



**HAL**  
open science

# Cartographie pangénomique à haut débit et en molécule unique de la réplication de l'ADN

Nikita Menezes Braganca

► **To cite this version:**

Nikita Menezes Braganca. Cartographie pangénomique à haut débit et en molécule unique de la réplication de l'ADN. Génomique, Transcriptomique et Protéomique [q-bio.GN]. Université Paris sciences et lettres, 2019. Français. NNT : 2019PSLEE040 . tel-02954741

**HAL Id: tel-02954741**

**<https://theses.hal.science/tel-02954741>**

Submitted on 1 Oct 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**THÈSE DE DOCTORAT**  
**DE L'UNIVERSITÉ PSL**

Préparée à l'École Normale Supérieure

**Cartographie pangénomique à haut débit et en molécule  
unique de la réplication de l'ADN**

Soutenue par

**Nikita MENEZES  
BRAGANCA**

Le 14 Octobre 2019

École doctorale n°

**École Doctorale  
Complexité du vivant  
ED515**

Spécialité

**Génomique**



**ENS**

Composition du jury :

M. Olivier HYRIEN École Normale Supérieure (IBENS)	<i>Directeur de thèse</i>
M. Auguste GENOVESIO École Normale Supérieure (IBENS)	<i>Co-directeur de thèse</i>
Mme Anne DANCKAERT Institut Pasteur	<i>Rapporteuse</i>
M. Benoit MIOTTO Institut Cochin	<i>Rapporteur</i>
Mme Kathrin MARHEINEKE I2BC, Paris Saclay	<i>Examinatrice</i>
M. Olivier ESPELI Collège de France	<i>Président du jury</i>
M. Florian MULLER Institut Pasteur	<i>Invité</i>



## Remerciements

*I will write this section in english since I was told that it is the most important one of the thesis and thus the one that most people will read.*

*How can I summarise these three years of PhD? When I started my PhD in 2016, it was more like “May the force be with you...” (Star Wars, 1977). Then, when I got stuck, like any other PhD student, it was more “After all, tomorrow is another day!” (Gone with the wind, 1939) thinking that things could only get better. Finally, on the last year “Houston, we have a problem!” (Apollo 13, 1995) would describe the last months that went by so quickly. But overall, “I consider myself as the luckiest (wo)man on the face of the earth” (The Pride of the Yankees, 1942) for tons of reasons.*

First I would like to thank Anne Danckaert, Olivier Espeli, Kathrin Marheineke, Benoit Miotto and Florian Mueller who accepted to be part of my jury. Thank you also to all the members of my thesis committee who were very supportive and encouraging during these 3 years : Ignacio Izzedine, Florian Muller and Alexandra Louis.

I would like to thank Olivier Hyrien, Master 2 and PhD supervisor, who supported me and also gave me the chance to work with him and his team. Thank you for being available anytime for discussing science!

*“Some people can’t believe in themselves until someone else believes in them first.”* (Good Will Hunting, 1997). I think Auguste Genovesio played that role when giving me the opportunity to join his team for my Master 1 internship. Thank you Auguste for believing in me more than I did. I was lucky to work at the interface between biology and computational biology and have both perspectives to move this project forward. While, being between two teams is not always easy, I learnt a lot from both. So thank you Olivier and Auguste for every scientific and non scientific discussion, lab meetings, correction of abstracts (even when I sent it last minute),

---

bottle of champagne/wine opening to celebrate publications/grants ..., lab diners, letting me take part in YRLS conference for 3 years, SPIBENS, allowing me to popularize science and many more things.

Thank you to all former and current members of Olivier's lab for your help, support and interesting discussions (Benoit, Magalie, Elizaveta, Laurent, Florence, Xia, Francesco, Bertrand and Etienne). Special thanks to Benoit and Magalie for pushing me, cheering me up during these 3 years and helping me during the preparation of my master thesis and my concours (yes, I do remember). Thank you Francesco (and Olivier) for coming up with this project and giving me the possibility to work with you. It was not always easy but we did it and published HOMARD (credits to Olivier for the name, which translates to lobster and facilitated illustrating the tool!). Thank you also to all our collaborators in Paris, Lyon and Bordeaux who contributed to the scientific discussion.

Of course thank you to the *CBB superteam* (our Whatsapp group). These 3 years would not have been the same without you all. To be fair, I will thank you by alphabetical order : Alice, Amira (will miss our end of the lab meeting hug), Benoît (“Petit Benoît” *pour les intimes*), Charles, Elise (my super plant care-taker), Felipe (thank you for your smile and for passing on your love for design), Elton (thank you for letting me annoy you for hours when I had questions), France (good luck for your defense on Friday), Guillaume (Dédé or Kiwi *pour les intimes*, you will hate me because everyone knows now!), Leila (thank you for your good vibes), Mathieu (Matéooo, I will miss your GIFs and your post-it and I am sure you are going to miss me because no one will be there to save your life as much as I did!), Maxime (Patator, the big boss of the babyfoot), Ouardia (the best sweet treat cook of the team), Raphaël, Sreetama, Shihav, Solène (still remembering your first days in the team and how you shouted your name when we went for the training), Tiphaine (thank you for your craziness, good mood and being our 16 personalities expert).

I also take the opportunity to thank all the people from the Genomic Section (best section of all time! Sorry, I had to say it!) but also from the 6th (Benoit, Ana, Sophia, Ignacio, Phuong), the 7th floor (Aurélien, Kamal, Ayush, Alexia, Frank) and the 9th floor (Marco, Chloé, Romain, Benjamin, Teddy, Mélissa, Laura), who made these 3 years even better. Sorry not to give all the names but this acknowledgement page would be endless if I did. Thank you also to all the administration staff (Lina, Virginie, Béatrice, Pierre-Emmanuel), SPIBENS board members (Nora, Auriane,

---

Yves, Patrick, Joanne, and the ones who followed), YRLS Organizing Committee members for the 3 past years and specially the OC from the 10th edition (Myriam (my super VP and BFF), Lucie, Sophia, Rémi, Marco, Pauline, Alice, Satish, Tania, Sophie, Serena, Cécile, Aura ...). Thank you the Universcience team Emilien, Jérôme and Stéphane for sharing his knowledge on how to popularize science and also for being very understanding during the writing of my thesis.

There are 4 persons I would like to thank particularly and that made these 3 years an amazing experience. Toni (no I did not forget you when listing all CBB superteam members), thank you for being always smiling, happy, positive and my Karaoke buddy. Kasia, I am already missing our walks and small chats, but I know we will meet soon for Brunch, waffles or shopping! Thank you as well for your support and being always positive. Caroline (my thesis writing buddy in the last months!) and Isa (Wonderwoman, she always finds a solution), thank you for your smile, being always in a good mood, encouraging and supportive.

I might have missed some names. It doesn't mean that I have forgotten you'll, it's just that I am already on my third page of acknowledgement ... So, I do apologize if your name is not listed.

*Je garde le meilleur pour la fin!* Mom, Dad and Sis', words are not enough to say how much I love you. I am thankful for having such a great family. These 3 years have been the most difficult ones and you have always been there when I needed it. You always supported and encouraged me, not only now, but my entire life. This thesis is the result of all the sacrifices you have done for me so far. What I became today is not thanks to me, but primarily thanks to you'll. For that I will be eternally grateful. I will stop here, I could write so many things but I have to keep this short. So, thank you Mom, Dad and Sis' for being there, always! *"Families are the compass that guides us. They are the inspiration to reach great heights, and our comfort when we occasionally falter."* (Brad Henry)

*I am leaving IBENS after having grown both personally and professionally but more importantly, I am leaving IBENS with wonderful memories and amazing friends!*



## Abstract

DNA replication is a vital process that ensures an accurate conveyance of the genetic information to the daughter cells. In eukaryotic organisms, genome replication is carried out by using multiple start sites, also known as replication origins. During phase S of the cell cycle, these origins are fired stochastically, resulting in bi-directional replication forks that propagate along the genome until the converging replication forks merge; this process is called termination. In metazoans, the mapping of replication remains challenging. Genome wide mapping of human replication origins performed using sequencing of Okazaki fragments (Ok-seq), initiation sites labelled with digoxigenin-dUTPs in a cell-free system (Ini-seq), isolated small nascent strands (SNS-seq) or replication bubbles (Bubble-seq), only modestly agree. This inconsistency can be due to the fact that these existing genome wide approaches use large cell populations that smooth out variability between chromosomal copies.

Thus, to get a better understanding of DNA replication and to uncover the cell-to-cell variability, the development of single molecule techniques is fundamental. DNA combing, a widespread technique used to map DNA replication at a single molecule level, has a very low throughput and tends to give faint signal in addition to uneven linearity of the DNA molecules. Thus, automated detection and mapping of the DNA fragments are arduous tasks. Therefore, single molecule techniques tend to be refractory to automation, forestalling genome-wide analysis.

To overcome these impediments, we repurposed an optical DNA mapping device based on microfluidics, the Bionano Genomics Irys system, for High-throughput Optical Mapping of Replicating DNA (HOMARD). Relying on the same labelling strategy as OMAR (Optical Mapping of DNA replication), our methodology labels fluorescently DNA replication tracks and nicking endonuclease sites (barcode) using two different fluorescent nucleotides (dUTP), in addition to a YOYO-1 intercalator DNA fiber staining. We typically collect, for a single run, over 34 000 images and



---

more than 63 000 Mbp of DNA. The advantages of such technology are the high quality of the images and the possibility to automatically align the DNA molecules on a reference genome by optical mapping. Our new open source tools, that required the adaptation of the provided proprietary software, empower us to map the intensity profiles extracted after having two essential preprocessing steps on the raw images. We can now simultaneously visualize the intensity profiles of all mapped DNA molecules, check the optical mapping performed and, in particular, see where the replication tracks are located genome-wide at a single molecule level.

We demonstrate the robustness of our approach by providing an ultra-high coverage (23,311 x) replication map of  $\lambda$  DNA in *Xenopus* egg extracts and the potential of the Irys system for DNA replication and other functional genomic studies apart from its standard use meaning genome assembly and structural variation analysis.

## Résumé

La réplication de l'ADN est un processus vital assurant une transmission fidèle de l'information génétique aux cellules filles. Dans les organismes eucaryotes, la réplication du génome s'effectue en utilisant plusieurs sites de d'initiation, également appelés origines de réplication. Pendant la phase S du cycle cellulaire, ces origines sont déclenchées stochastiquement et donnent naissance à des fourches de réplication bidirectionnelles qui se propagent le long du génome jusqu'à la fusion des fourches de réplication convergentes; ce processus passif s'appelle la terminaison. Chez les métazoaires, la cartographie de la réplication demeure difficile. Par exemple, la cartographie à l'échelle du génome des origines de réplication humaine réalisée en utilisant le séquençage des fragments d'Okazaki (Ok-seq), des sites d'initiation marqués à la digoxigénine-dUTP dans un système "sans cellule" (*cell-free*) (Ini-seq), des brins naissants (SNS-seq) ou des bulles de réplication (Bubble-seq), ne coïncident que modestement. Une explication possible de cette incohérence est que ces approches utilisent de grandes populations cellulaires ne donnant qu'une image moyenne de la réplication.

Ainsi, pour mieux comprendre la réplication de l'ADN et la variabilité intercellulaire, il est indispensable de développer des techniques en molécule unique, parmi lesquelles nous avons le peignage moléculaire. Cette technique répandue est utilisée pour cartographier la réplication de l'ADN mais son débit est très faible. Par ailleurs, cette approche tend à donner un faible rapport signal/bruit en plus d'obtenir des molécules d'ADN non linaires : l'automatisation de la détection et de la cartographie des fragments d'ADN devient alors une tâche ardue. Par conséquent, les techniques en molécule unique ont tendance à être réfractaire à l'automatisation, empêchant l'analyse du génome dans son ensemble.

Pour s'affranchir de ces limitations, nous avons réorienté un dispositif de cartographie optique de l'ADN basé sur la microfluidique, le système Irys conçu par

---

BionanoGenomics, pour le *High Throughput Optical Mapping of Replicating DNA* (HOMaRD). S'appuyant sur la même stratégie de marquage qu'OMAR (*Optical Mapping of DNA Replication*), notre méthodologie marque par fluorescence les segments répliqués de l'ADN et les sites d'endonucléation (code-barres) en utilisant deux nucléotides fluorescents différents, en plus d'un agent intercalant, le YOYO-1, permettant le marquage des fibres d'ADN. Ces fibres d'ADN sont véhiculés par électrophorèse dans les nanocanaux de la puce Irys où elles sont linéarisées et imagées automatiquement. Nous collectons typiquement, pour un seul passage, plus de 63 000 Mbp d'ADN et plus de 34 000 images réparties en "*scans*". Les avantages de cette technologie sont la qualité des images et la capacité à cartographier automatiquement les molécules d'ADN sur un génome de référence. Nos outils *open source*, qui ont nécessité l'adaptation des logiciels propriétaires, nous permettent de cartographier les profils d'intensité qui ont été extraits après deux étapes essentielles de pré-traitement des images brutes. Nous pouvons maintenant visualiser simultanément les profils d'intensité de toutes les molécules d'ADN cartographiées, vérifier la cartographie optique effectuée et, en particulier, voir où se situent les segments répliqués à l'échelle du génome en molécule unique.

Nous démontrons la robustesse de notre approche en fournissant une cartographie de la réplication avec une couverture ultra-élevée (23 311 x) de l'ADN du bactériophage  $\lambda$  répliqué dans les extraits d'œufs de Xénope et le potentiel du système Irys pour l'analyse de la réplication de l'ADN et pour d'autres études de génomiques fonctionnelles, en plus de son utilisation standard, à savoir l'assemblage génomique et l'analyse des variants structuraux.

## Liste des abréviations

ACS	ARS Consensus Sequence
ADN	Acide DésoxyriboNucléique
ARN	Acide Ribonucléique
ARS	Autonomous Replicating Sequence
BrdU	Bromodeoxyuridine
CDK	Cycline-dependent kinase
CldU	Chlorodéoxyuridine
CpG	Cytosine–phosphate–Guanine
CTR	Constant Timing Regions
DDK	Dbf4-Dependent Kinase
DHFR	Dihydrofolate reductase
DLS	Direct Label and Stain
dUTP	Deoxyuridine triphosphate
FISH	Fluorescence in situ hybridization
FOV	Field-Of-View
HCFS	Human Cell Free System
HOMaRD	High-Throughput Optical Mapping of Replicating DNA
IdU	Iododéoxyuridin
MCM	minichromosome maintenance protein complex
NGS	Next Generation Sequencing
OMAR	Optical Mapping of Replicating DNA
ORC	Origin Recognition Complexe
ORI	ORIGine de réplication
RFD	Replication Fork Directionality
TTR	Transition Timing Tegions
TIFF	Tag(ged) Image File Format



# Sommaire

<b>1</b>	<b>Introduction</b>	<b>17</b>
1.1	<b>La réplication de l'ADN</b>	20
1.1.1	Lever les zones d'ombres sur le programme de réplication de l'ADN : un challenge	20
1.1.2	Les résultats des études utilisant des méthodes en population ne s'accordent pas	21
1.1.3	Vers une analyse en molécule unique ...	22
1.1.4	... à très haut débit	23
1.2	<b>Problématique et objectif.</b>	24
1.3	<b>Plan.</b>	25
<b>2</b>	<b>De l'autoradiographie à l'imagerie haut débit de molécules d'ADN en cours de réplication</b>	<b>27</b>
2.1	<b>Les premières études sur la réplication de l'ADN</b>	30
2.1.1	L'autoradiographie des fibres d'ADN	30
2.1.2	Le modèle du réplicon	32
2.1.2.1	Le modèle du réplicon chez les procaryotes	32
2.1.2.2	La mise en évidence des initiateurs et réplicateurs chez les eucaryotes	33
2.1.3	Des origines de réplication aux efficacités variables	35
2.1.4	Le " <i>random-completion problem</i> " et le choix des origines de réplication	36
2.2	<b>Les approches <i>genome-wide</i> rejoignent le bal des méthodes d'analyse de la réplication : nouvelles perspectives</b>	38
2.2.1	Les techniques d'analyse pangénomique de la réplication de l'ADN	39
2.2.2	Le programme temporel de la réplication de l'ADN : une fonction importante et conservée	41

2.2.2.1	Avant que les méthodes <i>genome-wide</i> ne fassent leur apparition . . . . .	41
2.2.2.2	Le repli-seq : une approche qui a révolutionné l'étude du programme temporel de réplication . . . . .	41
2.2.2.3	L'organisation des domaines de réplication . . . . .	43
2.2.2.4	La régulation spatio-temporelle de la réplication de l'ADN . . . . .	43
2.2.3	Directionnalité des fourches de réplication . . . . .	44
2.2.3.1	Domaines N/U : directionnalité des fourches et programme de réplication . . . . .	44
2.2.3.2	OK-SEQ : séquençage à haut débit de fragments d'Okazaki . . . . .	45
2.2.4	Existe-il des séquences spécifiques associées aux origines de réplication chez les métazoaires ? . . . . .	48
<b>2.3</b>	<b>L'analyse de la réplication "single molecule" (en molécule unique)</b> . . . . .	<b>49</b>
2.3.1	Les premières méthodes en molécule unique. . . . .	49
2.3.1.1	Les biais et limitations . . . . .	50
2.3.2	Le peignage moléculaire pour l'étude de la réplication de l'ADN . . . . .	50
2.3.2.1	Le principe . . . . .	50
2.3.2.2	Les origines et fourches de réplication : qu'avons nous appris avec les méthodes en molécule unique? . . . . .	53
2.3.2.3	Les limitations du peignage moléculaire. . . . .	54
2.3.2.4	Une automatisation de l'analyse des données de peignage : CASA, FiberStudio et IDeFlx. . . . .	54
2.3.3	Vers du peignage moléculaire à haut débit : OMAR ( <i>Optical Mapping of Replicating DNA</i> ). . . . .	56
2.3.4	L'apport des analyses genome-wide en molécule-unique. . . . .	58
<b>2.4</b>	<b>L'objectif ultime : la cartographie pangénomique de la réplication de l'ADN en molécule unique avec une technique d'imagerie à haut débit.</b> . . . . .	<b>59</b>
2.4.1	Problématique. . . . .	59
2.4.2	Approche. . . . .	60
2.4.2.1	Les extraits d'œufs de Xénope et leur usage dans l'étude de la réplication de l'ADN. . . . .	60
2.4.2.2	Développement d'un pipeline dédié à la cartographie de la réplication. . . . .	61
<b>3</b>	<b>Un pipeline développé et dédié à la réplication de l'ADN : HOMaRD</b> . . . . .	<b>63</b>
<b>3.1</b>	<b>Méthodologie expérimentale et acquisition des données</b> . . . . .	<b>68</b>
3.1.1	Préparation des échantillons pour l'analyse de la réplication de l'ADN . . . . .	68
3.1.2	Irys System au service de la réplication de l'ADN : une méthode de peignage moléculaire automatisé . . . . .	69

3.1.2.1	Comment les données sont-elles organisées?	70
3.1.2.2	Des images brutes à la numérisation des molécules	70
3.1.2.3	Cartographie optique des molécules d'ADN	72
3.1.3	Comment analyser la réplication de l'ADN en utilisant les outils Bionano Genomics?	73
<b>3.2</b>	<b>Qualité des données en sortie du système Irys</b>	<b>75</b>
3.2.1	Contrôle qualité des données en sortie du logiciel Autodetect	75
3.2.1.1	Fichiers texte	75
3.2.1.2	Les images TIFF	79
3.2.2	La cartographie optique avec nos données	82
<b>3.3</b>	<b>Méthodologie développée pour l'analyse de la réplication de l'ADN à partir des données du système Irys</b>	<b>83</b>
3.3.1	Des étapes de <i>post-processing</i> indispensables	83
3.3.1.1	Correction de l'illumination inhomogène sur nos images brutes	83
3.3.1.2	Recalage d'images	85
3.3.2	Extraction des profils d'intensité	94
3.3.3	Cartographie des profils d'intensité	94
3.3.4	Visualisation des profils d'intensité en molécule unique	95
3.3.5	Visualisation des profils d'intensité en population	97
3.3.6	Pipeline BionanoGenomics vs. pipeline HOMARD	97
3.3.7	La détection des segments répliqués	99
3.3.7.1	Composition du signal réplcatif (rouge)	99
3.3.7.2	Approche : optimisation d'une fonction de coût pour déconvoluer le signal réplcatif brut	102
<b>4</b>	<b>Applications et résultats</b>	<b>109</b>
<b>4.1</b>	<b>Analyse de la réplication de l'ADN de Lambda dans des extraits d'œufs de Xénope</b>	<b>111</b>
4.1.1	Analyse de la sous-population des molécules fortement répliquées des concatémères 5'-3'	112
4.1.1.1	High-Throughput Optical Mapping of Replicating DNA	112
4.1.1.2	Conclusion	123
4.1.2	Analyse de la sous-population des molécules faiblement répliquées	123
4.1.2.1	Partitionnement ( <i>clustering</i> ) des différentes configurations de jonctions des concatémères de Lambda	125



4.1.2.2	Une initiation de la réplication préférentielle au niveau des jonctions des ADN de Lambda . . . . .	130
<b>4.2</b>	<b>Analyse de la réplication de l'ADN chez l'Homme . . . . .</b>	<b>134</b>
<b>4.3</b>	<b>Une nouvelle approche pour la détection des segments répliqués . . . . .</b>	<b>137</b>
4.3.1	Un doublement de fluorescence du YOYO-1 présent mais pas systématique dans notre échantillon de chromatine de sperme de Xénope . . . . .	138
4.3.2	Existe-il un doublement de fluorescence du signal YOYO-1 dans notre échantillon de levure? . . . . .	139
<b>5</b>	<b>Discussion et Perspectives . . . . .</b>	<b>143</b>
<b>5.1</b>	<b>La cartographie de la réplication de l'ADN en molécule unique et <i>genome-wide</i> : du rêve à la réalité? . . . . .</b>	<b>145</b>
5.1.1	Le développement d'HOMaRD, un défi relevé . . . . .	145
5.1.2	Les limitations technologiques et techniques . . . . .	146
5.1.3	Les limitations expérimentales . . . . .	148
<b>5.2</b>	<b>L'avenir de l'analyse de la réplication en molécule unique . . . . .</b>	<b>150</b>
5.2.1	Le système Irys vs. le nouveau système de Bionano Genomics®, Saphyr . . . . .	150
5.2.2	La technologie des nanopores (Oxford Nanopore Technologies®) . . . . .	152
5.2.2.1	Le séquençage avec la technologie des nanopores . . . . .	152
5.2.2.2	Hennion <i>et al.</i> et Muller <i>et al.</i> pionniers de l'analyse de la réplication de l'ADN par séquençage nanopore . . . . .	152
<b>5.3</b>	<b>Faut-il oublier la technologie des nanocanaux au profit de la technologie Nanopore? . . . . .</b>	<b>154</b>
<b>6</b>	<b>Annexes . . . . .</b>	<b>157</b>
<b>6.1</b>	<b>Annexe 1   <i>Post-processing</i> des canaux bleu (YOYO-1) et vert (code-barres) . . . . .</b>	<b>159</b>
<b>6.2</b>	<b>Annexe 2   Exemples de profils d'intensité pour les données de <i>S. cerevisiae</i> . . . . .</b>	<b>160</b>





## Chapitre1

# Introduction

---

<b>1.1</b>	<b>La réplication de l'ADN . . . . .</b>	<b>20</b>
1.1.1	Lever les zones d'ombres sur le programme de réplication de l'ADN : un challenge . . . . .	20
1.1.2	Les résultats des études utilisant des méthodes en population ne s'accordent pas . . . . .	21
1.1.3	Vers une analyse en molécule unique ... . . . .	22
1.1.4	... à très haut débit . . . . .	23
<b>1.2</b>	<b>Problématique et objectif. . . . .</b>	<b>24</b>
<b>1.3</b>	<b>Plan. . . . .</b>	<b>25</b>

---



*“In the year 2020 you will be able to go into the drug store, have your DNA sequence read in an hour or so, and given back to you on a compact disc so you can analyse it.”*

---

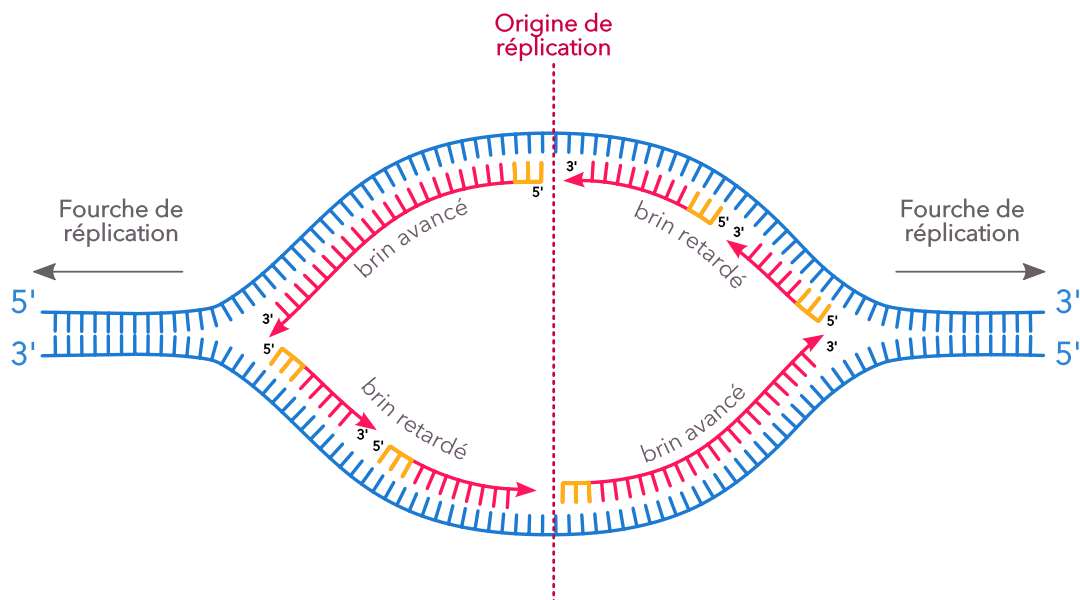
Walter Gilbert 1980 (Prix Nobel, Chimie)

En effet, Walter Gilbert n'est pas si loin de la réalité. Depuis le séquençage du premier génome humain, en 2003, les techniques de séquençage ont connu des révolutions technologiques majeures au moyen des avancées dans divers domaines tels que la physique ou encore les nanotechnologies. Il a fallu près de 15 ans et 3 milliard de dollars pour séquencer le premier génome humain (COLLINS et al. 2004), contre quelques heures et quelques centaines de dollars pour séquencer aujourd'hui n'importe quel génome et ce en utilisant différentes approches. Cependant, un simple CD-Rom ne peut contenir l'information générée par les séquenceurs. Outre, une amélioration de l'accessibilité à ces technologies avec des coûts réduits, le séquençage permet d'étudier la complexité de l'ADN, de comprendre son fonctionnement ainsi que les mécanismes dans lesquels il est en cause, et de déceler des mutations pouvant être impliquées dans certaines maladies.

## 1.1 La réplication de l'ADN

### 1.1.1 Lever les zones d'ombres sur le programme de réplication de l'ADN : un challenge

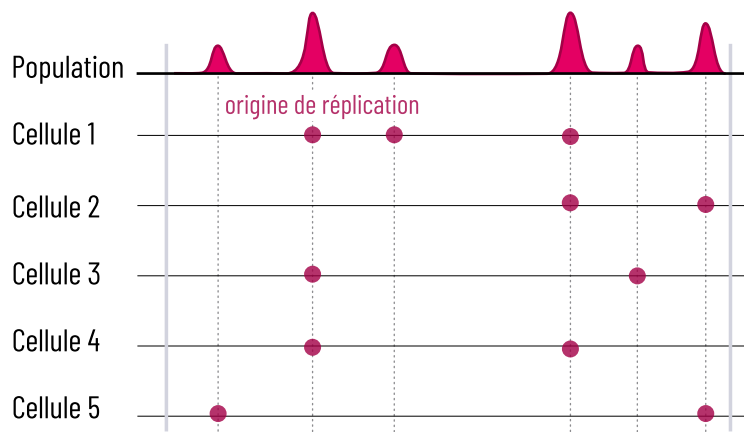
La réplication de l'ADN fait partie des processus complexes et cruciaux que nous cherchons toujours à décoder. Elle permet d'assurer la transmission fidèle et complète de l'information génétique d'une cellule mère à une cellule fille. Pour répliquer leur génome, les organismes eucaryotes utilisent de multiples sites d'initiation appelés origine de réplication (cf. Figure 1.1). Les origines sont activées stochastiquement lors de la phase S du cycle cellulaire. De chacune de ces origines va émaner deux fourches bidirectionnelles qui vont progresser le long de la molécule d'ADN à une vitesse d'environ 2 kb/min jusqu'à converger et fusionner avec les fourches adjacentes. Cette action, appelée terminaison de la réplication, est un phénomène passif. Une meilleure compréhension des facteurs régulant la réplication de l'ADN est essentielle. En effet, les perturbations de ce processus, ou stress réplcatif, menacent la stabilité du génome et ont été associées entre autre au cancer (HYRIEN, RAPPAILLES et al. 2013, HYRIEN 2015, DEWAR et WALTER 2017).



**Figure 1.1 – Bulle de réplication.** La structure schématisée ici est une bulle de réplication. Au centre de cette dernière se trouve une origine de réplication (ligne pointillée) d'où va émaner deux fourches de réplifications se déplaçant en sens opposé. En bleu sont représentés les brins d'ADN parentaux et en rouge les brins néoformés. Au cours de la réplication, la polymérisation se fait de 5' à 3' (unidirectionnelle). L'enzyme responsable de cette polymérisation est l'ADN polymérase qui nécessite une amorce ARN pour fonctionner (jaune). Le brin retardé est synthétisé de façon discontinue avec la formation des fragments d'Okazaki et le brin avancé est lui synthétisé de manière continue.

### 1.1.2 Les résultats des études utilisant des méthodes en population ne s'accordent pas

La cartographie de la réplication de l'ADN demeure encore aujourd'hui un défi et reste controversée. En effet, la cartographie pangénomique des origines de réplication chez l'Homme réalisée grâce à différentes méthodes telles que le séquençage des fragments Okazaki (OK-seq) (PETRYK et al. 2016), des sites d'initiation marqués à la digoxigénine-dUTP et isolés par immunoprécipitation (Ini-seq) (LANGLEY et al. 2016), des brins naissants isolés (SNS-seq) (BESNARD et al. 2012) ou des bulles de réplication (Bubble-seq) (MESNER et al. 2013), ne s'accordent que partiellement. Une des explications de cette incohérence est que ces approches utilisent des populations de cellules ne donnant qu'une image moyenne de la réplication. La hauteur des pics du profil moyen obtenu par une méthode populationnelle reflète l'efficacité des différentes origines de réplication (plus l'amplitude du pic est importante plus l'origine de réplication est efficace et inversement) (cf. Figure 1.2). Cela montre que chaque cellule utilise des groupes d'origines différents pour répliquer son génome supposant donc l'existence d'une variabilité inter-cellulaire.



**Figure 1.2 – Comparaison entre les méthodes en population et en molécules unique.** L'analyse de la réplication de l'ADN peut se faire soit par des méthodes populationnelles soit en molécule unique. Chaque cellule utilise des groupes d'origines différents pour répliquer son génome (CZAJKOWSKY et al. 2008). Or, les méthodes populationnelles ne nous donnent qu'une image moyenne de la réplication de l'ADN.



### 1.1.3 Vers une analyse en molécule unique ...

Les techniques en molécule unique permettent d'accéder à l'hétérogénéité entre les cellules (CZAJKOWSKY *et al.* 2008) et de visualiser des événements rares tels que l'arrêt de la progression des fourches de réplication. Le peignage moléculaire, technique communément utilisée dans le domaine de l'analyse de la réplication, consiste à étirer des molécules d'ADN marquées sur des lames de verre pré-traitées. Les cellules sont brièvement marquées par incorporation d'analogues nucléotidiques permettant ainsi de déduire la position des origines de réplication. Cependant, afin d'obtenir la position de ces origines sur le génome une étape supplémentaire est nécessaire : l'hybridation de sonde fluorescentes. Bien qu'utilisé en routine dans le domaine de la réplication, le peignage moléculaire présente certaines limitations :

1. il s'agit d'une méthode très bas débit en raison :
  - du marquage des régions d'intérêt par FISH (hybridation *in situ* par fluorescence). L'étape de dénaturation induit la décrochage et la perte de nombreuses fibres d'ADN.
  - de l'usage de sonde FISH pour identifier la région d'intérêt : cela implique que près de 99% des molécules sont non-identifiées sur une image.
2. la révélation des nucléotides fluorescents incorporés demande l'utilisation de différents anticorps induisant un signal pointillé. Cela cumulé au bruit présent dans l'image et aux molécules non linéairement étirées rendent l'automatisation et la détection des molécules ardues (cf. Figure 1.3).

Il faut compter plusieurs semaines, voir plusieurs mois afin de récolter et analyser un jeu de données statistiquement significatif.

Au sein du laboratoire, une méthode permettant le marquage de molécules d'ADN, OMAR (Optical Mapping of DNA Replication), a été élaborée et décrite en faisant usage d'un système *ex-vivo*, les extraits d'oeufs de Xénope (DE CARLI, GAGGIOLI *et al.* 2016) (cf. Sous-section 2.4.2.1). Cette approche consiste à marquer la réplication de l'ADN du bactériophage  $\lambda$  dans ce système à l'aide de dUTP (deoxyuridine triphosphate) fluorescents. Ensuite, à l'aide d'une endonucléase simple brin et l'incorporation de dUTP fluorescents au niveau des sites de coupure, un "code-barres" est ajouté à chacune des molécules d'ADN. Ce code-barres permet l'identification des fibres d'ADN peignées et leur cartographie sur un ADN de référence. L'usage de nucléotides directement fluorescents permet de contourner les étapes de révélation

avec des anticorps. Enfin, la dernière étape va consister à marquer les fibres à l'aide d'un agent intercalant, le YOYO-1.

Malgré ces améliorations non négligeables, l'automatisation de la détection des fibres d'ADN reste une tâche complexe en raison de l'irrégularité des surfaces de peignage et des molécules d'ADN non linéaires.

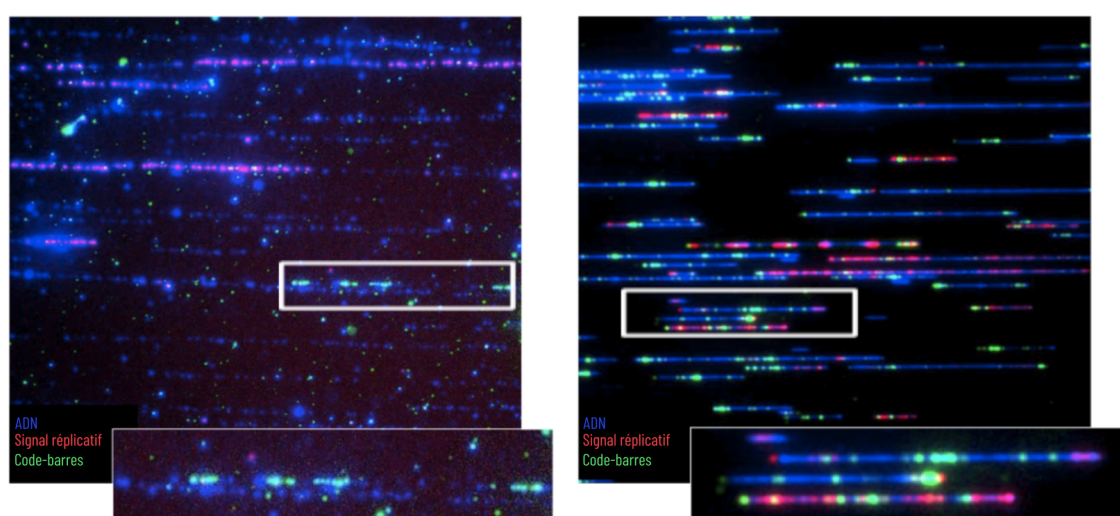
#### 1.1.4 ... à très haut débit

Une alternative à la technique de peignage moléculaire est l'usage de systèmes permettant l'étirement des fibres d'ADN dans des nanocanaux. Nous avons fait le choix de détourner à notre avantage un dispositif développé par Bionano Genomics®, le système Irys, originellement destiné à faire des assemblages *de novo* de génomes ou encore de l'analyse de variants structuraux. Cet appareil permet d'étirer par électrophorèse des centaines de milliers de molécules d'ADN marquées, dans des nanocanaux de 13 nm de diamètre. En routine, un agent intercalant, le YOYO-1, marque la molécule d'ADN (bleu) et des nucléotides fluorescents sont incorporés au niveau des coupures simples brins provoquées par une endonucléase pour obtenir le code-barres (vert). La présence d'un troisième laser dans le système Irys (rouge) rend possible la visualisation de la réplication de l'ADN. Afin d'observer les segments d'ADN répliqués, des dNTP fluorescents sont ajoutés lors du processus (cf. Figure 1.3).

Typiquement, un passage (ou *run*) d'un échantillon, génère, en une journée, plus de 32 000 images et permet d'imager automatiquement près de 1 million de molécules. Une fois cette étape réalisée, les outils bioinformatiques fournis par la compagnie permettent d'opérer une cartographie des molécules d'ADN.

## 1.2 Problématique et objectif.

Contrairement aux images obtenues en peignage moléculaire, nous avons des molécules linéaires ainsi qu’une augmentation du ratio signal/bruit (cf. Figure 1.3). Ces améliorations notables rendent possible l’automatisation de l’analyse de ces images. Cependant, les logiciels fournis par Bionano Genomics® sont propriétaires et ne peuvent être explorés et/ou modifiés. Par ailleurs, notre signal d’intérêt, à savoir le signal réplcatif, n’est pas géré par ces outils. Afin d’obtenir une meilleure compréhension du processus de réplcation de l’ADN, il nous a fallu explorer ces données afin de développer des outils bioinformatiques adaptés.



**Figure 1.3 – Peignage moléculaire vs. technologie des nanocanaux.** Le peignage moléculaire (à gauche) est une technique où des molécules d’ADN sont étirées sur des lames de verres pré-traitées (cf. Figure 2.7) alors que le système Irys est basé sur la technologie des nanocanaux permettant d’étirer automatiquement par électrophorèse des centaines de milliers de molécules (cf. Figure 3.4). Les molécules d’ADN (bleu) sont marquées au YOYO-1 (agent intercalant), le signal réplcatif (rouge) et le code-barres (vert) sont obtenus par incorporation de nucléotides fluorescents. Les images acquises avec le système Irys ne présentent quasiment pas de bruit de fond, les molécules sont rectilignes favorisant ainsi l’automatisation de l’analyse des images.

L’objectif de ma thèse a donc été d’implémenter des outils bioinformatique robustes dans le but de cartographier la réplcation de l’ADN en molécule unique et “*genome-wide*”.

Pour ce faire, nous avons décidé de nous appuyer sur les images et toutes informations utiles provenant du système Irys de Bionano Genomics® afin de pouvoir extraire les profils d'intensités de nos molécules et les cartographier. Aussi, un algorithme permettant la détection des segments répliqués dans nos signaux a été développé de manière à obtenir des informations quantitatives sur l'initiation de la réplication de l'ADN.

Grâce à nos outils, nous avons été à même d'explorer nos données de façon plus approfondie, d'abord avec des premiers résultats au niveau de la population avant d'aller vers la molécule unique. Nous avons montré que la méthode que nous avons développée, HOMaRD (*High-throughput Optical Mapping of Replicating DNA*), rend l'analyse de la réplication possible en faisant usage du système Irys mais ouvre aussi la voie à d'autres types d'analyses aussi bien en génomique fonctionnelle qu'en épigénétique.

### 1.3 Plan.

Nous verrons donc dans une première partie comment les méthodes d'analyse de la réplication ont évolué et à quels concepts elles ont donné naissance (cf. Chapitre 2). Puis nous décrirons le développement du pipeline HOMaRD en présentant les différentes étapes de *post-processing*, de cartographie pour enfin décrire l'approche de détection des segments d'ADN répliqués (cf. Chapitre 3). Dans une troisième partie, nous détaillerons les résultats obtenus sur les différents échantillons à savoir l'ADN du bactériophage  $\lambda$ , l'ADN de levure *S. cerevisiae* et enfin l'ADN humain (cf. Chapitre 4). Dans une dernière partie, nous concluons et verrons quelles sont les perspectives de ce projet (cf. Chapitre 5).



## De l'autoradiographie à l'imagerie haut débit de molécules d'ADN en cours de réplication

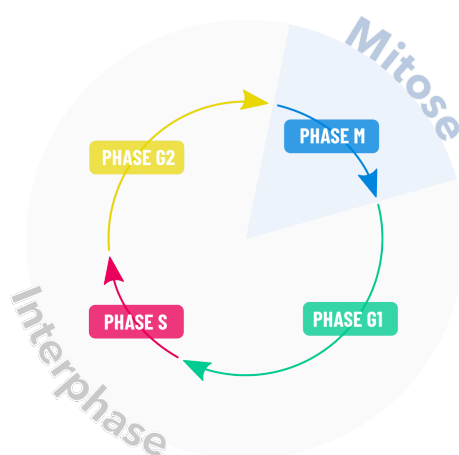
---

<b>2.1</b>	<b>Les premières études sur la réplication de l'ADN</b>	<b>30</b>
2.1.1	L'autoradiographie des fibres d'ADN	30
2.1.2	Le modèle du réplicon	32
2.1.3	Des origines de réplication aux efficacités variables	35
2.1.4	Le " <i>random-completion problem</i> " et le choix des origines de réplication	36
<b>2.2</b>	<b>Les approches <i>genome-wide</i> rejoignent le bal des méthodes d'analyse de la réplication : nouvelles perspectives</b>	<b>38</b>
2.2.1	Les techniques d'analyse pangénomique de la réplication de l'ADN	39
2.2.2	Le programme temporel de la réplication de l'ADN : une fonction importante et conservée	41
2.2.3	Directionnalité des fourches de réplication	44
2.2.4	Existe-il des séquences spécifiques associées aux origines de réplication chez les métazoaires?	48
<b>2.3</b>	<b>L'analyse de la réplication "<i>single molecule</i>"(en molécule unique)</b>	<b>49</b>
2.3.1	Les premières méthodes en molécule unique.	49
2.3.2	Le peignage moléculaire pour l'étude de la réplication de l'ADN	50
2.3.3	Vers du peignage moléculaire à haut débit : OMAR ( <i>Optical Mapping of Replicating DNA</i> ).	56
2.3.4	L'apport des analyses <i>genome-wide</i> en molécule-unique.	58
<b>2.4</b>	<b>L'objectif ultime : la cartographie pangénomique de la réplication de l'ADN en molécule unique avec une technique d'imagerie à haut débit.</b>	<b>59</b>

2.4.1	Problématique. . . . .	59
2.4.2	Approche. . . . .	60

---

Chez les eucaryotes, l'ADN (Acide DésoxyriboNucléique), molécule support de l'information génétique, est constitué de deux chaînes de nucléotides formant une double hélice. Cet ADN est transmis d'une cellule mère à ces deux cellules fille lors du cycle cellulaire qui est composé de quatre phases se succédant dans un ordre bien défini : G1 (Gap1), S (Synthèse), G2 (Gap2) et M (Mitose). Les phases G1 et G2 sont des phases de préparation à la duplication du génome et à la division cellulaire (respectivement), S correspond à la phase durant laquelle le matériel génétique est copié par le processus de réplication et enfin M renvoie à l'étape de division de la cellule.



**Figure 2.1 – Le cycle cellulaire.** Chez les eucaryotes, il existe quatre phases dans le cycle cellulaire qui sont la phase G1 correspondant à la phase de croissance, la phase S renvoyant à la phase de synthèse et donc de réplication de l'ADN, la phase G2 qui va être la phase de préparation à la mitose et enfin la phase M qui correspond à la phase de division cellulaire. Ce cycle cellulaire diffère selon les organismes et leur stade de développement.

Comme cela a été souligné dans l'introduction, chez les organismes eucaryotes la réplication de l'ADN se déclenche à de multiples origines de réplication. La détermination de ces origines a été extrêmement difficile et ce pour de nombreuses raisons telles que la taille des génomes eucaryotes pouvant aller jusqu'à plusieurs milliers de mégabases, les techniques utilisées mais aussi la rareté des intermédiaires de réplifications (HYRIEN 2015). Avec les avancées technologiques de ces dernières années, il est aujourd'hui possible de cartographier les origines de réplication par des techniques de séquençage à haut débit. Cependant, les études indépendantes et utilisant différentes approches ne s'accordent pas.

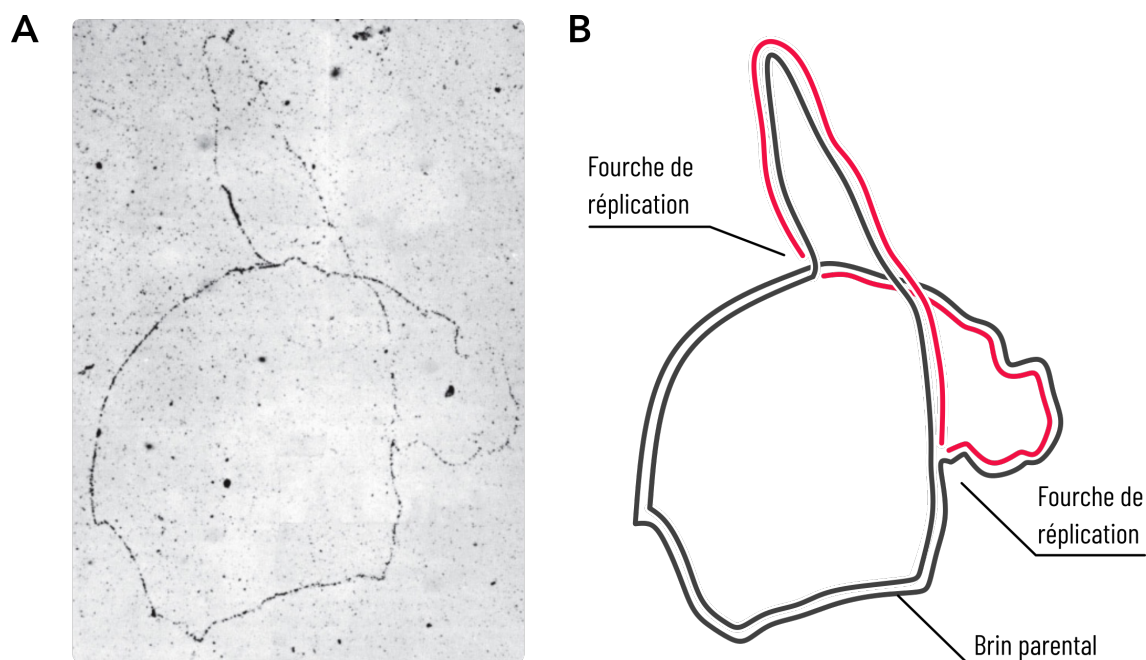
Dans ce chapitre, nous allons voir comment, depuis la découverte de la structure de l'ADN par Watson et Crick en 1953, les techniques utilisées pour l'analyse de la réplication de l'ADN ont fait évoluer nos connaissances sur le sujet et en quoi notre approche peut constituer une percée dans le domaine.



## 2.1 Les premières études sur la réplication de l'ADN

### 2.1.1 L'autoradiographie des fibres d'ADN

John Cairns décida d'appliquer la technique d'autoradiographie à l'analyse des fibres d'ADN dans le but d'étudier la structure et la réplication de l'ADN bactérien (CAIRNS 1963a, CAIRNS 1966). Son expérience consista à faire pousser *E. coli* en présence de thymidine radioactive permettant ainsi de visualiser le brin d'ADN nouvellement répliqué. Il montra que le chromosome bactérien d'*E. Coli* était une unique molécule d'ADN double brins circulaire (4,6 Mb) qui se répliquait à partir d'une origine de réplication avec deux fourches (cf. Figure 2.2).



**Figure 2.2 – Autoradiographie de l'ADN d'*E. coli*.** (A) ADN d'*E. coli* marquée avec de la thymidine tritiée et observée après autoradiographie. (B) Interprétation de la molécule d'ADN observée en (A) avec en gris foncé le brin parental et en rouge les brins néoformés. Adaptée de CAIRNS 1963b.

Cette découverte marqua un jalon dans le domaine de la réplication de l'ADN malgré les difficultés rencontrées pour la mise en place de cette technique comme le fait remarquer J.Cairns :

*“The autoradiography of DNA was a rather slow business. Exposure times were over two months. Therefore it took a couple of years to work out a technique for minimizing DNA breakage during extraction, and then it was only after quite a long search that I found any molecules that were sufficiently untangled to be interpretable.”*

---

J. Cairns

Quelques années plus tard, il a été confirmé que la réplication chez *E. coli* est bidirectionnelle : à partir d'une origine de réplication émanent deux fourches s'éloignant l'une de l'autre (MASTERS et BRODA 1971, PRESCOTT et KUEMPEL 1972). Cependant le positionnement de ces origines de réplication restait indéterminé.

Grâce à l'adaptation de la technique d'autoradiographie des fibres d'ADN à l'analyse de l'ADN eucaryote (J. HUBERMAN et A. RIGGS 1968), il a été démontré que la réplication de l'ADN chez ces organismes, en plus d'être bidirectionnelle, débutait à de multiples endroits à la différence des organismes procaryotes. Une première quantification de la distance séparant les sites d'initiation de la réplication ou origines de réplication (20 à 400 kb), ainsi qu'une estimation de la vitesse de progression des fourches (2 à 3 kb/min) ont pu être réalisées. Cette approche a également révélé que dans les cellules de mammifères ayant une phase S pouvant durer de 8 à 10h, des groupes de 5 à 10 réplicons (cf. Section 2.1.2) étaient répliqués de façon synchrone en une heure. Cela implique donc l'existence d'une activation séquentielle des *clusters* lors de la phase S du cycle cellulaire. Les différentes études menées par la suite permettent de retrouver des vitesses de progression des fourches relativement similaires mais ne s'accordent pas sur le nombre de réplicons activés au sein d'un même groupe (HAND 1975, WILLARD et S. A. LATT 1976).

Durant cette même période, Yurov et Liapunova (Yu B. YUROV et LIAPUNOVA 1977) prouvent l'existence, chez les mammifères, de réplicons allant jusqu'à 2Mb dont la réplication se faisait dans une fenêtre de temps plus longue lors de la phase S (environ 3h). Malgré une vitesse de progression des fourches quasi-constante observée dans les différents types cellulaires et dans l'ensemble des études qui ont été menées, les distances entre les origines, mais aussi la synchronicité de leur déclenchement sont

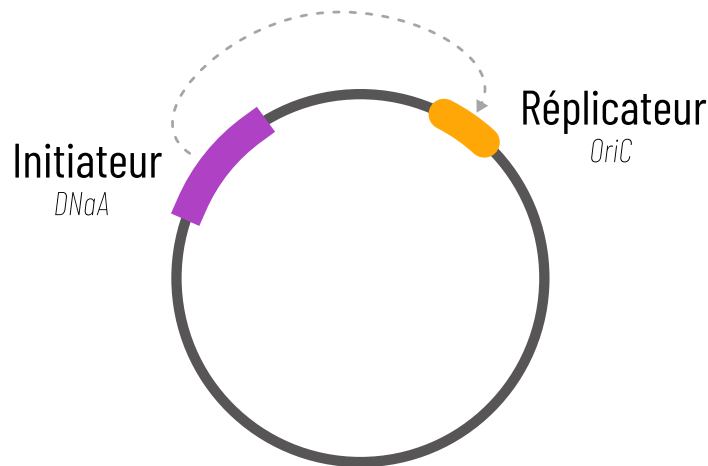
variables. Même si les raisons pouvant expliquer la présence d'une telle flexibilité restent obscures, certains chercheurs comme Berezney (BEREZNEY, Dharani DUBEY et J. HUBERMAN 2000) avancent que la cellule modifierait son programme de réplication suivant son stade de développement ou de différenciation et donc son environnement, afin de répondre à ses besoins.

De nouveaux concepts ont vu le jour grâce à la technique d'autoradiographie, cependant les fragments d'ADN n'étant pas identifiés il est difficile, voire impossible, de savoir si les origines de réplication observées sont déclenchées au niveau de séquences particulières ou non.

## 2.1.2 Le modèle du réplicon

### 2.1.2.1 Le modèle du réplicon chez les procaryotes

En parallèle des études précédentes, Jacob et Brenner proposent en 1963 (JACOB, BRENNER et CUZIN 1963) le modèle du réplicon chez les procaryotes. Le réplicon va correspondre à la séquence d'ADN qui va être répliquée à partir d'une origine de réplication et dont les fourchettes divergentes vont se rencontrer au niveau du site de terminaison. Un complexe protéique appelé initiateur reconnaît une séquence bien spécifique de l'ADN et le réplicateur permet l'ouverture de l'ADN dans cette région avant d'initier le processus de réplication. L'interaction entre ces deux éléments entraîne le recrutement de divers facteurs conduisant à l'initiation de la réplication. Le modèle décrit précédemment a été validé chez *E. coli* : le réplicateur, nommé Ori C, permet la réplication autonome du plasmide dans lequel il a été inséré (YASUDA et HIROTA 1977) et l'initiateur, appelé DnaA, a la capacité de se lier à Ori C (CHAKRABORTY et al. 1982) (cf. Figure 2.3).



**Figure 2.3 – Le modèle du réplicon.** Cette structure est composée d'un initiateur (protéine codée par le gène ici en violet) et un réplicateur (orange). L'interaction de l'initiateur avec le réplicateur va permettre l'initiation de la réplication de l'ADN. DNaA et OriC sont l'initiateur et le réplicateur chez *E. coli*.

### 2.1.2.2 La mise en évidence des initiateurs et répliqueurs chez les eucaryotes

#### L'IDENTIFICATION DES INITIATEURS CHEZ LES EUCARYOTES

L'initiation de la réplication de l'ADN peut se diviser en deux étapes :

- le *licencing* qui a lieu lors de la phase G1 et permettant de préparer les origines de réplication en cas de déclenchement.
- le *firing* qui réfère au déclenchement d'une fraction des origines de réplication préparées à l'étape précédente.

Elle est étroitement régulée de sorte que le processus de réplication ne se produit qu'une seule fois par cycle cellulaire (J. Julian BLOW et DUTTA 2005). Les étapes de *licencing* et *firing* sont interdépendantes puisque les origines de réplifications doivent être "préparées" lors du *licencing* avant d'être déclenchées et inversement. Par ailleurs, ces 2 évènements sont mutuellement exclusifs étant donné que les deux réactions sont associées à des stades spécifiques du cycle cellulaire sous le contrôle, entre autres, des kinases dépendantes de la cycline (CDK). Les CDK vont favoriser l'étape de *firing* et en même temps inhiber le *licencing* (SIDDQUI, FAN ON et J. F. DIFFLEY 2013).

L'initiateur eucaryote a été identifié pour la première fois chez la levure sous la forme d'un complexe à 6 sous-unités reconnaissant les origines de réplication connu sous le nom d'ORC (*Origin Recognition Complexe*) (S. P. BELL, KOBAYASHI et STILLMAN 2014). Son rôle est central puisque des mutations des gènes codant pour ORC entraînent des

anomalies au niveau du processus de la réplication (S. P. BELL, KOBAYASHI et STILLMAN 2014). Au cours de la phase G1 du cycle cellulaire, ORC va interagir avec les ARS (*Autonomously Replicating Sequences*) (cf. Paragraphe 2.1.2.2) avant que les protéines Cdc6 et Cdt1 ne se fixent. Le recrutement de ces dernières va conduire à la fixation d'une hélicase, MCM2-7 (*minichromosome maintenance protein complex*). Cet ensemble va former le complexe pré-réplicatif (pré-RC)(J. DIFFLEY et COCKER. 1992, J. F. DIFFLEY et al. 1994). Ensuite au cours de la phase S, MCM2-7 va être activé par les DDK (protéines kinases dépendantes de Dbf4) et les CDK (protéines kinases dépendantes des cyclines) ainsi que d'autres facteurs (FRAGKOS et al. 2015).

#### L'IDENTIFICATION DES RÉPLICATEURS CHEZ LES EUKARYOTES : QUAND ARS ET PLASMIDES S'ASSOCIENT ...

À la différence des initiateurs, bien caractérisés, les réplicateurs chez les eucaryotes font encore débat dans la communauté scientifique. En 1979, Stinchcomb et Struhl proposent une approche rendant possible l'identification des réplicateurs eucaryotes (STINCHCOMB, STRUHL et DAVIS 1979, Kevin STRUHL et al. 1978). Pour ce faire, l'ADN de *S. cerevisiae* est fragmenté et chacun des fragments obtenus va être cloné dans des plasmides ayant un marqueur de sélection. Si la cellule est en mesure de pousser dans un milieu sélectif, cela signifie qu'elle a soit incorporé un plasmide capable de se répliquer de façon autonome dans la cellule, soit qu'elle a intégré le marqueur de sélection dans un de ses chromosomes. Ainsi les colonies obtenues pourront être isolées et analysées dans le but de mieux caractériser les fragments clonés. Grâce à cette expérience, les premiers réplicateurs eucaryotes ont pu être isolés et identifiés chez la levure *S. cerevisiae*, les ARS. Celles-ci sont des séquences riches en A/T de 100 à 200 paires mais qui diffèrent entre elles par la séquence. Elles possèdent une séquence consensus de 11 paires de bases riche en Thymine, *ARS consensus sequence* (ACS) qui est nécessaire à leur fonctionnement mais ne constitue pas un élément suffisant. En effet, un élément non consensus situé en position 3' de l'ACS est aussi essentiel à l'activité des ARS (NEWLON et THEIS 1993). Eaton et al. montrent par séquençage et cartographie des sites de liaisons d'ORC (se liant aux ACS) et des nucléosomes, que des nucléosomes flanquent les ACS en 3' empêchant la liaison d'ORC. Ils suggèrent que les nucléosomes jouent un rôle déterminant dans la sélection et le fonctionnement des origines de réplication (EATON et al. 2010).

Afin d'aller plus loin dans l'analyse de la réplication, une autre technique, plus poussée que l'autoradiographie ou encore que l'expérience avec les ARS, est utilisée : l'électrophorèse 2D sur gel d'agarose.

### 2.1.3 Des origines de réplication aux efficacités variables

Bell et Byers ont été les premiers à démontrer l'utilité de l'électrophorèse sur gel 2D pour distinguer des intermédiaires de recombinaison ramifiés à partir de molécules linéaires, en se basant sur le différentiel de migration entre ces deux types de molécules, dans des conditions électrophorétiques adéquates (L. BELL et BYERS 1983). En 1987, Brewer et Fangman (BREWER et FANGMAN 1987) mais également Huberman (NAWOTKA et J A HUBERMAN 1988), ont des approches différentes mais adaptent tous deux l'utilisation du gel d'agarose bidimensionnel de Bell et Byers, dans le but de localiser les origines de réplication et de déterminer la direction du mouvement des fourches de réplication.

L'idée repose sur l'élaboration de conditions électrophorétiques permettant de discerner les molécules d'ADN linéaires non répliquées, des molécules non-linéaires en cours de réplication. La première dimension consiste en une séparation qui se fait à basse tension afin de séparer les molécules d'ADN selon leur masse. Pour la seconde dimension, la migration se fait sous haute tension dans un gel à plus forte concentration d'agarose de façon à ce que la mobilité d'une molécule non linéaire soit influencée par sa forme.

Dans le cas du gel 2D neutre/neutre de Brewer et Fangman, l'ADN génomique total est digéré à l'aide d'une enzyme de restriction. Une première migration sur un gel neutre est réalisée permettant une séparation des fragments de restriction en fonction de leur masse. Les fragments ramifiés de masse identique mais de formes différentes sont séparés orthogonalement par une électrophorèse dans des conditions neutres sous haute tension et haute concentration d'agarose. Les fragments de restrictions sont donc séparés selon leur forme : les intermédiaires ayant deux fourches divergentes (renvoyant à une bulle de réplication), une fourche divergente (en forme d'un simple Y, renvoyant à une réplication passive) ou encore deux fourches convergentes (en forme d'un double Y renvoyant à une terminaison).

Dans le cas du gel 2D neutre/alcalin d'Huberman, la première migration est la même que celle vue précédemment mais diffère par la seconde migration. Cette dernière est réalisée avec un angle de 90° dans des conditions alcalines ayant pour effet de dénaturer l'ADN induisant la séparation du brin parental et du brin naissant. Après avoir

transféré la membrane, des sondes sont utilisées pour identifier les configurations des brins naissants.

L'électrophorèse 2D sur gel est une technique qui a permis de donner naissance à de nouveaux concepts tels que le fait que sur les plasmides recombinés avec des fragments d'ADN de *S. cerevisiae*, la réplication de l'ADN est initiée uniquement au niveau des ARS. Il a également été démontré que dans un contexte chromosomal, les ARS sont activées à des moments différents au cours de la phase S du cycle cellulaires et avec des efficacités variables. Par ailleurs, une étude menée en 1988 par Linskens (LINSKENS et J A HUBERMAN 1988) sur les intermédiaires de réplication de l'ADN ribosomique chez la levure (120 tandems répétés de 9,1 kb), a montré que seule une faible proportion des ARS était déclenchée au sein de cette région génomique. Les ARS non utilisées, quant à elles, sont répliquées de façon passives par les fourches en provenance des origines de réplication voisines. Ces origines non-activées sont qualifiées de latentes.

#### 2.1.4 Le "*random-completion problem*" et le choix des origines de réplication

Czajkowski et *al.* ont analysé la réplication du chromosome VI de la levure *S. cerevisiae* par la technique de peignage moléculaire CZAJKOWSKY et al. 2008 et ont montré que pour répliquer leur génome, chaque cellule utilise un groupe d'origines différent. Il apparaît donc que l'activation des origines de réplication est probabiliste. Cependant, d'après les paramètres connus concernant la réplication de l'ADN chez le Xénope, à savoir la distance inter-origine moyenne et la vitesse moyenne de progression des fourches, il faudrait s'attendre à avoir des segments d'ADN non répliqués. Cela implique donc un allongement de la phase S afin de compléter le processus de réplication. Ce phénomène, observé par Laskey en 1985 (R A LASKEY 1985), a été baptisé le "*random completion problem*" (HYRIEN, MARHEINEKE et GOLDAR 2003, HYRIEN, RAPPAILLES et al. 2013). Appuyons-nous sur l'exemple du Xénope dont la phase S, d'une durée de  $\approx 20$  minutes, est tout de suite suivi de la phase M ( $\approx 10$  minutes) sans phase G1 ni G2. La vitesse de progression des fourches est de 0.5 kb par minute signifiant donc que deux fourches issues d'une même origine ne peuvent pas répliquer plus de 20 kb par cycle cellulaire. Dans le cas où nous supposons que les origines de réplication sont déclenchées de façon synchrone, les événements d'initiation ne peuvent pas être éloignés de leurs voisins de plus de 20 kb. Or, des études ont montré que le déclenchement des origines de réplication n'est pas synchrone pour ce qui est de la réplication de

plasmides (LUCAS *et al.* 2000) ou de chromatine de sperme de Xénope dans des extraits d'oeufs de Xénope (HERRICK *et al.* 2000, J Julian BLOW *et al.* 2001). Les origines doivent donc être moins espacées les unes des autres. Par ailleurs, si les origines de réplication étaient positionnées de façon aléatoire, la distribution des distances inter-origines suivrait une exponentielle décroissante et cela impliquerait qu'à la fin de la phase S un grand nombre de segments non répliqués persisterait. Ainsi, afin d'assurer la réplication complète d'un génome, l'espacement entre les évènements d'initiation doit être plus régulier malgré l'absence d'initiation séquence spécifique.

Il a été suggéré, pour résoudre ce "*random-completion problem*", que pour compléter le processus de réplication lors de la phase S, il y aurait une augmentation du taux d'initiation des origines de réplication (HYRIEN, RAPPAILLES *et al.* 2013). Une étude réalisée par peignage moléculaire chez la levure *S. pombe* avec des fibres d'ADN atteignant 1.5 à 2.5 Mb de long (KAYKOV *et NURSE* 2018) coïncide avec la solution énoncée précédemment. Kaylov et Nurse observent une augmentation graduelle du taux d'évènements d'initiation au cours de la phase S et un déclenchement de toutes les origines de réplication disponibles dans les segments d'ADN non répliqués en fin de phase S.



## 2.2 Les approches *genome-wide* rejoignent le bal des méthodes d'analyse de la réplication : nouvelles perspectives

Les décennies qui ont suivies la découverte de la structure de l'ADN, ont témoigné de la naissance de dizaines de nouveaux concepts relatifs au processus de la réplication de l'ADN. Ces derniers ont vu le jour grâce au développement de nouvelles méthodes permettant l'identification ainsi que la caractérisation des origines de réplifications chez différents organismes. Cependant, les techniques présentées dans la section précédente ne permettent d'analyser la réplication de l'ADN qu'au niveau de locus spécifiques et non pas au niveau du génome dans sa globalité. Avec l'apparition de la technologie des puces à ADN et des techniques de séquençage nouvelle génération (NGS, *Next Generation Sequencing*), les études pangénomique de la réplication de l'ADN se sont multipliées dans l'espoir d'avoir des réponses plus précises sur ce qui caractérise les origines de réplication en particulier chez les métazoaires (SCHEPERS et PAPIOR 2010, URBAN et al. 2015) mais aussi sur le programme de réplication (RHIND et GILBERT 2013).

À la différence des cartographies des origines de réplication qui sont plus résolutive, mais ne coïncident que très peu entre les différentes méthodes et études, le programme de réplication est lui reproductible, mais pas assez résolutif pour cartographier les origines individuelles.

Avant de voir comment les méthodes populationnelles ont permis d'accéder à ces informations, nous allons débiter par une présentation rapide des méthodes existantes dont certaines seront détaillées dans les sections suivantes.

### 2.2.1 Les techniques d'analyse pangénomique de la réplication de l'ADN

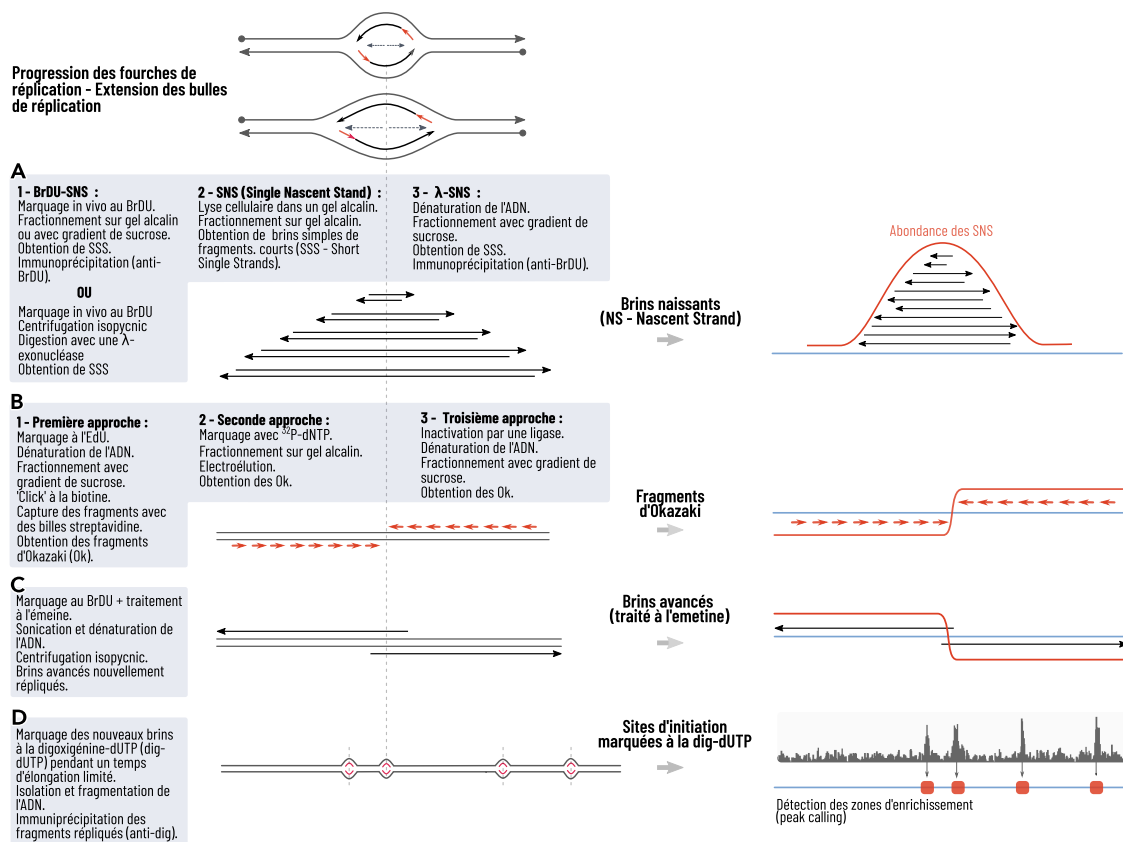
La réplication de l'ADN a été jusqu'à maintenant étudiée sur un échantillon de molécules limité en nombre et sur des locus particuliers avec des approches consistant à purifier des bulles de réplication, des brins naissants ou de fragments d'Okazaki. Ces méthodes associées à du séquençage haut-débit induisent une amélioration considérable du rendement donnant ainsi accès à un niveau de connaissance supplémentaire. Par exemple, l'étude des sites de liaisons des ORC chez la levure ou encore chez l'Homme a été possible grâce au couplage de l'immunoprécipitation de chromatine sur les séquences de liaison des ORC à du séquençage (MIOTTO, Ji et Kevin STRUHL 2016, DELLINO et al. 2013).

Les origines de réplifications peuvent être analysées sur l'ensemble d'un génome grâce à l'immunoprécipitation des brins naissants marqués au BrdU (bromodeoxyuridine) (BrdU-SNS) (MUKHOPADHYAY et al. 2014). Cette même étude peut être réalisée par une technique faisant usage d'une  $\lambda$ -exonucléase, une 5'-exonucléase qui élimine tous les brins d'ADN à l'exception des brins naissants ( $\lambda$ -SNS) (MARTIN et al. 2011, BESNARD et al. 2012). Dans le cas où l'ADN marqué au BrdU de chacune des sous-fractions de la phase S est immunoprécipité et séquencé (repli-seq, cf. Section 2.2.2.2)), nous sommes en mesure d'accéder au programme temporel de la réplication.

Une autre alternative aux brins naissants est le séquençage des bulles de réplication dont l'origine se trouve au centre de celles-ci (avec une vitesse de progression des fourches considérée comme constante) (bubble-seq, MESNER et al. 2013).

Une autre approche est basée sur la synthèse des brins avancés au niveau des origines et marqués au BrdU afin d'étudier la réplication au locus *DHFR* (dihydrofolate réductase) (HANDELI et al. 1989) où les cellules sont traitées avec de l'emetine afin d'empêcher la synthèse du brin retardé.

Le séquençage des fragments purifiés d'Okazaki (OK-seq, cf. Section 2.2.3.2), nous permet d'obtenir la directionnalité des fourches en plus d'identifier les sites d'initiation et de terminaison de la réplication (PETRYK et al. 2016). Une autre méthode *genome-wide* consiste à séquencer les sites d'initiation de la réplication en les marquant avec des digoxigénine-dUTP et les immunoprécipitants avec des anticorps anti-digoxigénine, dans un système cellulaire similaire au système *ex-vivo* du Xénope, le *human cell-free system* (HCFS) (ini-seq) (LANGLEY et al. 2016)(cf. Figure 2.4).



**Figure 2.4 – Schéma présentant des approches populationnelles pour l'analyse de la réplcation de l'ADN.** Sont schématisés ici les brins naissants (NS), les fragments d'Okazaki ainsi que les brins avancés au niveau des origines de réplcation (schéma à gauche). Les approches utilisées pour isoler ces mêmes structures sont présentées dans les cadres gris et à droite sont schématisées les méthodes d'analyse pour cartographier des origines de réplcation. Pour les méthodes basées sur l'abondance des SNS (*short nascent strand*) (**A**), l'ADN total est dénaturé et les SSS (*short single strand*) (0.5-3 kb) sont isolés sur gel d'agarose ou gradient de sucrose afin d'exclure les fragments d'Okazaki. L'enrichissement en SNS est obtenu par marquage au BrdU de l'ADN néosynthétisé et purification par immunoprécipitation ou centrifugation isopycnique (A1), par lyse cellulaire directement dans un gel d'agarose alcalin limitant ainsi les cassures (A2), ou en traitant les SSS avec une λ-exonucléase (5'-exonucléase) digérant l'ADN et non pas l'ARN conservant les brins naissants attachés à une amorce ARN (A3). Les origines de réplcation cartographiées sont déterminées grâce à l'abondance des SNS obtenue par PCR quantitative (VASSILEV et JOHNSON 1989), hybridation sur puce à ADN (LUCAS et al. 2000) ou séquençage haut débit (BESNARD et al. 2012). (**B**) Pour isoler les fragments d'Okazaki, il existe différentes approches : ces fragments sont marqués à l'EdU, couplés à de la biotine et isolés grâce à des billes de streptavidine (B1) (PETRYK et al. 2016), ou bien les cellules sont radiomarquées avant d'être isolées sur gel d'agarose alcalin (B2) (BURHANS et al. 1990), ou encore les fragments d'Okazaki sont accumulés après l'inactivation de la ligase, purifiés puis séquencés (B3) (Duncan J SMITH et WHITEHOUSE 2012). (**C**) Les brins avancés peuvent aussi permettre la cartographie des origines de réplcation en isolant les NS marqués au BrdU dans des cellules traitées à l'emetine (empêche la synthèse du brin retardé) (HANDELI et al. 1989). (**D**) Les sites d'initiation peuvent être marqués à la digoxigénine-dUTP et isolés par immunoprécipitation (billes couplées à de l'anti-digoxigénine) puis séquencés (LANGLEY et al. 2016). Adaptée de HYRIEN 2015, LANGLEY et al. 2016

## 2.2.2 Le programme temporel de la réplication de l'ADN : une fonction importante et conservée

Au cours de la phase S, les eucaryotes répliquent leur génome en suivant un programme temporel de réplication avec des origines de réplication se déclenchant à des temps différents. Le programme temporel de la réplication de l'ADN est présent chez tous les organismes eucaryotes suggérant qu'il s'agit d'une fonction importante et conservée. Cependant, ce processus demeure encore un mystère que la communauté scientifique tente d'élucider.

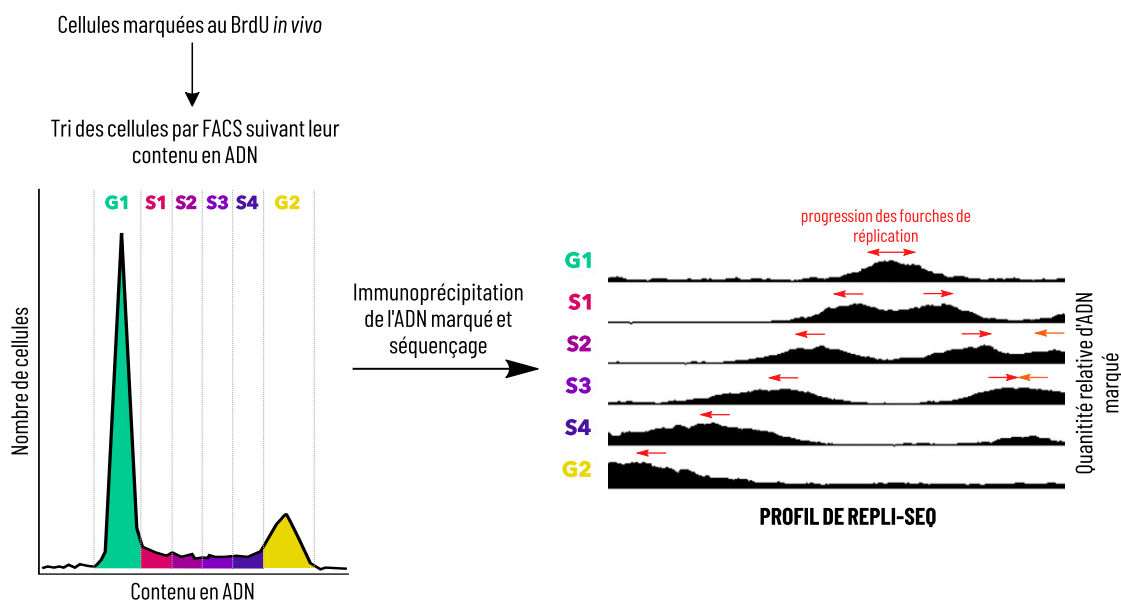
### 2.2.2.1 Avant que les méthodes *genome-wide* ne fassent leur apparition

En 1958, Taylor identifie les premières zone de réplication par étude cytogénétique de chromosomes. Ces derniers provenaient de cellules marquées de façon transitoire, à la thymidine radioactive, au cours de la phase S. Cette technique a mis en évidence l'existence de segments d'ADN ayant incorporé cette thymidine de façon synchrone (TAYLOR 1958). Une comparaison de ces segments avec les bandes chromosomique teintées au Giemsa, qui délimitent des régions d'euchromatine (conformation peu compacte de l'ADN et situé au centre du noyau) et d'hétérochromatine (conformation compacte de l'ADN et situé en périphérie du noyau) sur la longueur d'un chromosome, révèle une correspondance des bandes claires R avec les segments répliqués précocement et des bandes foncées G avec les segments répliqués tardivement. Cela nous indique donc que les bandes ne sont pas répliquées en même temps lors de la phase S (CRAIG et BICKMORE 1993, HOLMQUIST 1992, Samuel A. LATT 1977). Plus tard, des foyers de réplifications ont été identifiés par microscopie électronique de cellules de mammifères ayant été marqués en présence de BrdU et se répliquant à des moments différents de la phase S (JACKSON et POMBO 1998).

### 2.2.2.2 Le repli-seq : une approche qui a révolutionné l'étude du programme temporel de réplication

Le repli-seq consiste à marquer transitoirement, avec du BrdU, des cellules asynchrones en culture. Les cellules sont ensuite triées selon leur contenu en ADN en plusieurs fractions, divisant ainsi la phase S. L'étape suivante va consister à immunoprécipiter l'ADN marqué au BrdU de chacune des sous-fractions et à les analyser par séquençage à haut débit. Grâce à cette approche le fait que les origines

de réplication se déclenchent à des temps différents lors de la phase S a été consolidé. Chez les métazoaires, il existe un programme temporelle de la réplication organisé en domaines allant de 400 à 800 Kb répliqués par 5 à 10 origines de réplication en 60 minutes. Ces domaines sont répliqués en déclenchant de façon synchrone des origines suffisamment proches les unes des autres pour que la longueur de chaque domaine soit répliquée dans un laps de temps relativement court (environ 1h chez les mammifères)([DM GILBERT 2006](#)). Par ailleurs, outre la présence de domaines constitutifs, ce programme temporel connaît des changements important au cours du développement générant ainsi des profils spécifiques du type cellulaire. En réalité, les profils temporels de réplication sont une caractéristique reproductible de certains types de cellules à tel point qu'ils peuvent être utilisés pour les identifier ([POPE et al. 2011](#), [RYBA et al. 2011](#)). Ces domaines s'activent soit précocement au cours de la phase S, soit de façon plus tardive, et ce, dans une certaine fenêtre de temps. Étant donnée la faible variabilité de la vitesse moyenne des fourches de réplication au cours de la phase S, l'organisation de ces domaines serait médiée par l'efficacité des origines de réplication mais aussi par le moment auquel elles se déclenchent. De plus, il semblerait que les origines de réplication les plus efficaces soient activées précocement au cours de la phase S.



**Figure 2.5 – Répli-seq, une technique d'analyse du programme temporel de la réplication de l'ADN.**

La première étape va consister à marquer les cellules au BrdU avant de les trier par cytométrie de flux. La phase S peut être subdivisée en quatre parties. L'ADN de chacune de ces sous-fractions va être immunoprécipité avant d'être séquencé. Adaptée de [HANSEN et al. 2010](#)

### 2.2.2.3 L'organisation des domaines de réplication

Le programme de réplication est fortement corrélé avec l'organisation tridimensionnelle de la chromatine. Les domaines de réplication précoces se situent dans la partie la plus interne du noyau alors que les domaines de réplication tardifs sont plus localisés dans la zone riche en hétérochromatine en périphérie (BEREZNEY, Dharani DUBEY et J. HUBERMAN 2000). La technique de capture de conformation de la chromatine, ou Hi-C, est réalisée par liaison transversale (*cross-link*) de la chromatine cellulaire. La méthode de Hi-C permettant de cartographier les interactions *genome-wide*, a montré que la chromatine était organisée en 2 compartiments désagrégés spatialement et coïncidant avec les domaines précoces et tardifs (BURTON et al. 2009). Dans le programme temporel de réplication, nous pouvons distinguer deux types de régions : les TTR (Transition Timing Region) et les CTR (Constant Timing Region). Les zones de transitions (TTR) de 100 à 600 kb révèlent une modification progressive, de précoce à tardif, du programme temporel de réplication. Ces TTR vont s'intercaler entre les CTR qui correspondent à des domaines de réplifications précoces ou tardifs. Les CTR, en raison de leur taille, sont le site de déclenchement de multiples origines de réplication à la différence des TTR qui sont soit répliqués par des fourches unidirectionnelles provenant des CTR précoces vers les CTR tardif (FARKASH-AMAR et al. 2008), soit par déclenchement séquentiel d'origines de réplication (GUILBAUD et al. 2011).

### 2.2.2.4 La régulation spatio-temporelle de la réplication de l'ADN

La régulation du programme de réplication semble être régi par deux mécanismes bien distincts : d'une part, le placement des origines de réplication qui se déroule au cours de la phase G1 du cycle cellulaire et d'autre part, le déclenchement de ces origines au cours de la phase S. La probabilité de déclenchement des origines est contrôlée par l'existence de facteurs limitants tels que les protéines Sld2, Sld3, Cdc45 et DDK chez la levure, qui sont moins abondantes que le nombre d'origines activées au cours du cycle (MANTIERO et al. 2011), ou encore chez les mammifères où Cdc45 aurait une influence sur le nombre de fourches produites (P. G. WONG et al. 2011). Par ailleurs, ces mêmes facteurs limitants seraient recrutés par les origines de réplication les plus avides et donc les plus précoces. Il a également été montré que ces facteurs limitant d'initiation de la réplication sont recyclés lors de la phase S afin de déclencher les origines dont l'efficacité est plus faible rendant possible la réplication du génome dans son intégralité. La régulation de l'affinité des origines

pour les facteurs limitant reste encore incomprise. Une des hypothèses serait que le contexte chromatinien établi en phase G1 joue un rôle dans la capacité des origines à recruter les facteurs limitants lors de la phase S. Il est aussi plausible de penser que le taux de diffusion des activateurs diffèrent suivant les domaines (précoces ou tardifs) (GAUTHIER et BECHHOEFER 2009).

## 2.2.3 Directionnalité des fourches de réplication

### 2.2.3.1 Domaines N/U : directionnalité des fourches et programme de réplication

Les premiers séquençages du génome bactérien ont mis en évidence l'existence d'une composition en nucléotide asymétrique des brins avec un enrichissement en Guanine (G) et Thymidine (T) plutôt qu'en Adénine (A) et Cytosine (C) pour le brin sens (LOBRY 1996). Cela implique donc que cette constitution est corrélée avec le sens de réplication. Il est donc possible de définir des inversions de ratio GC et TA (*skew*) tels que  $SGC = (G - C)/(G + C)$  et  $STA = (T - A)/(T + A)$ . Chez les bactéries, ces inversions vont concorder avec la position d'une origine ou bien d'une terminaison (GRIGORIEV 1998). Dans la plupart des cas, le passage d'une valeur positive à négative de SGC et STA se fait au niveau des origines de réplication et dans le cas contraire il s'agira d'une terminaison (négatif à positif). Cette approche a également été appliquée aux métazoaires permettant la détermination de N-domaines qui traduit le changement linéaire de la direction des fourches situé entre deux zones d'initiation (TOUCHON et al. 2005, HYRIEN, RAPPAILLES et al. 2013). Les N-domaines ont été comparés avec le programme de réplication faisant ressortir une corrélation entre ces N-domaines et les U-domaines (du programme de réplication). Les bords des U-domaines vont correspondre à des zones d'initiation précoces alors que le centre va renvoyer à des zones d'initiation plus tardives (cela correspondrait à deux TTR entourant une CTR).

### 2.2.3.2 OK-SEQ : séquençage à haut débit de fragments d'Okazaki

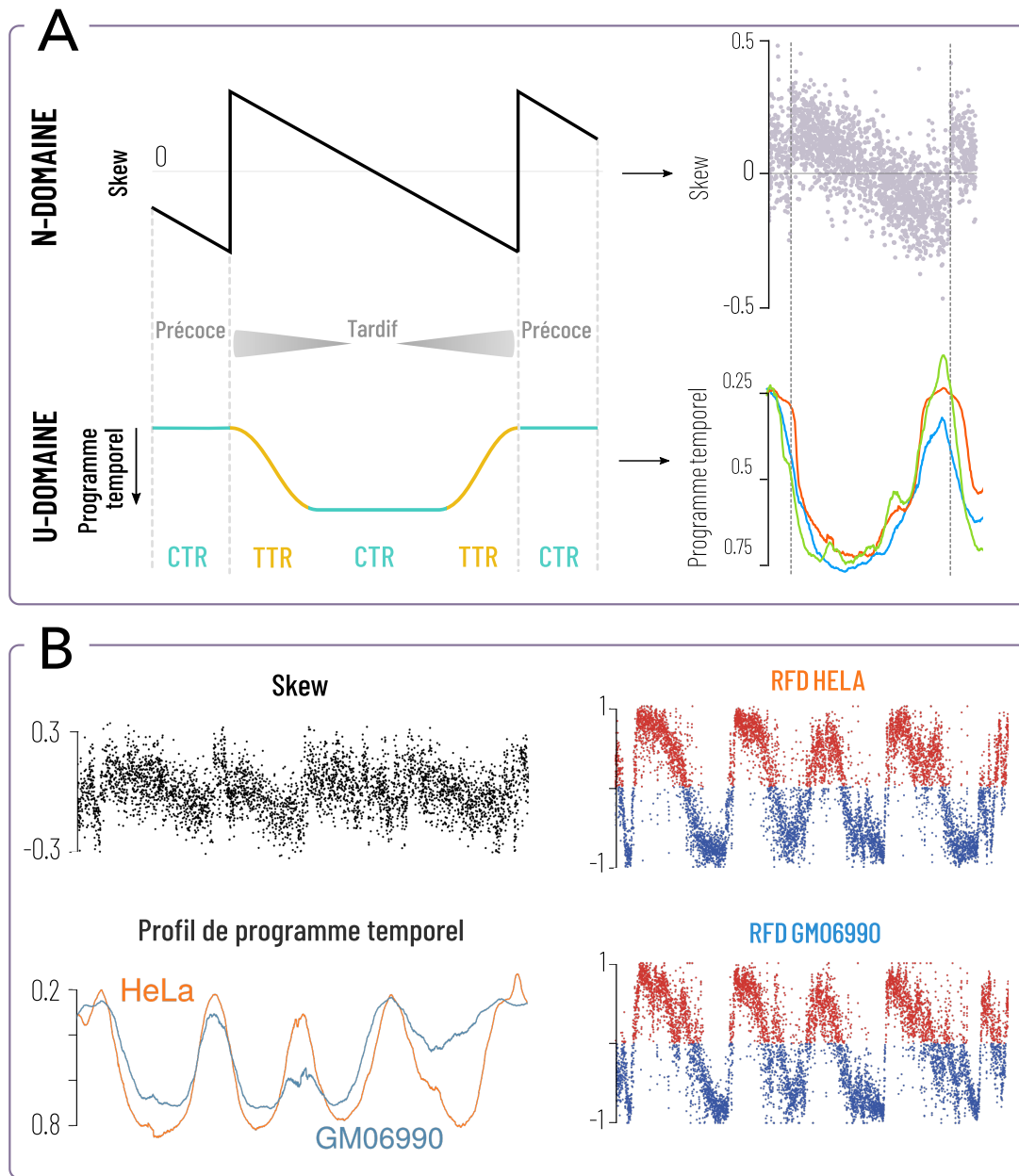
Jusqu'à présent, l'identification des origines de réplifications *genome-wide* chez les métazoaires s'est faite en effectuant un séquençages sur différents types d'échantillons tels que la capture des bulles de réplification (bubble-seq) (MESNER et al. 2013), la purification de brins naissants (SNS-seq) (VASSILEV et JOHNSON 1989), l'immunoprécipitation d'ADN marqué au BrdU (KARNANI et al. 2010) ou bien des ORC (origin recognition complex) (DELLINO et al. 2013) ou encore les sites d'initiation de la réplification marquées avec des digoxigénine-dUTP (Ini-seq) (LANGLEY et al. 2016). Le nombre d'origines de réplification allant de 12 000 à 250 000 origines, mais aussi leur localisation détectée avec ces approches, divergent. Les techniques présentées dans les sections précédentes sont des méthodes dont la résolutions est insuffisante pour accéder à la localisation des évènements d'initiation ou de terminaison le long des domaines N/U.

La méthode fondée sur l'étude des fragments d'Okazaki, développée chez *S. cerevisiae*, est la solution à ce problème de résolution permettant une analyse des fourches de réplification, de leur progression et de leur terminaison (Duncan J SMITH et WHITEHOUSE 2012, MCGUFFEE, Duncan J. SMITH et WHITEHOUSE 2013). La première étape va consister à isoler et séquencer les fragments d'Okazaki pour calculer la proportion de fourches provenant de la droite (R) et de la gauche (L). La directionnalité des fourches (RFD, *Replication Fork Directionality*) va donc correspondre à la différence entre les proportions de fourches L et R en train de répliquer au niveau d'un locus dans une population de cellules. Ainsi, la transition entre les fourches provenant de la droite et celle provenant de la gauche s'effectue au niveau d'une origine de réplification et qui est donc associée à une pente positive du profil de RDF.

Pour la première fois, une analyse approfondie de la progression des fourches de réplification, de l'initiation et de l'efficacité des origines de réplification à l'échelle du génome, mais aussi l'observation des évènements de terminaison, ont été possibles. Il a été montré, chez la levure, que les initiations se faisaient dans des régions exemptes de nucléosomes et que le processus de terminaison était un phénomène passif se produisant entre deux origines de réplification. Malheureusement, cette approche nécessite des mutations au niveau des gènes codant la ligase et les points de contrôle (*checkpoint*), permettant l'accumulation de fragments d'Okazaki (Duncan J SMITH et WHITEHOUSE 2012, MCGUFFEE, Duncan J. SMITH et WHITEHOUSE 2013).



Récemment a été décrite une nouvelle méthode, l'OK-seq (PETRYK *et al.* 2016), consistant à séquencer les fragments d'Okazaki sans nécessiter de modifications génétiques sur les lignées cellulaires comme cela a pu être le cas auparavant. Les profils d'OK-seq s'accordent avec le modèle des N/U domaines mais également les profils de Repli-seq (cf. Figure 2.6). Cette technique montre que l'initiation de la réplication se produit au niveau de zones d'initiation (150 kb). Les résultats obtenus par OK-seq pour ce qui est de l'analyse des zones d'initiations, concordent mieux avec ceux obtenus par le séquençage des bulles de réplication (bubble-seq) qu'avec les études réalisées par séquençage de brins naissants (SNS-seq).



**Figure 2.6 – Directionnalité des fourches de réplication et programme temporel d’activation des origines de réplication.** (A) : En haut est schématisé le N-domaine qui est déterminé par  $S = \text{STA} + \text{SGC}$  où  $\text{STA} = (\text{T}-\text{A})/(\text{T}+\text{A})$  and  $\text{SGC} = (\text{G}-\text{C})/(\text{G}+\text{C})$ . En bas, est schématisé un U-domaine où sont représentés les TTR (*transition timing regions*) se trouvant entre deux CTR (*constant timing regions*). Sur la droite sont donnés des exemples de N et U domaines du chromosome 10 chez l’homme. Les programmes temporels de différentes lignées cellulaires est présenté ici (orange=K562, bleu=GM06990, vert=BG02 ESC). (B) : Directionnalité des fourches de réplication (RFD) dans deux lignée cellulaire à savoir HeLa et GM06990 déterminée par la technique de séquençage des fragments d’Okazaki. Ici, les profils de RFD coïncident avec les U-domaines mais également avec les N-domaines. La RFD est calculée à partir du nombre de read ayant été cartographié sur le génome sur les brins Watson (bleu) et Crick (rouge) en appliquant la formule suivante  $\text{RFD} = (\text{C}-\text{W})/(\text{C}+\text{W})$ . Adaptée de [GUILBAUD et al. 2011](#) et [PETRYK et al. 2016](#)

## 2.2.4 Existe-il des séquences spécifiques associées aux origines de réplication chez les métazoaires ?

Des études se basant sur la cartographie des brins naissants à l'échelle du génome suggèrent que les origines de réplication se situent de préférence au niveau des séquences riches en GC (Guanine, Cytosine) mais sont également associées aux îlots CpG (BESNARD et al. 2012, PICARD et al. 2014, MUKHOPADHYAY et al. 2014). Les séquences d'ADN riches en G peuvent former des structures appelées des G-quadruplexes (composés de 4 brins) qui sont retrouvés dans 90 % des origines chez l'Homme d'après Besnard et al. (BESNARD et al. 2012). Karnani et al. (KARNANI et al. 2010) réalisent un séquençage de brins naissants et observent un enrichissement des sites d'initiation au niveau de séquences riches en A/T. Contrairement aux études présentées ci-avant, des approches, telles que l'OK-seq, montrent que les sites d'initiation de la réplication ne sont ni associés aux G-quadruplex ni aux îlots CpG malgré l'enrichissement des îlots CpG au niveau des bords des zones d'initiation (PETRYK et al. 2016). Des conclusions similaires sont tirées par les méthodes de cartographie des bulles de réplication (MESNER et al. 2013). Par ailleurs, des zones d'initiations de la réplication ont été identifiées dans ces mêmes études au lieu de sites de réplication spécifiques. Petryk et al. identifient par OK-seq des zones d'initiation allant jusqu'à 150 kb et dont la taille moyenne est de 30 kb.

## 2.3 L'analyse de la réplication "single molecule" (en molécule unique)

Les méthodes d'analyse populationnelles ont permis d'acquérir de nouvelles connaissances mais aussi de les approfondir malgré l'existence d'un désaccord entre les différentes approches. L'analyse en molécule unique permet d'accéder aux complexités et les subtilités qui sont "lissées" et occultées dans les approches populationnelles. Par ailleurs, les méthodes en molécule unique permettent d'accéder, entre autres, à la variabilité inter-cellulaire qui est indispensable à la compréhension de la régulation de la réplication de l'ADN.

### 2.3.1 Les premières méthodes en molécule unique.

L'observation et l'étude des bulles de réplication ainsi que la bidirectionnalité des fourches sont des paradigmes mis en évidence par les premières techniques en molécule unique telles que l'autoradiographie de molécules d'ADN (J. A. HUBERMAN et A. D. RIGGS 1966) ou encore l'usage de la microscopie électronique (MCKNIGHT et MILLER 1977). Puis une nouvelle méthode a vu le jour en 1993 grâce à Parra et Windle : l'étalement des fibres d'ADN (ou "*DNA fiber spreading*") qui consiste à lyser des cellules sur une lame de verre en faisant usage d'un détergent et en laissant l'ADN s'écouler le long de la lame (PARRA et WINDLE 1993). Ainsi, par hybridation in situ en fluorescence (FISH), il a été possible de positionner des locus particuliers sur le génome. Cependant, effectuer des mesures précises sur ce type de données est complexe puisque l'étirement de ces molécules peut être très variable et les molécules ont tendance à s'agglomérer.

Quelques années plus tard, le peignage moléculaire (ou "*DNA combing*") (A. BENSIMON et al. 1994, D. BENSIMON et al. 1995) a été développé. Cette approche consiste à étirer physiquement et aligner les fibres d'ADN sur des lamelles hydrophobes. Les deux principaux avantages de cette méthode sont, d'une part, que les molécules d'ADN sont étirées uniformément permettant d'effectuer des mesures directes mais également reproductibles sur la région d'intérêt, et d'autre part un grand nombre de molécules peuvent être peignées sur une seule surface, garantissant ainsi un ensemble de mesures statistiquement représentatif (HERRICK et al. 2000). La reproductibilité et la précision de cette technique en font un outil puissant.

### 2.3.1.1 Les biais et limitations

- La méthode d'autoradiographie n'est pas évidente à mettre en oeuvre en raison :
- de la fragilité de la molécule d'ADN qui a tendance à se fragmenter au moment de la préparation,
  - la façon peu consistante dont les molécules d'ADN sont étalées à la surface des lames
  - du nombre insuffisant de molécules analysables

Il existe de nombreux biais qui ne sont pas propres à l'autoradiographie des fibres d'ADN et peuvent être observés avec d'autres approches en molécule unique. Les sites d'initiation situés à proximités les uns des autres sont négligés si les fourches de réplication associées à celles-ci fusionnent avant même d'avoir été marquées. Si les fourches de réplication fusionnent au cours du marquage, le signal résultant va être interprété comme étant une seule fourche de réplication. Il peut aussi arriver que certains signaux correspondant à de grands réplicons soient ignorés en raison, par exemple, d'une cassure de la molécule. Malgré l'existence de ces biais, les premières études faisant usage de la technique d'autoradiographie des fibres d'ADN ont constitué la base du paradigme concernant la réplication de l'ADN chez les métazoaires.

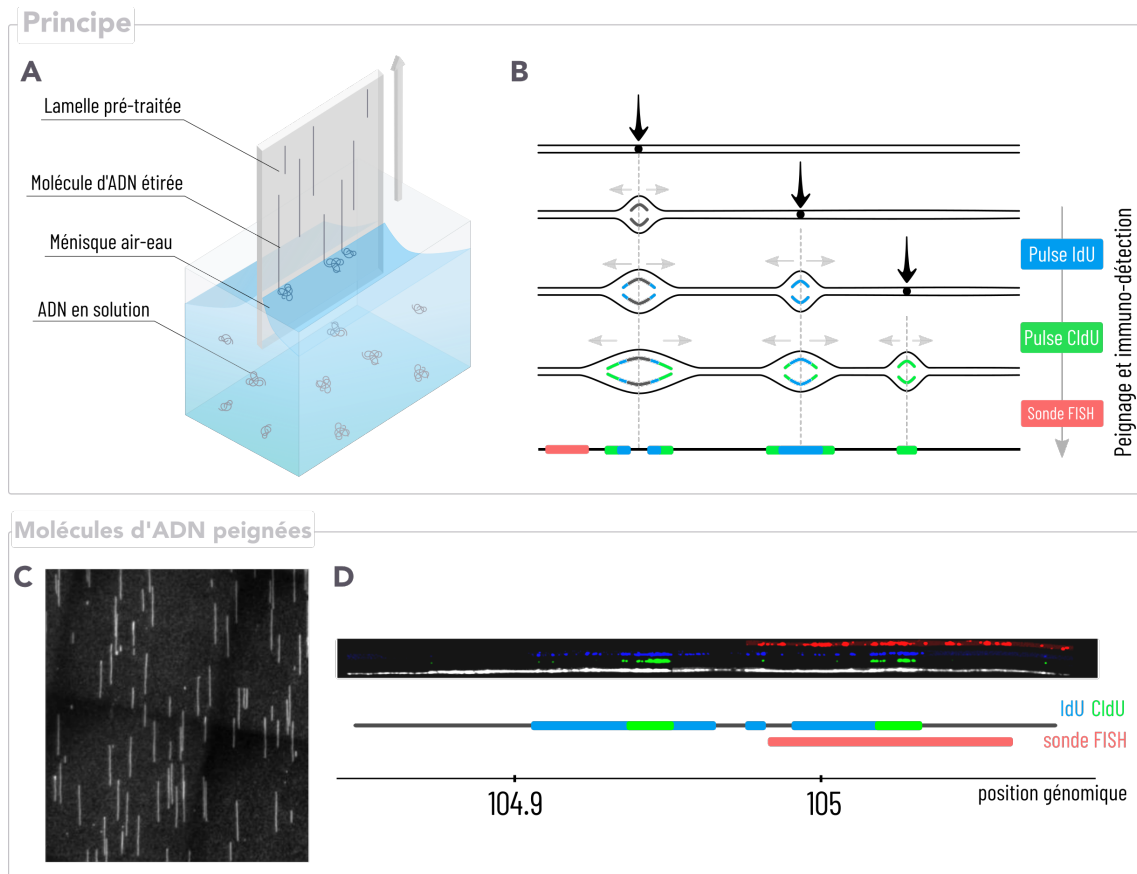
## 2.3.2 Le peignage moléculaire pour l'étude de la réplication de l'ADN

### 2.3.2.1 Le principe

La technique de peignage moléculaire a rapidement été adaptée à l'analyse de la réplication de l'ADN. Pour ce faire, les bulles de réplication sont marquées avec des analogues de la thymidine (digoxigenin-dUTP, biotin-dUTP) qui sont incorporés au niveau des segments d'ADN nouvellement synthétisés, dans des cellules en culture (HERRICK *et al.* 2000). Ces derniers peuvent être révélés grâce à une succession d'anticorps fluorescents tels que la streptavidine ou des anticorps anti-digoxigénine.

Cette approche a été adaptée aux cellules eucaryotes. Un marquage consécutif peut être réalisé avec deux analogues de la thymidine tels que l'iododéoxyuridine (IdU) et la chlorodéoxyuridine (CldU) qui ont la capacité à passer la membrane cytoplasmique des cellules eucaryotes contrairement à la digoxigenin-dUTP ou la biotin-dUTP (NORIO *et SCHILDKRAUT* 2001). Cela donne ainsi la faculté d'identifier et de distinguer les fourches de réplication, mais aussi les événements d'initiation ou de terminaison.

La vitesse de progression des fourches est aussi un paramètre pouvant être mesuré de façon précise (HERRICK et al. 2000). Des blocages éventuels de progression de fourches peuvent être détectés par une simple estimation du ratio IdU/CldU (TÉCHER et al. 2013). Un locus unique peut être étudié grâce à son identification par un code spécifique obtenu à l'aide de sondes révélés par FISH. En regroupant plusieurs fragments d'ADN, la dynamique de réplication à un locus donné peut être analysée. Ainsi, la technique de peignage moléculaire rend possible l'analyse de l'ensemble des molécules d'ADN présents sur une lamelle soit de façon anonyme afin d'obtenir des informations sur la réplication de l'ADN d'ordre plus général, soit les molécules sont identifiées à l'aide de sondes locus spécifique grâce à la technique de FISH. Un certain nombre de paramètres peuvent être extraits et analysés tels que la taille des bulles de réplication, la fraction d'ADN répliqué, les distances inter-origines et entre les bulles de réplication et enfin la densité des fourches de réplication et la vitesse de progression de ces dernières. Les molécules d'ADN peignées mesurent entre 200 et 700 kb de long (dans le meilleur des cas).



**Figure 2.7 – Principe du peignage moléculaire.** Une lame de verre pré-traitée est incubée dans une solution contenant les molécules d'ADN (A). Les extrémités libres des molécules d'ADN s'attachent à la surface et la lamelle est tirée vers le haut à vitesse constante. La force exercée par le ménisque permet d'étirer les molécules d'ADN uniformément. En (C) est présenté un exemple de résultat de peignage moléculaire avec de l'ADN du bactériophage  $\lambda$ . Il est possible de réaliser des pulses d'IdU (bleu) et de CldU (vert) et d'identifier le locus d'intérêt grâce à l'hybridation d'une sonde (rouge) (B). En (D) est présenté un exemple de molécule d'ADN humain (locus IGH) issu de cellules HeLa et marqué comme cela est expliqué en (B). En haut, les différentes couleurs de la molécule peignée sont séparées afin de faciliter la lecture. En dessous, est schématisé la molécule d'ADN peignée avec ses signaux ainsi que la sonde FISH permettant d'identifier le locus d'intérêt. Adaptée de [GUILBAUD et al. 2011](#).

### 2.3.2.2 Les origines et fourches de réplication : qu'avons nous appris avec les méthodes en molécule unique ?

La technique d'autoradiographie des fibres d'ADN a démontré que la réplication chez les eucaryotes était initiée au niveau de multiples origines de réplication. La conclusion principale tirée des études d'Huberman est relative à la vitesse de réplication des fourches. En effet, ils ont estimé une vitesse de progression des fourches de l'ordre de 2 à 3 kb / minutes (J. HUBERMAN et A. RIGGS 1968). D'autres études ont montré que la vitesse de progression des fourches chez les mammifères est comprise entre 0,6 et 3,6 kb par minute par la technique d'autoradiographie des fibres d'ADN et celle-ci a été confirmée par peignage moléculaire (Yury B. YUROV 1980, HYRIEN 2016). Une telle distribution de valeurs indique la présence de facteurs génétiques et/ou épigénétiques ayant justement une influence sur le mouvement des fourches de réplication. Les méthodes d'autoradiographie et de microscopie électronique ont aussi montré que la distance inter-origine changeait au cours de la différenciation cellulaire et du développement. En effet, des études ont montré que chez les embryons d'amphibiens ou encore de drosophiles la vitesse des fourches était plus faible et la distance inter-origine plus petite que dans les cellules somatiques (BLUMENTHAL, KRIEGSTEIN et HOGNESS 1974, CALLAN 1972). Cela nous indique donc l'existence d'une flexibilité importante de l'usage des origines de réplication et une réorganisation du programme de réplication en fonction du type tissulaire (D.D DUBEY et RAMAN 1987). En plus des difficultés techniques rencontrées pour l'analyse de la réplication chez les eucaryotes, se pose le problème de l'absence de dépendance de séquence pour ce qui est de l'activation des multiples origines de réplication rendant leur identification complexe. Grâce au peignage moléculaire, l'analyse du chromosome VI chez *S. cerevisiae* a montré que les origines de réplication, bien que déterminées par les ARS, sont déclenchées stochastiquement sans corrélation avec les origines adjacentes (CZAJKOWSKY et al. 2008).



### 2.3.2.3 Les limitations du peignage moléculaire.

Bien qu'utilisé en routine dans le domaine de la réplication, le peignage moléculaire présente un certain nombre de limitations (DE CARLI, GAGGIOLI et al. 2016). Le marquage des bulles de réplication est réalisé par incorporation d'analogues de la thymidine qui sont reconnus par des successions d'anticorps. Cela implique la formation d'un signal discontinu et difficile à analyser. En effet, dans le cas des molécules d'ADN avec des intermédiaires de réplication ayant été peignées et marquées, il est compliqué de savoir si un intervalle observé entre deux signaux réplcatifs correspond bien à un fragments d'ADN non répliqué ou bien, s'il s'agit en réalité d'un segment d'ADN répliqué (les deux signaux devant donc être fusionnés). Certains segments peuvent être faiblement marqué induisant un biais dans l'analyse des fragments d'ADN : ces segments peuvent être confondus avec des segments non répliqués.

Une seconde limitation de cette technique est l'usage de sonde FISH permettant l'identification des régions d'intérêts impliquant donc que près de 99% des molécules d'ADN peignées sont non identifiées sur une image.

Enfin, une dernière limitation va être l'automatisation de l'analyse des données en raison du bruit de fond présent dans les images mais aussi de la discontinuité du signal avec les problématiques présentées précédemment. Il faut donc compter plusieurs semaines, voir plusieurs mois afin de récolter et analyser un jeu de données statistiquement significatif.

### 2.3.2.4 Une automatisation de l'analyse des données de peignage : CASA, FiberStudio et IDeFlx.

Il existe à l'heure actuelle trois logiciels pour analyser les molécules d'ADN obtenus par peignage moléculaire. Ces derniers ne sont pas *open-source* et ne permettent donc pas d'apporter des modifications/améliorations pour d'autres types d'analyses.

## CASA : COMPUTER AIDED SCORING AND ANALYSIS.

Cet outil a été développé en 2011 par Wang et Chastain (WANG et al. 2011), permettant d'automatiquement détecter et mesurer des fibres d'ADN sur les images de peignage. CASA fonctionne sous le système d'exploitation Windows et fait usage de Matlab. Afin de détecter les molécules d'ADN à partir de l'image brute, une matrice de probabilité Hessienne est calculée. Puis, une approche de marche aléatoire est utilisée pour réaliser la détection des segments d'ADN (FRANGI et al. 1998, CHENG et al. 2009). Cependant les différents paramètres utilisés pour effectuer la détection sont définis pour une condition donnée faisant de cet outil, un logiciel semi-automatisé nécessitant une intervention humaine.

## IDeFIX

IDeFIX est un logiciel développé par la plateforme de l'Institut Génétique Moléculaire de Montpellier de T.Gostan et permettant l'analyse automatique des fibres d'ADN sur des images de peignage moléculaire. Les informations pouvant être extraites sont la vitesse individuelle des fourches, l'asymétrie de progression des fourches droite et gauche, les distances inter-origines ou encore la densité globale en fourche de réplication. Malheureusement ce logiciel est propriétaire et n'est pas disponible à la communauté scientifique.

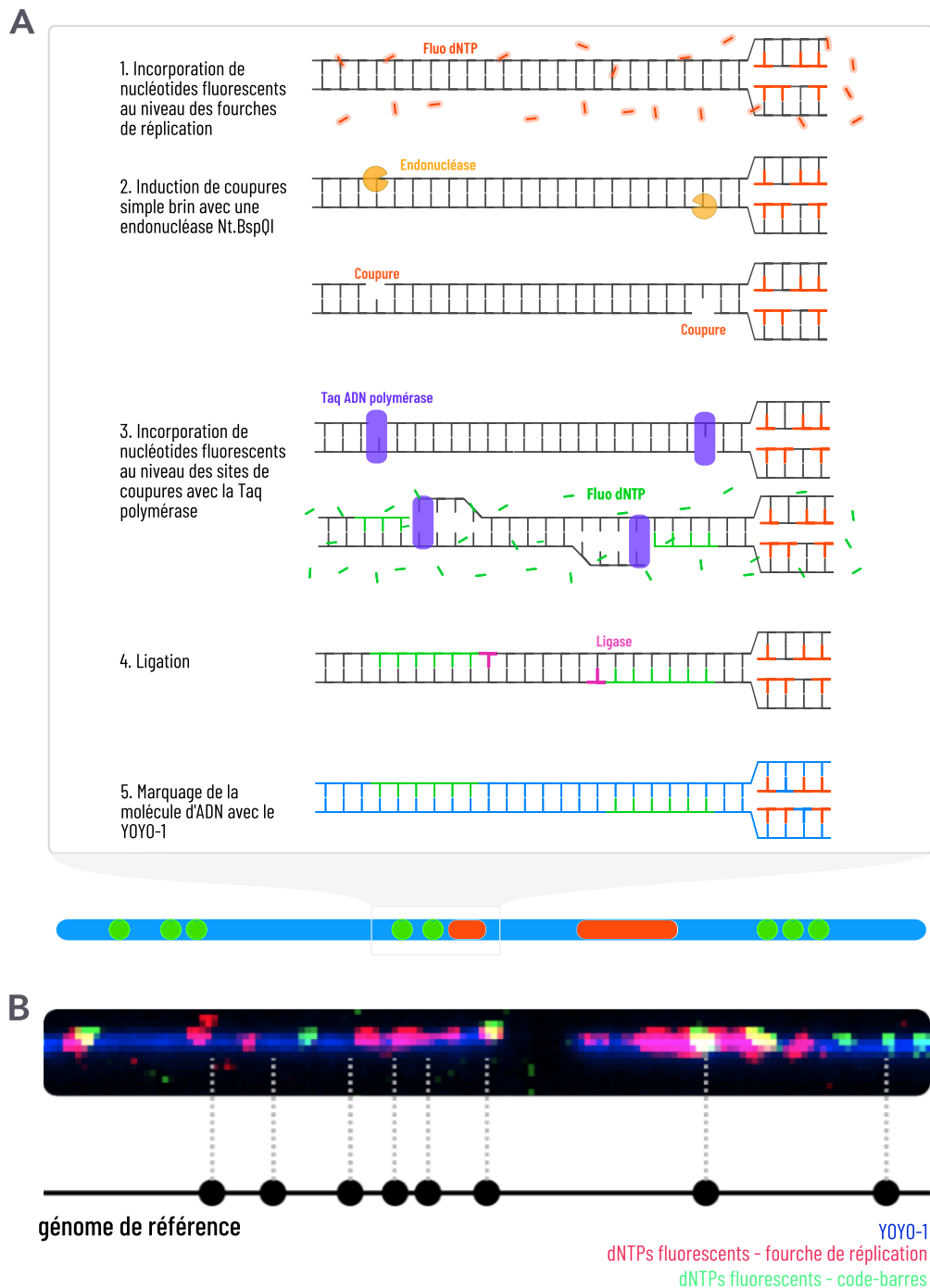
## FIBERSTUDIO DE GENOMIC VISION.

FiberStudio est un logiciel spécialement conçu pour l'analyse des résultats de peignage moléculaire et est développé par Genomic Vision. Il permet d'effectuer le même type d'analyse que CASA ou IDeFIX. Encore une fois, ce logiciel est propriétaire et n'est donc pas disponible à la communauté scientifique.

### 2.3.3 Vers du peignage moléculaire à haut débit : OMAR (*Optical MApping of Replicating DNA*).

Au sein du laboratoire, une méthode permettant le marquage de molécules d'ADN, OMAR (*Optical MApping of DNA Replication*) (DE CARLI, GAGGIOLI et al. 2016), a été élaborée et décrite en faisant usage d'un système *ex vivo*, les extraits d'œufs de Xénope dans lesquels a été répliqué de l'ADN du bactériophage  $\lambda$ .

Cette approche a permis de contourner certaines des limitations présentées précédemment. Il est désormais possible de marquer directement les segments répliqués en incorporant des analogues de la thymidine directement fluorescents dans les extraits d'œufs de Xénope, éliminant donc la révélation de ces segments avec des anticorps. Une autre amélioration non négligeable est l'identification de l'ensemble des molécules d'ADN présent sur une lamelle. Une endonucléase simple brin va permettre de réaliser une coupure de l'ADN au niveau d'une séquence spécifique comme décrit par Xiao et al. (XIAO et al. 2007). Ainsi, chaque molécule d'ADN va être identifiée par un code-barres autorisant sa cartographie sur un génome de référence (cf. Figure 2.3.3). En combinant ces deux améliorations, il est désormais possible de cartographier la réplication de l'ADN en molécule unique *genome-wide*. Cependant une des limitations n'est pas résolue : celle de l'automatisation de la détection des fibres d'ADN. En effet, la non linéarité des molécules, l'irrégularité des surfaces des lamelles utilisées pour réaliser le peignage moléculaire, mais aussi le bruit de fond présents dans les images générées sont réfractaires à une analyse automatique robuste.



**Figure 2.8 – OMAR, Optical Mapping of Replicating DNA.** (A) Schéma représentant les différentes étapes de marquage d'une molécule d'ADN. La première étape consiste à incorporer des nucléotides fluorescents au niveau des fourches de réplication (1). Pour obtenir le code-barres assurant l'identification de chacune de nos molécules, il faut réaliser un "nick labelling". Pour ce faire une endonucléase va reconnaître et effectuer une coupure au niveau de sites de restriction de 7 nucléotides sur un des deux brins de la molécule d'ADN (2). Grâce à l'action d'une TAQ polymérase, des nucléotides fluorescents sont incorporés aux sites de coupure (3) et va suivre une étape de ligation (4). Enfin, pour visualiser la molécule d'ADN dans son entièreté, un agent intercalant, le YOYO-1, va être utilisé(5). (B) : Exemple de molécule d'ADN ayant été marqué avec le YOYO-1 (bleu), les dNTP fluorescents pour la réplication (rouge) et le code-barres (vert). Grâce au code-barres la molécule peut être cartographiée sur un génome de référence.

### 2.3.4 L'apport des analyses genome-wide en molécule-unique.

Comme nous avons pu le voir, les techniques basées sur les populations et les techniques en molécule unique ne nous fournissent pas les mêmes informations sur le processus de réplication de l'ADN. Ainsi, ces deux approches sont complémentaires et permettent d'accéder à des paramètres différents. La plupart des méthodes populationnelles n'autorisent pas des quantifications précises de l'efficacité des origines de réplication. En effet, les techniques d'OK-seq, les techniques en molécule unique et le bubble-seq mettent en avant l'existence de zones d'initiations dont les origines sont multiples et peu efficaces alors que l'approche basée sur le séquençage des brins naissants met en évidence des zones étroites où les origines de réplication sont efficaces. Par ailleurs, les événements d'initiations, qu'ils soient spécifiques ou bien dispersés sur le génome, sont visualisables avec les techniques en molécule unique, contrairement aux techniques populationnelles cherchant les zones enrichies en intermédiaire de réplication, révélant ainsi uniquement les origines de réplication les plus efficaces. La vitesse de progression des fourches ne peut être mesurée par les techniques de SNS-seq ou de bubble-seq. Seule l'approche en molécule unique peut donner accès à une distribution précise de la vitesse de réplication. Un autre avantage majeur de l'analyse en molécule unique est la possibilité d'observer directement de multiples origines de réplifications sur une même molécule rendant possible l'étude de corrélations spatiales et temporelles entre les différentes origines de réplication.

## 2.4 L'objectif ultime : la cartographie pangénomique de la réplication de l'ADN en molécule unique avec une technique d'imagerie à haut débit.

Une compagnie américaine, Bionano Genomics<sup>®</sup>, a développé un dispositif, le système Irys, basé sur la technologie des nanocanaux permettant d'étirer des centaines de milliers de molécules d'ADN par électrophorèse et de les imager automatiquement. Les molécules d'ADN sont marquées à l'aide d'un agent intercalant, le YOYO-1, marquant la molécule d'ADN (bleu), et des nucléotides fluorescents sont incorporés au niveau des coupures simples brins provoquées par une endonucléase pour obtenir le code-barres (vert)(cf. Chapitre 1). Afin de visualiser des segments d'ADN répliqués, des nucléotides fluorescents sont incorporés lors du processus de réplication (rouge). La qualité des images acquises en sortie du système Irys est nettement supérieure qu'avec celle obtenue en peignage moléculaire : les images sont beaucoup moins bruitées, le ratio signal/bruit est meilleur et les molécules d'ADN marquées sont linéaires. Ces améliorations rendent l'automatisation de la détection des molécules et de l'analyse des données réalisables. Nous avons donc la capacité de cartographier la réplication de l'ADN sur l'ensemble du génome en molécule unique.

### 2.4.1 Problématique.

À partir des images originales générées par le système Irys, le logiciel Auto-Detect propriétaire de Bionano Genomics<sup>®</sup> extrait, entre autres, les coordonnées de début et de fin des molécules d'ADN dans chacune des images ainsi que les positions des code-barres (*nick-labels*) sur la molécule. Le code-barres permet la cartographie optique des molécule d'ADN par l'outil RefAligner également fourni par Bionano Genomics<sup>®</sup>. Le signal obtenu avec notre approche de marquage pour l'analyse de la réplication (rouge) n'est pas pris en charge par le système Irys et les outils bio-informatiques associés. L'analyse du canal de couleur rouge est habituellement le même que celui du canal de couleur vert à savoir la détection de points. Or, nous avons un signal répliatif discontinu et plus ou moins allongé. Étant donné que nous détournons le système de son usage initial, cela implique le développement d'algorithmes et d'outils spécifiques à nos données afin de permettre la cartographie de la réplication de l'ADN, *genome-wide*, en molécule unique.

## 2.4.2 Approche.

### 2.4.2.1 Les extraits d'œufs de Xénope et leur usage dans l'étude de la réplication de l'ADN.

Pour le développement de nos outils, nous avons fait le choix de travailler avec des extraits d'œufs de Xénope dans lesquels a été répliqué de l'ADN du bactériophage  $\lambda$ .

Le Xénope, amphibien dont l'espèce la plus connue est *Xenopus laevis*, est utilisé comme organisme modèle. Après la fertilisation des ovocytes de *Xenopus laevis*, une succession de 12 à 13 divisions cellulaires rapides et synchrones, où les phases M et S du cycle cellulaire s'alternent en absence des phase de "gap" G1 et G2 et des points de contrôle. Par ailleurs, cette phase de division présente très peu d'activité transcriptionnelle. Les embryons ainsi que les ovocytes de Xénope en raison de leur taille et de la possibilité d'en obtenir un grand nombre font de *Xenopus laevis* un excellent modèle d'étude dans des domaines allant de la biologie cellulaire à la biologie du développement.

Les extraits d'œufs de Xénope sont obtenus par broyage d'œufs matures de Xénope par centrifugation à faible vitesse appelée "*low speed supernatant*" (LSS). L'ajout de chromatine de spermatozoïde ou encore d'ADN de bactériophage  $\lambda$  au LSS va conduire à la formation d'une membrane nucléaire et la réplication semi-conservative de ces derniers. Les extraits d'œufs de Xénope constituent aujourd'hui l'un des seul système autorisant le processus de réplication de l'ADN de vertébrés *ex-vivo*, c'est à dire en dehors d'une cellule. Ce système permet d'imiter le processus de réplication qui se déroule *in vivo* chez les mammifères (J. Julian Blow et Ronald A. Laskey 1986).

*Pourquoi travailler sur un système ex-vivo ?* Les dNTP fluorescents utilisés dans la méthode OMAR, ne peuvent pas passer la membrane cellulaire. Si nous avions fait le choix de travailler sur le génome de la levure *S. cerevisiae*, plusieurs mois de mises au point de la méthode auraient été nécessaires.

Les différents génomes et échantillons sur lesquels nous avons travaillé seront présentés plus en détails dans le chapitre 4.

### 2.4.2.2 Développement d'un pipeline dédié à la cartographie de la réplication.

#### CORRECTION DES IMAGES BRUTES, EXTRACTION ET CARTOGRAPHIE DES PROFILS D'INTENSITÉ DES MOLÉCULES

Le but étant d'extraire les profils d'intensité des molécules d'ADN, de les cartographier pour ensuite effectuer la détection des segments répliqués, nous avons fait le choix de nous appuyer sur les fichiers de sortie du système Irys comprenant les images brutes avec les détections des molécules d'ADN et des code-barres d'une part, et d'autre part les fichiers liés à la cartographie optique permettant de localiser les molécules d'ADN sur un ADN de référence. Pour le développement de nos outils, deux étapes de *post-processing* majeures ont été nécessaires sur les images brutes : la correction de l'illumination (cf. Section 3.3.1.1), et le recalage des images permettant de superposer le signal de la réplication (rouge) et le code-barres (vert) sur la molécule d'ADN (bleu) (cf. Section 3.3.1.2). Une fois ces corrections faites, nous avons cartographié les profils d'intensité, étape qui a nécessité une analyse approfondie des fichiers de métadonnées et de l'ensemble des fichiers de sortie, afin de retrouver les différentes modifications à appliquer sur nos données (cf. Section 3.3.3).

#### DÉTECTION DES SEGMENTS D'ADN RÉPLIQUÉS.

Nous avons ensuite procédé au développement d'un algorithme permettant la détection des segments répliqués (rouge) sur les profils d'intensité (cf. Section 3.3.7). Suite aux observations faites sur nos données (DE CARLI, MENEZES et al. 2018), nous avons fait l'hypothèse que notre signal répliatif :

- est dépendant du contenu en nucléotides A-T : l'ajout de dUTP fluorescents en cours de réplication induit une intensité de fluorescence dépendante de la séquence de l'ADN.



- doit tenir compte de la Fonction d'Étalement du Point (FEP ou PSF - *Point Spread Function*). La diffraction de la lumière, qui détermine la limite de résolution d'un microscope, va rendre flou les objets ponctuels qui vont apparaître avec une taille et une forme minimale correspondant à la FEP. Dans notre cas, les molécules d'ADN observées sont en dessous de la limite de résolution du microscope et un pixel dans l'image va contenir environ 500 paires de bases (pb).
- du signal répliatif sous-jacent qui est proche d'un signal binaire : région de la molécule d'ADN répliquée ou non répliquée.

Après avoir testé différentes approches possibles, nous avons opté pour la reconstruction du signal de réplication par optimisation d'une fonction de coût en tenant compte des variables présentées ci-dessus et qu'il a été indispensable à déterminer (cf. Section 3.3.7). Une fois le développement des outils effectués, nous avons analysé le processus de réplication dans différents échantillons (cf. Chapitre 4).

## Un pipeline développé et dédié à la réplication de l'ADN : HOMaRD

---

<b>3.1</b>	<b>Méthodologie expérimentale et acquisition des données</b>	<b>68</b>
3.1.1	Préparation des échantillons pour l'analyse de la réplication de l'ADN	68
3.1.2	Irys System au service de la réplication de l'ADN : une méthode de peignage moléculaire automatisé	69
3.1.3	Comment analyser la réplication de l'ADN en utilisant les outils Bionano Genomics?	73
<b>3.2</b>	<b>Qualité des données en sortie du système Irys</b>	<b>75</b>
3.2.1	Contrôle qualité des données en sortie du logiciel Autodetect	75
3.2.2	La cartographie optique avec nos données	82
<b>3.3</b>	<b>Méthodologie développée pour l'analyse de la réplication de l'ADN à partir des données du système Irys</b>	<b>83</b>
3.3.1	Des étapes de <i>post-processing</i> indispensables	83
3.3.2	Extraction des profils d'intensité	94
3.3.3	Cartographie des profils d'intensité	94
3.3.4	Visualisation des profils d'intensité en molécule unique	95
3.3.5	Visualisation des profils d'intensité en population	97
3.3.6	Pipeline BionanoGenomics vs. pipeline HOMARD	97
3.3.7	La détection des segments répliqués	99

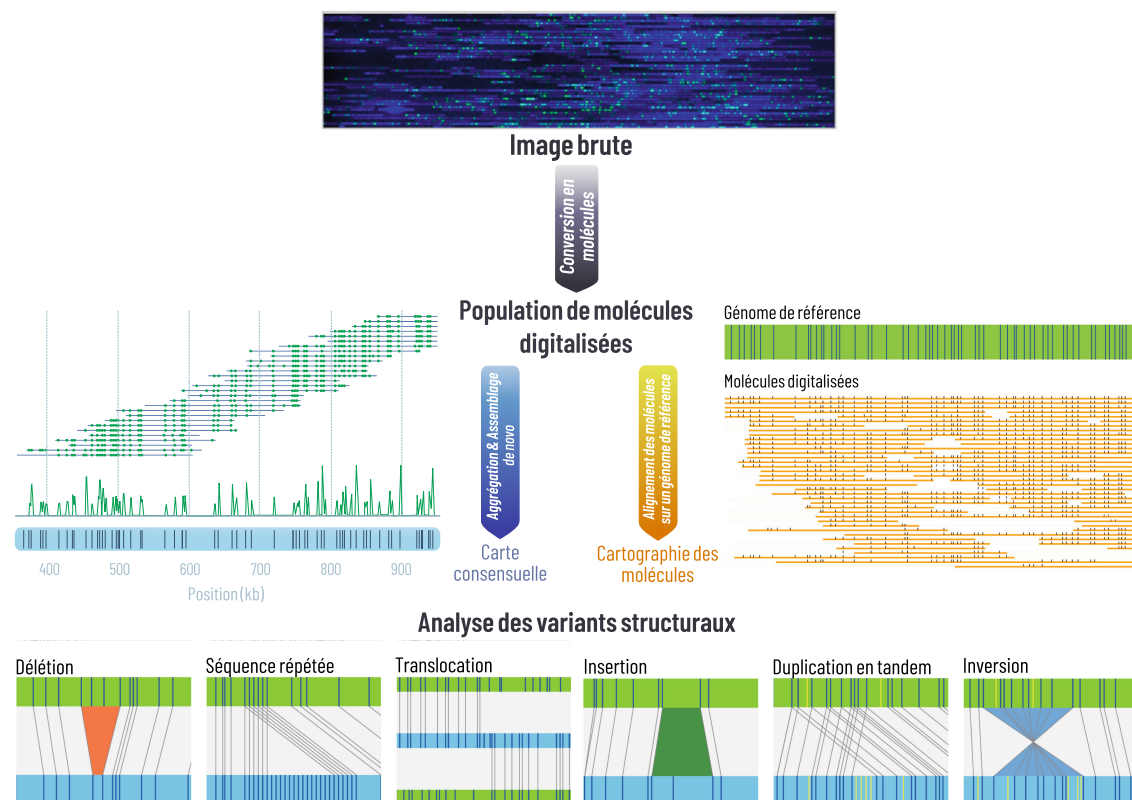
---



Le séquençage de lectures courtes pangénomique est aujourd'hui abordable et donc utilisable en routine. Cependant, il y a toujours des défis à relever parmi lesquels l'assemblage de génomes et la détection des variants structuraux (ALKAN, SAJJADIAN et EICHLER 2011, LEVY-SAKIN et EBENSTEIN 2013). Les génomes eucaryotes présentent un grand nombre de séquences répétées (environ 50% du génome humain). Les séquences à lecture courtes fournies par le séquençage NGS (*Next Generation Sequencing*) ne permettent pas une cartographie précise de ces répétitions. Les algorithmes d'alignement ne parviennent généralement pas à identifier l'emplacement génomique exact de ces séquences. Les variants structuraux, quant à eux, constituent la majeure partie des variations du génome. Les techniques NGS détectent de manière fiable les petites insertions ou délétions, mais n'ont qu'un pouvoir très limité lorsqu'il s'agit d'identifier ces variations sur plusieurs centaines de paires de bases, ou encore d'identifier les variations du nombre de copies (CNV, *Copy Number Variation*). La cartographie de génération suivante ou *Next Generation Mapping* (NGM) de Bionano Genomics® est la seule technologie permettant d'accéder rapidement à l'ensemble des types de variants structuraux (Hongzhi CAO et al. 2014, MAK et al. 2016). Cela est possible grâce aux fragments d'ADN ayant une longueur comprise entre 20 kb et 3 Gb contre quelques kilobases, dans le meilleur des cas, pour les techniques de séquençage NGS. Les technologies des nanopores, Pacbio et Illumina permettent d'obtenir, respectivement, des lectures de 2 Mb (PAYNE et al. 2018), 40 kb et 75 kb (SHELTON et al. 2015).

*Comment la technologie Irys fonctionne-t-elle ?* Les molécules d'ADN sont marquées, linéarisées et uniformément étirées dans les nanocanaux (DAS et al. 2010, V. MÜLLER et WESTERLUND 2017) pour ensuite être imagées automatiquement. Une endonucléase simple brin reconnaît et coupe une séquence de 7 paires de bases au niveau de laquelle vont être incorporés des nucléotides fluorescents (cf. Figure 2.3.3) permettant ainsi d'obtenir un code-barres environ 10 fois tous les 100 kb. Ces molécules sont ensuite alignées sur des cartes de référence d'un génome, ou bien permettent de construire ces cartes consensuelles (LAM et al. 2012, KRONENBERG et al. 2018). À partir des images brutes obtenues en sortie du système Irys, une première étape va consister à convertir ces images en molécules d'ADN conduisant donc à une digitalisation de l'information. Puis, ces mêmes molécules vont être agrégées et assemblées dans l'objectif de créer une carte consensuelle grâce à la technique de cartographie optique (alignement basé sur la distance séparant les points du code-barres). Le génome de référence va être digéré *in silico*, dans le cas d'un d'alignement des molécules digitalisées sur ce génome de référence. Ces cartes peuvent être créées à l'aide de différentes endonu-

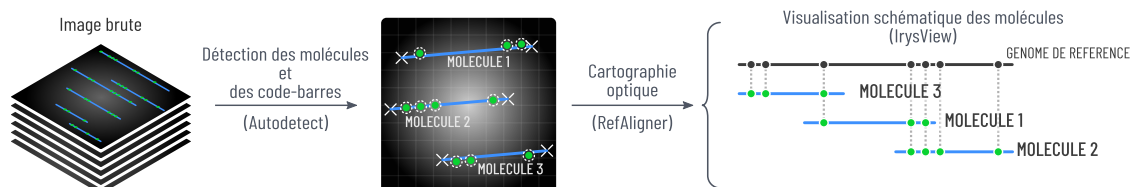
cléases pour obtenir une couverture plus large et une densité d'étiquetage plus élevée (HASTIE et al. 2013).



**Figure 3.1 – Des images brutes à l’analyse des variants structuraux.** À partir des images brutes générées par le système Irys (bleu : molécules d’ADN, vert : code-barres) des logiciels propriétaires (fournis par Bionano Genomics®) permettent leur conversion en molécules digitalisées. Les molécules une fois digitalisées vont soit permettre la réalisation d’un assemblage *de novo* (à gauche) soit vont être cartographiées sur un génome de référence (à droite). Des informations concernant les variants structuraux vont être retrouvées dans les fichiers de sortie produits par les logiciels propriétaires.

Le système Irys (<http://bionanogenomics.com>) est donc originellement conçu pour l’analyse de variants structuraux et la réalisation d’assemblage *de novo* grâce au long fragment d’ADN obtenus par cette technologie. En routine, les canaux de couleur utilisés avec le système Irys sont le canal bleu pour la visualisation des molécules d’ADN et le canal vert pour le code-barres. Le canal restant (rouge) ne sert qu’en cas de double marquage (*ibid.*) ou peut être utilisé à d’autres finalités telles que l’analyse de la méthylation de l’ADN (LEVY-SAKIN et EBENSTEIN 2013, MICHAELI et al. 2013, CHAN et al. 2018), les lésions de l’ADN (ZIRKIN et al. 2014) ou encore l’étude des empreintes génétiques des bactériophages (GRUNWALD et al. 2015). Nous avons fait le choix d’utiliser ce 3ème canal pour étudier la réplication de l’ADN. Il faut noter un point important, l’ensemble des outils fournis par Bionano Genomics®, à savoir AutoDetect, RefAligner et Irys View, sont propriétaires et constituent le pipeline pour le processus de cartographie (cf. Figure 3.2). Nous avons donc développé nos propres outils afin d’analyser notre

signal d'intérêt en se basant sur les fichiers de sortie des outils de Bionano Genomics®. Notre objectif étant de réaliser une analyse en molécule unique, nous avons extrait les profils d'intensité de chacune des molécules d'ADN détectées avec IrysView et cartographiées avec RefAligner pour ensuite détecter les tracts répliatifs sur le signal 1D extrait.



**Figure 3.2 – Pipeline d’analyse Bionano Genomics®.** À partir des images brutes obtenues en sortie du système Irys, un logiciel, Autodetect, va digitaliser les molécules imagées en détectant automatiquement des molécules d’ADN (bleu) ainsi que de leur code-barres (vert). Puis, à l’aide de RefAligner, ces molécules détectées vont être cartographiées. À droite est schématisé un alignement classique sur un génome de référence.

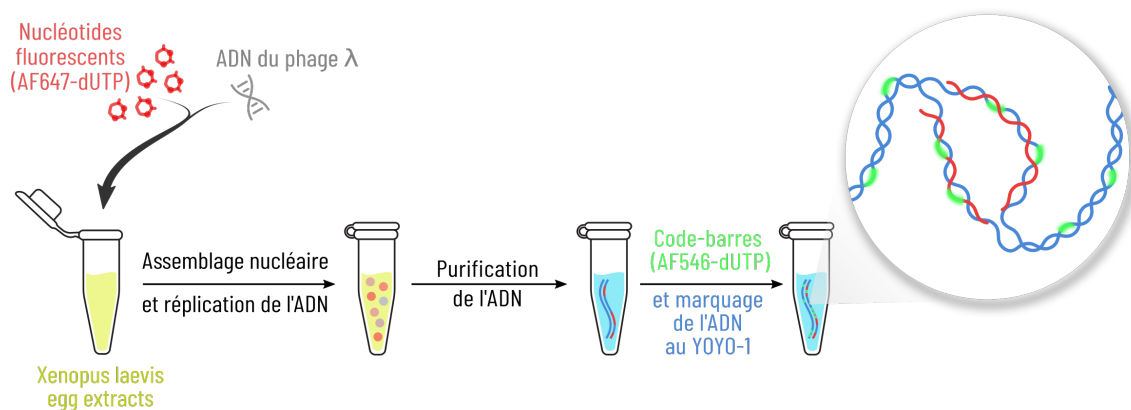
Dans cette partie nous allons décrire plus en détails comment nous sommes parvenus à partir des fichiers de sorties du système Irys, à développer un pipeline donnant la possibilité d’analyser la répliation de l’ADN en molécule unique.

## 3.1 Méthodologie expérimentale et acquisition des données

### 3.1.1 Préparation des échantillons pour l'analyse de la réplication de l'ADN

*La partie expérimentale a été réalisée par F. DE CARLI, biologiste du laboratoire d'O. HYRIEN.*

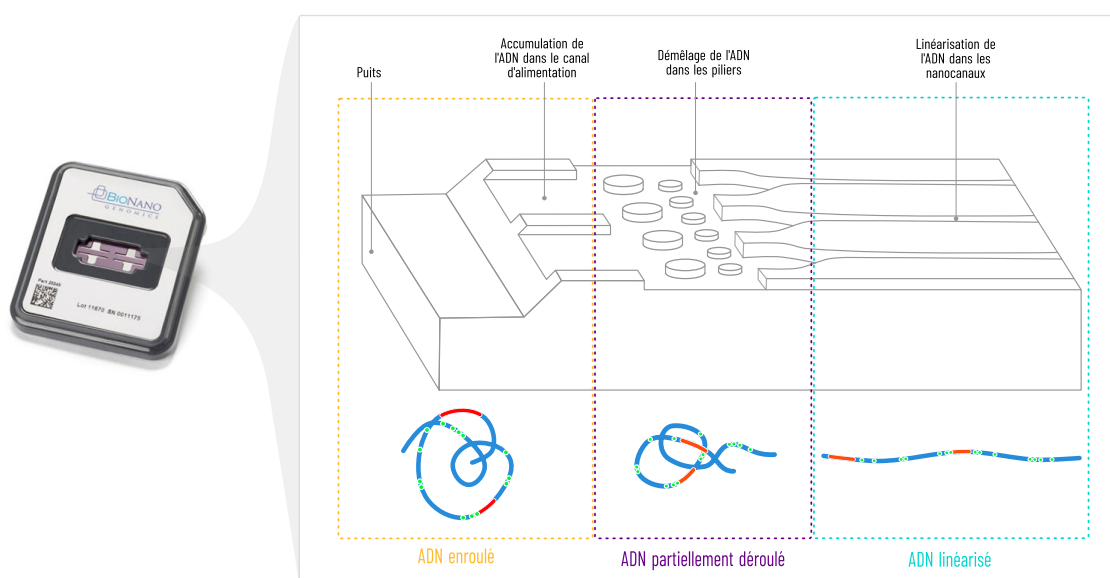
De Carli et *al.* décrivent les étapes de préparation des échantillons pour l'analyse de la réplication de l'ADN en molécule unique dans la méthode OMAR (Optical Mapping of DNA Replication) (DE CARLI, GAGGIOLI et al. 2016). OMAR a été développé pour le peignage moléculaire et nous avons souhaité étendre cette approche à la technologie des nanocanaux. Notre ADN d'intérêt, ici l'ADN de  $\lambda$ , (cf. Chapitre 4) est répliqué en présence de nucléotides fluorescents (AF647-dUTP, rouge) dans des extraits d'œufs de *Xénope*, avant l'étape de purification de l'ADN. Le code-barres est obtenu par incorporation de nucléotides fluorescents (AF546-dUTP, vert), à la suite de l'action d'une *nicking* endonucléase qui vient couper l'un des deux brins d'ADN au niveau d'une séquence de 7 paires de bases (sites Nt.BspQI). Enfin les fragments d'ADN sont marqués à l'aide d'un agent intercalant, le YOYO-1 (bleu) (cf. Figure 3.3)



**Figure 3.3 – Approche expérimentale.** L'ADN du bactériophage  $\lambda$  est répliqué durant 3 heures dans des extraits d'œufs de *Xenopus laevis* en présence de nucléotides fluorescents (AF647-dUTP, rouge). Une fois purifié, le code-barres est ajouté aux molécules d'ADN avec l'incorporation de nucléotides fluorescents (AF546-dUTP, vert) au niveau des sites de coupure et les molécules sont marquées à l'aide d'un agent intercalant, le YOYO-1.

### 3.1.2 Irys System au service de la réplication de l'ADN : une méthode de peignage moléculaire automatisé

L'échantillon d'ADN, répliqué et marqué, est déposé sur une puce du système Irys de Bionano Genomics®. Cette puce est constituée de 2 cellules de mesure (*flowcells*) comportant chacune près de 12 000 nanocanaux de 45 nm de diamètre (NAWY 2012). Dans la *flowcell*, les molécules d'ADN en solution véhiculées par électrophorèse dans les nanocanaux, vont être démêlées au niveau de la zone des piliers, passer dans les microcanaux avant d'entrer dans les nanocanaux (cf. Figure 3.4). Grâce à la combinaison des trois lasers, d'une caméra EM-CCD (réduction du bruit) et un système d'autofocus, le système Irys permet d'imager automatiquement et rapidement des centaines de milliers de molécules. La molécule d'ADN fait 2 nm de diamètre, il peut donc arriver que les molécules d'ADN se trouvent repliées ou forment un noeud. Il a été montré que ces événements étaient rares (7 % de la totalité du jeu de données) et ne constituaient pas un obstacle insurmontable en ce qui concerne la cartographie optique (REIFENBERGER, DORFMAN et Han CAO 2015). Par ailleurs, le code-barres identifiant la molécule se trouve modifié du fait d'un étirement anormal de cette dernière, rendant difficile voire impossible sa cartographie et ainsi permettant son exclusion du jeu de données final.

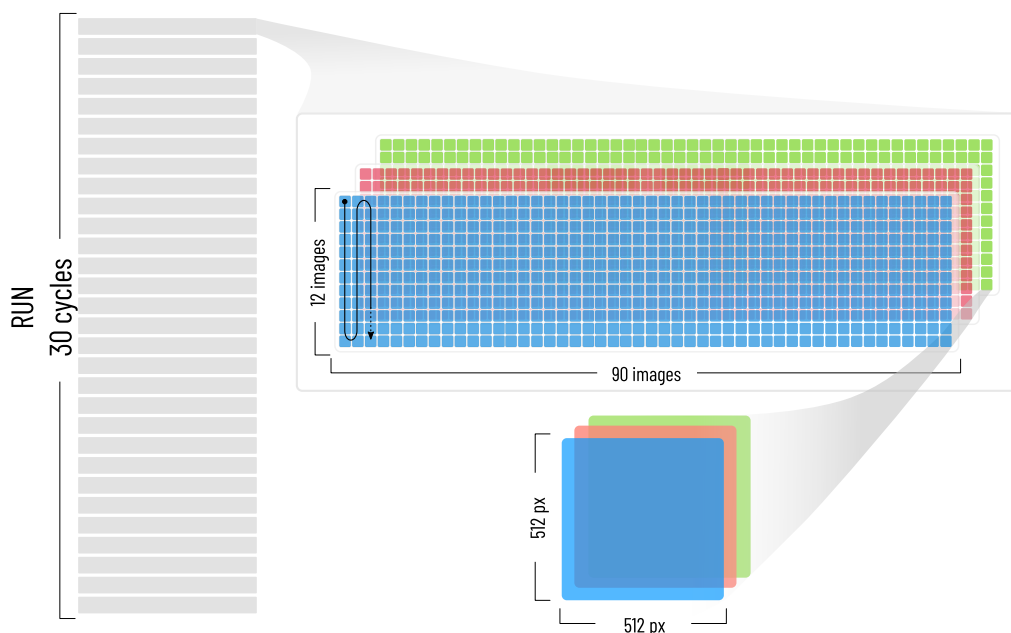


**Figure 3.4 – Puce du système Irys.** La puce est composée de deux cellules de mesure ou *flowcell*. Chacune d'elle est composée de près de 12 000 nanocanaux de 45 nm de diamètre. Les molécules d'ADN sont situées au niveau du puits. Lorsque l'électrophorèse est déclenchée, les molécules d'ADN marquées (YOYO-1 en bleu, réplication en rouge et code-barres en vert) arrivent dans une zone avec des piliers permettant le démêlage des molécules avant leur entrée dans les nanocanaux. Adaptée du site [www.bionanogenomics.com](http://www.bionanogenomics.com).



### 3.1.2.1 Comment les données sont-elles organisées ?

Un *run* (passage d'un échantillon dans la machine) est fait de 30 cycles (une électrophorèse correspondant à un cycle) et chacun d'eux est découpé en 1140 images (ou FOV pour *Field-Of-View*) de 512 x 512 pixels. Pour un jeu de données, il faut donc compter près de 34 200 images composées des 3 couleurs au format *Tagged Image File Format* (TIFF) (cf Figure 3.5). Une image couvre environ 140 nanocanaux et la longueur maximale d'une molécule d'ADN observable dans un unique champ est de 270 kb étirée à 85 % et un pixel va correspondre à environ 530 paires de bases. Les molécules les plus longues peuvent atteindre jusqu'à 3 Mb et s'étendre sur 12 images successives au maximum.



**Figure 3.5** – Organisation des images prises avec le système Irys. Un *run* est constitué de 30 cycles. Chaque de ces cycles correspond à une image TIFF composée de 1140 images (rouge, vert et bleu). Au moment de la prise d'images, la caméra prend 12 images verticalement et 90 images horizontalement et se déplace comme indiqué sur l'image (flèche noire). Les images générées font 512 x 512 pixels.

### 3.1.2.2 Des images brutes à la numérisation des molécules

Un outil propriétaire fourni par Bionano Genomics<sup>®</sup>, AutoDetect, donne la possibilité de détecter automatiquement les extrémités des molécules sur les images originales ainsi que la position des points constituant le code-barres. Chacune de ces molécules va être identifiée par un *Molecule Id* (identifiant unique) et consignée dans un tableau. À chaque cycle, en plus des images brutes, plusieurs fichiers par cycle sont générés dont 5 nécessaires au développement de nos outils (cf. Figure 3.6).

## FICHIERS .MOL

Les fichiers .mol contiennent l'ensemble des informations concernant les détections des molécules d'ADN faites sur les images brutes. Les paramètres qui nous sont indispensables pour la suite sont :

- l'identifiant de la molécule (*MoleculeId*) conservé dans l'ensemble des fichiers de sortie,
- le numéro du cycle où la molécule se trouve (*Scan*),
- les coordonnées de début et de fin des molécules (*XStart*, *YStart*, *XEnd* et *YEnd* ),
- les images sur lesquels les molécules commencent et se terminent (*FovStart*, *FovEnd*),
- les lignes (*RowStart*, *RowEnd*) et la colonne (*Column*) où ces molécules se trouvent (images 12 lignes x 90 colonnes)
- la longueur des molécules en pixels et en kilobases (*Length*, *LengthKb*)
- le chevauchement des molécules sur plusieurs champs ou non (*IsStitched*)

## FICHIERS .LAB

Les fichiers .lab sont ceux comprenant les données relatives aux code-barres et à leurs positions (*PositionOnMolecule*, *PositionOnMoleculeKb*) sur les molécules détectées. Les molécules identifiées par leur *MoleculeId* sont associées à leur code-barres dont les labels sont identifiées par des *LabelId*.

## FICHIERS .BNX

Les fichiers .bnx de Bionano Genomics® présentent une vue d'ensemble des données brutes sur les molécules et leur code-barres ainsi que les scores de qualité au cours d'une analyse (dans notre cas, uniquement le canal vert). Les données sont divisées en deux sections : l'en-tête décrivant le format spécifique des données, et le bloc d'information sur les molécules et les informations associées. Ce fichier combine les informations issues des fichiers .mol et .lab après avoir appliqué un filtre sur la qualité des sites marqués détectés.

### 3.1.2.3 Cartographie optique des molécules d'ADN

Le code-barres présent sur les molécules d'ADN rend possible leur alignement sur un ADN de référence par la technique de cartographie optique dont les pionniers sont Schwartz et son équipe (SCHWARTZ *et al.* 1993). À la différence du séquençage NGS où l'alignement sur le génome de référence se base sur la séquence elle-même, la cartographie optique est fondée sur les distances séparant les différents points du code-barres obtenu. Cette méthode produit des cartes physiques de motifs de séquences courtes (sites de reconnaissance d'enzymes de restriction) le long de fragments d'ADN d'une centaine à plusieurs milliers de kilobases. La cartographie optique, réalisée par l'outil RefAligner, fournit ainsi un outil à haut débit permettant d'ordonner et d'orienter les cartes physiques mais également de valider des assemblages génomiques. RefAligner va prendre en entrée les fichiers .bnx des molécules et .cmap du génome de référence. Une fois l'alignement des molécules réalisé, RefAligner va générer de nombreux fichiers dont les plus importants pour la suite du développement sont ceux présentés ci-après.

#### FICHIERS .CMAP

Le fichier CMAP de Bionano Genomics® est un fichier texte (délimité par des tabulations) de données brutes qui fournit les informations relatives à la localisation des sites marqués soit sur une carte de référence (*query* ou q.cmap) soit obtenus par digestion *in silico* d'une référence ou de données de séquence (*anchor* ou r.cmap). Il y a deux sections dans ce fichier : l'en-tête décrivant le format des données et les données d'alignement.

#### FICHIERS .XMAP

Il s'agit d'un fichier texte délimité par des tabulations correspondant à une comparaison croisée de l'alignement entre deux cartes : la référence qui a été digérée par l'endonucléase Nt.BspQI *in silico* (*anchor* ou r.cmap) et les code-barres des molécules (*query* ou q.cmap). Les données reportées dans le fichier renvoient aux coordonnées de début et de fin des code-barres des molécules ainsi que la position de chacun des sites du code-barres sur la carte de référence. Ce fichier contient deux sections : l'en-tête décrivant le format spécifique des données et le bloc d'information renseignant les alignements. Concernant l'alignement des molécules, en plus de leur

positionnement sur la carte de référence, nous avons l'orientation, une valeur de confiance et un pseudo-CIGAR qui va nous donner les insertions (I), les matches (M) et les délétions des sites marqués.

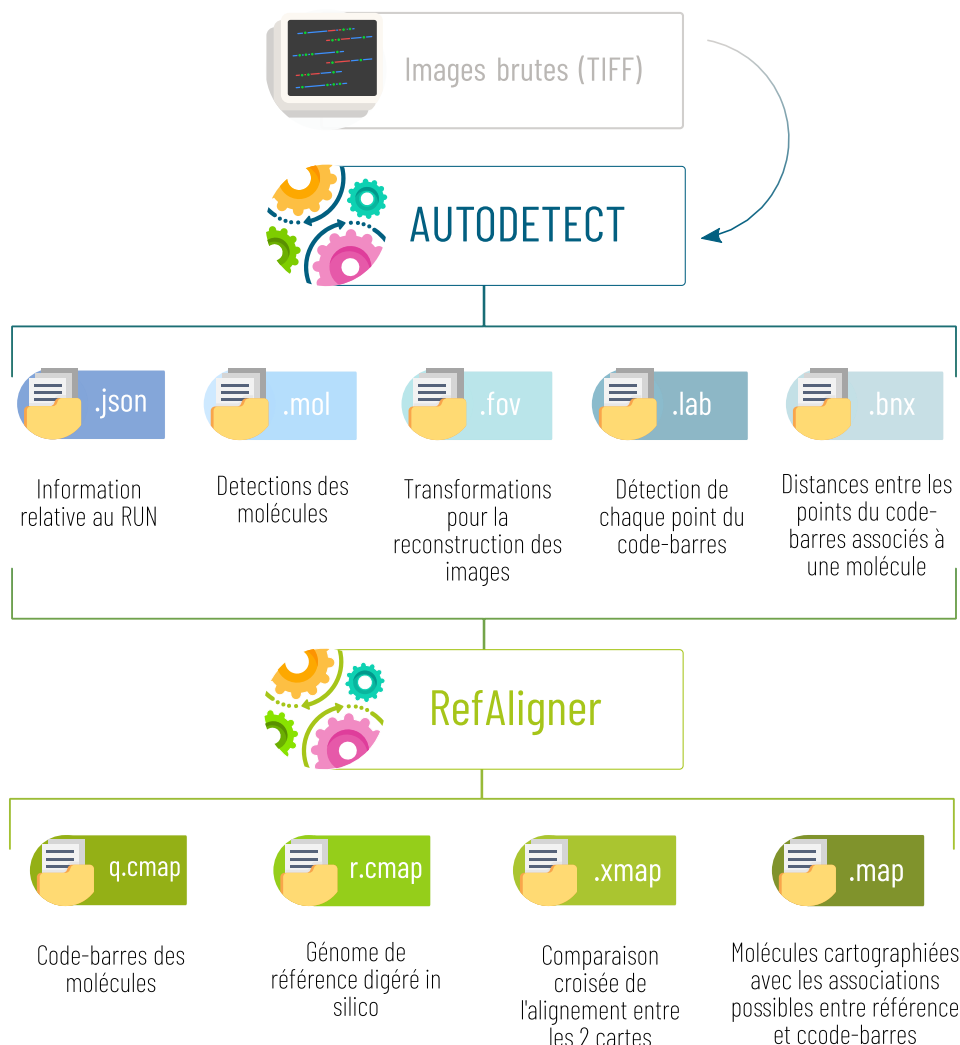
#### FICHIERS .MOLECULE QUALITY REPORT(MQR)

Ce fichier fournit un rapport sommaire sur la qualité des molécules et est généré d'après les résultats de l'alignement des molécules sur le génome de référence réalisé par l'outil RefAligner. Le MQR va identifier et produire le meilleur alignement de chaque molécule par rapport à la référence à condition que l'algorithme réponde aux critères minimaux de qualité de l'alignement.

### 3.1.3 Comment analyser la réplication de l'ADN en utilisant les outils Bionano Genomics?

Dans notre cas, nous avons dû apporter des modifications aussi bien au niveau des images qu'au niveau du fichier de métadonnées pour pouvoir utiliser les outils AutoDetect et RefAligner. En effet, il a d'abord fallu recréer des fichiers TIFF comprenant uniquement le signal YOYO-1 (canal de couleur bleue) et le code-barres (canal de couleur verte) dans cet ordre et donc supprimer le canal rouge du signal répliatif. Au niveau du fichier de métadonnées lu par AutoDetect, nous avons supprimé toutes les informations relatives à ce 3ème canal (rouge). De cette façon, la sortie d' AutoDetect ne nous donnera que les positions des code-barres et des extrémités des molécules d'ADN sans réaliser de détections sur le canal rouge, comme il est habituellement fait lors d'un double marquage.

*Pourquoi avons-nous eu besoin de faire cela sur notre jeu de données ?* Il est plus rapide de modifier les images et le fichier de métadonnées que de corriger manuellement les nombreux fichiers générés par AutoDetect contenant des centaines de milliers de lignes avec la possibilité d'introduire de multiples erreurs.



**Figure 3.6 – Fichiers de sortie des deux logiciels propriétaires de Bionano Genomics® : Autodetect et RefAligner.** Une fois les images brutes générées, Autodetect va se charger de la détection automatique des molécules d'ADN ainsi que de leur code-barres produisant ainsi un certain nombre de fichiers de sortie dont 5 fichiers qui nous sont utiles dans notre pipeline. Les fichiers produits par Autodetect vont être pris en entrée de RefAligner dans le but d'effectuer la cartographie optique. RefAligner va produire plusieurs fichiers de sorties dont 4 fichiers qui ont leur importance dans notre pipeline.

Il faut savoir que plusieurs mois de travail de "reverse-engineering" ont été nécessaires afin de comprendre le contenu des différents fichiers de sortie présentés ci-avant, la façon dont ils étaient générés et les liens entre les fichiers provenant d'Autodetect et de RefAligner. La documentation sur les fichiers de sortie proposée par Bionano Genomics® était beaucoup moins détaillée que celle d'aujourd'hui et aucune information n'était donnée sur les fichiers intermédiaires. Dans ce chapitre, n'ont été présentées que les informations essentielles et utiles à la compréhension du développement de nos outils bioinformatiques.

## 3.2 Qualité des données en sortie du système Irys

Comme cela a été souligné précédemment, nous avons pour objectif de créer notre propre pipeline d'analyse bioinformatique en nous appuyant sur les fichiers de sortie (images, détections et cartographie) du pipeline utilisé en routine par Bionano Genomics®. Avant d'aller plus loin, nous avons décidé de contrôler la qualité de nos données.

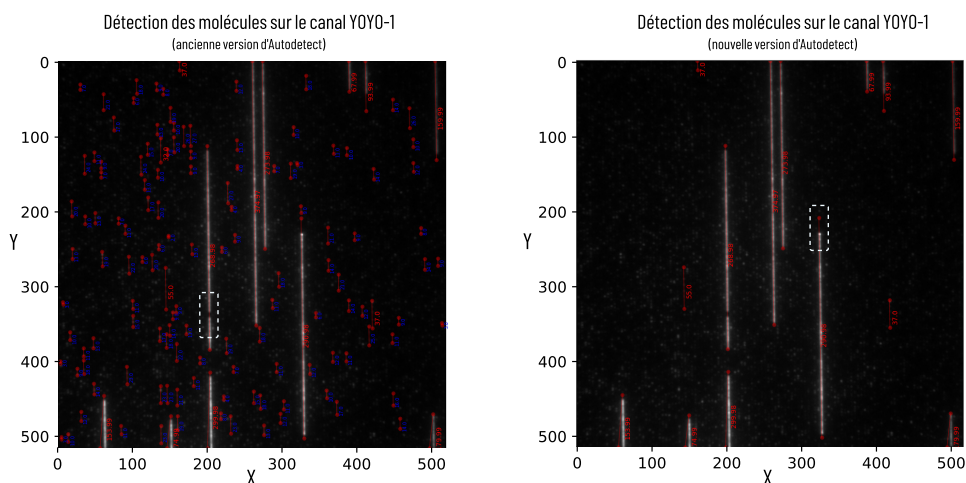
### 3.2.1 Contrôle qualité des données en sortie du logiciel Autodetect

Dans cette section, les quantifications sont réalisées sur les données d'ADN de bactériophage  $\lambda$  répliqué dans des extraits d'œufs de Xénope et marqué par la méthode décrite dans la Section 3.1.1. À l'issue du *run*, nous obtenons 61,1 Go de données incluant les fichiers textes (7.1 Go) ainsi que les images (54 Go). Ce jeu de données contient 903 891 molécules détectées. Ces dernières sont désignées à l'aide du *Molecule ID* qui sera conservé dans l'ensemble des fichiers en sortie d'Autodetect.

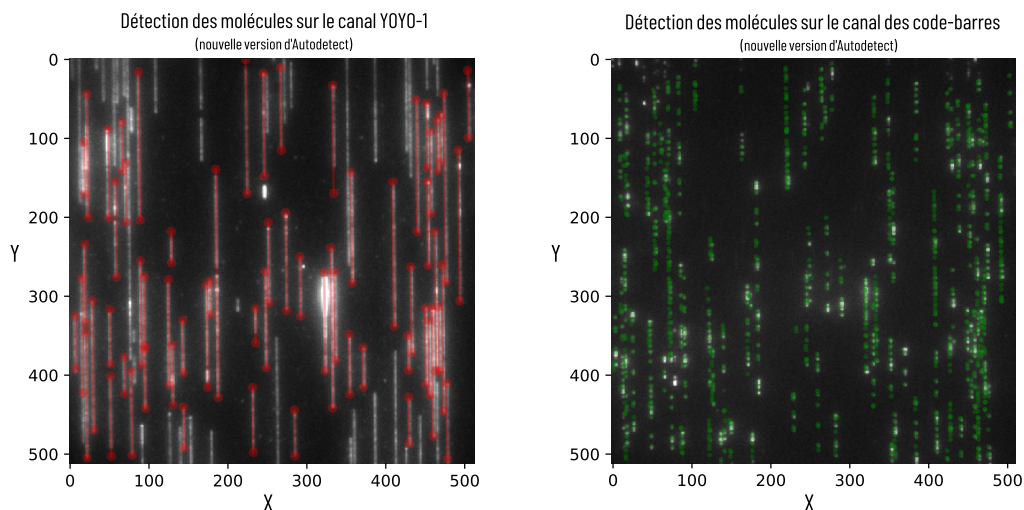
#### 3.2.1.1 Fichiers texte

LES MOLÉCULES D'ADN : DÉTECTION DES EXTRÉMITÉS, LEUR LINÉARITÉ DANS LES NANOCANAUX ET LES MOLÉCULES CHEVAUCHANTES

L'outil Autodetect va localiser les extrémités de chacune des molécules d'ADN (*XStart*, *YStart*, *XEnd* et *YEnd*). La première version de l'outil effectuait de fausses détections dans l'arrière-plan de nos images. Une localisation des extrémités des molécules plus précise est apparue avec une version plus avancée d'Autodetect. Malgré cela, il arrive que les détections ne soient pas extrêmement fiables : nous pouvons observer une sur-détection ou bien une sous détection au niveau des extrémités. Il arrive aussi qu'une unique molécule soit détectée alors qu'il y a plusieurs fragments d'ADN (cf. Figure 3.7).



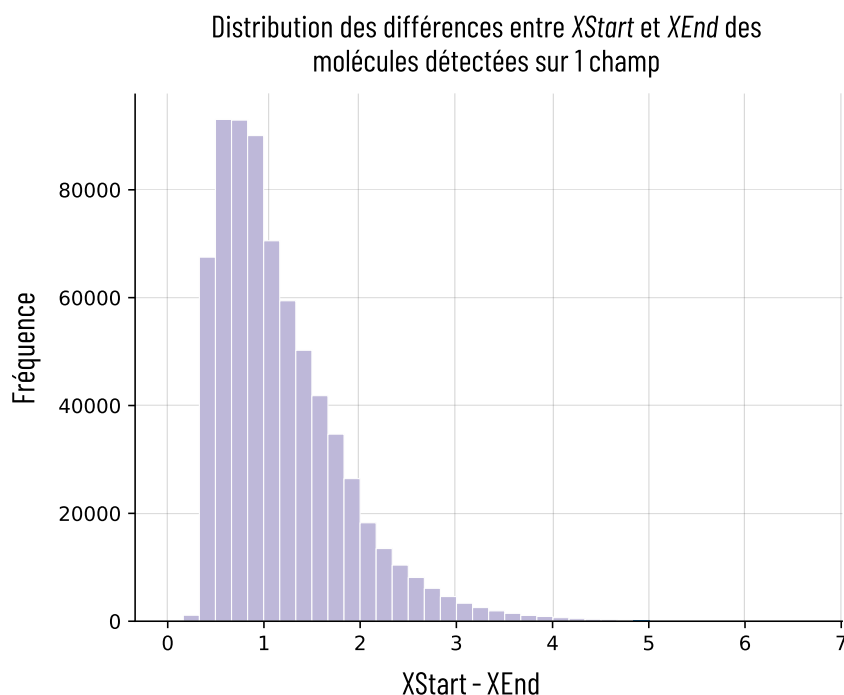
**Figure 3.7 – Comparaison des détections réalisées par les différentes versions d’Autodetect.** À gauche les détections faites par l’ancienne version d’Autodetect sur le canal bleu (YOYO-1) et à droite celles faites avec une nouvelle version du logiciel (traits rouges) sur la même image (niveau de gris). Dans la nouvelle version, les fragments faux positifs n’apparaissent plus. Aussi, il arrive que les extrémités des molécules d’ADN soient sur-détectées ou bien que des fragments soient considérés comme étant fusionnés (cadres pointillés). Les chiffres correspondent aux longueurs des molécules détectées (longueurs supérieures à 30px en rouge et longueurs inférieures à 30 px, en bleu).



**Figure 3.8 – Vérification des détections par Autodetect des molécules d’ADN et de leur code-barres.** Sur ces 2 images (512 x 512 px) sont représentées les détections faites pour les molécules d’ADN (à gauche) présent sur un seul champ et les code-barres associés (à droite). Les fragments de petites tailles (< 5 kb) ne semblent pas être détectés.

Nous avons également effectué une vérification visuelle de la détection des code-barres en utilisant le fichier .lab contenant l’ensemble des détections des labels pour le code-barres (cf. Figure 3.8). Au vu des résultats obtenus avec la nouvelle version d’Autodetect, nous avons considéré que ces détections étaient correctes et qu’aucune amélioration ne serait apportée sur la partie détections des molécules d’ADN et de leur code-barres.

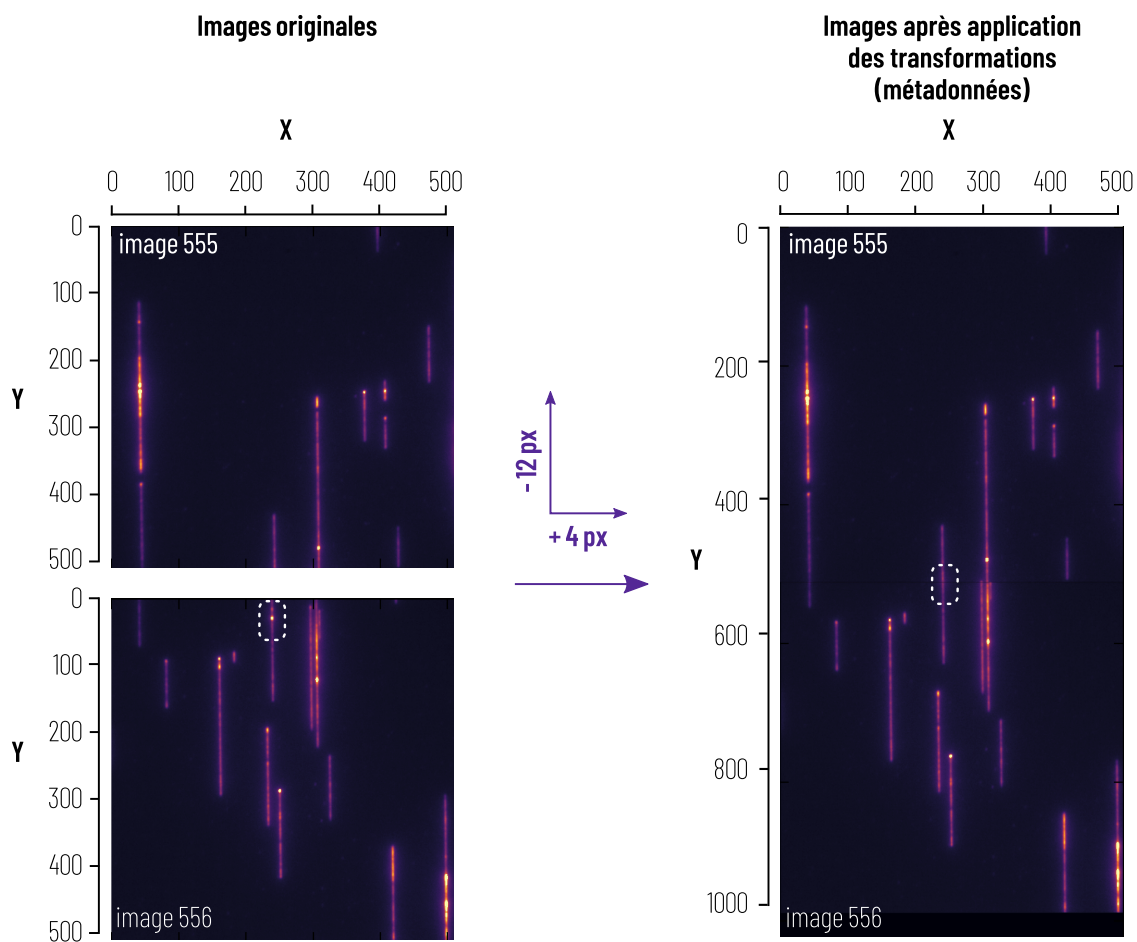
Dans le but d'extraire au mieux les profils d'intensité, nous avons évalué l'inclinaison des molécules d'ADN dans les nanocanaux. En effet, nous pourrions penser que ces dernières sont parfaitement rectilignes. Or, la différence entre le  $XStart$  et le  $XEnd$  de chacune de ces molécules nous donne une distribution de valeurs comprises entre 0 et 7 pixels pour les molécules d'ADN se trouvant dans un champ (cf. Figure 3.9). **Il sera indispensable de tenir compte de l'inclinaison des molécules d'ADN lors de l'extraction des profils d'intensité.**



**Figure 3.9 – Distribution des différences entre  $XStart$  et  $Xend$  des molécules (1 champ).** Au vu de cette distribution, les molécules sont inclinées dans les nanocanaux. Les différences sont comprises entre 0 et 7 pixels suggérant une possible erreur de détection dans le cas où nous avons des écarts supérieurs à 5 pixels (sachant qu'une molécule d'ADN fait entre 5 et 6 pixels de large et que cela ne concerne qu'un très faible nombre de molécules (environ 200 sur plus de 900 000 molécules)).

Dans le cas des molécules d'ADN chevauchant plusieurs champs, nous avons constaté la présence d'un décalage entre les différentes images. Parmi les fichiers de sortie, l'un d'eux semble donner les informations nécessaires à l'exécution du collage des images afin de récupérer l'information concernant les longues molécules. Lorsque nous appliquons les transformations données sur nos images nous ne parvenons pas à des résultats concluants (cf. Figure 3.10). **Nous avons donc fait le choix de mettre ce problème de côté dans la suite du projet et de développer nos outils sur les molécules présentes sur un seul champ.**

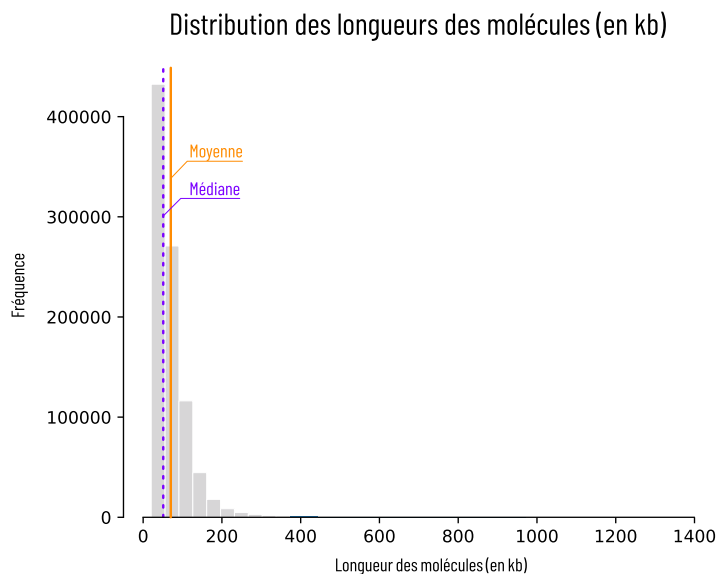




**Figure 3.10 – Assemblage de deux images avec des molécules chevauchantes.** Les transformations données dans un des fichiers de sortie d'Autodetect ont été appliquées à 2 des images du jeu de données. À gauche, les images originales successives faisant 512x512 pixels. Une fois les transformations appliquées à savoir une translation en X de 4 pixels et en Y de 12 pixels, nous obtenons l'image de droite. Le cadre pointillé met en avant le fait qu'avec une telle transformation nous perdons de l'information et que cette dernière est incorrecte.

#### AUTRES PARAMÈTRES

Les autres paramètres utiles dans les fichiers de sortie sont la longueur des molécules (en kilobases et en pixels), les images sur lesquelles commencent et se terminent les molécules détectées et le numéro du cycle correspondant. Par exemple, pour ce jeu de données, la longueur des molécules est comprise entre 0 et 1429 kb avec une médiane à 57.6 kb et une moyenne à 70.2 kb (cf. Figure 3.11). La grande majorité des molécules se trouve sur un champ (702 273 molécules, soit 77,7 %) et le reste des molécules se trouve sur au plus 6 champs (199 057 (22 %) sur 2 champs, 2 361 (0.3%) sur 3 champs, 17 sur 4 champs, 2 sur 5 champs et 1 sur 6 champs).



**Figure 3.11 – Distribution des longueurs de l'ensemble des molécules.** La grande majorité des molécules a une taille inférieure à 200 kb et nous avons une moyenne de taille de molécule de 70 kb et une médiane de  $\approx 57$  kb.

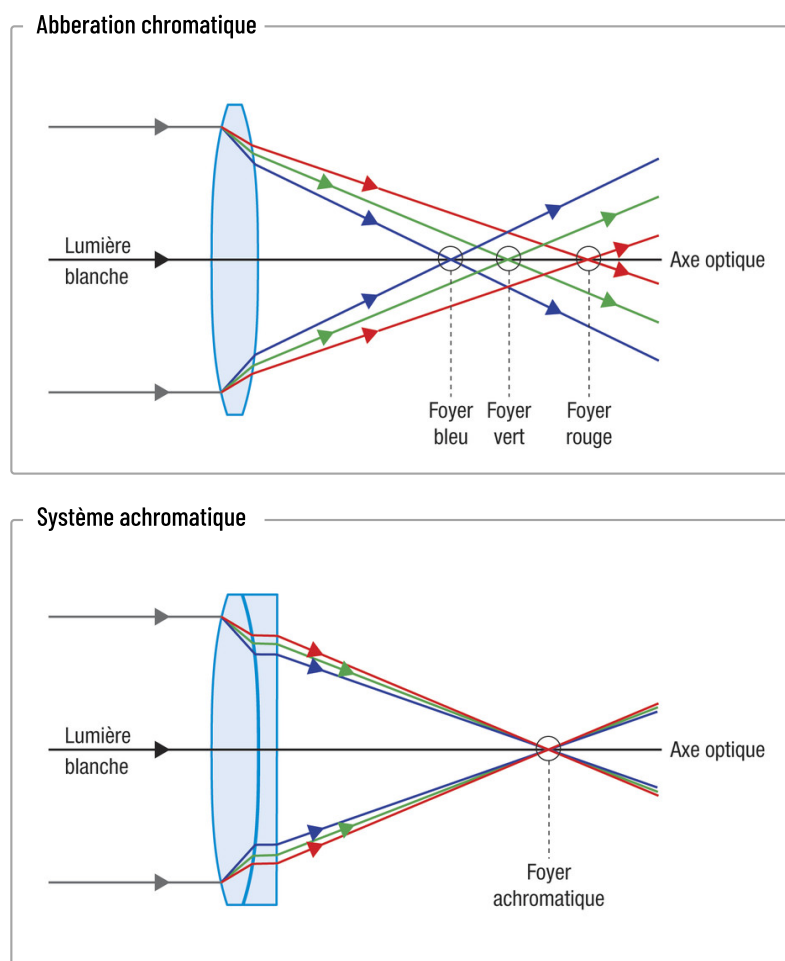
### 3.2.1.2 Les images TIFF

À l'issu d'un *run* près de 34 200 images de 512 x 512 pixels sont générées sur lesquelles nous avons pu constater quelques aberrations qui sont présentées dans cette section.

#### DÉCALAGE ENTRE LES DIFFÉRENTES COULEURS

En superposant les 3 signaux à savoir YOYO-1, code-barres et signal répliatif, nous remarquons un décalage entre les 3 couleurs. Il s'agit d'une aberration chromatique qui constitue l'un des principaux défauts des systèmes optiques. Ce sont des artefacts causés par la variation de l'indice de réfraction en fonction de la longueur d'onde. Ainsi, une lentille simple peut parfaitement focaliser pour une longueur d'onde donnée mais donnera un résultat flou pour les autres longueurs d'onde. En effet, les aberrations chromatiques résultent du fait que le verre des lentilles de microscopes présente des indices de réfractons variables suivant les longueurs d'onde de la lumière. Par conséquent, les longueurs d'onde courtes (bleu) vont focaliser près de l'arrière de l'objectif contrairement aux longueurs d'onde plus longues. Afin de contrôler ce défaut, il existe des systèmes achromatiques composés de deux lentilles de verres différentes et permettant la focalisation des différentes longueurs d'onde au même foyer (LoBIONDO, ABRAMOWITZ et FRIEDMAN 2011) (cf. Figure 3.12). Ces systèmes ne sont

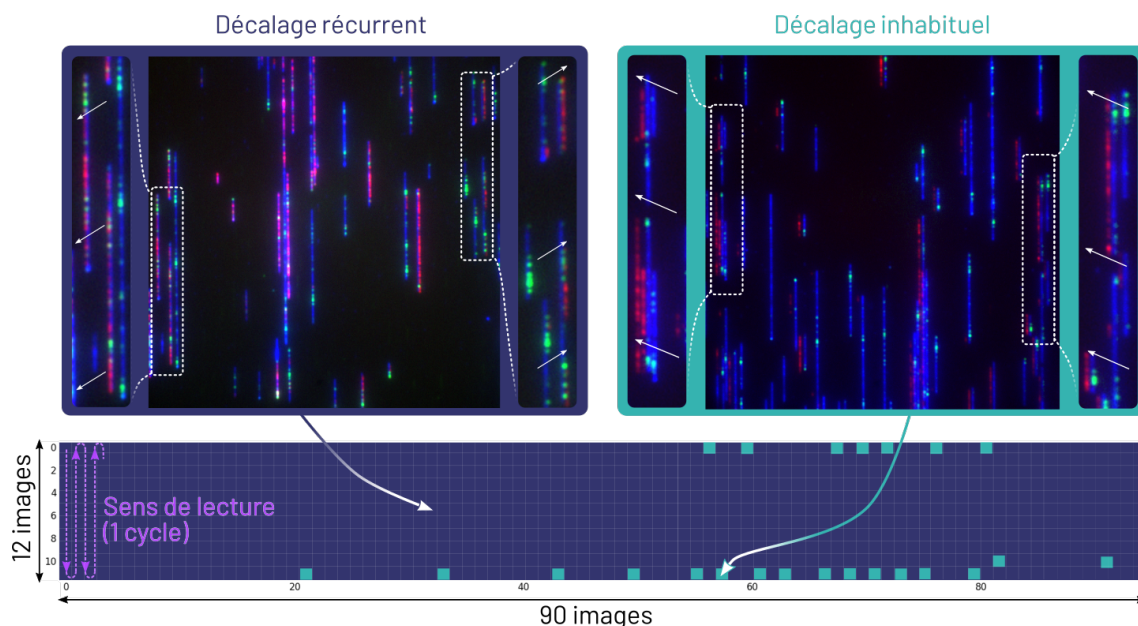
malheureusement pas utilisés dans le système Irys.



**Figure 3.12 – Aberration chromatique et moyen de correction.** Suivant la longueur d’onde nous avons une réfraction de la lumière variable induisant l’absence de convergence des faisceaux au niveau d’un unique foyer (schéma haut) : chaque couleur à son foyer. Une façon de corriger cela est l’usage d’un système achromatique permettant au faisceaux lumineux de focaliser au niveau du foyer achromatique (schéma bas). Cependant, ce système n’est pas implémenté dans le système Irys. Adaptée de Gilles Boisclair.

Le décalage observé est le même sur la grande majorité de nos images à savoir une mise à l’échelle et une translation en X et en Y. Cependant, certaines images d’un même cycle présentent un décalage différent. Après une annotation manuelle des images d’un cycle pris au hasard, il semblerait que cette anomalie n’apparaît qu’aux extrémités de la zone de prise d’images. De ce fait, nous pouvons penser que le système d’autofocus se trouve dérégulé dans ces régions au moment de la prise d’image (cf. Figure 3.13).

**Sachant que nous souhaitons utiliser les coordonnées des extrémités définies sur le signal YOYO-1 pour l’extraction des profils d’intensité du signal répliatif et du code-barres, une correction de ce décalage des couleurs est primordiale.**

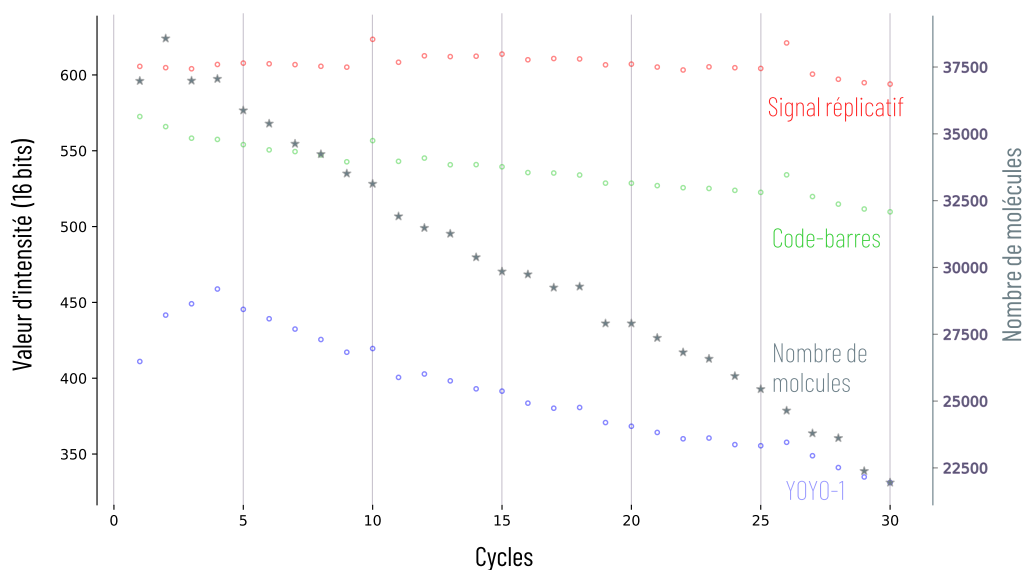


**Figure 3.13** – Décalage entre les signaux bleus, rouges et verts. Le décalage récurrent (en violet) est présent sur la quasi totalité d'un cycle. Cependant, il existe un décalage inhabituel où tous les signaux des molécules se décalent vers la gauche (en turquoise). La grille schématise les 1140 images d'un cycle et nous indique le sens de lecture de déplacement de la puce du système Irys au dessus de la caméra.

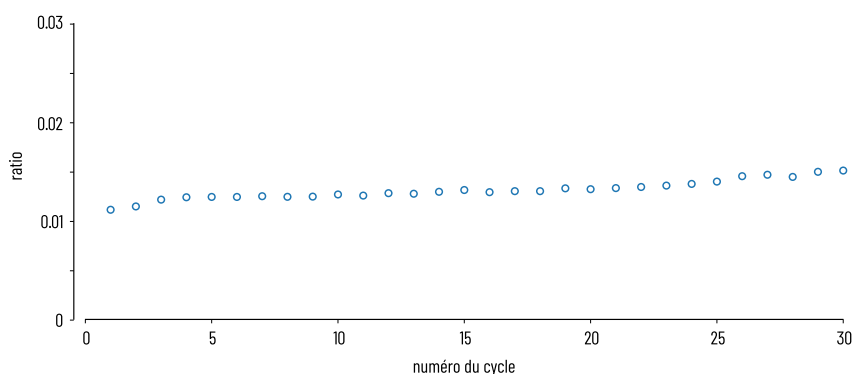
#### ILLUMINATION INHOMOGÈNE DE L'ARRIÈRE PLAN

Une image globale du *run* est délivrée à la fin de ce dernier, nous donnant ainsi un aperçu de la qualité générale du *run*. Le problème d'illumination étant récurrent sur les images prise en microscopie, nous avons fait le choix d'évaluer la qualité de l'arrière-plan des images. Nous avons calculé une image moyenne pour chaque cycle et ensuite déterminé la valeur moyenne de chacune de ces images (cf. Figure 3.14). Nous remarquons que cette valeur varie d'un cycle à un autre et cela peu importe le canal de couleur. Cependant, ces variations sont plus importantes sur le canal bleu (YOYO-1) avec des valeurs d'intensité moyennes comprises entre 331 et 459 et sur le canal vert (code-barres) avec des valeurs d'intensité moyennes comprises entre 510 et 573. La tendance observée pour le YOYO-1 (augmentation puis diminution des valeurs d'intensité) est sûrement liée à la quantité de molécules par cycle qui varie de la même façon.

**Nous avons donc prévu une étape de *post-processing* pour corriger au mieux cette illumination inhomogène.**



**Figure 3.14** – Variations de l'intensité moyenne des images médianes par cycle et par canal de couleur et quantification du nombre de molécules par cycle. Pour chaque cycle et chaque canal de couleur, la valeur moyenne de l'image médiane des 1140 images est calculée (axe de gauche). Le nombre de molécules d'ADN détectés à chaque cycle est également déterminé (axe de droite, étoiles grises). L'intensité moyenne pour les différents canaux de couleurs varie de la même façon que le nombre de molécules par cycle et plus particulièrement pour le canal bleu.



**Figure 3.15** – Ratio entre l'intensité moyenne des images médianes du signal YOYO-1 et le nombre de molécules par cycle. Nous pouvons remarquer que le ratio est quasiment constant au long du *run* impliquant donc que l'intensité moyenne des images médianes pour le signal YOYO-1 est proportionnelle au nombre de molécules dans un cycle donné.

### 3.2.2 La cartographie optique avec nos données

La cartographie des molécules d'ADN répliquées nous donne des résultats plus que satisfaisants. Sur les 903 891 molécules détectées nous avons conservé les molécules de plus de 90 kb et présents sur 1 champ, ce qui nous donne 120 793 molécules. Nous avons obtenu 69 203 molécules cartographiées c'est-à-dire 57,3 % des molécules après filtrage. **Nous avons un jeu de données suffisamment important pour développer notre pipeline d'analyse.**

### 3.3 Méthodologie développée pour l'analyse de la réplication de l'ADN à partir des données du système Irys

HOMARD, ou *High Throughput Mapping of Replicating DNA*, est le pipeline que nous avons conçu pour étudier les signaux réplcatifs obtenus pour divers échantillons. Dans le but d'effectuer une preuve de concept et s'assurer de la fiabilité de nos outils, nous avons fait le choix de travailler sur l'ADN du bactériophage  $\lambda$ . *Pourquoi ce choix ?* Il s'agit d'un génome de petite taille (48.5 kb) pouvant être répliqué dans les extraits d'œufs de Xénope. L'avantage de ces extraits est qu'ils fournissent un système *ex-vivo* permettant l'étude de divers processus parmi lesquels nous avons la réplication de l'ADN. De plus, nous avons une certaine connaissance de ce système (cf. Section 2.4.2.1), nous autorisant ainsi à comparer nos résultats avec la littérature.

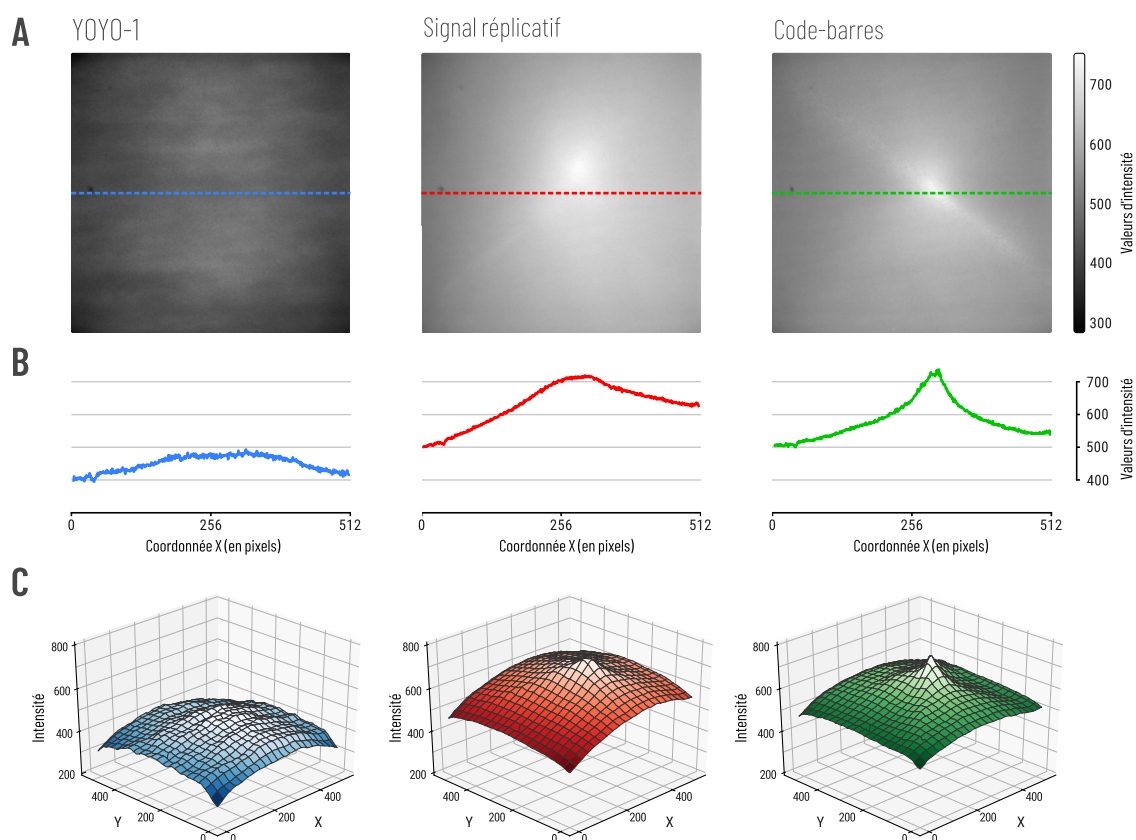
#### 3.3.1 Des étapes de *post-processing* indispensables

Après les observations faites sur nos données, deux corrections majeures doivent être apportées : la correction de l'inhomogénéité de l'illumination et le recalage des signaux réplcatifs (rouge) et du code-barres (vert) sur le signal YOYO-1 (bleu).

##### 3.3.1.1 Correction de l'illumination inhomogène sur nos images brutes

L'illumination inhomogène d'une image est tolérée dans les cas où nous souhaitons réaliser des analyses qualitatives. Cependant, lorsqu'il s'agit d'effectuer des analyses quantitatives, une correction de cette inhomogénéité est nécessaire. Dans le cas de la microscopie à fluorescence une approche appelée "*white referencing*" consiste à déterminer une image de référence. Pour cela, il faut générer une image blanche d'un échantillon uniformément fluorescent. Cette image blanche sera ensuite divisée ou bien soustraite des images acquises. Cette approche présente de nombreux inconvénients comme le fait que les conditions ne doivent pas changer au cours des acquisitions, ou encore que l'image blanche créée par l'utilisateur soit correcte. Il est possible de contourner ces problèmes en utilisant une approche "*data-driven*" (ou rétrospective), c'est-à-dire que la correction de l'illumination peut être réalisée à la suite de son acquisition et sur la base de l'information qu'elle contient. Par exemple, une image peut être floutée et soustraite à l'image d'origine pour corriger

l'illumination mais cela implique que nous supposons que le contenu de l'image est uniformément réparti. Ainsi, pour avoir une correction plus robuste, les méthodes rétrospectives utilisant de multiples images pour estimer une fonction d'illumination sont préférables. Il existe différentes approches pour ce qui est de la correction de l'illumination inhomogène mais elles ne semblent pas être adaptées pour les images obtenues par des techniques haut débit. Une approche assez directe consiste à calculer l'image moyenne de l'ensemble des images d'un jeu de données et d'appliquer un filtre médian (SINGH et al. 2014, JONES et al. 2006). Au vu de la quantité d'images et de la variabilité de la non-uniformité de l'illumination d'un cycle à un autre, nous nous sommes inspirés de l'approche décrite précédemment et avons fait le choix d'estimer une fonction d'illumination globale par cycle. Pour cela, nous avons calculé une image médiane par cycle pour chaque canal de couleur et soustrait cette fonction d'illumination à chacune des images originales TIFF (cf. Figure 3.16 et Figure 3.17).



**Figure 3.16 – Fonction d'illumination pour les 3 canaux de couleur du cycle 5 du jeu de données.**

Une image médiane est calculée à partir des 1140 images d'un cycle pour chaque couleur (A). Cette image médiane est par la suite soustraite de chacune des images du cycle en question. Nous pouvons voir que les variations d'intensité sont plus importantes pour les canaux rouge et vert (B et C).

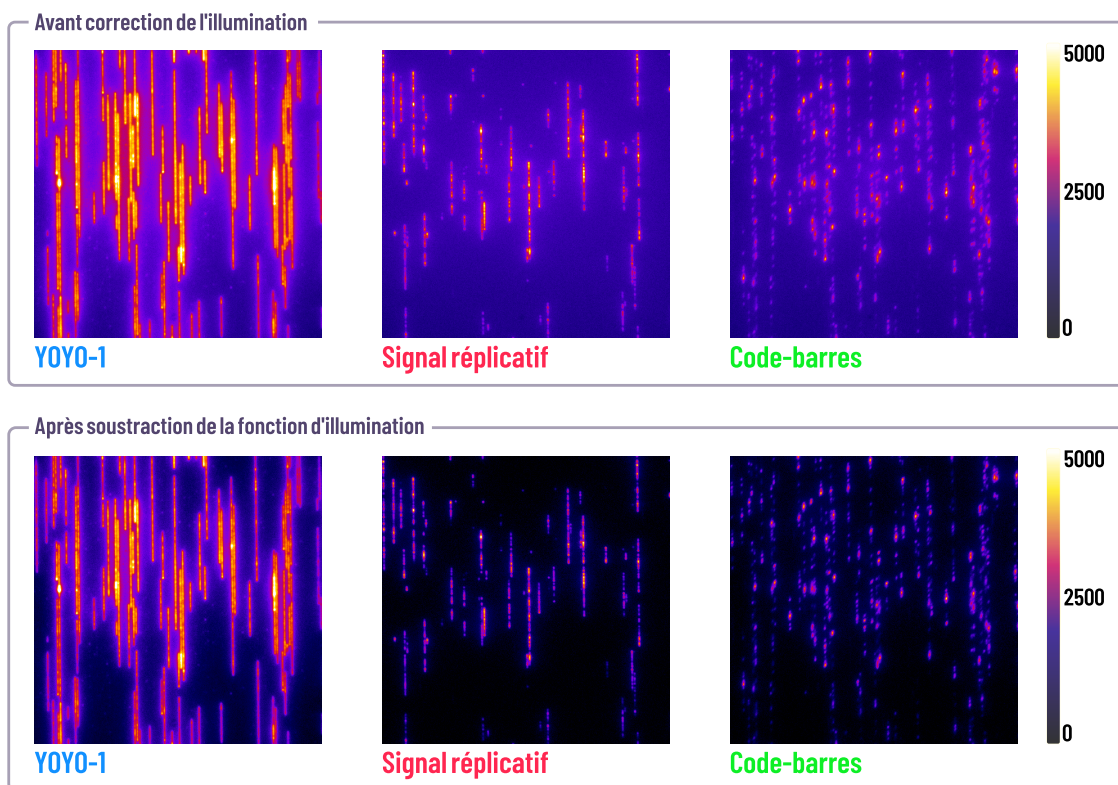


Figure 3.17 – Avant et après correction de l’inhomogénéité d’illumination pour les trois canaux de couleur. L’amélioration de l’arrière-plan des images est notable plus particulièrement pour les canaux du signal réplcatif et du code-barres par rapport au canal du YOYO-1.

### 3.3.1.2 Recalage d’images

L’objectif du recalage d’image est de superposer les pixels issus d’une image de référence sur ceux de l’image “cible” en les rapportant dans un même système de coordonnées. Nous pouvons retrouver cette méthode dans différents champs d’application tels que l’imagerie médicale (tomographie , IRM (Imagerie à Résonance Magnétique)... ) ou encore la vision par ordinateur. En raison des différents types de modifications possibles sur une image, il est impossible de concevoir une méthode universelle de recalage. Chacune des méthodes doit tenir compte, entre autres, des caractéristiques des données ou encore du bruit de fond. Néanmoins, les étapes classiques du recalage d’images sont les suivantes :

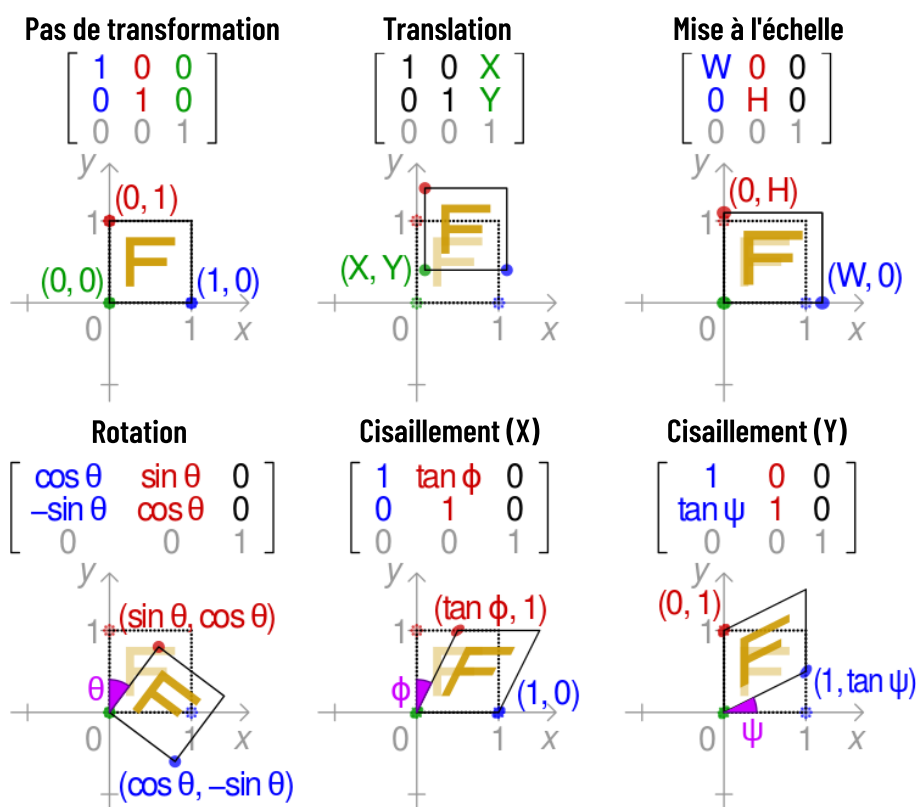
- l’extraction, de préférence automatique, des caractéristiques des objets d’intérêts tels que les contours, les intersections etc. . . .
- la mise en correspondance des caractéristiques c’est-à-dire qu’un lien est établie entre les caractéristiques détectées sur l’image de référence et l’image cible.
- l’estimation du modèle et donc des paramètres permettant la transformation.



- la transformation de l'image cible par le modèle déterminé à l'étape précédente.

La détection des extrémités réalisée par Autodetect sur le canal contenant les molécules d'ADN sera utilisée lors de l'extraction des profils d'intensité. C'est pourquoi nous devons réaligner les signaux rouges et verts sur le signal bleu. Notre cas est un peu particulier car les signaux à recaler sont 3 signaux de natures différentes : le signal du YOYO-1 est continu, le signal répliatif est discontinu ou peut être absent (molécules non répliquées à fortement répliquées avec la présence d'intermédiaires de réplication d'où cette discontinuité dans le type de signal) et le signal du code-barres qui est représenté par des taches circulaires. La matrice de transformation regroupant l'ensemble des paramètres que nous cherchons à déterminer est la matrice de transformation affine. La transformation affine est une transformation dans laquelle les rapports de distances et de colinéarité sont préservés. Si nous prenons comme exemple un point médian sur une ligne donnée, ce dernier restera le point médian de la ligne et tous les points situés sur une ligne droite resteront sur la ligne après une transformation affine appliquée. (Figure 3.18). Nous supposons que les transformations qui concernent nos images sont la translation (en X et en Y) et la mise à l'échelle.

**Notre objectif ici est de déterminer une matrice de transformation affine pour le signal répliatif et une seconde matrice pour le code-barres.** Notre stratégie va consister à minimiser la distance entre le signal bleu et rouge (et de même entre le signal bleu et vert). Dans un premier temps, nous allons associer les maxima locaux les plus proches entre les 2 signaux puis la somme de ces distances va être minimisée (fonction de coût) : à chaque itération, les paramètres de la matrice de transformation affine vont être réestimés et appliqués à l'image cible et ce jusqu'à ce que la fonction de coût atteigne un minimum.



**Figure 3.18 – Matrice de transformation affine.** La matrice est composée de 6 paramètres renvoyant à des transformations particulières : la translation, la mise à l'échelle, la rotation et enfin le cisaillement. Ceux que nous déterminons pour corriger nos images sont les paramètres de translation suivant les axes X et Y . Les paramètres de mise à l'échelle sont donnés dans les métadonnées en sortie d'Autodetect. Adaptée de <https://fr.wikipedia.org>.

## ÉTAPE 1 : DÉTERMINATION DES PAIRES DE POINTS ISSUES DU SIGNAL DE RÉFÉRENCE (YOYO-1) ET DU SIGNAL À RECALER (SIGNAL RÉPLICATIF)

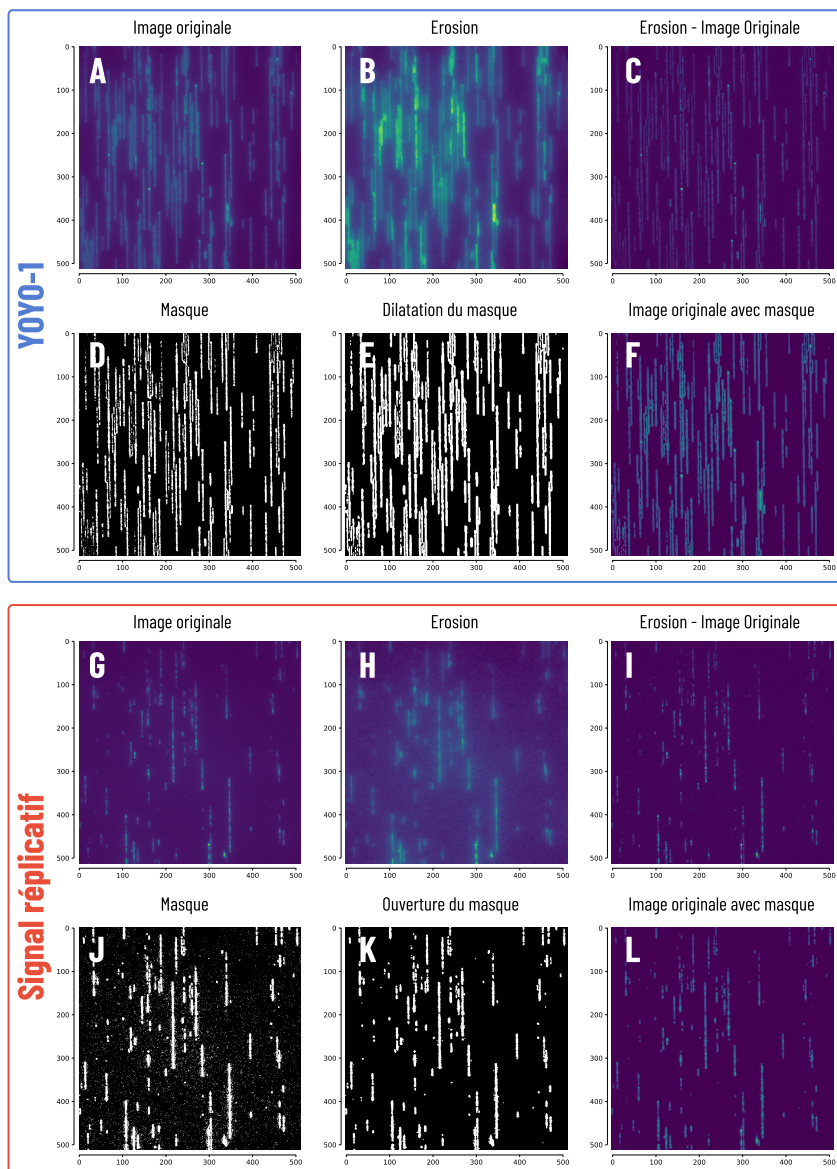
*Les explications ci-dessous sont données pour le recalage du signal rouge sur le signal bleu mais peuvent être étendues au recalage du signal vert sur le bleu.*

Dans un premier temps, nous avons déterminé les maxima locaux à conserver dans chacune de nos images d'intérêt. Une fois définis, des paires sont créées : un maximum pris dans un champs (FOV, *field-of-view*) du signal bleu sera associé au maximum le plus proche du même FOV pour le signal rouge.

Pour ce faire, nous utilisons la morphologie mathématique qui originellement provient de l'étude de la géométrie des milieux poreux dans les années soixante. En effet, les milieux poreux sont considérés comme étant binaires puisqu'un point appartient soit au pore lui-même ou bien au milieu entourant ce dernier. Cela a donc amené G.Matheron et J.Serra à formaliser cette approche pour l'analyse d'images binaires (MATHERON 1967, J. P. SERRA et JEAN 1982, J. SERRA 2010, H. TALBOT 2002). En traitement d'images, la morphologie mathématique est utilisée pour modifier une image par un élément structurant choisi, en appliquant des opérations de base telles que l'érosion ou la dilatation. Cette approche permet, entre autres, de réaliser du filtrage, c'est-à-dire supprimer ou bien conserver des éléments dans une image ayant des caractéristiques spécifiques telles que la forme, ou encore d'effectuer de la segmentation d'images.

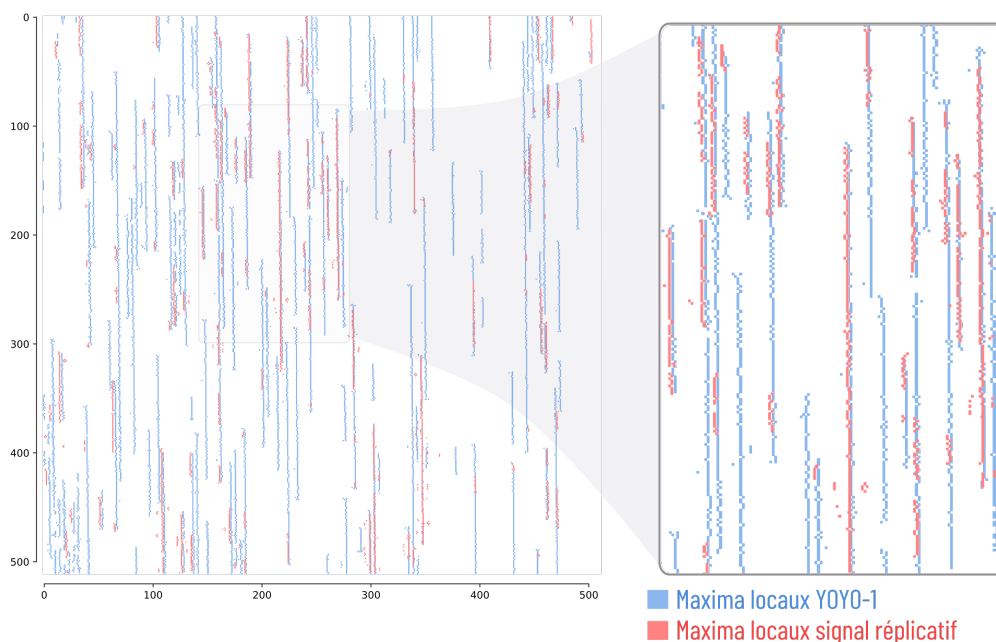
Dans notre cas, nous faisons usage de la morphologie mathématique afin de créer un masque pour chacune des images analysées afin d'exclure le bruit de fond autorisant ainsi une détection des maxima locaux plus précise au sein de nos signaux d'intérêt . Pour le canal bleu, nous réalisons une érosion de l'image en niveau de gris avec un élément structurant rectangulaire (dimension 4 x 5, composé de 1) et soustrayons de cette image, l'image originale, afin d'éliminer les halos autour des molécules. L'image résultante est ensuite binarisée permettant ainsi d'avoir un premier masque de notre signal. En regardant de près, nous pouvons voir que les masques des molécules sont fragmentés. Pour remédier à cela, une dilatation avec un élément structurant rectangulaire (dimension 5 x 2, composé de 1) est faite. Une fois le masque binaire final obtenu, il est appliqué à l'image originale par une simple multiplication de l'un par l'autre : les valeurs de pixels du bruit de fond sont maintenant nulles. À partir de cette dernière image, les maxima locaux sont détectés. Pour le canal rouge et vert, à la suite de la binarisation de l'image, au lieu d'appliquer une dilatation nous procédons à une ouverture. L'ouverture consiste en

une érosion suivie d'une dilatation. Ainsi les éléments de petite taille sont supprimés (bruit de fond) et les fragments du masque sont fusionnés.



**Figure 3.19** – Post-processing des canaux bleu (YOYO-1) et rouge (signal réplcatif). Dans chacun des cas, les images originales (A, G) se voient soustraire leur image après une étape d'érosion à l'aide d'un élément structurant rectangulaire (B, H), permettant ainsi d'éliminer le halo lumineux se trouvant autour des molécules (C, I). Les images obtenues sont ensuite binarisées (D, J). Le bruit de fond persistant et la fragmentation des masques des molécules nous ont amené à réaliser une dilatation dans le cas du signal YOYO-1 (E) et une ouverture dans le cas du signal réplcatif (K) (l'ouverture élimine les petits éléments présent dans l'arrière-plan). Enfin, nous multiplions les images originales par les masques respectifs (F, L). (cf. Figure 6.1 en annexe pour la même analyse sur le signal du code-barre)

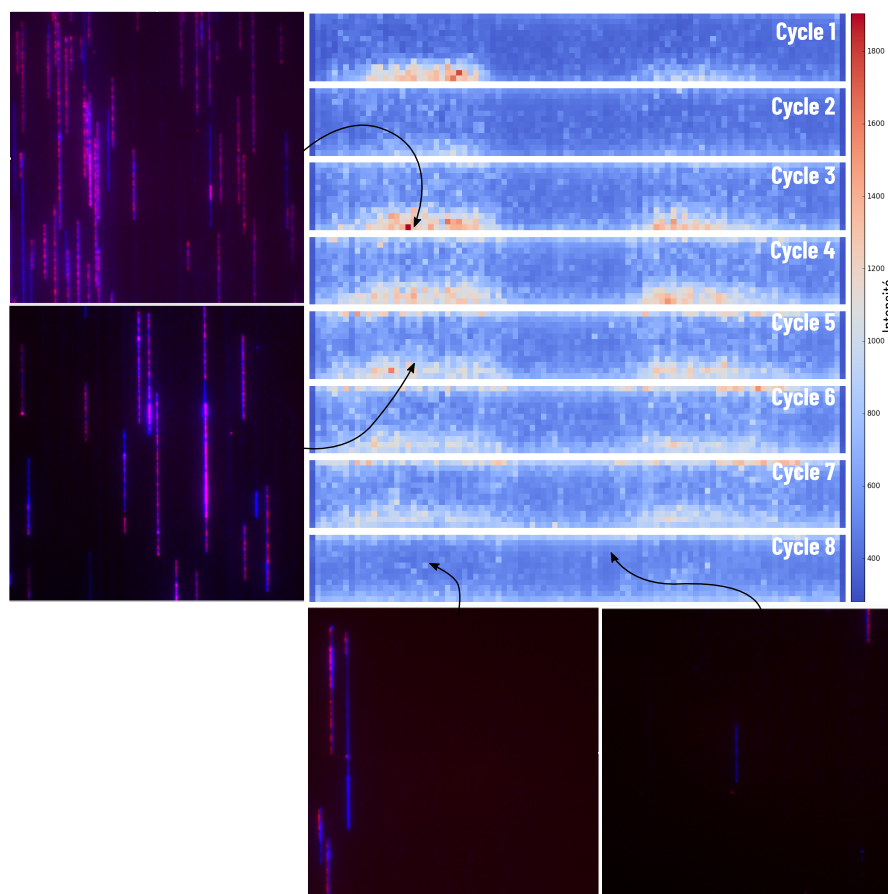
Une fois que nous avons obtenu l'ensemble des positions des maxima locaux, toutes les distances entre les différents points des 2 images (bleu et rouge) sont calculées et seules les distances inférieures à 6 pixels sont conservées. Les points sont alors appariés et vont constituer le paramètre d'entrée de la fonction de minimisation.



**Figure 3.20 – Détection des maxima locaux (YOYO-1 et signal répliatif).** Les maxima locaux sont détectés pour chacun des signaux à partir des images finales obtenues après post-processing. Ce sont les points rouges qui seront associés aux points bleus les plus proches et ce pour chacune des 171 images sélectionnées sur la base de leur intensité (cf. paragraphe 3.3.1.2)

## ÉTAPE 2 : ESTIMATION DE LA MATRICE DE TRANSFORMATION AFFINE

La matrice de transformation affine comporte 6 paramètres que nous cherchons à déterminer. Pour cela, nous avons fait le choix d'utiliser la fonction d'optimisation de la librairie *Scipy*. Afin d'estimer au mieux la matrice de transformation, les paramètres sont évalués sur les images possédant un grand nombre de molécules. Ainsi nous conservons, à chaque cycle, 15% des images (c'est-à-dire 171 images) ayant les intensités moyennes les plus élevées ayant été estimées sur le canal bleu (cf. Figure 3.21). Plus l'intensité est élevée, plus il y aura de molécules présentes dans ce FOV et donc plus il y aura de chance que ces molécules soient réparties plus ou moins uniformément sur l'image : cela permettrait une évaluation plus fiable des paramètres de la matrice puisque le décalage observé est de plus en plus important en partant du centre pour aller vers les bords de l'image.



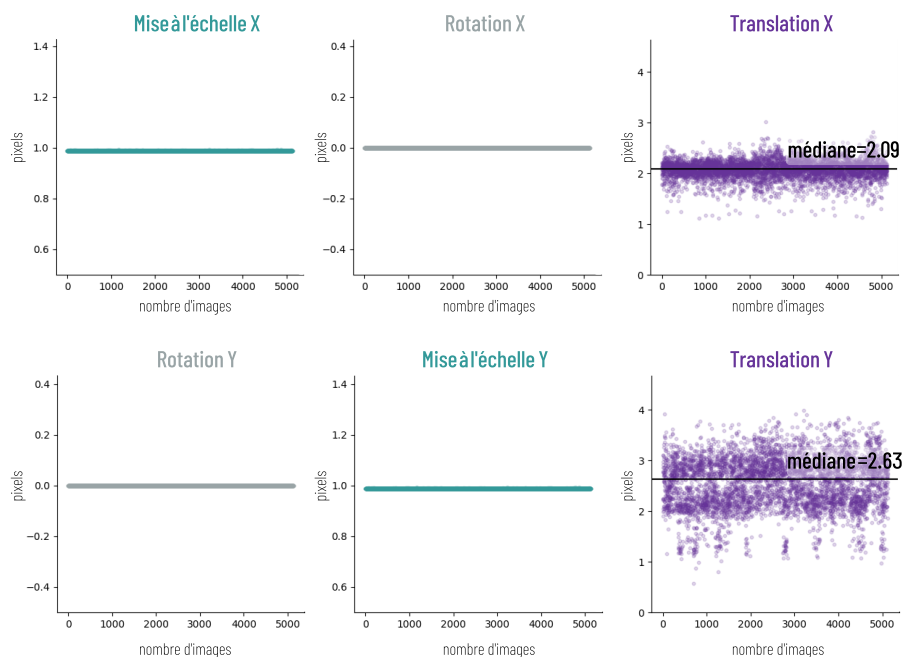
**Figure 3.21** – Répartition des molécules d'ADN suivant les cycles. Les valeurs moyennes de chacune des images de chaque cycle sont représentées. Les images contenant peu de molécules d'ADN sont représentées par des carrés bleus et inversement celles avec un grand nombre de molécules sont identifiées par un carré rouge. Le choix des images ayant de nombreuses molécules d'ADN est justifié du fait de la quantité de fragments d'ADN mais également de leur répartition dans l'image.

La fonction prend en entrée :

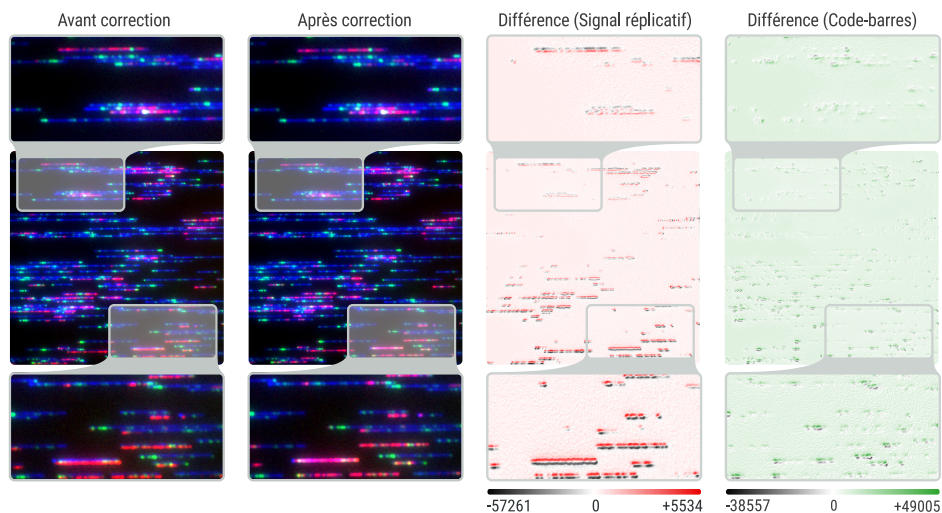
- la fonction de coût. À chaque itération, cette fonction va renvoyer la somme des distances des paires créées (cf. paragraphe 3.3.1.2).
- l'initialisation de la matrice (matrice identité). À chaque itération cette matrice de transformation va être modifiée.
- les paires de points entre la référence et la cible.
- la méthode utilisée ici SLSQP qui fait usage du *Sequential Least Squares Programming* pour minimiser une fonction de plusieurs variables avec n'importe quelle combinaison de bornes et de contraintes (KRAFT 1988).
- les bornes que nous définissons pour chacun des paramètres à déterminer.

Après avoir épluché l'ensemble des fichiers et les métadonnées obtenus à la suite d'un *run*, nous avons découvert des paramètres de mise à l'échelle que nous avons considérés dans le processus d'optimisation (signal répliatif : 0.99, code-barres : 0.996). Nous avons fixé les paramètres connus et n'estimons que les paramètres de translation selon les axes X et Y. La fonction renvoie une matrice de transformation affine pour chacune des images d'un *run* (cf. Figure 3.22). Nous assumons que l'aberration chromatique est la même sur l'ensemble des images. Au vu de la quantité d'images sur lesquels les paramètres sont estimés (171 images x 30 cycles), nous choisissons de prendre la valeur médiane pour chaque paramètre et définissons une matrice de transformation affine unique et robuste corrigeant l'ensemble des images d'un *run*. (À noter : cette matrice doit être estimée à chaque nouveau *run* et ne peut donc être utilisée d'un *run* à un autre.)

Une fois les paramètres définis, nous appliquons les transformations sur les images originale du canal rouge (signal répliatif) et du canal vert (code-barres)(cf. Figure 3.23). Nous pouvons voir que les différents signaux se superposent mais il semblerait que cette superposition soit imparfaite suivant l'axe Y. En effet, lorsque nous prenons les paires de points les plus proches dans le but de réaliser la minimisation de la somme des distances séparant ces paires, l'association des points n'est pas aussi fiable sur l'axe Y que sur l'axe X essentiellement parce que les signaux sont différents.



**Figure 3.22 – Paramètres de la matrice de transformation affine pour le signal répliatif.** Ces paramètres ont été calculée sur 5130 images d'un *run* (171 images x 30 cycles) avec la rotation en X et en Y nulle et la mise à l'échelle fixée à 0.99 (métadonnées). La valeur médiane des paramètres de translation en X et en Y ont été conservées afin de corriger l'ensemble des images du *run*. Les variations de valeurs sont plus importante en Y qu'en X en raison de l'imprécision des associations de points sur cet axe (dû au fait que les 2 signaux, répliatifs (discontinu) et YOYO-1 (continu) ne sont pas de même nature, il en est de même pour le signal du code-barres)



**Figure 3.23 – Recalage d'image.** La superposition des 3 couleurs pour les images originales mettent en avant l'existence d'une aberration chromatique et donc un décalage entre ces 3 couleurs (avant correction). L'ensemble des images d'un même *run* a été corrigé à l'aide d'une unique matrice de transformation pour faire en sorte que les signaux rouges et verts se superposent au signal bleu (après correction). Le décalage se produit sur les axes X et Y et est plus marqué sur les bords des images (zoom des molécules avant correction). La soustraction de l'image non corrigée à l'image corrigée (rouge pour le signal répliatif et vert pour le signal du code-barres) nous permet d'apprécier le gradient du décalage sur l'ensemble de l'image.



### 3.3.2 Extraction des profils d'intensité

Une fois les 2 corrections apportées, à savoir la correction de l'illumination et le recalage des images, nous pouvons procéder à l'extraction des profils d'intensité des molécules d'ADN détectées par Autodetect. Pour gérer au mieux nos données, nous utilisons la librairie *Pandas* permettant ainsi de conserver dans une trame de données *Pandas* (ou *dataframe*) toutes les informations utiles des fichiers originaux (*MoleculeID*, *XStart*, *XEnd*, etc ...) et en ajouter de nouvelles tels que les profils d'intensité.

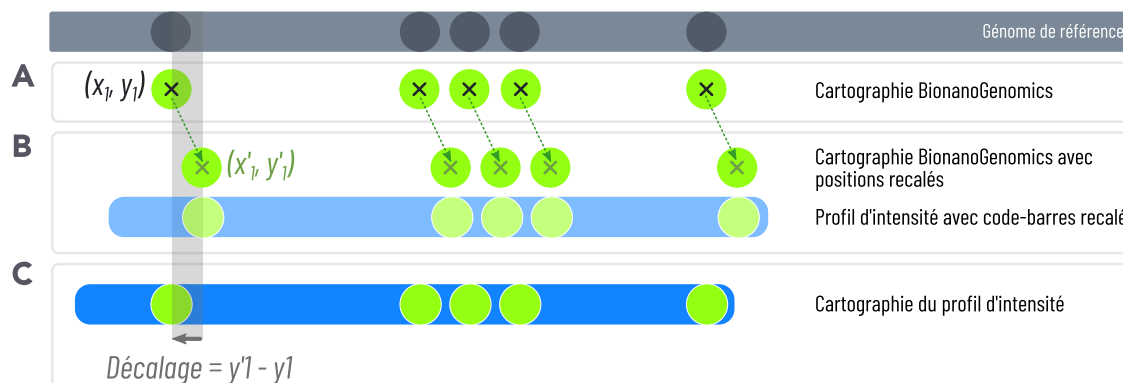
Grâce au recalage d'images, nous sommes en mesure d'extraire les profils d'intensité des trois signaux simultanément sur la base des coordonnées des extrémités des molécules obtenues par Autodetect. À chaque position le long de la molécule d'ADN, la moyenne de l'intensité est calculée sur une fenêtre de 3 pixels de largeur (perpendiculairement à la molécule). Nous avons fait ce choix afin d'éviter la prise en compte des intensités des molécules voisines étant donné que l'intensité des molécules s'étale sur 5 pixels de largeur. Par ailleurs, au vu des observations faites concernant la linéarité des molécules et pour extraire au mieux nos profils d'intensité, nous avons tenu compte de leur inclinaison dans les nanocanaux. Une fois l'extraction réalisée, les profils rouge, vert et bleu, de chacune des molécules, sont consignés dans un *dataframe*, en gardant le *MoleculeID* afin de regrouper ces informations avec celles déjà obtenues à l'aide des logiciels de Bionano Genomics®.

### 3.3.3 Cartographie des profils d'intensité

L'outil RefAligner ne se préoccupe que des distances séparant les différents labels du code-barres (ou “*nick*”) et retrouve ces distances sur le génome de référence. Dans un des fichiers de sortie d'Autodetect, nous disposons des coordonnées des *nicks* et dans un des fichiers de sortie de RefAligner, nous avons la position du premier *nick* correctement cartographiée sur le génome de référence. Etant donné les modifications apportées aux images, à savoir le recalage des signaux vert et rouge sur le signal bleu, nous devons évaluer le décalage entre le premier *nick* correctement cartographié et ce même *nick* après avoir subi les transformations dans le but de repositionner correctement les profils d'intensité sur le génome de référence. Il a donc fallu appliquer la matrice de transformation sur les positions des code-barres et par la suite calculer la différence entre l'ancienne et la nouvelle position du premier *nick*.

Aussi infime que soit cette correction, elle est indispensable (cf. Figure 3.24).

*Rappel : un pixel est équivalent à environ 500 paires de bases.*

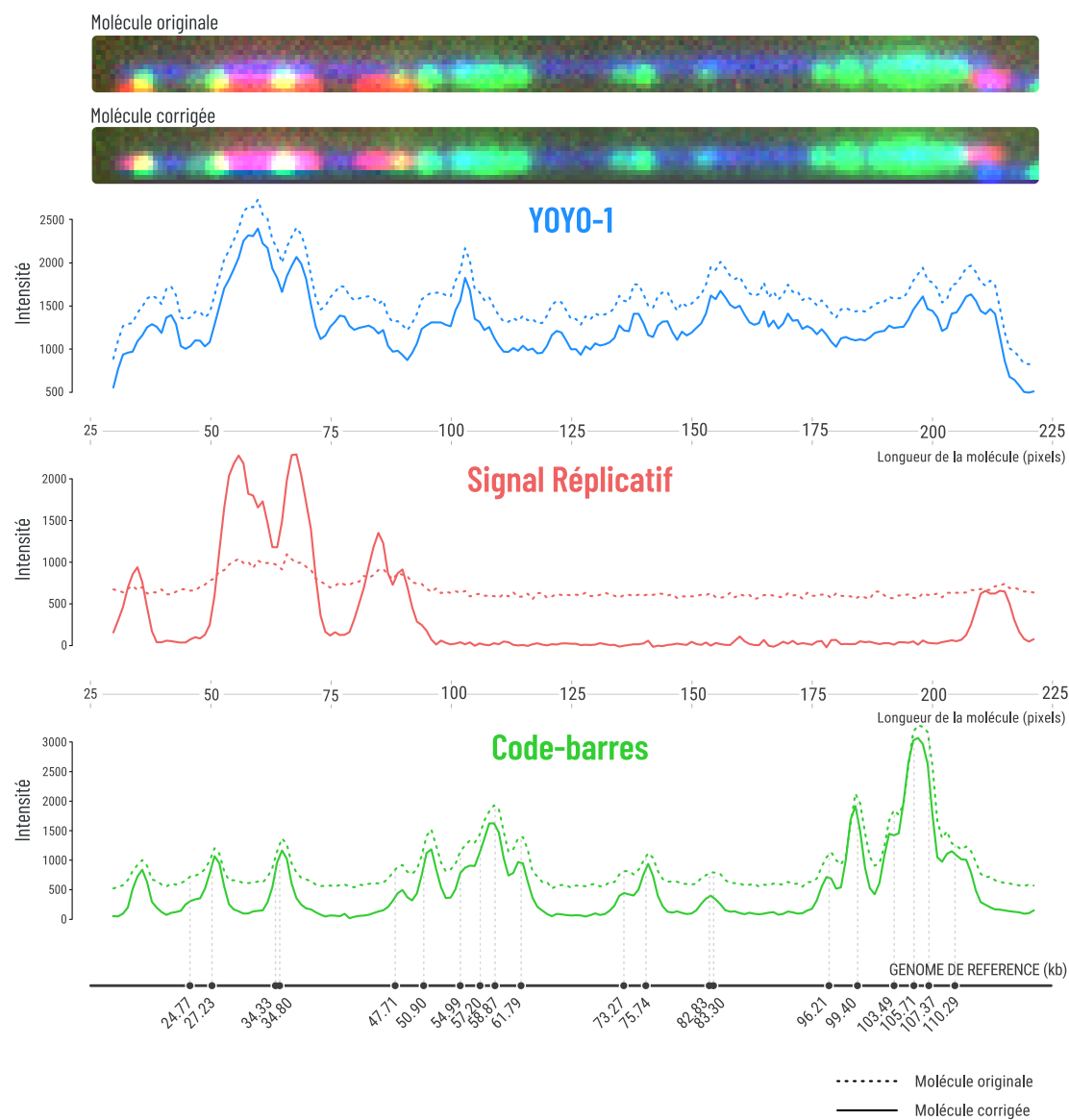


**Figure 3.24 – Cartographie des profils d'intensité : le principe.** RefAligner réalise la cartographie optique des molécules d'ADN grâce au code-barres. Nous disposons des coordonnées des positions des points du code-barres (“nick”) (A). Sachant que nous réalisons un recalage du signal vert sur le signal bleu, les coordonnées renvoyées par RefAligner se trouvent modifiées (B). Afin de cartographier nos profils d'intensité il est nécessaire de quantifier ce décalage (pour chacune des molécules) (C) permettant ainsi de replacer la molécule sur le génome de référence.

Une fois cette étape réalisée, nous sommes en mesure de positionner correctement le profil d'intensité sur l'ADN de référence et de visualiser chacun de ces profils individuellement.

### 3.3.4 Visualisation des profils d'intensité en molécule unique

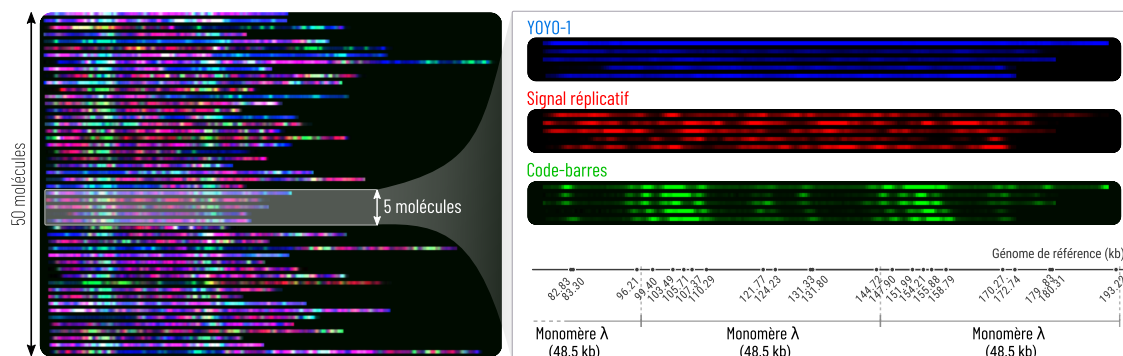
IrysView, un logiciel de visualisation mis à disposition par Bionano Genomics® et fonctionnant uniquement sous le système d'exploitation Windows, permet d'afficher une reconstruction de l'ensemble des molécules d'ADN ayant été cartographiées. À la différence de ce logiciel, pour une molécule donnée, nous sommes en mesure de visualiser les profils d'intensité des trois signaux et d'observer la qualité de la cartographie. Sur la figure 3.25, nous observons des améliorations notables essentiellement sur le signal réplication et sur celui du code-barres. La correction de l'illumination a abaissé la ligne de base des profils d'intensité en plus de permettre une amélioration du ratio signal sur bruit. Nous pouvons aussi noter que la seconde étape de post-processing, le recalage d'image, est une étape primordiale. En effet, en regardant de plus près l'extrémité droite du signal réplicatif, nous constatons qu'en absence du recalage des signaux nous aurions récupéré une information erronée : une partie du signal réplicatif n'était pas associée à la bonne molécule.



**Figure 3.25 – Visualisation d'un profil d'intensité d'un fragment d'ADN répliqué en molécule unique.** Un fragment d'ADN avec les trois couleurs est représenté avant et après avoir effectué les corrections (correction de l'illumination inhomogène et recalage d'image). En dessous, les profils d'intensité des différents signaux (trait pointillé : profil d'intensité avant correction, trait plein : profil d'intensité après correction) sont cartographiés sur un génome de référence. Pour les profils d'intensité, l'axe X est donné en pixel, pour le génome de référence les positions sont données en kilobases. Les traits en pointillés gris partant du profil d'intensité du code-barres le relient avec les positions correspondantes sur le génome de référence.

### 3.3.5 Visualisation des profils d'intensité en population

Nous sommes parvenus à cartographier nos profils d'intensité en molécule unique, il est donc maintenant possible d'obtenir une visualisation plus globale de nos données avec une visualisation en population. Cette dernière nous permet d'avoir une idée de la qualité de notre cartographie mais aussi de pouvoir réaliser des analyses populationnelles sur le signal répliatif (cf. Chapitre 4).



**Figure 3.26 – Visualisation des profils d'intensité des molécules d'ADN répliquées en population.**

Les molécules d'ADN ont été cartographiées sur un génome de référence comportant 30  $\lambda$  tous orientés *Forward* (cf. Section 4.1). À gauche, 50 molécules cartographiées sont représentées dans les 3 couleurs. À droite, les 3 signaux sont représentés séparément et le signal vert est aligné avec le génome de référence.

### 3.3.6 Pipeline BionanoGenomics vs. pipeline HOMARD

Ci-après est présenté un schéma récapitulatif du pipeline que nous avons développé en faisant usage des informations apportées par les différents logiciels fournis par la compagnie Bionano Genomics®. Nous sommes à présent en mesure de cartographier “*genome-wide*” nos profils d'intensité et en particulier les signaux répliatifs (cf. Figure 3.27), ce qui n'était pas possible précédemment.

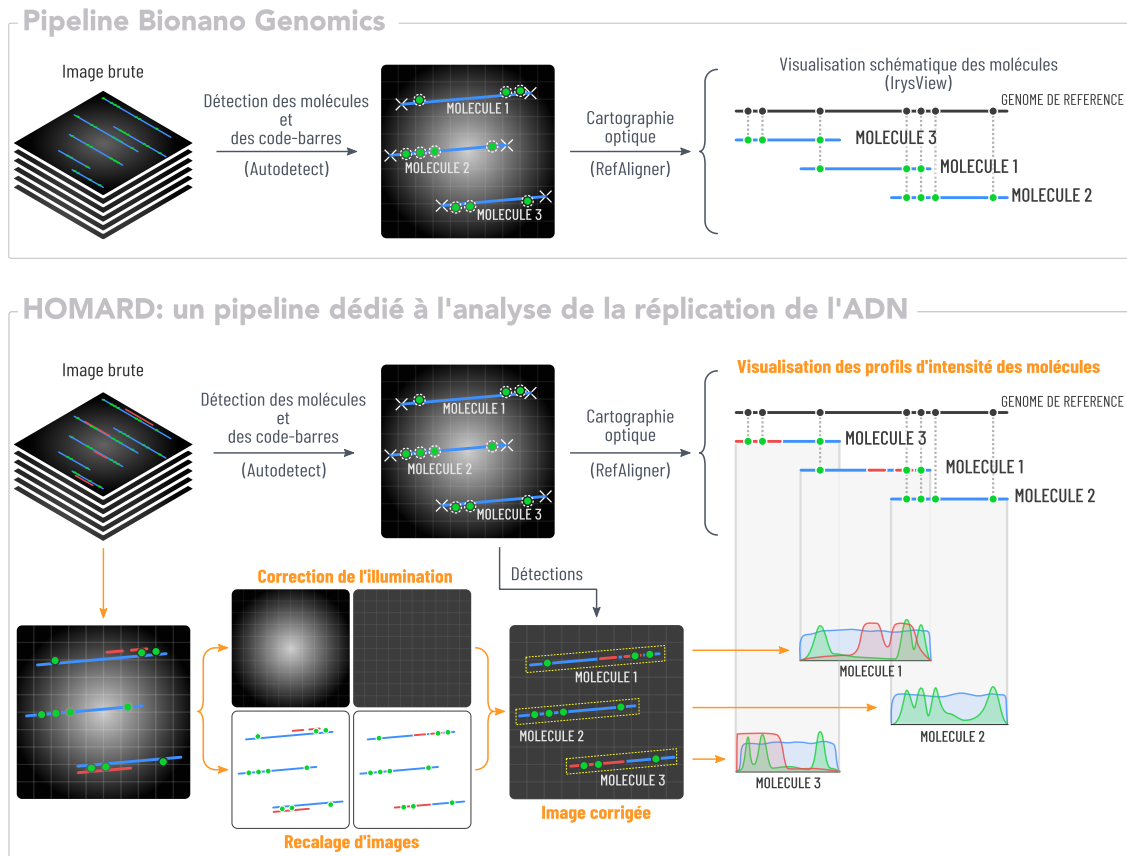


Figure 3.27 – Pipeline bioinformatique pour l’analyse de la réplication de l’ADN via le système Irys.

Le pipeline actuel de Bionano Genomics® permet à partir des images brutes de réaliser une détection automatique des molécules d’ADN ainsi que de leur code-barres associés grâce à Autodetect (logiciel propriétaire). Ensuite, RefAligner va permettre la cartographie optique de ces molécules sur un génome de référence. Les résultats sont visualisables sur le logiciel IrysView (fonctionne sous Windows) où la visualisation n’est que schématique. Le pipeline que nous avons développé en nous appuyant sur le pipeline de Bionano Genomics®, permet d’obtenir des profils d’intensité, de les cartographier et ainsi d’avoir une visualisation en molécule unique de la réplication de l’ADN. Pour ce faire une correction de l’illumination et le recalage des signaux rouges et verts sur le signal bleu ont été nécessaires. Dans notre cas la visualisation n’est pas schématique puisque nous avons accès au profils d’intensité et à la molécule d’origine.

### 3.3.7 La détection des segments répliqués

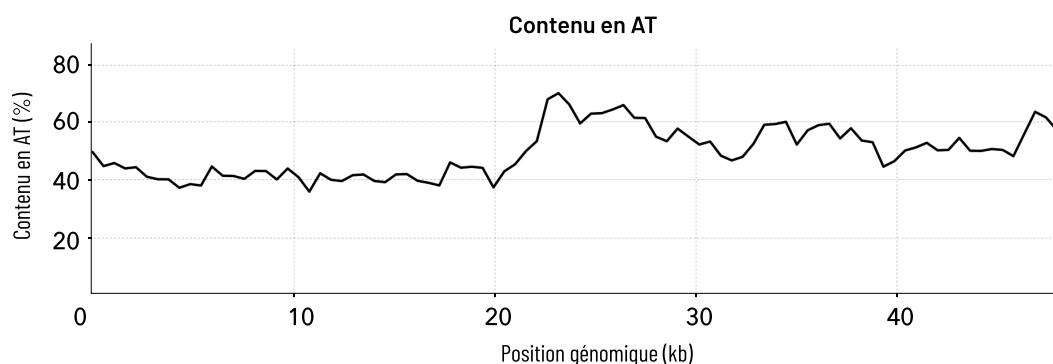
Nous avons maintenant la capacité d'extraire les profils d'intensité de centaines de milliers de molécules et de les cartographier. L'objectif consiste à présent à détecter automatiquement et de façon robuste les segments répliqués (région répliquée de la molécule). L'échantillon sur lequel nous avons travaillé pour cette détection est l'ADN de  $\lambda$  ayant répliqué dans des extraits d'œufs de Xénope (cf. Chapitre 4).

#### 3.3.7.1 Composition du signal réplcatif (rouge)

Dans un premier temps nous avons fait une hypothèse sur la composition de notre signal réplcatif rouge (cf. Section 4.1). Ce dernier serait dépendant de la Fonction d'Étalement du Point (FEP ou PSF - *Point Spread Function*), du contenu en A/T (puisque nous incorporons des dUTP fluorescents au niveau des fourches de réplication) et enfin du signal d'intérêt que nous cherchons à déterminer.

#### ESTIMATION DU CONTENU EN AT POUR L'ADN DE $\lambda$

Le contenu en AT de l'ADN de  $\lambda$  est calculé à partir de la séquence d'ADN localement avec une fenêtre glissante de 539 paires de bases afin d'atteindre la résolution de l'image. Nous obtenons ainsi la proportion de AT dans une fenêtre donnée (cf. Figure 3.28).



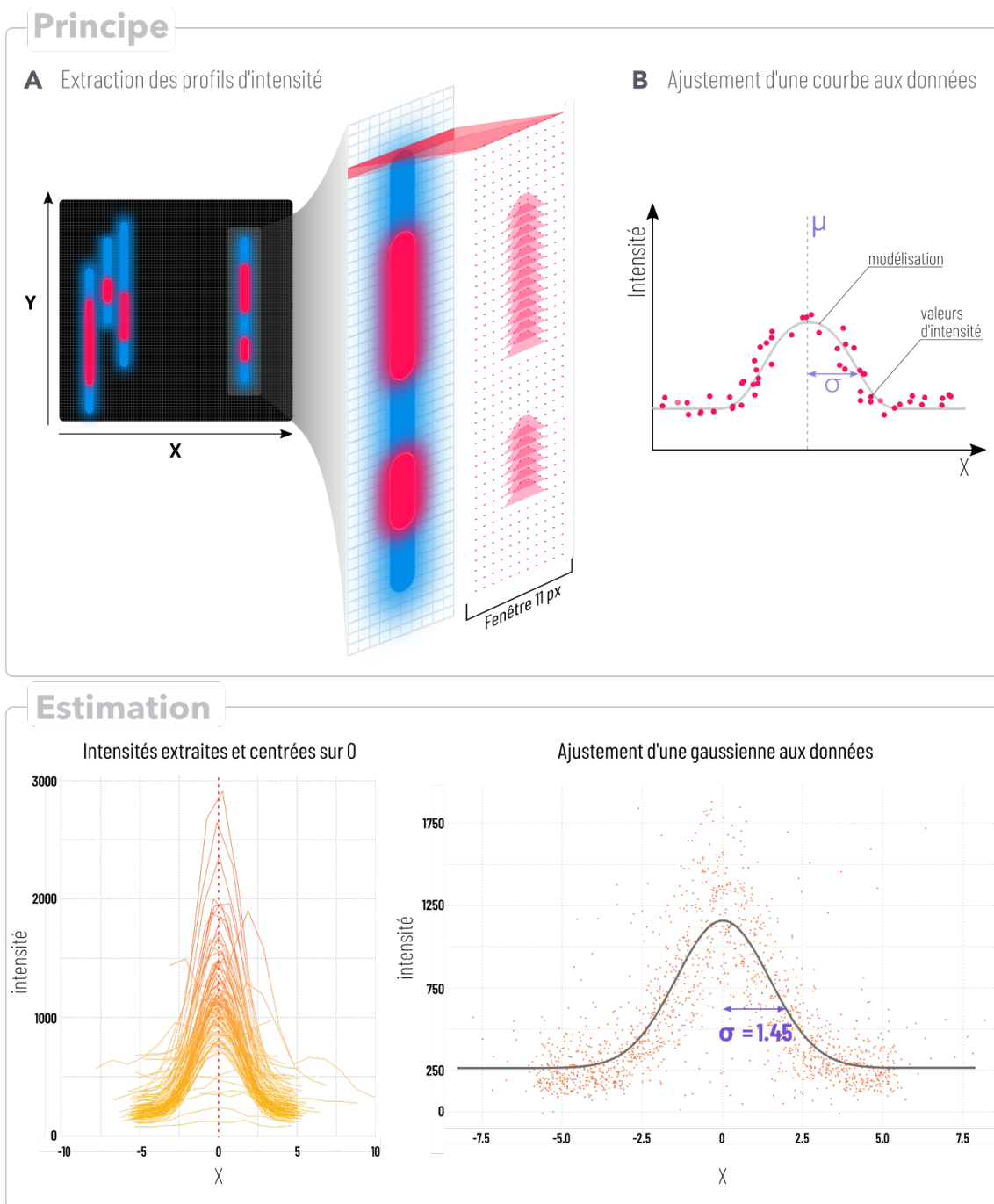
**Figure 3.28** – Contenu en AT de l'ADN de  $\lambda$ . Ce profil calculé à l'aide d'une fenêtre glissante de 539 paires de bases, représente la proportion de nucléotides A et T présents dans le génome de  $\lambda$  faisant 48502 paires de bases.

## ESTIMATION DE LA FEP DU SIGNAL RÉPLICATIF

La FEP est une fonction représentant la réponse d'un système d'acquisition d'image à une source ponctuelle. L'effet de cette dernière est décrit en appliquant une opération de convolution aux données. Ainsi, les opérations de déconvolution permettent d'inverser ce processus améliorant ainsi la résolution de l'image. Afin d'estimer la FEP sur nos données, il a fallu dans un premier temps conserver uniquement les molécules d'ADN isolées. En effet, si deux (ou plus) de molécules sont à proximité, l'intensité de ces dernières risque d'être influencée plus ou moins fortement. Près de 373 600 molécules de notre jeu de données ont été annotées comme étant isolées, c'est-à-dire n'ayant pas de molécules dans un voisinage de 5 pixels autour d'elles. Une fois ces molécules isolées sélectionnées, nous avons extrait les profils d'intensité du signal réplicatif perpendiculairement à la molécule en prenant une fenêtre de 11 pixels de large (cf. Figure 3.29, A) (typiquement une molécule d'ADN fait 5-6 pixels de large). L'ensemble des profils sont ensuite recentrés sur 0 avant de pouvoir évaluer notre FEP. Pour ce faire, nous avons ajusté une courbe à ces données (cf. Figure 3.29, B). Dans notre cas il s'agit d'une fonction gaussienne :

$$f(x) = A \exp\left(-\frac{(x-B)^2}{2c^2}\right) + d$$

tel que A correspond à l'amplitude de la "cloche", B la position du centre de la "cloche", C l'écart-type ( $\sigma$ ) et d une constante. Nous supposons que cette PSF est identique sur les axes X et Y et faisons donc le choix de déterminer le paramètre ( $\sigma$ ) sur l'axe X. Ainsi la valeur de l'écart-type est de 1.45 pixels signifiant donc qu'un signal ponctuel rouge s'étale sur environ 3 pixels.

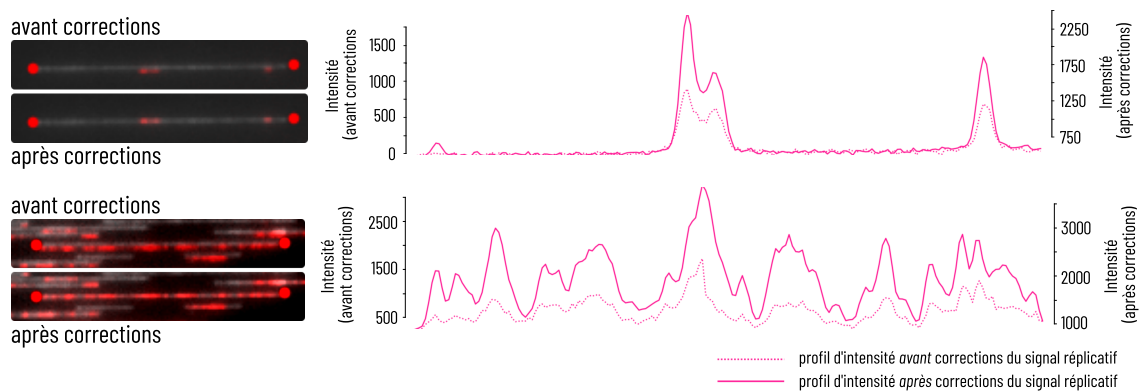


**Figure 3.29 – Estimation de la Fonction d'Étalement du Point (FEP).** Le principe : pour estimer au mieux la FEP, nous avons dans un premier temps sélectionné les molécules isolées puis extrait les profils d'intensité perpendiculairement aux molécules sur une fenêtre de 11 pixels (A). Puis, une gaussienne (modélisant le mieux la FEP) est ajustée aux données d'intensité de l'ensemble des molécules (B). Estimation de la FEP : pour ce faire nous avons commencé par centrer les profils sur 0 puis ajuster la courbe sur les données. Ici ne sont représentées que 100 des 373000 molécules. Une valeur de ( $\sigma$ ) de 1.45 pixels est déterminée et sera utilisée dans la détection des segments répliqués (cf. Section 3.3.7



### 3.3.7.2 Approche : optimisation d'une fonction de coût pour déconvoluer le signal répliatif brut

Définir un simple seuillage sur nos données n'est pas une solution pouvant donner des résultats robustes et fiables au vu des variations d'intensité entre les molécules (cf. Figure 3.30). C'est pourquoi, à partir des paramètres présentés ci-avant, nous avons tenté une approche permettant de "déconvoluer" le signal afin de retomber sur l'information qui nous intéresse, à savoir le signal répliatif "binaire" caché.



**Figure 3.30 – Exemples de profils d'intensité du signal répliatif.** Ici sont présentés deux exemples de profils d'intensité du signal répliatif avant (trait pointillé) et après (trait plein) avoir appliqué les corrections (correction de l'illumination et recalage des signaux). Nous pouvons voir que fixer une valeur seuil dans le premier exemple (molécule du haut) pourrait donner un résultat représentatif des segments répliqués. Mais appliquer ce même seuillage sur la seconde molécule (bas) risque de donner un résultat erroné.

#### LE MODÈLE

Soient,

$X = \{x_i, \dots, x_n\}$	le signal "binaire" caché
$Y = \{y_i, \dots, y_n\}$	le signal répliatif observé ou reconstruit
$Y = f(X)$	le modèle d'observation
$f_i(X)$	$i^{\text{ème}}$ élément du vecteur $f(X)$
$\hat{X}$	l'estimation du signal "binaire" caché

$$f(X) = (AT \times X) * FEP \quad \text{le modèle} \quad (3.1)$$

$$\epsilon(X) = \frac{1}{n} \sum_{i=1}^n (x_i^2(x_i - 1)^2) \quad \text{la pénalité} \quad (3.2)$$

$$\hat{X} = \min_X \sum_{i=1}^n (y_i - f_i(X))^2 + \lambda \epsilon(X) \quad \text{l'estimation du signal caché} \quad (3.3)$$

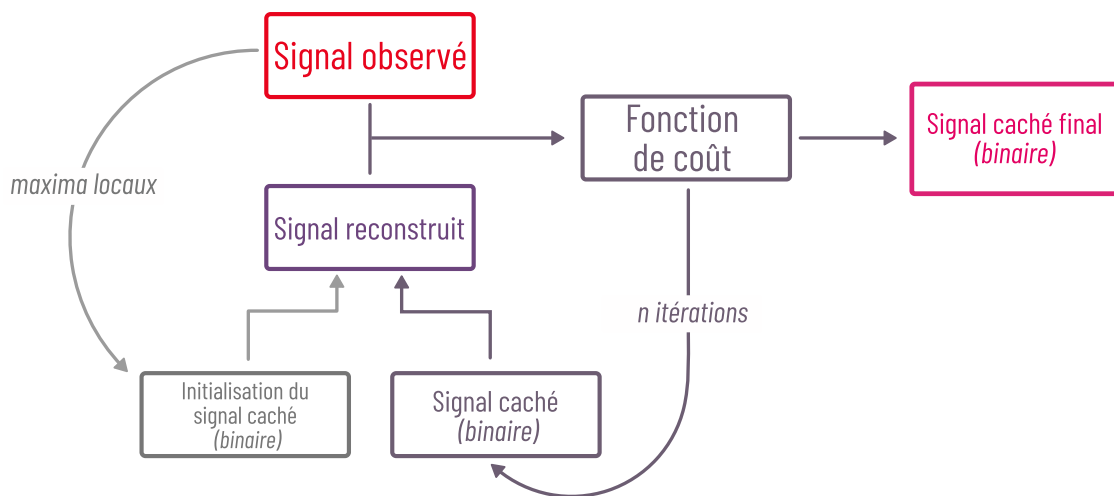
où

- dans l'équation 3.1,  $AT$  correspond au contenu en AT,  $AT \times X$  est un produit réalisé terme à terme et  $FEP$  est une gaussienne ayant comme paramètres ceux déterminés par ajustement de courbe (cf. paragraphe 3.3.7.1)
- dans l'équation 3.3,  $\lambda$  est une constante. Cette constante permet de renforcer la pénalité (cf. équation 3.2). Nous avons, après plusieurs essais avec différentes valeurs de  $\lambda$ , défini  $\lambda = 2$ .

La pénalité (cf. équation 3.2) est une fonction permettant de contraindre l'optimiseur à renvoyer des valeurs égales ou très proches de 0 et de 1. En effet, sans cette pénalité la fonction d'optimisation de la librairie *Scipy* autorise les valeurs du signal caché à fluctuer entre 0 et 1. Or, nous faisons l'hypothèse que notre signal caché est binaire (signal "binaire" caché), de ce fait nous devons "forcer" l'optimiseur à maintenir les valeurs proches ou égales à 0 et 1 en ajoutant cette pénalité aux signaux "binaires" cachés ne répondant pas à nos critères. Lorsque les valeurs du signal "binaire" caché sont égales à 0 ou 1, cette fonction n'entraîne pas une augmentation du coût total (pénalité = 0) lors de la minimisation. Par contre, lorsque les valeurs fluctuent entre 0 et 1, le coût total est impacté.

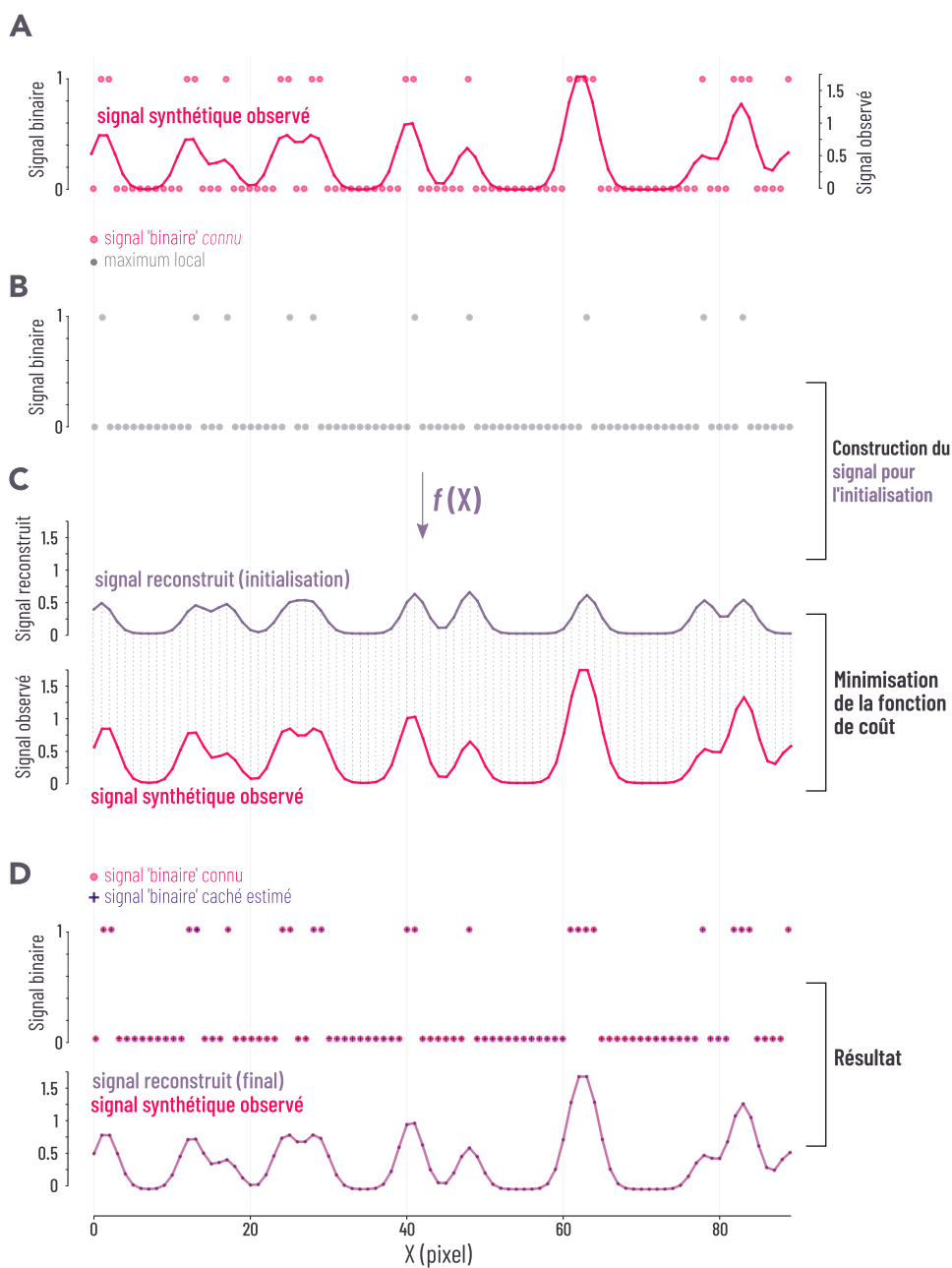
## PROCESSUS D'OPTIMISATION POUR LA DÉTECTION AUTOMATIQUE DES SEGMENTS RÉPLIQUÉS

Afin de déterminer les segments répliqués de nos profils d'intensité du signal répliatif (signal observé), nous avons procédé de la façon suivante. Nous disposons du signal observé et du signal reconstruit à partir des maxima locaux détectés dans le signal observé. Ces maxima vont correspondre au signal "binaire" caché pour l'initialisation. À chaque itération, le signal "binaire" caché se trouve modifié jusqu'à ce que le signal reconstruit et le signal observé coïncident (cf. Figure 3.31).



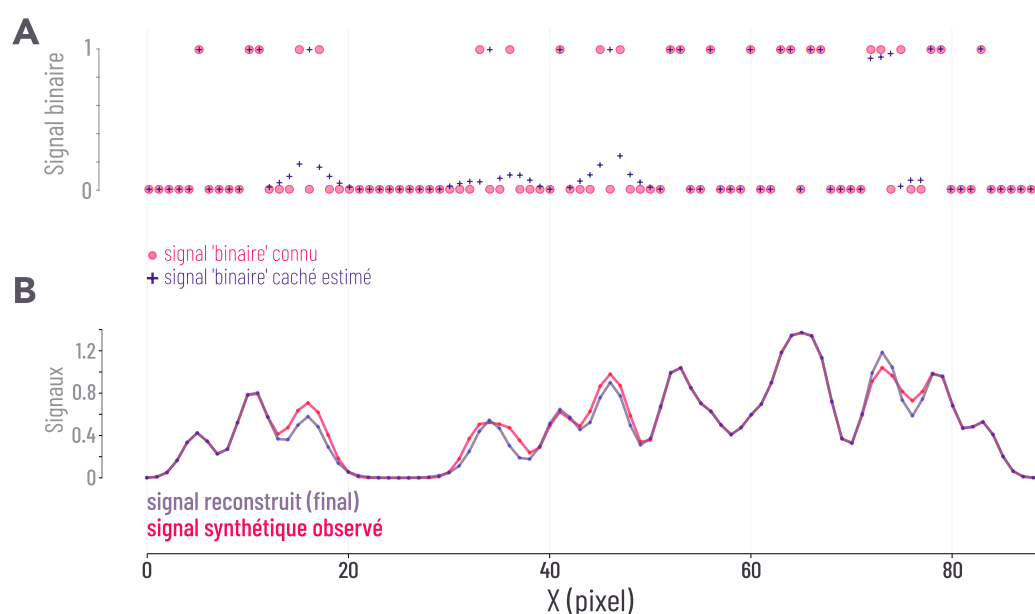
**Figure 3.31** – Schéma explicatif du processus d'optimisation pour la détection automatique des segments répliqués. À partir du signal observé (qui peut être soit un signal reconstruit généré à partir d'un signal binaire connu (signal synthétique observé)(cf. Figure 3.32), soit le signal répliatif original (signal original observé)), nous détectons les maxima locaux. À partir de ces maxima, nous générons un signal reconstruit qui servira à l'initialisation lors de l'optimisation. Au premier tour, un coût est calculé entre le signal observé et le signal reconstruit (cf. équation 3.3). Le signal 'binaire' ayant servi à l'initialisation (maxima locaux) va être modifié, un nouveau signal reconstruit va être généré et un coût sera calculé. Il y aura  $n$  itérations jusqu'à ce que le signal reconstruit et le signal observé coïncident permettant ainsi d'obtenir l'estimation finale du signal "binaire" caché.

Afin de vérifier que notre optimisation fonctionne, nous avons tout d'abord testé notre algorithme sur un signal dont nous connaissons le signal binaire (signal "binaire" connu). À partir de ce signal "binaire" connu, le signal synthétique observé est généré et nous effectuons la minimisation de la fonction de coût (cf. Figure 3.32).



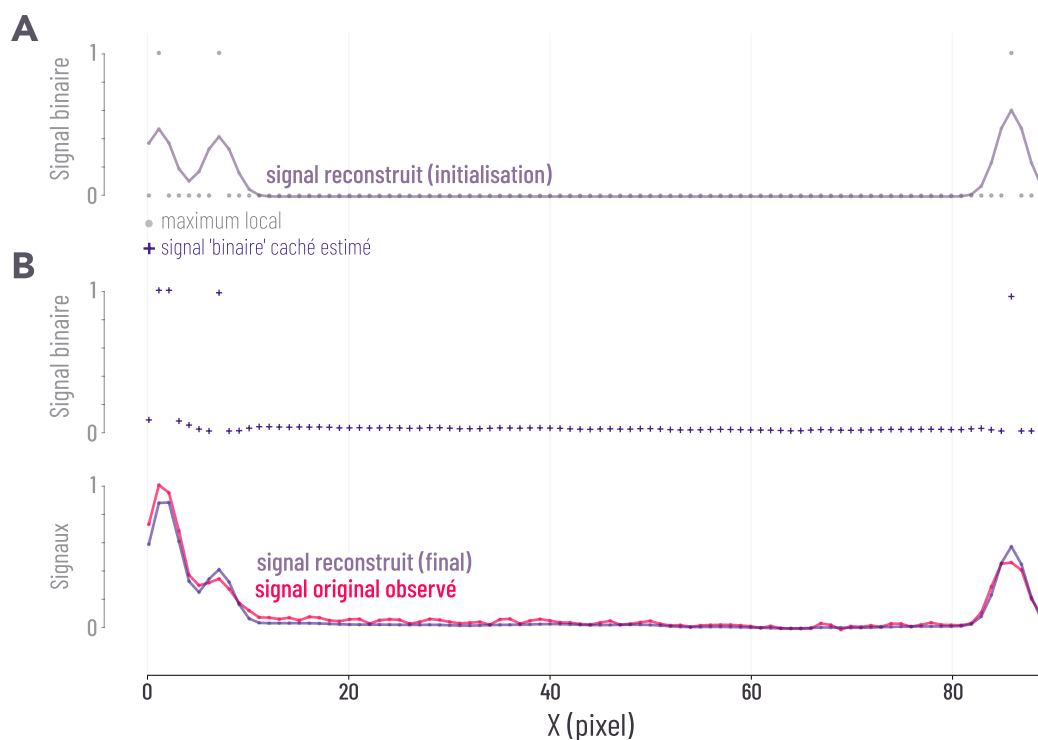
**Figure 3.32 – Détermination des segments répliqués sur un signal synthétique observé.** Nous avons procédé comme expliqué dans la figure 3.31. **(A)** À partir d'un signal "binaire" connu, nous avons généré le signal synthétique observé (rouge). **(B)** Sur ce signal synthétique observé, nous détectons les maxima locaux à partir desquels le signal reconstruit est créé et servira à l'initialisation lors de l'optimisation. **(C)** Nous minimisons ensuite la fonction de coût afin d'estimer le signal "binaire" caché **(D)** À la suite de l'optimisation, nous parvenons à retrouver le signal binaire connu (superposition du signal "binaire" connu et de celui estimé).

Avec notre algorithme, nous parvenons à retrouver le signal "binaire" caché (cf. Figure 3.32). Suite à ce résultat, nous avons réalisé la détection sur des signaux plus complexes (cf. Figure 3.33). Nous remarquons que pour ce type de signal, le signal "binaire" caché estimé ne coïncide pas totalement avec le signal "binaire" *connu*. Nous avons donc fait le choix d'appliquer notre algorithme sur des molécules faiblement répliquées (cf. Figure 3.34) dont la complexité est plus proche du signal analysé dans la figure 3.32. Nous obtenons un résultat satisfaisant sur l'exemple présenté dans la figure 3.34 puisque le signal "binaire" caché estimé semble refléter le signal observé. Il est important de préciser que les profils d'intensité du signal réplcatif sont normalisés entre 0 et 1.



**Figure 3.33 – Détermination des segments répliqués sur un signal synthétique observé complexe.**

Nous avons procédé comme expliqué dans la figure 3.31 mais cette fois-ci avec un signal complexe par rapport à celui analysé dans la figure 3.32. Dans ce cas le signal "binaire" caché estimé ((A), croix violettes) ne se superpose pas parfaitement au signal "binaire" *connu* ((A), points rouges). Il en est de même avec le signal reconstruit et le signal observé (B).



**Figure 3.34 – Détermination des segments répliqués sur un signal original observé issu d'une molécule faiblement répliquée.** (A) À partir du signal original observé (montré en B, signal rouge), nous détectons les maxima locaux qui vont nous permettre de créer le signal reconstruit servant à l'initialisation. (B) À la suite de l'optimisation, le signal original observé et le signal reconstruit se superposent de façon satisfaisante.

Dans ce chapitre, nous avons présenté les outils bioinformatiques que nous avons développés afin d'analyser le signal répliatif (signal rouge) dans différents échantillons (cf. Chapitre 4). Le pipeline HOMaRD permet d'extraire les centaines de milliers de profils d'intensité dans les 3 couleurs (ADN, code-barres, signal répliatif) et de cartographier ces profils d'intensité sur un génome de référence. Ainsi, il est possible d'obtenir une information au niveau de la molécule unique mais également d'agrèger les profils d'intensité pour une analyse en population. Pour ce qui est de la détection des segments répliqués, le modèle que nous proposons sera utilisé pour l'analyse des molécules d'ADN dont le taux de répliation est faible (cf. Chapitre 4.1.2).



## Chapitre4

# Applications et résultats

---

<b>4.1</b>	<b>Analyse de la réplication de l'ADN de Lambda dans des extraits d'œufs de Xénope</b>	<b>111</b>
4.1.1	Analyse de la sous-population des molécules fortement répliquées des concatémères 5'-3'	112
4.1.2	Analyse de la sous-population des molécules faiblement répliquées	123
<b>4.2</b>	<b>Analyse de la réplication de l'ADN chez l'Homme</b>	<b>134</b>
<b>4.3</b>	<b>Une nouvelle approche pour la détection des segments répliqués</b>	<b>137</b>
4.3.1	Un doublement de fluorescence du YOYO-1 présent mais pas systématique dans notre échantillon de chromatine de sperme de Xénope	138
4.3.2	Existe-il un doublement de fluorescence du signal YOYO-1 dans notre échantillon de levure?	139

---





## 4.1 Analyse de la réplication de l'ADN de Lambda dans des extraits d'œufs de Xénope

Dans le but de développer nos outils bioinformatiques nous avons fait le choix de travailler sur un modèle simple et sur un génome de petite taille, l'ADN du bactériophage  $\lambda$  (48 502 pb) ayant été répliqué dans des extraits d'œufs de Xénope. Un avantage important de cette approche est qu'elle permet la réplication de pratiquement n'importe quel ADN tel que l'ADN de  $\lambda$  (cf. Section 2.4.2.1). Le bactériophage  $\lambda$  est un outil prisé des biologistes en raison de sa facilité de manipulation génétique et biochimique (ENQUIST et SKALKA 1978). Il s'agit d'un virus procaryote infectant *Escherichia coli* et qui a été découvert en 1951 par Esther Lederberg (R. M. LEDERBERG et J. LEDERBERG 1951). Le bactériophage  $\lambda$  possède un ADN double brin faisant 48 502 paires de bases avec des extrémités cohésives. Ces extrémités simple-brins sont complémentaires et peuvent facilement fusionner afin de former des multimères linéaires ou des molécules circulaires. L'ADN de  $\lambda$  répliqué dans des œufs de Xénope est exposé au processus de jonction des extrémités non homologues (ou non-homologous end joining (NHEJ)). Il s'agit d'un mécanisme impliqué dans la réparation des cassures double-brins de l'ADN pouvant survenir aussi bien chez les procaryotes que chez les eucaryotes. Ce mécanisme n'assure pas la restauration de la séquence initiale de l'ADN induisant donc un changement de l'information génétique (mutations). Dans notre échantillon, l'ADN de  $\lambda$  va former rapidement des concatémères composés d'au moins 2 fragments. Ainsi les conformations de paires possibles que nous allons retrouver dans nos échantillons sont des concatémères *Foward-Foward*, *Forward-Reverse*, *Reverse-Foward* et *Forward-Reverse* (cf. Figure 4.1).

## Types de jonctions Lambda



**Figure 4.1 – Orientations des concatémères du bactériophage  $\lambda$ .** Sont présentés les 4 différents types de jonctions possibles pour les concatémères de  $\lambda$  sachant que le *Forward-Forward* et le *Reverse-Reverse* représente une seule et même jonction.

Pour notre publication (DE CARLI, MENEZES et al. 2018), en plus de décrire notre pipeline, nous avons analysé une sous-population de molécules, à savoir les concatémères *Forward-Forward* fortement répliqués. Dans la sous-section suivante, seront présentés les résultats obtenus pour l’analyse de la sous-population de molécule d’ADN de  $\lambda$  faiblement répliquées.

#### 4.1.1 Analyse de la sous-population des molécules fortement répliquées des concatémères 5’-3’

##### 4.1.1.1 High-Throughput Optical Mapping of Replicating DNA

Ci-après, notre publication décrivant nos outils d’analyse de la réplication de l’ADN en faisant usage du système Irys, sur un échantillon de l’ADN de  $\lambda$  répliqué dans des extraits d’œufs de Xénope.

*De Carli, F.\*, Menezes, N.\*, Berrabah, W., Barbe, V., Genovesio, A., et Hyrien, O. (2018). High-Throughput Optical Mapping of Replicating DNA. Small Methods, 2(9), 1800146.*

\* co-premiers auteurs.

# High-Throughput Optical Mapping of Replicating DNA

Francesco De Carli, Nikita Menezes, Wahiba Berrabah, Valérie Barbe, Auguste Genovesio,\* and Olivier Hyrien\*

DNA replication is a crucial process for the universal ability of living organisms to reproduce. Genome-wide methods to map DNA replication use large cell populations, which smoothes out variability between chromosomal copies. Single-molecule methods may reveal this variability but remain refractory to automation, precluding genome-wide analyses. Here, the Bionano Genomics Irys system, an optical DNA mapping device, is repurposed for high-throughput optical mapping of replicating DNA (HOMARD). HOMARD combines direct fluorescent labeling of replication tracks and nicking endonuclease sites with DNA linearization in nanochannel arrays and dedicated image processing. The robustness of this approach is demonstrated by an ultrahigh coverage (23 311×) replication map of bacteriophage  $\lambda$ -DNA in *Xenopus* egg extracts. This coverage, by far the highest ever reported for a single-molecule replication study, confirms with unprecedented statistical significance the lack of sequence preference for replication initiation in this system. The tools developed here open the way to genome-wide analysis of DNA replication at the single-molecule level and may be readily adapted for similar studies of other epigenetic features.

are today recognized as a major threat to genome stability and have been associated with cancer and other diseases.<sup>[5]</sup> However, mapping genome replication in metazoans has long remained challenging. For example, genome-wide mapping of human replication origins has been achieved by multiple approaches with only modest agreement.<sup>[3,6]</sup> One possible source of discrepancy is the incomplete purification of sequenced replication intermediates. A further limitation is that genome-wide replication mapping so far has been restricted to cell populations. As there is considerable cell-to-cell variability, ensemble profiles are at best a smoothed average of individual replication patterns. Single-molecule (SM) techniques must be used to fully evaluate cell-to-cell heterogeneity and uncover spatial correlations in replication patterns. SM techniques are also required to visualize rare pathological events such as fork stalling. Furthermore, SM replication signals are free from con-

taminating DNA molecules. Thus, the development of SM methods is crucial to progress in our understanding of DNA replication.

Unfortunately, none of the existing SM techniques has lent itself to full and robust automation. Consequently, their throughput has remained drastically low. Typically, cells are pulse-labeled with tagged nucleotide analogs, genomic DNA is purified, stretched on glass coverslips by DNA combing<sup>[7]</sup> or other techniques,<sup>[8]</sup> and the tagged nucleotides are detected by several layers of antibodies.<sup>[7,8]</sup> The labeled DNA stretches reveal fork progression during the pulse and allow to infer origin positions and fork speeds, but contain no sequence information. This information can be obtained by hybridizing fluorescent DNA probes,<sup>[9,10]</sup> but only very few DNA molecules contain the probed locus. Months of work is required to collect and analyze a statistically significant sample of molecules for a single mammalian locus.<sup>[11,12]</sup>

We recently described a novel, straightforward method for optical mapping of DNA replication (OMAR) using replication-competent *Xenopus* egg extracts, a system that relies on the same proteins as, and recapitulates most aspects of, cellular DNA replication.<sup>[13]</sup> Bacteriophage  $\lambda$ -DNA was replicated in egg extracts in the presence of a fluorescent deoxyuridine triphosphate (dUTP), purified, nicked with a site-specific nicking endonuclease (NE), nick-labeled with another fluorescent dUTP, stained with YOYO-1, and combed. Direct epifluorescence revealed, in three distinct colors, the DNA molecules,

## 1. Introduction

DNA replication is a crucial biological process that ensures accurate transmission of genetic information to daughter cells.<sup>[1]</sup> Eukaryotic organisms replicate their genome from multiple start sites, termed replication origins, that are stochastically activated in S phase to establish bidirectional replication forks that progress along the template until converging forks merge.<sup>[1–4]</sup> Understanding the regulation of DNA replication is essential as perturbations of this process, referred to as replication stress,

Dr. F. De Carli, N. Menezes, Dr. A. Genovesio, Dr. O. Hyrien  
IBENS

Département de Biologie  
Ecole Normale Supérieure  
CNRS

Inserm

PSL Research University  
46 rue d'Ulm, F-75005 Paris, France


E-mail: auguste.genovesio@ens.fr; hyrien@biologie.ens.fr

W. Berrabah, Dr. V. Barbe

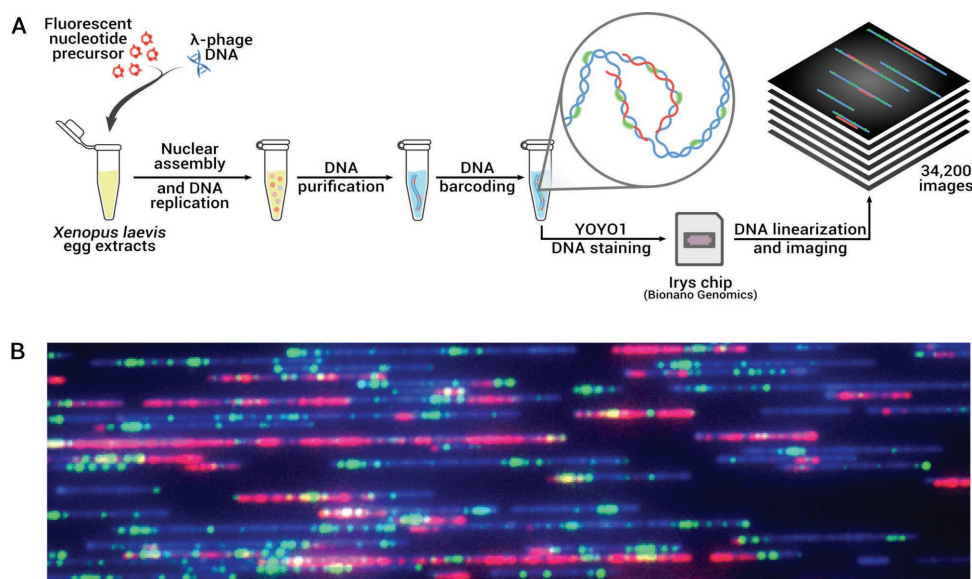
Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA)  
Génoscope

Institut de biologie François-Jacob

2 rue Gaston Crémieux, CP 5706, 91057 EVRY cedex, France

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/smt.201800146>.

DOI: 10.1002/smt.201800146



**Figure 1.** Experimental approach. A) Bacteriophage  $\lambda$ -DNA (48.5 kb) was replicated for 3 h in *Xenopus laevis* egg extracts supplemented with AF647-dUTP (red). Purified DNA was nick-labeled at Nt.BspQI sites with AF546-dUTP (green), stained with YOYO-1 (blue), stretched in nanochannels, and imaged using the Irys system (Bionano Genomics Inc.) yielding 34 200 FOVs containing a total of 63 890 Mb of DNA molecules (28 860 Mb in molecules > 90 kb). B) An exemplary merged image of stretched, 3-color fluorescent DNA molecules spanning an entire FOV (270 kb) in the direction of stretching.

their replication tracts, and their NE sites, allowing alignment to a reference genome.<sup>[13]</sup> The irregular surface deposition of combed DNA molecules, nevertheless, remained refractory to robust automated analysis. However, elongation of DNA in parallel nanochannel arrays has recently emerged as a potential alternative to DNA combing<sup>[14]</sup> and a commercial automated platform, the Bionano Genomics Irys system, demonstrated ability to assemble large numbers of nick-labeled DNA molecules into genomic maps.<sup>[15,16]</sup>

Here, we have exploited the Irys system for high-throughput optical mapping of replicating DNA (HOMARD) using the same fluorescent labeling strategy as in OMAR. We developed all necessary computational tools to robustly extract replication signal with high throughput. We report the high-quality imaging, detection, and mapping of 43493 barcoded, replicating DNA molecules in *Xenopus* egg extracts. The results demonstrate the feasibility of high-throughput mapping of DNA replication at the SM level and provide an ultrahigh (23 311 $\times$ ) replication coverage of in vitro reconstituted  $\lambda$ -DNA minichromosomes, revealing replication landscape with ground-breaking precision.

## 2. Results

### 2.1. Linearizing and Imaging DNA Replication Intermediates in Nanochannel Arrays

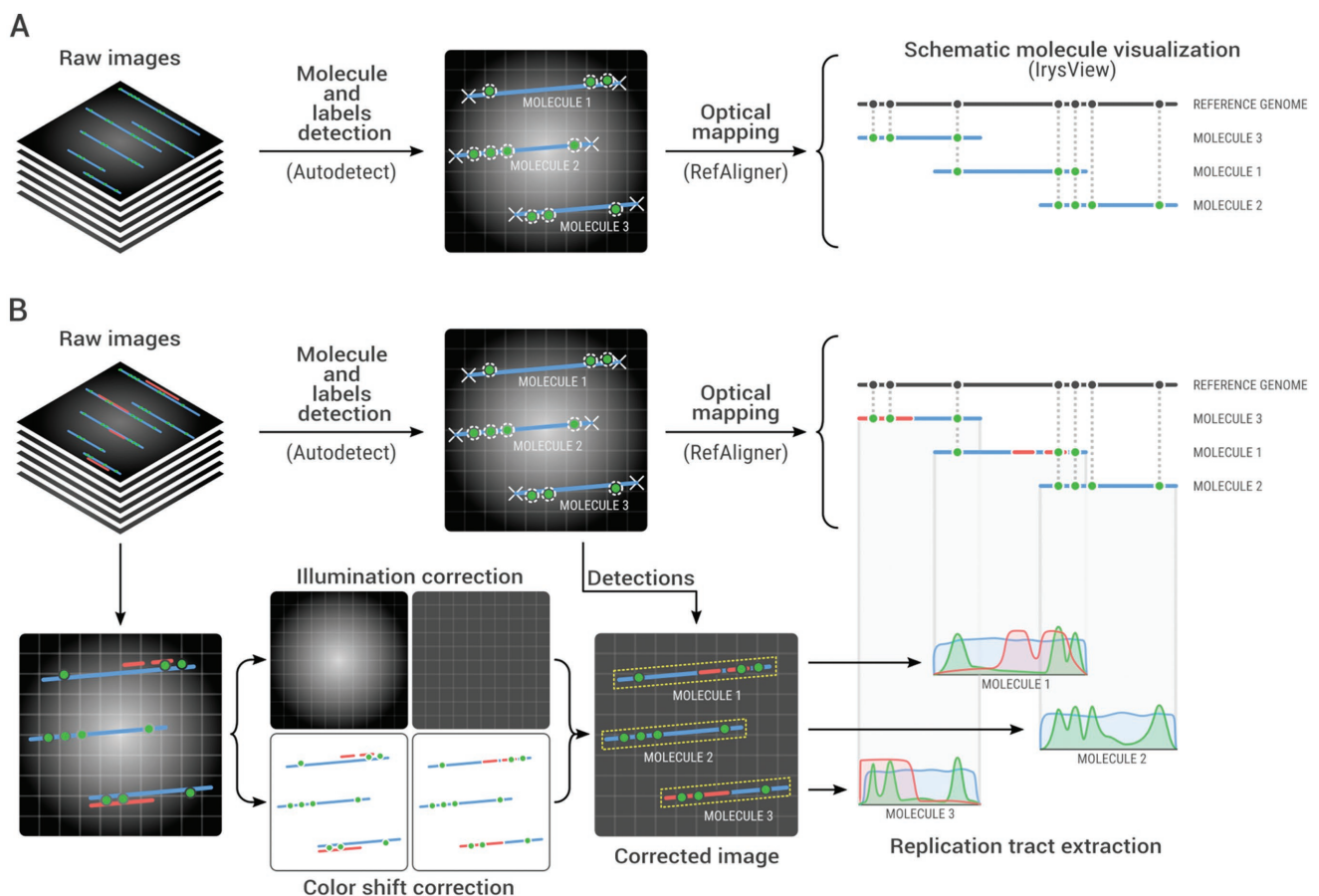
The experimental strategy for high-throughput visualization of nick-labeled DNA replication intermediates is shown in Figure 1A. Bacteriophage  $\lambda$ -DNA (48.5 kb) was replicated in *Xenopus* egg extracts in the presence of AF647-dUTP (red), purified, nick-labeled at Nt.BstQI sites with AF546-dUTP

(green), stained with YOYO-1 (blue) and run on a Bionano Genomics IrysChip. Automated imaging yielded 34 200 fields-of-view (FOVs) in 3 colors each per sample. A single FOV spans  $\approx$ 140 nanochannels of equivalent length to  $\approx$ 270 kb of B-form DNA stretched to  $\approx$ 85% of crystal length. Molecules of up to  $\approx$ 3 Mb may be imaged through 12 consecutively stitched FOVs.

A typical FOV section (Figure 1B) shows a mixture of unreplicated and replicating DNA molecules. A broad distribution of molecular sizes was observed, because linear  $\lambda$ -DNA monomers are actively end-joined in concatamers in egg extracts.<sup>[13]</sup> Replicating DNA molecules showed alternating replicated and unreplicated DNA stretches 5–50 kb in size, as previously observed by DNA combing.<sup>[13]</sup> Compared to DNA combing<sup>[13]</sup> or to earlier nanochannel experiments,<sup>[14]</sup> the advantages of the Irys system were the robust and highly uniform linearization of DNA, the lower background and the smoother signals, allowing high-throughput data collection and analysis.

### 2.2. A Robust Pipeline for High-Throughput Replication Tract Imaging

The standard Irys visualization pipeline is shown in Figure 2A. Raw images are analyzed with Autodetect to get DNA molecules' coordinates (start; end) and barcode label positions. Schematized molecules are aligned to the reference genome using IrysView and RefAligner. Standard genome mapping applications of the Irys system only require two-color images (DNA, blue + barcode, green), although the third color (red) has been occasionally used for dual red/green nick-labeling experiments.<sup>[17]</sup>

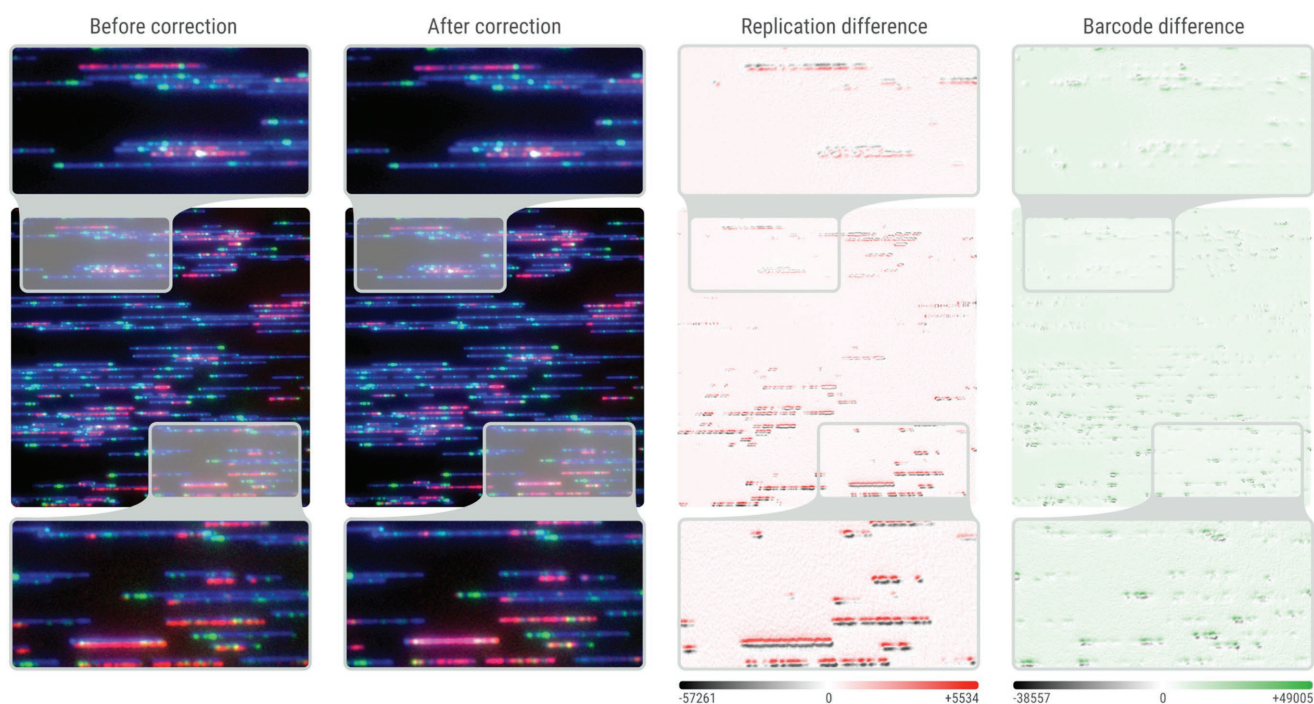


**Figure 2.** Bioinformatics for high-throughput optical mapping of replicating DNA (HOMARD). A) The current workflow for Irys data analysis and visualization consists of 3 steps. First, the DNA molecules and their labels (barcode) are automatically detected using Autodetect; second, individual barcoded DNA molecules are aligned to the reference genome using RefAligner; third, schematic results are visualized using IrysView. B) A custom pipeline was developed to output SM intensity profiles (bottom pipeline). Raw 3-color channels images (34 200 per run) are corrected for chromatic shifts between the three channels (blue = DNA, red = replication, green = barcode) and for illumination inhomogeneities. Corrected images display red replication tracts and green NE labels realigned onto their original DNA molecule (blue) over a homogeneous low background. Then, the Autodetect and RefAligner coordinates of each molecule are used to automatically extract the three color intensity profiles and align them to the reference genome.

Here we took advantage of the third channel's availability to image replication tracks and we developed a dedicated pipeline for HOMARD grounded on the original Bionano pipeline (Figure 2B). After DNA molecules and barcodes were detected and mapped using Autodetect and RefAligner, custom Python scripts were developed to correct images for inhomogeneous illumination and for chromatic shifts as well as to extract 3-color intensity profiles for individual DNA molecules at high accuracy and throughput.

The raw images presented systematic inhomogeneous illumination with stronger intensity values at their center. This effect was particularly clear when computing the median at each location ( $x,y$ ) of the blue, red, and green image channels of each scan (Figure S1, Supporting Information). Those median images per scan showed channel-specific intensity ranges and shapes. For each scan, the blue channel had lower pixel intensities and a flatter shape than the red and green ones. As median images slightly varied between scans, we corrected the raw images by subtracting the median image of the corresponding scan, which successfully flattened images of the entire run.

Once the illumination biases were corrected, the composite images produced by the Irys system revealed a significant chromatic shift. The blue (DNA), green (barcode), and red (replication) signals of each molecule were not exactly aligned to each other (Figure 3). The misalignments were low at image center but gradually increased toward the image borders and were more pronounced between red and blue than between green and blue. This observation was consistent with transverse chromatic shift. As molecules were automatically detected on the blue (YOYO-1) channel, it was necessary to rectify the coordinates of the red (AF647) and green (barcode) pixels in order to assign them to the correct DNA molecule. To align the channels, we used an affine registration approach where channels were assumed to be related to one another by an affine transform. Coefficients of a transformation matrix between channels were identified as minimizing the sum of squared error of a pair of feature point sets (see the Experimental Section). Such an affine transform comprises all global distortions that preserve lines, such as magnification, rotation, and shift. The misalignment was satisfactorily corrected for most image pairs, as shown in Figure 3.



**Figure 3.** Chromatic shift correction. Raw 3-color images present chromatic shifts between the red, green, and blue channels. As a result, replication and barcode signals are shifted with respect to the DNA signal (before correction). For all 34 200 images of a run, red and green images are automatically corrected to match the unaltered blue image (after correction). Chromatic shifts occur on both X and Y axes, and are stronger at image borders (zoomed-in molecules inside grey rectangles). Subtracting the uncorrected from the corrected red (replication difference) and green (barcode difference) images highlights the spatial chromatic shift gradient.

### 2.3. High-Throughput Mapping and Profiling of Replicating DNA Molecules

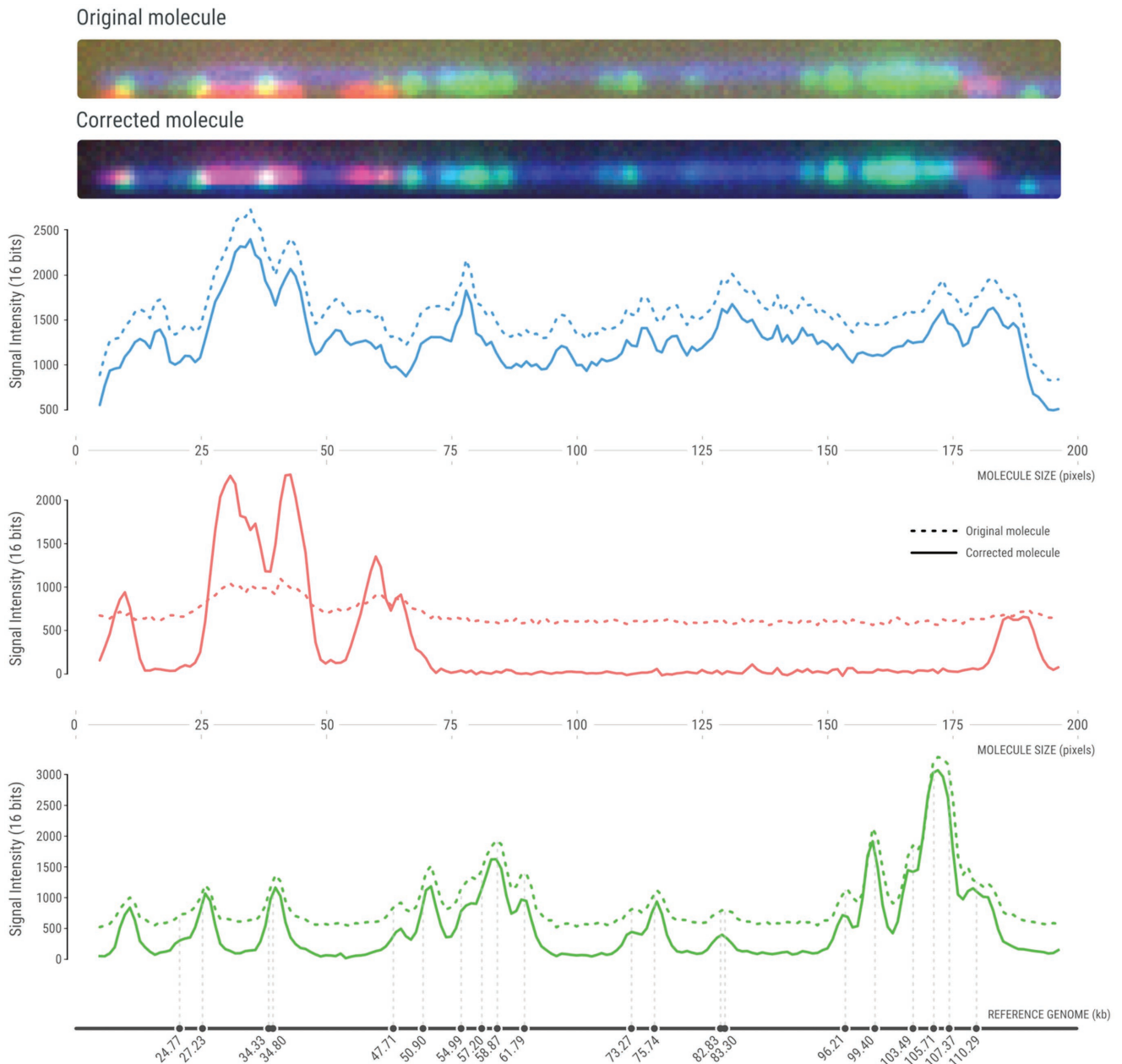
Once all images were corrected for color shift and illumination biases, we used 1) the molecule coordinates provided by Autodetect to extract SM intensity profiles from them (taking into account the tilting of DNA molecules) and 2) the optical mapping results from RefAligner to align them onto their genomic coordinate (see the Experimental Section).

Autodetect reported 884 832 DNA molecules  $\geq 20$  kb totaling 63 245 Mb. Only DNA molecules  $\geq 90$  kb with label signal-to-noise ratio (SNR)  $\geq 2.75$  (static) were used for mapping (210 147 molecules totaling 28 860 Mb). Because  $\lambda$ -DNA monomers are end-joined in complex concatamers in egg extracts,<sup>[13]</sup> we built a synthetic reference genome composed of 30  $\lambda$ -DNA molecules in a head-to-tail tandem array (1455 Kb). Using RefAligner, 33.1% ( $N = 69203$ ) of the selected 209 051 molecules were mapped to the synthetic genome. The remaining molecules corresponded to more complex concatamers containing head-to-head and tail-to-tail junctions between  $\lambda$ -DNA monomers and/or incomplete monomers. For this report, we focused on molecules entirely contained in a single FOV ( $N = 120 793$ ), of which 36.0% were mapped ( $N = 43 493$ ).

Similar stretching of replicated and unreplicated segments is mandatory for unbiased mapping and analysis of replication intermediates. To check if any bias could have occurred at this level, we partitioned the 120 793 single-FOV molecules into unreplicated ( $N = 90 831$ ) and replicating ( $N = 29 962$ ) molecules based on their computed mean red signal intensity.

The two datasets showed similar molecule size distributions and barcode label densities. Replicating intermediates were mapped to their in silico reference genome with only slightly lower efficiency (29.4%) than unreplicated molecules (38.2%). Therefore, 1) AF647-labeled tracts did not prevent nick-labeling with AF546-dUTP; 2) replication forks did not prevent DNA entry into nanochannels; 3) the close juxtaposition of sister duplexes in nanochannels did not result in different stretching of replicated and unreplicated DNA segments.

**Figure 4** shows intensity profile extraction and genome positioning of an exemplary, 120 kb long replicative DNA molecule. The raw image presented a downward shift of the red and green signals relative to the blue signal. Chromatic shift correction realigned the signals which increased the signal-to-noise ratio of the red and green intensity profiles. Illumination correction lowered the baseline of the three profiles, due to subtraction of the median background image, which further increased the signal-to-noise ratios. Note that without chromatic correction, the rightmost replication bubble on Figure 4 would have been missed. A complementary example of false signal collection from an adjacent DNA molecule in the absence of chromatic shift correction is shown in Figure S2 (Supporting Information). The molecule's profiles were aligned to the reference genome by RefAligner based on the corrected green profile. Note that the green profile confirms the head-to-tail tandem orientation of two  $\lambda$ -DNA monomers. The two monomers were identically stretched despite the presence of replication bubbles on the first monomer but not the second one (Figure 4).



**Figure 4.** Single-molecule intensity profiles from a replicative DNA molecule. Image crops and fluorescent intensity profiles (blue, DNA; red, replication; green, barcodes) of a representative replicative DNA molecule before (top image, dotted line profiles) and after (bottom image, solid line profiles) chromatic shift and illumination correction. The length of the molecule is indicated in pixels under the blue and red profiles. The reference genome coordinate is indicated in kb under the green profile, with vertical dashed lines connecting the green profile peaks with their in silico map position.

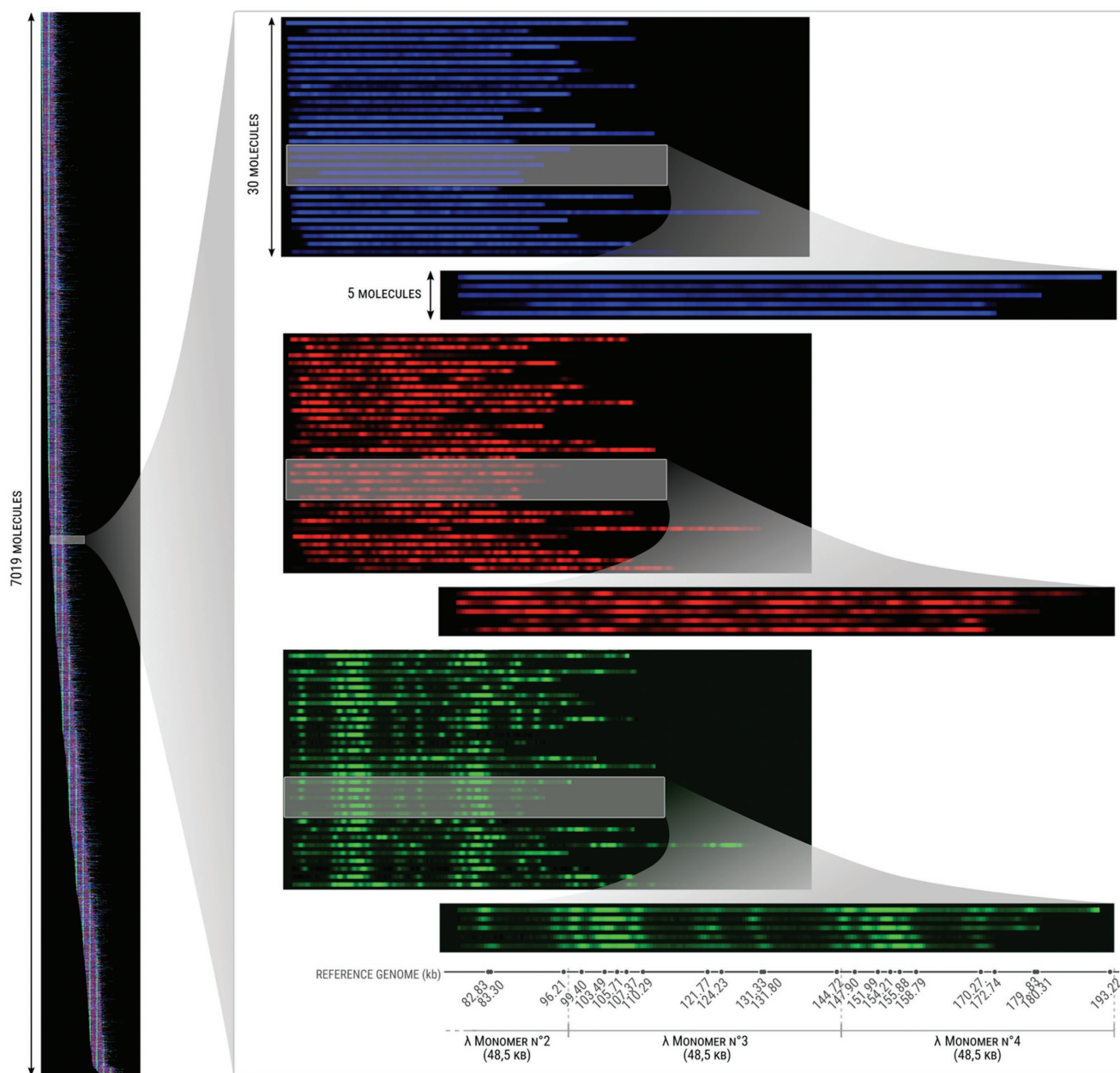
We then created a software to simultaneously visualize the color-coded intensity profiles of all mapped molecules. **Figure 5** shows the full collection of 7019 replicating  $\lambda$ -DNA concatamers mapped to (and included in) their synthetic reference genome and ordered by molecule start position. The correct genome alignment of the concatamers is visible from the repetitive pattern of vertical lines in the green channel (barcode). Half of the DNA molecules mapped to the first 7  $\lambda$ -DNA monomers of the synthetic genome and all the molecules mapped to the first 14 monomers. Enlarged views of the aligned molecules

illustrate, as expected, their continuous DNA signal (blue), discontinuous and irregular replication signal (red), and repetitive barcode signal (green) aligned with Nt.BstQI sites on the synthetic genome map.

#### 2.4. Sequence-Dependence of $\lambda$ -DNA Replication Signal

Several studies reported long ago that replication initiation is independent of DNA sequence in *Xenopus* eggs, egg

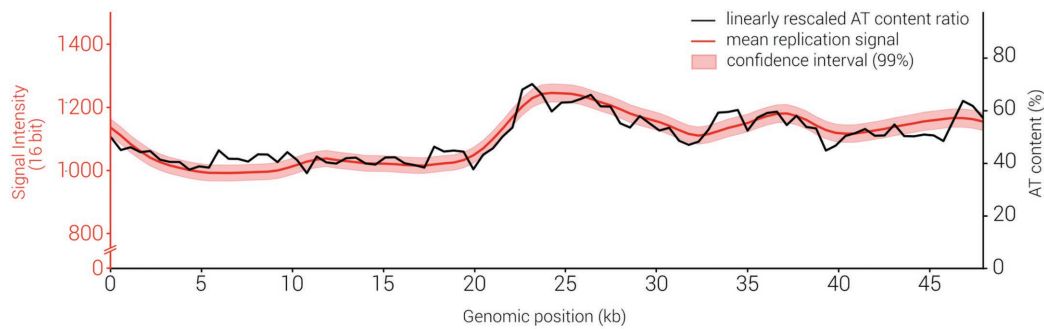




**Figure 5.** High-throughput, single-molecule replication analysis. Left: for each of the 7019 replicating  $\lambda$ -DNA concatamers mapped to the synthetic, head-to-tail concatameric  $\lambda$  genome, the intensity signals collected from the three channels were used to create a horizontal, synthetic tricolor profile of 3 pixels height. The 7019 tricolor profiles were mapped to their genomic location and vertically ranked by molecule start position. Blue, DNA; red, replication; green, barcodes. A repetitive pattern of green (barcode) vertical lines is visible, as expected for the successful optical mapping of  $\lambda$ -DNA concatamers. All molecules were mapped to the first 14 monomers of the 30 monomer-long synthetic genome. Right: consecutive zoomed-in views of selected regions indicated by shaded rectangles, with the three colors separately shown. The corresponding section of the synthetic genome map is aligned below the enlarged green molecule images. Nt.BstQI sites are indicated by black dots. Numbers below each dot indicate their map position on the synthetic  $\lambda$ -DNA concatamer genome.

extracts, and early embryos.<sup>[18]</sup> We recently confirmed this conclusion by OMAR analysis of  $\lambda$ -DNA in egg extracts.<sup>[13]</sup> However, it was reported that the *Xenopus* origin recognition complex (ORC) preferentially binds to asymmetric adenine + thymine (AT)-rich sequences<sup>[19]</sup> and a standard DNA combing study reported a twofold preference for replication initiation at asymmetric AT-rich DNA sequences in  $\lambda$ -DNA in egg extracts.<sup>[20]</sup> Here, the mapping of 7019 replicating  $\lambda$ -DNA

concatamers, totaling 23 311  $\lambda$ -DNA monomers, allowed us to reinvestigate this question with unprecedented precision. Since AF647-dUTP incorporation only occurs opposite to adenines on the DNA template, the signal from replicated tracts, if they are evenly distributed along the template, was expected to be strictly proportional to the local AT-content. Accordingly, we found that the profile of the mean red signal intensity of the 23 311  $\lambda$ -DNA monomers closely followed the AT-content



**Figure 6.** The mean replication signal of replicating  $\lambda$ -DNA molecules closely follows the local AT-content. Black line, AT-content of the  $\lambda$ -DNA sequence computed by nonoverlapping 537 bp windows. Thin red line, mean replication signal (on a 16-bit scale) of 23 311  $\lambda$ -DNA monomers from 7019 replicating concatamers. Thick red line, 99% confidence interval after Bonferroni correction.

profile (**Figure 6**: Pearson correlation  $R = 89\%$ ). Only small (<10%) albeit statistically significant ( $P < 10^{-8}$ ) deviations were observed between the two profiles. These results confirm the lack of strong sequence preference previously observed with less precise methods.<sup>[18]</sup> We did not observe preferential initiation in the AT-rich moiety of  $\lambda$ -DNA, which might have been previously observed<sup>[20]</sup> due to measurement of bromodeoxyuridine incorporation in the absence of normalization by AT-content. Nevertheless, our results also raise the possibility that replication tracks are not distributed with absolute uniformity. Further analysis will reveal if these small deviations reflect an unanticipated subtle sequence dependence of  $\lambda$ -DNA replication initiation or elongation in *Xenopus* egg extracts.

### 3. Discussion

Despite the importance of SM information for understanding DNA replication, current SM replication mapping methods have remained refractory to robust automation and are therefore low throughput.<sup>[11,12]</sup> We recently described a new method for OMAR that combines triple-color fluorescent labeling of DNA, replication tracks, and nicking endonuclease sites with DNA combing.<sup>[13]</sup> This method considerably simplified the labeling and mapping of DNA replication intermediates with respect to standard, antibody-based DNA combing, but remained difficult to automate. In this work, we leveraged the potential of OMAR labeling strategy by combining it with the Bionano Genomics Irys system's power to automatically image and robustly map nick-labeled DNA molecules at high throughput. Compared to standard SM replication mapping techniques, HOMARD drastically accelerated data collection (from several months to a few days) and enormously increased dataset size (from a few hundreds to >200 000 molecules in a single experiment). This was made possible thanks to the superior DNA linearization and imaging capacities as well as the dedicated tools, either provided by the Irys system for molecule detection and mapping, or additionally developed here for correction of uneven illumination and chromatic shifts and for extraction and alignment of replication signals.

Importantly, we demonstrate that DNA replication intermediates can be driven and extended into Irys nanochannel arrays

with comparable efficiency to linear, nonreplicating DNA molecules. Furthermore, we observe a similar stretching of unreplicated and replicated DNA along replication intermediates, which was mandatory to map them based on distance measurements between consecutive nick-labels. Thus, chasing replication forks following replicative track labeling was not required to allow proper entry and stretching of DNA into nanochannels.

The standard Bionano Genomics pipeline can be used to detect DNA replication intermediates and their nick labels but is not suited to detect labeled replication tracks. Although the Autodetect software can detect both dotted (green nick labels) and elongated (blue DNA molecules) objects, our attempts to detect red replication tracts using either option were unsatisfactory. When using the molecule's detection option, small red tracts were missed. When using the nick-label option, the multiple detected red dots did not accurately represent the true extension of the replicative signals. We therefore developed scripts to correct for illumination defects and chromatic shifts, and to automatically output tricolor intensity profiles of multiple molecules aligned to their genomic map. In doing so, we found that image correction prior to intensity profile extraction increased the SNR and avoided signal omission or erroneous collection from adjacent molecules. Finally, our approach facilitates downstream analyses by compressing 2D signals to 1D intensity profiles.

Purified DNA added to *Xenopus* egg extracts is chromatinized and replicated in a manner that extraordinarily faithfully mimics DNA replication in the early embryo.<sup>[21]</sup> Taking advantage of this system, we produced an ultrahigh coverage (23 311 $\times$ ) single-molecule replication map of eukaryotic minichromosomes made of tandem  $\lambda$ -DNA concatamers. The mean replicative signal from local incorporation of fluorescent dUTP was correlated with unprecedented precision to the local AT-content, suggesting potential for optical sequencing. Because individual molecules are not fully replicated, this almost exact correlation is consistent with a lack of any sequence preference for replication bubble location, as observed in previous studies in *Xenopus* eggs and egg extracts and embryos,<sup>[13,18]</sup> although other studies suggested preferential initiation at asymmetric AT-rich elements.<sup>[19,20]</sup> Asymmetric AT-rich sequences constitute a strong nucleosome excluding signal<sup>[22]</sup> and the establishment of nucleosome-free regions is currently speculated to be a general property of replication initiation in all eukaryotes.<sup>[23]</sup>

Even though ORC may preferentially bind AT-rich nucleosome free gaps in egg extracts,<sup>[19]</sup> our results are consistent with the notion that the ultimate distribution of potential initiation sites is determined by the location of MCM2-7 complexes spread away around ORC during origin licensing rather than by ORC itself.<sup>[18,24,25]</sup> Further analyses will reveal whether the small yet significant deviations between replicative signal and AT-content reflect a subtle sequence dependence of replication initiation or elongation and will explore potential functional correlations between neighboring origins and forks.

HOMARD requires fluorescent metabolic labeling of replication forks. This is convenient to perform in *Xenopus* egg extracts by direct addition of fluorescent dNTPs. Attempts to use click-chemistry to fluorescently label EdU (5-ethynyl-2'-deoxyuridine) incorporated during DNA replication in human cells resulted in excessive DNA breakage for nanochannel analysis.<sup>[14]</sup> By contrast, permeabilization procedures that allow fluorescent dNTP entry into living mammalian cells<sup>[14,26]</sup> have been successfully used for DNA combing<sup>[27]</sup> and nanochannel<sup>[14]</sup> analysis. Furthermore, a human cell-free system that supports fluorescent dNTP labeling has been used to study replication by DNA combing and other methods.<sup>[28–30]</sup> The nick-labeling strategy employed here is routinely used for optical mapping of the human genome.<sup>[15,16]</sup> Therefore, extension of HOMARD to the human genome should be straightforward.

The Irys system is currently limited to three fluorescent channels. Two channels are required for DNA and nick label detection, which leaves only one channel for replicative labeling. A single color pulse of cells can interrogate early-firing origins in synchronized cells, but analysis of the entire S phase requires a sequential labeling scheme to orient replication forks and distinguish elongation from initiation and termination tracts. Consecutive <sup>3</sup>H-thymidine pulses of different concentrations were historically used to demonstrate bidirectional replication by DNA fiber autoradiography in *E. coli*<sup>[31]</sup> and in mammalian cells.<sup>[32]</sup> More recently, DNA combing of ATTO 633-dUTP-labeled cells showed oriented tracts of decreasing intensity consistent with consumption of the label during fork progression.<sup>[27]</sup> Finally, distinguishing replicated and unreplicated DNA segments by their different YOYO-1 fluorescent intensity<sup>[13,33]</sup> may allow to orient pulsed fluorescent dNMP tracks. Monocolor replication analysis is therefore feasible but will necessitate dedicated automated analysis tools that we plan to develop. Additional fluorescent channel incorporation in future optical mapping devices would further extend the range of possibilities.

## 4. Conclusion

To conclude, we demonstrate that the Bionano Genomics Irys system, a powerful optical DNA mapping device so far used for genome assembly and structural variation analysis, can be harnessed for high-throughput, single molecule analysis of DNA replication. We provide by far the highest coverage single-molecule replication map ever produced and confirm with unprecedented statistical significance the lack of sequence preference for replication initiation in *Xenopus* egg extracts. This work paves the way for future genome-wide, single-molecule

analyses of eukaryotic DNA replication in multiple organisms. In addition, the tools we developed may be used for genome-wide, single-molecule studies of other epigenetic features such as cytosine methylation.

## 5. Experimental Section

**Preparation of Labeled DNA Replication Intermediates for Nanochannel Analysis:** Replication of  $\lambda$ -DNA in the presence of  $20 \times 10^{-6}$  M AlexaFluor647-aha-dUTP (AF647-dUTP) in *Xenopus* egg extracts and purification of the DNA in agarose plugs was as described previously,<sup>[13]</sup> except that pelleted nuclei from  $5 \times 50 \mu\text{L}$  replication reactions were pooled together in a single agarose plug and that proteinase K digestion was performed within a fivefold larger volume of the same digestion buffer. Subsequent processing steps differed from our previous protocol.<sup>[13]</sup> The agarose plug was rinsed five times in TE (Tris-Cl  $10 \times 10^{-3}$  M, ethylene diamine tetracetate (EDTA)  $1 \times 10^{-3}$  M, pH 8.0), molten at 70 °C for 2 min and digested at 43 °C with one unit agarase and dialysed against TE for 6 h. Nicking at Nt.BspQI sites, nick-translation with AF546-dUTP, YOYO-1 DNA staining, and sample run were then performed according to Bionano Genomics specifications for the Irys system. The procedures for experimental use and care of *X. laevis* frogs were approved by the Comité d'éthique Charles Darwin n°5 and the Ministère de l'Éducation Nationale, de l'Enseignement et de la Recherche (Project authorization APAFIS#696).

**Data Collection:** DNA molecules were linearized and imaged using the Bionano Genomics Irys system. Each Irys nanofluidic chip contains two flow cells,  $\approx 13\,000$  nanochannels each. DNA molecules were driven by electrophoresis into the nanochannels where they were rectilinearly stretched by physical confinement. Molecules were held still by stopping electrophoresis, automatically imaged (1140 images per scan) and electrophoresis was resumed to image new DNA molecules. The cycle was repeated 30 times yielding a total 34 200 images per run per color. DNA molecules (outlined by YOYO-1 staining) and locations of fluorescent nick-labels along each molecule were detected using the AutoDetect software (v 2.1.4.9159) provided by Bionano Genomics.

**Illumination and Chromatic Shift Corrections:** To correct raw images for inhomogeneous illumination, single pixel median of red, green, and blue images were computed for each of the 30 scans composing a run and subtracted from each of the 1140 images of the same scan. To correct for chromatic shift, the 15% brightest, illumination-corrected images of each scan (171 out of 1140 scan images) were selected, which were enriched in foreground (i.e., molecules) and thus more informative. From this image subset (5130 out of 34 200 run images) the sets of points corresponding to local maxima for each of the three channels were extracted and the closest points were paired for both the red/blue and the green/blue pairs. Then, for each image and each color pair, the affine transformation matrix was determined that minimized the distances between the paired points. Matrix parameters clustered around very similar median values for each scan of a run. Therefore, the median parameter values of the total run were computed to define a unique transformation matrix that was then applied to the 34 200 images of the run.

**Optical Mapping:** To map  $\lambda$ -DNA concatamers, a synthetic reference genome was built composed of 30  $\lambda$ -DNA genomes in head-to-tail concatamer (48 502 bp each, 1 455 060 bp in total). The fasta genome was in silico digested to Bionano cmap using Knickers software (v 1.5.5). Optical mapping was performed using the IrysView Genomic Analysis Viewer (v 2.5.1.29842) and RefAligner (r5122) from Bionano Genomics with the following parameters:

```
-nosplit 2 -BestRef 1 -biaswt 0 -Mfast 0 -FP 1.5 -FN 0.15 -sf 0.2 -sd 0.0
-A 5 -outlier 1e-3 -outlierMax 40 -endoutlier 1e-4 -S -1000 -sr 0.03 -se 0.2
-MaxSF 0.25 -MaxSE 0.5 -resbias 4 64 -maxmem 27 -M 3 3 -minlen 90 -T
1e-7 -maxthreads 4 -hashgen 5 3 2.4 1.5 0.05 5.0 1 1 3 -hash -hashdelta
10 -hashoffset 1 -hashmaxmem 64 -insertThreads 6 -matype 0 -PVres
2 -PVendoutlier -AlignRes 2.0 -rres 0.9 -resEstimate -ScanScaling
```

2 -RepeatMask 5 0.01 -RepeatRec 0.7 0.6 1.4 -maxEnd 50 -usecolor 1 -stdout -stderr.

Only DNA molecules  $\geq 90$  kb with label SNR  $\geq 2.75$  (static) were used for mapping.

**Intensity Profile Extraction:** Once the three color channels were corrected for color shifts, the intensity profiles along the length of each molecule were automatically extracted using both X and Y molecule coordinates identified by Autodetect (Xstart, Ystart—Xend, Yend) and a 3 px width-window, in order to correct for the observed slight tilting of molecules with respect to image borders and to fully enclose each molecule's contour without overlapping adjacent molecules. Results were exported in a plain table (MoleculeIntensityProfile.csv) and appended to the original Molecules.mol Bionano file for further use.

**Alignment of Single-Molecule Intensity Profiles to the Reference Genome:** To position SM intensity profiles to their genomic location, information was linked from several Bionano Genomics' output files to intensity profiles contained in MoleculeIntensityProfile.csv produced at the previous step. The ID of the first correctly mapped molecule's nick label (FirstLabel\_ID) was recovered and its genomic position (GenomPosStart) from the MoleculeQualityReport.xmap file generated after optical mapping and its distance from the molecule start from the Labels.lab file (FirstLabel\_Position) generated by Autodetect and corrected for chromatic shift. genomic coordinates of all molecule intensity profiles were then recovered by aligning the corrected FirstLabel\_Position onto the GenomPosStart.

**Statistical Testing:** To test if the local mean replication signal and AT-content of replicating  $\lambda$ -DNA monomers were linearly related to each other, the AT-content profile of the  $\lambda$  genome was computed with a 533 bp sliding window (estimated number of bp per pixel) and the following statistical test was devised. As per the central limit theorem, the sample size being very large, the mean of the replicative signal follows a Gaussian distribution at each position. A Gaussian confidence interval at 1% risk was then obtained from the estimated mean and standard deviation of the sample mean at each position. This confidence interval was subsequently adjusted for multiple testing with the Bonferroni correction.

**Availability:** The software developed for this study is available in the GitHub repository (<https://github.com/biocompibens/HOMaRD>). The single molecule data that support our findings are available from the corresponding authors upon reasonable request.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

F.D.C. and N.M. contributed equally to this work. The authors thank Gaël Millot (Institut Pasteur) for bringing our attention to the Bionano technology at an early stage, Erwan Denis, Ghislaine Magdelenat, and Caroline Belser (Genoscope) for their contribution to the production and analysis of optical mapping data, Felipe Delestro (IBENS) for help with Figure preparation, Benoît le Tallec (IBENS) and other members of the O.H. lab for helpful discussions. This work was supported by the Ligue Nationale Contre le Cancer (Comité de Paris), the Association pour la Recherche sur le Cancer, the Agence Nationale de la Recherche (ANR-15-CE12-0011-01), the Fondation pour la Recherche Médicale (FRM DE1201512344404), the Cancéropôle Ile-de-France and the INCa (PL-BIO), the programme "Investissements d'Avenir" launched by the French Government and implemented by the ANR (ANR-10-LABX-54 MEMOLIFE and ANR-10-IDEX-0001-02 PSL\*Research University), and the France Génomique national infrastructure, funded as part of the « Investissements d'Avenir » program (ANR-10-INBS-09).

## Conflict of Interest

The authors declare no conflict of interest.

## Keywords

DNA barcoding, DNA replication, nanochannel arrays, single-molecule analysis, Xenopus egg extracts

Received: May 9, 2018

Revised: July 6, 2018

Published online:

- [1] M. L. DePamphilis, S. D. Bell, *Genome Duplication*, Garland Science, London, New-York **2011**.
- [2] O. Hyrien, A. Rappailles, G. Guilbaud, A. Baker, C. L. Chen, A. Goldar, N. Petryk, M. Kahli, E. Ma, Y. d'Aubenton-Carafa, B. Audit, C. Thermes, A. Arneodo, *J. Mol. Biol.* **2013**, 425, 4673.
- [3] O. Hyrien, *J. Cell Biol.* **2015**, 208, 147.
- [4] J. M. Dewar, J. C. Walter, *Nat. Rev. Mol. Cell Biol.* **2017**, 18, 507.
- [5] S. Munoz, J. Mendez, *Chromosoma* **2017**, 126, 1.
- [6] N. Petryk, M. Kahli, Y. d'Aubenton-Carafa, Y. Jaszczyszyn, Y. Shen, M. Silvain, C. Thermes, C. L. Chen, O. Hyrien, *Nat. Commun.* **2016**, 7, 10208.
- [7] J. Herrick, P. Stanislawski, O. Hyrien, A. Bensimon, *J. Mol. Biol.* **2000**, 300, 1133.
- [8] D. A. Jackson, A. Pombo, *J. Cell Biol.* **1998**, 140, 1285.
- [9] P. Pasero, A. Bensimon, E. Schwob, *Genes Dev.* **2002**, 16, 2479.
- [10] P. Norio, C. L. Schildkraut, *Science* **2001**, 294, 2361.
- [11] R. Lebofsky, R. Heilig, M. Sonneleitner, J. Weissenbach, A. Bensimon, *Mol. Biol. Cell* **2006**, 17, 5337.
- [12] A. Demczuk, M. G. Gauthier, I. Veras, S. Kosiyatrakul, C. L. Schildkraut, M. Busslinger, J. Bechhoefer, P. Norio, *PLoS Biol.* **2012**, 10, e1001360.
- [13] F. De Carli, V. Gaggioli, G. A. Millot, O. Hyrien, *Int. J. Dev. Biol.* **2016**, 60, 297.
- [14] J. Lacroix, S. Pelofy, C. Blatche, M. J. Pillaire, S. Huet, C. Chapuis, J. S. Hoffmann, A. Bancaud, *Small* **2016**, 12, 5963.
- [15] E. T. Lam, A. Hastie, C. Lin, D. Ehrlich, S. K. Das, M. D. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, P. Y. Kwok, *Nat. Biotechnol.* **2012**, 30, 771.
- [16] H. Cao, A. R. Hastie, D. Cao, E. T. Lam, Y. Sun, H. Huang, X. Liu, L. Lin, W. Andrews, S. Chan, S. Huang, X. Tong, M. Requa, T. Anantharaman, A. Krogh, H. Yang, H. Cao, X. Xu, *GigaScience* **2014**, 3, 34.
- [17] A. R. Hastie, L. Dong, A. Smith, J. Finklestein, E. T. Lam, N. Huo, H. Cao, P. Y. Kwok, K. R. Deal, J. Dvorak, M. C. Luo, Y. Gu, M. Xiao, *PLoS One* **2013**, 8, e55864.
- [18] O. Hyrien, K. Marheineke, A. Goldar, *BioEssays* **2003**, 25, 116.
- [19] D. Kong, T. R. Coleman, M. L. DePamphilis, *EMBO J.* **2003**, 22, 3441.
- [20] S. Stanojic, J. M. Lemaitre, K. Brodolin, E. Danis, M. Mechali, *Mol. Cell. Biol.* **2008**, 28, 5265.
- [21] J. J. Blow, R. A. Laskey, *Int. J. Dev. Biol.* **2016**, 60, 201.
- [22] K. Struhl, E. Segal, *Nat. Struct. Mol. Biol.* **2013**, 20, 267.
- [23] M. L. Eaton, K. Galani, S. Kang, S. P. Bell, D. M. MacAlpine, *Genes Dev.* **2010**, 24, 748.
- [24] K. J. Harvey, J. Newport, *Mol. Cell. Biol.* **2003**, 23, 6769.
- [25] I. Lucas, M. Chevrier-Miller, J. M. Sogo, O. Hyrien, *J. Mol. Biol.* **2000**, 296, 769.

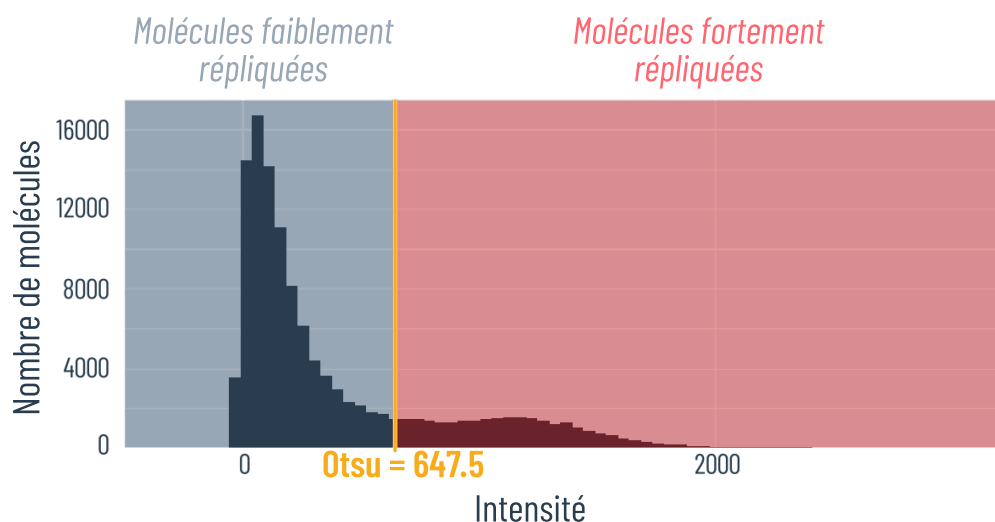
- [26] A. Maya-Mendoza, P. Olivares-Chauvet, F. Kohlmeier, D. A. Jackson, *Methods* **2012**, *57*, 140.
- [27] W. Xiang, M. J. Roberti, J. K. Hering, S. Huet, S. Alexander, J. Ellenberg, *J. Cell Biol.* **2018**, *217*, 1973.
- [28] T. Krude, *Cell Cycle* **2006**, *5*, 2115.
- [29] K. Marheineke, O. Hyrien, T. Krude, *Nucleic Acids Res.* **2005**, *33*, 6931.
- [30] A. R. Langley, S. Graf, J. C. Smith, T. Krude, *Nucleic Acids Res.* **2016**, *44*, 10230.
- [31] D. M. Prescott, P. L. Kuempel, *Proc. Natl. Acad. Sci. USA* **1972**, *69*, 2842.
- [32] J. A. Huberman, A. Tsai, *J. Mol. Biol.* **1973**, *75*, 5.
- [33] H. Yardimci, A. B. Loveland, S. Habuchi, A. M. van Oijen, J. C. Walter, *Mol. Cell* **2010**, *40*, 834.

#### 4.1.1.2 Conclusion

Dans cette publication, nous démontrons la robustesse de notre approche en fournissant une carte de la réplication de l'ADN du bactériophage  $\lambda$  dans les extraits d'œufs de *Xénope* avec un débit ultra-élevé (23 311 x). Grâce à cette couverture sans précédent et, de ce fait, à la puissance statistique, nous sommes parvenu à confirmer de manière quasi-incontestable que l'initiation de la réplication chez cet organisme n'était pas dépendante de la séquence comme certaines études avaient pu le montrer par le passé (STANOJCIC et al. 2008). Nous avons également réalisé l'analyse d'un réplicat biologique (n'apparaissant pas dans la publication) et avons tiré les mêmes conclusions. Par ailleurs, en plus de son utilisation standard dans le contexte de l'assemblage *de novo* de génome et l'analyse des variants structuraux, nous mettons en évidence le potentiel du système Irys pour d'autres types d'analyses. HOMaRD ouvre la voie à l'analyse pangénomique en molécule unique de la réplication de l'ADN, mais aussi à d'autres études de génomiques fonctionnelles ou encore d'épigénétique.

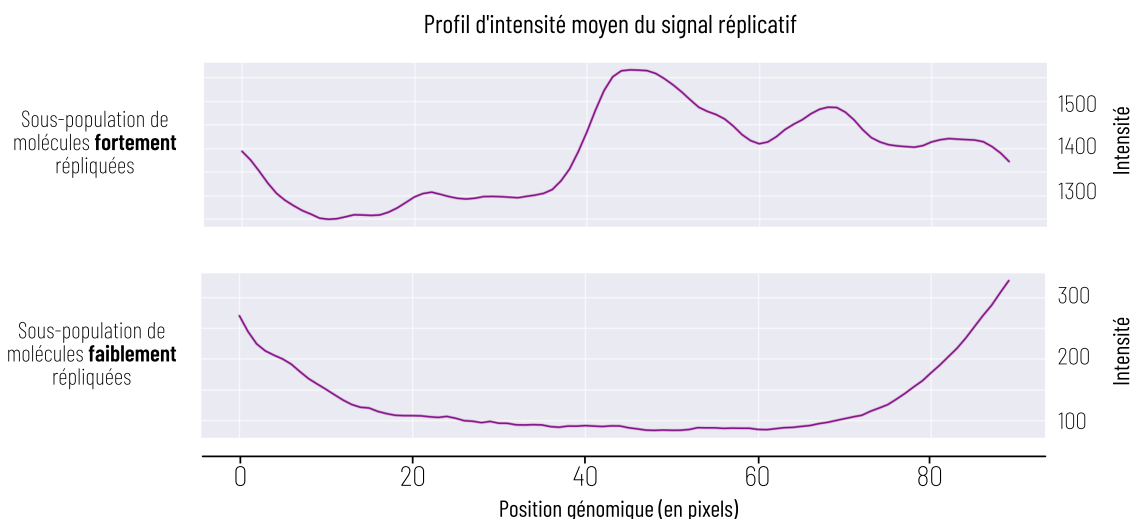
#### 4.1.2 Analyse de la sous-population des molécules faiblement répliquées

Nous nous sommes ensuite intéressés aux molécules d'ADN faiblement répliquées (cf. Figure 4.2). Nous avons séparé nos molécules d'ADN faiblement et fortement répliquées en appliquant un seuillage d'Otsu (OTSU 1979) sur la distribution des intensités moyennes du signal répliatif. Cette approche est généralement appliquée pour binariser une image et effectuer un seuillage automatique basé sur la forme de l'histogramme de l'image en question. L'algorithme derrière cette approche suppose que l'image n'est composée que de deux classes de pixels et consiste à minimiser la variance intra-classe. La valeur de seuillage d'Otsu appliquée à notre distribution nous permet ainsi d'obtenir nos deux sous-populations.



**Figure 4.2** – Distribution des moyennes des valeurs d'intensité des molécules d'ADN et détermination de la valeur de seuillage par la méthode d'Otsu. Pour déterminer les 2 sous-populations de molécules (faiblement et fortement répliquées), nous avons calculé les moyennes des valeurs d'intensité de chacune des molécules. À partir de la distribution, une valeur de seuillage a été déterminée par la méthode d'Otsu (seuil = 647,5).

Lorsque nous avons observé les profils moyens du signal réplcatif des deux sous-populations, nous avons constaté une différence assez flagrante. En effet, le profil moyen du signal réplcatif pour la sous-population faiblement répliquée suggère que la réplication est préférentielle au niveau des jonctions entre les molécules d'ADN de  $\lambda$ . Cette observation n'a jamais été faite auparavant. En effet, De Carli analyse par peignage moléculaire 167 molécules d'ADN de  $\lambda$  et montre que la densité d'initiation est homogène le long du génome de  $\lambda$  répliqué dans des extraits d'œufs de Xénope (DE CARLI, GAGGIOLI et al. 2016). Stanojcic et al. proposent que l'initiation de la réplication chez le Xénope se produit préférentiellement dans les régions riches en A/T (STANOJCIC et al. 2008) et cela sur un échantillon de 94 molécules. Étant donné la couverture que nous obtenons avec notre approche, nous avons exploré de façon plus approfondie cette population afin de vérifier si l'observation que nous avons faite au niveau de l'analyse en population (réplication préférentielle aux jonctions) était confirmée au niveau de la molécule unique.



**Figure 4.3 – Profils d’intensité moyen des 2 sous-populations.** Les profils d’intensité moyens des deux sous-populations (molécules faiblement (bas) et fortement (haut) répliquées) diffèrent aussi bien au niveau de leur forme générale qu’au niveau des valeurs d’intensité. Du fait de la forme en U du profil moyen d’intensité du signal réplcatif (bas), nous pouvons dire que la réplication de l’ADN du bactériophage  $\lambda$  débute au niveau des jonctions des concatémères.

#### 4.1.2.1 Partitionnement (*clustering*) des différentes configurations de jonctions des concatémères de Lambda

La cartographie de nos molécules pour ce jeu de données a été réalisée sur un "génomme de référence" constitué de tandems directs. Nous supposons donc que nos molécules cartographiées ne sont composées que d’ADN de  $\lambda$  *Forward-Forward*. Or, lorsque nous regardons de plus près, il existe des molécules d’ADN qui présentent l’ADN de  $\lambda$  dans des orientations différentes. RefAligner n’est donc pas suffisamment stringent pour éviter la cartographie de ces molécules d’ADN. Afin d’étudier la réplication de l’ADN au niveau des jonctions des  $\lambda$ , un partitionnement de nos données a été nécessaire. Nous partons avec l’*a priori* que nous ne connaissons pas les orientations de l’ADN de  $\lambda$  dans nos molécules et qu’au moins le premier  $\lambda$  des molécules cartographiées est orienté *Forward*.



## ETAPE 1 : ISOLER LES PROFILS D'INTENSITÉ

La première étape a donc consisté à isoler tous les ADN de  $\lambda$  entiers présents dans les molécules cartographiées en conservant leur position au sein de la molécule et l'identifiant de la molécule à laquelle ils appartiennent. Nous parvenons sur la sous-population des molécules faiblement répliquées à récupérer 16 085  $\lambda$ s.

ETAPE 2 : PARTITIONNER LES PROFILS D'INTENSITÉ DU CODE-BARRES PAR L'ALGORITHME DU *K-MEANS*

À partir de là, nous avons regroupé chaque molécule de  $\lambda$  à une orientation donnée, à savoir soit une orientation *Forward*, soit *Reverse*, en nous basant sur le signal du code-barres.

Pour effectuer le clustering, nous avons appliqué la méthode des K-moyennes (ou *K-means*) sur les profils d'intensité du code-barres (faisant 90 pixels) normalisés (aire sous la courbe égale à 1)(cf. Figure 4.5).

Le *K-means* est un algorithme non supervisé de clustering non hiérarchique (John A. HARTIGAN 1975, J A HARTIGAN et M. A. WONG 1979). Il commence par sélectionner  $k$  points pour l'initialisation des centroïdes. Le nombre de  $k$  points peut être sélectionné arbitrairement ou bien par une autre approche qui consiste à lancer *K-Means* avec différentes valeurs de  $k$  et calculer la variance des différents *clusters* (courbe "elbow", cf. Figure 4.4). La variance va correspondre à la somme des distances entre chaque centroïde d'un *cluster* et les différentes observations incluses dans le même *cluster*. Ainsi, le nombre de *clusters*  $k$  que nous cherchons à déterminer doit être choisi de telle sorte que les *clusters* retenus minimisent la distance entre leurs centres (centroïdes) et les observations dans le même *cluster*. Ensuite, deux étapes sont répétées itérativement :

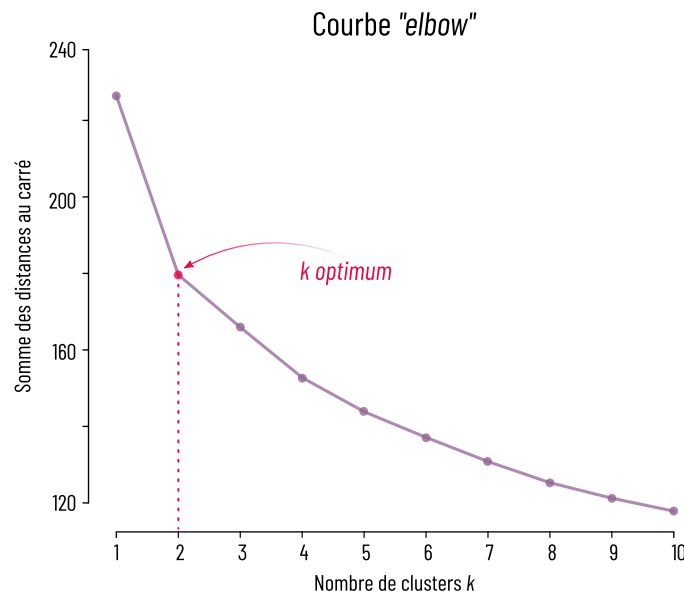
1. l'étape d'affectation : chacun des  $m$  points de notre ensemble de données est assigné à un *cluster* représenté par le plus proche des  $k$  centroïdes. Pour ce faire, la distance euclidienne de chacun des points au centroïde est calculée et les points sont liés au centroïde le plus proche.
2. l'étape de "mise à jour" : de l'étape précédente, nous avons un ensemble de points qui sont assignés à un *cluster*. Pour chacun de ces ensembles, nous calculons une moyenne qui nous permet de déclarer un nouveau centroïde du *cluster*.

À la suite de chaque itération, les centroïdes se déplacent, et la distance totale entre chaque point et le centroïde qui lui est assigné diminue. Les deux étapes décrites précédemment s'altèrent jusqu'à la convergence. La convergence de l'algorithme *K-Means* peut être due :

- soit au nombre d'itérations fixé à l'avance et dans ce cas l'algorithme réalisera les itérations et s'arrêtera qu'importe la forme des *clusters* trouvés
- soit à la stabilisation des centres de *clusters*.

L'algorithme de *K-means*, pour un même jeu de données, peut effectuer des partitionnements différents. En effet, la première étape renvoyant à l'initialisation des centroïdes est aléatoire. De ce fait, l'algorithme définira des *clusters* différents en fonction de cette première étape d'initialisation. Ainsi, l'algorithme du *K-Means* va avoir tendance à converger vers un optimum local. Afin de palier à cela, le *K-means* doit être lancé plusieurs fois sur le jeu de données (avec le même nombre  $k$  et des initialisations différentes) afin de voir la composition des *clusters* qui se forment pour ne garder que le résultat qui semble refléter le mieux nos données.

Pour notre *K-means*, étant donné que nous nous attendons à obtenir deux groupes, la valeur de  $k$  est définie à 2. Cependant, au vu de notre courbe "elbow", nous avons également testé le partitionnement avec un  $k$  égal à 3 (cf. Figure 4.4).



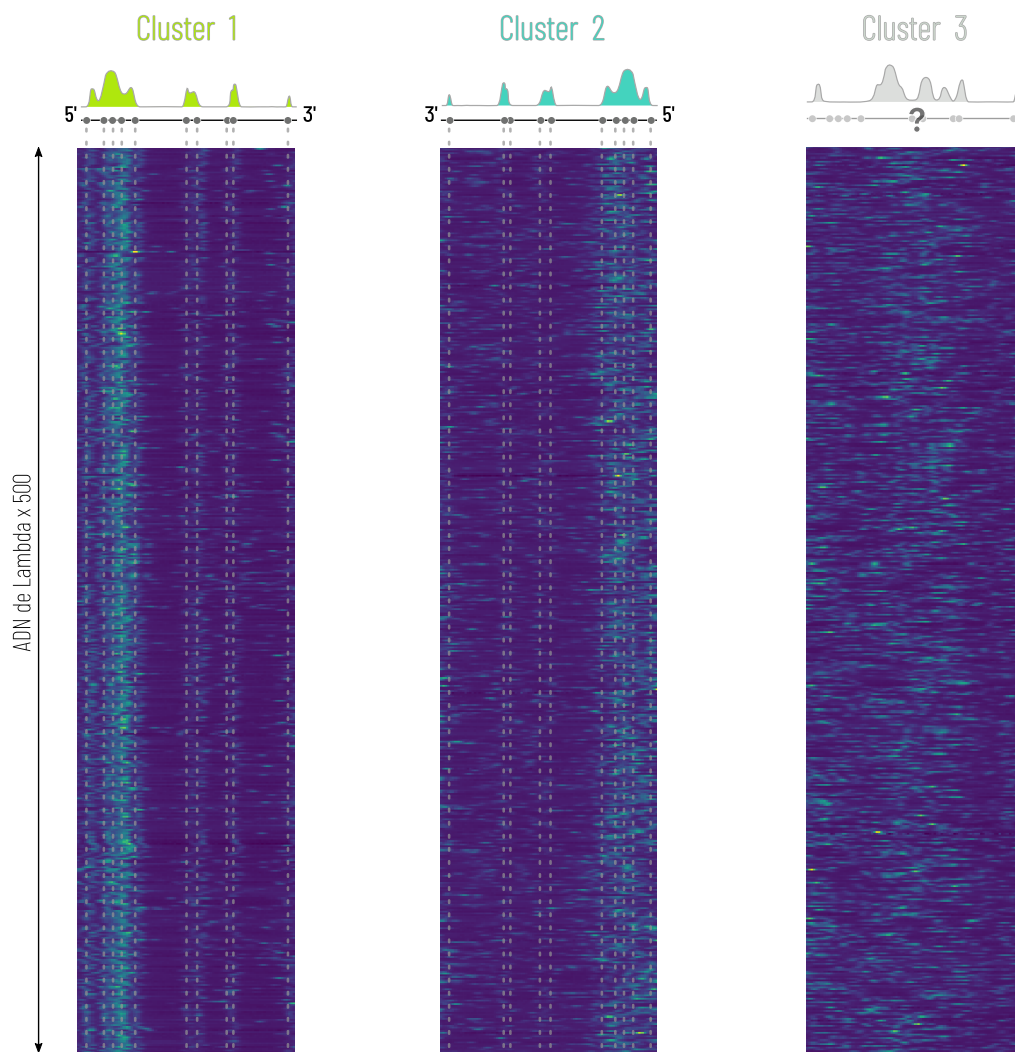
**Figure 4.4 – Courbe "elbow", détermination du  $k$  nombre de clusters.** Afin de déterminer le nombre de clusters  $k$  optimum pour effectuer la classification, nous avons calculé la somme des distances au carré pour des valeurs de  $k$  allant de 1 à 10. Sur ce graphique, ayant la forme d'un bras (de l'épaule en 1 à la main en 10), le point représentant le coude correspond au nombre optimum de clusters (d'où le nom *courbe "elbow"*).



**Figure 4.5 – Partitionnement des ADN de  $\lambda$ .** De (A) à (D) sont schématisés les étapes nécessaires au partitionnement de nos données. (A) Nous prenons comme exemple une molécule composée de 4 ADN de  $\lambda$ s (3 jonctions) avec le profil d'intensité du code-barres associé. Cette molécule a été cartographiée sur le génome de référence (tandem direct, 5'3'). À partir de là, les profils d'intensité sont isolés : la molécule est découpée en 4 fragments correspondant aux 4  $\lambda$ s qui sont tous anonymes (orientation inconnue). (B) En appliquant l'algorithme du K-means sur nos données, avec  $k$  (nombre de *clusters*) égal à 3, nous sommes capable de labelliser chacun de ces fragments (C). Le principe du K-means est présenté ici dans le cas de données ayant 2 dimensions mais dans notre cas les profils d'intensité des  $\lambda$ s font 90 dimensions. Le *cluster* vert renvoie aux fragments orientés *Forward* (fragments n°1 et n°2), le *cluster* turquoise correspond aux fragments orientés *Reverse* (fragment n°3) et enfin le dernier *cluster* (gris) correspond aux fragments n'appartenant à aucune de ces deux classes (fragment n°4). Enfin, nous reconstituons les différentes jonctions présentes au sein de notre molécule initiales (D), ici au nombre de trois : jonction *Forward-Forward*, *Forward-Reverse* et "inconnue".

## RÉSULTATS DU PARTITIONNEMENT DU JEU DE DONNÉES

Il se trouve que le nombre de *clusters* est de 3 avec l'ADN de  $\lambda$  orienté *Forward*, *Reverse* et le "reste" correspondant aux fragments ne pouvant être classifiés dans aucune des deux catégories (cf. Figure 4.6) avec , respectivement dans chaque classe, 8663 (55%), 2713 (16%) et 4706 (29%) fragments. Nous pouvons à présent reconstruire les jonctions et les identifier(cf. Figure 4.5).

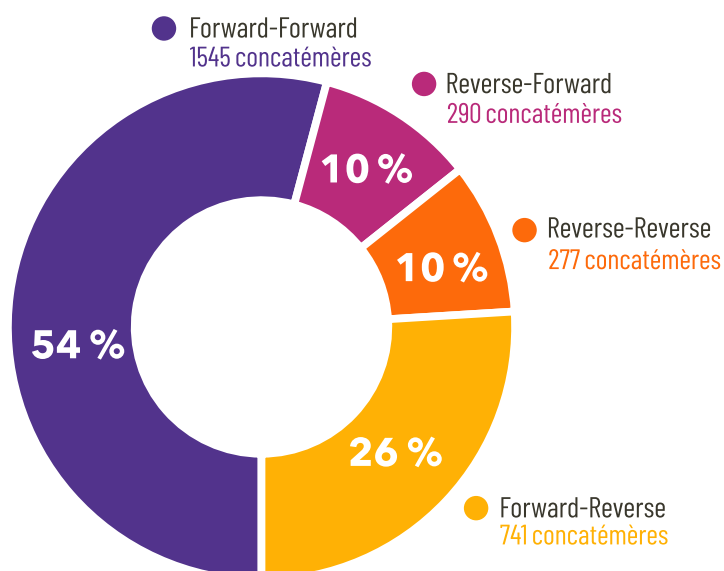


**Figure 4.6 – Résultat du partitionnement.** À la suite de la classification réalisée par l'algorithme du K-means avec  $k$  (nombre de *clusters*) = 3, nous obtenons le partitionnement suivant : le cluster 1 (vert) correspond aux  $\lambda$ s orientés *Forward*, le cluster 2 (turquoise) correspondant aux  $\lambda$ s orientés *Reverse* et enfin le cluster 3 (gris) renvoyant aux  $\lambda$ s n'appartenant à aucune des deux premières classes. Les images contiennent les profils d'intensité du code-barres de 500 fragments de molécules pour chacun des *clusters*.

#### 4.1.2.2 Une initiation de la réplication préférentielle au niveau des jonctions des ADN de Lambda

##### UN RÉSULTAT UNIQUE POUR DES TYPES DE JONCTION DIFFÉRENTS

Les jonctions des  $\lambda$ s présents dans chacune des molécules ont été recréées et 2853 jonctions sont récupérées (cf. Figure 4.7). Nous avons effectué la détection des segments répliqués sur ces différents concatémères. Le contenu en AT pris en compte dans la détections a été adapté en fonction des jonctions. Dans le cas des jonctions *Forward-Forward*, nous avons une symétrie quasi-parfaite pour le profil moyen du signal répliatif (brut) mais aussi pour les détections des segments répliqués. Pour ce qui est des jonctions *Forward-Reverse* et *Reverse-Forward*, nous avons une forme de U, moins prononcée, mais présente. Dans ces différentes jonctions, au moins un des  $\lambda$ s est orienté *Forward* comme le génome de référence. Par contre, la jonction *Reverse-Reverse* ne devrait normalement pas apparaître puisque la cartographie est effectuée sur un génome de référence composé de tandem direct. De plus, nous remarquons que le profil moyen du signal du code-barres pour cette jonction n'est pas aussi précis que pour les trois autres jonctions. Aucune conclusion ne sera donc tirée de cette dernière jonction (cf. Figure 4.8).



**Figure 4.7 – Proportions de jonctions récupérées.** À la suite de la classification nous recréons les différentes jonctions présentes dans chacune des molécules d'ADN cartographiées et considérées comme faiblement répliquées. Nous avons 2853 jonctions avec une majorité de jonctions *Forward-Forward*(1545 sur 2853), ce qui est attendu.

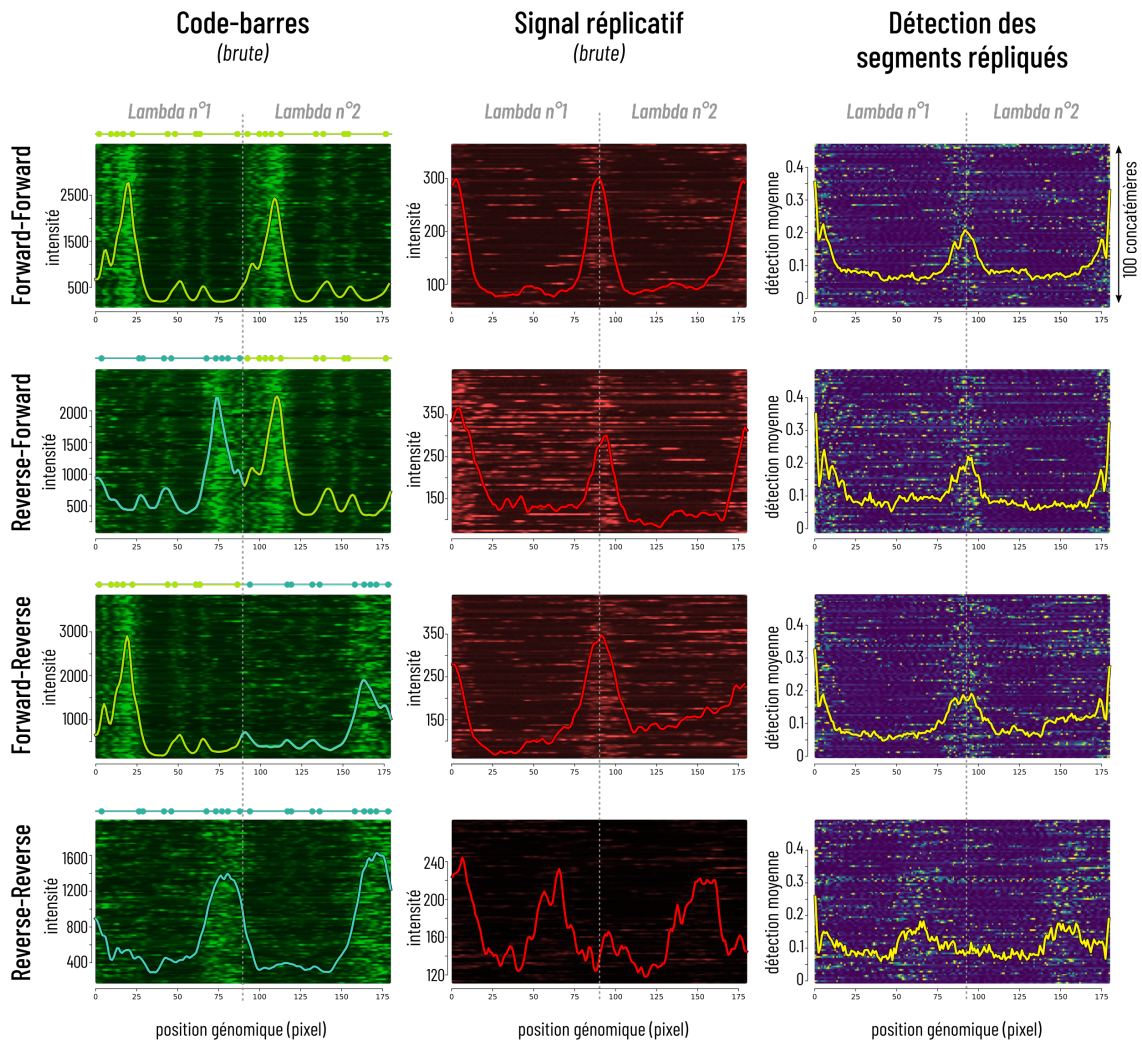


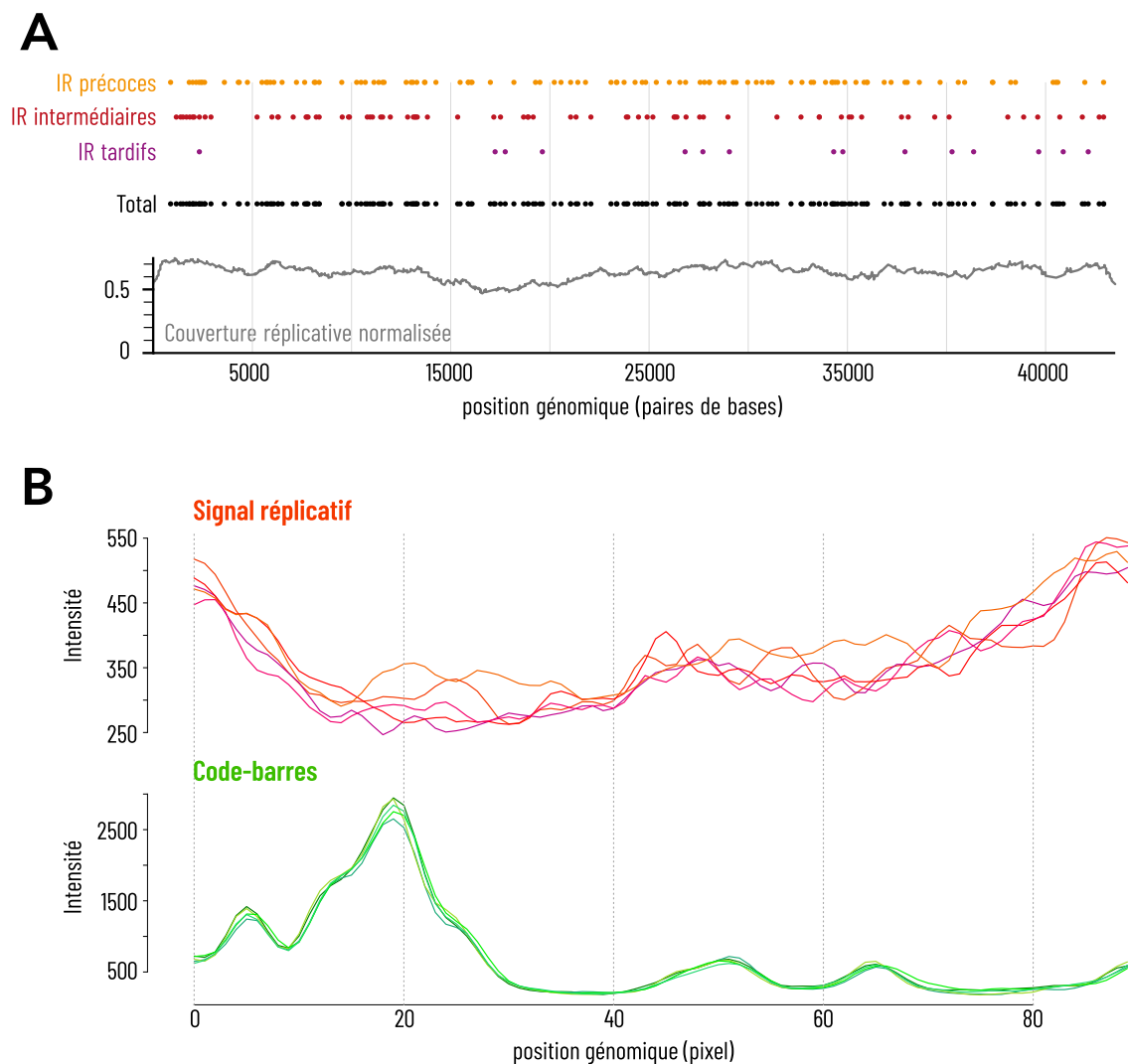
Figure 4.8 – Profils moyen d’intensité pour les différents types de jonctions. 100 concatémères sont visualisés pour chaque type de jonction pour le signal code-barres (vert), le signal réplcatif (rouge) et les détections des segments répliqués. Les images sont associées aux profils d’intensité moyens évalués sur l’ensemble des concatémères présents dans chaque groupe de jonction (jonctions *Forward-Forward* : 1545 concatémères, *Reverse-Forward* : 290 concatémères, *Reverse-Reverse* : 277 concatémères, *Forward-Reverse* : 741 concatémères). Le signal du code-barres nous permet de réaliser une vérification visuelle des concatémères recréés. Les profils moyens pour le signal réplcatif brute et la détection des segments répliqués ont des tendances similaires avec une forme en U et donc une préférence de l’initiation de la réplication au niveau des jonctions pour les concatémères *Forward-Forward*, *Forward-Reverse* et *Reverse-Forward*. Par contre, la forme de U du signal n’est pas observée sur la jonction *Reverse-Reverse*.

## COMMENT EXPLIQUER CE PROFIL RÉPLICATIF EN FORME DE U ?

Comme expliqué au début de cette section, l'ADN de  $\lambda$  répliqué dans des extraits d'œufs de Xénope est exposé au processus de jonction des extrémités non homologues. Ainsi, dans notre échantillon, l'ADN de  $\lambda$  va rapidement former des concatémères constitués d'au moins deux fragments.

D'après les résultats que nous avons obtenus, la réplication de l'ADN de  $\lambda$  débute au niveau des jonctions mais elle n'est pas spécifique du type de jonction indiquant donc qu'elle n'est pas dépendante de la séquence. Il s'agit d'un phénomène qui est observé pour la première fois. En effet, les études précédentes étaient réalisées soit sur des monomères de  $\lambda$  (STANOJIC *et al.* 2008) soit n'a pas été observé dans le jeu de données analysé (DE CARLI, GAGGIOLI *et al.* 2016). En prenant 167 monomères de  $\lambda$  orientés *Forward* (comme le fait De Carli), nous constatons la forme en U au niveau du profil d'intensité moyen du signal répliatif. Le nombre de molécules analysées n'influence donc pas les observations que nous avons réalisées sur nos données en utilisant la technologie des nanocanaux (cf. Figure 4.9).

Une explication possible de cette forme en U serait la présence au niveau de ces jonctions de marques épigénétiques pouvant faire appel à la machinerie nécessaire au processus de réplication de l'ADN. Si ces marques sont transitoires et diffèrent selon les expériences, cela peut expliquer pourquoi les résultats obtenus par peignage moléculaire (*ibid.*) diffèrent de ceux que nous avons obtenus grâce au système Irys.



**Figure 4.9 – Comparaison des observations faites sur la réplication de l'ADN de  $\lambda$  par peignage moléculaire et nanocanaux. (A)** Les positions des évènements d'initiation le long de l'ADN de  $\lambda$  sont présentés pour l'ensemble des intermédiaires de réplication (IR)(au total 198 initiations et 167 molécules analysées). De Carli montre ici que la réplication de l'ADN de  $\lambda$  ne démarre pas dans des zones préférentielles. **(B)** Nous avons analysé le profil d'intensité moyen de 167 monomères de  $\lambda$  (orienté *Forward*, cf. courbes vertes représentant les profils moyens du signal du code-barres) pris au hasard dans l'ensemble des molécules (faiblement et fortement répliquées) et avons renouvelé ce calcul 5 fois. Malgré le faible échantillons, à chaque fois nous retrouvons cette forme en U impliquant que la réplication de l'ADN de  $\lambda$  est préférentiellement sur les extrémités. Adapté de [DE CARLI, GAGGIOLI et al. 2016](#) pour (A).

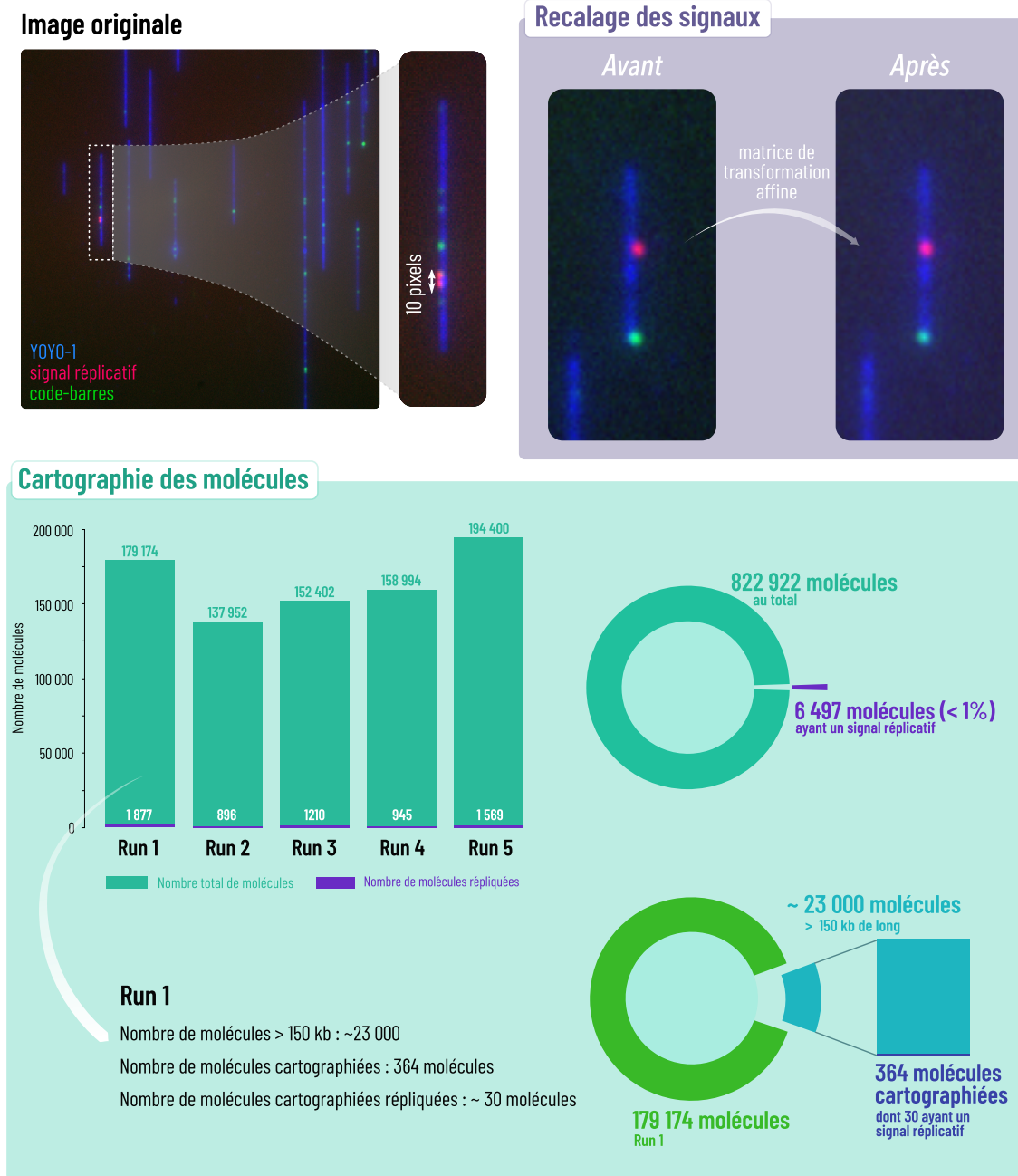


## 4.2 Analyse de la réplication de l'ADN chez l'Homme

*Travail réalisé en collaboration avec Torsten KRUDE (Département de Zoologie, Université de Cambridge (Royaume-Uni)).*

Un système rappelant celui des extraits d'œufs de Xénope est utilisé pour générer les données présentées ci-après : le *human cell free system (HCFS)* (KRUDE 2000). Ce système a été pensé dans le but d'analyser biochimiquement l'initiation de la réplication dans les noyaux des cellules somatiques humaines ou bien de mammifères. Le développement d'un tel système est justifié puisqu'il existe des différences au niveau de la régulation du cycle cellulaire entre les cellules somatiques ou bien embryonnaires de mammifères, ou encore chez la levure. En effet, chez les Amphibiens par exemple, le cycle cellulaire embryonnaire précoce est caractérisé par une alternance rapide de phases S et M et donc une absence de phase G1 et G2. Ce système permet la réplication semi-conservatrice de l'ADN dans des noyaux en phase G1 avec la nécessité de kinases cycline-dépendantes pour l'initiation et l'assemblage du complexe pré-RC avant l'initiation. Par ailleurs, dans ce système seul un très faible pourcentage de l'ADN génomique est répliqué. Le système HCFS a été utilisé dans le cadre de l'étude de la réplication chez l'Homme par peignage moléculaire (MARHEINEKE, GOLDAR et al. 2009) ainsi que par séquençage de sites d'initiation marqués avec de la digoxigénine-dUTP et immunoprécipités avec des anticorps anti-digoxigénine (Ini-seq) (LANGLEY et al. 2016).

L'échantillon analysé ci-après contient de l'ADN génomique d'un million de noyaux d'EJ30 (cellules cancéreuses) en phase G1 tardive ayant été incubé dans le HCFS en présence de dUTP fluorescents durant 60 minutes. Plusieurs fractions de ce même échantillon ont été passés dans le système Irys afin d'obtenir une couverture suffisamment importante. Nous avons étudié ici 5 *runs* provenant de ce même échantillon et appliqué les différentes étapes du pipeline HOMaRD (cf. Figure 4.10).



**Figure 4.10 – Analyse des données générées sur le génome humain.** Un exemple d’image de notre échantillon est donné afin d’apprécier le signal réplcatif obtenu. Contrairement au signal attendu, nous avons un signal réplcatif d’environ 10 pixels équivalent à  $\approx 5$ kb. Malgré la faible quantité de signal (dont dépend l’estimation de la matrice de transformation affine (cf. Section 3.3.1.2), nous parvenons à recalcr les images et faire en sorte que les 3 signaux se superposent (cadre violet). Nous avons analysé 5 runs effectués à partir du même échantillon. Moins de 1% des molécules imagées présentent un signal réplcatif. En analysant plus en détail le run 1, nous avons constaté que sur  $\approx 180\,000$  molécules seules 23 000 d’entre elles pouvaient être considérées pour la cartographie. Sur ces 23 000 molécules 364 ont été cartographiées avec succès (1,6%) dont 30 molécules présentent un signal réplcatif (0,1%).

À notre grande surprise, nous ne sommes pas parvenus à obtenir les signaux attendus pour le signal répliatif (cf. Image originale sur la Figure 4.10). En effet, avec une incorporation de dUTP durant 60 minutes et une vitesse moyenne d'environ 300 pb/min pour ce système, nous nous attendions à obtenir des signaux répliatifs pouvant aller jusqu'à 180 kb (soit environ 36 pixels) (MARHEINEKE, HYRIEN et KRUDE 2005). Or, les signaux que nous observons font au maximum 10 pixels ( $\approx 5$  kb) et extrêmement peu fréquents. Nous avons tout de même extrait les profils d'intensité des molécules pour les 5 *runs* et constaté que moins de 1% de nos données (6497 sur 822 922 molécules) présentaient un signal répliatif (signal répliatif conservé lorsque l'intensité maximum des molécules était supérieur à 150). Pour ce qui est de la cartographie, nous l'avons réalisé sur notre premier *run* (179 174 molécules) qui contenait  $\approx 23$  000 molécules de plus de 150 kb de long (pré-requis pour l'étape de cartographie). Sur ces 23 000 molécules, seulement 364 ont été cartographiées parmi lesquels 30 présentaient un signal répliatif.

Au vu de ces résultats, nous avons fait le choix de ne pas poursuivre l'analyse de cet échantillon. Une première explication possible pour l'obtention de tels résultats est un problème survenu au cours de l'expérience. Nous pouvons également nous demander pourquoi la taux de molécules cartographiées est si faible : seulement 364 molécules ont été cartographiées sur les 23 000 soit un peu plus d'1 % des molécules. Cela pourrait être dû à un paramétrage incorrecte de RefAligner (pour effectuer notre cartographie nous avons utilisé les paramètres par défaut pour ce qui est de la cartographie optique pour le génome humain). Afin de pouvoir analyser la réplication de l'ADN en faisant usage du HCFS il faudra attendre la préparation de nouveaux échantillons. Par ailleurs, en plus des difficultés expérimentales, nous avons rencontré des problèmes au niveau des fichiers de sortie du système Irys. En effet, les logiciels ayant subit un certain nombre de mises à jour, une réadaptation de notre outil a été nécessaire, mais l'étape permettant la liaison entre les profils d'intensité des molécules et leur position sur le génome de référence n'a pas été ré-implémentée.

### 4.3 Une nouvelle approche pour la détection des segments répliqués

Les nucléotides fluorescents utilisés pour marquer les segments d'ADN en cours de réplication dans les systèmes *ex vivo* (extraits d'œufs de Xénope) ou *in vitro* (HCFS (cf. Section 4.2), etc) ne pénètrent pas passivement dans les cellules. Des protocoles de perméabilisation transitoire (par exemple [MAYA-MENDOZA et al. 2010](#)) sont donc nécessaires, ce qui rend extrêmement difficile le contrôle précis de la quantité de nucléotides ayant pénétré dans les cellules. Le marquage de la réplication *in vivo* dans des cellules eucaryotes avec les analogues de la thymidine utilisés en peignage moléculaire (IdU ou CldU, cf. Section 2.3.2.1)) est quant à lui difficilement envisageable, les étapes d'immunodétection n'étant *a priori* pas réalisables dans les nanocanaux. Afin de marquer la réplication *in vivo* et de façon compatible avec les nanocanaux, il a alors été envisagé d'utiliser le 5-ethynyl-2'-déoxyuridine (EdU), un analogue de la thymidine qui diffuse dans de nombreux tissus et cellules et qui peut être rendu "visible" par couplage chimique (réaction de "click-chemistry") à des groupements azides fluorescents ([SALIC et MITCHISON 2008](#)). Cependant, Florence Proux, ingénieure au laboratoire, a montré que la réaction de "click", qui utilise des ions cuivre endommageant l'ADN, ne permet pas d'obtenir des molécules suffisamment longues pour être positionnées de façon fiable sur un génome de référence.

Nous avons alors tenté de visualiser les segments répliqués par une autre approche basée directement sur le signal du YOYO-1. De Carli et *al.* ([DE CARLI, GAGGIOLI et al. 2016](#)) avaient en effet observé un doublement de fluorescence du signal YOYO-1 au niveau des segments répliqués, sur des données obtenues par peignage moléculaire. Ces résultats sont cohérents avec les expériences menées par Yardimci et *al.* ([YARDIMCI et al. 2012](#)) qui ont étudié la réplication de l'ADN du phage  $\lambda$  dans des extrait d'œufs de Xénope en faisant usage d'un système microfluidique différent du système Irys. Ils démontrent que les segments néorépliqués ayant incorporé la digoxigénine-dUTP (reconnue par l'ADN polymérase comme étant de la désoxythymidine triphosphate (dTTP) et révélée à l'aide d'anticorps anti-dig) colocalisent avec les segments présentant un doublement de fluorescence au niveau de la molécule d'ADN (marqué au SYTOX orange, un marqueur induisant une augmentation de l'intensité de fluorescence (450 fois) lors de sa liaison avec un ADN double brin([YAN et al. 2000](#))).

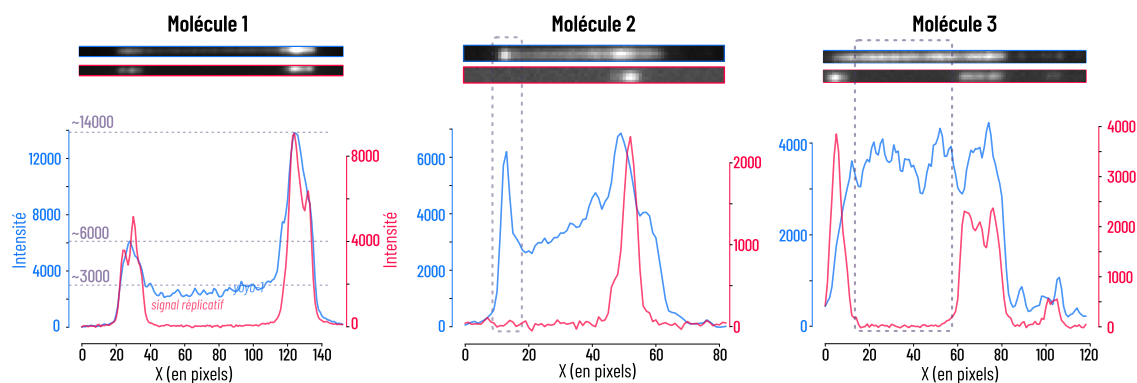
### 4.3.1 Un doublement de fluorescence du YOYO-1 présent mais pas systématique dans notre échantillon de chromatine de sperme de Xénope

Si nous parvenons à montrer que le doublement de fluorescence correspond systématiquement aux régions répliquées, cela pourrait faciliter l'analyse de la réplification chez tout type d'organisme.

Tout d'abord, nous avons analysé ce potentiel doublement de fluorescence au niveau des régions répliquées en faisant répliquer de la chromatine de sperme de Xénope dans des extraits d'œufs de Xénope en présence de nucléotides fluorescents (cf. Section 4.3.1). Nous pouvons donc comparer l'intensité du signal YOYO-1 au niveau des segments répliqués (i.e. marqués par les nucléotides fluorescents) et non-répliqués.

L'échantillon de chromatine de sperme de Xénope répliqué dans des extraits d'œufs de Xénope en présence de nucléotides fluorescents est passé dans le système Irys et nous avons extrait les profils d'intensité à l'aide de notre pipeline HOMaRD (cf. Chapitre 3). Pour rappel, l'objectif ici est de voir si notre signal répliatif (servant donc de contrôle) coïncide avec un doublement de fluorescence du signal YOYO-1.

Nous avons analysé manuellement 100 molécules prises au hasard parmi environ 80 000 molécules présentant un signal répliatif (valeur de seuil fixé à 500). Sur les 100 molécules, seules 38 d'entre elles présentent une intensité forte du signal YOYO-1 au niveau des segments répliqués (cf. Figure 4.11, molécule 1) et sur les 62 autres nous observons une quasi-absence de corrélation entre les 2 signaux. Les molécules 2 et 3 de la figure 4.11 montrent des zones (cadres en pointillés) où le signal répliatif est absent alors que l'intensité du signal YOYO-1 est élevée.



**Figure 4.11 – Analyse du doublement de fluorescence du YOYO-1, échantillon Xénope.** Trois exemples de molécules d'ADN issues de notre échantillon de chromatine de sperme de Xénope sont présentés (images brutes et profils d'intensité associées). Le signal YOYO-1 (bleu) et réplicatif (rouge) de la molécule 1 (à gauche) coïncident. Si nous considérons la ligne de base du profil d'intensité à 3000, le premier pic du signal YOYO-1 correspond à un doublement de fluorescence avec une intensité de 6000 alors que le second est à 14 000 soit prêt de 5 fois l'intensité de base. Les deux autres exemples montrent des cas où nos deux signaux ne corrélaient pas (molécule 2 (centre) et 3 (à droite)). En effet, pour la molécule 2 nous avons un pic d'intensité du YOYO-1 (cadre en pointillés) et une absence de signal réplicatif dans cette même zone. Il en est de même pour la molécule 3.

### 4.3.2 Existe-il un doublement de fluorescence du signal YOYO-1 dans notre échantillon de levure ?

Nous avons également souhaité voir si nous parvenions à mettre en évidence un éventuel doublement de la fluorescence du YOYO-1 au niveau des régions répliquées chez la levure *S. cerevisiae*. Comme cela a été présenté précédemment, la levure a joué un rôle important dans la compréhension du processus de réplication de l'ADN et est un organisme modèle pour l'étude de la réplication de l'ADN (cf. Chapitre 2). La levure *S. cerevisiae* présente près de 400 ARS confirmées (<http://cerevisiae.oridb.org>) dont la localisation est connue. Les cellules ont été synchronisées en G1 puis relarguées en phase S pendant 35 minutes en présence d'hydroxyurée, une drogue qui entraîne un ralentissement des fourches de réplication. Ainsi, seules auront été répliqués les segments d'ADN correspondant aux ARS qui se sont déclenchées.

Nous avons appliqué les différentes étapes du pipeline que nous avons développé afin d'analyser nos données. Les échantillons de levure ont été marqués uniquement à l'aide de l'agent intercalant et les molécules sont identifiées par leur code-barres. L'analyse populationnelle (cf. Figure 4.12) du signal YOYO-1 ne fait pas ressortir d'enrichissement particulier au niveau de certaines ARS (c'est à dire les ARS les plus précoces) que ce soit au niveau de l'image ou au niveau du profil moyen. Lorsque nous sommes passés à l'analyse des profils d'intensité individuels (cf. Figure 6.2), nous nous sommes rendus compte que les images présentaient des aberrations de forme et d'intensité différentes dont la localisation dans l'image est variable (cf. Figure 4.13). Ces aberrations induisent des variations d'intensité qui sont proches de celles attendues si nous avions un doublement de l'intensité rendant difficile l'analyse notre jeu de données.

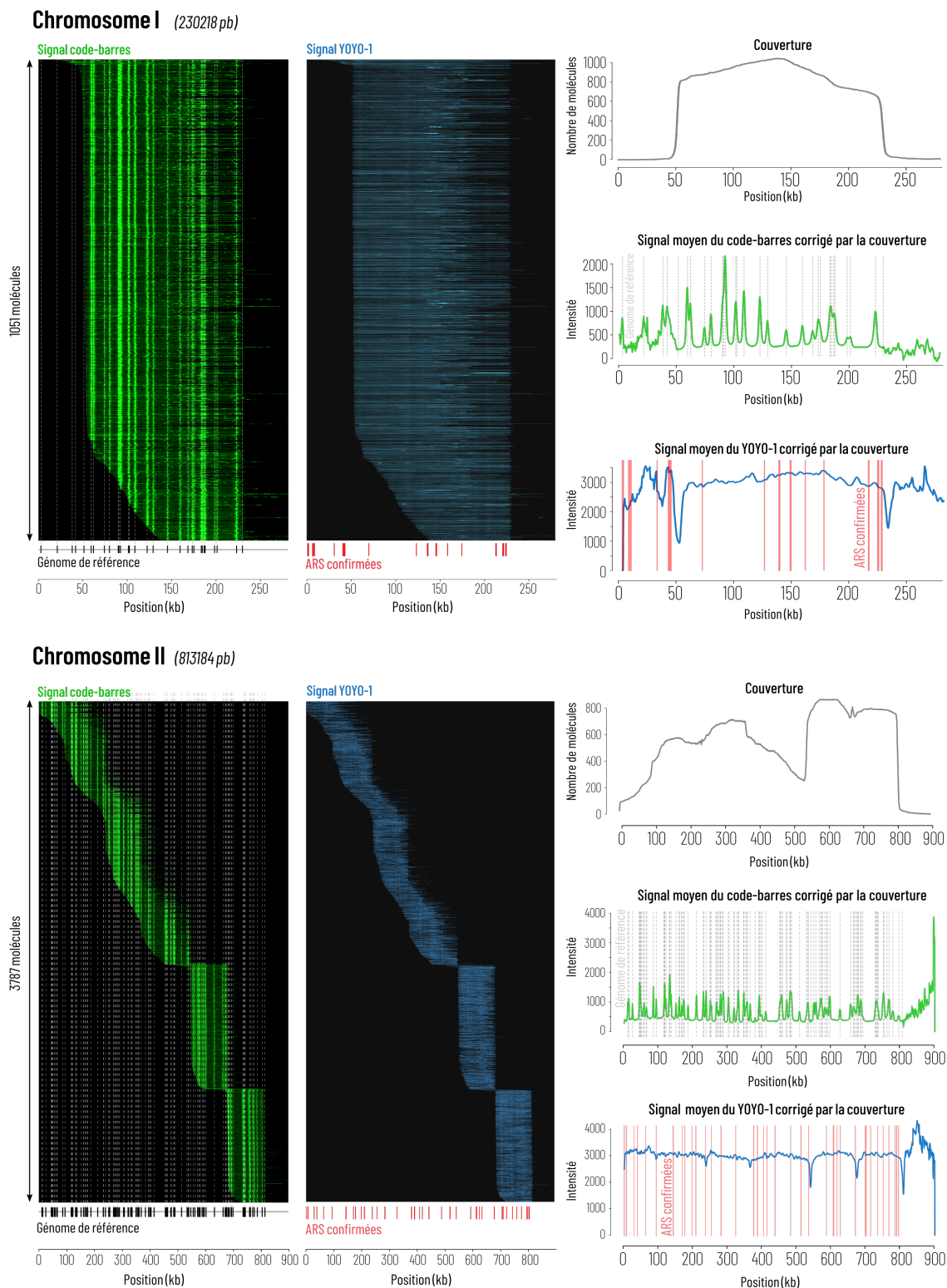
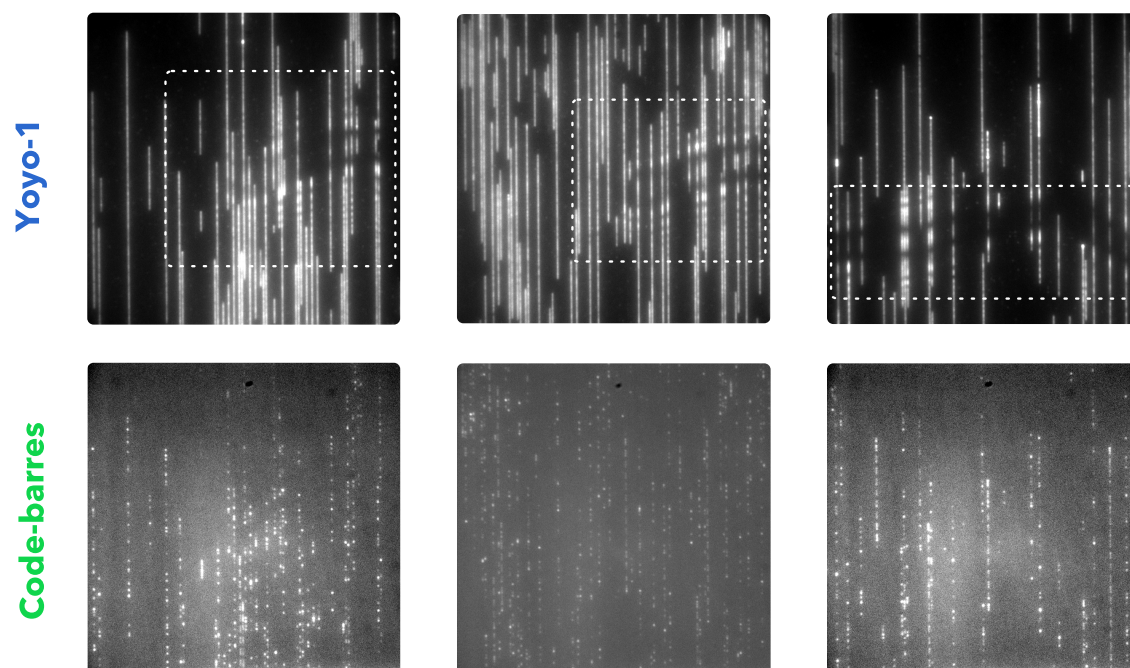


Figure 4.12 – Analyse populationnelle des chromosomes I et II de la levure *S. cerevisiae*. Les molécules d’ADN de levure ont été cartographiées sur les différents chromosomes et une visualisation de ces molécules est réalisée pour les chromosomes I (1051 molécules) et II (3787 molécules) : l’image du signal des code-barres ainsi que le profil moyen de l’intensité de ce dernier (corrigé par la couverture), nous permettent d’apprécier visuellement la qualité de la cartographie optique. Pour ce qui est de l’image su signal YOYO-1, nous remarquons qu’il n’y a pas de motif (doublement d’intensité) qui ressort au niveau des ARS confirmés. Cette même observation peut être faite sur le profil moyen des intensités du signal réplicatif (corrigé par la couverture).





**Figure 4.13 – Aberrations observées sur les données de levure.** Sont présentées ici trois exemples d'images pour les canaux de couleur bleu (YOYO-1, haut) et vert (code-barres, bas). Nous constatons la présence d'aberrations (cadres blancs pointillés) sur les images du signal YOYO-1 dont la nature est inconnue. Ces aberrations varient en fonction des images et ne sont pas systématiquement localisées au même endroit. Par ailleurs, ces aberrations sont absentes des images présentant le signal des code-barres.

Au vu des résultats obtenus sur nos échantillons et plus particulièrement sur celui du Xénope, il paraît difficile de se reposer uniquement sur le "doublement" de fluorescence du signal YOYO-1 pour étudier la réplication de l'ADN. Aussi, il est important de noter que lorsque le signal YOYO-1 coïncide avec le signal réplicatif, nous n'observons pas nécessairement un doublement de fluorescence dans notre "échantillon Xénope" (cf. Figure 4.11, molécule 1, second pic avec une intensité 5 fois plus importante que le signal de base).

## Discussion et Perspectives

---

<b>5.1</b>	<b>La cartographie de la réplication de l'ADN en molécule unique et <i>genome-wide</i> : du rêve à la réalité ?</b>	<b>145</b>
5.1.1	Le développement d'HOMaRD, un défi relevé	145
5.1.2	Les limitations technologiques et techniques	146
5.1.3	Les limitations expérimentales	148
<b>5.2</b>	<b>L'avenir de l'analyse de la réplication en molécule unique</b>	<b>150</b>
5.2.1	Le système Irys vs. le nouveau système de Bionano Genomics®, Saphyr	150
5.2.2	La technologie des nanopores (Oxford Nanopore Technologies®)	152
<b>5.3</b>	<b>Faut-il oublier la technologie des nanocanaux au profit de la technologie Nanopore ?</b>	<b>154</b>

---



## 5.1 La cartographie de la réplication de l'ADN en molécule unique et *genome-wide* : du rêve à la réalité?

### 5.1.1 Le développement d'HOMaRD, un défi relevé

Après plusieurs décennies passées à l'étude de la réplication de l'ADN par des techniques en molécule unique difficiles à mettre en oeuvre, nous entrons dans une nouvelle ère, où il est possible de réaliser ces mêmes analyses en faisant appel à de nouveaux outils, parmi lesquels nous avons le système Irys. Le débit obtenu est sans précédent : l'analyse de la réplication de l'ADN peut être réalisée *genome-wide* en une journée avec près d'un million de molécules imagées et localisées automatiquement contre quelques centaines de molécules obtenues en plusieurs mois de travail avec la technique de peignage moléculaire par exemple.

Le développement d'HOMaRD a été réalisé en deux temps : la première étape a été de cartographier les profils d'intensité extraits des images brutes obtenues en sortie du système Irys et la seconde a consisté à développer un algorithme permettant la détection automatique des segments répliqués (cf. Chapitre 3.27).

Les images brutes générées par le système Irys ont nécessité deux étapes de *post-processing* indispensables, à savoir la correction de l'inhomogénéité de nos images et le recalage des signaux réplikatifs et du code-barres sur le signal du YOYO-1. Nous avons pu constater que l'algorithme que nous avons développé pour le recalage des images est robuste puisque même en présence de peu de signal réplikatif, le recalage est plus que satisfaisant (cf. Figure 4.10). De plus, il faut garder en mémoire que ce recalage est dépendant du *run*, nous obligeant donc à ré-estimer la matrice de transformation affine à chaque nouvelle acquisition de données. Il a donc été important que notre approche soit robuste et généralisable.

Suite à ces corrections sur nos images brutes, notre second défi a été de comprendre et de développer les outils bioinformatique adéquats dans le but d'analyser le processus de la réplication de l'ADN. En effet, notre pipeline d'analyse repose en grande partie sur les détections exécutées par Autodetect et la cartographie optique réalisées par RefAligner. Plusieurs mois ont été nécessaires pour appréhender les dizaines de fichiers générés par ces différents logiciels et outils fournis par Bionano Genomics®. Ces étapes de "*reverse engineering*" ont été des étapes cruciales pour briser le verrou technologique et accéder aux informations clés autorisant la carto-

graphie de la réplication de l'ADN. Une fois ces fichiers décodés nous avons été en mesure de cartographier les profils d'intensité et d'obtenir une visualisation aussi bien au niveau de la population qu'au niveau de la molécule unique pour les trois signaux étudiés.

L'étape suivante a consisté à détecter automatiquement les segments répliqués de chacune des molécules. Pour ce faire, nous avons testé plusieurs approches avant de nous focaliser sur l'optimisation d'une fonction de coût permettant d'obtenir en sortie le signal répliatif sous-jacent (signal "binaire" caché). En effet, nous avons fait l'hypothèse que notre signal était constitué de la FEP, du signal répliatif "vrai" et dépendant du contenu en AT. Au vu de ces informations sur la composition de notre signal, réaliser un seuillage nous a paru être une approche non adaptée, d'où notre choix.

### 5.1.2 Les limitations technologiques et techniques

À partir de là, nous avons été en mesure de cartographier la réplication de l'ADN de  $\lambda$  répliqué dans des extraits d'œufs de Xénope et d'analyser 2 sous-populations (molécules faiblement et fortement répliquées)(cf. Section 4.1). L'étude de la réplication en molécule unique avec le système Irys du génome humain s'est avérée infructueuse (cf. Section 4.2). Pour faire en sorte d'obtenir des données analysables, il faudrait réaliser de nouveaux échantillons, vérifier les étapes de cartographie optique mais également ré-adapter notre pipeline. En effet, du fait de la mise à jour des logiciels Autodetect et en particulier RefAligner, notre pipeline ne parvient plus à faire le lien entre la nouvelle position des profils d'intensité (cf. Section 3.3.3) sur le génome de référence et le profil d'intensité lui-même.

Une des limitations majeures de notre pipeline est qu'il est totalement dépendant des fichiers de sortie des logiciels de Bionano Genomics® et des images brutes obtenus. Ainsi, si nous n'avons plus accès aux fichiers de sorties d'Autodetect et de RefAligner, le pipeline sera obsolète et les détections des molécules ainsi que du code-barres sur les images brutes devront être pensées et codées de A à Z. Pour ce qui est de la cartographie optique, des alternatives existent (LEUNG et al. 2017) permettant donc de contourner l'utilisation de RefAligner. À terme, se détacher totalement des logiciels de détection et de cartographie constitue une meilleure option afin d'avoir la main sur le pipeline dans son intégralité et non de façon partielle comme cela est le cas à l'heure actuelle. Aussi, l'intégration des molécules longues (molécules cheveu-

chant au moins 2 images consécutives), qui ne sont pas prises en compte dans notre pipeline pour l'instant, se donnerait possiblement des résultats plus satisfaisants concernant la détection que ceux obtenus avec Autodetect (cf. Figure 3.10).

Comme n'importe quel outil bioinformatique, HOMaRD est perfectible. Parmi les points à améliorer, nous avons la partie concernant la détection des segments répliqués. Nous avons réalisé la détection des segments répliqués sur la population des molécules d'ADN faiblement répliquées (cf. Figure 3.34). Notre algorithme ne permet pas pour le moment de détecter des segments répliqués sur des signaux complexes parmi lesquels les molécules d'ADN fortement répliquées (cf. Figures 3.32, 3.33).

Une première explication serait que notre modèle actuel (cf. Section 3.3.7.2) est peut-être simpliste pour des cas complexes. Au sein d'une molécule fortement répliquée, l'intensité des pixels pourrait avoir une influence plus importante sur les pixels voisins, impactant sûrement la détection des segments répliqués et donc le modèle. Il serait également intéressant de prendre en compte l'intensité des molécules dans l'environnement direct de la molécule analysée qui pourrait avoir un effet sur la détection des segments. Aussi, les segments que nous détectons sur les molécules sont parfois des pixels seuls proches de groupes de pixels (cf. Figure 3.33). Nous devons donc déterminer un seuil pour lequel nous allons fusionner les zones détectées afin de former de "véritables segments" d'ADN répliqués.

Nous ne sommes donc pas encore en mesure d'extraire de façon robuste des paramètres précieux pour l'analyse de la réplication tels que les longueurs des segments répliqués dans les différentes sous-populations (molécules faiblement et fortement répliquées), les distances entre ces segments

### 5.1.3 Les limitations expérimentales

Comme nous l'avons démontré, la technologie des nanocanaux permet une cartographie à haut débit de la réplication de molécules d'ADN répliquées dans des extraits d'œufs de Xénope. Et malgré les premiers résultats peu encourageants obtenus sur nos échantillons humains (cf. Section 4.2), le système HCFS développé par T. Krude (KRUDE 2000) reste une alternative intéressante et prometteuse pour l'étude de la réplication chez l'Homme en nanocanaux puisqu'il est possible d'obtenir de longues molécules d'ADN ayant incorporé des nucléotides fluorescents au niveau des sites d'initiation. Appropriée pour les systèmes *ex vivo*, cette technologie semble plus difficile à utiliser pour un marquage réplcatif *in vivo* :

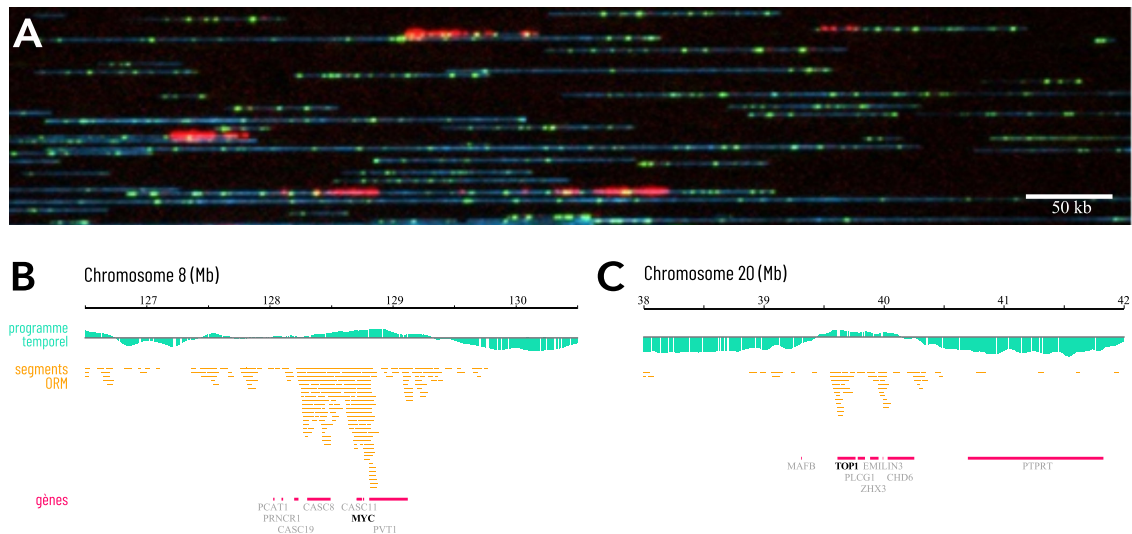
- les nucléotides fluorescents ne pénètrent pas passivement dans les cellules eucaryotes,
- les nucléosides tels que l'IdU ou le CldU, qui pénètrent passivement dans les cellules, ne sont détectables qu'après dénaturation de l'ADN et immunofluorescence, une procédure incompatible avec les nanocanaux,
- l'utilisation d'un autre nucléoside analogue de la thymidine, l'EdU, détectable par "*click-chemistry*" sans dénaturation de l'ADN, s'est révélée impossible car l'ADN est cassé lors de cette réaction chimique.

Cependant, il existe une alternative qui consiste à l'incorporer des nucléotides fluorescents au niveau des segments neo-répliqués à la suite d'une transfection des cellules humaines. Les premières analyses de la réplication de l'ADN humain utilisant la technologie des nanocanaux ont été publiées sur BioRxiv par Klein et *al.* (KLEIN et al. 2017, BioRxiv). Ils cartographient les origines de réplication en combinant la technique de cartographie développée par Bionano Genomics (LAM et al. 2012) via le système Saphyr, une évolution récente du système Irys (cf. Section 5.2.1) et le marquage pulsé de cellules HeLa avec des nucléotides fluorescents (*in vivo*) (cf. Figure 5.1). Ces cellules HeLa sont synchronisées et bloquées en phase S du cycle cellulaire avec de l'aphidicoline qui est un inhibiteur spécifique de l'ADN polymérase induisant une pause du cycle cellulaire en G1/S. Les cellules sont ensuite transfectées avec des nucléotides fluorescents, Alexa647-dUTP. Ainsi les origines précoces et se déclenchant juste après le relargage sont marquées et du fait de la déplétion rapide en nucléotides transfectés, les origines de réplication tardives ne le sont pas.

Avec le système Saphyr, ils parviennent à obtenir une couverture de 290 X du génome humain. Les segments répliqués sont détectés par Autodetect comme sur le canal des code-barres (détection de points) et les détections sont filtrées de la manière suivante :

- les molécules contenant un signal rouge avec au moins 5 pixels voisins également rouge dans une fenêtre de 20 kb sont conservés.
- les segments d'ADN sont ensuite séparés si la distance entre deux segments est supérieure ou égale à 30 kb.

Les segments d'ADN répliqués font 48 kb en moyenne avec un enrichissement au niveau des locus MYC et Top1 connus pour être répliqués par des origines de réplication précoces (TAO et al. 2000, KELLER et al. 2002). Les positions des segments détectés avec cette approche sont comparées avec celles obtenues via des approches populationnelles. Ils montrent une colocalisation de leurs segments avec les zones d'initiation déterminées par OK-seq (PETRYK et al. 2016) ainsi qu'avec les résultats obtenus par Ini-seq (LANGLEY et al. 2016) mais dans une moindre mesure avec les données de SNS-seq (PICARD et al. 2014).



**Figure 5.1 – Cartographie de la réplication de l'ADN avec le système Saphyr.** (A) Des cellules HeLa sont synchronisées en phase début de phase S du cycle cellulaire avec de l'aphidicoline permettant l'incorporation de nucléotides fluorescents (rouge) au niveau des sites d'initiation précoces. Les molécules d'ADN sont marqués avec du YOYO-1 (bleu) et le code-barres à l'aide d'une enzyme de restriction simple brin et incorporation de nucléotides fluorescents au niveau des sites de coupures (vert). Un champs fait ici 650 kb. (B)(C) La distribution des segments répliqués (segments ORM, *Optical Replication Mapping* (orange)) au niveau des locus des gènes MYC et (B) et TOP1 (C) est présentée en dessous des profils de programme temporel de la réplication pour ces mêmes locus. Adaptée de KLEIN et al. 2017



Enfin, nous avons envisagé qu'il serait possible de détecter directement les régions répliquées en mesurant l'intensité de fluorescence du YOYO-1. Applicable quel que soit le type d'organisme, cette méthode dispenserait de tout marquage réplificatif. Nous avons cependant constaté que le doublement de l'intensité de fluorescence du YOYO-1 ne semble pas être un paramètre suffisamment fiable pour détecter les segments dupliqués. Même s'il reste possible qu'un autre agent intercalant, tel que le SYTOX orange, soit plus approprié, cette approche est pour l'heure abandonnée.

## 5.2 L'avenir de l'analyse de la réplication en molécule unique

### 5.2.1 Le système Irys vs. le nouveau système de Bionano Genomics®, Saphyr

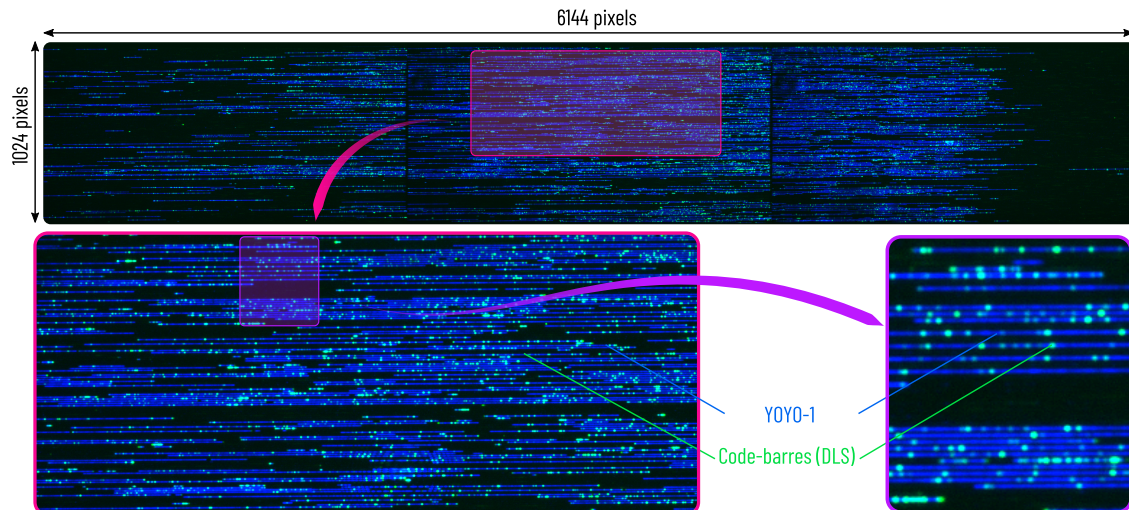
Saphyr est le dernier dispositif développé par Bionano Genomics®, depuis le système Irys. Les molécules imagées vont de 100 kb à quelques mégabases et son débit est nettement supérieur avec 1300 Gb de données générées par flowcell, soit environ 25 à 130 fois plus de données qu'avec le système Irys (10-50 Gb par flowcell). La taille des fichiers générés s'en voit donc impactée mais également le contenu et l'organisation des fichiers de sortie des logiciels Autodetect et RefAligner. Ces derniers ne sont plus les mêmes que ceux d'Irys, en particulier les fichiers intermédiaires dont nous faisons usage dans notre pipeline. En ce qui concerne les images brutes obtenues, la définition mais aussi la résolution ont changé (cf. Figure 5.2) : Saphyr génère 137 images de 1024 x 6144 pixels par cycle contre 1140 images de 512 x 512 par cycle. Nous remarquons que l'espacement entre les molécules de différents nanocanaux est plus net et que le signal ne semble pas s'éteindre sur les molécules avoisinantes. L'influence de l'intensité de ces molécules est amoindrie sur les molécules présentes dans leur environnement direct.

Aussi, l'approche pour effectuer le code-barres était initialement basée sur l'usage d'une enzyme de restriction simple brin. Cependant, lorsque les sites de coupures sont trop proches cela introduit systématiquement des ruptures de l'ADN, limitant ainsi la contiguïté des cartes. Aujourd'hui, la technique utilisée est le *Direct Label and Stain* (DLS) laissant intactes les molécules d'ADN. Le protocole expérimental est simplifié puisqu'il n'y a qu'une seule réaction enzymatique pour le marquage qui est suivi d'un nettoyage et d'une coloration, et plus de réparation des sites de coupure. Grâce au DLS les assemblages *de novo* sont plus précis. L'autre impact de l'utilisation du DLS est sur le coût grâce à l'amélioration considérable du débit :

désormais, il ne faut compter que 500 \$ pour cartographier entièrement un génome humain (ou un génome de taille similaire).

Une autre nouveauté concerne la visualisation et l'analyse des résultats qui se fait en ligne via Bionano Access et non plus par Irys View (qui fonctionne uniquement sur le système d'exploitation Windows).

Bionano Genomics® risque à terme de ne plus prendre en charge le système Irys puisque Saphyr est devenu le fleuron de la compagnie entrée en bourse depuis peu. De plus, le Génoscope, en possession des 2 systèmes, ne souhaite plus maintenir l'entretien du système Irys en plus de celui du Saphyr car trop coûteux. Au lieu d'adapter le pipeline existant (HOMaRD), il serait préférable de développer un outil bioinformatique à partir des images brutes et de se défaire des logiciels de Bionano Genomics®.



**Figure 5.2 – Exemple d'image obtenue avec le système Saphyr.** L'image TIFF avec les deux couleurs (YOYO-1 et code-barres obtenu par la technique de DLS) est un exemple d'image brute obtenue en utilisant le système Saphyr. Sa définition est de 1024 x 6144 pixels. À la différence des images obtenues avec le système Irys, nous constatons une meilleure séparation des molécules (espacement entre les molécules des nanocanaux adjacents) et le décalage des couleurs paraît ne pas être aussi présent qu'avec le système Irys.

## 5.2.2 La technologie des nanopores (Oxford Nanopore Technologies®)

En parallèle des nanocanaux, la technologie des nanopores a été utilisée au sein du laboratoire dans le but d'étudier la réplication à l'échelle du génome mais également pour s'affranchir d'une part des différentes contraintes imposées par les méthodes en molécule unique et d'autre part des limitations posées par les analyses populationnelles.

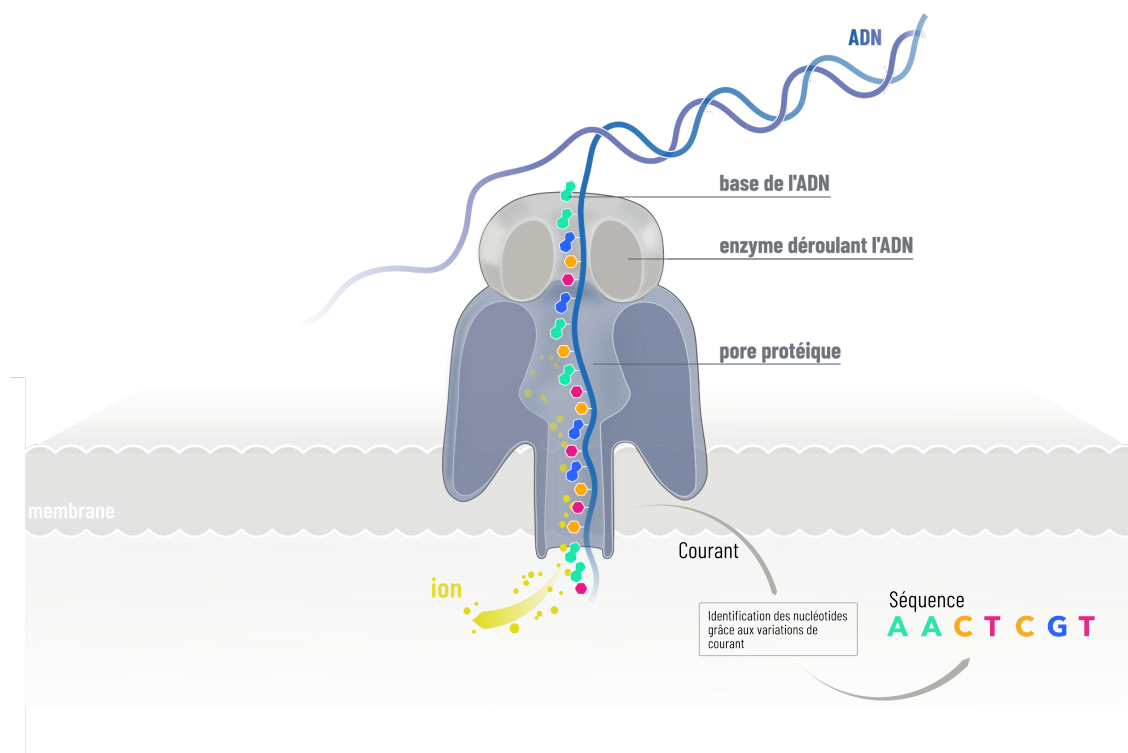
### 5.2.2.1 Le séquençage avec la technologie des nanopores

Les divers appareils développés par Oxford Nanopore Technologies (MinION, GridION et PromethION) font usage de nanopores protéiques ancrés dans une membrane synthétique. Un courant ionique traverse le nanopore en établissant une tension à travers cette membrane. Lorsque le brin d'ADN passe à travers le pore, cet événement crée une perturbation du courant qui dépend des bases présentes à l'intérieur du pore (cf. Figure 5.3). Les signaux électriques enregistrés sont ensuite interprétés par des algorithmes et traduits en séquence d'ADN. La longueur des lectures peut atteindre dans le meilleur des cas 2 Mb avec une taille moyenne des fragments de l'ordre de la dizaine de kilobases. Le développement de cette technologie est extrêmement rapide puisque le PromethION peut générer jusqu'à 7 Tb de données pour quelques centaines d'euros. La technologie Nanopore est aujourd'hui utilisée pour séquencer l'ADN mais également l'ARN.

### 5.2.2.2 Hennion *et al.* et Muller *et al.* pionniers de l'analyse de la réplication de l'ADN par séquençage nanopore

Deux méthodes récentes permettent d'analyser la réplication de l'ADN par séquençage nanopore. Développées par Muller *et al.* (C. A. MÜLLER *et al.* 2019) et par Hennion *et al.* au laboratoire (HENNION *et al.* 2019), toutes deux proposent une approche consistant à identifier un analogue de la thymidine, le BrdU, incorporé au cours de la réplication.

Muller *et al.* ont développé le *DNAscent* permettant d'identifier sur des données de séquençage nanopore les régions de l'ADN de levure *S. cerevisiae* ayant incorporé le BrdU *in vivo*. Grâce à leur algorithme, ils sont en mesure de localiser les origines de réplication et de déterminer l'orientation des fourches de réplication sur près



**Figure 5.3 – Principe du séquençage Nanopore.** Le nanopore est un pore protéique ancré dans une membrane sur lequel va venir se lier une enzyme permettant à l'ADN de se dérouler. Le passage de l'ADN induit une modification de l'intensité du courant qui dépend des bases présentes à l'intérieur du pore, permettant leur identification.

de 100 000 molécules de 20 à 160 kb de long couvrant l'ensemble du génome. Par ailleurs, parmi les origines de réplication identifiées, 80% d'entre elles se situent au niveau d'origines connues (les ARS) et les 20% restants sont constitués d'origines dispersées sur l'ensemble du génome, qui n'avaient pas pu être mises en évidence par les méthodes populationnelles d'analyse de la réplication.

Hennion et *al.* proposent une approche similaire qui diffère au niveau de la stratégie de marquage (*ibid.*). Ils développent un outil, le Fork-seq, permettant la cartographie de la réplication de l'ADN sur l'ensemble du génome en molécule unique avec une résolution de 200 nucléotides. Cette méthode se base également sur la détection du BrdU. Un marquage pulsé au BrdU est réalisé, *in vivo*, chez la levure *S. cerevisiae*. L'échantillon est ensuite séquençé en utilisant la technologie des nanopores et des fragments d'ADN compris entre 10 et 140 kb sont analysés. À l'aide des algorithmes développés, l'étude de la directionnalité et de la vitesse de progression des fourches est réalisable. À la différence des observations faites avec DNAscent, Hennion et *al.* identifient que moins de 7% des origines sont dispersées.

### 5.3 Faut-il oublier la technologie des nanocanaux au profit de la technologie Nanopore ?

La technologie des nanocanaux et celle des nanopores permettent toutes deux d'accéder à la cartographie de la réplication de l'ADN *genome-wide* et en molécule unique en très peu de temps. Avec la quantité et la qualité des données générées à partir de ces appareils, nous avons dans l'espoir d'accéder à de nouvelles informations et faire de nouvelles découvertes concernant la réplication de l'ADN. Cependant, avant de parvenir à cela, ces deux approches ont nécessité, et nécessitent toujours le développement d'outils bioinformatiques robustes, adaptables et spécifiques de nos données. Ces technologies évoluant très rapidement, il faut être réactif et capable d'ajuster les outils développés comme, par exemple, lorsque les formats des fichiers de sortie sont mis à jour et modifiés.

La limitation première du côté des systèmes Saphyr et Irys est notre dépendance par rapport aux fichiers de sortie de logiciels Autodetect et RefAligner qui peuvent à terme constituer un frein à l'analyse des données. En effet, plus la quantité de données générées sera grande plus les informations dans les fichiers de sortie risquent d'être réduites à l'essentiel. Une autre limitation à laquelle nous avons dû faire face avec le système Irys et qui sera peut être rencontrée avec le système Saphyr concerne l'analyse des images et des profils 1D (une dimension). D'un point de vue expérimental, comme nous avons pu le voir précédemment, le marquage des segments répliqués *in vivo* est complexe et en absence de canaux supplémentaires il est impossible de distinguer des différents intermédiaires de réplication à moins d'adopter l'approche de Klein et *al.* (KLEIN et al. 2017) en travaillant sur des cellules synchronisées en début de phase S. Du côté des nanopores, les deux limitations majeures sont la longueur des molécules et la faible précision de l'identification des nucléotides ("*base-calling*") qui est à l'heure actuelle de 62% (NOAKES et al. 2019). Les améliorations apportées sont faites sur les algorithmes d'identification, sur les nanopores et les moteurs associés, sur les méthodes biochimiques pour réaliser une relecture des molécules ou encore sur le voltage utilisé lors du *run* (*ibid.*).

Les systèmes de Bionano Genomics sont, à l'heure actuelle, les seuls nous permettant d'obtenir des molécules longues pouvant atteindre plusieurs mégabases permettant ainsi l'analyse des corrélations spatiales à longue distance entre les intermédiaires de réplication. Toutefois, la technologie des nanopores évolue rapidement et à terme il sera sûrement possible d'obtenir des longueurs de fragments proches de celles obtenues avec les nanocanaux.



## Chapitre 6

# Annexes

---

6.1	Annexe 1   <i>Post-processing</i> des canaux bleu (YOYO-1) et vert (code-barres) . . . . .	159
6.2	Annexe 2   Exemples de profils d'intensité pour les données de <i>S. cerevisiae</i> . . . . .	160

---





## 6.1 Annexe 1 | Post-processing des canaux bleu (YOYO-1) et vert (code-barres)

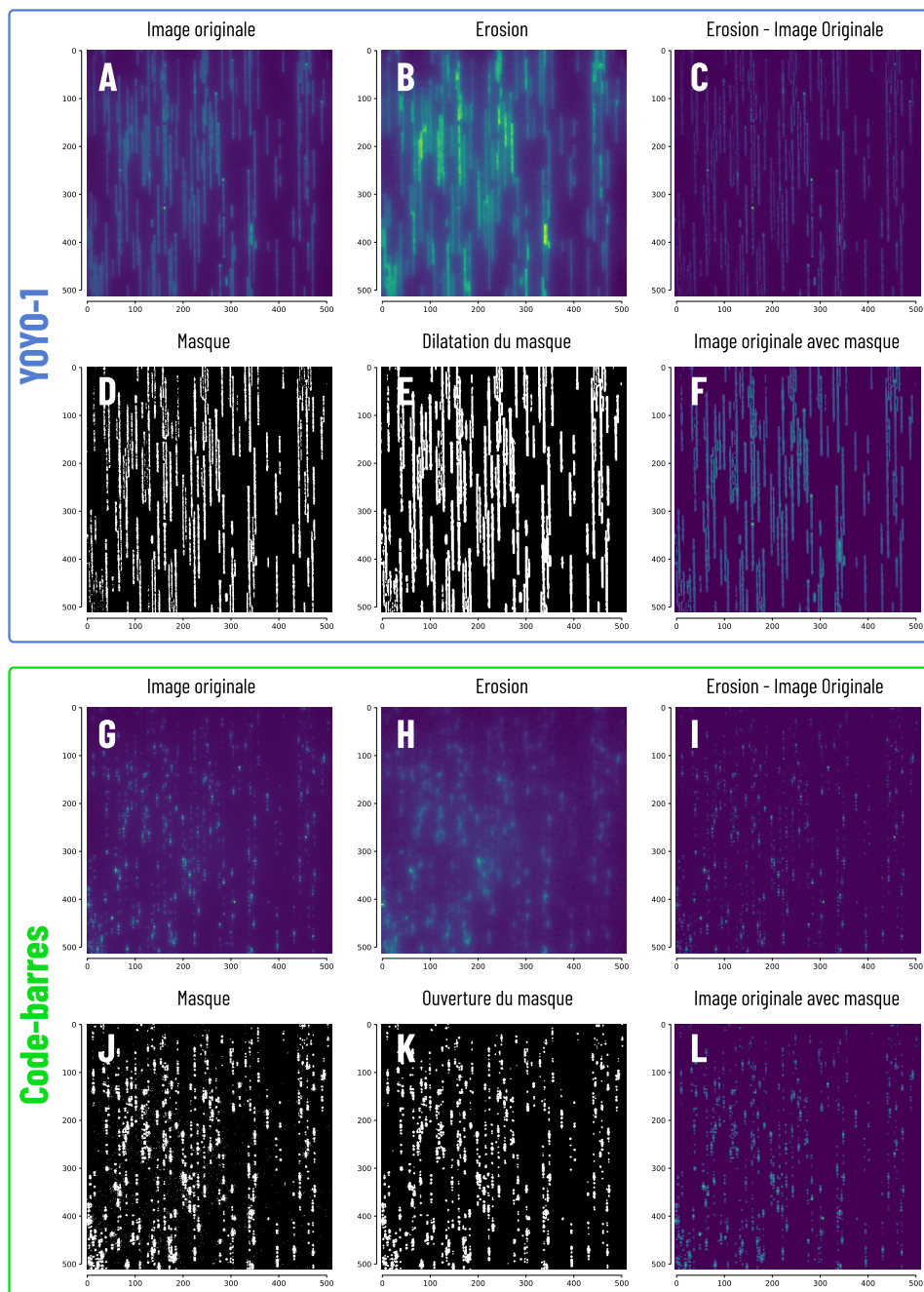
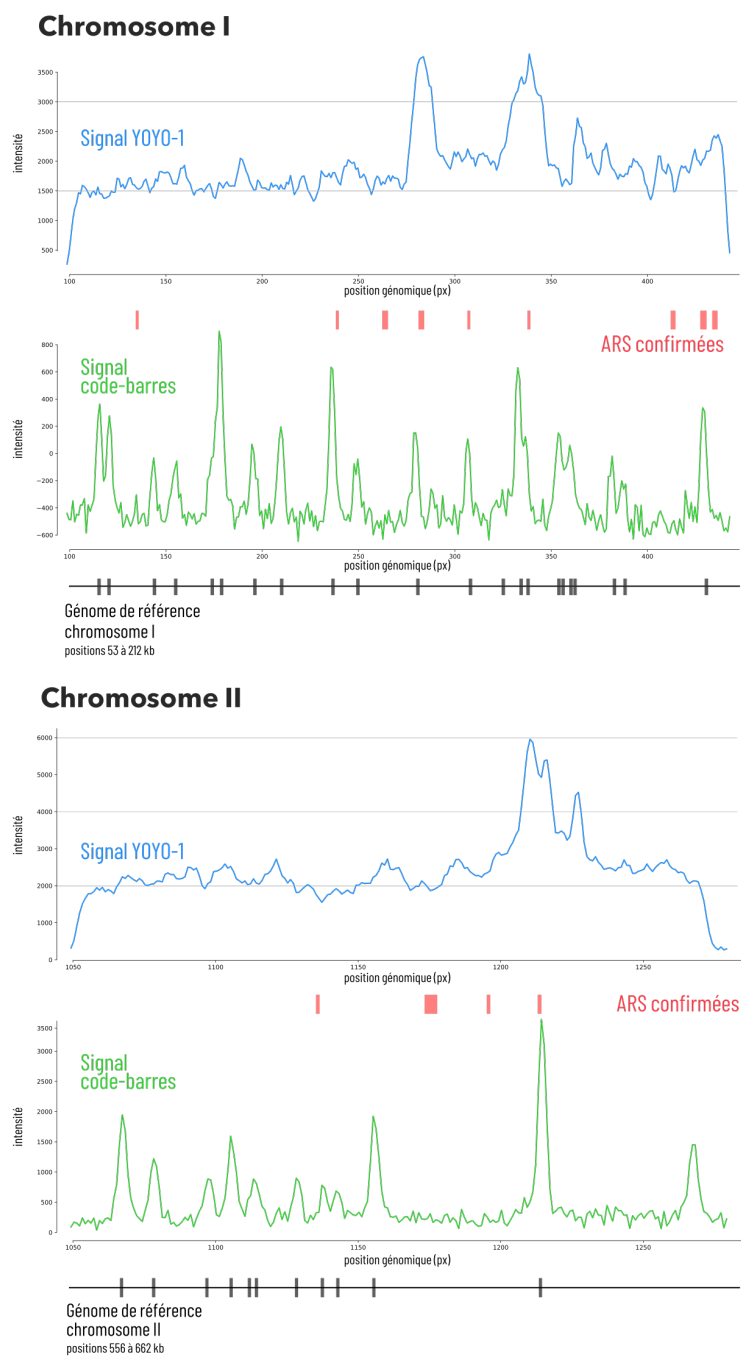


Figure 6.1 – Post-processing des canaux bleu (YOYO-1) et et vert (code-barres). Dans chacun des cas, les images originales (A, G) se voient soustraire leur image après une étape d'érosion à l'aide d'un élément structurant rectangulaire (B, H), permettant ainsi d'éliminer le halo lumineux se trouvant autour des molécules (C, I). Les images obtenues sont ensuite binarisées (D, J). Le bruit de fond persistant et la fragmentation des masques des molécules nous ont amené à réaliser une dilatation dans le cas du signal YOYO-1 (E) et une ouverture dans le cas du code-barres (K) (l'ouverture élimine les petits éléments présent dans l'arrière-plan). Enfin, nous multiplions les images originales par les masques respectifs (F, L).

## 6.2 Annexe 2 | Exemples de profils d'intensité pour les données de *S. cerevisiae*



**Figure 6.2 – Exemples de profils d'intensité pour les données de *S. cerevisiae*.** Ci-dessus sont présentés 2 exemples de profils d'intensité du YOYO-1 et du code-barres pour les chromosomes I et II. Nous observons un doublement de la fluorescence du signal YOYO-1 qui paraît coïncider avec 2 ARS confirmées pour l'exemple de la molécule cartographiée sur le chromosome I et avec une ARS confirmée pour l'exemple du chromosome II. Le profil d'intensité du code-barres nous permet d'apprécier la qualité de la cartographie des 2 molécules.





# Liste des figures

1.1	Bulle de réplication . . . . .	20
1.2	Comparaison entre les méthodes en population et en molécules unique	21
1.3	Peignage moléculaire (OMAR) vs. technologie des nanocanaux . . . . .	24
2.1	Le cycle cellulaire . . . . .	29
2.2	Autoradiographie de l'ADN d' <i>E.coli</i> . . . . .	30
2.3	Le modèle du réplicon . . . . .	33
2.4	Schéma présentant des approches populationnelles pour l'analyse de la réplication de l'ADN . . . . .	40
2.5	Répli-seq, une technique d'analyse du programme temporel de la ré- plication de l'ADN . . . . .	42
2.6	Directionnalité des fourches de réplication et programme temporel d'ac- tivation des origines de réplication . . . . .	47
2.7	Principe du peignage moléculaire . . . . .	52
2.8	OMAR, Optical MApping of Replicating DNA . . . . .	57
3.1	Des images brutes à l'analyse des variants structuraux . . . . .	66
3.2	Pipeline d'analyse Bionano Genomics® . . . . .	67
3.3	Approche expérimentale . . . . .	68
3.4	Puce du système Irys. . . . .	69
3.5	Organisation des images prises avec le système Irys . . . . .	70
3.6	Fichiers de sortie des deux logiciels propriétaires de Bionano Genomics® : Autodetect et RefAligner . . . . .	74
3.7	Comparaison des détections réalisées par les différentes versions d'Auto- todetect. . . . .	76
3.8	Vérification des détections par Autodetect des molécules d'ADN et de leur code-barres . . . . .	76
3.9	Distribution des différences entre <i>XStart</i> et <i>Xend</i> des molécules (1 champ). . . . .	77
3.10	Assemblage de deux images avec des molécules chevauchantes. . . . .	78

3.11 Distribution des longueurs de l'ensemble des molécules. . . . .	79
3.12 Aberration chromatique et moyen de correction. . . . .	80
3.13 Décalage entre les signaux bleus, rouges et verts. . . . .	81
3.14 Variations de l'intensité moyenne des images médianes par cycle et par canal de couleur et quantification du nombre de molécules par cycle . .	82
3.15 Ratio entre l'intensité moyenne des images médianes du signal YOYO-1 et le nombre de molécules par cycle . . . . .	82
3.16 Fonction d'illumination pour les 3 canaux de couleur du cycle 5 du jeu de données . . . . .	84
3.17 Correction de l'inhomogénéité d'illumination pour les trois canaux de couleur . . . . .	85
3.18 Matrice de transformation affine. . . . .	87
3.19 Post-processing des canaux bleu (YOYO-1) et rouge (signal réplcatif)	89
3.20 Détection des maxima locaux (YOYO-1 et signal réplcatif) . . . . .	90
3.21 Répartition des molécules d'ADN suivant les cycles. . . . .	91
3.22 Paramètres de la matrice de transformation affine pour le signal réplcatif	93
3.23 Recalage d'image . . . . .	93
3.24 Cartographie des profils d'intensité : le principe . . . . .	95
3.25 Visualisation d'un profil d'intensité d'un fragment d'ADN réplquée en molécule unique. . . . .	96
3.26 Visualisation des profils d'intensité des molécules d'ADN réplquées en population . . . . .	97
3.27 Pipeline bioinformatique pour l'analyse de la réplcation de l'ADN via le système Irys. . . . .	98
3.28 Contenu en AT de l'ADN de $\lambda$ . . . . .	99
3.29 Estimation de la Fonction d'Étalement du Point (FEP) . . . . .	101
3.30 Exemples de profils d'intensité du signal réplcatif . . . . .	102
3.31 Schéma explicatif du processus d'optimisation pour la détection automatique des segments réplqués . . . . .	104
3.32 Détermination des segments réplqués sur un signal synthétique observé	105
3.33 Détermination des segments réplqués sur un signal synthétique observé complexe . . . . .	106
3.34 Détermination des segments réplqués sur un signal original observé issu d'une molécule faiblement réplquée . . . . .	107
4.1 Orientations des concatémères du bactériophage $\lambda$ . . . . .	112

---

4.2	Distribution des moyennes des valeurs d'intensité des molécules d'ADN et détermination de la valeur de seuillage par la méthode d'Otsu . . . . .	124
4.3	Profils d'intensité moyens des 2 sous-populations . . . . .	125
4.4	Courbe "elbow", détermination du k nombre de <i>clusters</i> . . . . .	127
4.5	Partitionnement des ADN de $\lambda$ . . . . .	128
4.6	Résultat du partitionnement . . . . .	129
4.7	Proportions de jonctions récupérées . . . . .	130
4.8	Profils moyen d'intensité pour les différents types de jonctions . . . . .	131
4.9	Comparaison des observations faites sur la réplication de l'ADN de $\lambda$ par peignage moléculaire et nanocanaux . . . . .	133
4.10	Analyse des données générées sur le génome humain . . . . .	135
4.11	Analyse du doublement de fluorescence du YOYO-1, échantillon Xénope	139
4.12	Analyse populationnelle des chromosomes I et II de la levure <i>S. cerevisiae</i>	141
4.13	Aberrations observées sur les données de levure . . . . .	142
5.1	Cartographie de la réplication de l'ADN avec le système Saphyr . . . . .	149
5.2	Exemple d'image obtenue avec le système Saphyr . . . . .	151
5.3	Principe du séquençage Nanopore . . . . .	153
6.1	<i>Post-processing</i> des canaux bleu (YOYO-1) et vert (code-barres) . . . . .	159
6.2	Exemples de profils d'intensité pour les données de <i>S. cerevisiae</i> . . . . .	160



## Bibliographie

- ALKAN, Can, Saba SAJJADIAN et Evan E. EICHLER (2011). « Limitations of next-generation genome sequence assembly ». In : *Nature Methods* 8.1, p. 61–65. ISSN : 15487091. DOI : 10.1038/nmeth.1527. URL : <http://www.nature.com>.
- BELL, Leslie et Breck BYERS (1983). « Separation of branched from linear DNA by two-dimensional gel electrophoresis ». In : *Analytical Biochemistry* 130.2, p. 527–535. ISSN : 00032697. DOI : 10.1016/0003-2697(83)90628-0. URL : <https://linkinghub.elsevier.com/retrieve/pii/0003269783906280>.
- BELL, Stephen P, Ryuji KOBAYASHI et Bruce STILLMAN (2014). « Yeast Origin Recognition Complex Functions in Transcription Silencing Replication Sequence comparison ». In : *Science* 262.5141, p. 1844–1849.
- BENSIMON, Aaron et al. (1994). « Alignment and Sensitive Detection of DNA by a moving interface ». In : *Science* 265.19, p. 2096–2098. URL : <http://www.sciencemag.org/content/265/5181/2096.full.pdf>.
- BENSIMON, D. et al. (1995). « Stretching DNA with a receding meniscus : Experiments and models ». In : *Physical Review Letters* 74.23, p. 4754–4757. ISSN : 00319007. DOI : 10.1103/PhysRevLett.74.4754. URL : <https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.74.4754>.
- BEREZNEY, Ronald, Dharani DUBEY et Joel HUBERMAN (2000). « Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci ». In : *Chromosoma* 108.8, p. 471–484. ISSN : 00095915. DOI : 10.1007/s004120050399.
- BESNARD, Emilie et al. (2012). « Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs ». In : *Nature Structural and Molecular Biology* 19.8, p. 837–844. ISSN : 15459993. DOI : 10.1038/nsmb.2339. URL : <http://www.replicationdomain.com/>.
- BLOW, J. Julian et Anindya DUTTA (2005). « Preventing re-replication of chromosomal DNA ». In : *Nature Reviews Molecular Cell Biology* 6.6, p. 476–486. ISSN : 1471-0072. DOI : 10.1038/nrm1663. URL : <http://www.nature.com/articles/nrm1663>.

- BLOW, J. Julian et Ronald A. LASKEY (1986). « Initiation of DNA replication in nuclei and purified DNA by a cell-free extract of *Xenopus* eggs ». In : *Cell* 47.4, p. 577–587. ISSN : 00928674. DOI : 10.1016/0092-8674(86)90622-7.
- BLOW, J Julian et al. (2001). « Replication origins in *Xenopus* egg extract are 5-15 kilobases apart and are activated in clusters that fire at different times ». In : *Journal of Cell Biology* 152.1, p. 15–25. ISSN : 00219525. DOI : 10.1083/jcb.152.1.15. URL : <http://www.jcb.org/cgi/content/full/152/1/15>.
- BLUMENTHAL, Alan B., Henry J. KRIEGSTEIN et David S. HOGNESS (1974). « Chromosomes *Drosophila melanogaster* The Units of DNA Replication ». In : *Cold Spring Harb Symp Quant Biol*. DOI : 10.1101/SQB.1974.038.01.024. URL : <http://symposium.cshlp.org/content/38/205.refs.html><http://symposium.cshlp.org/subscriptions>.
- BREWER, Bonita J et Walton L FANGMAN (1987). « The localization of replication origins on ARS plasmids in *S. cerevisiae* ». In : *Cell* 51.3, p. 463–471. ISSN : 00928674. DOI : 10.1016/0092-8674(87)90642-8. URL : <http://www.ncbi.nlm.nih.gov/pubmed/2822257>.
- BURHANS, William C. et al. (1990). « Identification of an origin of bidirectional DNA replication in mammalian chromosomes ». In : *Cell* 62.5, p. 955–965. ISSN : 00928674. DOI : 10.1016/0092-8674(90)90270-O. URL : <https://linkinghub.elsevier.com/retrieve/pii/009286749090270O>.
- BURTON, Dennis R. et al. (2009). « Broad and potent neutralizing antibodies from an african donor reveal a new HIV-1 vaccine target ». In : *Science* 326.5950, p. 285–289. ISSN : 00368075. DOI : 10.1126/science.1178746. URL : [www.sciencemag.org/cgi/content/full/1178746/DC1](http://www.sciencemag.org/cgi/content/full/1178746/DC1).
- CAIRNS, John (1963a). « The bacterial chromosome and its manner of replication as seen by autoradiography ». In : *Journal of Molecular Biology* 6.3, 208–IN5. ISSN : 00222836. DOI : 10.1016/S0022-2836(63)80070-4. URL : [http://dx.doi.org/10.1016/S0022-2836\(63\)80070-4](http://dx.doi.org/10.1016/S0022-2836(63)80070-4).
- (1963b). « The Chromosome of *Escherichia coli* ». In : *Cold Spring Harbor Symposia on Quantitative Biology* 28.0, p. 43–46. ISSN : 0091-7451. DOI : 10.1101/SQB.1963.028.01.011. URL : <http://symposium.cshlp.org/cgi/doi/10.1101/SQB.1963.028.01.011>.
- (1966). « Autoradiography of HeLa cell DNA. » In : *Journal of molecular biology* 15.1, p. 372–3. ISSN : 0022-2836. URL : <http://www.ncbi.nlm.nih.gov/pubmed/5912048>.
- CALLAN, H G (1972). « Replication of DNA in the chromosomes of eukaryotes ». In : *Proc R Soc Lond B Biol Sci* 181.62, p. 19–41. URL : [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=4402332](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=4402332).
- CAO, Hongzhi et al. (2014). *Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology*. Rapp. tech. 1. DOI : 10.1186/2047-217X-3-34. URL : <http://www.gigasciencejournal.com/content/3/1/34>.

- CHAKRABORTY, T et al. (1982). *Purification of the E. coli dnaA gene product. ; Purification of the E. coli dnaA gene product.* Rapp. tech. 12, p. 1545–1549. DOI : 10.1002/j.1460-2075.1982.tb01353.x. URL : <https://www.embopress.org/doi/pdf/10.1002/j.1460-2075.1982.tb01353.x>.
- CHAN, Saki et al. (2018). « Structural variation detection and analysis using bionano optical mapping ». In : *Methods in Molecular Biology* 1833, p. 193–203. ISSN : 10643745. DOI : 10.1007/978-1-4939-8666-8{\\_}16.
- CHENG, Jie Zhi et al. (2009). « Detection of arterial calcification in mammograms by random walks ». In : *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5636 LNCS.CC, p. 713–724. ISSN : 03029743. DOI : 10.1007/978-3-642-02498-6{\\_}59.
- COLLINS, F. S. et al. (2004). « Finishing the euchromatic sequence of the human genome ». In : *Nature* 431.7011, p. 931–945. ISSN : 00280836. DOI : 10.1038/nature03001. URL : <http://www.genome.gov/10000923.%20http://www.nature.com/articles/nature03001>.
- CRAIG, Jeffrey et Wendy BICKMORE (1993). « Chromosome Bands - flavours to savour ». In : *BioEssays*, p. 1–6.
- CZAJKOWSKY, Daniel M et al. (2008). « DNA Combing Reveals Intrinsic Temporal Disorder in the Replication of Yeast Chromosome VI ». In : *Journal of Molecular Biology* 375.1, p. 12–19. ISSN : 00222836. DOI : 10.1016/j.jmb.2007.10.046. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2151843/pdf/nihms35685.pdf>.
- DAS, Somes K. et al. (2010). « Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes ». In : *Nucleic Acids Research* 38.18. ISSN : 03051048. DOI : 10.1093/nar/gkq673. URL : <https://academic.oup.com/nar/article-abstract/38/18/e177/1069400>.
- DE CARLI, Francesco, Vincent GAGGIOLI et al. (2016). « Single-molecule, antibody-free fluorescent visualisation of replication tracts along barcoded DNA molecules ». In : *International Journal of Developmental Biology* 60.7-9, p. 297–304. ISSN : 02146282. DOI : 10.1387/ijdb.160139oh. URL : <http://www.intjdevbiol.com/paper.php?doi=160139oh>.
- DE CARLI, Francesco, Nikita MENEZES et al. (2018). « High-Throughput Optical Mapping of Replicating DNA ». In : *Small Methods* 2.9, p. 1800146. ISSN : 23669608. DOI : 10.1002/smt.201800146. URL : <http://doi.wiley.com/10.1002/smt.201800146>.
- DELLINO, GI et al. (2013). « Genome-wide mapping of human DNA-replication origins : Levels of transcription at ORC1 sites regulate origin selection and replication timing ». In : *Genome Research* 23.1, p. 1–11. ISSN : 1088-9051. DOI : 10.1101/gr.142331.112.
- DEWAR, James M et Johannes C WALTER (2017). « Mechanisms of DNA replication termination ». In : *Nature Publishing Group* 18.8, p. 507–516. ISSN : 1471-0072. DOI : 10.1038/nrm.2017.42. URL : <http://dx.doi.org/10.1038/nrm.2017.42>.

- DIFFLEY, J.F. et J.H. COCKER. (1992). « Protein-DNA interactions at a yeast replication origin. » In : *Nature* 359, p. 710–713.
- DIFFLEY, John F.X. et al. (1994). « Two steps in the assembly of complexes at yeast replication origins in vivo ». In : *Cell* 78.2, p. 303–316. ISSN : 00928674. DOI : 10.1016/0092-8674(94)90299-2.
- DM GILBERT, SM Gasser (2006). « Nuclear structure and DNA replication ». In : *DePamphilis ML (ed) DNA replication and human disease. Cold Spring Harbor Laboratory, Cold Spring Harbor.*
- DUBEY, D.D et Rajiva RAMAN (1987). « Factors influencing replicon organization in tissues having different S-phase durations in the mole rat, *Bandicota bengalensis\** ». In : *Chromosoma (Berl.)* P. 285–289. URL : <https://link.springer.com/content/pdf/10.1007%2FBF00294785.pdf>.
- EATON, Matthew L et al. (2010). « Conserved nucleosome positioning defines replication origins ». In : *Genes and Development* 24.8, p. 748–753. ISSN : 08909369. DOI : 10.1101/gad.1913210. URL : <http://www.genesdev.org>.
- ENQUIST, L. W. et A. M. SKALKA (1978). « Replication of bacteriophage lambda DNA ». In : *Trends in Biochemical Sciences* 3.4, p. 279–283. ISSN : 09680004. DOI : 10.1016/S0968-0004(78)96033-4.
- FARKASH-AMAR, Shlomit et al. (2008). « Global organization of replication time zones of the mouse genome ». In : *Genome Research* 18.10, p. 1562–1570. ISSN : 10889051. DOI : 10.1101/gr.079566.108.
- FRAGKOS, Michalis et al. (2015). « DNA replication in eukaryotic cells requires the accurate synthesis of large amounts of DNA ». In : *Nature Publishing Group* 16. DOI : 10.1038/nrm4002. URL : <http://www.nature.com/insb.bib.cnrs.fr/articles/nrm4002.pdf>.
- FRANGI, Alejandro F. et al. (1998). « Multiscale vessel enhancement filtering ». In : October, p. 130–137. DOI : 10.1007/bfb0056195. URL : <http://link.springer.com/10.1007/BFb0056195>.
- GAUTHIER, Michel G. et John BECHHOEFER (2009). « Control of DNA replication by anomalous reaction-diffusion kinetics ». In : *Physical Review Letters* 102.15. ISSN : 00319007. DOI : 10.1103/PhysRevLett.102.158104. URL : <https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.102.158104>.
- GRIGORIEV, Andrei (1998). « Analyzing genomes with cumulative skew diagrams ». In : *Nucleic Acids Research* 26.10, p. 2286–2290. ISSN : 03051048. DOI : 10.1093/nar/26.10.2286. URL : <http://mol.genes.nig.ac.jp/ecoli>.
- GRUNWALD, Assaf et al. (2015). « Bacteriophage strain typing by rapid single molecule analysis ». In : *Nucleic Acids Research* 43.18. ISSN : 13624962. DOI : 10.1093/nar/gkv563. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4605287/pdf/gkv563.pdf>.

- GUILBAUD, Guillaume et al. (2011). « Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome ». In : *PLoS Computational Biology* 7.12. Sous la dir. de Christopher E. PEARSON, e1002322. ISSN : 1553734X. DOI : 10.1371/journal.pcbi.1002322. URL : <https://dx.plos.org/10.1371/journal.pcbi.1002322>.
- H. TALBOT, R. Beare (2002). « The birth of Mathematical Morphology.(with G. Mathéron) ». In : *Mathematical Morphology*, p. 1–16.
- HAND, Roger (1975). « Altered Patterns of Initiation during Inhibition of Protein Synthesis ». In : *The Journal of Cell Biology* 67, p. 761–773. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2111654/pdf/jc673761.pdf>.
- HANDELI, Shlomo et al. (1989). *Mapping replication units in animal cells*. Rapp. tech. 6, p. 909–920. DOI : 10.1016/0092-8674(89)90329-2.
- HANSEN, R. S. et al. (2010). « Sequencing newly replicated DNA reveals widespread plasticity in human replication timing ». In : *Proceedings of the National Academy of Sciences* 107.1, p. 139–144. ISSN : 0027-8424. DOI : 10.1073/pnas.0912402107.
- HARTIGAN, J A et M A WONG (1979). « A K-Means Clustering Algorithm ». In : *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1, p. 100–108. URL : [https://www.labri.fr/perso/bpinaud/userfiles/downloads/hartigan\\_1979\\_kmeans.pdf](https://www.labri.fr/perso/bpinaud/userfiles/downloads/hartigan_1979_kmeans.pdf).
- HARTIGAN, John A. (1975). « Wiley Series in Probability and Mathematical Statistics ». In : p. 354–359. DOI : 10.1002/9781118165485.scard.
- HASTIE, Alex R. et al. (2013). « Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo Sequence Assembly of the Complex *Aegilops tauschii* Genome ». In : *PLoS ONE* 8.2, p. 55864. ISSN : 19326203. DOI : 10.1371/journal.pone.0055864. URL : [www.plosone.org](http://www.plosone.org).
- HENNION, Magali et al. (2019). « Site-specific and dispersed replication initiation in *S. cerevisiae* quantified by nanopore sequencing ». In : *bioRxiv*, p. 1–18.
- HERRICK, J et al. (2000). « Replication fork density increases during DNA synthesis in *X. laevis* egg extracts. » In : *Journal of molecular biology* 300.5, p. 1133–42. ISSN : 0022-2836. DOI : 10.1006/jmbi.2000.3930. URL : <https://linkinghub.elsevier.com/retrieve/pii/S0022283600939305%20http://www.ncbi.nlm.nih.gov/pubmed/10903859>.
- HOLMQUIST, G P (1992). « Chromosome bands, their chromatin flavors, and their functional features ». In : *American Journal of Human Genetics* 51.1, p. 17–37. ISSN : 0002-9297. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1682890/pdf/ajhg00065-0022.pdf>.
- HUBERMAN, J. A. et A. D. RIGGS (1966). *Autoradiography of chromosomal DNA fibers from Chinese hamster cells*. Rapp. tech. 3, p. 599–606. DOI : 10.1073/pnas.55.3.599. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC224194/pdf/pnas00142-0149.pdf>.

- HUBERMAN, Joel et Arthur RIGGS (1968). « On the Mechanism of DNA Replication in Mammalian Chromosomes ». In : *J. Mol. Biol* 32, p. 327–341.
- HYRIEN, Olivier (2015). « Peaks cloaked in the mist : The landscape of mammalian replication origins ». In : *Journal of Cell Biology* 208.2, p. 147–160. ISSN : 15408140. DOI : 10.1083/jcb.201407004. URL : [www.jcb.org/cgi/doi/10.1083/jcb.201407004JCB147](http://www.jcb.org/cgi/doi/10.1083/jcb.201407004JCB147).
- (2016). « Up and down the slope : Replication timing and fork directionality gradients in eukaryotic genomes ». In : *The Initiation of DNA Replication in Eukaryotes*. Cham : Springer International Publishing, p. 65–85. ISBN : 9783319246963. DOI : 10.1007/978-3-319-24696-3\_{\\_}4. URL : [http://link.springer.com/10.1007/978-3-319-24696-3\\_4](http://link.springer.com/10.1007/978-3-319-24696-3_4).
- HYRIEN, Olivier, Kathrin MARHEINEKE et Arach GOLDAR (2003). « Paradoxes of eukaryotic DNA replication : MCM proteins and the random completion problem ». In : *BioEssays* 25.2, p. 116–125. ISSN : 02659247. DOI : 10.1002/bies.10208.
- HYRIEN, Olivier, Aurélien RAPPAILLES et al. (2013). « From Simple Bacterial and Archaeal Replicons to Replication N/U-Domains ». In : *Journal of Molecular Biology* 425, p. 4673–4689. DOI : 10.1016/j.jmb.2013.09.021. URL : <http://dx..>
- JACKSON, Dean A et Ana POMBO (1998). « Replicon clusters are stable units of chromosome structure : Evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells ». In : *Journal of Cell Biology* 140.6, p. 1285–1295. ISSN : 00219525. DOI : 10.1083/jcb.140.6.1285. URL : <http://www.jcb.org>.
- JACOB, F., S. BRENNER et F. CUZIN (1963). « On the Regulation of DNA Replication in Bacteria ». In : *Cold Spring Harbor Symposia on Quantitative Biology* 28.0, p. 329–348. ISSN : 0091-7451. DOI : 10.1101/sqb.1963.028.01.048.
- JONES, T.R. et al. (2006). « Methods for high-content, high-throughput image-based cell screening ». In : *Proceedings of the First MICCAI Workshop on Microscopic Image Analysis with Applications in Biology*, p. 65–72. URL : <https://personal.broadinstitute.org/anne/publications/JonesMIAABPreprint.pdf%20http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.3889&rep=rep1&type=pdf>.
- KARNANI, Neerja et al. (2010). « Genomic Study of Replication Initiation in Human Chromosomes Reveals the Influence of Transcription Regulation and Chromatin Structure on Origin Selection ». In : *Molecular Biology of the Cell* 21, p. 393–404. DOI : 10.1091/mbc.E09. URL : <http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E09>.
- KAYKOV, Atanas et Paul NURSE (2018). « Analysis of fission yeast single DNA molecules on the megabase scale using DNA combing ». In : *Methods in Molecular Biology* 1721, p. 9–24. ISSN : 10643745. DOI : 10.1007/978-1-4939-7546-4\_{\\_}2.
- KELLER, Christian et al. (2002). « The origin recognition complex marks a replication origin in the human TOP1 gene promoter ». In : *Journal of Biological Chemistry* 277.35, p. 31430–31440. ISSN : 00219258. DOI : 10.1074/jbc.M202165200. URL : <http://www.jbc.org/>.

- KLEIN, Kyle et al. (2017). *Genome-Wide Identification of Early-Firing Human Replication Origins by Optical Replication Mapping*. DOI : <https://doi.org/10.1101/214841>. URL : <https://www.biorxiv.org/content/early/2017/11/06/214841>.
- KRAFT, Dieter (1988). *A software package for sequential quadratic programming*. T. 88. 28. Köln, p. 33. URL : <https://www.tib.eu/en/search/id/TIBKAT%3A016896521/A-software-package-for-sequential-quadratic-programming/>.
- KRONENBERG, Zev N. et al. (2018). « High-resolution comparative analysis of great ape genomes ». In : *Science* 360.6393, eaar6343. ISSN : 0036-8075. DOI : 10.1126/science.aar6343.
- KRUDE, Torsten (2000). *Initiation of human DNA replication in vitro using nuclei from cells arrested at an initiation-competent state*. Rapp. tech. 18, p. 13699–13707. DOI : 10.1074/jbc.275.18.13699. URL : <http://www.jbc.org/>.
- LAM, Ernest T et al. (2012). « Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly ». In : *Nature Biotechnology* 30.8, p. 771–776. ISSN : 1087-0156. DOI : 10.1038/nbt.2303. URL : <http://www.nature.com/articles/nbt.2303>.
- LANGLEY, Alexander R et al. (2016). « Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq) ». In : *Nucleic Acids Research* 44.21, p. 10230–10247. DOI : 10.1093/nar/gkw760.
- LASKEY, R A (1985). « Chromosome replication in early development of *Xenopus laevis* ». In : *Journal of Embryology and Experimental Morphology* 89.SUPPL. P. 285–296. ISSN : 0022-0752. URL : <https://dev.biologists.org/content/develop/89/Supplement/285.full.pdf>.
- LATT, Samuel A. (1977). « Fluorescent probes of chromosome structure and replication ». In : *Canadian Journal of Genetics and Cytology* 19.4, p. 603–623. ISSN : 0008-4093. DOI : 10.1139/g77-065.
- LEDERBERG, R M et Joshua LEDERBERG (1951). « Genetic studies of lysogenicity in *Echerichia coli*. » In : *Genetics* 38. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1209586/pdf/51.pdf>.
- LEUNG, Alden King Yung et al. (2017). « OMBlast : Alignment tool for optical mapping using a seed-and-extend approach ». In : *Bioinformatics* 33.3, p. 311–319. ISSN : 14602059. DOI : 10.1093/bioinformatics/btw620. URL : <https://github.com/aldenleung/OMBlast>.
- LEVY-SAKIN, Michal et Yuval EBENSTEIN (2013). « Beyond sequencing : optical mapping of DNA in the age of nanotechnology and nanoscopy ». In : *Current Opinion in Biotechnology* 24.4, p. 690–698. ISSN : 0958-1669. DOI : 10.1016/J.COPBIO.2013.01.009. URL : <https://www.sciencedirect.com/science/article/abs/pii/S0958166913000128?via%3Dihub>.

- LINSKENS, M H et J A HUBERMAN (1988). « Organization of replication of ribosomal DNA in *Saccharomyces cerevisiae*. » In : *Molecular and Cellular Biology* 8.11, p. 4927–4935. ISSN : 0270-7306. DOI : 10.1128/mcb.8.11.4927. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC365586/pdf/molcellb00071-0349.pdf>.
- LOBIONDO, Joseph, Mortimer ABRAMOWITZ et Marc M. FRIEDMAN (2011). « Microscope objectives ». In : *Current Protocols in Cytometry SUPPL.* 58, p. 1–15. ISSN : 19349300. DOI : 10.1002/0471142956.cy0202s58.
- LOBRY, J.R. (1996). « Asymmetric substitution patterns in the two DNA strands of bacteria ». In : *Molecular Biology and Evolution* 13.5, p. 660–665. ISSN : 0737-4038. DOI : 10.1093/oxfordjournals.molbev.a025626. URL : <https://academic.oup.com/mbe/article-abstract/13/5/660/1083035>.
- LUCAS, Isabelle et al. (2000). « Mechanisms ensuring rapid and complete DNA replication despite random initiation in *Xenopus* early embryos ». In : *Journal of Molecular Biology* 296.3, p. 769–786. ISSN : 00222836. DOI : 10.1006/jmbi.2000.3500.
- MAK, Angel C.Y. et al. (2016). « Genome-wide structural variation detection by genome mapping on nanochannel arrays ». In : *Genetics* 202.1, p. 351–362. ISSN : 19432631. DOI : 10.1534/genetics.115.183483. URL : [www.genetics.org/lookup/suppl/](http://www.genetics.org/lookup/suppl/).
- MANTIERO, Davide et al. (2011). « Limiting replication initiation factors execute the temporal programme of origin firing in budding yeast ». In : *EMBO Journal* 30.23, p. 4805–4814. ISSN : 02614189. DOI : 10.1038/emboj.2011.404. URL : [www.embojournal.org](http://www.embojournal.org).
- MARHEINEKE, Kathrin, Arach GOLDAR et al. (2009). « Use of DNA Combing to Study DNA Replication in *Xenopus* and Human Cell-Free Systems ». In : *Methods in Molecular Biology*. ISSN : 00414131. DOI : 10.1007/978-1-60327-817-5.
- MARHEINEKE, Kathrin, Olivier HYRIEN et Torsten KRUDE (2005). « Visualization of bidirectional initiation of chromosomal DNA replication in a human cell free system ». In : *Nucleic Acids Research* 33.21, p. 6931–6941. ISSN : 03051048. DOI : 10.1093/nar/gki994.
- MARTIN, Melvenia M. et al. (2011). « Genome-wide depletion of replication initiation events in highly transcribed regions ». In : *Genome Research* 21.11, p. 1822–1832. ISSN : 1549-5469. DOI : 10.1101/gr.124644.111.
- MASTERS, Millicent et Paul BRODA (1971). « Evidence for the Bidirectional Replication of the *Escherichia coli* Chromosome ». In : *Nature New Biology* 232.31, p. 137–140. ISSN : 0090-0028. DOI : 10.1038/newbio232137a0. URL : <http://www.nature.com/articles/newbio232137a0>.
- MATHERON, G (1967). *Éléments pour une Théorie des Milieux Poreux*. Paris : Masson et Cie. URL : <https://www.worldcat.org/title/elements-pour-une-theorie-des-milieux-poreux/oclc/7503513>.



- MAYA-MENDOZA, A et al. (2010). « S Phase Progression in Human Cells Is Dictated by the Genetic Continuity of DNA Foci ». In : *PLoS Genet* 6.4, p. 1000900. DOI : 10.1371/journal.pgen.1000900. URL : [www.plosgenetics.org](http://www.plosgenetics.org).
- McGUFFEE, Sean R., Duncan J. SMITH et Iestyn WHITEHOUSE (2013). « Quantitative, Genome-Wide Analysis of Eukaryotic Replication Initiation and Termination ». In : *Molecular Cell* 50.1, p. 123–135. ISSN : 10972765. DOI : 10.1016/j.molcel.2013.03.004. URL : <http://dx.doi.org/10.1016/j.molcel.2013.03.004>.
- McKNIGHT, Steven L. et Oscar L. MILLER (1977). « Electron microscopic analysis of chromatin replication in the cellular blastoderm drosophila melanogaster embryo ». In : *Cell* 12.3, p. 795–804. ISSN : 00928674. DOI : 10.1016/0092-8674(77)90278-1. URL : <https://linkinghub.elsevier.com/retrieve/pii/0092867477902781>.
- MESNER, Larry D. et al. (2013). « Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins ». In : *Genome Research* 23.11, p. 1774–1788. ISSN : 10889051. DOI : 10.1101/gr.155218.113.
- MICHAELI, Yael et al. (2013). « Optical detection of epigenetic marks : Sensitive quantification and direct imaging of individual hydroxymethylcytosine bases ». In : *Chemical Communications* 49.77, p. 8599–8601. ISSN : 1364548X. DOI : 10.1039/c3cc42543f.
- MIOTTO, Benoit, Zhe JI et Kevin STRUHL (2016). « Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers ». In : *Proceedings of the National Academy of Sciences* 113.33, E4810–E4819. ISSN : 0027-8424. DOI : 10.1073/pnas.1609060113. URL : <https://www.pnas.org/content/pnas/113/33/E4810.full.pdf>.
- MUKHOPADHYAY, Rituparna et al. (2014). « Allele-Specific Genome-wide Profiling in Human Primary Erythroblasts Reveal Replication Program Organization ». In : *PLoS Genetics* 10.5, p. 1004319. ISSN : 15537404. DOI : 10.1371/journal.pgen.1004319. URL : [www.plosgenetics.org](http://www.plosgenetics.org).
- MÜLLER, Carolin A. et al. (2019). « Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads ». In : *Nature Methods* 16.5, p. 429–436. ISSN : 15487105. DOI : 10.1038/s41592-019-0394-y. URL : <https://doi.org/10.1038/s41592-019-0394-y>.
- MÜLLER, Vilhelm. et Fredrik WESTERLUND (2017). « Optical DNA mapping in nanofluidic devices : principles and applications. » In : *Lab on a Chip* 4. DOI : 10.1039/C6LC01439A.
- NAWOTKA, Kevin A et J A HUBERMAN (1988). « Two-dimensional gel electrophoretic method for mapping DNA replicons. » In : *Molecular and Cellular Biology* 8.4, p. 1408–1413. ISSN : 0270-7306. DOI : 10.1128/mcb.8.4.1408. URL : <http://mcb.asm.org/>.
- NAWY, Tal (2012). « DNA : stretch for the camera ». In : *Nature Methods* 9.9, p. 863–863. ISSN : 1548-7091. DOI : 10.1038/nmeth.2166. URL : <http://dx.doi.org/10.1038/nmeth.2166>.

- NEWLON, Carol S. et James F. THEIS (1993). « The structure and function of yeast ARS elements ». In : *Current Opinion in Genetics and Development* 3.5, p. 752–758. ISSN : 0959437X. DOI : 10.1016/S0959-437X(05)80094-2. URL : <https://linkinghub.elsevier.com/retrieve/pii/S0959437X05800942>.
- NOAKES, Matthew T et al. (2019). « Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage ». In : *Nature Biotechnology* 37.6, p. 651–656. ISSN : 15461696. DOI : 10.1038/s41587-019-0096-0. URL : <https://doi.org/10.1038/s41587-019-0096-0>.
- NORIO, Paolo et Carl L SCHILDKRAUT (2001). *Visualization of DNA replication on individual Epstein-Barr virus episomes*. Rapp. tech. 5550, p. 2361–2364. DOI : 10.1126/science.1064603. URL : <http://science.sciencemag.org/>.
- OTSU, N (1979). « Threshold Selection Method From Gray-Level Histograms. » In : *IEEE Trans Syst Man Cybern* SMC-9.1, p. 62–66. ISSN : 0018-9472. DOI : 10.1109/TSMC.1979.4310076.
- PARRA, I et B WINDLE (1993). « High resolution visual mapping of stretched DNA by fluorescent hybridization. » In : *Nature genetics* 5.1, p. 17–21. ISSN : 1061-4036. DOI : 10.1038/ng0993-17. URL : <http://www.nature.com/ng/journal/v5/n1/pdf/ng0993-17.pdf>.
- PAYNE, Alexander et al. (2018). « Whale watching with BulkVis : A graphical viewer for Oxford Nanopore bulk fast5 files ». In : *bioRxiv*, p. 312256. DOI : 10.1101/312256. URL : <https://www.biorxiv.org/content/10.1101/312256v1>.
- PETRYK, Nataliya et al. (2016). « Replication landscape of the human genome ». In : *Nature Communications*. DOI : 10.1038/ncomms10208. URL : [www.nature.com/naturecommunications](http://www.nature.com/naturecommunications).
- PICARD, Franck et al. (2014). « The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells ». In : *PLoS Genetics* 10.5, p. 1004282. ISSN : 15537404. DOI : 10.1371/journal.pgen.1004282. URL : [www.plosgenetics.org](http://www.plosgenetics.org).
- POPE, Benjamin D et al. (2011). « DNA replication timing is maintained genome-wide in primary human myoblasts independent of D4Z4 contraction in FSH muscular dystrophy ». In : *PLoS ONE* 6.11, p. 27413. ISSN : 19326203. DOI : 10.1371/journal.pone.0027413. URL : [www.plosone.org](http://www.plosone.org).
- PRESCOTT, D M et P L KUEMPEL (1972). « Bidirectional replication of the chromosome in Escherichia coli. » In : *Proceedings of the National Academy of Sciences of the United States of America* 69.10, p. 2842–5. ISSN : 0027-8424. URL : <http://www.ncbi.nlm.nih.gov/pubmed/4562743><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC389658>.

- REIFENBERGER, Jeffrey G, Kevin D DORFMAN et Han CAO (2015). « Topological events in single molecules of E. coli DNA confined in nanochannels ». In : *Cite this : Analyst* 140, p. 4887. DOI : 10.1039/c5an00343a. URL : [www.rsc.org/analyst](http://www.rsc.org/analyst).
- RHIND, Nicholas et David M GILBERT (2013). « DNA Replication Timing ». In : *Cold Spring Harbor perspectives in biology*. DOI : 10.1101/cshperspect.a010132. URL : [www.cshperspectives.org](http://www.cshperspectives.org).
- RYBA, Tyrone et al. (2011). « Genome-scale analysis of replication timing : From bench to bioinformatics ». In : *Nature Protocols* 6.6, p. 870–895. ISSN : 17542189. DOI : 10.1038/nprot.2011.328.
- SALIC, Adrian et Timothy J MITCHISON (2008). *A chemical method for fast and sensitive detection of DNA synthesis in vivo*. Rapp. tech. 7, p. 2415–2420. URL : [www.pnas.org/cgi/doi/10.1073/pnas.0712168105](http://www.pnas.org/cgi/doi/10.1073/pnas.0712168105).
- SCHEPERS, Aloys et Peer PAPIOR (2010). *Why are we where we are ? Understanding replication origins and initiation sites in eukaryotes using ChIP-approaches*. DOI : 10.1007/s10577-009-9087-1. URL : <https://link.springer.com/content/pdf/10.1007%2Fs10577-009-9087-1.pdf>.
- SCHWARTZ, D. et al. (1993). « Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping ». In : *Science* 262.5130, p. 110–114. ISSN : 0036-8075. DOI : 10.1126/science.8211116. URL : <http://www.sciencemag.org/cgi/doi/10.1126/science.8211116>.
- SERRA, Jean (2010). « Introduction to Mathematical Morphology ». In : *Image Processing and Mathematical Morphology*, p. 1–9. DOI : 10.1201/9781420089448-c1.
- SERRA, Jean Paul. et JEAN (1982). *Image analysis and mathematical morphology*. Academic Press. ISBN : 0126372403. URL : <https://dl.acm.org/citation.cfm?id=1098652>.
- SHELTON, Jennifer Marie et al. (2015). « Tools and pipelines for BioNano data : molecule assembly pipeline and FASTA super scaffolding tool ». In : *bioRxiv*, p. 020966. ISSN : 1471-2164. DOI : 10.1101/020966. URL : <http://www.biorxiv.org/content/early/2015/06/15/020966.abstract>.
- SIDDIQUI, Khalid, Kin FAN ON et John F.X. DIFFLEY (2013). « Regulating DNA replication in Eukarya ». In : *Cold Spring Harbor Perspectives in Biology* 5.4, p. 1–17. ISSN : 19430264. DOI : 10.1101/cshperspect.a012922.
- SINGH, S. et al. (2014). « Pipeline for illumination correction of images for high-throughput microscopy ». In : *Journal of Microscopy* 256.3, p. 231–236. ISSN : 13652818. DOI : 10.1111/jmi.12178. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4359755/pdf/jmi0256-0231.pdf>.
- SMITH, Duncan J et Iestyn WHITEHOUSE (2012). « Intrinsic coupling of lagging-strand synthesis to chromatin assembly ». In : *Nature* 483.7390, p. 434–438. ISSN : 00280836. DOI : 10.1038/nature10895. URL : <https://www.nature.com/articles/nature10895.pdf>.

- STANOJCIC, Slavica et al. (2008). « In *Xenopus* egg extracts, DNA replication initiates preferentially at or near asymmetric AT sequences. » In : *Molecular and cellular biology* 28.17, p. 5265–74. ISSN : 1098-5549. DOI : 10.1128/MCB.00181-08. URL : <http://www.ncbi.nlm.nih.gov/pubmed/18573882><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2519731>.
- STINCHCOMB, D, K STRUHL et R DAVIS (1979). « Isolation and characterisation of a yeast chromosomal replicator ». In : *Nature* 282.5734, p. 39–43. ISSN : 0028-0836. DOI : 10.1038/282039a0. URL : <https://www-nature-com.insb.bib.cnrs.fr/articles/282039a0.pdf>.
- STRUHL, Kevin et al. (1978). *High-frequency transformation of yeast : Autonomous replication of hybrid DNA molecules*. Rapp. tech. 3, p. 1035–1039. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC383183/pdf/pnas00003-0033.pdf>.
- TAO, Liang et al. (2000). « Major DNA replication initiation sites in the c-myc locus in human cells ». In : *Journal of Cellular Biochemistry* 78.3, p. 442–457. ISSN : 07302312. DOI : 10.1002/1097-4644(20000901)78:3<442::AID-JCB9>3.0.CO;2-1.
- TAYLOR, J. H. (1958). « The mode of chromosome duplication in *Crepis capillaris* ». In : *Experimental Cell Research* 15.2, p. 350–357. ISSN : 00144827. DOI : 10.1016/0014-4827(58)90036-3. URL : <https://linkinghub.elsevier.com/retrieve/pii/0014482758900363>.
- TÉCHER, Hervé et al. (2013). « Replication dynamics : Biases and robustness of DNA fiber analysis ». In : *Journal of Molecular Biology* 425.23, p. 4845–4855. ISSN : 00222836. DOI : 10.1016/j.jmb.2013.03.040. URL : <http://dx.doi.org/10.1016/j.jmb.2013.03.040>.
- TOUCHON, Marie et al. (2005). « Replication-associated strand asymmetries in mammalian genomes : Toward detection of replication origins ». In : *Proceedings of the National Academy of Sciences* 102.28, p. 9836–9841. ISSN : 0027-8424. DOI : 10.1073/pnas.0500577102. URL : [www.pnas.org/cgi/doi/10.1073/pnas.0500577102](http://www.pnas.org/cgi/doi/10.1073/pnas.0500577102).
- URBAN, John M et al. (2015). « The hunt for origins of DNA replication in multicellular eukaryotes ». In : *F1000 Prime Reports*. DOI : 10.12703/P7-30. URL : <http://f1000.com/prime/reports/b/7/30>.
- VASSILEV, Lyubomir et Edward M JOHNSON (1989). « Mapping initiation sites of DNA replication in vivo using polymerase chain reaction amplification of nascent strand segments ». In : *Nucleic Acids Research*. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC334878/pdf/nar00136-0127.pdf>.
- WANG, Yaping et al. (2011). « Automated DNA fiber tracking and measurement ». In : *Proceedings - International Symposium on Biomedical Imaging May 2014*, p. 1349–1352. ISSN : 19457928. DOI : 10.1109/ISBI.2011.5872650.
- WILLARD, H. F. et S. A. LATT (1976). *Analysis of deoxyribonucleic acid replication in human X chromosomes by fluorescence microscopy*. Rapp. tech., p. 213–227. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1685019/pdf/ajhg00213-0013.pdf>.

- WONG, Philip G et al. (2011). « Cdc45 limits replicon usage from a low density of precs in mammalian cells ». In : *PLoS ONE* 6.3, p. 17533. ISSN : 19326203. DOI : 10.1371/journal.pone.0017533. URL : [www.plosone.org](http://www.plosone.org).
- XIAO, Ming et al. (2007). « Rapid DNA mapping by fluorescent single molecule detection ». In : *Nucleic Acids Research* 35.3. ISSN : 03051048. DOI : 10.1093/nar/gkl1044. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1807959/pdf/gkl1044.pdf>.
- YAN, Xiaomei et al. (2000). « Development of a Mechanism-Based, DNA Staining Protocol Using SYTOX Orange Nucleic Acid Stain and DNA Fragment Sizing Flow Cytometry ». In : *Analytical Biochemistry* 286.1, p. 138–148. ISSN : 0003-2697. DOI : 10.1006/ABIO.2000.4789. URL : <https://www.sciencedirect.com/science/article/pii/S0003269700947894?via%3Dihub>.
- YARDIMCI, Hasan et al. (2012). « Single-molecule analysis of DNA replication in *Xenopus* egg extracts ». In : *Methods* 57.2, p. 179–186. ISSN : 10462023. DOI : 10.1016/j.ymeth.2012.03.033. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3427465/pdf/nihms375430.pdf>.
- YASUDA, Seiichi et Yukinori HIROTA (1977). « Cloning and mapping of the replication origin of *Escherichia coli*. » In : *Proceedings of the National Academy of Sciences* 74.12, p. 5458–5462. ISSN : 0027-8424. DOI : 10.1073/pnas.74.12.5458. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431763/pdf/pnas00043-0266.pdf>.
- YUROV, Yu B. et Natalia A. LIAPUNOVA (1977). « The units of DNA replication in the mammalian chromosomes : Evidence for a large size of replication units ». In : *Chromosoma* 60.3, p. 253–267. ISSN : 00095915. DOI : 10.1007/BF00329774.
- YUROV, Yury B. (1980). « Rate of DNA replication fork movement within a single mammalian cell ». In : *Journal of Molecular Biology* 136.3, p. 339–342. ISSN : 00222836. DOI : 10.1016/0022-2836(80)90378-2. URL : <https://linkinghub.elsevier.com/retrieve/pii/0022283680903782>.
- ZIRKIN, Shahar et al. (2014). « Lighting up individual DNA damage sites by in vitro repair synthesis ». In : *Journal of the American Chemical Society* 136.21, p. 7771–7776. ISSN : 15205126. DOI : 10.1021/ja503677n. URL : <http://pubs.acs.org/doi/10.1021/ja503677n>.



## RÉSUMÉ

---

La réplication de l'ADN est un processus vital qui assure la transmission l'information génétique aux cellules filles. Chez les eucaryotes, la réplication du génome s'effectue en utilisant de multiples origines de réplication. Chez les métazoaires, la cartographie de la réplication demeure difficile. Les cartographies pangénomiques des origines de réplication chez l'Homme réalisées à l'aide de techniques de séquençage, ne s'accordent que modérément. Une explication possible de ces incohérences est que ces approches utilisent de grandes populations cellulaires qui ne nous donne qu'une image moyenne de la réplication. Ainsi, pour mieux comprendre la réplication de l'ADN et accéder à cette variabilité inter-cellulaire, il est fondamental de développer des techniques en molécule unique telle que le peignage moléculaire. Cependant, cette dernière est réfractaire à l'automatisation et empêche l'analyse pangénomique de la réplication. Pour surmonter ces obstacles, nous avons ré-employé un dispositif de cartographie optique basé sur de la microfluidique, le système Bionano Genomics Irys, pour le High Throughput Optical MAPPING of Replicating DNA (HOMARD). Nous recueillons généralement, pour un "run", plus de 34 000 images et plus de 63 000 Mpb d'ADN. Nos nouveaux outils open source, qui ont nécessité l'adaptation du logiciel propriétaire fourni, nous permettent de visualiser simultanément les profils d'intensité de l'ensemble des molécules d'ADN cartographiées, de vérifier la qualité de la cartographie réalisée et, en particulier, de voir où sont situés les segments répliqués au niveau du génome en molécule unique. Nous démontrons la robustesse de notre approche en fournissant, avec une couverture sans précédent (23 311 x), une carte de la réplication de l'ADN bactériophage dupliqué dans des extraits d'œufs *Xenopus* et mettons en évidence le potentiel du système Irys pour l'étude de la réplication de l'ADN et autres études de génomiques fonctionnelles, en plus de son utilisation standard.

## MOTS CLÉS

---

Réplication de l'ADN, Traitement d'images, Cartographie Optique

## ABSTRACT

---

DNA replication is a vital process ensuring accurate conveyance of the genetic information to the daughter cells. In eukaryotic organisms, genome replication is carried out by using multiple start sites, also known as replication origins. In metazoans, the mapping of replication remains challenging. Genome wide mapping of human replication origins performed using sequencing techniques only modestly agree. These existing genome wide approaches use large cell populations that smooth out variability between chromosomal copies that could explain this inconsistency. Thus, to get a better understanding of DNA replication and to uncover the cell-to-cell variability, the development of single molecule techniques is fundamental. DNA combing, a widespread technique used to map DNA replication at a single molecule level, is refractory to automation, forestalling genome-wide analysis. To overcome these impediments, we repurposed an optical DNA mapping device based on microfluidics, the Bionano Genomics Irys system, for High-throughput Optical MAPPING of Replicating DNA (HOMARD). We typically collect, for a single run, over 34 000 images and more than 63 000 Mbp of DNA. Our new open source tools, that required the adaptation of the provided proprietary software, empower us to simultaneously visualize the intensity profiles of all mapped DNA molecules, check the optical mapping performed and, in particular, see where the replication tracks are located genome-wide at a single molecule level. We demonstrate the robustness of our approach by providing an ultra-high coverage (23,311 x) replication map of bacteriophage DNA in *Xenopus* egg extracts and the potential of the Irys system for DNA replication and other functional genomic studies apart from its standard use.

## KEYWORDS

---

DNA replication, Image analysis, Optical Mapping