Optimisation et automatisation d'outils de bio-imagerie 3D pour caractériser la morphologie nucléaire et l'organisation de la chromatine

Tristan Dubos

Thèse présentée pour obtenir le grade de docteur délivré par l'Université Clermont Auvergne

École doctorale : Sciences Pour l'Ingénieur

Spécialité Bio-informatique

Optimisation et automatisation d'outils de bio-imagerie 3D pour caractériser la morphologie nucléaire et l'organisation de la chromatine

---

présentée et soutenue publiquement par

Dubos Tristan

le 17 septembre 2021

Directeurs de thèse : **Sophie Desset, Emilie Péry, Frédéric Chausse**

Jury

Célia Baroux,	Maître de Conférences, Université Zurich	Rapportrice
Nicolas Passat,	Professeur, Université Reims Champagne-Ardenne	Rapporteur
David Legland,	Ingénieur de recherche, INRAE Nantes	Examinateur
Sophie Desset,	Ingénieure de recherche, GReD	Co-directrice
Emilie Péry,	Maître de Conférences, Institut Pascal	Co-directrice
Frédéric Chausse,	Professeur, Institut Pascal	Co-directeur

THESE

Remerciements

Je voudrais commencer par remercier mes co-directeurs de thèse. Un grand merci à Sophie Desset, Emilie Péry, Christophe totout et Frédéric Chausse. Sophie, tu es à l'initiative de ce projet de thèse ; merci pour ton accueil, ton optimisme et pour le temps consacré à mon encadrement au cours de nos longues discussions. Merci de m'avoir fait découvrir le monde de la microscopie et de m'avoir accordé ta confiance. Merci aussi à toi et à Stéphane d'avoir partagé vos coins à champignons. Emilie et Frédéric vous avez accepté de rejoindre mes encadrants en cours de thèse ; merci pour les échanges sans limites et le partage des idées qui m'ont été très profitables lorsque je rencontrais des points de blocage. Christophe, merci pour avoir pensé à moi lors de l'initiation de ce projet de thèse. Merci pour ton dynamisme et ta passion pour la recherche.

Je voudrais aussi adresser un merci particulier à toute l'équipe pour la bonne ambiance sans laquelle je n'aurais pas pu surmonter mes coups de mou ! Merci à Aline pour ses nombreuses attentions sucrées. Merci à Manu pour m'avoir « toléré » durant 3 ans. Merci Aurélia et Sam, merci encore de fredonner tes chansons « phare » (Salade de fruit ! ! ! !). Merci à Sylvie et Sylviane (les inséparables) qui m'ont permis d'accéder au bâtiment de Math ! Merci à Simon pour sa bienveillance et surtout son sérieux ! Et enfin merci à mon p'tit Corto pour ces supers moments salés dans le bureau, les lacs en été et tous les moments passés ensemble ! Je remercie toutes les personnes du laboratoire que j'ai croisées durant ces 3 ans. Merci à tous les membres des Graduates pour les soirées thésards. Merci à Mélu pour les moments de détente en musique (ta reconversion musicale est sur la bonne voie ! ! ! :)). Un grand merci à Coto Margaux pour les pauses sur la terrasse et les bières de réconfort ! Merci à Marie-Jo C. Martinez pour toute son énergie et ses cours en portugais :)

Merci à toutes les personnes avec qui j'ai collaboré en particulier Cédric et Pierre pour nos passionnantes discussions en informatique. Merci encore à Gisèle et François pour m'avoir formé et fait découvrir la bio-informatique.

Merci à tous ceux qui m'ont permis de décompresser quand cela était nécessaire : aux acharnés de la pétanque, Gab, JP, Tiff et Gaël ! Merci à Coco pour toutes nos sorties escalades présentes passées et futures ;) !

Merci à Plume et Lila, mes petites boules de poils ! Merci pour tout Axel (tu n'entres dans aucune case, je te mets avec les animaux !)

Enfin je remercie mes parents pour m'avoir soutenu tout au long de ce projet. Merci à petit nuage pour toutes les attentions de ces derniers temps. Merci aux frangins de Jacou Dav, Cel et Bixou.

Résumé

Le noyau est un organite essentiel des cellules eucaryotes qui présente une morphologie dynamique et dont le contenu est organisé en domaines. Nous avons développé le plugin ImageJ appelé NucleusJ pour le calcul de paramètres caractérisant la forme et la taille des noyaux ainsi que les objets intra-nucléaires tels que les domaines de chromatine et leur position par rapport à la périphérie nucléaire [Poulet et al., 2015]. Une nouvelle version complètement automatisée appelée NucleusJ2.0 a été publiée [Dubos et al., 2020] comportant de nouvelles fonctionnalités et des alternatives aux étapes manuelles. Une étape dite d'*autocrop* a été implémentée pour capturer automatiquement un grand nombre de noyaux à différentes profondeurs dans une pile d'images grand champ. Nous avons complété la méthode de segmentation du noyau, un seuillage Otsu modifié, en adaptant des algorithmes de définition de l'enveloppe convexe en 3D tels que le *gift wrapping* (marche de Jarvis) et la méthode du parcours de Graham avant de les intégrer dans NucleusJ2.0. Afin de s'affranchir d'une étape semi-automatique inhérente à l'étape de ligne de partage des eaux utilisée pour décrire le contenu en chromatine, NODeJ a été développé pour segmenter automatiquement les domaines nucléaires de forte intensité en appliquant la méthode du gradient. De plus, pour améliorer l'efficacité de notre flux d'analyse d'images, une nouvelle bibliothèque appelée Simple OMERO Client a été conçue afin d'utiliser NucleusJ2.0 directement sur des images stockées au sein de la plateforme OMERO tout en bénéficiant de la puissance de calcul d'un serveur distant. Tous ces développements ont été utilisés collectivement dans un projet de preuve de concept où la structure tridimensionnelle du spermatozoïde est évaluée comme indicateur de sa qualité. Enfin, dans une phase plus exploratoire, un modèle U-Net a été entraîné à l'aide d'images segmentées par NucleusJ2.0 afin d'explorer les nouvelles méthodes innovantes de *deep learning*. Les développements présentés dans ce manuscrit ont été conçus pour optimiser le stockage et l'analyse d'images 3D de noyaux et pour évoluer vers des analyses à plus haut débit avec le souhait d'assurer un haut niveau de répétabilité, de reproductibilité et d'accessibilité. De tels développements ouvriront la voie à une meilleure description du noyau en 3D et, nous l'espérons, bénéficieront au plus grand nombre de chercheurs dans le domaine de l'imagerie 3D dédiée à la microscopie afin de faire progresser nos connaissances sur le rôle du noyau dans la régulation de l'expression des gènes.

Abstract

The nucleus is a compartmentalized organelle of the eukaryotic cells with a dynamic morphology and containing distinct chromosomal domains and nuclear bodies. To link nuclear structure and function, we have developed a simple and user-friendly ImageJ plugin called NucleusJ to quantify in 3D the nuclear morphology as well as positioning and organization of nuclear domains. From confocal images, the workflow applies for a batch of images a modified Otsu thresholding method to segment the nuclear space and a 3D watershed algorithm to delimit chromatin domains by partitioning the nucleus. Quantitative parameters are computed including shape and size of nuclei as well as intra-nuclear objects such as chromatin domains and their position in respect to the nuclear periphery [Poulet et al., 2015]. New improvements have been added to NucleusJ and a new version has been released [Dubos et al., 2020]. First, the current version involves manual intervention to delimit a bounding volume including each considered nucleus. To overcome this limitation and automatically capture large numbers of nuclei at various depths in the original image, the so-called *autocrop* procedure has been implemented to automatically detect the nuclei from wide field images. Spatial positions of the cropped nuclei are then recorded in order to estimate distance maps as a new estimator of spatial distribution of the nuclei in a whole tissue context. Second, as the main step in our workflow is to delimit the nucleus or intra-nuclear objects from the background alternatives to the 3D watershed which cannot be fully automated have been explored. To this aim, we adapted convex hull algorithms in 3D such as the *gift wrapping* (Jarvis march) and the Graham scan methods based on discrete geometry and integrated into the NucleusJ workflow. Moreover, to get rid of the semi-automated step inherent to the 3D watershed step, NODeJ was developed to automatically segment high-intensity nuclear domains such as chromocenters by applying the gradient method on the mask of a nucleus to enhance the contrast of intra-nuclear objects and allow their segmentation. Finally, to improve the efficiency of our image analysis workflow and move to high-throughput image analysis, a new library called OMERO Service Client was designed in order to directly load NucleusJ on images stored in a well-organized manner within the OMERO platform while benefiting from the computing power of a remote server. These developments were collectively used in a proof-of-concept project to evaluate the relevance of nuclear morphology and chromatin organization traits computed by NucleusJ to predict the viability from human spermatozoa from large field images stored in OMERO. Finally, in a more exploratory phase, a U-Net model was trained using ground truth datasets gained from NucleusJ segmentation in order to explore new innovative methods to be implemented in the future. The developments presented in this manuscript have been designed to optimize the 3D image storage and analysis of nuclei and to move towards higher throughput analyses. Our workflow was designed with the aim to ensure a high standard of repeatability, reproducibility and accessibility. Such developments will open the way to a better description of the 3D nucleus that we hope will benefit the largest number of researcher in the field and advance our knowledge on the role of the nucleus in the regulation of gene expression.

Liste des abréviations

2D 2 dimensions

3D 3 dimensions

A. thaliana Arabidopsis thaliana

ADN Acide désoxyribonucléique

API Interface de Programmation d'Applications

ARN Acide ribonucléique

ARNr Acide ribonucléique ribosomique

AuBi Auvergne BioInformatique

Cc Chromocentres

CLI Command Line Interface

CNN Convolutional Neural Network

CRWN Crowded Nuclei

DAPI 4',6-diamidino-2-phénylindole

DIC Differential Interference Contrast

ER Réticulum endoplasmique

FiSH Fluorescence in Situ Hybridization

FITC Isothiocyanate de fluorescéine

GC Cellule de garde

GFP Green Fluorescent Protein

INDEPTH Impact of nuclear domains on gene expression and plant traits

INM Inner Nuclear Menbrane

KASH Klarsicht/ANC-1/Syne Homology

LAD Lamina-Associated Domains

LINC Linker of Nucleoskeleton and Cytoskeleton

NEAP Nuclear Envelope Anchored Protein

NJ1 NucleusJ

NJ2 NucleusJ2.0

NODeJ Nuclear Object Detection

NPC Nuclear Pore Complex

ONM Outer Nuclear Menbrane

PC Cellule de pavement

PI Propidium

PLAD Plant Lamina-Associated Domains

RNA-seq RNA sequencing

ROI Region Of Interest

SOC Service OMERO Client

SUN Sad1/UNc84

Table des matières

Introduction générale

La microscopie en biologie est un domaine en constant progrès générant des jeux de données d'images dont la complexité et la taille sont de plus en plus importantes. Cet afflux de données s'accompagne de grands challenges pour le traitement et le stockage des images produites. Ainsi, après l'étape d'acquisition, il est nécessaire d'appliquer divers traitements aux images afin d'en extraire les informations biologiques et de les quantifier. A ce titre, la segmentation, qui consiste à détecter la ou les régions d'intérêt, constitue l'une des étapes essentielles dans l'analyse d'images. L'imagerie au service de la biologie doit être considérée comme un processus continu. En effet, il est nécessaire de concevoir l'expérience biologique comme un tout de la préparation de l'échantillon, à l'acquisition jusqu'à la mise en place des outils permettant l'analyse des images et leur stockage de façon ordonnée et pérenne. La motivation première de ma thèse s'inscrit donc dans ce continuum et vise à automatiser l'analyse d'image de noyaux en 3 dimensions (3D) en utilisant principalement la plante *Arabidopsis thaliana* comme modèle. Cet organisme à la croissance rapide, dont le génome de petite taille est organisé en 5 chromosomes est séquencé depuis près de 20 ans, dispose de nombreuses ressources disponibles pour la communauté scientifique comme par exemple de grandes collections de mutants ce qui en fait un excellent organisme modèle en génétique. Cette plante est utilisée dans notre équipe de recherche afin de mieux comprendre la régulation de l'expression des gènes. Ainsi, les développements informatiques que j'ai réalisés durant mes années de thèse permettent de caractériser le noyau de façon quantitative afin de comprendre comment l'organisation tridimensionnelle du noyau affecte l'expression des gènes. Au départ de ce travail, malgré de multiples efforts pour standardiser la production des échantillons biologiques [Desset et al., 2018] et la mise en place de méthodes semi-automatiques d'analyse des images avec la création du logiciel NucleusJ, un plugin d'ImageJ [Poulet et al., 2015], la caractérisation des images de noyaux 3D restait encore très manuelle et nécessitait des temps d'analyse importants. L'analyse d'images n'avait donc pas atteint à ce stade le haut débit et les résultats produits restaient difficiles à reproduire car dépendants de l'utilisateur.

Chapitre 1

Contexte

Le présent manuscrit s'inscrit dans une logique de mise en place d'une analyse d'images à haut débit avec le souhait de standardiser nos méthodes d'analyse. Cet objectif, qui peut apparaître très technologique, est doublé de plusieurs enjeux scientifiques. En effet, derrière ce travail de développement dédié à l'analyse d'images, notre objectif est de pouvoir effectuer des études en architecture nucléaire de la cellule eucaryote c'est-à-dire l'étude de l'organisation spatiale d'un chromosome voire d'un gène dans le noyau. La position d'un gène dans le noyau est un élément majeur déterminant son niveau d'expression notamment en agissant sur son niveau de transcription. A titre d'exemple, l'exposition d'une plante à la lumière induit le repositionnement nucléaire d'un ensemble de gènes et active leur expression qui est essentielle à l'activité photosynthétique et la survie de la plante [Feng et al., 2014]. De plus, chez l'humain, la présence d'un chromosome 21 supplémentaire dans une maladie génétique telle que le syndrome de Down, conduit à une perturbation transcriptionnelle à l'échelle du génome [Letourneau et al., 2014].

J'ai donc articulé cette première partie de mon manuscrit de façon à décrire nos connaissances sur l'organisation 3D du noyau, son rôle dans l'organisation du génome et la régulation de la transcription. J'ai ensuite décrit les outils de microscopie et d'analyse d'images 3D les plus couramment utilisés ainsi que leur spécificité en termes de production d'images. Enfin, j'ai proposé une classification des stratégies d'analyse d'image des méthodes classiques utilisées de longue date aux méthodes plus récentes basées sur l'intelligence artificielle afin de mieux positionner mon travail.

1.1 Étude de l'architecture du noyau et son contenu en ADN

Le noyau est un compartiment cellulaire abritant la majeure partie de l'information génétique d'une cellule eucaryote. Il est apparu il y a environ 1,5 milliard d'années sans que son origine exacte ne soit complètement résolue [Thomas, 2010]. Les recherches sur l'origine évolutive du noyau sont basées sur la comparaison entre eucaryotes et procaryotes dont le matériel génétique est libre dans la cellule et ces études ont été relancées lors de la découverte des Archées il y a une cinquantaine d'années [Martin, 2005]. En effet, comme les eucaryotes, les Archées possèdent un noyau. Toutefois, le fait que les Archées puissent représenter un intermédiaire entre procaryotes et eucaryotes n'est pas complètement clair [Eme et al., 2017]. La présence d'un noyau chez les eucaryotes et les Archées questionne quant aux origines membranaires mobilisées lors de la mise en place de l'enveloppe nucléaire et plusieurs hypothèses non-exclusives incluant une invagination de la membrane plasmique, la fusion de vésicules ou l'implication du réticulum endoplasmique restent discutées [Martin, 2005].

Le noyau est un organite central dans la cellule eucaryote qui contient plusieurs sous-régions ou domaines regroupés sous le terme de corps nucléaires [Mao et al., 2011]. Le noyau joue un rôle essentiel dans la protection et le maintien de l'intégrité des chromosomes, participe activement au bon déroulement de la division cellulaire et donc au développement et à la reproduction des organismes eucaryotes. Il est au cœur de processus biologiques essentiels et contient toute la machinerie nécessaire à la réplication de l'ADN, la transcription des gènes et leur régulation. Depuis longtemps, les biologistes ont également noté que le noyau peut adopter une forme et un volume variables (Figure 1.1) [Skinner and Johnson, 2017]. L'existence d'une relation entre sa forme et l'organisation du contenu nucléaire au cours du développement, dans différents tissus ou encore au cours de certaines maladies humaines a permis d'utiliser sa morphologie comme biomarqueur dans certaines pathologies comme le cancer [Zink et al., 2004]. Le noyau est donc pour l'ensemble de ces raisons un objet d'étude important aussi bien en biologie fondamentale qu'en santé humaine. Après une description générale des différentes structures qui se trouvent au sein du noyau, depuis les différents constituants de la périphérie jusqu'au contenu en chromatine, je présenterai ensuite la thématique de recherche de l'équipe « Dynamique de la chromatine et du noyau 3D » (CODED) afin de mieux positionner le travail réalisé au cours de mon projet de thèse.

1.1.1 Le noyau est un compartiment organisé

1.1.1.1 Le noyau est un organite central de la cellule eucaryote

Le noyau est délimité par l'enveloppe nucléaire qui est une structure multifonctionnelle. L'enveloppe nucléaire compartimente la transcription et la réplication dans le noyau et la traduction dans le cytoplasme tout en permettant de nombreux échanges nucléo-cytoplasmiques au travers des pores nucléaires et en assurant des interactions à la fois à l'extérieur et à l'intérieur du noyau grâce à sa double membrane lipidique (Figure 1.2).

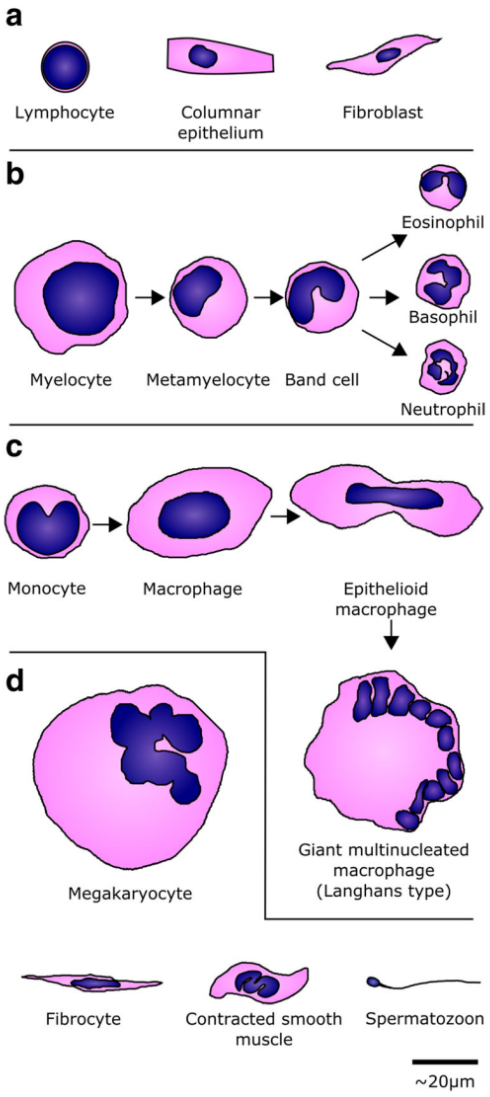


FIGURE 1.1: **Variabilité de la morphologie nucléaire** : exemples de types cellulaires chez l'homme, avec le noyau représenté en bleu et le cytoplasme en rose. En **a** des formes de noyaux arrondis. En **b** les stades de la lignée des granulocytes et **c** des monocytes, décrivant des noyaux lobés. En **d** d'autres exemples de formes allongées dans les fibroblastes ou encore le noyau très condensé du spermatozoïde. Source [Skinner and Johnson, 2017].

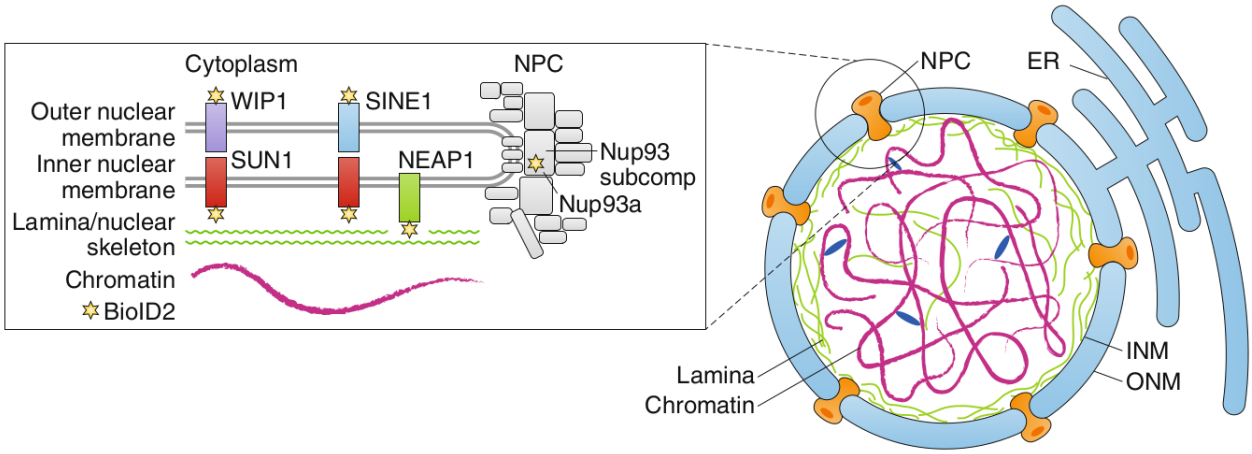


FIGURE 1.2: **Structure schématique de la périphérie nucléaire chez *Arabidopsis thaliana*** : L'enveloppe nucléaire est une double membrane constituée d'une membrane externe (ONM) contenant des protéines comme WIP1 et SINE1 deux protéines à domaine KASH qui interagissent avec le cytosquelette et la membrane interne (INM) contenant la protéine SUN1 une protéine à domaine SUN ainsi que NEAP1 une protéine qui pourrait interagir avec des composants à l'intérieur du noyau. Les protéines à domaine SUN et KASH constituent le complexe LINC. La membrane nucléaire permet les échanges entre le noyau et le cytoplasme grâce aux pores nucléaires (NPC) constitués de nombreuses nucléoporines comme Nup93. Le nucléosquelette est situé immédiatement sous l'INM et est en contact avec la chromatine. Source [Graumann and Evans, 2020].

1 - **La membrane externe** (ONM sur la Figure 1.2) est en continuité avec le réticulum endoplasmique (ER sur la Figure 1.2) et interagit avec d'autres composants de la cellule comme le cytosquelette. L'un des acteurs centraux de cette connexion est le complexe LINC (*LInker of Nucleoskeleton and Cytoskeleton*) qui, grâce aux protéines KASH (Klarsicht/ANC-1/Syne Homology), assure indirectement l'ancrage du cytosquelette à l'enveloppe nucléaire [Zhou et al., 2012]. Notre équipe a d'ailleurs permis de mieux comprendre la fonction du complexe LINC chez les plantes en caractérisant une nouvelle protéine KASH chez Arabidopsis thaliana [Graumann, 2014]. La liaison aux protéines du cytosquelette pendant la mitose est essentielle car elle permet la ségrégation correcte des chromosomes au cours des divisions cellulaires pour générer deux cellules filles viables. Chez les cellules végétales, cette connexion avec le cytosquelette contribue aux propriétés mécaniques du noyau et permet la migration nucléaire par exemple lors des alternances jour-nuit [Higa et al., 2014] ou encore lors de la polarisation du noyau dans certains types cellulaires comme lors de la différenciation des cellules de garde et de pavement de l'épiderme des feuilles [Yang et al., 2020]. L'enveloppe nucléaire est aussi un lieu d'échanges avec le cytoplasme de macromolécules comme les ARN messagers ou les protéines via les pores nucléaires (NPC sur la Figure 1.2). Les pores nucléaires sont organisés en octamère d'un complexe multiprotéique formé d'environ une trentaine de nucléoporines [Tamura et al., 2010]. Il existerait plusieurs centaines à plusieurs milliers de pores nucléaires par noyau régulant les échanges nucléocytoplasmiques [Fiserova et al., 2009].

2 - **La membrane interne** (INM sur la Figure 1.2) permet la transmission de ce lien mécanique. En effet, comme mentionné ci-dessus, si le complexe LINC interagit avec le cytosquelette côté cytoplasmique grâce aux protéines à domaine KASH, ces dernières interagissent directement avec les protéines à domaine SUN (Sad1/UNc84) qui sont, elles, ancrées dans la membrane interne et connues pour interagir avec le nucléosquelette coté intra-nucléaire [Haque et al., 2006, Graumann, 2014] (Figure 1.2). Ainsi, le complexe LINC assurerait la transmission de signaux d'origine extracellulaire vers l'intérieur du noyau grâce à un continuum entre le cytosquelette, la membrane nucléaire et le nucléosquelette [Goswami et al., 2020].

3 - **Le nucléosquelette** est une structure de type fibrillaire située immédiatement sous l'enveloppe nucléaire interne. Cette structure aussi appelée lamina interagit avec les chromosomes et contribue à leur position dans l'espace 3D du noyau avec potentiellement des conséquences sur l'expression du génome [Bickmore, 2013]. Chez les animaux, les lamines sont les constituants essentiels du nucléosquelette et des mutations dans les lamines, en particulier dans la lamine A, sont associées à un grand nombre de maladies humaines appelées laminopathies, rappelant des pathologies associées à des altérations du complexe LINC. Les orthologues les plus probables des lamines chez les plantes sont les protéines NUCLEAR MATRIX CONSTITUENT PROTEIN (NMCP) découvertes chez la carotte et ses orthologues d'*A. thaliana* CROWDED NUCLEI (CRWN) [Dittmer et al., 2007]. L'interaction des protéines CRWNs avec SUN1 et SUN2 a été démontrée et cette interaction pourrait participer au recrutement des protéines CRWN à la périphérie du noyau [Graumann, 2014]. Comme les mutants de la lamine animale qui présentent de petits noyaux déformés, les mutants *crwn1* et *crwn4* ont des noyaux petits, plus arrondis [Wang et al., 2013]. Un autre composant du nucléosquelette végétal est KAKU4, découvert dans un crible pour des mutants avec des noyaux plus petits et sphériques [Goto et al., 2014]. Chez *A. thaliana*, KAKU4 interagit avec CRWN1 et CRWN4 et sa surexpression entraîne une surcroissance de l'enveloppe nucléaire [Goto et al., 2014]. CRWN1 interagit directement avec la chromatine, comme le confirme la récente analyse d'immunoprécipitation de la chromatine effectuée à l'échelle du génome (ChIP-Seq) réalisée par [Hu et al., 2019] et qui permet d'identifier les régions de la chromatine reconnues par CRWN1. Ces séquences ont été appelées Plant Lamina-Associated Domains (PLAD) par analogie aux LADs bien caractérisés chez les cellules d'insectes ou les cellules animales. Les séquences PLADs sont pour la plupart des régions chromosomiques associées à une faible transcription et comprennent des éléments génétiques mobiles (les éléments transposables) qui sont proposés être des points d'ancrage plus forts pour la liaison de CRWN1 [Hu et al., 2019].

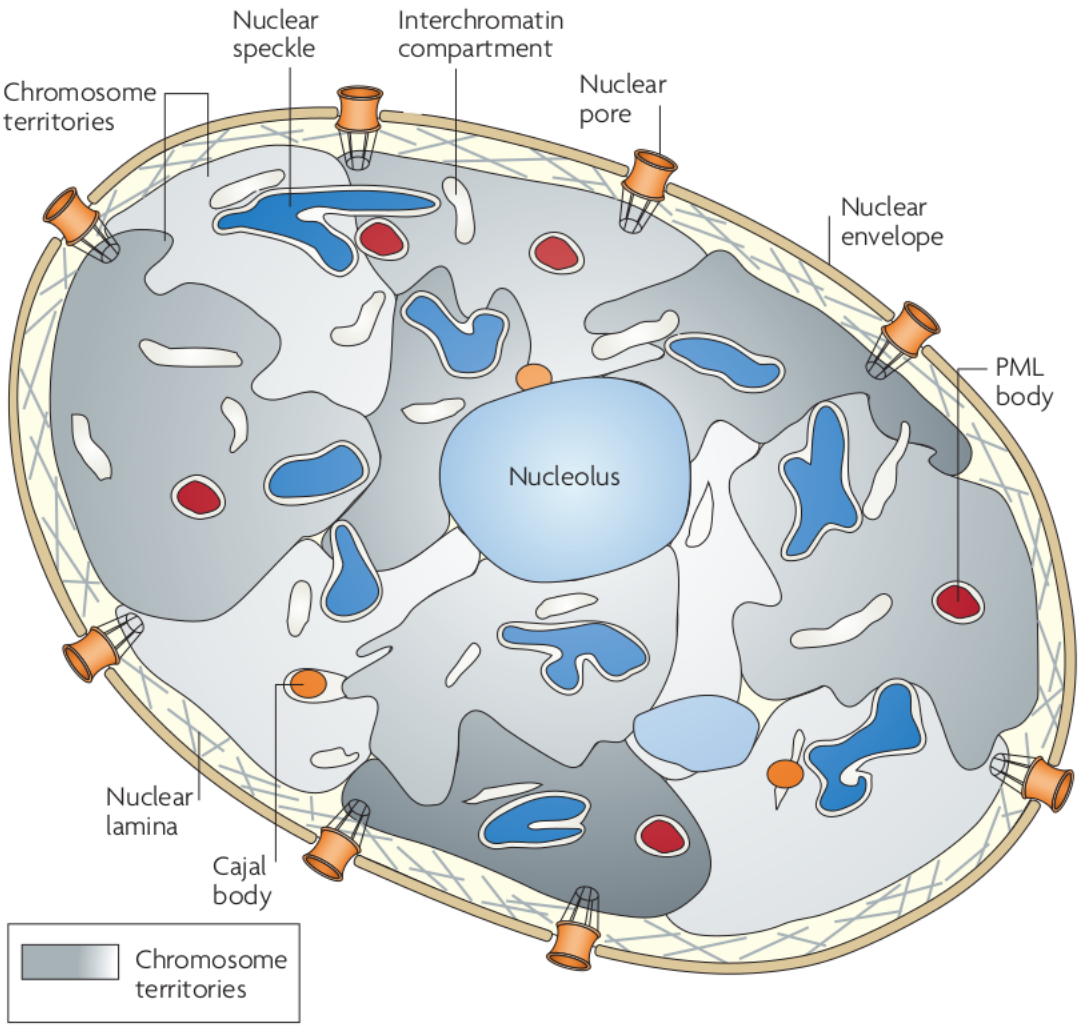


FIGURE 1.3: **Organisation en compartiments nucléaires dans un noyau de cellule de mammifère** : Adapté de [Lanctôt et al., 2007]. L'enveloppe nucléaire délimite le contour du noyau et contient les pores nucléaires (en orange). Le réseau de lamines est représenté en gris sous l'enveloppe nucléaire et permet de structurer le noyau et d'interagir avec la chromatine. Au sein du noyau chaque chromosome occupe un territoire chromosomique (différents niveaux de gris). La transcription, est en général active aux frontières des territoires chromosomiques et forme des transcription factories. Quelques domaines nucléaires comme le nucléole, les Nuclear speckles, les PML bodies et les Cajal bodies sont représentés et localisés à l'intérieur des territoires chromosomiques.

4 - **Le nucléoplasme**, qui correspond à l'espace intra-nucléaire, est un compartiment complexe dont l'organisation tridimensionnelle est le sujet d'un intérêt croissant ces dernières années notamment grâce à des consortium internationaux comme INC pour les cellules d'insecte, animale et humaine (International Nucleome Consortium ; https://inc-cost.eu/) ou INDEPTH pour les cellules végétales (Impact of nuclear domains on gene expression and plant traits ; https://www.brookes.ac.uk/indepth/). En effet, l'organisation du matériel génétique dans le noyau n'est pas aléatoire et joue un rôle important pour la régulation de la transcription, un processus essentielle dans l'établissement des caractères génétiques [Rodriguez-Granados et al., 2016].

1.1.1.2 L'ADN nucléaire est organisé en chromatine

Après la découverte de la double hélice d'ADN au siècle dernier, l'une des avancées les plus importantes de ces vingt dernières années en génétique est sans doute la découverte d'un ensemble de mécanismes visant notamment à emballer et compacter l'ADN en chromatine grâce aux protéines histones. Ainsi, 147 paires de bases (bp) de la double hélice d'ADN s'enroulent autour d'un octamère d'histone pour former le nucléosome, l'élément de base de la chromatine. La chromatine est souvent présentée comme un collier de perles où chaque perle représente un nucléosome et où la répartition des nucléosomes peut être plus ou moins dense. Cette organisation permet à la chromatine d'adopter différentes conformations plus ou moins compactes. Ces différents niveaux de compaction font également intervenir des modifications chimiques sur l'ADN ou sur les histones qui changent les propriétés physico-chimiques de la chromatine et modifient sa capacité d'interaction avec d'autres protéines non-histones ou des facteurs de transcription [Probst et al., 2009]. La chromatine est un point central de la régulation de l'expression des gènes permettant ou non leur expression. Schématiquement, il existe plusieurs niveaux de compaction de la chromatine initialement décrits il y a environ une centaine d'années sur la base d'observations cytologiques [Heitz, 1928] : l'hétérochromatine très compactée et défavorable à l'expression des gènes et l'euchromatine constituant un état plus relâché et favorable à l'expression des gènes. L'étude de l'ensemble de ces facteurs, qui organisent la chromatine dans l'espace nucléaire et influencent l'expression des gènes, est devenue un domaine à part entière de la génétique appelé **l'épigénétique**. L'épigénétique peut se définir comme l'ensemble des facteurs affectant l'expression d'un gène sans modifier sa séquence d'ADN, qui sont transmissibles à une cellule fille après la division cellulaire, réversibles et définissent le niveau de transcription d'un gène [Probst et al., 2009]. Ces mécanismes regroupent les modifications de l'ADN comme la méthylation, les modifications post-traductionnelles des histones et les

variants d'histones, les petits ARN non codants, le remodelage de la chromatine et l'architecture nucléaire [Finn and Misteli, 2019].

1.1.1.3 Le noyau à l'interphase est organisé en sous domaines fonctionnels

Entre deux divisions cellulaires, le matériel génétique passe d'un statut hyper condensé (un ensemble de chromosomes métaphasiques) à un état plus relâché (le noyau à l'interphase). A l'interphase, de nombreux sous-compartiments se forment [Costa-Nunes et al., 2014].

Les chromosomes dans le noyau interphasique s'organisent en territoires chromosomiques comme l'avait observé Carl Rabl en étudiant des cellules animales au microscope [Rabl, 1885]. La mise en place des territoires chromosomiques pourrait dépendre de la densité en gènes (modèle radial), les régions les plus riches en gènes étant les plus centrales et éloignées de la périphérie nucléaire et inversement pour les régions pauvres en gènes qui seraient elles proches de la périphérie [Bickmore and Van Steensel, 2013]. L'organisation en territoires chromosomiques pourrait également être déterminée par des interactions spécifiques entre différentes régions chromosomiques. L'idée sous-jacente dans ces deux modèles est que l'organisation des chromosomes à l'interphase n'est pas aléatoire. Chez Arabidopsis, l'existence de territoires chromosomiques a également été confirmée par des expériences de FiSH (Fluorescence *In Situ* Hybridization) en utilisant un ensemble de sondes spécifiques de chacun des 5 chromosomes marquées avec des fluorochromes de couleurs différentes [Pecinka et al., 2004]. Arabidopsis est une espèce modèle diploïde (2n=10 chromosomes) avec un petit génome complètement séquencé (∼150Mb soit assez proche de celui de la drosophile) et comportant peu de séquences répétées (10% du génome). Dans une cellule somatique diploïde à l'interphase, 6 à 10 structures appelées des chromocentres peuvent être identifiées par coloration d'un noyau avec un intercalant de l'ADN comme le DAPI (Figure 1.4, [Fransz et al., 2002]). Ces domaines correspondent aux régions centromériques et péricentromériques des chromosomes ainsi qu'aux régions fortement compactées comme les gènes d'ARN ribosomiques 45S (ARNr 45S). Compte tenu de l'organisation en 5 chromosomes, ce résultat suggère que certains chromosomes homologues sont appariés au niveau des régions centromériques et péricentromériques et ne formeraient qu'un seul chromocentre.

Le nucléole est le siège de la transcription des ADN ribosomiques codant pour les ARNr qui s'accumulent dans le nucléole. Cette accumulation impacte l'organisation nucléaire en excluant l'ADN génomique ce qui se traduit lors de la coloration des noyaux au DAPI par une région dépourvue de coloration (Figure 1.4). Territoires chromosomiques, nucléole et chromocentres ne sont que quelques exemples de domaines nucléaires

mais beaucoup d'autres ont été décrits chez différents eucaryotes [Mao et al., 2011, Costa-Nunes et al., 2014].

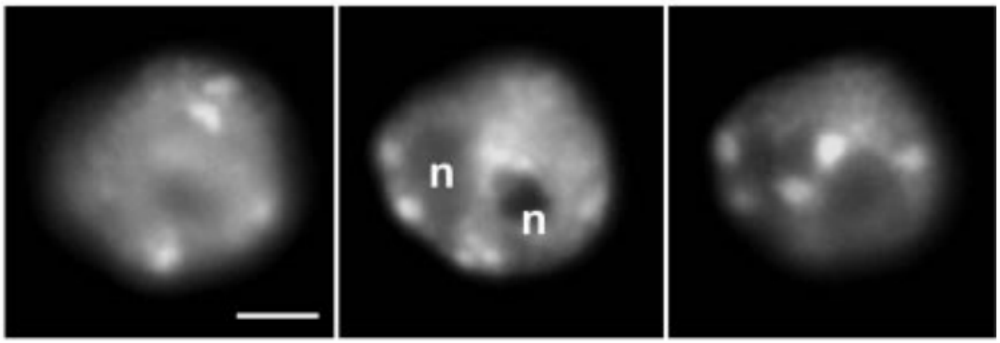


FIGURE 1.4: **Images de noyaux à l'interphase** : Noyaux d'*A.thaliana* colorés par un intercalant de l'ADN (DAPI). Echelle 2 μm. Les zones de forte intensité identifient les chromocentres alors que les zones dépourvues de coloration correspondent aux nucléoles (n). Extrait de [Fransz et al., 2002].

1.1.2 Étude de l'organisation nucléaire chez *Arabidopsis thaliana*

Pour mieux comprendre comment le noyau régule l'organisation de la chromatine et donc l'expression des gènes, plusieurs stratégies et différentes approches méthodologiques peuvent être envisagées [Fraser et al., 2015]. Une première stratégie vise à caractériser les acteurs régulant l'organisation de la chromatine soit au niveau local (au niveau du nucléosome) soit à un niveau d'ordre supérieur (au niveau de l'architecture nucléaire). Une seconde stratégie est de déterminer la contribution de la périphérie du noyau dans cette organisation. En effet, la périphérie où les corps nucléaires peuvent représenter des points d'ancrage de la chromatine et structurer le reste du génome comme nous l'avons évoqué pour les LADs/PLADs. Notre équipe avait également noté la proximité des chromocentres avec la périphérie du noyau, une observation qui est à l'origine des développements en microscopie et analyse d'images réalisées par notre équipe [Poulet et al., 2015, Poulet et al., 2017]. C'est également cette seconde approche dédiée à l'imagerie 3D qui est à l'origine de mon projet de thèse. En effet, lorsqu'il s'agit de s'intéresser au noyau 3D, la microscopie à sections optiques est une approche de choix. L'observation peut se faire sur tissu vivant ou fixé, sur noyaux isolés ou dans un contexte tissulaire [Dumur et al., 2019]. L'analyse des piles d'images permet alors de définir quantitativement les variations par exemple entre un génotype sauvage et un génotype mutant et cette analyse devra être si possible automatisée afin de travailler sur un grand nombre de noyaux. C'est dans ce contexte que mon projet a démarré en 2018.

Contexte du projet de thèse

Notre constat initial est que la périphérie nucléaire contrôle l'expression des gènes en régulant leur position dans l'espace nucléaire. Nous avons donc fait l'hypothèse que des modifications de morphologie du noyau pourraient perturber cette organisation et affecter l'expression des gènes. C'est en suivant cette hypothèse de travail que notre équipe s'intéresse depuis quelques années à la périphérie nucléaire et cherche à identifier et caractériser certaines protéines affectant la morphologie du noyau en utilisant la plante modèle *Arabidopsis thaliana*. Les premières études de l'équipe se sont concentrées sur les protéines SUN [Graumann, 2014, Poulet et al., 2017], puis sur les protéines KAKU4 et CRWN (CROWDED NUCLEI) composant le nucléosquelette ([Dubos et al., 2020], [Mermet et al 2021]). Pour réaliser nos études nous combinons des approches d'imagerie 3D, de génétique, de biologie moléculaire et de bio-informatique pour mieux décrire l'importance de la périphérie du noyau. Les approches bio-informatique et les techniques de séquençages à haut-débit de type Hi-C [**?**] (une technique permettant de déterminer les interactions entre régions de chromatine à l'échelle du génome) par exemple permettent de détecter les changements organisationnels des séquences chromosomiques en évaluant et en quantifiant leur proximité dans la cellule. Le second exemple est le résultat des techniques de RNA-seq qui nous permet d'évaluer le niveau de transcription des gènes dans la plante. Ces approches donnent des résultats normalisant le contenu en ADN des cellules à l'échelle d'un ou plusieurs organismes entiers, masquant parfois les spécificités à l'échelle d'un tissu ou d'un type cellulaire. Enfin l'une des méthodes employées classiquement en génétique est d'étudier la fonction de gènes candidats en caractérisant le phénotype de plantes mutées pour l'un, l'autre ou plusieurs de ces gènes. L'impact de ces mutations sur la morphologie des noyaux et l'organisation de la chromatine, par exemple sur le nombre et la taille des chromocentres, sont autant d'indicateurs qui nous permettront de mieux caractériser nos mutants. Ainsi, les outils d'analyse d'images développés au cours de ma thèse permettent une analyse spatiale de la morphologie nucléaire de façon rapide et automatisée. Cette méthodologie nous permet d'envisager d'analyser la dynamique de l'organisation de la chromatine par exemple au cours de la germination de la graine ou au cours du développement précoce de la plante. L'imagerie devient alors un réel outil de phénotypage au service du généticien.

1.2 Principe de microscopie en trois dimensions

Inventée à la fin du XVI^{e} siècle, la microscopie optique s'est rapidement imposée comme une technique d'observation indispensable aux études en biologie. Elle a notamment été largement appliquée à l'observation du noyau et c'est par ce biais que Boveri a découvert les chromosomes (*Boveri–Sutton chromosome theory*) au début du siècle dernier ou qu'Emil Heitz a suggéré l'existence de l'hétérochromatine. Un microscope optique est composé d'une source de rayonnement, d'optique pour transmettre et agrandir l'objet observé et d'un détecteur pour visualiser l'image. La fluorescence désigne l'émission de photons par une molécule, appelée fluorophore ou fluorochrome, immédiatement après une excitation lumineuse (Figure 1.5). Ce phénomène peut être observé en microscopie photonique, lorsque que la source lumineuse d'excitation est focalisée par l'objectif. La fluorescence observée est celle émise en direction de l'objectif. Nous utilisons des filtres dans le trajet lumineux qui permettent le tri des différentes longueurs d'onde. Ce type de microscopie dite à épifluorescence, présente un inconvénient majeur car elle produit du flou ; ce phénomène s'explique par le fait que la fluorescence est émise par toute l'épaisseur de l'échantillon, et pas seulement par le plan focal. Les

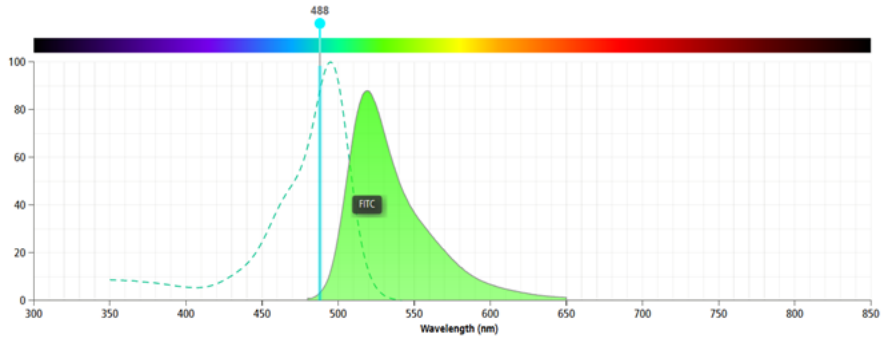


FIGURE 1.5: **Spectres d'excitation et d'émission du FITC** (isothiocyanate de fluorescéine) : La fluorescéine est la première molécule fluorescente qui a été très utilisée en biologie. Le pic d'excitation du FITC est de 495 nm (spectre en pointillés bleus) et son pic d'émission de 521 nm (spectre en vert). Classiquement, un laser de longueur d'onde 488nm est utilisé pour exciter le FITC (ligne verticale bleue ciel). Source : https://www.bdbiosciences.com/en-us/resources/bd-spectrum-viewer

techniques de sectionnement optique permettent de supprimer les contributions lumineuses hors plan focal, autrement dit le flou (Figure 1.6 A). La première d'entre elles, la microscopie confocale date du milieu du $XX^{ème}$ siècle (Marvin Minsky, brevet US3013467A 1957) et repose sur un dispositif dans le trajet lumineux, un sténopé plus communément appelé pinhole ou trou d'épingle, qui ne laisse passer que les photons venant du plan focal, d'où le nom de microscopie confocale ou monofocale (Figure 1.6 B). Une autre technique de sectionnement optique procède par prises de vues successives d'un même champ, en projetant une grille dans

le plan focal qui se déplace entre chaque image du plan. L'image est ensuite reconstituée à partir de l'ensemble des images (3 à 5 images) par soustraction des contributions hors plan focal repérées par les barreaux de la grille. Il s'agit de lumière structurée (Figure 1.7).

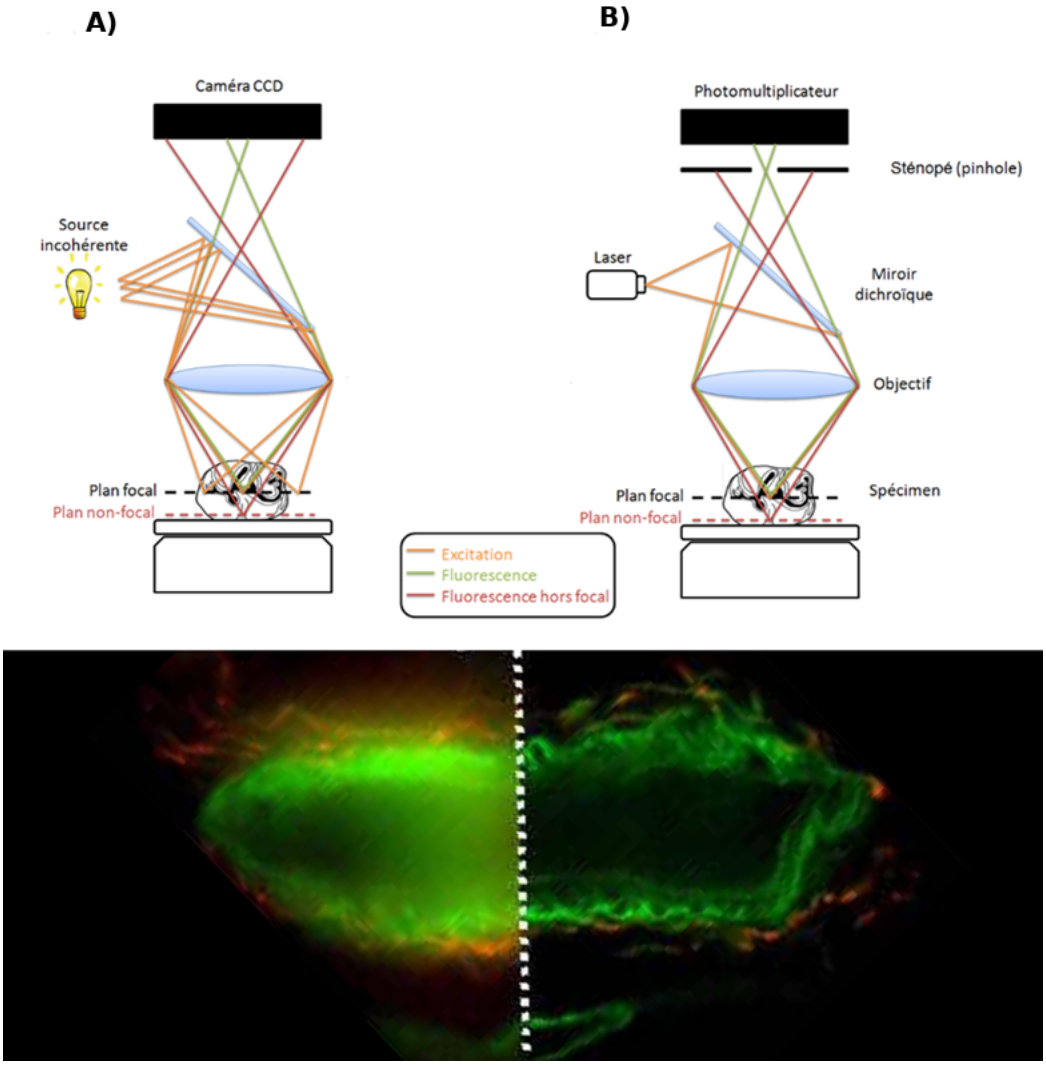


FIGURE 1.6: **Principe du sectionnement optique par un sténopé**. A) la fluorescence est émise par tous les plans illuminés ce qui procure l'impression de flou lors de son observation en microscopie épifluorescence. B) en intercalant un obstacle dans le trajet lumineux percé d'un sténopé ou trou d'épingle (pinhole), seule la lumière émise par le plan focal est détectée : (Sources : haut : thèse Saïma Ben Had / Bas : Damien Schapman, Université de Rouen).

La microscopie à sectionnement optique combine plusieurs domaines technologiques : les molécules fluorescentes, l'optique, et l'informatique. Les avancées continues dans ces trois domaines sont à l'origine de sa popularité dans les laboratoires de biologie, avec des innovations qui en repoussent sans cesse les limites :

- Il existe une très grande diversité de fluorochromes, couvrant tout le spectre des rayonnements électromagnétiques de l'ultra-violet aux infra-rouges, avec des propriétés variables quant à leur stabilité,

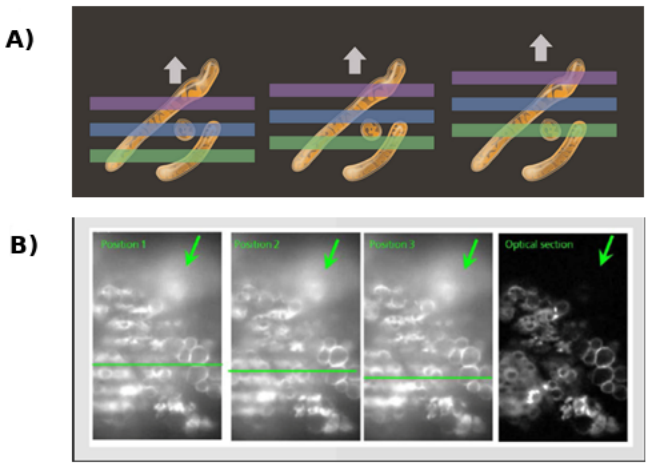


FIGURE 1.7: **Principe de la microscopie à éclairage structuré.** A Schéma du principe : Une grille (lignes de couleur) est interposée dans le trajet lumineux et son ombre projetée sur le plan focal. Trois images d'un même plan sont capturées avec un éclairage structuré, décalé en phase, par déplacement latéral de la grille. Les images sont ensuite retraitées pour éliminer les contributions lumineuses hors plan focal. https://www.microscope.healthcare.nikon.com/fr_AMS/products/super-resolution-microscopes/n-sim-s/the-principle-of-structured-illumination-microscopy. B 3 images acquises avec différentes positions de la grille (à gauche) et image retraitée (à droite). (Zeiss)

brillance, rendement quantique et pouvant être couplés chimiquement pour colorer des molécules d'intérêt ou des compartiments cellulaires.

- Les dispositifs optiques qui s'ajoutent aux microscopes permettent des sectionnements optiques par plusieurs techniques, chacune d'elles permettant de nouvelles applications. Couplées à des dispositifs informatiques, certaines méthodes d'acquisition permettent de repousser la limite de diffraction, se sont les technologies dites de super résolution. De plus, de nouvelles générations de détecteurs, ponctuels ou vidéo, permettent de produire des images de meilleure qualité ou plus riches en informations, et des astuces de montages associées à des calculs informatiques permettent de nouvelles performances de sensibilité ou résolution (Airyscan, Zeiss), de précision pour la quantification (Hybrid, Leica), ou la rapidité de reconnaissance spectrale (module spectral, Zeiss).
- L'informatique est indispensable à l'acquisition d'images pour piloter le microscope ou reconstituer une image, qu'elle soit acquise point par point en confocal ou pour soustraire la lumière hors plan focal d'un grand champ pour la lumière structurée. L'augmentation de la puissance de calcul et des capacités de mémoire ont permis des progrès très significatifs pour les temps d'acquisition, du chargement d'images à l'écran avec une grande flexibilité pour l'affichage, du traitement d'images à la volée, et les capacités de stockage facilitent la gestion de fichiers images de plusieurs de GigaOctets (Go).

L'usage de la microscopie à sectionnement optique s'est particulièrement répandu dans les sciences biologiques. En effet, toutes les techniques d'histologie sont réalisables avec des fluorophores et peuvent donc bénéficier des apports de la microscopie confocale, en particulier de la possibilité de reconstruire les objets en 3 dimensions, l'acquisition dynamique d'échantillons vivants, la détection simultanée de plusieurs molécules grâce à la diversité des fluorophores.

Parallèlement au développement de la microscopie confocale se multiplient les solutions d'analyse d'images. Ce terme regroupe à la fois les traitements permettant d'améliorer le rendu visuel et ceux qui extraient une information quantitative. Ce dernier domaine est particulièrement exigeant en qualité d'images et impose de connaître ses contraintes au moment de l'acquisition. C'est pourquoi nous décrirons les images numériques produites par les microscopes et les paramètres d'acquisitions essentiels pour obtenir des images de la qualité attendue avant d'aborder les grands principes de traitement d'images.

1.2.1 De l'acquisition à l'image numérique

L'une des étapes clefs lors de l'acquisition est la conversion de l'image en données numériques. Le terme d'image numérique désigne, dans son sens le plus général, toute image qui a été acquise, traitée et sauvegardée sous une forme codée par des valeurs numériques. La numérisation est le processus qui permet de passer de l'état d'image physique d'un objet caractérisé par l'aspect continu du signal à l'état d'image numérique caractérisé par l'aspect discret : il s'agit d'une convolution de l'objet en image au travers d'un processus de discrétisation. Du fait de la discrétisation, l'intensité lumineuse d'une image numérique ne peut prendre que des valeurs quantifiées en un nombre fini de points distincts ce qui permettra une exploitation ultérieure par des outils logiciels sur ordinateur. Une image est un ensemble de pixels (contraction du terme anglo-saxon « picture elements ») caractérisés par leur intensité lumineuse variable. Le voxel (contraction de « volumetric pixel ») est un pixel en 3D auquel est rajouté une dimension z, décrivant la distance entre 2 plans. Le voxel s'avère particulièrement adapté à la reconstitution de volumes à partir d'une série ou « pile » d'images organisées en différentes sections/coupes. Le voxel est très utilisé en Imagerie Médicale 3D (Scanner, IRM...) et en biologie (microscopie confocale, Super Resolution Microscopie...) et le plus souvent de type noir et blanc ou monochrome. L'image la plus simple a 1 bit qui pour chaque pixel ne comporte que 2^1 possibilités : 0 pour le noir et 1 pour le blanc. Il s'agit d'une image binaire. Dans le cas d'une image en noir et blanc décrite en 8 bits, qui est un format très courant en analyse d'images, il y aura 2^8 soit 256 possibilités de niveaux de gris : de 0 pour le noir à 255 niveaux de gris jusqu'au blanc. L'histogramme d'une image est la

représentation de la fonction discrète qui à chaque niveau de gris de l'image associe le nombre de pixels ayant ce niveau de gris. Ainsi, l'histogramme d'une image en 256 niveaux de gris sera représenté par 256 valeurs en abscisses, et le nombre de pixels de l'image pour chaque niveau de gris en ordonnées (Figure 1.8). En résumé, une image numérique résulte d'une double discrétisation, un échantillonnage dans l'espace (nombre et taille de pixels (2D) ou voxels (3D) dans l'image numérique) et d'une quantification de l'intensité (valeur des pixels). Le processus de discrétisation permet ainsi d'extraire une description quantitative des structures ou de processus biologiques. Il permet ainsi de décrire : le nombre d'objets, leurs formes, leurs dimensions ou encore leurs dispositions spatiales grâce à des paramètres géométriques.

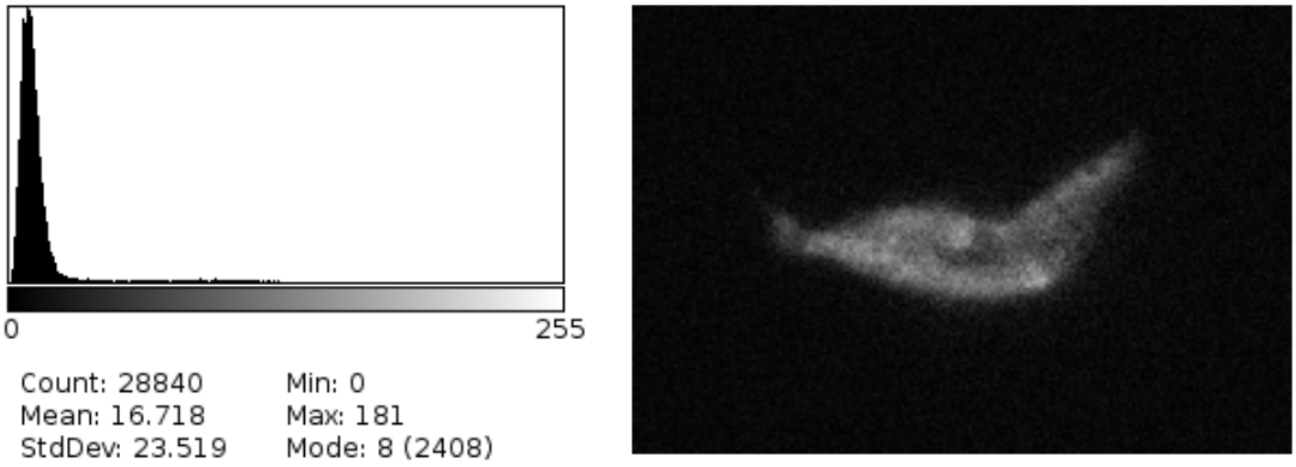


FIGURE 1.8: **Description de l'intensité des pixels d'une image**. Représentation en histogramme des valeurs de pixels issus de imageJ provenant de l'image 2D en 8 bits de droite. Des valeurs statistiques décrivent le nombre de pixels de l'image : count = 28 840 pixels pour une image qui dans ce cas comporte au total 206 x 140 pixels. L'histogramme est également caractérisé par les valeurs d'intensités maximum (Max), minimum (Min), moyen (Mean) et un écart type (StdDev).

1.2.2 Les grandes caractéristiques descriptives d'une image

La numérisation d'une image peut être considérée comme un tableau informatique, dans lequel chaque case décrit une intensité au travers d'une valeur numérique. Les spécificités de chaque image peuvent être décrites par les paramètres suivants :

- La définition de l'image : il s'agit de la taille des images exprimée en pixels.
- La taille des pixels : elle indique la distance entre deux pixels successifs dans le plan XY. Elle est parfois appelée résolution de l'image mais ce terme peut entraîner des confusions en microscopie avec la résolution optique théorique des objectifs.
- Le pas (ou résolution en z) : il indique la distance entre deux plans z successifs pour les piles d'images.

- L'histogramme : il représente la distribution des intensités des pixels/voxels dans l'image.
- La dynamique : elle rend compte de la gamme d'intensité présente dans l'image. Les valeurs d'intensité ont des correspondances en niveaux de gris dans l'image.

Un fichier image peut contenir plusieurs tableaux de voxels comme une pile de plans successifs, mais aussi d'autres dimensions :

- Le canal : plusieurs fluorochromes peuvent être détectés sur un même échantillon, donnant lieu à plusieurs piles d'images distinctes possédant les mêmes coordonnées de voxels. Un canal désigne une fenêtre de longueurs d'onde sélectionnée avant le détecteur.
- Le temps : Le même échantillon peut être imagé plusieurs fois pour suivre un phénomène dynamique. Ces paramètres sont toujours un choix de l'expérimentateur au moment de l'acquisition de l'image. Ce sont eux qui déterminent la qualité d'une image en vue de son analyse.

1.2.3 La qualité des acquisitions : un compromis

Il faut trouver pour chaque expérience le meilleur compromis pour ses échantillons en fonction de la question posée et de l'usage de l'image. Il existe néanmoins des règles à respecter pour produire des images pertinentes et permettre des analyses. Dans l'article [Dumur et al., 2019], les auteurs proposent d'anticiper par un plan d'expérience (*experiment design*) toutes les étapes pouvant impacter les images. Ces recommandations faites pour étudier le noyau chez les plantes peuvent être généralisées pour toutes les expériences de microsocopie et sont particulièrement cruciales dans les techniques de super résolution. En effet, si il est possible de s'accorder facilement sur la qualité visuelle d'une image, lorsque l'objectif est d'en extraire une information quantitative, cette image doit répondre à des règles plus exigeantes. Cependant, au moment de la production d'images, l'expérimentateur rencontrera des contraintes qui le conduiront à s'écarter plus ou moins de l'image quantitative idéale :

- La définition de l'image et la taille des pixels : la durée d'acquisition sera directement impactée par le choix de l'expérimentateur. Des temps trop longs d'acquisition peuvent être incompatibles avec le temps dont dispose l'expérimentateur sur le microscope, sur la durée de vie du fluorophore ou de l'échantillon.
- La dynamique : il est recherché en général dans une image une dynamique possédant les plus faibles valeurs d'intensité et les plus grandes, sans saturation. Des signaux faibles peuvent pousser l'expérimentateur à choisir des fortes puissances de la lumière d'excitation ou des temps longs d'exposition

(lumière structurée ou *spinning disc*) ou des vitesses lentes du balayage par le laser. Il y aura là aussi un impact direct sur la vitesse d'acquisition et le blanchiment possible du fluorochrome.

- Le nombre de dimensions : le nombre de canaux à imager va impacter également la vitesse d'acquisition puisqu'il est rarement possible d'imager simultanément plusieurs canaux. D'autre part, la question de superposition des spectres d'émission et d'excitation, de différence d'intensité entre eux, crée une difficulté supplémentaire dans le réglage des paramètres ou des vérifications préalables sur des échantillons témoins. Enfin, les vitesses d'acquisition doivent être cohérentes avec la dynamique du phénomène étudié.
- La nature de l'échantillon biologique étudié est aussi à prendre en compte : un échantillon vivant ne supportera pas toujours des temps longs d'acquisition par exemple.

Toutes ces règles ne sont pas toujours bien maîtrisées par les expérimentateurs mais pourtant nécessaires pour la production d'informations quantitatives à partir d'images. C'est pour ces raisons que nous voyons apparaître de récentes publications portant sur les bonnes pratiques en microscopie, tant pour l'acquisition que l'analyse qui l'accompagne [Dumur et al., 2019] [Jonkman et al., 2020], [Miura and Nørrelykke, 2021]. Il est donc important pour la reproductibilité des résultats en biologie de faire suivre avec l'image tous les paramètres (métadonnées) qui ont permis sa production, depuis la préparation de l'échantillon jusqu'aux paramètres d'acquisition. Du coté informatique, nous avons la plateforme OMERO (base de données destinée au stockage et à l'annotation d'images) qui propose un service de gestion des métadonnées des images.

1.3 Méthodes de segmentation en 3 dimensions

1.3.1 Méthodes de segmentation

La segmentation est une opération de traitement de l'image visant à classifier les pixels de l'image en différents groupes. Dans le cas où nous voulons séparer deux groupes de pixels, dans notre cas différencier le noyau d'une cellule du fond de l'image, c'est une étape de binarisation. L'image après binarisation sera de même dimension que l'image de départ et les pixels composant l'objet seront représentés en blanc (valeur de pixels = 1) et les autres en noir (valeur de pixels = 0). Ce format simplifie l'image et a pour principal avantage de réduire la taille de stockage de l'image.

Afin d'isoler les régions d'intérêt ou ROI en anglais pour *Region Of Interest*, différentes méthodes peuvent être appliquées à l'image. Dans le cas du métier de bio-informaticien nous sélectionnons ces algorithmes basés sur des principes mathématiques et statistiques en fonction des objets que nous souhaitons détecter. Nous exposerons dans cette section une liste non exhaustive que nous avons utilisée dans les développements logiciels décrits dans les chapitres résultats. Nous présenterons les grandes familles d'approches et les méthodes couramment utilisées en traitement d'image pour isoler les objets d'intérêt et détaillerons les principes généraux.

1.3.2 Segmentation manuelle

Le cerveau humain est capable de manière instantanée de regrouper les pixels sur une image pour différencier l'ensemble des objets d'intérêt. Dans le cas d'une simple photographie, le contraste des différentes couleurs affichées permet aisément de segmenter les objets manuellement. Toutefois dans le cadre de l'analyse d'images en biologie, une expertise est nécessaire pour détecter les ROI. En effet comme les images issues de microscopie sont pour la plupart en niveaux de gris, il faut bien connaître les différentes composantes du tissu ou de l'organisme marqués afin de les distinguer correctement. Cependant les capacités visuelles de l'espèce humaine a ses limites, et la distinction du contraste entre deux niveaux de gris varie entre individus. La segmentation manuelle est par conséquent subjective et peut varier d'un opérateur à l'autre. La délimitation des objets de manière complètement manuelle engendre donc des difficultés dans la reproductibilité des résultats. Pour limiter le biais, de nombreux outils dédiés au traitement d'images comme l'adaptation de weka en plugin imageJ [Arganda-Carreras et al., 2017], ilastik [Berg et al., 2019] ou encore AnnotatorJ

[Hollandi et al., 2020], qui intègrent des méthodes statistiques ou de machine learning pour aider l'utilisateur, sont apparus ces dernières années. Dans le cadre de mon travail de thèse, le format images 3D (piles d'images) ajoute une difficulté supplémentaire dans le processus de segmentation manuelle. En effet, cette troisième dimension ne facilite pas la détermination de la limite des objets en dimension Z. La possibilité de changer de repère sur les outils de segmentation est une solution adaptée mais elle augmente considérablement le temps d'analyse par image.

1.3.3 Approche de segmentation automatisée par méthode mathématique

Cette famille de méthodes de segmentation se base sur la distribution des valeurs des pixels/voxels composant les images [Gonzalez, R.C., and Woods, 1992]. Les histogrammes sont très souvent utilisés pour représenter la distribution des valeurs de pixels/voxels. Ces méthodes s'appuient majoritairement sur des principes statistiques pour constituer des groupes de pixel/voxels. Voici quelques grandes familles d'algorithmes utilisés pour la segmentation notamment au cours de mon travail de thèse.

- **Segmentation basée sur les régions** : cette technique consiste dans un premier temps à choisir des pixels/voxels de manière aléatoire dans une image. Ensuite de manière itérative, les pixels/voxels voisins sont examinés et sont regroupés lorsqu'ils partagent des caractéristiques communes. Les groupements de pixels/voxels ainsi formés sont appelés des composantes connexes (connected components [Chang et al., 2004]). Lorsque les groupements de pixels ne grossissent plus, de nouveaux pixels/voxels sont tirés au sort parmi ceux qui n'ont pas été regroupés avant d'examiner leurs similarités avec leurs voisins. Ces étapes sont réitérées jusqu'à ce que l'ensemble des pixels de l'image appartiennent à une composante connexe.

- **Segmentation basée sur les gradients** [Gonzalez, R.C., and Woods, 1992] : cette approche vise à calculer la variation de chaque valeur de pixels/voxels dans une image au travers d'une fonction mathématique. Cette méthode mathématique est présente dans de nombreux algorithmes de segmentation. Par exemple si elle est couplée à une fonction seuillage elle peut permettre la segmentation des contours des objets dans l'image. La forme la plus simple de la réalisation du gradient est la soustraction des valeurs des pixels à leurs voisin. Cette opération aura alors pour conséquence de faire ressortir les chutes de valeur d'intensité des pixels.

- **Isodata algorithme** : un des premiers algorithmes de cette famille est proposé par [Ridler and Calvard, 1978]. Il consiste à choisir de manière “aléatoire” (souvent en utilisant la moyenne ou la médiane des valeurs de l'image) une première valeur de seuil afin de la faire varier de manière itérative jusqu'à minimiser la variabilité entre les différentes classes. Le seuillage de Otsu [Nobuyuki Otsu, 1979] consiste quant à lui à sélectionner le seuil pour lequel la variance des valeurs est minimale au sein de chaque classe. (Figure 1.9).

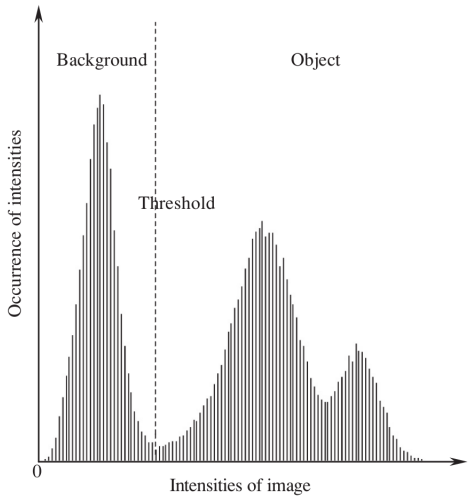


FIGURE 1.9: **Seuillage par la méthode d'Otsu** : Représentation d'un histogramme d'une image 8 bits (256 niveaux d'intensité en abscisse) lors d'un seuillage de Otsu. La ligne pointillée correspond au seuil (*threshold*) permettant de diviser les pixels/voxels en 2 classes : une contenant le fond de l'image (*background*) et l'autre les objets (*object*) d'intérêt. Cette classification en deux groupes permettra ensuite une binarisation : 0 (background) et 1 (object).

- **Les lignes de partage des eaux** : cette famille d'algorithmes de segmentation se base sur des concepts provenant du domaine de l'étude de la topographie [Beucher and Lantuejoul, 1979]. Elle consiste à considérer les valeurs de gris d'une image comme des reliefs, où les intensités de pixels/voxels les plus faibles seront associés aux altitudes basses et les intensités fortes aux sommets. Une fois les pixels/voxels regroupés par valeurs extrêmes, une phase « d'inondation » est effectuée afin de tester les pixels/voxels de proche en proche dans l'image. Ils sont ajoutés à la composante connexe s'ils respectent une condition comme par exemple la croissance des valeurs d'intensités. Une fois les maxima atteints, la décroissance des valeurs débute (comparable à la descente d'une vallée), une nouvelle composante connexe est créée pour regrouper les pixels. Une ligne séparatrice des deux composantes connexes d'une unité de large est mémorisée. Cette dernière correspondant à la ligne de partage des eaux (Figure 1.10).

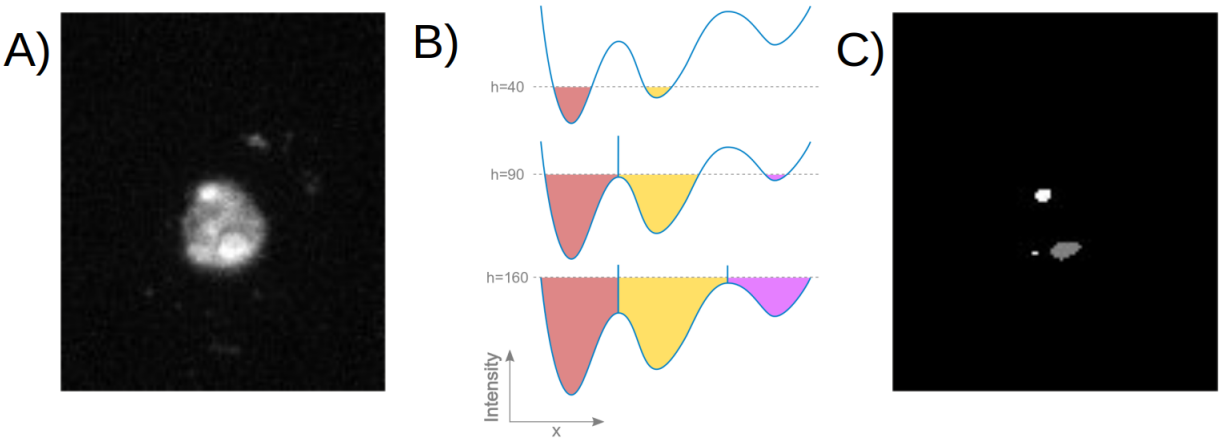


FIGURE 1.10: **Segmentation par la ligne de partage des eaux** : A) A partir d'une image de noyau coloré au DAPI et présentant des régions d'intensités variables comme par exemple les régions correspondant aux chromocentres (spots blanc), B) l'algorithme va définir les valeurs d'intensité des voxels de l'image. Le schéma montre ici la définition des valeurs basses de voxels qui définissent des minima régionaux. Les minima régionaux vont servir de point d'initiation (« source ») pour la « montée des eaux » qui progresse du schéma en haut vers celui du bas. Une frontière sera créée lorsque les eaux provenant de différents minima (de différentes couleurs) sont sur le point de fusionner. Enfin, « l'inondation finale » définira ici 3 régions distinctes en fonctions des seuils d'intensités définis. C) Résultat de la segmentation par la ligne de partage des eaux de l'image en A).

Toutes ces méthodes ont été éprouvées et largement utilisées dans le domaine de l'analyse d'image. Toutefois elles présentent le désavantage d'être relativement dépendantes des images à traiter et très souvent l'utilisateur devra tester différentes méthodes avant de retenir la plus adaptée.

1.3.4 Segmentation et apprentissage automatique

Cette famille de méthodes plus connue sous le nom de "machine learning" est en plein essor ces 10 dernières années, avec un effort de la recherche sur la thématique d'une sous catégorie nommée le "deep learning" ou apprentissage profond [von Chamier et al., 2019]. Ceci s'explique majoritairement par les progrès technologiques de l'informatique en termes de vitesse de calculs et notamment le développement des processeurs graphiques. Ces évolutions nous amènent dans l'ère du « big data » où il est maintenant possible de générer des jeux de données de grande taille. Le grand principe de ces méthodes d'apprentissage consiste à générer un modèle. Dans le contexte de jeux de données de tailles importantes il est impossible d'expertiser entièrement à la main, seul un sous-échantillon des données sera utilisé comme jeux de données d'entraînement. Une fois le modèle généré il sera ensuite appliqué à l'ensemble des données.

Dans la méthodologie de l'apprentissage profond, nous pouvons subdiviser l'approche en deux catégories :

l'apprentissage supervisé et non supervisé [Huan Liu, 2004]. La différence principale réside dans l'annotation des données pour l'apprentissage. En effet dans le cas d'un apprentissage supervisé, les données sont annotées par un expert (étape de labellisation) contrairement à l'apprentissage non supervisé où seul le nombre de classe d'objet souhaité en sortie est indiqué.

Pour une donnée comme une image, il est en effet possible de répertorier des informations de manière quantitative ou qualitative comme illustré avec les deux exemples suivants :

- Un premier cas de figure ***est un apprentissage supervisé basé sur une approche quantitative***. Dans ce cas, considérons la segmentation des noyaux, le jeu de données d'apprentissage fourni en entrée au modèle sera constitué par des labellisations quantitatives. En termes simples, il s'agit de définir les pixels/voxels comme appartenant ou non au noyau. L'utilisateur devra donc délimiter chaque noyau manuellement afin de classer les pixels/voxels en deux catégories. Cette tâche peut s'avérer difficile notamment pour les images en niveau de gris par exemple, lorsqu'il s'agit de définir le bord d'un noyau (frontière noyau/fond de l'image) qui est parfois diffuse. La qualité des données et la précision de l'expert seront très influent sur le modèle généré et impacteront le résultat final de segmentation des noyaux. Dans le cas des données en 3D, la labellisation est d'autant plus complexe et longue ce qui constitue un problème majeur.
- Un second exemple est ***l'apprentissage supervisé basé sur une approche qualitative***. Dans ce cas, le modèle sera entraîné à compter le nombre de noyaux au sein de notre image, mais la précision de la délimitation n'aura pas d'importance. La difficulté dans ce type de modèle vient lorsque certains des objets partagent des caractéristiques proches en termes de signaux : par exemple la cellule végétale contient d'autres organelles que le noyau, comme les chloroplastes qui sont connus pour émettre de la fluorescence qui pourra être détectée au même titre que celle du noyau introduisant la comptabilisation d'objets supplémentaires. Nous noterons enfin que dans ce cas de figure l'avantage majeur est la simplicité/rapidité de l'annotation du jeux de données d'entraînement.

Dans le cas de ***l'apprentissage non supervisé*** il n'y a pas de labellisation des données à réaliser en amont de l'entraînement du modèle. L'avantage majeur de cette approche est que le modèle est généré sans *a priori*. Il limite donc les éventuels biais introduits par les experts lors de la labellisation des données. Avec cette stratégie, il est donc possible de faire ressortir des informations jamais observées à cause du trop grand volume de données à traiter. A contrario, cette approche est assez imprévisible et peut faire ressortir des

résultats non informatifs en créant par exemple des sous-catégories inattendues [Alloghani, 2020].

Les approches supervisées et non supervisées seront donc choisies en fonction de la problématique abordée et de la complexité des données associées. De manière générale, les deux approches peuvent être utilisées pour un problème donné, et l'expérimentateur orientera le choix vers l'une ou l'autre des stratégies en fonction des avantages et inconvénients cités ci-dessus.

Enfin, il existe une alternative à ces deux premières méthodes qui est ***l'apprentissage semi-supervisé***. C'est donc la génération d'un modèle à partir d'un mélange de données annotées et non annotées. Le modèle en cours d'entraînement demandera à l'utilisateur de labelliser certaines images afin d'orienter l'entraînement. Cette méthode est intéressante lorsque les jeux de données sont très conséquents et que le temps humain requis pour la labellisation des informations est très important.

1.3.5 Réseau de neurones artificiels

Dans les années 1950, deux neurologues Warren McCulloch et Walter Pitts s'inspirent des observations faites sur le fonctionnement des neurones biologiques dans l'objectif de proposer une modélisation similaire à celle de l'intelligence humaine. Leurs travaux présentent pour la première fois le neurone artificiel qui correspond à une fonction mathématique qui reçoit une ou plusieurs informations en entrée et la transforme en un résultat en sortie [W. S. McCULLOCH and PITTS, 2007]. Les opérations mathématiques du neurone peuvent être simples comme agréger les valeurs en entrée ou encore classer les données entrantes en fonction d'une valeur seuil. Le choix de la fonction est dans un premier temps manuel afin d'adapter le modèle aux données et leurs éventuelles évolutions. C'est pourquoi Franck Rosenblatt en 1957 proposera le perceptron (Figure 1.11 A)) qui est une méthode d'apprentissage supervisé qui classe de manière binaire en utilisant un système de pondération des données entrantes. Cependant, le perceptron ne peut traiter des problèmes de classification non linéaires, c'est pourquoi est apparu dans la fin des années 90 le perceptron multicouche correspondant aux premiers réseaux de neurones artificiels (Figure 1.11 B)). Le principe est de positionner les neurones sur différentes couches afin qu'ils puissent interagir avec les entrées et sorties des neurones des couches voisines, mécanismes schématisant le comportement des dendrites des neurones biologiques.

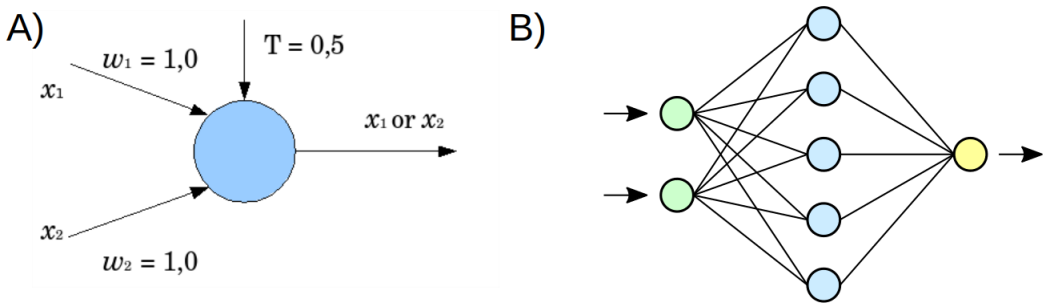


FIGURE 1.11: **Représentation schématique d'un neurone formel et d'un réseau de neurones** (source : https://fr.wikipedia.org/wiki/). A) Neurone formel possèdant plusieurs entrées, avec x_1 et x_2 qui représentent les valeurs numériques en entrées et w_1 et w_2 leurs poids respectifs associés. L'exemple donné ici est une fonction mathématique qui permet de filtrer les valeurs avec l'utilisation d'un seuil noté T de valeur 0,5 donnant en sortie la valeur x_1 ou x_2. B) Réseau de neurones avec 3 couches : les neurones en vert représentent la couche de neurones traitant l'entrée des valeurs, la couche bleue de neurones formels dont la fonction mathématique est appliquée aux données, et enfin la couche jaune représentant la donnée en sortie.

L'apprentissage profond ou Deep Learning, introduit par Kunihiko Fukushima [Fukushima, 1979], est la branche de l'apprentissage automatique la plus récente qui se base sur les réseaux de neurones. Les performances de ces modèles reposent totalement sur la quantité mais aussi la qualité des données annotées. Les excellentes performances particulièrement pour les applications en vision/reconnaissance d'objets des réseaux d'apprentissages profonds proviennent des progrès en informatique dans deux domaines :

- Tout d'abord dans la possibilité de stocker des bases de données annotées de grand volume comme par exemple la base de données ImageNet [Russakovsky et al., 2015] (https://image-net.org/challenges/LSVRC/2010/) qui contient désormais des millions d'images avec des objets annotés manuellement.
- D'autre part avec l'amélioration des performances des processeurs graphiques qui permettent une accélération du calcul.

Le challenge de reconnaissance et de labélisation d'objet par ordinateur est complexe par la diversité d'objets possibles mais aussi du fait des différentes formes et couleurs qu'un même objet peut prendre (angle de vue, distance de capture de l'objet, exposition lumineuse). Les réseaux à neurones convolutifs ou en anglais convolutional neural network (CNN) se sont imposés dans le domaine en proposant les meilleures solutions au challenge ImageNet 2012 ou encore le concours ICPR 2012 pour la détection de cellules cancéreuses. Dans ces concours, les tâches d'apprentissage et les réseaux qui en découlent consistent surtout à labelliser des objets au sein d'images en 2D plus qu'en la détection précise de ROI dans une image. C'est pourquoi en 2015 le réseau U-Net [Ronneberger et al., 2015] propose une nouvelle architecture (Figure 1.12) basée sur les CNN

pour la segmentation d'objets de nature biomédicales. Cette nouvelle méthodologie a pour avantage majeur de proposer des résultats précis avec des jeux de données de taille réduite. Pour cela, les concepteurs ont couplé aux étapes de *pooling* classique des étapes de sur-échantillonnage augmentant la résolution de l'image en sortie.

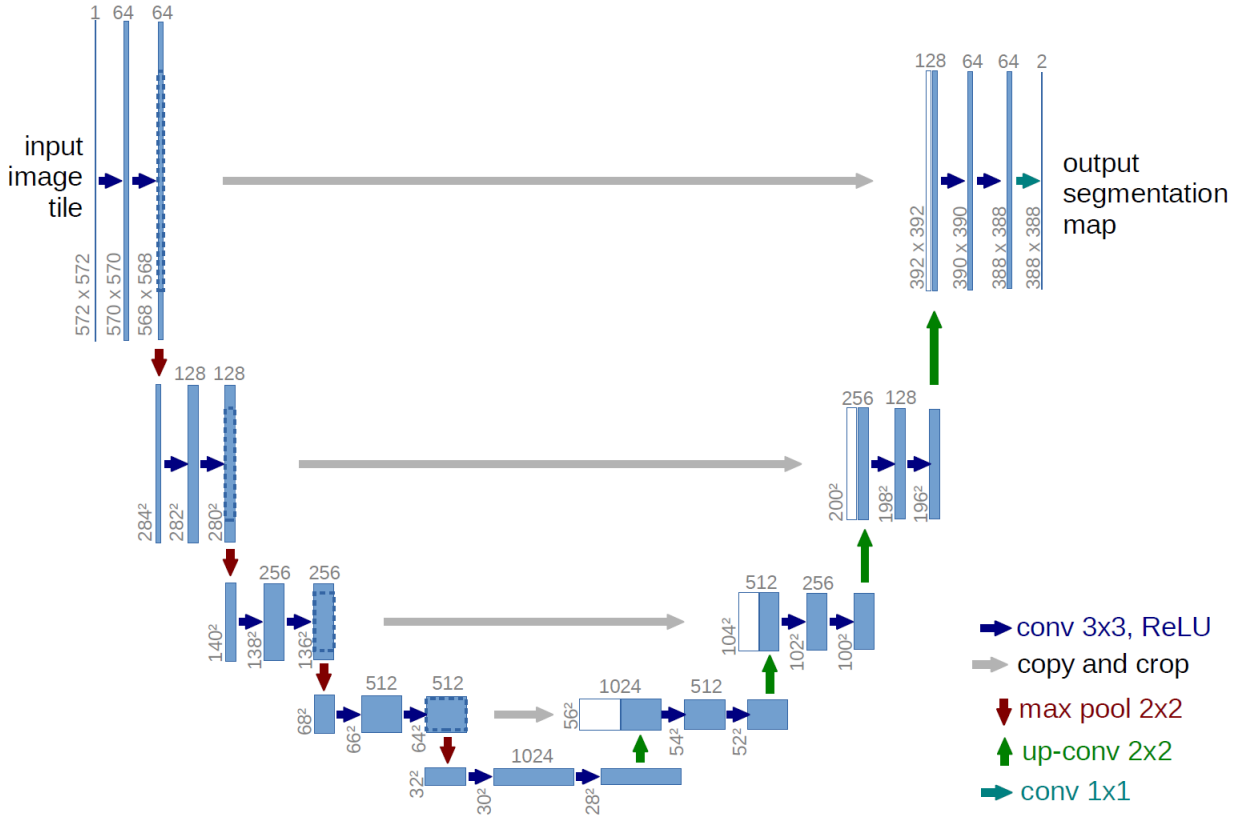
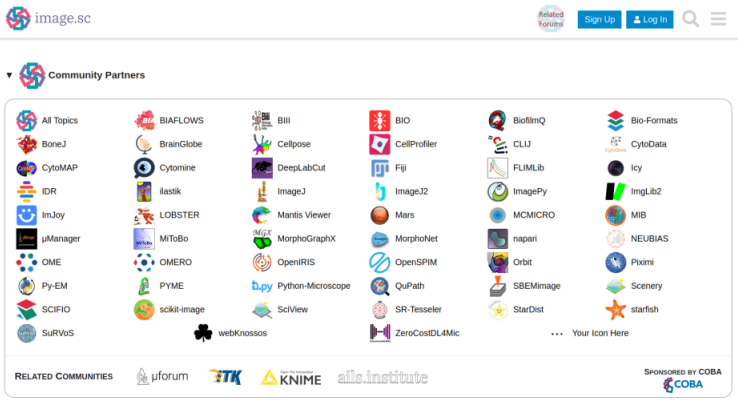


FIGURE 1.12: **Représentation schématique de l'architecture du réseau de neurones artificiels U-Net** : U-Net est composé d'une étape de convolution puis de déconvolution pour donner en résultat l'image segmentée (avec ici deux labels en sortie). Le nom de U-Net vient de la forme en U du modèle. Source [Ronneberger et al., 2015]

1.4 Les solutions logicielles dédiées à la caractérisation morphologique des noyaux

Face à la génération d'images de résolution et de tailles croissantes, les solutions logicielles dédiées à l'analyse d'images se multiplient ces dernières années avec l'apparition des plus connues comme CellProfiler [Kamentsky et al., 2011], Knime4Bio [Lindenbaum et al., 2011], Icy [De Chaumont et al., 2012] ou Imaris https://imaris.oxinst.com/. Bien d'autres outils existent comme le montre la Figure 1.13 extraite du site image.sc qui est dédié aux nombreuses questions relatives à l'analyse d'images et l'utilisation des outils.

A)



B)

Site Statistics

	Last 24 hours	Last 7 days	Last 30 days	All Time
Topics	30	119	516	24.1k
Posts	242	921	4.1k	172k
Users	28	97	384	17.0k
Active Users	375	825	1.7k	—
Likes	85	409	2.0k	63.8k

FIGURE 1.13: **Description du forum image.sc** : A) Ensemble des logiciels partenaires de la plateforme. B) Statistiques de l'activité du forum. Capture d'image extraite du site https://forum.image.sc/

L'ensemble de ces plateformes se basent sur des langages informatiques différents mais ont une philosophie commune qui est l'interopérabilité des systèmes intégrant pour les plus récentes des méthodes d'apprentissage profond. Elles ont pour point commun la possibilité d'utiliser un logiciel pionnier en analyse d'images : ImageJ [Collins, 2007] (ou sa version plus récente intégrant des plugins majeurs : Fiji [Schindelin et al., 2009]). La plateforme ImageJ est un projet open source développé en langage Java qui a pour point fort d'être un outil facilement portable sur les différents systèmes d'exploitation. Les contributions de la communauté de

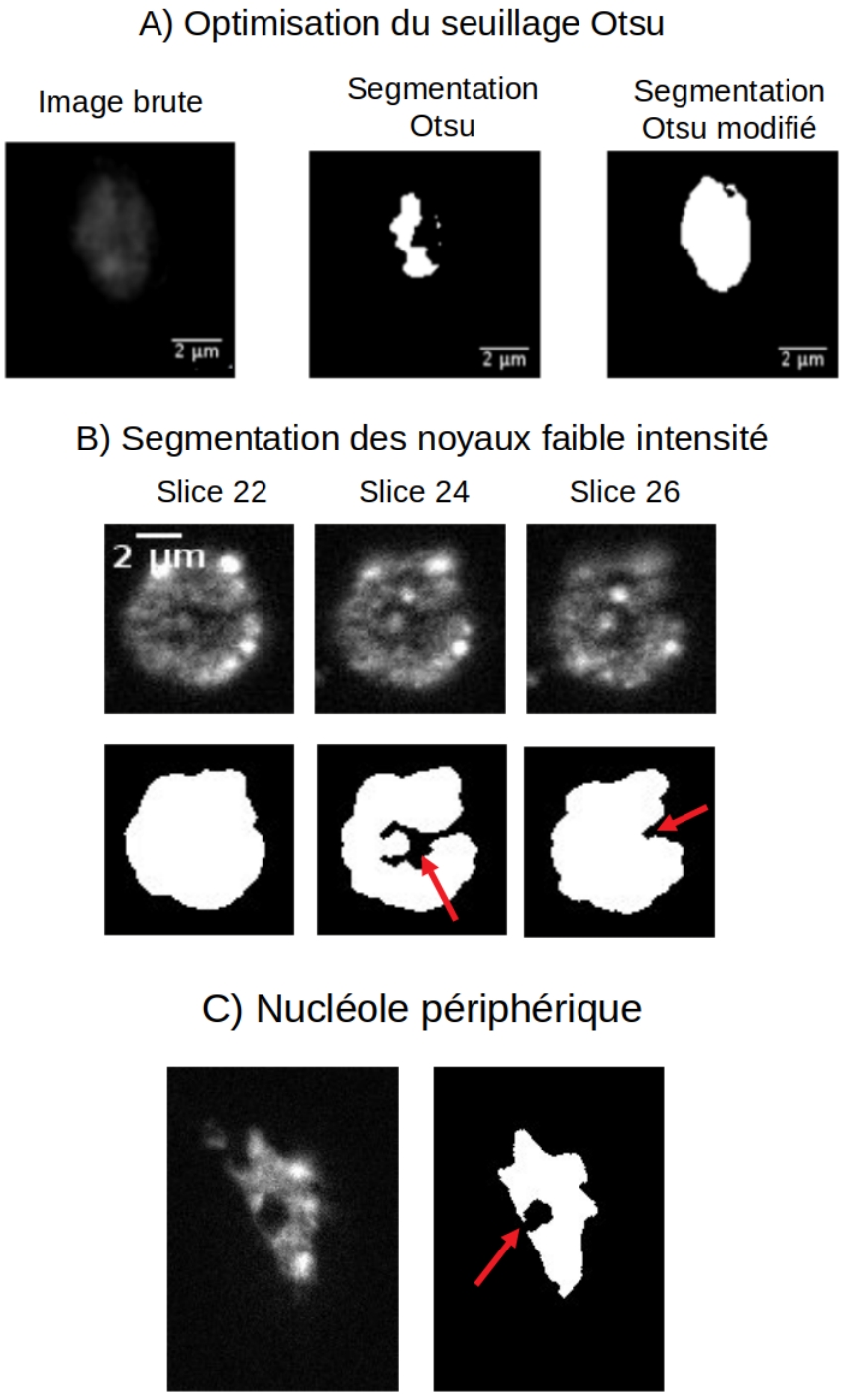


FIGURE 1.14: **Exemples de segmentations de noyaux d'*A. thaliana*** : résultats de la segmentation d'un noyau de faible intensité par un seuillage Otsu classique et par la méthode proposée par NucleusJ : Otsu modifié qui sélectionne le seuil Otsu pour lequel la sphéricité du noyau segmenté est maximale A). Exemples de segmentation contenant des artefacts suite à l'utilisation de la méthode Otsu modifié. Dans le cas de noyaux de faibles intensités en B) et C) de l'exclusion du nucléole dans le masque du noyau. Les flèches rouge représentent des cas de l'absence de segmentations du nucléoles.

développeurs se font sous forme de plugins, et aujourd'hui l'ensemble des fonctions mathématiques dédiées au traitement d'images que nous avons citées précédemment y sont disponibles. Ces deux derniers points font d'ImageJ une des plateformes de traitement d'images en biologie la plus complète. Malgré la diversité des possibilités des méthodes mathématiques et de plugins qu'elle propose, cette plateforme reste difficilement accessible aux biologistes pour la mise en place de chaînes de traitements automatiques. Les principales raisons sont les suivantes :

- La majorité des plugins proposent des méthodes mathématiques dédiées à l'analyse d'images plus que des solutions dédiées à un problème d'analyse spécifique. Il est donc souvent nécessaire de connaître et de sélectionner les méthodes puis de les chaîner afin de réaliser la bonne séquence de traitement de l'image pour en extraire l'information souhaitée.
- Bien que ImageJ embarque un langage propre de macro qui facilite la création de chaîne de traitement, beaucoup de fonctionnalités ne sont pas disponibles via le langage macro. Il est dans ce cas obligatoire d'avoir des compétences en programmation pour accéder au code source en Java (lorsqu'il est disponible).
- De nombreuses méthodes mathématiques ont été développées à une période où les microscopes produisaient majoritairement des images en 2D. Or, toutes ces méthodes n'ont pas été actualisées pour prendre en charge des images en 3D.

C'est en 2013 que l'équipe s'est intéressée à ces problématiques d'analyses d'images, suite à l'achat d'un microscope à section optique équipé d'un module OptiGrid dans le but d'étudier la morphologie nucléaire chez *A. thaliana*. Axel Poulet, doctorant en bio-informatique, a développé un plugin sur la plateforme ImageJ appelé NucleusJ [Poulet et al., 2015]. Ce plugin, initialement conçu pour des images de noyaux 3D individuels (qui étaient obtenues par microscopie confocale avant l'achat du microscope), prend en charge des piles d'images contenant un noyau par fichier et permet la segmentation du noyau et de la chromatine (Figure 1.16 D). Les paramètres en sortie d'analyse permettent de discriminer les populations cellulaires sur des aspects de morphologie nucléaire et d'organisation de la chromatine avec notamment la caractérisation des chromocentres (Figure 1.15 2).

Les raisons d'un développement d'une chaîne de traitement et non de la réutilisation de méthodes déjà existantes, s'expliquent par plusieurs spécificités liées à notre modèle biologique *A. thaliana* (illustrées Figure 1.15) :

- **Le nucléole** : c'est un domaine nucléaire représentant jusqu'à 30% du volume du noyau mais ne contenant que très peu d'ADN. Le nucléole n'est donc pas marqué par les intercalants fluorescents comme le DAPI qui cible spécifiquement la chromatine et peut donc fausser l'estimation du volume nucléaire selon la méthode utilisée.
- **Les chromocentres** qui correspondent aux regroupements des régions hétérochromatiques et qui au contraire du nucléole sont fortement marqués par le DAPI du fait de leur forte densité en ADN.
- **La grande dynamique d'intensité de marquage** : la morphologie nucléaire dans notre modèle biologique est décrite par une large gamme de forme, de taille et d'intensité de coloration des noyaux au sein d'une image grand champ (Figure 1.16 *Raw images*).

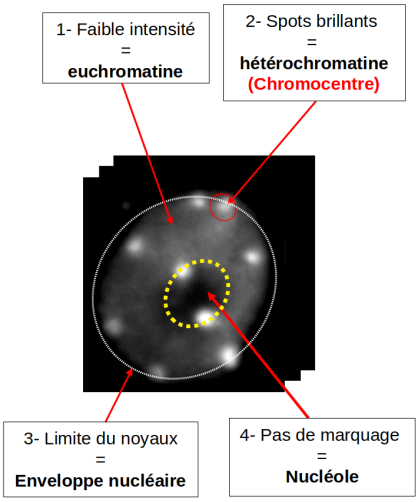


FIGURE 1.15: **Description des composantes du noyaux d'*A. thaliana*** : Noyau à l'interphase d'une cellule d'*A. thaliana* coloré au DAPI. Nous avons considéré 4 composantes dans nos analyses d'image en fonction de la coloration au DAPI : 1) l'euchromatine, 2) l'hétérochromatine, 3) la limite de coloration et 4) le nucléole.

Afin de répondre à ces difficultés NucleusJ propose à partir d'images contenant un seul noyau de réaliser dans un premier temps la segmentation du masque du noyau basée sur la recherche d'un seuil d'Otsu qui maximise la sphéricité de l'objet segmenté. Cette méthode permet de segmenter automatiquement des noyaux avec une plus grande diversité d'intensité et notamment les noyaux de faible intensité comme dans la Figure 1.14 A). Pour faciliter la segmentation des chromocentres, NucleusJ propose une méthode de ligne de partage des eaux en 3D (Figure 1.10), qui contient toutefois une étape de seuillage manuelle par l'utilisateur.

Les solutions proposées par le plugin NucleusJ pour accélérer les étapes manuelles sont cependant insuffisantes dans le cas d'analyse contenant plusieurs centaines de noyaux. Les étapes manuelles ralentissant le processus d'analyse sont les suivantes :

- Dans un premier temps, l'utilisateur devra constituer une collection d'imagettes ne contenant qu'un seul noyau à partir images grands champs, Figure 1.16) : étape de crop.
- Deuxièmement, après l'étape de segmentation du masque du noyau, il est nécessaire de contrôler manuellement le bon déroulement de la segmentation afin d'éliminer les images contenant des invaginations provenant de gros nucléoles (Figure 1.14 B) et C)). Cette sélection manuelle conduit souvent à écarter les plus gros noyaux, ce qui introduit un biais d'analyse par l'élimination des populations cellulaires avec des gros nucléoles comme c'est le cas pour les cellules de pavements de l'épiderme qui est notre tissu de prédilection lors de la collecte des images.
- Troisièmement, l'étape de segmentation des chromocentres pour décrire l'organisation de la chromatine qui fait appel à une méthode de ligne de partage des eaux est limitée en terme de débit car elle nécessite d'appliquer un seuil sélectionné par l'utilisateur. Ce seuil pourra varier d'un utilisateur à un autre et une telle méthode pose donc des problèmes de répétabilité entre opérateurs.

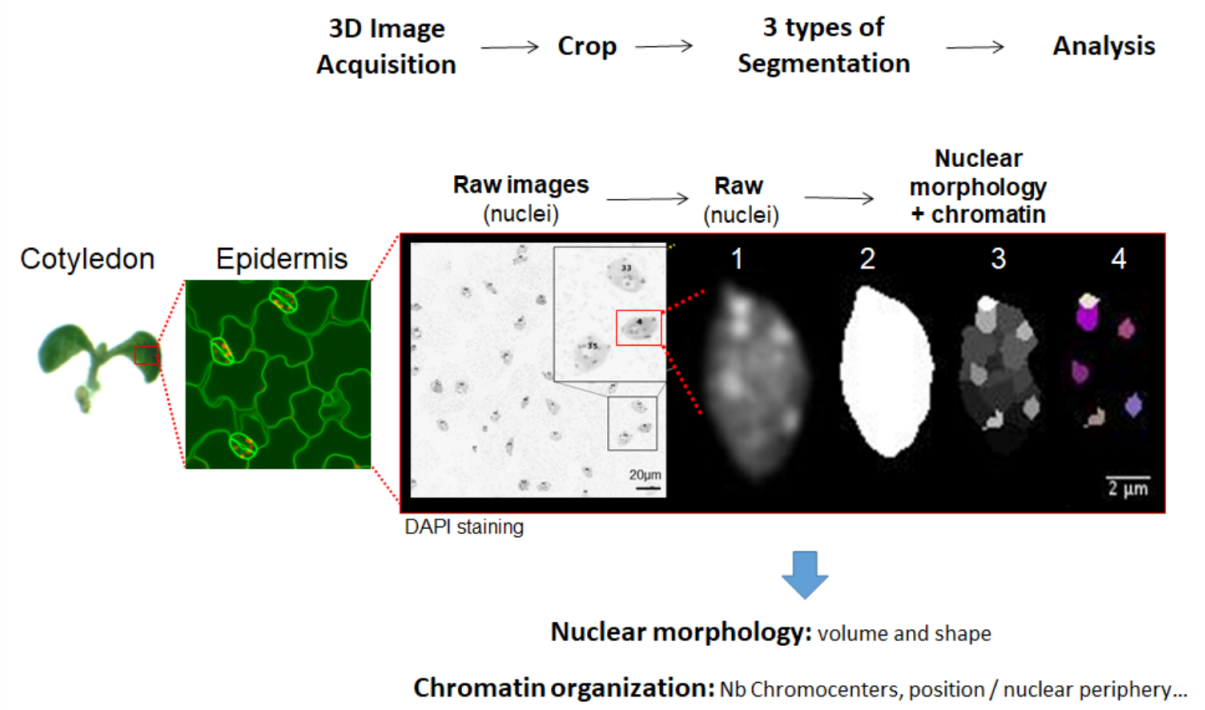


FIGURE 1.16: **Pipeline du logiciel NucleusJ permettant l'analyse 3D des noyaux d'*A. thaliana*.** [Poulet et al., 2015]. Chaque image complexe d'un champ de noyaux d'épiderme du cotylédon d'une plantule *A. thaliana* contient de 20 à 100 noyaux. 2) Chaque noyau est individualisé (crop) manuellement par l'utilisateur. 3) La segmentation du noyau est basée sur la méthode d'Otsu. 4) La segmentation des chromocentres est basée sur l'algorithme de ligne de partage des eaux, ici appliqué à des images 3D. 5) Au final, l'expérimentateur détermine manuellement le seuil à appliquer afin d'obtenir une segmentation reflétant l'image initiale : sur cette image 6 chromocentres ont été validés manuellement par le biologiste.

Chapitre 2

Solutions dédiées à l'étude de la morphologie nucléaire : Nucleus2.0

NucleusJ (NJ1), le premier outil créé dans l'équipe pour analyser la morphologie des noyaux et leur contenu en chromatine, a été conçu pour analyser des images acquises noyau par noyau avec un microscope confocal (*CLSM-Confocal Laser Scanning Micrsocpy*) [Poulet et al., 2015]. La qualité des images obtenues à ce moment-là nécessitait un prétraitement par une déconvolution. L'étape de déconvolution qui consiste à corriger les altérations de l'image lors de sa numérisation venait s'ajouter au temps de l'acquistion via le microscope confocal ce qui ralentissait les analyses. Un microscope à sectionnement optique par lumière structurée acquis par l'équipe (Leica DM 6000 équipé d'une optigrid) permet la capture d'images de champs larges contenant plusieurs noyaux et ne nécessitant pas de déconvolution avant leur analyse par NucleusJ. La capture des images de champs larges facilite le repérage du type cellulaire comme les cellules de garde et de pavement dans l'épiderme de feuille et permet donc de classer les noyaux en sous-populations (Figure 2.2). Les images à champ large donnent également accès à des informations sur la distribution spatiale des noyaux ou encore sur leur orientation dans le tissu. Ce paramètre spatial est particulièrement intéressant pour l'étude de notre tissu d'intérêt dans l'équipe qui est un épithélium de feuille ou cotylédon, un tissu polarisé, c'est-à-dire composé de cellules asymétriques et très organisées entre elles [Wada, 2018]. Pour finir, l'étape d'acquisition étant plus rapide, elle permet l'augmentation du volume de données et naturellement l'obtention de résultats statistiquement plus robustes. En effet, particulièrement dans le domaine de la biologie, la variabilité individuelle au sein d'une population, que ce soit à l'échelle d'un organisme, d'un organe ou encore d'une cellule,

est un souci récurrent.

La première version de NucleusJ (NJ1) permettait de calculer 15 paramètres décrivant la morphologie des noyaux, et de façon beaucoup plus originale, l'organisation de la chromatine en étudiant les chromocentres. L'outil a été validé sur des populations de noyaux d'origines cellulaires diverses (poils racinaires et cotylédon) et il a ouvert la voie au phénotypage de mutants de différentes protéines nucléaires étudiées par l'équipe [Poulet et al., 2017, Benoit et al., 2017]. Les premiers efforts de l'équipe ont porté sur l'amélioration des techniques de préparation des échantillons et l'acquisition des images. A mon arrivée, la technique de production d'images était rapide et complètement standardisée [Desset et al., 2018]. Pour une condition biologique donnée, plusieurs centaines d'images de noyaux sont produites. Cependant, l'utilisation de la microscopie orientée vers un plus haut débit a décalé la problématique du temps nécessaire pour générer les données au temps nécessaire pour leur analyse. C'est pourquoi le premier objectif de ma thèse a été d'automatiser le plugin. Cette première partie introduit donc les améliorations apportées par la seconde version du plugin NucleusJ2.0, et sera suivie des résultats obtenus présentés par l'article : Automated 3D bio-imaging analysis of nuclear organization by NucleusJ 2.0. Suite à l'article, je présenterai les améliorations intégrées après sa publication.

2.1 Automatisation de la segmentation du masque du noyau

L'algorithme de segmentation des noyaux individualisés, NJ1, propose une solution dynamique qui s'adapte bien au noyaux de petite taille et de forte intensité mais commet des erreurs pour les cas suivants :

- Les images de masques de noyaux dont l'intensité moyenne est faible (Figure 1.14 B)). Ces noyaux segmentés présentent des invaginations artificielles en bordure et dans le centre de l'image.
- Les images de masques de noyaux contenant de gros nucléoles excentrés créent des cavités qui sont des artefacts et qui ne sont pas comptabilisés dans le volume du noyau. Ce type de cas est sur-représenté dans les noyaux les plus gros comme ceux des cellules de pavement et créer des biais dans les analyses statistiques.

Ces deux types d'erreur de segmentation représentent 30% des noyaux analysés dans nos tissus de cotylédons et nous obligent à filtrer nos données avant analyse statistique. Pour éviter cette étape de tri des images, nous avons implémenté une étape de calcul d'enveloppe convexe appelé gift wrapping à partir de la segmentation d'Otsu modifé.

Nous avons spécifiquement implémenté l'algorithme de la marche de Jarvis [Jarvis, 1973] qui est le suivant :

Algorithm 1: Adaptation de l'algorithme de la marche de Jarvis

Data: L : Ensemble des points bordant le masque issus de Otsu modifié

Data: D : 2 x (rayon du masque issu de Otsu modifié)

Data: L' : liste de points

repeat

  Data: p0 : point d'abscisse minimale dans L

  Data: p' : le point courant

  for *p dans L* **do**

    if *angle entre p' p est le plus incliné vers la gauche ET distance euclidienne entre p et p' < D*

    then

      Data: Ajouter p' dans L'

      Data: p' = p

until *p' = p0*;

L'algorithme répète cette procédure dans tous les plans de l'image en 3D du masque du noyau obtenus par la segmentation Otsu modifié, c'est-à-dire XY, XZ et YZ. Suite à cela, l'union de l'ensemble des plans segmentés est réalisée. Afin d'optimiser les temps de calcul du *gift wrapping*, nous avons limité l'exploration dans le nuage de points à une distance arbitraire. Cette dernière est calculée à partir du rayon du masque du noyau issu de la segmentation par la méthode de Otsu modifiée.

La segmentation par la méthode de gift wrapping montre des résultats satisfaisants pour les noyaux contenant de gros nucléoles comme nous pouvons le voir dans les exemples donnés en Figures 2.1 A), B) et C). Les objets segmentés présentent des volumes supérieurs car ils comblent les invaginations artéfactuelles en bordure de noyau. Ceci a d'ailleurs pu être démontré par l'utilisation de billes fluorescentes de taille standardisée utilisées au laboratoire pour calibrer les microscopes. Dans ce cas, plus la sphéricité est proche de 1 et moins cette augmentation du volume par la méthode du gift wrapping est forte (Figure 2B [Dubos et al., 2020]). Toutefois, nous observons une segmentation trop importante des noyaux ayant une forme allongée comme dans l'exemple D) de la Figure 2.1. Cependant, ces cas sont rares dans nos échantillons (moins de 0,1% sur les 680 noyaux Col-0 étudiés).

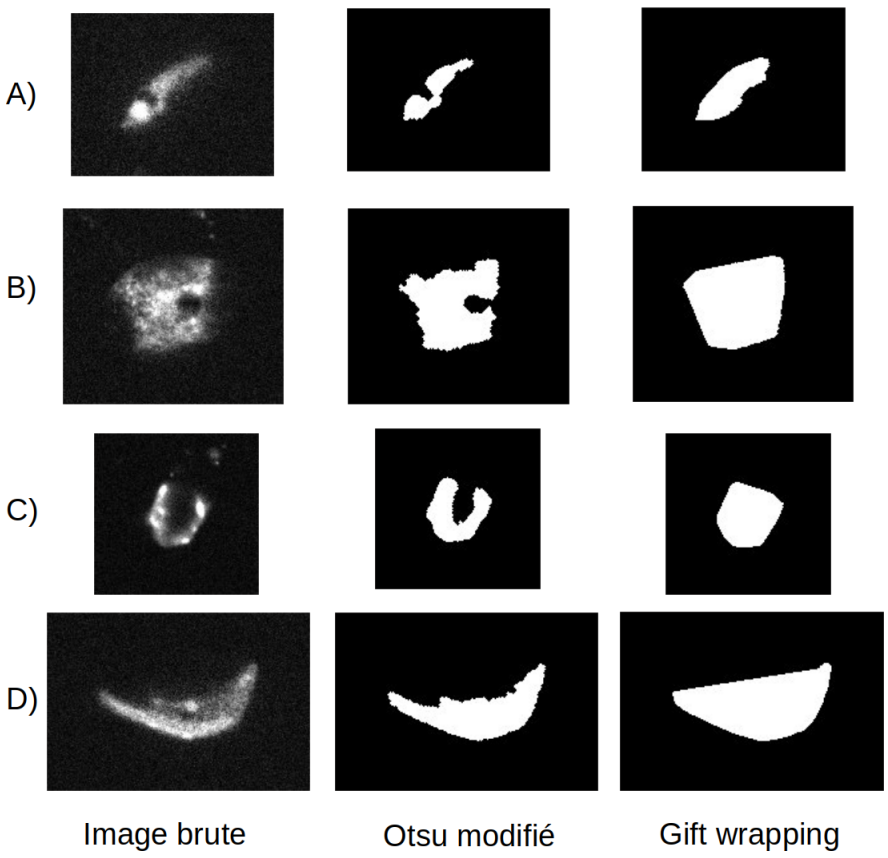


FIGURE 2.1: **Exemples de plan de segmentation par l'utilisation du gift wrapping** : Comparaison des images brutes, des noyaux segmentés par la méthode d'Otsu modifié (NJ1) et la segmentation par la méthode du gift wrapping (NJ2). A), B) et C) sont des exemples d'anomalies de segmentation liées à la présence du nucléole et D) liées à la forme allongée du noyau.

2.2 Automatisation de l'étape d'isolement des noyaux issus d'une image grands champs : l'*autocrop*

Le tri des noyaux ayant de gros nucléoles ou des défauts de segmentation en périphérie nucléaire n'est plus nécessaire grâce la mise en place du calcul de l'enveloppe convexe comme nous venons de le voir. Cet algorithme fonctionne avec une image en entrée contenant un noyau isolé dans un fichier. Ce format n'est plus d'actualité dans nos protocoles expérimentaux comme nous l'avons vu dans la Figure 2.2 B) notre microscope de prédilection génère des images 3D de champ large contenant plusieurs noyaux. L'évolution de la méthode d'acquisition engendre donc une étape très chronophage de sélection/découpage manuel de chaque noyau dans des fichiers séparés. Ce temps est estimé entre 1 à 2 heures par pile d'images de 10 à 60 noyaux. Nous avons donc implémenté dans NucleusJ2.0 [Dubos et al., 2020] une nouvelle méthode appelée *autocrop*, qui automatise cette étape, basée sur la détection de composantes connexes suite à une segmentation par la

méthode d'Otsu [Nobuyuki Otsu, 1979]. Une étape de filtre des composantes connexes est nécessaire pour filtrer les chloroplastes qui sont aussi révélés par le marquage au DAPI.

L'algorithme de l'*autocrop* est le suivant :

Algorithm 2: Algorithme de l'*autocrop*

Data: Img : Images 3D contenant 1 ou plusieurs noyaux

Segmentation Otsu de Img

Data: Composante : liste des composantes connexes dans Img

for *C dans Composante* **do**

  if *C* > *Volume minimum ET C* < *Volume maximum* **then**

    Création d'un fichier

L'*autocrop* a été développé avec de nombreuses options modifiables (Annexe 6.2) afin de pouvoir s'adapter à des images 3D d'autres modèles. L'ensemble des possibilités est fournie dans la documentation https://gitlab.com/DesTristus/NucleusJ2.0/-/wikis/Autocrop. Par exemple, lors de la sélection des composantes connexes, il est possible de filtrer sur la taille minimale de chaque composante connexe et par défaut, l'algorithme utilise 1 unité (1 unité = 1 100 μm^3 dans nos images). Cette limite inférieure nous permet de contre-sélectionner les chloroplastes qui sont de très petites structures par rapport au noyau et que nous souhaitons écarter de notre analyse. Toutefois, ce seuil reste ajustable pour pouvoir analyser des cellules qui aurait des tailles moyennes supérieures. Un second exemple de flexibilité de l'algorithme est illustré lors des études en Fluorescence *in situ* Hybridization (FiSH) du papier NucleusJ2.0 nous pouvons choisir dans quel canal calculer le seuil d'Otsu pour la détection des composantes connexes. Une fois les objets détectés, les coordonnées des sous-images (ou imagettes) définies par l'*autocrop* seront appliquées à l'ensemble des canaux de l'image. Cette méthode est particulièrement intéressante pour la co-localisation de signaux issus de plusieurs canaux de fluorescence par exemple CFP et YFP. Une expérience de FiSH est décrite Figure 1-E de l'article [Dubos et al., 2020]. L'algorithme permet ainsi de choisir dans quel canal sera calculé le seuil d'Otsu pour la détection des composantes connexes. Une fois les objets détectés, les coordonnées des sous-images définies lors de l'*autocrop* seront appliquées à l'ensemble des canaux de l'image.

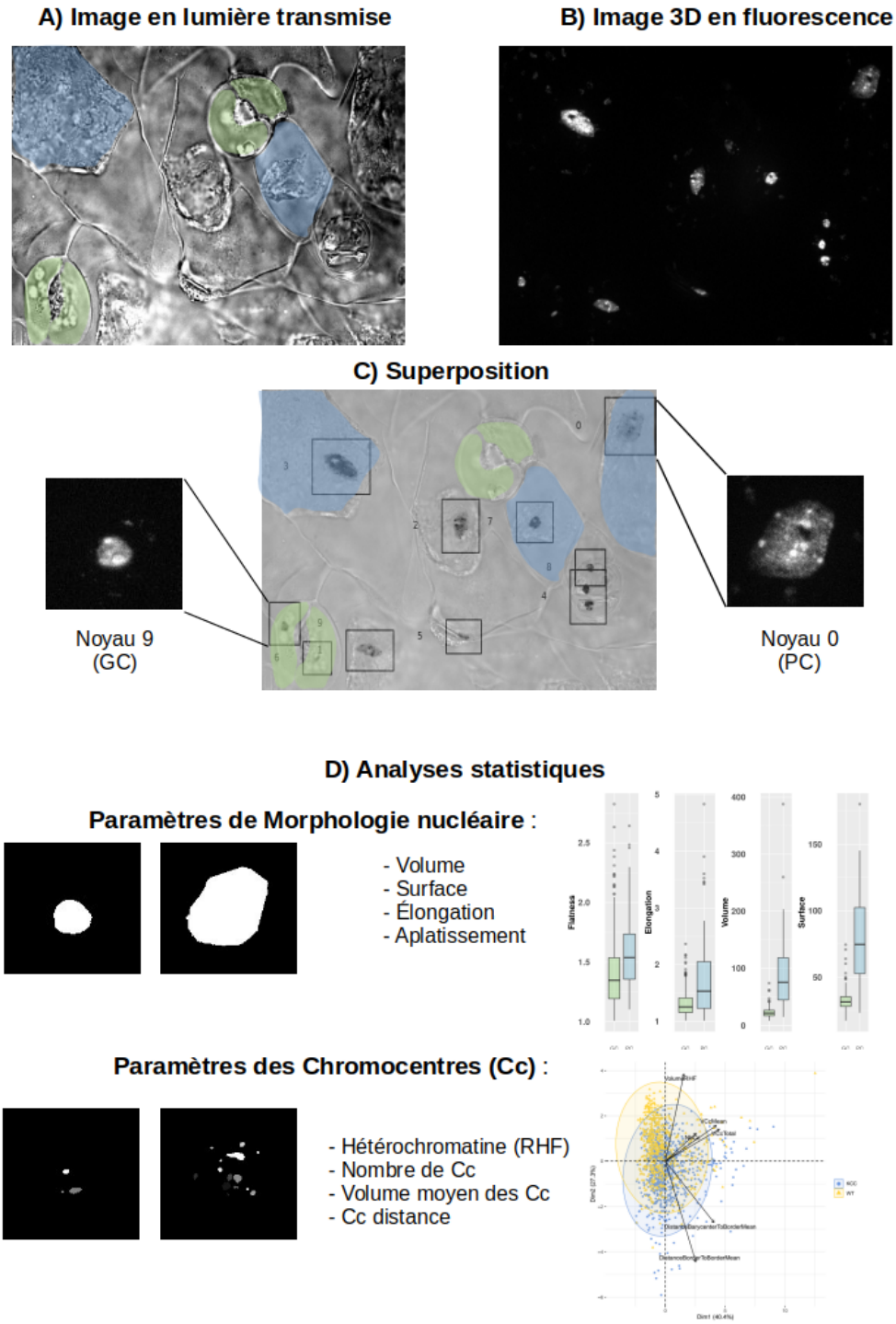


FIGURE 2.2: **Description de l'analyse de la morphologie nucléaire** : A) acquisition de portions de tissus en lumière transmise pour l'identification des types cellulaires (en bleu les cellules de pavement (PC), en vert les cellules de garde (GC)). B) Image grand champ des noyaux par fluorescence 3D suite à une coloration au DAPI. C) Annotation manuelle des types de cellules après fusion de la projection issue de l'image 3D (dont les couleurs ont été inversées pour faciliter l'annotation) depuis laquelle ont été isolés les noyaux (un par fichier, exemple noyaux 0 et 9) étape appelée crop. D) Segmentation du masque du noyau et des chromocentres, à l'issue des segmentations les paramètres géométriques permettant de discriminer les populations cellulaires sont utilisés pour représenter les résultats sous forme de graphes (boxplots, PCA...).

2.3 Validation biologique

Afin de valider l'automatisation des étapes d'autocrop et de segmentation par *gift wrapping* inclues dans NJ2, nous avons réalisé une étude à grande échelle avec un jeu de données de 266 images 3D grand champ, soit plus de 3000 noyaux d'épiderme de feuille générés à partir de 8 génotypes. Cette étude s'est concentrée sur les composants du nucléosquelette et plus particulèrement sur les gènes KAKU4, CRWN1 et CRWN4. Notre équipe disposait d'une lignée sauvage (sans mutation) et de mutants correspondant à la perte de fonction de chaque gène mais également de combinaison entre ces mutants. Nous avions donc à notre disposition le génotype sauvage, les simples mutants *kaku4*, *crwn1* et *crwn4*, les doubles mutants *kaku4 crwn1*, *kaku4 crwn4* et *crwn1 crwn4* et le triple mutant *kaku4 crwn1 crwn4* [Dubos et al., 2020], les résultats du triple mutant *k4c1c4* sont présentés car ce triple mutant présente des différences significatives par rapport au génotype sauvage. J'ai également effectué l'étude de l'organisation de la chromatine, soit après une coloration au DAPI (analyse des chromocentres), soit après une analyse FISH (étude des gènes codant pour les ARN ribosomiques et du marqueur 180bp présent dans les régions centromériques de chaque chromosomes).

2.4 Organisation, stockage et partage de jeux de données

Le partage de données d'imagerie 3D en biologie est une étape importante encore assez peu courante. Partager des jeux de données expertisés permet de comparer des résultats entre différents laboratoires et permet de calibrer et tester les algorithmes dédiés à l'analyse d'image (*benchmarking*). Pour cela, il faut disposer d'outils d'analyse communs et/ou de données communes. C'est pourquoi l'outil NucleusJ2.0 est open source https://gitlab.com/DesTristus/NucleusJ2.0, et nous avons cherché à rendre publique nos données de manière pérenne dans une base de données OMERO [Goldberg et al., 2005] : https://omero.bio.fsu.edu/webclient/userdata/?experimenter=-1. Nous proposons une traçabilité des résultats en mettant à disposition 1) les images brutes, 2) les images segmentées, 3) les fichiers de résultats décrivant les paramètres, les variables géométriques (volume ou encore la surface) mais aussi la version du logiciel utilisé ainsi que les paramètres d'analyses. Cette contribution nous paraissait importante lors de l'initiation de mes développements. Je me suis rendu compte de la difficulté de trouver des jeux de données pour tester mes algorithmes. Lors de mes recherches bibliographiques, il m'est apparu une nette fracture entre le monde de la biologie et de l'informatique/calcul digital. J'ai constaté deux cas de figures :

- Dans les publications de traitement du signal, de nombreux jeux de données *in silico* dédiés à la mesure théorique de paramètres issus de méthodes mathématiques sont souvent disponibles. Cependant, ces données "parfaites" ne reflètent pas la réalité, elles n'incluent pas la convolution de l'image lors de sa numérisation, ce qui entraîne des altérations de l'image.
- A l'inverse, dans les publications de logiciels ou d'algorithmes dédiés à l'analyse d'image, les résultats obtenus sont souvent spécifiques du jeu de données étudié, lequel contient très souvent un nombre faible d'images et est souvent incomplet. Appliquer de telles méthodes sur nos images se solde souvent par un échec s'expliquant par des images avec des caractéristiques différentes (en termes de dynamique des intensités et/ou de bruit de fond).

Ce constat nous a amené dans une démarche visant à associer l'ensemble des jeux de données ayant permis de valider notre outil. Une telle démarche permettra à un utilisateur de reproduire dans son environnement informatique les résultats obtenus dans notre laboratoire. Cette étape est souvent un pré-requis avant l'analyse d'un nouveau jeu de données dans un autre laboratoire. Nous avons donc rendu public à la communauté scientifique un premier jeux de données contenant des billes fluorescentes de taille théorique connue (1μm, 2.5μm et 4μm de diamètre) https://omero.bio.fsu.edu/webclient/userdata/?experimenter=902. Ces billes utilisées en métrologie lors de la calibration des microscopes nous ont permis de valider nos paramètres mathématiques (volume et surface) en situation réelle. La seconde partie du jeu de données que nous avons expertisé est composé de noyaux d'épithélium de cotylédons d'*A. thaliana* provenant d'un génotype sauvage et du mutant *k4c1c4*. Il nous a permis de valider notre processus d'analyse d'un point de vue algorithmique avec l'automatisation de l'analyse, mais aussi d'un point de vue biologique en montrant des différences attendues entre les populations par comparaison des paramètres morphologiques et des paramètres liés à l'organisation de la chromatine obtenus.

2.5 Développements informatiques associés à l'utilisation de NucleusJ2.0

Les méthodes de programmation en langage java ont évolué depuis la première version NJ1 avec notamment l'apparition de Maven https://maven.apache.org/ qui est un outil de gestion et d'automatisation de production. Cette plateforme est très utilisée par la communauté de développeurs de ImageJ et facilite

le partage et l'intégration du code. Ceci a permis d'intégrer plus facilement de nouvelles dépendances dans NucleusJ2.0, nous avons par exemple certaines fonctionnalités de morpholibJ [Legland et al., 2016] pour le calcul de paramètres morphologiques ou encore bio-format https://www.openmicroscopy.org/bio-formats/ permettant de lire les images en formats propriétaires issus des microscopes. Maven permet également de produire une version logicielle complète incluant toutes les dépendances. Cette fonctionnalité représente deux avantages majeurs :

- Une **facilité d'installation** pour les utilisateurs biologistes qui ne nécessite uniquement un fichier au format jar à intégrer dans ImageJ.
- Une **simplification de la maintenance des fonctionnalités** du plugin, avec la possibilité de tester rapidement la compatibilité des dépendances en cas de changement de version.

Une grosse partie de mon travail de début de thèse à consisté à prendre en main ces outils et méthodes de développement informatique pour réorganiser le code de NucleusJ et ensuite y intégrer les nouvelles fonctionnalités et proposer NucleusJ2.0.

2.6 Article : "Automated 3D bio-imaging analysis of nuclear organization by NucleusJ 2.0"

Nucleus

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/kncl20

Automated 3D bio-imaging analysis of nuclear organization by NucleusJ 2.0

Tristan Dubos , Axel Poulet , Céline Gonthier-Gueret , Guillaume Mougeot , Emmanuel Vanrobays , Yanru Li , Sylvie Tutois , Emilie Pery , Frédéric Chausse , Aline V. Probst , Christophe Tatout & Sophie Desset

To cite this article: Tristan Dubos , Axel Poulet , Céline Gonthier-Gueret , Guillaume Mougeot , Emmanuel Vanrobays , Yanru Li , Sylvie Tutois , Emilie Pery , Frédéric Chausse , Aline V. Probst , Christophe Tatout & Sophie Desset (2020) Automated 3D bio-imaging analysis of nuclear organization by NucleusJ 2.0, Nucleus, 11:1, 315-329, DOI: 10.1080/19491034.2020.1845012

To link to this article: https://doi.org/10.1080/19491034.2020.1845012

NUCLEUS
2020, VOL. 11, NO. 1, 315–329
https://doi.org/10.1080/19491034.2020.1845012

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

Automated 3D bio-imaging analysis of nuclear organization by NucleusJ 2.0

Tristan Dubos [a], Axel Poulet[b], Céline Gonthier-Gueret [c], Guillaume Mougeot [a,d], Emmanuel Vanrobays [a], Yanru Li[e], Sylvie Tutois [a], Emilie Pery [f], Frédéric Chausse [f], Aline V. Probst [a], Christophe Tatout [a], and Sophie Desset [a]

[a]GReD, CNRS, INSERM, Université Clermont Auvergne, Clermont-Ferrand, France58; [b]Department of Molecular, Cellular & Developmental Biology, Yale University, New Haven, CT, USA; [c]Neurodol, INSERM, Université Clermont Auvergne, Clermont-Ferrand, France; [d]Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK; [e]Department of Plant and Microbial Biology, Zürich-Basel Plant Science Center, University of Zürich, Zürich, Switzerland; [f]Institut Pascal, Université Clermont Auvergne, Clermont-Ferrand, France

ABSTRACT

NucleusJ 1.0, an ImageJ plugin, is a useful tool to analyze nuclear morphology and chromatin organization in plant and animal cells. NucleusJ 2.0 is a new release of NucleusJ, in which image processing is achieved more quickly using a command-lineuser interface. Starting with large collection of 3D nuclei, segmentation can be performed by the previously developed Otsu-modified method or by a new 3D gift-wrapping method, taking better account of nuclear indentations and unstained nucleoli. These two complementary methods are compared for their accuracy by using three types of datasets available to the community at https://www.brookes.ac.uk/indepth/images/. Finally, NucleusJ 2.0 was evaluated using original plant genetic material by assessing its efficiency on nuclei stained with DNA dyes or after 3D-DNA Fluorescence *in situ* hybridization. With these improvements, NucleusJ 2.0 permits the generation of large user-curated datasets that will be useful for software benchmarking or to train convolution neural networks.

ARTICLE HISTORY
Received 18 August 2020
Revised 18 October 2020
Accepted 27 October 2020

KEYWORDS
Three-dimensional microscopy imaging; image analysis; plant nucleus; nuclear organization; 3D DNA FISH

Introduction

Investigation of nuclear morphology and its impact on chromatin organization is an active field [1]. Identification of key players that determine nuclear morphology at the nuclear periphery is a complex task, for which the model plant *Arabidopsis thaliana* offers an amenable genetic system. As in animals, the plant nucleus is delimited by a double membrane interrupted by nuclear pores, which allow exchanges with the cytoplasm. The outer nuclear membrane (ONM) is connected to the endoplasmic reticulum and the cytoskeleton, while the inner nuclear membrane (INM) is linked with a filamentous network constituting the nucleoskeleton [2]. The Linker of Nucleoskeleton and Cytoskeleton (LINC) complex provides a junction between the interior of the nucleus and the cytoplasm by means of SUN (Sad1 and Unc-84 homology) proteins anchored in the INM that interact with the nucleoskeleton [3,4] and KASH (Klarsicht, Anc-1 and Syne homology) proteins anchored in the ONM and connected to the cytoskeleton, respectively [5,6]. In animals, the nucleoskeleton is made of lamins. Plants do not have lamin orthologs in term of sequence homology, but the CROWDED NUCLEI (CRWN) proteins, which contain long coiled-coil regions like lamins, are likely to have similar functions in nuclear morphology and chromatin organization [7,8] as well as in regulation of gene transcription [9,10]. Other proteins are thought to be involved in the constitution, or anchoring, of the nucleoskeleton at the nuclear periphery, such as KAKU4 [11] and NUCLEAR ENVELOPE ASSOCIATED PROTEINS (NEAPs) [12]. Most of these protein components have been shown to impact nuclear morphology, nuclei becoming usually smaller and more spherical in mutant backgrounds [13], allowing genetic screens for mutants with altered nuclear morphology [11]. To better understand the impact of the nuclear periphery on nuclear

CONTACT Sophie Desset sophie.desset@uca.fr GReD, CNRS, INSERM, Université Clermont Auvergne, Clermont-Ferrand, France
Supplemental data for this article can be accessed here.

organization, molecular techniques applied to the whole genome such as chromatin conformation capture (Hi-C) or Chromatin Immunoprecipitation (ChIP-Seq) have been applied in plants and brought a new vision of the 3D genome [14–16].

Bio-imaging is a complementary approach, in particular when coupled to quantitative image analysis [17]. However, the throughput of bio-imaging limits its routine use in 3D image processing [18]. To this aim, more automated methods are needed and efficient 3D segmentation is required to delimit the boundary of objects such as the nucleus or its chromatin organization [19]. Investigation of the cell nucleus has strongly benefited from these applications, as the nucleus is a spatial structure for which morphology can be modified in diseases [20,21]. Segmentation applied to nuclear organization is also a research tool to investigate the genetic determinants of nuclear size and shape and also chromatin organization [22,23]. There are various softwares packages dedicated to nuclear segmentation [24]. These are usually developed for the detection of nuclei in a wide-field stack and to compute morphology parameters, but only a few of these go as far as the analysis of the content and organization of chromatin [25,26]. Moreover, most of them are optimized for one given tissue or cell type but they are not functional on images containing a large diversity of nuclear size and shape as found in plants [27]. These limitations motivated the development of an ImageJ plugin called NucleusJ dedicated to the analysis of nuclear organization of 3D plant nuclei [23]. Within NucleusJ 1.0, segmentation methods to compute nuclear morphology (a modified Otsu threshold method) and chromatin organization (a 3D watershed method) were chosen as the most relevant methods for nuclear segmentation for 3D nuclei. Although initially developed for plant nuclei stained with DNA dyes [23,28,29], NucleusJ 1.0 can also be used for other cell types [20] and adapted to segment Fluorescence *in situ* hybridization (FISH) signals [29]. However, each NucleusJ 1.0 analysis is time-consuming, the segmentation threshold is user-dependent and nuclear segmentation failed for a substantial fraction of nuclei.

Here, we introduce the optimized NucleusJ version termed NucleusJ 2.0. To increase the number of nuclei considered in a single analysis, a method was introduced to delimit an automatic bounding volume (autocrop) around each nucleus of a 3D wide-field stack containing 10 to a hundred nuclei. Each of the collected nuclei can then be segmented through two complementary methods, either based on the Otsu threshold method or on edge-detection through a 3D gift-wrapping method. From the segmented objects, NucleusJ 2.0 computes new nuclear morphology parameters using a revised and more accurate method of nuclear surface calculation. The accuracy of the measurements performed with NucleusJ 2.0 was confirmed with digitized spheres and multicolor fluorescent beads of standardized sizes. NucleusJ 2.0 was then used to characterize nuclei stained with DNA dyes or labeled with 3D-DNA FISH in whole-mount tissue of a plant mutant with strong alteration of nuclear morphology and chromatin organization. Finally, computation efficiency of NucleusJ 2.0 has been optimized to include a command line version that can be used on distant servers at computing centers.

Material and methods

Plant materials

All mutant and wild type (WT) *Arabidopsis thaliana* plants were from the Columbia-0 (Col-0) ecotype. Mutant lines were T-DNA insertions obtained from The Nottingham Arabidopsis Stock Center (NASC): *kaku4-2* (SALK_076754), *crwn1-2* (SALK_041774) and *crwn4-1* (SALK_079296). The triple mutant *kaku4-2 crwn1-2 crwn4-1* (*k4c1c4*) was obtained by genetic crosses. Cotyledons for image acquisitions were grown from sterilized seeds sown on germination medium containing 0.8% w/v agar, 1% w/v sucrose and Murashige & Skoog salts (M0255; Duchefa Biochemie, Netherlands), grown at 23°C and harvested 13 days after germination (dag). For phenotypic evaluation, seedlings were grown on soil in an Arabilab growth chamber at 20°C. In both cases, seeds were subjected to 2 days of stratification at 4°C in the dark and then grown under 16 h light/8 h dark cycles. The leaf area of the 21-day-

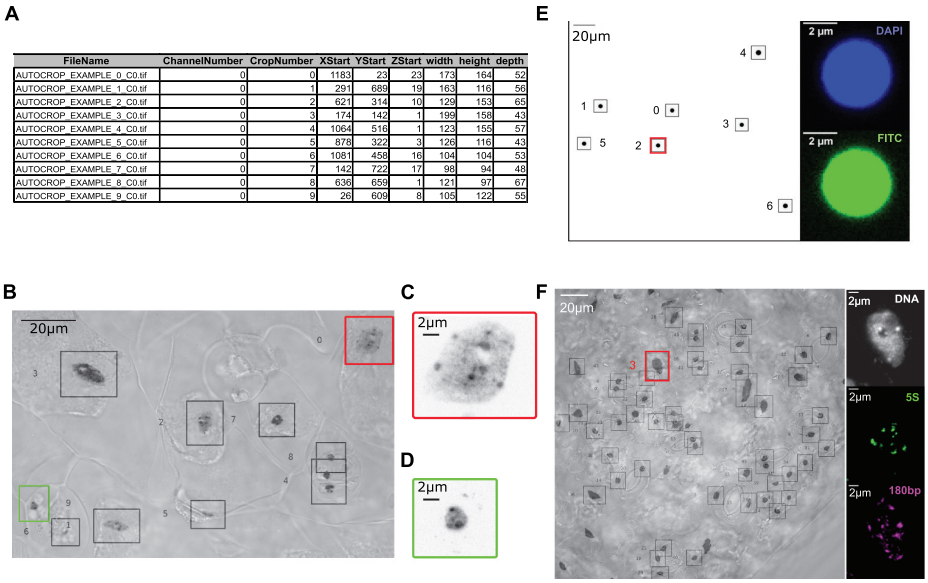
old plants was determined with the ImageJ software using the SIOX (Simple Interaction Object eXtraction) plugin [30]. 13 dag cotyledons were used to determine the number of stomates, guard cell, and pavement cell nuclei. To this aim a maximum Z-projection of a wide-field stack stained with Hoechst 33,258 and a single plane image under transmission light using Differential Interference Contrast (DIC) were combined (overlay) as described in Figure 1.

Digitized spheres and microspheres

Digitized spheres of various radii of 5, 10, 20, 30, 40, and 50 voxels were designed as binary objects (0 for the background, 1 for the object) using isotropic object voxels (*i.e.* cubes where X, Y, Z-axis values are 1, 1, 1) and background voxels that do not belong to the object (Supplemental table 1). Fluorescent microspheres are standardized polystyrene beads commonly used for alignment and calibration of confocal microscopes. Slides containing fluorescent microspheres of 1, 2.5, and 4 µm diameter (Invitrogen) were used in this study (Supplemental table 2).

Sample preparation, DNA staining and 3D DNA-FISH

3D images were obtained from whole mount preparations of 13 dag cotyledons as previously described [29,31]. Whole mount preparations were then used either for DNA staining using Hoechst-staining procedure (Hoechst 33 258 overnight at 1 µg/ml final) or for 3D-DNA Fluorescence *In Situ* Hybridization (3D DNA



FileName	ChannelNumber	CropNumber	XStart	YStart	ZStart	width	height	depth
AUTOCROP_EXAMPLE_0_C0.tif	0	0	1183	23	23	173	164	52
AUTOCROP_EXAMPLE_1_C0.tif	0	1	291	689	19	163	116	56
AUTOCROP_EXAMPLE_2_C0.tif	0	2	621	314	10	129	153	65
AUTOCROP_EXAMPLE_3_C0.tif	0	3	174	142	1	199	158	43
AUTOCROP_EXAMPLE_4_C0.tif	0	4	1064	516	1	123	155	57
AUTOCROP_EXAMPLE_5_C0.tif	0	5	878	322	3	126	116	43
AUTOCROP_EXAMPLE_6_C0.tif	0	6	1081	458	16	104	104	53
AUTOCROP_EXAMPLE_7_C0.tif	0	7	142	722	17	98	94	48
AUTOCROP_EXAMPLE_8_C0.tif	0	8	636	659	1	121	97	67
AUTOCROP_EXAMPLE_9_C0.tif	0	9	26	609	8	105	122	55

Figure 1. Application of NucleusJ 2.0 autocrop method on a wide-field stack.

(a) Output table from a typical experiment 'Autocrop_example' using Arabidopsis cotyledons stained with Hoechst 33,258 DNA dye. (b) Overlay of a maximum Z-projection of the Autocrop_example wide-field stack gained from a sample of cotyledon stained with Hoechst 33,258 and a single plane image under transmission light using Differential Interference Contrast (DIC). Coordinates (X, Y and Z start) of the nine nuclei are described in Figure 1a. DIC allows identifying nuclei of guard cells (autocrop boxes #0 and # 6, respectively, bordered in red and green), which are located at stomates and pavement cells outside of stomates. Autocrop automatically draws bounding boxes (from 0 to n) of the overlay and inverted Lookup Table (LUT) to easily look at the nuclei position. This example also highlights rare cases where the bounding box contains two different nuclei (autocrop box #4).(c) and d) are close-up images of sub-regions of the Z-projection illustrating, respectively, a large pavement cell nucleus (autocrop box #0 in red) and a smaller and rounder guard cell nucleus (autocrop box #6 in green). (e) Multi-channel autocrop on 4 µm microspheres. Maximum Z-projection of a wide-field stack (left) with seven microspheres (#0 to 6). Selected Z-slice (right) of the same microsphere (autocrop box#2 in red) in two fluorescent channels; DAPI channel (top), FITC channel (bottom).(f) Multi-channel autocrop on plant nuclei. Overlay of a maximum Z-projection of the autocrop in DAPI channel wide-field stack (left) with 46 nuclei. Selected Z-slice (right) of nucleus #3 (red box) in three fluorescent channels; DAPI channel (gray level), Cy5 for 5S rDNA probe (green level) and Cy3 for 180bp probe (magenta level).

Table 1. OMERO-FSU Datasets.

DATASET NAMES	ACQUISITION SYSTEM	DATASET NAMES	LINK	KEY FILTER	VALUE FILTER	
#1 DIGITIZED SPHERES	Wide-field + OptoGrid	#1a DIGITIZED SPHERES – RAW	dataset-4902	RAW	DIGITIZED	6
		#1b DIGITIZED SPHERES – SEGMENTED	dataset-4901	SEGMENTATION	OTSU	6
					GIFT-WRAPPING	6
#2 FLUORESCENT MICROSPHERES	Wide-field + OptoGrid	#2a FLUORESCENT MICROSPHERES – ACQUISITION	dataset-4903	RAW	WIDE FIELD STACK	17
		#2b FLUORESCENT MICROSPHERES – AUTOCROP	dataset-4951	PICTURE TYPE	Z MAX PROJECTION	17
		#2 c FLUORESCENT MICROSPHERES – RAW	dataset-4952	RAW	DAPI CROP	81
					FITC CROP	97
		#2d FLUORESCENT MICROSPHERES – SEGMENTED	dataset-4953	SEGMENTATION	OTSU	157
					GIFT-WRAPPING	157
#3 NUCLEAR MORPHOLOGY	Wide-field + OptoGrid	#3a NUCLEAR MORPHOLOGY – ACQUISITION	dataset-4954	RAW	WIDE FIELD STACK	63
		#3b NUCLEAR MORPHOLOGY – AUTOCROP	dataset-4955	PICTURE TYPE	DIC	63
					Z MAX PROJECTION	63
					OVERLAY	63
		#3 c NUCLEAR MORPHOLOGY – RAW	dataset-4956	RAW	DAPI CROP	1554
		#3d NUCLEAR MORPHOLOGY – SEGMENTED	dataset-4957	SEGMENTATION	OTSU	1554
					GIFT-WRAPPING	1553
		#3e NUCLEAR MORPHOLOGY – RAW BAD CROPS	dataset-5009	RAW	BAD CROP	102
#4 CHROMATIN ORGANIZATION	Wide-field + OptoGrid	#4 CHROMATIN ORGANIZATION – RAW	dataset-5018	SEGMENTATION	THRESHOLD WATERSHED	315
#5 180BP_5S DNA FISH	Wide-field + OptoGrid	#5a 180BP_5S DNA FISH – ACQUISITION	dataset-5010	RAW	WIDE FIELD HYPERSTACK	8
		#5b 180BP_5S DNA FISH – AUTOCROP	dataset-5011	PICTURE TYPE	DIC	4
					Z MAX PROJECTION	4
					OVERLAY	4
		#5 c 180BP_5S DNA FISH – RAW	dataset-5012	RAW	CY3 CROP	160
					CY5 CROP	149
					DAPI CROP	160
		#5d 180BP_5S DNA FISH – GIFT SEGMENTED	dataset-5015	SEGMENTATION	GIFT-WRAPPING	159
		#5e 180BP_5S DNA FISH – FISH SIGNALS	dataset-5016	SEGMENTATION	CY3 THRESHOLD WATERSHED	129
					CY5 THRESHOLD WATERSHED	159
#6 5S DNA FISH	Confocal	#6a SINGLE DNA FISH – RAW	dataset-5019	RAW	DAPI CROP	80
					CY3 CROP	80
		#6b SINGLE DNA FISH – GIFT SEGMENTED	dataset-5021	SEGMENTATION	GIFT WRAPPING	80
		#6 c SINGLE DNA FISH – FISH SIGNALS	dataset-5022	SEGMENTATION	CY3 THRESHOLD WATERSHED	75

Six types of image datasets (Dataset names) gained from wide-field or confocal microscopy (Acquisition systems) were stored as six main directories at OMERO-FSU under the name IDP3006_Dubos-Desset _2020. For each dataset, images were organized in acquisition, autocrop, raw and segmented sub-directories and can be directly accessed using the web link included into Table 1 (Link). OMERO allows to screen for key-value pairs (Key and value filters). Number of images are indicated in the last column. The six datasets and their image processing represent a total of 7,313 images.

FISH). 3D DNA FISH experiments were performed as previously described [32]. 5S rDNA probe was produced from pCT4.2 vector [33] amplified with 5 CY5-dUTP and directly-labeled oligonucleotides (/CY5/ CCCAAATTTTGACCTTTAAG) and (/CY5/ GTCGACAAAAAGTCAATGGA) or with 5'CY3-dUTP and the same oligonucleotides without fluorescent labels. 180pb satellite repeat probe was designed as an LNA-oligonucleotide (/5TYE563/ GTATGATTGAGTATAAGAACTTAAACCG – Qiagen) [34].

Microscope and image acquisition

Microscopic observations were performed by structured illumination microscopy to produce wide-field stacks using an Optigrid module (Leica-microsystems MAAF DM 16000B). All stacks were acquired using an X63 oil N.A. 1.4 objective and a digital CMOS Camera (ORCA-Flash4.0 V2 C11440-22 CU – Hamamatsu) at an optimal resolution such that lateral and axial resolution were respectively XY = 0.103 µm and Z = 0.2 µm. For better resolution, some confocal images were acquired with a Zeiss LSM 800 with an X63 oil N.A. 1.4 objective and a voxel size XY = 0.60 µm and Z = 0.2 µm. Final image numbers are given for each dataset in Table 1.

Datasets storage and availability

Datasets were stored at OMERO-Florida State University (OMERO-FSU), a public repository under the accession number IDP3006 Dubos–Desset Nucleus 2020 that can be accessed through the INDEPTH COST-Action (CA16212) website at https://www.brookes.ac.uk/indepth/images/. The INDEPTH image webpage provides a guideline to access and download the datasets that are freely available for research purposes. More tutorials to use OMERO are available at the OMERO webpage at https://www.openmicroscopy.org/.

Six types of datasets were produced for this study (#1 to #6) (Table 1). Each dataset was stored at OMERO-FSU under the accession number IDP3006_Dubos-Desset_2020 where it was organized in four folders: acquisition (wide-field stacks or confocal images), autocrop (results of autocrop processing), raw (3D images containing a single nucleus per stack) and segmented (raw image after segmentation) except for dataset#1, which was generated *in silico* and for that reason no acquisition nor autocrop was performed. OMERO allows screening for a subset of each dataset using key-pair values as described in Table 1. A training dataset is also available at OMERO-FSU under the accession number IDP2002 Dubos – Desset 2020.

NucleusJ 2.0 algorithm development

Documentation: NucleusJ 2.0 is an ImageJ plugin in Java language released as a jar file for the ImageJ platform. Installation and usage guides are available at https://gitlab.com/DesTristus/NucleusJ2.0. The software contains a 3D autocrop module and three independent segmentation methods hereafter referred to as the modified Otsu, 3D gift-wrapping and 3D watershed methods. NucleusJ was described in a previous publication [23]. New functionalities or improvements are detailed below.

Autocrop: From any given 2D or 3D wide-field image a simple Otsu threshold [35] was applied to obtain a binary image *i.e.* transforming all the voxels of the image to a value of 0 for the background voxels and 1 for each voxel from the object (nucleus). Volume of each connected voxel (*i.e.* connected component) was computed using the MorpholibJ library [36] and connected components above 1 µm were conserved. Finally, for each connected component, a coordinate box was designed by adding as a default parameter 20 voxels at the most extreme voxels (*i.e.* boundary voxels) coordinated in each X, Y and Z dimensions. Note that the number of added voxels and hence the size of the coordinate box can be modified according to the specific application. An optional and configurable step is available when multiple coordinate boxes are produced for a single nucleus. This step groups boxes when they display a 50% shared surface. Finally, a maximum Z-projection of the initial wide-field image was automatically generated as a.*tif* file. In this 2D image, each nucleus was numbered and

surrounded by its coordinate box. A tabulated file containing the list of nuclei and coordinate boxes was produced as a *.txt* file. Documentation is provided as Supplemental file 1.

Gift-wrapping: To provide an independent and complementary method to the modified Otsu threshold method available in NucleusJ 1.0, the Jarvis march algorithm [37] was implemented into NucleusJ 2.0 (Supplemental file 2). For the sake of simplicity, the method was designed in 2D and implemented slice by slice to the whole 3D object. Hence, three axes *i.e.* XY, XZ, and YZ were used to decompose the 3D volume in 2D slices. The method then computed the union of each possible plan in XY XZ YZ. For each slice, in order to tune the 3D gift-wrapping algorithm, and to fill the shape artifacts as well as possible, a parameter of maximal threshold distance *td* was applied between two vertices defining the final boundary. The best threshold distance was determined experimentally as the half of the estimated radius of a sphere with a volume equivalent to the object one (Supplemental file 2).

SurfaceArea calculation: When analyzing 3D objects, the surface area plays an important role for shape description. In NucleusJ 1.0, the surface area parameter was computed for any given object as the sum of all voxel boundary exposed to the background. This was improved in NucleusJ 2.0 to better take into account the contribution of each surface element area using the discrete geometry technique [38]. The first step of the area calculation was to determine the image gradient of the raw image f, which was estimated from finite differences in the anisotropic image (Supplemental file 3). The algorithm then browsed each boundary voxel of the segmented object. For each boundary voxel, the contribution of each surface element area was then computed for the final area.

Nuclear morphology parameters in NucleusJ 2.0: NucleusJ 2.0 was revisited for its functionalities to segment nuclei. Description of the quantitative parameters generated by NucleusJ 1.0 can be found in supplemental materials of [23]. Parameters computed by NucleusJ 2.0 are described in Supplemental file 4.

Statistics

Statistical analyses were performed using various R packages [39] for Principal Component Analysis (PCA) ggplot2 [40], factoextra (https://rpkgs.datanovia.com/factoextra/) and FactoMineR [41] were used. Student t-test was used to compare means between WT and mutant backgrounds.

Results

Automatic selection of 3D nuclei from wide-field stacks using a 3D autocrop process

Any kind of 3D bio-image analysis starts with the capture of images of best quality. During this initial step, 3D stacks containing the objects of interest, such as a cell nucleus, are collected either one by one when using a confocal microscope or by collecting multiple nuclei at a time using wide-field microscopy. In the latter, bounding boxes surrounding the 3D nuclei have to be defined to delimit and extract (crop) the appropriate volume containing each nucleus. Automatic tracking of segmented objects would strongly reduce this tedious manual step. Setting-up such middle to high-throughput 3D tracking process is a timely objective when using wide-field stacks to rapidly buildup large image datasets [42].

While ImageJ offers 3D crop and 2D autocrop plugins [42,43] automated and scalable 3D autocrop plugins to extract multiple nuclei from wide-field stacks is missing. Furthermore, detection of multiple objects should take into account variable fluorescence intensity between objects that does not allow the same segmentation threshold to be applied to the whole stack.

Our initial motivation was to implement a simple and rapid method to automatically identify and isolate large numbers of 3D nuclei from wide-field stacks, a process hereafter called autocrop. Once the objects are delimited in the wide-field stack, more sophisticated segmentation methods could be applied for each single object regardless of their fluorescence intensity. The autocrop basic principle relies on a simple Otsu threshold method [35] applied in 3D to the wide-field stack (Materials and Methods). To avoid the selection of too many objects from the background, a simple and scalable size filter was introduced. In our plant model, this has been useful to filter-out chloroplasts that as autofluorescent objects are considered as noise in our analyses. A second scalable filter allows limiting the number of multiple boxes

for a given object, which was proven to be helpful especially for large nuclei. Default autocrop parameters can be modified prior to the analysis through a simple configuration file (*config file*). More details are available in the autocrop documentation (Supplemental file 1). The method is illustrated below for 3D stacks although it can also be used for 2D images.

The autocrop process with default settings was applied for microspheres and wide-field stacks of whole mount cotyledons of WT stained with DNA dyes (Materials and Methods; Supplemental tables 2–3). The autocrop produces i) a collection of 3D images containing a single object stored in a dedicated folder, ii) a table containing the spatial coordinates of the bounding box (X, Y and Z) to trace back their positions in the original wide-field stack, iii) the bounding box volumes (width, height, and depth) and iv) an inverted color 2D Z-projection using maximum intensity projection, in which each object is numbered (Figure 1(a-c)). The number of cropped objects obtained after the autocrop process depends on the density and size of the objects within the original wide-field stack. For instance, 91 successful crops were obtained from 3 wide-field images for 1 µm microspheres, while only 46 crops were obtained for 4 µm microspheres (Supplemental table 2). When imaging the epidermis of an Arabidopsis cotyledon, a typical experiment described in Supplemental table 3 that starts with 12 WT plants generates 35 wide-field stacks (average of 3 images per cotyledon) allows generating 786 crops using the autocrop method (average of 22 crops per wide-field stack). Once the autocrop is performed, an overlay of the Z-projection allows tracing back the nuclei in their tissue context and annotating them regarding their origin (*i.e.* guard cell or pavement cell nuclei) or their pertinence (some objects that are not nuclei are detected) (Figure 1(b-d)). In this experiment, the Z-projection inspection allowed discarding 92 abnormal nuclei. Finally, 694 nuclei (88% of the initial crops) were usable to further analysis of their nuclear morphology.

An interesting functionality of the autocrop is to generate multiple autocrops from a given wide-field stack using the same coordinate table to select identical boxes in different wavelength channels. This application is illustrated with microspheres of 4 µm labeled with fluorophores emitting in the DAPI (blue) and FITC (green) channels (Figure 1(e)) and with plant nuclei labeled with DNA dyes and two fluorescent probes emitting in two distinct channels (figure 1(f)). This functionality allows capturing DAPI and probe channels in typical Fluorescence *in situ* hybridization (FISH) experiments from the same image.

In our hands, the 3D autocrop method applied to wide-field stacks is an efficient automated process to detect position and isolate large numbers of nuclei that can then be subjected to further image analysis.

3D gift-wrapping as a complementary method for nuclear segmentation of plant nuclei

There is an increasing demand in biology to describe nuclear morphology (shape and size) or to evaluate the organization of chromatin domains [18,44]. A first step in such an image analysis is to delimit the nucleus from the background, a process called segmentation. We previously developed a modified Otsu threshold segmentation process delimiting the nuclear boundaries according to their range of gray scale values [23]. However, despite the improvement of the original Otsu method, 10–20% of nuclei with poor segmentation still needed to be discarded after manual curation of the images (Figure 2(a)). Standardization of sample preparation and image acquisition protocol [31] did not overcome this bias.

Careful analysis identified two main characteristics of the nuclei that were poorly segmented. First, the nucleolus that is not stained by DNA dyes such as DAPI or Hoechst is often excluded during the segmentation step. This phenomenon is common for large nuclei containing large nucleoli. A simple fill step was then introduced before segmentation. However, this did not solve the abnormal segmentation when the nucleolus is close to the nuclear periphery (Figure 2(a)). Second, very high signal intensity was observed for some chromocenters of large nuclei leading to the detection of chromocenters not as an object within a nucleus

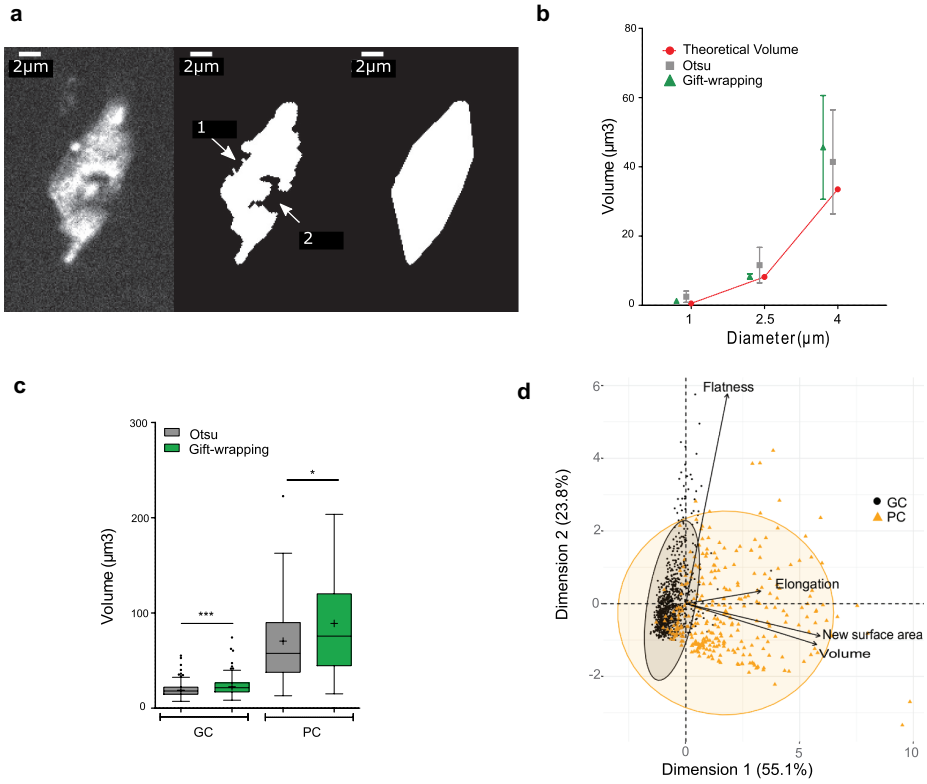


Figure 2. Evaluation of a 3D gift-wrapping method of segmentation.

(a) Example of a nucleus raw slice (left) after Otsu-modified segmentation (middle) and gift-wrapping segmentation (right). Artefactual indentation at the nuclear border (arrow #1); Nucleolus border (arrow #2).(b) Comparison of Otsu and gift–wrapping methods using standardized microspheres. Microsphere volume of 1, 2.5 and 4 µm diameter (n = 28, 24, and 15, respectively) were computed by the Otsu (green triangle) and gift-wrapping (gray square) methods and compared to theoretical volumes (red circle) (Supplemental table 2).(c) Comparison of nuclear volumes after segmentation of plant nuclei by the Otsu or gift-wrapping methods. Nuclei were split into two categories: guard cells (n = 375) and pavement cells (n = 127) (Supplemental table 4) were segmented by the two methods and volumes of the segmented nuclei were computed by NucleusJ 2.0. Modified Otsu method (gray); gift-wrapping (green). Student t-test P-value: *** <0.0001, * = 0.0046.(d) Principal component analysis of morphology parameters (Flatness, Elongation, New surface area and Volume) obtained after segmentation by the gift-wrapping (left) or Otsu (right) methods of the same nuclei as in Figure 2c. Guard cell nuclei (GC, black) and pavement cell nuclei (PC, orange).

but as a distinct object (*i.e.* a nucleus). This results in segmentation of the nucleus into several small objects. In summary, poor segmentation of plant nuclei was strongly biased toward larger nuclei with high contrast between chromocenters and nucleolus that generated concavity artifacts and holes. This also induced a bias in our analysis as a substantial fraction of the larger nuclei was discarded from the analysis.

To improve our 3D segmentation process, an alternative method, hereafter called gift-wrapping, was implemented in 3D using an edge-based segmentation [37]. The algorithm segments one 2D slice at once in all orientations of the 3D object before building the 3D segmentation (Materials and Methods).

The gift-wrapping method was implemented in NucleusJ 2.0, which now produces three folders *i.e.* one folder for nuclei segmented by the Otsu modified method, one segmented by the gift-wrapping method and a third one containing the nuclei that cannot be segmented. This last category is classified as 'bad crop' (Supplemental table 3). First, we confirmed that the method did not alter the segmentation of small convex objects with a homogeneous signal, like microspheres (Figure 2(b), Supplemental table 2). Then, a dataset of 502 nuclei from WT plants stained with Hoechst was used to evaluate the appropriateness of the method (Supplemental table 4). Shape artifacts due to the nucleolus or to low staining intensity were

efficiently removed (Figure 2(a)). The limit of this method is its accuracy, as the measured nuclear volume is significantly increased especially for small objects such as guard cell nuclei (Figure 2 (c)). This over-segmentation also leads to the loss of 0.02% of the nuclei (Supplemental table 4) located at the boundaries of the 3D images that are then considered as incomplete by the software. Although the gift-wrapping method increases the size of segmented objects, PCA analysis confirmed that it does not introduce bias in the analysis and that the two nuclei populations of different cellular origin are still nicely distinguished (Figure 2(d)).

The 3D gift-wrapping segmentation process therefore resolves some drawbacks of our previous Otsu modified method by decreasing the bias previously observed for the segmentation of larger nuclei. It provides an alternative, more efficient, segmentation method for plant nuclei.

Computer parameter optimization

Describing nuclear morphology requires an accurate estimation of the nuclear volume and surface area. NucleusJ 1.0 takes advantage of MorphoLibJ, a library that calculates the volume as the sum of voxels included into the segmented nucleus, multiplied by the volume of an individual voxel (voxel calibration) and computes the surface area through a modified version of the Crofton formula [36].

To compute the surface area, surfaces of all voxels delimiting the boundary of the object are summed [23]. However, this calculation was the more error-prone, with the error increasing for

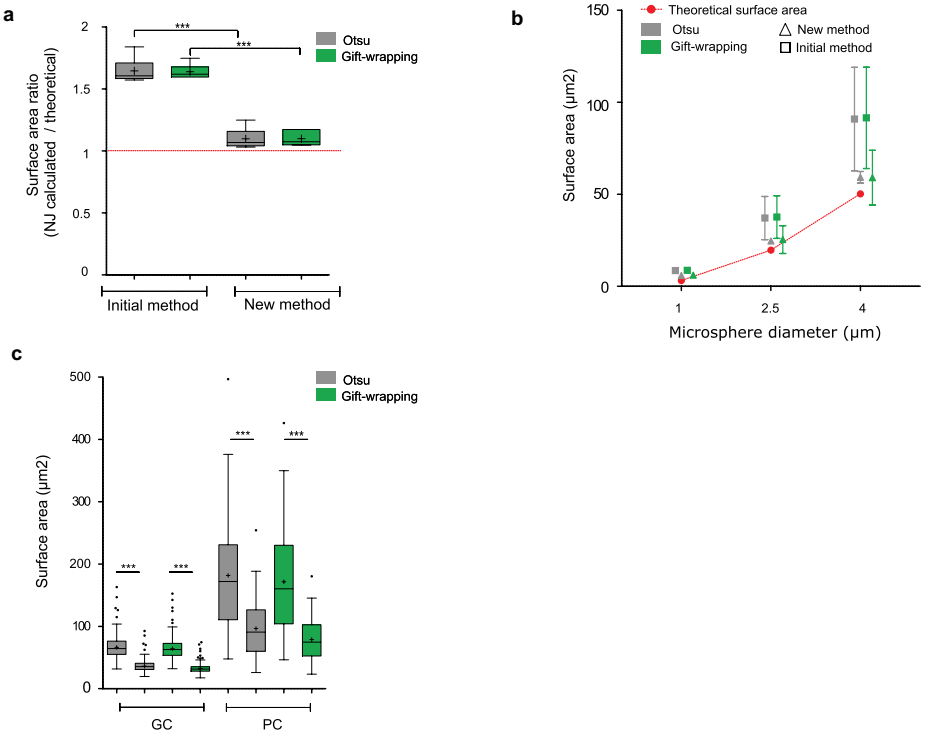


Figure 3. Evaluation of a new method of Surface Area calculation implemented into NucleusJ 2.0.

(a) Digitized spheres of various radii of 5, 10, 20, 30, 40, and 50 voxels were generated and used to calculate the surface area (Supplemental table 1) with the initial and newly developed calculation method. Data are presented as a ratio between the observed and theoretical size of the digitized spheres. Student t-test P-value: *** <0.0001.(b) Fluorescent Microspheres of diameter of 1, 2.5, and 4 µm (n = 28, 24, and 15, respectively) were imaged using the wide-field microscope with an optigrid module, subjected to autocrop and segmented by the two NucleusJ 2.0 segmentation methods. Surface area gained from Otsu and gift-wrapping segmentation was then computed with the initial and new method of calculation (Supplemental table2). Red: theoretical surface area; gray: Otsu method; green: gift-wrapping method; triangle: new method; square: initial method. (c) Plant nuclei. Surface area of nuclei from cotyledon epidermis of WT plants were segmented by the Otsu or gift-wrapping methods. Nuclei from guard cells (GC, n = 375) and pavement cells (PC, n = 127) (Supplemental table 4). Student t-test P-value: *** <0.0001.

small-segmented objects consisting of only a few voxels. This approximation led to significant over-estimation of the surface area, when artificial objects or standardized microspheres were measured with NucleusJ 1.0 (Figure 3(a)).

To correct the surface area computation, a new method (Materials and Methods) was implemented in NucleusJ 2.0 to calculate the surface element (surfel) contribution of each voxel using an iterative convolution method based on *on-surface convolution* [38]. The algorithm browses each boundary voxel delimiting the 3D object from the background of the segmented image. For each boundary voxel, the contribution of each surfel to the final area is computed. While the surface area calculation remains imperfect, it gives more meaningful results for digitized spheres of variable radius with a ratio between calculated and theoretical surface area of 1.068 and 1.084 instead of 1.608 and 1.636, respectively, for the Otsu and gift-wrapping segmentation methods (Figure 3(a)). The accuracy of our surface calculation was also assessed using standardized microspheres of three distinct diameters ranging from 1 to 4 µm. For these, the new surface area calculation was closer to the theoretical values (Figure 3(b)). Surface area was also assessed for a dataset of 502 nuclei from WT plants stained with Hoechst indicating that the surface area was over-estimated when the surface area was calculated with NucleusJ 1.0 (Figure 3(c)).

Taken together, a new surface area calculation is now introduced in NucleusJ 2.0 to obtain a more realistic value of this key parameter.

Application of NucleusJ 2.0 to quantify alterations of nuclear morphology and nuclear organization

To better understand the contribution of nuclear morphology to chromatin organization, one possible approach is to compare WT and mutant plants. *crwn* mutants are well described and show altered nuclear morphology, chromatin organization and gene expression [7–10]. *CRWNs* are a small gene family composed of four members with CRWN1 proteins having major effects on nuclear morphology (smaller and rounder nuclei in *crwn1* mutants) and CRWN4 having major effects on chromatin organization (dispersed chromocenters and 5S rRNA gene signals in *crwn4* mutants). Additive phenotypes are observed in the double *crwn1 crwn4* mutant [8]. Furthermore, an additional protein of the plant nuclear lamina called KAKU4 was found to co-immunoprecipitate with CRWN1 and CRWN4 [11].

We, therefore, used NucleusJ 2.0 to compare WT *Col*-0 seedlings (hereafter WT) and a *kaku4-2 crwn 1–2 crwn 4–1* mutant in the *Col*-0 genetic background (hereafter called *k4c1c4* mutant). The *k4c1c4* mutant was chosen for its strong impact on plant growth (Supplemental Figure 1(a)). We acquired 35 and 28 wide-field stacks, respectively, from 12 WT and 10 mutant plants stained with Hoechst. After the autocrop and segmentation process, WT (n = 502) and *k4c1c4* mutant (n = 672) nuclei were annotated manually as guard (GC) or pavement cell (PC) nuclei after examination of the Z-projection (Supplemental tables 3–4). Despite the reduced plant size of the triple mutant, we found an increased number of stomates (7.5 and 10.11, respectively, in WT and *k4c1c4* mutant; Supplemental table 5) and observed a significant increase in nuclear density both in guard and pavement cells (Supplemental Figure 1b). The dataset was then used to assess nuclear morphology and chromocenter organization. Results gained from our new gift-wrapping method are presented in Figure 4. The nuclei of the *k4c1c4* mutant are strongly affected for most of the nuclear parameters and show reduced nuclear size, reduced elongation (Figure 4(a)) and fusion of chromocenters leading to an increased chromocenter volume (Figure 4(b)). This is very similar to previous results obtained with *crwn1 crwn2* mutant [23]. NucleusJ can also be used to quantify 3D DNA FISH signals. To illustrate this application, two new datasets of 3D-DNA-FISH images were used to investigate the effect of the triple *k4c1c4* mutant. The distinct fluorescent channels were collected corresponding to signals from the DNA dye and the probes for 180bp and 5S rDNA repeats (Supplemental tables 7–8). The *k4c1c4* mutant shows a reduced number of 180bp signals, however, situated at similar positions close to the nuclear periphery as in WT nuclei (Figure 5(a) and 5(c)). In contrast, 5S rDNA loci, while their number is very similar to the WT, are

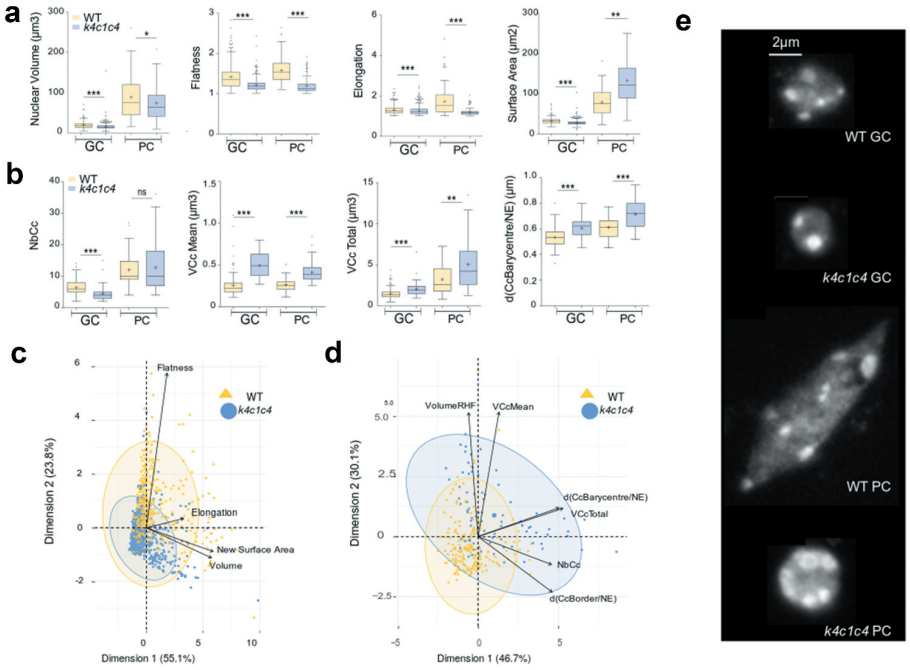


Figure 4. NucleusJ2.0 analysis of the *k4c1c4* mutant with 1altered nuclear morphology and chromatin organization.

Nuclear morphology parameters were computed by the gift-wrapping segmentation method using an initial dataset of WT (n = 663) and *k4c1c4* mutant (n = 881) nuclei described in Supplemental table 4.(a) Nuclear morphology parameters computed by NucleusJ 2.0 on GC and PC cells: nuclear volume (µm^3), flatness, elongation and surface area (µm^2). WT (n = 502) and *k4c1c4* mutant (n = 672) (Supplemental table 4). Student t-test P-value: * <0.01, ** <0.001 and *** <0.0001.(b) Chromatin organization parameters computed by NucleusJ 2.0: number of chromocentres (NbCc), Mean of chromocentre volume (VCc Mean, µm^3), total volume of chromocentres (VCc Total, µm^3), and distance between chromocentre barycenter and the nuclear envelope (d(CcBarycentre/NE), µm). WT (n = 186) and *k4c1c4* mutant (n = 81) (Supplemental table 6). Student t-test P-value: ns > 0.01, ** <0.001 and *** <0.0001.(c) and d) Principal component analysis using morphological parameters (n = 1544; Supplemental table 4) and chromatin organization parameters (n = 267; Supplemental table 6).(e) Typical images chosen with parameters close to the median values of morphological parameters and chromatin organization parameters. Z-projection of raw nuclei. Scale Bar 2 µm. GC: guard cell, PC: pavement cell.

smaller in size and localize closer to the nuclear envelope (Figure 5(b) and 5(d)). Each of the NucleusJ 2.0 parameters describing the *k4c1c4* mutant is illustrated in Supplementary Figure 2 indicating a very strong effect of the mutant background on the organization of the 180bp satellite repeats.

In summary, we demonstrate that the new version of NucleusJ 2.0 can also be efficiently used to quantify alterations of nuclear domains as revealed by DNA dyes (*i.e.* chromocenters) or by 3D-DNA FISH (180bp satellite repeat and 5S rDNA arrays).

Discussion

Nuclei morphology is intrinsically linked to biological processes like gene regulation or development and has been widely used as a marker for human disease [21]. Microscopy imaging provides a classical mean to investigate nuclear morphology variations with segmentation as a critical step in the image analysis process. Although large numbers of segmentation methods are available, no universal segmentation technique will work for all images and all model systems and only few of them are suitable for big data analyses.

We provide evidence that our methods to calculate nuclear morphology parameters are accurate by using virtual objects like digitized spheres and standardized microbeads usually used to calibrate confocal microscopes and cell sorters.

The first step in NucleusJ is to describe the nuclear morphology. We have now implemented two complementary segmentation methods into NucleusJ 2.0 that give slightly different results. As a simple and fast method, the Otsu method relies on the application of a threshold to distinguish the object from the background, but

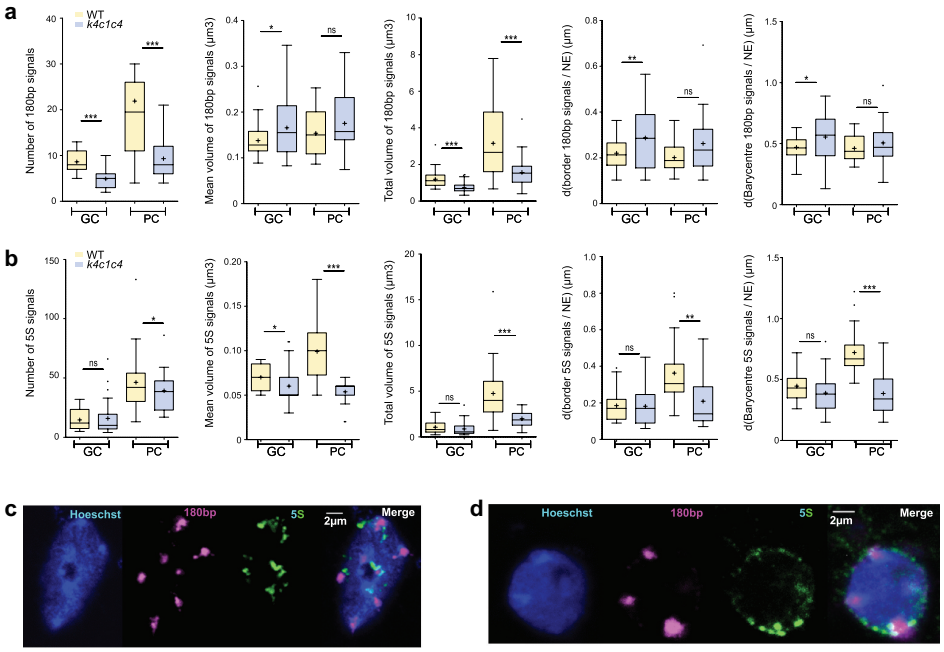


Figure 5. Analysis of aspect and position of 180pb and 5S rDNA repeats revealed by 3D-DNA FISH using NucleusJ 2.0.

NucleusJ 2.0 parameters applied to A) 180bp signals and B) 5S rDNA. Parameters: number of DNA FISH signals, mean volume of FISH signal (µm^3), total volume of FISH signal (µm^3), distance between FISH signal border and the nuclear envelope (d(FISH signal Border/NE), µm) and distance between FISH signal barycenter and the nuclear envelope (d(Barycenter of FISH signal/NE), µm) (Supplemental tables 7–8). Student t-test P-value: ns >0.01, * <0.01, ** <0.001 and *** <0.0001. C) Typical 3D DNA FISH Z-projection of pavement cell nuclei of WT (n = 65 for 180bp and n = 32 for 5S) and D) *k4c1c4* mutant (n = 95 for 180bp and n = 48 for 5S) (Supplemental tables 7–8). From left to right: Hoechst (DNA, blue), 5 TYE 563 LNA probe (180bp, purple), CY5 PCR probe (5S, green) and merge. Scale Bar 2 µm.

defining the most relevant threshold for each single nucleus of a wide-field stack is not a simple task as local variations have to be taken into consideration. This method is also sensitive to nuclear indentations and low labeling area at the position of the nucleolus often located close to the nuclear periphery. The gift-wrapping method was shown to be less sensitive to these factors. It has however the drawback of slightly increasing the object volume and it requires more computing resources than the Otsu method. This is one of the reasons to offer a command line version of NucleusJ 2.0 to use more powerful servers at computing centers. Taken together, the two methods are considered complementary to each other and present two possible alternatives to generate segmented datasets.

The autocrop process will contribute to middle throughput 3D image analysis. It automatically produces an average of 21 isolated 3D nuclei in about 6 minutes reducing substantially the time-intensive manual process. Starting from these large populations of nuclei, NucleusJ 2.0 then describes automatically the 3D nuclear morphology.

The second step in NucleusJ 2.0 was to describe chromatin organization and a triple *kaku4-2 crwn 1–2 crwn 4–1* mutant was used to illustrate the potential of our new software. In the triple mutant some nuclear morphology parameters are altered: guard cells show reduced nuclear volume, while pavement cells show an increased surface area and both types of nuclei revealed reduced flatness in the mutant background.

We illustrated the use of NucleusJ 2.0 to measure FISH signals in a quantitative manner in order to estimate the volume occupied by the FISH signals. Then, the border or the barycenter of the signal are used to compute the positions of the FISH signal in respect to the nuclear periphery. This offers an original method to quantify chromatin compaction while most softwares often define FISH signals as standardized spots that can only be used to compute distances. Using

this method we observed that the 180bp signals are more dispersed while the 5S rDNA loci seem more condensed in the mutant background and preferentially locate at the nuclear periphery. All these parameters were detected successfully by the new software update and are consistent with previous observations [8].

Segmentation of nuclear domains remains challenging to characterize nuclear architecture and NucleusJ 2.0 can provide a solution. In future, further optimization of nuclear domain detection and quantification are required as this still relies on a 3D watershed process, which is time-consuming and user-dependent. Recent progress in Artificial Intelligence is opening up new opportunities to improve 3D bio-imaging. Convolutional Neural Networks (CNN) such as those developed for U-Net, Ilastik, or Star dist [45–47] have been successfully applied for 3D bio-imaging. For the 3D watershed process, preliminary tests using a U-Net CNN [46] currently under development are promising. We expect that the present study, which provides several well-annotated training datasets, will help to pave the way for the development of CNN methods, which we expect will transform the way we acquire and analyze 3D images creating an automated high-performance tool well designed for big-data analysis. Here it is important to highlight the need for large, high-quality datasets to train new CNNs. When looking at image repositories, large-annotated 3D nuclear datasets, digitized objects or standardized microspheres are not easily accessible to benchmark new software or train new networks. This work supported by the COST-Action CA162121 is a first attempt to provide such datasets to the community through a fully public repository with free access to the data.

Acknowledgments

The work of CG, TD, SD, AVP and CT was supported by CNRS, INSERM, Université Clermont Auvergne (UCA), 16-IDEX-0001 CAP 20-25 challenge 1, Temporary Assignment to the CNRS (CT), Pack Ambition Recherche project *Noyau-HD* from the Region Auvergne Rhone Alpes and the COST-Action INDEPTH (CA16212). All pictures were acquired at the CLIC microscopy facility (CLermont Imagerie Confocale). We would like to acknowledge Eric Richards and Kentaro Tamura who made available initial mutant seeds, Rémy Malgouyres for his advices in discrete geometry and Pierre Pouchin for his assistance for using ImageJ and OMERO. Finally, we thank David Evans for critical reading of the manuscript.

Disclosure statement

No potential conflicts of interest were disclosed.

Funding

European Cooperation in Science and Technology [CA16212]; Initiative Science-Innovation-Territoires -Économie [16-IDEX-0001 CAP 20-25 challenge 1]; Région Auvergne-Rhône-Alpes [NOYAU-HD].

ORCID

Tristan Dubos http://orcid.org/0000-0002-4265-2379
Céline Gonthier-Gueret http://orcid.org/0000-0002-4449-559X
Guillaume Mougeot http://orcid.org/0000-0003-3576-7300
Emmanuel Vanrobays http://orcid.org/0000-0002-3209-6743
Sylvie Tutois http://orcid.org/0000-0003-3429-2470
Emilie Pery http://orcid.org/0000-0001-6198-9973
Frédéric Chausse http://orcid.org/0000-0001-7794-1587
Aline V. Probst http://orcid.org/0000-0001-9534-8058
Christophe Tatout http://orcid.org/0000-0001-5215-2338
Sophie Desset http://orcid.org/0000-0002-4897-4977

References

[1] Skinner BM, Johnson EEP. Nuclear morphologies: their diversity and functional relevance. Chromosoma. 2017;126:195–212.
[2] Meier I, Richards EJ, Evans DE. Cell biology of the plant nucleus. Annu Rev Plant Biol. 2017;68:139–172.
[3] Graumann K, Runions J, Evans DE. Characterization of SUN-domain proteins at the higher plant nuclear envelope. Plant J. 2010;61:134–144.
[4] Graumann K, Vanrobays E, Tutois S, et al. Characterization of two distinct subfamilies of SUN-domain proteins in Arabidopsis and their interactions with the novel KASH-domain protein AtTIK. J Exp Bot. 2014;65:6499–6512.
[5] Tamura K, Iwabuchi K, Fukao Y, et al. Myosin XI-i links the nuclear membrane to the cytoskeleton to control nuclear movement and shape in Arabidopsis. Curr Biol. 2013;23:1776–1781.
[6] Zhou X, Graumann K, Evans DE, et al. Novel plant SUN–KASH bridges are involved in RanGAP anchoring and nuclear shape determination. J Cell Biol. 2012;196:203–211.

[7] Dittmer TA, Stacey NJ, Sugimoto-Shirasu K, et al. LITTLE NUCLEI genes affecting nuclear morphology in Arabidopsis thaliana. Plant Cell. 2007;19:2793–2803.
[8] Wang H, Dittmer TA, Richards EJ. Arabidopsis CROWDED NUCLEI (CRWN) proteins are required for nuclear size control and heterochromatin organization. BMC Plant Biol. 2013;13:200.
[9] Choi J, Strickler SR, Richards EJ. Loss of CRWN nuclear proteins induces cell death and salicylic acid defense signaling. Plant Physiol. 2019;179:1315–1329.
[10] Guo T, Mao X, Zhang H, et al. Lamin-like proteins negatively regulate plant immunity through NAC WITH TRANSMEMBRANE MOTIF1-LIKE9 and NONEXPRESSOR OF PR GENES1 in Arabidopsis thaliana. Mol Plant. 2017;10:1334–1348.
[11] Goto C, Tamura K, Fukao Y, et al. The novel nuclear envelope protein KAKU4 modulates nuclear morphology in Arabidopsis. Plant Cell. 2014;26:2143–2155.
[12] Pawar V, Poulet A, Détourné G, et al. A novel family of plant nuclear envelope-associated proteins. J Exp Bot. 2016;67:5699–5710.
[13] Tatout C, Evans DE, Vanrobays E, et al. The plant LINC complex at the nuclear envelope. Chromosome Res. 2014;22:241–252.
[14] Bi X, Cheng Y-J, Hu B, et al. Nonrandom domain organization of the Arabidopsis genome at the nuclear periphery. Genome Res. 2017;27:1–12.
[15] Grob S, Schmid MW, Grossniklaus U. Hi-C analysis in arabidopsis identifies the KNOT, a structure with similarities to the flamenco locus of drosophila. Mol Cell. 2014;55:678–693.
[16] Hu B, Wang N, Bi X, et al. Plant lamin-like proteins mediate chromatin tethering at the nuclear periphery. Genome Biol. 2019;20:87.
[17] Eliceiri KW, Berthold MR, Goldberg IG, et al. Biological imaging software tools. Nat Methods. 2012;9:697–710.
[18] Dumur T, Duncan S, Graumann K, et al. Probing the 3D architecture of the plant nucleus with microscopy approaches: challenges and solutions. Nucleus. 2019;10:181–212.
[19] Kaur D, Kaur Y Various image segmentation techniques: a review. 2014; 6.
[20] Kemeny S, Tatout C, Salaun G, et al. Spatial organization of chromosome territories in the interphase nucleus of trisomy 21 cells. Chromosoma. 20178;127: 247-259.
[21] Zink D, Fischer AH, Nickerson JA. Nuclear structure in cancer cells. Nat Rev Cancer. 2004;4:677–687.
[22] Andrey P, Kiêu K, Kress C, et al. Statistical analysis of 3D images detects regular spatial distributions of centromeres and chromocenters in animal and plant nuclei. PLoS Comput Biol. 2010;6:e1000853.
[23] Poulet A, Arganda-Carreras I, Legland D, et al. NucleusJ: an ImageJ plugin for quantifying 3D images of interphase nuclei. Bioinformatics. 2015;31:1144–1146.
[24] Xing F, Yang L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. IEEE Rev Biomed Eng. 2016;9:234–263.
[25] Iannuccelli E, Mompart F, Gellin J, et al. NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-FISH experiments. Bioinformatics. 2010;26:696–697.
[26] Ollion J, Cochennec J, Loll F, et al. TANGO: a generic tool for high-throughput 3D image analysis for studying nuclear organization. Bioinformatics. 2013;29:1840–1841.
[27] Chytilova E, Macas J, Sliwinska E, et al. Nuclear dynamics in Arabidopsis thaliana. Mol Biol Cell. 2000;11:2733–2741.
[28] Benoit M, Simon L, Desset S, et al. Replication-coupled histone H3.1 deposition determines nucleosome composition and heterochromatin dynamics during Arabidopsis seedling development. New Phytol. 2019;221:385–398.
[29] Poulet A, Duc C, Voisin M, et al. The LINC complex contributes to heterochromatin organisation and transcriptional gene silencing in plants. J Cell Sci. 2017;130:590–601.
[30] Friedland G, Jantz K, Rojas R SIOX: simple interactive object extraction in still images. Seventh IEEE International Symposium on Multimedia (ISM'05), Irvine (CA). 2005. p. 253–259.
[31] Desset S, Poulet A, Tatout C. Quantitative 3D analysis of nuclear morphology and heterochromatin organization from whole-mount plant tissue using nucleus. J Methods Mol Biol. 2018;1675:615–632.
[32] Bauwens S, Katsanis K, Van Montagu M, et al. Procedure for whole mount fluorescence in situ hybridization of interphase nuclei on Arabidopsis thaliana. Plant J. 1994;6:123–131.
[33] Campell BR, Song Y, Posch TE, et al. Sequence and organization of 5S ribosomal RNA-encoding genes of Arabidopsis thaliana. Gene. 1992;112:225–228.
[34] Simon L, Probst AV. High-affinity LNA–DNA mixmer probes for detection of chromosome-specific polymorphisms of 5S rDNA repeats in arabidopsis thaliana. In: Bemer M, Baroux C, editors. Plant Chromatin Dynamics. New York: Springer New York; 2018. p. 481–491.
[35] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979;9:62–66.
[36] Legland D, Arganda-Carreras I, Andrey P. MorphoLibJ: integrated library and plugins for mathematical morphology with ImageJ. Bioinformatics. 2016;32: 3532–3534. Oxford Academic.
[37] Jarvis RA. On the identification of the convex hull of a finite set of points in the plane. Inf Process Lett. 1973;2:18–21.

[38] Fourey S, Magouyres R. Normals estimation for digital surfaces based on convolutions. Comput Graphics. 2009;33:2–10.

[39] The R Core Team. R: A language and environment for statistical computing. r foundation for statistical computing. 2015;Version 3.2.3.

[40] Wickham H. ggplot2: elegant graphics for data analysis [internet]. New York:Springer-Verlag. 2009. cited 2020 Aug 3. Available from: https://www.springer.com/gp/book/9780387981413

[41] Lê S, Josse J, Husson F, others. FactoMineR: an R package for multivariate analysis. J Stat Softw. 2008;25:1–18.

[42] Schindelin J, Rueden CT, Hiner MC, et al. The ImageJ ecosystem: an open platform for biomedical image analysis. Mol Reprod Dev. 2015;82:518–529.

[43] Schneider CA, Rasband WS, Eliceiri KW. NIH image to imageJ: 25 years of image analysis. Nat Methods. 2012;9:671.

[44] Parry G, Aline PV, Baroux C, et al. Meeting report - INDEPTH kick-off meeting. Journal of cell science. 2018;131.

[45] Berg S, Kutra D, Kroeger T, et al. ilastik: interactive machine learning for (bio)image analysis. Nat Methods. 2019;16:1226–1232, Nature Publishing Group

[46] Çiçek Ö, Abdulkadir A, Lienkamp SS, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu M, et al., editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture notes in computer science. Vol. 9901. Cham: Springer; 2016. p. 424–432.

[47] Schmidt U, Weigert M, Broaddus C, et al. Cell detection with star-convex polygons. arXiv:180603535 [cs]. 2018;11071:265–273.

2.7 Résultats annexes et devenir de l'outil

Cette partie décrit les développements annexes dédiés à la seule automatisation du plugin. En effet, le plugin initial a été modifié pour améliorer son fonctionnement, faciliter son utilisation et son partage, et préciser certains calculs de paramètres. Les modifications sont expliquées dans la publication mais j'ai souhaité revenir sur notre démarche et nos motivations. L'ensemble des approches informatiques mise en place dans le développement du logiciel et sa maintenance seront davantage détaillés ainsi que les diverses améliorations ajoutées depuis l'apparition de l'outil afin de maintenir son utilisation par la communauté.

2.7.1 Outil en ligne de commande et intégration continue

Le contexte d'analyse de volume important de données a orienté le développement de NucleusJ2.0 vers la possibilité de l'utiliser en ligne de commande (CLI). La méthode de calcul utilisée, notamment le *gift wrapping*, fait apparaître des limites dans l'utilisation du logiciel car elle nécessite des ressources informatiques importantes (plusieurs Go de mémoire vive) et de nombreux processeurs. La problématique majeure se résume donc à pouvoir analyser des quantités de données conséquentes (plus de 10 Go d'images) dans des temps acceptables (1 minute par noyau). Dans ce cadre, les serveurs représentent des solutions adaptées permettant d'analyser les données rapidement mais demandent des adaptations pour permettre l'appel du logiciel en ligne de commande. NucleusJ2.0 intègre donc cette fonctionnalité pour qu'il soit utilisable sur des structures informatiques de type serveur et adapté au calcul d'un grand volume de données. Enfin, nous proposons l'utilisation d'un fichier de configuration pour la modification des paramètres d'analyse et la traçabilité des résultats.

En parallèle, sur un autre aspect technique, il m'est apparu nécessaire de mettre en place une méthodologie informatique visant à maintenir facilement NucleusJ2.0 tout en permettant des développements ou des optimisations ultérieures. C'est pourquoi, j'ai mis en place avec l'aide de Rémi Martin (stagiaire IUT $2^{ème}$ année, été 2020) l'intégration continue de l'outil via la plateforme gitlab https://docs.gitlab.com/ee/ci/ Cette méthode consiste à valider automatiquement l'ensemble des nouvelles modifications ajoutées au logiciel et donc de vérifier l'intégrité des améliorations apportées. Ce processus est d'autant plus utile que l'outil repose sur de nombreuses dépendances maintenues par des communautés extérieures (exemples : MorpholibJ [Legland et al., 2016] ou Bio-formats). Il est donc très difficile de suivre l'ensemble des nouvelles versions des dépendances et de mesurer l'impact des modifications sur l'intégrité de notre logiciel.

2.7.2 Connexion à la base de données OMERO

L'analyse de la morphologie nucléaire dans l'équipe est faite en routine, permettant la caractérisation rapide de génotypes mutants. La génération massive d'images 3D liée à la multiplication des expériences par l'équipe, nous a amenés avec Sophie Desset, à réfléchir à un meilleur système de gestion des données en termes d'organisation de stockage. Au début de ma thèse, nous avons donc décidé d'une nomenclature pour le nom de nos fichiers, afin de pouvoir identifier et trier nos images dans le futur. Les noms des images contiennent donc un identifiant unique, basé sur l'heure de l'acquisition du fichier (récupéré automatiquement dans les métadonnées), ainsi que les informations des marqueurs fluorescents utilisés, l'organisme, le tissu, ou encore le ou les gènes mutés. Ce travail est une plus-value pour l'intégration automatique de nos données dans un gestionnaire de base de données et un gain de temps pour les méta-analyses futures (par exemple l'analyse d'images de noyaux avec un phénotype commun ou provenant d'un même génotype mutant, d'un même tissu etc). La table présente en Figure 2.3 résume les 10 informations contenues dans le nom de nos images.

	FEATURE	FORMAT	CONTENT	EXAMPLE
1	Date	YYYYMMJJ	date of the slide production	20190709
2	ID	10 numbers	ID produce at the time of image production, metadata of the image file (ImageJ plugin available)	1562746680.71
3	GENUS/species	Family (1 letter) Species (2 letters)	Genus-species	Ath
4	Genotype	ex: SUN1-wt, SUN1-1, genes are separated by --	Gene(s) , allele(s). TAIR accession number (AT4G41310) for a gene, SALK accession number for the allele...	Col0-HON4-wt
5	Tissue	3 letter code	Tissue (cotyledon, leaf, ...), type of preparation (isolated nuclei, squash, ...) whole mount is default state	Cot
6	AgeStage	letter/number code	Age of the seedling (J13) or developmental stage (leaf...)	D13
7	GrowthCondition	Standard= STD. A code should be created for any other condition	Describe standard clture conditions : in vitro ATG medium+sucrose, 23°C 16h day 8h night ...	STD
8	TypeExp	LIVE, FIXE, DFISH	techniques such as FISH,immuno or DAPI/Hoechst staining, Live cell or fixed tissue	FIXE-DIC
9	MolSpecFluo1,2,3	Target-Technique-fluo, in the order of used chanels, separeted by --	Probe (ex 180 pb), method of labelling (PCR, NT,...), fluorochrome	H258
10	Plant/Field	1 letter for the Plant number; 1 number for the acquisition field	Important for further analysis of the dataset	L1

FIGURE 2.3: **Table de la nomenclature du nom des fichiers** : Dans l'exemple le nom d'image sera : 20190709_1562746680.71_Ath_Col0-HON4-wt_Cot_D13_STD_FIXE-DIC_H258_L1 .

Les données associées aux expériences qui ont permis de valider l'outil NucleusJ2.0 ont amené un besoin de structurer le stockage des fichiers bruts et de résultats, afin de les publier et de les rendre accessibles à la communauté scientifique. Notre choix s'est porté sur le logiciel de base de données open source OMERO

[Goldberg et al., 2005]. Notre choix pour cette plateforme s'explique par les raisons suivantes :

- Le GReD, et plus précisément notre informaticien Pierre Pouchin, a mis en place il y a 10 ans OMERO. Aujourd'hui, le GReD dispose d'environ 40 To d'images provenant d'une dizaine d'équipes. Nous avons pleinement bénéficié des compétences de Pierre pour cette partie du travail.
- La plateforme OMERO propose un système facile dans son installation et sa maintenance informatique. Une équipe de plusieurs informaticiens, basée en Ecosse, se consacre au maintien et au développement/évolution du logiciel depuis plus de 15 ans (publication initiale datant de 2005), ce qui rend l'outil pérenne. La structuration sous forme de serveur permet une maintenance facile mais aussi une accessibilité aux données à distance. Enfin, un système d'encapsulation (ou *wrapper*), permet l'interopérabilité d'un ensemble de langages informatiques sur la plateforme, ce qui facilite la réutilisation de fonctionnalité/pipeline développé par d'autre intervenants de la communauté.
- Les fonctionnalités basiques proposées par OMERO répondent bien au besoin de management de données à l'échelle d'une communauté scientifique. En effet, il est facile de partager, visionner ou encore annoter des images après leur intégration dans OMERO. Cette structuration favorise donc le contexte de méta-analyse en simplifiant le groupement d'images ayant une annotation commune.

Actuellement, nous disposons de 3 instances OMERO :

- un serveur local OMERO-GReD
- un serveur distant hébergé par la plateforme de Bio-Informatique du centre de calcul Clermontois (AuBi) OMERO-AuBi
- un serveur OMERO-FSU basé à Florida State University.

Les deux premières instances sont administrées par Pierre Pouchin et non ouvertes vers l'extérieur. La troisième est complètement ouverte au public et a été créée dans le cadre d'un projet Européen COST-Action coordonné par notre équipe (https://www.brookes.ac.uk/indepth/). Dans un futur proche, OMERO-GReD sera transféré vers OMERO-AuBi qui dispose de capacités de stockage et de calcul très supérieures à celles des serveurs du GReD. Notre stratégie sera donc de stocker l'ensemble de nos images au centre de calcul et de déposer les données publiées à OMERO-FSU.

Au cours de ce travail, nous avons voulu rendre possible l'analyse des images par NucleusJ2.0 directement à partir de la plateforme OMERO, sans avoir besoin de rapatrier les images à traiter sur une machine locale. Durant la mise en place de ce processus, nous avons constaté l'absence de certaines fonctionnalités permettant

la communication entre OMERO et ImageJ, nous avons décidé de développer une nouvelle bibliothèque. Avec l'aide de Pierre Pouchin et un élève ingénieur de $2^{ème}$ de année d'école informatique, Rémi Valarcher (semestre1 2020), une liste de méthodes facilitant l'intégration des outils ImageJ en langage Java sur la plateforme OMERO a été construite. Cette bibliothèque a été nommée Simple Omero Client (SOC) (https://github.com/GReD-Clermont/simple-omero-client. Les tests et la mise en pratique des fonctionnalités de la bibliothèque ont été faites au travers de l'intégration de NucleusJ2.0 (Figure 2.4) sur une instance OMERO. Actuellement, il est possible de lancer les analyses de NucleusJ2.0 sur le serveur OMERO-AuBi qui récupère directement les données dans la base de données hébergée sur la plateforme AuBi. Après analyse, les résultats générés par NucleusJ2.0 sont retournés et stockés automatiquement dans la base de données OMERO.

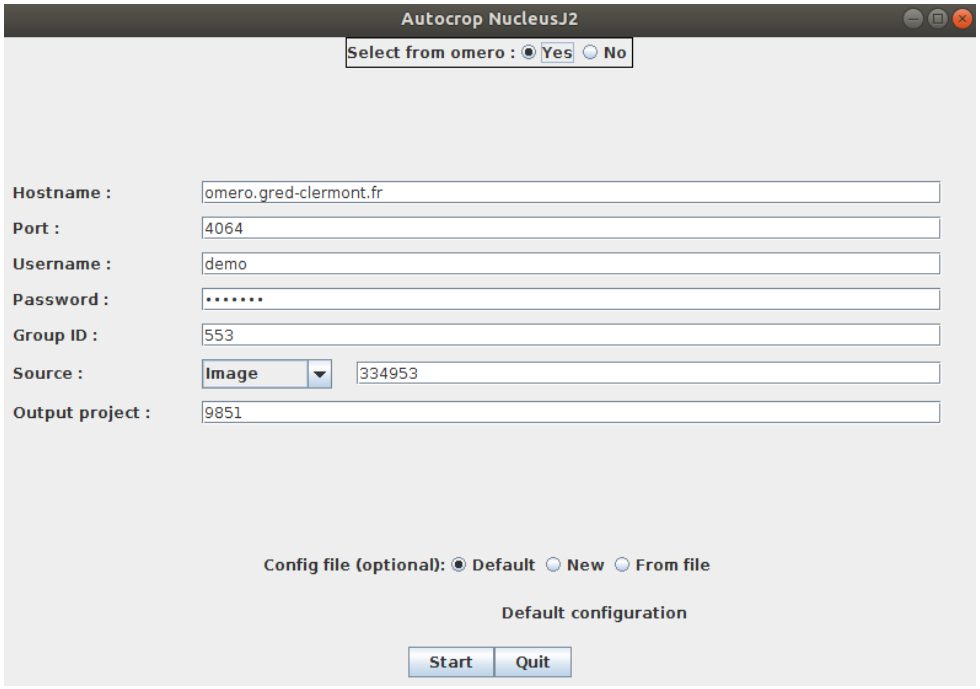


FIGURE 2.4: **Intégration de SOC dans NucleusJ2.0** : capture de la boîte de dialogue de l'autocrop de NucleusJ2.0 permettant la communication avec la plateforme OMERO. La boîte de dialogue donne accès aux champs de configuration du serveur de stockage ainsi qu'aux champs déterminant les paramètres d'analyses avec en entrée l'identifiant de l'image et en sortie l'identifiant du répertoire (output project) où sont stockés les résultats dans OMERO.

2.7.3 Optimisation des méthodes de calcul

Comme nous l'avons vu précédemment dans la section 2.1, l'implémentation d'une méthode de calcul d'enveloppe convexe appelée *gift wrapping* [Jarvis, 1973] répond à la problématique de segmentation du masque du noyau. Cet algorithme initialement décrit sur des images 2D a été adapté pour traiter nos images 3D

[Dubos et al., 2020]. Les modifications pour prendre en charge l'analyse d'images 3D ont pour conséquences des temps de calcul corrélés à la taille de l'objet segmenté (Figure 2.5). Par exemple à partir d'un volume de 100 µm^3, le temps passé pour segmenter le noyau est de l'ordre de 200 minutes. Ce temps de traitement n'est pas acceptable pour des utilisateurs ne disposant pas de ressources de calculs dédiées. Il faut en totalité 57 heures pour traiter les 100 noyaux sélectionnés dans les données présentées (Figure 2.5).
Une première solution pour réduire ce temps d'analyse a été de lancer NucleusJ2.0 en ligne commande (CLI). Cette fonctionnalité rend possible l'utilisation des ressources de d'OMERO-AuBi et la possibilité d'accélérer les calculs en les parallélisant grâce à la fonctionnalité parallel de bash. Cette démarche nécessite cependant des connaissances pour l'utilisateur du langage en CLI et restreint son utilisation aux environnements Unix. Toutefois, l'utilisation du CLI se démocratise de plus en plus. Il est de plus en plus facile d'installer des machines virtuelles LINUX sous l'OS windows, et l'OS Mac qui dérive d'UNIX permet facilement de passer en lignes de commandes.

Nous avons également envisagé plusieurs autres solutions pour accélérer les calculs. Tout d'abord, une option a été introduite afin de modifier la distance (D) dans l'exploration des points lors de l'extension de l'enveloppe convexe (Section 2.1 algorithme 1). Initialement, ce paramètre qui avait été défini de manière arbitraire, se basait sur le diamètre de l'objet à segmenter issu du calcul de Otsu modifié (NJ1). Il existait donc une adaptation de cette constante en fonction du diamètre du noyau à segmenter, ce qui était un atout lors de la segmentation de données hétérogènes. Lors du calcul de l'enveloppe convexe, l'algorithme parcourt la totalité des points bordant l'objet obtenu après segmentation par le seuillage Otsu modifié. Afin d'optimiser le temps de calcul, le nombre de points parcourus a été réduit afin de ne conserver que les points extrêmes dans chaque plan. Cette modification a été introduite avec l'aide d'Alexandre Rongier (stagiaire IUT d'informatique, été 2021). Le *gift wrapping* est calculé à partir de la segmentation Otsu modifié, il est dépendant du nombre de points/sommets soit une complexité de O(n^2) avec n représentant le nombre de points. Nous avons donc débuté les tests pour l'accélération du calcul en intégrant une bibliothèque de calcul de l'enveloppe convexe (QuickHull) implémenté en langage C et disponible en java à l'adresse https://www.cs.ubc.ca/~lloyd/java/quickhull3d.html. Une des méthodes présente dans cette librairie permet la simplification de l'objet segmenté en ne gardant que les sommets du polygone sur chaque plan 2D. Cette approche permet de réduire de 5 fois le temps calcul en diminuant le nombre de points explorés. Après implémentation, nous avons testé cet algorithme sur un noyau de Col-0 du jeu de données de la publication de NucleusJ2.0 pour lequel la segmentation avec le *gift wrapping* demande 15 minutes et qui passe à 3 minutes

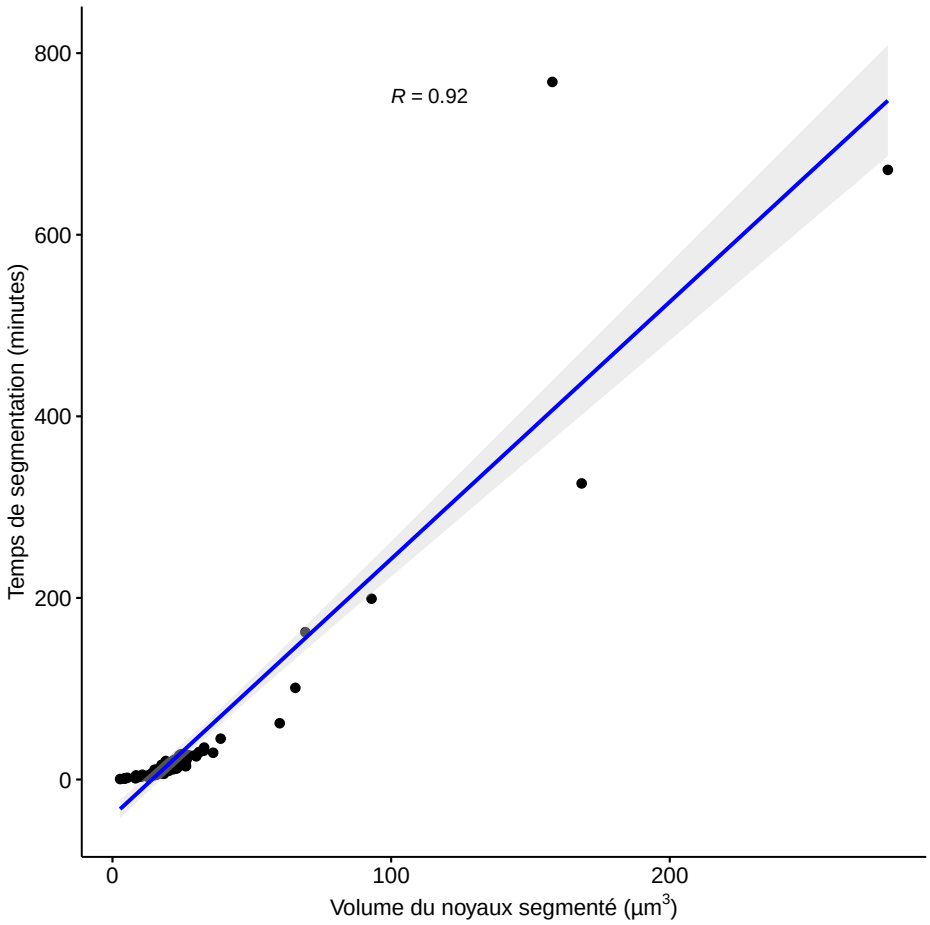


FIGURE 2.5: **Nuage de points représentant 100 noyaux segmentés issus du jeu de données Col-0** [Dubos et al., 2020]. L'axe des ordonnées représente le temps de calcul en minutes du *gift wrapping* pour un noyau donné et en abscisse son volume. Une corrélation de 0,92 entre le temps de segmentation et le volume de l'objet segmenté est observée.

lorsque la bibliothèque QuickHull est utilisée.

En parallèle de ces expérimentations sur l'amélioration du temps de calcul de la méthode du *gift wrapping*, nous avons exploré un autre algorithme de calcul d'enveloppe convexe : le parcours de Graham [Graham, 1972]. Ce dernier, dont la complexité de calcul est de $O(nlog(n))$ où n représente le nombre de points parcourus, fait que cet algorithme est théoriquement moins gourmand en temps de calcul. Nous avons utilisé un code implémenté en java du parcours de Graham qui est disponible et rendu open source par Bart Kiers sur son répertoire github : https://github.com/bkiers/GrahamScan. Pour tester l'implémentation, nous avons donc analysé les 100 noyaux de Col-0 sélectionnés pour leur volume (Figure 2.5). La différence de temps de calcul est drastique puisque dans ce cas elle ne dure que 4 minutes contre 57 heures pour le *gift wrapping*. Afin de vérifier nos résultats, nous avons analysé 680 noyaux Col-0 de la publication de NucleusJ2.0 et comparé

les deux méthodes (*gift wrapping* / parcours de Graham) grâce aux paramètres décrivant la morphologie nucléaire en sortie de NucleusJ2.0. La méthode de Graham segmente globalement des objets plus petits et ce de manière constante comme décrit par l'allure normale de l'histogramme des ratio de volumes (Figure 2.6 A). Ainsi pour 108 noyaux (15% des 680 noyaux), l'estimation du volume diffère de plus de 5% selon la méthode utilisée. Enfin, les analyses en composante principale calculées à partir des paramètres de morphologie nucléaire ne montrent aucune différence entre les deux méthodes (Figure 2.6 B) et C)). Nous arrivons dans les deux cas à distinguer les deux types cellulaires : les cellules de garde (GC) et les cellules de pavement (PC) composant notre échantillon.

L'implémentation de la méthode de Graham répond donc bien à nos besoins d'accélération du calcul de la segmentation de nos objets avec un temps moyen de 2 secondes par noyau. Les différences avec la méthode initiale (Gift wrapping) sont mineures et uniformes au sein des images traitées. Nous pouvons grâce à cette nouvelle méthode atteindre nos objectifs clefs de biologie, visant à discriminer des populations cellulaires à partir des paramètres de morphologie nucléaire. C'est pourquoi, cette nouvelle méthode a été intégrée dans la version 2.0.0 de NucleusJ2.0 le 19/05/2021.

2.8 Bilan

Cette partie de mon travail de doctorat est quantitativement la plus importante. Tout d'abord, il a fallu réorganiser le code de NJ1, réaliser l'intégration des dépendances avec Maven, mettre en place l'utilisation en ligne de commande, implémenter de nouvelles méthodes comme l'*autocrop*, le *gift wrapping* et la méthode de Graham. Il a fallu ensuite organiser le stockage des données dans OMERO et relier OMERO avec ImageJ pour bénéficier de toute la puissance de calcul du serveur OMERO-AuBi. Toutefois, ce travail très technologique autorise aujourd'hui une utilisation plus sereine du plugin dans sa version NucleusJ2.0, et je l'espère, une diffusion plus large de notre outil dans la communauté scientifique dans le futur.

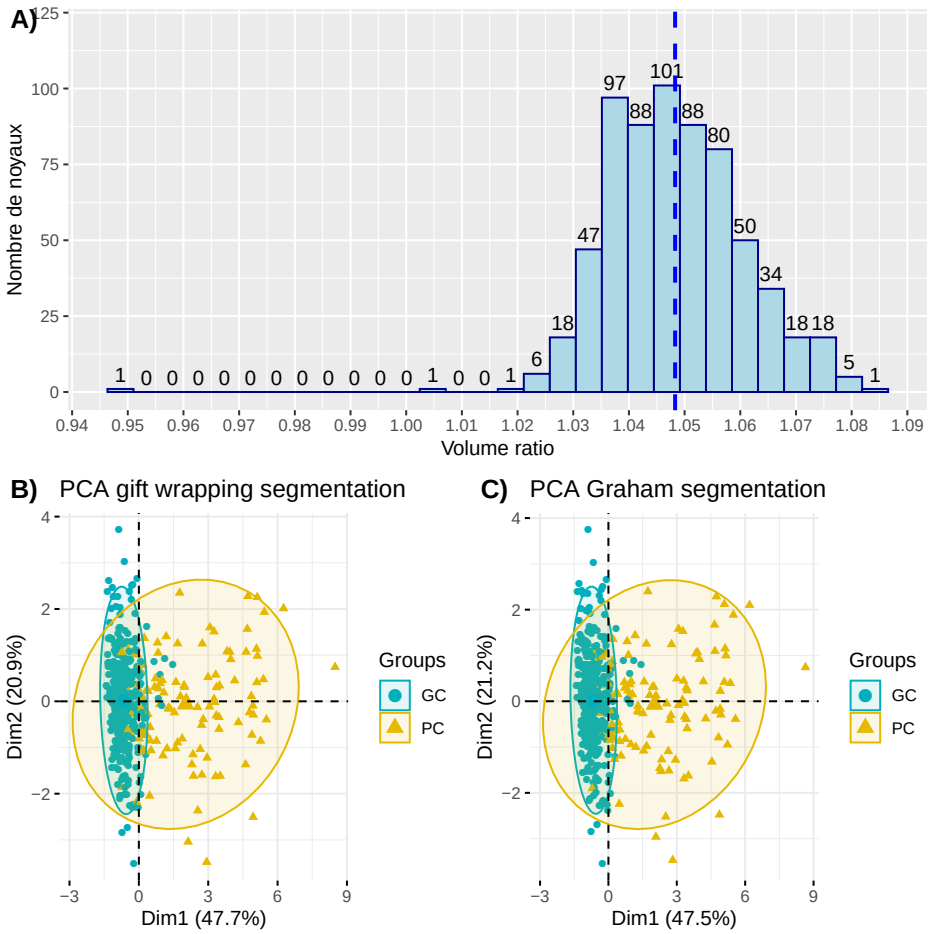


FIGURE 2.6: **Comparaison des résultats de segmentation de la méthode de Graham à partir de 680 noyaux de Col-0** [Dubos et al., 2020]. A) histogramme de distribution des ratios entre les volumes des noyaux segmentés par *gift wrapping* (méthode initiale) et volumes obtenus suite à la segmentation de la méthode de Graham. B) analyse en composante principale des individus générés à partir des paramètres de morphologie nucléaire du *gift wrapping* et C) de la méthode de Graham. Les deux populations cellulaires sont représentées en jaune pour les cellules de pavement (PC) et en vert pour les cellules de garde (GC).

Chapitre 3

Segmentation des compartiments intra-nucléaires : NODeJ

Nous venons de voir que NucleusJ2.0 permet l'automatisation complète du processus d'analyse de la morphologie nucléaire à partir d'images grand champ, et qu'il comporte une nouvelle méthodologie de segmentation qui évite une étape fastidieuse, qui était nécessaire pour écarter les gros noyaux avec des invaginations artéfactuelles. Lors de sa publication en 2020, il restait encore une étape de l'analyse pour laquelle nous n'avions pas trouvé de solution automatisée : la segmentation des chromocentres. Cette segmentation assurée par la méthode de partage des eaux 3D (Watershed 3D) est l'une des originalités de l'outil NJ1, elle permet de décrire l'organisation de la chromatine par la détection d'objets plus intenses (chromocentres) ou d'objets révélés par FISH (Figure 2F [Dubos et al., 2020]). Cependant, la technique de partage des eaux retenue dans la version initiale NJ1, nécessite un seuillage manuel par l'utilisateur très chronophage, et entraîne une reproductibilité peu satisfaisante, puisque le seuil est défini par l'utilisateur. Cette procédure est une limite majeure pour l'analyse de composition en chromatine des noyaux à haut débit.

3.1 Exploration des méthodes disponibles

De nombreuses solutions logicielles sont proposées pour détecter de petites particules dans les images. Il paraissait naturel d'explorer en priorité la solution proposée par la communauté imageJ qui est le plugin ComDet https://imagej.net/plugins/spots-colocalization-comdet. Cependant, les essais sur nos images 3D n'ont pas fonctionné pour les raisons suivantes :

- L'outil prend en charge des images 3D mais détecte les objets plan par plan en 2D.
- Il est nécessaire de fixer les paramètres connus à l'avance sur nos objets tels que l'intensité ou la taille d'objets (Figure 3.1 A).
- Les objets que détectent le plugin demanderaient des adaptations car l'algorithme détecte les objets sur l'ensemble de l'image. Dans l'exemple de résultats Figure 3.1 B on constate que l'on détecte le noyau, une composant du noyau (ne correspondant pas à un chromocentre), et des objets extra-nucléaires (correspondant certainement à du bruit de fond).

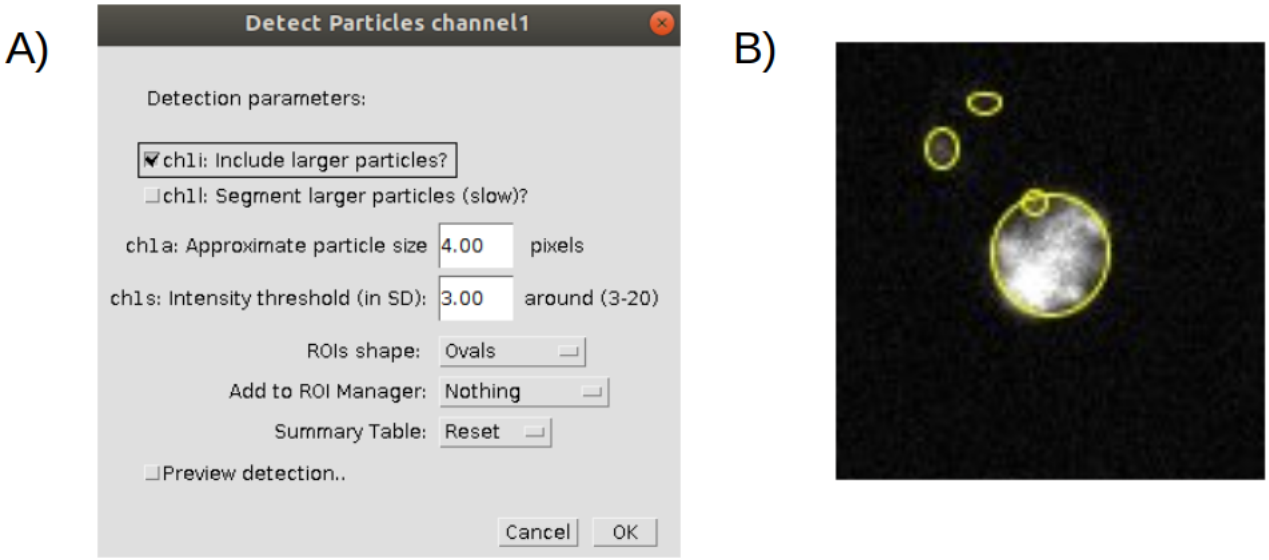


FIGURE 3.1: **Utilisation de ComDet pour la détection des chromocentres** : A la fenêtre de paramétrage du plugin ComDet dans imageJ. B Résultats de l'analyse par ComDet sur une image en 2D d'*A. thaliana*.

Malgré l'existence du plugin imageJ permettant la description de certains objets intranucléaires comme les spots de FiSH ComDet, aucune méthode ni aucun outil spécifiquement dédiés aux chromocentres ne

correspondent à nos attentes. La segmentation des spots de FiSH s'inscrit le plus souvent dans une démarche de dénombrement d'objets détectés (notion de spot), plutôt que dans une caractérisation morphologique/quantitative détaillée. D'autres méthodes provenant d'autres plateformes que imageJ sont disponibles mais leurs implémentations se sont avérées peu convaincantes pour les raisons suivantes :

- La méthodologie utilisée est plus souvent adaptée aux images en 2D et se transpose très mal en 3D pour des raisons tant informatiques que mathématiques comme dans le cas de ComDet (Figure 3.1). De plus, la plupart des outils implémentés pour des images 3D ne fonctionnent pas, comme pour le cas du logiciel fish-quant https://code.google.com/archive/p/fish-quant/downloads qui est signalé comme obsolète. Nous trouvons aussi des codes qui ne sont pas à jour et demandent de nombreuses manipulations pour leur installation ou leur utilisation. Par exemple, dans le cas de FISHcount https://github.com/JIC-CSB/FISHcount nous avons générer un environnement virtuel en utilisant Docker (https://www.docker.com/) pour gérer certaines dépendances.
- Des solutions développées dans d'autres langages de programmation comme spade https://gitlab.inria.fr/ncedilni/spade en *python*, nécessitent des adaptations du code dans l'optique de proposer une solution informatique unique à l'utilisateur au travers de NucleusJ2.0. La multiplication des langages (Java et python) complexifie la maintenance de l'outil à long terme.

Cette exploration montre une problématique malheureusement assez classique dans le domaine de la bioinformatique : les limites de la réutilisation des développements pré-existants. J'ai alors cherché à repartir de différents algorithmes, comme l'utilisation de plusieurs seuils d'Otsu consécutifs (1 pour délimiter le noyau, 1 pour définir le nucléole et 1 pour définir les chromocentres) sur les voxels appartenant au masque du noyau. Cette méthode s'est révélée efficace en 2D mais très décevante en 3D du fait de la dynamique des intensités des voxels composant nos objets. Enfin, la solution est venue d'une collaboration avec Axel Poulet et la mise en place d'une segmentation basée sur le calcul de gradient qui a permis le développement d'un nouveau plugin ImageJ appelé NODeJ soumis pour publication mi-mai dans Bioinformatics.

3.2 Article : NODeJ : an imageJ plugin for 2D and 3D segmentation of nuclear objects

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Notes

OXFORD

Subject Section

NODeJ: an ImageJ plugin for 2D and 3D segmentation of nuclear objects

Tristan Dubos [1,2,†], Axel Poulet [3,†], Emilie Pery [2], Frédéric Chausse [2], Christophe Tatout [1], Sophie Desset [1,*], Josien C. van Wolfswinkel [3,*] and Yannick Jacob [3,*]

[1] GReD, CNRS, INSERM, Université Clermont Auvergne, Clermont-Ferrand, France. [2] Institut Pascal, Université Clermont Auvergne, Clermont-Ferrand, France. [3] Yale University, Department of Molecular, Cellular and Developmental Biology, Faculty of Arts and Sciences, 260 Whitney Avenue, New Haven,CT 06511, USA.

[†] Equal contribution. [*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Here, we present an automated method, Nuclear Object DetectionJ (NODeJ), developed as an imageJ plugin, to segment and analyze high intensity domains in nuclei from 2D or 3D images. NODeJ applies the gradient method on the mask of a nucleus to enhance the contrast of intra-nuclear objects and allow their segmentation.

Availability and implementation: NODeJ is written in Java and provided as an imageJ plugin and a CLI option. Code, documentation and further information can be found at https://gitlab.com/axpoulet/image2danalysis. The images used in this application note are available here: https://www.brookes.ac.uk/indepth/images/ and https://doi.org/10.15454/1HSOIE.

Contact: sophie.desset@udamail.fr, josien.van.wolfswinkel@yale.edu and yannick.jacob@yale.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The nucleus is a dynamic and complex structure that changes morphology and organization of its DNA content during development. The spatial arrangement of the chromatin and other subnuclear structures within the nucleus has fundamental consequences for the accessibility and activity of regions of the genome (Jacob et al., 2010 andBickmore and Van Steensel, 2013). Two-dimensional (2D) or three-dimensional (3D) images are widely used to investigate nucleus morphology and chromatin organization, including NucleusJ2.0 (Dubos et al., 2020) and ilastik (Berg et al., 2019). With these tools, any further segmentation of subnuclear structures relies on a semi-automated procedure that requires user input, leading to limitations in data throughput and potential user bias. In this report, we describe NODeJ as a new tool to automatically segment visible domains in the nucleus. The relevant parameters for the objects detected are then computed using the NucleusJ2.0 method implemented within NODeJ.

2 Methods

NODeJ processes images of nuclei as acquired by live imaging of strains/ecotypes expressing fluorescent reporters, or from fixed tissues or isolated nuclei stained with DNA dyes. The program can be run as an imageJ plugin or via command line. Our method is based on the gradient algorithm for object boundary detection (Gonzalez and Woods, 2008). For each image, NODeJ takes as an input the raw image of the nucleus as well as the mask of this nucleus (assuming one nucleus per image) and computes the image gradient. It then applies a threshold value computed on the image gradient to segment the final object (Fig.1). To derive the gradient (Fig.1C) from the raw image (Fig.1A), the program computes a new value δv_x for each voxel v_x inside the mask of the nucleus (Fig.1B), defined as:

$$\delta v_x = \frac{1}{n} \sum_{neigh=x-n}^{x+n} v_x - v_{neigh}$$

The size of neighborhood n is automatically adjusted to the size of the nucleus (for small nuclei $volume < 50\mu m^3$ $n = 3$ and for large nuclei $n = 7$). If a voxel is outside the mask, the algorithm ignores v_{neigh} and goes to the n+1 v_{neigh}. Once the image gradient is obtained, its signal is

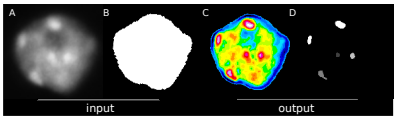


Fig. 1: NODeJ workflow. A. Raw image of a plant nucleus (*A. thaliana*) at interphase stained with DAPI (from Arpòn et al., 2021). B. Image of the segmented nucleus. C. Image of the gradient obtain with NODeJ. High voxel values are shown in red and low values are shown in blue. D. The resulting segmented image, in which each object (connected component) can be analyzed individually.

smoothed using a Gaussian blur filter from imageJ (Schneider et al., 2012), and the threshold value t is computed as $t = \overline{x} + s * f$ with f being a factor defined by the size of the nucleus (1.5 for small nuclei and 2.5 for large nuclei), s being the standard deviation and $\overline{x}$ being the average of the voxel values from the gradient image. Finally, the connected components of the binary image, obtained by applying the threshold t, are defined using the library morpholiBJ (Legland et al., 2016) (Fig.1D).

3 Results

To demonstrate the performance of NODeJ, we processed 1749 nuclei from three different *A. thaliana* data sets (Poulet et al., 2015, Dubos et al., 2020, Arpòn et al., 2021) to identify and analyze characteristic heterochromatin domains, known as chromocenters (Fransz et al., 2002). The data set from Arpòn et al., 2021 contains a heterogeneous population of nuclei from various tissues with different levels of endoreduplication (i.e. ploidy level) and therefore different numbers of chromocenters. Using NODeJ, we obtained a distribution centered between 7 and 10 chromocenters per nucleus (Fig.2A) and most chromocenters had a volume < $1 \mu m^3$ (Fig.2B), which is in agreement with previously published results (Arpòn et al., 2021). Next, we tested whether NODeJ is able to detect known characteristics of the nuclear periphery mutants *crwn* (Wang et al., 2013) and *kaku4 crwn1 crwn4* (Dubos et al., 2020) present in the data sets of Poulet et al., 2015 and Dubos et al., 2020. NODeJ, correctly detected the previously described changes in heterochromatin organization, such as decreased number of chromocenters, increased volume, and increased distance between chromocenters and the nuclear periphery (Fig.2C and D) (Wang et al., 2013, Dubos et al., 2020).

4 Conclusion

NODeJ is a novel tool for automated subsegmentation of nuclear images, and is able to accurately identify heterochromatin domains from a diverse set of *A. thaliana* nuclei. NODeJ was developed as an extension of NucleusJ2.0 and allows for efficient automated analysis of subnuclear structures by eliminating the semi-automated steps inherent to the use of NucleusJ2.0, resulting in reduced processing time and bias in the analysis. This is also valuable for the preparation of training data sets for machine learning. Further, we believe that the utility of this tool can be extended to other data sets such as those obtained by nucleolar labeling or DNA fluorescent in situ hybridization (DNA-FISH).

Funding

This project was supported by grants #R35GM128661 and #R35GM128619 from the National Institutes of Health to Y.J and J.C.vW respectively.

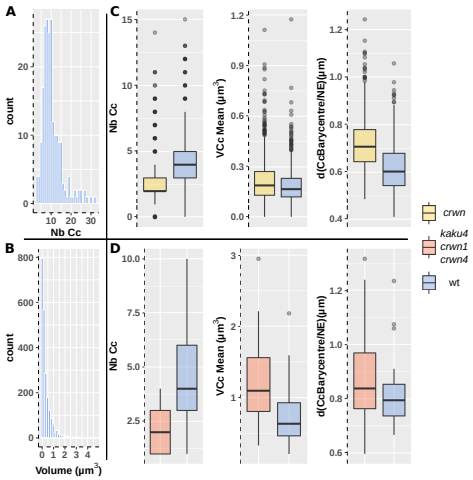


Fig. 2: Analysis of *A. thaliana* data sets with NODeJ. A. and B. Results from Arpòn et al., 2021 describing chromocenters from isolated wild type nuclei (n=212) extracted from whole plants (Supplementary data). The histograms show the repartition of the chromocenter number by nucleus (A) and their volume (B). C. and D. Results from the analysis of two data sets from mutants (*crwn* and *kaku4 crwn1 crwn4* triple mutant) known to alter chromatin organization. C. *crwn* mutants (n=39) and wild type plants (n=38) from Poulet et al., 2015 (Supplementary data). D. *kaku4 crwn1 crwn4* triple mutants (n=851) and wild type plants (n=609) from Dubos et al., 2020 (Supplementary data). Boxplots show: number of chromocenters (Nb Cc), mean chromocenter volume (VCc Mean in μm^3) and distance of chromocenter barycenter to the nuclear envelope (d(CcBarycenter/NE) in μm) (Supplementary tables 1, 2 and 3 describe the computed parameters).

Funding support was also provided by 16-IDEX-0001 CAP 20-25 challenge 1, Pack Ambition Recherche project Noyau-HD from the Region Auvergne Rhone-Alpes and the COST-Action INDEPTH (CA16212) to C.T.

References

Arpòn, J., K. Sakai, V. Gaudin, and P. Andrey
2021. Spatial modeling of biological patterns shows multiscale organization of Arabidopsis thaliana heterochromatin. *Scientific Reports*, 11(1):1–17.

Berg, S., D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, and A. Kreshuk
2019. Ilastik: Interactive Machine Learning for (Bio)Image Analysis. *Nature Methods*, 16(12):1226–1232.

Bickmore, W. A. and B. Van Steensel
2013. Genome architecture: Domain organization of interphase chromosomes. *Cell*, 152(6):1270–1284.

Dubos, T., A. Poulet, C. Gonthier-Gueret, G. Mougeot, E. Vanrobays, Y. Li, S. Tutois, E. Pery, F. Chausse, A. V. Probst, C. Tatout, and S. Desset
2020. Automated 3D bio-imaging analysis of nuclear organization by NucleusJ 2.0. *Nucleus*, 11(1):315–329.

Fransz, P., J. H. De Jong, M. Lysak, M. R. Castiglione, and I. Schubert
2002. Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14584–14589.

Gonzalez, R. C. and R. E. Woods
2008. *CV book*, volume 3rd Editio.

Jacob, Y., H. Stroud, C. Leblanc, S. Feng, L. Zhuo, E. Caro, C. Hassel, C. Gutierrez, S. D. Michaels, and S. E. Jacobsen
2010. Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. *Nature*, 466(7309):987–991.

Legland, D., I. Arganda-Carreras, and P. Andrey
2016. MorphoLibJ: Integrated library and plugins for mathematical morphology with ImageJ. *Bioinformatics*, 32(22):3532–3534.

Poulet, A., I. Arganda-Carreras, D. Legland, A. V. Probst, P. Andrey, and C. Tatout
2015. NucleusJ: An ImageJ plugin for quantifying 3D images of interphase nuclei. *Bioinformatics*, 31(7):1144–1146.

Schneider, C. A., W. S. Rasband, and K. W. Eliceiri
2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7):671–675.

Wang, H., T. A. Dittmer, and E. J. Richards
2013. Arabidopsis CROWDED NUCLEI (CRWN) proteins are required for nuclear size control and heterochromatin organization. *BMC Plant Biology*, 13(1):1–13.

3.3 Bilan

Avec NODeJ, nous avons maintenant la possibilité d'appliquer un *workflow* complet allant de l'*autocrop* qui génère les imagettes individuelles, à la segmentation du noyau qui génère les paramètres de morphologie nucléaire et enfin à l'organisation de la chromatine et ce de façon beaucoup plus automatisée qu'avec NJ1. Dans le futur, il faudra intégrer NODeJ dans NJ2 afin de permettre à l'utilisateur d'appliquer soit une segmentation en appliquant la ligne de partage des eaux (NJ1) soit le gradient descendant (NODeJ).

Chapitre 4

Application de NucleusJ2.0 dans un environnement intégré dans OMERO sur un modèle de spermatozoïdes humains

Comme nous l'avons vu, l'étude de la morphologie des noyaux et de leur contenu est au cœur de nombreux processus biologiques. C'est pourquoi, il nous paraît important de pouvoir partager NucleusJ comme outil d'analyse dans d'autres modèles que les plantes. Mon équipe a déjà établi des collaborations en ce sens sur des modèles animaux, comme celui du développement cardiaque chez la drosophile (Guillaume Junion, GReD), ou la maturation des spermatozoïdes chez la souris (Ayhan Kocer, GReD). Un autre exemple est l'analyse nucléaire qui a permis une publication dans un modèle de trisomie humaine [Kemeny et al., 2018]. NucleusJ a été utilisé pour analyser des images de FiSH montrant que le chromosome 21 surnuméraire provoquait des modifications majeures dans le noyau des lymphocytes, avec un déplacement vers la périphérie de certains territoires chromosomiques (HSA1 et HSA17) et une augmentation de la compaction d'au moins deux chromosomes (HSA1 et HSA17).

4.1 Le modèle du spermatozoïde humain

En 2019, nous avons commencé une nouvelle collaboration avec une équipe du laboratoire IMOST et le service de reproduction du CHU de Clermont-Ferrand dans un projet financé par la Ligue contre le Cancer que porte le Dr Hanaë Pons. L'étude de la morphologie nucléaire est au cœur de ce projet qui vise notamment à examiner les modifications de la structure tridimensionnelle du noyau des spermatozoïdes humains dans un modèle de thérapie par irradiation (IRA-thérapie). Les éventuelles modifications de la chromatine pendant ces processus seraient alors proposées comme un indicateur de leur qualité.

L'IRA-thérapie est un traitement par iode Radioactif (I^{131}) pour soigner le cancer de la thyroïde, la 3^e cause de cancer chez l'homme en âge de procréer. Ce traitement peut altérer la spermatogenèse avec le risque d'une stérilité. Les recommandations préconisent une cryoconservation préventive des gamètes. A ce jour, l'impact de la cryoconservation et de ce type d'irradiation sur la qualité des spermatozoïdes humains est encore mal connu. Afin de mieux quantifier l'impact de l'IRA-thérapie, l'équipe d'Hanaë Pons utilise un modèle d'exposition radioactive de spermatozoïdes humains in vitro mimant celles des gonades dans le cadre d'une irradiation par l'$Iode^{131}$. La morphologie du noyau et l'organisation de la chromatine doivent être étudiées par imagerie 3D en utilisant NucleusJ2.0 en même temps que les paramètres spermatiques standards, les altérations majeures de l'ADN et la taille des télomères par des approches moléculaires et cellulaires.

L'organisation chromatinienne du spermatozoïde fait l'objet d'études nombreuses, notamment avec des expériences de FISH qui ciblent des régions d'hétérochromatine comme les centromères ou les télomères, ou qui visent à localiser les différents territoires chromosomiques. Le modèle d'organisation initialement proposé par [Zalensky et al., 1995] est aujourd'hui contesté grâce aux progrès techniques qui permettent d'étudier la position statistique des télomères ou des centromères de plusieurs chromosomes en même temps et dans un grand nombre de cellules [Ioannou et al., 2017]. L'organisation des centromères regroupés en chromocentres et éloignés de la périphérie, reliés aux télomères se trouvant eux au plus proche de la membrane nucléaire, ne fait plus consensus avec un nouveau modèle d'une organisation davantage distribuée dans le spermatozoïde (Figure 4.1). L'organisation des télomères entre eux est également discutée, mais leur position périphérique semble acquise.

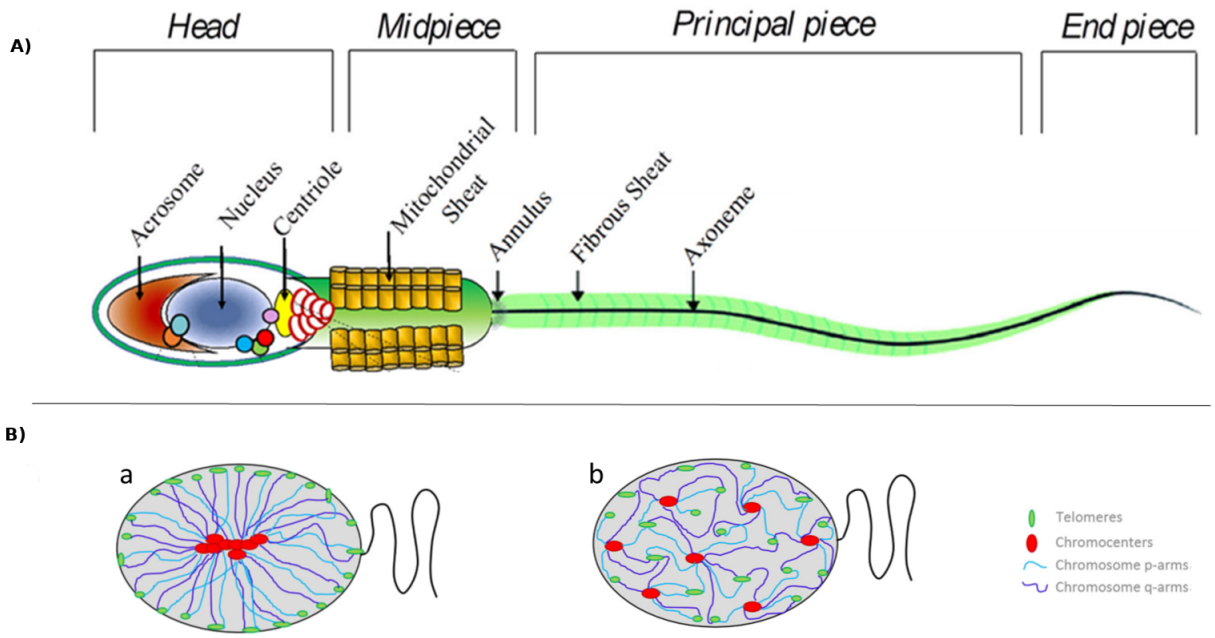


FIGURE 4.1: **Le spermatozoïde et son organisation chromatinienne** : A) Schéma d'un spermatozoïde humain d'après [Manfrevola et al., 2021]. Les petits disques dans le noyau représentent les différents complexes LINC formés à partir de deux protéines KASH différentes et 4 protéines sun. B) Modèles d'organisation de la chromatine dans le spermatozoïde humain (d'après [Ioannou et al., 2017] : (a) modèle « Hairpin loop » dans lequel les centromères (rouge) sont alignés dans le spermatozoïde avec chaque bras chromosomique (bleu) qui s'allonge jusqu'au télomère (vert) en (b) nouveau modèle dit « segmenté » ou les télomères et centromères sont distribués tout le long du spermatozoïde, en périphérie nucléaire où proche de la périphérie, avec des localisations et regroupements préférentiels ou fréquents.

4.2 Production des images

Les échantillons de sperme ont été fixés et colorés par un intercalant avant d'être imagés par notre microscope Leica DM6000 MMAF équipé d'un Optigrid (objectif 63x N.A 1.4). Dans les expériences préliminaires menées en 2019 et 2020, le protocole de coloration (type de fixateur et des d'intercalants) et les conditions d'acquisition d'images (densité spermatique, nombre de champs) ont été définis en collaboration avec une étudiante de master 1, Maissa Andrieux (été 2020), sur la base d'analyses des images que j'ai effectuées avec NucleusJ2.0. Nous avons fait le choix de pouvoir discriminer les spermatozoïdes vivants ou morts selon un protocole établi pour le sperme de bélier [Makarevich et al., 2010]. Les conditions expérimentales retenues après mes analyses sont une coloration rapide des spermatozoïdes morts à l'iodure de propidium (PI), suivie d'une fixation et de la coloration de l'étalement total avec du DAPI. En 2021, un nouveau jeu de données a été produit par une étudiante de master 2, Marine Compagnon, avec 10 donneurs pour savoir si la description des noyaux par notre méthode pouvait discriminer une population de spermatozoïdes frais d'une population congelée. Peu de données sont disponibles dans la littérature sur la structure des noyaux après la congélation.

4.3 Séquence d'analyse par NucleusJ2.0 via la bibliothèque de communication Simple Omero Client

Nous avons expérimenté à cette occasion le fonctionnement de la bibliothèque de communication *Simple Omero Client* (Section 2.7.2) entre l'instance OMERO et notre serveur de calcul pour analyser directement les images sauvegardées dans OMERO. Les résultats obtenus sont directement sauvés après analyse dans la base de donnée OMERO au cours du processus, ce qui permet une traçabilité parfaite entre les jeux de données et les résultats.

L'API OMERO n'est cependant pas très souple et ne permet que deux niveaux de hiérarchie (projet/jeu de données). Ainsi, chacun des 10 donneurs représente un projet distinct et les sous-dossiers correspondent aux différents jeux de données pour ce donneur. Chaque projet contient à la fois les jeux de données d'entrée (images chargées dans OMERO) et les données de sortie (images segmentées générées par NucleusJ2.0). La séquence d'analyse par NucleusJ2.0 pour chaque donneur ainsi que l'organisation des données dans OMERO est décrite dans la Figure 4.2.

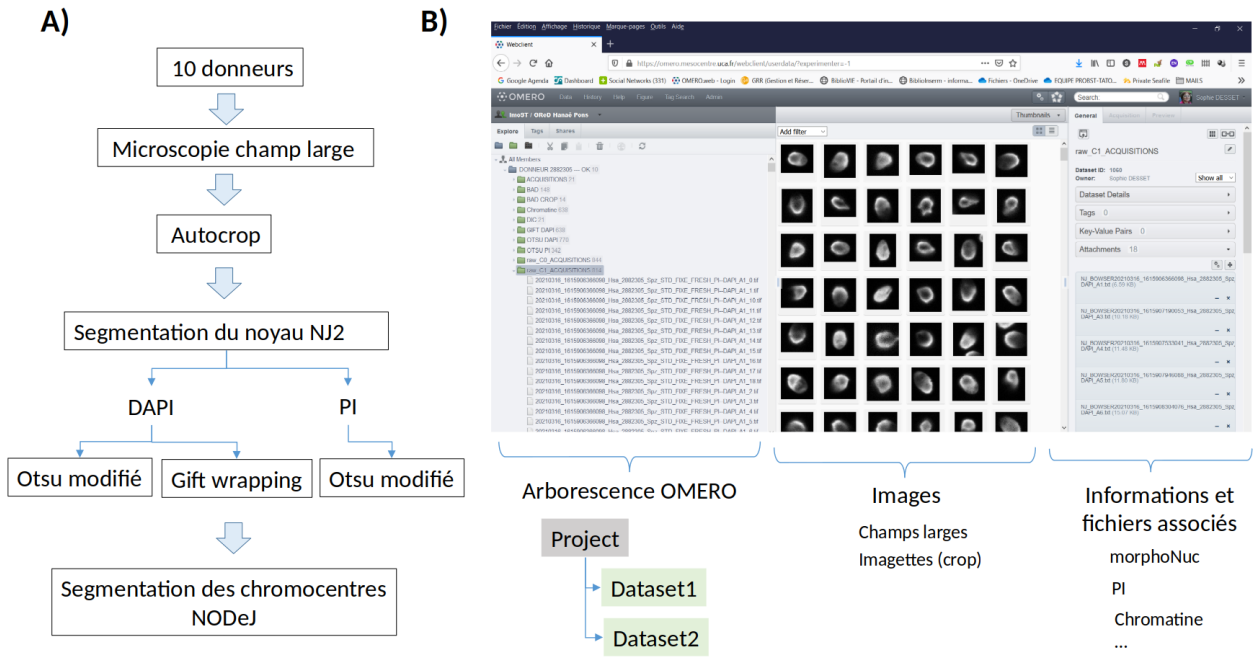


FIGURE 4.2: **Workflow de l'analyse et Organisation des jeux de données dans OMERO**. A) A partir de chaque donneur, une dizaine d'images grands champs sont générées. L'étape d'autocrop est réalisée et suivie par la segmentation par NucleusJ2.0 pour déterminer la morphologie nucléaire (Otsu et *gift wrapping*) et la viabilité des spermatozoïdes (Otsu). Enfin l'organisation de la chromatine est déterminée par NODeJ. B) Organisation des images et des fichiers de résultats dans OMERO. Un « projet » est créé pour chaque donneur, désigné par un numéro correspondant à un dossier en gris. Les sous-dossiers (« datasets » selon la nomenclature OMERO) sont en vert et contiennent les fichiers images (volet de gauche). La fenêtre centrale permet d'avoir une vision globale des images. Les fichiers .csv contenant les calculs de paramètres générés par NucleusJ2.0 ou NODeJ sont directement attachés au sous-dossier correspondant (volet de droite).

Ce projet a permis de tester sur un grand jeu de données les performances de NucleusJ2.0 dans l'environnement OMERO grâce à la bibliothèque *Simple Omero Client*. Cette dernière facilite la gestion des fichiers et des dossiers sans risque d'erreur. La bibliothèque permet d'effectuer les calculs sans mobiliser son poste personnel et sans utiliser de disques durs externes. En revanche, le temps de calcul reste encore relativement long du fait des temps de transfert de fichiers entre serveur de calcul et le serveur d'OMERO (durant l'envoi des images brutes ou les retours de résultats), bien que ceci soit transparent pour l'utilisateur. Ceci devra être optimisé dans le futur.

Les données traitées sont résumées dans le tableau Figure 4.3. Pour chaque donneur les images sont collectées avant (Fresh) et après (Frozen) cryopréservation. NucleusJ2.0 écarte automatiquement les noyaux correctement segmentés par Otsu et pour lesquels le *gift wrapping* échoue ce sont les « Bad Crop ». Cela concerne en moyenne 10% des noyaux correctement segmentés par la méthode Otsu modifié. Il est à noter que la réalisation des expériences (mesure des paramètres biologiques et de la structure tridimensionnelle des noyaux) jusqu'à l'analyse des contenus a été menée entièrement au cours des 4 mois effectifs de stage par une

étudiante novice, tandis qu'au début de ma thèse, trois personnes expérimentées ont été nécessaires à plein temps pendant 4 semaines pour analyser la seule morphologie nucléaire du même nombre de noyaux.

Non du dataset	Acquisition 3D	Raw image C=1	Otsu segmentation	*Gift wrapping*	Chromatine
FRESH	114	3307	2994	2492	2492
FROZEN	115	3212	3033	2707	2707
BAD		1395	669	931	

FIGURE 4.3: Nombre d'images (Acquisition 3D) dans chaque dataset pour 10 donneurs dans l'étude de comparaison entre sperme frais (FRESH) et congelé (FROZEN). BAD désigne les images éliminées de noyaux isolés après autocrop (Raw image C=1) après tri visuel ou au cours du traitement par NucleusJ2.0 (Bad crop, les noyaux segmentés par Otsu modifié qui échouent).

4.4 Critères de tri des noyaux

Après l'étape de l'*autocrop* environ 20% des fichiers individuels de noyaux sont éliminés. Cette estimation s'avère variable d'un donneur à l'autre (de 4 à 37%) et peut refléter le taux d'agrégation des spermatozoïdes, la présence plus ou moins importante de débris cellulaires, la saturation de certaines images au moment de l'acquisition, ou des anomalies (Figure 4.4 A,B,C et D).

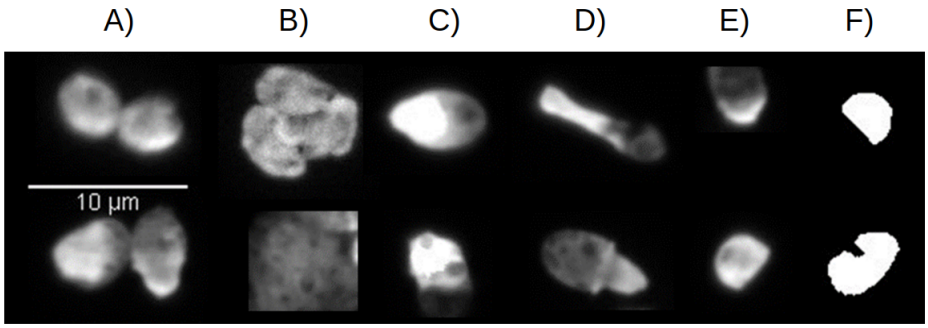


FIGURE 4.4: **Exemple de noyaux colorés au DAPI écartés de l'étude**. A) Spermatozoïdes agrégés. B) Débris cellulaires. C) Images de spermatozoïdes présentant une saturation. D) Spermatozoïdes avec des formes anormales. E) Noyaux incomplets. F) Segmentations incomplètes.

Ce chiffre traduit la difficulté à détecter un noyau de spermatozoïde par la méthode Otsu lors de l'*autocrop* en raison de la très grande variation d'intensité de fluorescence au sein d'un même noyau. Nous observons sur la Figure 4.4 E que certains noyaux incomplet sont détectés au moment de l'*autocrop* et se trouvent coupés au moment du crop. Parfois, la segmentation ultérieure aboutira à un noyau incomplet, comme s'il était coupé au milieu (Figure 4.4 F). Enfin, certaines images de noyaux présentent des plans avec une invagination comme cela avait été observé pour les nucléoles dans les noyaux de plantes mais dans le cas des spermatozoïdes lié à la présence de vacuoles dans le noyau du spermatozoïde. L'existence de ces vacuoles nous a conduis à choisir

d'utiliser la méthode du *gift wrapping* pour la segmentation nucléaire. Ces noyaux (environ 20%) incomplets sont écartés de l'étude après l'étape segmentation avant de réaliser l'analyse par NODeJ.

4.5 Résultats biologiques

4.5.1 Structures chromatiniennes détectées par NODeJ

La Figure 4.5 A est une illustration d'une séquence complète d'analyse par NucleusJ2.0/NODeJ d'une image de spermatozoïde. La segmentation du contenu nculéaire par NODeJ permet de détecter les zones les plus intenses qui s'apparentent à de l'hétérochromatine. Sur l'ensemble des images, nous observons une structure typique en anneau ou en arc, située du côté de la queue du spermatozoïde. Cet anneau est positionné face à l'anneau nucléaire, une structure protéique responsable de la stabilité de l'ADN du spermatozoïde. Cet anneau est localisé à la base du noyau au niveau de la jonction entre la tête du spermatozoïde et le flagelle (Figure 4.1 A). Dans l'article de [Barone JG, De Lara J, Cummings KB, 1994], les auteurs ont démontré que lorsque le noyau spermatique humain se décondense, l'ADN reste ancré à cette structure [Ward and Coffey, 1989]. La structure détectée est parfois interrompue par des zones moins intenses, possiblement décondensées. Parfois des structures plus sphériques sont détectées à chaque extrémité de l'arc le long de la périphérie (Figure 4.5 C). Sans davantage de précision sur la nature de ces différentes régions, nous avons choisi de suivre à la fois le paramètre du volume total d'hétérochromatine détecté par NODeJ (VtCc, paramètre volume total de chromocentre) et d'extraire les paramètres de l'arc qui se trouve être le plus volumineux parmi les zones d'hétérochromatine détectées dans chaque noyau (volume et distances).

4.5.2 Définition des populations de spermatozoïdes morts ou vivants

Dans le cadre d'une étude de population globale par imagerie, il nous a paru important de pouvoir nous assurer de l'état de viabilité des spermatozoïdes observés. En effet, il est possible que l'irradiation n'ait pas le même impact sur la chromatine des spermatozoides vivants ou morts. Afin de distinguer, au moins *a posteriori*, les spermatozoïdes morts des vivants, nous avons choisi d'utiliser un double marquage DAPI (pour identifier les spermatozoïdes) et l'iodure de propidium (statut vivant/mort). Ainsi, deux populations peuvent être observées : l'une, colorée par les deux intercalants, est composée de spermatozoïdes morts (segmentation possible des images par NucleusJ2.0 et forte intensité du propidium (PI)) ; l'autre, colorée seulement par le DAPI, distingue les spermatozoïdes vivants (absence de résultat de segmentation obtenue sur les images du

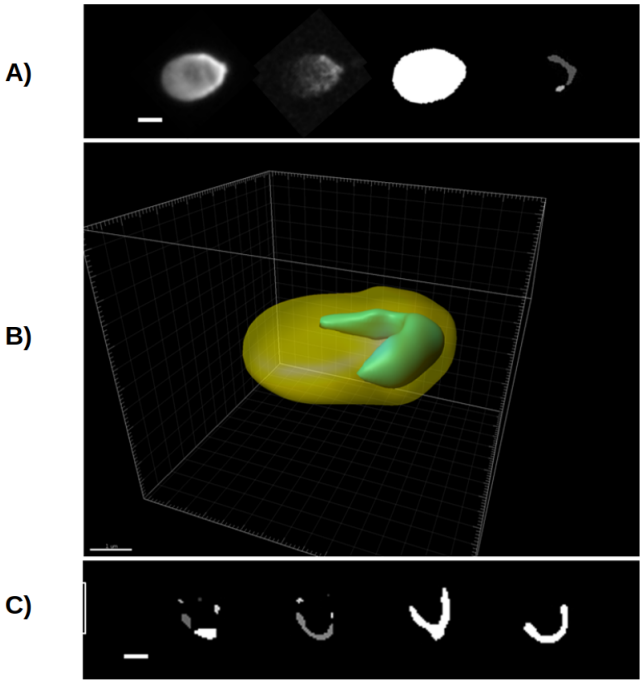


FIGURE 4.5: **Illustration de l'analyse des spermatozoïdes par NucleusJ2.0/NODeJ**. A) Séquence d'analyse à partir d'un spermatozoïde coloré par deux intercalants. Images de Z projections de gauche à droite : DAPI, PI, segmentation du noyau par NucleusJ2, domaine intranucléaire segmenté par NODeJ. B) vue 3D du noyau du même spermatozoïde réalisée avec le logiciel Imaris. En segmentation du noyau (jaune) et le domaine intra-nucléaire (vert). C) Exemples de segmentations obtenues par NODeJ à partir d'un sperme frais du même donneur. Echelles : A et C, 2 μm. B, 1 μm.

canal PI). Toutefois, l'exclusion du PI par les spermatozoïdes vivants est souvent partielle et n'est qu'un reflet de la vitalité. Cet intercalant pénètre plus ou moins dans les cellules aux membranes fragilisées. Il existe donc une population intermédiaire dont le statut n'est pas strictement définissable mais où la pénétration du PI suit un gradient d'intensité. Après la segmentation des images du canal PI par la méthode Otsu, nous avons classé les noyaux selon leur intensité moyenne après marquage au PI. Nous avons fait le choix de sélectionner parmi eux :

- Les 20% de spermatozoïdes les plus intenses pour définir la population des spermatozoïdes morts.
- Pour les vivants, nous avons conservé les images de tous les spermatozoïdes excluant le PI. Lorsque les effectifs étaient inférieurs à 10% (un seul donneur avec un sperme de faible vitalité), nous avons complété cette population avec des spermatozoïdes faiblement colorés au PI pour atteindre 10% de la population totale.

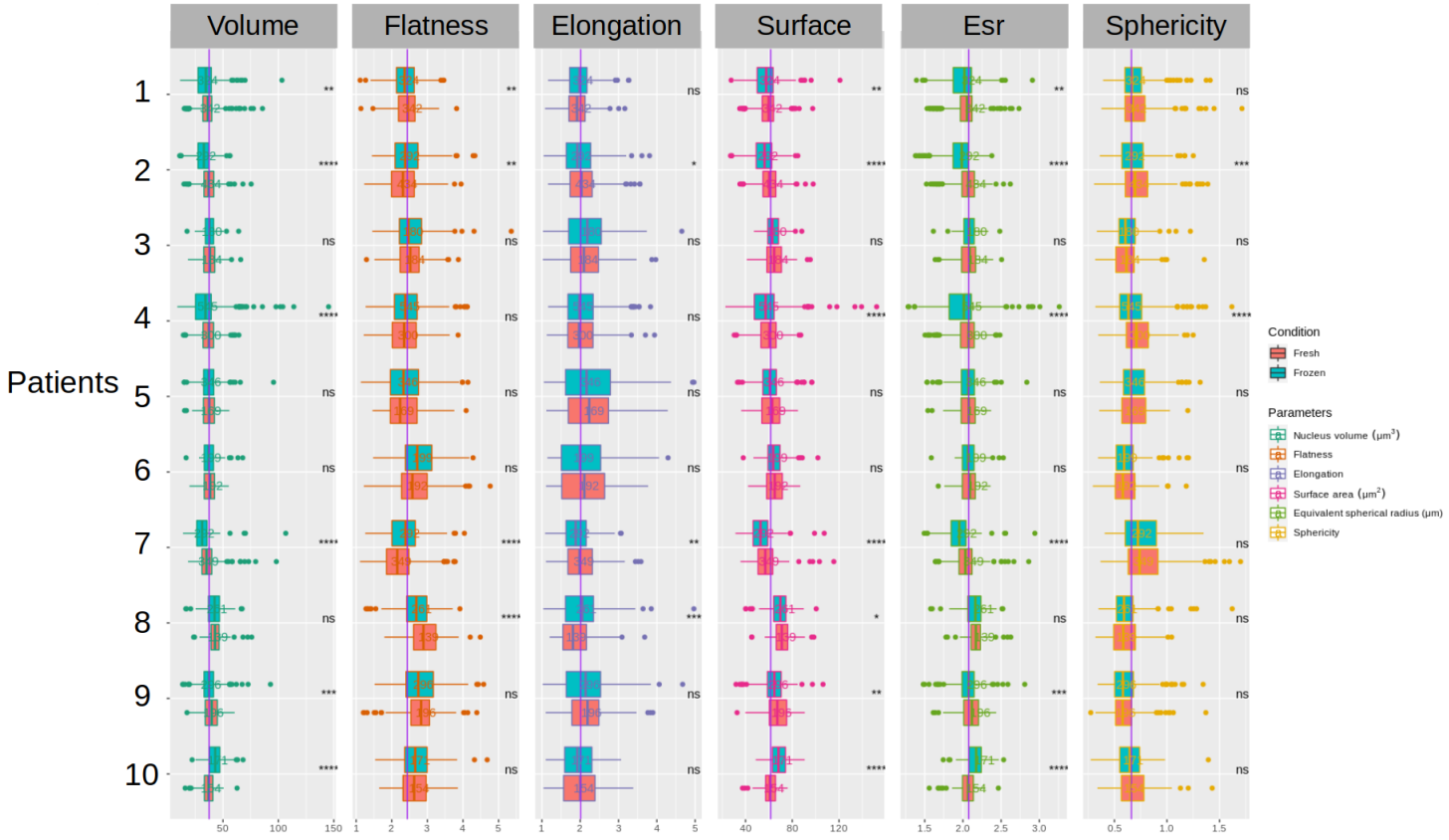


FIGURE 4.6: **Impact de la congélation sur les paramètres spermatiques, morphologiques et de qualités nucléaires**. Variation interindividuelle de l'impact de la congélation pour les différents paramètres de morphologie nucléaire. Les valeurs correspondent aux médianes +/- intervalle de confiance (IC) des paramètres morphologiques nucléaires. SAC : surface de l'aire corrigée ; Esr : équivalent sphérique du rayon.

4.5.3 Comparaison de la structure tridimensionnelle des noyaux de sperme frais ou congelé

L'objectif de cette première étude est la validation d'un nouveau marqueur, la structure tridimensionnelle du noyau spermatique par bio-imagerie. Nous souhaitions déterminer si ce marqueur permettrait d'évaluer la qualité des spermatozoïdes humains. L'organisation 3D du noyau spermatique est donc comparée après congélation-décongélation à l'état frais. Cette analyse a été réalisée en parallèle à l'évaluation de marqueurs nucléaires usuels : fragmentation de l'ADN, décondensation de la chromatine et taille des télomères spermatiques chez 10 patients présentant diverses qualités spermatiques.

Le premier constat est que la morphologie nucléaire semble inchangée entre les conditions « frais » et « congelé ». Il existe une très grande variabilité interindividuelle pour les paramètres de morphologie (Annexe 6.3) mais globalement, en calculant la taille d'effet des populations totales, aucun des paramètres explorés n'est modifié de façon significative (Figure 4.6). Les variations entre donneurs reflètent sans doute la qualité hétérogène des 10 échantillons disponibles dans notre étude.

4.5.4 Contenu en chromatine

L'effet de la congélation-décongélation sur la structure de la chromatine est en revanche beaucoup plus marqué, avec un fort impact sur le volume de la structure condensée (ci-après appelée hétérochromatine-like Figure 4.7 A) observée dans les noyaux qui diminue après congélation/décongélation. De façon remarquable cet effet s'observe chez tous les donneurs (Figure 4.6).

Ces résultats sont cohérents avec la littérature qui constate une décondensation de la chromatine après la congélation [Boitrelle et al., 2012] [Hammadeh et al., 1999]. Sur les images des spermatozoïdes congelés, l'arc de chromatine condensée devient fragmenté et segmenté en plusieurs morceaux par NODeJ, ce qui se traduit par une augmentation du nombre de « chromocentres » détectés (NODeJ étant conçu à l'origine pour détecter des chromocentres).

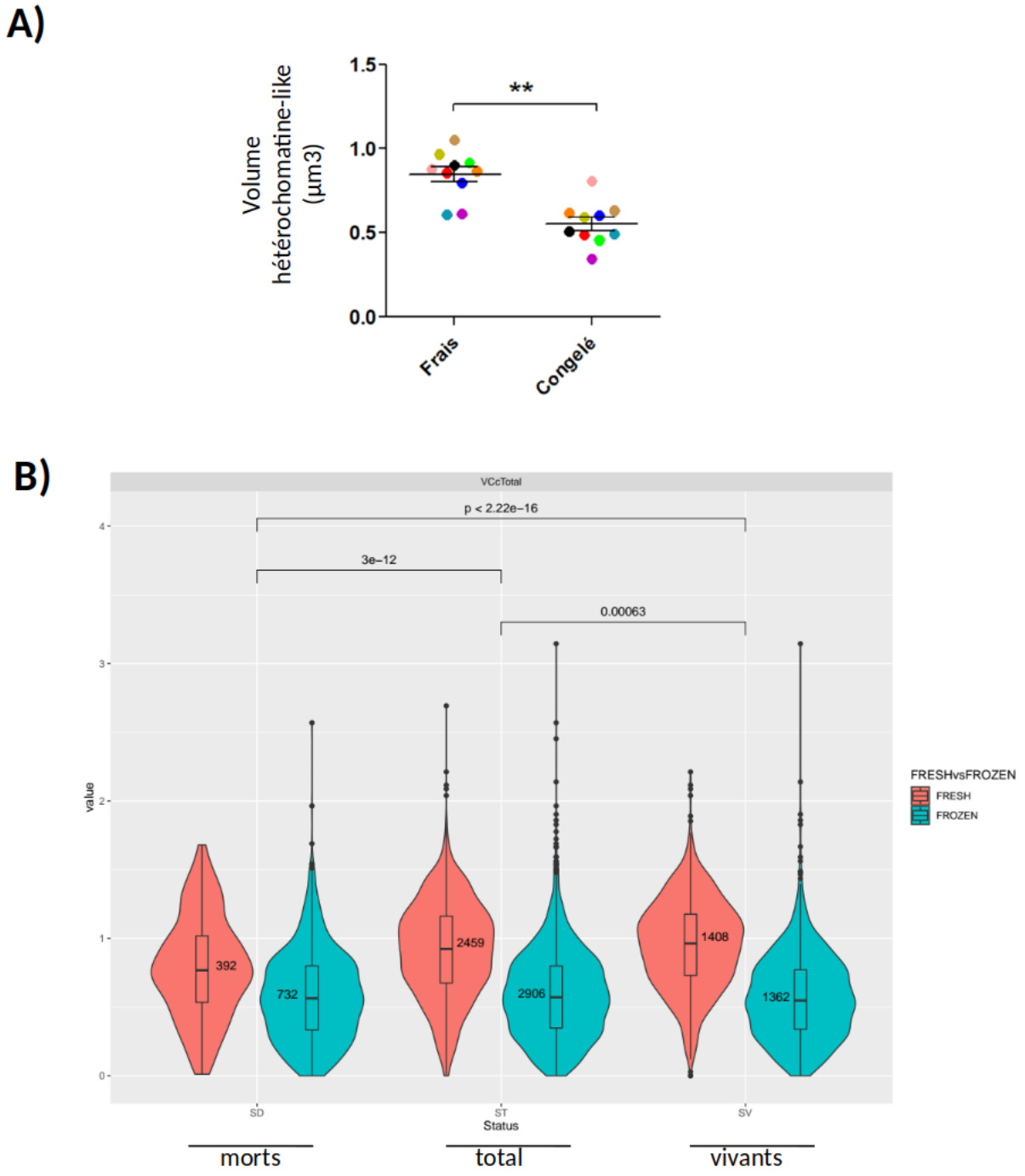


FIGURE 4.7: **Impact de la congélation sur le volume de l'hétérochromatine**. A) Moyenne du volume de l'hétérochromatine-like chez chaque donneur avant et après congélation/décongélation. Les valeurs représentées par des points correspondent aux moyennes +/- SEM et chaque couleur représente un patient différent. ** indique p<0.01 . B) Violin plots comparant les volumes de l'hétérochromatine dans les spermatozoïdes frais (rouge) et congelés (vert) réalisés à partir des populations de spermatozoïdes morts (morts), de la population totale (total) ou des spermatozoïdes vivants (vivants).

4.5.5 Comportement des populations de spermatozoïdes « vivants », « morts », et totaux

Pour nous assurer que ces résultats n'étaient pas un effet de la diminution de la viabilité au cours de la congélation/décongélation, nous avons comparé les paramètres de chromatine pour les populations de spermatozoïdes vivants ou morts selon les critères de sélection du marqueur PI, à la population générale. Les paramètres mesurés dans la population générale sont représentatifs des spermatozoïdes vivants, les deux populations se comportent de façon identique

4.6 Bilan

Cette étude valide, sur un autre modèle que les noyaux de plantes les outils d'analyse d'images, que ce soit pour générer les imagettes (*autocrop*), la morphologie nucléaire (*gift wrapping*), le contenu des noyaux en chromatine (NODeJ) ou la bibliothèque de communication avec OMERO. Nous montrons ici que l'analyse de la structure tridimensionnelle peut être un indicateur pour décrire une population de spermatozoïdes frais ou congelés. Ce type d'analyse sera continué dans le reste de l'étude pour étudier la structure de la chromatine dans des spermatozoïdes irradiés. Il faudra sans doute tenter de caractériser la structure hétérochromatine-like, peut-être en effectuant des expériences de multi-FiSH, afin de déterminer si plusieurs chromosomes contribuent à cette structure ou si cette région coïncide avec un territoire chromosomique particulier. Une telle analyse avait été réalisée lors de nos travaux sur la trisomie 21 [Kemeny et al., 2018]. Enfin, la procédure actuelle pourra sans doute être simplifiée dans le futur. Nous avons vu en effet que la population totale de spermatozoïdes est un reflet de la population de cellules vivantes. Il ne sera désormais plus nécessaire de faire les colorations avec l'iodure de propidium qui augmentait le temps de manipulation des échantillons et doublait celui des acquisitions.

Chapitre 5

Entraînement d'un réseau de neurones de type U-NET

Parallèlement à la recherche d'une solution pour segmenter le contenu des noyaux, nous avons décidé après plusieurs discussions avec des chercheurs de l'Institut Pascal de mettre en œuvre une solution de segmentation des noyaux en faisant appel aux réseaux de neurones profonds. Les méthodes d'apprentissage sont adaptées aux tâches d'automatisation de processus et spécifiquement à la segmentation d'objets qui est une étape majeure pour nos analyses. Du point de vue informatique, cette technologie est devenue accessible par la publication de nombreuses bibliothèques, avec toutefois des contraintes de matériel ou d'interopérabilité qui limitent leurs usages notamment pour les biologistes. Les réseaux de neurones sont connus pour leur capacité à résoudre des problèmes jugés faciles par un humain mais difficiles à décrire d'un point de vue informatique. Dans notre cas, la détection des zones de forte intensité dans les noyaux, décrivant les chromocentres, sont faciles à délimiter pour l'opérateur, ce qui n'est le cas pas pour la machine par une approche mathématique classique. Concernant les prérequis pour l'usage de l'apprentissage profond, nous disposions déjà d'un jeu de données de plusieurs centaines d'images 3D expertisé manuellement après traitement par NucleusJ2.0. C'est pour l'ensemble de ces raisons qu'un travail de bibliographie en vue du choix de méthodes adaptées à nos problématiques d'analyses, a été engagé courant 2019 avec un stagiaire de licence 2, Rémi Caudron (été 2019). Ce travail nous a orienté vers le réseau U-Net [Ronneberger et al., 2015] et les premiers résultats de segmentation rapidement obtenus nous ont encouragés à poursuivre dans cette voie.

Pour nous familiariser avec U-Net, les premiers essais de segmentation se sont focalisés sur la segmentation

des noyaux. Concernant le jeu d'entraînement de U-Net, nous avons choisi de recourir à nos données analysées par NucleusJ2.0 plutôt qu'à une labellisation manuelle. Ces choix ont été motivés essentiellement par les raisons suivantes :

- La segmentation d'un noyau est moins complexe que la segmentation des chromocentres dont le nombre, la position au sein du noyau ou encore l'intensité sont variables d'un noyau à l'autre.
- Dans le temps imparti nous n'avions pas la possibilité d'expertiser un jeu de données segmentées manuellement. Dans la majorité des cas, il nous est impossible de juger si la segmentation manuelle des opérateurs est de meilleure qualité que les méthodes de *gift wrapping* ou d'Otsu modifié proposées dans NucleusJ2.0.

De nombreuses questions ont émergé durant l'exploration des méthodes d'apprentissage profond. Ces questions concernent le choix des méthodes mais aussi la nature et les qualités requises pour constituer un jeu d'entraînement. Ce chapitre expose les résultats préliminaires obtenus avec U-Net qui montrent l'impact de la constitution d'un jeu de données sur le résultat de la segmentation de noyaux. Un second travail non inclus dans ce manuscrit a été d'établir une revue bibliographique des modèles disponibles en analyse du noyau 3D en collaboration avec le doctorant Guillaume Mougeot.

5.1 Choix du réseau de neurones

Le regain d'intérêt de la communauté scientifique pour l'apprentissage automatisé et particulièrement l'apprentissage profond a fait apparaître depuis une dizaine d'années de très nombreuses bibliothèques comme Tensorflow/Keras [Chollet et al., 2015] [Abadi et al., 2016] ou plus récemment Pytorch [Paszke et al., 2017]. De nombreux modèles ont été générés pour la segmentation, ils continuent de l'être dans de nombreux domaines d'application comme la vision par ordinateur, le diagnostic médical ou encore la robotique. Un des algorithmes majeurs appliqués aux images de biologie est apparu en 2015 et il est encore considéré comme une référence dans notre domaine, probablement pour des raisons de facilité de mise en œuvre. Il s'agit de U-Net. Ce réseau de neurones à convolution, développé spécifiquement pour l'analyse d'images, a montré durant des années son efficacité bien que ce type de réseau ne propose pas dans sa version initiale le traitement d'images en 3D. Ce modèle de référence a donc été choisi comme point de départ pour segmenter les noyaux de plantes. Notre choix pour U-Net se justifie par les points suivants :

- Le nombre important d'études portant sur l'utilisation de U-Net en fait un outil très documenté (plus de

28 000 citations). C'est un outil facile d'installation/utilisation pour un non expert. Cette référence bien connue de la communauté se présente donc comme un point d'entrée logique pour débuter l'exploration des méthodes d'apprentissage profond.

- U-Net a été développé pour la segmentation d'images de cellules en culture obtenues à partir de microscopie en contraste de phase (DIC), c'est-à-dire des images relativement similaires à nos données.
- Du point de vue technique, nous étions confrontés initialement à des contraintes en ressources informatiques avec, entre autres, l'absence de carte graphique dédiée au calcul à notre disposition. Or l'utilisation de la bibliothèque tensorflow est possible avec des ressources informatiques classiques, qui dans mon cas, correspondent à un processeur intel xeon, 32Go de RAM, AMD Radeo Pro WX4100 et ce dans des temps de calcul qui restent acceptables.

Enfin, nous avons constaté que les réseaux disponibles, dont U-Net, sont majoritairement dédiés à l'analyse des images en 2D ou montrent des performances peu satisfaisantes en 3D sur nos images. Nous avons donc décidé de transformer nos images initialement de 3D en n images 2D (n correspondant au nombre de couches d'une pile d'images) afin de répondre aux contraintes du réseau.

Durant la mise en place du code j'ai utilisé la plateforme Jupyter-notebook (https://jupyterbook.org/intro.html) pour tester et mettre en place notre réseau U-Net. Une fois le code fonctionnel, plusieurs scripts python (disponibles en ligne https://gitlab.com/DesTristus/machine_leaning) ont été développés pour entraîner des modèles U-Net ou appliquer un modèle généré sur un ensemble d'images. Nous avons utilisé le coefficient de Jaccard pour l'évaluation du réseau au cours de l'entraînement, qui correspond au rapport de l'intersection entre les ensembles A et B sur les différences entre les ensembles A et B, avec A correspondant aux objets segmentés par le réseau et B aux images données pour l'apprentissage :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

5.2 Choix du jeu de données

Après le choix du réseau U-Net, le second choix portait sur les données à utiliser pour l'entraînement. D'un point de vue qualitatif, il s'agira de déterminer quelles seront les images segmentées qui serviront de référence (vérité de terrain ou *ground truth*), et d'un point de vue quantitatif, quel devra être le nombre d'images

nécessaires pour l'entraînement du réseau. Ces deux points sont abordés de façon empirique dans la littérature, c'est pourquoi cette partie décrit la démarche utilisée pour constituer un jeu de données représentatif de nos données pour entraîner le réseau.

Concernant la vérité de terrain pour des images, les bonnes pratiques de la communauté privilégient une expertise manuelle. Dans le cas des images de nos noyaux en 3D, cette tâche peut s'avérer très chronophage suivant le volume de données nécessaires (15 minutes en moyenne par image 3D). Nous avons choisi dans cette première exploration d'utiliser la segmentation automatique issue de NucleusJ2.0 comme référence. Nous avons utilisé les 680 noyaux de l'écotype Col-0 analysés dans la publication [Dubos et al., 2020] comme jeu de données de référence. Ce jeu de données contient une grande gamme de taille de noyaux dont certains étaient difficiles à segmenter par la méthode d'Otsu modifié, du fait de la position du nucléole proche de la périphérie nucléaire. Cette diversité nous est apparue comme adaptée pour l'apprentissage du modèle U-Net. Nous disposons également des 680 noyaux segmentés par la méthode du *gift wrapping*.

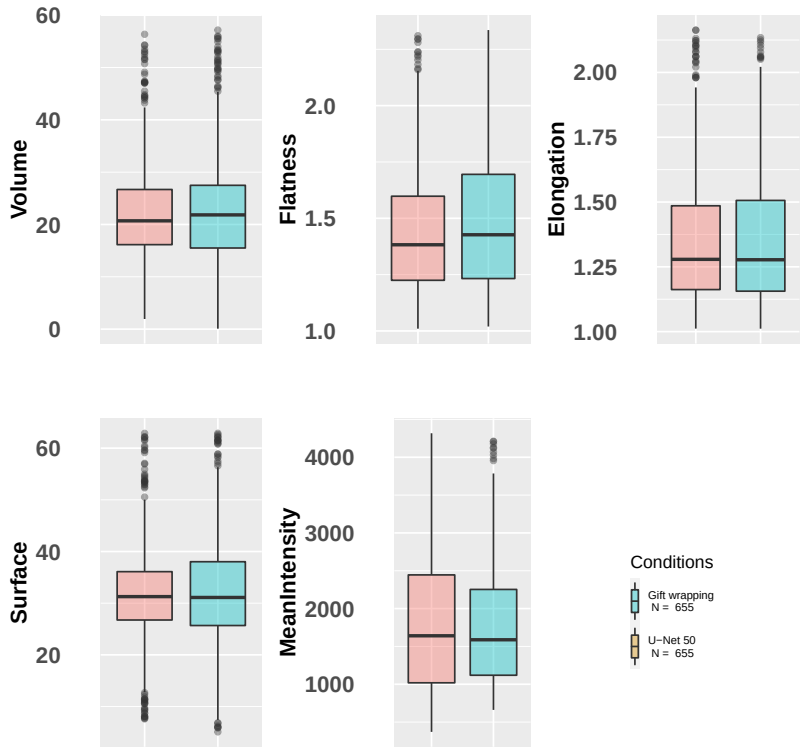


FIGURE 5.1: **Comparaison des paramètres de morphologie nucléaire** : résultats des segmentations par U-Net_50 et *gift wrapping* (Volume, Flatness, Elongation, Surface et MeanIntensity) pour 680 noyaux de Col-0 issus de [Dubos et al., 2020].

5.2.1 Entrainement de U-Net avec 50 images 3D

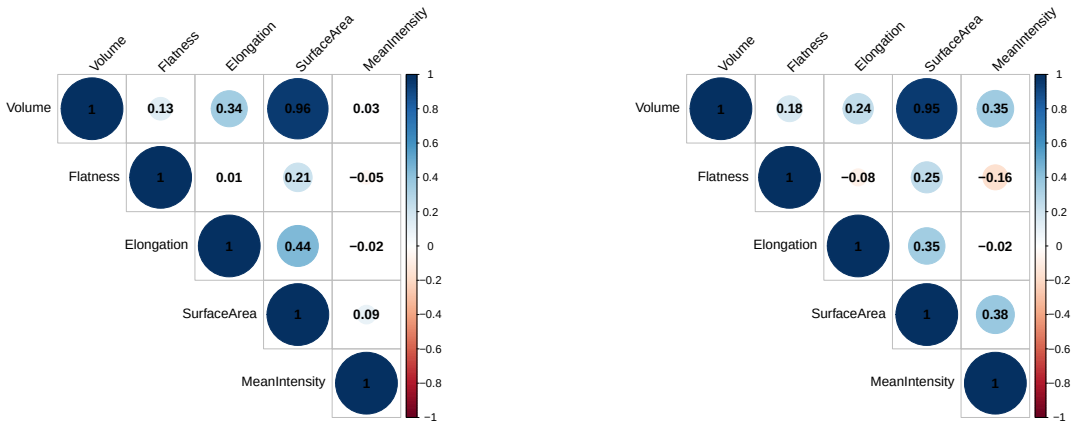
Comme les méthodologies d'apprentissage profond sont très gourmandes en ressources informatiques et que nous ne disposions pas de carte graphique (GPU) dédiée durant la phase initiale d'expérimentation, il a été choisi arbitrairement de limiter à 50 images 3D soit 2644 images 2D (moyenne de 53 plans par images). Le modèle issu de cet entraînement est nommé ensuite U-Net_50.

Le premier apprentissage a été réalisé grâce à un tirage aléatoire de 50 noyaux parmi les 680 noyaux Col-0 à notre disposition. Avec ce jeu de données, la performance (*accuracy*) du modèle est estimée à 99,3%. Ce chiffre très élevé n'est sans doute pas représentatif de la pertinence de notre modèle car il y a de nombreuses couches de fond dans nos images de noyaux nécessaire au bon fonctionnement de la segmentation d'Otsu. Ces couches induisent une erreur dans le calcul de l'indice de Jaccard qui lisse la statistique du taux de réussite de l'apprentissage. Il conviendra de recalculer cet indice en s'affranchissant de ces plans non informatifs.

La fonctionnalité de NucleusJ2.0 qui permet de calculer les paramètres de morphologie nucléaire à partir des images segmentées a été appliquée sur les images segmentées par le modèle U-Net_50. Afin de classer puis comparer les résultats obtenus par les deux méthodes de segmentation, nous avons ensuite calculé le ratio des volumes obtenus par l'approche classique de *gift wrapping* et par la méthode U-Net :

- un ratio de volume nucléaire *gift wrapping*/U-Net_50 inférieur à 0,8 détermine une classe de noyaux qui seront appelés « DOWN »
- un ratio supérieur à 1,2, une seconde classe de noyaux qui seront « UP »
- les noyaux entre ces deux classes sont considérés comme « neutres ».

La comparaison des résultats de segmentation des 655 noyaux par le modèle U-Net_50 et ceux du *gift wrapping*, ne montre aucune différence statistique pour l'ensemble des paramètres de morphologie nucléaire Figure 5.1. Toutefois, une corrélation apparaît dans le modèle U-Net_50 entre l'intensité moyenne des noyaux et deux paramètres : la surface (cor=0,38) et le volume (cor=0,35) (Figure 5.2 (b)). De telles corrélations n'étaient pas attendues car elles suggèrent que la segmentation dépendrait de la taille du noyau et de son intensité. Nous avons examiné le détail des paramètres obtenus par chacune des deux méthodes de segmentation pour la classe DOWN et après vérification manuelle des images nous constatons que ces noyaux sont majoritairement de faible intensité. Cette classe de noyau est peu représentée dans le jeu de données d'entraînement et pourrait expliquer la difficulté du modèle U-Net_50 à segmenter les noyaux de faible intensité.



(a) *Gift wrapping* (b) Unet-50

FIGURE 5.2: **Matrices de corrélation des paramètres de morphologie nucléaire**. (a) segmentation par la méthode de référence *gift wrapping* et en (b) segmentation à partir du modèle U-Net entraîné avec 2644 images 2D provenant de 50 images 3D. Le jeu de données correspond à 655 noyaux de Col-0 issus de [Dubos et al., 2020]

Pour vérifier ce constat visuel nous avons réalisé des tirages aléatoires de 50 noyaux parmi les 655 noyaux pour constituer plusieurs jeux de données d'entraînement. La Figure 5.3 permet de constater la variabilité de la corrélation entre l'intensité moyenne du noyau et son volume en fonction du tirage au sort des noyaux.

A ce stade, deux questions se posent concernant le jeu de donnés d'apprentissage :

- Quelle est la taille minimale du jeu de données dont nous avons besoin pour entraîner un modèle U-Net ? Un effectif de 50 noyaux ne semblent pas être suffisant pour couvrir toute la gamme de diversité présente dans notre jeu de données.
- Quelles seraient les métriques ou paramètres utilisables pour estimer le nombre de classes de noyaux présents dans nos données ?

5.2.2 Entraînement de U-Net avec 129 images 3D

L'état de l'art et les bonnes pratiques de la communauté de recherche en apprentissage machine ne donnent pas de règles générales concernant la taille du jeu de données pour entraîner un modèle U-Net. Toutefois, au travers de la pratique des utilisateurs, plusieurs milliers d'images annotées sont souvent utilisés pour pouvoir reconnaître un objet donné. Ce nombre augmente de manière non linéaire en fonction du nombre et

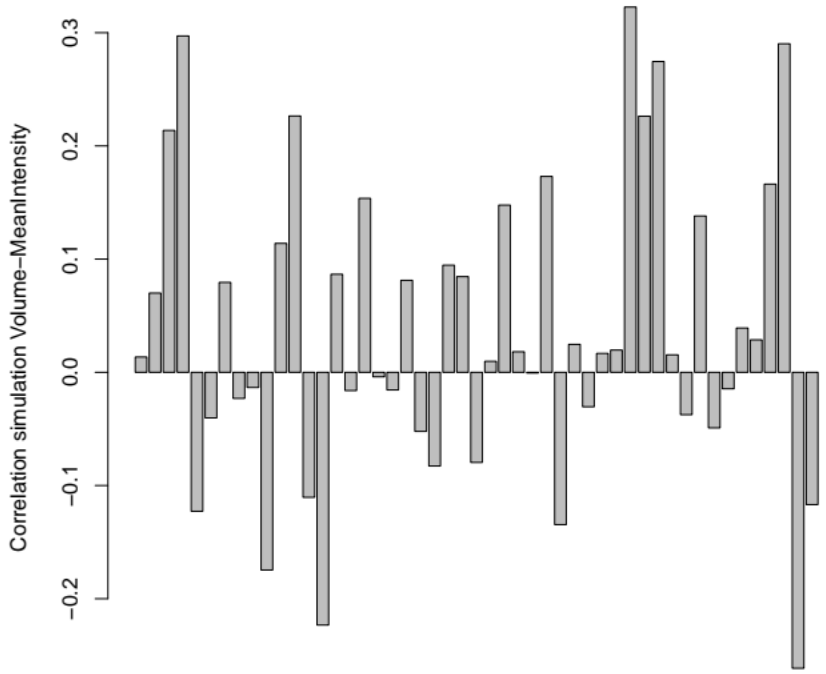


FIGURE 5.3: Evaluation de la correlation Volume/MeanIntensity pour 48 tirages aléatoires de 50 noyaux parmi 655 noyaux Col-0.

de la complexité de forme des objets à détecter au sein d'une image. Le nombre de données dépend donc de la complexité du problème à résoudre mais aussi de l'architecture du réseau utilisé. Dans notre cas, avec l'utilisation de U-Net, nous montrons que le nombre de 50 images 3D sélectionnées de manière arbitraire est insuffisant, et la vérification manuelle des segmentations indique une sous-représentation des images de noyaux avec des intensités faibles. Cette classe de noyaux est encore plus problématique lorsqu'elle concerne des petits noyaux dans des contextes où le rapport signal/bruit est important. Enfin, nous ne pouvons pas certifier que l'ensemble des paramètres que nous générons nous permet de connaître et de décrire toutes les classes existantes de noyaux dans nos données.

Nous avons alors cherché des solutions afin d'améliorer le modèle pour qu'il puisse prendre en compte la classe de noyaux de faible intensité. Pour cela, j'ai augmenté le jeu de données d'entraînement, initialement de 50 noyaux, en ajoutant 79 noyaux de faible intensité qui sont correctement segmentés avec la méthode du *gift wrapping*. L'apprentissage d'un modèle avec U-Net et ces 129 noyaux sera appelé U-Net_129_SORT.

Nous avons également réalisé un troisième modèle entraîné à partir de 129 noyaux tirés aléatoirement : U-Net-129_RANDOM. 129 images 3D représentent 6837 images 2D (moyenne de 53 plans par image). Nous disposons donc maintenant de 3 modèles que nous pouvons comparer pour leur efficacité à segmenter les 680 noyaux de notre jeu de données.

- Les volumes des noyaux segmentés par U-Net_129_SORT et U-Net-129_RANDOM montrent respectivement 30 et 25% de noyaux avec des ratio de volumes qui diffèrent de plus de 20% avec les résultats obtenus par la méthode du *gift wrapping* (UP + DOWN).
- U-Net_129_SORT se distingue de U-Net_129_RANDOM par une diminution de la corrélation entre intensité moyenne / volume du noyau ou surface du noyau avec respectivement r=0,18 et r=0,24 et de 0,2 et 0,27.
- Avec le modèle U-Net_129_SORT, 25 noyaux (5% de la totalité des noyaux) sont classés comme DOWN.
- Enfin, avec U-Net_129_SORT, 144 noyaux (25% du total) sont classés comme UP contre 76 noyaux (13% du total) pour le modèle U-Net_129_RANDOM. Ces 76 noyaux sont d'ailleurs inclus dans l'ensemble des 144 noyaux UP issus de U-Net_129_SORT.

Les deux modèles U-Net_129_SORT et U-Net_129_RANDOM ont des résultats de segmentation qui diffèrent du fait de la proportion variable des classes de noyaux (des noyaux de faible intensité) en données d'apprentissage. L'examen visuel des résultats de segmentation reste subjectif et pour nombre de noyaux il ne permet pas de conclure sur le résultat se trouvant le plus proche de la vérité. Toutefois, si nous regardons les segmentations sur l'ensemble de la population étudiée, il n'y a pas de décalage majeur dans le volume des objets segmentés Figure 5.4 .

Malgré l'ancienneté toute relative du réseau U-Net [Ronneberger et al., 2015], ce réseau semble générer des résultats préliminaires satisfaisants sur nos données biologiques. En effet, nous sommes capables via ces modèles de discriminer les populations de noyaux appartenant aux deux populations de cellules présentes dans l'épiderme de cotylédon d'*A. thaliana* : les cellules de garde et de pavement (Figure 5.5). Ces premiers résultats suggèrent que 50 images 3D tirées de façon aléatoire ne seraient pas suffisantes pour couvrir la diversité des noyaux et donc pour entraîner efficacement le réseau U-Net. Les modèles U-Net_129_SORT et U-net_129_RANDOM donnent les résultats les plus satisfaisants des 3 modèles générés et s'approchent de ceux obtenus par la méthode de segmentation de référence *gift wrapping*. Cependant, nous sommes conscients que ces segmentations peuvent être biaisées car elles ne se basent pas sur une expertise manuelle.

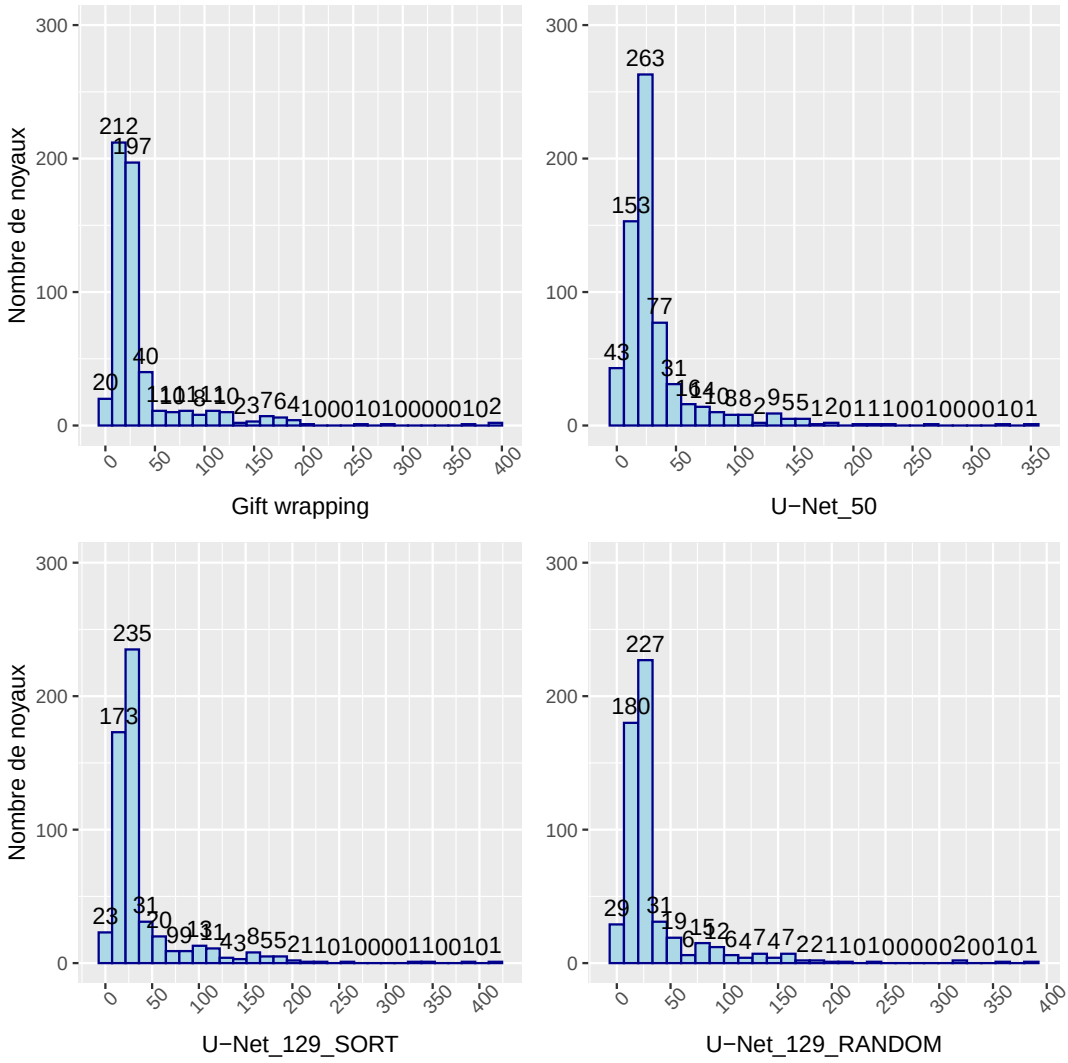


FIGURE 5.4: Histogramme du volume des noyaux calculé à partir de NucleusJ2.0 pour les segmentations obtenues par le *gift wrapping*, le modèle U-Net_50,U-Net_129_SORT et U-Net_129_RANDOM.

5.2.3 Utilisation d'une segmentation manuelle

Avec l'aide de Guillaume Mougeot, nouveau doctorant dans l'équipe, nous avons récemment entrepris de segmenter manuellement des noyaux grâce à l'outil Ilastik [Berg et al., 2019]. Cette segmentation manuelle correspondrait mieux au standard « ground truth » décrit dans la littérature pour constituer les jeux de données d'entraînement. La comparaison de la segmentation manuelle et de celle obtenue sur les mêmes noyaux par la méthode du *gift wrapping* montre une grande disparité Figure 5.6. Ceci s'explique majoritairement par la difficulté d'expertiser visuellement la limite du masque du noyau dans les cas où le nucléole est en périphérie nucléaire. La différence est d'autant plus grande que la méthode de *gift wrapping* surestime la taille de l'objet expliquant les cas extrêmes où le ratio des volumes varie d'un facteur. Toutefois, si nous

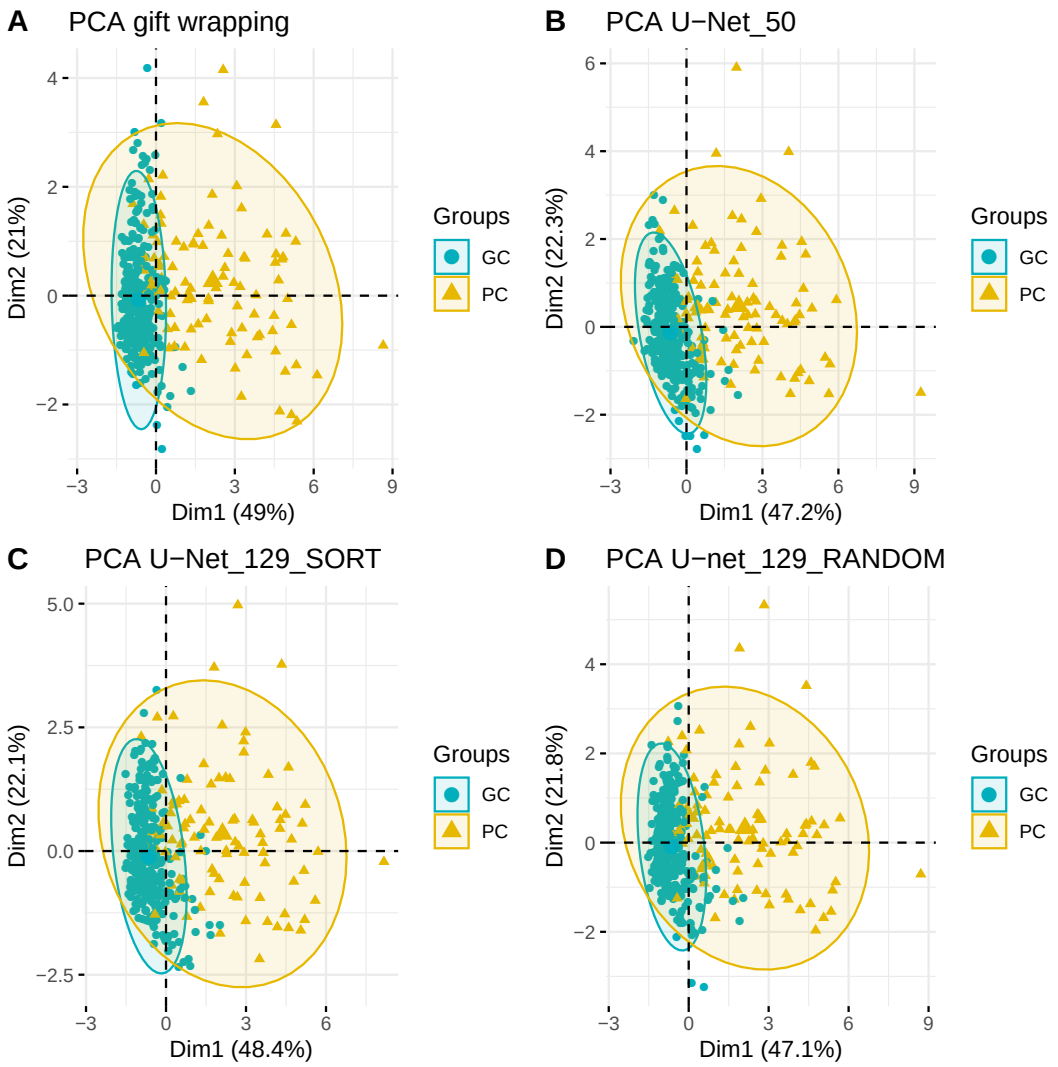


FIGURE 5.5: Analyse en composante principale de 655 noyaux de cellules de garde et de cellules de pavement obtenue A) à partir des paramètres issus de NucleusJ2.0 pour les segmentations obtenues par *gift wrapping*, B) le modèle U-Net_50 , C) U-Net_129_SORT et D) U-Net_129_RANDOM.

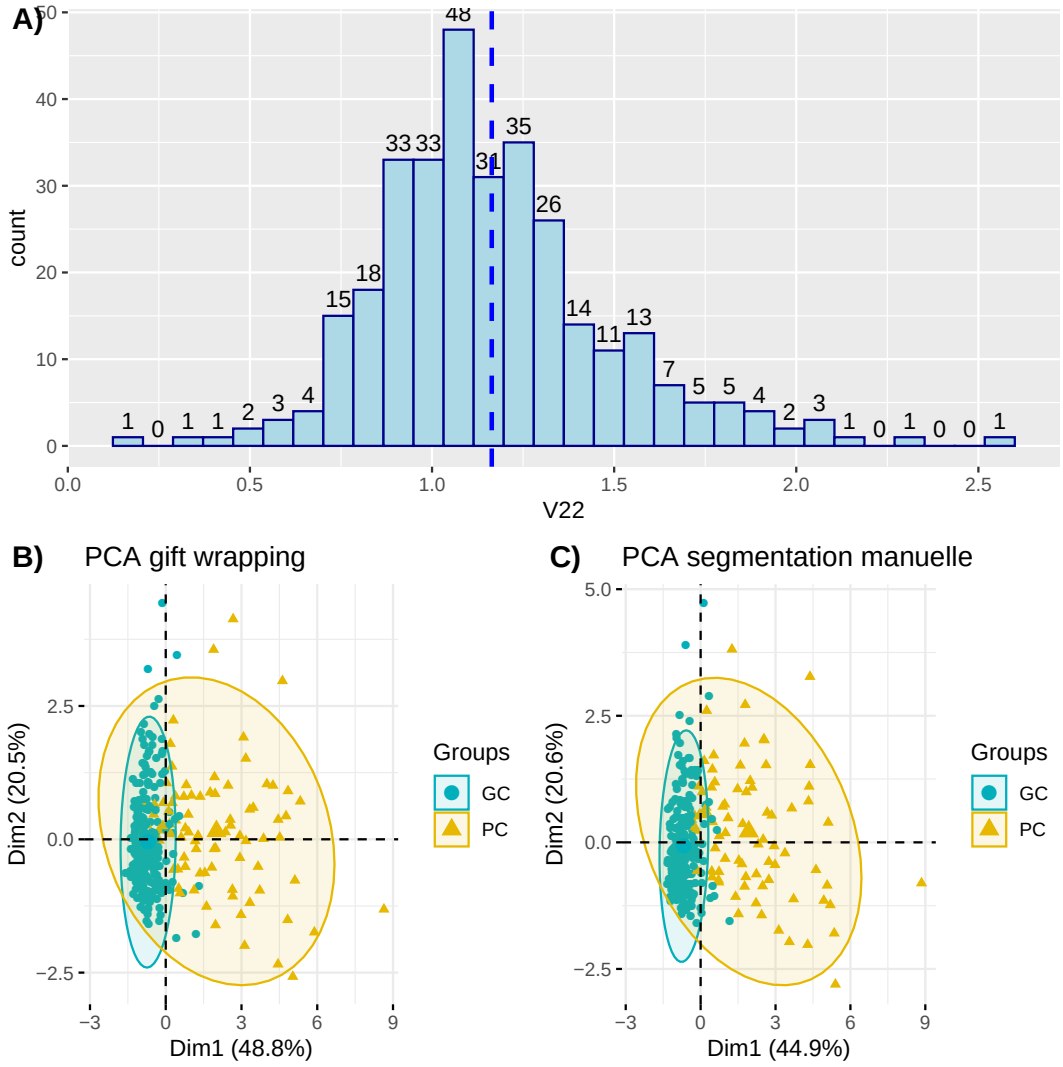


FIGURE 5.6: Description des résultats de segmentation à partir de 318 noyaux Col-0. A) Histogramme de distribution des ratios entre les volumes des noyaux segmentés par *gift wrapping* (méthode de référence) et volumes obtenus suite à la segmentation manuelle. Analyse en composantes principales des individus générés à partir des paramètres de morphologie nucléaire B) du *gift wrapping* et C) de la segmentation manuelle. Les deux populations cellulaires sont représentées en jaune pour les cellules de pavement (PC) et en vert pour les cellules de garde (GC).

comparons les segmentations des noyaux en globalité avec les paramètres de morphologie nucléaire générés par NucleusJ2.0, nous arrivons à discriminer nos populations cellulaires autant avec la segmentation manuelle qu'avec *gift wrapping* Figure 5.6 B et C .

5.3 Bilan

Cette partie constitue une étude préliminaire de l'utilisation d'un réseau de neurones de type U-Net pour segmenter des noyaux. Nos premiers essais montrent l'impact de la taille du jeu de données d'entraînement sur les modèles générés. Ainsi, un effectif de 50 images 3D ne semblent pas suffisant pour couvrir l'ensemble des classes de noyaux présentes dans nos échantillons, et l'intensité moyenne du signal des noyaux influence le modèle généré. L'enrichissement du jeu de données de 79 images supplémentaires a permis d'améliorer le modèle ou en tout cas de diminuer l'impact des noyaux faiblement colorés sur la pertinence du modèle. Il conviendrait maintenant d'estimer le coût du modèle avec l'indice de Jaccard, mais en s'affranchissant des couches non informatives de l'image. Une fois que nous disposerons de cet indicateur, nous pourrons reprendre cette analyse et faire varier le nombre de noyaux, par exemple en incrémentant de 50 le nombre de noyaux constituant le jeu d'entraînement. Nous avons maintenant créé un nouveau jeu de données expertisé manuellement pour respecter les préconisations de l'état de l'art dans le domaine. L'aspect tridimensionnel de nos données montre cependant les difficultés pour délimiter précisément la périphérie nucléaire des noyaux en présence de gros nucléoles. Cette démarche est indispensable pour vérifier l'influence de la méthode d'enveloppe convexe sur les modèles générés. L'existence de ce nouveau jeu de données permettra ainsi de tester plusieurs types de jeux de données d'entraînement segmentés manuellement par Ilastik ou automatiquement par NJ2. U-Net représente un premier modèle exploré mais d'autres modèles pourront ensuite être testés. Ce travail de « benchmarking » des modèles adaptés à la segmentation des noyaux 3D est actuellement en cours et constituera la première étape du doctorat de Guillaume Mougeot.

Chapitre 6

Discussion et perspectives

Les travaux décrits dans cette thèse proposent l'automatisation du logiciel NucleusJ [Poulet et al., 2015] dédié au traitement d'images biologiques générées par microscopie 3D. Nous proposons une nouvelle version appelée NucleusJ2.0, qui s'appuie comme la version précédente sur la plateforme d'analyse d'images imageJ. Nous avons démontré ses performances dans une publication en 2020 dans la revue Nucleus [Dubos et al., 2020]. Nous avons utilisé dans cet article des images 3D de noyaux d'un mutant d'*A. thaliana* ayant des défauts de la morphologie nucléaire et de leur contenu en ADN pour démontrer l'efficacité de l'automatisation de la méthode. En plus de l'outil open source, nous avons mis à disposition une ressource de plusieurs milliers d'images de noyaux de tissus de plantes ainsi que leurs segmentations associées. Ces images sont adressées aux communautés scientifiques de bio-imagerie, de biologie et aux bio-analystes. Nous avons aussi associé les jeux de données d'objets de calibrations de microscope, des sphères fluorescentes, utilisées pour l'amélioration de la précision des calculs du paramètre de la surface. L'ensemble de ces images sont stockées et partagées dans une base de données communautaire OMERO, mise en place à l'initiative d'un financement européen COST-Action INDEPTH (*Impact of Nuclear Domains On Gene Expression Plant Traits*). L'association des outils d'analyse avec la publication d'un jeu de données expertisé est une aide à la promotion des outils. Elle permet à un utilisateur potentiel de les tester pour vérifier son bon fonctionnement et de comparer ses propres données. Elle répond également à une demande croissante de la communauté scientifique, que ce soit pour la disponibilité des données brutes en vue d'une répétabilité des analyses, ou de leur utilisation ultérieure avec d'autres techniques, comme les algorithme d'apprentissage. Cette demande s'est structurée récemment dans des démarches de science ouverte, que promeuvent les tutelles des sciences de la vie comme l'Inserm et

le CNRS, ou l'Inrae, et la gestion des données est à présent exigée par les financeurs lors des réponses aux appels d'offre. Pour répondre à ce besoin, nous avons pu tester la pertinence de l'API OMERO également recommandée dans la littérature de bonnes pratiques (Heddleston et al., 2021) ou au cœur du projet de structuration d'envergure nationale MuDiS4LS porté par l'Institut Français de Bioinformatique.

Dans un second temps, nous avons proposé des solutions pour la segmentation automatique du contenu de la chromatine des noyaux laquelle nécessitait encore une intervention manuelle de l'utilisateur pour le choix du seuil de la méthode de ligne de partage des eaux en 3D. En collaboration avec Axel Poulet, j'ai pu montrer comment notre nouveau plugin NODeJ, développé lui aussi sur la plateforme imageJ, pouvait automatiser la segmentation des chromocentres au travers d'un second article soumis à Bioinformatics. Dans cette publication nous réanalysons un ensemble de jeux de données contenant des images de mutants d'*A. thaliana* connus pour contenir des modifications de la chromatine et notamment une variation du nombre de chromocentres. Ainsi, nous montrons la rapidité et la simplicité d'utilisation de NODeJ sur le jeu de données publié avec NucleusJ2.0 par une analyse automatique de 1400 images de noyaux en moins d'une heure.

Enfin, pour améliorer l'efficacité de notre flux d'analyse d'images, une nouvelle bibliothèque appelée Simple OMERO Client a été conçue afin d'utiliser NucleusJ2.0 directement sur des images stockées au sein de la plateforme OMERO, tout en bénéficiant de la puissance de calcul d'un serveur distant. L'ensemble de nos développements ont été utilisés collectivement dans un projet de preuve de concept, où la structure tridimensionnelle des spermatozoïdes est évaluée en tant que critère de leur qualité. Dans un modèle comparant les spermatozoïdes frais et congelés, nous avons pu montrer que si leur variation de volume est différente d'un donneur à un autre et ne représente pas un bio-marqueur de qualité, en revanche, leur contenu en régions condensées, ressemblant à l'hétérochromatine des chromocentres, pourrait contribuer à mieux les caractériser. Ce modèle montre que NucleusJ2.0 peut-être utile au-delà du modèle plantes notamment en santé, un point important pour notre équipe et le maintien de sa labellisation INSERM.

Dans la dernière partie, nous avons décrit notre démarche dans l'application de méthodes d'apprentissage profond sur nos données. Compte tenu des difficultés techniques déjà exposées, nous nous sommes donc orientés vers l'utilisation d'un réseau de neurones à convolution appelé U-Net et nous avons constaté l'absence d'indication de la communauté scientifique sur la quantité et la qualité de données nécessaires pour entraîner le réseau U-Net. Ainsi, durant ces expériences préliminaires nous avons testé plusieurs volumes de données d'apprentissage pour évaluer l'efficacité du réseau sur nos données. Les réseaux générés nous ont montré des premier résultats encourageants, nous permettant de discriminer morphologiquement nos populations

cellulaires à partir des paramètres géométriques calculés, grâce aux images segmentées en sorties. La démarche que nous proposons est de décrire les jeux de données d'entraînement en s'assurant de la représentation de toute la diversité des objets présents dans les images. Les paramètres mesurés par NucleusJ2.0 nous ont servi de critères pour caractériser une population typique de noyaux chez *A. thaliana* et de constituer un jeu de données représentatif des classes d'objets de nos populations nucléaires. Toutefois, il est important d'identifier l'ensemble des classes et de les représenter dans de bonnes proportions pour éviter des biais d'apprentissage.

6.1 NucleusJ2.0 et NODeJ : une automatisation pour la caractérisation de la morphologie nucléaire et de la chromatine

6.1.1 Validations biologiques des outils

Les analyses de noyaux issus de mutants chez la plante modèle *A. thaliana* valident la pertinence du calcul des paramètres obtenus avec NucleusJ2.0 et NODeJ, décrivant la morphologie nucléaire et leur composition en chromatine. Nous avons montré également que ces deux outils peuvent détecter d'autres objets nucléaires comme des régions répétées rendues fluorescentes par FiSH. Grâce au chaînage et à l'automatisation de ces méthodes de segmentations, nous pouvons identifier plus rapidement des phénotypes comportant des modifications du noyau ou de la chromatine, afin de mieux décrire les altérations observées chez des plantes mutantes. Ceci aura des conséquences sur les approches de génétique et participera à mieux valider la fonction de gènes notamment impliqués dans la régulation de la transcription. Nous avons appliqué notre processus d'analyse sur un modèle de spermatozoïdes chez l'homme, qui confirme l'adaptabilité de nos outils à d'autres type de données. Toutefois, dans les deux cas, les données ont été produites sur les mêmes technologies de microscopes, il serait donc intéressant de tester les limites de nos outils avec des images produites par d'autres méthodes d'acquisitions. Dans le futur, mon équipe espère explorer plusieurs autres types de microscopie comme l'imagerie sur les cellules vivantes avec le *spinning disc* afin d'étudier la dynamique de certains domaines nucléaires au cours du développement ou lors de stress biotique. La réalisation d'images en super résolution via la microscopie STED (*Stimulated-Emission-Depletion*) permettrait également d'analyser plus finement les compartiments nucléaires comme le nucléosquelette et le pore nucléaire. Enfin, il est également prévu de réaliser des images en microscopie électronique (*Serial Block Face scanning Electron Microscopy*) en collaboration avec l'équipe d'Oxford dans le cadre du doctorat de Guillaume Mougeot.

6.1.2 Perspectives d'évolution des algorithmes

6.1.2.1 Optimisation de l'étape automatisation autocrop-segmentation

Nous avons maintenant un outil qui sélectionne automatiquement les noyaux dans une image grand champ et les isole dans des imagettes. Cette étape permet de rechercher le seuil Otsu optimal de chaque noyau, nécessaire dans le cas des images grand champ, qui contiennent des noyaux de différentes intensités. Avec cette procédure, nous rencontrons des cas où deux noyaux se retrouvent dans le même fichier à cause de leur proximité. Lorsque cela se produit, seuls les paramètres du noyau le plus gros sont retenus. C'est un phénomène rare pour nos images, sans incidence sur nos analyses. Cependant, le développement d'un algorithme analysant plusieurs objets au sein d'une même image ouvrirait la possibilité d'analyser des images où les noyaux ont un plus fort taux de regroupement, ce qui est fréquent dans les modèles de culture cellulaire ou de tissus d'origine animale. De plus, ce développement permettrait à terme l'analyse en une étape des images grand champ contenant plusieurs objets que nous produisons désormais dans notre routine. Un tel développement permettrait d'ajouter des paramètres décrivant les informations spatiales des noyaux dans le tissu (profondeur, orientation du noyau, distance entre noyaux etc).

6.1.2.2 Détection d'autres objets intra-nucléaires

Le développement récent de NODeJ permet désormais la caractérisation automatisée des domaines de chromatine comme les chromocentres à partir du masque du noyau. Les résultats que nous obtenons en sortie décrivent le nombre, la position, le volume moyen ou le volume total des chromocentres. Par soustraction de ces informations au reste du volume du masque du noyau, nous quantifions le contenu en chromatine décondensée : l'euchromatine. Actuellement, le nucléole est inclus dans le volume d'euchromatine alors qu'il ne contient pas ou très peu de chromatine. Or, l'absence de marquage du nucléole ne nous a pas permis pour l'instant de quantifier ce domaine nucléaire. L'utilisation de la méthode des gradients est une piste pour détecter ces zones de faibles intensités. Nous sommes en train de produire des images de noyaux contenant de la fibrillarine couplée à la GFP pour visualiser le nucléole au sein de noyaux colorés au DAPI qui permettraient la validation de l'algorithme. En cas de succès de ces développements, nous aurions à terme un outil décrivant les compartiments majoritaires du noyau dans nos préparations : l'hétérochromatine (les chromocentres), le nucléole et par déduction l'euchromatine.

6.1.2.3 Amélioration de la précision dans la détection du noyau

La détection des noyaux en microscopie se base majoritairement sur le marquage de l'ADN avec l'utilisation du DAPI. Ainsi, grâce à cette méthode, nous délimitons le volume des noyaux par segmentation de la limite de la coloration au DAPI, sans inclure la double membrane externe. Pour vérifier si cette approximation du volume nucléaire est correcte, nous pourrions essayer d'utiliser une des méthodes classiques de marquage de la périphérie nucléaire utilisée chez les autres organismes modèles qui consiste à segmenter le noyau après marquage du nucléosquelette. Cette méthodologie semble néanmoins difficile à mettre en œuvre pour plusieurs raisons :

- L'utilisation d'anticorps contre les lamines : dans les tissus de plantes, les anticorps pénètrent difficilement à cause de la paroi végétale. Ce type de marquage ne pourra être appliqué qu'aux noyaux isolés. La seconde limite provient de la difficulté de produire de bons anticorps contre des protéines végétales.
- L'utilisation d'une protéine chimère par transgenèse de la périphérie nucléaire telle que SUN, NEAP1, KAKU4 ou CRWN couplée à une molécule fluorescente de type GFP. Les essais déjà réalisés en ce sens dans l'équipe n'ont pas permis non plus de répondre à la question, à cause de difficultés techniques illustrées dans la Figure 6.1 : d'une part nous observons souvent dans ces modèles transgéniques des protubérances membranaires artéfactuelles et un signal de fluorescence discontinu, ce qui rend difficile la segmentation des images. D'autre part, les signaux sont insuffisants pour obtenir le contour complet du noyau, particulièrement dans les premiers et derniers plans (Figure 6.1).

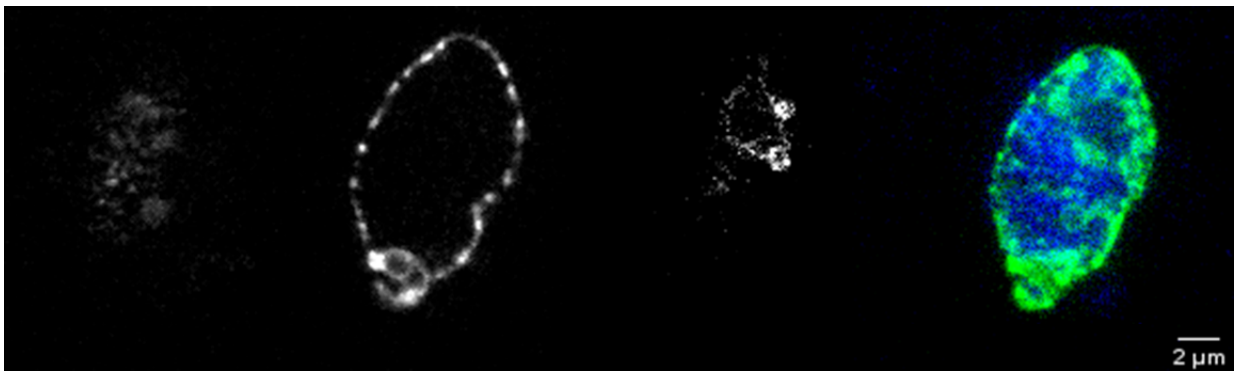


FIGURE 6.1: **Visualisation de la périphérie nucléaire**. Noyau de cotylédon fixé coloré au DAPI provenant d'une plante exprimant SUN1-GFP sous contrôle du promoteur 35s [Oda et al ; Plant J 2011], pile de 49 images acquises en lumière structurée objx63, NA1,4,(xyz : 0,1x0,1x0,2 μm). A gauche, 3 plans (5,22,44) du canal GFP ; à droite, Z projection des deux canaux (bleu : DAPI et vert GFP).

Enfin, l'amélioration de la segmentation du masque du noyau dans NucleusJ2.0 par l'intégration de l'al-

gorithme du *gift wrapping* en appliquant la marche de Jarvis puis, plus récemment, le parcours de Graham, fournit des résultats pour la caractérisation de la morphologie nucléaire dans des temps d'analyse très satisfaisants, et résout les difficultés initiales observées avec la méthode d'Otsu modifiée, liées à la présence du nucléole. Cependant, ces méthodes de calculs de l'enveloppe convexe surestiment le volume de certains noyaux ayant des formes allongées (Figure 2.1 D)). Il semble donc nécessaire de mesurer l'impact de ce biais par exemple sur des populations de noyaux de tissus racinaire différencié d'*A. thaliana*, dont la forme est moins sphérique afin de déterminer les limites de l'outil. Pour réaliser cette étude, nous disposons de jeux de données de noyaux issus de poils racinaires [Benoit et al., 2017].

6.2 Développements informatiques dédiés à l'analyse d'images

L'évolution de la microscopie à haut débit nécessite désormais des logiciels performants pour la quantification de signaux biologiques. La cadence de la production de données engendre une demande croissante de solutions logicielles facilement accessibles, déployables et permettant la reproductibilité des résultats et la traçabilité des données. Ce flux de données nécessite aussi des solutions de stockage adaptées et des systèmes d'analyses optimisés en temps de calcul. Malgré la diversité des bibliothèques regroupant des fonctionnalités mathématiques dédiées à la bio-imagerie et le nombre de plate-formes qui sont apparues ces dernières années, le domaine de la bio-analyse nécessite encore des connaissances pluridisciplinaires complexes pour leurs mise en place et déploiement en laboratoire. Ainsi, les développements informatiques de la communauté qui se consacre aux traitements des images est très active et tend à s'organiser par exemple autour de forum (https://forum.image.sc/) ou de consortium nationaux comme le GDR Imabio (http://imabio-cnrs.fr/) ou européens comme le COST-Action NEUBIAS (http://eubias.org/NEUBIAS/). Les développements en informatique se poursuivent pour adapter les algorithmes aux nouvelles technologies des microscopes, avec notamment le passage des méthodes initialement consacrées à l'analyse d'images 2D à des images de microscopie en 3D. Ce travail se fait de manière à intégrer de nouvelles technologies informatiques telles que l'utilisation des processeurs graphiques, ou les serveurs dédiés aux calculs pour réduire les temps d'analyse ainsi que d'environnements assurant une meilleure reproductibilité comme docker ou conda. Pour finir ces outils, qu'ils soient matériels ou logiciels, sont des prérequis essentiels pour mettre à disposition les nouvelles méthodologies basées sur l'apprentissage profond, qui sont gourmandes en ressources informatiques.

Les plugins NucleusJ2.0 et NODeJ reposent sur des dépendances qui sont en constante évolution pour

intégrer les nouvelles méthodes citées ci-dessus. Nous avons constaté, il y a quelques mois, des changements dans la méthode du filtre gaussien proposé par ImageJ suite à une mise à jour de notre code, entraînant des difficultés de détection des noyaux par l'autocrop dans une partie de nos images. L'identification et la correction de ce type de changement est une tâche chronophage pour le maintien des outils. Afin de prévenir ces changements extérieurs à nos développements, il semble donc indispensable de mettre en place des tests fonctionnels afin de s'assurer de la pérennité des plugins mais également pour suivre de façon plus efficace les versions des dépendances. Pour cela, nous sommes en train de sélectionner des données expertisées manuellement au travers desquelles nous pourrons vérifier automatiquement nos résultats avant de proposer une nouvelle version aux utilisateurs.

6.3 Vers une automatisation de la segmentation par les méthodes d'apprentissage profond

6.3.1 Choix de l'utilisation du modèle U-Net

Les méthodes d'apprentissage profond sont à l'heure actuelle encore difficiles à prendre en main par un public non initié car elles nécessitent des compétences en programmation pour les mettre en œuvre, mais elles requièrent aussi de maîtriser l'environnement logiciel qui évolue très rapidement. Dans de nombreuses publications seul le cœur du code est disponible, ce qui pose des difficultés de reproductibilité des résultats car il est nécessaire de réimplémenter les algorithmes [Dacrema et al., 2019]. La simplicité de mise en place de U-Net est donc un atout majeur de ce réseau malgré son ancienneté, et explique sans aucun doute son succès malgré ses limites, en particulier pour la 3D. D'autres solutions existent notamment avec des modèles pré-entrainés voire même des service web comme Imjoy (https://imjoy.io/#/) NucleAIzer (www.nucleaizer.org), CellPose (www.cellpose.org) ou DeepCell (www.deepcell.org). Le domaine est en évolution rapide et de nombreux nouveaux outils de type *web-services* se mettent en place pour partager les algorithmes et modèles.

6.3.2 Les impacts du jeu de données d'entraînement

Les résultats préliminaires de la segmentation des masques de noyaux par l'utilisation d'un réseau U-Net, montrent des résultats proches de ceux obtenus par NucleusJ2.0 pour la caractérisation de la morphologie

nucléaire. Ces résultats sont cependant à prendre avec précaution puisque pour entraîner notre modèle nous n'avons pas utilisé comme vérité de terrain des données segmentées manuellement comme le conseillent les bonnes pratiques de la communauté, mais les résultats de segmentation issus de NucleusJ2.0. Nous avons donc un risque d'avoir entraîné le réseau à appliquer une méthode d'enveloppe convexe sur les images que nous lui soumettons. Par exemple, les données en sortie risquent de contenir les biais de sur-estimation des masques de noyaux du *gift wrapping* expliqués dans la section 2.1. Nous avons donc commencé un travail de segmentation manuelle des noyaux qui servira de vérité de terrain au modèle U-Net et aux autres modèles qui seront testés. Les premières segmentations de noyaux par le logiciel ilastik sont comparées aux résultats avec les réseaux entraînés avec les données issues de NucleusJ2.0, et montrent une grande disparité d'objets avec des ratios de volume pouvant aller jusqu'à un facteur 3. La comparaison manuelle des noyaux ayant des résultats de segmentation les plus divergents montre qu'il est difficile de délimiter visuellement le bord de nos objets dans le cas de gros nucléoles. Nous avons vu qu'il est important de garder une représentation de la diversité d'un jeu de données dans le jeu d'entraînement. L'exclusion de telles images de notre jeu de données d'entraînement nous semblerait donc risqué car cela pourrait conduire à une faiblesse du modèle et à une mauvaise segmentation des noyaux de faible intensité comme nous l'avons vu dans la Figure 5.2. La solution que nous envisageons est de constituer un jeu de données mixte, constitué des segmentations issues de NucleusJ2.0 pour les noyaux avec de gros nucléoles, et des segmentations manuelles où le contour sera plus précis. Nous ajouterons tous les autres cas où la segmentation par *gift wrapping* nous paraît satisfaisante. Cette méthode permet ainsi de limiter les biais de la segmentation manuelle de l'opérateur et accélérerait la constitution d'un jeu de données d'entraînement de plus grande taille.

6.4 Conclusion

NucleusJ2.0 et NODeJ proposent une automatisation des processus d'analyse d'images permettant un nouveau pas dans la description de la structure tridimensionnelle des noyaux. Ces outils donnent les moyens opérationnels d'un phénotypage haut débit, en utilisant la microscopie 3 dimensions, pour la caractérisation de mutants de la morphologie nucléaire et de l'organisation de la chromatine. Enfin l'utilisation des résultats de ces méthodes de segmentations conventionnelles serviront dans la mise en place de modèles d'apprentissage profond en permettant une labellisation rapide des images.

Annexes

Parameter name	Default value	Value type	Corresponding number in user interface	Description
xCropBoxSize	40	Int	1	Number of voxels added to xmin and xmax of the connected component defining the final box size in x
yCropBoxSize	40	Int	2	Number of voxels added to ymin and ymax of the connected component defining the final box size in y
zCropBoxSize	20	Int	3	Number of voxels added to zmin and zmax of the connected component defining the final box size in z
minVolumeNucleus	1	Int	4	Minimum volume of detected object
maxVolumeNucleus	2147483647	Int	5	Maximum volume of detected object
thresholdOSTUcomputing	20	Int	6	Minimal default OTSU threshold
channelToComputeThreshold	0	Int	7	Channel number used to compute OTSU threshold (Channel 1 is 0 etc)
slicesOTSUcomputing	0	Int	8	Slice start used to compute OTSU threshold
boxesPercentSurfaceToFilter	50	Int	9	Surface percent of boxes to groups them
boxesRegroupement	true	boolean	true/false	Activation of boxes regroupement
xcal	1	Double	11	X calibration value
ycal	1	Double	12	Y calibration value
zcal	1	Double	13	Z calibration value

FIGURE 6.2: **Liste des paramètres de l'*autocrop***. Description de l'ensemble des paramètres disponibles pour l'utilisation de l'*autocrop* de NucleusJ2.0. Extrait de la documentation en ligne https://gitlab.com/DesTristus/NucleusJ2.0/-/wikis/Autocrop

Parameter name	Description
NucleusFileName	Nucleus image file name
Volume	Volume of segmented object (sum of white voxels)
Flatness	Computed by ratio between major and minor axes of the segmented object
Elongation	Computed by ratio between major and medium axes of the object
Sphericity	Computed by ratio between surface area and volume
Esr	Equivalent spherical radius, estimated radius computed from the volume
SurfaceArea	Segmented object surface area
MeanIntensityNucleus	Mean voxels intensity of segmented image
MeanIntensityBackground	Mean intensity of the background
StandardDeviation	Standard deviation of the segmented object intensity
MinIntensity	Minimum voxel intensity of the segmented object
MaxIntensity	Maximum voxel intensity of the segmented object
MedianIntensityImage	Median voxel intensity of raw image
MedianIntensityNucleus	Median voxel intensity of segmented object
MedianIntensityBackground	Median voxel intensity of background
ImageSize	Number of voxel
OTSUThreshold	OTSU threshold used to segment the image

FIGURE 6.3: **Liste des paramètres décrivant la morphologie nucléaire**. Description de l'ensemble des paramètres disponibles à l'issu de l'analyse de segmentation disponible dans NucleusJ2.0 (*gift wrapping* et Otsu modifié). Extrait de la documentation en ligne https://gitlab.com/DesTristus/NucleusJ2.0/-/wikis/Nucleus-segmentation

Parameter name	Description
Title	Image name
NbCc	mean volume of chromocenter(s) per nucleus
VccMean	mean volume of chromocenter(s) per nucleus
VccTotal	total volume of chromocenter(s) per nucleus
DistanceBorderToBorderMean	mean distance of chromocenter(s) border to nuclear periphery.
DistanceBarycenterToBorderMean	mean distance of chromocenter(s) barycenter to nuclear periphery.
IntensityRHF	total chromocenter intensity / nuclear intensity.
VolumeRHF = total chromocenter volume / n	total chromocenter volume / nuclear volume.

FIGURE 6.4: **Liste des paramètres décrivant les chromocentres**. Ensemble des paramètres disponibles à l'issu de l'analyse de segmentation disponible dans NODeJ. Le fichier de résultat est au format tabulé où chaque ligne correspond aux paramètres d'un chromocentre détecté.

Parameter name	Description
Titre	Image name and chromocenter number
Volume	volume of chromocenter
BorderToBorderDistance	distance between chromocenter border to nuclear periphery
BarycenterToBorderDistanceNucleus	distance between chromocenter barycenter to nuclear periphery.

FIGURE 6.5: **Liste des paramètres décrivant les chromocentres au sein d'un noyau**. Ensemble des paramètres disponibles à l'issu de l'analyse de segmentation disponible dans NODeJ. Le fichier de résultat est au format tabulé où chaque ligne correspond aux paramètres d'un noyaux analysé.

Bibliographie

[Abadi et al., 2016] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems.

[Alloghani, 2020] Alloghani, M. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. *Biomedical Signal Processing and Artificial Intelligence in Healthcare*, (January) :51–89.

[Arganda-Carreras et al., 2017] Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K. W., Schindelin, J., Cardona, A., and Seung, H. S. (2017). Trainable Weka Segmentation : A machine learning tool for microscopy pixel classification. *Bioinformatics*, 33(15) :2424–2426.

[Barone JG, De Lara J, Cummings KB, 1994] Barone JG, De Lara J, Cummings KB, W. W. J. A. (1994). DNA organization in human spermatozoa.

[Benoit et al., 2017] Benoit, M., Voisin, M., Desset, S., Vanrobays, E., Evans, D. E., Tatout, C., Poulet, A., Duc, C. C., Probst, A. V., Tutois, S., Voisin, M., Desset, S., Tutois, S., Vanrobays, E., Benoit, M., Evans, D. E., Probst, A. V., and Tatout, C. (2017). The LINC complex contributes to heterochromatin organisation and transcriptional gene silencing in plants. *Journal of Cell Science*, 130(3) :590–601.

[Berg et al., 2019] Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., Eren, K., Cervantes, J. I., Xu, B., Beuttenmueller, F., Wolny, A., Zhang, C., Koethe, U., Hamprecht, F. A., and Kreshuk, A. (2019). Ilastik : Interactive Machine Learning for (Bio)Image Analysis. *Nature Methods*, 16(12) :1226–1232.

[Beucher and Lantuejoul, 1979] Beucher, S. and Lantuejoul, C. (1979). Use of Watersheds in Contour Detection.

[Bickmore, 2013] Bickmore, W. A. (2013). The spatial organization of the human genome. *Annual Review of Genomics and Human Genetics*, 14 :67–84.

[Bickmore and Van Steensel, 2013] Bickmore, W. A. and Van Steensel, B. (2013). Genome architecture : Domain organization of interphase chromosomes. *Cell*, 152(6) :1270–1284.

[Boitrelle et al., 2012] Boitrelle, F., Albert, M., Theillac, C., Ferfouri, F., Bergere, M., Vialard, F., Wainer, R., Bailly, M., and Selva, J. (2012). Cryopreservation of human spermatozoa decreases the number of motile normal spermatozoa, induces nuclear vacuolization and chromatin decondensation. *Journal of Andrology*, 33(6) :1371–1378.

[Chang et al., 2004] Chang, F., Chen, C. J., and Lu, C. J. (2004). A linear-time component-labeling algorithm using contour tracing technique. *Computer Vision and Image Understanding*, 93(2) :206–220.

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras. *"oui"*.

[Collins, 2007] Collins, T. J. (2007). ImageJ for microscopy. *BioTechniques*, 43(1 Suppl) :25–30.

[Costa-Nunes et al., 2014] Costa-Nunes, P., Vitins, A., and Pontes, O. (2014). Connecting the dots of RNA-directed DNA methylation in Arabidopsis thaliana. *Chromosome Research*, 22(2) :225–240.

[Dacrema et al., 2019] Dacrema, M. F., Cremonesi, P., and Jannach, D. (2019). Are we really making much progress ? A worrying analysis of recent neural recommendation approaches. *RecSys 2019 - 13th ACM Conference on Recommender Systems*, pages 101–109.

[De Chaumont et al., 2012] De Chaumont, F., Dallongeville, S., Chenouard, N., Hervé, N., Pop, S., Provoost, T., Meas-Yedid, V., Pankajakshan, P., Lecomte, T., Le Montagner, Y., Lagache, T., Dufour, A., and Olivo-Marin, J. C. (2012). Icy : An open bioimage informatics platform for extended reproducible research. *Nature Methods*, 9(7) :690–696.

[Desset et al., 2018] Desset, S., Poulet, A., and Tatout, C. (2018). Quantitative 3D Analysis of Nuclear Morphology and Heterochromatin Organization from Whole-Mount Plant Tissue Using NucleusJ. *Methods in molecular biology (Clifton, N.J.)*, 1675 :615–632.

[Dittmer et al., 2007] Dittmer, T. A., Stacey, N. J., Sugimoto-Shirasu, K., and Richards, E. J. (2007). LITTLE NUCLEI genes affecting nuclear morphology in Arabidopsis thaliana. *Plant Cell*, 19(9) :2793–2803.

[Dubos et al., 2020] Dubos, T., Poulet, A., Gonthier-Gueret, C., Mougeot, G., Vanrobays, E., Li, Y., Tutois, S., Pery, E., Chausse, F., Probst, A. V., Tatout, C., and Desset, S. (2020). Automated 3D bio-imaging analysis of nuclear organization by NucleusJ 2.0. *Nucleus*, 11(1) :315–329.

[Dumur et al., 2019] Dumur, T., Duncan, S., Graumann, K., Desset, S., Randall, R. S., Scheid, O. M., Bass, H. W., Prodanov, D., Tatout, C., and Baroux, C. (2019). Probing the 3D architecture of the plant nucleus with microscopy approaches : challenges and solutions. *Nucleus*, 10(1) :181–212.

[Eme et al., 2017] Eme, L., Spang, A., Lombard, J., Stairs, C. W., and Ettema, T. J. (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, 15(12) :711–723.

[Feng et al., 2014] Feng, C. M., Qiu, Y., Van Buskirk, E. K., Yang, E. J., and Chen, M. (2014). Light-regulated gene repositioning in Arabidopsis. *Nature communications*, 5 :3027.

[Finn and Misteli, 2019] Finn, E. H. and Misteli, T. (2019). Molecular basis and biological function of variability in spatial genome organization. *Science*, 365(6457).

[Fiserova et al., 2009] Fiserova, J., Kiseleva, E., and Goldberg, M. W. (2009). Nuclear envelope and nuclear pore complex structure and organization in tobacco BY-2 cells. *Plant Journal*, 59(2) :243–255.

[Fransz et al., 2002] Fransz, P., De Jong, J. H., Lysak, M., Castiglione, M. R., and Schubert, I. (2002). Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22) :14584–14589.

[Fraser et al., 2015] Fraser, J., Williamson, I., Bickmore, W. A., and Dostie, J. (2015). An Overview of Genome Organization and How We Got There : from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*, 79(3) :347–372.

[Fukushima, 1979] Fukushima, K. (1979). A self-organizing neural network model for a mechanism ofpattern recognition unaffected by shift in position. *Trans. IECE*, J62-A(10) :658–665.

[Goldberg et al., 2005] Goldberg, I. G., Allan, C., Burel, J. M., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P. K., and Swedlow, J. R. (2005). The Open Microscopy Environment (OME) Data Model and XML file : open tools for informatics and quantitative analysis in biological imaging. *Genome biology*, 6(5).

[Gonzalez, R.C., and Woods, 1992] Gonzalez, R.C., and Woods, R. E. (1992). *Digital Image Processing.*

[Goswami et al., 2020] Goswami, R., Asnacios, A., Hamant, O., and Chabouté, M. E. (2020). Is the plant nucleus a mechanical rheostat ? *Current Opinion in Plant Biology*, 57 :155–163.

[Goto et al., 2014] Goto, C., Tamura, K., Fukao, Y., Shimada, T., and Hara-Nishimura, I. (2014). The novel nuclear envelope protein KAKU4 modulates nuclear morphology in Arabidopsis. *Plant Cell*, 26(5) :2143–2155.

[Graham, 1972] Graham, R. L. (1972). An efficient algorith for determining the convex hull of a finite planar set.

[Graumann, 2014] Graumann, K. (2014). Evidence for LINC1-SUN associations at the plant nuclear periphery. *PLoS ONE*, 9(3) :1–7.

[Graumann and Evans, 2020] Graumann, K. and Evans, D. E. (2020). Growing the nuclear envelope proteome. *Nature Plants*, 6(7) :740–741.

[Hammadeh et al., 1999] Hammadeh, M. E., Askari, A. S., Georg, T., Rosenbaum, P., and Schmidt, W. (1999). Effect of freeze-thawing procedure on chromatin stability, morphological alteration and membrane integrity of human spermatozoa in fertile and subfertile men. *International Journal of Andrology*, 22(3) :155–162.

[Haque et al., 2006] Haque, F., Lloyd, D. J., Smallwood, D. T., Dent, C. L., Shanahan, C. M., Fry, A. M., Trembath, R. C., and Shackleton, S. (2006). SUN1 Interacts with Nuclear Lamin A and Cytoplasmic Nesprins To Provide a Physical Connection between the Nuclear Lamina and the Cytoskeleton. *Molecular and Cellular Biology*, 26(10) :3738–3751.

[Heitz, 1928] Heitz, E. (1928). Das {Heterochromatin} der {Moose}. *Jahrb Wiss Botanik 69 : 762–818.*

[Higa et al., 2014] Higa, T., Suetsugu, N., and Wada, M. (2014). Plant nuclear photorelocation movement. *Journal of Experimental Botany*, 65(11) :2873–2881.

[Hollandi et al., 2020] Hollandi, R., Diósdi, Á., Hollandi, G., Moshkov, N., and Horváth, P. (2020). Annotator J : An image J plugin to ease hand annotation of cellular compartments. *Molecular Biology of the Cell*, 31(20) :2179–2186.

[Hu et al., 2019] Hu, B., Wang, N., Bi, X., Karaaslan, E. S., Weber, A. L., Zhu, W., Berendzen, K. W., and Liu, C. (2019). Plant lamin-like proteins mediate chromatin tethering at the nuclear periphery. *Genome Biology*, 20(1) :1–18.

[Huan Liu, 2004] Huan Liu, L. Y. (2004). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4) :491 – 502.

[Ioannou et al., 2017] Ioannou, D., Millan, N. M., Jordan, E., and Tempest, H. G. (2017). A new model of sperm nuclear architecture following assessment of the organization of centromeres and telomeres in three-dimensions. *Scientific Reports*, 7(December 2016) :1–14.

[Jarvis, 1973] Jarvis, R. A. (1973). On the identification of the convex hull of a finite set of points in the plane. *Information Processing Letters*, 2(1) :18–21.

[Jonkman et al., 2020] Jonkman, J., Brown, C. M., Wright, G. D., Anderson, K. I., and North, A. J. (2020). Tutorial : guidance for quantitative confocal microscopy. *Nature Protocols*, 15(5) :1585–1611.

[Kamentsky et al., 2011] Kamentsky, L., Jones, T. R., Fraser, A., Bray, M. A., Logan, D. J., Madden, K. L., Ljosa, V., Rueden, C., Eliceiri, K. W., and Carpenter, A. E. (2011). Improved structure, function and compatibility for cellprofiler : Modular high-throughput image analysis software. *Bioinformatics*, 27(8) :1179–1180.

[Kemeny et al., 2018] Kemeny, S., Tatout, C., Salaun, G., Pebrel-Richard, C., Goumy, C., Ollier, N., Maurin, E., Pereira, B., Vago, P., and Gouas, L. (2018). Spatial organization of chromosome territories in the interphase nucleus of trisomy 21 cells. *Chromosoma*, 127(2) :247–259.

[Lanctôt et al., 2007] Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G., and Cremer, T. (2007). Dynamic genome architecture in the nuclear space : Regulation of gene expression in three dimensions. *Nature Reviews Genetics*, 8(2) :104–115.

[Legland et al., 2016] Legland, D., Arganda-Carreras, I., and Andrey, P. (2016). MorphoLibJ : Integrated library and plugins for mathematical morphology with ImageJ. *Bioinformatics*, 32(22) :3532–3534.

[Letourneau et al., 2014] Letourneau, A., Santoni, F. A., Bonilla, X., Sailani, M. R., Gonzalez, D., Kind, J., Chevalier, C., Thurman, R., Sandstrom, R. S., Hibaoui, Y., Garieri, M., Popadin, K., Falconnet, E., Gagnebin, M., Gehrig, C., Vannier, A., Guipponi, M., Farinelli, L., Robyr, D., Migliavacca, E., Borel, C., Deutsch, S., Feki, A., Stamatoyannopoulos, J. A., Herault, Y., Van Steensel, B., Guigo, R., and Antonarakis, S. E. (2014). Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature*, 508(7496) :345–350.

[Lindenbaum et al., 2011] Lindenbaum, P., Le scouarnec, S., Portero, V., and Redon, R. (2011). Knime4Bio : A set of custom nodes for the interpretation of next-generation sequencing data with KNIME. *Bioinfor-*

matics, 27(22) :3200–3201.

[Makarevich et al., 2010] Makarevich, A. V., Kubovicova, E., Sirotkin, A. V., and Pivko, J. (2010). Demonstration of the effect of epidermal growth factor on ram sperm parameters using two fluorescent assays. *Veterinarni Medicina*, 55(12) :581–589.

[Manfrevola et al., 2021] Manfrevola, F., Guillou, F., Fasano, S., Pierantoni, R., and Chianese, R. (2021). Lincking the nuclear envelope to sperm architecture. *Genes*, 12(5).

[Mao et al., 2011] Mao, Y. S., Zhang, B., and Spector, D. L. (2011). Biogenesis and function of nuclear bodies. *Trends Genet*, 27(8) :295–306.

[Martin, 2005] Martin, W. (2005). Archaebacteria (Archaea) and the origin of the eukaryotic nucleus. *Current Opinion in Microbiology*, 8(6) :630–637.

[Miura and Nørrelykke, 2021] Miura, K. and Nørrelykke, S. F. (2021). Reproducible image handling and analysis. *The EMBO Journal*, 40(3) :1–14.

[Nobuyuki Otsu, 1979] Nobuyuki Otsu (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern*, 9(1) :62–66.

[Paszke et al., 2017] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

[Pecinka et al., 2004] Pecinka, A., Schubert, V., Meister, A., Kreth, G., Klatte, M., Lysak, M. A., Fuchs, J., and Schubert, I. (2004). Chromosome territory arrangement and homologous pairing in nuclei of Arabidopsis thaliana are predominantly random except for NOR-bearing chromosomes. *Chromosoma*, 113(5) :258–269.

[Poulet et al., 2015] Poulet, A., Arganda-Carreras, I., Legland, D., Probst, A. V., Andrey, P., and Tatout, C. (2015). NucleusJ : An ImageJ plugin for quantifying 3D images of interphase nuclei. *Bioinformatics*, 31(7) :1144–1146.

[Poulet et al., 2017] Poulet, A., Probst, A. V., Graumann, K., Tatout, C., and Evans, D. (2017). Exploring the evolution of the proteins of the plant nuclear envelope. *Nucleus*, 8(1) :46–59.

[Probst et al., 2009] Probst, A. V., Dunleavy, E., and Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. *Nature Reviews Molecular Cell Biology*, 10(3) :192–206.

[Rabl, 1885] Rabl, C. (1885). Uber Zelltheilung. *Morphol. Jahrb.*, 10 :214–330.

[Ridler and Calvard, 1978] Ridler, T. and Calvard, S. (1978). Picture Thresholding Using An Interactive Selection Method. *IEEE Transactions on Systems, Man and Cybernetics*, smc-8(8) :630–632.

[Rodriguez-Granados et al., 2016] Rodriguez-Granados, N. Y., Ramirez-Prado, J. S., Veluchamy, A., Latrasse, D., Raynaud, C., Crespi, M., Ariel, F., and Benhamed, M. (2016). Put your 3D glasses on : Plant chromatin is on show. *Journal of Experimental Botany*, 67(11) :3205–3221.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net : Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351 :234–241.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3) :211–252.

[Schindelin et al., 2009] Schindelin, J., Arganda-Carrera, I., Frise, E., Verena, K., Mark, L., Tobias, P., Stephan, P., Curtis, R., Stephan, S., Benjamin, S., Jean-Yves, T., Daniel, J. W., Volker, H., Kevin, E., Pavel, T., and Albert, C. (2009). Fiji - an Open platform for biological image analysis. *Nature Methods*, 9(7).

[Skinner and Johnson, 2017] Skinner, B. M. and Johnson, E. E. (2017). Nuclear morphologies : their diversity and functional relevance. *Chromosoma*, 126(2) :195–212.

[Tamura et al., 2010] Tamura, K., Fukao, Y., Iwamoto, M., Haraguchi, T., and Hara-Nishimura, I. (2010). Identification and characterization of nuclear pore complex components in Arabidopsis thaliana. *Plant Cell*, 22(12) :4084–4097.

[Thomas, 2010] Thomas, C. (2010). Origin of the cell nucleus, mitosis and sex : roles of intracellular coevolution. *Biology Direct*, 5 :1–78.

[von Chamier et al., 2019] von Chamier, L., Laine, R. F., and Henriques, R. (2019). Artificial intelligence for microscopy : What you should know. *Biochemical Society Transactions*, 47(4) :1029–1040.

[W. S. McCULLOCH and PITTS, 2007] W. S. McCULLOCH and PITTS, W. H. (2007). WHAT THE FROG'S EYE TELLS THE FROG'S BRAIN. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 59(1) :1–29.

[Wada, 2018] Wada, M. (2018). Nuclear movement and positioning in plant cells. *Seminars in Cell and Developmental Biology*, 82 :17–24.

[Wang et al., 2013] Wang, H., Dittmer, T. A., and Richards, E. J. (2013). Arabidopsis CROWDED NUCLEI (CRWN) proteins are required for nuclear size control and heterochromatin organization. *BMC Plant Biology*, 13(1) :1–13.

[Ward and Coffey, 1989] Ward, W. S. and Coffey, D. S. (1989). Identification of a sperm nuclear annulus : a sperm DNA anchor. *Biology of Reproduction*, 41(2) :361–370.

[Yang et al., 2020] Yang, K., Wang, L., Le, J., and Dong, J. (2020). Cell polarity : Regulators and mechanisms in plants. *Journal of Integrative Plant Biology*, 62(1) :132–147.

[Zalensky et al., 1995] Zalensky, A. O., Allen, M. J., Kobayashi, A., Zalenskaya, I. A., Balhorn, R., and Bradbury, E. M. (1995). Well-defined genome architecture in the human sperm nucleus. *Chromosoma*, 103(9) :577–590.

[Zhou et al., 2012] Zhou, X., Graumann, K., Evans, D. E., and Meier, I. (2012). Novel plant SUN-KASH bridges are involved in RanGAP anchoring and nuclear shape determination. *Journal of Cell Biology*, 196(2) :203–211.

[Zink et al., 2004] Zink, D., Fischer, A. H., and Nickerson, J. A. (2004). Nuclear structure in cancer cells. *Nature Reviews Cancer*, 4(9) :677–687.