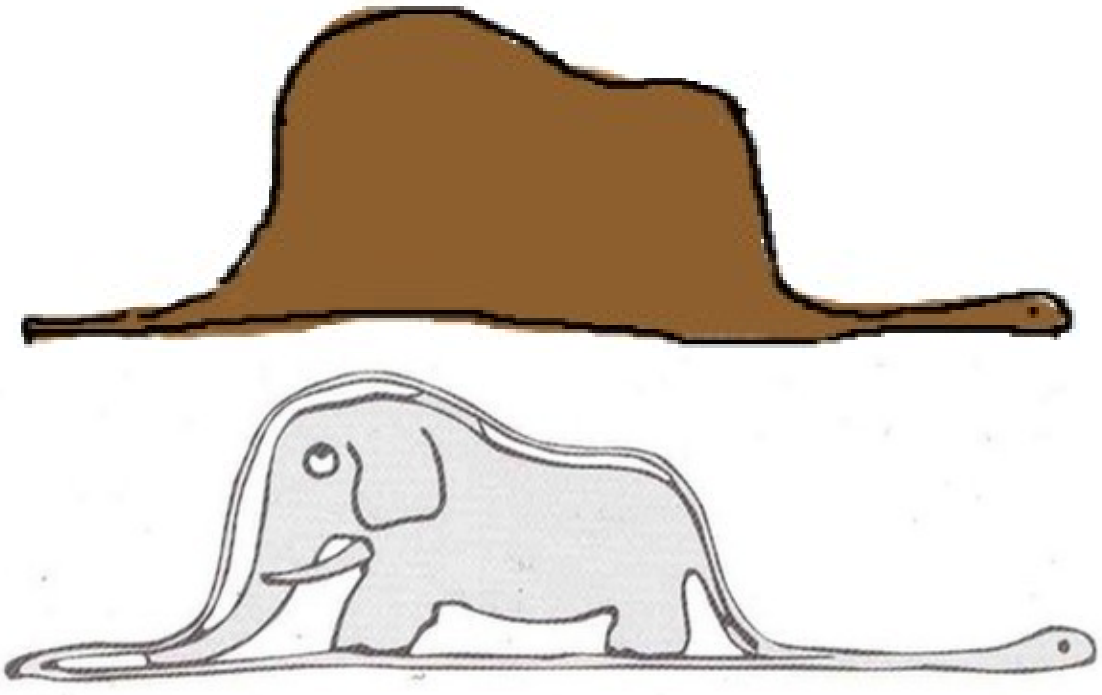
Master1 EBE, UPMC

Superresolution insights into the nuclear localization of the c-Myc oncogene

February – June 2014

Maxime Woringer (maxime.woringer@ens.fr)
Under the supervision of Lana Bosanac,
Tjian Lab,
University of California, Berkeley



Superresolution insights into the nuclear localization of the c-Myc oncogene

Maxime Woringer

June 14, 2014

Contents

Introduction

The proto-oncogene c-Myc is one of most often disregulated protein involved in cancers. It was also the first characterized oncogene, more than thirty years ago. Therefore, extensive literature describes multiple functions of the c-Myc protein. Indeed, c-Myc is involved in mechanisms as diverse as differentiation (it is one of the four transcription factors isolated by the Yamanaka lab to generate induced pluripotent stem cells – iPSc [1]), chromatin remodelling or cell cycle and apoptosis.

Furthermore, no inclusive framework accounting for these functions had been proposed until recently. Indeed, the mechanisms by which c-Myc promotes reprogramming either towards tumorigenesis or differentiation are a field of hectic activity, but the mechanisms remain elusive.

Recently proposed mechanisms see c-Myc as a "global genome amplifier" [2] that regulate with little specificity as much as 15% of the genome, modulates RNA polymerase II pausing at promoters and localizes at euchromatin regions. Another approach describes c-Myc as a "global chromatin remodeller" [3], a factor able to trigger the creation of an open-chromatin environment suitable for gene activation for instance through the recruitment of the histone acetylase GCN5 [4]. Moreover, c-Myc exhibits an *in vitro* affinity for the so-called and almost ubiquituous "E-box" `CACGTG` DNA sequence.

All in all, traditional biochemistry and bioinformatics have not succeeded in proposing a rea-

sonable model for c-Myc activity and interaction. Indeed, these techniques have trouble to identify preferential interactions and binding sites, suggesting that some factors regulating these interactions remain unknown. At the global nuclear scale, the mechanisms by which c-Myc regulate the cell are yet to be unveiled.

Although the nuclear protein landscape is well studied, little is known about the influence of nuclear organization and architecture on transcription factor activity [5]. The nucleus is indeed a crowded bioreactor of complex geometry that can catalyze the formation of local gradients and enrichments amenable of interfering with protein interactions [6]. Indeed, PolII labile clusters have recently been described, and the functions of these spatial regulations still await investigation [7].

Since growing evidence suggest a very disperse but yet non homogeneous action mechanism of c-Myc across the genome, it is of high interest to map its positions and dynamics in the mammalian nucleus. We used a combination of superresolution imaging and biochemistry to first characterize c-Myc nuclear patterns and to assay its biological significance.

Overviews *The* MYC *gene and c-Myc protein*[1]: The human *MYC* gene is located on chromosome 8 and cromprises 3 exons. Several splicing variants have been reported (UCSC genome browser), but notably, they almost always carry the 5' final part of the *MYC* gene.

c-Myc proteins are the translation products of the *MYC* gene. c-Myc can be found in two main isoforms of respective molecular masses of 50.6k and 48.9k, the two differing by a mere 15 amino-acids difference at the N-ter end. This truncation (there is a non-canonical translation starting site CUG 45 bp upstream the canonical AUG) has been linked to several cancers. The usual numbering of the amino-acids of c-Myc starts at the canonical translation starting site. Although we performed our work using the full-length isoform of c-Myc, we stick to this nomenclature, and designate the N-ter 15 aminoacids by -14 to 0. c-Myc is a transcription factor member of the bHLH-LZ family (it harbors a basic-Helix-Loop-Helix-Leucine-Zipper domain at the C-ter end, Figure 1).

Several interacting domains have been characterized (Figure 1). The main ones are:

- The *MBI* and *MBII* domains are part of the transactivating domain. MBII is responsible for the recruitment of the TRRAP protein and subsequently for the recruitment of the GCN5 histone acetyltransferase.

- The *bHLH* domain is required for c-Myc to interact with DNA. DNA-binding is conditioned by dimerization with the MAX protein (or other membrers of the bHLH family). Cristal structure of the c-Myc/MAX heterodimer in complex with DNA and site-directed mutagenesis show that the basic domain is involved in the interaction with DNA. The *HLH* domain is involved in dimerization.

- Another DNA binding domain (termed *MBIV*) has also been described [9].

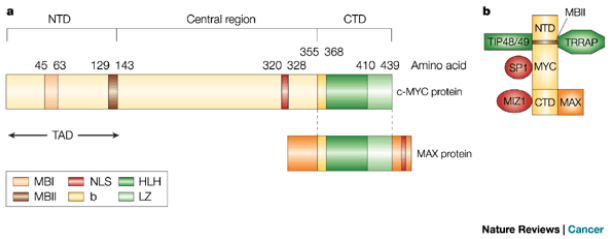


Figure 1: Structure of the c-Myc protein and main documented interactions. Except from [10], Figure 1. NTD, CTD: N-terminal domain and C-terminal domain, resp. *(a).* linear organization of c-Myc and MAX. *(b).* c-Myc partners and their interacting domains.

Both over- and underexpression of c-Myc have been reported to cause dramatic effects to the cell. Transfection of 500 ng of constructs expressing c-Myc and Ras are used as a transformation assay [9]. Conversely, underexpression of c-Myc triggers

[1]. We limit our presentation to the relevant features for our study, although extensive literature exist on the subject (such as [8]).

major changes in histone modifications (especially acetylation) [11]. This feature is a challenge for microscopy where transient transfections are usually used.

1 Methods

1.1 Microscopy

Resolving spatial structures of nuclear factors can be done with a confocal microscope, allowing for the determination of heterogeneities with a micrometer scale (that is, 10% of the radius of a standard nucleus). However, many nuclear factors seem approximately homogeneous at that scale, including DNA and RNA polymerase II, even though they exhibit clusturing at higher magnifications [7, 12]. Furthermore, the micron-scale is not relevant for most of the mechanisms involved in transcription regulation, and it has been shown by various methods that some proteins in the nucleus exhibit confinment at scales around 100 nm [6].

In addition, the resolution of colocalizations in the nucleus is also very limited with a confocal microscope. We thus used so-called "superresolution" imaging techniques [13].

1.1.1 Confocal

Although the resolution of certain nuclear substructures require superresolution, some other were of large size, and thus visible with a standard confocal microscope (Zeiss LSM 710). We especially used the confocal to compare staining intensities and to quantify gradients, since a very low pinhole aperture allows to image only a slice of given section, correcting for potential inhomogeneous cell height.

Images were acquired with a 40x Plan-Apochromat oil objective of 1.4 numerical aperture (DIC M27, Zeiss) and quantification was realized using the ImageJ software [14].

1.1.2 Superresolution microscopy

Confocal microscopy (and also traditional bright field microscopy) resolution is intrisically limited by light diffraction. Indeed, light beams do not behave as described in geometrical optics theory where a lens can focus parallel rays in a point of null volume. Focalization of a light beam is physically limited, and the image of a point light source (such as a single fluorescent molecule) through a set of lenses (a microscope) will give a gaussian-like blur (called point spread function, or PSF), whose standard deviation only depends on the wavelength (λ) and the numerical aperture (NA) of the objective: $\frac{\lambda}{NA}$. Thus, two molecules closer than this distance cannot be resolved if imaged at the same time.

Consequently, resolving two molecules closer than the diffraction limit require to image first one of the molecules, and after it bleaches, image the second one (Figure 2a). Each isolated PSF can then be fitted with a gaussian to determine its centroid, allowing localization of the underneath fluorophore with a nanometric resolution (usually 15-30 nm, Figure 2c). This resolution only depends on the number of photons collected ($N_{photons}$): the pointing accuracy is given by the standard error on the mean and scales as $\frac{1}{\sqrt{N_{photons}}}$. The coordinates of the individually fitted PSF from thousands of images from the same sample are then combined to build a "pointillist", superresolution picture (Figure 2c). They can be mathematically described as a "spatial point pattern" (see Section 4.2.1. We used two main techniques to observe single molecules and thus perform superresolution microscopy:

- *PALM* (photoactivated localization microscopy, [15]) uses photoactivatable/photoconvertible fluorescent proteins as probe.
- *dSTORM* (direct stochastic optical reconstruction microscopy): The protein of interest is stained with an organic dye (such as Alexa

647 [16], TMR, SiR [17], etc.) through specific antibodies or HaloTags (see below).

These two techniques allow to image a single focal plane. However, precise spatial correlations, and other important features such as the fractal dimension of the reconstructed point pattern are inaccessible with 2D data. 3D acquisition (a 1 µm thick slice resolved in 3D) are realized by placing a deformable mirror between the objective and the camera [18] (see Figure 2b). This adaptive optics device allows: (1) To precisely "shape" the PSF to correct for the aberrations introduced by a non-perfect optical path. This ensures that the observed PSF is as closed as possible from the theoretical, diffraction-limited Airy function. (2) To purposely introduce aberrations. One of them, astigmatism, differentially shapes the PSF respective to its distance to the focal plane (in z). The shaped PSF are progressively horizontally (respectively vertically) elongated as the observed molecule lies above (respectively below) the focal plane (Figure 2c.3). These 2D gaussian-like functions can be fitted using a previously described code, allowing to recover the real z coordinate of the detected molecule.

The subsequent fitting of the PSF in the movies was realized using a modified version of MTT [19], a set of Matlab scripts. The program fits each local maximum in the image with a 2D gaussian ($g(x,y) = Ae^{-(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})}$) and records the A, σ_x and σ_y parameters in addition to to the x and y localization. The images are then drift-corrected by tracking over time the positions of a bead visible in the field of view (Figure 2c2). We wrote a script to integrate the various steps of the analysis, where minimal user input is required.

1.1.3 Single particle tracking

We performed 2D single particle tracking using the microscope and setup described in section 1.1.2 and Figure 2b. Live cells transfected with a very low amount of fusion protein fused to a HaloTag were

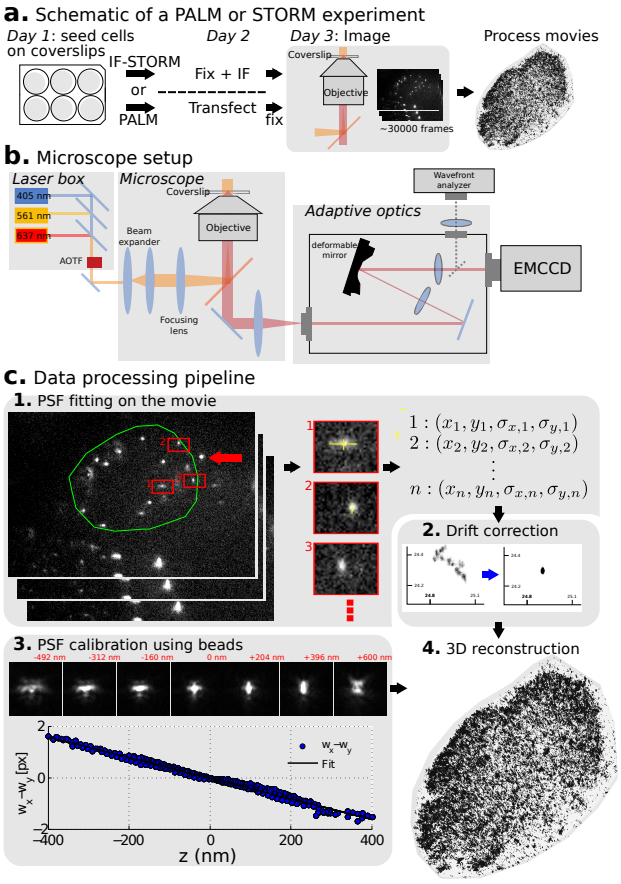


Figure 2: Data acquisition and processing during a PALM or STORM experiment. (a). Pipeline of a PALM or STORM experiment. IF-STORM: Immunofluorescence followed by STORM imaging(b). Microscope setup. The adaptive optics part (on the right) give the ability to shape PSFs and thus to image in 3D. (c).1: Local maxima are extracted from the movie and a 2D gaussian is fitted to the spot. The point is discarded if the fit quality is too low. The parameters of the gaussian are then extracted and the image is drift-corrected. 2: Thanks to a calibration on a bead (3), real z coordinates are back-calculated, leading to the 3D map (4).

stained with TMR and imaged in conditions where at most one particle per frame is visible in the nucleus and $\sim$ 10000 frames were recorded at 100 frames per second (10 ms/frame).

The data was subsequently analyzed using custom code available in the lab [20]. This code is able to detect particles in very noisy conditions, where PSFs appear convolved by motion blurs. The algorithm applies stringent conditions to ensure that one trace represents only one particle.

1.1.4 Labeling strategy

Since each labeling strategy has its own specificities, we made use of several techniques to achieve acceptable data quality: full nucleus superresolu-

tion pictures were achieved with the organic dye Alexa647 conjugated to an antibody in fixed cells or with transfected Dendra2-tagged constructs. SPT was performed using low levels of transfection of c-Myc-HaloTag and live staining with TMR. The rationale behind the choice of fluorophores is explicited in 4.1.1.

Additional methods, including tranfection, immunofluorescence, imaging conditions, mathematical analysis and formalization of the data are described in section 4.1.

1.2 Cell culture

1.2.1 NTera2 cell line and differentiation

Previous studies document a crucial involvement of c-Myc in both differentiation and iPSC induction. However, most of the available models to study differentiation and induction are technically difficult to handle. Furthermore, imaging imposes additional constraints on the choice of the animal model. Most of the mouse-derived cells show very high level of autofluorescence, human ES cells are known to attach very weakly to coverslips, making the handling and preparation of microscopy experiments more challenging. On the other hands, the traditional cell lines used for imaging such as HeLa or U2OS are not very suitable to study c-Myc regulation. Indeed, since c-Mys is disregulated in most of the cancers, and provided that many if not most cell lines derive from tumor cells, it is necessary to assay to what extent c-Myc is disregulated before starting real experiments. For instance, in U2OS, c-Myc levels are five times higher than in normal cells [21].

We thus used the NTera2 (NT2) cell line available in the lab. This cell line present the interesting properties that one can trigger differentiation from an "embryonic carcinoma" (EC) towards a neuron-like cell fate upon addition of retinoic acid to the medium [22].

Cells were differentiated by adding 10 µg/ml retinoic acid (RA) to the medium (Figure 3).

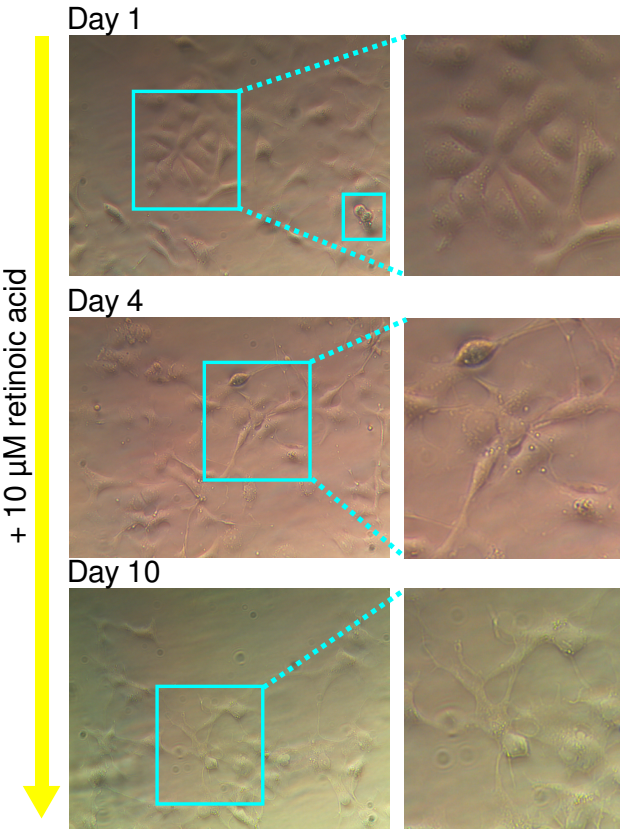


Figure 3: NT2 cells differentiate in less than 10 days upon addition of retinoic acid. 1 µM of retinoic acid was added to the medium. After a couple of days, the cells lose their elliptic shape and exhibit a neuron-like phenotype. Cells are approximately 15-20 µm width.

Medium was changed every two other day. Previous characterization of this cell line in the lab indicate that cells achieve differentiation in less than seven days (as assayed by the drop in Oct4 levels, Lana Bosanac, unpublished results). Differentiated cells were thus imaged 7-10 after adding RA to the medium.

Unless specified, all experiments were conducted on NT2 cells.

1.2.2 U2OS Dendra2-Rpb1

Additional experiments were conducted on a human osteosarcoma, U2OS-derived cell line (gift of Xavier Darzacq, Paris, France). This cell line expresses a Dendra2-tagged α-amanitin resistant Rpb1 subunit of the RNA polymerase II, allowing replacement of the endogeneous PolII by its tagged counterpart upon addition of α-amanitin to the medium.

1.3 Constructs and mutagenesis for microscopy

1.3.1 *MYC* cloning

To our knowledge, no commercially available plasmid include the alternative translation site from the CUG codon. Thus, we cloned the *MYC* gene from the cDNA. Starting from whole RNA extracts from NT2 cells, a reverse transcription (SuperScript™ II, Invitrogen) followed by PCR gave the required insert (Figure 4a), that was then cloned into several vectors (see below). To ensure expression from an exogenous vector, we replaced the non-canonical TSS starting codon CUG by the traditional AUG.

1.3.2 Fluorescent fusions

In order to realize microscopy acquisition, c-Myc has to be tagged with a fluorescent protein or dye. In the case of superresolution experiments, specific dyes and proteins are required.

We tried two labeling strategies in parallel, one suitable to do STORM imaging, the other for PALM (Figure 4b). All fusions involve the C-ter end of c-Myc:

SPT we fused a genetically encoded HaloTag® [23] to the C-ter end of c-Myc (*MYC* was inserted between the *SgfI* and *SacI* restriction site of the pFC17A HaloTag® CMVd3 Flexi® vector, Promega). The HaloTag is an engineered haloalkane dehalogenase that binds covalently under physiological conditions small organic molecules called Halo-ligands. Halo-ligands fused to variable organic dyes can then be purchased, ensuring specific labeling of the HaloTag-ed fusion proteins.

PALM c-Myc was fused to the photoconvertible protein Dendra2 ([24], *HindIII* and *PmeI* restriction sites into the pDendra2-N vector, Clontech)

In addition to these two main constructs, two additional cloning steps generated a GFP and a YFP c-Myc fusion, particularly suitable for FRAP (Fluorescence Recovery After Photobleaching) experiments (pGFP-N and pYFP-N, Clontech, same restriction sites as for the Dendra2 vector).

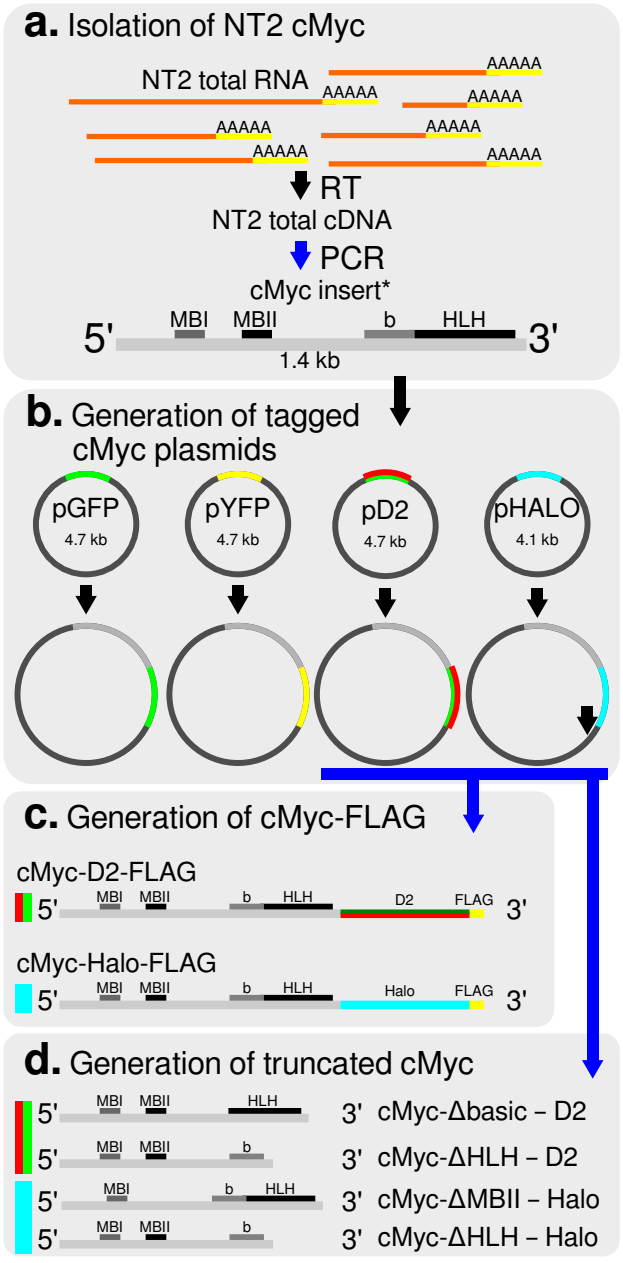


Figure 4: Cloning strategy employed to derive the vectors needed for microscopy and biochemistry assays. Blue arrows denote PCR step (followed by ligation in the case of the generation of the mutants. Unless specified, black arrows indicate traditional restriction/ligation cloning.

To facilitate the biochemistry experiments required to assay the impact of the fused tag on the c-Myc protein, we inserted by site-directed mutagenesis a FLAG-tag on the C-ter of either the Dendra2 or the HaloTag sequence (Figure 4c).

1.3.3 Site-directed mutagenesis

In order to investigate the domain(s) driving the *in-situ* localization of c-Myc, we tried to derive several mutants, exhibiting deletions of full domains of c-Myc. Deleted sites were chosen according to published reference (see Introfuction), and based on preliminary observations (section 2.4).

The following vectors were engineered, leading to truncated proteins (Figure 4d): c-Myc-ΔMBII (deletion of aminoacids 129–143), c-Myc-Δbasic (amino acids 355–367), c-Myc-ΔHLH (amino-acids 368–410). We tried to generate the double deletion of both the basic region and the helix-loop-helix domain (c-Myc-ΔbHLH mutant) without success.

The deleted segments (from 36 to 126 nt) were too long to proceed with the canonical site-directed mutagenesis involving a pair of overlapping primers containing the mutation of interest, followed by PCR, *DpnI* treatment and transformation. We thus performed a PCR of the backbone (e.g.: to generate ΔMBII, a *forward primer* starting at the 3' end of the MBII domain, and a reverse primer starting at the beginning of the MBII domain, oriented backward), followed by digestion of the bacterial template with *DpnI*. 5' ends were phosphorylated and a blunt self-ligation was performed before transformation.

1.4 Functional assays

In order to assay whether fusion of the tag at the C-ter end of c-Myc impairs its function, we looked for documented interaction by Co-Immunoprecipitation (CoIP). Two to three 10 cm dishes of NT2 cells were cultured and each of them was transfected with 3 µg of the corresponding plasmid using XFect (Clontech). Cells were harvested and frozen 24 to 36 h later. Nuclear extracts were then used for the CoIP. The antibodies we used used are detailed in Supplementary Table 2.

1.5 Generation of CRISPR cell lines

Since the expression level of c-Myc is highly regulated *in vivo*, a characteristic we observed during our transfections, we seeked for a way to express c-Myc at endogenous levels. To do so, we tried to generate a NT2 cell line where at least of one of the alleles of the *MYC* gene is replaced by a tagged version.

Genome editing, and especially knock-ins cell lines in mammalian cells are challenging. Indeed, the methods require to provide a vector containing homology with the targetting region, and to trigger homologous recombination to integrate the sequence into the genome.

Efficent gene editing methods [25] for human cells include recombinant adeno-associated virus (rAAV, [26]), Zinc Finger Nucleases (ZFN, [27]), transcription activator-like effector-based nucleases (TALENs), and clustered regularly interspaced short palindromic repeats (CRISPR, [28, 29]). All these techniques rely on providing the sequence to insert flanked by homologies matching the locus to insert in a donor vector. Under normal conditions, the homologous recombination events required to insert the exogenous target sequence of the donor vector are very rare. This issue can be circumvented by introducing double strand breaks (DSBs) that trigger DNA repair, including homologous recombination. If the DSBs preferentially occur at the locus of interest, they increase specificity and limit off-target recombination, two main concerns in genome editing.

1.5.1 CRISPR outline

Unlike TALENs and ZFNs, CRISPR/Cas9 genome editing do not require to assemble a synthetic DNA binding domain in the protein, that require to assemble a high number (dozens) of repeated fragments. A challenging technique where traditional restriction/ligation cannot be applied, and where specific assembly reactions such as the

Golden Gate reaction [30] are required. Furthermore CRISPR/Cas9 do not rely on virus handling to produce targetting.

Instead, the CRISPR/Cas9 specificity is provided by a small RNA sequence complementary of a portion of the gene of interest. Thus genome editing can be achieved by transfecting the small RNA (called single guide RNA, sgRNA), the Cas9 RNA-dependent nuclease and the donor vector.

More precisely, the CRISPR/Cas9 system derives from bacteria, in which it can be seen as an adaptive defense against foreign DNA. During a first infection, the bacterium integrates fragments of the foreign DNA at a specific locus. When transcribed, the integrated fragment forms the variable part of a longer RNA (called CRISPR RNA, or crRNA) recognized by the endonuclease Cas9. This variable part then allows Cas9 targetting at the foreign DNA where it can perform its endonuclease activity, provoking DSBs and cleaving the foreign fragment. In the endogeneous bacterial system, two single-strand RNAs (a crRNA and a trans-activating crRNA (or tracrRNA) that partially hybridizes with the non variable part of the crRNA, are required to form a catalytically active complex (Figure 5a, left). Fortunately, these two RNAs can be combined in a single fragment [31] called single guide RNA (sgRNA, or simply gRNA) (Figure 5a, right). This sgRNA has recently been optimized for better expression and targetting efficiency [32], and this latter strategy is used in the lab.

With this system, virtually any DNA sequence can be targetted. The only limitation is the presence of a so-called protospacer adjacent motif (PAM sequence, [33]) at the 3' end of the variable part of the RNA. Targetting and cleaving of Cas9 is conditioned by the presence of this (highly represented in the genome) trinucleotide (NGG for the Cas9 used in the lab, N can be any nucleotide) [34]. When interacting with a sgRNA (or a crRNA:trRNA duplex), Cas9 undergoes a conformational rearrangement [35] and exhibits a >10 fold increased affinity for DNA [34]. Homology of ~ 20 to 24 nucleotides are usually provided, although some experiments show that some regions of the variable sequences tolerate mismatches [36].

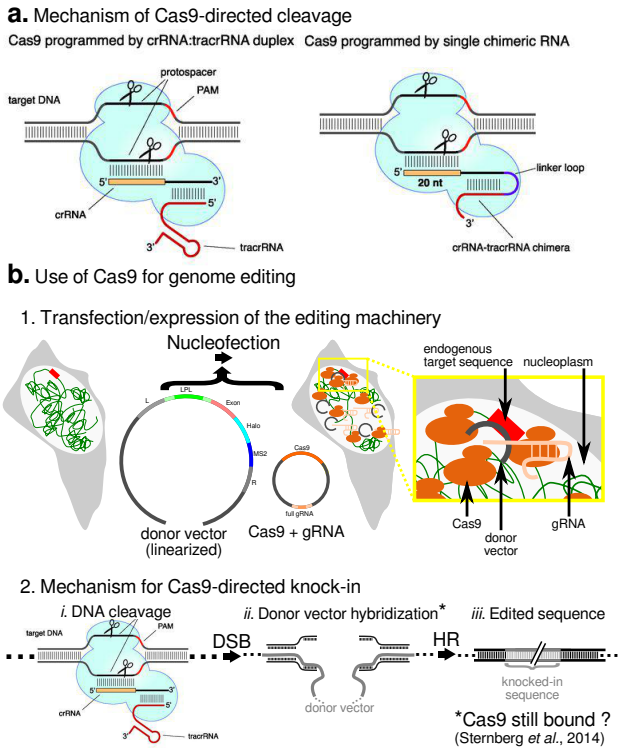


Figure 5: Use of the CRISPR/Cas9 for genome editing. Parts of the figure are derived from [31], figure 5.

1.5.2 Targetting strategy

We tried to generate the following alterations to the *MYC* endogenous locus (Figure 6):

- c-Myc-HaloTag, suitable for STORM, and localization of a subpopulation of the c-Myc pool with high accuracy,
- c-Myc-Dendra2, suitable for PALM and long term imaging.
- c-Myc-HaloTag-MS2, to monitor single mRNA *in vivo*. Indeed, we inserted 24 repeats of the MS2 stem loop ([37]), whose RNA is bound by the MCP protein. Thus, transfection with a MCP protein fused to either a HaloTag or a fluoresencent protein allow to see the mRNAs.

A gRNA was designed in the last intron of the *MYC* gene with the help of the MIT's CRISPR design tool (`http://crispr.mit.edu/`, Feng Zhang

lab). This online tool generates candidate variable sequences for the gRNA with a properly located PAM sequence, checks for the uniqueness of the target RNA, accounting for the tolerated mismatches, and presents a list of the most probable off-target sites (within and without genes). We chose a gRNA with minimal in-ORF off-targetting. The gRNA vector was synthetized as described in Figure 7e. We made sure that a proper insertion would disrupt the PAM sequence used to target Cas9, thus avoiding multiple cleavages of a properly edited sequence.

The donor vector was designed to knock-in the fluorescent tag at the C-ter of c-Myc, similarly as ([38, 2], as other locations were very likely to disrupt the protein expression, function or regulation ([8]). Therefore, we designed homologies against the last intron, the last exon and the 3'UTR (Figure 7 middle).

Although the reported targetting efficiencies of the CRISPR/Cas9 system are very high, it is still necessary to select for integrations with a drug. We thus inserted a puromycin resistance gene (with its own constitutive promoter), flanked by two LoxP cassettes in the last intron of c-Myc, allowing for excision of the marker after selection of the cell line. Indeed, the cassette is removed upon transfection of the Cre recombinase using viral transfection. The final construct only carries the protein tag Halo/Dendra2, all the other markers have been excised (Figure 7, bottom).

1.5.3 Construct building

As described in (Figure 7), the donor vector requires to assemble 7 (for the Dendra2 and Halo vectors) to 8 fragments (for the Dendra2-MS2 and Halo-MS2 vectors). Thus, a faster strategy than the traditional PCR-restriction-ligation pipeline is required. On the other hand, the MS2 sequence consists of 24 repeats of 54 nucleotides, which cannot be amplified by PCR [2].

[2]. Bionumbers, accession numbers 109211 and 104511.

To assemble such large vectors, multiple techniques are available, one of them, termed recombineering [39], involves *in vivo* homologous recombination in *E. coli*. Although potentially efficient, we designed a two step Gibson assembly to generate the Halo donor vector followed by traditional cloning to integrate the Dendra2 and the MS2 sequences.

The Gibson assembly [40] is an *in vitro* reaction combining in the same tube an exonuclease, a DNA polymerase, a DNA ligase and the linear DNA fragments to be assembled, each of those carrying relatively small homology sequences (15-30 nt.) with the two fragments they should be assembled with. During the reaction, the exonuclease creates sticky ends, allowing efficient annealing of the homology arms while the polymerase and the ligase close the junction, resulting in the assembling of the fragments.

Even though the Gibson assembly can ligate >10 fragments [40] together in a single reaction, the efficiency decreases drastically with the length of the homology sequence provided. We thus proceeded to two successive assembly reactions (Figure 7a.1 and a.2), combining respectively 4 and 3 fragments. ~ 50 bacterial clones were screened by colony PCR to detect the correctly assembled vector.

To transform the ligation and assembly products, we generated chemically competent bacteria according to the Inoue protocol ([41, 42]), and achieved a competency of $> 10^8$ colonies per microgram of PUC19 plasmid.

1.5.4 Transfections and screening

We tried to generated the cell lines according to the following protocol:

- *Day 1, transfection*: NT2 cells were transfected with both the gRNA vector(s) ($\sim$ 1.5 µg) and the donor vector ($\sim$ 4 µg) using Nucleofection (Lonza). We also tried to transfect using XFect (Clontech), without success.
- *Day 2, transfection efficiency check*: Cells

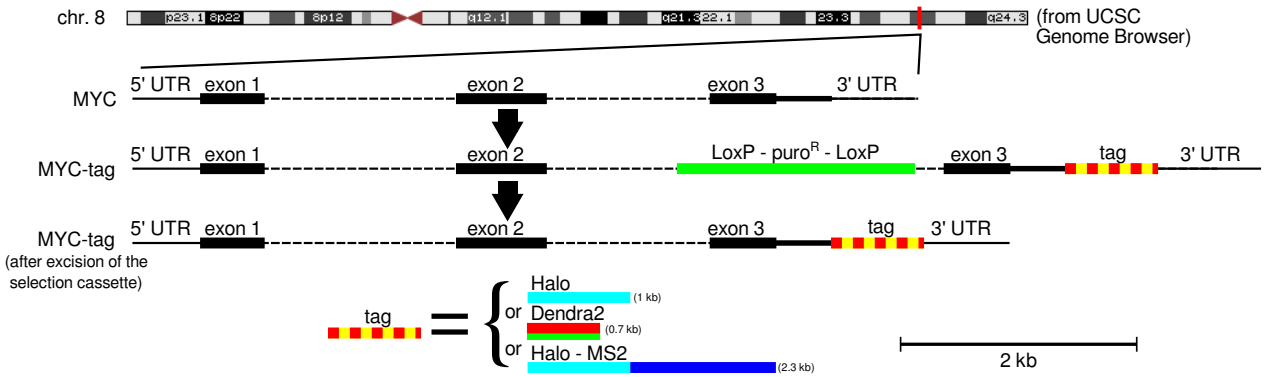


Figure 6: Targetting the *MYC* gene. *(top).* location of the *MYC* gene on chromosome 8 and detail of the three exons, four introns *MYC* gene. *(middle).* *MYC* structure after editing. *(bottom).* final structure of the *MYC* gene after selection and excision of the puromycin resistance cassette. Note that a few nucleotides scarf remains in the last intron, as a remnant of the LoxP cassette (not represented here).

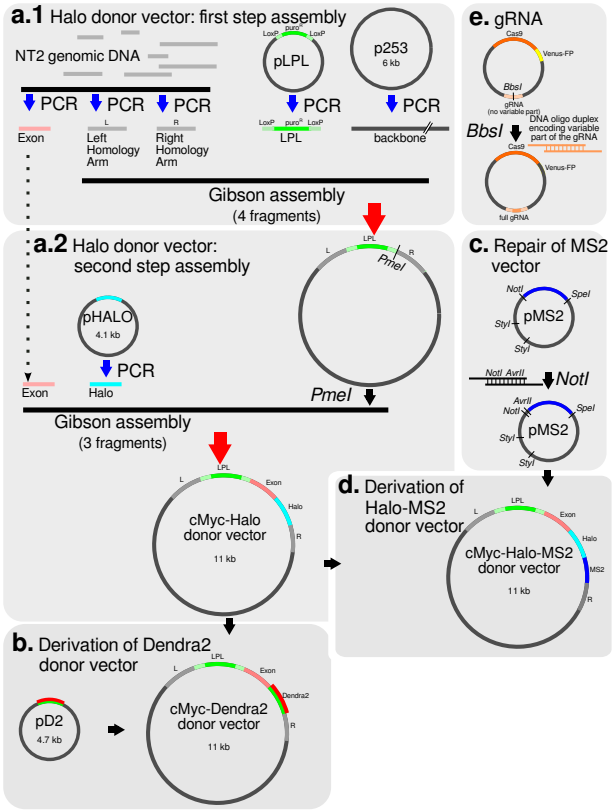


Figure 7: Cloning stategy to edit *MYC*. (a, b, c and d). Assembly of the donor vector. (e). Generation of the sgRNA. Black arrows stand for restriction/ligation cloning, blue arrows denote PCR, red arrows Gibson assemblies. Restriction sites are indicatic in italic. Relative sizes of vectors/sequences are not respected.

were checked under a fluorescence microscope for the expression of Cas9. Indeed, in the gRNA-Cas9-Venus construct, the Cas9 protein carries a "venus" fluorescent protein (Figure 7e).

- *Day 3-15, puromycin selection*: An increasing amount of puromycin is applied to the culture. Cells are checked daily for colony formation.

- *Day 15, fluorescence checking*: Cells were checked under a fluorescence microscope for expression of the tag (either Dendra2 or Halo). Fluorescent colonies (if any) are picked. In case of very low fluorescence, cells were sorted by FACS (Fluorescence-activated cell sorting).

- *Day $>$ 15, characterization of the cell line*: see Supplementary methods, 4.1.3.

This protocol implements several steps of screening. The first step is realized by drug selection. Since the donor vector contains a puromycin-resistance gene coupled to a strong promoter, drug selection allows to screen for integration into the NT2 genome. To perform this drug selection, we first established a killing curve to determine the concentration that would kill $>$ 95% of the non resistant cells in 36 h. The second step is the fluorescence screening, that highlights cells where the tagged c-Myc is inserted in frame into an active gene. Finally, colony genomic PCR are performed to screen for the insertion at the correct locus.

1.6 Absolute quantification of protein levels

Since one of the key characteristics of c-Myc is its high regulation in cells (see Introduction), we implemented a method (described in [43]) to measure absolute levels of tagged proteins in fixed single cells. This method relies to the comparison with a

fluorescence microscope of the the fluorescence of the pool of tagged proteins with a reference situated right next to it (Figure 8).

Concretely, cells expressing a Dendra2-tagged protein of interest are plated on coverslips (Figure 8), fixed and imaged in PBS, ensuring a homogenous medium inside and outside of the cell. Purified Dendra2 is added to the medium at a known concentration, providing a side-by-side reference to optically quantify the protein expression (Figure 8b and c). The use of a confocal microscope and the side-by-side reference avoid biases induced by off-focus fluorescence of the medium and variability in the optical setup. Opposite to the initial method, we performed the measurement on fixed cells, since photophysical parameters of fluorescent proteins are affected by the solvent (pH, redox parameters, etc., [44]).

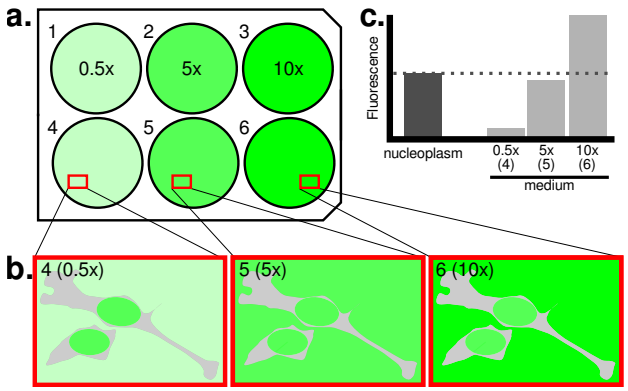


Figure 8: Estimation of the protein concentration per cell. (a and b). As the Dendra2 concentration in the medium is gradually increased, cell nuclei progressively disappear respective to the background. (c) represents a quantification of (b). Nuclear Dendra2 concentration is determined when the observed fluorescence of the nucleus and the medium are equal. It is risky to estimate this value when the concentrations are different by normalizing fluorescence signals, since the cameras in use can show strong non-linearities in their response.

1.6.1 Protein expression

The authors who previously described the method purchased purified GFP. However, since our super-resolution strategy involves Dendra2, so does our clonings and the design of the CRISPR cell lines. To avoid the generation of additional cell lines, we purified the Dendra2 protein.

The Dendra2 cloning sequence was cloned into a bacterial expression vector (Figure 9a). This cloning lead to a construct where 6 histidines were added to the C-ter end of Dendra2. This sequence is under a T7 promoter repressed by a LacO/LacI system. The 6xHis sequence exhibits high affinity for nickel, allowing efficient purification by immobilized metal affinity chromatography (IMAC): the 6xHis tag binds to Ni-NTA resin with high affinity.

This construct was transformed into the BL21 cell line, an *E. coli*-derived bacteria strain constitutively expressing the phage T7 polymerase and some mammalian rare codons. Cells were grown in LB until OD_{600} reaches ~ 0.5, and IPTG (Isopropyl β-D-1-thiogalactopyranoside), an allolactose analog was added to the culture. IPTG binds to the LacI protein, thus freeing the LacO sequence and allowing the T7 polymerase to express the fluorescent protein.

We determined by Coomassie staining the time when optimal expression of Dendra2 was achieved in a small culture volume (Figure 9b), and proceeded to the large scale expression the next day.

We grew, harvested and purified Dendra2 in the dark whenever possible, or under low, indirect light to avoid bleaching of the fluorescent protein. Bacteria were grown in a one liter LB flask at 37°C and harvested 3 hours after IPTG induction. The next day, bacteria were lysed and sonicated, releasing a greenish cytoplasm indicating of proper Dendra2 expression. Lysates were incubated with Ni-NTA resin, multiple washes were realized, the compounds specifically bound to the resin were eluted with imidazole and 1 ml fractions were collected. 5 µl of each fraction was analyzed with SDS-PAGE (SDS-polyacrylamide gel electrophoresis) and revealed with Coomassie blue (Figure 9c). The most concentrated fractions (fraction 3–6) were then pooled, dialyzed overnight into a storage buffer, frozen in liquid nitrogen and kept in the dark at -80°C. Total Dendra2 concentration was determined thanks to a Bradford assay (Bio-Rad).

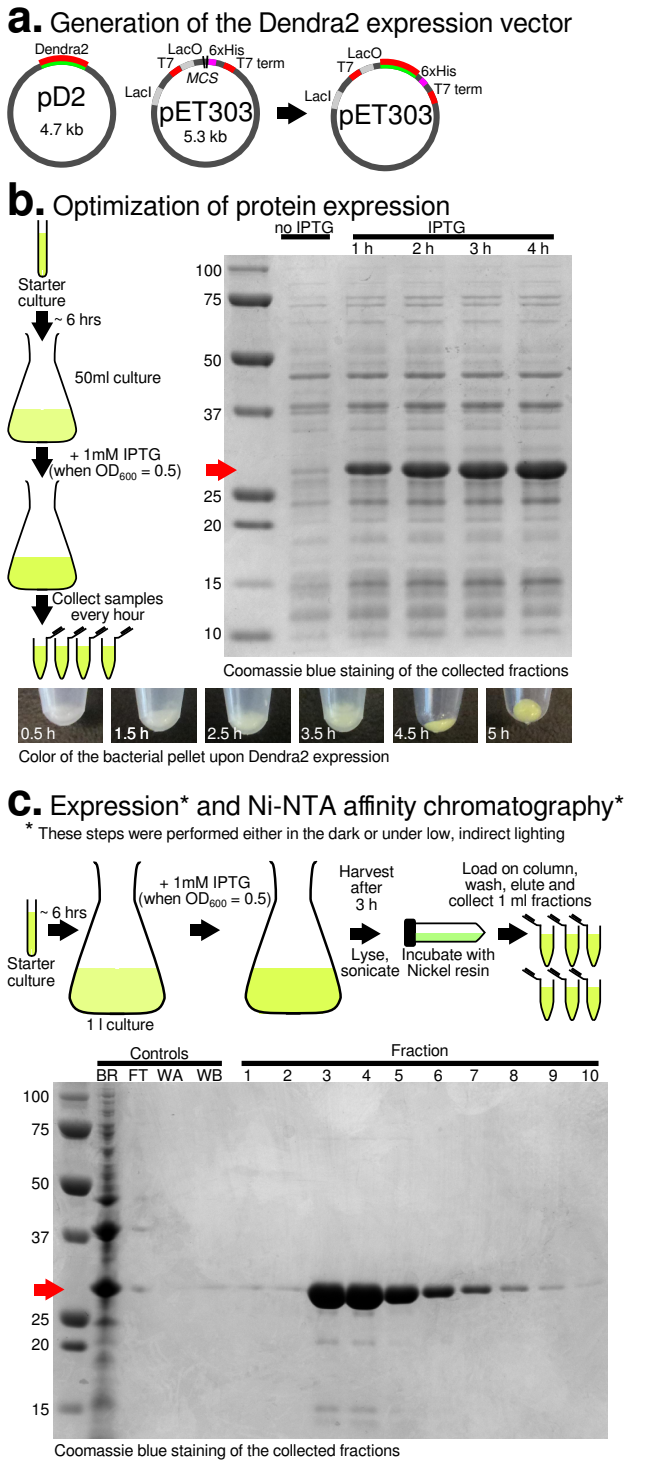


Figure 9: Purification of the Dendra2 protein. 12% polyacrylamide gels. BR: lysate before Ni-NTA resin purification, FT: flow-through the resin, WA and WB: samples from two successive washes (containing 30 mM and 40 mM imidazole, respectively). red arrow (26k) indicates the expected size. Size marker in kilodaltons. IPTG: Isopropyl β-D-1-thiogalactopyranoside. OD_{600}: optical density assayed at 600 nm. Ni-NTA: Nickel (Ni^{2+})-nitrilotriacetic acid.

1.6.2 Fluorescent Dendra2 concentration determination

Although the combination of the Coomassie staining (to estimate the fraction of Dendra2 within the eluted proteins) and the Bradford assay give an accurate estimate of the concentration of the Dendra2 protein, these techniques fail to estimate both the number of properly folded proteins, and more importantly, the number of proteins actually fluorescent. Indeed, to compare nuclear fluorescence to the purified Dendra2 solution, it is necessary not only to evaluate the concentration of the Dendra2 protein, but to estimate the number of fluorescently active proteins.

Two main mechanisms can lead to non-fluerescent Dendra2. The first one is misfolding. It has been shown with GFP, a fluorescent protein exhibiting a similar β-sheet rich, barreel-like structure, that the complete folding of the protein can take more than thirty minutes [45, 46]. A proper folding is crucial to obtain a working protein, since the gain of fluorescence during the folding requires a self-catalyzed intra-chain reaction to form the fluorophore [47]. Second, when properly folded, the protein can bleach. In that case, the fluorophore can be altered without major change in the conformation of the full protein. Two methods to assay the actual concentration of fluorescent proteins are present in Supplementary Methods 4.1.4

1.6.3 Validation using the Rpb1-D2 cell line

To validate our technique, we applied the quantification technique to a U2OS (human osteosarcoma) cell line where the catalytic subunit of the RNA polymerase II (PolII) was replaced by a Dendra2-tagged, α-amanitin resistant form (described in [7]). Xavier Darzacq (École normale supérieure, Paris, France) kindly shipped us some cells. We chose this cell line for various reasons:

- Although still debated, the number of PolII per cell has been estimated. In budding yeast, estimates range from 2000 to 30000 [2]. Another estimate (using comparison of catalytic efficiency compared to a purified reference) yields to $\sim$ 75000 PolII per HeLa cell [48].

On the other hand, quantitative Western blots suggest >300000 PolII/HeLa cell [49]. Recently published superresolution microscopy based estimates in U2OS-derived stable cell-line expressing PolII-SNAP suggest a number of PolII around 80000 molecules per cell [50]. Other unpublished estimates range between 30000 and 360000 PolII/cell (Xavier Darzacq, personal communication). These figures correspond to a concentration of 0.1 to 1 µM PolII/cell (assuming a 600 µm^3 nuclear volume).

- PolII counts have been estimated using single molecule techniques ([50], unpublished results from the Darzacq lab),
- Purified Rpb1 and appropriate antibodies were available in the lab, allowing for a biochemical quantification of the PolII expression levels, either by quantitative Western blot or by dot blot and comparison with a standard.

Once the PolII concentration had been determined, we tried to compute the average volume of a U2OS nucleus from a confocal z stack to convert the concentration into number of molecules per cell.

2 Results

2.1 Setting-up a system to assay absolute protein number

We first tried to implement a system allowing for absolute quantification of protein numbers in single cells, as described in section 1.6. To perform relative fluorescence intensity of fluorescent medium versus nucleus, we first purified the Dendra2 thanks to purposely inserted FLAG-tag and nickel-affinity column purification. This method lead to $\sim$ 4 ml of protein at a concentration of $\sim$ 3.4 mg/ml (Figure 10c), that is $\sim$ 130 µM (molecular weight of Dendra2-FLAG is around 27k). We tried to measure the amount of actually fluorescent proteins within this solution, either by absorbance or by fluorescence. Unfortunately, at the time of the report, we were still lacking a standard fluorophore as a reference, and had no time to assay the absorbance. The following computations thus assume a significant amount ($\sim$ 100%) of actually fluorescent proteins within the purified solution. This assumption is lead by the precautions we took to avoid bleaching the fluorophore, and based on previous observations on GFP, that describe its high stability during standard Ni-NTA purification.

As a control, we assayed the molecule number of RNA polymerase II (PolII) in U2OS. This engineered cell line expresses a Dendra2-tagged α-amanitin resistant Rpb1 (Figure 10a). Upon addition of α-amanitin, the Dendra2-tagged Rpb1 replaces the endogenous one without major perturbation of the expression level. Thus, assaying the number of molecules within this system can provide an order of magnitude of the number of PolII in U2OS cell.

Preliminary observations on a serial dilution of Dendra2 fluorescent medium are presented in (Figure 10b). A first observation is that the expression level of PolII and its localization shows little variability. Flurthermore, and although presenting severe inconsistensies, the nucleus "appears" when changing the medium from 10 µM Dendra2 to 2.5 µM, suggesting that the actual value lies between these extreme limits. To further improve this estimate, we plotted unnormalized fluorescence profiles under two medium concentrations. The ratios between in cell and out-of-the-cell fluorescences are consistent with a concentration of $\sim$ 4-5 µM.

We tried to determine the volume of a U2OS using a confocal z stack without success. We thus take the rough estimate of a nuclear volume of $\sim$600 µm^3. This leads to an estimated count count of 1.4 to 1.8 million of polymerases per nucleus. A value a four times higher than previously published concentrations.

However, this data is highly preliminary, and

should be considered as a proof of concept rather than a reliable estimation method until more testing has been performed. Indeed, when diluting the concentration of Dendra2 in the medium, both the fluorescence of the medium and the nucleus were observed to decrease, opposite to what was expected (a background fluorescence going down and the nuclear level remaining constant). This issue has not been troubleshooted yet, but cannot compromise our first rough estimate, because the order of the relative fluorescences cannot be inverted.

Furthermore, more characterization of the purified Dendra2 protein is needed to determine the fraction of actually fluorescent proteins. Three of these assays (native gel, absorbance and quantum yield) will be performed within the next few weeks.

Finally, from a biochemical point of view, more characterization of the U2OS is needed (quantitative Western Blot with stringent lysis conditions, quantification of the relative expression of the tagged versus untagged polymerase).

2.2 c-Myc do not show major disregulation in NT2 cells

As c-Myc is disregulated in many cancers, and provided that NT2 is a cancer cell line, we sequenced parts of the *MYC* coding sequence and checked for differences with respect to reference genomes (GRCh38 and HuRef). Because of the cloning strategies in use, we couldn't check the 5' end of the construct (nucleotides 1 to 69, corresponding to the 45 nucleotides of the alternative translation starting site and the 24 following nucleotides). No difference was found respective to reference genomes.

Since our study focuses on the role of cMyc in differentiation and gene regulation, we considered assessing key protein levels (c-Myc, Sox2, Oct4, Rpb1, TRRAP, Max) across differentiation. Our first preliminary data suggest that c-Myc and MAX levels are unmodified upon differentiation (Figure 11. In this experiment, we cannot rule out the possibility of an increased c-Myc expression. If

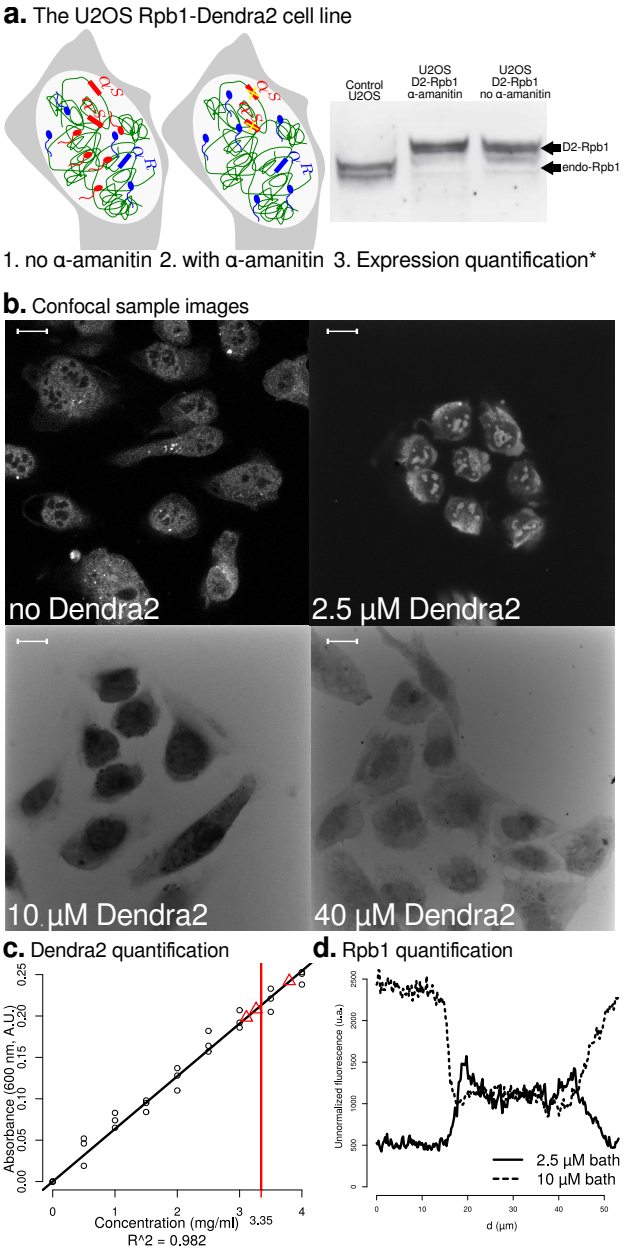


Figure 10: Quantifying PolII quantification in U2OS cells. *(a, left).* Schematic of the endogeneous replacement of Rpb1 by its tagged counterpart upon α-amanitin treatment. *(a, right).* Quantification of the effectiveness of the replacement and the relative expression levels of the endogeneous Rpb1 versus the tagged one. This gel is reproduced from [7], with permission of the lab. *(b).* Scale bar represents 20 µm. *(c).* Bradford assay estimating the concentration of the purified protein. /(d). Consistent profiles across a nucleus upon two different concentrations of Dendra2 in the extracellular medium. Although the unnormalized fluorescence of the nucleus is constant upon dilution of the medium (around 1200 u.a), the background fluorescence is divided by approximaltely 5 (2500 to 500, consistent with the fluorophore dilution, suggesting that within this range, the camera shows an approximately linear response).

confirmed, this behaviour would be consistent at least with observatons of hematopoietic stem cell differentiation into a commited progenitor [51, 52]. Although we prepared the nuclear extract and had qPCR primers to perform a more precise and com-

prehensive study, we didn't have time to implement those strategies.

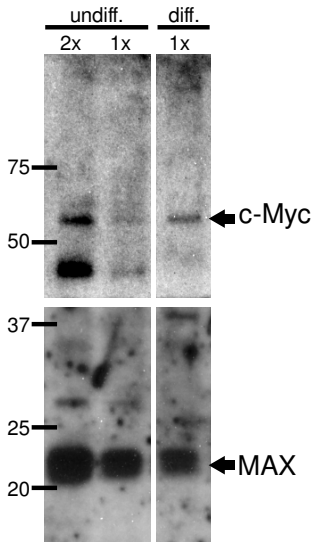


Figure 11: No major change in neither c-Myc nor MAX is visible during differentiation. 1x and 2x symbols denote the relative concentration when loading the gel. Marker in kilodaltons. undiff.: undifferentiated cells, diff.: differentiated cells.

2.3 C-ter tagging of c-Myc seems not to impair its functions

Nuclear factors localizations and dynamics are driven by their specific interactions with nuclear components [20, 5]. It is thus crucial to assay whether the fused tag impairs the function and localization of the fusion protein. We thus performed the following assays and verifications:

- The tags were cloned in C-ter of c-Myc to avoid disruption of any known regulatory function of c-Myc (as reviewed in [8], and an EGFP C-ter fusion has been successfully described [2]).

- We chose the Dendra2 fluorescent protein, an engineered monomeric form of the Dendra protein, that has been shown to exhibit little interference with the cell ([24, 7]), especially with respect to aggregation and cluster induction.

- Since the Max interacting helix-loop-helix (HLH) of c-Myc is located in the C-terminal region of the protein, we were especially worried about disrupting this crucial interaction. A disruption of the interaction with Max would mean no DNA binding, and a loss of many other partners. We assayed this interaction by Co-Immunoprecipitation of the c-Myc/Max duplex in undifferentiated and differentiated cells using an anti-Max antibody for the IP. That way, the endogenous and the transfected proteins can be distinguished on the western blot as they exhibit a $\sim$ 30k difference in molecular mass. Results for the differentiated case are presented Figure 12b. At the time of the first CoIP, the Halo construct was not ready. Hence, results are not presented. A problem (likely related to the transfection efficiency) occured in the undifferentiated case, and the assay is being repeated, and we are including the Halo fusion protein within the tested samples.

- Another important control was to determine whether the localization of the endogenous versus the tagged protein were identical, and what was the acceptable transfection level. We were especially interested in the aggregation pattern induced by Dendra2 (GFP can significantly decrease the dissociation constants of some engineered systems, Lukasz Bugaj, Schaeffer Lab, personal communication). We tried to perform double color imaging, where the cells were transfected either with a Dendra2-conjugated c-Myc or with a Halo-conjugated fusion protein and stained adequately. Secondly, the total c-Myc population was stained with a primary antibody. Unfortunately, this experiment lead no result due to transfection problems. We compare instead single-color stainings in two distinct experiments (12b). We are in the process of repeating the assay.

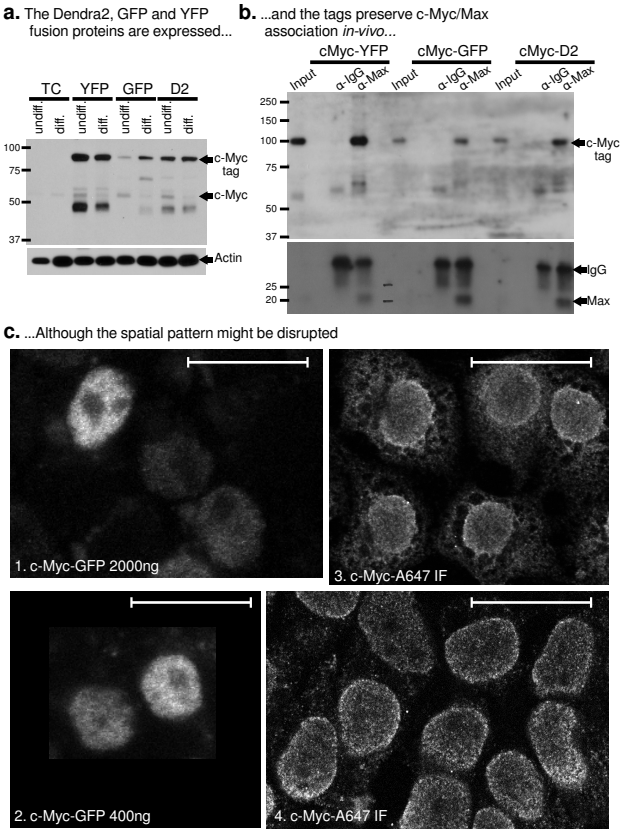


Figure 12: Functional assessment of the differences in interaction and interaction of the tagged c-Myc. *(c, left).* Fixed cell confocal imaging of c-Myc-GFPtransfected cells showing a homogeneous pattern in the nucleus *(right).* Immunofluorescence staining of untranfected cellsScale bar. indicates 20 µm.

2.4 c-Myc seems to accumulate at the periphery of the nucleus in NT2 cells

We first performed an immunofluorescence to assess c-Myc localization (Figure 13a). As previously reported, the protein is mostly nuclear, but also exhibits a significant cytoplasmic localization (observations from at least two independent samples and stainings). Strikingly, c-Myc seems to accumulate to the periphery of the nucleus.

We tried to further investigate this localization. We first wondered whether c-Myc accumulation was precisely located at the nuclear envelope, or if it was accumulating at some distance from it. To assess this question, we performed a dual color staining with DAPI, hoping that the drop in the DAPI signal would precisely identify the nuclear envelope. However, this drop appears to be smooth (spread over several microns). We thus tried to specifically stain the nuclear envelope by transfecting 1 µg/25 mm coverslip of a laminA-GFP. Indeed, laminA accumulation at the nuclear envelope is a well known feature. However, the construct did not localized at the envelope, but rather showed a mostly homogeneous nuclear signal (data not shown), and even a very low pinhole aperture could hardly identify a slight accumulation at the envelope. Finally, since the 647 nm laser (used for c-Myc imaging) and the 405 nm laser (DAPI) are different, the optics show significant chromatic aberrations that result in a drift of a few hundred of nanometers of one color respective to the other (visible in Figure 13c): c-Myc looks out of the DAPI signal on one side). Without further calibration, these chromatic aberrations prevent a precise measurement of a nuclear envelope accumulation.

We then tried to quantify the accumulation. Indeed, a significant concentration gradient can make a significant difference in terms of reactivity, by deplacing thermodynamical equilibria, thus favoring some enzymatic reactions at some points of the nucleus. We first intented to plot profiles across the nucleus to measure the enrichment (Figure 13a, bottom, left). This measure qualitatively confirmed the enrichment, whereas DAPI profiles were approximately homogeneous across the same profiles. However, since the profile plots the value of single pixels, the measure is very noisy, and is hard to use for quantification. Furthermore, the measured enrichment highly depends on the choice of the profile (Figure 13a, bottom, left).

We tried to circumvent this issue by computing the average enrichment respective to the nuclear envelope. Nuclei contours were manually identified as polygons based on the DAPI staining (Figure 13b.1), and the distance to the polygon was computed for each pixel (Figure 13b.3) (if we define a polygon as a union of n segments defined by n points $(x_i)_{i\in[1,n]}$, i.e.: $P = \cup_{i=1}^{n-1} [x_i, x_{i+1}]$, the distance *dist* of a point p to a polygon P is the minimum euclidian distance from the point to the

segment: $dist(p, P) = \min_{i \in [1, n-1]} dist\left(p, [x_i, x_{i+1}]\right)$). The pixel intensities are then averaged respective to their distance to the polygon and normalized for the decreasing volume considered as the radii reach the middle of the cell (Figure 13b.4). Although being a noisy measurement, this method confirmed the enrichment at the periphery. However the signal was too noisy allow quantification, and was invisible in some cells (Figure 13b.6). We believe that this is due to a measurement limitation, and not to biological variability, since the recorded signal was very noisy due to the low light input coming from the almost closed pinhole of the confocal.

2.5 Superresolution provides more insights on c-Myc fine organization

To overcome the limitations of confocal microscopy, we moved to superresolution microscopy to further characterize c-Myc localization.

2.5.1 Acquisitions

We first acquired (immunofluorescence-STORM) IF-STORM picture of c-Myc in undifferentiated and differentiated NT2 cells. Acceptable acquisition quality was achieved for $>$ 5 cells over four rounds of acquisition. In those cells, we obtained $>$100k detections and a pointing accuracy $<$15 nm in x and y, and $<$50 nm in z. Once again, the accumulation at the nuclear periphery was the most striking characteristic of the 3D maps, and was conserved upon differentiation (Figure 14).

At that point of the characterization, it is necessary to assay how real is this pattern

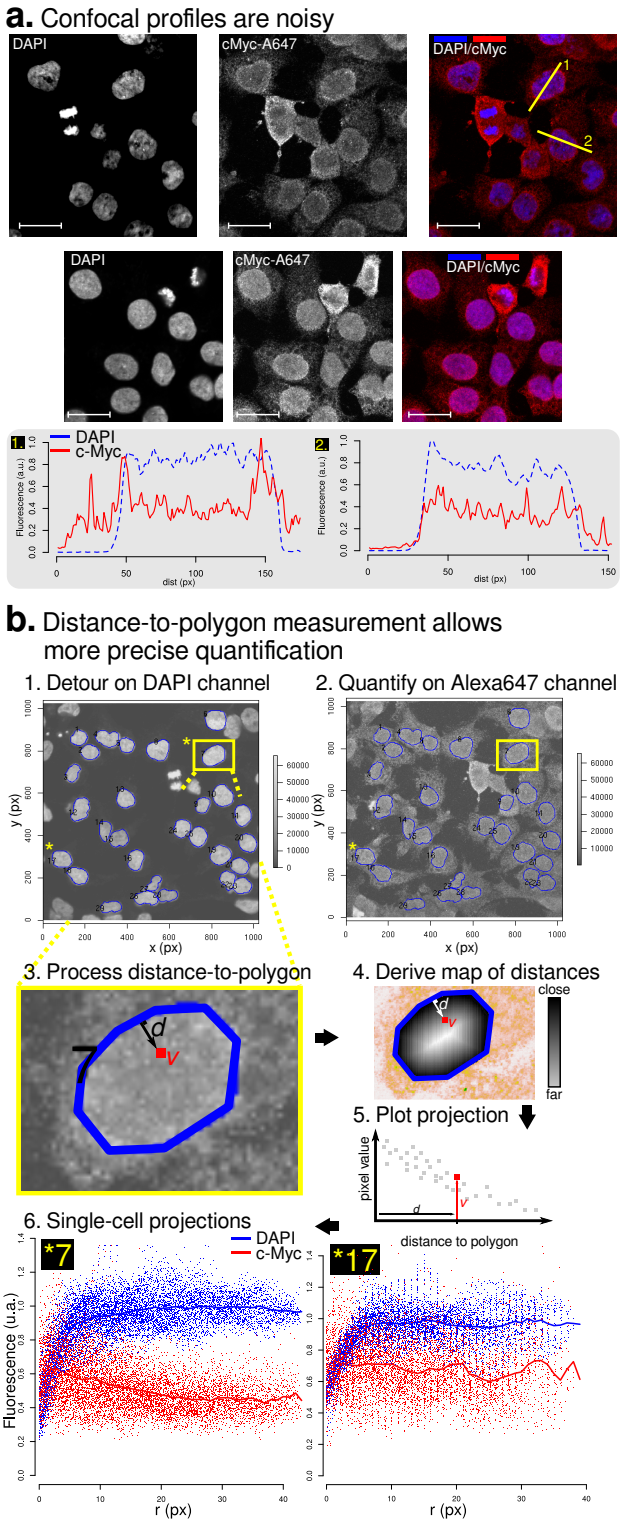


Figure 13: Quantification of the c-Myc gradient observed in confocal microscopy. Scale bar represents 20 µm.

2.5.2 Distance to envelope profiles

We applied a similar analysis as detailed in the previous section to the IF-STORM data. The data are projected along the z axis, and such is the polyhedron defining the nuclear envelope. Concretely, this analysis is not exactly equivalent to a 3D analysis. Indeed, as the radii increase, the probability to have an off-focus portion of nuclear envelope closer to the considered point than the in-focus nuclear envelope increase. Thus, the distances are overestimated for the high radii. In other words, the actual drop in density is sharper than the observed. The results are presented in figure (Figure 14a.4 and b.4) (additional data were

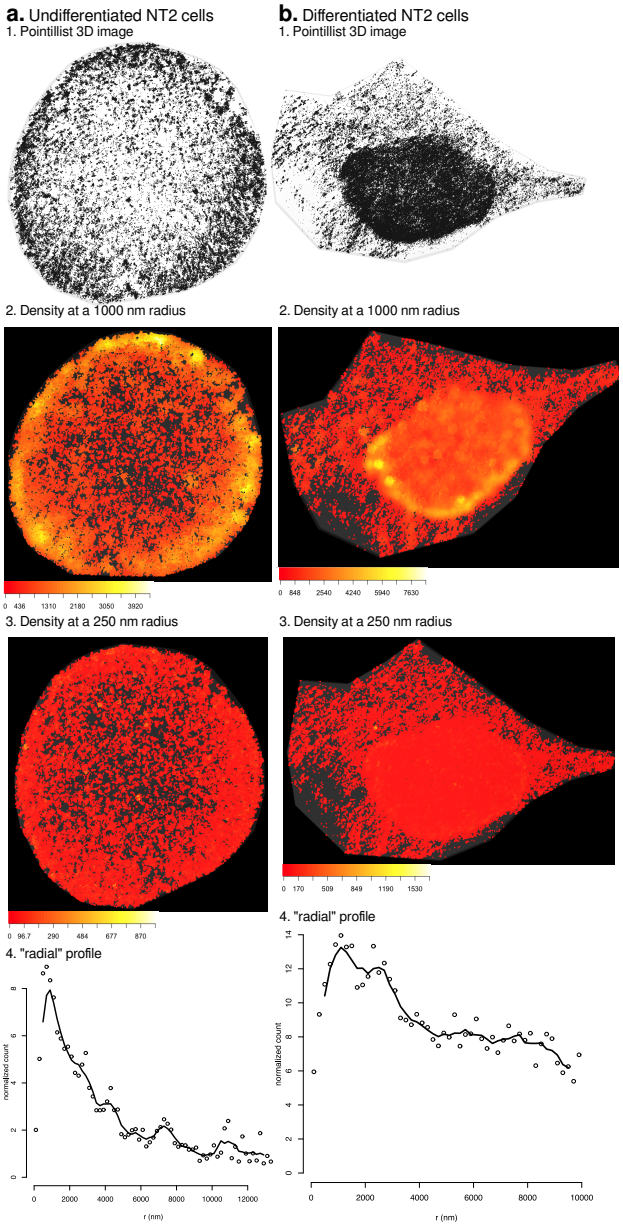


Figure 14: **a.** c-Myc IF-STORM of a nucleus of an undifferentiated NT2 cell. **b.** same as a, but with a differentiated cell. The full cell is shown, the region of high density of detections is the nucleus. **c.** Distance histograms were computed using a 100 nm binning. The superimposed smoothed curve represents a rolling average over five points.

collected and reproduced the presented pattern). The patterns appear here much more striking, due to the better quality of the data, and the size of the dataset (>50k points versus a few thousands, noisy pixel values).

This quantification allowed to draw several conclusions. First, the enrichment reaches 5 fold in the undifferentiated cell, an increase that could have biologically relevant implications. Second, although the gradient show different slopes between the undifferentiated cell and the differentiated one, we think this parameter is more related to the quality of the acquisition rather than a difference in the cells. To test this hypothesis, a larger dataset of improved quality is required. This analysis provides a quantitative estimate of the c-Myc gradient, and allows to derive the main trend governing c-Myc pattern. We then tried to look at finer details, and we are currently running simulations to detect potential c-Myc clusters.

2.6 Relation with chromatin

Provided the peripheral accumulation of c-Myc, we investigated the mechanisms driving its symmetric localization and accumulation. We tested several hypotheses, all of them originate from previous knowledge of c-Myc structure and properties:

2.6.1 Active genes

As a global genome enhancer and chromatin remodeller, c-Myc has been described to interact with active genes [53] by ChIP-seq. We first tried to do dual color superresolution imaging using immunofluorescence of the histone modification H3K3me3 on top of a transfection of c-Myc-D2 or c-Myc-Halo. However, our attempts remained unsuccessful for several reasons (low transfection efficiency and cross-talk between the fluorophores used – although this combination of fluorophores has already been used in the past, Ignacio Izeddin, unpublished results). We then moved to separate two colors IF-STORM imaging. We compared the "radial" profiles of c-Myc stained and H3K4me3 stained NT2 cells in undifferentiated and differentiated cells. For the undifferentiated case, and because of a lack of time, we used a dataset previously acquired in the lab (staining and imaging under similar conditions as we did the imaging, dataset consisting of two undifferentiated and two differentiated cells, data from Lana Bosanac, unpublished results).

Although noisy (Figure 15), and as expected, the H3K4me3 profile does not show any enrichment at the periphery, consistent with an enrichment of

heterochromatinized and gene poor regions at the periphery of the nucleus [54].

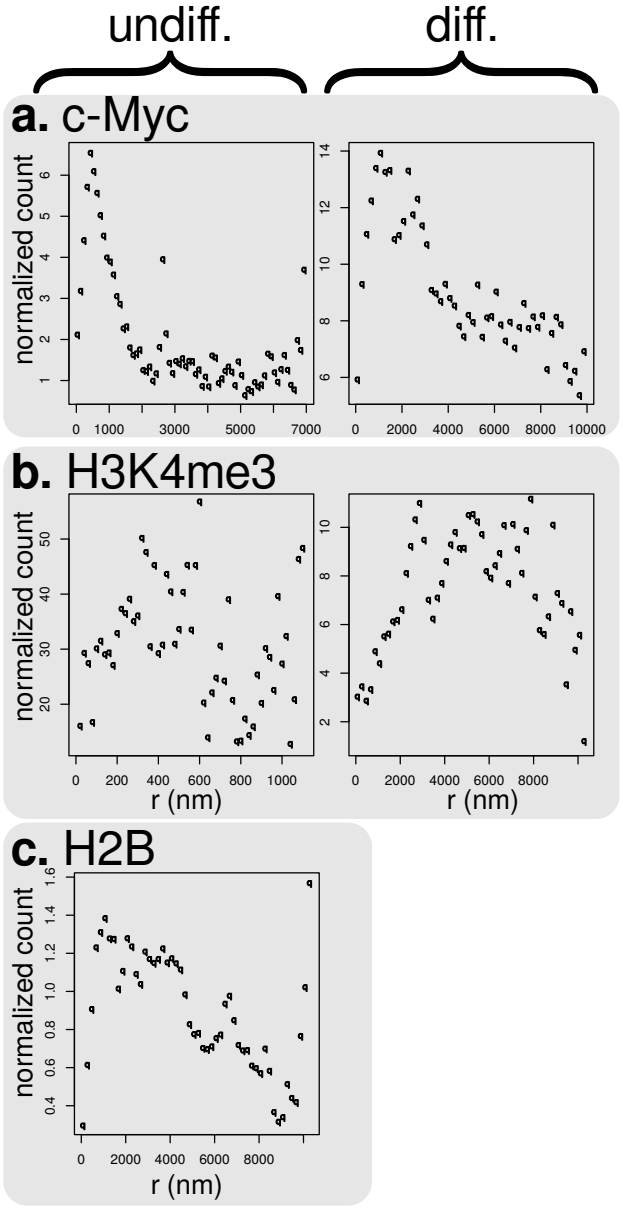


Figure 15: Quantification of the peripheral accumulation of c-Myc, H2B and H3K4me3 across differentiation. H2B data come from transfection of an H2B-Dendra2 construct. H3K4me3 and c-Myc from IF-STORM. x axis: distance to nuclear border, y axis: normalized count.

2.6.2 Chromatin

In vitro experiments show a strong affinity of the c-Myc/Max dimer for the E-box motif (CACGTG) and some other degenerate forms [55]. In a first approximation, one can expect an E-box every $\left(\frac{1}{4}\right)^6 = 4096$ bp, that is about 1.5 million E-boxes per diploid genome (more precise approximation, accounting for the sequence and codon biases could also be derived [56]). If spread uniformly in a ~ 600 μm^3 nucleus, there is ~ 2500 E-boxes/μm^3, or an average distance of 74 nm between two E-boxes. As another approximation, we considered this number as a global affinity for chromatin, and thus looked for a correlation between the chromatin profile and the c-Myc profile.

Although correlation of c-Myc levels with DAPI staining, recently published references use H2B as a proxy to approach chromatin densities. We present the "radial" profile derived from their dataset ($\sim$ 450000 detections of a U2OS transfected H2B-Dendra2 described in [12], kind gift of Vincent Récamier, Ignacio Izeddin and Xavier Darzacq), and tried to replicate their data in NT2 cells (Figure 15).

Opposite to the active marks, H2B and c-Myc seem to exhibit a similar trend, suggesting that a model of global affinity for chromatin cannot be ruled out. In Supplementary results 4.3.3, we investigate potential correlations with the genomic patter of E-boxes and c-Myc spatial localization.

3 Discussion

During this internship, we were able to apply the superresolution pipeline available in the lab to c-Myc through the test of adequate antibodies and generation of the appropriate constructs. We also proposed improvements to the data analysis pipeline. One combines most of the scripts and interfaces previously used scripts for the PALM/STORM reconstruction, reducing the user input required. The second is a straightforward on-movie method to estimate the pointing accuracy. We are in the process of integrating this control in the processing pipeline so that it will be automatically applied to all the acquisistions.

We used a combination of imaging techniques (confocal, PALM, STORM, SPT, FRAP), biochemistry and molecular cloning to investigate c-Myc nuclear localization and its functional significance. Our preliminary STORM with primary antibody results in differentiated and undifferentiated

NT2 cells indicate that c-Myc robustly accumulates at the periphery of the nucleus, an unexpected pattern regarding previously published literature: on the one hand, repeated independent studies performing ChIP-seq and quantitative ChIP-seq describe a very clear colocalization of c-Myc with the histones active marks such as H3K4me3 [53]. On the other hand, a number of experiments ranging from electron microscopy to chromosome painting demonstrate that DNA sequences are not randomly organized into the nucleus. The chromosomes tend to adopt a radial distribution following their richness in genes: the gene-poor chromosome 19 is robustly found at the periphery of the nucleus. Finally, preliminary data available in the lab provide additional arguments tending to prove that active genes (H3K4me3-marked genes) are enriched at the center of the nucleus.

These known feature tend to challenge our experiments, where an opposite pattern is observed. How can a protein be on the one hand enriched at very specific active regions in the genome, and on the other hand localize at the gene-poor periphery of the nucleus?

One possible explanation is to refute one of the datasets. We are currently trying to image c-Myc with another antibody, or after cMyc-Halo transfection to make sure that the observed pattern is due to specific antibody binding. Another possibility is that for some reason (complexation, post-translatonal modification), peripheral c-Myc is less soluble than the central one, leading to subdetection of c-Myc when the ChIP has been performed. Our working hypothesis for the course of this discussion is to consider those two patterns as real.

A second series of hypotheses resides in the fact that STORM and ChIP-seq measure two totally different types of events at different scales. In ChIP-seq, non-specific interactions are lost in the noise, whereas rare but locus-specific strong interactions (such as a possible complex histone-TRRAP-GCN5-c-Myc) will be captured, as ChIP-seq realizes an ensemble average over hundreds of thousands of cells. In STORM superresolution microscopy, cells are fixed and the x, y, and z coordinates of a (supposedly significant) fraction of the tagged proteins is recorded, providing a snapshot of c-Myc localization. Those images however do not give any information on the affinity or kinetics of the binding. In our case, we don't know the affinity of the binding at the nuclear periphery, but one can imagine highly transient and dynamic, low affinity interactions at the nuclear periphery whereas long-term, specific interactions located more in the center of the nucleus are lost in the noise: a snapshot cannot distinguish between highly dynamic, low affinity interactions and fixed interaction at a given locus.

From an other perspective, the mechanistic origin of the c-Myc gradient raises many questions. First, what is the nuclear structure underlying c-Myc "unspecific" binding? Indeed, it seems difficult to tether proteins without (even dynamically) binding them to a substrate that fill space. One obvious candidate structure spanning the entire nucleus is DNA, that also exhibit peripheral enrichment [12]. But it is not the only one. Indeed, recent findings suggest that other nuclear protein structures can form gels in the nucleus and be a substrate where diffusing can adsorb and be tethered. Such proteins notably involve PolII [57] (through its repeated, unstructured C-terminal domain) and histone (through their tails). More generally, many abundant nuclear proteins exhibit so-called low complexity domains that can both exhibit space filling properties and thus act as a surface for aspecific adsorption. These surfaces can be characterized by their fractal dimension, a measure accessible with 3D superresolution microscopy [12]. For instance, in U2OS cells, the fractal dimension of chromatin has been estimated as 2.7. Thus, estimating c-Myc fractal dimension could help to identify the network structuring its localization.

Second, does this gradient show any functional significance? If DNA targets are involved in creating the c-Myc gradient, then it is likely that an-

swering the latter question will require to investigate the evolutionary history of spatial chromatin organization.

However, it might be possible to imagine a model where no space-filling structure is required to tether c-Myc, and one can design a model to exhibit that kind of gradient by combining free diffusion and dynamic binding at the nuclear envelope. This model is presented as a supplementary discussion (4.5), although the analytical solutions remain to be determined.

Finally, it is noteworthy that single particle tracking experiment can provide valuable insights to discriminate the previous hypotheses [3]. Indeed, SPT provides both the position and the dynamics of single molecules, and some models have been proposed to discriminate between specific and aspecific binding events. It becomes then possible to analyze the collected traces with respect to the density of specific binding events across the nucleus versus the density of traces. In this experiment, homogeneous (or center-enriched) specific binding events combined to global enrichment in the number of traces at the nuclear periphery would support our hypothesis reconciling STORM and ChIP-seq experiments.

Acknowledgements

I would like to thank Lana for her wonderful supervision, support, dedication and incredible patience, especially during the long hours of teaching at the microscope. Furthermore, thank to Frank for teaching me the design of CRISPR/Cas9 cell lines from A to Z. I want also thank Claudia and Sharon for supervising me for the CoIP and Western blots and the Dendra2 purification, respectively. I also want to thank Benjamin for all the discussions we had, especially about the design of the absolute quantification of protein numbers, Gina for all her cloning tips and recipes, Spencer for very fruitful chemical discussions and his everyday support, Lea and Kevin for their precious advices and Janeen for taking care of all the visa paperwork.

I must also acknowldge the whole Darzacq lab for continuous advices and discussions, providing unpublished datasets and the U2OS cell line.

Finally, I am very thankful to Lukasz Bugaj, PhD student in the David Schaffer lab for discussions about the Cry2 tagging, Milos for his support and advices, and more generally to all the Tjian lab for its incredible support and kindness.

♡ Thank you ♡,

Ever tried.
Ever failed.
No matter.
Try again.
Fail again.
Fail better.

Samuel Beckett

[3] Although we tried to acquire such a dataset (presented in the supplementary results, its quality is too low to be safely interpreted.

References

[1] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, Aug 2006. [DOI:10.1016/j.cell.2006.07.024] [PubMed:16904174].

[2] Z. Nie, G. Hu, G. Wei, K. Cui, A. Yamane, W. Resch, R. Wang, D. R. Green, L. Tessarollo, R. Casellas, K. Zhao, and D. Levens. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151(1):68–79, Sep 2012. [PubMed Central:PMC3471363] [DOI:10.1016/j.cell.2012.08.033] [PubMed:23021216].

[3] P. S. Knoepfler, X. Y. Zhang, P. F. Cheng, P. R. Gafken, S. B. McMahon, and R. N. Eisenman. Myc influences global chromatin structure. *EMBO J.*, 25(12):2723–2734, Jun 2006.

[4] S. He, Z. Liu, D. Y. Oh, and C. J. Thiele. MYCN and the epigenome. *Front Oncol*, 3:1, 2013.

[5] M Woringer, X Darzacq, and I Izeddin. Geometry of the nucleus: a perspective on gene expression regulation. *Current Opinion in Chemical Biology*, 20(1), June 2014. In press.

[6] A. Bancaud, S. Huet, N. Daigle, J. Mozziconacci, J. Beaudouin, and J. Ellenberg. Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. *EMBO J*, 28(24):3785–3798, Dec 2009. [PubMed Central:PMC2797059] [DOI:10.1038/emboj.2009.340] [PubMed:19927119].

[7] I. I. Cisse, I. Izeddin, S. Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan, and X. Darzacq. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science*, 341(6146):664–667, Aug 2013. [DOI:10.1126/science.1239053] [PubMed:23828889].

[8] J. Liu and D. Levens. Making myc. *Curr. Top. Microbiol. Immunol.*, 302:1–32, 2006.

[9] V. H. Cowling, S. Chandriani, M. L. Whitfield, and M. D. Cole. A conserved Myc protein domain, MBIV, regulates DNA binding, apoptosis, transformation, and G2 arrest. *Mol. Cell. Biol.*, 26(11):4226–4239, Jun 2006.

[10] S. Pelengaris, M. Khan, and G. Evan. c-MYC: more than just a matter of life and death. *Nat. Rev. Cancer*, 2(10):764–776, Oct 2002.

[11] R. Cotterman, V. X. Jin, S. R. Krig, J. M. Lemen, A. Wey, P. J. Farnham, and P. S. Knoepfler. N-Myc regulates a widespread euchromatic program in the human genome partially independent of its role as a classical transcription factor. *Cancer Res.*, 68(23):9654–9662, Dec 2008.

[12] V Récamier, I Izeddin, L Bosanac, M Dahan, F Proux, and X Darzacq. Single cell correlation fractal dimension of chromatin: A framework to interpret 3D single molecule super-resolution. *Nucleus*, 5(1), 2014.

[13] Bo Huang, Mark Bates, and Xiaowei Zhuang. Super-resolution fluorescence microscopy. *Annual Review of Biochemistry*, 78(1):993–1016, 2009. PMID: 19489737.

[14] Michael D Abràmoff, Paulo J Magalhães, and Sunanda J Ram. Image processing with imagej. *Biophotonics international*, 11(7):36–43, 2004.

[15] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793):1642–1645, Sep 2006. [DOI:10.1126/science.1127344] [PubMed:16902090].

[16] Nataliya Panchuk-Voloshina, Rosaria P. Haugland, Janell Bishop-Stewart, Mahesh K. Bhalgat, Paul J. Millard, Fei Mao, Wai-Yee Leung, and Richard P. Haugland. Alexa dyes, a series of new fluorescent dyes that yield exceptionally bright, photostable conjugates. *Journal of Histochemistry Cytochemistry*, 47(9):1179–1188, 1999.

[17] G. Lukinavi?ius, K. Umezawa, N. Olivier, A. Honigmann, G. Yang, T. Plass, V. Mueller, L. Reymond, I. R. Correa, Z. G. Luo, C. Schultz, E. A. Lemke, P. Heppenstall, C. Eggeling, S. Manley, and K. Johnsson. A near-infrared fluorophore for live-cell super-resolution microscopy of cellular proteins. *Nat Chem*, 5(2):132–139, Feb 2013. [DOI:10.1038/nchem.1546] [PubMed:23344448].

[18] I. Izeddin, M. El Beheiry, J. Andilla, D. Ciepielewski, X. Darzacq, and M. Dahan. PSF shaping using adaptive optics for three-dimensional single-molecule super-resolution imaging and tracking. *Opt Express*, 20(5):4957–4967, Feb 2012. [PubMed:22418300].

[19] A. Serge, N. Bertaux, H. Rigneault, and D. Marguet. Dynamic multiple-target tracing to probe spatiotemporal cartography of cell membranes. *Nat. Methods*, 5(8):687–694, Aug 2008. [DOI:10.1038/nmeth.1233] [PubMed:18604216].

[20] L Bosanac, I Izeddin, V Récamier, C Dugast-Darzacq, O Bensaude, C Spriet, L Heliot, and X Darzacq. The Positive Transcription Elongation Factor (P-TEFb) nuclear target search is guided by RNA polymerase II CTD and chaperoned by the 7SK snRNP. *manuscript in preparation*, 2014.

[21] K. M. Niforou, A. K. Anagnostopoulos, K. Vougas, C. Kittas, V. G. Gorgoulis, and G. T. Tsangaris. The proteome profile of the human osteosarcoma U2OS cell line. *Cancer Genomics Proteomics*, 5(1):63–78, 2008.

[22] P Andrews. *Cell Biology, a laboratory handbook*, chapter TERA2 and its NTERA2 subline: pluripotent humane embryonal carcinoma cells, pages 183–190. Elsevier, 2003.

[23] G. V. Los and K. Wood. The HaloTag: a novel technology for cell imaging and protein analysis. *Methods Mol. Biol.*, 356:195–208, 2007. [PubMed:16988404].

[24] N. G. Gurskaya, V. V. Verkhusha, A. S. Shcheglov, D. B. Staroverov, T. V. Chepurnykh, A. F. Fradkov, S. Lukyanov, and K. A. Lukyanov. Engineering of a monomeric green-to-red photoactivatable fluorescent protein induced by blue light. *Nat. Biotechnol.*, 24(4):461–465, Apr 2006. [DOI:10.1038/nbt1191] [PubMed:16550175].

[25] H. Kim and J. S. Kim. A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.*, 15(5):321–334, May 2014. [DOI:10.1038/nrg3686] [PubMed:24690881].

[26] P. Collin, O. Nashchekina, R. Walker, and J. Pines. The spindle assembly checkpoint works like a rheostat rather than a toggle switch. *Nat. Cell Biol.*, 15(11):1378–1385, Nov 2013. [PubMed Central:PMC3836401] [DOI:10.1038/ncb2855] [PubMed:24096242].

[27] F. D. Urnov, E. J. Rebar, M. C. Holmes, H. S. Zhang, and P. D. Gregory. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.*, 11(9):636–646, Sep 2010.

[28] Y. Niu, B. Shen, Y. Cui, Y. Chen, J. Wang, L. Wang, Y. Kang, X. Zhao, W. Si, W. Li, A. P. Xiang, J. Zhou, X. Guo, Y. Bi, C. Si, B. Hu, G. Dong, H. Wang, Z. Zhou, T. Li, T. Tan, X. Pu, F. Wang, S. Ji, Q. Zhou, X. Huang, W. Ji, and J. Sha. Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell*, 156(4):836–843, Feb 2014. [DOI:10.1016/j.cell.2014.01.027] [PubMed:24486104].

[29] C. Mussolino, J. Alzubi, E. J. Fine, R. Morbitzer, T. J. Cradick, T. Lahaye, G. Bao, and T. Cathomen. TALENs facilitate targeted genome editing in human cells with high specificity and low cytotoxicity. *Nucleic Acids Res.*, May 2014. [DOI:10.1093/nar/gku305] [PubMed:24792154].

[30] T. Cermak, E. L. Doyle, M. Christian, L. Wang, Y. Zhang, C. Schmidt, J. A. Baller, N. V. Somia, A. J. Bogdanove, and D. F. Voytas. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.*, 39(12):e82, Jul 2011. [PubMed Central:PMC3130291] [DOI:10.1093/nar/gkr218] [PubMed:21493687].

[31] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, Aug 2012. [DOI:10.1126/science.1225829] [PubMed:22745249].

[32] B. Chen, L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G. W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, and B. Huang. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, 155(7):1479–1491, Dec 2013. [PubMed Central:PMC3918502] [DOI:10.1016/j.cell.2013.12.001] [PubMed:24360272].

[33] F. J. Mojica, C. Diez-Villasenor, J. Garcia-Martinez, and C. Almendros. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, Engl.)*, 155(Pt 3):733–740, Mar 2009. [DOI:10.1099/mic.0.023960-0] [PubMed:19246744].

[34] S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, 507(7490):62–67, Mar 2014. [DOI:10.1038/nature13011] [PubMed:24476820].

[35] M. Jinek, F. Jiang, D. W. Taylor, S. H. Sternberg, E. Kaya, E. Ma, C. Anders, M. Hauer, K. Zhou, S. Lin, M. Kaplan, A. T. Iavarone, E. Charpentier, E. Nogales, and J. A. Doudna. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*, 343(6176):1247997, Mar 2014. [DOI:10.1126/science.1247997] [PubMed:24505130].

[36] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. RNA-guided human genome engineering via Cas9. *Science*, 339(6121):823–826, Feb 2013. [PubMed Central:PMC3712628] [DOI:10.1126/science.1232033] [PubMed:23287722].

[37] Timothy T. Weil, Richard M. Parton, and Ilan Davis. Making the message clear: visualizing mrna localization. *Trends in Cell Biology*, 20(7):380 – 390, 2010.

[38] I. Izeddin, V. Récamier, L. Bosanac, I. I. Cisse, L. Boudarène, C. Dugast-Darzacq, F. Proux, O. Bénichou, R. Voituriez, O. Bensaude, et al. Distinct target search modes for c-Myc and P-TEFb revealed by single molecule tracking in live cells. *Under review*, 2014.

[39] James A Sawitzke, Lynn C Thomason, Nina Costantino, Mikhail Bubunenko, Simanti Datta, and Donald L Court. Recombineering:< i> in vivo</i>

genetic engineering in< i> e. coli, s. enterica</i>, and beyond. *Methods in enzymology*, 421:171–199, 2007.

[40] D. G. Gibson, L. Young, R. Y. Chuang, J. C. Venter, C. A. Hutchison, and H. O. Smith. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, 6(5):343–345, May 2009. [DOI:10.1038/nmeth.1318] [PubMed:19363495].

[41] Joe Sambrook. *Molecular Cloning: A Laboratory Manual, Third Edition (3 volume set)*. Cold Spring Harbor Laboratory Press, 2001.

[42] H. Inoue, H. Nojima, and H. Okayama. High efficiency transformation of Escherichia coli with plasmids. *Gene*, 96(1):23–28, Nov 1990.

[43] A. H. Morrison, M. Scheeler, J. Dubuis, and T. Gregor. Quantifying the Bicoid morphogen gradient in living fly embryos. *Cold Spring Harb Protoc*, 2012(4):398–406, Apr 2012. [PubMed Central:PMC4001741] [DOI:10.1101/pdb.top068536] [PubMed:22474658].

[44] Benjamien Moeyaert. Structural insight into fluorescent protein properties by rational design and characterization of photoconvertible, photoswitchable and phototoxic fluorescent proteins. Master's thesis, Katholieke Universiteit Leuven, 2010.

[45] Roger Heim, Andrew B Cubitt, and Roger Y Tsien. Improved green fluorescence. 1995.

[46] Y. Makino, K. Amada, H. Taguchi, and M. Yoshida. Chaperonin-mediated folding of green fluorescent protein. *J. Biol. Chem.*, 272(19):12468–12474, May 1997. [PubMed:9139695].

[47] B. T. Andrews, S. Gosavi, J. M. Finke, J. N. Onuchic, and P. A. Jennings. The dual-basin landscape in GFP folding. *Proc. Natl. Acad. Sci. U.S.A.*, 105(34):12283–12288, Aug 2008. [PubMed Central:PMC2527903] [DOI:10.1073/pnas.0804039105] [PubMed:18713871].

[48] D. A. Jackson, F. J. Iborra, E. M. Manders, and P. R. Cook. Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Mol. Biol. Cell*, 9(6):1523–1536, Jun 1998.

[49] H. Kimura, Y. Tao, R. G. Roeder, and P. R. Cook. Quantitation of RNA polymerase II and its transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure. *Mol. Cell. Biol.*, 19(8):5383–5392, Aug 1999.

[50] Z. W. Zhao, R. Roy, J. C. Gebhardt, D. M. Suter, A. R. Chapman, and X. S. Xie. Spatial organization of RNA polymerase II inside a mammalian cell nucleus revealed by reflected light-sheet superresolution microscopy. *Proc. Natl. Acad. Sci. U.S.A.*, 111(2):681–686, Jan 2014.

[51] A. Wilson, M. J. Murphy, T. Oskarsson, K. Kaloulis, M. D. Bettess, G. M. Oser, A. C. Pasche, C. Knabenhans, H. R. Macdonald, and A. Trumpp. c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev.*, 18(22):2747–2763, Nov 2004. [PubMed Central:PMC528895] [DOI:10.1101/gad.313104] [PubMed:15545632].

[52] Brian S Garrison and Derrick J Rossi. Controlling stem cell fate one substrate at a time. *Nature immunology*, 11(3):193, 2010.

[53] E. Guccione, F. Martinato, G. Finocchiaro, L. Luzi, L. Tizzoni, V. Dall' Olio, G. Zardo, C. Nervi, L. Bernard, and B. Amati. Myc-binding-site recognition in the human genome is determined by chromatin context. *Nat. Cell Biol.*, 8(7):764–770, Jul 2006.

[54] G. Kreth, J. Finsterle, J. von Hase, M. Cremer, and C. Cremer. Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. *Biophys. J.*, 86(5):2803–2812, May 2004.

[55] A. J. Walhout, J. M. Gubbels, R. Bernards, P. C. van der Vliet, and H. T. Timmers. c-Myc/Max heterodimers bind cooperatively to the E-box sequences located in the first intron of the rat ornithine decarboxylase (ODC) gene. *Nucleic Acids Res.*, 25(8):1493–1501, Apr 1997.

[56] S Robin, F Rodolphe, and S Schbath. *ADN, mots et modèles*. Belin, 2003.

[57] I. Kwon, M. Kato, S. Xiang, L. Wu, P. Theodoropoulos, H. Mirzaei, T. Han, S. Xie, J. L. Corden, and S. L. McKnight. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell*, 155(5):1049–1060, Nov 2013.

[58] S. van de Linde, A. Loschberger, T. Klein, M. Heidbreder, S. Wolter, M. Heilemann, and M. Sauer. Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nat Protoc*, 6(7):991–1009, Jul 2011. [DOI:10.1038/nprot.2011.336] [PubMed:21720313].

[59] S. H. Lee, J. Y. Shin, A. Lee, and C. Bustamante. Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). *Proc. Natl. Acad. Sci. U.S.A.*, 109(43):17436–17441, Oct 2012. [PubMed Central:PMC3491528] [DOI:10.1073/pnas.1215175109] [PubMed:23045631].

[60] H. Ochman, A. S. Gerber, and D. L. Hartl. Genetic applications of an inverse polymerase chain reaction. *Genetics*, 120(3):621–623, Nov 1988. [PubMed Central:PMC1203539] [PubMed:2852134].

[61] V. Adam, K. Nienhaus, D. Bourgeois, and G. U. Nienhaus. Structural basis of enhanced photoconversion yield in green fluorescent protein-like protein Dendra2. *Biochemistry*, 48(22):4905–4915, Jun 2009. [DOI:10.1021/bi900383a] [PubMed:19371086].

[62] Robert Sjöback, Jan Nygren, and Mikael Kubista. Absorption and fluorescence properties of fluorescein. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 51(6):L7–L21, 1995.

[63] William W Ward. Biochemical and physical properties of green fluorescent protein. 2005.

[64] Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns (Statistics in Practice)*. Wiley-Interscience, 1 edition, March 2008.

[65] Brian RIpley. *Spatial Statitics*. WIley, 1981.

[66] Adrian Baddeley. *Package 'spatstat'*. 2013.

[67] Adrian Baddeley and Rolf Turner. Spatstat: An r package for analyzing spatial point patterns. In *Journal of Statistical Software*, volume 12, pages 1–42, 2005.

[68] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[69] J. Chen, Z. Zhang, L. Li, B. C. Chen, A. Revyakin, B. Hajj, W. Legant, M. Dahan, T. Lionnet, E. Betzig, R. Tjian, and Z. Liu. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell*, 156(6):1274–1285, Mar 2014.

[70] O. Bénichou, C. Chevalier, J. Klafter, B. Meyer, and R. Voituriez. Geometry-controlled kinetics. *Nat Chem*, 2(6):472–477, Jun 2010.
•• In this theoretical work, the authors derive analytically the distribution of times that it takes for a diffusing molecule to reach a given target, i.e. the first passage time (FPT), in realistic confined geometries. Bénichou and coworkers demonstrate the emergence of two universality classes, namely compact and non-compact exploration, for any diffusion process. The authors show that only two parameters (the fractal dimension d_f and the dimension of the walk d_w) determine the universality class of the search process and, consequently, the stark dependence or lack of thereof of the FPT with the initial distance between the searcher and its target. [DOI:10.1038/nchem.622] [PubMed:20489716].

Geometry of the nucleus: a perspective on gene expression regulation

Maxime Woringer[1], Xavier Darzacq[1,*], and Ignacio Izeddin[1,*]

[1]Functional Imaging of Transcription; École normale supérieure; Institut de Biologie de l'ENS (IBENS); Inserm U1024; CNRS UMR 8197; Paris, France

March 30, 2014

Contents

Abstract

Gene expression control results from the combined interactions of the nearly hundred proteins forming the pre-initiation complex, thousands of transcription regulators, and genomic DNA. In the recent years, new technologies have revealed several key aspects of nuclear spatial organization that showed a fine interplay between the function of nuclear proteins, their 3D organization, and their dynamics. Here we review several concepts that link biochemical reactivity in the nucleus to its 3D spatial organization. We present the analogies between the emerging understanding of nuclear organization in the field of cell biology, and the more established disciplines of heterogeneous catalysis and the physics of random walks. We provide several recent examples showing how nuclear geometry affects protein reactivity in the nucleus.

Highlights

- Analogies between gene expression, target-search and heterogeneous catalysis.
- The fractal geometry of the nucleus architecture drives TFs diffusion.
- Diffusion of TFs is influenced both by geometrical factors and chemical parameters.
- TFs show diverse patterns of space exploration, influencing gene expression.

Introduction

Regulation of eukaryotic transcription and control of gene expression are two key questions in today's cellular and molecular biology [1]. The understanding of their physical and chemical principles is essential in many areas of applied science. Clear examples are cancer research, biological engineering, regenerative medicine or pharmacology.

Gene expression is regulated by transcription factors (TFs) interacting at specific *loci* to trigger gene activation. Through this interaction, the assembly of the pre-initiation complex (PIC) at promoters' sites leads to RNA polymerase II (Pol II) engagement in elongation. Our current understanding of this process includes the high mobility of diffusing TFs reaching for specific DNA sequences (referred as target-search) and the combinatorial assembly of the PIC. However, the spatial and geometric constraints that encompass protein-DNA and protein-protein interactions are often overlooked and not properly understood [2]. In addition, all biomolecular processes relevant to gene expression take place in a crowded and complex environment where regulation mechanisms operate at different levels of complexity.

The target-search of TFs in the nucleus are governed by diffusive processes. And while in yeast it has been shown that the search time of upstream TFs determines the gene activation rate [3], pure Brownian diffusion of TFs falls short to fully describe the efficiency and complexity of the gene expression process [4, 5, 6, 7]. Gene expression must thus be regulated by several other parameters spanning from exploration of the nuclear space to exploration of the space of protein conformations: variation of global and local concentrations, diversity in the target-search patterns and in space exploration, regulated docking affecting the conformation of both TF and its substrate.

The problems of target-search and reactivity have been formalized in different fields. Since more than a century, chemists have investigated the field of heterogeneous catalysis [8], accounting for diffusion and reaction on surfaces of reduced dimensionality. Likewise, following the seminal work of Pierre-Gilles de Gennes [9, 10], physicists have developed formalisms accounting for the diffusivity of molecules in random or disordered systems [11], potentially modifying their reactivity.

In this review we evaluate recent achievements in the understanding of the influence of geometrical factors on the regulation of transcription. We survey and compare the different formalisms used in biology, chemistry and physics in order to draw their similarities and differences. We aim to foster cross-disciplinary interactions among these fields, potentially leading to a more unified usage of these concepts.

1 Available space in the nucleus

While the mechanisms behind the regulation of gene expression are far from being fully understood, its very first step requires two or more biomolecules to interact at a given moment of time in a given position of the space. In a first approximation to this problem, we can consider the nucleus as a closed container in which a number of reactants diffuse prior to engage in a chemical reaction. In this idealized system, the kinetics

*To whom correspondence should be addressed: Ignacio Izeddin izeddin@biologie.ens.fr, Xavier Darzacq darzacq@biologie.ens.fr

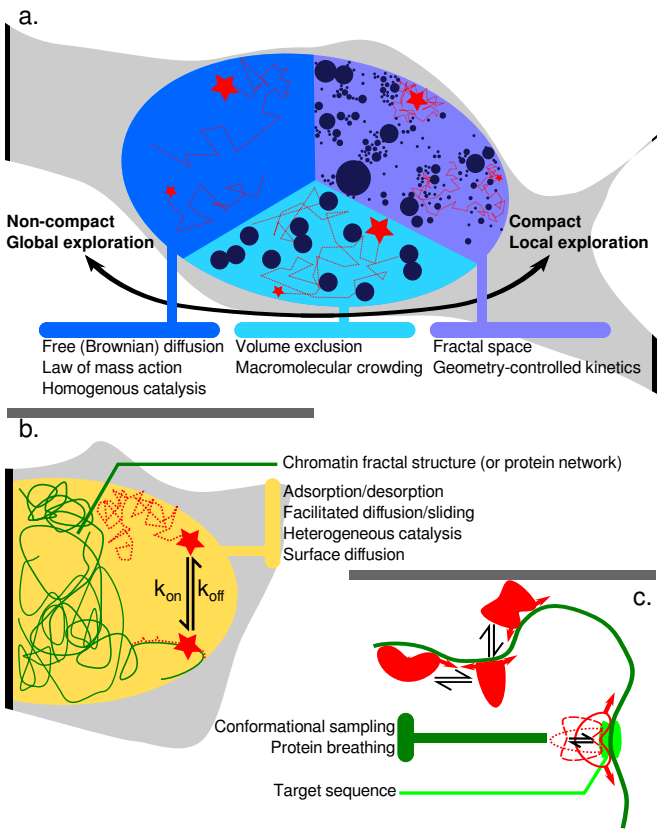


Figure 1: **TFs exploration patterns in the nucleus are highly diverse**. **(a).** Diffusion of TFs can occur in a space of *reduced volume*. This spans from free diffusion (thus performing a non-compact, global walk) –left–, to diffusion in a fractal medium, showing obstacles at all scales and realizing a recurrent, compact walk, –right–. **(b).** Diffusion of TFs can occur on a space of *reduced dimensionality*, here represented through binding and facilitated diffusion on a nuclear macromolecular network (such as DNA or proteins). TFs oscillate between 3D and less-than-3D diffusion. **(c).** TFs diffuse in the *conformational space*, and sample available conformations, exhibiting "protein breathing".

of the reaction can simply be derived from the law of mass action (given that the system were in equilibrium). As such, the reaction rate is proportional to the product of the concentrations of the participating molecules. To evaluate the reaction kinetics when a small number of reactants are involved, as often the case in gene expression [12], the first step is to assess the probability of encounter between reactants. In this scenario, the diffusion properties of the molecules, given by the Einstein-Smoluchowski equation, determine the first-encounter time [12, 13].

With such a simplified model of gene expression, it is easy to imagine the role of crowding, molecular exclusion, and local concentration in the kinetics of this process (Figure 1), and by extension in all the biochemistry of the cell. High molecular weight components in the nucleus, such as prominently but not exclusively chromatin, effectively reduce the accessible volume in which TFs are free to diffuse, potentially regulating the process of gene expression. A "rule of thumb" for the volume of a DNA is 1 nm^3/bp [1]. Thus, neglecting adsorbed water, the volume of human DNA is $\sim$ 2x3x10^9 = 6x10^9 nm^3. Similarly, the exclusion volume of nucleosomes can be computed[2], leading to an estimated volume of chromatin of $\sim$ 25 μm^3, which is a fraction of 12% of the volume of a human nucleus ($\sim$ 6 µm diameter [3]). Other estimates (10% in [15], 20%-50% in [16]) give similar order of magnitude. Such exclusion volume would at most change by a mere factor of two the rate of homogenous biochemical reactions. We must thus take into consideration other characteristics such as the complex geometry of nuclear organization or the heterogeneity of local molecular concentration. The former, as discussed below renders the calculations of exclusion volume invalid; regarding the latter, many nuclear components do not show a homogeneous spatial distribution in the nucleus [17], and it has been shown that the local concentration of Pol II is regulated, giving rise to significant differences at the local level throughout the nucleoplasm [18].

2 The complex geometry of the nucleus affects diffusion

An additional layer of complexity can be added to the target-search problem of TFs when taking into consideration the complexity of DNA packing in the nucleus. DNA exhibits a hierarchy of structures that spans from the molecular level up to the size of the nucleus. This includes coiling, wrapping, supercoiling, etc. of the DNA polymer but also the non-random organization of the genetic information in the nucleus and the existence of chromosomal territories [19, 20, 1, 21]. In recent years, growingly solid experimental evidence demonstrates that chromatin exhibits characteristics of a fractal structure [22, 16, 23] with a measurable fractal dimension (see Table 1, Figure 2 and [24]), which had been hypothesized almost thirty years ago [25, 26].

With these considerations in mind, the question of how much volume is excluded by chromatin becomes crucial. Mathematically, the complementary of a fractal displays the dimensionality of the fractal-embedding space (3D in our case) [27]. A single-point diffusing molecule in the complementary space would therefore display the same characteristics than in a three-dimensional volume. On the other hand, a particle with finite size can have an accessible space that is a fractal.

Even though computing the exclusion volume of a fractal (characterized by its fractal dimension d_f) requires strong assumptions, extensive work in the field of heterogeneous catalysis provides analytical and computational tools to address this question [28, 29, 30, 11]. Most of the current models in the field take two parameters into account: the fractal scaling regime ($\delta_{min}, \delta_{max}$) (i.e. the range of scales where the object can be regarded as fractal) and the size δ of the diffusing molecule. Exclusion volumes and diffusion properties of the molecules can then be derived. Under these assumptions, the available volume A for a diffusing molecule scales as a power of its size ($A \propto \delta^{2-d_f}$ [8]). Thus, the relevant parameter to estimate diffusible space is no longer the volume of nucleus constituents but its fractal dimension d_f.

An important question to elucidate is how the fractal structure effectively influences the diffusion of TFs. From a theoretical point of view, diffusion in a fractal structure is characterized by a deviation from the free, Brownian diffusion (Figure 1a, left) to an anomalous, subdiffusive behaviour (Figure 1a right), for instance observed by computing the mean square displacement (MSD) on single particle tracking (SPT) experiments (Table 1). In the context of the nucleus, several studies report anomalous diffusion [31, 16, 32], thus suggesting a fractal organization of the nucleus as one possible explanatory mechanism.

[1]. Bionumbers `http://bionumbers.hms.harvard.edu/`, accession number: 103778.

[2]. Cristal structure of the human nucleosome core, DOI: 10.2210/pdb2cv5/pdb, NDB ID: PD0676, derived from [14] and Bionumbers, accession numbers: 102977 and 102987.

[3]. Bionumbers, accession number: 105995.

3 Diffusion of TFs is altered by chemical interactions

Even though diffusion of a TF in the chromatin exclusion volume, a complex, possibly fractal medium, is an accurate representation of the nucleus, target-search models usually consider the fractal chromatin as an inert surface. In this scenario, apparent diffusion coefficients are only determined by the size of the TF (throughout exclusion volume and the scaling of diffusion coefficients with the radius), leaving little room for regulation since TFs exhibit very similar Stokes radii, in the order of a few nanometers. These models are also inconsistent with recent SPT observations, where TFs of comparable sizes show different exploratory behaviors [32], which cannot be fully accounted for by the fractal organization described above.

Indeed, such models neglect the widely described regulated interactions of TFs with DNA and other proteins [33, 34, 35]. Binding and unbinding rates (k_{on} and k_{off}) of these interactions can dramatically affect the apparent diffusion coefficient of molecules, a phenomenon recently evidenced in single-molecules studies in living cells [36, 37, 38, 32]. On the other hand, in the context of heterogeneous catalysis, the adsorption of reactants in intricate geometries has been well characterized. In this framework, molecules undergo successive binding/unbinding events on a surface (referred as chemisorption). During this process, both the TF and the adsorbed surface (DNA or protein network) experience conformational rearrangements [39], modifications that are analogous to the enzyme-substrate co-adaptation described in Koshland's induced fit model [40].

In addition, adsorbed TFs are not necessarily statically trapped: they can diffuse on the adsorbent, thus switching from a 3D space exploration to a "surface" of reduced dimensionality. This mechanism is known as facilitated diffusion in biology (see [41, 42] for theoretical considerations, and [43, 44, 45] for experimental evidence) and can be seen as a beautiful example of heterogeneous catalysis in living matter. Indeed, diffusion on a surface of reduced dimensionality increases encounter probabilities, thus reactivity. From a physical point of view, and following the nomenclature introduced by de Gennes [9], TFs can switch from a "non-compact" to a "compact" exploration (cf. Figure 1a, right and Figure 2) [46]. In a compact exploration, the molecule oversamples the explored space and visits a previously accessed site multiple times, thus performing a "recurrent walk" [47].

It is noteworthy to point out that facilitated diffusion can occur within any structure of reduced dimensionality. The adsorbent structure for TFs can be chromatin (of fractal dimension between two and three), but could also be any protein domain susceptible of forming a network in the nucleus, such as the C-terminal domain (CTD) of Pol II, histone tails, nuclear lamina, etc. Indeed, interacting proteins can form gels [48] or polymeric networks [49]. Furthermore, live cell experiments suggest the coexistence of intricate networks influencing the diffusion of TFs [32].

In addition to such geometry-controlled diffusion, taking into account biological reactivity is of particular relevance. Numerous post-translational modifications (such as phosphorylation, ubiquitylation or multimerization) affect TFs [40]. These regulations trigger dramatic changes in the space-exploring properties of the TF (plausibly switching between compact and non-compact modes of exploration).

4 TFs undergo explorations in the conformational space

When the TF finally reaches its target, the consequent reaction (whose final step can be transcription initiation) is a stochastic process [3, 50, 51]. In bacteria, the *lac* repressor repeatedly slides over its *lac* operator before binding [45]. Also, experiments on transcription elongation by Pol II show that, once bound to its target DNA sequence, elongation exhibits a high failure rate larger than 90% [52]. All in all, these examples indicate that the problem of transcription regulation cannot be reduced to a target-search process, even though it is an important first step in a complex sequence of events.

The bound TF has to overcome an activation energy barrier (E_a) to proceed to the final step of the reaction. At a molecular scale, the protein can be seen as a polymer diffusing in a narrow potential landscape. This landscape is determined by the electrostatic fields between residues that give the protein its shape [53]. Through thermal noise, the protein constantly explores different conformations within the potential landscape. The final reaction occurs when the protein reaches a conformation that drives it down into the adequate potential well.

From this perspective, attempts to characterize the "target size" [54] of the target-search process (or effective cross section of interaction) is reduced to a chimera. Such a size reflects the conformational sampling of the protein in a space of very high dimensionality (defined by the positions of the aminoacids in the protein) more than its diffusive motion. Rather than a size, this measure should be considered as a reaction probability reflecting the potential landscape sampling of the protein.

Conclusion

In this review, we have presented several formalisms used to describe diffusion in complex geometries, chemical adsorption, facilitated diffusion and molecular docking. Although each of them originated from unrelated works in the fields of biology, physics and chemistry, we highlight their common cornerstones in order to gain insight into eukaryotic gene expression regulation. Even though concepts still lack unification, we believe that in the near future, delving in the parallelisms between these fields will be fundamental to a deeper understanding of transcription.

In the nucleus, each TF senses a (sometimes dramatically) different environment depending on its physical and chemical properties, paving the way for highly diverse regulation of gene expression. Compact, local explorers can exhibit inhomogeneous concentrations throughout the nucleus, enabling concentration-based regulation processes. On the other hand, non-compact, global explorers such as c-Myc [32] can mediate global effects on the genome, which is consistent with its described role as a "global genome amplifier" [55] and "global chromatin remodeler" [56].

Furthermore, protein-DNA and protein-protein interactions are highly regulated and dynamic. A TF constantly switching between chromatin-bound and unbound states can jump from a DNA chain to another, thus escaping simple 1D sliding: it will diffuse on a surface of fractal dimension higher than one. Post-translational modification of the TF affinity for a biomolecular network in the nucleus (such as DNA, Pol II CTD, etc.) can lead to fundamental differences in diffusive behaviour, possibly influencing the patterns of gene expression.

When the TF has found its "geometrical" target, a second, conformational target-search takes place before the TF proceeds through the chemical reaction. This conformational

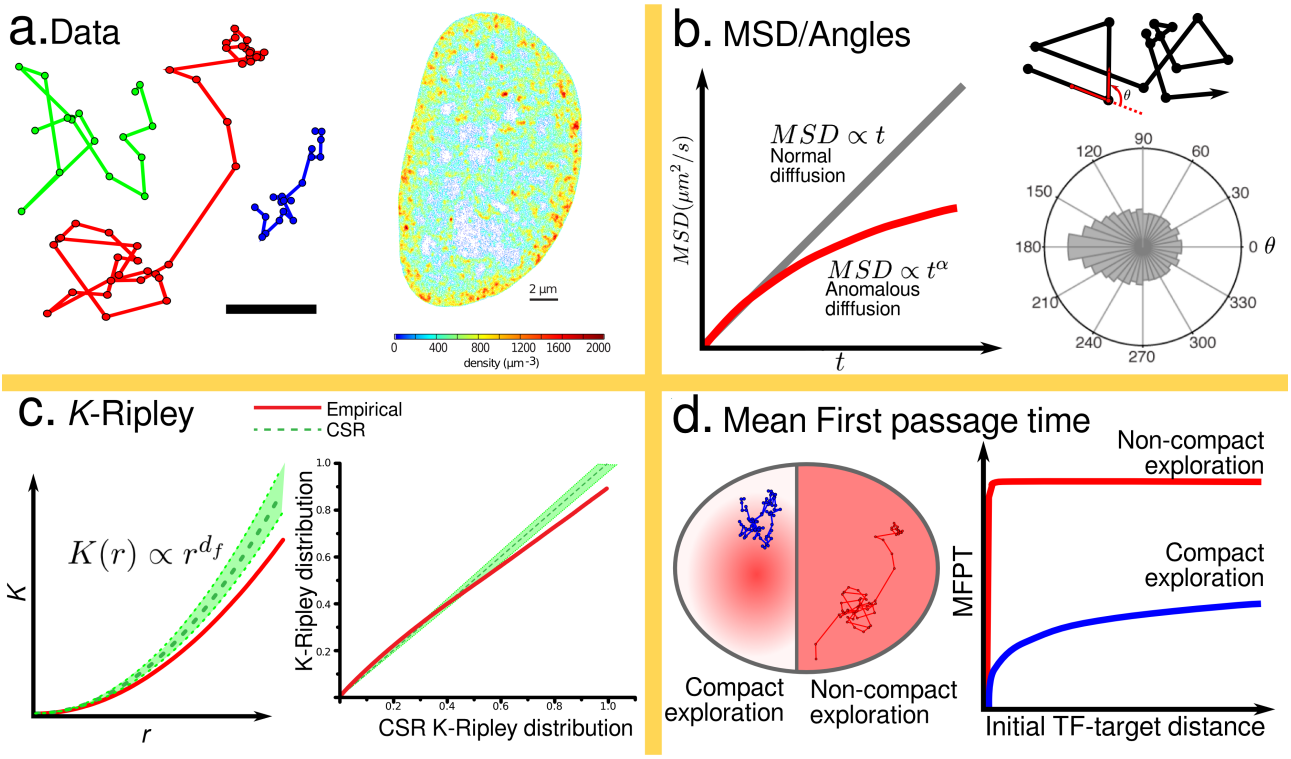


Figure 2: **Useful representations to analyze single molecule microscopy experiments**.
(a). Samples of single molecule data acquired during various types of experiments. *(left)* Single particle traces, 10 ms resolution. Scale bar: 500 nm *(right)* PALM –photoactivated localization microscopy– reconstruction of histone H2B fused to the photoconvertible fluorescent protein Dendra2.
(b). Representation of single particle tracking (SPT) data. *(left)* Mean square displacement (MSD). *(right)* –top– computation of angles between successive steps and the subsequent histogram –bottom–.
(c). Representation of PALM/STORM data. *(left)* Empirical K-Ripley function compared to the null model of complete spatial randomness (CSR). *(right)* Same data, but the empirical histogram is now plotted against the average of CSR, allowing a better determination of the fractal dimension d_f. Envelopes represent 95% confidence intervals over simulations of the CSR.
(d). Analytical tools allow for the computation of mean first passage times (MFPT) of a particle looking for a specific target and starting from a given position. In the case of compact exploration, the TF is a local explorer, recurrently visiting neighbouring sites. Conversely, in the non-compact case, distance from the target is not a relevant parameter and each point is visited with equal probability.

search is realized in a parameter space of high dimensionality. This dimensionality is further increased if we consider the ordered, combinatorial binding of coactivators to the TF.

All these space-exploring behaviors, assemblage routes, and regulatory processes are far from being mutually exclusive. Complex gene expression regulation in the nucleus actually arises from the coexistence of biochemical and biophysical mechanisms acting at all levels of gene expression. Nonetheless, from a genomic perspective, this complexity is required to tune the expression of $\sim$ 20 000 genes at a single gene resolution all along highly diverse processes such as cell cycle or differentiation. Conversely, from a TF's point of view, the nucleus should be regarded as a multiverse, where different proteins experience different landscapes with multiple scales, while being in the same space. Thus, the words of the French surrealist Paul Éluard seem more than appropriate: « il y a un autre monde mais il est dans celui-ci » (there is another world, but it is in this one).

Acknowledgements

References

[1] E. Heard and W. Bickmore. The ins and outs of gene regulation and chromosome territory organisation. *Curr Opin Cell Biol*, 19(3):311–316, Jun 2007. [DOI:10.1016/j.ceb.2007.04.016] [PubMed:17467967].

[2] D. Normanno, M. Dahan, and X. Darzacq. Intra-nuclear mobility and target search mechanisms of transcription factors: a single-molecule perspective on gene expression. *Biochim Biophys Acta*, 1819:482–493, Jun 2012.

[3] D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao, and R. H. Singer. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science*, 332(6028):475–478, Apr 2011. [PubMed Central:PMC3152976] [DOI:10.1126/science.1202142] [PubMed:21512033].

[4] S. A. Isaacson, D. M. McQueen, and C. S. Peskin. The influence of volume exclusion by chromatin on the time required to find specific DNA binding sites by diffusion. *Proc Natl Acad Sci USA*, 108(9):3815–3820, Mar 2011.
•• In this work, the authors develop a mathematical model to investigate the influence of the exclusion volume of chromatin on the target-search time of individual molecules. Using three-dimensional structured illumination microscopy images of a DAPI stained nucleus (from Schermelleh et al., Science, 2008), Isaacson and colleagues build an effective potential energy field that tends to impede the diffusing molecule to enter high-density chromatin regions. With this model, they find that when the target is located in regions of low chromatin density, a moderate value of volume exclusivity (the strenght of the potential field) leads to faster target-search times. Interestingly, when the DAPI image is randomized (but conserving the density inhomogeneities), this reduction of the time required to find the target is lost. This indicates that the efficiency improvement of the target-search process due to volume exclusion is related to the spatial correlation of chromatin, i.e. its geometry in the nucleus. [PubMed Central:PMC3048119] [DOI:10.1073/pnas.1018821108] [PubMed:21300894].

[5] T Chou and M. R D'Orsogna. First passage time problems in biology. *Submitted*, 2014.

[6] C. Chevalier, O. Bénichou, B. Meyer, and R. Voituriez. First-passage quantities of brownian motion in a bounded domain with multiple targets: a unified approach. *Journal of Physics A: Mathematical and Theoretical*, 44(2):025002+, January 2011.

[7] A. D. Riggs, S. Bourgeois, and M. Cohn. The *lac* repressor-operator interaction. 3. kinetic studies. *J Mol Biol*, 53(3):401–417, November 1970.

[8] U. Ozkan. *Design of heterogeneous catalysts*. John Wiley & Sons, 2009.

[9] P. G. de Gennes. Kinetics of diffusion-controlled processes in dense polymer systems. I. nonentangled regimes. *The Journal of Chemical Physics*, 76(6):3316–3321, 1982.

[10] P. G. de Gennes. Kinetics of diffusion-controlled processes in dense polymer systems. II. effects of entanglements. *The Journal of Chemical Physics*, 76(6):3322–3326, 1982.

[11] K. Lindenberg, W.-S. Sheu, and R. Kopelman. Scaling properties of diffusion-limited reactions on fractal and euclidean geometries. *J Stat Phys*, 65(5-6):1269–1283, 1991.

[12] S.A. Rice. *Diffusion-Limited Reactions*. Comprehensive Chemical Kinetics. Elsevier Science, 1985.

[13] D. ben Avraham and S. Havlin. *Diffusion and Reactions in Fractals and Disordered Systems*. Cambridge University Press, January 2005.

[14] Y. Tsunaka, N. Kajimura, S. Tate, and K. Morikawa. Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Res*, 33(10):3424–3434, 2005. [PubMed Central:PMC1150222] [DOI:10.1093/nar/gki663] [PubMed:15951514].

[15] K. Rippe. Dynamic organization of the cell nucleus. *Curr Opin Genet Dev*, 17(5):373–380, Oct 2007. [DOI:10.1016/j.gde.2007.08.007] [PubMed:17913491].

[16] A. Bancaud, S. Huet, N. Daigle, J. Mozziconacci, J. Beaudouin, and J. Ellenberg. Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. *EMBO J*, 28(24):3785–3798, Dec 2009. [PubMed Central:PMC2797059] [DOI:10.1038/emboj.2009.340] [PubMed:19927119].

[17] J. M. Dowen, S. Bilodeau, D. A. Orlando, M. R. Hubner, B. J. Abraham, D. L. Spector, and R. A. Young. Multiple structural maintenance of chromosome complexes at transcriptional regulatory elements. *Stem Cell Reports*, 1(5):371–378, 2013. [PubMed Central:PMC3841252] [DOI:10.1016/j.stemcr.2013.09.002] [PubMed:24286025].

[18] I. I. Cisse, I. Izeddin, S. Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan, and X. Darzacq. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science*, 341(6146):664–667, Aug 2013. [DOI:10.1126/science.1239053] [PubMed:23828889].

[19] J. H. Gibcus and J. Dekker. The hierarchy of the 3D genome. *Mol Cell*, 49(5):773–782, Mar 2013. [PubMed Central:PMC3741673] [DOI:10.1016/j.molcel.2013.02.011] [PubMed:23473598].

[20] G. Cavalli and T. Misteli. Functional implications of genome topology. *Nat. Struct. Mol. Biol.*, 20(3):290–299, Mar 2013. [DOI:10.1038/nsmb.2474] [PubMed:23463314].

[21] H. Albiez, M. Cremer, C. Tiberi, L. Vecchio, L. Schermelleh, S. Dittrich, K. Küpper, B. Joffe, T. Thormeyer, J. von Hase, et al. Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. *Chromosome Res*, 14(7):707–733, November 2006.

[22] S. Huet, C. Lavelle, H. Ranchon, P. Carrivain, J. M. Victor, and A. Bancaud. Relevance and limitations of crowding, fractal, and polymer models to describe nuclear architecture. *Int Rev Cell Mol Biol*, 307:443–479, 2014. [DOI:10.1016/B978-0-12-800046-5.00013-8] [PubMed:24380602].

[23] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009. [PubMed Central:PMC2858594] [DOI:10.1126/science.1181369] [PubMed:19815776].

[24] V Récamier, I Izeddin, L Bosanac, M Dahan, F Proux, and X Darzacq. Single cell correlation fractal dimension of chromatin: A framework to interpret 3D single molecule super-resolution. *Nucleus*, 5(1), 2014.
• In this work, the authors use photoactivated localization microscopy (PALM) combined with adaptive optics to retrieve the 3D position of hundreds of thousands of H2B histone molecules of the chromatin structure in the nucleus of human cells, with a localization accuracy of $\sim$ 15 nm in XY and $\sim$ 30 nm in Z. Besides observing chromatin density inhomogoneity at the single cell level with molecular resolution, they develop a computational approach to retrieve the so-called K-Ripley distribution of these histone molecules, corrected for the enclosing envelope of the nucleus and for bias of detection along the optical axis. The K-Ripley distribution is the distribution of distances between every two H2B molecules of the chromatin structure. When compared to a randomized distribution of points in the same nuclear envelope (complete spatial randomness –CSR– test), they observe that the ratio between them scales with a power law within the observable range (30 nm to 3 μm), which is indicative of a fractal distribution of chromatin. From this computation, they retrieve a direct measure of the fractal dimension of chromatin d_f=2.7, a consistent value across the cell population. DOI:10.4161/nucl.28227.

[25] Y. Grosberg, S. K Nechaev, and E. Shakhnovich. The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de physique*, 49(12):2095–2100, 1988.

[26] A. Grosberg, I. Rabin, S. Khavlin, and A Nir. Self-similarity in the structure of DNA: why are introns needed? *Biofizika*, 38(1):75–83, 1992.

[27] A. L. Barabási and H. E. Stanley. *Fractal Concepts in Surface Growth*. Cambridge University Press, Cambridge, UK, 1995.

[28] E. Albano. Reaction kinetics in fractals. In Robert A. Meyers, editor, *Encyclopedia of Complexity and Systems Science*, pages 7565–7583. Springer New York, 2009.

[29] P. Pfeifer and D. Avnir. Chemistry in noninteger dimensions between two and three. I. fractal theory of heterogeneous surfaces. *J Chem Phys*, 79(7):3558–3565, 1983.

[30] M-O Coppens and G. Froment. Knudsen diffusion in porous catalysts with a fractal internal surface. *Fractals*, 3(04):807–820, 1995.

[31] G. Guigas, C. Kalla, and M. Weiss. Probing the nanoscale viscoelasticity of intracellular fluids in living cells. *Biophys J*, 93(1):316–323, Jul 2007. [PubMed Central:PMC1914431] [DOI:10.1529/biophysj.106.099267] [PubMed:17416631].

[32] I. Izeddin, V. Récamier, L. Bosanac, I. I. Cisse, L. Boudarène, C. Dugast-Darzacq, F. Proux, O. Bénichou, R. Voituriez, O. Bensaude, et al. Distinct target search modes for c-Myc and P-TEFb revealed by single molecule tracking in live cells. *Under review*, 2014.
• Using single-molecule tracking of transcription factors (TFs) in live human cells, the authors show how different TFs (P-TEFb and c-Myc) explore their surrounding space in a fundamentally different manner, thus coining the concept of "protein-specific geometry of the nucleus". A combination of analytical approaches and numerical simulations is used to interpret the experimental data in the context of compact and non-compact exploration. With this approach, the authors provide a novel framework to study nuclear proteins with a wide range of mobility, and a better understanding of the basis for spatial cooperation and assembly of macromolecular complexes in the nucleus.

[33] I. Kwon, M. Kato, S. Xiang, L. Wu, P. Theodoropoulos, H. Mirzaei, T. Han, S. Xie, J. L. Corden, and S. L. McKnight. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell*, 155(5):1049 – 1060, 2013.
•• When translocating the products of the low-complexity (LC) domains of three RNA-binding proteins (FUS/EWS/TAF15), they function as transcription activation domains. The authors use this system to demonstrate that the network of polymeric fibers formed from these domains binds to the C-terminal domain (CTD) of RNA polymerase II. By mutation analysis, they also show that the binding between the CTD and the polymers correlates with the strength of transcriptional activity. Moreover, the binding between the CTD and the fibers formed by the LC domains is reversible upon phosphorylation of the CTD, suggesting a model of how RNA polymerase II is recruited to active genes and released for elongation after phosphorylation of the CTD. This and the previous works from the McKnight lab show how protein (polymeric) networks in the nucleus can induce a certain degree of organization implicated in the regulation of transcriptional activity.

[34] G. D. Stormo and Y. Zhao. Determining the specificity of protein-DNA interactions. *Nat Rev Genet*, 11(11):751–760, Nov 2010. [DOI:10.1038/nrg2845] [PubMed:20877328].

[35] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*, 79:233–269, 2010. [PubMed Central:PMC3285485] [DOI:10.1146/annurev-biochem-060408-091030] [PubMed:20334529].

[36] D. Mazza, A. Abernathy, N. Golob, T. Morisaki, and J. G. McNally. A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Res*,

40(15):e119, Aug 2012. [PubMed Central:PMC3424588] [DOI:10.1093/nar/gks701] [PubMed:22844090].

[37] J. C. Gebhardt, D. M. Suter, R. Roy, Z. W. Zhao, A. R. Chapman, S. Basu, T. Maniatis, and X. S. Xie. Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nat Methods*, 10(5):421–426, May 2013. [PubMed Central:PMC3664538] [DOI:10.1038/nmeth.2411] [PubMed:23524394].

[38] J. Chen, Z. Zhang, L. Li, B.-C. Chen, A. Revyakin, B. Hajj, W. Legant, M. Dahan, T. Lionnet, E. Betzig, et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell*, 156(6):1274 – 1285, 2014.

[39] T. Sunami and H. Kono. Local conformational changes in the DNA interfaces of proteins. *PLoS ONE*, 8(2):e56080, 2013. [PubMed Central:PMC3571985] [DOI:10.1371/journal.pone.0056080] [PubMed:23418514].

[40] G. Collier and V. Ortiz. Emerging computational approaches for the study of protein allostery. *Arch Biochem Biophys*, 538(1):6 – 15, 2013.

[41] O. G. Berg, R. B. Winter, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20(24):6929–6948, Nov 1981. [PubMed:7317363].

[42] O. Bénichou, C. Chevalier, B. Meyer, and R. Voituriez. Facilitated diffusion of proteins on chromatin. *Phys Rev Lett*, 106:038102, Jan 2011.

[43] J. Gorman, F. Wang, S. Redding, A. J. Plys, T. Fazio, S. Wind, E. E. Alani, and E. C. Greene. Single-molecule imaging reveals target-search mechanisms during DNA mismatch repair. *Proc Natl Acad Sci USA*, 109(45):E3074–E3083, 2012.

[44] J. Gorman and E. C. Greene. Visualizing one-dimensional diffusion of proteins along DNA. *Nat Struct Mol Biol*, 15(8):768–774, Aug 2008. [DOI:10.1038/nsmb.1441] [PubMed:18679428].

[45] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf. The *lac* repressor displays facilitated diffusion in living cells. *Science*, 336(6088):1595–1598, Jun 2012.
•• The authors provide the first experimental observation of facilitated diffusion in living cells. Using single-molecule imaging of the lac repressor in bacteria, the authors report a combination of 3D diffusion plus 1D sliding on DNA. Using a system with two identical lac operator sites (the specific target of the lac repressor) and placing them on DNA at increasing spacer distances, they provide a measurement for the sliding length of the TF to be 45 ± 10 base pairs. Furthermore, using blockroads for sliding at both sites of the operator, Hammar and colleagues determine that the repressor slides several times over its operator before binding in more than 90% of the cases. [DOI:10.1126/science.1221648] [PubMed:22723426].

[46] O. Bénichou, C. Chevalier, J. Klafter, B. Meyer, and R. Voituriez. Geometry-controlled kinetics. *Nat Chem*, 2(6):472–477, Jun 2010.
•• In this theoretical work, the authors derive analytically the distribution of times that it takes for a diffusing molecule to reach a given target, i.e. the first passage time (FPT), in realistic confined geometries. Bénichou and coworkers demonstrate the emergence of two universality classes, namely compact and non-compact exploration, for any diffusion process. The authors show that only two parameters (the fractal dimension d_f and the dimension of the walk d_w) determine the universality class of the search process and, consequently, the stark dependence or lack of thereof of the FPT with the initial distance between the searcher and its target. [DOI:10.1038/nchem.622] [PubMed:20489716].

[47] G. Pólya. Über eine aufgabe der wahrscheinlichkeitsrechnung betreffend die irrfahrt im straßennetz. *Math Ann*, 84(1-2):149–160, 1921.

[48] W. A. Petka, J. L. Harden, K. P. McGrath, D. Wirtz, and D. A. Tirrell. Reversible hydrogels from self-assembling artificial proteins. *Science*, 281(5375):389–392, 1998.

[49] M. Kato, T. W. Han, S. Xie, K. Shi, X. Du, L. C. Wu, H. Mirzaei, E. J. Goldsmith, J. Longgood, J. Pei, et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell*, 149(4):753–767, May 2012. [DOI:10.1016/j.cell.2012.04.017] [PubMed:22579281].

[50] A. Sanchez and I. Golding. Genetic determinants and cellular constraints in noisy gene expression. *Science*, 342(6163):1188–1193, Dec 2013. [DOI:10.1126/science.1242975] [PubMed:24311680].

[51] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, Oct 2008. [PubMed Central:PMC3118044] [DOI:10.1016/j.cell.2008.09.050] [PubMed:18957198].

[52] X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, and R. H. Singer. In vivo dynamics of RNA polymerase II transcription. *Nat Struct Mol Biol*, 14(9):796–806, Sep 2007. [DOI:10.1038/nsmb1280] [PubMed:17676063].

[53] X. Salvatella. Understanding protein dynamics using conformational ensembles. *Adv Exp Med Biol*, 805:67–85, 2014. [DOI:10.1007/978-3-319-02970-2$_3$] [PubMed:24446357].

[54] S. E. Halford and J. F. Marko. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res*, 32(10):3040–3052, 2004. [PubMed Central:PMC434431] [DOI:10.1093/nar/gkh624] [PubMed:15178741].

[55] Z. Nie, G. Hu, G. Wei, K. Cui, A. Yamane, W. Resch, R. Wang, D. R. Green, L. Tessarollo, R. Casellas, and et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151:68–79, Sep 2012.

[56] S. He, Z. Liu, D.-Y. Oh, and C. J. Thiele. MYCN and the epigenome. *Front Oncol*, 3:1, Jan 2013.

[57] S. Burov, S. M. Tabei, T. Huynh, M. P. Murrell, L. H. Philipson, S. A. Rice, M. L. Gardel, N. F. Scherer, and A. R. Dinner. Distribution of directional change as a signature of complex dynamics. *Proc Natl Acad Sci USA*, 110(49):19689–19694, Dec 2013. [PubMed Central:PMC3856831] [DOI:10.1073/pnas.1319473110] [PubMed:24248363].

[58] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns (Statistics in Practice)*. Wiley-Interscience, 1 edition, March 2008.

Measure	Origin	Formula	Explanation	Provides	Main Bias
MSD (mean square displacement) Fig. 1b, left	SPT	$MSD(t) = \langle [x(t) - x(0)]^2 \rangle = 2dD_c t^\alpha \propto t^{2/d_w}$ - Where $x(t)$ stands for the particle coordinates at time t, - d is the dimension of the space (1, 2 or 3), - D_c is the diffusion coefficient of the molecule, - d_w is the dimension of the walk	- The MSD is widely used to determine (apparent) diffusion coefficients. But its scaling α coefficient also contains information about the way the particle explores the space. - When $\alpha > 1$, the particle undergoes a superdiffusive or directed motion. If $\alpha = 1$, it is a Brownian diffusion. $\alpha < 1$ is indicative of a subdiffusive behaviour, which is the signature of a constrain in space or time in the diffusion of the particle. - See [32] and [28].	D_c, α, d_w	- In a heterogenous mix of different diffusive behaviours, overly detection of slow versus fast moving particles may induce a bias on the shape of the $MSD(t)$ curve. - Very hard to distinguish transitions in the dynamics of the diffusing particle.
Angular distribution Fig. 1b, right	SPT	Histogram of angles between consecutive translocations.	- Bridges the dynamics properties of the particle (measured traces) with its geometrical properties (the way it explores the space). - Offers information about the underlying space where the diffusion takes place. - See [32, 57].		Requires a high number of translocations to build the histogram and its evolution with increasing Δt.
K-ripley Fig. 1c	PALM / STORM	Distribution of number of neighbours within radius r from every detected particle, ie. the cumulative histogram of distances between points.	- In an unconfined geometry (no edge effect) with points uniformely dispersed (complete spatial randomness), $K(r)$ scales as πr^2 or πr^3 for 2D and 3D, respectively. - $K(r)$ or its derivatives give insight about the size of clusters (if any), or the fractal dimension when it scales as a power law ($K(r) \propto r^{d_f}$). - See [24] and [58].	d_f, CSR test	- Requires 3D data to derive fractal dimensions larger than $d_f = 2$. - Can be tested against CSR. - Data have to be border-corrected in case of confined geometries.
MFPT (mean first passage time) Fig. 1d	d_f, d_w, enclosing geometry	$\frac{\langle T_{TS} \rangle}{\langle T \rangle_T} = \Pi_{TS} \propto \begin{cases} 1 - \kappa(\frac{1}{r})^{d_f - d_w} & \text{(non-compact)} \\ (\frac{r}{R})^{d_w - d_f} & \text{(compact)} \end{cases}$ - $\langle T_{TS} \rangle$ denotes the mean first passage time of a molecule starting its walk from S and reaching a target T at a distance r. $\langle \bar{T} \rangle_T$ denotes is the mean of $\langle T_{TS} \rangle$ respective to all starting points in the geometry. - Π_{TS} is a function of the geometry and of the position of both the starting site and the target. - κ is a constant and R is the characteristic scale of the geometry.	This gives insights on the distance-dependence of the mean first passage time. - Where compact refers to $d_w > d_f$ and non-compact to $d_w < d_f$ (see 3). - For analytical derivation/simulations, see [46].		The geometrical factor Π_{TS} has to be computed numerically. Requires prior knowledge of the geometry of the enclosing volume.

Table 1: Useful computational and mathematical tools to analyze superresolution microscopy datasets and their main bias. Illustrations can be found in Figure 2. SPT: Single particle tracking, PALM: Photoactivated localization microscopy. STORM: Stochastic optical reconstruction microscopy. CSR: complete spatial randomness

4 Supplementary materials

4.1 Supplementary methods

4.1.1 Microscopy

Superresolution techniques

PALM This process of imaging is implemented with photoactivatable/photoconvertible fluorescent proteins: the proteins are initially in a dark state at the wavelength λ_1 (such as 561 nm) of observation. They are then converted to a florescent state (usually using a low intensity second laser of wavelength λ_2, for instance violet 405 nm) and bleached by the first laser, allowing to repeat the process with some of the remaining unconverted unbleached molecules.

dSTORM When stained, the proteins of interest are all fluorescent, and it is not possible to resolve single molecules. However, strong laser excitation can switch off most of the fluorophores and allow to resolve single molecules: some of them are bleached and will never be seen again, but a fraction will be sent to a triplet, non fluorescent state for an arbitrary amount of time, and will eventually become fluorescent again, ready to image. Single-molecule regimes occur naturally in-vivo because the reducing cytoplasm favors transitions from and to the triplet state. In the case of fixed cells, a "STORM" buffer containing a reducing agent and an oxygen-scavenging enzymatic reaction was used.

Imaging Concretely, cells were seeded and grown on 25 mm ethanol-cleaned coverslips >24 h before imaging. Depending on the experiment, cells were either kept live, stained with Halo-ligand dyes and/or fixed with 4% paraformaldehyde and stained with a standard immunofluorescence protocol. Medium was changed to either PBS (fixed sample, PALM), phenol-free DMEM (live sample) or STORM buffer (50 mM Tris-HCL, pH 8.0, 10 mM NaCl, 10% glucose, 100 mM β-mercaptoethylamine and an oxygen scavenging reaction: 0.6 mg/ml glucose oxydase, 0.17 mg/ml catalase, [58] and Nikon superresolution imaging brochure) and imaged with a 100x immersion oil objective on a 37°C clamped stage. Fluorescent beads (TeraSpeck, Life Technologies) we sparsely placed on the coverslip allowing for PSF shaping calibration and to correct the stage drift during superresolution acquisition. Approximately 30000 frames were acquired under inclined-total internal reflection fluorescence microscopy (TIRF) with a laser wavelength matching the excitation maximum of the fluorophore at a laser power of 100-300 mW.

Choice of fluorophores and labelling strategy Since c-Myc expression level is highly relevant to the homeostasis of the cell, we tried to limit c-Myc overexpression as much as possible. This was achieved by using immunofluorescence staining instead of transfections whenever possible, and by trying to generate an endogenously tagged *MYC* cell line (see section 1.5).

Fluorophores the fluorophores used during this internship are detailed in Table 1. We chose this set of fluorophores among a large available palette [58] for the following reasons:

- Alexa647 is a far red fluorophore, compatible both with TMR and Dendra2 (and the labeling strategy – immunofluorescence – doesn't interfere with the labeling of the two other molecules, potentially allowing multicolor imaging). It was used conjugated to a primary or a secondary antibody.
- TMR is a bright organic dye that can be conjugated to HaloTag fusion proteins in live cells.
- The Dendra2 protein doesn't require any staining, and its photophysics ensures

that each fluorophore is statistically detected at most once before bleaching [59], it is thus suitable to detect submicon-size clusters in a unbiased way.

Table 1: Fluorophores used during the internship. Fluor.: fluorophore, Abs./Em.: absorption/emission maxima (nm), Org.: organic dye, Prot.: proteic, genetically encoded fluorescent protein, gr.: green, unconverted form, re.: red, photoconverted form.

Fluor.	Type	Abs./Em.	Usage
Alexa647	Org.	650/665	IF-STORM
TMR	Org.	554/580	SPT/STORM
Dendra2	Prot.	490/507 gr. 553/573 re.	PALM

Immunofluorescence (IF) we performed IF on cells and used the following primary antibodies: c-Myc (mouse epitope, ref. sc-40, Santa Cruz Biotechnology – SCBT –, conjugated with Alexa647), histone H3 trimetyl-K4 (H3K4me3, rabbit antibody ab8580, Abcam), secondary antibodies were purchased conjugated with Alexa 647 (Life Technologies). We tried to assess the specificity of the c-Myc staining by RNAi (sc-29226, Santa Cruz Biotechnology). The RNAi was performed in one round of starvation using Lipofectamin (Invitrogen).

Transfections $\sim$ 500 ng/well of one (or more) of the constructs described below were transfected (XFect, Clontech) on cells 24 hrs after seeding on a 6-well plate. 2.5-4.5 µg of CMV or PUC19 were cotransfected to reach a total of 3-5 µg transfected DNA, according to the manufacturer.

Determination of the pointing accuracy We assayed the pointing accuracy on one of the acquisitions according to the protocol described in Figure 16. In this protocol, repeated imaging of a bead of fixed position and located next to the cell of interest is used to estimate the pointing incertainties in 3D (σ_x, σ_y and σ_z). Its intensity is rescaled to the average intensity of a detection, and noise extracted from another region of the image is added to the synthetic PSF. This artificial data is then processed as a normal movie, and finally the resolution in 3D is computed. This method is likely to underestimate the pointing accuracy because in that case, the movie cannot be drift-corrected. This method lead to a resolution of 12 nm in x, 17 nm in y and 50 nm in z, which is in the higher range of the publications we reviewed [13, 5].

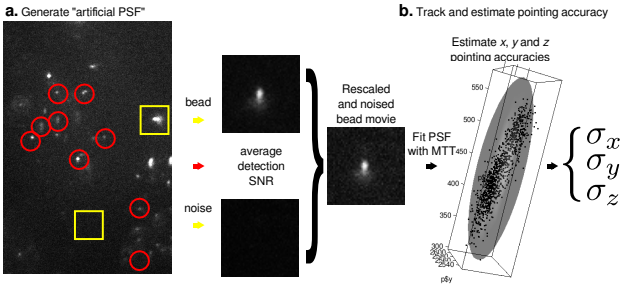


Figure 16: Procedure to determine the 3D pointing accuracy on real data. (a). yellow: a bead and a region of noise are extracted from a subset of $\sim$ 1000 frames from a movie of interest. red: for an important number of spots, both the peak intensity and the signal-to-noise ratio (SNR) are computed. The bead aveage intensity is then rescaled according to the estimated SNR and average peak. Noise is then added to the image to compensate for the proportional downscaling of the noise when normalizaing the signals. The synthetic movie is then fitted with MTT and a 3D gaussian is fitted to the resulting cluster of detections.

4.1.2 Functional assays

Table 2: Antibodies used for the functional characterization of the tagged versions of c-Myc, Abbreviations in parenthesis denote the organism where the antibody was raised: R.: rabbit, G.: goat, M.: mouse. Prov: provider, SCBT: Santa Cruz Biotechnology, Inc., RL: Rockland, Inc, CoIP: Co-Immunoprecipitation, WB: Western Blot., 2^{ary}: secondary antibody, HRP: conjugated to horseraddish peroxydase.

Antibody	Prov.	Reference	Used for
anti-MAX (R.)	SCBT	sc-765	CoIP and WB
anti-cMyc (R.)	Abcam	ab-3072	WB
anti-TRRAP (G.)	SCBT	sc-5405	not used yet (WE
anti-FLAG			not used yet (Col
anti-Rabbit-IgG	RL.	18-8816-33	2^{ary}, HRP, WB.

The anti-rabbit-IgG used is specific of the non SDS-denaturated (non-reduced) rabbit IgG, allowing to detect the primary antibody used for blotting without detecting the IgGs used for the CoIP.

4.1.3 CRISPR cell line characterization

In case of isolation of cell lines showing integration at the right site, the following characterizations would have been performed:

- *Southern blot* : a probe is designed against the tagged *MYC* sequence, and the colony PCR product is hybridized against the probe.
- *Inverse PCR* [60]: this technique allows to map unspecific integration sites that can superimpose the adequate insertion site. NT2 genomic DNA is digested in fragments of a few kilobases in length, and these fragments are circularized using a ligation in diluted conditions. PCR is then performed using a forward and a reverse primer located inside the *MYC* gene, amplifying the flanking regions of the gene. The amplicons are then cloned, sequenced and the results are finally mapped to the human genome.

4.1.4 Titration of the amount of fluorescent proteins

We review here two methods to precisely assay the quantity of protein actually fluorescent in a solution.

Quantum yield-based approach It is clear that traditional protein assays, such as native gels or absorbance quantitation might not account for bleached proteins. We thus designed an assay derived from the quantum yield determination protocol used in chemistry. Indeed, actually non fluorescent proteins show a quantum yield (QY, that is the ratio between the number of photons received divided by the number of photons emitted) close to zero, whereas the quantum yield of Dendra2 has been experimentally determined (green form: 0.50 [4] and 0.55 for the red form, buffer conditions not detailed. 0.45 for the red form in 100 mM potassium phosphate buffer, pH 5, and 0.61 in 100 mM potassium carbonate buffer, pH 9 [61], and 0.55 [44], buffer conditions not detailed. Knowing the actual quantum yield of Dendra2 and measuring the apparent quantum yield of our purified solution, it becomes possible to back-calculate the real concentration of the purified solution.

Since the photophysical parameters (such as absorption/emission spectra and quantum yields) are highly environment dependent ([44]), we reproduced the buffer conditions used by the authors who determined such parameters.

Measuring the ratio between the absorbed and emitted photons is challenging, since one has to account for the absorption of the cuvette, changes in the beam intensity as the lamp gets old, undesirable self-quenching due to concentration effects, etc.

A simple solution to overcome these issues is make the measurement relative to a fluorophore of similar spectral properties (to avoid sensitivity variability of the detector across different wavelengths) and to use a gradient of concentrations, to avoid both self-quenching and non-linearities of the detector. Under this setup, the quantum yield is given by the slope of measured fluorescence versus concentration. Concretely, the measurements have to be corrected for the refractive index of the medium, and the relative quantum yield is given by the formula:

$$\Phi_x = \Phi_{standard} \frac{grad_x}{grad_{standard}} \frac{\eta_x^2}{\eta_{standard}^2}$$

where Φ_x and $\Phi_{standard}$ denote the QY of the purified protein and the fluorescent standard (respectively), η_x and $\eta_{standard}$ stand for the refractive index of the purified protein and the standard, respectively. In our case, the ratio was neglected since we work with diluted solutions.

Samples were diluted at a concentration around 1 µM and read using a fluorescence plate reader

[4]. According to Evrogen's Dendra2 "detailed description" on their website.

(EnVision Multilabel Plate Reader, PerkinElmer) s in a transparent 96-well plate. We used both fluorescein ([62]), QY=0.92 in 0.01 M NaOH, Abs/Em=494/521 nm, and Alexa 488 QY=0.92 in PBS, Abs/Em=495/519 nm [16] and [5]).

Extinction coefficient approach As an orthogonal method of estimation, we determined the extinction coefficient of Dendra2, since it has been reported for GFP that the extinction coefficient is only obtained with properly folded and fluorescent proteins [63]. The same buffer as for the quantum yield determination was used, and the protein was diluted around a concentration of 10 µM, according to [61].

4.2 Data analysis

Once collected and processed to extract the detection coordinates, the data were analyzed as follows and implementations are presented.

4.2.1 Mathematical background

Superresolution microscopy produces a set of 2D or 3D coordinates located inside a region of $\mathbb{R}^2$ or $\mathbb{R}^3$. This region is assumed to be of sufficient regularity. An incertainty $\mathcal{I} = (\sigma_x, \sigma_y, \sigma_z)$ is associated to each point. Here we assume that all the points share the same incertainties.

Mathematically, such a datatype can be described in the formalism of spatial point patterns [[[64, 65]. We present here the few tools we used to handle the data.

Spatial point pattern Although one can define a spatial point pattern in several way, we give the following definition. Given:

- a set $\mathcal{B}$, here $\mathbb{R}^2$ or $\mathbb{R}^3$
- N a random variable (describing the number of points in the pattern),
- N spatial random variables $X_i, i \in [1, N], \Omega(X_i) \subset \mathcal{B}$,

a spatial point pattern PP can be unformally defined as the collection of the X_i, that said $PP = \{X_i, i \in [1, N]\}$. The points x of PP (that said, where a point is drawn out of the $X_i : [X = x], x \in \mathcal{B}$) are called *events* so that they can be distinguished from the the elements of $\mathcal{B}$.

Given a spatial point pattern PP, one can consider the following cases:

- PP is *finite* (each subset of the space $\mathcal{B}$ contains a finite number of points)
- PP is *simple* (two events cannot occur at the same point)

We will only consider the case where those two properties are satified. Furthermore:

- PP is *homogeneous* iff the distribution of the X_i do not change respective to translation, i.e $g_{X_i} = g_{X_{i,+x}}$, where g_a denotes the distribution probability of the random variable \$.\$, and \$+xa is notation for the random variable defined by $P[X_{i,+x} = k] = P[X_i = x + k]$
- PP is *isotropic* iff the distribution of the X_i do not change respective to rotation, i.e $g_{X_i} = g_{X_{i,rot}}$, where *rot* stands for a rotation of the random variable.

From these definitions, it arises that a PP that is bound to a finite region of space (that is $N(A) = 0, \forall A \subset \mathcal{B} \setminus D$, D of finite measure) cannot be homogeneous. In our case any PP considered will have $N(A) = 0$ (no event) when A is out of the nucleus. On the other hand, we will assume the PP isotropic (following **récamier**), an assumption required for the computation of border corrections.

A last definition is the one of a *marked* point pattern, where the X_i have values in $\mathcal{B} \times \mathcal{M}$ and $\mathcal{M}$ can be a countable or uncountable set accounting for additional properties of the events. For instance, the description of a 3D, two color PALM/STORM

[5]. Life Technologies, sheet for Alexa Fluor dyes.

experiment involving PolII and c-Myccan be realized as the following:

- N events (detection of molecules) located inside the nucleus $\mathcal{N} \subset \mathbb{R}^3$.
- Each event X_i takes values in $\mathcal{N} \times \{$c-Myc, $-$PolII$\}$, and should be understood as: "there is a molecule of c-Myc/PolII at a given position of the nucleus.

We can then define several canonical point patterns and some associated properties and functions.

First order statistic: intensity The intensity $\lambda(x)$ of a point of space x extends the notion of the mean of a traditional random variable. It describes the expected number of events that lie at a given position of space. If we describe $\mathbb{E}(N(B(x,r)))$ the expected number of points inside a ball of center x and radius r. $\lambda(x)$ is then defined as the expected number of points in x, that is $\lambda(x) = \lim_{r\to 0} \frac{N(B(x,r))}{l(B(x,r))}$, where $l(.)$ denotes the measure of the ball (its volume or surface in our case). Note that this quantity is defined everywhere in the case of finite PP.

Practically, this definition is useless, since we observe a realization of the underlying random variable: for an empirical PP, the $\lambda(x)$ converges to a sum of Diracs (corresponding to the position of the observed events). This issue is overcomes by applying a smoothing to the data to obtain a continuous density. The estimated intensity can then be expressed as a convolution: $\lambda(x_0) = \int_{\mathcal{B}} \sum_{i=1}^{N(B)} \mathbf{1}[X_i = x] I(x - x_0)dx$, where I gives the weight of a point given its distance \$x-x$_0$\$to one event

It should be noted that the intensity is a local parameter, the intensity depends on the position x of space.

Second order statistic: K-Ripley function Similarly to the first order statistic, second order statistic extend the notion of variance to a PP.

One of the main second order statistics is the K-Ripley $K(r)$ function. This function denotes the average number of neighbours located within a radius r of an event. Although very simple, this function can provide many indications:

Enrichment (as compared to its ditribution under a reference PP) at low radii characterize clusturing whereas enrichment at high radii denotes exclusion. Furthermore, this distribution allows to determine fine thresholds and behaviours, such as: strong clusturing at a low radius followed by exclusion at higher radii.

Cluster sizes the normalized K-Ripley function can give insights into cluster sizes: the peaks of the function lie between the radius and the diameter of the clusters.

Fractal dimensions see below.

Many other functions exist and can be reviewed in [65] and [64]. All of them compute pairwise "relations" between points.

Fractal dimension Many structures in the nucleus exhibit self-similarity (for instance, DNA forms loops, coils, supercoils, domains, etc.), an organization that can usefully be described using fractal mathematics [5].

An object can be described as fractal when his physical properties lie in between the traditional topolgical dimensions. For instance, a highly tortuous curve (such as DNA) can exhibit space-filling properties, although it has a mathematically null volume. These objects are characterized by their fractal dimension, that especially accounts for this space-filling property.

Such space-filling behaviour has many consequences in terms of diffusion, reactivity and kinetics, and has been widely investigated in various fields of chemistry and physics.

Mathematically, the fractal dimension can be defined as the scaling of the volume/surface of an object respective to its radius. For instance, surf $\propto$

r^2 for 2D objects. For fractal objects, this dimension is not an integer. More rigorously, the most accepted definition of a fractal dimension is Hausdorff fractal dimension, that derives from Hausdorff's measure. This definition, however, cannot be used in practice, as it requires to consider all the possible paving of the volume with spheres.

A practical way of measuring fractal dimensions is to compute the boxcounting dimension (also called the Minkowski-Bouligand dimension), that is equivalement to the Hausdorff fractal dimension. The boxcounting dimension describes the scaling between the number of "cubic boxes" required to encompass the measured object respective to the size of the box. For instance, paving a square of edge $n * \epsilon$ requires $N = n^2$ cubic boxes of edge ϵ. For a cube, it is n^3. Thus, the boxcounting dimension is given by $\lim_{\epsilon \to 0} \frac{\log(N(\epsilon)}{\log(1/\epsilon)}$.

Although tractable, this method is very computationnally intensive in the cases where some edge corrections have to be performed . Fortunately, other approaches to compute fractal dimensions are available and more tractable. One of them, termed the correlation fractal dimension, derives from the K-Ripley function. Indeed, the scaling of $K(r)$ with respect to r denotes the way the PP fills space. If $K(r)$ grows like r^2, it is more or less enclosed in a plane (more generally, in a 2D manifold), if it scales as r^3, it uniformly fills space. Thus, the scaling of the K-Ripley function, or correlation fractal dimension can be considered as an approximation of the fractal dimension. Mathematically, the correlation fractal dimension exhibits similar properties and converges toward the Hausdorff dimension.

Edge correction Empirical PP are always limited to an observation window. In our experiments, the best case scenario is the limit of the cell. The reference point patterns, however, often have unlimited spatial extension (se below). Thus, limits of the observation window introduce a bias when trying to estimate either intensity or second order statistics, because the events lying outside the observation window will never be observed. For instance, if we compare the K-Ripley of an homogeneous point distribution (termed later as a Poisson process) in a borderless versus a case with border, $K(r)_{\text{no border}}$ shows a monotonic growth whereas $K(r)_{\text{border}}$ plateaus as r reaches the border. Subsequently, one shoudltake into account the limits of the observation window, and correct for the detections that will never be observed.

Several corrections can be performed:

An erosion of the zone of study one introduce a "buffer zone", where points are considered as "arrival points" and not as "starting points" in the computation of the K-Ripley for instance. This solution, however, gets more and more suboptimal as the dimensionality of the space increases, a phenomenon usually referred as the "curse of dimensionality". In our case, eroding the nucleus would force us to discard a lot of observations. Furthermore, this method is inaccurate when the intensity of the PP is a function of the distance to the border, as it has been recently described for H2B.

We focus here on the isotropic correction: under the assumption of isotropy the events are evenly distributed around each event. In that case, the number of events located on a circle of given radius and centered around an event that are missed "missed" because of the border is proportional to the fraction of circle outside the border. This correction has been successfully applied to superresolution in [12]. Other corrections have been previously developes, such as a homogeneous correction [64].

Canonical point patterns We present here some traditional point patterns. These were used as reference in the analysis we performed later. Demonstrations and further details can be found in [65].

Binomial A binomial point process represents the uniform distribution of a given number of points inside a closed set W. For a binomial point process showing n_{tot} events, the number of points for a given set A located inside W is given by $\mathbb{E}(N(A)) = n_{tot}\frac{\ell(A)}{\ell(W,)}$ where ℓ designs the area/volume of the considered region.

Poisson In a Poisson point process, the expected number of points in a given region follows a Poisson distribution of parameter $\lambda\ell(A)$. This is equivalent to a uniform distribition over $\mathbb{B}$.

The intensity, as described upper, does not have to be constant, and one can also describe a non-homogeneous Poisson process, where the expected number of detections is obtained by integrating the local intensities $\lambda(x)$.

In the case of a homogeneous Poisson process without borders, the K-Ripley distribution can be derived analytically: $K(r) = \lambda\pi r^2$ in 2D, $K(r) = \lambda\pi r^3$ in 3D.

Note that when conditioning a Poisson process by the total number of detections, one find back a binomial process. The Poisson process is used a reference, null model of uniformly ditributed events.

Poisson cluster Finally, we introduce a last model for the purpose of the following analysis. A Poisson cluster process can virually be divided in three steps. First, the location of *parent* clusters is determined according to a Poisson process. Then, each parent cluster gives rise to a random number of *children*, that are then distributed according a second probability law. A common Poisson cluster process, represented *fig is a homogeneous Poisson process where the *children* are normally distributed around the cluster center.

4.2.2 An implementation: Rpalm3d

All the point pattern theory described in the previous section finds an implementation in the excellent package *spatstat* [66, 67] for `R` [68]. This package presents however very limited 3D capabilities. For instance, the window of observation is limited to parallepiped, and the K-Ripley functions are not usable because they are too slow (on a $\sim$ 100000 points PP, a 3D K-Ripley was stop after 24 hrs of computation without completing). A full 3D version of *spatstat*, called *spatstat 2* has been announced several years ago [6], but was never released. Furthermore, we contacted Rolf Turner, one of the former developers of *spatstat*, who confirmed that no release was planned in the close future. We thus implemented a few functions to analyze the point patterns and to easily process PALM data. These functions are packaged in a `R` library temporarily called `Rpalm3d` [7] (Figure 17).

The code is mostly `R`, and the critical parts were written in `C` for performance reasons.

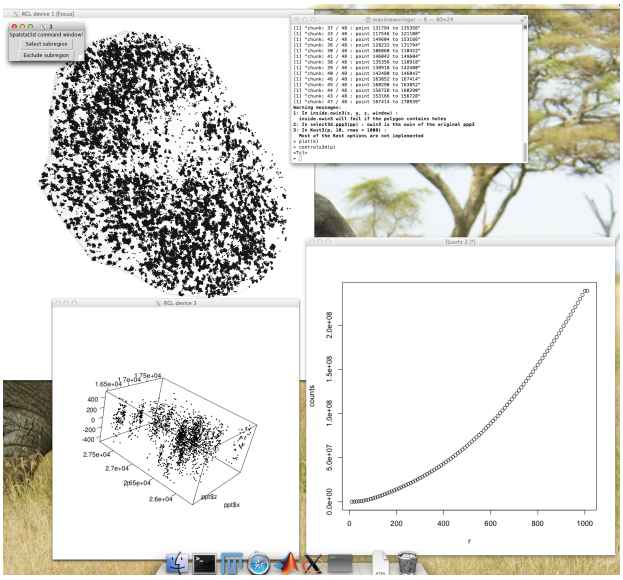


Figure 17: A screenshot showing a few features of `Rpalm3d`: 3D visualization and subregion selection, K-Ripley plot and the console to type commands.

Interface We first determined a way to describe the border of the nucleus. We use a polyhedron that is built from the union of triangles. Although the implementation is general, we only used polyhedrons generated by extruding a 2D polyhedron.

[6]. Spatstat website, page "Spatstat Version 2, Preliminary Announcement", dated: 6th revision, 9th February 2010.

[7]. This package is freely (as in speech) available upon request at maxime.woringer@ens.fr.

This description suggests/implies a geometrical point of view on the problem. We thus implemented several algorithm originating or inspired from computational geometry. On the one hand, computational geometry algorithms are tailored for speed and parallelization. On the other hand, they often include hard-to-tackle singular cases and are often prone to numerical instabilities. There is thus a trade-off of this paradigm versus the previously described algorithms that were matrix-based.

The whole package is designed as an extension of *spatstat*. It uses *rgl* to plot the 3D representations. In addition to the few analytical tests available for 3D spatial point patterns, we implemented a few simulation methods to produce either binomial or Poisson point patterns which intensity $\lambda(x, y, z))$ can be split as $\lambda(x, y, z) = f_1(x) * f_2(y) * f_3(z)$. This method in particular allows for simulation of homogeneous point patterns ($\lambda = cst$) and for realistic (i.e.: empirical) z detection profile for PALM/STORM experiments, provided the three empirical distributions f_1, f_2 and f_3. These point patterns were simulated according to the rejection sampling method. In his simple form, points are drawn uniformly across the domain of interest, and each point is either kept or discarded with a probability derived from the input distirbution.

Local density The maximum likelihood estimator of the intensity of a homogeneous Poisson process is the count of events per volume unit. However, even the computation of the volume of a 3D polyhedron is still a debated subject. In our case, we implemented a very classic algorithm where the polyhedron is split into pyramids that all share one common vertex. In that configuration, the final volume is the oriented sum of the volumes of pyramids.

K-Ripley We optimized the computation of K-Ripleys by splitting the nuclear volume in cubes. Indeed to compute $K(r), r \in [0, r_{max}]$, the nucleus is split into non overlapping cubes of edge r_{max}. That way, all the relevant neighbours for the K-Ripley are located within the 26 cubes adjacent to the one considered. When computing K-Ripley on 3D PALM/STORM pictures with r_{max}=1µm, the slice is about 1.2x15x10=180 µm^3. The theoretical speed improvement (neglecting preprocessing overhead) is 10 fold.

Border correction Since previously published literature mostly considers isotropic correction, we focused on that one. In 3D, border correction requires to compute the surface of the sphere inside the polyhedron. Previous implementations either relied on heavy numerical integration and convolutions [12]. We proposed last year a fully FFT-based implementation that could potentially accelerate computations. However, the algorithm was very RAM-demanding, CPU- and GPU-demanding and its accuracy couldn't be tested. Furthermore, the previous algorithm was also time consuming, as it took $\sim$ 3h30 to compute a single K-Ripley.

Here we propose a geometrical, approximated computation of the Ripley function. Indeed, the determination of the fraction of the sphere outside the nucleus is estimated by uniformly sampling a sphere of given radius. Provided that the number of tested points is high enough, the fraction of points outside the border (out of the nucleus) converges towards the fraction out of the nucleus.

Concretely, we implemented a fast, approximate point-in-polyhedron test derived from a 2D version. An approximate algorithm is acceptable because the high number of tested points (more than 100 points per sphere, a figure that should be multiplied by the number of points). In that case 1-2% of uncapture are released and unhandled.

4.3 Supplementary results

4.3.1 Accuracy of `Rpalm3d` for first and second-order estimations

We first characterized the correctness of our implementations on previously studied protein. The

histone 2B (H2B) 3D-PALM structure has recently been resolved and characterized using the same methods as we did [12]. Although our firsts tests on unpublished H2B PALM datasets (provided by Ignacio Izeddin, Laboratoire d'Imagerie Fonctionnelle de la Transcription, Paris) were consistent with the published data, we were unable to obtain coherent results when we moved to the exact dataset used in the paper (PALM data provided by Vincent Récamier and Xavier Darzacq). Finally, more careful tests on small synthetic datasets that can be analyzed with the limited 3D capabilities of *spatstat* revealed unacceptably wrong behaviour. Debugging he code would have required too much time to be achieved during the course of this internship. Thus, we tried to perform the analysis in 2D using *spatstat*. This analysis is still in progress

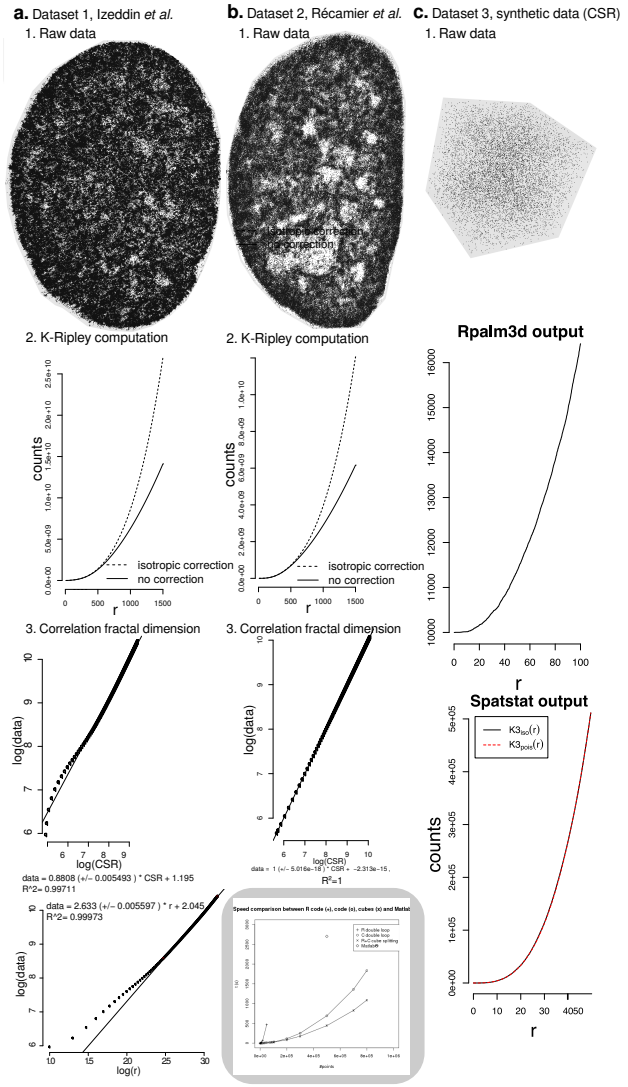


Figure 18: Comparison of `Rpalm3d` on three datasets. *(a).* Consistent estimate of a fractal dimension of 2.6 using a H2B dataset. *(b).* Inconsistent estimate of a fractal dimension of 3. *(c).* Inaccuracies on synthetic data. The sacling and the overall shape of the curve is wrong. *(d).* comarison of the running speed of the various implementations.

4.3.2 Detection of c-Myc clusters

One the first order, radial accumulation of c-Myc has been quantified, it becomes possible to look at finer structure within the pattern. One important question is whether c-Myc shows clusturing, as it has been described for PolII [7] and Sox2, another Yamanaka factor [69]). Identifying clusters is not trivial within STORM datasets is not trivial, because the intrinsic photophysics of the employed dyes causes the fluorophores to be detected multiple times, thus showing clusters corresponding to multiple detections.

However, precise quantification of the pointing accuracy can help to distinguish between biologically-relevant clusters and multiple detections.

As described in the methods **??**, we performed an estimate of the pointing accuracy in x, y and z on one of the acquired dataset. This estimate allowed us to investigate c-Myc clusturing by providing a more realistic model than the traditional complete spatial randomness (CSR, a homogeneous Poisson process 4.2.1). As the time of finalizing the report, we are still in the process of running the simulations.

4.3.3 Relation of c-Myc respective to E-boxes

In the previous paragraph, we assumed that the E-boxes were uniformly spread along the genome. This is obviously an approximation, and we tried to infer a more precise spatial distribution. We first recorded the positions of each E-box in the hg38, and plotted their distribution across the genome. We didn't have time to perform any statistical analysis (comparison with the anlytical derivation of the variance of the exponential law), but one can see striking heterogeneities in the `CACGTG` pat-

tern, with some regions showing a > 5 folds enrichment. Furthermore, we computed the distribution of distances between two successive detections. In the random case (no sequence bias), the distances between consecutive events follow an exponential distribution. In our case, two main points can be noted:

- Both the overall and the per-chromosome distance distributions show a significant deviation from an exponential law, as assayed by the log(count) *vs.* successive distance.
- If some chromosomic regions seem enriched, some chromosomes show significant enrichment. We focused on per-chromosome average because the radial position of chromosome has been documented for several cell types [54]. For instance, chromosome 18 is gene poor and is known to localize at the periphery of the nucleus. Conversely, the gene-rich chromosome 19 has approximately the same size, and is mostly located in the center of the nucleus (initially assayed by "chromosome painting").

To determine whether chromosome-wide E-box enrichment bias can account for the c-Myc distribution, we compared the per-chromosom E-box enrichment (as the average number of E-box per base pair) and the rank of the chromosome respective to its distance to the center of the cell. We particularly focus on chromosomes 18 and 19. Unsurprisingly, no pattern emerges, indicative of either a lack of correlation or a lack of resolution.

4.3.4 A collection of c-Myc mutants to functionally investigate its inhomogeneous nuclear localization.

As an orthogonal approach to determine the factors driving c-Myc nuclear patterns, we generated truncated version of the tagged c-Myc construct, allowing superresolution imaging of theses binding deficient proteins.

As a first, harsh screen, we generated three first constructs, where we deleted the entire interaction domains as follows: cMyc-ΔMBII-Dendra2 (deletion of the TRRAP interaction domain MBII), cMyc-Δbasic-Halo (DNA binding domain deleted), cMyc-ΔHLH-Dendra2 and cMyc-ΔHLH-Halo (HLH Max-interacting domain deleted).

We then tried to assay the intranuclear localisation of tagged proteins after transfection by confocal and superresolution microscopy. Unfortunately, we ran into several transfection problems that prevented us from getting any data from these constructs. We are still in the process of imaging these mutants.

4.4 Dynamics of c-Myc

Since the static confocal pictures and the superresolution reconstructions were providing more questions than answer, especially regarding the previously described colocalization with histones 3 trimethyl K4, we wondered whether this difference could be accounted for by different dynamics. Indeed, one can envision a mechanism by which c-Myc would spend most of its time at the borders of the nucleus, a result of a weak interaction for a very high number of sites (for instance the E-boxes, leading to a ChIP-seq signal very close to the noise level) combined to specific, long term interactions with specific partners, for instance a recruitement to modified histones by the GCN5/TRRAP complex, leading to sharp peaks on a more limited number of targets.

To test this hypothesis, we performed a series of live cell experiments. The main questions were to first characterize c-Myc diffusive behaviour in NT2 cells, mostly following a methodology previously described [38, 20].

Since we acquired the data very lately, and that the dataset quality is not outstanding, we do not derive any figure from the plots, but simply show the quality of the plots that can be obtained.

4.4.1 Fluorescence recovery after photobleaching (FRAP)

In FRAP experiments, a focused laser bleaches one specific region of the cytoplasm. The fluorescence at the bleach spot then progressively comes back to the nucleoplasmic level as new fluorescent proteins diffuse and replace the bleached proteins at their (more or less specific) binding sites. That way, and under conditions where diffusion is not the factor limiting fluorescencen recovery, FRAP allows to compute a k_{off} of the protein for its binding site. Moreover, slightly more complex models can be derived to account for several types of binding, for instance to describe a population exhibiting both specific and aspecific binding (and thus a difference in residence times at the binding sites) This measure can give insights on the exploration patterns of c-Myc at a the full nuclear scale and therefore be used to compare various c-Myc mutants. Indeed, how trnascription factors explore space is a critical parameter influencing the way the protein can act as a regulator. Space-exploring patterns have been characterized as diverse as highly confined, performing a strongly recurrent walk around a given region, thus creating an increase in the local concentration. On the other hand, some transcription factors have been described to explore nuclear space in a much less recurrent manner, with a probability less than one to come back to its starting point.

We performed preliminary FRAP experiments in unperturbed cell transfected with c-Myc-GFP in order to compare the bound fraction apparent diffusion coefficients of several mutants.

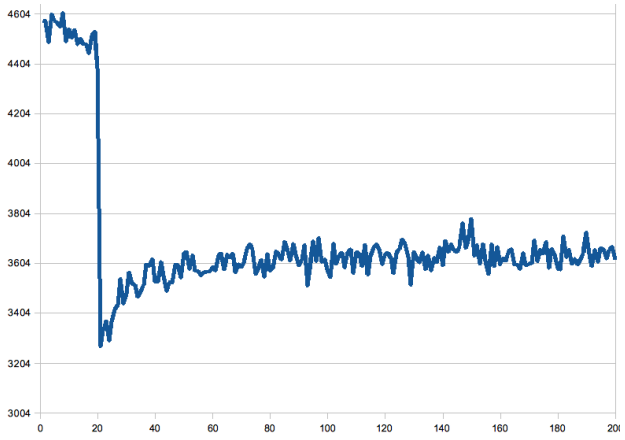


Figure 19: preliminary c-Myc FRAP data showing high immobile fraction. x: time in ms, y: fluorescence (u.a.)

4.4.2 Identifying the distribution of transient high confinement of c-Myc using single particle tracking (SPT).

A counterpart to the ensemble FRAP measurements is single particle tracking, where individual molecule traces are recorded, allowing to build single-molecule mean square displacement (MSD) and extract crucial parameters governing the exploration pattern of the protein, such as the compacity of its movement [70, 5]. However, MSD determinations are limited by the undersampling of high diffusion coefficient (they are more likely to go out of focus than slow-moving molecules) that tend to introduce artifacts that resemble anomalous diffusion. To circumvent this issue, other methods for data analysis have been proposed, including the computation of angles between two successive translocations (Burov, Izeddin, Woringer). Since we are mainly interested in the gradient respective to the distance to the nuclear periphery, we focused on quantifying the number of very confined traces in different portions of the nucleus. We are still in the process of analyzing the acquired data.

A first look at the obtained data show that they reproduce previously observed dynamics [38].

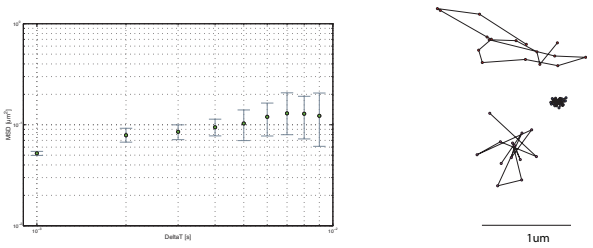


Figure 20: Preliminary single particle data showing a preliminary MSD and a sample of the source traces.

4.4.3 Engineering of a cell line expressing tagged versions of c-Myc using the CRISPR system

An engineered cell line where one or both of the c-Myc alleles are tag would be a great tool to further investigate its cellular functions. Indeed, an endogeneously expressed gene releases the constraints and the stress induced by transfections, allows to monitor protein localization over extended periods of time (such as differentiation). Furthermore, expression at an endogeneous level allows for fine probing of cellular transcription regulation. Finally, it would make it easier to monitor several proteins simultaneously (such as c-Myc interacting partners TRRAP or GCN5). However, our first attempt failed and no cell survived selection. We then tried the following conditions, detailed in Table 3:

Several tranfections gave puromycin-resistant colonies. However, we were unable to detect fluorescent colonies under a confocal microscope. We then moved to FACS to enrich the fluorescent-positive cells. FACS was able to sort a fluorescent-positive population, and we isolated ~2500 Dendra2-positive cells (Figure 21b). This population (either Dendra2-positive of Halo-positive) was very distinct from the control cells and represented a low amount of 0.5-1% of the cells. Unfortunately, none of the $\sim$ 2500 cells sorted during our first attempt.

We performed two other sorts, one with the Dendra2-transfected cells again (Figure 21c), and one with the Halo-transfected cells (Figure 21c). That time, we also sorted the Dendra2-negative cells as a control. The negatively sorted cells also died, suggesting unoptimized sorting conditions. We didn't try to optimize further the FACS conditions due to time constraints.

1% of fluorescent-positive cells is very low compared to the integration rates usually observed in other cells. Our last hypothesis is a very low transfection efficiency of the linear plasmid. We are thus trying to co-transfect to additional gRNA targetted against the transfected donor vector to linearize it *in-vivo*, in order to overcome a possibly low transfection efficiency of linear DNA.

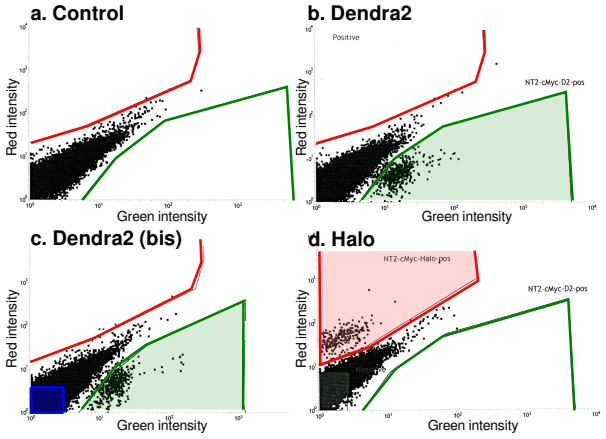


Figure 21: Result of the FACS sorting, showing a distinct population of either GFP positive or Halo-positive cells.

4.5 Supplementary discussion

4.5.1 An alternative model to explain c-Myc gradient

Our current model to explain c-Myc accumulation at the periphery of the membrane require a dense underlying structure at least partially tethering the protein. Since the accumulation spans a few microns inside the nucleus, the nuclear envelope cannot be this substrate. However, the constraint for a substrate can be waived (at least theoretically) without the need of overwhelmingly strong assumptions. We present here a model where a protein freely diffuses in the nucleoplasm. When hitting the nuclear envelope, the protein is trapped with a given probability, and released with another probability. This simple model is sufficient to create a gradient from the nuclear envelope towards the center of the nucleus.

We now formalize this model as discrete and in one dimension where the 0 on the x axis locates the nuclear envelope, and positive values point to the center of the nucleus. We proceed to the following simplifications (although they can be easily removed without changing the analytical paradigm):

Table 3: Different conditions assessed to generate the cell line, KpnI: donor vector linearized using KpnI, gradient: puromycin started at 0.25µg/ml and then gradually increased until 1 µg/ml.

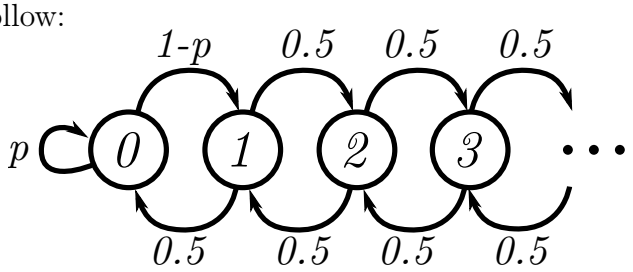
Construct	Linear vector	Selection	Transfection	Outcome	Diagnosis
Halo	yes	1 µg/ml	Nucleo.	no colonies	Selection too strong
Halo	no	gradient	XFect	no colonies	Low transfection efficiency
Halo	no	gradient	Nucleo	cell death	Contamination? (added P/S
Dendra2	no	gradient	Nucleo.	1. colonies, killed by FACS 2. backup killed by FACS	FACS too harsh, FACS too harsh
Halo	yes	gradient	Nucleo	1. colonies, killed by FACS 2. backup, from 1.	FACS too harsh Contamination
Dendra2-MS2	yes	gradient	Nucleo	**still under selection**	

- The nucleus radius is infinitely large compared to the slope of the gradient. Thus, the distribution of the protein should be an ever-decreasing function of the distance to 0.
- The nucleus is discretized at a given step.
- We do not consider the temporal aspect of the binding/unbinding to the nuclear envelope.
- Space is assumed euclidian (thus diffusion is Brownian). It is possible that this model holds also in the case of more complex geometries (such as fractal).
- The protein moves at each time point either one step left or right. At the membrane, it only can move right or stay at zero.

We note: X the discrete random variable describing the position of the particle at one given time, and dX the movement of the particle between two time points -1, 0 or 1 in our case. Thus, our model can be formalized as follow:

$$\begin{cases} P(dX = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} | x = 0) = \begin{pmatrix} 0 \\ p \\ 1-p \end{pmatrix} \\ P(dX = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} | x \neq 0) = \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \end{pmatrix} \end{cases}$$

A much more tractable form can be derived, and this model can be formalized and analyzed as a Markov chain to derive the equilibrium distribution of the molecules. This simple model could easily be integrated into a temporal framework by switching to the so-called continuous time Markov chain (where the newly introduced time between two events of the initial chain is exponentially distributed). The n-states Markov chain is defined as follow:



And we denote $N_T =^t (n_{1,T}, n_{2,T}, n_{3,T}, \dots)$ the number of molecules in each state at time T.

Once formalized as a Markov chain (formally: a graph), the adjacency matrix P of the graph can immediately derived:

$$P = \begin{pmatrix} p & 1-p & 0 & \dots & \\ 1/2 & 0 & 1/2 & 0 & \\ 0 & 1/2 & 0 & 1/2 & 0 \\ \vdots & & & \ddots & \end{pmatrix}$$

From this matrix point of view, a population of molecules distributed as N_T will evolve following the linear relation $N_{T+1} = PN_T$. From this perspective, the equilibrium distribution N_{eq} is reached when $N_{T+1} = PN_T$. That said, the right eigenvector of P is the stable distribution of this chain. Neither $N = (1, 1, 1, \dots)$ nor $N = (1, 0, 0, \dots)$ are eigenvectors of P, implying the existence of a gradient along the x axis. Further computations are required to characterize the distribution. However, this model is extremely simple, and

does not account for a limited number of binding sites at the nuclear envelope, or for anomalous diffusion when unbound.