *LBR* and *MTF2* are upregulated during the first step of transformation, while the other CRMs are deregulated upon c-*MYC* overexpression. Inspection of the publically available MYC cistrome of HeLa cells [19] revealed the presence of high-confidence (see "Methods") MYC-binding sites in the *PRMT3* and *GTF3C4* promoters. Considering that these two genes are induced after MYC overexpression, *GTF3C4* and *PRMT3* are most probably direct targets of MYC in the BJ system.

We conclude from these results that: (i) LBR and MTF2, both involved in transcription repression (Additional file 5: Table S1), are potential regulators of the immortalization process and/or cooperate with the oncogene in the second step; and (ii) PRMT3, GTF3C4, AKAP1, MBD5, and GTF3C4 are new players in the tumorigenesis process which mediate the MYC-dependent effect on chromatin remodeling.

In silico prediction of key TFs involved in deregulated gene expression during cell transformation

To reconstruct the alterations in the activity of TFs during the steps that lead to cell transformation in the BJ model, we integrated information on TF–TG associations

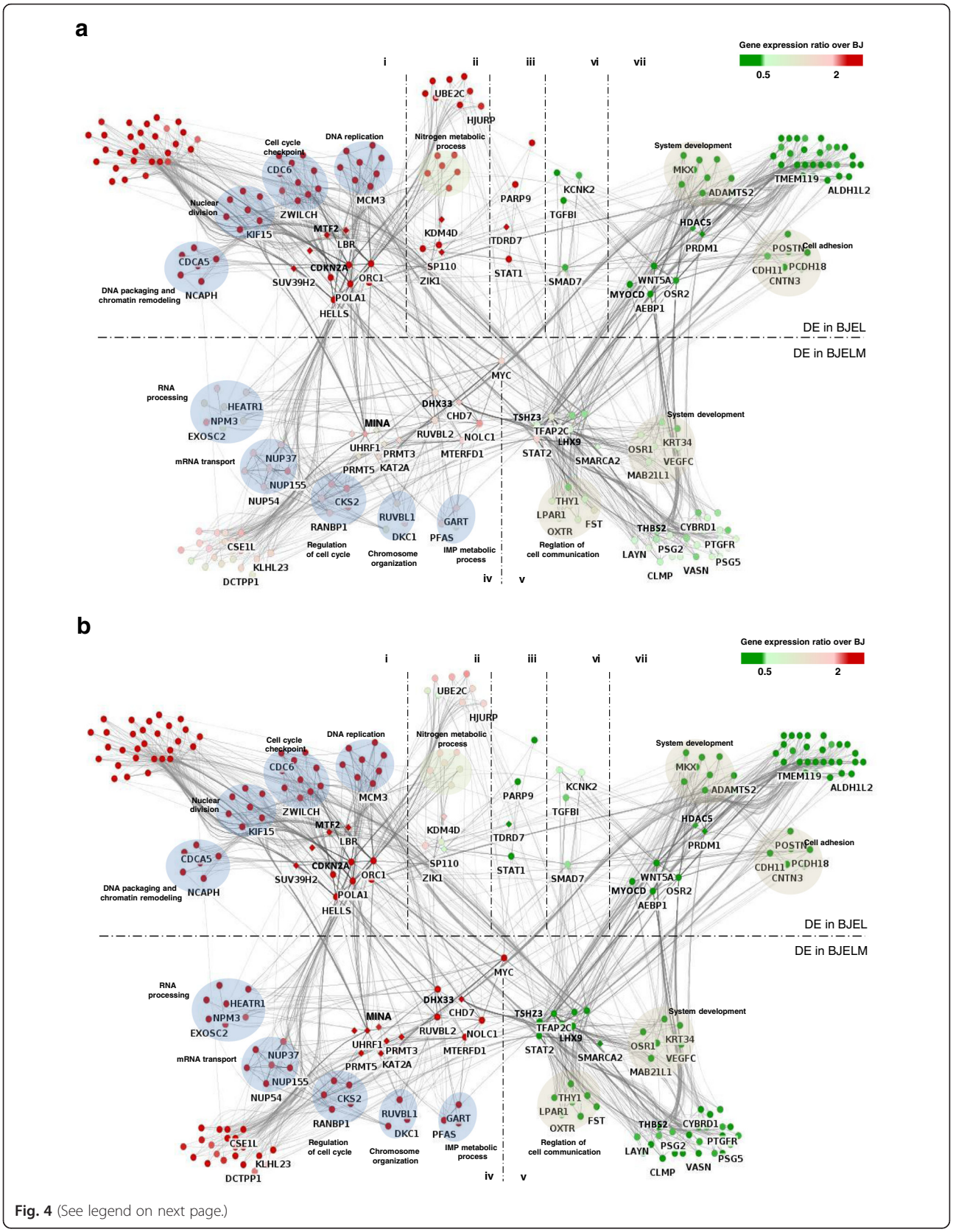


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 Gene regulatory network (GRN) of the BJ stepwise transformation system. **a** GRN of immortalized BJEL cells. **b** GRN of transformed BJELM cells. Chromatin remodelers/modulators are represented as *diamond-shaped nodes*, while other genes, highly connected "hubs", and "bottlenecks" are represented as *circles*. The differential expression levels at immortalization and during the transformation steps were colored per node in a heat map format such that the dynamic changes could be visualized. *Dashed lines* separate the GRN into seven segments corresponding to seven (i to vii) gene co-expression pathways. Functionally related genes are *circled* under an enriched GO term (DAVID, $p < 0.05$)

described in the CellNet database [11, 20]. This led to the identification of 142 TFs (Additional file 7: Table S2), of which 42 are differentially expressed during stepwise tumorigenesis (Fig. 2a). Sorting these TFs for their association with particular co-expression pathways led to a clustering of groups of TFs that were preferentially associated with one or more co-expression pathways (Fig. 2b). According to the hierarchical clustering of TF-specific association with co-expression pathways, we could distinguish at least three subgroups: TFs that are preferentially associated with pathways i and iv, with pathways ii and iii, or with pathways v and vii. Importantly, these distinct groups of co-expression pathway-associated TFs are apparently involved in regulating specific cell biological functions, as revealed by the corresponding GO analysis (Fig. 2c). Specifically, TFs associated with pathways v and vii are involved in regulation of cell differentiation and tissue development, while co-expression pathway i-associated TFs are primarily involved in telomere maintenance and chromatin remodeling.

Notably, co-expression pathway ii comprises genes involved in immune and defense responses; STAT1 is among the TFs that are specifically associated with these co-regulated genes. That *STAT1* is upregulated in pre-transformed cells may reflect the established role of this factor in cell autonomous anti-tumor immune response [21] (Fig. 1; Additional file 3: Figure S3; pathway ii). In addition, STAT1 is known to negatively regulate angiogenesis, tumorigenicity, and metastasis of tumor cells [22] and suppresses mouse mammary gland tumorigenesis [23]. Downregulation of *STAT1* by c-*MYC* overexpression observed in the current study is also detected in Burkitt's lymphoma [24], supporting the concept that immune escape of tumor cells could be promoted by activation of a cellular oncogene [24].

Several functionally related (GO term 45595: regulation of cell differentiation) TFs, such as MYOCD, TWIST1, TBX5, and SMAD7, which are known to be involved in cancer development and/or sustainment [25–27], are specifically associated regulators of genes that are down-regulated along the cell transformation process in our model system (pathways v, vi, and vii). In particular, myocardin (MYOCD), a transcriptional co-factor for smooth muscle cell-specific genes that has been shown to block human mesenchymal transformation [28], was down-regulated in pre-transformed cells. Thus, decreased *MYOCD* expression may contribute to an increased proliferative potential of pre-transformed and transformed cells. In addition, it is likely to contribute to the concomitant loss of fibroblast identity and gain of stem cell identity as revealed by cell identity analysis using CellNet [11] (Additional file 8: Figure S5).

Another functionally related group of TFs that are associated with co-expression pathways i and iv are involved in chromatin remodeling. Among these are UHRF1, HELLS, and HDAC2, all of which are known to affect the tumorigenesis process [29, 30]. Remarkably, RUVBL2/TIP49, a member of the same group that is upregulated in BJELM cells and is known to interact with c-MYC, has been reported to regulate β-catenin-mediated neoplastic transformation [31].

Altogether, the observed associations reveal that the stepwise tumorigenesis model recapitulates aberrations of several regulatory systems, ranging from cell autonomous immune responses to chromatin remodeling and cell (de)differentiation, all of which are features previously reported to be altered in human cancer.

Cell transformation significantly impacts on chromatin state dynamics

Given the known deregulation of cancer epigenomes due to mis-expression or mutation of epigenetic factors [32, 33], the de-regulation of CRMs in the BJ system, and the fact that c-MYC recruits a variety of epigenetic factors and chromatin remodelers to its targets [34, 35], we performed a genome-wide analysis of chromatin state transitions for all three steps of the cell transformation. We used chromatin immunoprecipitation (ChIP) coupled with massive parallel sequencing (ChIP-seq) for several functionally interpretable histone modifications, including H3K27me3 (inactive promoters), H3K4me3, H3K9ac (active promoters), and H3K27ac (active promoters and enhancers). We also determined the chromatin association of RNA Pol II, which is generally enriched at the transcriptional start sites (TSSs) of active promoters.

In view of the dynamic nature of gene expression observed during the tumorigenesis process, we focused on elucidating histone modification patterns at TSSs. To identify gene promoters with a similar pattern, we performed unsupervised clustering of all the possible combinations of histone marks and RNA Pol II-normalized ChIP signals (see "Methods"; Additional file 9: Figure S6)

within 1.5 kb up- and downstream of each TSS for all DEGs, comprising a total of 7616 transcripts (Additional file 10). We detected 26 different combinations of histone marks at DEG promoters and classified them into seven chromatin states (Additional file 11: Figure S7): (a) active (RNA Pol II and at least two histone marks of active transcription are present); (b) weakly active (RNA Pol II and only one histone mark of active transcription); (c) transcription-prone (at least two histone marks of active transcription but no RNA Pol II); (d) bivalent (any of states a to c but accompanied by repressive H3K27me3 marks); (e) ambiguous (only one histone mark or RNA Pol II alone); (f) empty (no signal); and (g) repressed (only H3K27me3). Further, the dynamic changes in chromatin states at the promoters of DEGs through the stepwise transformation process and all possible combinations of chromatin state evolution (chromatin state transitions) were assessed, giving rise to 135 chromatin state clusters, and integrated with the transcriptome changes along the transformation process (Fig. 3).

The majority of clusters revealed highly dynamic chromatin patterns, suggesting that chromatin state regulation of DEGs is tightly linked to, and possibly controls to a significant degree, DEG expression and, thus, the acquisition of the pre-transformed and transformed cell states. The differential regulation of CRMs indicates a tight linkage between, and mutual regulation of, DEGs (including TFs) and CRMs. Interestingly, genes sharing the same co-expression pathway could be further subdivided according to their chromatin state transitions into groups of genes with related functionalities. For example, co-expression pathway iv comprises genes overexpressed at the second step of transformation associated with chromatin patterns, such as gain in activating H3K4me3 and H3K9ac marks in the absence of repressive H3K27me3; this group of genes is involved in rRNA and noncoding RNA processing and chromatin organization. In contrast, a second group sharing the same co-expression pathway, which loses H3K27me3 with a concomitant gain of H3K4me3 and H3K9ac marks, is predominantly linked to organic acid transport (Fig. 3c). Thus, groups of functionally related genes can be distinguished at the chromatin level despite their similar expression patterns.

Reconstruction of GRNs

To provide a comprehensive view of the signal transduction events governing the cell transformation in the stepwise human cellular tumorigenesis model, we reconstructed a GRN integrating gene interaction data from publically available databases with gene expression data from our experiments (see "Methods" for details). This integration process resulted in the establishment of a comprehensive GRN of 1265 nodes and 5327 edges.

To explore the functionally most relevant aspects of the reconstructed network, we reduced its complexity by applying topological criteria to identify highly connected ("hubs") and key connector nodes ("bottlenecks") that are relevant to the investigated signal transduction processes [14]. The reduced network of 253 nodes and 2657 edges (Fig. 4a, b) shows the connectivity between the major nodes, which are possibly functionally involved in the cell transformation process. The network is divided into two parts, showing key regulatory nodes differentially expressed on the first step of transformation in the upper part, while those changing expression levels after c-*MYC* overexpression are depicted in the lower part. Dashed lines in Fig. 4 split the network landscape into seven sections to place co-expressed genes in proximity to each other. The flow of signal goes from the BJEL state (upper part) through the MYC to other TFs and TGs in the BJELM state (lower part). In addition, functionally related highly connected nodes are grouped together in the context of the corresponding enriched GO terms to reveal subprograms, such as regulation of cell adhesion, cell communication, or RNA processing, all of which are hallmarks of cell transformation. In the centre of the network we placed the bottleneck genes, which are supposed to direct the flow of signaling information from the functionally related hubs to the target genes (not shown in the reduced network). The reconstructed gene network pointed towards bottleneck genes that are key factors, like the RNA biogenesis-linked *NOLC1*, *DHX33*, and *CHD7*, as potential key regulators of cell transformation and direct downstream targets of c-MYC, based on the ChIP-seq analysis of publically available data sets ("Methods", Additional file 12). These genes are pivotal for RNA metabolic processes [36–38]. Interestingly, previous studies have shown a correlation between the upregulation of these genes and tumor progression and, indeed, marked increases in rRNA synthesis is a general attribute of many types of cancers [39, 40], suggesting that changes in rRNA synthesis may be a prerequisite alteration in cell transformation. DHX33 has been reported to be an important mediator of rRNA synthesis and cell growth [41]. Furthermore, following the fact that rDNA transcription is greatly influenced by the *RAS*, *MYC*, and *NPM* oncogenes, *DHX33* upregulation was shown to be required for enhanced transcription during RAS activation and for RAS-initiated tumor progression [37]. The observations that *DHX33* is overexpressed in our system following *cMYC* overexpression and has a MYC binding site in its promoter (GSM1088663) suggest that DHX33 is a mediator of MYC signaling. Other key factors that became apparent from these GRNs include TSHZ3, previously reported to correspond to a novel potential tumor suppressor [42], and LHX9, which is aberrantly

methylated and downregulated in malignant gliomas of childhood [43]. Thus, the present reconstructed GRN is a rich source of (novel) regulators of tumorigenesis that could be further studied in suitable (in vivo) systems.

Validation of predicted factors

With the aim of evaluating the biological relevance of the reconstructed GRNs, we assessed the role of the TFs DHX33, NOLC1, and CHD7 as well as that of the CRMs PRMT3 and GTF3C4, all of which act "downstream" of MYC, in cell transformation. Specifically, we used the CRISPR-Cas9 technology to inactivate these genes in BJELM cells and evaluated the consequence of this perturbation on the tumorigenic properties that had been acquired in these cells by the overexpression of c-*MYC* relative to the isogenic non-tumorigenic BJEL precursor cells. For this we used a well-established tumorigenesis assay, namely the acquisition of anchorage-independent growth on soft agar; this assay is widely used as a predictor of tumorigenicity and is considered one of the most stringent assays for studying the malignant transformation of cells [44]. In fact, as illustrated in Fig. 5a, normal BJ and immortalized BJEL cells are not able to grow in an anchorage-independent manner, while the overexpression of c-*MYC* conferred onto BJELM cancer cells the ability to proliferate under these conditions (Fig. 5a, b). Importantly, CRISPR-Cas9-mediated individual inactivation of all tested "downstream" factors of Myc (*CHD7*, *DHX33*, *GTF3C4*, *NOLC1*, and *PRMT3* genes) resulted in a drastic drop in the ability of BJELM cells to form colonies on soft agar, while BJELM cells, as well as mock-transfected BJELM cells ("siGLO"), showed efficient colony formation in soft agar (Fig. 5a). Together our data reveal that each of these factors plays a critical role in mediating key oncogenic effects of *MYC* overexpression in this isogenic model system.

Discussion

Isogenic cellular tumorigenesis models are versatile tools for systems biology studies

Any comparative analysis of normal and tumor cells with the aim of identifying the mutational and deregulatory events that cause cancer is seriously limited, and may even be impossible, if using established cell lines or patent-matched normal and tumor samples. Established cell lines have acquired extensive genetic alterations to support continuous growth in culture ("immortalization") and to escape senescence and/or other failsafe programs [45]. In addition, cancer cells are genetically unstable and carry many genetic abnormalities accumulated due to various conditions, including infections during tissue culture. When using normal and tumor tissue sections from the same patient, the tumor history is generally unknown and both genome instability and clonal heterogeneity/selection limit any comparative data analysis. Yet, it is important to understand the deregulation which occurs at multiple gene regulatory levels when a cell converts into a tumor cell by a minimal set of genetic alterations. Moreover, cancer genomics provides us with sets of "driver" and "passenger" genes, whose implication, alone and in combination, in the tumorigenic process is only known for a very small subset. Thus, there is a need for a model system which can be engineered and recapitulates the basic features of a tumor cell. Such a system was originally developed by Hahn and Weinberg [3] and has been used in this study to decipher the regulatory levels and gene regulatory networks (GRNs) that are altered by "simple" engineered tumorigenesis of primary human cells.

This system is virtually isogenic, thus granting the possibility to dissect GRNs reflecting system deregulation due to the introduction of defined genetic elements. In the present system the overexpressed catalytic subunit of hTERT protects BJ cells from telomere erosion [3, 46].

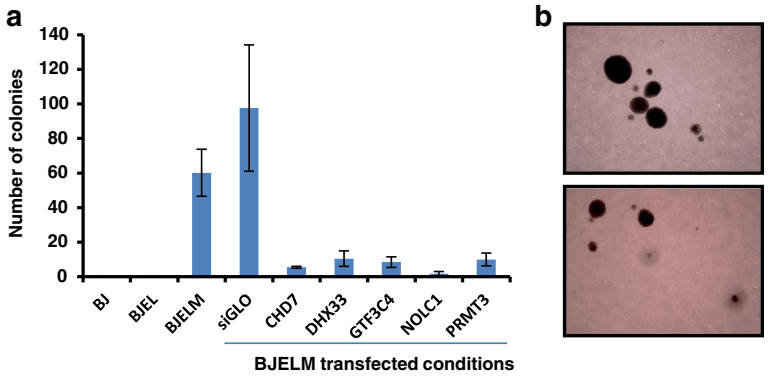


Fig. 5 Validation of predicted factors. **a** Test for anchorage-independent growth on soft agar. All BJELM transfected conditions, except for the control, exhibit drastic decreases in the capacity to form colonies on soft agar. **b** Colonies formed by BJELM cells after 3 weeks of incubation on soft agar. The error bars represent the standard deviation between the biological replicates

In addition, large T and small t-antigen expressed from the SV40 early region inactivate the tumor suppressors RB and P53, thus allowing the cells to evade antiproliferative and apoptotic signals [47]. Finally, overexpression of c-*MYC*, often upregulated through either a stabilization mutation or gene amplification in a wide variety of human cancers, transforms the cells into bona fide cancer cells [48]. Though such a system may seem reduced and simple compared with tumorigenesis in the animal, it nevertheless enabled us to decipher the underlying deregulated gene networks, including alterations of TF activities, and to identify transformation-associated deregulation of epigenome modifiers. As could be expected, GO analysis of DEGs yielded GO terms indicative of cell transformation. Indeed, a marked increase in rRNA synthesis is a general attribute of many cancers [40, 41] and rRNA transcription was shown to be stimulated by c-MYC [49]. This correlates with our observations showing MYC-mediated upregulation of genes functionally related to RNA biogenesis, such as *DHX33*, *HEATR1*, *NOLC1*, and others (Fig. 4a, b, RNA processing functional group). Notably, disease-oriented functional annotation clustering showed that DEGs in the stepwise BJ transformation system comprise genes that are implicated in different types of cancer, such as breast or bladder cancer. In addition, we used cBioPortal [50, 51] to see if genes identified in this study can be correlated with publically available datasets of human cancers (cBioPortal is an exploratory analysis tool that, among others, hosts TCGA (The Cancer Genome Atlas) datasets ready for network analysis). From the cross-cancer alteration analysis under the simultaneous query of MYC, NOLC1, DHX33, and CHD7, a large number of cancers possess alterations in these genes (Additional file 13: Figure S8). In particular breast cancer, neuroendocrine prostate cancer, and ovarian serous cystadenocarcinoma have the highest rate of alteration of these genes in tumor samples (62.1, 53.3, and 45 % of cases, respectively), suggesting that our model recapitulates some traits of real tumor samples. This indicates that the BJ model can be used to determine key gene regulatory principles of the transformation process which can also be observed in "real" human tumors. Moreover, the availability of CRISPR technologies facilitates the engineering of such isogenic systems from primary human cells to model the process of tumorigenesis and assess the contribution of (combinations of) aberrations by introduction of genetic elements which are found deregulated or mutated in human tumors.

Deregulation of CRMs and the epigenome landscape in tumorigenesis: mutual inter-relationship

Increasing evidence suggests that many epigenetic regulatory proteins are deregulated in cancer and that histone mark patterns are globally changed within the cancer epigenome [32, 52]. Our observations support this as a number of CRMs are differentially expressed during cell transformation, including the classes of "writers" and "erasers". Most of them have been reported to play a role in tumorigenesis and their expression patterns in transformed BJELM cells are similar to those in several types of cancer (Additional file 4: Figure S4; Additional file 5: Table S1), indicating that the BJ stepwise transformation system is capable of recapitulating the deregulation of molecular pathways seen in "real" cancer and possibly can identify new regulators of tumorigenesis. In this respect, we point out several CRMs that have not been previously associated with tumorigenic cell transformation.

Deregulation of CRMs in the BJ model, which does not suffer from genome instability, reveals the epigenetic consequences of hTERT, SV-40 T and t antigen, and MYC introduction and, thus, the mutual interconnection between transcription factor deregulation and epigenome alteration on the pathway towards tumorigenesis. This would not be possible by comparing non-isogenic normal and cancer cell lines, as genome instability of cancer cells leads to merging of effects due to the introduction of exogenous elements and those coming from genome aberrations already existing in the tumor cell line.

GRNs of tumorigenesis

Cellular phenotypes are determined by the temporal regulation and dynamics of networks of co-regulated genes. Thus, elucidating GRNs is crucial for understanding of normal and cancer cell functioning. Until today only a few studies have addressed this issue—for example, the elucidation of the P53 regulatory network [53] or the analysis of locus expression signatures from retroviral insertion-induced tumorigenesis [54]. To perform a systematic analysis of GRNs underlying tumorigenesis, we used a novel combinatorial approach by (i) integrating transcriptome data during transformation steps with chromatin state dynamics, (ii) complementing this with an analysis of CRMs involved in this process, and (iii) inferring key transformation-related TFs by using a database of established TF–TG associations from multiple human lineages. The reconstructed GRNs reveal a crosstalk between the elements perturbing the normal system through TFs and CRMs as "transformation mediators" to the executor nodes, thus giving a comprehensive view of the molecular chain of events. The present approach could be applied to dissecting other processes, like cell differentiation or cell fate reprogramming, and the decryption of cause-and-consequence mechanistic links.

Conclusions

In the current study we reconstructed GRNs that are altered during the process of stepwise human cellular tumorigenesis, providing a rich source of (novel) regulators

of tumorigenesis. Using the reconstructed network, we predict and validate several transcription factors as being key players in the establishment of tumorigenic traits of transformed cells. Our data reveal the importance of CRMs in oncogene-induced tumorigenesis and identify new CRMs involved in this process.

Availability of data and materials

SNP arrays, Affymetrix microarrays and Illumina platform ChIP-seq data sets supporting the results of this article are available in the Gene Expression Omnibus repository under the accession number GSE72533 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72533).

Additional files

Additional file 1: Figure S1. SNP analysis. **a** Statistics of changes in single-nucleotide polymorphisms (*SNP*), loss of heterozygosity (*LOH*), and copy number (*CN*) in the BJ stepwise transformation system in comparison with a previously used MCF10A-derived breast cancer tumorigenesis system (*MCF10A*, non-tumorigenic breast cell line; *MCF10AT*, premalignant breast cell line generated by HRAS transformation of MCF10A; *MCF10CA1a*, poorly-differentiated malignant breast cell line derived from a MCF10AT xenograft). Note the nearly isogenic character of the BJ stepwise transformation system and the high genetic divergence in the MCF10A system. **b** Chromosome diagrams illustrating changes in copy number in immortalized BJEL and transformed BJELM cells relative to normal BJ cells. *Red* and *blue triangles* correspond to loss and gain in copy number, respectively. (PDF 2.96 mb)

Additional file 2: Figure S2. Description and validation of the BJ stepwise tumorigenesis system. **a** The expression pattern of co-regulated genes based on comparative transcriptomics of primary, immortalized, and transformed cells. **b** Statistics of changes in differentially expressed genes during cell transformation. **c** RT-qPCR validation of exogenous (*cMYC*, *TERT*) and transformation-relevant (*CCND2*, *THBS1*) gene expression. **d** Western blot analysis of whole cell extracts of BJ, BJEL, and BJELM cells confirms overexpression of T antigen and MYC-ER during the immortalization and tumorigenic steps, respectively. **e** Validation of the determination of TRAIL sensitivity during the transformation process. TRAIL-induced apoptosis was observed specifically in BJELM cells, while BJ and BJEL cells showed resistance to TRAIL treatment. **f** Reproducibility between replicates evaluated by calculation of the Pearson correlation coefficient. **g** Reproducibility between replicates evaluated by calculation of the skewness parameter between BJEL and BJELM replicates relative to BJ replicates. (PDF 2.95 mb)

Additional file 3: Figure S3. Gene Ontology (GO) analysis of differentially expressed genes in the stepwise tumorigenesis system. **a** Functional clustering by biological pathway using DAVID for each set of co-regulated genes (pathways i to vii). **b** Disease-related GO (DAVID) of differentially expressed genes. The *x-axis* (*p* value) is given as $-\log(p\ \text{value})$. Illustrated GO terms have p value <0.05. (PDF 2.96 mb)

Additional file 4: Figure S4. Gene expression levels of chromatin remodelers/modulators (CRMs) differentially expressed in the stepwise tumorigenesis system. CRMs are specified on the *left* together with the corresponding co-expression pathways. The heatmap shows the ratio of expression in BJEL or BJELM cells relative to the expression in BJ cells. The corresponding color code is shown on the *right*. (PDF 2.95 mb)

Additional file 5: Table S1. Chromatin remodelers/modulators differentially expressed during cell transformation. The table describes their reported function and involvement in cancer and provides the corresponding references (listed in Additional file 6). Abbreviations: *BRD* bromodomain, *CHD* chromodomain, *HAT* histoneacetyltransferases, *HMT* histonemethyltransferase, *TDRD* Tudor domain. (PDF 563 kb)

Additional file 6. Supplementary data references. (DOCX 21 kb)

Additional file 7: Table S2. Transcription factors preferentially associated with specific co-expression pathways and which originate from deregulated gene programming during tumorigenesis. Some of these TFs are differentially expressed as well and the co-expression pathway they belong to is shown in the last column. (PDF 102 kb)

Additional file 8: Figure S5. Classification heatmap model showing the loss of fibroblast identity by BJ fibroblasts during the transformation, while gaining traits of embryonic stem cells. The analysis was performed using the CellNet tool. The *color key* shows the similarity between the training system and study samples. *Yellow* and *black* indicate high and low levels of resemblance, respectively. *b.r.* biological replicate. (PDF 2.96 mb)

Additional file 9: Figure S6. A two-step normalization procedure required for proper multiprofile comparison. **a** To account for technical aspects like antibody efficiency and sequencing depth, we used Epimetheus, a two-step normalization procedure in which (i) the raw count intensity in ChIP-seq datasets produced with antibodies targeting the same factor are corrected following a quantile normalization procedure; then (ii) normalized ChIP-seq profile read counts corresponding to a variety of factors are brought to the same scale via a z-score normalization correction. **b** The effect of the quantile normalization on H3K9ac datasets assessed in all three cell lines of the stepwise transformation model. Notice that BJELM cells display lower intensity levels of the H3K9ac mark in the *LBR* promoter (*blue arrow*) relative to BJ and BJEL cells, while *LBR* gene expression is upregulated in BJEL and BJELM cells; after quantile correction, the levels in the BJELM dataset appears the same as in the BJEL dataset, both higher than in the BJ dataset. Furthermore, notice the higher background (region under the *red brace*) in the raw profiles of the BJEL dataset (in comparison with BJ and BJELM), which is brought to the same level in all three datasets after normalization. (PDF 520 kb)

Additional file 10. Transcriptome data summary provided in Excel format. (XLS 1293 kb)

Additional file 11: Figure S7. Chromatin state annotation. **a** Statistical analysis of chromatin states (initial and combined) at the TSSs of DEGs. **b** Chromatin state classification. **c** Normalized ChIP signal intensities at the TSSs ±500 bp, ordered from first to 26th chromatin state as in **a** in BJ cells. (PDF 2.95 mb)

Additional file 12. MYC target genes as determined by the analysis of publically available ChIP-seq profiles and defined by the association of the MACS peaks (p value threshold = −30) with the TSS of genes using a 10-kb distance as an association criterion. (XLS 199 kb)

Additional file 13: Figure S8. Cross-cancer alteration summary for CHD7, DHX33, NOLC1, and MYC among 123 cancer types; 99 cancer types that have alterations in these genes are displayed in the histogram. In 49 types of cancer, alterations in these genes occur in more than 10 % of cases. In particular, breast cancer, neuroendocrine prostate cancer, and ovarian serous cystadenocarcinoma have the highest rate of amplification of these genes in tumor samples (55.2, 50.5, and 44.1 % of cases, respectively). (PDF 2990 kb)

Abbreviations

ChIP: chromatin immunoprecipitation; CRM: chromatin remodeler/modulator; DMEM: Dulbecco's modified Eagle's medium; FCS: fetal calf serum; GO: Gene Ontology; GRN: gene regulatory network; RNA Pol II: RNA polymerase II; SNP: single-nucleotide polymorphism; TRAIL: tumor necrosis factor-related apoptosis-inducing ligand; TF: transcription factor; TF–TG: transcription factor–target gene; TSS: transcription start site.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HG designed the study. VM performed the experimental work. VM, MAMP, and MAMS performed data analysis. VM and HG wrote the manuscript and revised it together with MAMP. All authors read and approved the final manuscript.

Acknowledgements
We are very grateful to Dr William C. Hahn and Dr Robert A. Weinberg for providing plasmids and cells. We thank former members of our laboratory, Dr Pattabhiraman Shankaranarayanan and Dr Jelena Vjetrovic, for providing BJELM transformed cells. We thank Dr Valeria Pavet-Portal for help in setting up the stepwise cell system, bioinformaticians Matthias Blum and Pierre-Etienne Cholley for expert advice, Benjamin Billore and Michele Lieb for technical support, and all other members of our laboratory for discussions. We thank as well the IGBMC Microarray and Sequencing platform (France Génomique consortium—ANR-10-INBS-0009).

Funding
This work was supported by funds from the Alliance Nationale pour les Sciences de la Vie et de la Santé (Aviesan)—Institut Thématique Multi-organismes Cancer (ITMO Cancer)—Institut National du Cancer (INCa) and the Ligue National Contre le Cancer (Equipe Labellisée).

Received: 23 December 2015 Accepted: 19 April 2016
Published online: 19 May 2016

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
2. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719–24.
3. Hahn WC, Hahn WC, Counter CM, Counter CM, Lundberg AS, Lundberg AS, et al. Creation of human tumour cells with defined genetic elements. Nature. 1999;400(6743):464–8.
4. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125(2):315–26.
5. Vjetrovic J, Shankaranarayanan P, Mendoza-Parra MA, Gronemeyer H. Senescence-secreted factors activate Myc and sensitize pretransformed cells to TRAIL-induced apoptosis. Aging Cell. 2014;13(3):487–96.
6. Draghici S. Data analysis tools for DNA Microarrays. Chapman & Hall/CRC Mathematical and Computational Biology. 2003.
7. Mendoza-Parra MA, Van Gool W, Saleem MAM, Ceschin DG, Gronemeyer H. A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Res. 2013;41(21), e196.
8. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009;25(15):1952–8.
9. Mendoza-Parra M-A, Nowicka M, Van Gool W, Gronemeyer H. Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. BMC Genomics. 2013;14(1):834.
10. Gao J, Ade AS, Tarcea VG, Weymouth TE, Mirel BR, Jagadish HV, et al. Integrating and annotating the interactome using the MiMI plugin for cytoscape. Bioinformatics. 2009;25(1):137–8.
11. Cahan P, Li H, Morris SA, Lummertz da Rocha E, Daley GQ, Collins JJ. Cell Net: network biology applied to stem cell engineering. Cell. 2014;158(4):903–15.
12. Kim K-P, Schöler HR. Cell Net–where your cells are standing. Cell. 2014;158(4):699–701.
13. Lin CY, Chin CH, Wu HH, Chen SH, Ho CW, Ko MT. Hubba: hub objects analyzer–a framework of interactome hubs identification for network biology. Nucleic Acids Res. 2008;36(Web Server issue):438–43.
14. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007;3(4):713–20.
15. Barsky A, Gardy JL, Hancock REW, Munzner T. Cerebral: A Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. Bioinformatics. 2007;23(8):1040–2.
16. Pavet V, Shlyakhtina Y, He T, Ceschin DG, Kohonen P, Perälä M, et al. Plasminogen activator urokinase expression reveals TRAIL responsiveness and supports fractional survival of cancer cells. Cell Death Dis. 2014;5, e1043.
17. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
18. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
19. Walz S, Lorenzin F, Morton J, Wiese KE, von Eyss B, Herold S, et al. Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. Nature. 2014;511(7510):483–7.
20. Morris SA, Cahan P, Li H, Zhao AM, San Roman AK, Shivdasani RA, et al. Dissecting engineered cell types and enhancing cell fate conversion via Cell Net. Cell. 2014;158(4):889–902.
21. Messina NL, Banks KM, Vidacs E, Martin BP, Long F, Christiansen AJ, et al. Modulation of antitumour immune responses by intratumoural Stat1 expression. Immunol Cell Biol. 2013;91(9):556–67.
22. Huang S, Bucana CD, Van Arsdall M, Fidler IJ. Stat1 negatively regulates angiogenesis, tumorigenicity and metastasis of tumor cells. Oncogene. 2002;21(16):2504–12.
23. Koromilas AE, Sexl V. The tumor suppressor function of STAT1 in breast cancer. Jak-Stat. 2013;2(2), e23353.
24. Schlee M, Hölzel M, Bernard S, Mailhammer R, Schuhmacher M, Reschke J, et al. c-MYC activation impairs the NF-κB and the interferon response: Implications for the pathogenesis of Burkitt's lymphoma. Int J Cancer. 2007;120(7):1387–95.
25. Stolfi C, De Simone V, Colantoni A, Franzè E, Ribichini E, Fantini MC, et al. A functional role for Smad7 in sustaining colon cancer cell growth and survival. Cell Death Dis. 2014;5, e1073.
26. Yu J, Ma X, Cheung KF, Li X, Tian L, Wang S, et al. Epigenetic inactivation of T-box transcription factor 5, a novel tumor suppressor gene, is associated with colon cancer. Oncogene. 2010;29(49):6464–74.
27. Lee K-W, Yeo S-Y, Sung CO, Kim S-H. Twist1 is a key regulator of cancer-associated fibroblasts. Cancer Res. 2015;75(1):73–85.
28. Milyavsky M, Shats I, Cholostoy A, Brosh R, Buganim Y, Weisz L, et al. Inactivation of myocardin and p16 during malignant transformation contributes to a differentiation defect. Cancer Cell. 2007;11(2):133–46.
29. Benavente CA, Finkelstein D, Johnson DA, Ashery-padan R, Dyer MA. Chromatin remodelers HELLS and UHRF1 mediate the epigenetic deregulation of genes that drive retinoblastoma tumor progression. Oncotarget. 2014;5(20):9594–608.
30. Jung KH, Noh JH, Kim JK, Eun JW, Bae HJ, Xie HJ, et al. HDAC2 overexpression confers oncogenic potential to human lung cancer cells by deregulating expression of apoptosis and cell cycle proteins. J Cell Biochem. 2012;113(6):2167–77.
31. Feng Y, Lee N, Fearon ER. TIP49 regulates β-catenin-mediated neoplastic transformation and T-cell factor target gene induction via effects on chromatin remodeling. Cancer Res. 2003;63:8726–34.
32. Chaligné R, Popova T, Saleem MM, Gentien D, Ban K, Piolot T, et al. The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer. Genome Res. 2015;25:488–503.
33. Chi P, Allis CD, Wang GG. Covalent histone modifications: miswritten, misinterpreted, and miserased in human cancers. Nat Rev Cancer. 2010;10(7):457–69.
34. Martinato F, Cesaroni M, Amati B, Guccione E. Analysis of Myc-induced histone modifications on target chromatin. PLoS One. 2008;3(11), e3650.
35. Frank SR, Schroeder M, Fernandez P, Taubert S, Amati B. Binding of c-Myc to chromatin mediates mitogen-induced acetylation of histone H4 and gene activation. Genes Dev. 2001;15(16):2069–82.
36. Gilder AS, Do PM, Carrero ZI, Cosman AM, Broome HJ, Velma V, et al. Coilin participates in the suppression of RNA polymerase I in response to cisplatin-induced DNA damage. Mol Biol Cell. 2011;22(7):1070–9.
37. Zhang Y, Saporita AJ, Weber JD. P19ARF and $RasV^{12}$ offer opposing regulation of DHX33 translation to dictate tumor cell fate. Mol Cell Biol. 2013;33(8):1594–607.
38. Zentner GE, Hurd EA, Schnetz MP, Handoko L, Wang C, Wang Z, et al. CHD7 functions in the nucleolus as a positive regulator of ribosomal RNA biogenesis. Hum Mol Genet. 2010;19(18):3491–501.
39. Williamson D, Lu Y-J, Fang C, Pritchard-Jones K, Janet S. Nascent pre-rRNA overexpression correlates with an adverse prognosis in alveolar rhabdomyosarcoma. Genes Chromosomes Cancer. 2006;45:839–45.
40. Grummt I. Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. Genes Dev. 2003;17(14):1691–702.
41. Zhang Y, Forys JT, Miceli AP, Gwinn AS, Weber D. Identification of DHX33 as a mediator of rRNA synthesis and cell growth. Mol Cell Biol. 2011;31(23):4676–91.

42. Yamamoto M, Cid E, Bru S, Yamamoto F. Rare and frequent promoter methylation, respectively, of TSHZ2 and 3 genes that are both downregulated in expression in breast and prostate cancers. PLoS One. 2011;6(3):1–10.
43. Vladimirova V, Mikeska T, Waha A, Soerensen N, Xu J, Reynolds PC, et al. Aberrant methylation and reduced expression of LHX9 in malignant gliomas of childhood. Neoplasia. 2009;11(7):700–11.
44. Jiang Y, Rom WN, Yie T, Chi CX. Induction of tumor suppression and glandular differentiation of A549 lung carcinoma cells by dominant-negative IGF-I receptor. Oncogene. 1999;18:6071–7.
45. Schmitt CA. Senescence, apoptosis and therapy–cutting the lifelines of cancer. Nat Rev Cancer. 2003;3(4):286–95.
46. Shay JW, Wright WE. Telomerase: a target for cancer therapeutics. Cancer Cell. 2002;2(4):257–65.
47. Ali SH, DeCaprio JA. Cellular transformation by SV40 large T antigen: interaction with host proteins. Semin Cancer Biol. 2001;11(1):15–23.
48. Nesbit CE, Tersak JM, Prochownik EV. MYC oncogenes and human neoplastic disease. Oncogene. 1999;18(19):3004–16.
49. Grandori C, Gomez-Roman N, Felton-Edkins ZA, Ngouenet C, Galloway DA, Eisenman RN, et al. c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. Nat Cell Biol. 2005;7(3):311–8.
50. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
51. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):l1.
52. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2014;128(4):683–92.
53. Song W, Wang J, Yang Y, Jing N, Zhang X, Chen L, et al. Rewiring drug-activated p53-regulatory network from suppressing to promoting tumorigenesis. J Mol Cell Biol. 2012;4(4):197–206.
54. Lee E, de Ridder J, Kool J, Wessels LFA, Bussemaker HJ. Identifying regulatory mechanisms underlying tumorigenesis using locus expression signature analysis. Proc Natl Acad Sci U S A. 2014;111(15):5747–52.

PUBLICATION N°2

Reconstructed cell fate-regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis

Marco-Antonio Mendoza-Parra, Valeriya Malysheva, Mohamed Ashick Mohamed Saleem, Michele Lieb, Aurelie Godel, and Hinrich Gronemeyer

Genome Research 2016 Sep 20. pii: gr.208926.116 [Epub ahead of print]

Reconstructed cell fate–regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis

Marco-Antonio Mendoza-Parra, Valeriya Malysheva, Mohamed Ashick Mohamed Saleem, et al.

Genome Res. published online September 20, 2016
Access the most recent version at doi:10.1101/gr.208926.116

Reconstructed cell fate–regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis

Marco-Antonio Mendoza-Parra, Valeriya Malysheva, Mohamed Ashick Mohamed Saleem, Michele Lieb, Aurelie Godel, and Hinrich Gronemeyer

Equipe Labellisée Ligue Contre le Cancer, Department of Functional Genomics and Cancer, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Centre National de la Recherche Scientifique, UMR7104, Institut National de la Santé et de la Recherche Médicale, U964, Université de Strasbourg, Illkirch, France

Cell lineages, which shape the 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. Significant knowledge about these events and the involved key drivers comes from homogeneous cell differentiation models. Even a single chemical trigger, such as the morphogen all–*trans* retinoic acid (RA), can induce the complex network of gene-regulatory decisions that matures a stem/precursor cell to a particular step within a given lineage. Here we have dissected the GRNs involved in the RA-induced neuronal or endodermal cell fate specification by integrating dynamic RXRA binding, chromatin accessibility, epigenetic promoter epigenetic status, and the transcriptional activity inferred from RNA polymerase II mapping and transcription profiling. Our data reveal how RA induces a network of transcription factors (TFs), which direct the temporal organization of cognate GRNs, thereby driving neuronal/endodermal cell fate specification. Modeling signal transduction propagation using the reconstructed GRNs indicated critical TFs for neuronal cell fate specification, which were confirmed by CRISPR/Cas9-mediated genome editing. Overall, this study demonstrates that a systems view of cell fate specification combined with computational signal transduction models provides the necessary insight in cellular plasticity for cell fate engineering. The present integrated approach can be used to monitor the in vitro capacity of (engineered) cells/tissues to establish cell lineages for regenerative medicine.

[Supplemental material is available for this article.]

The life of cells in multicellular organisms is directed by dynamic gene programs, which guide and define lineage progression from pluripotent to differentiated states through series of temporal decisions. Knowledge of these programs and decisions reveals not only how cells acquire physiological functionalities, it also provides key information for therapy, as deviations from this blueprint can lead to disease. Moreover, the possibility to interfere with cell programming by treating stem cells or reprogramming somatic cells may generate specific autologous cell types for regenerative medicine in a personal medicine context.

Cell lineages derive from series of subsequent programming decisions. Cell differentiation models, particularly those where the series of transitions within a lineage is initiated by a single chemical trigger like all-*trans* retinoic acid (RA), significantly facilitated the study of cell fate acquisition. The use of RA (rather than complex culture conditions) as a defined trigger of regulatory events is essential to elucidate the dynamically regulated "downstream" gene networks. In this context, our study of F9 embryo carcinoma (EC) cells provided a first detailed view of RA-induced gene program diversification through a plethora of regulatory decisions (Mendoza-Parra et al. 2011).

EC cells can differentiate into all three primary germ layers (Soprano et al. 2007). While F9 cells differentiate into primitive endoderm when treated with RA in monolayer, parietal or visceral endodermal differentiation is observed when RA is either complemented with cyclic AMP or when cells are cultured as embryoid bodies in suspension. P19 EC cells differentiate into either skeletal muscle or neuronal cell types upon treatment with dimethlysulfoxide or RA, respectively. Thus, RA can induce cell fate commitment toward two distinct primary germ layers. However, the temporal evolution of the corresponding gene programs and the regulatory mechanisms remained elusive.

RA signaling is initiated by its binding to retinoid receptor heterodimers (RAR/RXR), members of the nuclear receptor (NR) family of ligand-regulated TFs (Laudet and Gronemeyer 2002). Upon ligand binding, RAR/RXR recruits coactivator complexes leading to the transcriptional activation of target genes (TGs) (Gronemeyer et al. 2004; Rosenfeld et al. 2006). The complexity of the RA signaling is largely increased by the expression of three RXR and three RAR isotypes (alpha, beta, and gamma), as each RAR/RXR combination could regulate cognate gene programs (Chiba et al. 1997). Interestingly, particular isotype-selective RAR ligands (Alvarez et al. 2014) induced specific cell fate transitions: F9 cells show similar morphological cell differentiation phenotypes when treated with RA or the RARG-selective ligand BMS961, but not with the RARA-selective ligand BMS753. In contrast, in P19 cells BMS753 and RA induce the same morphological differentiation, while BMS961 has no such effect (Taneja et al. 1996). These observations strongly support a critical role of RAR

Corresponding authors: marco@igbmc.fr, hg@igbmc.u-strasbg.fr
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.208926.116.

isotypes in the establishment of different cell fate commitment processes.

Given that RARA/G isotypes are expressed similarly in both EC cells (Supplemental Fig. S1), we reconstructed the dynamics of GRNs that are at the basis of the cell fate decisions in F9 and P19 cells by characterizing common and cell-specific RA-induced gene programs (Supplemental Fig. S2). We subsequently developed a computational signal transduction model that was used to (1) verify the temporal transcriptional coherence of the reconstructed GRN, and (2) predict potential downstream TFs that drive neuronal cell fate commitment. Using CRISPR/dCas9 (D10/N863A) technology, we activated the transcription of several predicted factors and assessed their capacity to induce the acquisition of neuronal identity. Overall, this study provides a detailed view of the complex regulatory wirings that are commonly initiated in both EC model systems but lead to distinct cells fates and which can be engineered for redirecting cell fate decisions.

Results

RA induces both common and cell fate-specific programs in F9 and P19 cells

As RA induces a neuronal cell fate of P19 cells, while driving endodermal differentiation of F9 cells, we first defined common and cell-specific RA-induced programs in these models. We used previously established monolayer cultures (Monzo et al. 2012) for efficient morphological P19 cell differentiation by RA and showed that this process is driven by RARA by using RAR isoform-specific agonists (Supplemental Fig. S3A). Neuronal cell fate commitment was confirmed by the induction of neurogenin 1 (*Neurog1*) and *Neurod1* (Supplemental Fig. S3B). Analysis of the global transcriptome changes during P19 cell differentiation revealed a previously reported progressive increase of differentially expressed genes (DEGs) (Wei et al. 2002). Indeed, after 2 h of RA treatment, only 51 genes showed an induction of ≥1.8-fold, while >1000 were induced after 72 h (Supplemental Fig. S4).

A comparison of the temporal transcriptome changes during endodermal F9 (Mendoza-Parra et al. 2011) and neuronal P19 cell differentiation revealed that >60% of genes are commonly regulated in both cell lines, albeit with different kinetics in some cases (Fig. 1). F9 cells present a higher number of DEGs in the first hours of RA treatment (Fig. 1A), but most of these early responders are also observed in P19 cells at later time points. In keeping with the progressive expression of the differentiated phenotype, divergent cell type-selective gene expression increased toward later time points, such that at 72 h, only <30% of the genes differentially expressed in P19 were similarly regulated in F9 cells (Fig. 1B). Gene Ontology (GO) analysis classified the commonly RA-regulated genes as involved in retinoid binding or cell fate commitment. Among them were classical RA-induced genes (e.g., *Rarb*, *Cyp26a1*, or *Hoxa1*), while pluripotency factors were down-regulated (Fig. 1C). As

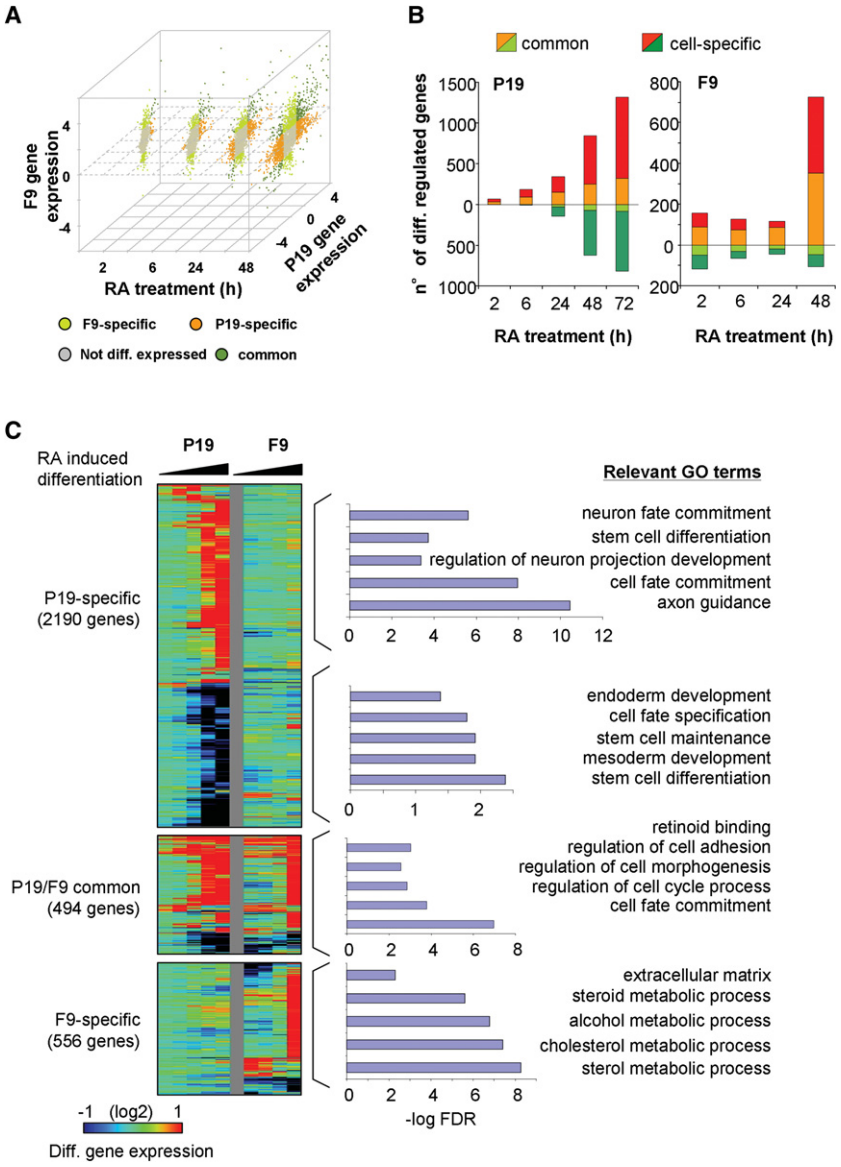


Figure 1. Common and specific RA-induced differentiation programs characterized in F9 and P19 embryonal carcinoma cells. (*A*) Scatterplot illustrating transcriptome changes in F9 and P19 EC cells at different time points during RA-induced differentiation. Gene expression levels relative to the undifferentiated state were classified as common, EC-specific, or not differentially expressed, based on a defined fold change threshold (up-regulated genes, fold change > 1.8; down-regulated genes, fold change < 0.5) at a given time point. (*B*) Differential gene expression levels in both model systems were used for computing the number of differentially regulated genes (*y*-axis) at various time points covering the first 72 h of RA treatment (*x*-axis). DEGs were classified as either commonly or cell-specifically expressed. This classification takes into consideration the gene expression response over all evaluated time points, in contrast to *A*, where a classification per time point is performed. (*C*) Temporal changes in transcriptional expression in either F9 or P19 EC cells are displayed for common and cell type-specific genes. Relevant GO terms per common or cell type-specific group of genes are displayed.

expected, P19-specific RA-induced genes are enriched for GO terms like neuronal fate commitment, while down-regulated genes are enriched for terms like endoderm or mesoderm development and stem cell maintenance, which are repressed during neuronal cell fate acquisition.

Chromatin state dynamics during neuronal and endodermal differentiation correlate with gene coexpression patterns

While the above transcriptome profiling revealed the RA-induced changes, an understanding of the corresponding regulatory mechanisms requires additional analyses of the RA-modulated key players and the information on epigenome and chromatin structure changes. To this end, we mapped RXRA binding sites to identify cognate TGs and complemented this readout with the characterization of epigenetic marks indicative for active and repressed transcription, open chromatin regions, and RNAPII binding at regulated genes. Our combinatorial analysis of the generated data sets demonstrated the existence of genomic regions preferentially enriched for repressive marks (H3K27me3), bivalent/poised (H3K27me3 and H3K4me3), or active promoter regions (H3K4me3 and/or RNAPII), but also for candidate enhancer regions where open chromatin sites co-occurred with RXRA binding (Fig. 2A).

An example of the temporal connection between these various regulatory events is the *HoxA* cluster, where the progressive loss of the repressive H3K27me3 mark during RA-induced differentiation both in P19 and F9 cells correlates with a gain in FAIRE, RXRA, and RNAPII enrichment patterns (Fig. 2B). These progressive changes of chromatin accessibility/TF association and gain of marks for active transcription with concomitant loss of "repressive" marks correlated with the collinear mechanism for transcription activation of *Hox* genes, previously described in other systems (Kashyap et al. 2011; Montavon and Duboule 2013).

To evaluate the coherence between epigenetic status and transcriptional activity, temporal transcriptomes were analyzed in the context of gene coexpression paths with the Dynamic Regulatory Events Miner (DREM) (Ernst et al. 2007). This analysis gave rise to a total of six coexpression paths (Mendoza-Parra et al. 2011) for the endodermal differentiation and 10 coexpression paths for the neuronal cell fate acquisition (Fig. 2C).

Assuming that genes with similar temporal expression patterns share common temporal alterations of epigenome and RNAPII recruitment patterns, we assessed the enrichment of H3K27me3, H3K4me3, and/or RNAPII at the promoter regions of genes differentially expressed in both model systems and displayed it in a coexpression path context. To accurately define temporal enrichment patterns, we first normalized the ChIP-seq profiles using a novel two-step normalization procedure (Supplemental Methods; Supplemental Fig. S5).

We observed in general a positive correlation between the temporal evolution of gene coexpression paths and normalized H3K4me3 and RNAPII enrichment patterns at promoter regions of concerned genes, while a negative correlation was seen with the repressive H3K27me3 mark (Fig. 2D). Given the presence of both common and endodermal (F9)/neurogenesis (P19)-specific gene programming in each path, we analyzed these programs separately (Supplemental Fig. S6). As expected, the evolution of the chromatin states of gene promoters from the common program was highly similar in F9 and P19, while the states of fate-specific programs showed significant temporal divergence. In coexpression paths with a similar epigenetic landscape in both cell lines (path1 in F9; path1, 2, and 4 in P19), RA induction led to a temporal increase in the ratios of "active" over "repressive" chromatin in a F9/P19-specific manner, coinciding with increased gene expression.

In contrast, genes of other paths showed already in the non-induced state distinct epigenetic and/or RNAPII association characteristics (paths2, 4, and 5 in F9; path3 and partially path1 in P19 cells). Paths composed of genes gradually repressed during differentiation in an endodermal (F9)/neurogenesis (P19)-specific manner frequently gained in "repressive" chromatin (path9/10 of P19 cells; path6 in F9 cells). Importantly, the temporal evolution of specific genes fully reflected the global promoter characteristics within these paths, as for the commonly regulated *Rarb* or *Pou5f1* and the P19-specific *Neurog1* or *Tal2* gene promoters (Supplemental Fig. S7).

Altogether, these data support the concept that RA-induced common and fate-specific temporal changes in gene programming closely correlate with changes in the ratios of "active" and "repressive" chromatin marks at the cognate promoter regions.

Dissection of common and divergent target gene programming in neuronal and endodermal lineage-committed cells by RAR isotype-specific ligands

To identify core GRNs for the cell fate transitions, we established P19 transcriptomes after treatment with RAR subtype-specific agonists. Gene coexpression paths were nearly identical for the RARA-specific agonist (BMS753) and RA (Fig. 3A), in keeping with the common induction of a neuronal fate (Supplemental Fig. S3). No such effect was seen with RARB or RARG-specific agonists (Fig. 3A; Supplemental Figs. S8, S9). In F9 cells, both RA and the RARG agonist (BMS961) induced endodermal differentiation, as revealed by corresponding gene expression changes (Supplemental Fig. S8; Mendoza-Parra et al. 2011). Despite the similar response kinetics of RA and BMS753, the RARA agonist did not regulate the same number of genes as RA, suggesting that only a fraction of the RA responsive genes in P19 cells is required for phenotypic differentiation. Apparently, the BMS753-regulon corresponds to a minimal regulatory network, but the regulatory input of RA is more complex and extends beyond known differentiation features.

To reveal the direct RAR-RXR heterodimer TGs, we compared the proximal binding of RXRA (<10 kb distance) and the co-occurrence of open chromatin regions with RA or RAR subtype-specific agonist-regulated genes. From 695 RA-induced genes with FAIRE and RXRA sites in proximity, 44% responded to BMS753 but <2% to BMS961 or BMS641 (Fig. 3B). A similar analysis for F9 cells showed that from 327 RA–up-regulated genes displaying FAIRE and RXRA sites in proximity, about half (166 genes) responded also to the RARG-specific agonist BMS961, while significantly less (~25% and <4%) responded to RARA or RARB agonists, corroborating our previous findings (Mendoza-Parra and Gronemeyer 2013). Together, our results provide a complete gene regulatory framework accounting for the observations (Taneja et al. 1996) that RARA triggers neuronal differentiation of P19, while RARG induces endodermal differentiation of F9 cells.

To link the appearance of FAIRE and RXRA sites to transcription activation, we classified genes according to their temporal induction during RA or BMS753 treatment and proximal FAIRE and RXRA co-occurrence (SOTA, self-organization tree algorithm) (Fig. 3C; Supplemental Fig. S10). This methodology classified the transcriptional activation of P19 genes in six temporal patterns. Importantly, each class of the RA-induced P19 RXRA target genes

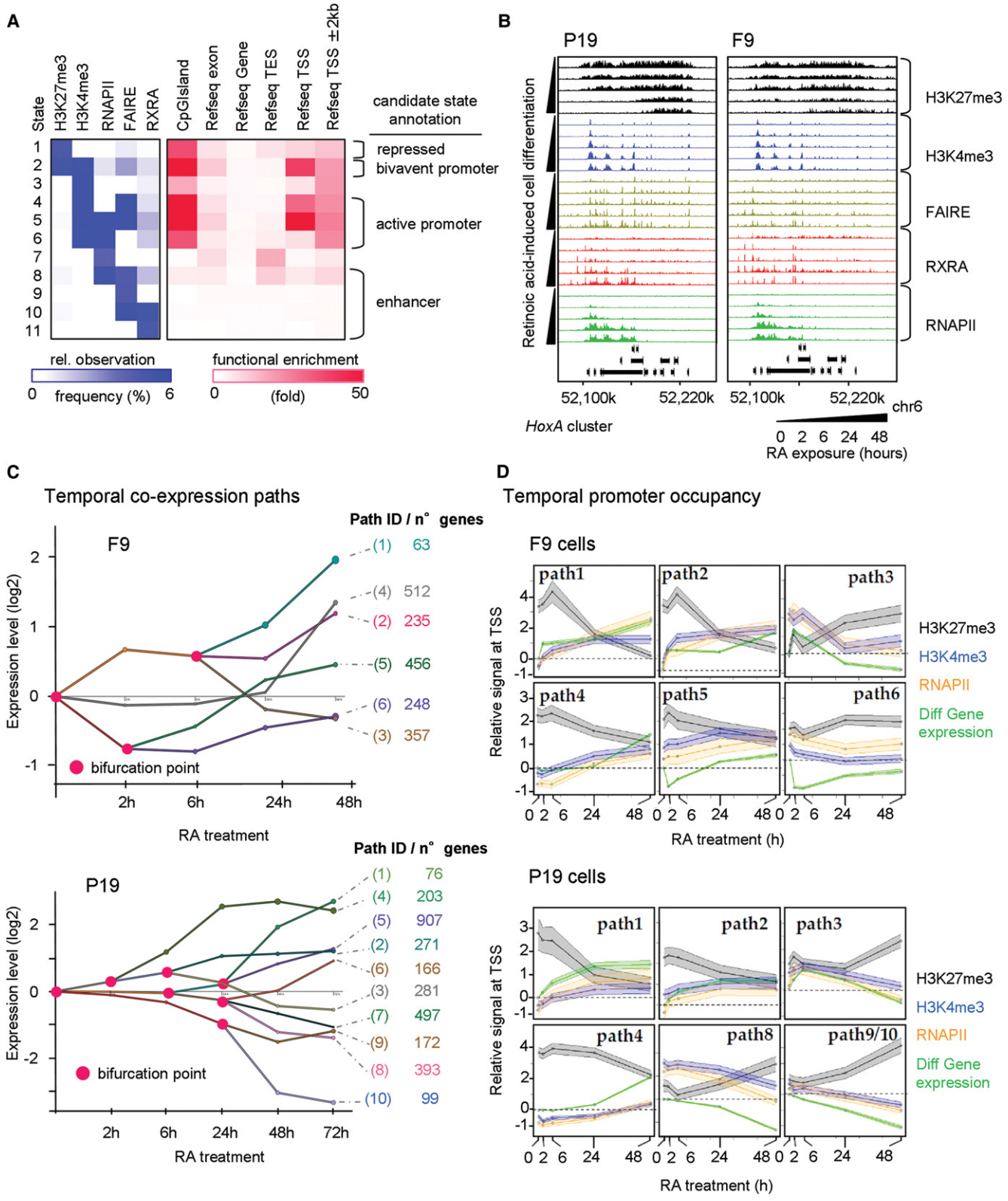


Figure 2. Multiparametric view of retinoid-induced cell fate transitions. (*A*) Chromatin state analysis performed over all profiled factors at all time points in P19 and F9 cells. Based on the predicted states resulting from a combination of all studied factors (*left* panel, relative observation frequency); four major candidate states were inferred: repressed, bivalent or active promoter, and enhancer-related states. This classification is supported by their functional enrichment levels associated with the described genomic annotations (*right* panel). (*B*) The *HoxA* cluster at Chromosome 6 displaying temporal changes in the enrichment of H3K27me3 and H3K4me3, the chromatin accessibility status (FAIRE-seq), the recruitment of the RXRA, and the transcriptional activity revealed by the profiling of the RNAPII. (*C*) Stratification of the temporal transcriptome profiling during RA-induced F9 (*upper* panel) or P19 cell differentiation (*lower* panel) in gene coexpression paths, accompanied by relevant bifurcation points (pink circles). Numbers of genes composing each of the coexpression paths are displayed (*right*). (*D*) Dynamics of promoter chromatin states during RA-induced F9 (*upper* panel) or P19 cell differentiation (*lower* panel). Gene promoters of the coexpression paths displayed in *C* are analyzed for temporal enrichment of (1) the repressive histone modification mark H3K27me3 (black), (2) the active histone modification mark H3K4me3 (blue), and (3) RNAPII (orange). Changes of mRNA levels relative to the noninduced condition are also displayed ("Diff Gene expression"; green). The *y*-axis corresponds to the average relative enrichment level derived from Epimetheus normalization (Supplemental Fig. S5). The shaded area corresponds to a 95% confidence interval.

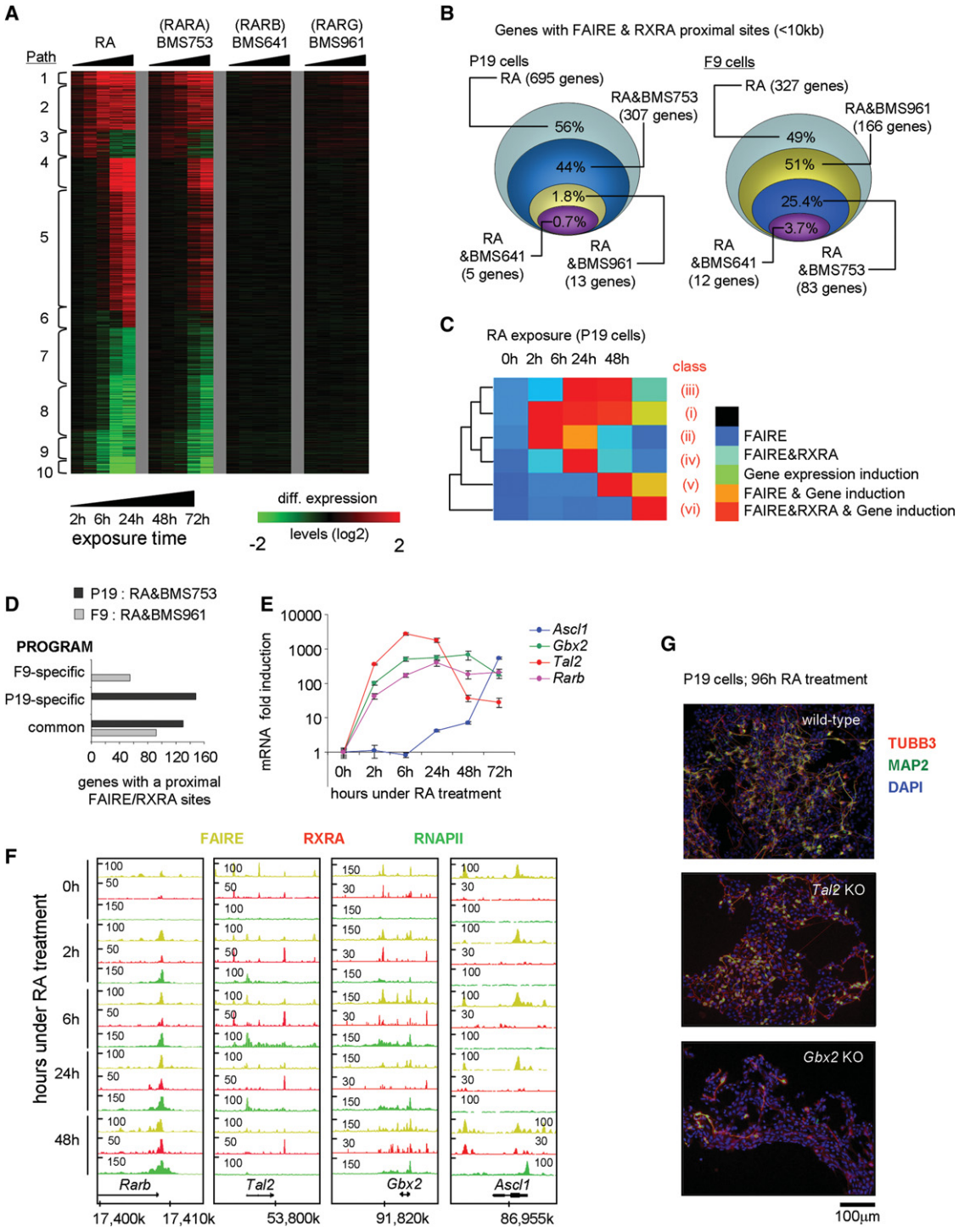


Figure 3. Different RAR subtypes induce chromatin alteration in RA-responsive genes of P19 and F9 cells. (*A*) Heat map illustrating the transcriptional responses of genes comprising the 10 coexpression paths characterized in P19 cells during RA-induced differentiation or in the presence of the indicated RAR isotype-specific agonists. (*B*) DEGs during RA-induced differentiation in P19 or F9 cells that present FAIRE and RXRA binding in proximity (<10 kb from the TSS) are compared with their corresponding transcriptional response in the presence of RAR isotype-specific agonists. (*C*) Heat map illustrating temporal SOTA classification of P19 genes positive for RXRA binding, and/or display altered chromatin structure (FAIRE-seq), and/or are induced in response to RA. This classification gave rise to the identification of six classes of genes with different temporal induction patterns (Supplemental Fig. S5). (*D*) Number of DEGs F9 or P19 cells commonly regulated by RA and BMS753 or RA and BMS961 and presenting a proximal FAIRE and RXRA binding site, stratified for the cell-specific (P19, F9) and common programs. (*E*) RT-qPCR revealing the temporal RA-induced mRNA expression profiles of bona fide RA target genes. (*F*) FAIRE-seq, RXRA, and RNAPII ChIP-seq profiles for the factors assessed in *E*. *Rarb, Gbx2*, and *Tal2* are early responding genes, while *Ascl1* gets significantly induced only after 24 h of RA induction. (*G*) Immunofluorescence micrograph of wild-type and CRISPR/Cas9-inactivated *Tal2* or *Gbx2* P19 cells after 96 h of RA treatment. Cells were stained for the neuronal markers TUBB3 (red) and MAP2 (green); nuclei were stained with DAPI (blue). *Gbx2*-inactivated cells present a lower frequency of double-stained TUBB3/MAP2 cells and shorter axon-like extension than *Tal2*-inactivated or wild-type cells.

contains a great number of genes that are equally induced in F9 cells, irrespective of the divergent cell fate acquisition (Fig. 3D; Supplemental Fig. S10). Among those are not only early induced prototypical TGs, like *Rarb* (Fig. 3E,F), *Foxa1* (Tan et al. 2010; Mendoza-Parra et al. 2011), and *Hoxa1*, but also late-induced direct TGs, such as *Pbx1, Pbx2, Cdh2, Sox6, and Sox11* (Supplemental Fig. S10). This shows that, despite significantly advanced divergent differentiation, RA still continues to induce an identical subset of TGs irrespective of endodermal or neuronal differentiation.

As expected, the P19-specific direct RXRA targets comprise factors involved in neurogenesis, mostly expressed at late time points during differentiation (*Ascl1* [Fig. 3E,F; Voronova et al. 2011; Huang et al. 2012, 2015]; *Gata3* [Martinez-Monedero et al. 2008]). Interestingly, however, the expression of some P19-specific TGs was already affected during the first hours of RA-treatment, among them, the TFs *Gbx2* (Bouillet et al. 1995; Inoue et al. 2012; Nakayama et al. 2013), and *Tal2*, which is essential for midbrain neurogenesis (Achim et al. 2013) and contains an intronic RA response element (Kobayashi et al. 2014, 2015). We identified two additional RXRA binding sites proximal to *Tal2*—a constitutive RXRA binding site ~3 kb downstream from the coding region and a second site upstream of the transcription start site (TSS) (~5 kb), which is similarly occupied in the absence of ligand but persists only until 6 h after initiating RA treatment (Fig. 3F).

To evaluate the importance of TAL2 and GBX2 for RA-induced neuronal commitment, we used CRISPR/Cas9-mediated gene inactivation (Supplemental Fig. S11A). *Tal2*-gene inactivation did not impair the expression of other neuronal-specific factors like ASCL1, NEUROD1, POU3F4, or NEUROG1 (Supplemental Fig. S11B). In contrary, *Gbx2*-inactivation reduced their expression severely, suggesting that GBX2 rather than TAL2 is a critical mediator of RA-induced neuronal commitment. This has been further supported by immunohistochemical analysis of the neuron-specific tubulin, beta 3 class III (TUBB3) and the microtubule-associated factor MAP2 (Fig. 3G; Supplemental Fig. S9), as RA induction of *Gbx2*-inactivated cells resulted in dramatically reduced numbers of TUBB3 and MAP2-stained cells, concomitantly with a major reduction of axonal extensions.

A network of TFs drives cell fate lineage decisions

The above integrative approach identified direct RXRA TGs, several of which are TFs. Conceptually, these genes could initiate TF-guided signal transduction cascades, ultimately generating the differentiated phenotype. To identify TFs relevant for the RA-induced neuronal fate of P19 cells, we established DREM-predicted coexpression paths (Fig. 2C). DREM evaluates the enrichment of coexpression paths for TGs associated with given TFs retrieved from TF-TG collections (Fig. 4A). Indeed, correlating RXRA binding/FAIRE site annotations with DREM-based gene coexpression analysis revealed the presence of RA target genes in the early path1-5, compliant with the inductive role of RXR-RAR heterodimers (Fig. 4B).

To identify additional relevant TFs, we reconstructed the RA-induced TF-TG networks involved in neuronal (P19) and endodermal (F9) differentiation by integrating the GRN interactions that constitute CellNet (Morris et al. 2014) into the DREM analysis (Fig. 4A). We identified multiple TFs associated with several coexpression paths but also path-specific TFs (Fig. 4C). Several of them were differentially expressed upon exposure to RA or RARA-specific agonists, supporting a direct implication in the predicted bifurcation (Fig. 4D). The negatively regulated coexpression path 10 associated with the self-renewal and pluripotency factors NANOG, POU5F1, ZFP42, SOX2, or SALL4 or with GBX2 and OTX2, TFs expressed very early during neuroectoderm development (Millet et al. 1999). Note that RA induction of GBX2 negatively regulates the expression of OTX2 in the anterior brain (Li and Joyner 2001; Inoue et al. 2012), corroborating their inverse expression patterns (Fig. 4D). Similarly, the early induced path1 is enriched for homeobox TF (HOXB1, HOXD1)-TGs but also for targets of ASCL1, OLIG2, and POU3F4, which are specifically expressed in neural tissues and, moreover, impose a neuronal fate on MEFs (Vierbuchen et al. 2010). Furthermore, the intermediate to late-induced path4 is enriched for MEIS2, PBX1, TBX2, or HOXA1, the latter being essential for neuronal commitment of mouse embryonic stem cells (Martinez-Ceballos and Gudas 2008). Integrating the CellNet TF-TG regulatory network information into the endodermal differentiation model (F9) revealed a set of TFs specifically involved in endodermal gene programming ("F9-specific") (Supplemental Fig. S12). However, we found a surprisingly large number of TFs that are commonly involved in both RA-induced endodermal and neuronal differentiation. A comparison of the GRNs inferred from these analyses is provided below.

Generation of comprehensive RA-driven signal transduction networks for neuronal and endodermal cell fates

To provide a comparative view of the signal transduction cascades driving the differential cell fates induced by RA in F9 and P19 cells, we integrated the CellNet TF-TG relationships (Morris et al. 2014), complemented by direct RA target and DREM analysis data, resulting in the reconstruction of a comprehensive GRN (2981 nodes, 44,931 edges) (Fig. 5A; Supplemental File S1). Two major nodes (blue squares) represent the initial RXRA/RAR signal interpreter in P19 or F9 cells. Each of them is associated with its direct targets of the common or fate-specific programs.

As CellNet was established using different cell types, it comprises also TF-TG interactions that are irrelevant for RA-dependent gene regulation. To exclude such interactions, we developed a computational approach that evaluates the coherence of the TF-TG relationships with the temporal evolution of transcription activation (Fig. 5B). Specifically, all interconnections from nodes not differentially expressed or originating from nodes not related to the initial cue were excluded, reducing the reconstructed GRN to 1931 nodes and 11,625 edges. The temporal evolution of common and fate-specific networks is evident from the superposition of RA-dependent gene expression patterns at the first four time points of the reconstructed GRN (Fig. 5C; Supplemental File S1) and from the increasing fraction of transduced nodes for each lineage-specific program (Fig. 5D).

The reconstructed network reveals also the RAR isotype-selective induction of endodermal or neuronal fates. Indeed, the RARA-specific agonist BMS753 fully recapitulates the neurogenic RA-response of P19 in both common and P19-specific gene regulatory programs, while only a minor fraction of this program is regulated in F9 (Fig. 5E,F; Supplemental File S1). Similarly, the RARG-specific BMS961 activates endodermal programming as RA in F9 but remains as ineffective in P19 as the RARB-agonist BMS641 in both cell fate programs. Further reduction by applying topological criteria generated a network (80 nodes, 626 edges) (Supplemental Fig. S17) with major nodes distributed in four subnetworks: two implicated in cell differentiation (pluripotency, HOX factors) and two neuronal/endodermal regulatory programs.

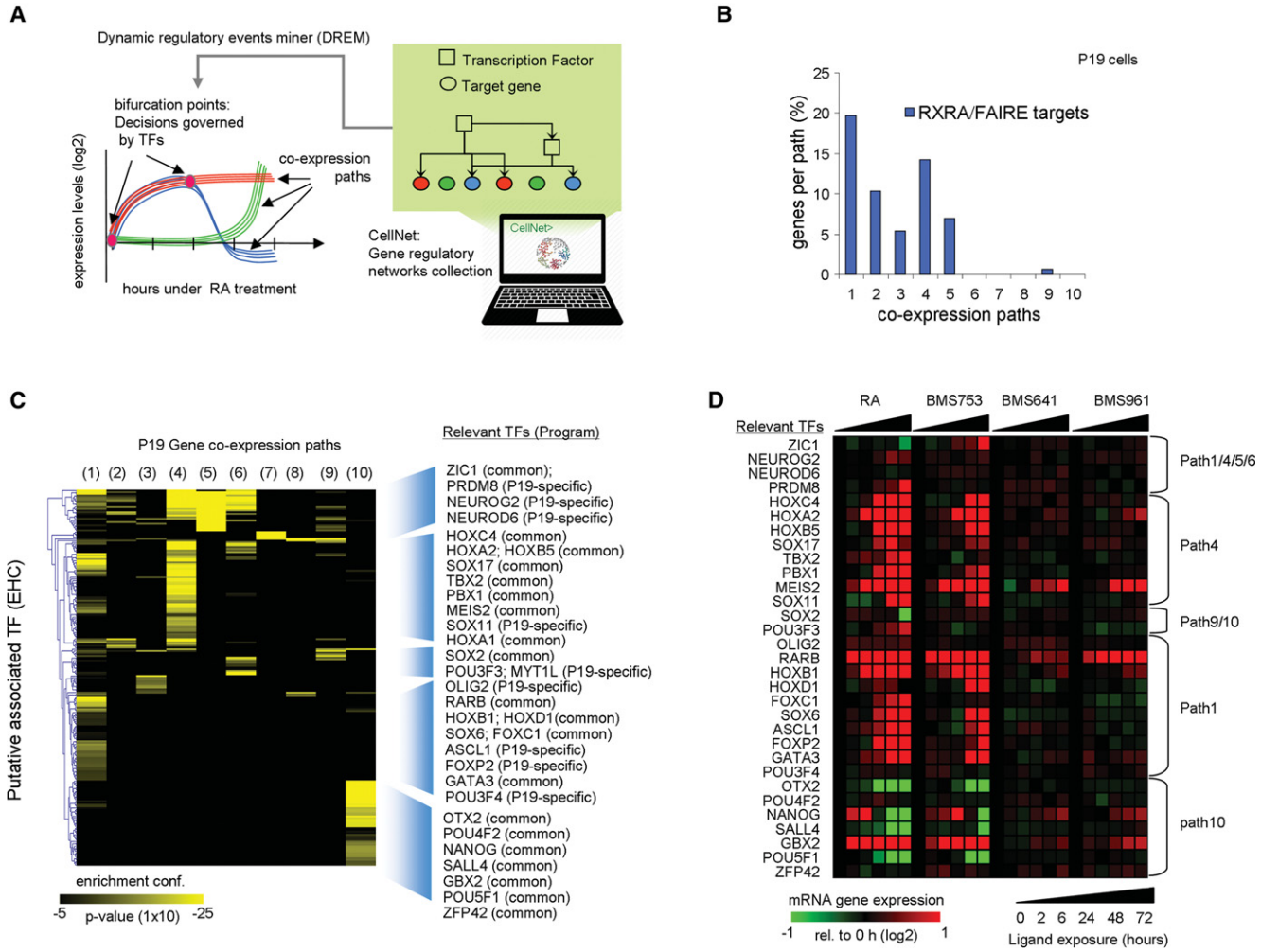


Figure 4. Inferring relevant TF-TG relationships during RA-induced neuronal cell fate transition. (*A*) Scheme of the strategy applied by DREM to integrate static TF-TG regulatory relationships with the temporal gene coexpression paths. In addition, the information provided by the CellNet collection and the RXRA binding/FAIRE site information have been integrated. (*B*) Bar graph depicting the fraction of RXRA and FAIRE TGs in each coexpression path assessed in P19 cells during the RA-induced differentiation. (*C*) TFs' enrichment per coexpression path in P19 cells as predicted by DREM (hypergeometric distribution probability). The heat map illustrates the enrichment confidence per coexpression path for each of the most confident TF associations (CellNet database) further classified by Euclidean hierarchical clustering (EHC). On the *right*, the identity of relevant TFs per cluster and their implication in common and endodermal (F9)-/neuronal (P19)-specific gene programs is displayed. (*D*) mRNA expression levels of relevant TFs displayed in *C* assessed with either RA or RAR-specific agonists.

In summary, the reconstructed GRN reconstitutes a scenario in which cascades of TF-driven common and specific regulatory programs are responsible for acquisition of endodermal and neuronal fates. Thus, cell fate specification is predefined by a given cellular context even when the same chemical trigger is used for program initiation.

Identification of "master regulators" from a hierarchical analysis of the GRN

The reconstructed GRN for neuronal/endodermal fates reveals common and cell fate-specific factors, which instruct the two RA-induced differentiation programs. The neurogenic GRN contains several known neuronal TFs, but the majority of these are activated late. To identify early key TFs ("master regulators") critical for cell fate commitment, we simulated the capacity of each of the 1087 nodes of the P19-specific program to propagate the transcriptional regulatory cascade toward the latest time point, corresponding to the ultimate biological readout (Fig. 6A; Supplemental Fig. S13). This analysis predicted less than 75 nodes as master regulators of the neurogenic program (Fig. 6B; Supplemental Fig. S13). Among them, several known neuronal TFs, like NEUROD1, NEUROG2, POU3F2, or MYT1L, reconstitute <20% of the P19 program, while other "early" factors, like ASCL1 (Huang et al. 2012, 2015), NR2F2 (Zhou et al. 2015), or NR4A2 (Park et al. 2006) reconstitute >60% (Fig. 6B). Importantly, this analysis identified additional TFs (e.g., GBX2, TAL2, TSHZ1, DMRT1, LHX2) with the capacity to reconstitute >50% of the P19-specific program. Moreover, the reconstructed GRN revealed direct and indirect links between many of these factors and connection to the neuronal factors ASCL1, NEUROG1, NEUROG2, and/or POU3F2 (Fig. 6C).

To evaluate the relevance of predicted TF-TG relationships, we used the CRISPR/dCas9 transcription activation strategy to induce expression of endogenous factors (Konermann et al. 2015).

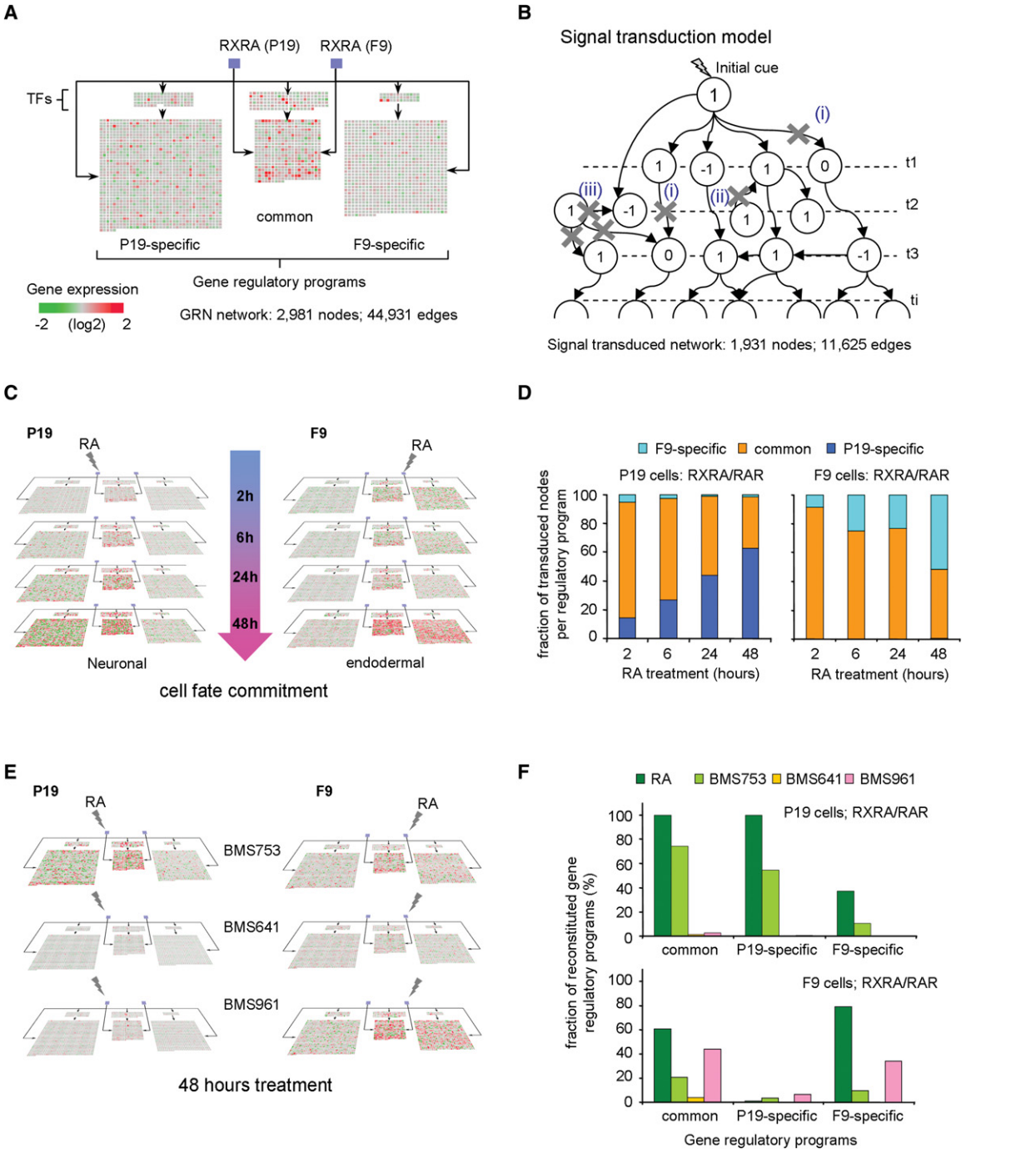


Figure 5. Temporal signal propagation in RA-induced GRNs for neuronal and endodermal cell fate decisions. (*A*) Structure of the reconstructed GRN displaying genes that are selectively or commonly regulated during neuronal and endodermal cell differentiation. For illustration purposes, all edges were removed; arrows indicate the direct regulation of each of these programs by TFs that are bona fide direct RA responsive genes (blue squares; black arrows). Gene expression changes are illustrated as heat maps. (*B*) Signal transduction model aiming at evaluating the coherence between the reconstructed GRN and the temporal gene expression changes. The starting node where the initial cue activates the signal transduction is depicted, as well as the downstream node interconnections required for its propagation. The temporal transcriptional state for each node is defined as 1, 0, or −1 (up-regulated, nonresponsive, or down-regulated, respectively). The model excludes signal cascade progression branches (illustrated by crosses) when (1) the state of a node remains nonresponsive; (2) the directionality of the TF-TG relationship is opposite to the temporal signal flux; or (3) the TF-TG relationships are not part of the main signal transduction propagation branches. (*C*) Temporal transcriptional evolution of the reconstructed GRNs in P19 or F9 RA-induced cell differentiation. Note that common programs dominate at early time points, while the neuronal/endodermal programs take over at late time points. (*D*) Fraction of transduced nodes per regulatory program for both model systems (F9-specific, common, P19-specific), as assessed by the signal transduction model. As illustrated in *C*, the common gene regulatory program is activated early (>80% in both cell lines after 2 h of RA treatment), while the cell fate-specific program is set up progressively (~60% of specific programs in either of the model systems after 48 h of RA). (*E*) Responsiveness of common and neuronal-/endodermal-specific GRNs described in *A* to agonists selective for the three RAR isotypes. (*F*) Fraction of reconstituted gene regulatory programs (GRPs) (after 72 h of RA treatment) in both model systems when either the RA or RAR-specific agonists-derived transcriptomes are used for modeling signal transduction propagation.

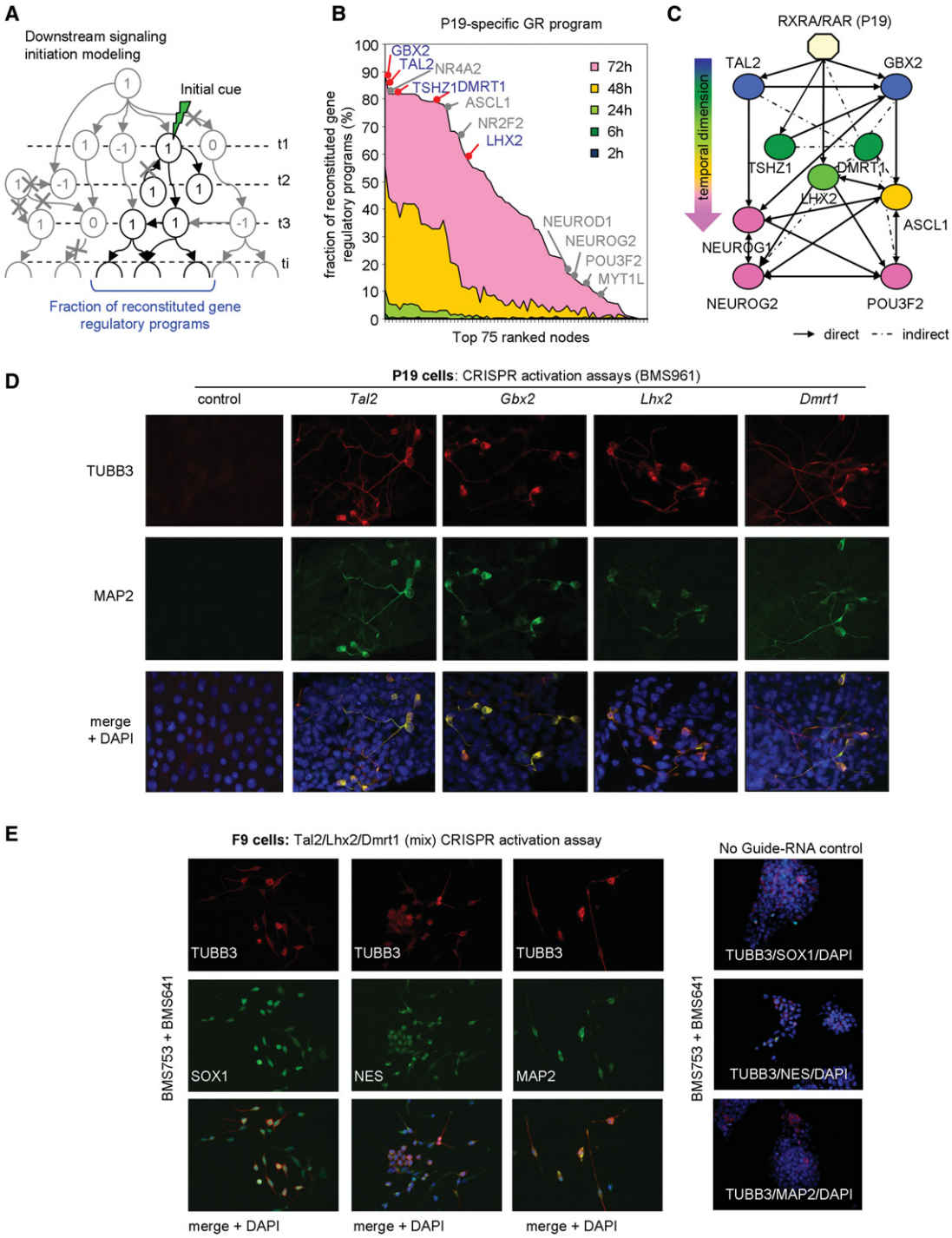


Figure 6. Predicting master regulators of neurogenesis by modeling signal transduction propagation. (*A*) Scheme of the signal transduction propagation model initiated at a downstream layer in the reconstructed GRN. (*B*) 1087 nodes comprising the P19-specific GRP (*x*-axis) ranked according to their performance in reconstituting the ultimate level of the P19-specific program (*y*-axis). Previously known neuronal factors are depicted in association with their position in the ranking (gray). Less characterized factors with significant signal propagation performance toward the final level are in blue. (*C*) Transcriptional regulatory relationships among the newly predicted factors in *B* are depicted in the context of their interconnections with relevant neuronal markers. Their relative temporal transcriptional response under RA-driven conditions is indicated (color coded). (*D*) Immunofluorescence micrographs illustrating the presence of the neuronal markers TUBB3 (red) and MAP2 (green), in P19 cells after CRISPR/dCas9 (D10/N863A)-mediated transcription activation of *Tal2, Gbx2, Lhx2*, or *Dmrt1* treated with the RARG-specific agonist BMS961 or vehicle. (*E*) Immunofluorescence micrographs revealing the presence of the neuronal markers TUBB3 (red) and MAP2, SOX1, or Nestin (NES; green) in F9 cells after CRISPR/dCas9 (D10/N863A)-mediated transcription activation of *Tal2, Lhx2*, and *Dmrt1* treated with BMS961 and the RARB-specific BMS641. In the *right* panel, a mock-CRISPR/dCas9 (D10/N863A) transfection assay (no guide RNA) in F9 cells under identical treatment conditions is displayed.

Specifically, we used guide RNAs to target the *Tal2*, *Gbx2*, *Lhx2*, or *Dmrt1* promoters for VP64-mediated transcription activation. To study if the common regulatory program is required for efficient cell fate specification, we performed the activation assays in the presence or absence of the RARG-specific agonist BMS961. This ligand does not induce neuronal differentiation of P19 cells (Supplemental Fig. S9) but activates components of the common program. *Tal2* activation (>200-fold in the presence of BMS961) resulted in induced mRNA expression of *Gbx2* (greater than sevenfold), *Lhx2* (>3.5-fold), and of the neuronal factors *Pou3f2*, *Neurog2*, and *Neurog1* (>3.5-fold). Similarly, *Gbx2*, *Lhx2*, or *Dmrt1* activation resulted in increased expression of known neuronal factors (Supplemental Fig. S14). The BMS961-enhanced response of most neuronal factors supported our hypothesis that the common program is required for/supports the fate-selective programs. In all cases, the engineered activation of these factors (*Tal2*, *Gbx2*, *Lhx2*, or *Dmrt1*) induced the response of the above neuronal markers and led to a positive immunostaining for the neuronal markers TUBB3 and MAP2 (Fig. 6D).

To ultimately demonstrate the potential of the identified neurogenic key factors to impose a neurogenic fate onto a differently committed cell, we used the CRISPR/dCas9 (D10/N863A) strategy to induce in F9 cells the expression of known neurogenic TFs and master regulators predicted by our transcription propagation approach. As illustrated in Supplemental Figure S15A, inefficient CRISPR/dCas9 (D10/N863A)-mediated activation of neuronal factors in F9 cells was observed in the absence of retinoids. We therefore hypothesized that the activation of the common gene program is required for efficient CRISPR/dCas9 (D10/N863A)-mediated induction of these factors in F9 cells. Indeed, exposing CRISPR/dCas9 (D10/N863A)-transfected cells to ATRA (Supplemental Fig. S15A) or RAR subtype-specific agonists (Supplemental Fig. S16A) resulted in dramatically increased expression of the neurogenic factors. This is also supported by the presence (ATRA, BMS753) and absence (EtOH) of morphological changes in CRISPR/dCas9 (D10/N863A)-transfected cells (Supplemental Fig. S15B). Together, this suggested that activation of a subset of the RA-induced program(s) is required for optimal CRISPR/dCas9 (D10/N863A)-mediated transcription activation, possibly due to modulation of promoter accessibility. Using this combinatorial approach, induction of neurogenesis-specific genes was seen upon CRISPR/dCas9 (D10/N863A)-mediated activation of cognate genes for both known neurogenic factors (ASCL1, NEUROG2, POU3F2, MYT1L, OLIG2) (Supplemental Fig. S16A) and the new ones predicted in the present study (TAL2, LHX2, DMRT1) (Supplemental Fig. S16B). In all cases, F9 neuronal transdifferentiation was confirmed by immunostaining for TUBB3, SOX1, Nestin, and MAP2 (Fig. 6E). Together, these results demonstrated that the use of signal propagation models from reconstructed GRNs identifies novel (and confirms known) key TFs involved in cell fate acquisition.

The EC GRNs are relevant for mouse embryonic stem cell differentiation

To explore the relevance of our observations and networks for RA-driven mouse ESC differentiation, we have analyzed publicly available temporal studies (GSE30176 [Lin et al. 2011]; GSE34279 [Gaertner et al. 2012]). Reconstruction of its dynamic regulatory map resulted in 14 coexpression paths (Fig. 7A). The integration of the CellNet TF-TG collection predicted several self-renewal TFs enriched in the most down-regulated group of genes, as well as factors like OTX2, GBX2, TSHZ1, or DMRT1, identified here as relevant components of the RA-induced neuronal differentiation. Other coexpression paths are also enriched for components identified in the P19 model, revealing major similarities.

Comparing RA-regulated genes in mouse ES and EC cells revealed that >75% of these genes are commonly up-regulated in ES and P19 cells; about half of those are also induced in F9 cells (Fig. 7B,C). Similarly, >65% of the genes repressed in ES are also repressed in P19 cells, again supporting a similar response to RA (Fig. 7B; Supplemental Fig. S18). Despite these similarities, each of the systems contained sets of additional DEGs. GO analysis for each of the observed sets of common P19 and ES up-regulated genes retrieved neuronal fate-related terms, while up-regulated genes shared by all three systems were specifically enriched for RA metabolic processes (Fig. 7D).

Unexpectedly, the transcriptional response of ES cells contributed significantly to both the common and the specific (P19/neuronal; F9 endodermal) gene regulatory programs (GRPs) (Fig. 7E,F), corroborating earlier reports of nonhomogeneous RA-induced differentiation of mouse ES cells (Sartore et al. 2011). Indeed, improved differentiation protocols involve complex cocktails of factors to increase the yield and purity of neuronal precursors (Ying et al. 2003; Abranches et al. 2009).

Discussion

Cell fate transitions are fundamental for the genesis of multicellular organisms, and aberrations from this body plan can generate pathologies. One such process is neurogenesis, a highly complex phenomenon that involves a plethora of instructive signals, including cell-to-cell communication and extrinsic chemical signals, which during organogenesis generate regionally organized cells with diverse functionality.

Interestingly, the blueprint of neurogenesis, which includes the principal architecture of the brain, is already encoded within neuronal stem cells. Indeed, 3D cultures of cerebral organoids have been developed from ES or iPS cells (Lancaster et al. 2013). Notably, neurogenesis occurs also in the adult mammalian brain (Eriksson et al. 1998; Ming and Song 2011), and the plasticity of cell fates in adult tissues prompted critical reflection about concepts of stemness, cell differentiation, and regeneration (Sanchez Alvarado and Yamanaka 2014). However, while some key TFs can be sufficient for cell reprogramming (Weintraub et al. 1989; Zhou et al. 2008; Ieda et al. 2010; Sekiya and Suzuki 2011), our knowledge about the temporal evolution and regulation of gene networks, which specify cell fates and plasticity, has remained fragmentary. Therefore, we have initiated a study to define the temporal regulation of gene programs that are initiated by a single compound, the morphogen all-*trans* retinoic acid, in P19 cells, which are committed to undergo neuronal differentiation. The involvement of RA in the developing nervous system and the adult brain, including its role in regeneration, is well-documented (Vergara et al. 2005). We have compared these programs with those responsible for RA-induced endodermal differentiation of F9 cells (Mendoza-Parra et al. 2011) and defined common and cell-specific programs, as well as subnetworks initiated by nodes critical for lineage identity. The results of this analysis were used to instruct cells adapting a neuronal fate by a combination of subtype-specific retinoids and CRISPR/dCas-mediated activation of endogenous genes.

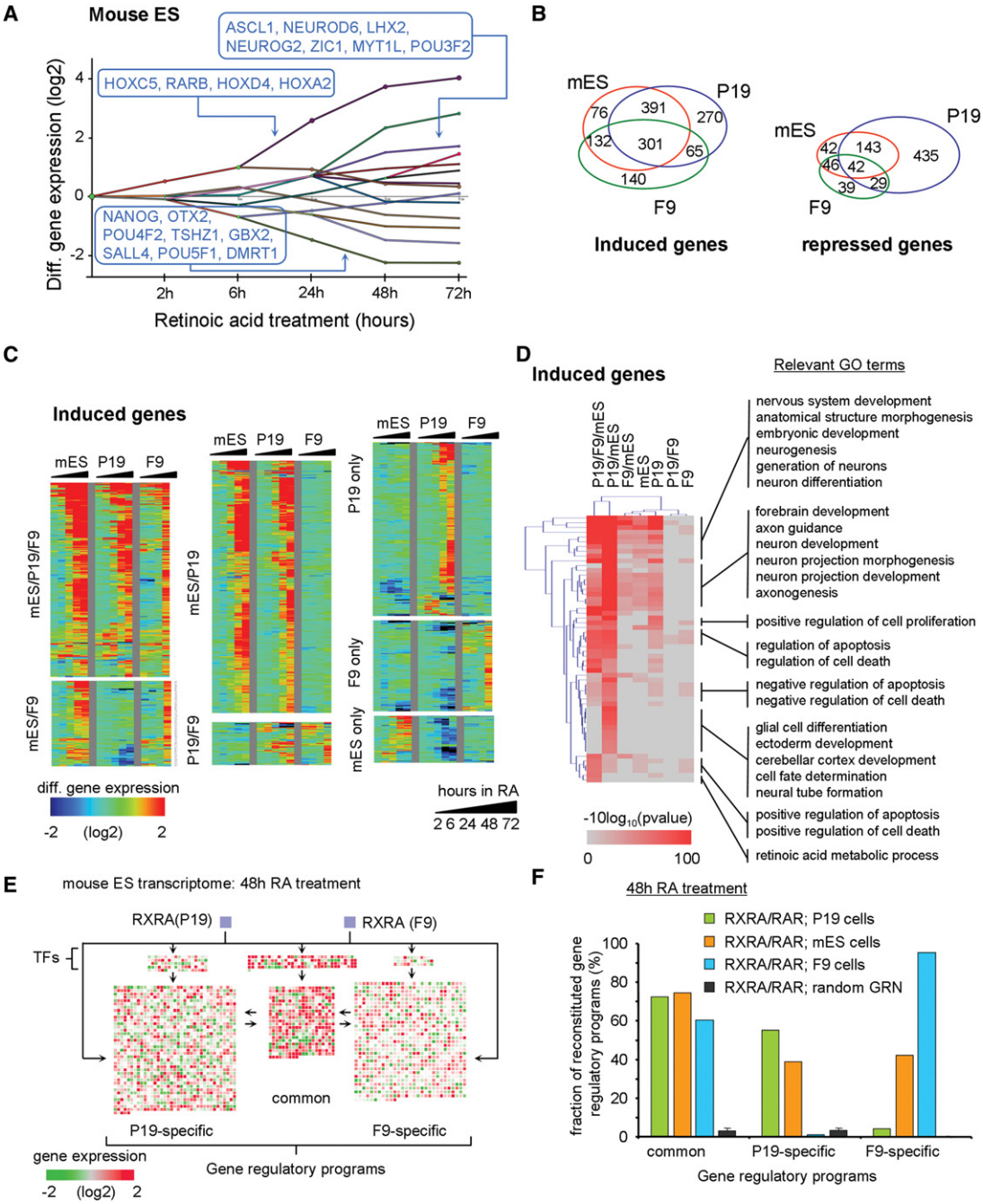


Figure 7. Relevance of the inferred GRP in EC cells in comparison to the mouse ES model system. (*A*) Dynamic regulatory map reconstructed from publicly available temporal transcriptome data of RA-treated mES cells. (*B*) Venn diagram illustrating the number of DEGs shared with either P19 or F9 cells during the RA-induced program (all time points included). (*C*) Temporal mRNA gene expression levels (heat map; induced genes) associated with each of the cell model systems and displayed based on the classification in *B* (for repressed genes, see Supplemental Fig. S18). (*D*) GO analysis of induced genes displayed in *B*. (*E*) Genes expressed in mouse ES cells after 48 h of RA treatment revealing common and F9-/P19-specific programs and color-coded according to their expression levels relative to the noninduced state. Genes composing all three GRPs are regulated in ES cells, despite the expected neuronal cell fate commitment. (*F*) Fraction of reconstituted GRPs in all three cell systems (after 48 h of RA treatment). Note that in mouse ES cells, both the P19- and F9-specific programs are induced at a level of ~40%; this contrasts with the much more specific neuronal and endodermal programs in P19 or F9 cells, respectively.

RA induces modular gene programs in committed EC cells

A comparison of RA-induced neuronal and endodermal GRNs revealed common, endodermal-, and neuronal-specific programs; most of the well-known RA-targets (e.g., *Rarb*, *Hox* genes) belong to the common program. The specific programs can be activated by RARA (neuronal) and RARG (endodermal)-selective retinoids (Alvarez et al. 2014), which both activate the common program (Fig. 5E). Given that RA regulates multiple embryonic (e.g., limb development) and cell physiological (e.g., differentiation,

apoptosis) phenomena in different compartments (e.g., hematopoietic system, skin) at different developmental stages (e.g., embryogenesis, organogenesis, adult homeostasis), the overall RA-program is likely composed of common and specific modules. Thus, genes supporting stemness (*Sox2, Nanog, Myc*) are commonly repressed in both EC cell lines, as differentiated cells lose pluripotency. The coordinately regulated *Hox* genes may provide spatiotemporal information to the neuronal and endodermal progeny; for example, the self-organizing capacity observed for ES/iPS cell-derived cerebral organoids (Lancaster et al. 2013) may be linked to the ability of *Hox* genes to define the body plan.

We noted that the common program does not operate in isolation, as it enables CRISPR-activated key genes (Fig. 6D) to induce neuronal differentiation. This indicates intimate links between the cell fate-specific and common programs, which may be of importance for identifying conditions that support/improve the efficiency/functionality of engineered ES/iPS cells for regenerative purposes. It is likely that similar scenarios exist for other nuclear receptors/TFs with similar pleiotropic action as retinoid receptors. It would be interesting to compare in this respect the common and specific gene programs induced by retinoids and vitamin D during hematopoiesis.

The molecular origin of the divergent cell-specific gene programs in P19 and F9 cells remains elusive. While it is clear that different RAR isotypes trigger neurogenic (P19, RARA) and endodermal (F9, RARG) differentiation, we have so far not been able to identify RAR subtype-selective pioneer principles (Zaret and Carroll 2011). Thus, it is unlikely that an RAR subtype-specific gene-regulatory event drives lineage specification; rather, it appears that P19 and F9 cells are already committed. This is supported by the differential epigenetic makeup of P19- and F9-specific genes. In general, activated P19-specific genes lose repressive H3K27me3 marks (with or without gaining H3K4me3 marks) in P19 but not in F9 cells, and vice versa (see Supplemental Fig. S6). Genes that became repressed in one EC cell line showed generally increased levels of H3K27me3 with or without loss of H3K4me3; no such effect was seen in the other EC cell line. However, genes of the common program showed similar epigenetic changes, irrespective of the epigenetic status of genes from the neuronal-/endodermal-specific program.

Notably, the commitment of P19 and F9 cells to their respective lineage was not irreversible, as we could transdifferentiate F9 cells into neurons by activating the common RA-induced program together with the CRISPR/Cas9-mediated induction of endogenous F9 genes that were identified as master regulators of the neuronal program using our novel signal propagation approach (Fig. 6). Notably, activation of the common program was requisite for transdifferentiation.

The RA-regulated programs of ES and EC cells share common and divergent features

A comparative analysis of the gene programs initiated by RA in P19, F9, and ES cells (Lin et al. 2011; Gaertner et al. 2012) yielded the initially surprising result that the ES program was a composite of both EC cells rather than a mimic of the neurogenic P19 program (Supplemental Fig. S17). However, this result reflects that (1) only a fraction of ES cells develop into neurons, (2) sophisticated ES culture conditions are required for efficient differentiation in vitro (Studer 2014), and (3) exogenous RA addresses simultaneously all accessible developmental programs in ES cells, including endodermal ones, thus justifying our choice of committed P19 cells for defining the neurogenic GRN.

A novel in vitro signal propagation approach to identify master regulators

Validation of the RA-dependent neuronal GRN in P19 revealed unexpected results. For example, inactivation of the early induced *Tal2* had no obvious consequences on neurogenesis (Fig. 3G), while inactivation of the similarly expressed *Gbx2* strongly impaired neurogenesis. However, even though not required, CRISPR-mediated activation of endogenous *Tal2* was sufficient to drive neurogenesis (together with the common program), as did the activation of *Gbx2* (see Fig. 6E). Thus, the program is composed of both necessary and sufficient actors, including significant functional redundancy.

One of the questions that derives from the present definition of the neuronal network refers to its plasticity in supporting transdifferentiation. Fibroblasts can be converted to electrophysiologically responsive, marker-positive neurons by exogenously expressed ASCL1, POU3F2, and MYT1L (Wapinski et al. 2013); similar results were obtained by overexpressing two neurogenins in human iPS cells (Busskamp et al. 2014). All these factors are activated rather late in the RA-induced GRN following complex regulatory events (Fig. 6C). This suggests two scenarios: (1) either the complex history of temporally organized gene regulatory events is necessary, as it generates a spatiotemporal "memory" for the development, functional specification, and structural organization of all the cells that constitute a functional CNS, and the transdifferentiation experiments reveal only a testable fraction of this scenario; or (2) the cellular plasticity allows for virtually any cell fate conversion given the correct set of conditions and factors is provided (see also Sanchez Alvarado and Yamanaka 2014). Validating these scenarios experimentally requires blueprints of the developmental programs driving differentiation of CNS compartments and cell types in vivo and an assessment of how this program can be recapitulated in the structures of cerebral organoids.

The value of reconstructing networks

We demonstrate here that by reconstructing the cellular network corresponding to induced cell fate transitions, it is possible to infer relevant factors, their interdependency, and hierarchical position. Particularly useful was the approach to validate nodes and connectivities that were imported from heterologous settings by monitoring their temporal coherence with the current expression data and confirming the functional relevance of predicted key factors by CRISPR-based approaches. By evaluating the potential of a factor to generate the final nodes of the network, we identified several known (e.g., NR4A2, ASCL1, NR2F2) and novel (TAL2, GBX2, LHX2, DMRT1) key factors involved in retinoid-induced neurogenesis (see Fig. 6B). Note that identification of DMRT1 as a potential neuronal differentiation factor previously involved enormous transcriptome profiling efforts (Yamamizu et al. 2013).

Modeling temporal signal propagation in reconstructed GRNs is a general approach to reveal transcriptional interconnection and identify master regulators in any system. Indeed, for validating the corresponding Cytoscape plugin, we applied it to diverse phenomena, including differentiation, reprogramming, and tumorigenesis, supporting its general utility (MA Mendoza-Parra, PE Cholley, J Moehlin, M Lieb, and H Gronemeyer, unpubl.). We thus believe that the comprehensive approach described here is not limited to understanding the molecular circuits

underlying physiological and, when altered, pathological cell fate transition. It provides, moreover, a comprehensive way to monitor the ability of stem, reprogrammed, or transdifferentiated cells to properly adopt a desired cell fate.

Methods

Cell culture

F9 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum (FCS) and 4.5 g/L glucose; P19 cells were grown in DMEM supplemented with 1 g/L glucose, 5% FCS, and 5% delipidated FCS. Both media contained 40 µg/mL Gentamicin. F9 or P19 EC cells were cultured in monolayer on gelatin-coated culture plates (0.1%). For cell differentiation assays, RA was added to plates to a final concentration of 1 µM for different exposure times. For treatment with RAR subtype-specific agonists, cells were incubated with BMS961 (RARG-specific; 0.1 µM), BMS753 (RARA-specific; 1 µM), and/or BMS641 (RARB-specific; 0.1 µM).

RT-qPCR and transcriptomics

Total RNA was extracted from EC cells treated with either RA or RAR-specific agonists, using the GenElute Mammalian Total RNA Miniprep kit (Sigma). Two micrograms of the extracted RNA were used for reverse transcription (AMV-RTase, Roche; Oligo [dT], New England Biolabs; 1 h at 42°C and 10 min at 94°C). Transcribed cDNA was diluted 10-fold and used for real-time quantitative PCR (Roche LC480) (primers, Supplemental Methods).

For transcriptomics analysis, AffymetrixGeneChip Mouse Gene 1.0 ST arrays were used (Supplemental Methods). For comparing transcriptomes, we normalized all raw CELL files with the Affymetrix software Expression Console.

Chromatin immunoprecipitation assays

ChIP assays were performed according to standard procedures (Supplemental Methods). All ChIP and FAIRE assays were validated using positive and negative controls. ChIP validation assays were performed by quantitative real-time PCR using the Qiagen Quantitect kit.

Massive parallel sequencing and quality control

qPCR-validated ChIPs were quantified (Qubit dsDNA HS kit; Invitrogen); multiplexed sequencing libraries were prepared from 10 ng of the ChIPed material (Supplemental Methods).

Sequence-aligned files were qualified for enrichment using the NGS-QC Generator (Mendoza-Parra et al. 2013b). Briefly, this methodology computes enrichment quality descriptors discretized in a scale ranging from "AAA" (Best) to "DDD" (worst). Based on this quantitative method, all ChIP-seq and FAIRE data sets described in this study presented quality grades higher than "CCC"; integrative studies were thus performed exclusively with high-quality data sets.

Enrichment pattern detection and intensity profile normalization

Relevant binding sites in all ChIP-seq and FAIRE-seq data sets were identified with MeDiChISeq (Mendoza-Parra et al. 2013a); multiprofile comparisons were done after quantile normalization (Supplemental Methods; Mendoza-Parra et al. 2012).

Dynamic regulatory maps and RA-driven GRN reconstruction

We reconstructed GRNs by combining several layers of information. First, we identified direct TGs as those containing (1) a proximal RXRA and FAIRE enrichment event (<10 kb distance), and (2) responding to both RA and the corresponding BMS-specific agonist. Downstream regulatory processes were reconstructed by integrating the TF-TG collection of CellNet (Cahan et al. 2014; Kim and Scholer 2014) in the RA-regulated EC GRPs deduced by DREM (Supplemental Methods).

The integration in Cytoscape (version 2.8) of the RXRA-direct targets per cell type with the downstream regulatory networks assessed from the DREM/CellNet approach generated a GRN composed of 2981 nodes and 44,931 edges, organized in common or EC-specific regulated programs. GRN complexity was reduced by applying topological metrics (Yu et al. 2007; Chin et al. 2014). The ultimate reduced GRN was composed of 80 nodes and 626 edges, with a ranking color code (heat map) displaying the hub importance metrics (Supplemental Fig. S17). The organization of reduced GRN and its visualization were performed with the Cytoscape package Cerebral (Supplemental File S1; Barsky et al. 2007).

Modeling signal transduction progression in reconstructed GRNs

To validate the relevance of the TF-TGs relationships composing the reconstructed F9/P19 GRN, we developed a computational framework for modeling signal propagation within the network. It takes as initial information: (1) the topology of the reconstructed network in which the TF-TG directionality is essential; (2) the temporal transcriptional information associated with each of the nodes composing the network; and (3) the node from which the signal transduction is initiated, (starting node) to follow the temporal evolution of signal(s) until the ultimate time points of the experimental data set (final nodes). In this context, the signal propagation model evaluates in the first round the transcriptional response at the first time point (e.g., 2 h of RA treatment) of the TGs associated with the starting node. In the second round, the model defines starting nodes, initially defined by the user as well as those with a differential transcriptional behavior in the first round. In this manner, the second round evaluates the interconnections (edges) between the newly defined starting nodes and their corresponding targets by evaluating their transcriptional behavior at the second time point (e.g., 6 h of RA treatment). Such analysis over all available transcriptional time points reveals the coherence between the TF-TGs relationships and the temporal transcriptional information. Finally, the number of retrieved nodes at the end of the signal transduction model is compared with the expected user-provided list of final nodes. The signal propagation was performed multiple times using a randomized network as a control.

The GRN reduction (Fig. 5), the prediction of factors driving the neuronal program (Fig. 6), as well as the evaluation over mouse ES data sets (Fig. 7) have been performed using an in-house R script (Supplemental File S2); a Cytoscape plugin is in preparation.

Targeted gene knockouts with the CRISPR/*Cas9* system

Cells were transfected with pairs of double-nickase plasmids encoding the Cas9D10A mutation and a 20-nt guide RNA (Santa Cruz Biotech). Single cell-derived cultures were treated with ATRA, and loss-of-expression from the targeted genes was validated by qPCR relative to control cultures (Supplemental Methods).

CRISPR/dCas9 (D10/N863A) transcriptional activation and immunohistochemical staining

EC cells were transfected with CRISPR/dCas9 (D10/N863A) activation plasmids (Santa Cruz Biotech) using lipofection and treated with ATRA, RAR-specific agonists, or ethanol, complemented with antibiotics. Six days later, cells were fixed, permeabilized, and immunostained as specified (Supplemental Methods).

Data access

Affymetrix microarrays and Illumina platform ChIP-seq and FAIRE-seq data described in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE68291.

Acknowledgments

We thank all the members of our laboratory and the IGBMC sequencing and microarray platform for discussion. Special thanks go to Anna Podlesny for sharing expertise in ICC assays. This study was supported by AVIESAN-ITMO Cancer, the Ligue National Contre le Cancer, the Institut National du Cancer (INCa), and the Agence Nationale de la Recherche (ANRT-07-PCVI-0031-01, ANR-10-LABX-0030-INRT, and ANR-10-IDEX-0002-02).

References

Abranches E, Silva M, Pradier L, Schulz H, Hummel O, Henrique D, Bekman E. 2009. Neural differentiation of embryonic stem cells *in vitro*: a road map to neurogenesis in the embryo. *PLoS One* **4:** e6286.

Achim K, Peltopuro P, Lahti L, Tsai HH, Zachariah A, Astrand M, Salminen M, Rowitch D, Partanen J. 2013. The role of *Tal2* and *Tal1* in the differentiation of midbrain GABAergic neuron precursors. *Biol Open* **2:** 990–997.

Alvarez R, Vaz B, Gronemeyer H, de Lera AR. 2014. Functions, therapeutic applications, and synthesis of retinoids and carotenoids. *Chem Rev* **114:** 1–125.

Barsky A, Gardy JL, Hancock RE, Munzner T. 2007. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* **23:** 1040–1042.

Bouillet P, Chazaud C, Oulad-Abdelghani M, Dolle P, Chambon P. 1995. Sequence and expression pattern of the *Stra7* (*Gbx-2*) homeobox-containing gene induced by retinoic acid in P19 embryonal carcinoma cells. *Dev Dyn* **204:** 372–382.

Busskamp V, Lewis NE, Guye P, Ng AH, Shipman SL, Byrne SM, Sanjana NE, Murn J, Li Y, Li S, et al. 2014. Rapid neurogenesis through transcriptional activation in human stem cells. *Mol Syst Biol* **10:** 760.

Cahan P, Li H, Morris SA, Lummertz da Rocha E, Daley GQ, Collins JJ. 2014. CellNet: network biology applied to stem cell engineering. *Cell* **158:** 903–915.

Chiba H, Clifford J, Metzger D, Chambon P. 1997. Specific and redundant functions of retinoid X Receptor/Retinoic acid receptor heterodimers in differentiation, proliferation, and apoptosis of F9 embryonal carcinoma cells. *J Cell Biol* **139:** 735–747.

Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. 2014. *cytoHubba*: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol* **8 (**Suppl 4**):** S11.

Eriksson PS, Perfilieva E, Bjork-Eriksson T, Alborn AM, Nordborg C, Peterson DA, Gage FH. 1998. Neurogenesis in the adult human hippocampus. *Nat Med* **4:** 1313–1317.

Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z. 2007. Reconstructing dynamic regulatory maps. *Mol Syst Biol* **3:** 74.

Gaertner B, Johnston J, Chen K, Wallaschek N, Paulson A, Garruss AS, Gaudenz K, De Kumar B, Krumlauf R, Zeitlinger J. 2012. Poised RNA polymerase II changes over developmental time and prepares genes for future expression. *Cell Rep* **2:** 1670–1683.

Gronemeyer H, Gustafsson JA, Laudet V. 2004. Principles for modulation of the nuclear receptor superfamily. *Nat Rev Drug Discov* **3:** 950–964.

Huang HS, Turner DL, Thompson RC, Uhler MD. 2012. Ascl1-induced neuronal differentiation of P19 cells requires expression of a specific inhibitor protein of cyclic AMP-dependent protein kinase. *J Neurochem* **120:** 667–683.

Huang HS, Redmond TM, Kubish GM, Gupta S, Thompson RC, Turner DL, Uhler MD. 2015. Transcriptional regulatory events initiated by Ascl1 and Neurog2 during neuronal differentiation of P19 embryonic carcinoma cells. *J Mol Neurosci* **55:** 648–705.

Ieda M, Fu JD, Delgado-Olguin P, Vedantham V, Hayashi Y, Bruneau BG, Srivastava D. 2010. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142:** 375–386.

Inoue F, Kurokawa D, Takahashi M, Aizawa S. 2012. Gbx2 directly restricts *Otx2* expression to forebrain and midbrain, competing with class III POU factors. *Mol Cell Biol* **32:** 2618–2627.

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

Kim KP, Scholer HR. 2014. CellNet—where your cells are standing. *Cell* **158:** 699–701.

Kobayashi T, Komori R, Ishida K, Kino K, Tanuma S, Miyazawa H. 2014. *Tal2* expression is induced by all-*trans* retinoic acid in P19 cells prior to acquisition of neural fate. *Sci Rep* **4:** 4935.

Kobayashi T, Suzuki M, Morikawa M, Kino K, Tanuma SI, Miyazawa H. 2015. Transcriptional regulation of *Tal2* gene by all-*trans* retinoic acid (atRA) in P19 cells. *Biol Pharm Bull* **38:** 248–256.

Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, et al. 2015. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517:** 583–588.

Lancaster MA, Renner M, Martin CA, Wenzel D, Bicknell LS, Hurles ME, Homfray T, Penninger JM, Jackson AP, Knoblich JA. 2013. Cerebral organoids model human brain development and microcephaly. *Nature* **501:** 373–379.

Laudet V, Gronemeyer H. 2002. *The nuclear receptor factsbook*. Academic Press, San Diego.

Li JY, Joyner AL. 2001. *Otx2* and *Gbx2* are required for refinement and not induction of mid-hindbrain gene expression. *Development* **128:** 4979–4991.

Lin C, Garrett AS, De Kumar B, Smith ER, Gogol M, Seidel C, Krumlauf R, Shilatifard A. 2011. Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC). *Genes Dev* **25:** 1486–1498.

Martinez-Ceballos E, Gudas LJ. 2008. Hoxa1 is required for the retinoic acid-induced differentiation of embryonic stem cells into neurons. *J Neurosci Res* **86:** 2809–2819.

Martinez-Monedero R, Yi E, Oshima K, Glowatzki E, Edge AS. 2008. Differentiation of inner ear stem cells to functional sensory neurons. *Dev Neurobiol* **68:** 669–684.

Mendoza-Parra MA, Gronemeyer H. 2013. Genome-wide studies of nuclear receptors in cell fate decisions. *Semin Cell Dev Biol* **24:** 706–715.

Mendoza-Parra MA, Walia M, Sankar M, Gronemeyer H. 2011. Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. *Mol Syst Biol* **7:** 538.

Mendoza-Parra MA, Sankar M, Walia M, 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 MA, Nowicka M, Van Gool W, Gronemeyer H. 2013a. Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. *BMC Genomics* **14:** 834.

Mendoza-Parra MA, Van Gool W, Mohamed Saleem MA, Ceschin DG, Gronemeyer H. 2013b. A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res* **41:** e196.

Millet S, Campbell K, Epstein DJ, Losos K, Harris E, Joyner AL. 1999. A role for *Gbx2* in repression of *Otx2* and positioning the mid/hindbrain organizer. *Nature* **401:** 161–164.

Ming GL, Song H. 2011. Adult neurogenesis in the mammalian brain: significant answers and significant questions. *Neuron* **70:** 687–702.

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

Monzo HJ, Park TI, Montgomery JM, Faull RL, Dragunow M, Curtis MA. 2012. A method for generating high-yield enriched neuronal cultures from P19 embryonal carcinoma cells. *J Neurosci Methods* **204:** 87–103.

Morris SA, Cahan P, Li H, Zhao AM, San Roman AK, Shivdasani RA, Collins JJ, Daley GQ. 2014. Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158:** 889–902.

Nakayama Y, Kikuta H, Kanai M, Yoshikawa K, Kawamura A, Kobayashi K, Wang Z, Khan A, Kawakami K, Yamasu K. 2013. Gbx2 functions as a transcriptional repressor to regulate the specification and morphogenesis of the mid-hindbrain junction in a dosage- and stage-dependent manner. *Mech Dev* **130:** 532–552.

Park CH, Kang JS, Shin YH, Chang MY, Chung S, Koh HC, Zhu MH, Oh SB, Lee YS, Panagiotakos G, et al. 2006. Acquisition of *in vitro* and *in vivo*

functionality of Nurr1-induced dopamine neurons. *FASEB J* **20:** 2553–2555.
Rosenfeld MG, Lunyak VV, Glass CK. 2006. Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes Dev* **20:** 1405–1428.
Sanchez Alvarado A, Yamanaka S. 2014. Rethinking differentiation: stem cells, regeneration, and plasticity. *Cell* **157:** 110–119.
Sartore RC, Campos PB, Trujillo CA, Ramalho BL, Negraes PD, Paulsen BS, Meletti T, Costa ES, Chicaybam L, Bonamino MH, et al. 2011. Retinoic acid-treated pluripotent stem cells undergoing neurogenesis present increased aneuploidy and micronuclei formation. *PLoS One* **6:** e20667.
Sekiya S, Suzuki A. 2011. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475:** 390–393.
Soprano DR, Teets BW, Soprano KJ. 2007. Role of retinoic acid in the differentiation of embryonal carcinoma and embryonic stem cells. *Vitam Horm* **75:** 69–95.
Studer L. 2014. The nervous system. In *Essentials of stem cell biology* (ed. Lanza R, Atala A), pp. 163–184. Academic Press, Amsterdam.
Tan Y, Xie Z, Ding M, Wang Z, Yu Q, Meng L, Zhu H, Huang X, Yu L, Meng X, et al. 2010. Increased levels of FoxA1 transcription factor in pluripotent P19 embryonal carcinoma cells stimulate neural differentiation. *Stem Cells Dev* **19:** 1365–1374.
Taneja R, Roy B, Plassat JL, Zusi CF, Ostrowski J, Reczek PR, Chambon P. 1996. Cell-type and promoter-context dependent retinoic acid receptor (RAR) redundancies for *RARβ2* and *Hoxa-1* activation in F9 and P19 cells can be artefactually generated by gene knockouts. *Proc Natl Acad Sci* **93:** 6197–6202.
Vergara MN, Arsenijevic Y, Del Rio-Tsonis K. 2005. CNS regeneration: a morphogen's tale. *J Neurobiol* **64:** 491–507.
Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Sudhof TC, Wernig M. 2010. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463:** 1035–1041.
Voronova A, Fischer A, Ryan T, Al Madhoun A, Skerjanc IS. 2011. Ascl1/Mash1 is a novel target of Gli2 during Gli2-induced neurogenesis in P19 EC cells. *PLoS One* **6:** e19174.
Wapinski OL, Vierbuchen T, Qu K, Lee QY, Chanda S, Fuentes DR, Giresi PG, Ng YH, Marro S, Neff NF, et al. 2013. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell* **155:** 621–635.
Wei Y, Harris T, Childs G. 2002. Global gene expression patterns during neural differentiation of P19 embryonic carcinoma cells. *Differentiation* **70:** 204–219.
Weintraub H, Tapscott SJ, Davis RL, Thayer MJ, Adam MA, Lassar AB, Miller AD. 1989. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci* **86:** 5434–5438.
Yamamizu K, Piao Y, Sharov AA, Zsiros V, Yu H, Nakazawa K, Schlessinger D, Ko MS. 2013. Identification of transcription factors for lineage-specific ESC differentiation. *Stem Cell Rep* **1:** 545–559.
Ying QL, Stavridis M, Griffiths D, Li M, Smith A. 2003. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat Biotechnol* **21:** 183–186.
Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3:** e59.
Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25:** 2227–2241.
Zhou Q, Brown J, Kanarek A, Rajagopal J, Melton DA. 2008. *In vivo* reprogramming of adult pancreatic exocrine cells to β-cells. *Nature* **455:** 627–632.
Zhou X, Liu F, Tian M, Xu Z, Liang Q, Wang C, Li J, Liu Z, Tang K, He M, et al. 2015. Transcription factors COUP-TFI and COUP-TFII are required for the production of granule cells in the mouse olfactory bulb. *Development* **142:** 1593–1605.

Received April 24, 2016; accepted in revised form September 16, 2016.

PUBLICATION N°3

LOGIQA: a database dedicated to long-range genome interactions quality assessment.

Marco-Antonio Mendoza-Parra, Matthias Blum, Valeriya Malysheva, Pierre-Etienne Cholley and Hinrich Gronemeyer

BMC Genomics 16; 17:355

Mendoza-Parra et al. BMC Genomics (2016) 17:355
DOI 10.1186/s12864-016-2642-1

BMC Genomics

DATABASE Open Access

LOGIQA: a database dedicated to long-range genome interactions quality assessment

Marco-Antonio Mendoza-Parra[1,2,3,4,5*], Matthias Blum[1,2,3,4,5], Valeriya Malysheva[1,2,3,4,5], Pierre-Etienne Cholley[1,2,3,4,5] and Hinrich Gronemeyer[1,2,3,4,5*]

Abstract

Background: Proximity ligation-mediated methods are essential to study the impact of three-dimensional chromatin organization on gene programming. Albeit significant progress has been made in the development of computational tools that assess long-range chromatin interactions, next to nothing is known about the quality of the generated datasets.

Method: We have developed LOGIQA (www.ngs-qc.org/logiqa), a database hosting quality scores for long-range genome interaction assays, accessible through a user-friendly web-based environment.

Results: Currently, LOGIQA harbors QC scores for >900 datasets, which provides a global view of their relative quality and reveals the impact of genome size, coverage and other technical aspects. LOGIQA provides a user-friendly dataset query panel and a genome viewer to assess local genome-interaction maps at different resolution and quality-assessment conditions.

Conclusions: LOGIQA is the first database hosting quality scores dedicated to long-range chromatin interaction assays, which in addition provides a platform for visualizing genome interactions made available by the scientific community.

Keywords: HiC, Quality, Chromatin architecture

Background

Today massive parallel DNA sequencing is used not only to decrypt the digital nature of genomes but, in combination with a variety of molecular biology techniques, it provides functional insights into a plethora of regulatory levels and functions, including epigenomics and protein-genome interactions (e.g., ChIP-seq, MeDIP-seq), global transcriptional activity (e.g., RNA-seq, GRO-seq, Ribo-seq), protein-RNA interactions (e.g., CLIP/RIP-seq), chromatin accessibility (e.g., DNase-seq, FAIRE-seq, ATAC-seq, MNase-seq) and the 3-dimensional chromatin organisation [HiC [1], ChIA-PET [2, 3]].

While data acquisition is not anymore an issue, today's challenge is the availability of user-friendly computational solutions to interrogate and integrate - in a comparative manner - billions of data points from different types of functional genomics datasets. In fact, large consortia, like ENCODE, modENCODE, IHEC, NIH Epigenomics Roadmap provide enormous amounts of functional genomics data [4]. In addition, a great number of laboratories perform functional genomics studies in a diverse set of systems covering a large number of molecular targets, such that the number of genomics data linked to various cell/(patho)physiological functions increase exponentially in public repositories like the Gene Expression Omnibus (GEO [5]). However, despite the fact that these repositories contain huge amounts of functional genomics information their exploitation is seriously limited by (i) the lack of information on the quality of these datasets and (ii) the limited toolbox of exploratory computational resources.

* Correspondence: marco@igbmc.fr; hg@igbmc.u-strasbg.fr
[1]Equipe Labellisée Ligue Contre le Cancer, Illkirch, France
Full list of author information is available at the end of the article

In this context, we have developed previously a quality control system dedicated to ChIP-seq and enrichment-related datasets [6] (www.ngs-qc.org). Here we describe LOGIQA (www.ngs-qc.org/logiqa), a database hosting quality scores for long-range genome interaction assays accessible through a user-friendly web-based environment dedicated to quality-scored visualization of long-range interaction maps.

Construction and content

Principles used for quality assessment

LOGIQA is based on the principles applied by the NGS-QC Generator to compute quality descriptors [6]; specifically this involves the assessment of multiple random samplings over long-range interaction readouts to infer numerical local and global quality scores (Fig. 1). In fact, the working hypothesis is that under ideal conditions,

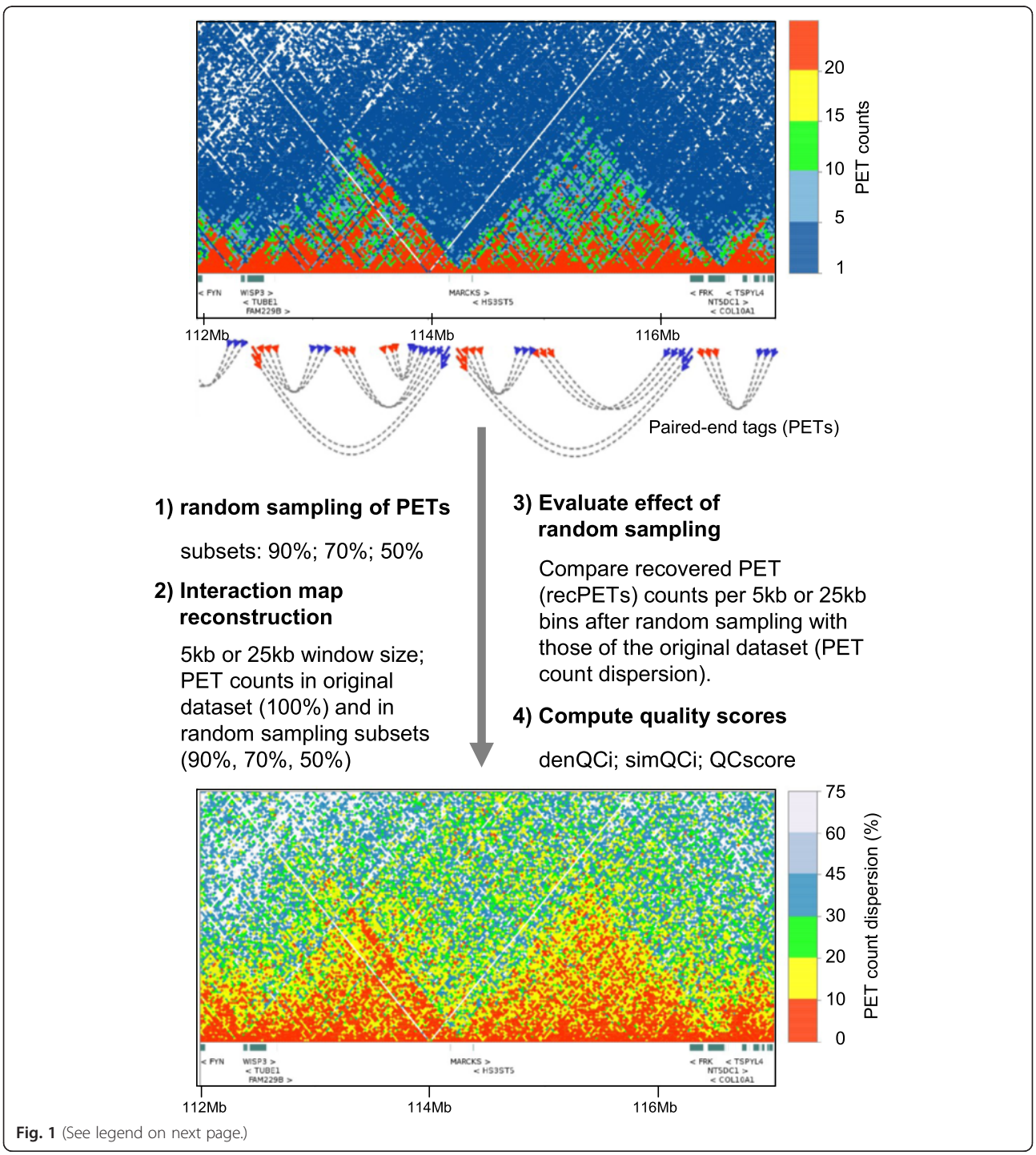


Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Principles in use for Quality Assessment. Total mapped paired-end tags (PETs) are first classified in intra-chromosomal and inter-chromosomal events. For quality assessment, only intra-chromosomal PETs spanning genome distances longer than 10 kb - referred here as filtered PETs - are considered. Random sub-sampling generates PET subsets corresponding to 90, 70 and 50 % of the original filtered PETs and the numbers of PETs in 5 kb or 25 kb size genomic windows is quantified. By comparing each of the PET counts/window in the various random subsets with that observed on the original dataset, the fraction of recovered PET counts (recPETs) after random sub-sampling and the dispersion from the theoretically expected values are calculated. Note that the expected values correspond to a decrease in the number of PET counts per window that is proportional to the random sub-sampling (e.g. recPETs/window =50 % when 50 % of filtered PETs are random sub-sampled). By evaluating the fraction of genomic windows with recPET count dispersions lower than a defined confidence interval (default value 10 %) global quality descriptors like the density and similarity quality indicators (denQCi, and simQCi respectively), as well as the global QCscore are computed. Overall these quality descriptors reflect the fractions of the observed long-range chromatin interactions (>10 kb), which are considered reproducible. On top of the panel: a chromatin interaction map derived from a HiC assay is depicted on the context of the observed PET counts (heatmap scale). On the bottom: After LOGIQA data treatment, the chromatin interaction map displays the inferred PET counts dispersion (in percent; heatmap scale). Notably, the bottom panel recapitulates the genomic contacts observed on the top panel, but in addition it provides a further information concerning their reproducibility over the multiple random sub-sampling assays accomplished during quality assessment

the reconstructed chromatin interaction maps from a subset of the mapped paired-end tags (PETs) should present the same patterns than those observed in the original map. Obviously, multiple factors can lead to a deviation from this optimal situation; one of them is the sequencing depth. Indeed, sequencing depths below a "saturation point", as previously described for ChIP-sequencing assays[7], will lead to a decreased accuracy of chromatin interaction patterns. Importantly, applying this concept to long-range chromatin interaction assays provides a direct relationship between the sequencing depth and the confidence in predicting chromatin interactions. This confidence is herein referred to as the quality of the dataset under study.

Technically, we first selected unique PETs (excluding potential PCR-generated "clonal" reads), which participate in intra-chromosomal interactions longer than 10 kb. We thereby excluded PETs resulting from short-range chromatin interactions, which dominate chromatin interactomes (forming the diagonal in interaction maps) and would bias the quality assessment due to their over-representation. Indeed, Removal of PETs spanning >10 kb or >25 kb led to a direct correlation between the amounts of PETs per dataset and their associated QCscores (Additional file 1: Figure S1A). This correlated also with an improved visual quality and visibility of Topologically Associating Domains (TADs) in chromatin interaction maps (Additional file 1: Figure S1B). Next we established randomly sampled interaction PET subsets for defined fractions of the original population (90 %, 70 %, 50 %; described hereafter as s90, s70 or s50). After random sampling, intra-chromosomal interaction maps were reconstructed by assessing the number of PET counts within 5 kb or 25 kb bins. These two analytical windows enable quality assessment at two different resolutions and facilitate the comparison of different types of datasets; this concerns particularly HiC assays that are generated with different restriction enzymes or ChIA-PET assays involving sonication-sheared chromatin.

Finally, global and local quality scores were computed by comparing the recovered PET counts per 5 kb or 25 kb bin after random sampling with those observed in the original dataset (Fig. 2a).

Computing local and global quality indicators

Technically quality assessment is performed by first computing the recovered PET counts after random sampling as follows:

$$recPETcounts = \left(\frac{samPETcounts}{oPETcounts}\right) * 100$$

where *samPETcounts* correspond to PET counts assessed after random sampling and *oPETcounts* correspond to those retrieved with the original dataset. Then it is used for computing the difference between the observed recovered PET counts after random sampling relative to that ideally expected (samd; which is equivalent to the random sampling density (90 %, 70 % or 50 %)):

$$\partial PETcounts = samd - recPETcounts$$

The recovered PET count dispersion (δPETcounts) per genomic window is referred to as the local QC indicator, such that each evaluated genomic region (5 kb or 25 kb window) can be expressed by this quantitative readout assessed for a given random sampling subset analysis. Importantly, representing genome interaction maps in the context of PET count dispersions (*δPETcounts*) transforms the display into a uniform scale for comparing datasets generated at variable PET sequencing levels (e.g. PET count dispersion: 5-50 %).

Finally, while *δPETcounts* interaction maps provide a visual display of the quality associated to a given genomic region, they do not allow evaluation of the quality of the entire dataset. Therefore, we defined the following global quality descriptors:

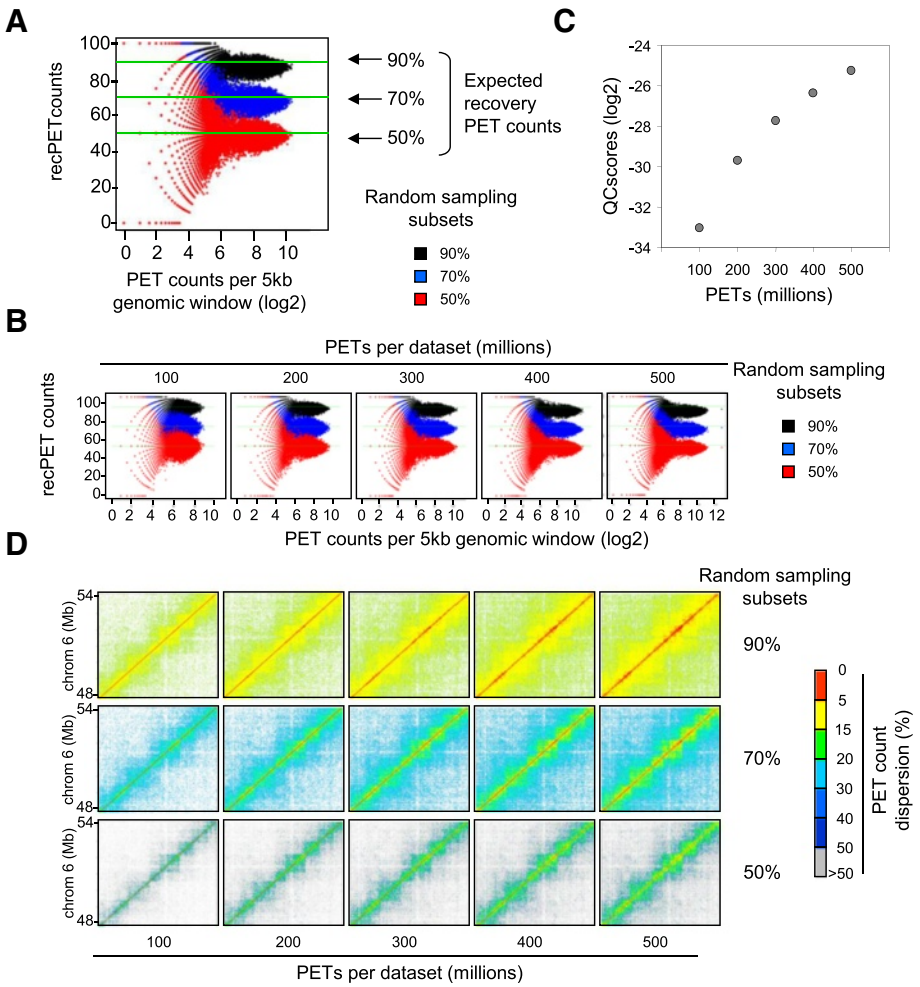


Fig. 2 Assessing quality descriptors over long-range genome interaction assays. **a** Scatter-plot illustrating the fraction of PET counts recovered after random subsampling (Y-axis) relative to the original PET counts in 5 kb genome windows (X-axis). Note that genome windows with high PET counts contain PET levels close to the expected value; in contrast, the lower the PET counts, the higher is the deviation from this theoretically expected level. **b** Recovery scatter-plots assessed from datasets with increasing PET count levels (from 100 to 500 millions). Note that we generated these datasets by random sub-sampling of a large metafile (>600 million reads). **c** QCscores computed from datasets presenting increasing PET count levels (from 100 to 500 millions). The illustrated QCscores, computed from five independent replicates, present variation coefficients below 3 % (see Additional file 1: Figure S2). **d** Local displays illustrating chromatin interactions (chromosome 6, mm9) evaluated in the context of PET count dispersion levels (percentage) per genomic window (5 kb) relative to the expected recovery levels. Note that short-range genomic interactions (diagonal) show the lowest dispersion levels

Density quality indicators (denQCi)

The fraction of genomic regions (5 kb or 25 kb window) in the random sampled datasets presenting *δPETcounts* lower than a defined threshold; which in the context of this study has been fixed at 10 %. Specifically, LOGIQA presents denQCi values computed for 90 %, 70 % and 50 % random samplings (denQC.90, denQC.70 and denQC.50 respectively).

Similarity quality indicators (simQCi)

The ratio between two denQCis is used to evaluate their degree of similarity. Specifically, LOGIQA presents simQCi values computed for denQC.90 and denQC.70 relative to denQC.50 (simQC.90/50 and simQC.70/50 respectively).

Note that **denQCi** aims at quantifying the proportion of genomic regions that fluctuates in less than 10 % for a given random sampling. In fact, an s90 random sampling presents generally less variation from the original dataset, while the s50 subset will have the highest deviation. The **simQCi** measures the relative difference between denQC indicators computed at different random subsampling conditions. For instance, simQC.90/50 compares the denQC at 90 % to that computed at 50 % subsampling. In an ideal situation (saturation of the interactome readout), the fraction of genome interactions

affected by the random sampling is identical at 90 % and 50 % and would yield a simQC = 1. While none of the evaluated datasets are at saturation, the closer this indicator is to 1, the lower is the difference of the denQC indicators between the two random sub-samplings and the higher is the dataset quality.

Intuitively, high quality datasets generally contain a high amount of genomics regions that are "robust" to the most severe 50 % random sub-sampling (i.e., they will display high denQC.50 levels); they will also show low differences between denQCis assessed at various random sub-sampling conditions (i.e., their simQC.90/50 and simQC.s70/50 will be close to 1). To integrate these two aspects on a single readout, we defined a global **QCscore**, which summarizes the previous metrics (denQCi and simQCi) into a single quality descriptor according to the following formula:

$$QCscore = \left(\frac{denQC.50}{simQC.90/50}\right) * \left(\frac{denQC.50}{simQC.70/50}\right)$$

The **QCscore** provides a quality readout, in which the influence of both the denQC.50 and the simQCis computed for s90 relative to s50 (simQC.90/50), and s70 relative to s50 (simQC.70/s50) are represented.

Quality scores computed for a variety of long-range chromatin interaction assays

Because of its universal principle, LOGIQA allows to compute quality scores for chromatin interaction datasets generated from a variety of techniques. Indeed, LOGIQA hosts currently QC scores for >250 publicly available HiC (including several variants of the original protocol, like in situ or capture HiC), but also several ChIA-PET (>50) and 4C-seq (>900) datasets.

Utility

Quality score validations

One of the principal motivations for the development of the present quality score system was to provide a numerical quality descriptor that can predict the optimal sequencing depth for long-range chromatin interaction assays. In fact, even though chromatin interaction assays are expected to require high sequencing depth [8, 9], to date there is no quantitative approach that can compare multiple HiC or similar assays in the context of their relative sequencing depths. The QCscores computed by LOGIQA solve this problem. To illustrate this point, we have constructed a HiC metafile composed of more than 600 million PETs and established subsets by random sampling (100, 200, 300, 400 and 500 million PETs), which were used for calibration of a quality scale. This calibration system reveals a direct negative correlation between sequencing depth and the deviation of the recovered PET count levels from the original dataset after random sampling (Fig. 2b; note the enlarged dispersions of the 100 million vs. the 500 million PET datasets) which translates into a gain of global QCscores for high PET counts (Fig. 2c). Importantly, the reproducibility of the computed global QCscores has been validated from multiple independent random samplings, for which the coefficient of variation was systematically <10 % (Additional file 1: Figure S2). This calibration revealed also the influence of the sequencing depth on PET count dispersion in a selected genome region, as illustrated for chromosome 6 in Fig. 2d, where the chromatin interaction maps reconstructed from different total PET counts are compared using a color-code for PET count dispersion.

We next computed the quality scores for datasets that were reported to be of superior quality due to a modification of the technology, referred to as in situ HiC [10]. Specifically, these assays involve cell in situ proximity ligation, which reduces the frequency of random intermolecular ligation. In this context, we compared QC scores computed for 126 HiC and 87 in situ HiC datasets in the context of their total sequenced PETs. The QC scores of the in situ HiC datasets were generally among the top for a given PET range (Fig. 3a, e) even though there was no clear separation in the quality of HiC and in situ HiC. Rather, it appears that the quality of HiC is more variable than that of in situ HiC, which were generally performed with lower total PETs (Fig. 3b). Our comparative analysis supported also the notion that there are less inter-chromosomal PETs in in situ HiC, as we observed on average more than 70 % intra-chromosomal PETs for in situ HiCs, while significantly less were seen in HiCs (Fig. 3c). Given that LOGIQA computes QC scores on the basis of intra-chromosomal PETs that span a genomic distance of above 10 kb (referred to as "filtered PETs"), we compared the two HiC technologies in the context of filtered PETs. We noted that in situ HiC assays generated on average significantly higher amounts of filtered PETs (~40 %) than HiC (~25 %) assays (Fig. 3d).

Albeit increasing the PET coverage can compensate for reduced QC scores, we were rather interested in comparing the QC scores of HiC and in situ HiC at comparable PET coverage (and thus similar sequencing costs). Notably, mean QC scores around −30 were attained by in situ HiC at a total PET coverage of 50 M to 100 M, while for HiC 100 M to 200 M PETs were required to reach this score (Fig. 3e; dashed green line).

To demonstrate that the global QC score is a meaningful value also for local quality assessment we generated local genome interaction maps (chromosome 6, hg19) generated from two datasets with similar numbers of filtered PETs (~120 million) but significantly different

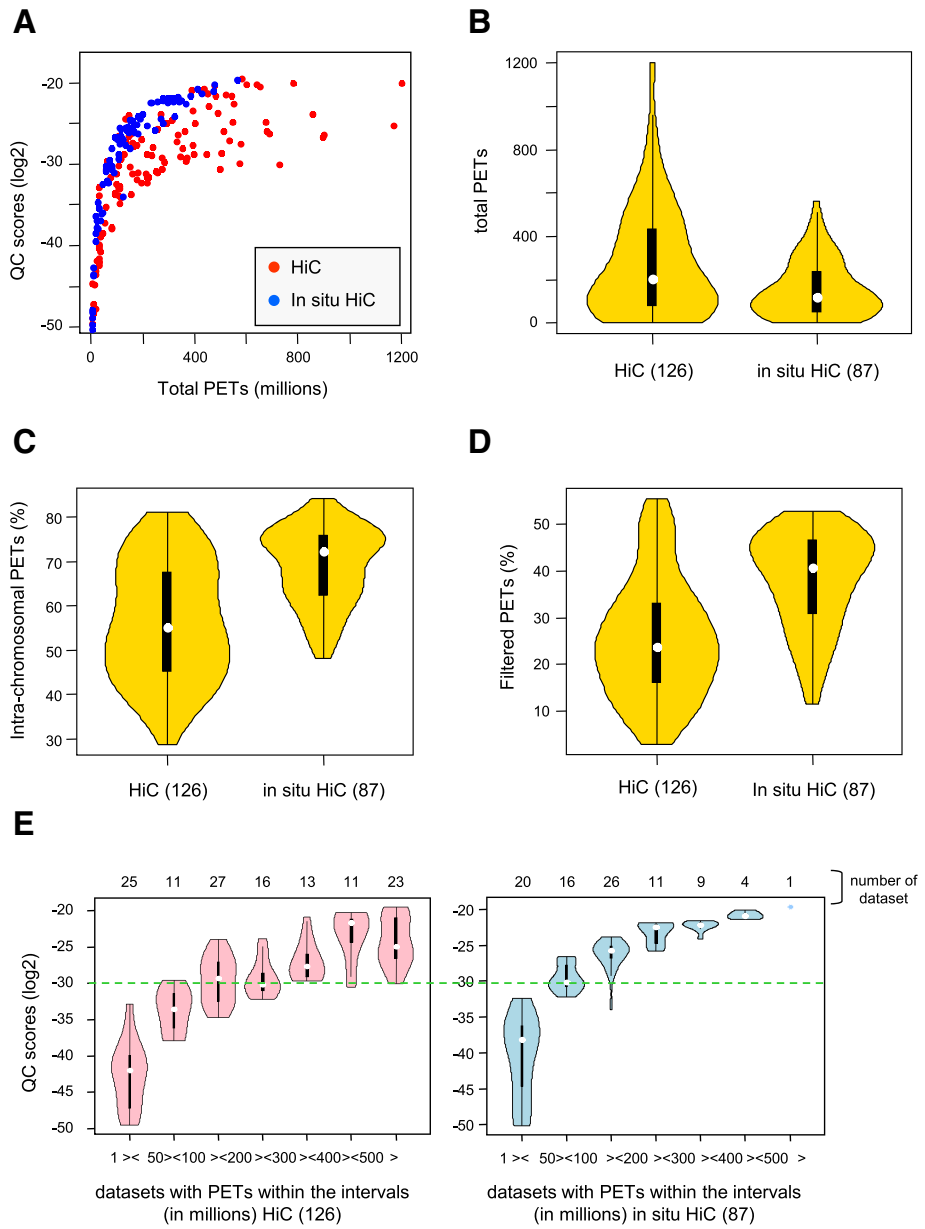


Fig. 3 Quality scores assessed on 76 HiC and 71 in situ HiC assays evaluated in the context of total sequenced PETs. **a** Global Quality scores computed for HiC and in situ HiC assays relative to the total PETs. **b**, **c** and **d** Violin plots illustrating the number of total PETs (**b**), and the fraction of intra-chromosomal (**c**) and intra-chromosomal (**d**) filtered PETs. **e** Violin plots displaying the QC scores for HiC and in situ HiC datasets stratified for identical total PET intervals. The dashed horizontal green line demarcates the median QC score assessed for in situ assays with less than 200 million PETs

global QC scores (Fig. 4). Importantly, the in situ HiC data formed clearly defined topological domains (TADs) for the illustrated region, which corresponds to the human histone gene cluster 1, while the dataset generated by classical HiC appeared less well defined. The visual perception of this difference is further enhanced when the graphic displays were generated from randomly sub-sampled fractions of the two original PET datasets. In fact, when 50 % of the PETs were used for reconstructing the chromatin interactomes, the TAD pattern was readily detectable by visual inspection in the in situ HiC assay for PET dispersion levels <10 %, while the classical HiC assay had PET dispersion levels >20 and a very blurred graphical presentation, in which no TADs could be identified.

Taken together, in situ HiC generates higher amounts of intra-chromosomal PETs and delivers at similar PET coverage better QC scores than HiC. Thus, the present comparative study with large populations of HiC datasets demonstrates the utility of the quality scores computed by LOGIQA.

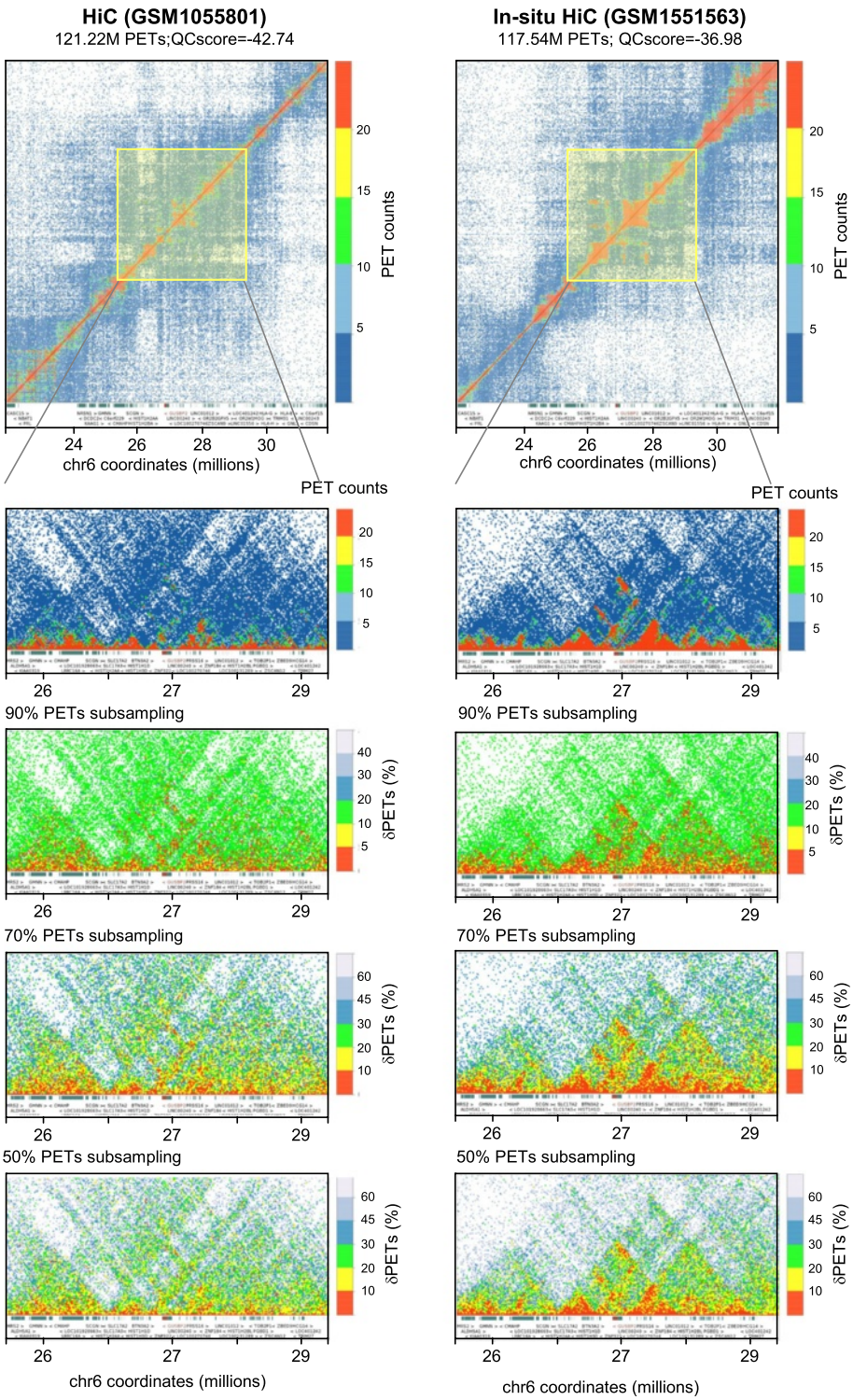


Fig. 4 Chromosome 6 interaction maps displayed for two datasets presenting similar number of filtered PETs but different global QC scores. The illustrated HiC (GSM1055801) and in situ HiC (GSM1551563) datasets comprise about 120 million filtered PETs, nevertheless their global QC scores are different (higher quality for in situ than for the classical HiC assay). In both cases, large genome interaction views (top panels: 10 million bp), as well as closer views (5 million bp) clearly demonstrate the presence of more clearly defined topological domains in the in situ HiC dataset. Note that for the close-ups, both the PET count displays from the original datasets, as well as PET count dispersion displays (dPETs) of the random sub-samplings clearly illustrate the differences in quality of the interaction patterns

Quality scores as quantitative means for revealing heterogeneity among datasets

The LOGIQA database provides a global view of the relative quality of all long-range chromatin interaction assays, thus revealing the impact of the methodology, sequencing-depth and other technical/performance aspects that are specific to each individual assay. To illustrate the last point, we compiled the QC scores of multiple ChIA-PET, HiC and in situ HiC assays and displayed them relative to the filtered PETs used in the assays (Fig. 5, central panel). We then displayed contact maps for two pairs of datasets with largely distinct QC scores but similar filtered PET density - one pair comprised a ChIA-PET and a HiC (about 9 M filtered PETs) and the other an in situ HiC and a classical HiC (about 120 M filtered PETs). The illustrated maps correspond to the same region of chromosome 6 in which either the total PET counts or the PET count dispersions at 70 % sub-sampling are displayed (top and bottom panels, respectively, in each of the blue-framed boxes). It is very obvious from these displays that the in situ HiC GSM1551536 (top right) displays more confident chromatin interaction patterns than the HiC GSM1055801 (bottom right) and indeed, LOGIQA attributed a global QC score of −36.98 to the in situ HiC but only −42.74 to the HiC assay. Remarkably, the target-driven ChIA-PET GSM811037 presented a rather similar global QC score (−43.71) as HiC GSM1055801 even though a very low number of filtered PETs were obtained in this assay (~9 million) and TAD structures are clearly discernible in the connectivity maps (Top left), albeit with lower confidence than in the in situ HiC GSM1551536. In stark contrast to the ChIA-PET the connectivity map of HiC GSM927076 (Bottom left) that was generated with similar number of PETs does not reveal any TAD structures and received from LOGIQA the rather poor global QC score −52.75.

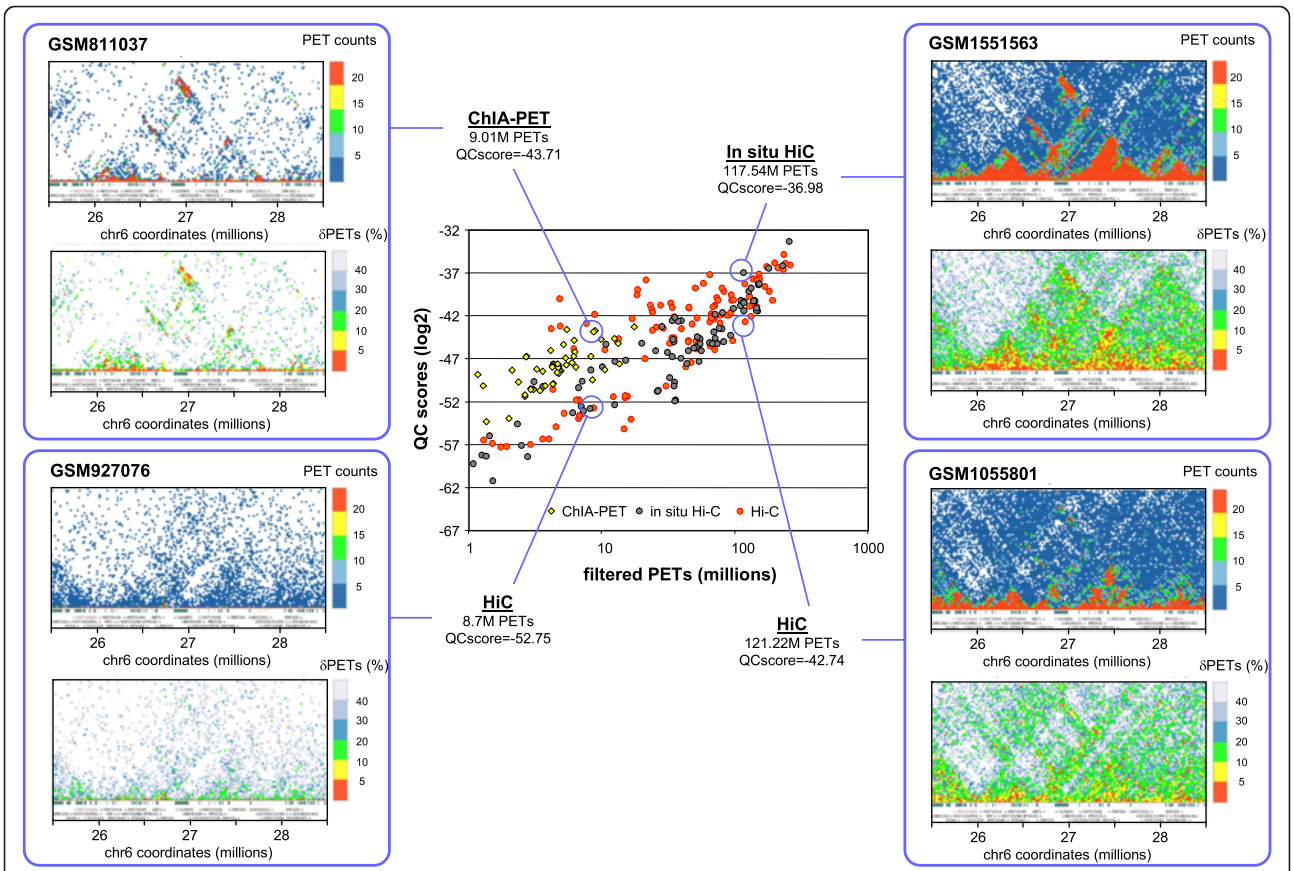


Fig. 5 Comparison of a variety of long-range chromatin interaction datasets in the context of the sequenced paired-end tags (PETs). (*Center*) Scatter-plot illustrating the global quality scores for several long-range chromatin interaction assays in the context of the associated PET counts. (*Left and right panels*) To highlight the power of discrimination provided by global QC scores, the indicated datasets, chosen to represent low (left panels; ~9M PETs - GSM811037 & GSM927076) and high (*right panels*; ~120 PETs – GSM1551563 & GSM1055801) filtered PET count conditions, are illustrated in a local context (the panels show the histone gene cluster on chromosome 6). Local interaction maps generated by LOGIQA are depicted in PET counts (*top*) or PET count dispersion (bottom; %δPETs retrieved after 70 % random PET sub-sampling). Filtered PETs correspond to the number of intra-chromosomal contacts spanning a minimal genome distance of 10 kb

Overall, Fig. 5 clearly illustrates very convincingly the comparative power of the numerical QC scores computed by LOGIQA and their coherence with the visual impression obtained from chromatin contact maps.

While LOGIQA contains also quality indicators for more than 600 4C-seq assays, it is important to note that these values were computed differently. Since 4C-seq assays query all potential genomic interactions associated to a given genomic region - commonly referred to as "bait" - it resembles ChIP-seq assays, in which a target factor is used to define specific sites within the genome. Consequently, we performed quality assessments of 4C-seq similarly as for ChIP-seq assays using the NGS-QC Generator algorithm (for details see [6] or www.ngs-qc.org).

LOGIQA provides a unique web access interface

In contrast to other computational solutions dedicated to visualize HiC or related datasets [10], users do not require to install any software to use LOGIQA. Furthermore, while a few other databases that host publicly available HiC and related assays became recently available [11, 12], LOGIQA is to our knowledge the first database of quality descriptors for a large collection of publicly available datasets. LOGIQA is a fully functional web-based system, which provides to users the quality scores for currently more than 900 publicly available datasets covering mouse, human and drosophila on a variety of long-range chromatin interaction assays. Specifically, global QC scores for all evaluated datasets are

Fig. 6 LOGIQA: A database hosting local and global quality scores assessed of long-range interaction assays. (*Top right*). Scatter-plot illustrating quality scores (y-axis, QC-score) for >160 HiC assays in the context of their associated paired-end tag (PET) counts (x-axis, log10). (*Top left*). Illustration of the LOGIQA query panel. (*Bottom panel*). Screenshot of the visualization tool displaying local chromatin interaction events depicted by their PET counts dispersion levels (heatmap scale)

available in a scatter-plot format relative to their related PET counts, revealing the impact of genome size, sequencing-depth, and technical performance on the robustness and thus, quality of the data sets (Fig. 6 and Additional file 1: Figure S3).

To facilitate the retrieval of datasets, LOGIQA provides a user-friendly query panel covering items like species, type of experiment (e.g. in situ HiC), use of restriction enzyme for chromatin fragmentation, target molecule for ChIA-PET assays, name of (an) author(s), minimal/maximal PET counts to be retrieved, as well as a keyword search for the abstract of the corresponding publication(s).

Finally, LOGIQA provides a dedicated genome viewer, in which users can either select a defined gene (with user-defined upstream and downstream extensions), or provide genome coordinates (Fig. 6 and Additional file 1: Figure S4). The visualisation module displays either local QC dispersion readouts (for 70, 50 or 90 % random sampling conditions) or PET counts. The user can modify in both cases the associated heatmap scale and the genome window resolution (5 or 25 kb windows) (Additional file 1: Figure S5).

Discussion and conclusions

Multiple features, which are at least in part interdependent, affect what can be considered as 'quality' of a long-range chromatin interaction assay. It is obvious that several experimental steps and procedures can be performed under more or less optimal conditions and that this will influence the final dataset. Some of the variables are purely experimental (crosslinking, restriction digest, end repair and biotin labelling in HiC; crosslinking, sonication and IP/antibody quality in ChIA-PET; generation of the sequencing library as well as sequencing coverage); others are bioinformatic (read alignment stringency). In this context, previous studies suggested that quality assessment in chromatin interaction assays could be performed by evaluating the alignment statistics, the frequency of dangling-end or self-circle PETs to reveal potential experimental problems during sample preparation, the levels of duplicated PETs as indicator of library complexity and PCR amplification bias, the fraction of intra over inter-chromosomal interactions and the frequency of long-range versus short-range intra-chromosomal interactions (see also [13]).

LOGIQA provides users with the possibility to retrieve the total PET counts, the fraction of unique PETs and number of intra and inter-chromosomal events. However, these are criteria that are more or less subjective, non-quantitative and non-cumulative; different users may value them differently. For example, while HiC assays may be judged subjectively as 'good' because they contain a high frequency of intra-chromosomal events, the variable ratio of long/short interaction PETs is generally not assessed. The quality assessment of LOGIQA fills this gap by computing the frequency of genomic contacts, which are in addition tested for "robustness" by random sub-sampling.

LOGIQA is based on the concept that we have previously presented for the assessment of quality scores for ChIP-seq and related assays [6]. The use of random sub-sampling of mapped PETs follows the same principle as for mapped reads from ChIP-seq assays. Specifically, this methodology is based on the concept of a "sequencing saturation point", beyond which no new enrichments can be identified [7, 14]. This concept has been initially evaluated in a retrospective manner in ChIP-sequencing assays by assessing the number of significant binding sites retrieved when only a subset of the original sequenced reads is used for profile reconstruction (read random sub-sampling approach; [15]). In a similar manner we have shown empirically that in ChIP-sequencing assays genomic regions with high intensity levels followed a proportional decrease after mapped read sub-sampling [6].

LOGIQA is an independent tool that complements the NGS-QC database with quality score information associated to long-range chromatin interaction assays. In fact, the study of chromatin interactomes is rapidly gaining popularity in scientific community, as revealed by >170 publications indexed in Medline (November 2015) and >500 datasets deposited in GEO. While these numbers are small compared to several thousand ChIP-seq and related datasets, there is an obvious need of establishing quality standards for both types of datasets. Since our first release of the NGS-QC Generator tool in 2013, we have processed more than 30,000 public datasets and we expect to cover virtually all ChIP-seq datasets by 2016. Similarly, LOGIQA will be expanded to cover all available HiC datasets and other type of datasets, like ChIA-PET. Ultimately, we will provide to users a cross-visualisation platform that displays datasets processed by the NGS-QC Generator together with those retrieved by LOGIQA such that users can explore long-range chromatin interaction maps in the context of available ChIP-seq and related datasets. Together, LOGIQA and NGS-QC Generator represent powerful tools for quality-guided exploration of public repositories dedicated to functional genomics datasets.

Availability and requirements

Database availability

LOGIQA is available trough a dedicated web access : www.ngs-qc.org/logiqa.

Ethics approval and consent to participate

Not applicable

Consent for publication
Not applicable

Additional file

Additional file 1: Figure S1. Influence of the short-range PET distance on the assessment of LOGIQA QCscores. **Figure S2.** Global QC scores reproducibility evaluated over multiple PETs' random sub-sampling. **Figure S3.** Global overview of the LOGIQA web application. **Figure S4.** Visualization panel (Interaction map). **Figure S5.** Genome interaction maps for the dataset GSM1551643 assessed at 5kb and 25kb bins resolution. (PDF 1952 kb)

Abbreviations
ChIP-seq: chromatin immunoprecitipation combined with massive parallel sequencing; LOGIQA: Long-range Genome Interactions Quality Assessment; PETs: paired-end tags.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MAMP developed the concept for the assessment of QC scores over long-range chromatin interaction datasets. MB implemented the computational requirements for both datasets processing and web access. VM contributed to the identification of datasets for the validation of the QC concept. P-EC participated in the processing of 4C-seq datasets. HG coordinated the project together with MAMP. MAMP and HG wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
NGS-QC Generator tool/database and LOGIQA were developed in the laboratory of Hinrich Gronemeyer, which is supported by the AVIESAN-ITMO Cancer, the Ligue National Contre le Cancer (HG; Equipe Labellisée) and the Institut National du Cancer (INCa). The support of the Fondation pour la Recherche Médicale (FRM) for the position of a bioinformatics engineer is acknowledged.

Author details
[1]Equipe Labellisée Ligue Contre le Cancer, Illkirch, France. [2]Department of Functional Genomics and Cancer, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Illkirch, France. [3]Centre National de la Recherche Scientifique UMR 7104, Illkirch, France. [4]Institut National de la Santé et de la Recherche Médicale U964, Illkirch, France. [5]University of Strasbourg, Illkirch, France.

Received: 20 January 2016 Accepted: 22 April 2016
Published online: 16 May 2016

References
1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
2. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009;462:58–64.
3. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet. 2011;43:630–8.
4. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: Roadmap for regulation. Nature. 2015;518:314–6.
5. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013;41:D991–995.
6. Mendoza-Parra MA, Van Gool W, Mohamed Saleem MA, Ceschin DG, Gronemeyer H. A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Res. 2013;41:e196.
7. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008;26:1351–9.
8. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58:268–76.
9. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.
10. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
11. Li C, Dong X, Fan H, Wang C, Ding G, Li Y. The 3DGD: a database of genome 3D structure. Bioinformatics. 2014;30:1640–2.
12. Teng L, He B, Wang J, Tan K. 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics. 2015;31:2560–4.
13. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.
14. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10:669–80.
15. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.

PUBLICATION N°4

Chromatin structure dynamics directs cell fate acquisition (in preparation)

Valeriya Malysheva, Marco-Antonio Mendoza-Parra*, Matthias Blum and Hinrich Gronemeyer*

Equipe Labellisée Ligue Contre le Cancer,

Department of Functional Genomics and Cancer,

Institut de Génétique et de Biologie Moléculaire et Cellulaire,

Centre National de la Recherche Scientifique, UMR7104,

Institut National de la Santé et de la Recherche Médicale, U964,

Université de Strasbourg,

Illkirch, France

*Corresponding authors:

Marco Antonio Mendoza-Parra

E-mail: marco@igbmc.fr

Hinrich Gronemeyer

E-mail: hg@igbmc.fr

Phone: +(33) 3 88 65 34 73

Fax: +(33) 3 88 65 34 37

Cell fate transitions are fundamental processes in the ontogeny of multicellular organisms and aberrations can generate pathologies. While cell fate acquisition is a highly complex phenomenon that involves a plethora of intrinsic and extrinsic instructive signals that direct the lineage progression of stem cells, the regulatory circuitry to generate, for example, the early basic architecture and functions of an organ acts rather cell autonomously, as cerebral organoids have been generated *in vitro* from ES or iPS cells[1]. We have previously defined the dynamic gene-regulatory networks underlying endodermal and neuronal differentiation induced by the morphogen all-*trans* retinoic acid (RA)[2]. Here we assessed the contribution of the chromatin interactome[3] to commitment and selective acquisition of these two cell fates. We observed a previously unrecognized highly dynamic re-wiring of chromatin domains during cell differentiation. Long-range chromatin interactions are massively reorganized, erasing up to 95% of the interactome of undifferentiated cells and establishing new interactions already 6 hours after RA treatment. Integration of chromatin interactions together with temporal epigenetic and transcriptomic data indicated key regulatory elements that respond to the initial signal. Our data reveal an enormous capacity of the morphogen to reorganize long-range chromatin interactions as a means to "read" distant epigenetic signals to drive cell fate acquisition and suggest that the differential establishment of chromatin contacts directs the acquisition of the two cell fates.

Examination of the higher-order chromatin structure at sub-chromosomal scale, considering that chromosomes are composed of cell-invariant TADs[4,5], revealed a dramatic global chromatin reorganization in both F9 and P19 stem cells during the first 48 hours of RA-induced cell differentiation (Fig. 1a, b). We noted an increasing number of new TADs at the last time point, while their sizes remained largely constant (Extended Data Fig. 1a). In addition, numerous chromatin structure changes occurred within the domains. In keeping with previous studies[6] we observed that between time points large portions of interactions increase or decrease within stable domains (Fig. 1c, d). This suggests that a subset of TADs undergo concerted, domain–wide rearrangements and/or changes in interaction frequencies as an early response to the morphogen. Interestingly, the comparison of the initial chromatin architectures of F9 and P19 cells indicated large differences in domain structures, suggesting that these cells are lineage-committed already at the non-differentiated state.

Unexpectedly, we observed massive reorganization of long-range chromatin interactions (Fig. 2a) during differentiation of F9 and P19 cells along the endodermal and neuronal lineages, respectively. Several dynamic trends can be seen: initial long-range interactions are almost completely erased and replaced with transient loops after 6 hours of RA treatment and finally new long-range chromatin interactions are established at 48 hours of differentiation (Fig. 2b, c). Interestingly, we remarked a phenomenological trait that during endodermal differentiation the length of interactions connected to genes decreases, while during neuronal-like differentiation P19 cells tend to gain longer interactions (Extended Data Fig. 1b). This global rearrangement is not only due to the change in contact preferences of distinct regions, but also due to appearance of entirely new ones that form long-range interactions (Fig. 2a). Common interactions between F9 and P19 represent the minimum of all observed interactions at any time point (Extended Data Table 1), with the majority of them being erased during differentiation.

In attempt to understand the different RA-induced cell fate acquisitions of F9 and P19 cells, we asked whether the promoters of (key regulatory) genes of F9 and P19 specific programs, decorticated in our previous study[2], can be distinguished by their divergent chromatin connectivity. Surprisingly, the promoters (defined as 3kb around the transcriptional start site, TSS) of genes belonging to the F9-specific program participated in long-range chromatin interactions in both cell lineages; the same was true for the P19-specific program (Fig. 3a). Moreover, in both cell lineages, independently of the program, some of the contacts were formed with regions marked by the presence of RXR, identifying them as direct targets in both cell types. However, for a given promoter the identity of these loops appeared to be strikingly different between F9 and P19 lineages, as shown by the example of *Tal2*, which is specifically expressed only during neuronal differentiation. While the TAD borders in this particular region appear to be rather similar between two cell lineages, the special preferences of *Tal2* promoter connections are evident (Fig. 3b), as the *Tal2* promoter interacted preferentially with upstream regions in F9, while it generated only down-stream interactions in P19 cells. Integration with temporal epigenetic data indicated that in F9 cells ultra-long-range interactions (> 5 Mb) of Tal2 connect to regions that are depleted of open chromatin or marked as repressed (ChromHMM annotation, see Methods), while the opposite was observed in P19, where Tal2 interacts with chromatin regions annotated as open chromatin (Fig. 3c). Long-range chromatin interactions (60 Kb to 5 Mb) reveal a divergence between F9 and P19 cells for the occupancy of the distal anchor sites by RXR and chromatin accessibility. In particular,

for the short-range loops '*a'* and '*b'*, which pre-exist before induction in both in F9 and P19 and both marked as accessible but repressed by H3K27me3 chromatin before treatment, only in P19 cells the distal site anchors in a region marked by RXR binding during the differentiation. No RXR binding is seen in F9 cells. In addition, the P19-specific loop '*f'* is marked by RXR presence, while the RXR is absent in this region in F9 and the loop is not formed. This corroborates the concept of the positive gene regulatory role of chromatin interactions anchored to holo-RAR/RXR-bound enhancers. However, the presence of the RA-induced loop '*d'* in both in F9 and P19 cells, which connects to a region bound by RXR in F9 but not in P19 cells, is not readily explainable with a model where holo-RAR/RXR provides *a priori* positive transcription-regulatory input. Further scrutiny of this anchor site will reveal if the transcription activation domains of the heterodimer are incapacitated, for example by co-binding of another TF or by swapping of the RXR partner[7], or alternatively, if holo-RAR/RXR can act both as repressor and activator of transcription in a locus/loop-specific context. In addition, the increasingly repressive chromatin region where *Tal2* is embedded, may affect the efficiency of a TF activation domain (note the differential abundance of H3K27me3 marks in F9 and 919 in the lower panels of Fig. 3c). Note also that the complexity of this relatively simple interactome of a single gene promoter is further increased by the presence of several loops that anchor at sites of open, accessible chromatin ('*c*', '*h*' and '*i*'); These loops could, in principle, provide additional regulatory input on *Tal2* gene expression via TFs that interact with these regions. Clearly, in addition to validating 'key loops' with higher precision by 3C-related approaches, the impact of these various sites on the RA-dependent regulation of *Tal2* and other key factors[2] needs to be assess by gain and loss-of-function experiments. Experiments using CRISPR–mediated mutation of TF binding sites are ongoing.

One of the major challenges of the present study, as for functional genomics in general, is the meaningful integration of the different types of datasets with the aim of understanding the molecular features of the particular biological system and to predict its response to effectors. Towards this goal we have developed a regulatory network approach that integrates in addition to the classical transcription factor-target gene (TF-TG) relationships, such as CellNet[8], the information derived from TF ChIP-seq data present in the public domain and extracted and quality-graded by the NGS QC approach[9] (www.ngs-qc.org; comprising >41,000 non-selected ChIP-seq data sets) and complemented these data with our FAIRE-seq and HiC information (for details see Methods). Briefly, we match experimentally (ChIP-seq) identified TF and 'open' chromatin (i.e., FAIRE-seq

positive) sites retrieved within anchor regions (hereafter referred to as Genomic Associated Platforms, GAPs) of highly confident (1%FDR) loops emanating from the promoters (3kb around the TSS) of differentially expressed genes (DEGs). Using this approach we could identify long-range chromatin interactions of DEG promoters with GAPs and the potential TF(s) involved in the regulation of DEG genes through chromatin interaction. The resulting reconstructed temporal transcriptional regulatory landscapes of RA-driven neuronal/endodermal cell differentiation comprised a large number of GAPs, which acted as direct mediators to link TFs and cognate DEGs (Fig 4a). This transcriptional regulatory landscape is composed of 19,661 nodes (i.e. genes; TFs; GAPs) and 53,910 interactions, representing (i) RXRa short (less than 10kb around TSS) and long-range interaction events; (ii) TFs associated to GAPs; (iii) GAPs associated to DEGs in both cell differentiation model, as well as (iv) TF-TG associations retrieved as part of the CellNet collection [8].

The ability of this extended GRN (eGRN) to reconstitute the temporal transcriptional regulatory cascade deriving into neuronal/endodermal cell fate acquisition was validated by using of a 'signal propagation strategy'[2] in which each node is evaluated by its capacity to induce (all or parts of) the cell-fate specific programs (Fig. 4b). In this manner, the node corresponding to RXR (denoted as RXR-P19) presented the highest yield for the propagation towards the P19-specific gene regulatory program (84%), in agreement with this nature of master regulator during the RA-driven neuronal cell fate induction. Importantly, this analysis predicted multiple other factors - most of them presented in our previous study[2], but in addition multiple GAPs were shown to present significant yields for driving transcriptional regulatory cascades towards neurogenesis (Fig. 4c). This observation can be explained by their association with major master TFs, supporting a model in which several of them might act in a long-range chromatin interaction manner.

An example of such propagation of initial RA signal through GAPs towards the TFs is illustrated in the subset of the reconstructed network (Fig 4d). Through the long-range chromatin interactions with GAPs, the expression of *Pax6* - one of key TFs in the development of neural tissues (reviewed in [10]) - is activated under RA treatment (Fig. 4e). The propagation of signal continues through CellNet-predicted interaction with *Neurod1*. Neurod1 in turn binds to a GAP on Chromosome 18 according to ChIP-seq datasets imported from the public domain. In turn, the latter GAP interacts through looping with the promoter of the TF-encoding *Zfp516*; note that *Zfp516* RNA is specifically upregulated in P19 cells (Fig.

4d, e). Overall this example shows the connectivity between TFs, GAPs and target genes and illustrates the propagation of the signal in such a eGRN for neuronal differentiation. The validation of key GAPs predicted by the presented above approach using CRISPR–mediated mutation are ongoing.

METHODS SUMMARY

Cell culture. F9 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum (FCS) and 4,5 g/l glucose; P19 cells were grown in DMEM supplemented with 1 g/l glucose, 5% FCS and 5% delipidated FCS. Both media contained with 40 μg/ml Gentamicin. F9 or P19 EC cells were cultured in monolayer on gelatine-coated culture plates (0.1%). For cell differentiation assays, RA was added to plates to a final concentration of 1μM for different exposure times.

Transcriptome and Epigenome assays. The data of transcriptome dynamics and chromatin immunoprecipitation assays used in the current study has been assessed in our previous study[2] and are available from the Gene expression Omnibus database (GSE68291).

HiC. The original HiC protocol has been improved, increasing the ligation yields and modifying the steps that favor chromatin de-crosslinking (see details in Extended Data Methods), while keeping the conventional HiC workflow[11].

Chromatin structure, epigenome and transcriptome integration. We have annotated open-chromatin regions - defined by FAIRE-seq assay - retrieved on the promoters'-associated distal GAPs. Furthermore, FAIRE localization sites were then compared with a comprehensive collection of TF ChIP-seq assays retrieved from the public domain[9]. Note that the TF collection in use in this study includes a large amount of datasets in addition to those provided by the ENCODE consortium, thus representing a comprehensive comparative study regarding not only the number of datasets used but also with respect the diversity cellular systems. Transcriptome, RXR binding sites from ChIP-seq, TF annotations from public datasets and HiC long-range chromatin interactions were integrated and visualized using the Cytoscape platform (version 2.8.3). The signal propagation was performed multiple times using a randomized network as control.

REFERENCES

(1). Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373-379 (2013).
(2). Mendoza-Parra, M. A. *et al.* Reconstructed cell fate-regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis. *Genome Res* (2016).
(3). Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110-1121 (2016).
(4). Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
(5). Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385 (2012).
(6). Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).
(7). Mendoza-Parra, M. A., Walia, M., Sankar, M. & Gronemeyer, H. Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. *Molecular systems biology* **7**, 538 (2011).
(8). Cahan, P. *et al.* CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903-915 (2014).
(9). Mendoza-Parra, M. A., Van Gool, W., Mohamed Saleem, M. A., Ceschin, D. G. & Gronemeyer, H. A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res* **41**, e196 (2013).
(10). Simpson, T. I. & Price, D. J. Pax6; a pleiotropic player in development. *Bioessays* **24**, 1041-1051 (2002).
(11). Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-276 (2012).

FIGURE LEGENDS

Figure 1. Dynamics of chromatin associating domains (TADs) in F9 (a) and P19 (b) with corresponding examples of specific TADs. (c) and (d) show the examples of intra-domain changes in interactions frequencies during the differentiation in case of stable TADs. (e) Comparison of P19 and F9 chromatin domains at non-differentiated state. Yellow arrows point at the differences in domain architecture between different conditions. In (a,b and e) HiC maps show normalized frequencies of interactions. In (c and d) the difference of normalized interactions maps is shown.

Figure 2. Dynamics of long-range chromatin interactions along the differentiation process of F9 and P19 cell lineages. (a) shows the main trends of interactions temporal dynamics. (b) and (c) show in details different dynamics patterns of interactions in F9 and P19, respectively.

Figure 3. Cell type-specific long-range chromatin interactions. (a) Comparison of long-range interactions of genes of common and specific F9 and P19 regulatory programs

after 6 and 48 hours of RA treatment. Direct targets (repressed and induced) indicate on interactions of DEGs with the sites that possess RXR binding signal annotated in F9 and/or P19. Interactions of indirect targets do not possess RXR binding signal, but do change their expression in response to RA treatment. (b) Selective directional preferences of *Tal2* interactions in F9 and P19 cells. HiC map shows normalized interaction frequencies. (c) Integration of long-range (+/- 50kb) and ultra-long range (+/- 5Mb) interactions with corresponding epigenetic landscapes. Precise chromatin states of the distal regions in case of ultra-long range interactions were defined by ChromHMM.

Figure 4. Reconstruction of extended Gene Regulatory Network (eGRN). (a) Schematic representation of the integration principles. (b) Temporal signal propagation model fro evaluation of coherence between the reconstructed eGRN and the temporal gene expression changes. (c) Predicting key GAPs and TFs by signal propagation model initiated at a downstream layers of reconstructed eGRN. Nodes and GAPs are ranked according to their performance in reconstructing the ultimate level of the P19-specific program. GAPs with relatively high yield of reconstruction are marked in red. The performance of reconstruction key regulatory TFs is precised in blue. (d) Subset of the reconstructed eGRN showing the example of signal propagation through the connections of TFs *Pax6*, *Neurod1* and *Zfp516* and GAPs.(e) *Pax6* and *Zfp516* interactome in the epigenetic context.

Extended Data Figure 1. Dynamics of genome interactome. (a) TADs size variations during F9 and P19 differentiation. (b) Changes in length of chromatin interactions during cellular transformation. (c) and (d) Show the proportions of genome involved in interactions, involvement of new interacting GAPs and disappearance of others along the differentiation process. Statistically significant differences has been confirmed by Kolmogorov-Smirnov test, p-value < 0.001

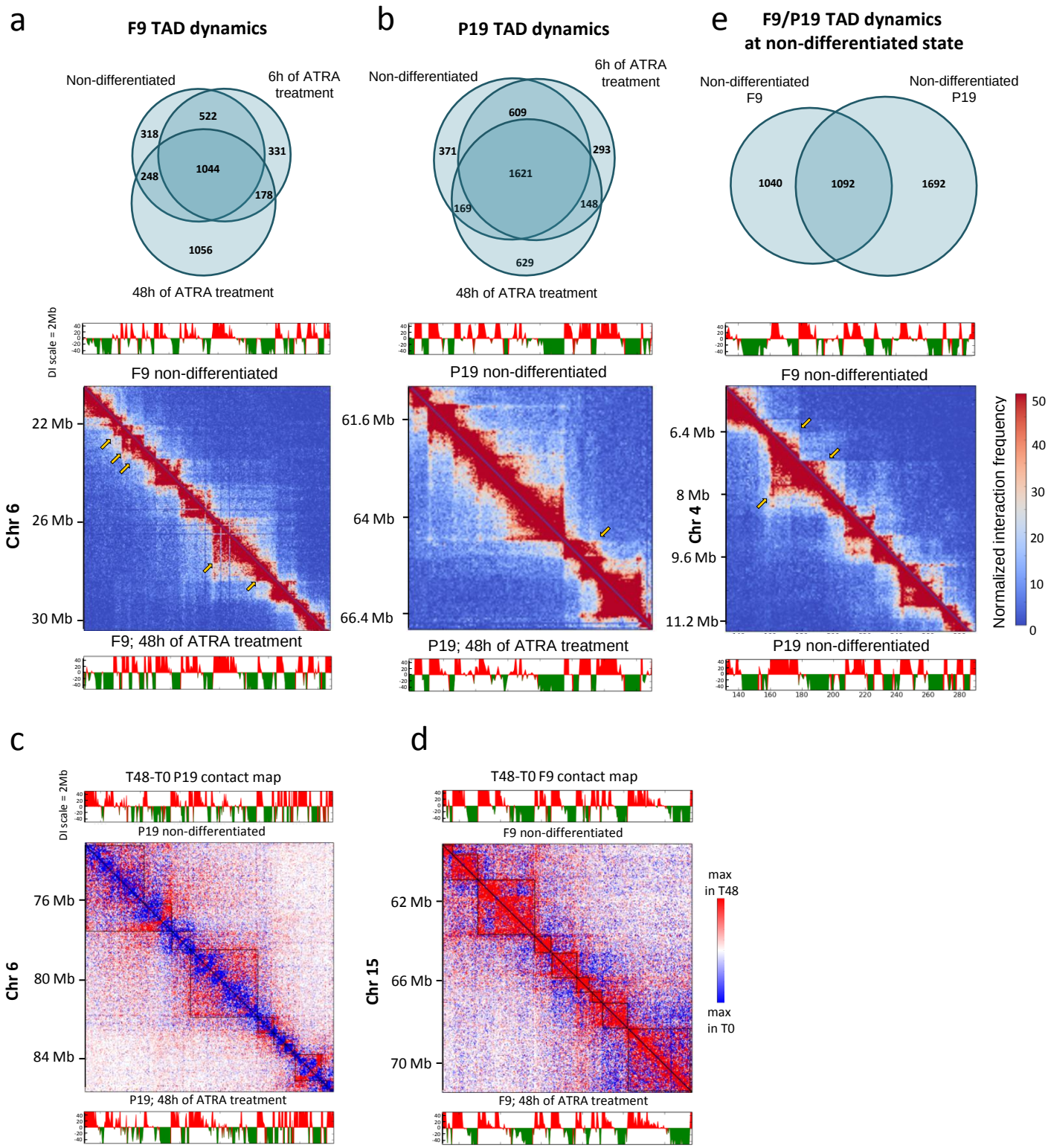

a
F9 TAD dynamics
Non-differentiated
6h of ATRA treatment
48h of ATRA treatment
522
318
331
1044
248
178
1056
b
P19 TAD dynamics
Non-differentiated
6h of ATRA treatment
48h of ATRA treatment
609
371
293
1621
169
148
629
e
F9/P19 TAD dynamics at non-differentiated state
Non-differentiated F9
Non-differentiated P19
1040
1092
1692
DI scale = 2Mb
F9 non-differentiated
P19 non-differentiated
F9 non-differentiated
Chr 6
22 Mb
26 Mb
30 Mb
61.6 Mb
64 Mb
66.4 Mb
Chr 4
6.4 Mb
8 Mb
9.6 Mb
11.2 Mb
F9; 48h of ATRA treatment
P19; 48h of ATRA treatment
P19 non-differentiated
Normalized interaction frequency
50
40
30
20
10
0
c
T48-T0 P19 contact map
DI scale = 2Mb
P19 non-differentiated
Chr 6
76 Mb
80 Mb
84 Mb
P19; 48h of ATRA treatment
d
T48-T0 F9 contact map
F9 non-differentiated
Chr 15
62 Mb
66 Mb
70 Mb
F9; 48h of ATRA treatment
max in T48
max in T0

Figure 1

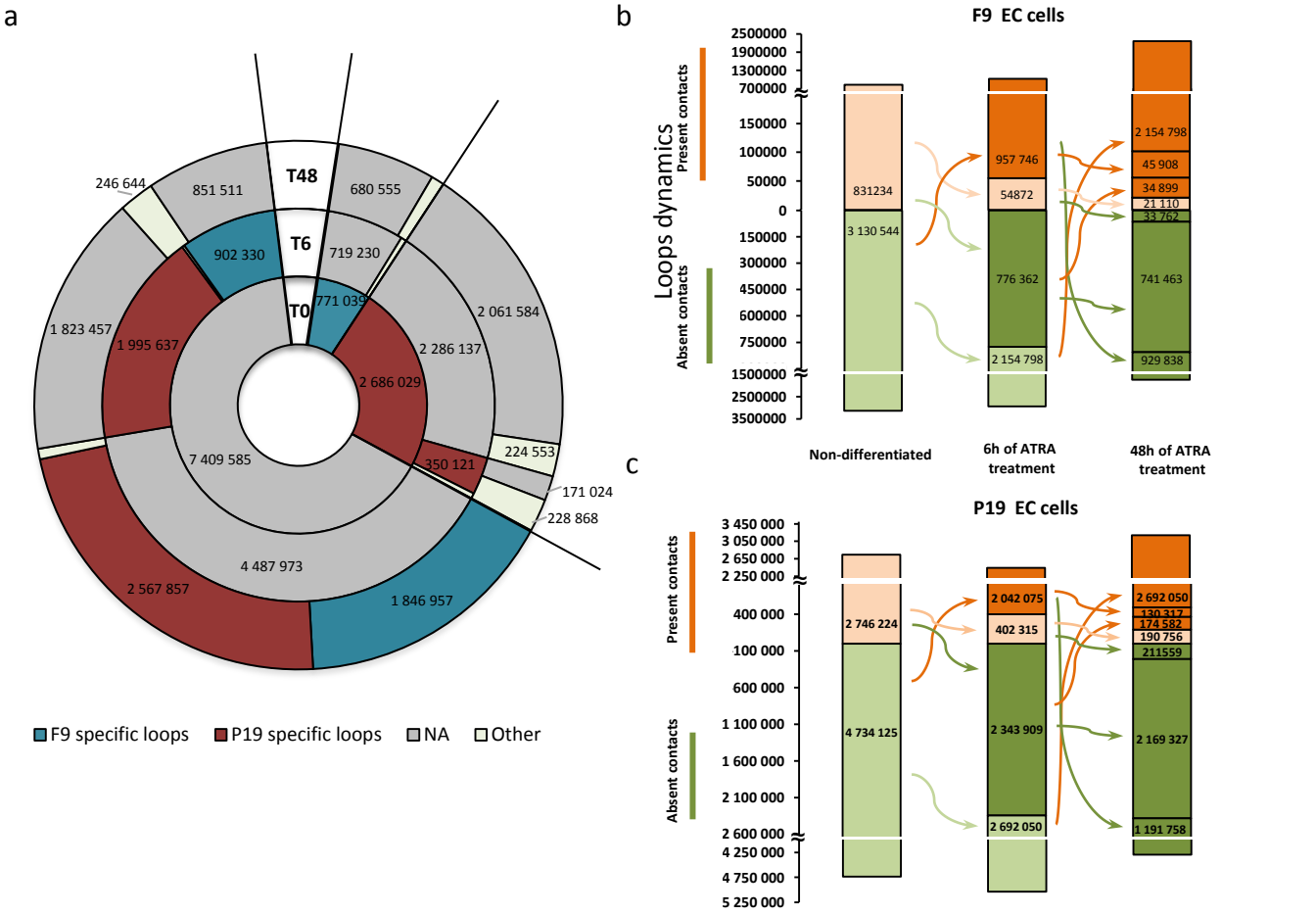

a
T48
T6
T0
246 644
851 511
680 555
902 330
719 230
771 039
1 823 457
2 061 584
1 995 637
2 286 137
2 686 029
7 409 585
224 553
350 121
171 024
228 868
4 487 973
2 567 857
1 846 957
F9 specific loops
P19 specific loops
NA
Other
b
F9 EC cells
Loops dynamics
Present contacts
Absent contacts
2500000
1900000
1300000
700000
150000
100000
50000
0
150000
300000
450000
600000
750000
1500000
2500000
3500000
831234
3 130 544
957 746
54872
776 362
2 154 798
2 154 798
45 908
34 899
21 110
33 762
741 463
929 838
Non-differentiated
6h of ATRA treatment
48h of ATRA treatment
c
P19 EC cells
Present contacts
Absent contacts
3 450 000
3 050 000
2 650 000
2 250 000
400 000
100 000
600 000
1 100 000
1 600 000
2 100 000
2 600 000
4 250 000
4 750 000
5 250 000
2 746 224
4 734 125
2 042 075
402 315
2 343 909
2 692 050
2 692 050
130 317
174 582
190 756
211559
2 169 327
1 191 758

Figure 2

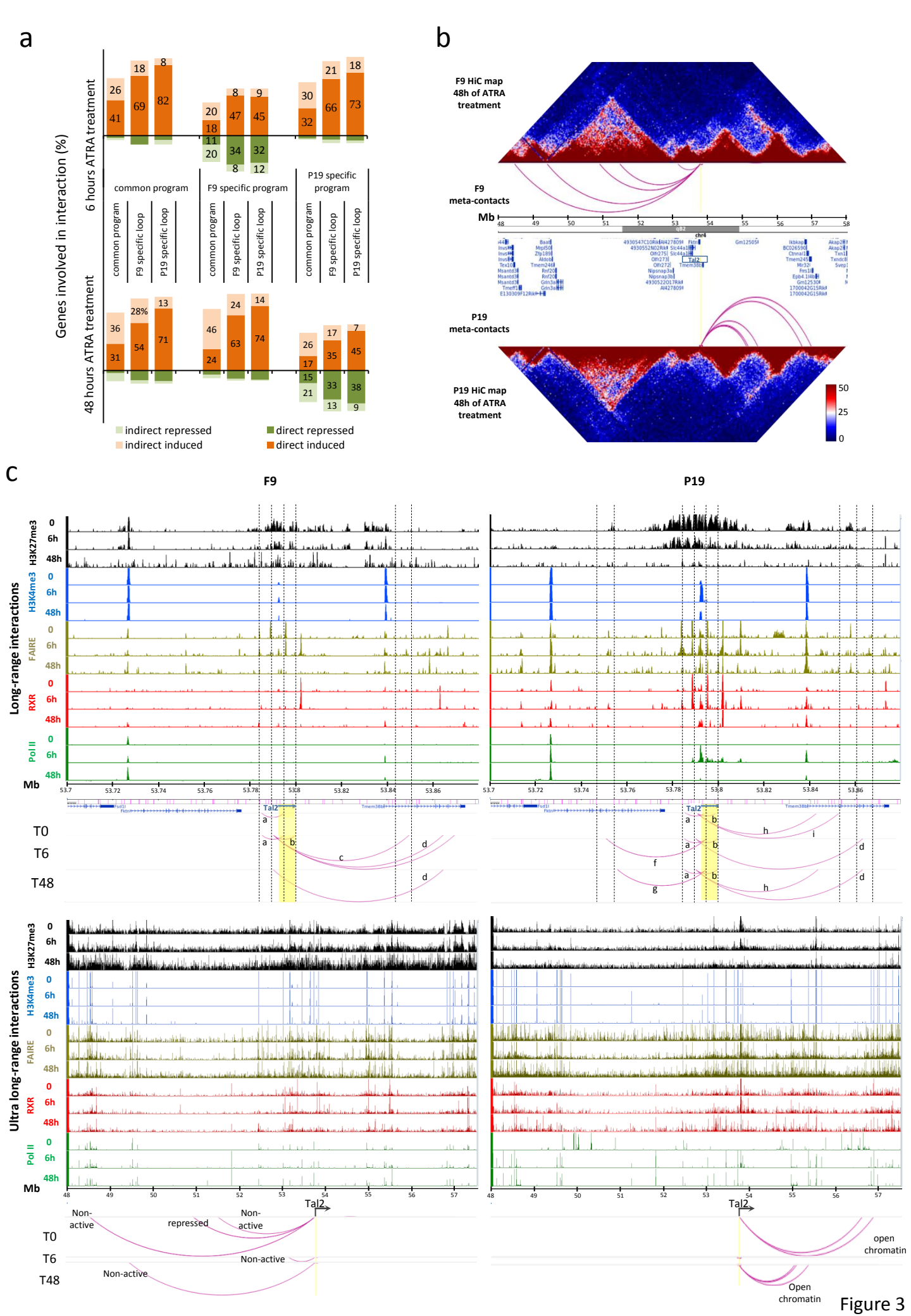

a
Genes involved in interaction (%)
6 hours ATRA treatment
48 hours ATRA treatment
common program
F9 specific program
P19 specific program
common program
F9 specific loop
P19 specific loop
indirect repressed
indirect induced
direct repressed
direct induced
b
F9 HiC map 48h of ATRA treatment
F9 meta-contacts
P19 meta-contacts
P19 HiC map 48h of ATRA treatment
Mb
c
F9
P19
Long-range interactions
Ultra long-range interactions
H3K27me3
H3K4me3
FAIRE
RXR
Pol II
0
6h
48h
Mb
Tal2
T0
T6
T48
Non-active
repressed
open chromatin
Open chromatin

Figure 3

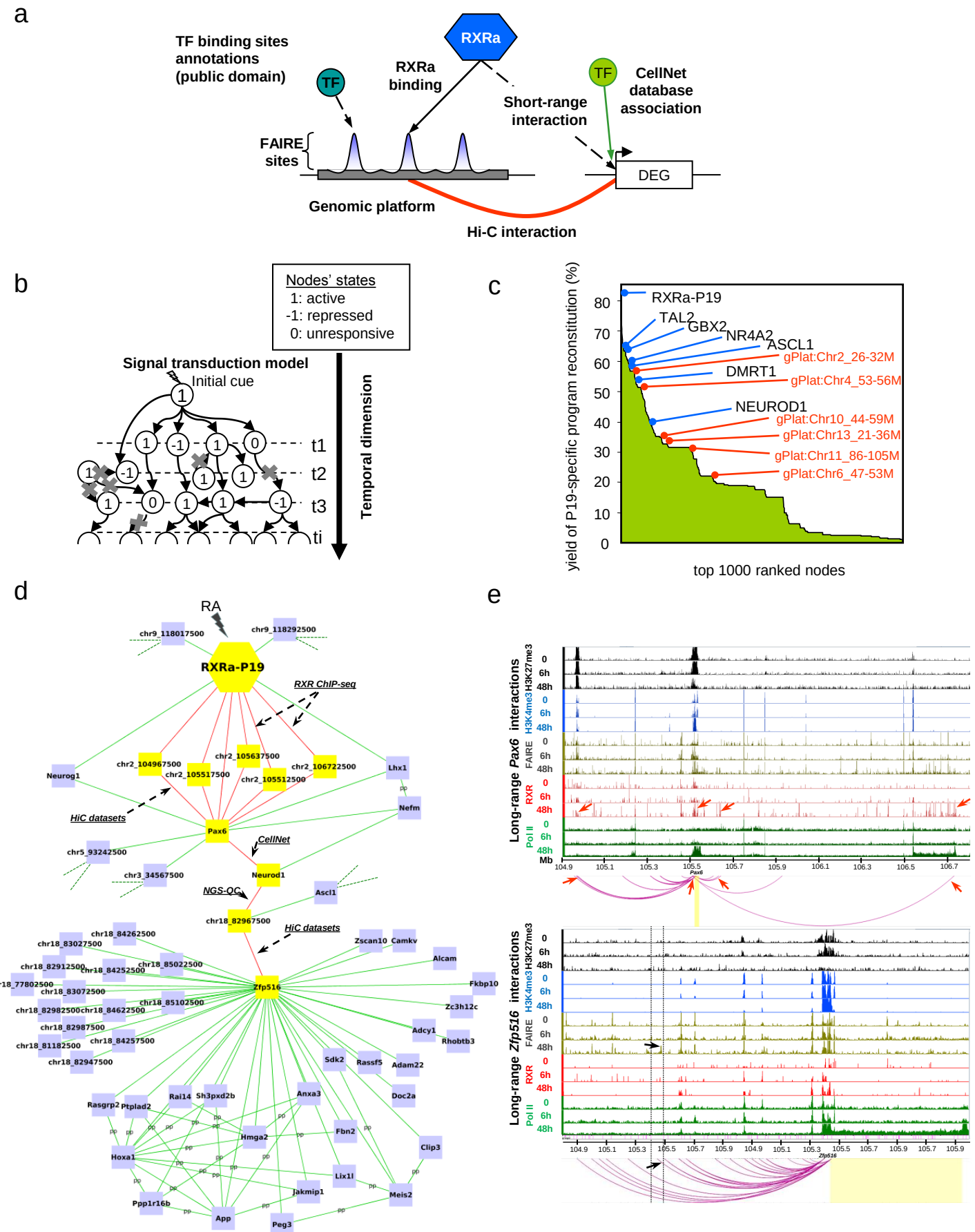
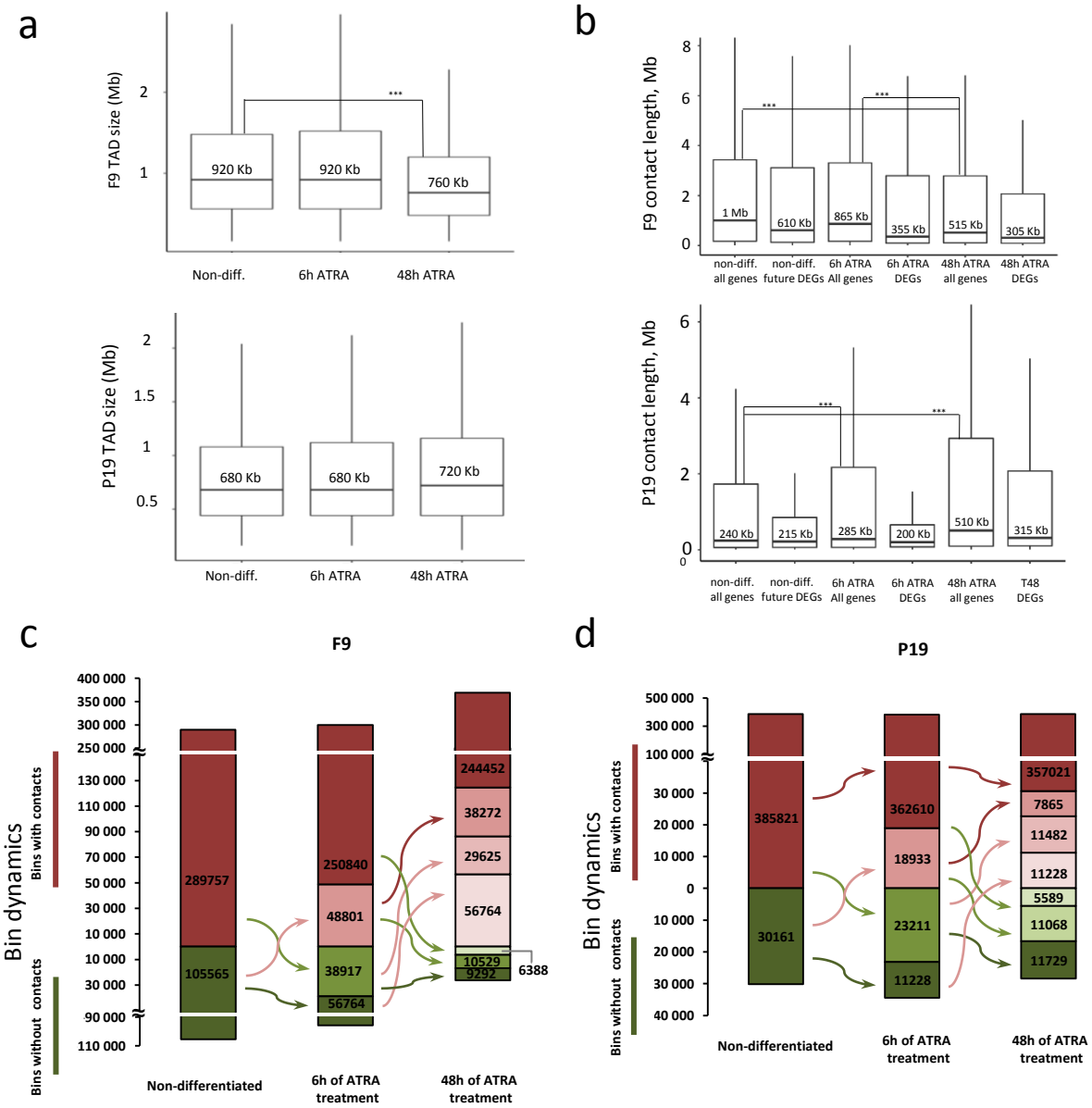

a
TF binding sites annotations (public domain)
RXRa
RXRa binding
TF
Short-range interaction
TF
CellNet database association
FAIRE sites
DEG
Genomic platform
Hi-C interaction
b
Nodes' states
1: active
-1: repressed
0: unresponsive
Signal transduction model
Initial cue
Temporal dimension
t1
t2
t3
ti
c
yield of P19-specific program reconstitution (%)
top 1000 ranked nodes
RXRa-P19
TAL2
GBX2
NR4A2
ASCL1
gPlat:Chr2_26-32M
DMRT1
gPlat:Chr4_53-56M
NEUROD1
gPlat:Chr10_44-59M
gPlat:Chr13_21-36M
gPlat:Chr11_86-105M
gPlat:Chr6_47-53M
d
RA
RXRa-P19
RXR ChIP-seq
HiC datasets
CellNet
NGS-QC
HiC datasets
Pax6
Zfp516
Neurog1
Lhx1
Nefm
Neurod1
Ascl1
Zscan10
Camkv
Alcam
Fkbp10
Zc3h12c
Adcy1
Rhobtb3
Sdk2
Rassf5
Adam22
Doc2a
Anxa3
Fbn2
Clip3
Lix1l
Meis2
Jakmip1
App
Peg3
Ppp1r16b
Hoxa1
Hmga2
Rasgrp2
Ptplad2
Rai14
Sh3pxd2b
e
Long-range Pax6 interactions
Long-range Zfp516 interactions
H3K27me3
H3K4me3
FAIRE
RXR
Pol II
0
6h
48h
Mb
Pax6
Zfp516

Figure 4



Extended Data Figure 1

PUBLICATION N°5

Chromatin dynamics during tumorigenic transformation (in preparation)

Valeriya Malysheva, Matthias Blum, Marco-Antonio Mendoza-Parra and Hinrich Gronemeyer*

Equipe Labellisée Ligue Contre le Cancer,

Department of Functional Genomics and Cancer,

Institut de Génétique et de Biologie Moléculaire et Cellulaire,

Centre National de la Recherche Scientifique, UMR7104,

Institut National de la Santé et de la Recherche Médicale, U964,

Université de Strasbourg,

Illkirch, France

Hinrich Gronemeyer

E-mail: hg@igbmc.fr

Phone: +(33) 3 88 65 34 73

Fax: +(33) 3 88 65 34 37

Running title: Chromatin interactome dynamics during cell transformation

Key words: functional genomics / isogenic stepwise human primary cell transformation / chromatin interactome dynamics / HiC / cell fate transition

ABSTRACT

The evolution of tumors involves alterations at multiple regulatory levels, due to mutations in key factors, such as transcription factors, proto-oncogenes, epigenome/chromatin architecture modulators, or metabolic enzymes/key factors. Investigating the very initial steps of tumorigenesis is hampered by the potential existence of previous tumor clones in clinical samples and the consequences of genome instability in established tumors at diagnosis. Here we have set out to understand the net consequences of cell immortalization and c-Myc protooncogen-induced tumorigenesis on the global chromatin structure of normal primary human cells in a stepwise tumorigenesis model. Our results reveal a dramatic global re-wiring of the chromatin during tumorigenesis.

INTRODUCTION

Cancer is a heterogenic disease, which originates from somatically acquired (sets of) mutations, can be accelerated by genetic predisposition, showing very different latencies and tissue preferences. The systemic changes that can be initiated by single or up to an estimated 6 mutations have been described as hallmarks of cancer [1]. The mutations can affects different layers of gene regulation, generally affecting master regulators, such as transcription factors, regulators of cell proliferation, cell death/survival regulators, regulatory metabolic enzymes, components of signalling pathways, epigenetic modulators or factors shaping cellular structures like the actomyosin skeleton or the 3D architecture of chromatin. At diagnosis, in particular solid tumors often reveal a heavy burden of mutations, some of which are causally related to the tumorigenic events in this particular tumor clone ('drivers') [2], while others ('passengers') accumulate most likely as consequence of acquired genetic instability at later stages of the tumorigenic process. Moreover, several features of the genome structure and dynamics make it particularly prone to the acquisition of genetic aberrations, such as the somatic rearrangement of immunoglobulin genes in leukemia, which bears the risk of the generating oncogenic fusions (e.g. IGH/MYC [3,4]) or the androgen-dependent vulnerability of the TMPRSS2 locus to fuse to *ERG* genes in prostate cancertomlins [5]. This development of complex mutational burden in cancer cells makes it extremely difficult, if not impossible, to investigate the temporal order of events and the effects of the driver mutations on the (de)regulation of the various regulatory signalling platforms in the cell.

To monitor the effects of a minimal amount of steps that transforms a normal human cell to a tumor cell, while remaining genetically stable, stepwise primary cell transformation systems have been created [6]. At early passages these cells remain genetically stable, particularly if compared to established cell lines [7], and the net effects of the genetic events leading to a tumor cell can be studied individually at every step. Using this approach we have recently revealed the altered gene regulatory networks during tumorigenesis and discovered novel deregulated chromatin modulators [7]. Thus, given the inherent limitations of mouse models, this approach –albeit *in vitro*, is the only possibility to monitor in a human system the effects of defined oncogenic events on a particular regulatory platform.

Here we embarked on a study asking for the changes in the chromatin architecture, which may be affected by cell immortalization due to expression of exogenous large and small T from the SV40 early region and the subsequent tumorigenic transformation by the overexpression of the c-*Myc* transcription factor (TF). c-*Myc*, as well as its other family members (N-*Myc*, L-*Myc*, S-*Myc*), all of which act as heterodimers with a number of partners (Max, Mxi, Mad3, Mad4, Mnt/Rox), can be a powerful oncogene if mis-expressed, as it is the case in leukemia due to chromosomal translocations and non-physiological regulation/expression in many other cancers [8]. Myc is an exceptionally pleiotypic TF, as it has been reported to be critically involved in (uncontrolled) cell growth and proliferation [9-11], angiogenesis [12], stem cell renewal, maintenance and differentiation [13-15], genome instability [16] and response to DNA damage [17]. This may be linked to effects of Myc on the global chromatin structure [18] and cancer cells can display very different chromatin interactions at the 8q24 locus harbouring the c-*Myc* gene [19-23].

The 3-dimensional structure of cancer cell chromatin has become an interest of recent research but the focus has been so far on the effect of frequent chromosomal translocations (e.g., *BCR-ABL, MYC-IGH*) or on mutations in key architectural factors, like the subunits of the cohesion complex, which were found in a diverse set of cancers [24]. One of the insights gained from these studies is that the distribution of chromosomal alterations is related to the positioning of these alterations in the 3D chromatin architecture [25]. Only very recently comparative direct global 3D chromatin structure studies between a particular cancer and the normal cells of origin have been reported, as for prostate cancer [26]. Yet, in all these studies normal tissue is compared with very late stages of the tumorigenic evolution, including the development of multiple clonal cancer cell lineages and major chromosomal aberration (i.e., loss/gain of parts of chromosomes/alleles

including LOH, generation double minutes, chromosomal translocations) due to genomic instability.

Here we have defined the changes in chromatin architecture during each of the steps, immortalization and c-Myc overexpression in human BJ fibroblasts, which do not show any of the major consequences of genome instability. We describe the very early alterations in chromatin architecture due to two precisely defined immortalizing and oncogenic insults.

RESULTS

Domain level chromatin dynamics

We examined higher-order chromatin structure at a sub-chromosomal scale and observe that introduction of transforming genetic elements induce extensive chromatin reorganization in each stage of stepwise tumorigenesis (Figure 1A). Though the overall organization of the chromatin in topologically associating domains (TADs) is conserved and TADs exhibit a rather stable size (Figure 1A, B), specific domains are seriously affected such that up to 27% of them are unique in every cell line. Interestingly, the immortalisation by the SV40 early region that expresses the large T antigen which cause inhibition of the p53 and Rb-family of tumor suppressors and the small T antigen which action on the pp2A phosphatase [27] induces the complete reorganization of some chromatin domains which is seen only in immortalized cells, while others are progressivel generated in a stepwise manner along the differentiation process (Figure 1C). It will be interesting to correlate the transcriptional activities of genes in these affected domains and of the known targets of the SV40 early region with the altered chromatin re-wiring at these loci. Moreover, the integration epigenetic information that we have previously reported [7] may reveal altered functional characteristics of these regions.

In addition, numerous chromatin structure changes occur within the stable domains. In agreement with previous studies [28] we observe large portions of interactions to increase or decrease across the stable domain (Figure 1D, E) between different transformation states. Altogether this indicates that immortalizing and oncogenic signals induce concerted domain–wide rearrangements and extensive changes in the chromatin interactome.

Dynamics of long-range chromatin interactions

Even though the global organization of chromatin into TADs is largely maintained, we observe massive reorganization of long-range chromatin interactions within TADs during the stepwise cellular transformation. Of a major importance is the observation that the interactomes of normal, immortalized and *bona fide* cancer cells are almost exclusive with only minor overlaps (Figure 2A). At the same time, immortalized cells are closer to normal cells (10% of shared contacts) while cancer cells interactome is unique with only 2.4% of interactions left from previous stages of transformation. Interestingly, the distance of gene-centric interactions under the MYC overexpression dramatically increased in comparison to the interactions in normal and immortalized cells (Figure 2B). These results indicate a major role of MYC as a regulating factor of chromatin organization.

COMMENTS AND PERSPECTIVES

The capacities of MYC as a transcription factor, capable of inducing such a global reorganization of chromatin is astonishing. However it goes in line with previous numerous studies, showing that MYC acts (among other mechanisms) through regulation of chromatin remodelers [29-32]. In our previous study we described a number of CRMs that were previously unrecognized as the mediators of MYC tumorigenic action [7]. Thus, one could expect that MYC impairs the interactome of normal cells by changing the accessibility of DNA and rewiring large regions of chromatin. However, the scale of such reorganization was previously unknown.

In this respect we are currently integrating chromatin structure data of the current study with our previously described transcriptome and epigenetic landscape (GSE72533) [7] coupled with the analysis of chromatin accessibility (FAIRE-seq) for each step of tumorigenic transformation. This will reveal the mechanisms through which MYC is acting as a global chromatin remodeler inducing the acquisition of aberrant (tumorigenic) cell fate.

METHODS

Cell culture. Primary human diploid BJ foreskin fibroblasts were obtained from the American Type Culture Collection (ATCC). Genetically defined cells of BJ stepwise

system (BJ and BJEL) - were generously provided by Drs. Hahn and Weinberg. BJELM cells were produced previously in our laboratory by retroviral transfection of BJEL cell with pBabe-MYC-ER[46]. Cells were cultured in monolayer conditions in Dulbecco's modified Eagle's medium (DMEM)/M199 (4:1) (with 1 g/l glucose) supplemented with 10% of heat inactivated fetal calf serum (FCS) and gentamicin. The medium for BJEL was supplemented with G-418 (400 μg/μl) and of hygromycin (100 μg/μl). The medium for BJELM was supplemented with G-418 (400 μg/μl), hygromycin (100 μg/μl) and puromycin (0,5 μg/ml) and continuously grown with 10^{-6}M 4-hydroxytamoxyfen (4-OHT).

FAIRE-seq. Isolation of active regulatory elements was performed as previously described [33] using 0.5 mln cells. See details in extended Data Methods. All FAIRE assays were validated using positive and negative controls by quantitative real-time PCR (qPCR, Roche LC480) using Quantitect kit (Qiagen).

Massive parallel sequencing and quality control. qPCR-validated FAIRE assays were quantified (Qubit dsDNA HS kit; Invitrogen); 10ng of the material was used for preparing multiplexed sequencing libraries (Supplemental Methods). Sequence-aligned files were qualified for enrichment using the NGS-QC Generator [34]. Briefly, this methodology computes enrichment quality descriptors discretized in a scale ranging from "AAA" (best) to "DDD" (worst). Based on this quantitative method, all FAIRE datasets described in this study presented quality grades at least "BBB"; integrative studies were thus performed exclusively with high quality datasets.

Enrichment pattern detection and intensity profile normalization. Relevant binding sites in all ChIP-Seq and FAIRE-Seq datasets were identified with MeDiChISeq [35]; multi-profile comparisons were done after quantile normalization [36] (Supplemental Methods).

Transcriptome and Epigenome assays. The data of transcriptome dynamics and chromatin immunoprecitpitation assays used in the current study has been assessed in our previous study [7] and are available from the Gene expression Omnibus database (GSE72533).

HiC. The original HiC protocol has been improved, increasing the ligation yields and modifying the steps that favor chromatin decrosllinking (see details in Extended Data

Methods), while keeping conventional HiC workflow [37]. Per HiC essay 10-20 mln cells has been used.

REFERENCES

(1). Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).
(2). Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724 (2009).
(3). Thangavelu, M. *et al.* Clinical, morphologic, and cytogenetic characteristics of patients with lymphoid malignancies characterized by both t(14;18)(q32;q21) and t(8;14)(q24;q32) or t(8;22)(q24;q11). *Genes Chromosomes Cancer* **2**, 147-158 (1990).
(4). Lee, W. M. The myc family of nuclear proto-oncogenes. *Cancer Treat Res* **47**, 37-71 (1989).
(5). Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648 (2005).
(6). Hahn, W. C. & Weinberg, R. A. Modelling the molecular circuitry of cancer. *Nat Rev Cancer* **2**, 331-341 (2002).
(7). Malysheva, V., Mendoza-Parra, M. A., Saleem, M. A. & Gronemeyer, H. Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis. *Genome Med* **8**, 57 (2016).
(8). Adhikary, S. & Eilers, M. Transcriptional regulation and transformation by Myc proteins. *Nat Rev Mol Cell Biol* **6**, 635-645 (2005).
(9). Johnston, L. A., Prober, D. A., Edgar, B. A., Eisenman, R. N. & Gallant, P. Drosophila myc regulates cellular growth during development. *Cell* **98**, 779-790 (1999).
(10). Trumpp, A. *et al.* c-Myc regulates mammalian body size by controlling cell number but not cell size. *Nature* **414**, 768-773 (2001).
(11). Mateyak, M. K., Obaya, A. J., Adachi, S. & Sedivy, J. M. Phenotypes of c-Myc-deficient rat fibroblasts isolated by targeted homologous recombination. *Cell Growth Differ* **8**, 1039-1048 (1997).
(12). Baudino, T. A. *et al.* c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression. *Genes Dev* **16**, 2530-2543 (2002).
(13). Arnold, I. & Watt, F. M. c-Myc activation in transgenic mouse epidermis results in mobilization of stem cells and differentiation of their progeny. *Curr Biol* **11**, 558-568 (2001).
(14). Frye, M., Gardner, C., Li, E. R., Arnold, I. & Watt, F. M. Evidence that Myc activation depletes the epidermal stem cell compartment by modulating adhesive interactions with the local microenvironment. *Development* **130**, 2793-2808 (2003).
(15). Wilson, A. *et al.* c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev* **18**, 2747-2763 (2004).
(16). Felsher, D. W. & Bishop, J. M. Transient excess of MYC activity can elicit genomic instability and tumorigenesis. *Proc Natl Acad Sci U S A* **96**, 3940-3944 (1999).
(17). Herold, S. *et al.* Negative regulation of the mammalian UV response by Myc through association with Miz-1. *Mol Cell* **10**, 509-521 (2002).
(18). Knoepfler, P. S. *et al.* Myc influences global chromatin structure. *EMBO J* **25**, 2723-2734 (2006).
(19). Jia, L. *et al.* Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet* **5**, e1000597 (2009).
(20). Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A* **107**, 9742-9746 (2010).

(21). Pomerantz, M. M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882-884 (2009).
(22). Du, M. *et al.* Prostate cancer risk locus at 8q24 as a regulatory hub by physical interactions with multiple genomic loci across the genome. *Hum Mol Genet* **24**, 154-166 (2015).
(23). He, H. *et al.* Multiple functional variants in long-range enhancer elements contribute to the risk of SNP rs965513 in thyroid cancer. *Proc Natl Acad Sci U S A* **112**, 6128-6133 (2015).
(24). Corces, M. R. & Corces, V. G. The three-dimensional cancer genome. *Curr Opin Genet Dev* **36**, 1-7 (2016).
(25). Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L. A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nature Biotechnology* **29**, 1109-1113 (2011).
(26). Taberlay, P. C. *et al.* Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res* **26**, 719-731 (2016).
(27). Ahuja, D., Saenz-Robles, M. T. & Pipas, J. M. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. *Oncogene* **24**, 7729-7745 (2005).
(28). Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).
(29). Romero, O. A. *et al.* MAX inactivation in small cell lung cancer disrupts MYC-SWI/SNF programs and is synthetic lethal with BRG1. *Cancer discovery* **4**, 292-303 (2014).
(30). Delmore, J. E. *et al.* BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell* **146**, 904-917 (2011).
(31). Dawson, M. A. *et al.* Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* **478**, 529-533 (2011).
(32). Rickman, D. S. *et al.* Oncogene-mediated alterations in chromatin conformation. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 9083-9088 (2012).
(33). Simon, J. M., Giresi, P. G., Davis, I. J. & Lieb, J. D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* **7**, 256-267 (2012).
(34). Mendoza-Parra, M. A., Van Gool, W., Mohamed Saleem, M. A., Ceschin, D. G. & Gronemeyer, H. A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res* **41**, e196 (2013).
(35). Mendoza-Parra, M. A., Nowicka, M., Van Gool, W. & Gronemeyer, H. Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. *BMC Genomics* **14**, 834 (2013).
(36). Mendoza-Parra, M. A., Sankar, M., Walia, M. & Gronemeyer, H. POLYPHEMUS: R package for comparative analysis of RNA polymerase II ChIP-seq profiles by non-linear normalization. *Nucleic Acids Res* **40**, e30 (2012).
(37). Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-276 (2012).

FIGURE LEGENDS

Figure 1. Dynamics of chromatin associating domains (TADs) in stepwise tumorigenesis. (a) Statistical comparison of unique and common TADs between BJ, BJEL and BJELM cells (b) TADs size stability during the cell transformation. Statistically

significant differences has been confirmed by Kolmogorov-Smirnov test, p-value < 0.001. (c) Corresponding examples of TADs for BJ, BJEL or BJELM TADs. (d) and (e) show the intra-domain changes in interactions frequencies during the differentiation in case of stable TADs. Yellow arrows point at the differences in domain architecture between different conditions. In (c and d) HiC maps show normalized frequencies of interactions. In (e) the difference of normalized interactions maps is shown.

Figure 2. Dynamics of long-range chromatin interactions along the differentiation process of F9 and P19 cell lineages. (a) The main trends of interactions temporal dynamics. (b) Changes in length of chromatin interactions during cellular transformation. Statistically significant differences has been confirmed by Kolmogorov-Smirnov test, p-value < 0.001.

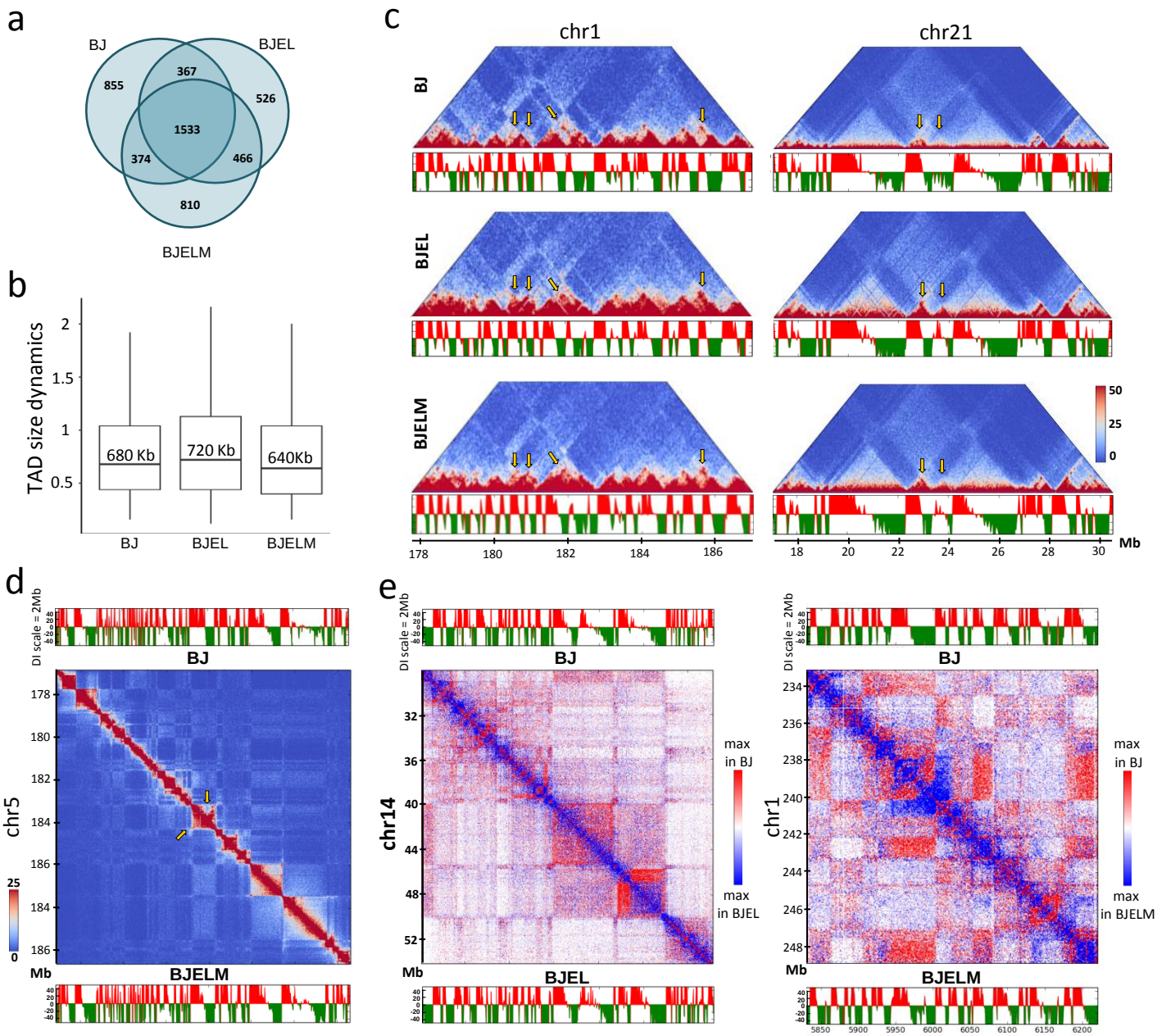

a
BJ
BJEL
BJELM
855
367
526
1533
374
466
810
b
TAD size dynamics
2
1.5
1
0.5
680 Kb
720 Kb
640Kb
BJ
BJEL
BJELM
c
chr1
chr21
BJ
BJEL
BJELM
50
25
0
178
180
182
184
186
18
20
22
24
26
28
30
Mb
d
DI scale = 2Mb
BJ
chr5
178
180
182
184
186
25
0
Mb
BJELM
e
DI scale = 2Mb
BJ
chr14
32
36
40
44
48
52
max
in BJ
max
in BJEL
Mb
BJEL
DI scale = 2Mb
BJ
chr1
234
236
238
240
242
244
246
248
max
in BJ
max
in BJELM
Mb
BJELM
5850
5900
5950
6000
6050
6100
6150
6200

Figure 1

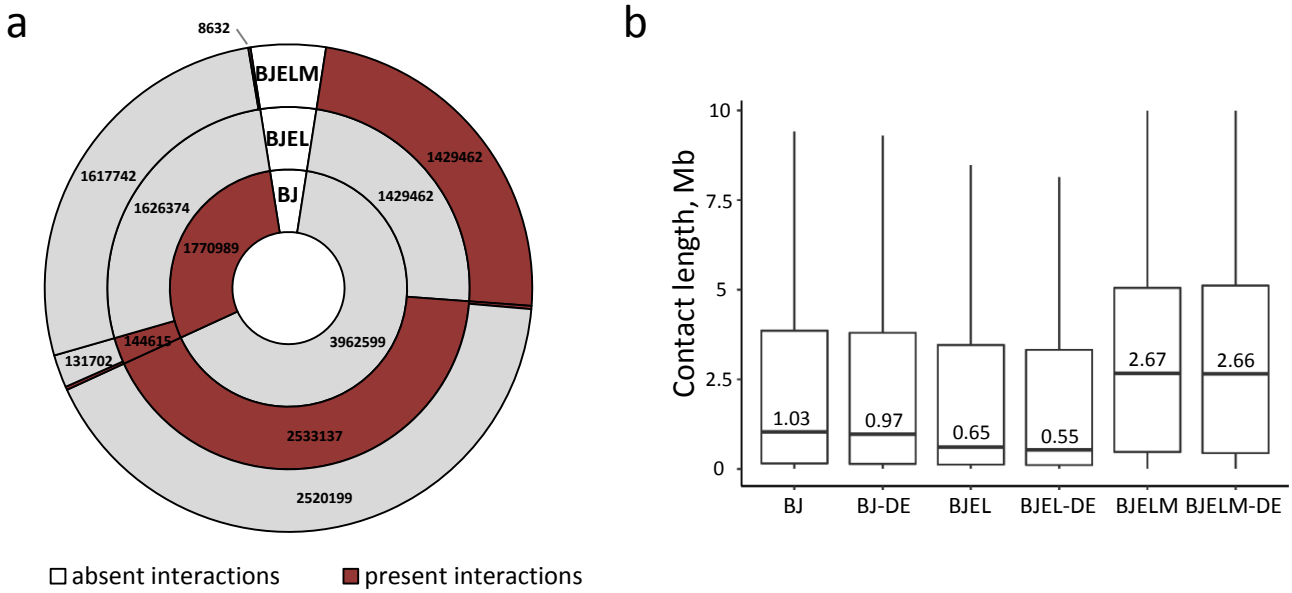

a
8632
BJELM
BJEL
BJ
1617742
1626374
1770989
1429462
1429462
144615
131702
3962599
2533137
2520199
absent interactions
present interactions
b
Contact length, Mb
10
7.5
5
2.5
0
1.03
0.97
0.65
0.55
2.67
2.66
BJ
BJ-DE
BJEL
BJEL-DE
BJELM
BJELM-DE

Figure 2

DISCUSSION

DISCUSSION

FROM SIMPLE SIGNALS TO COMPLEX REGULATORY SYSTEMS

How can a higher organism, like a mammal, acquire the enormous complexity in cell diversity (at least several hundred different cell types), functional cooperation (signal exchanges between organs, like through the (neuro)endocrine system) and compartmentalization (organelles within cells and organs within the organism) upon development from a single fertilized oocyte? At the very beginning, decisions are imposed (e.g., by the mother in flies or gravity in plants) during early embryogenesis, which generate asymmetry. For example, in *Drosophila* it is a maternally transduced signal, the local asymmetric deposition of *bicoid* RNA, which imposes polarity onto the egg. Diffusion of this RNA and its translation generates a Bicoid gradient that is interpreted by other (transiently expressed) transcription factors, the gap genes, and subsequently by the pair-rule genes. This cascade of (sometimes interacting) TFs ultimately defines segment identities within the embryo. Thus, local molecular information provided by the mother is transformed by simple physical means into increasingly complex, temporally organized molecular information that initiates the generation of complex structures.

Obviously this process has from a certain moment on to be temporally coordinated and needs increasingly complex regulatory mechanisms that define cell and organ identities and functionalities, which are established successively early in embryogenesis. These processes involve systems of high molecular complexity and selectivity to properly interpret genetic information. They involve for example pleiotypic actors (i.e., TFs, regulatory RNAs), regulatory mechanisms and the corresponding machineries (i.e., epigenetic modification together with the cognate enzymes, machineries and targeting principles) and structural organization (i.e., ribosomes, chromatin, organelles, organs). All this information is encoded in nuclear and mitochondrial DNA and includes the information for the (self-)organization of these multiple structures.

1. HOW IT BEGINS: INITIATION OF DECISIONS THAT DETERMINE CELL FATES

Transcription factors are believed to be master regulators of cell fate determination/acquisition. However, the fact that some cells have master regulators differentially expressed, while other cells not, raises the question why and how these differences are established, suggesting that there are mechanisms of regulation of master regulators. Together with many other types of studies chromatin conformation capture-based experiments suggest that the evolution of the chromatin modification and architecture has spatially connected (super)enhancers that are likely to have co-evolved with the cognate TFs in common modulable compartments, thus providing a dynamic superstructure of high plasticity and sufficient complexity to organize the expression of GRNs with cell type-relevant genetic information.

While this view provides a scaffold for the temporal organization of cell type-selective GRNs, this does not solve the question of why in some cells a particular regulatory element would be active or repressed [viii], but not in other cells.

As any chemical/biological event has a physical basis that generally leads to energy minimization of the system [ix], one could assume that the differences in TF expression between the cells of an organism should derive *ab initio* from a simple physical cause. These differences have to be established early in embryo development, when for the first time a heterogeneity/asymmetry of the cells comprising the embryo is observed. Assuming that the first commitment is established at this stage, the initial driver signal has already pre-defined different sets of key regulators [x] that will subsequently shape the fate of a (group of) cell(s). But what is (are) the principle(s) linking the physical or chemical signal to the landscape of the regulatory elements that will define the gene regulatory network of the cell?

[viii] or, alternatively, accessible to the cognate TF or not, perhaps due to epigenetic modification. None of all these reflections provides a solution to the "chicken-or-egg" conundrum.

[ix] Energy minimization may have multiple manifestations, concerning, for example, surface energy [135,259], the membrane composition (fluidity, 'rafts') and charges, interactions (covalent, electrostatic, dipolar, hydrogen bonding, hydrophobic, etc.) between the atoms of molecules (in case of protein structure such interactions can lead to protein conformational changes resulting in lower energy states), …

[x] whatever these regulators are in physical or chemical nature

Let's assume that an external signal – a key fate regulator - is generated to specify a particular cell fate. Candidates for such "simple" signals that serve as initial driver could be (the gradient of) a chemical compound around/in the embryo, physical constraints e.g. from surrounding cells, differences in the membrane charge or simply gravity. The orchestration of these signals will afterwards spatially organize embryo cell fates such that they adopt/adapt to "patterns" (see above for the case of *Drosophila* and *bicoid*). Yet, one of the most fundamental questions is: what is the molecular mechanism underlying the formation of these patterns and how are the gene-regulatory programs established and regulated?

The most influential ideas in this field are embedded in the "Reaction-Diffusion" model formulated by Alan Turing [212] and the "Positional Information" model of Lewis Wolpert [213,214]. Turing came up with the intrinsically non-intuitive idea that diffusion itself can create a pattern; he developed a two-equation system, a simple mathematical model of interacting morphogens (short-range activator and long-range inhibitor) that could spontaneously produce a pattern/dissimilarity within a uniform field of cells/nuclei. The formation of palatal ridges is an example of such reaction-diffusion system [215]. The space between ridges is self-controlled by the activator-inhibitor pair of Fgf and Shh.

In contrast to Turing Wolpert didn't try to find a way of self-organization of an organism but rather defined how a complex organism can arise from initial heterogeneity or polarity of a driver signal across the tissue. In this model the morphogen concentrations can act as specific positional coordinates (raising the term "positional information") that differ in space and are interpreted by molecular actors of the cell that sense morphogen concentrations, thus defining its fate. The proof of this concept came with the discovery of the maternally initiated *bicoid* gradient in *Drosophila* [216] that provides distinct inputs to the gap gene network, which in turn converts these smooth spatial differences into more discrete molecular patterns and provides the positional information for the next level of gene regulation (by the segment polarity genes).

While conceptually different, these two main models of early embryo development have often been considered as opposing ideas. However, these models are in fact complementary and work together in several possible cooperation modes, with reaction-diffusion providing self-organized regularity and positional information being a flexible way to interpret regional differences and tune them into proper pattern formation during evolution (reviewed by [217]).

Whereas several examples of cause-consequence relationship between the initial signal and the resulting cellular phenotype are known (i.e., endocrine chemical signals/hormones specifying cell types and cellular responses; TFs driving myogenesis, osteogenesis, adipogenesis or hematopoiesis), we know very little about the establishment and dynamics of gene regulatory networks initiated by external or/and internal signal(s) and which interpret the information provided by the signal. In the context of the current study we have reconstructed two types of gene regulatory networks, one for the process of cell differentiation initiated by the chemical morphogen retinoic acid and a second for a cellular model of tumorigenesis in response to the introduction of defined genetic elements.

2. EPIGENOME

From the breakthrough discoveries and conceptual advances in epigenetics, molecular hallmarks of epigenetic control emerged that are important for cell-type identity, cellular reprogramming and tumorigenesis [218]. One of the key features of chromatin marks is their reversibility, which is an important aspect for the development of epigenetic drugs. In the frame of the current project our results have shown that chromatin undergoes global epigenetic restructuration during the stepwise transformation process and suggested the direct implication of chromatin remodelers in the tumorigenesis. In this respect we discovered several CRMs [xi] that have not been previously associated with tumorigenic cell transformation [26].

However most of the studies are generally showing a descriptive correlation between the transcriptome and epigenome dynamics, keeping the causality question open. Epigenetic modifications could be consequence of signaling pathways (directed by TFs that recruit epigenetic writers or erasers); thus epigenetic modifications would be a step within the signaling cascade, which has the capacity of signal diversification by regulating sets of genes in response to a single trigger. For example, ER signal diversification into gene sub-programs occurs due to the interplay between two ER-recruited epigenetic factors, the histone acetyltransferases CBP and P300 and the methyltransferase PRMT4/CARM1 – the net result is the diversification of estrogen-regulated gene sub-programs [219]. This case of signal diversification highlights a mechanism of complexity generation from a simple single chemical signal – estrogen. Notably this signal itself is the result of a complex chain of events (chemical synthesis – steroidigenesis -

[xi] in particular, PRMT3 and GTF3C4

in specified cells) and uses another principle of diversification, namely the endocrine principle, where a specified organ generates a chemical signal that is transported though the bloodstream and captured by specified proteins (hormone receptors); the endocrine system itself uses chemical synthesis as means of diversification, to ultimately provide highly specific chemical structure-based information to target cells/organs that express the (co-evolved) cognate receptors.

Another question is how the epigenetic information (marks) is selectively targeted to specific sites in the genome. Some epigenetic writers associate "off" and/or "on" the DNA/chromatin with TFs, which recognize specific DNA sequences and transport the chromatin remodelers to their target sites, thus modifying the (surrounding) epigenetic landscape. Indeed, in the presence of agonists some nuclear (*holo*) receptors [xii] recruit CoAs [xiii], which in turn recruit HATs [xiv]. Some non-liganded (*apo*) receptors [xv] can bind CoRs [xvi], which associate with HDACs. However, in most cases – like for the extensively studied PRC2 complex, which deposits H3K27me3 marks through its EZH2 subunit – the mechanism(s) of targeting and the modulation of this process by altered chromatin accessibility and histone modification have remained largely elusive, albeit several options are being discussed [220].

A conceptually different mechanism involves non-coding RNAs, which similarly to TFs could be able to target cargos to sequence specific loci [221,222]. Several groups have reported the involvement of small RNAs in interacting with, and presumably directing of chromatin modifying activities to genomic targets [223,224].

Several of the above mentioned points can now be rather easily addressed by using the CRISPR technology in gain and loss-of-function approaches [xvii].

[xii] e.g., the estrogen receptor (homodimer)
[xiii] like members of the SRC family
[xiv] like CBP or P300
[xv] e.g., the retinoic acid receptor (heterodimer with RXR)
[xvi] like NCoR or SMRT
[xvii] for example, by deletion of a particular epigenetic writer, or by targeting of an eraser or writer to a selective site

3. 3D ORGANIZATION: CAUSE OR EFFECT?

Integrative chromatin structure/transcriptome/epigenome studies have undoubtedly revealed a strong correlation between linear (DNA accessibility, chromatin modifications) and spatial chromatin architecture, modification and gene expression. However, the causality of events remains unclear: the open questions are (i) does the conformational change bring forth the changes in gene expression patterns, or (ii) does the process of RNA transcription change the conformation of the involved loci; (iii) does this "crosstalk" go in both directions or could there be a dominant sequence of events that defines this relationship and finally (iv) what are the factors/features that regulate the chromatin conformation dynamics?

In the case of Polycomb complex (PRC)-mediated repression the epigenetic landscape has been extensively studied [225]. Proteins of this complex regulate stem cell pluripotency [226] by repressing hundreds of genes through the assembly of the PRC1 and PRC2 complexes, which results in the dynamic deposition of H3K27me3 and H2AK119ub1 marks and concomitant chromatin condensation. To determine whether PRC1 components have a causal role in the regulation of gene networks, Schoenfelder et al. [227] performed the KO of RING1A alone and with RING1B in mouse ESCs. While the loss of RING1A weakened the PRC1 network contacts, double KO disrupted the Hox gene network followed by massive de-repression of Hox genes. In contrast, the pluripotency network was not perturbed by ablation of RING1. These data show that the PRC1 complex and in particular RING1 proteins are central to the maintenance of hard-wired target gene networks and chromatin organization can be a cause rather than an effect, at least in the above case.

At the same time there are examples of the opposite situation when transcription affects genome conformation. It is long known that RNA polymerase II acts as a molecular motor and, with the help of other enzymes/factors [xviii] is able to separate DNA strands, displace nucleosomes, and arrange the local chromatin into a more open conformation [228]. These effects are generally not widespread, and are mostly limited to the 'looping out' of chromatin, affecting antisense transcripts and nearby genes, only rarely spreading along entire topological domains. However, inhibition of RNA transcription in a cell by the RNA Polymerase II/III inhibitor α-amanitin,

[xviii] such as the helicase family

RNA Polymerase I inhibitor actinomycin D or CDK*9* inhibitor flavopiridol has profound effects on nuclear structure. Flavopiridol causes disintegration of the nucleolus and triggers a widespread re-localization of several proteins and RNA species, such as the spliceosome complex or the small nucleolar ribonucleoproteins of the dark nucleolar caps [229]. Thus, spatial location truly matters for both genes and regulatory elements. Locally, looping interactions are one of the most important ways to modulate gene activity, through enhancer/silencer elements or co-localization with a transcription factory. On a larger scale topological domains are regulating large sections (hundreds of kilobases or even megabases) of chromatin and potentially function as delimiters of enhancers action landscape.

To reveal the role of chromatin structure dynamics in cell fate decision processes and try to answer the question of causality, we applied a systems biology approach integrating the transcriptome, epigenome and chromatin structure from temporal series of experimental data during early steps of cell differentiation and stepwise cell transformation process.

We observe a previously unrecognized highly dynamic re-wiring of chromatin domains during cell differentiation and tumorigenesis (see Publication N° 5, Malysheva et al. 2016. 'Chromatin dynamics during tumorigenic transformation', manuscript in preparation). Long-range chromatin interactions are massively reorganized. Integration of chromatin interactions together with temporal epigenetic and transcriptome data indicated key regulatory elements that respond to the initial signal. Corresponding validation experiments are ongoing. Our data reveal an enormous capacity of the morphogen to reorganize long-range chromatin interactions as a means to "read" distant epigenetic signals to drive cell fate acquisition and suggest that the differential establishment of chromatin contacts directs the acquisition of the two cell fates (see Publication N° 4, Malysheva et al. 2016. 'Chromatin structure dynamics directs cell fate acquisition', manuscript in preparation).

4. LIMITATIONS OF PROXIMITY LIGATION METHODS

The current research on the regulation of cell fate processes extends largely beyond a mere identification of the involved genes and asks for the annotation of gene-regulatory elements as positive and negative *trans*-regulatory DNA elements, such as (super)enhancers, enhanceosomes, locus control regions or insulators, as the proper association of those elements

to the corresponding genes is a crucial step. In the pre-C era this was done using the linear proximity criteria, using the assumption that an enhancer regulates the most proximal gene, generally using a custom threshold of 10 or 50 kb; if the TF is known one can also establish cumulative distribution functions for differentially regulated genes relative to non-regulated control genes to monitor "preferred" distances for the definition of thresholds. However, ChIA-PET and 3C-based studies showed that this assumption of proximity criteria is very naïve and we know now that enhancer – promoter interactions can span even Mb distances [230,231]. Moreover, it is estimated that hundreds of thousands of enhancers exist in the human genome [92,232], vastly outnumbering protein-encoding genes. Thus, single enhancers can regulate multiple genes and vice-versa one promoter can interact with several enhancers, with estimated number of 4 enhancers per gene per cell type [233]. To add even more complexity, some intergenic enhancers can act as alternative promoters [234] while some enhancers may have a dual function and act as insulator, rendering the distinction of regulatory elements landscape very difficult [235].

Chromosome conformation-based techniques combined with integrative epigenetic studies could clarify this blurry situation. However, these techniques possess a number of limitations, which have to be carefully considered to assure proper association of regulatory elements with their cognate genes. These limitations can be of both, technical and biological origin. Some of the technical limitations originate from the requirement of chemical crosslinking in the C protocols and the effects of crosslinking on the interactomes map, as discussed below.

Limitations originating from crosslinking. Formaldehyde is a zero-length homobifunctional crosslinker [236] that exists in aqueous solution predominantly as methylene glycol [xix] with residual carbonyl formaldehyde [xx]; this equilibrium is pH, concentration and temperature-dependent and long-standing methylene glycol polymerizes to polyoxymethylene glycol; all these factors may affect crosslinking efficiency. In presence of extracts/cells/tissue it reacts with proteins, glytcoproteins, nucleic acids and polysaccharides. The most reactive sites are primary amines (e.g., lysine), purines (in DNA, e.g., cysteine) and the subsequent crosslinking of these functional groups to less reactive groups, such as primary amides (e.g., glutamine, asparagine), guanidine groups (e.g., arginine) and tyrosine ring carbons is a favored process. Depending on fixation

[xix] hydrated formaldehyde, which penetrates tissue and cells rapidly

[xx] which fixes tissues slowly

time and conditions smaller or larger cross-linked protein-DNA complexes can be created; most likely also the local composition and component density (e.g., of proteins and RNAs) affects the crosslinking reaction. While nucleosomes are present at relatively uniform density, structural and regulatory proteins occupy selectively certain regions of the genome, introducing variability in the size of complexes. This in turn influences the range of interaction partners, such that elements in larger complexes are likely to interact with more targets than elements in smaller complexes. One possible way to confront these effects is to sequence individual complexes in a go, and look specifically at the number of DNA fragments, and the types of interactions they establish within and outside of the complex.

Restriction fragment-dependent limitations. Another source of technical limitations originates from the use of restriction enzymes for chromatin digestion. Two characteristics of restriction enzyme are the frequency of cutting and the distribution of restriction sites within the genome; these features will invariably define the theoretically possible resolution of the essay. The more frequent are the restriction sites, the higher is the resolution of the essay. For genome-wide studies the most frequently used enzymes are the 6bp-recognizing HindIII, which generates average fragment sizes of 3kb, and the 4bp-recognizing DpnII or MboI, which cut the DNA on average every 300-400bp. Essentially, the larger the fragment, the less genes and regulatory elements can be resolved. Furthermore, some of the genes may not contain the restriction site and in these cases the enhancer – promoter interactions are impossible to define with the current method. For example, ~5000 of genes in mouse genome do not have HindIII sites and cannot be resolved when using this enzyme. Albeit the resolution of the essay with 4bp-recognizing enzymes is conceptually much better, the high frequency of cutting and thus the larger amount of DNA fragments demands a much higher depth of sequencing, as the number of possible ligation events increases in a non-linear manner. Thus, the final choice of the enzyme is a compromise between the resolution required for a particular study and the affordable depth of sequencing, particularly for the studies conducted with organisms of large genome sizes.

Additional biological limitations are imposed by the intrinsic characteristics of gene structures. Indeed, irrespectively of the method used for fragmentation of crosslinked chromatin - with restriction enzymes or by physical means, like sonication - genes that share promoters, overlap on different DNA strands or are juxtaposed cannot be resolved, neither with 3C-based methods

nor by ChIP-seq. Altogether these factors constrain the complexity of the library and must be taken into consideration during data analysis.

Limitations due to chromatin dynamics. Furthermore, due to the dynamic nature of chromatin and the fact that at single cell level transcriptional activity occurs in bursts, regulatory interactions may exist only in a small percentage (~2-3%) of a given cell population, thus making them difficult to detect by analysis of chromatin conformation data alone. Indeed, in "bulk C essays" the regulatory signal present in a minority of cells will be diluted by the absence of that signal in other cells and may remain undetected. These considerations may explain the discrepancy between the results obtained from single-cell microscopy-based experiments and the interaction frequencies detected by molecular assays. With the development of a single-cell Hi-C it is now possible to determine how much of this discrepancy is technology-dependent and how much single cell assays resolve (stochastic, programmed) cell-to-cell variability that is not seen in cell populations.

5. CHALLENGES IN GENE REGULATORY NETWORK RECONSTRUCTION

One of the approaches of network reconstruction is the integration of information that can be obtained from (various types of) interaction databases (CellNet, MiMI, etc.). This way it is possible to pinpoint the interacting partners/factors of genes/proteins of interest, e.g. as differentially expressed genes, and then to define high-degree nodes in this network. With this approach a more comprehensive network can be established than when restricting the analysis to only those interactions that occur between query nodes. This way the network gets 'enriched' with potentially functionally relevant information, which facilitates the definition of relevant sub-networks and/or non-differentially expressed nodes that are topologically important in the network but would otherwise not be identified. This being said, such integration of external information necessitates validation, which can be done by a 'signal propagation' test [207] and system perturbation (e.g., CRISPR-based gain or loss-of-function or similar).

However, identification of hubs in these networks can be biased towards favoring nodes that are in general highly connected. This may concern promiscuous, ubiquitous or well-studied nodes, as nodes with many interactions in the query database have a higher probability of being included in the network. A more targeted analysis is needed to determine key regulatory nodes in

the network. The 'signal propagation' approach that we developed recently [207] is particularly useful, as it verifies the correct flux of information from the initial signal to the final pattern of target nodes. Indeed, this procedure 'cleans' the network from potential artefactual interactions, which do occur in the context of other cells/conditions but are irrelevant for the actual study. Moreover, one of the salient features of the 'signal propagation' concept is that it ranks nodes according to their ability to generate the final pattern of nodes or, in other words, it identifies the key nodes/master regulators within the network.

A limitation of the use of the graph theory for the analysis of biochemical networks is the static feature of graphs. All real biological networks are dynamic, as the activity of nodes and their links within the biochemical networks change over time. Thus the abstraction to graphs can mask temporal aspects in the flux of information. Moreover, static sub-networks are not sensitive to the amount of initial substrates or enzymes or, for example, to the gradient of a chemical signal, while real systems respond to these quantitative parameters. Thus static network generalizes the outcome assuming a certain threshold of presence/absence of gene/protein that oversimplifies the fine-tuning in the propagation of certain types of signals [xxi]. Nevertheless, while static graph representation of a system is a prerequisite for building detailed networks [237], dynamic modeling approaches (*e.g.*, Petri nets[238]) can be used to simulate network dynamics; to establish such dynamic networks the graph representations can serve as useful skeletons of the model. As modeling of the dynamics of biochemical networks recapitulates more accurately ***in silico*** the dynamic features of a biological system, it will be more valuable for developing quantitative hypotheses.

However, the challenge with building dynamic models of biochemical/biological networks is that they require the integration of kinetic and quantitative parameters, which are difficult to obtain experimentally and are frequently not available/reported. Another obstacle is the computational resources necessary for dynamic analyses, as time and memory requirements for computation increase exponentially with the number of steps in a path and/or the number of nodes in a graph. This computational bottleneck discourages most laboratories from calculating the static and dynamic properties of large regulatory biochemical networks. One of the ways to overcome this challenge is sampling [239] and/or parallelization of algorithms [240]. Another

[xxi] i.e., the response of the cell to the *bicoid* gradient will *a priori* not be reflected in a static network

approach involves the abstraction from exact chemical constants and factors and evaluate the signal propagation efficiency using a Boolean concept [241]. Finally, a comparative analysis with the temporal transcriptome changes may significantly improve the identification of key regulatory factors.

In summary, we are just starting to decipher the rules of the dynamics of complex biochemical systems and to develop concepts for *in silico* modeling, in which the graph theory plays an important role for organizing the accumulated knowledge. GRN reconstructions have proven their utility (i) for providing an overview of the organization of different types of biochemical networks across species, (ii) for the analysis of multivariate data when lists of genes or proteins can be placed in the context of prior knowledge, (iii) for the development of hypotheses about the cooperation of multiple factors, including to generate complex phenotypes and (iv) for the identification of 'master regulators' involved in the investigated biological processes.

6. THOUGHTS ABOUT THE BEGINNING, NON-EQUALITY AND DIVERSIFICATION

Reflecting on the above described multiplicity of interconnected mechanisms [xxii] and large number of different types of actors [xxiii] that are modulating regulatory systems at various levels [xxiv], the question arises if there could be a common unifying principle that defines the initiation and diversification of cellular systems in higher organism. Indeed, one of the most profound questions in nature [is] – how complexity arises from initial simplicity [242], or in other words, how the complex structures of an adult derive from a (simple) fertilized egg.

It appears that nature has invented different pathways to generate this complexity. As discussed above, in *Drosophila* the maternally transduced *Bcd* RNA establishes a gradient of the morphogen and TF Bcd that defines the anterior pole of the embryo and controls transcription of target genes in a concentration-dependent manner. Another type of gradients, like the differential nuclear localization of the TF Dorsal, defines a dorsal-ventral axis and thus, additional positional

[xxii] information transfer and diversification through signaling involving a multitude of levels like the metabolomes, transcriptomes, epigenomes, chromatin interactomes, etc.

[xxiii] chemical signals like hormones, TFs, epigenetic modulators, actors that spatially organize chromatin and its dynamics

[xxiv] transcription and translation, epigenetics, diverse types of ncRNA functions, metabolism, catabolism, etc.

information. These TFs gradients are interpreted by target genes that execute the positional information resulting in increasingly complex segmentation, which is then complemented by the information exchanged between adjacent (non-identical) cells to shape the final pattern. Taken together, in this species a maternally defined physical gradient provides (positional) information to make cells non-identical.

In mice no morphogen gradient has been described to operate at the very early times of embryogenesis. Moreover, the first unsolved question is why a fertilized egg starts to divide [xxv]. Most likely, the actors and mechanisms involved in cell division are maintained operative in the egg and activated upon formation of the zygote. In contrast to *Drosophila*, where 13 rounds of extremely rapid chromosome duplications and mitosis occur in nuclei in the absence of cytokinesis/cellularization[xxvi], mouse zygotes undergo 4 cell divisions (**Figure 7**), which occur without significant cell growth (leading to smaller cells); the first cell division occurs 4h to 10h post fertilization, is twice as long [xxvii] as the second and starts in the male pronucleus. Zygotic gene activity starts in the long G2 phase [xxviii]of the second division. The resulting 16 identical cells are developing under physical laws, particularly that of surface energy minimization, which results in the formation of a ball-like structure. During the 5th division, two cell populations are formed, polarized external cells (giving rise to trophectoderm) and apolar internal cells that will give rise to the inner cell mass and then segregate into the epiblast and primitive endoderm. Upon implantation in the uterus (most likely receiving maternal signals) the embryo undergoes gastrulation during which the three embryonic layers are committed and organized in three dimensions. Taken this information together, physical laws [xxix] govern the first rounds of cell division to generate a ball of identical cells, with the 5th division introducing non-identity. The fact that this coincides with implantation infers instructive signals from the mother; this is conceptually similar as the maternal information instructing the *Drosophila* embryo.

Taking everything together, different organisms have developed different ways to generate complexity from the initial simplicity of the fertilized egg. However, it appears that – at least in

[xxv] this questions bears resemblance to the question concerning the initiation of the Big Bang
[xxvi] This syncytium contains thousands of nuclei
[xxvii] 120 min
[xxviii] 12 to 16 h
[xxix] in particular minimization of surface energy

invertebrates - instructive maternal signals are required to generate initial non-equality of cells. In the mouse an alternative possibility to generate non-identical cells is that energy minimization of surface tension at the 8-cell stage is sufficient for (stochastic) polarization [xxx] of some cells, which then divide to generate two outer polar cells, while the less superficial cells divide to give one outer and one inner cell, thus leading to a compartmentalization that forms the blastocyst. Once non-identity of cells and cell communication are established additional cell autonomous and non-autonomous actions can be established resulting in further diversification and pattern formation of the organism. This may involve both physical [xxxi] and morphogenetic [xxxii] mechanisms [243] but also self-organizing principles, such as the (re)formation of TADs during mitosis after the S phase.

In this context it is worth noting the insight of Alan Turing in how "… a system, although originally it may be quite homogenous [xxxiii], may later develop a pattern or structure due to an instability of the homogeneous equilibrium, which is triggered off by random disturbances [xxxiv] ,, and that "this theory does not make any new hypotheses; it merely suggests that certain well-known physical laws are sufficient to account for many of the facts" [212]. He even suggested that hormones may be morphogens – a notion which is firmly justified by the fact that all-*trans* retinoic acid is now regarded as a morphogen for limb formation and possible other morphogenetic processes.

It will be interesting to consider cell or chromatin states or cell patterns from the point of physical laws, as extrapolated from Turing's reflections. For him these entities are "stable" but some (stochastic) deviations from this stability are essential to drive the entity into another stable state, much so in a Waddington sense [245], thus predicting that the process of cell fate acquisition involves an (induced) instability that leads to a new stable (lower energy) equilibrium of the cell.

[xxx] to maximize intercellular contact as a consequence of physical compaction, which initiation the formation of cell junctions

[xxxi] cell responses due to mechanosensing or shear forces, the letter of which contribute to tissue organization during cardiovascular development

[xxxii] diffusion-reaction model of Turing for *Drosophila* embryogenesis but also for limb development triggered by the "zone of polarizing activity (ZPA)" which generates a directionally instructive gradient of the morphogen retinoic acid

[xxxiii] like the syncytium of the *Drosophila* embryo or the 8-cell stage of a mouse embryo

[xxxiv] like the stochastic polarization of some mouse embryo cells at the 8-cell stage

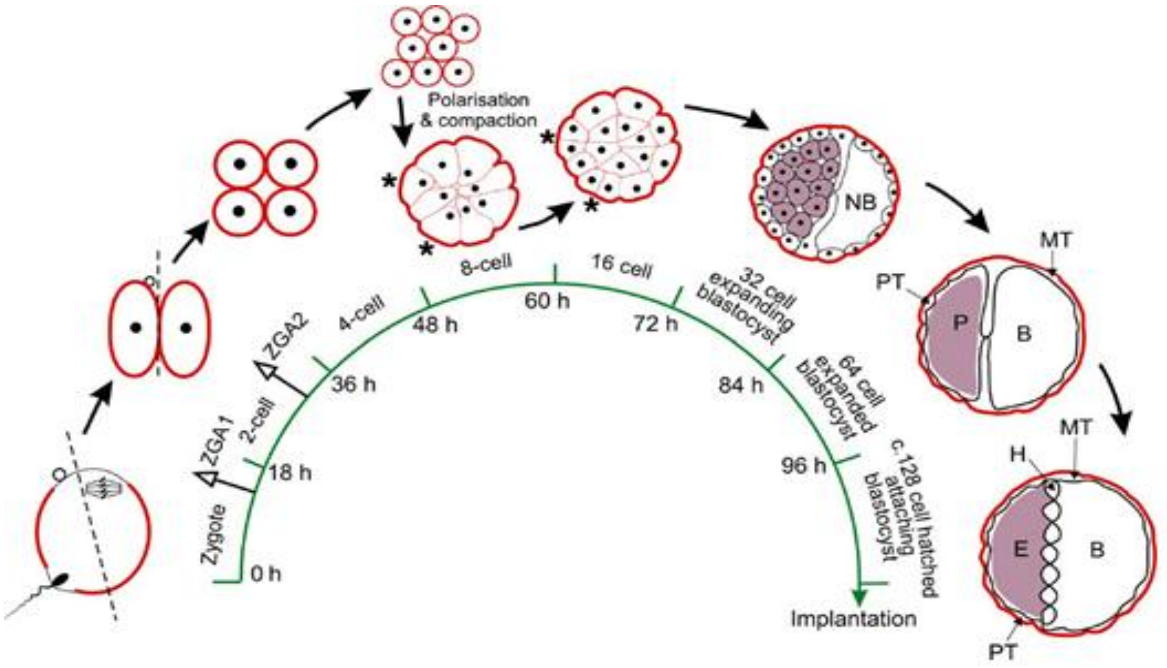


Figure 7. **Schematic outline of some key events during pre-implantation mouse development.**

During the 8-cell stage, individual cells polarise, maximise intercellular contact (compaction) and initiate junctional formation. Some polarised 8-cells (*) have "more superficial positions than others and divide conservatively to generate two outer polar cells at the 16-cell stage, whereas the less superficial 8-cells divide to give one outer polar and one inner non-polar cell. After a further round of cell divisions to the 32-cell stage, formation of the nascent blastocoel (NB) occurs and the inner cells (purple) are separated from the nascent blastocoel by trophoblastic processes. With full expansion of the blastocoel (B), division to the 64-cell give two committed tissue lineages: inner pluriblast (P, purple) and outer trophoblast, which is designated mural (MT) adjacent to blastocoel and polar (PT) over the pluriblastic cells that make up the inner cell mass (ICM). Over the next 12–24 h the embryo's cells continue to increase in number, the blastocoel expands further, the trophoblastic processes overlying the ICM withdraw, and a layer of hypoblast (H) cells derived from the pluriblast cells is evident on the surface of the ICM. The remaining cells of the ICM are now called epiblast (E). The embryo sheds its outer acellular coating (the zona pellucida—not shown) and initiates attachment to the uterine epithelium. [Taken from [244]]

Identifying the causes and sensitive targets of such a destabilization [xxxv] could facilitate changing cell fates, as we have demonstrated in the case of *trans*-differentiation of F9 cells towards the neuronal lineage [207]. Thus as the cell fate appears to be flexible (either in natural conditions as in case of newt's eye [9,10] or experimentally forced) and sensitive to the environment the cell fate *sensu stricto* doesn't exist as the cell development not only directed by internal (epi)genomic information but also tuned by external stimuli.

[xxxv] Probably hormones, TFs, epigenetic modulators, and similar actors are potential sources of such destabilization

PERSPECTIVES AND CONCLUSIONS

Despite all the efforts in developmental biology the question of embryo development and cell lineage establishment is open. Why, in response to what stimulus and how can different parts of an early embryo start to acquire their specific traits? How is the chromatin architecture set up during early embryogenesis and what drives this process; is it a slow process or is it already fully established [xxxvi] or is this a completely dynamic process; if the latter is the case what are the instructive signals? How this information is transmitted to the cell progeny? What is the memory capacity of the chromatin organization and whether it acts in concert with epigenetic memory? Single cell HiC, CHiC and HiChIP should help us to understand the mechanism of these processes.

RNA world. Additional layers of information can be added to get closer to a holistic analysis of a cell. This includes RNA analyses (miR expression and other regulatory ncRNAs and their targets) in the data integration effort. It is important also to reveal whether and how different types of RNA (e.g. lncRNA, eRNA) instruct the structuring of chromatin [246].

Metabolomics. The metabolomics should not be ignored in the integrative studies as metabolites can act as co-factors of regulators, like in the case of iron- and α-ketoglutarate (α-KG) dependent JmJC-domain containing proteins. Indeed, in gliomas tumor-derived IDH1 and IDH2 mutations reduce α-KG and accumulate an α-KG antagonist, 2-hydroxyglutarate, leading to genome-wide histone and DNA methylation alterations [247–249]. Moreover, succinate dehydrogenase mutations in paragangliomas result in a hypermethylator phenotype, associated with downregulation of key genes involved in neuroendocrine differentiation [250]. In addition, succinate accumulation in SDH-deficient mouse chromaffin cells leads to DNA hypermethylation by inhibition of 2-OG-dependent histone and DNA demethylases establishing a migratory phenotype [250]. These results reveal the interplay between the Krebs cycle, epigenomic changes, and cancer and argue for the need of data integration, which should be as comprehensive as possible.

De novo genome assembly and haplotype phasing. Finally, C-type data could have applications beyond genome structure studies. There have been promising studies using long-range chromatin interactions from HiC datasets for the purpose of de novo genome assembly and haplotype

[xxxvi] e.g., as a scaffolding matrix – the Mb TADs – at the 4 cell stage and then gets refined

phasing [251–255]. The main concept in these approaches is to link contigs by Hi-C contacts in order to assemble the scaffolds of entire chromosomes in the genome. This approach is less expensive, as it involves high-throughput short read sequencing, which is more affordable than expensive long read technologies, such as traditional Sanger sequencing, PacBio real time single molecule [256] or recently developed nanopore sequencing [257]. The automation of the assembly process is a necessary step [254] which would enable a higher number of species genome to be assembled at high quality.

Identifying the functionality of disease associated SNPs. Knowing the spatial organization of chromatin in a large number of different cell types would also enable a better understanding of the regulatory role(s) of non-coding regulatory DNA elements, associated with important traits or diseases in Genome-Wide Association Studies (GWAS). To reveal the full interaction landscape of regulatory elements and distal SNPs one could imagine to apply the capture-HiC from another angle and instead of using a promoter-centric approach, use the regulatory elements themselves as baits. That would allow monitoring the entire spectrum of regulatory interactions of any particular region; in addition to unraveling the functionality of disease-associated SNPs, this would be particularly attractive for structure-function studies on super-enhancers, enhanceosomes or locus-control regions.

To give a comprehensive response to mentioned above questions we need to improve the existing methods of analysis as well as the quality evaluation of the produced datasets as a necessary key step in integrative studies. As very little attention is devoted to the quality of the chromatin (epigenetic) structure datasets generated, we invested in the development of a method for their quality assessment [105]. However further effort should be done to make this tool accessible for the quality evaluation of the user-generated data sets in real time.

Qualitative and quantitative technological improvements will facilitate integrative studies. With the progress in DNA sequencing technologies and advances in molecular biology, C-based experiments will undergo quantitative and qualitative improvements. In this respect we expect an increase in the number of cell types, tissues and organisms for which the chromatin interactomes will be established to relate common and cell/developmental stage/disease-specific architectural chromatin features with gene-regulatory events. Functional insight will particularly come from comparative temporal analyses of various (patho)physiological processes (stress response, cell

cycle, embryo development, tumorigenesis), conditions (wild types vs. KO) and related species. To understand cell-to-cell variations more single cell studies have to be performed. As the spatial regulation is not functioning on its own but in a tight crosstalk with the epigenome more integrative systems biology studies are required, with 3D FISH, super-resolution microscopy and rapidly evolving CRISPR technologies helping in establishing the causal link between regulatory factors and chromatin architecture and revealing the link between regulatory and structural long-range chromatin interactions. Pluripotency factors have been shown to play an important role in chromatin organization (discussed by [258]); nevertheless, future studies will be necessary to distinguish between the direct effects of a loss or gain-of-function of these factors on genome organization and secondary effects due to changes in gene expression of other factors of the chromatin landscape.

Ways to improve the resolution of chromatin interactome assays. From the qualitative side, the progress in sequencing technologies will help to increase the resolution of HiC and CHiC methods. However, C-type experiments have the resolution of restriction enzymes used in a study, thus new methods such as sonication/tagmentation-based HiC and other methods that do not rely on fixation-digestion-ligation approaches need to be developed. All these improvements will help us gaining a much better molecular insight in the processes at the different functional levels (i.e., TF-dependent signaling, epigenome function, chromatin organization and 3D architecture) that ultimately define the acquisition of a particular cell fate. Later some of these methods could be applied in a personalized medicine perspective to understand the aberrations in an individual cancer sample, and choose the combination of therapies that might have the highest likelihood of success. High-resolution CHiC methods are crucial in this application.

Altogether, the most fascinating perspective that derives from the studies of us and others of the decryption of the molecular basis of, and the mechanism(s) for the acquisition of a specific cell fates is the possibility to follow all the processes in their full complexity from the initial driving force(s) to the ultimate functional organism, and describe it as a roadmap of gene regulatory networks, in which each edge that links the genes/proteins would be explained from the (bio)chemical and (bio)physical points of view. We are thus experiencing a most exciting era of biosciences: the possibility to integrate the knowledge from various disciplines with the support of powerful computational resources and sophisticated tools to approach an understanding of the ontogenesis and homeostasis of multicellular organisms.

REFERENCES

1. Waddington, C. H. *THE STRATEGY OF THE GENES A Discussion of Some Aspects of Theoretical Biology*. *George Allen & Unwin* (1957).
2. Gurdon, J. B., Elsdale, T. R. & Fischberg, M. Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei. *Nature* **182,** 64–65 (1958).
3. Gurdon, J. & Road, P. The Developmental Capacity of Nuclei taken from Intestinal Epithelium Cells of Feeding Tadpoles. *J. Embryol. exp. Morph.* **10,** 622–640 (1962).
4. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51,** 987–1000 (1987).
5. Kulessa, H., Frampton, J. & Graf, T. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, ' and erythroblasts. *GENES Dev.* 1250–1262 (1995).
6. Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117,** 663–676 (2004).
7. Halder, G., Callaerts, P. & Gehring, W. J. Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. *Science (80-.).* **267,** 1788–1792 (1995).
8. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131,** 861–872 (2007).
9. Tsonis, P. A., Madhavan, M., Tancous, E. E. & Del Rio-Tsonis, K. A newt's eye view of lens regeneration. *Int. J. Dev. Biol.* **48,** 975–980 (2004).
10. Maki, N. *et al.* Expression profiles during dedifferentiation in newt lens regeneration revealed by expressed sequence tags. *Mol. Vis.* **16,** 72–78 (2010).
11. Weintraub, H. *et al.* Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 5434–8 (1989).
12. Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J. & Melton, D. A. In vivo reprogramming of adult pancreatic exocrine cells to β-cells. *Nature* **455,** 627–632 (2008).
13. Ieda, M. *et al.* Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142,** 375–386 (2010).
14. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475,** 390–393 (2011).
15. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
16. Hanahan, D. & Weinberg, R. a. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).
17. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **68,** 820–3 (1971).
18. Amos, W. Mutation biases and mutation rate variation around very short human microsatellites revealed by human-chimpanzee-orangutan genomic sequence alignments. *J. Mol. Evol.* **71,** 192–201 (2010).
19. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).

20. Fearon, E. F. & Vogelstein, B. A genetic model for Colorectal Tumorigenesis. *Cell Rev.* **61,** 759–767 (1990).
21. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148,** 59–71 (2012).
22. Sjoblom, T. *et al.* The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science).* **314,** 268-74 (2006).
23. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
24. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–9 (2012).
25. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–4 (2012).
26. Malysheva, V., Mendoza-Parra, M. A., Saleem, M.-A. M. & Gronemeyer, H. Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis. *Genome Med* **8,** 1–16 (2016).
27. Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301,** 89–92 (1983).
28. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37,** 853–62 (2005).
29. Liu, F., Wang, L., Perna, F. & Nimer, S. D. Beyond transcription factors: how oncogenic signalling reshapes the epigenetic landscape. *Nat. Rev. Cancer* **16,** 359–372 (2016).
30. You, J. S. & Jones, P. a. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell* **22,** 9–20 (2012).
31. Ley, T. J. *et al.* DNMT3A Mutations in Acute Myeloid Leukemia. *N. Engl. J. Med.* **363,** 2424–2433 (2010).
32. Morin, R. D. *et al.* Somatic mutation of EZH2 (Y641) in Follicular and Diffuse Large B-cell Lymphomas of Germinal Center Origin. *Nat. Genet.* **42,** 181–185 (2010).
33. Morin, R. D. *et al.* Frequent mutation of histone modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2012).
34. Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25,** 1499–1507 (2015).
35. Gilbertson, R. J. Mapping cancer origins. *Cell* **145,** 25–29 (2011).
36. Blanpain, C. Tracing the cellular origin of cancer. *Nat. Cell Biol.* **15,** 126–134 (2013).
37. Layek, R., Datta, A., Bittner, M. & Dougherty, E. R. Cancer therapy design based on pathway logic. *Bioinformatics* **27,** 548–555 (2011).
38. Dawson, M. A. *et al.* Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* **478,** 529–533 (2011).
39. Shih, C. & Weinberg, R. A. Isolation of a transforming sequence from a human bladder carcinoma cell line. *Cell* **29,** 161–9 (1982).

40. Goldfarb, M., Shimizu, K., Perucho, M. & Wigler, M. Isolation and preliminary characterization of a human transforming gene from T24 bladder carcinoma cells. *Nature* **296,** 404–409 (1982).
41. Pulciani, S. *et al.* Oncogenes in human tumor cell lines: molecular cloning of a transforming gene from human bladder carcinoma cells. *Proc. Natl. Acad. Sci. U. S. A.* **79,** 2845–2849 (1982).
42. Boxer, L. M. & Dang, C. V. Translocations involving c-myc and c-myc function. *Oncogene* **20,** 5595–5610 (2001).
43. Kress, T. R., Sabò, A. & Amati, B. MYC: connecting selective transcriptional control to global RNA production. *Nat. Rev. Cancer* **15,** 593–607 (2015).
44. Klapproth, K. & Wirth, T. Advances in the understanding of MYC-induced lymphomagenesis. *Br. J. Haematol.* **149,** 484–497 (2010).
45. Eilers, M. & Eisenman, R. N. Myc ' s broad reach. *Genes Dev.* 2755–2766 (2008). doi:10.1101/gad.1712408.Freely
46. Wang, Y. *et al.* Synthetic lethal targeting of MYC by activation of the DR5 death receptor pathway. *Cancer Cell* **5,** 501–512 (2004).
47. Ricci, M. S. *et al.* Direct repression of FLIP expression by c-myc is a major determinant of TRAIL sensitivity. *Mol. Cell. Biol.* **24,** 8541–8555 (2004).
48. Flinn, E. M. *et al.* Recruitment of Gcn5-containing complexes during c-Myc-dependent gene activation: Structure and function aspects. *J. Biol. Chem.* **277,** 23399–23406 (2002).
49. Eisenman, R. N. Deconstructing Myc. *Genes Dev.* **15,** 2023–2030 (2001).
50. Gartel, A. L. & Shchors, K. Mechanisms of c-myc-mediated transcriptional repression of growth arrest genes. *Exp. Cell Res.* **283,** 17–21 (2003).
51. Brenner, C. *et al.* Myc represses transcription through recruitment of DNA methyltransferase corepressor. *EMBO J.* **247600509,** 336–346 (2005).
52. Vervoorts, J., Lüscher-Firzlaff, J. & Lüscher, B. The ins and outs of MYC regulation by posttranslational mechanisms. *J. Biol. Chem.* **281,** 34725–34729 (2006).
53. Welcker, M. *et al.* The Fbw7 tumor suppressor regulates glycogen synthase kinase 3 phosphorylation-dependent c-Myc protein degradation. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9085–90 (2004).
54. Yada, M. *et al.* Phosphorylation-dependent degradation of c-Myc is mediated by the F-box protein Fbw7. *Embo J.* **23,** 2116–2125 (2004).
55. Popov, N. *et al.* The ubiquitin-specific protease USP28 is required for MYC stability. *Nat. Cell Biol.* **9,** 765–774 (2007).
56. Adhikary, S. *et al.* The ubiquitin ligase HectH9 regulates transcriptional activation by Myc and is essential for tumor cell proliferation. *Cell* **123,** 409–421 (2005).
57. Sears, R. *et al.* Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability. *Genes Dev.* **14,** 2501–2514 (2000).
58. Wang, Y., Quon, K. C., Knee, D. A., Nesterov, A. & Kraft, A. S. RAS , MYC , and Sensitivity to Tumor Necrosis Factor-A – Related Apoptosis- Inducing Ligand – Induced Apoptosis In Response : *Cancer Res.* 1615–1617 (2005).

59. Soucek, L. *et al.* Modelling Myc inhibition as a cancer therapy. *Nature* **455,** 679–83 (2008).
60. Sodir, N. M. *et al.* Endogenous Myc maintains the tumor microenvironment. *Genes Dev.* **25,** 907–916 (2011).
61. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448,** 313–317 (2007).
62. Lobo, N. a, Shimono, Y., Qian, D. & Clarke, M. F. The biology of cancer stem cells. *Annu. Rev. Cell Dev. Biol.* **23,** 675–699 (2007).
63. Kinzler, K. W. & Vogelstein, B. Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature* **386,** 761, 763 (1997).
64. Ahuja, D., Sáenz-Robles, M. T. & Pipas, J. M. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. *Oncogene* **24,** 7729–7745 (2005).
65. Hahn, W. C. *et al.* Creation of human tumour cells with defined genetic elements. *Nature* **400,** 464–8 (1999).
66. Hahn, W. C. *et al.* Enumeration of the simian virus 40 early region elements necessary for human cell transformation. *Mol Cell Biol* **22,** 2111–2123 (2002).
67. Nesterov, A. Oncogenic Ras Sensitizes Normal Human Cells to Tumor Necrosis Factor- -Related Apoptosis-Inducing Ligand-Induced Apoptosis. *Cancer Res.* **64,** 3922–3927 (2004).
68. Pavet, V. *et al.* Plasminogen activator urokinase expression reveals TRAIL responsiveness and supports fractional survival of cancer cells. *Cell Death Dis.* **5,** e1043 (2014).
69. Cavalli, G. & Misteli, T. Functional implications of genome topology. *Nat. Struct. Mol. Biol.* **20,** 290–9 (2013).
70. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8A resolution resolution. *Nature* **389,** 251 (1997).
71. Robinson, P. J. J., Fairall, L., Huynh, V. A. T. & Rhodes, D. EM measurements define the dimensions of the '30-nm' chromatin fiber: evidence for a compact, interdigitated structure. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 6506–11 (2006).
72. Widom, J. & Klug, A. Structure of the 300A chromatin filament: X-ray diffraction from oriented samples. *Cell* **43,** 207–213 (1985).
73. Woodcock, C. L. & Ghosh, R. P. Chromatin Higher-order Structure and Dynamics. *Cold Spring Harb. Perspect. Biol.* **2,** a000596 (2010).
74. Williams, S. P. *et al.* Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length. *Biophys. J.* **49,** 233–48 (1986).
75. Luger, K., Dechassa, M. L. & Tremethick, D. J. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.* **13,** 436–47 (2012).
76. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–80 (2012).
77. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485,** 381–385 (2012).

78. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–472 (2012).
79. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
80. Cmarko, D. *et al.* Ultrastructural analysis of transcription and splicing in the cell nucleus after bromo-UTP microinjection. *Mol. Biol. Cell* **10,** 211–23 (1999).
81. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* **326,** 289–93 (2009).
82. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14,** 390–403 (2013).
83. Liu, C. & Weigel, D. Chromatin in 3D: progress and prospects for plants. *Genome Biol* **16,** 170 (2015).
84. Ciabrelli, F. & Cavalli, G. Chromatin-driven behavior of topologically associating domains. *J. Mol. Biol.* **427,** 608–625 (2015).
85. Guillaume, A. *et al.* Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **340,** (2013).
86. Lupianez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161,** 1012–1025 (2015).
87. Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24,** 390–400 (2014).
88. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153,** 1281–1295 (2013).
89. Gurudatta, B. V., Yang, J., Van Bortle, K., Donlin-Asp, P. G. & Corces, V. G. Dynamic changes in the genomic localization of DNA replication-related element binding factor during the cell cycle. *Cell Cycle* **12,** 1605–1615 (2013).
90. Berlivet, S. *et al.* Clustering of Tissue-Specific Sub-TADs Accompanies the Regulation of HoxA Genes in Developing Limbs. *PLoS Genet.* **9,** (2013).
91. Bonora, G., Plath, K. & Denholtz, M. A mechanistic link between gene regulation and genome architecture in mammalian development. *Curr. Opin. Genet. Dev.* **27,** 92–101 (2014).
92. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488,** 116–120 (2012).
93. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci USA* **111,** 996–1001 (2014).
94. Sofueva, S. *et al.* Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J.* **32,** 3119–29 (2013).
95. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15,** 1–12 (2016).
96. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2,** 1–22 (2010).
97. Van Steensel, B. Chromatin: constructing the big picture. *EMBO J.* **30,** 1885–95 (2011).
98. Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10,** 211–219 (2001).

99. Solovei, I. *et al.* Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell* **137,** 356–68 (2009).

100. Moter, A. & Göbel, U. B. Fluorescence in situ hybridization (FISH) for direct visualization of microorganisms. *J. Microbiol. Methods* **41,** 85–112 (2000).

101. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science (80-.).* **295,** 1306–11 (2002).

102. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38,** 1341–1347 (2006).

103. Van de Werken, H. J. G. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods* **9,** 969–972 (2012).

104. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).

105. Mendoza-Parra, M.-A., Blum, M., Malysheva, V., Cholley, P.-E. & Gronemeyer, H. LOGIQA: a database dedicated to long-range genome interactions quality assessment. *BMC Genomics* **17,** 355 (2016).

106. Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res* **24,** 1854–1868 (2014).

107. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502,** 59–64 (2013).

108. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124,** 207–219 (2006).

109. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).

110. Handoko, L. *et al.* CTCF-Mediated Functional Chromatin Interactome in Pluripotent Cells. *Nat. Genet.* **43,** 630–638 (2012).

111. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* **22,** 490–503 (2012).

112. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).

113. DeMare, L. E. *et al.* The genomic landscape of cohesin-Associated chromatin interactions. *Genome Res.* **23,** 1224–1234 (2013).

114. Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter– enhancer long-range associations Yubo. *Nature* **504,** 306–310 (2013).

115. Heidari, N., Phanstiel, D. & He, C. Genome-wide map of regulatory interactions in the human genome. *Genome Res.* **24,** 1905–17 (2014).

116. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163,** 1611–1627 (2015).

117. Mumbach, M. R. *et al.* HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* (2016). doi:doi:10.1038/nmeth.3999
118. Filion, G. J. *et al.* Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* **143,** 212–224 (2010).
119. Escamilla-Del-Arenal, M., Da Rocha, S. T. & Heard, E. Evolutionary diversity and developmental regulation of X-chromosome inactivation. *Hum. Genet.* **130,** 307–327 (2011).
120. Chaligné, R. *et al.* The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer. *Genome Res.* **25,** 488–503 (2015).
121. Galupa, R. & Heard, E. X-chromosome inactivation: New insights into cis and trans regulation. *Curr. Opin. Genet. Dev.* **31,** 57–66 (2015).
122. Giorgetti, L. *et al.* Chromosome Conformation and Transcription. *Cell* **157,** 950–963 (2014).
123. Muller, H. J. & Altenburg, E. The frequency of translocations produced by X-rays in Drosophila. *Genetics* **15,** 283–311 (1929).
124. McClintock, B. Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* **16,** 13–47 (1951).
125. Lyon, M. F. X-chromosome inactivation and developmental patterns in mammals. *Biol. Rev. Camb. Philos. Soc.* **47,** 1–35 (1972).
126. Surani, M. A., Barton, S. C. & Norris, M. L. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature* **308,** 548–550 (1984).
127. Mcgrath, J. & Solter, D. Completion of Mouse Embryogenesis Requires Both the Maternal and Paternal Genomes. *Cell* **37,** 179–183 (1984).
128. Mandel, J. L. & Chambon, P. DNA methylation: organ specific variations in the methylation pattern within and around ovalbumin and other chicken genes. *Nucleic Acids Res.* **7,** 2081–2103 (1979).
129. Shen, C.-K. & Maniatis, T. Tissue-specific DNA methylation in a cluster of rabbit beta-like globin genes. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 6634–6638 (1980).
130. Sakai, T. *et al.* Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *Am. J. Hum. Genet.* **48,** 880–8 (1991).
131. Feinberg, A. P. & Vogelstein, B. Hypomethylation of ras oncogenes in primary human cancers. *Biochem. Biophys. Res. Commun.* **111,** 47–54 (1983).
132. Oshimo, Y. *et al.* Promoter methylation of cyclin D2 gene in gastric carcinoma. *Int. J. Oncol.* **23,** 1663–1670 (2003).
133. Badal, V., Chuang, L., Tan, E. & Badal, S. CpG methylation of human papillomavirus type 16 DNA in cervical cancer cell lines and in clinical specimens: genomic hypomethylation correlates with carcinogenic. *J. Virol.* **77,** 6227–6234 (2003).
134. Adorján, P. *et al.* Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.* **30,** e21 (2002).
135. Goelz, S., Vogelstein, B. & Feinberg, A. Hypomethylation of DNA from benign and malignant human colon neoplasms. *Science (80-.).* **228,** 187–190 (1985).

136. Feinberg, A. P. & Kuo, K. C. Reduced Genomic 5-Methylcytosine content in human colonic neoplasi and content as afraction oftotal. *Cancer Res.* 1159–1161 (1988).
137. Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **51,** 786–94 (1964).
138. Jeppesen, P. & Turner, B. M. The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell* **74,** 281–289 (1993).
139. Bone, J. R., Lavender, J., Ron, R. & Turner, B. M. Acetylated histone H4 on the male X chromosome is associated with dosage compensation in Drosophila. *Genes Dev.* 96–104 (1994).
140. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39,** 311–8 (2007).
141. Hon, G. C., Hawkins, R. D. & Ren, B. Predictive chromatin signatures in the mammalian genome. *Hum. Mol. Genet.* **18,** 12–14 (2009).
142. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21931–21936 (2010).
143. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8,** 532–8 (2006).
144. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125,** 315–326 (2006).
145. Rugg-Gunn, P. J. *et al.* Cell-Surface Proteomics Identifies Lineage-Specific Markers of Embryo-Derived Stem Cells. *Dev. Cell* **22,** 887–901 (2012).
146. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149,** 590–604 (2012).
147. Alder, O. *et al.* Ring1B and Suv39h1 delineate distinct chromatin states at bivalent genes during early mouse lineage commitment. *Development* **137,** 2483–2492 (2010).
148. Dahl, J. A., Reiner, A. H., Klungland, A., Wakayama, T. & Collas, P. Histone H3 lysine 27 methylation asymmetry on developmentally-regulated promoters distinguish the first two lineages in mouse preimplantation embryos. *PLoS One* **5,** (2010).
149. Vastenhouw, N. L. *et al.* Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464,** 922–926 (2010).
150. Akkers, R. C. *et al.* A Hierarchy of H3K4me3 and H3K27me3 Acquisition in Spatial Gene Regulation in Xenopus Embryos. *Dev. Cell* **17,** 425–434 (2009).
151. Sachs, M. *et al.* Bivalent Chromatin Marks Developmental Regulatory Genes in the Mouse Embryonic Germline in Vivo. *Cell Rep.* **3,** 1777–1784 (2013).
152. Wei, Y., Schatten, H. & Sun, Q. Y. Environmental epigenetic inheritance through gametes and implications for human reproduction. *Hum. Reprod. Update* **21,** 194–208 (2015).
153. Hahn, M. A. *et al.* Loss of the polycomb mark from bivalent promoters leads to activation of cancer-promoting genes in colorectal tumors. *Cancer Res.* **74,** 3617–3629 (2014).
154. Battich, N., Stoeger, T. & Pelkmans, L. Control of Transcript Variability in Single Mammalian Cells. *Cell* **163,** 1596–1610 (2015).

155. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–5 (2015).
156. Morris, K. V & Mattick, J. S. The rise of regulatory RNA. *Nat. Rev. Genet.* **15,** 423–37 (2014).
157. Spellman, P. T. & Rubin, G. M. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.* **1,** 5 (2002).
158. Caron, H. The Human Transcriptome Map: Clustering of Highly Expressed Genes in Chromosomal Domains. *Science (80-.).* **291,** 1289–1292 (2013).
159. Lan, X. *et al.* Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic Acids Res.* **40,** 7690–7704 (2012).
160. Ferraiuolo, M. a, Sanyal, A., Naumova, N., Dekker, J. & Dostie, J. From cells to chromatin: capturing snapshots of genome organization with 5C technology. *Methods* **58,** 255–267 (2012).
161. Kieffer-Kwon, K. R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155,** 1507–1520 (2013).
162. Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36,** 1065–1071 (2004).
163. Ragoczy, T., Bender, M. A., Telling, A., Byron, R. & Groudine, M. The locus control region is required for association of the murine α-globin locus with engaged transcription factories during erythroid maturation. *Genes Dev.* **20,** 1447–1457 (2006).
164. Jackson, D. A., Hassan, A. B., Errington, R. J. & Cook, P. R. Visualization of focal sites of transcription within human nuclei. *EMBO J.* **12,** 1059–65 (1993).
165. Sutherland, H. & Bickmore, W. A. Transcription factories: gene expression in unions? *Nat. Rev. Genet.* **10,** 457–466 (2009).
166. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42,** 53–61 (2010).
167. Negre, N. *et al.* A cis-regulatory map of the Drosophila genome. *Nature* **471,** 527–531 (2011).
168. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13,** R48 (2012).
169. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. An Integrative Systems Medicine Approach to Mapping Human Metabolic Diseases. *Nat Rev Genet* **12,** 56–68 (2011).
170. Euler, L. Solutio problematis ad geometrian situs pertinentis. *Coment. Acad. Sci. Petropolitanae* **8,** 128–140 (1736).
171. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world'networks. *Nature* **393,** 440–442 (1998).
172. Barabasi, A.-L. & Reka, A. Emergence of Scaling in Random Networks. *Science (80-.).* **286,** 509–512 (1999).
173. Barabási, A. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5,** 101–113 (2004).
174. Ma'ayan, A., Blitzer, R. D. & Iyengar, R. Toward predictive models of mammalian cells. *Annu. Rev. Biophys. Biomol. Struct.* **34,** 319–49 (2005).

175. Jeong, H., Mason, S. P., Barabási, a L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).
176. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. Hierarchical Organization of Modularity in Metabolic Networks. *Science (80-.).* **297,** 1551–1555 (2002).
177. Borneman, A. R. *et al.* Target hub proteins serve as master regulators of development in yeast. *Genes Dev.* **20,** 435–448 (2006).
178. Dyer, M. D., Murali, T. M. & Sobral, B. W. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog.* **4,** (2008).
179. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* **3,** 713–720 (2007).
180. Cahan, P. *et al.* CellNet: Network Biology Applied to Stem Cell Engineering. *Cell* **158,** 903–915 (2014).
181. Przytycka, T. M., Singh, M. & Slonim, D. K. Toward the dynamic interactome: It's about time. *Brief. Bioinform.* **11,** 15–29 (2010).
182. Rachlin, J., Cohen, D. D., Cantor, C. & Kasif, S. Biological context networks: a mosaic view of the interactome. *Mol. Syst. Biol.* **2,** 66 (2006).
183. Agarwal, S., Deane, C. M., Porter, M. A. & Jones, N. S. Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Comput. Biol.* **6,** 1–12 (2010).
184. Gao, S. & Wang, X. Identification of highly synchronized subnetworks from gene expression data. *BMC Bioinformatics* **14 Suppl 9,** S5 (2013).
185. Zinman, G. E. *et al.* ModuleBlast: Identifying activated sub-networks within and across species. *Nucleic Acids Res.* **43,** e20 (2015).
186. Soul, J., Hardingham, T. E., Boot-Handford, R. P. & Schwartz, J.-M. PhenomeExpress: a refined network analysis of expression datasets by inclusion of known disease phenotypes. *Sci. Rep.* **5,** 8117 (2015).
187. Lin, C. Y. *et al.* Hubba: hub objects analyzer--a framework of interactome hubs identification for network biology. *Nucleic Acids Res.* **36,** 438–443 (2008).
188. Hernandez-Toro, J., Prieto, C. & De Las Rivas, J. APID2NET: Unified interactome graphic analyzer. *Bioinformatics* **23,** 2495–2497 (2007).
189. Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3,** 140 (2007).
190. Assenov, Y., Ramirez, F., Schelhorn, S. E. S. E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24,** 282–284 (2008).
191. Doncheva, N. T., Assenov, Y., Domingues, F. S. & Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **7,** 670–85 (2012).
192. Scardoni, G. *et al.* Biological network analysis with CentiScaPe: centralities and experimental dataset integration. *F1000Research* **3,** 1–8 (2014).
193. Aderem, A. *et al.* A Systems Biology Approach to Infectious Disease Research : Innovating the Pathogen-Host Research Paradigm. *MBio* **2,** 1–4 (2011).

194. Werner, H. M. J., Mills, G. B. & Ram, P. T. Cancer Systems Biology: a peek into the future of patient care? *Nat. Rev. Clin. Oncol.* **11,** 167–176 (2014).

195. Torkamani, A. & Schork, N. J. Identification of rare cancer driver mutations by network reconstruction. *Genome Res.* **19,** 1570–1578 (2009).

196. Mine, K. L. *et al.* Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. *Nat. Commun.* **4,** 1806 (2013).

197. Svoboda, M. *et al.* AID/APOBEC-network reconstruction identifies pathways associated with survival in ovarian cancer. *BMC Genomics* **17,** 643 (2016).

198. Gardner, T. S., di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science (80-.).* **301,** 102–5 (2003).

199. ENCODE Project *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

200. Göndör, A. & Ohlsson, R. Chromosome crosstalk in three dimensions. *Nature* **461,** 212–7 (2009).

201. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164,** 1110–1121 (2016).

202. Wei, Z. *et al.* Klf4 organizes long-range chromosomal interactions with the OCT4 locus inreprogramming andpluripotency. *Cell Stem Cell* **13,** 36–47 (2013).

203. De Wit, E. *et al.* The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501,** 227–31 (2013).

204. Soprano, D. R., Teets, B. W. & Soprano, K. J. Role of Retinoic Acid in the Differentiation of Embryonal Carcinoma and Embryonic Stem Cells. *Vitam. Horm.* **75,** 69–95 (2007).

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

206. Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501,** (2013).

207. Mendoza-Parra, M.-A. *et al.* Reconstructed cell fate-regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis. *Genome Res.* gr.208926.116 (2016). doi:10.1101/GR.208926.116

208. Corces, M. R. & Corces, V. G. The three-dimensional cancer genome. *Curr. Opin. Genet. Dev.* **36,** 1–7 (2016).

209. Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L. a. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.* **29,** 1109–1113 (2011).

210. Taberlay, P. C. *et al.* Three-dimensional disorganisation of the cancer genome occurs coincident with long range genetic and epigenetic alterations. *Genome Res.* gr.201517.115 (2010). doi:10.1101/gr.201517.115

211. Vjetrovic, J., Shankaranarayanan, P., Mendoza-Parra, M. a. & Gronemeyer, H. Senescence-secreted factors activate Myc and sensitize pretransformed cells to TRAIL-induced apoptosis. *Aging Cell* **13,** 487–496 (2014).

212. Turing, A. M. The chemical basis of morphogenesis. *Philos. Trans. R. Soc. London.* **237,** 37–72 (1952).

213. Wolpert, L. Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.* **25,** 1–47 (1969).
214. Wolpert, L. Positional information revisited. *Development* **107,** 3–12 (1989).
215. Economou, A. D. *et al.* Periodic stripe formation by a Turing mechanism operating at growth zones in the mammalian palate. *Nat. Genet.* **44,** 348–51 (2012).
216. Driever, W. & Nusslein-Volhard, C. The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell* **54,** 95–104 (1988).
217. Green, J. B. A. & Sharpe, J. Positional information and reaction-diffusion: two big ideas in developmental biology combine. *Development* **142,** 1203–1211 (2015).
218. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17,** 1–14 (2016).
219. Ceschin, D. G. *et al.* Methylation specifies distinct estrogen-induced binding site repertoires of CBP to chromatin. *Genes Dev.* **25,** 1132–1146 (2011).
220. Van Kruijsbergen, I., Hontelez, S. & Veenstra, G. J. C. Recruiting polycomb to chromatin. *Int. J. Biochem. Cell Biol.* **67,** 177–187 (2015).
221. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482,** 339–346 (2012).
222. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution - Trashing old rules to forge new ones. *Cell* **157,** 77–94 (2014).
223. Volpe, T. *et al.* Regulation oh heterochromatic silencing and histone H3 Lysine-9 by RNAi. *Science (80-.).* **297,** 1833–1837 (2002).
224. Mochizuki, K., Fine, N. A., Fujisawa, T. & Gorovsky, M. A. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in Tetrahymena. *Cell* **110,** 689–699 (2002).
225. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat. Struct. Mol. Biol.* **20,** 1147–55 (2013).
226. Sparmann, A. & van Lohuizen, M. Polycomb silencers control cell fate, development and cancer. *Nat. Rev. Cancer* **6,** 846–856 (2006).
227. Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat. Genet.* **47,** 1179–86 (2015).
228. Ljungman, M. The transcription stress response. *Cell Cycle* **6,** 2252–2257 (2007).
229. Bensaude, O. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription* **2,** 103–108 (2011).
230. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).
231. Williamson, I., Hill, R. E. & Bickmore, W. A. Enhancers: From Developmental Genetics to the Genetics of Common Human Disease. *Dev. Cell* **21,** 17–19 (2011).
232. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14,** 288–95 (2013).

233. De Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502,** 499–506 (2013).
234. Kowalczyk, M. S. *et al.* Intragenic Enhancers Act as Alternative Promoters. *Mol. Cell* **45,** 447–458 (2012).
235. Kolovos, P., Knoch, T. a, Grosveld, F. G., Cook, P. R. & Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin* **5,** 1 (2012).
236. Fox, C. H., Johnson, F. B., Whiting, J. & Roller, P. P. Formaldehyde Fixation. *J. Histochem. Cytochem.* **33,** 845–853 (1985).
237. Eungdamrong, N. J. & Iyengar, R. Computational approaches for modeling regulatory cellular networks. *Trends Cell Biol.* **14,** 661–669 (2004).
238. Hardy, S. & Robillard, P. N. Petri net-based method for the analysis of the dynamics of signal propagation in signaling pathways. *Bioinformatics* **24,** 209–217 (2008).
239. Kashtan, N., Itzkovitz, S., Milo, R. & Alon, U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* **20,** 1746–1758 (2004).
240. Chin, G., Chavarria, D. G., Nakamura, G. C. & Sofia, H. J. BioGraphE: high-performance bionetwork analysis using the Biological Graph Environment. *BMC Bioinformatics* **9 Suppl 6,** S6 (2008).
241. Li, S., Assmann, S. M. & Albert, R. Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *PLoS Biol.* **4,** 1732–1748 (2006).
242. Nüsslein-Volhard, C. Gradients that organize embryo development. *Sci. Am.* **275,** 54–55; 58–61 (1996).
243. Meinhardt, H. Biological pattern formation: new observations provide support for theoretical predictions. *Bioessays* **16,** 627–632 (1994).
244. Johnson, M. H. & McConnell, J. M. L. Lineage allocation and cell polarity during mouse embryogenesis. *Semin. Cell Dev. Biol.* **15,** 583–597 (2004).
245. Allen, M. Compelled by the Diagram: Thinking through C. H. Waddington's Epigenetic Landscape. *Contemp. Hist. Presence Vis. Cult.* **4,** 119 (2015).
246. Böhmdorfer, G. & Wierzbicki, A. T. Control of Chromatin Structure by Long Noncoding RNA. *Trends Cell Biol.* **25,** 623–632 (2015).
247. Xu, W. *et al.* Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of ??-ketoglutarate-dependent dioxygenases. *Cancer Cell* **19,** 17–30 (2011).
248. McCarthy, N. Metabolism: Unmasking an oncometabolite. *Nat. Rev. Cancer* **12,** 229–229 (2012).
249. Shim, E. H. *et al.* L-2-hydroxyglutarate: An epigenetic modifier and putative oncometabolite in renal cancer. *Cancer Discov.* **4,** 1290–1298 (2014).
250. Letouzé, E. *et al.* SDH Mutations Establish a Hypermethylator Phenotype in Paraganglioma. *Cancer Cell* **23,** 739–752 (2013).
251. Selvaraj, S., R Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31,** 1111–8 (2013).
252. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.* **31,** 1143–1147 (2013).
253. Korbel, J. O. & Lee, C. Genome assembly and haplotyping with Hi-C. *Nat Biotech* **31,** 1099–1101 (2013).

254. Marie-Nelly, H. *et al.* High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5,** 5695 (2014).

255. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31,** 1119–1125 (2013).

256. English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* **7,** 1–12 (2012).

257. Laszlo, A. H. *et al.* Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol.* **32,** 829–33 (2014).

258. Novo, C. L. & Rugg-Gunn, P. J. Chromatin organization in pluripotent cells: Emerging approaches to study and disrupt function. *Brief. Funct. Genomics* **15,** 305–314 (2016).

259. Maître, J.-L., Niwayama, R., Turlier, H., Nédélec, F. & Hiiragi, T. Pulsatile cell-autonomous contractility drives compaction in the mouse embryo. *Nat. Cell Biol.* **17,** 849–855 (2015).

260. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* (2016). doi:10.1038/nrm.2016.104

APPENDIX I: FRENCH THESIS ABSTRACT

RECONSTRUCTION DES RESEAUX DE REGULATION GENIQUES RESPONSABLES DU DESTIN CELLULAIRE

ETAT DE L'ART

D'importants progrès ont vu le jour ces dix dernières années dans l'identification d'altérations génétiques impliquées dans le développement de nombreux cancers, et agissant à différents niveaux de la régulation génique. Alors que de nombreuses études se sont concentrées sur le rôle prépondérant des voies de signalisation dans ce phénomène, des recherches récentes ont montré que les cellules cancéreuses ont un épigénome qui diffère de manière drastique de celui de cellules normales. Ces modifications épigénétiques, qui incluent la méthylation de l'ADN, les modifications post-traductionnelles des histones et les variants d'histones, influencent la transcription des gènes. Durant la tumorigenèse une dérégulation globale du transcriptome est observée, ainsi que des changements drastiques du paysage chromatinien. Comment sont opérés ces changements, et comment ils régulent la transformation de l'identité cellulaire sont deux questions majeures aujourd'hui.

Les interactions longue distance de la chromatine ont potentiellement une fonction importante dans la pathogenèse du cancer [1]. En effet, certains oncogènes régulateurs de la transcription peuvent induire des changements de la structure de la chromatine, conduisant à des altérations génomiques [2], et à l'expression aberrante de facteurs de transcription. Toutefois les différentes relations entre la dérégulation globale de l'architecture de la chromatine, du transcriptome, et de l'épigénome lors des changements de l'identité cellulaire (différenciation cellulaire ou transformation tumorale) restent à élucider.

QUESTIONS POSEES

L'étude présentée ici révèle les interactions mises en jeu entre l'épigénome et le transcriptome lors de changements du destin cellulaire tels que la différenciation et la tumorigenèse. Ainsi nous avons adressé les questions suivantes : (i) comment l'expression globale des gènes et l'organisation de l'épigénome sont modifiées lors de ses évènements cellulaires; (ii) quelle est la fonction régulatrice des protéines de remodelage de la chromatine pendant la tumorigenèse ; (iii) quelles sont les altérations globales de l'architecture chromatinienne durant la différenciation et la transformation cellulaire ; et (iv) comment le transcriptome, l'épigénome et la structure globale de la chromatine se coordonnent durant l'acquisition de

l'identité cellulaire dans des lignées de cellules normales, quelle est leur degré de plasticité, et comment est altérée cette coordination dans des cellules cancéreuses.

APPROCHES EXPERIMENTALES

La première étape vers l'analyse des relations complexes entre transcriptome, épigénome, et structure de la chromatine dans l'acquisition de l'identité cellulaire, a été de caractériser les modifications dynamiques de l'expression des gènes, et de l'état de la chromatine en utilisant trois systèmes cellulaires différents. Les deux premiers modèles cellulaires utilisés sont des lignées de cellules de carcinome embryonnaire F9 et P19 qui se différencient respectivement en cellules neuronales et endodermales après traitement à l'ATRA (all-trans retinoic acid). En parallèle nous avons utilisé un modèle cellulaire isogénique pour l'induction par étape de la tumorigenèse [3], dans le but d'étudier l'organisation dynamique de la chromatine lors de la transformation tumorale des cellules. Celle-ci est induite par l'introduction de gènes responsables de l'immortalisation (hTERT : la sous-unité catalytique de la télomérase, et SV40 : antigène t) et de la transformation (l'oncogène *c-MYC*) des cellules souches en cellules primaires humaines.

Ce dernier modèle a été sélectionné pour étudier les changements transcriptomiques, épigénomiques, et de l'organisation de la chromatine accompagnant la transformation de cellules humaines normales en cellules tumorales. En effet une telle étude comparative n'est pas facilement transposable à des cellules tumorales primaires ou à des lignées cellulaires cancéreuses bien établies qui sont caractérisées par un nombre très important d'altérations génétiques, alors que notre système présente un nombre plus limité de modifications génétiques. De plus, dans ce modèle les cellules sont quasi isogéniques, ce qui permet de comparer les cellules immortalisées et tumorales directement avec les cellules normales.
Afin de réaliser une analyse systématique des réseaux de régulation géniques impliqués dans la transformation cellulaire physiologique (différenciation) et pathologique (tumorigenèse), nous avons utilisé une nouvelle approche combinatoire visant à (i) intégrer les données transcriptomiques aux données de l'état de la chromatine durant les différentes étapes de transformation, (ii) identifier les facteurs de transcription clés dans la transformation cellulaire en utilisant des bases de données montrant des associations établies entre facteurs de transcription et gènes cibles, et, dans le cas de la tumorigenèse (iii) compléter l'étude avec une analyse des protéines de remodelage et de modulation de la chromatine (CRM pour Chromatin Remodelers and Modulators) impliquées dans ce processus.

Dans le but d'établir l'interactome de la chromatine dans ces systèmes, nous avons réalisé des captures de conformation chromosomique sur tout le génome (HiC) à différents temps de traitement par l'ATRA, dans les cellules F9 et P19, et à chaque étape de la transformation cellulaire. L'intégration des données de l'organisation de la chromatine avec les données épigénomiques et transcriptomiques a permis de faire une caractérisation fonctionnelle des interactions longue distance de la chromatine en connectant les facteurs régulateurs clés aux éléments régulateurs de l'ADN.

RESULTATS

La reconstruction des réseaux de régulation géniques, modifiés durant les étapes de tumorigenèse dans des lignées cellulaires humaines, nous a permis d'identifier un large nombre de nouveaux régulateurs de ce phénomène (Malysheva Valeriya, Marco-Antonio Mendoza-Parra, Mohamed Ashick Mohamed Saleem and Hinrich Gronemeyer. Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis. *Genome Medicine*. (2016) **8,** 1–16 2016). Grâce à l'intégration des données de séquençage haut-débit, nous avons pu prédire puis valider le rôle clé de plusieurs facteurs de transcription dans l'établissement de l'identité tumorale des cellules transformées (Figures 1). Notre analyse a aussi indiqué que les CRMs sont largement impliqués dans la transformation tumorale induite par l'expression d'oncogènes, et a identifié de nouveaux CRMs ayant une fonction dans ce procédé (Figure 2).

Enfin nous avons caractérisé la dynamique de l'architecture chromatinienne lors de la transformation des cellules normales en cellules tumorales (Figure 3) (Malysheva Valeriya, Marco-Antonio Mendoza-Parra, Matthias Blum and Hinrich Gronemeyer. Chromatin dynamics during tumorigenic transformation. *Manuscrit en cours de préparation*). Dans ces travaux nous décrivons les altérations globales de l'architecture de la chromatine qui sont établies très tôt lors de la transformation cellulaire. L'analyse des changements drastiques de l'interactome de la chromatine observés pendant la tumorigenèse après l'étape d'immortalisation cellulaire, et la transformation oncogénique par l'activité de c-MYC, est actuellement en cours. Cette analyse inclut l'intégration de données de la structure de la chromatine avec nos données transcriptomiques et épigénomiques précédemment décrites (Malysheva et al. 2016), ainsi qu'avec des données sur l'accessibilité de la chromatine (FAIRE-seq). Le but de cette étude est de mieux comprendre l'impact des facteurs de la tumorigenèse sur la structure de la chromatine, et en particulier les mécanismes par lesquel le

facteur c-MYC agit comme un facteur global du remodelage de la chromatine lors de la transformation tumorigénique des cellules.

A notre connaissance il s'agit de la première étude intégrative de réseaux de régulation géniques lors de la différenciation cellulaire, et de la tumorigenèse par étape, dans un système virtuellement isogénique.

Cette approche systématique pour caractériser la différenciation des cellules F9 et P19 révèle comment ATRA active un réseau spécifique de facteurs de transcription qui va guider l'organisation temporelle de réseaux de régulation géniques et induire la différenciation neuronale / endodermale des cellules. La modélisation de la transduction du signal en utilisant des réseaux de régulation géniques reconstruits à partir d'études génomiques globales a mis en lumière les facteurs de transcriptions clés dans la spécification de l'identité neuronale des cellules. Leurs fonctions ont ensuite été validées par édition du génome en utilisant le système CRISPR/Cas9 (Marco-Antonio Mendoza-Parra, Valeriya Malysheva, Mohamed Ashick Mohamed Saleem, Michele Lieb, Aurelie Godel, and Hinrich Gronemeyer. Reconstructed cell fate-regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis. *Genome Research.* (2016). doi:10.1101/GR.208926.116).

L'analyse des données HiC a montré que les cellules F9 et P19 subissent des changements importants dans leur organisation chromatinienne après traitement avec l'ATRA. De façon surprenante nous avons observé des modifications hautement dynamiques d'interactions entre domaines chromatiniens, ce qui n'avait jamais été reporté jusqu'alors (Figure 4). Ainsi un traitement par un seul composé chimique est capable d'induire une réorganisation globale de la chromatine. Les validations expérimentales sont actuellement en cours. Elles consistent à invalider les régions régulatrices de l'ADN en induisant des mutations par la technologie CRISPR.

La capacité d'un réseau de régulation génique à reconstituer la cascade de régulation transcriptionnelle aboutissant à la differentiation neuronale ou endodermique, est validée en utilisant la stratégie de "la propagation du signal" (Mendoza-Parra et al. 2016). Cette méthode évalue la capacité de chaque intersection à induire tout ou partie des programmes types cellulaires – spécifiques (Figure 4). Il est important de noter que cette analyse a permis de prédire de nombreux facteurs clés de la differentiation cellulaire, dont la pluspart sont présentés dans la publication de Mendoza-Parra et al. (2016). De plus nous avons mis en

lumière le rôle de plusieurs GAPs (Genome Associated Platforms) dans l'activation des cascades transcriptionelles pour la neurogenèse (Figure 5).

Pour conclure nos données ont révélé les grandes capacités d'un seul morphogène à réorganiser les interactions longues distances de la chromatine lors de l'acquisition du destin cellulaire. Nous suggérons que ces changements de points de contact de la chromatine influencent l'acquisition de l'identité cellulaire (Malysheva Valeriya, Marco-Antonio Mendoza-Parra*, Matthias Blum and Hinrich Gronemeyer*. Chromatin structure dynamics directs cell fate acquisition. *Manuscrit en cours de préparation*).

CONCLUSIONS ET PERSPECTIVES

Nous avons reconstruit les réseaux de régulation géniques altérés durant les étapes de tumorigenèse dans des cellules humaines. L'analyse de ces réseaux nous a permis de prédire puis valider le rôle clé de plusieurs facteurs de transcription dans l'acquisition du caractère tumoral des cellules transformées. Notre étude suggère que les CRMs sont directement impliqués dans ce phénomène, et de nouveaux CRMs critiques pour la transformation tumorale ont été identifiés. Des expériences complémentaires sont maintenant requises afin d'identifier des cibles thérapeutiques potentielles parmi ces facteurs.

Notre étude sur la différenciation cellulaire a montré qu'un seul composé chimique, tel que le morphogène ATRA, peut activer des réseaux complexes de régulation géniques permettant d'induire la différenciation d'une cellule souche/précurseur en une cellule avec une identité spécifique (dépendante de son origine). Notre approche systématique pour caractériser l'acquisition de l'identité cellulaire, combinée à la modélisation de la transduction du signal, renforce nos connaissances sur les mécanismes responsables de la plasticité cellulaire, ce qui pourrait être utilisé pour induire artificiellement la différenciation cellulaire.

Nous poursuivons aujourd'hui nos efforts en terme de bioinformatique afin de délivrer une analyse fonctionnelle de la dynamique des interactions longue distance de la chromatine en intégrant les données relatives à l'organisation de la chromatine, aux données épigénomiques et transcriptomiques. Pour conclure ce travail apporte une compréhension globale du rôle, mais aussi des interactions entre transcriptome, épigénome et organisation chromatinienne qui tous façonnent l'identité cellulaire.

REFERENCES

[1] R. Jager, G. Migliorini, M. Henrion, R. Kandaswamy, H.E. Speedy et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nature Communications 6:6178, 2015

[2] D.S. Rickman, T.D. Soong, B. Moss, J.M. Mosquera, J. Dlabal, S, Terry et al. Oncogene-mediated alterations in chromatin conformation. Proc Natl Acad Sci USA, 109:9083-8, 2012

[3] W. C. Hahn, W. C. Hahn, C. M. Counter, C. M. Counter, a S. Lundberg, a S. Lundberg, R. L. Beijersbergen, R. L. Beijersbergen, M. W. Brooks, M. W. Brooks, R. a Weinberg, and R. a Weinberg, "Creation of human tumour cells with defined genetic elements.," *Nature*, vol. 400, no. 6743, pp. 464–8, 1999.

[4] M.A. Mendoza-Parra, M Walia, M. Sankar, H. Gronemeyer. "Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics" *Mol Syst Biol* 7: 538, 2011

[5] V. Malysheva, M.A. Mendoza-Parra, M.A. M. Saleem and H. Gronemeyer. Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis. Genome Medicine 8:57, 2016.

[6] M.A. Mendoza-Parra, V. Malysheva , M.A. M. Saleem, M. Lieb, A. Godel, H. Gronemeyer. Reconstructed cell fate-regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis. Genome research, in revision

LEGENDES DES FIGURES

Figure 1. Réseau de régulation génique (GRN pour Gene Regulatory Network) du système de transformation cellulaire par étape BJ. a. GRN de cellules BJEL immortalisées. **b**, GRN de cellules BJELM transformées. Les intersections où sont localisés les facteurs de remodelage ou de modulation de la chromatine sont représentés par des losanges. Les noeuds les plus connectés, et ceux qui ont un effet d'entonnoir sont représentés par des cercles. A chaque intersection, les niveaux d'expression différentielle à l'étape d'immortalisation cellulaire (cellules BJEL), puis à l'étape de transformation tumorigénique (cellules BJELM) sont représentés par un gradient de couleurs, permettant de visualiser les changements d'expression de façon dynamique. Les traits en pointillés séparent les GRNs en 7 parties (i à vii) regroupant des gènes co-exprimés. Les gènes avec des fonctions similaires (d'après la classification GO pour Gene Ontology) sont groupés dans un même cercle (outils bioinformatique DAVID, $p < 0.05$).

Figure 2. Validations des prédictions. a. Test de croissance indépendante à l'ancrage en milieu agar mou. Toutes les conditions de cellules BJELM transfectées, mis à par le contrôle, montrent une diminution drastique de leur capacité à former des colonies en milieu agar mou.

b. Colonies formées par les cellules BJELM après 3 semaines d'incubation sur un milieu agar mou.

Figure 3. Dynamique des domaines d'association de la chromatine (TADs) lors de la tumorigenèse par étape. a. Comparaison statistique de TADs uniques et communs entre les cellules BJ, BJEL, et BEJLM. **b.** Stabilité de la taille des TADs lors de la transformation des cellules. Les différences statistiquement significatives ont été confirmées par le test Kolmogorov-Smirnov, p-value < 0.001. **c.** Exemples correspondant aux TADs les cellules BJ, BJEL ou BJELM. **d et e.** Visualisation des changements des fréquences d'interactions intra-domaine dans le cas de TADs stables lors de la différenciation cellulaire. Les flèches jaunes mettent en évidence les différences dans l'architecture des domaines entre les différentes conditions. Dans c et d les cartes HiC montrent la fréquence normalisée des interactions. Dans e la différence entre les cartes d'interactions normalisées est présentée.

Figure 4. Dynamique des domaines d'association de la chromatine (TADs) dans les cellules F9 (a) et P19 (b). c et d. Exemples de changements de la fréquence d'interactions intra-domaines lors de la différenciation cellulaire dans le cas de TADs stables. **e.** Comparaison des domaines chromatiniens dans les cellules P19 et F9 non-différenciées. Les flèches jaunes mettent en évidence les différences dans l'architecture des domaines entre les différentes conditions. Dans a, b et e, les cartes HiC montrent la fréquence normalisée des interactions. Dans c et d, la différence entre les cartes d'interactions normalisées est présentée.

Figure 5. Reconstruction d'un réseau de régulation génique étendue (eGRN pour extended GRN). a. Représentation schématique des principes d'intégration. **b.** Modèle temporel de transduction du signal pour l'évaluation de la cohérence entre les eGRNs reconstruits et les changements temporels d'expression génique. **c.** Prédiction des facteurs de transcription et GAPs (Genome Associated Platforms) clés par le modèle de propagation du signal. Les nœuds et les GAPs sont classifiés en fonction de leur capacité à induire la totalité des programmes de différenciation P19-spécifiques. Les GAPs présentant de fortes capacités de reconstruction sont indiqués en rouge. Les facteurs de transcriptions présentant de fortes capacités de reconstruction sont indiqués en bleu. **d.** Portion de eGRN reconstruite montrant l'exemple de la propagation du signal au travers des connections des facteurs de transcription *Pax6*, *Neurod1*, *Zfp516* et GAPs. **e.** Interactomes de *Pax6* et *Zfp516* dans un contexte épigénétique.

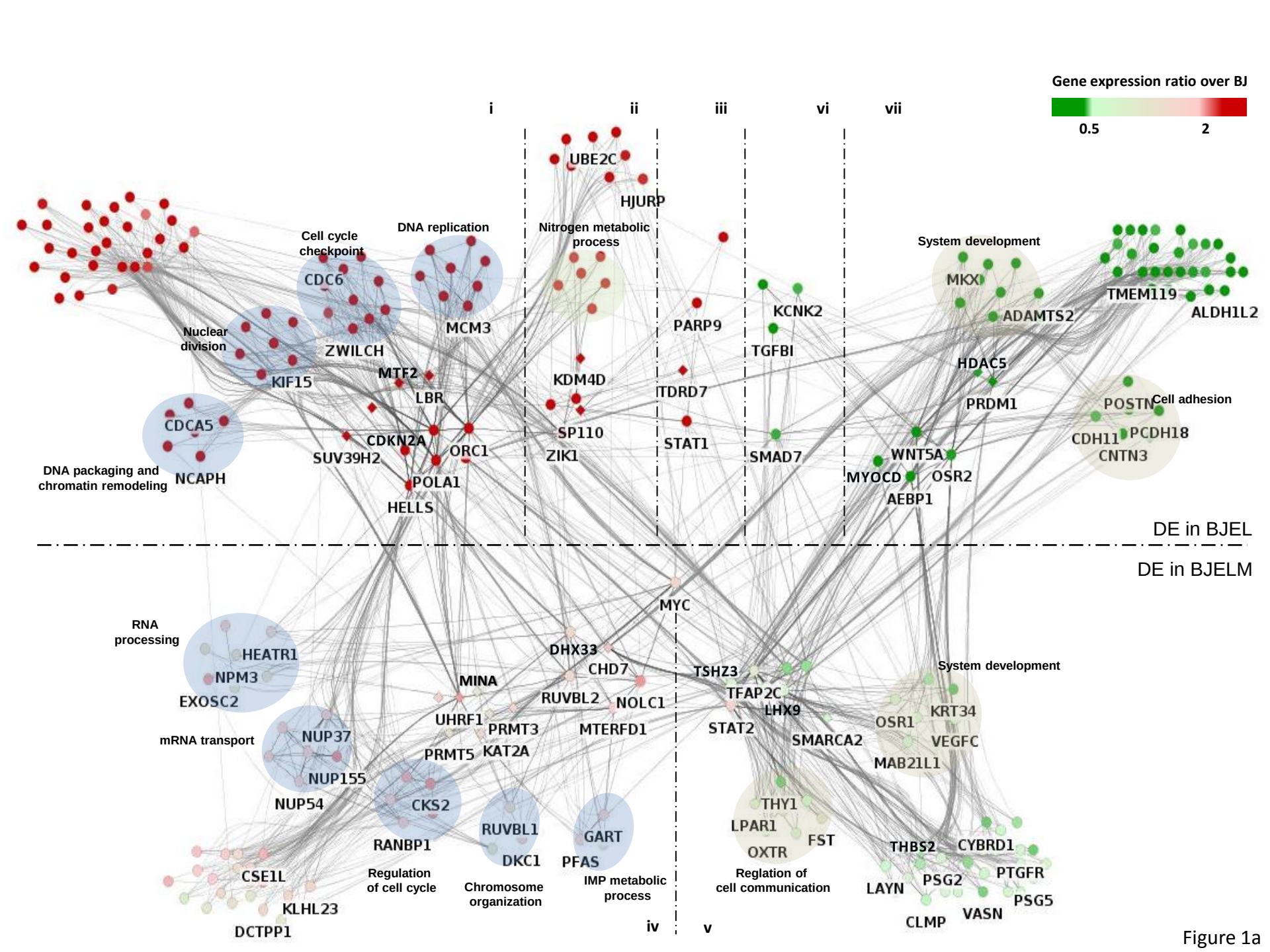

Gene expression ratio over BJ
0.5
2
i
ii
iii
vi
vii
UBE2C
HJURP
Cell cycle checkpoint
DNA replication
Nitrogen metabolic process
System development
CDC6
MCM3
MKX
TMEM119
ALDH1L2
ADAMTS2
Nuclear division
ZWILCH
KIF15
MTF2
LBR
KDM4D
PARP9
KCNK2
TGFBI
HDAC5
PRDM1
TDRD7
STAT1
SMAD7
POSTN
Cell adhesion
CDH11
PCDH18
CNTN3
CDCA5
NCAPH
DNA packaging and chromatin remodeling
CDKN2A
SUV39H2
ORC1
POLA1
HELLS
SP110
ZIK1
WNT5A
MYOCD
OSR2
AEBP1
DE in BJEL
DE in BJELM
MYC
RNA processing
HEATR1
NPM3
EXOSC2
DHX33
CHD7
MINA
RUVBL2
NOLC1
UHRF1
PRMT3
MTERFD1
PRMT5
KAT2A
TSHZ3
TFAP2C
LHX9
STAT2
SMARCA2
System development
OSR1
KRT34
VEGFC
MAB21L1
mRNA transport
NUP37
NUP155
NUP54
CKS2
RANBP1
RUVBL1
DKC1
GART
PFAS
THY1
LPAR1
OXTR
FST
THBS2
CYBRD1
PTGFR
LAYN
PSG2
PSG5
CLMP
VASN
CSE1L
KLHL23
DCTPP1
Regulation of cell cycle
Chromosome organization
IMP metabolic process
Reglation of cell communication
iv
v

Figure 1a

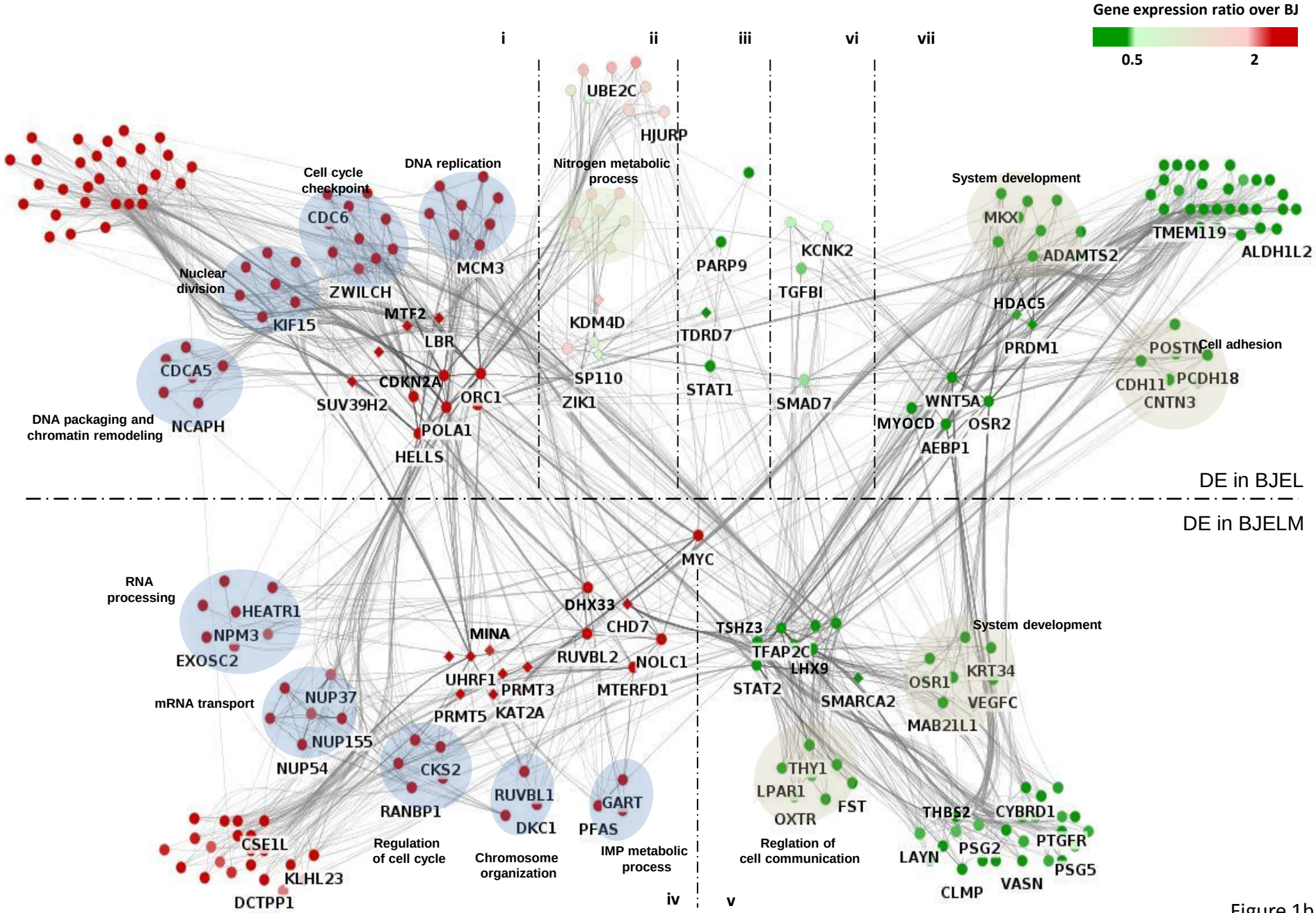

Gene expression ratio over BJ
0.5
2
i
ii
iii
vi
vii
UBE2C
HJURP
Nitrogen metabolic process
DNA replication
Cell cycle checkpoint
CDC6
MCM3
ZWILCH
Nuclear division
KIF15
MTF2
LBR
CDCA5
CDKN2A
ORC1
SUV39H2
NCAPH
DNA packaging and chromatin remodeling
POLA1
HELLS
KDM4D
SP110
ZIK1
PARP9
TDRD7
STAT1
KCNK2
TGFBI
SMAD7
System development
MKX
ADAMTS2
HDAC5
PRDM1
WNT5A
MYOCD
OSR2
AEBP1
TMEM119
ALDH1L2
POSTN
Cell adhesion
CDH11
PCDH18
CNTN3
DE in BJEL
DE in BJELM
MYC
RNA processing
HEATR1
NPM3
EXOSC2
DHX33
CHD7
MINA
RUVBL2
NOLC1
UHRF1
PRMT3
PRMT5
KAT2A
MTERFD1
mRNA transport
NUP37
NUP155
NUP54
CKS2
RANBP1
RUVBL1
DKC1
GART
PFAS
CSE1L
KLHL23
DCTPP1
Regulation of cell cycle
Chromosome organization
IMP metabolic process
iv
v
TSHZ3
TFAP2C
LHX9
STAT2
SMARCA2
THY1
LPAR1
FST
OXTR
Reglation of cell communication
System development
OSR1
KRT34
VEGFC
MAB21L1
THBS2
CYBRD1
PTGFR
LAYN
PSG2
PSG5
CLMP
VASN

Figure 1b

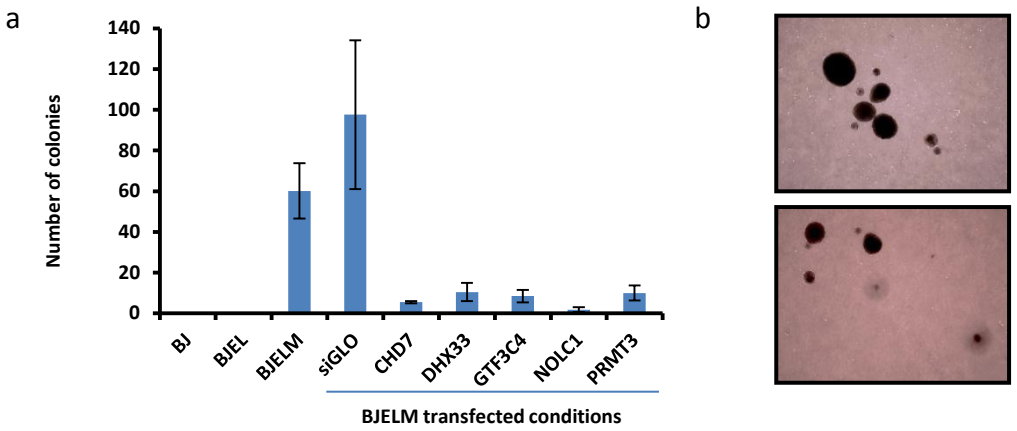

a
b
Number of colonies
140
120
100
80
60
40
20
0
BJ
BJEL
BJELM
siGLO
CHD7
DHX33
GTF3C4
NOLC1
PRMT3
BJELM transfected conditions

Figure 2

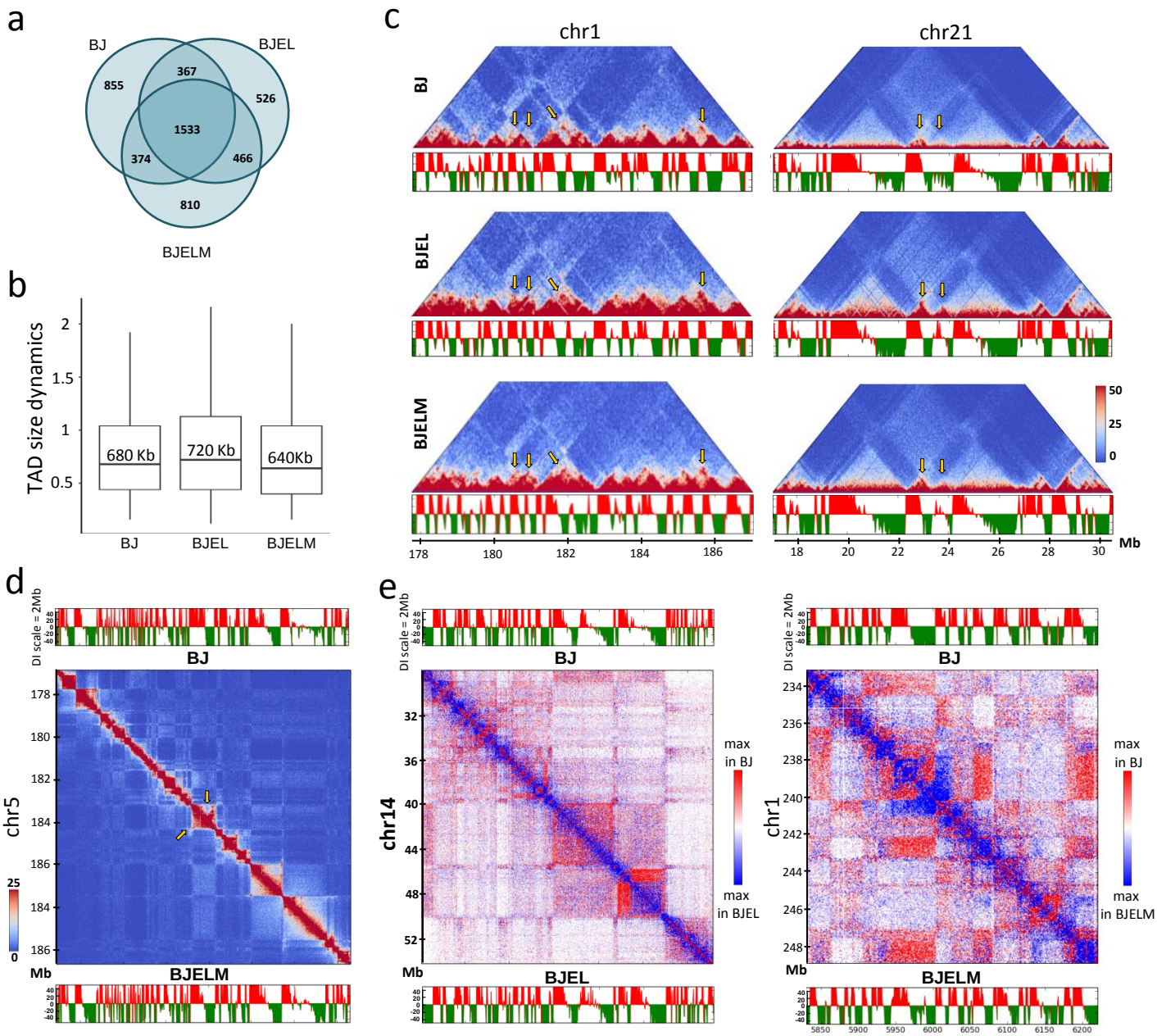

a
BJ
BJEL
BJELM
855
367
526
1533
374
466
810
b
TAD size dynamics
2
1.5
1
0.5
680 Kb
720 Kb
640Kb
BJ
BJEL
BJELM
c
chr1
chr21
BJ
BJEL
BJELM
50
25
0
178
180
182
184
186
18
20
22
24
26
28
30
Mb
d
DI scale = 2Mb
BJ
chr5
178
180
182
184
186
25
0
Mb
BJELM
e
DI scale = 2Mb
BJ
chr14
32
36
40
44
48
52
max
in BJ
max
in BJEL
Mb
BJEL
DI scale = 2Mb
BJ
chr1
234
236
238
240
242
244
246
248
max
in BJ
max
in BJELM
Mb
BJELM
5850
5900
5950
6000
6050
6100
6150
6200

Figure 3

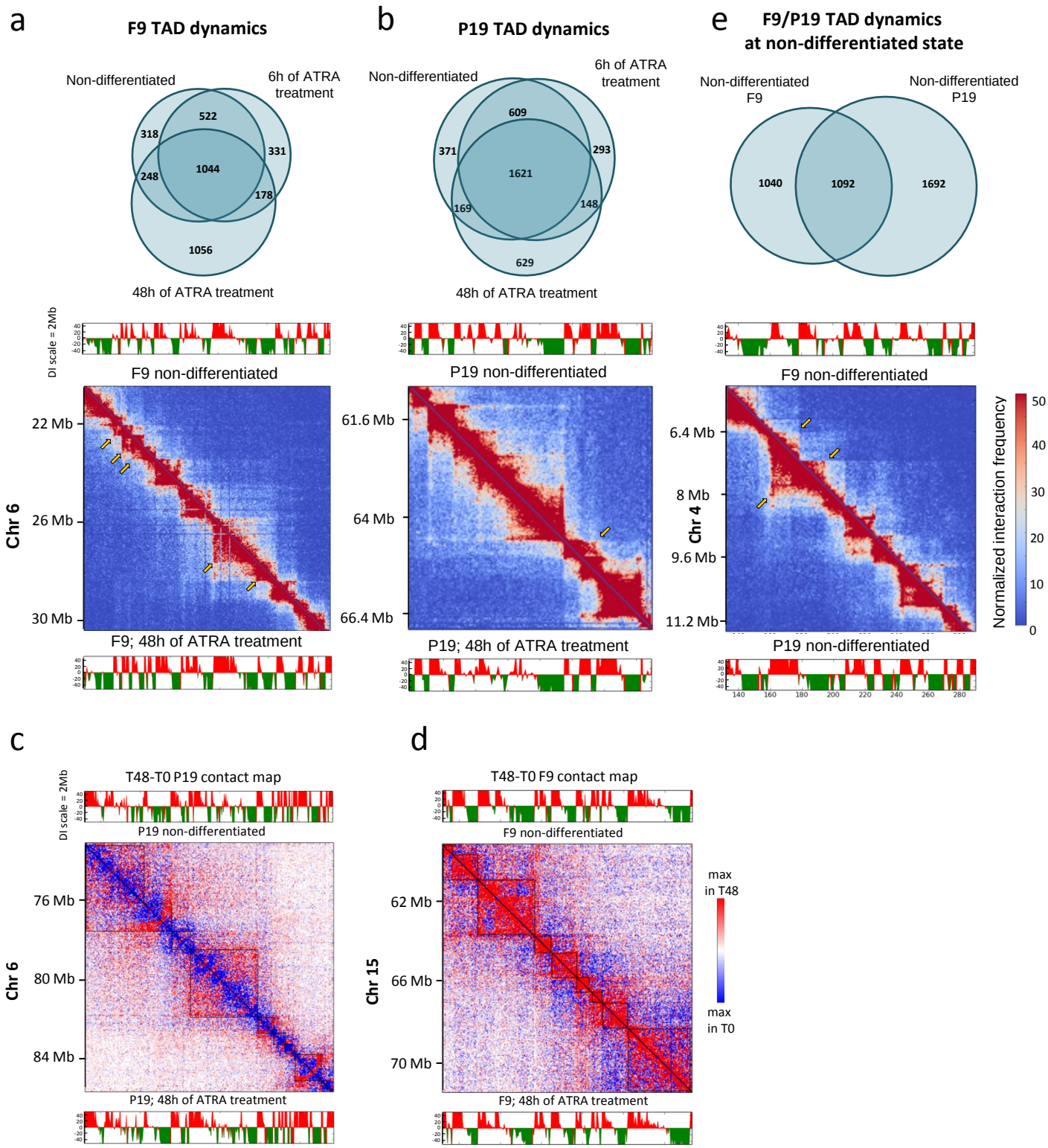

a
F9 TAD dynamics
Non-differentiated
6h of ATRA treatment
48h of ATRA treatment
522
318
331
1044
248
178
1056
b
P19 TAD dynamics
Non-differentiated
6h of ATRA treatment
48h of ATRA treatment
609
371
293
1621
169
148
629
e
F9/P19 TAD dynamics at non-differentiated state
Non-differentiated F9
Non-differentiated P19
1040
1092
1692
DI scale = 2Mb
F9 non-differentiated
P19 non-differentiated
F9 non-differentiated
Chr 6
22 Mb
26 Mb
30 Mb
61.6 Mb
64 Mb
66.4 Mb
Chr 4
6.4 Mb
8 Mb
9.6 Mb
11.2 Mb
F9; 48h of ATRA treatment
P19; 48h of ATRA treatment
P19 non-differentiated
Normalized interaction frequency
50
40
30
20
10
0
c
T48-T0 P19 contact map
DI scale = 2Mb
P19 non-differentiated
Chr 6
76 Mb
80 Mb
84 Mb
P19; 48h of ATRA treatment
d
T48-T0 F9 contact map
F9 non-differentiated
Chr 15
62 Mb
66 Mb
70 Mb
F9; 48h of ATRA treatment
max in T48
max in T0

Figure 4

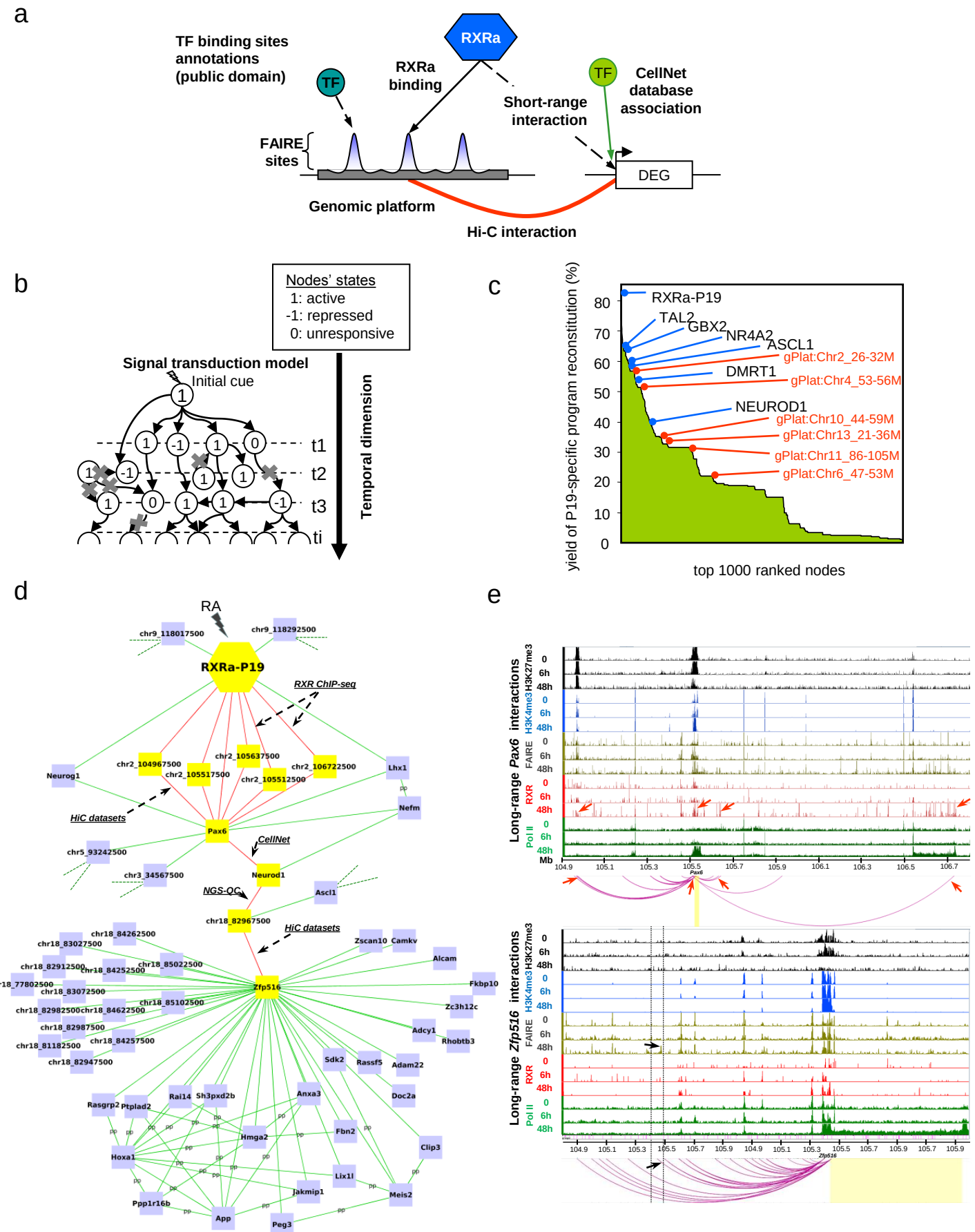

a
TF binding sites annotations (public domain)
RXRa
RXRa binding
TF
Short-range interaction
TF
CellNet database association
FAIRE sites
DEG
Genomic platform
Hi-C interaction
b
Nodes' states
1: active
-1: repressed
0: unresponsive
Signal transduction model
Initial cue
Temporal dimension
t1
t2
t3
ti
c
yield of P19-specific program reconstitution (%)
top 1000 ranked nodes
RXRa-P19
TAL2
GBX2
NR4A2
ASCL1
gPlat:Chr2_26-32M
DMRT1
gPlat:Chr4_53-56M
NEUROD1
gPlat:Chr10_44-59M
gPlat:Chr13_21-36M
gPlat:Chr11_86-105M
gPlat:Chr6_47-53M
d
RA
RXRa-P19
RXR ChIP-seq
HiC datasets
CellNet
NGS-QC
HiC datasets
Pax6
Zfp516
Neurog1
Lhx1
Nefm
Neurod1
Ascl1
Zscan10
Camkv
Alcam
Fkbp10
Zc3h12c
Adcy1
Rhobtb3
Adam22
Rassf5
Sdk2
Doc2a
Anxa3
Sh3pxd2b
Rai14
Ptplad2
Rasgrp2
Fbn2
Hmga2
Clip3
Hoxa1
Lix1l
Jakmip1
Meis2
Ppp1r16b
App
Peg3
e
Long-range Pax6 interactions
Long-range Zfp516 interactions
H3K27me3
H3K4me3
FAIRE
RXR
Pol II
Mb

Figure 5

PUBLICATIONS

1. Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis. Malysheva V., Mendoza-Parra M. A., Mohamed Saleem M. A. and Gronemeyer H., Genome Medicine. (2016) 8:57.

2. Mendoza-Parra M. A., Malysheva V., Mohamed Saleem M. A., Lieb M., Godel A., and Gronemeyer H.. Reconstructed cell fate-regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis. *Genome Research.* (2016). doi:10.1101/GR.208926.1164.

3. LOGIQA: A database dedicated to Long-range Genome Interactions Quality Assessment Mendoza-Parra M. A., Blum M., Malysheva V., Cholley P.-E. and Gronemeyer H. BMC Genomics. (2016) 17:355

4. Chromatin structure dynamics directs cell fate acquisition. Malysheva V., Mendoza-Parra M. A.*, Blum M. and Gronemeyer H*. *Manuscript in preparation*

5. Chromatin dynamics during tumorigenic transformation. Malysheva V., Blum M., Mendoza-Parra M. A. and Gronemeyer H*. *Manuscript in preparation*

COMMUNICATIONS

1. Presentation and poster – 3rd international FASEB Conference on Retinoids. June 19-24, 2016. West Palm Beach, Florida, USA. Chromatin architecture in cell fate decision processes. (Malysheva V., Mendoza-Parra M. A., Blum M. and Gronemeyer H*). Best Poster Prize.

2. Poster – 6th EMBO meeting. September 5-8, 2015. Birmingham, UK. Reconstructing gene regulatory networks of tumorigenesis. (Malysheva V., Mendoza-Parra M. A., Mohamed Saleem M. A. and Gronemeyer H.)

3. Poster - EMBO conference: From Functional Genomics to Systems Biology. November 8-11, 2014. EMBL, Heidelberg, Germany. Systems Biology of Retinoic Acid-induced cell fate transitions. (Mendoza-Parra M. A., Malysheva V., Mohamed Saleem M. A. and Gronemeyer H.)

4. Poster - Conference: Chromatin and Epigenetics: From Omics to Single Cells. October 14-15, 2014. IGBMC, Strasbourg, France. Systems Biology of Retinoic Acid-induced cell fate transitions. (Mendoza-Parra M. A., Malysheva V., Mohamed Saleem M. A. and Gronemeyer H.)

5. Poster – IGBMC Poster Session, "Changes in chromatin structure during tumorigenesis: Setting up the study" (Malysheva V., Mendoza-Parra M. A. and Gronemeyer H.). March, 2014**.** Prize for the best poster voted by the scientific committee.

Valeriya MALYSHEVA

RECONSTRUCTION DES RESEAUX DE REGULATION GENIQUES RESPONSABLES DU DESTIN CELLULAIRE

Résumé

L'établissement de l'identité cellulaire est un phénomène très complexe qui implique pléthore de signaux instructifs intrinsèques et extrinsèques. Cependant, malgré les progrès importants qui ont été faits pour l'identification des régulateurs clés, les liens mécanistiques entre facteurs de transcription, épigénome, et structure de la chromatine lors de la différenciation cellulaire, et de la transformation tumorigénique des cellules, sont peu connus. Pour résoudre ces problématiques nous avons utilisé deux modèles de transition de l'identité cellulaire : la différenciation neuronale et endodermique induites par un même morphogène, l'acide rétinoïque. Concernant la transformation tumorale des cellules nous avons utilisé un système de tumorigenèse par étape de cellules primaires humaines. Nous avons conduit des études intégratives incluant des données transcriptomiques, épigénomiques, et des données concernant l'architecture de la chromatine. Notre approche systématique pour caractériser l'acquisition de l'identité cellulaire, combinée à la modélisation de la transduction du signal, renforce donc nos connaissances sur les mécanismes responsables de la plasticité cellulaire. Une meilleure compréhension des mécanismes régulateurs de l'identité cellulaire non seulement nous éclaire sur les relations de cause à effet entre les différents niveaux de régulation dans la cellule, mais aussi ouvre de nouvelles possibilités en terme de transdifférenciation dirigée.

Mots clés : Identité cellulaire, Tumorigenèse, Biologie des systèmes, Réseaux de régulation géniques.

Résumé en anglais

The cell fate acquisition is a highly complex phenomenon that involves a plethora of intrinsic and extrinsic instructive signals. However, despite the important progress in identification of key regulatory factors of this process, the mechanistic links between transcription factors, epigenome and chromatin structure which coordinate the regulation of cell differentiation and deregulation of gene networks during cell transformation are largely unknown. To address these questions for two model systems of cell fate transitions, namely the neuronal and endodermal cell differentiation induced by the morphogen retinoic acid and the stepwise tumorigenesis of primary human cells, we conducted integrative transcriptome, epigenome and chromatin architecture studies. Through extensive integration with thousands of available genomic data sets, we deciphered the gene regulatory networks of these processes and revealed new insights in the molecular circuitry of cell fate acquisition. The understanding of regulatory mechanisms that underlie the cell fate decision processes not only brings the fundamental understanding of cause-and-consequence relationships inside the cell, but also open the doors to the directed trans-differentiation.

Key words: Cell fate, Tumorigenesis, Systems Biology, Gene Regulatory Networks