



The versatility of high-content high-throughput time-lapse screening data : developing generic methods for data re-use and comparative analyses

Alice Schoenauer Sebag

► To cite this version:

Alice Schoenauer Sebag. The versatility of high-content high-throughput time-lapse screening data : developing generic methods for data re-use and comparative analyses. Other. Ecole Nationale Supérieure des Mines de Paris, 2015. English. NNT : 2015ENMP0035 . tel-01297853

HAL Id: tel-01297853

<https://pastel.hal.science/tel-01297853>

Submitted on 5 Apr 2016

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

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

École doctorale n° 432: Sciences des métiers de l'ingénieur

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité doctorale "Bio-informatique"

présentée et soutenue publiquement par

Alice Schoenauer Sebag

le 4 décembre 2015

Développement de méthodes pour les données de cribles temporels à haut contenu et haut débit: versatilité et analyses comparatives

The versatility of high-content high-throughput time-lapse screening data: developing generic methods for data re-use and comparative analyses

Directeur de thèse : **Jean-Philippe Vert et Robert Barouki**

Co-encadrant de thèse : **Thomas Walter**

Jury

M. Wolfgang HUBER,	Docteur, EMBL	Rapporteur
M. Miguel LUENGO-OROZ,	Docteur, Univ. Polit. de Madrid	Rapporteur
M. Bernard SALLES,	Professeur, INRA	Examineur
M. Jean-Philippe VERT,	Ing. en chef des Mines, MINES ParisTech	Examineur
M. Thomas WALTER,	Docteur, MINES ParisTech	Examineur
M. Robert BAROUKI,	Professeur, Univ. Paris Descartes	Examineur

MINES ParisTech

Centre de Bio-Informatique (CBIO)

35 rue Saint-Honoré, 77300 Fontainebleau, France

Declaration of Authorship

I, Alice SCHOENAUER SEBAG, declare that this thesis titled, 'The versatility of high-content high-throughput time-lapse screening data: developing generic methods for data re-use and comparative analyses' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

MINES PARISTECH

Abstract

Department of Computational Biology
Mines Paristech

Doctor of Philosophy

**The versatility of high-content high-throughput time-lapse screening data:
developing generic methods for data re-use and comparative analyses**

by Alice SCHOENAUER SEBAG

Biological screens test large sets of experimental conditions with respect to their specific biological effect on living systems.

Live cell imaging is an excellent tool to study in detail the consequences of chemical perturbation on a given biological process. However, the analysis of live cell screens demands the combination of robust computer vision methods and quality control procedures, and efficient statistical approaches for the detection of significant effects.

This thesis addresses these challenges by developing analytical methods for High Throughput time-lapse microscopy screening data. The developed frameworks are applied to publicly available HCS data, demonstrating their applicability and the benefits of HCS data remining. The first multivariate workflow for the study of single cell motility in such large-scale data is detailed in Chapter 2. Chapter 3 presents this workflow application to previously published data, and the development of a new distance for drug target inference by in silico comparisons of parallel siRNA and drug screens. Finally, chapter 4 presents a complete methodological pipeline for performing HT time-lapse screens in Environmental Toxicology.

Acknowledgements

Thank you. I just feel so grateful.

Mariah Carey

Vittoria ! Vittoria !/L'alba vindice appar [...] !

Luigi Illica et Giuseppe Giacosa

Comme parfaitement résumé ci-dessus, j'aimerais remercier toutes les personnes qui m'ont permis de finir – la tête haute, entourée et peut-être un peu meilleure qu'au début – plus patiente, plus rigoureuse, plus précise, plus détendue entre autres.

Je voudrais, pour parler de vous, inventer des mots plus vibrants, plus respectueux et plus tendres. André Gide

Et la première personne que je voudrais remercier pour m'avoir guidée jusque là est Thomas. Pour sa patience avec moi, sa gentillesse, son écoute. Pour avoir tant pris soin de mon état au point d'être aussi triste ou excité que moi des résultats ou des manip. Pour sa rigueur et son amour de la précision, des bullet points tous identiques et de la visualisation de données qui répond à ce qu'on lui demande. Pour tous les déjeuners où il a renoncé à la cantine pour m'accompagner, pour ces voyages qu'il m'a permis de faire (Louvain, Stockholm, Tübingen, Dublin, New York, Vienne, Munich, Lake Tahoe), pour ta présence, ton soutien inépuisable, ton temps et ta bonté. Pour avoir supporté mon obstination et mon humeur de gueux par les temps pluvieux (et parfois même ensoleillés). Merci.

I would like to thank the jury for their time and interest in my work.

Je voudrais aussi remercier Jean-Philippe pour m'avoir accueillie au CBIO, puis toute son aide et ses encouragements ML-esques, ainsi que sur la suite. Pour le free ride de « NIPS » 2012 ! Rien n'aurait été possible non plus sans Robert qui a accepté de prendre en thèse cet OVNI un peu matheux, et Céline (RT) grâce à qui j'ai appris les rudiments de la biologie cellulaire. Que d'espoirs partagés, que de temps à s'esquinter les yeux en salle de microscopie, que de déceptions aussi : merci pour ton temps et ta patience ! Rien non plus sans Olivier et Céline (D) qui m'ont tout appris sur le microscope, et avec qui j'ai aussi partagé de longs moments à tout régler, re-régler, dérégler, ... Et Martine bien

sûr, qui malgré sa todo liste infinie a toujours trouvé le temps de me montrer quelque chose ou de nourrir nos cellules.

Merci aux « grands », Anne-Claire et Emile, pour les retours d'expérience, Nelle pour toute son aide avec Python – toujours prête à filer un coup de main même au milieu d'une grosse deadline, Elsa, Matahi. Véronique pour son aide, et tout le CBIO. Alix, Ludmila et Eléonore, Sophie et son équipe pour leur accueil très chaleureux pendant les quelques mois que j'ai passés dans leur bureau. Merci à toute l'unité 1124 !

Merci à Cyril Kao qui a aidé ce projet à prendre forme, au Corps des ponts, des eaux et des forêts de l'avoir accepté, et au Ministère de l'Ecologie, du Développement Durable et de l'Energie pour l'avoir financé. Merci à Sandra Plancade, Marco Cuturi, Rudi Höfler, Christoph Sommer, Wolfgang Huber, Beate Neumann pour leur aide et leurs réponses à mes questions durant ces trois dernières années. Merci à Luc Tamisier, à l'assistance CBIO ainsi qu'à Jean Duong. Merci à Shift en général et Eric et Jawish en particulier pour leur flexibilité ces derniers mois, et Pauline pour avoir supporté avec le sourire les détails administratifs !

En premier lieu aussi je voudrais remercier mes parents, ma sœur et Jean. Le lycée, la prépa, sautes d'humeurs (doux euphémisme) et crises existentielles, ils ont à peu près tout enduré et ils m'ont toujours soutenue... Papa dispo pour faire des exos de maths le samedi de 14h à 20h et 21h à 23h, pour discuter optimisation à la piscine, pour m'aider à déménager, toujours prêt à sauver mon monde. Maman à le refaire, en commençant par la fonction objectif : également là pour la méditation et la reprise à zéro de la formalisation du problème... Mathilde, dispo pour déjeuner, un ciné, Garnier, me remonter le moral, se moquer des imbéciles, un sms scato et des blagues niveau CM2, tout ce dont j'ai eu et aurai besoin pour moins me prendre au sérieux, réaliser que rien n'est tragique. Jean qui depuis 1998 (date du premier billet d'opéra que j'aie gardé) m'a ouvert un monde qui est une de mes raisons d'être, ce qui me redonne le sens quand je crois l'avoir perdu : Garnier Bastille TCE Pleyel Philharmonie Cité de la musique Chaillot Châtelet : autant d'endroits où je me suis construite grâce à toi, qui m'ont permis de continuer à sourire ou à bosser en me disant que ce meilleur est toujours à venir !

En premier lieu toujours je voudrais remercier mes amis.

Ceux de (presque) toujours : Nico pour tous les Skypes à des heures presque indues, pour son soutien à toute épreuve malgré la distance, Loulou pour sa présence critique et les soirées de la rue des Quatre Fils, Gégé pour toutes nos soirées londoniennes ou sur What'sApp, sans oublier nos folles manucures, Fred, Dimi, Rapha, Flo.

Mes coupains chéris à la folie pour tous nos dîners entre Paris, Munich, la Baule et Montpellier : Margaux, Nicolas et Roro, dîners où je suis parfois (souvent ?) arrivée avec le sourire d'un boule-dogue ou dans un état à ramasser à la petite cuillère, ce que vous avez toujours fait avec patience, délicatesse, amour et une pincée de taquinerie. Pour tous les dîners où je voulais manger une feuille d'épinard sans sauce accompagnée d'eau gazeuse et où vous m'avez supportée...

Kambiz ! Mein Lieblingskankun, merci pour ta présence. Laure, comme tu me manques depuis que tu habites à cinq minutes de l'océan Pacifique... Merci Ramon pour ton soutien infaillible et ton intérêt pour la motilité et l'amotilité. Le crew de choc, le crew de mon premier Memphis et des Chat Noir, de la Keszino, des livres blancs de la Défense et de la chevauchée des Walkyries à la CK : Vincent, Max, Anna, Charles, Timothée et Titi, merci d'avoir été là ! Mes copines rochelaises : Noémie et son regard éclairé sur ma vie, Olivia et son grand coeur aussi cheesylove que son mix que j'écoute tout le temps, Greg ;). Keurkeurkeur

Merci à Sandra, Charlotte et Judith pour leur soutien ces derniers mois. Mes copains opéresques : Pierre Juliette et Sylvain, merci pour votre soutien sans faille et vos taquineries Kangoo-related ! Thank you Andrei for your total confidence and support. Ma Clairounette pour tous ces moments partagés : aropiens, Gibus, piknik elektronik sous la pluie entre autres.

Et ceux que j'aurai pu oublier...

Contents

Declaration of Authorship	i
Abstract	ii
Acknowledgements	iii
.	iii
.	iv
.	iv
Contents	vi
List of Figures	ix
List of Tables	xiii
1 Introduction	1
1.1 Data sets	5
1.1.1 Mitochondria data set	6
1.1.2 PCNA data set	6
1.1.3 Drug screen	8
1.1.4 Xenobiotic screen	8
1.2 Software	8
2 A generic methodological framework for studying single cell motility	10
2.1 Studying single cell motility in a HT setup	11
2.2 MotIW overview	12
2.2.1 Segmentation and tracking	13
2.2.1.1 Segmentation	14
2.2.1.2 Cell tracking by supervised learning	14
2.2.1.3 Validation of MotIW cell tracking model	16
2.2.2 Trajectory features	18
2.2.2.1 Particle motion features	18
2.2.2.2 Other global features	21
2.2.2.3 Averaged local features	21
2.2.2.4 Feature set evaluation	21

2.2.3	Statistical procedure	23
2.2.3.1	Trajectory quality control	23
2.2.3.2	False discovery rate control	24
2.2.3.3	Formal statistical procedure	24
2.3	Validation on a simulated screen	26
2.3.1	Screen simulation	26
2.3.2	Application to a simulated screen	27
3	High-content screening data as a resource	29
3.1	Data re-use in Bioimage Informatics	30
3.2	MotIW reveals modes of movement and genes involved in nuclear motility	32
3.2.1	Hit list	32
3.2.1.1	Functional analysis	33
3.2.1.2	Intersection with other published motility gene lists . . .	34
3.2.2	Cell trajectory ontology	35
3.3	Cell cycle length study	37
3.3.1	Complete cell cycle detection	38
3.3.2	Cell cycle length hit list	39
3.3.3	Discussion	40
3.4	Functional inference by in silico comparison of small-molecule and siRNA screens	45
3.4.1	Materials and methods	46
3.4.1.1	Experimental work	46
3.4.1.2	Object segmentation	47
3.4.1.3	Object classification and phenotypic scores	47
3.4.1.4	Quality control	48
3.4.1.5	Selection of Mitocheck experiments for target inference .	48
3.4.1.6	Detection of drug screen hit experiments	49
3.4.1.7	Other analyses	50
3.4.2	Phenotypic profile distances	51
3.4.2.1	Euclidean distance on phenotypic scores	52
3.4.2.2	Phenotypic trajectory distance	53
3.4.2.3	Sinkhorn divergence	53
3.4.2.4	Distance quality evaluation	57
3.4.3	Applications	59
3.4.3.1	Small molecule similarity evaluation	60
3.4.3.2	Target pathway inference	66
3.4.4	Discussion	67
4	Xenobiotic screen	71
4.1	Materials and methods	74
4.1.1	Experimental work	74
4.1.1.1	Chemicals	74
4.1.1.2	Cell culture	74
4.1.1.3	Cell transfection and clonal selection	75
4.1.1.4	Production of 96 well-plate for imaging	75

4.1.1.5	Time-lapse imaging	76
4.1.1.6	Phototoxicity assays	77
4.1.1.7	Chemical dose choice	77
4.1.2	Bioinformatics methods	78
4.1.2.1	Web-based user interface for result visualization	78
4.1.2.2	Quality control	78
4.1.2.3	Object segmentation	79
4.1.2.4	Object feature extraction	81
4.1.2.5	Object classification	81
4.1.2.6	Object tracking and trajectory feature extraction	83
4.2	Results	83
4.2.1	Preliminary choices	83
4.2.2	Motility study	83
4.2.3	Phenotypic study	86
4.2.3.1	Phenotypic class selection	86
4.2.3.2	Results	86
4.3	Discussion	87
5	Conclusion	90
A	Appendices	93
A.1	Cell cycle gene list	93
A.2	Functional inference by in silico comparison of small-molecule and siRNA screens	95
A.2.1	Choice of λ parameter	95
A.2.2	Phenotypic scores of JNJ7706621	97
A.2.3	Two-dimensional hierarchical clustering of drug screen condition distance to Mitocheck siRNAs for different distances	98
A.2.4	Two-dimensional hierarchical clustering of drug screen hit condition distance to Mitocheck siRNAs for different distances	100
A.3	Literature review	105
A.4	Phenotypic study	107

List of Figures

1.1	Images of a wild-type <i>C. elegans</i> worm (top left, nuclear staining, [Biol., 2005]), 12 <i>D. melanogaster</i> larvae (top right, colors: spatial repartition of different mRNAs, [Lecuyer et al., 2007]) and human breast cancer cells (bottom, red: DNA, green: cytoplasmic membrane, our data)	3
1.2	Image from a control video of the Mitocheck dataset (white: histone 2B) .	7
1.3	Image from a control video of the PCNA dataset (white: histone 2B, green: PCNA)	7
2.1	Overview of MotIW	12
2.2	Illustration of angular match features.	15
2.3	Details of tracking precision and recall according to event types	18
2.4	A cell trajectory with notations	19
2.5	Convex hull of the example track from figure 2.4	22
2.6	Heatmap showing trajectory feature similarities on a subset of the Mitocheck dataset (1.1 million trajectories coming from detected motility hit experiments according to MotIW). The dengrograms were obtained using the <i>Ward</i> method and the euclidean distance between feature correlations.	22
2.7	Simulated trajectories: stop-and-go (green), flip-directed (red), random (orange), fast random (purple), curbed-directed (blue)	26
3.1	Evaluation of k-means clustering quality as a function of the number of clusters (average and standard deviation on 10 algorithm initializations). The same protocol was applied to a subset of the Mitocheck dataset, and two samples of the same dimensions, respectively drawn from the Uniform and the Normal distributions.	36
3.2	Comparison of cluster distributions between controls (Ctrl) and experiments (Exp) for the eight trajectory clusters which were identified in the Mitocheck dataset. The clusters are in the same order as in figure 3.3. . .	37
3.3	Characterization of our ontology of trajectories. Each column is a single cell trajectory ; trajectories are grouped by cluster label. 1,000 trajectories were randomly selected per trajectory cluster.	38
3.4	Examples of object divisions from the Mitocheck dataset	38
3.5	Approach to cell cycle study	39
3.6	Histograms showing cell cycle length for complete (top) and incomplete (bottom) trajectories, for two experiments of the Mitocheck dataset concerning ARSF which were detected as significantly different from controls for cell cycle length.	41

3.7	Histograms showing cell cycle length for complete (top) and incomplete (bottom) trajectories, for two experiments of the Mitocheck dataset concerning CACNA1D (left) and DIMT1 (right), which were detected as significantly different from controls for cell cycle length.	42
3.8	Example of the time evolutions of nuclear size ("roisize", top left and bottom) and nuclear intensity ("total intensity", top right) for all complete trajectories of a control experiment from the Mitocheck dataset. As discussed in the text, no clear slope break is seen for most trajectories for any of the two indicators, hence preventing the delimitation of cell cycle phases using only this information.	43
3.9	DNA intensity and size as provided by H2B-GFP information is not sufficient to differentiate between the different cell cycle phases. Data and labelling come from the PCNA dataset.	44
3.10	Precision and recall per class as provided by Cell Cognition. Compared with the original classifier as published in [Walter et al., 2010], classes <i>ADCCM</i> (Asymmetric Distribution of Condensed Chromosome Masses) and <i>Out of focus</i> were added. More nuclei were furthermore included for training in most classes. <i>Shape1</i> (resp. <i>Shape3</i> , <i>MetaphaseAlignment</i>) corresponds to binucleated (resp. polylobed, metaphase alignment problem) nuclei.	47
3.11	Number of hit genes per category. As hit detection is univariate, a gene can be in more than one category.	49
3.12	Distributions of phenotypic scores from the drug screen experiments. Each boxplot corresponds to the distribution of control phenotypic scores, whereas each red dot is an experiment in which cells were exposed to a drug. . . .	50
3.13	Cost matrix for phenotypic Sinkhorn divergence	56
3.14	Convergence of Sinkhorn divergence as a function of lambda. Divergences were computed between drug screen experiments and Mitocheck hit experiments for different values of lambda, and the distribution of their relative variation to the divergences computed for $\lambda = 30$ are showed here. . . .	57
3.15	Separation between Mitocheck hit categories (left) for $\lambda = 0.1$. Global Sinkhorn divergences between Mitocheck hit experiments were computed for $\lambda = 0.1$, and multi-dimensional scaling was used for representing them in two dimensions in the first two lines. Divergences between these experiments and the drug screen were included and their multi-dimension scaling is shown on the right plot.	58
3.16	Separation between Mitocheck hit categories (left) for $\lambda = 10$. Global Sinkhorn divergences between Mitocheck hit experiments were computed for $\lambda = 10$, and multi-dimensional scaling was used for representing them in two dimensions in the first two lines. Divergences between these experiments and the drug screen were included and their multi-dimension scaling is shown on the right plot.	58
3.17	Mean separability and replicability scores of investigated distances on all conditions (left) and hit conditions only (right - bars represent standard deviations).	59
3.18	Drug screen condition - Mitocheck siRNA two-dimensional hierarchical clustering using global Sinkhorn divergence. Ward method was used in combination with the Euclidean distance.	60

3.19	Drug screen hit condition - Mitocheck siRNA two-dimensional hierarchical clustering using global Sinkhorn divergence. Ward method was used in combination with the Euclidean distance.	62
3.20	Visualization of condition clustering for phenotypic score distance. A black dot means that the conditions belong to the same cluster, a white dot that they do not.	64
3.21	Visualization of condition clustering for global Sinkhorn divergence. A black dot means that the conditions belong to the same cluster, a white dot that they do not.	65
4.1	Xenobiotic screen complete workflow	74
4.2	Illustration of the experimental settings	76
4.3	Example of a plate setup. <i>Cl1</i> : clone number, <i>Indpt 10</i> : CO_2 -independent cell medium with 10% FCS.	76
4.4	Diagram of the databases for experimental metadata storage	79
4.5	Trajectory feature heatmaps corresponding to control wells. Each line corresponds to an experiment, whose plate and name are indicated. Each column corresponds to a trajectory feature. A robust normalization (with median and inter-quartile range) was applied, using all plate, which still permits to see that control responses vary from well to well and plate to plate.	84
4.6	Interphase (up) and frozen(down) distances. Colors are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. <i>Legend</i> : B: BPA, D: DMSO, E: Endo, M: MeHg, N: Nonane, P: PCB, R: nothing, T(red): TCDD, T(magenta): TGF- β 1	88
A.1	Separation between Mitocheck hit categories for $\lambda = 0.1$. Global Sinkhorn divergences between Mitocheck hit experiments were computed for $\lambda = 0.1$, and multi-dimensional scaling was used for representing them in two dimensions in the first two lines. Divergences between theses experiments and the drug screen were included and their multi-dimension scaling is showed on the last line (grey: controls).	95
A.2	Separation between Mitocheck hit categories for $\lambda = 10$. Global Sinkhorn divergences between Mitocheck hit experiments were computed for $\lambda = 10$, and multi-dimensional scaling was used for representing them in two dimensions in the first two lines. Divergences between theses experiments and the drug screen were included and their multi-dimension scaling is showed on the last line (grey: controls).	96
A.3	Phenotypic scores of JNJ7706621 experiments, as a function of plate (left, middle, right) and dose (abscissa). The redder a square, the further away from control phenotypic scores.	97
A.4	Drug screen condition - Mitocheck siRNA two-dimensional hierarchical clustering using sum of time Sinkhorn divergence. Ward method was used in combination with the Euclidean distance.	98
A.5	Drug screen condition - Mitocheck siRNA two-dimensional hierarchical clustering using phenotypic trajectory distance. Ward method was used in combination with the Euclidean distance.	98
A.6	Drug screen condition - Mitocheck siRNA two-dimensional hierarchical clustering using Euclidean distance of phenotypic scores. Ward method was used in combination with the Euclidean distance.	99

A.7	Drug screen hit condition - Mitochondrial siRNA two-dimensional hierarchical clustering using sum of time Sinkhorn divergence. Ward method was used in combination with the Euclidean distance.	100
A.8	Corresponding visualization of condition clustering for time Sinkhorn divergence. A black dot means that the conditions belong to the same cluster, a white dot that they do not.	101
A.9	Drug screen hit condition - Mitochondrial siRNA two-dimensional hierarchical clustering using phenotypic trajectory distance. Centroid method was used in combination with the Euclidean distance.	102
A.10	Corresponding visualization of condition clustering for phenotypic trajectory distance. A black dot means that the conditions belong to the same cluster, a white dot that they do not.	103
A.11	Drug screen hit condition - Mitochondrial siRNA two-dimensional hierarchical clustering using Euclidean distance of phenotypic scores. Ward method was used in combination with the Euclidean distance.	104
A.12	Apoptosis distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates. . . .	108
A.13	Frozen distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates. . . .	109
A.14	Interphase distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates. . . .	110
A.15	Metaphase distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates. . . .	111
A.16	Micronucleated distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates. . . .	112
A.17	Prometaphase distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates. . . .	113

List of Tables

1	Abbreviations	xiv
2	Main definitions	xiv
2.1	Mean recall and precision on all types of matches E (10-fold cross-validation)	17
2.2	Cell trajectory features and their formulas. Notations: $(m_t)_{t=1\dots T}$, time sequence of cell 2D positions. T, track time duration. P, total track length	19
2.3	Results from the application of MotIW to simulated data	27
3.1	Existing medium- to high-throughput studies of cell motility	34
3.2	Hit list intersections	34
3.3	Selected drugs and dose ranges. All drugs were tested for 11 doses.	46
3.4	Known protein targets of hit drugs (bold: present in Mitocheck hit experiments). Drugs are grouped by target similarity. Source: DrugBank [Wishart et al., 2008] unless specified.	63
3.5	Rank of known drug targets, when applicable. Condition group are identical to that in the text.	66
3.6	Rank of known drug targets, when applicable. Condition group are identical to that in the text.	68
4.1	Nuclear morphology classes with examples. The <i>artefact</i> and <i>cluster</i> examples are shown with segmentation contours.	82
4.2	Chemical dilutions	83
4.3	<i>Rank product</i> p-values for different phenotypic distances (<0.05)	87
A.1	Cell cycle gene list. In bold are the three genes for which we found an extension of cell cycle length.	94
A.2	Human xenobiotic levels	105

TABLE 1: Abbreviations

ADCCM	Asymmetric Distribution of Condensed Chromosome Masses
AURKA,B	Aurora kinase A, B
BPA	bisphenol-A
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
ds	double stranded
Endo	α -endosulfan
FACS	fluorescence-activated cell sorter
FCS	fetal calf serum
H2B	histone 2B
HC	high-content
HCS	high-content screening
HT	high-throughput
MeHg	methylmercury
MSD	mean squared displacement
PCB153	2,2',4,4',5,5'-Hexachlorobiphenyl
PCNA	proliferating cell nuclear antigen
RNA	ribonucleic acid
RT-qPCR	real-time quantitative polymerase chain reaction
TCDD	2,3,7,8-Tetrachlorodibenzo-p-dioxin
TOP1	topoisomerase (DNA) I

TABLE 2: Main definitions

Condition	Combination of a dose and a chemical
Xenobiotic	Any chemical which is foreign to an organism, i.e. is neither produced nor expected to be found in it.

A Mine et Baba, Tita et Pigeo

Chapter 1

Introduction

Résumé - Introduction (see *infra* for English text)

De récents progrès en chimie organique et en biologie moléculaire ont permis la constitution de géantes librairies de molécules, qui sont respectivement des médicaments potentiels et des composés conçus pour aboutir à la sur- ou sous-expression d'un gène d'un organisme donné. Ces produits nécessitent d'être testés : il importe de vérifier la toxicologie comme l'effet attendu des médicaments potentiels. D'autre part, l'existence de librairies de petites molécules comme de siRNAs permettent de tester systématiquement la fonction des gènes d'une espèce.

Des milliers de nouveaux produits chimiques sont par ailleurs synthétisés à des fins industrielles chaque année. Il importe également de les tester, ce qui fait de la toxicologie environnementale un troisième champ d'application majeur pour les cribles biologiques.

Un crible biologique est un ensemble d'expériences conçu pour tester en parallèle les effets de plusieurs composés sur une action biologique spécifique dans un organisme donné. Une expérience de vidéomicroscopie à épifluorescence est une expérience durant laquelle de multiples images d'échantillons fluorescents sont acquises au cours du temps. Cela fournit beaucoup d'informations supplémentaires par rapport à la prise d'une unique image, comme par exemple l'observation d'événements rares ou encore l'ordonnancement des événements observés.

Les données de cribles biologiques à haut débit réalisés avec de la vidéomicroscopie sont donc très riches ; leur analyse nécessite la mise au point de techniques statistiques multivariées robustes. La question est donc de savoir comment développer optimalement des méthodes analytiques pour de telles données. Nous présenterons dans le chapitre 2 le premier workflow pour l'étude de la motilité cellulaire individuelle dans de telles données. Le chapitre 3 appliquera ce workflow ainsi que d'autres aux données du projet Mitochek [Neumann et al., 2010], démontrant l'utilité de la ré-utilisation de telles données. Enfin, le chapitre 4 développera une approche méthodologique pour l'analyse de données de vidéomicroscopie à fin de criblage en toxicologie environnementale.

Biological screens

Progress in organic chemistry and molecular biology have led to the constitution of giant libraries, respectively of putative drugs and potential biologically active small molecules, and engineered organisms or proteins for gene silencing or overexpression. As an example, the Biomolecular Screening Facility of the EPFL (Lausanne, Switzerland) has a collection of 65,000 compounds and 130,000 small interfering RNAs (siRNAs¹), while most bio-technological companies offer to ship custom genome-wide siRNA libraries over a fortnight. Pharmaceutical industry libraries are impressive as well, most containing more than one million compounds.

Putative drugs need to be tested for the expected biological effect and against undesirable secondary effects. Understanding their mode-of-action, which screening experiments can help to do, is also a major concern during drug discovery, and often its rate-limiting step [Eggert et al., 2004]. Screens have therefore become a major component of drug discovery processes [Wei et al., 2012]. On the other hand, the development of biomolecular engineering, and small molecule and siRNA libraries at (more) affordable costs, has led to a significant increase in functional genomic screens, which test for gene function and gene relations. For example, [Giaever et al., 2002] exhaustively engineered gene-deletion mutants of the yeast *Saccharomyces cerevisiae*, while [Gwack et al., 2006] performed a genome-wide siRNA screen in the fruit fly *Drosophila melanogaster*.

Finally, progress in organic chemistry has also led to the explosion of the number of new molecules which are synthesized each year for industrial purposes (e.g. pesticides, plastics, food additives). This calls for the development of systematic testing experiments, both for the desired action and against undesirable effects on living organisms. Hence a third major field of application for biological screens is Environmental Toxicology.

Biological screens are experiments which are designed for testing a set of compounds for a specific biological action in a given organism. The latter can be any organism in which the action is easily detected, such as a fish (*Danio rerio*), a fly (*Drosophila melanogaster*), a worm (*Caenorhabditis elegans*) or a human cell (*Homo sapiens*). The biological action which is tested can go from a simple univariate assessment of cell death, to the multivariate quantification of an effect on a complex cellular phenotype such as cell division or motility. This will determine the screen **content**, which would respectively be **low** and **high**. Another important parameter of biological screens are their **throughput**. It ranges from **low**, for example when testing a dozen of carefully selected compounds, to **high** for hundreds of thousands of siRNAs in the case of a genome-wide screen.

¹Small interfering RNAs are short double-stranded RNA molecules which interferes with the expression of genes that present complementary nucleotide sequences.

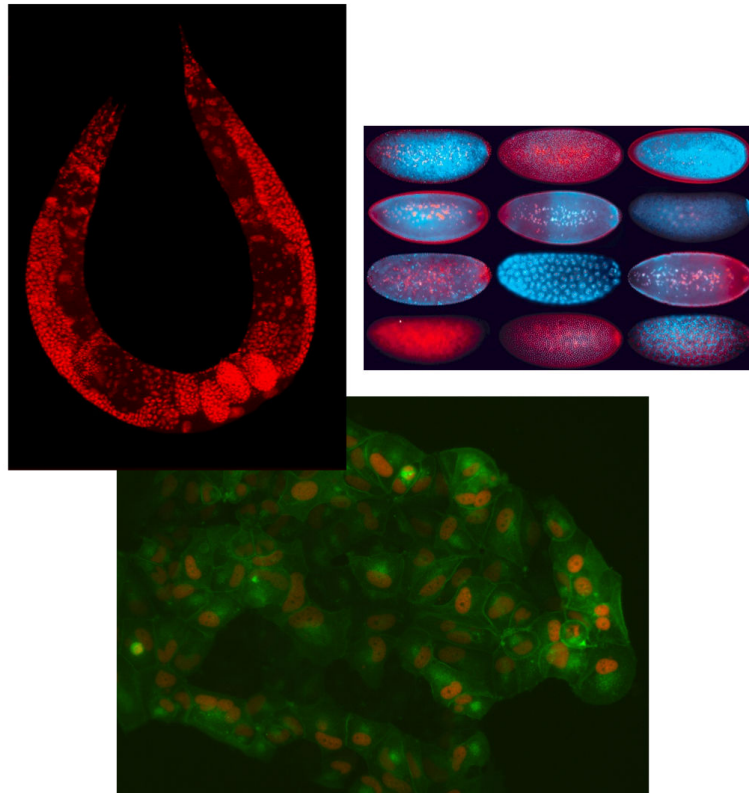


FIGURE 1.1: Images of a wild-type *C. elegans* worm (top left, nuclear staining, [Biol., 2005]), 12 *D. melanogaster* larvae (top right, colors: spatial repartition of different mRNAs, [Lecuyer et al., 2007]) and human breast cancer cells (bottom, red: DNA, green: cytoplasmic membrane, our data)

Time-lapse microscopy

Time-lapse microscopy experiments are experiments in which images are regularly acquired over time. Hence they produce rich 3 or 4-dimensional datasets: they are high-content (HC) almost by definition. Most time-lapse data comes from fluorescent samples. Fluorescent labelling has the advantage of being presently very affordable, (mostly) non-toxic to living organisms and flexible. It makes it possible to easily follow a single protein of interest, which would have been delicate in bright-field images. Furthermore, although it demands either the addition of some reagent when acquiring for a few minutes or hours, or cell genetic modification when acquiring repeatedly for a few days, fluorescence images are easier to segment and analyze than bright field images. This generalized use of fluorescent proteins was permitted by their recent discovery, which started by the green fluorescent protein (GFP) [Chalfie et al., 1994], [Muzzey and van Oudenaarden, 2009].

The use of time-lapse microscopy rather complicates data acquisition: samples need to be maintained in the appropriate atmosphere at the appropriate temperature, and one should be certain to avoid phototoxicity from repeated light exposure. Nevertheless, they provide a wealth of information which it is not possible to access otherwise. First

of all, time-lapse microscopy experiments seem natural for studying dynamic processes such as cell division or cell motility. Furthermore, they enable to visualize very transient events, which are barely observed in endpoints assays, such as early anaphases [Neumann et al., 2010]. Time-lapse microscopy experiments also permit to establish causality links between phenotypes [Perlman et al., 2004]. Indeed, it makes it possible to observe the order in which phenotypes occur, therefore enabling to determine which one leads to the other. It also makes it possible to study cell population heterogeneity in response to gene silencing or chemical exposure: tracking cells over time permits to perform *in silico* cell alignment, and therefore to determine if there exists cell subpopulations with different phenotypic stories. This is not surprising given the stochastic nature of gene expression [Raj and van Oudenaarden, 2008], and can already been found in the Event Order Maps of [Neumann et al., 2010] (although it was not formalized in that direction in the latter).

Despite more complex experimental procedures, time-lapse microscopy therefore has a real advantage over endpoint assays when studying complex phenotypes: they truly permit to functionally dig into the consequences of either gene silencing or chemical exposure. As such, it started being used approximately 15 years ago, although some pioneering studies date back from the 1980s (e.g. [Sulston et al., 1983], see [Muzzey and van Oudenaarden, 2009] for a review on quantitative time-lapse fluorescence microscopy in single cells). We hereafter refer to time-lapse fluorescence microscopy data by time-lapse data.

Analysis of high-throughput time-lapse screening data

High-throughput (HT) screening experiments have only been made recently possible by the development of screening robots, automated microscopes and measurement devices, and relevant software. These tools are necessary for preparing and performing the experiments. In the case of low-content experiments, result analysis remains simple, although it should not be forgotten that statistics on large datasets should not be done as on small ones (see paragraph on the control of false discovery rate in section 2.2.3.2). Quality control procedures shall be included in the pipeline as well.

On top of quality control and large dataset statistics, robust and efficient multivariate analytical methods are necessary to deal with HC HT screening experiments. This applies to HT time-lapse screening data as well. It is indeed a specific type of HC HT screening data, in which one data dimension is time. Most of the time, analyzing such datasets demands to combine computer vision methods to multivariate statistical algorithms for significant effect detection. These methods should be tailored to the biological process which is studied, but they all have in common the challenges to be robust when faced

with noise, and scalable as dataset sizes increase. The question is to know how to develop analytical methods for optimally exploiting such datasets.

A first biological process which seems natural to be studied with time-lapse microscopy is single cell motility. A systematic functional genomics approach to cell motility is all the more needed since all the involved proteins and pathways are not yet known. Nevertheless, it was never studied using multivariate statistical tests in HT settings. During this thesis, we therefore developed a generic methodological workflow for studying single cell motility in HT time-lapse screening data, which will be presented in chapter 2. As will be detailed, an *ad-hoc* statistical procedure indeed had to be developed.

The generic quality of this workflow enabled it to be applied to an existing dataset, the Mitocheck dataset (see section 1.1.1 for a presentation of this dataset). This revealed the quantity of unexploited information in this dataset, and more generally the wealth of existing HT time-lapse screening data which can be re-used to different purposes than the original experimental design. Proofs of this constitutes chapter 3, which presents how an ontology of single cell trajectories could be extracted from the Mitocheck dataset (section 3.2.2). The latter was also used for detecting cell cycle genes (section 3.3). In section 3.4, it also permitted to perform drug target inference on an unpublished time-lapse drug screen, which made it necessary to develop a new distance.

Finally, as was mentioned in the beginning of this introduction, one of the most important applications of screening is in Toxicology. Moreover, HT time-lapse screening experiments have never been performed in Environmental Toxicology. We therefore developed a robust methodological workflow and its visualization Web-interface, for conducting and analyzing HT time-lapse screening data in Environmental Toxicology. This composes chapter 4.

Before diving into the main matters, all the datasets which were analyzed in the course of this thesis, as well as the software which we used to this end, are briefly described in the following section.

1.1 Data sets

Four time-lapse datasets were used in our work, which will briefly be presented in this section: the Mitocheck dataset, which is the first genome-wide siRNA time-lapse screen, a PCNA dataset for the study of cell cycle phases, an unpublished drug screen in similar settings to those of the Mitocheck dataset, and an unpublished xenobiotic screen.

1.1.1 Mitochek data set

The main dataset which we used is a previously published genome-wide data set of time-resolved records of cellular phenotype responses to gene silencing, which were generated for virtually all protein-coding genes [Neumann et al., 2010]. It is publicly available at mitochek.org.

For this, arrays of transfection cocktails containing small interfering RNA (siRNA) were spotted directly into live cell-imaging chambers in a 384 format. HeLa cells (ATCC® CCL-2™) stably expressing the core histone 2B tagged with GFP were seeded on top of the arrays, and imaged 18 h after the transfection for 48 h with a time-lapse of 30 min (Plan10x, NA 0.4; Olympus - see fig. 1.2 for an example). Imaging chambers were sealed during imaging. Each microarray contained 8 negative controls (scrambled: not targeting any gene) and 12 positive controls showing different phenotypes. 22,612 protein-coding genes have been targeted by at least 2 siRNAs each, in total 51,767 siRNAs. For each siRNA, there is data from at least 3 technical replicates, which created 182,191 quality controlled time-lapse experiments in total. Due to updates in the genome annotation, some reagents could not be mapped to the current ENSEMBL version. In total, the data set contains data for 17,816 protein-coding genes in 144,909 quality controlled time-lapse experiments.

HeLa cells are epithelial cancer cells which were derived from the adenocarcinoma of Henrietta Lacks in 1951. Being the first human cell line to survive *ex vivo* for more than a few days, they are very appreciated from cell biologists as they are easy to grow and transfect. Indeed, this cell line is mentioned in approximately 0.3% of PubMed abstracts although 64% of its genome has a copy number greater than three [Adey et al., 2013]. It is not motile as can be the case of other cell lines which are widely used in migration studies such as the epithelial metastatic breast-cancer derived MDA-MB-231 cell line (ATCC® HTB-26™), or the epithelial metastatic lung-cancer derived NCI-H1299 (ATCC® CRL-5803™). Indeed, it was the 16th slowest out of 54 in the first World Cell Race [Maiuri et al., 2012]. Gene silencing in this background therefore makes it easier to identify migration suppressors, that is, genes whose silencing will enhance cell motility, rather than migration enhancers, given that HeLa basal cell motility is rather low.

1.1.2 PCNA data set

In section 3.3, another published dataset is mentioned, which is related to the study of cell cycle phases. It was published with [Held et al., 2010] and is publicly available with annotations².

²<http://www.cellcognition.org/downloads/data>

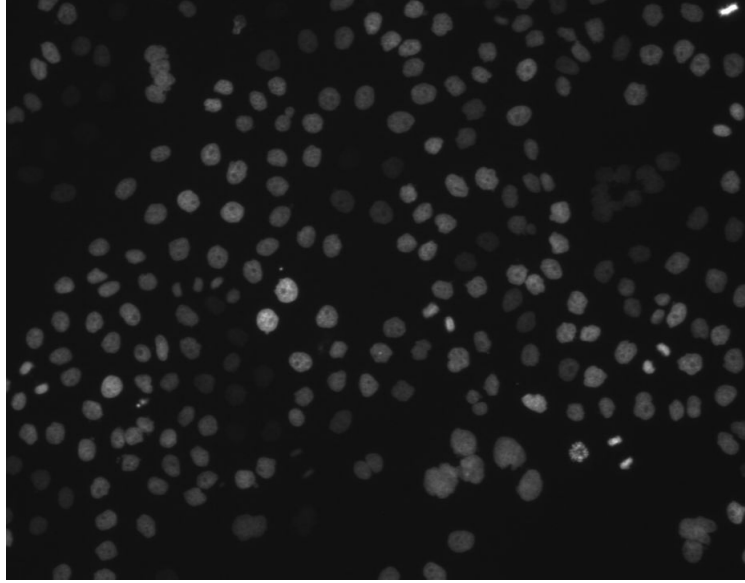


FIGURE 1.2: Image from a control video of the Mitocheck dataset (white: histone 2B)

Briefly, HeLa cells were stably transfected for a red fluorescent chromatin protein (histone 2B fused to mCherry, H2B-mCherry) and a green fluorescent DNA replication factory (proliferating nuclear antigen fused to GFP, PCNA-mEGFP). Cells were seeded on LabTek chambered coverslips for live microscopy, and imaged for 48h with a time-lapse of 6 min (Plan10x, NA 0.5; Nikon - see fig. 1.3 for an example). Cells were maintained at 37°C in humidified atmosphere of 5% CO₂ during imaging. Following this, cell nuclei were segmented using local adaptative thresholding, improved by a split-and-merge approach as described, and samples for the different cell cycle phases were manually annotated using the open-source software Cell Cognition [Held et al., 2010].

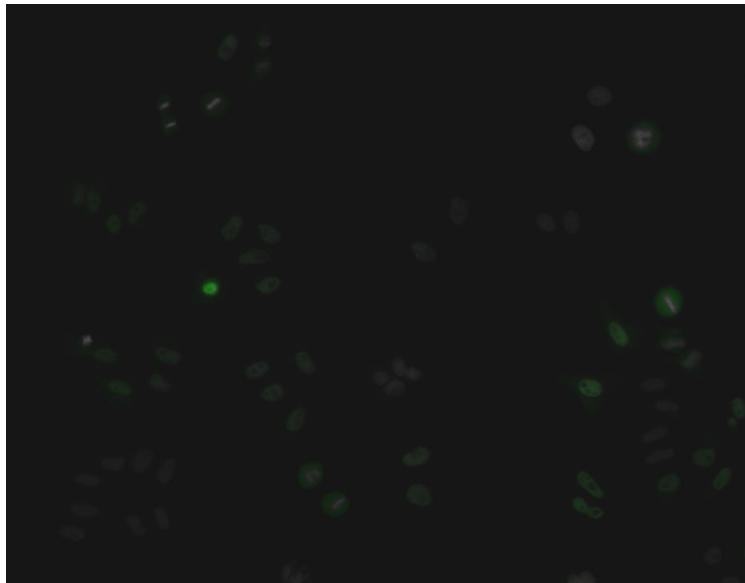


FIGURE 1.3: Image from a control video of the PCNA dataset (white: histone 2B, green: PCNA)

1.1.3 Drug screen

In section 3.4, we analyze an unpublished time-lapse drug screen³. In this dataset, 25 drugs were screened for their effect on HeLa cells in similar experimental settings to that of the Mitocheck dataset.

Experiments were not conducted in the context of this PhD. They were performed at the Advanced Light Microscopy facility of the EMBL (Heidelberg, Germany) by Beate Neumann, Jutta Bulkescher and Thomas Walter. Briefly, HeLa cells were stably transfected for a green fluorescent chromatin protein (H2B-GFP). Cells were seeded on 384-well plates for live microscopy ^{**}h prior to imaging. Drug exposure occurred ^{***}h prior to imaging. Finally, cells were imaged for 48h with a time-lapse of 30 min (Plan10x, ^{**} NA; [MICROSCOPE BRAND]). Cells were maintained at 37°C in humidified atmosphere of 5% CO₂ during imaging.

1.1.4 Xenobiotic screen

In chapter 4, we analyze an unpublished time-lapse xenobiotic screen, for which we performed the experiments. In this dataset, 5 xenobiotics were screened for their effect on MCF-7 cells.

Briefly, MCF-7 cells (ATCC® Catalog N°HTB-22™) were stably transfected for a red fluorescent chromatin protein (H2B-mCherry) and a green fluorescent membrane protein (myrPalm fused to GFP, myrPalm-GFP). Cells were seeded on 96-well plates 24h prior to exposure, and they were exposed to xenobiotics 24h prior to imaging. Finally, cells were imaged for 48h with a time-lapse of 15 min (Plan10x, 0.3 M27; Zeiss - see bottom of fig. 1.1 for an example). Imaging plates were sealed during imaging. Our experimental procedures will be detailed in section 4.1.

1.2 Software

We use CellCognition [Held et al., 2010]⁴ for segmentation and object feature extraction in all projects. To store, manage and access screening data, we use a previously published data format CellH5 [Sommer et al., 2013]. All scripts are written in the programming language Python 2.7⁵ using scipy [Jones et al., 2001], numpy, scikit-learn,

³Manuscript in preparation

⁴<http://cellcognition.org>

⁵<http://www.python.org>

fastcluster [Müllner, 2013], rpy2 and statsmodels, and all plots were generated by matplotlib [Hunter, 2007]. The Web-based user interface which was used for data visualization and sharing is based on Django⁶, Linux-Apache web-server, mod_wsgi and SQLite⁷. The R statistical function stats.p_adjust was used for adjusting p-values according to the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. Finally, CPLEX⁸ was used for optimization in the tracking procedure.

⁶<https://www.djangoproject.com/>

⁷<https://sqlite.org/>

⁸<http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/>

Chapter 2

A generic methodological framework for studying single cell motility

This chapter was published in [Schoenauer Sebag et al., 2015].

Résumé - Un cadre méthodologique général pour l'étude de la motilité cellulaire individuelle (see *infra* for English text)

Il existe de nombreux tests de motilité cellulaire, dont la majorité fournissent des informations à l'échelle d'une population de cellules (comme le test de la blessure). Toutefois, l'information à l'échelle individuelle est cruciale, car elle permet de détecter l'existence de sous-populations de cellules en terme comportemental.

Une seule autre étude de motilité cellulaire individuelle à haut débit a par conséquent été publiée, qui se base sur les empreintes des cellules sur un tapis de polymère [van Roosmalen et al., 2015]. Dans ce chapitre, nous présentons la première approche méthodologique pour étudier la motilité cellulaire individuelle dans des données de vidéomicroscopie d'un crible à grande échelle.

Ce workflow, MotIW (pour *Motility Integrated Workflow*), est constitué des étapes suivantes. Après l'acquisition des données, la segmentation et la description des objets sont réalisées à l'aide du logiciel libre *Cell Cognition* [Held et al., 2010]. Le suivi cellulaire est ensuite réalisé. Pour ce faire, nous nous sommes inspirés de [Lou and Hamprecht, 2011], qui formule le suivi de cellules entre deux images consécutives comme un problème d'apprentissage structuré. Le suivi cellulaire permet de résumer chaque expérience comme un ensemble de trajectoires de cellules, qui sont décrites à l'aide d'un ensemble original de 15 descripteurs. Enfin, les distributions de ces descripteurs permettent de caractériser une expérience : elles sont utilisées dans un test statistique multi-varié que nous avons conçu afin de déterminer si la motilité cellulaire individuelle y est différente de celle des expériences contrôles. Le workflow permet donc d'aller d'un ensemble de molécules test à une liste des molécules modifiant significativement la motilité cellulaire. Enfin, nous montrons dans la dernière partie l'intérêt et le pouvoir de notre méthode en l'appliquant à un jeu de données simulé.

2.1 Studying single cell motility in a HT setup

Cell *migration* describes "any directed cell movement within the body", as according to [Kramer et al., 2013]. On the other hand, cell *motility* more broadly encompasses any cell movement which is active, i.e. energy-consuming. Cell motility plays a key role in many physiological processes including embryonic development or immune response [Friedl and Weigelin, 2008], and is also involved in pathological processes such as fibrosis and metastasis. The latter is dependent on the ability of cancer cells to migrate, both as single cells and collectively [Decaestecker et al., 2007], [Yilmaz and Christofori, 2010], which highlights the need to understand the molecular basis of both processes.

Cell motility assays

Many *in vitro* assays have been specifically designed to study cell motility [Decaestecker et al., 2007], [Kramer et al., 2013]. The most classic methods are wound healing assay, cell exclusion zone assay, and trans-well migration assay. They measure the ability of cells to migrate into some free space (the wound, the exclusion zone) or a new chamber, in a limited amount of time. These assays have the advantage of being widely known; as an example, the Boyden assay is a trans-well migration assay which was introduced in 1962 [Boyden, 1962]. More recently, particle-coated plates were developed, in which particles are phagocytosed by cells as they move [Albrecht-Buehler, 1977]. In this assay, a single picture is taken at the end of the experiment, showing the path that was cleared by the moving cell. Analysis of these images is therefore equivalent to the analysis of the temporal projection of single cell trajectories.

Live-imaging data can be obtained from experiments using any of these methods (except the trans-well migration assay in its classical version). Data at single cell level could therefore theoretically be obtained. But the classic assays are most of the time used as endpoint assays: the experimenter is focused on getting aggregated data at the level of the cell population (e.g. wound closure time or percentage of cells staying in the upper compartment of the well).

However, single cell characteristics are relevant: they enable to detect patterns which are not visible at the population level, such as the existence of cell subpopulations with regard to their motility behaviours [Mokhtari et al., 2013], [Wong et al., 2014], [Maiuri et al., 2015], [Schoenauer Sebag et al., 2015]. This finds an application in drug design, as one might want to target specific populations [Perlman et al., 2004], [Singh et al., 2010]. The little use of single cell motility assays is probably due to both cultural and technical reasons: people are used to proceeding at population level, and univariate

analysis is easier than multivariate analysis. Furthermore, single cell motility studies in live cell imaging data have so far been limited to low-to-medium throughput [Lara et al., 2011], [Maiuri et al., 2015]. While this limitation has in principle be alleviated recently [Neumann et al., 2010], live cell imaging still remains a relatively expensive technique and produces large amounts of data, thus requiring an appropriate infrastructure, both for imaging and IT.

HT motility studies

As a consequence, there are only few automatic workflows for the comprehensive analysis of single cell trajectories including tracking, statistical analysis and data mining, applicable to HCS data: [van Roosmalen et al., 2015] analyses temporal projections of single cell trajectories, as observed by cell imprints (see above). Information on membrane dynamics is indirectly inferred from these data, but the data is not informative about direct movement features, such as instantaneous speed, curvature or, more importantly, resting time and speed variations. The methods published in [Mokhtari et al., 2013] are based on manual tracking and do therefore not scale easily to HCS.

In this chapter, we present MotIW (**M**otility study **I**ntegrated **W**orkflow). A generic methodological framework, MotIW enables to quantitatively study cell motility at single cell resolution in HT time-lapse data in an unsupervised way. It consists of cell tracking, cell trajectory mapping to an original feature space, and outlier experiment detection according to a new statistical procedure (cf figure 2.1). We show the power of our method in section 2.3 by applying MotIW to simulated data, which allows us to estimate recall and precision to be expected on real data. We then apply this workflow to a previously published genome-wide screen by RNA interference (RNAi) and live cell imaging, the Mitocheck dataset, in section 3.2 of chapter 3.

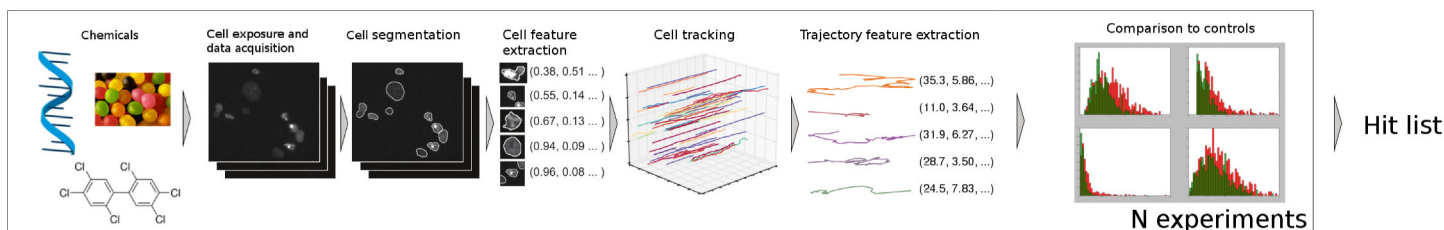


FIGURE 2.1: Overview of MotIW

2.2 MotIW overview

In this section, we present MotIW, our workflow for the automatic and quantitative analysis of single cell motility in video sets from time-lapse microscopy-based screens.

Figure 2.1 summarizes its different steps. Briefly, for each video nuclei are segmented and features are extracted as published previously [Walter et al., 2010], [Held et al., 2010]. Cells are tracked using a machine-learning based tracking procedure, described in section 2.2.1. The trajectories are then mapped to a feature space described in section 2.2.2. Presented in section 2.2.3, an original statistical procedure then enables the detection of experiments in which single cell motility is significantly different than that in control movies. Finally, section 2.3 describes the simulation of trajectories which allows us to validate the performance of the workflow.

2.2.1 Segmentation and tracking

The first step of the workflow is the establishment of single cell trajectories: we want to follow individual cells over time and record their spatial displacements. Many methods have been proposed in the image analysis and computer vision communities for the automatic tracking of individual objects. When it comes to the tracking of biological objects, such as cells or particles (e.g. single molecules or vesicles), there are two main approaches. On the one hand, it is possible to associate pre-segmented objects in consecutive frames, e.g. [Lou and Hamprecht, 2011], [Meijering et al., 2012], [Chenouard et al., 2014]. On the other hand, the deformable model approach relies on identifying and modeling objects in the first frame, and linking them to objects in consecutive frames by updating the models, e.g. [Zimmer et al., 2002]. Methods also differ in the amount of prior knowledge which is hard-coded in the tracking model. For example, [Li et al., 2008] assumes that there are three types of cell motion: Brownian motion, migration at a constant speed, and migration at a constant acceleration. At the other extreme, [Sbalzarini and Koumoutsakos, 2005] formulates tracking as an optimization problem, where a cost function of particle matches, typically depending on distance and intensity moments, is minimized. In particular, no constraint - except for maximal speed - on the type of movement is imposed.

We chose to keep segmentation and tracking steps independent: objects are identified before the establishment of object temporal correspondences by our tracking model. While the deformable model approach is in principle appealing as it jointly solves segmentation and tracking, it relies on a high time resolution and is therefore less generally applicable. Furthermore, as segmentation and tracking are not independent in this approach, the resulting method is necessarily less modular.

2.2.1.1 Segmentation

Segmentation of nuclei is in principle a relatively simple problem, as nuclei appear as bright objects on a dark background. The main difficulty arises when two or more nuclei come in close proximity to each other and are therefore segmented as one single object. Classically, this problem is solved by splitting objects after the first segmentation, e.g. [Held et al., 2010]. Briefly, a distance map of the binary segmentation is calculated, where we assign to each pixel its distance to the closest background pixel. If objects are touching, this typically generates important concavities in the resulting binary shape, thereby producing prominent maxima in the distance map. The final split is generated by calculating the Watershed transformation on the inverted distance map. To avoid false splits, the distance map is typically preprocessed by either morphological or linear filtering. The problem of this strategy is that non-convex shapes may be also split, and consequently the detection of multi-nucleated morphologies will be more difficult [Walter et al., 2010]. Nevertheless, we chose to start from this segmentation strategy, as implemented in CellCognition, as the main purpose of this study is to track nuclei and to analyze spatial trajectories, which will be eased by splitting the nuclei.

2.2.1.2 Cell tracking by supervised learning

Cell tracking should be able to face several challenges which are common in videos from high content screens. They include high population density in each picture, high phenotypic inter-cell variability, and possibly low time resolution between successive images. Furthermore, the algorithm has to handle apparitions, disparitions, divisions and fusions. Cells can indeed disappear, e.g. when they move outside the field of view or lose adhesion. They can also appear, for instance if they enter the field of view or more rarely if the expression of their fluorescent marker increases. Finally, they can fuse or seem to fuse, for example when a nucleus moves on top of another, or if two nuclei are still connected by chromosome bridges.

To be applicable in a screening context, we cannot *a priori* model cell motion, as such hypotheses are bound to break in the presence of phenotypes. Indeed, the impact of chemical exposure on cell motion is not known. As we also wished to avoid any manual parameter tuning, we extended a non-parametric structured learning approach from [Lou and Hamprecht, 2011].

We first characterize each cell nucleus in each image by a set of 239 shape and texture features on the one hand ([Walter et al., 2010], [Held et al., 2010]), and geometric features on the other hand (its distance to the border, its position in the image, and

the orientation of its main axis). The goal of cell tracking in this approach is to match cells in successive images, by assigning them the most likely instant temporal behaviour in the set $\mathbf{E} = \{\text{move}, \text{appear}, \text{disappear}, \text{split in 2 or 3}, \text{merge at 2 or 3}\}$. All possible matches between cells in consecutive frames are exhaustively considered, subject to distance thresholding. Match features are:

- the absolute difference of shape and texture features if the event is *move*, *split*, *merge*, the object features otherwise
- the geometrical distance between object at time t , $Obj_{i,t}$ and object at time $t+1$, $Obj_{j,t+1}$, if the event is *move*, *split*, *merge*, the minimal distance to the image border otherwise,
- the angle between $Obj_{i,t}$ and the elements of $Obj_{j,t+1}$, if the event is a *split* (angle α on fig. 2.2),
- the angle between the main axis of $Obj_{i,t}$ and $Obj_{j,t+1}$ weighted by their average eccentricity, if the event is a *move* (angle β on fig. 2.2).

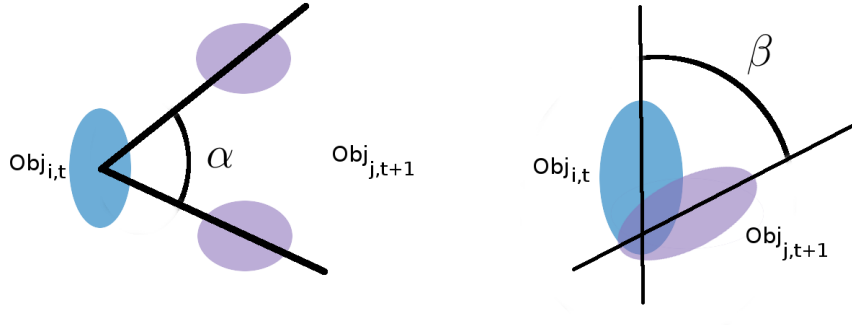


FIGURE 2.2: Illustration of angular match features.

The optimal object matching $\hat{z}(t)$ comes down to bi-partite graph matching: it is solved by maximizing a likelihood function L which depends on the weights w of match features and the match features $f_{i,j}^e$, subject to the constraint that all objects are matched in both frames (cf equation 2.3).

$$\hat{z}(t) = \arg \max_{z(t)} \mathbf{L}(z(t); w) \quad (2.1)$$

$$s.t. \quad \forall i \quad \sum_{Obj_{j,t+1}^e} z_{i,j}^e(t) = 1 \quad (2.2)$$

$$and \quad \forall j \quad \sum_{Obj_{i,t}^e} z_{i,j}^e(t) = 1 \quad (2.3)$$

where

$$\mathbf{L}(z(t); w) = \sum_{\substack{e \in \mathbf{E} \\ Obj_{i,t} \\ Obj_{j,t+1}}} \langle w^e, f_{i,j}^e \rangle z_{i,j}^e(t)$$

The weights w are learned by a structured support vector machine using annotated trajectories, following the formulation of [Lou and Hamprecht, 2011] (drawing on [Tsochan-taridis et al., 2005]). The likelihood maximization, an integer linear programming (ILP) problem, is solved by IBM Cplex.

The extension compared to [Lou and Hamprecht, 2011] lies in the choice of match features. Furthermore, we enabled the tracking model to learn from partial annotations of different experiments¹. This permits the user to integrate examples from both control and non-control experiments in the training set, which is crucial to guarantee that the model can efficiently track cells in all conditions. We also added three object division and fusion to \mathbf{E} . This is important in a screening context, where aberrant cell divisions may occur. We also implemented a more time-efficient computation of match hypotheses using kd-trees.

2.2.1.3 Validation of MotIW cell tracking model

To validate MotIW’s cell tracking procedure, we compare it to Cell Cognition’s constrained nearest-neighbour (CNN) tracking algorithm, and to [Jaqaman et al., 2008] as implemented in Cell Profiler [Carpenter et al., 2006]. We have chosen these two approaches for benchmarking, as they are available in popular High Content Screening software. [Jaqaman et al., 2008] views tracking as a linear assignment problem (LAP). It starts by computing 1-to-1 matches in consecutive frames, which produces tracklets. It then connects these tracklets by solving an optimization problem which is global both in time and space (2D-space and time duration of the experiment). For performing this optimization, that is, for choosing when to perform tracklet merges, splits, appearances and disappearances, it uses user-defined costs.

Our training set consists of approximately 32,000 matches, among which 0.5% *appear*, 0.5% *disappear*, 1% *merge* and 2% *split*. Data was taken from the Mitocheck data set, and in particular from both control experiments **and** experiments as selected by [Neumann et al., 2010] for being significantly different from controls regarding nuclear morphology. This ensures that the algorithm also works in the presence of phenotypes. One may

¹However, this is not learning from partial annotations in the sense of [Lou and Hamprecht, 2012]. Indeed, in our implementation of [Lou and Hamprecht, 2011], the user chooses a subset of cells which has to be annotated on all movie frames. In [Lou and Hamprecht, 2012], the user can choose both a subset of movie frames and a subset of cells (s)he wishes to annotate on those frames.

nevertheless be willing to perform cell tracking before having performed any statistical analysis. In this case, learning from control experiments only slightly diminishes the cell tracker performances, albeit non significantly.

To establish the data set, we first performed tracking with CellCognition’s CNN tracking and manually corrected for mistakes made by the algorithm. As in most cases, even such a simple tracker is able to find the correct assignment, we found this procedure much less time consuming than annotating all correspondences from scratch. As shown in Table 2.1, MotIW outperforms the other two methods as measured by the average accuracy on the five movement types. Note that these are not the overall accuracies of correct assignments, which are much higher for all three methods. As can be seen on fig. 2.3, all three methods show similarly good performances on *move* events and have therefore similar overall (pooled) accuracies. The contribution of the learning approach is most important for the other events, such as cell division, when object matching is less trivial.

In particular, it is interesting to see that this method even outperforms [Jaqaman et al., 2008], which relies on an optimization scheme in space and time, i.e. optimizes not only the assignments between two consecutive frames, but on the entire video sequence. While this might seem surprising at first sight, it is explicable by the fact that the latter approach was developed for particle tracking. Particles have only few distinguishing features, such as intensity and size. It is therefore feasible to manually define and tune a cost function for particle tracking based on these features only. It is nevertheless hardly feasible for larger feature sets, which are necessary to track more complex objects in terms of texture and shape.

In the future, it will be interesting to see whether this result can still be improved by an optimization scheme in time. Another promising strategy for future investigation will be to couple segmentation with tracking without relying on a deformable model approach. In particular, we could argue that split algorithms can provide us with alternative hypotheses on segmentation. The best combination of segmentation and tracking can then be found by global optimization in time.

Altogether, we conclude that the tracking procedure is sufficiently accurate to generate single cell trajectories.

TABLE 2.1: Mean recall and precision on all types of matches **E** (10-fold cross-validation)

Algorithm	Mean recall (%)	Mean precision(%)
CNN	72.7	62.8
[Jaqaman et al., 2008]	78.3	73.0
MotIW	91.1	91.5

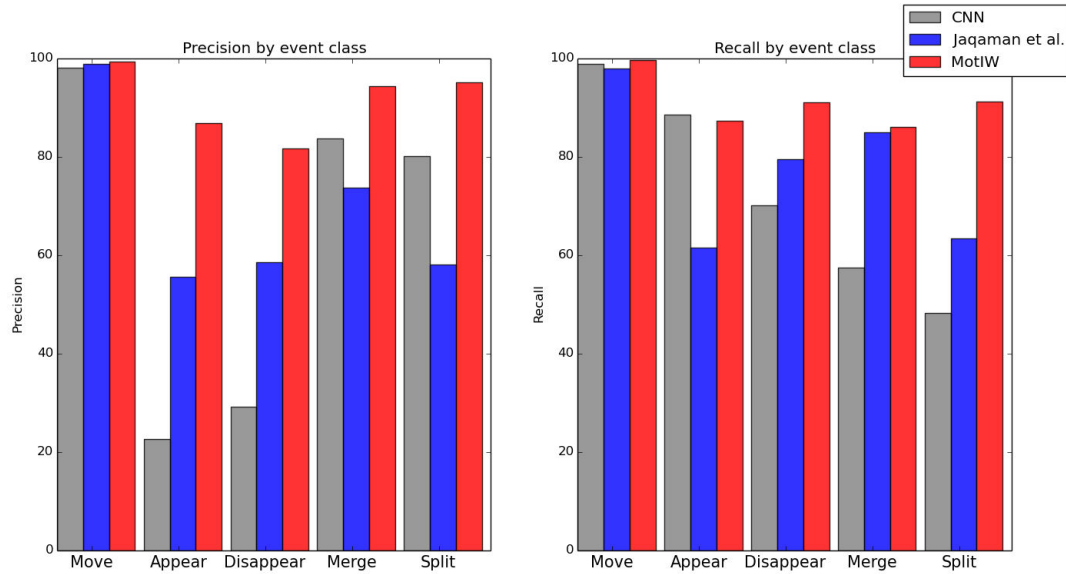


FIGURE 2.3: Details of tracking precision and recall according to event types

2.2.2 Trajectory features

Once cell tracking has been performed, each experiment is summarized as a set of cell trajectories in the two-dimensional space over time. Instead of analyzing these point sequences directly, we calculate for each trajectory a set of relevant features which will allow us to represent each trajectory to a point in this feature space. For instance, cell speed is an important characteristic of cell motion, and it is certainly the most studied one. However, it makes sense to also describe other aspects of movement. For instance, if we consider the example trajectory of fig. 2.4, one might also be interested in quantifying the percentage of time which is spent in each englobing blue ball, or whether diffusion would have been an appropriate model for this particular cell. A multivariate study of cell trajectories is therefore relevant to capture all the information which they contain. For this, a set of 15 features was assembled, partly from previous publications on quantitative motility analysis, partly newly designed.

Robust and precise features are needed to account for the partial stochasticity of cell migratory behaviour. We use three types of features, as detailed in table 2.2.

2.2.2.1 Particle motion features

This group of features encompasses the diffusion coefficient and the movement type, which were in the first place used to study particle motion (see [Ferrari et al., 2001], [Sbalzarini and Koumoutsakos, 2005] for one of its applications to single particle motion in Biology).

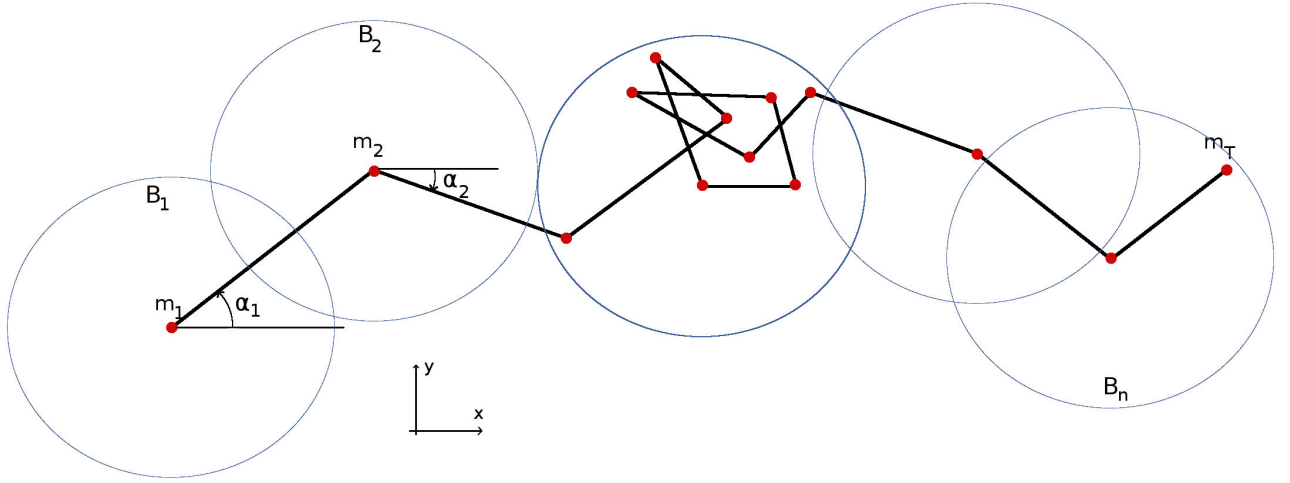


FIGURE 2.4: A cell trajectory with notations

TABLE 2.2: Cell trajectory features and their formulas. Notations: $(m_t)_{t=1\dots T}$, time sequence of cell 2D positions. T, track time duration. P, total track length

Particle motion features	
Diffusion coefficient	According to [Sbalzarini and Koumoutsakos, 2005]
Diffusion adequation	Correlation between MSD(t) and t
Movement type	According to [Sbalzarini and Koumoutsakos, 2005]
Englobing ball number	See text
Track entropy	See text
Other global features	
Convex hull area	-
Effective path length	$L = \ m_T - m_1\ _2$
Effective speed	L/\sqrt{T}
Largest move	-
Straightness index	$\sqrt{T}L/P$
Track curvature	See text
Averaged local features	
Mean squared displacement (MSD)	$\frac{1}{T-1} \sum \ m_t - m_{t-1}\ _2^2$
Mean signed turning angle	$\arctan\left(\frac{\sum \sin(\alpha_{t+1} - \alpha_t)}{\sum \cos(\alpha_{t+1} - \alpha_t)}\right)$

Let us consider a particle, and note m_t its position at timepoint t . The moment of order p of this particle, $\langle d^p \rangle$, can be computed according to the following formula:

$$\langle d^p \rangle = \langle \|m_t - m_{t-1}\|_2^p \rangle_t$$

For large t , it is proportionate to t^{γ_p} for most dispersive processes [Ferrari et al., 2001].

Assuming that γ_p is proportionate to p (i.e. that the particle movement is strongly self-similar), the constant $\gamma = \gamma_p/p$ (hereafter the particle's *Movement type*) quantifies how directed the particle motion is. If γ is equal to 1, the movement is perfectly directed, whereas if γ is equal to 0.5, it is perfectly diffusive. Between 0.5 and 1, the movement is super-diffusive, whereas below 0.5 it is called sub-diffusive.

Furthermore, assuming $\gamma = 0.5$, the constant linking $\langle d^2 \rangle$ (that is, the mean squared displacement) and t can be computed: it is the *Diffusion coefficient*. The *Diffusion adequation* is the correlation coefficient between $\langle d^2 \rangle$ and t , hence measuring how well the diffusive model applies to the track at hand.

Here, we present two newly designed features to characterize the alternance between periods of diffusive motion and periods of directed motion: the track entropy and the englobing ball number.

It has been observed that cell motion in 2D alternates between diffusive and directed motions (in the absence of any perturbation or chemical gradient). The feature track *Entropy* was designed to measure how the time sequence of 2D cell positions $m_t = (x_t, y_t)$ distributes in balls of radius r . This will be computed greedily by recursively searching the center of a new ball of radius r among the set of remaining track points, that will contain the biggest number of them. This feature is calculated according to the following procedure, for each track of time duration T :

1. $S = \{1, \dots, T\}$
2. while $S \neq \{\}$:
 - i. do $t^* \leftarrow \arg \max_S \text{card}(B_r(t))$
 where $B_r(t) = \{i \mid \|m_i - m_t\|_2 \leq r \text{ and } \min(\|m_{i-1} - m_t\|_2, \|m_{i+1} - m_t\|_2) \leq r\}$
 - ii. do $S \leftarrow S \setminus B_r(t^*)$
3. Compute the track *Entropy* according to the following formula:

$$\text{Entropy}_r = -\frac{1}{T} \sum_{B_r} \frac{\text{card}(B_r)}{T} \log\left(\frac{\text{card}(B_r)}{T}\right) \quad (2.4)$$

The track *Entropy* measures the entropy of the distribution of track positions in balls of radius r . To deal with cells whose trajectories are concentrated in space, but were not concentrated in time, the constraint is imposed that these balls shall contain only consecutive positions in time. The englobing *Ball number* is the number of balls of

radius r that contain all track positions. It is normalized by the square root of T to be independent of the track time-length T .

Different radii may be relevant for different data (depending of, e.g., the experiment time-lapse, the pixel size or the cell type). We chose to use two different radii r_1 and r_2 with $r_1 < r_2$, to incorporate information about cell trajectories on two different time-scales. r_1 and r_2 were manually chosen, such that for the Mitocheck data set, the corresponding features are neither constant over a large number of trajectories nor too correlated. They respectively correspond to approximately $2.5\mu m$ and $12\mu m$. In the following, the features *Entropy i* and *Ball number i* correspond to radius r_i .

2.2.2.2 Other global features

We further defined the following global descriptors of cell trajectories: the cell's *Largest move* along the trajectory, its track *Convex hull area* and its average *Track curvature*. The track *Convex hull area* is the area of the convex hull containing all track points, as coloured in green on fig. 2.5. It is normalized by the square root of the track time-length. Following [Naffar-Abu-Amara et al., 2008], this feature enables us to have an idea of the area which the cell has visited during its trajectory, although it does not exactly indicate the area which its cytoplasm has covered. Finally, for each trajectory and each time-point t , an orthogonal regression is performed on $\{(x_i, y_i) | i \in \{t, \dots, t + \Delta_t\}\}$ using orthogonal distance regression ($\Delta_t = 10$). The mean *Track curvature* of the trajectory is the average of all regression sums of squares.

2.2.2.3 Averaged local features

Finally, two features are averaged local features, which are the cell means squared displacement (*MSD*) and its *Mean signed turning angle* (see table 2.2 for the formulae).

2.2.2.4 Feature set evaluation

Track time-length is an irrelevant random variable for studying single cell motility, which could bias some features. Therefore, we ensured that they are not significantly correlated with this parameter: the correlation between track time-lengths and features is maximal for the *Effective space length*, where it is equal to approximately 30% (on a subset of the Mitocheck dataset, data not shown).

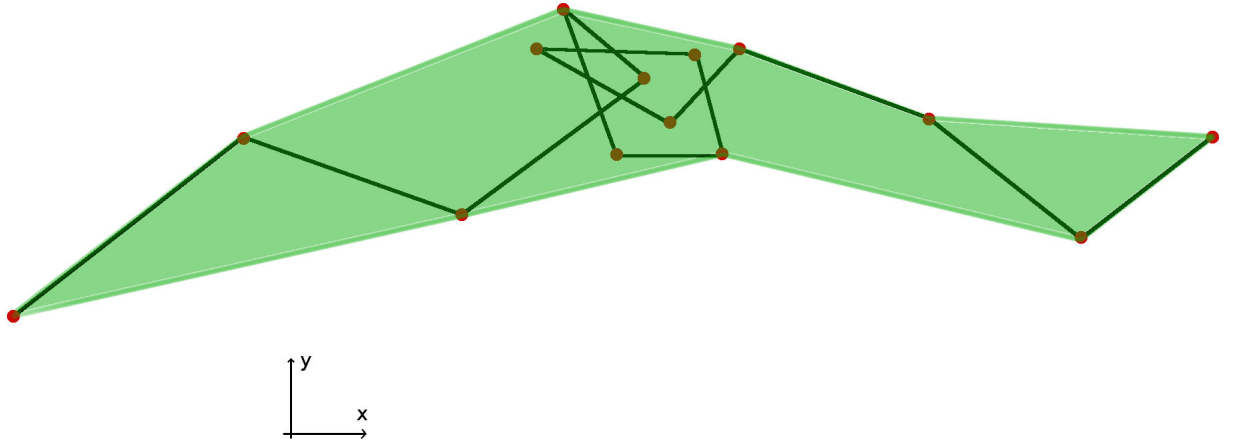


FIGURE 2.5: Convex hull of the example track from figure 2.4

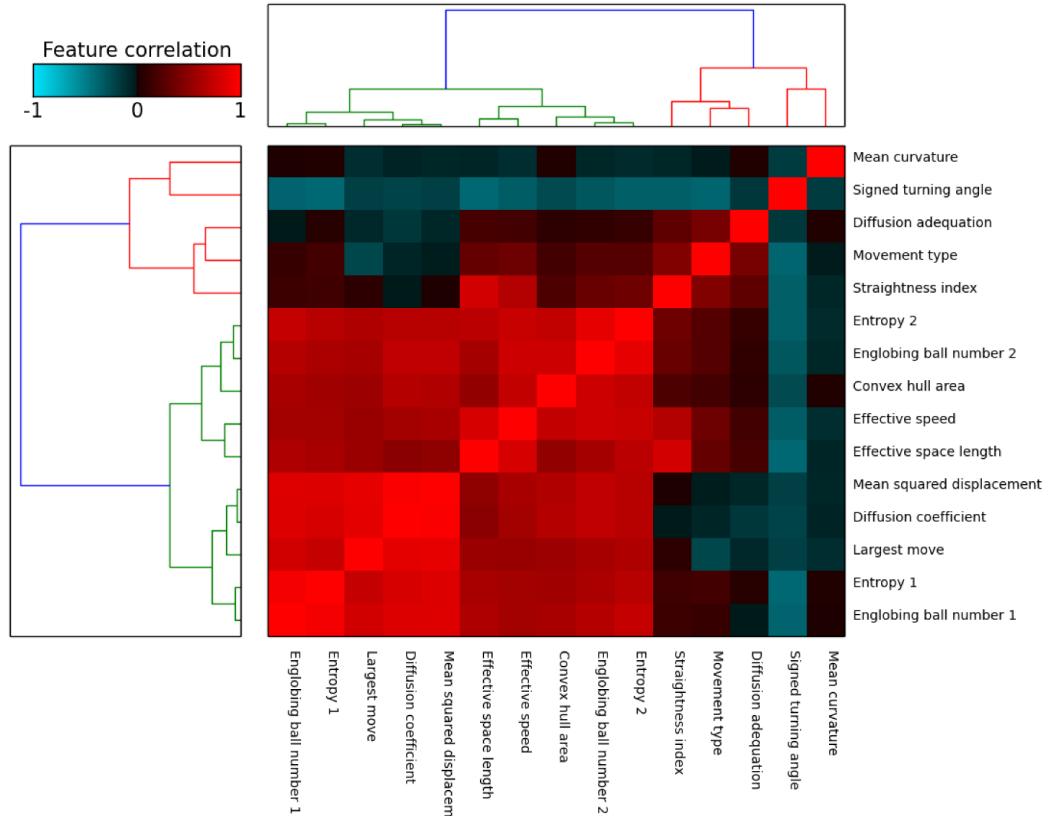
FIGURE 2.6: Heatmap showing trajectory feature similarities on a subset of the Mittocheck dataset (1.1 million trajectories coming from detected motility hit experiments according to MotIW). The dendrograms were obtained using the *Ward* method and the euclidean distance between feature correlations.

Figure 2.6 shows the correlation matrix for the extracted features. One group of highly correlated features are visible in the bottom-left corner of the heatmap, which encompasses speed-related features. The existence of two feature subgroups within this group can be explained by the following observation: the first group of features, from *Ball number 1* to *MSD*, is linked to cell instantaneous displacements, whereas the second group,

from *Effective speed* to *Entropy 2*, is linked to its displacements on the whole trajectory.

The other correlations can as well be explained by feature definitions. As an example, the anti-correlation between *Mean signed turning angle* and *Movement type* can be interpreted as follows: a low signed turning angle is indicative of correlated motion, which is super-diffusive and translates into a high *Movement type*.

Fig. 2.6 indicates that there are less degrees of freedom than features, which was verified by a principal component analysis (PCA). On the same trajectory subset, approximately 95% of the variance is explained by the first seven principal components.

2.2.3 Statistical procedure

2.2.3.1 Trajectory quality control

Prior to statistical analysis, a trajectory quality control is performed. First of all, trajectories resulting from object fusion are discarded. Indeed, trajectories resulting from a fusion are most of the times cell cluster trajectories, rather than cell trajectories. Trajectories which are shorter than 10 frames are also discarded, to ensure that features such as the diffusion coefficient are computed on a sufficiently large number of points. Finally, because we are interested by single cell motility rather than collective motility, all trajectories with more than 5 neighbours in a perimeter of 50 pixels are deleted. This trajectory quality control ensures that cell clusters are not considered, and increases the dataset robustness. This quality controls eliminates $11.6 \pm 13.3\%$ (median \pm interquartile range) of cell trajectories per experiment, as estimated on a random subset of 1,000 experiments from the Mitochek dataset.

HT screening data is organized in batches of experiments which have been acquired simultaneously. Each batch includes a set of negative controls, i.e. conditions where no effect is expected. Due to a non-negligible batch effect, an experiment can only be compared with controls of the same batch in most of the cases.

Let us consider an experiment i . Following to trajectory feature extraction, it can be summarized as a set of Θ feature distributions ($\Theta = 15$). The comparison of these distributions with those of controls from the same batch B_i , using Kolmogorov-Smirnov 2-sample test, provides a list of p-values $(p_\theta)_{\theta=1\dots\Theta}$.

A final statistic S_i combining the p-values of all features is obtained by Fisher's formula :

$$S_i = -2 \sum_{\theta} \ln(p_\theta) \quad (2.5)$$

As shown on fig. 2.6, the features are not independent. Therefore, the distribution of this statistic under the null hypothesis does not follow a chi-squared law with 2Θ degrees of freedom. To assess which values of this statistic should be considered as indicative of altered motility, a sample of the distribution of S under the null hypothesis is then computed by comparing the control experiments which were not used in the experiment-controls comparisons, with the other controls from the same batch.

In the absence of an explicit form for the null distribution, this sample allows to quantify the intra-batch variations of single cell motility features. The variations can be due to technical artefacts or biological variability. Then, the comparison of the distribution of S statistics obtained from control-experiment comparisons, to the distribution obtained from control-control comparisons, permits the computation of empirical p-values. This enables the detection of hit experiments with regard to single cell motility.

2.2.3.2 False discovery rate control

False discoveries are controlled using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. Indeed, in the current case, approximately 150,000 statistical tests will be performed. Selecting experiments whose p-value is below 0.05 rigorously means that there will be 5% false discoveries. Without any adjustment, 150,000 tests will produce 7,500 false discoveries, which is not acceptable.

There are different ways to adjust p-values. If the statistical tests are independent, the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] is a procedure to adjust p-values and control the false discovery rate. If the tests are not independent, there are other procedures (see for example [Benjamini and Yekutieli, 2001] or [Delattre and Roquain, 2015]). Here, we assume in a first approximation that the tests are independent. Rigorously speaking, they are not: experiments which are performed on the same day share some dependence, which may be the influence of the temperature, the pressure, the passage number of the cells or another experimental variable. We do the hypothesis that we can neglect this local dependence (384 experiments versus 150,000).

2.2.3.3 Formal statistical procedure

This procedure is repeated n times to ensure that the final p-value of an experiment i does not depend on the choice of a specific subset of control experiments in its batch. Here is its formalized description:

1. **Compute a sample of statistic (2.5) under null hypothesis from control-control comparisons.**

For each batch b ,

For k in $\{1, \dots, C_b(C_b - 1)/2\}$, where C_b is the number of controls of batch b that passed the quality control

a. Randomly split the control experiments in two groups $A_{b,k}$ of cardinal 2, and $B_{b,k}$ of cardinal $C_b - 2$

b. For each control j of $A_{b,k}$, compute the statistic $S_{b,k,j}^0$ (2.5) by comparing it to the pooled group of controls $B_{b,k}$

2. **Compute statistics from experiment-control comparisons.** For computation time feasibility, only $n = 5$ repetitions corresponding to n splits of the controls set $(A_{b,k}, B_{b,k})$ are selected on each batch for experiment-control comparisons.

a. For each repetition k in $\{1, \dots, n\}$:

For each experiment i belonging to a batch b , compute the statistic $S_{k,i}$ (2.5) by comparing it to the pooled group of controls $B_{b,k}$

b. Combine distinct iterations. In order to be conservative, we chose the following approach:

$$S_i = \max_{k \in \{1 \dots n\}} S_{k,i} \quad (2.6)$$

3. For each experiment i , compute the p-value p_i :

$$p_i = \max \left(\frac{\text{card}(\{(b, k, j) | S_{b,k,j}^0 \geq S_i\})}{\text{card}(\{(b, k, j)\})}, \frac{1}{\text{card}(\{(b, k, j)\})} \right)$$

4. For each experiment i , compute the adjusted p-value p'_i to control the false discovery rate (Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995])

Each experiment is characterized by an adjusted p-value ; significantly different experiments are those whose adjusted p-values are below a certain threshold. An experimental condition is selected as being a hit if 50% or more of its replicate experiments are significantly different. This is a way of ensuring reproducibility. It amounts to representing an experimental condition by the median of its replicate scores. Since the mean is sensitive to outliers, using the mean of its replicate scores instead of their median would lead to too many false positives.

2.3 Validation on a simulated screen

2.3.1 Screen simulation

In order to evaluate the performance of our workflow on data for which the groundtruth is known, we designed a process to simulate a HT screening experiment.

In a first step, five types of single cell movements were designed, in agreement with qualitative observations from the dataset: *random*, *fast-random*, *curbed-directed*, *flip-directed* and *stop-and-go* (see fig. 2.7).

Let (d_t, ϕ_t) be the polar coordinates of the difference vector $m_{t-1} - m_t$ of any two consecutive points. For *random* movement, ϕ_t is chosen at random and the distance $d_t = \|m_t - m_{t-1}\|_2$ is drawn from a normal distribution, whose parameters are estimated from the data. The same holds for *fast-random* with increased distance d_t . For the *curbed-directed* movement type, d_t follows again a normal distribution as for *random* movement, but the angle is calculated as $\phi_t + \epsilon$ with $\phi_t = \phi_{t-1} + \Delta\phi_t$, where $\Delta\phi_t$ and ϵ follow normal distributions, whose parameters are set manually to visually match some observed trajectories.

Flip-directed and *stop-and-go* are two composite types of movement, where the cells alternate between different states. The dwelling times in the two states are random integers with manually fixed ranges (which can be different for the two states) and are drawn independently for each trajectory. *Flip-directed* movement corresponds to directed movement (ϕ_t is drawn from a normal distribution) with a 180 degree flip for every state transition. Finally, *stop-and-go* movement alternates between slow random movement (where ϕ_t is drawn from a uniform distribution) and fast directed movement (where ϕ_t is drawn from a normal distribution).

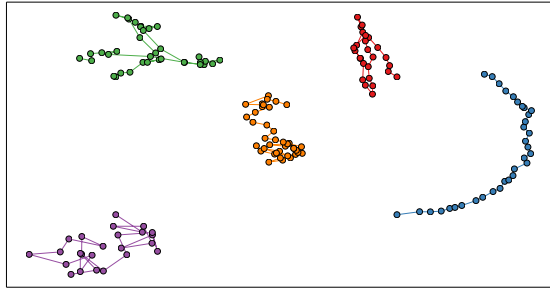


FIGURE 2.7: Simulated trajectories: stop-and-go (green), flip-directed (red), random (orange), fast random (purple), curbed-directed (blue)

In a second step, we want to simulate movies (controls and experiments), i.e. sets of trajectories. For this, we define five movie types with different proportions of single

cell movement types (cf supra). *Normal* movies account both for control movies and experiments in which cell motility is similar to that of controls. They contain on average 80% of *random* trajectories, and a mix of the four other trajectory types. This reflects our observation that in real data, experiments and controls typically contain all possible types of cell trajectories and that phenotypes are characterized in a shift in percentage. All other movie types contain (on average) from 50 to 65% *random* trajectories, the rest being completed according to the movie type. For example, movie type *fast* is composed of 30% of *fast-random* trajectories, 60% of *random* trajectories, and a mix of the three other trajectory types.

The total number of trajectories in each movie was drawn at random from real data in the following way: first, a batch is randomly chosen in the data set. Then, we assign a permutation of the real trajectory numbers from the experiments of the picked batch to the simulated positions. This enables to include potential batch effects in our simulated data. Furthermore, they match those from the dataset on which we developed this workflow, namely the Mitocheck dataset. The number of trajectories of each movement type in each movie is drawn from the corresponding movie type multinomial distribution, where the percentages were defined as described above.

The third step was the simulation of approximately 50,000 experimental conditions, which were distributed on 130 plates, and performed in triplicate as in Mitocheck experimental setup [Neumann et al., 2010]. For sake of simplicity, triplicates were supposed to belong to the same movie type. On each plate, between 5% and 15% of the experiments were selected to be other than *normal* movies.

2.3.2 Application to a simulated screen

Our workflow successfully recognized more than 98% of the experiments, as detailed in table 2.3.

TABLE 2.3: Results from the application of MotIW to simulated data

	Recall (%)	Precision (%)
Outlier experiment detection	99.2	98.9
Outlier condition detection	99.5	100.0
Trajectory clustering	91.4 ± 2.1	89.4 ± 4.8

Our simulation pipeline was also used to estimate how useful the trajectory feature set is to capture the differences between different types of trajectory motion. 500 samples of each trajectory type were simulated, and their features extracted. A PCA was performed, after which we retained the eighth first principal components, which explain approximately

95% of the data set variance. Finally, k-means was applied to the data set with $k = 5$. Many simulation parameters (e.g. each track length) are chosen at random, and k-means' results depend on its initialization: the procedure was therefore repeated 10 times. The results are presented in table 2.3. Although distinguishing trajectory types is subject to some errors, it shows that the whole pipeline is robust enough to identify experiments in which cell motility is significantly different.

We therefore conclude that MotIW is capable of identifying trajectory clusters in an unsupervised way. Hence, we are confident that this methodological workflow will allow the identification of migratory behaviors in HT live cell imaging data sets. Moreover, we believe that the general strategy of identifying hit experiments prior to cluster analysis might be useful for unsupervised approaches to HCS data.

This strategy can indeed also be seen as a way of intelligently downsampling the data to a reasonable size. Compared to random sampling, it has the advantage of enriching the data set in extreme cases. In a supervised setting, this occurs as a natural result from the annotation process: the numbers of samples in each class of the training set rarely reflect their proportion in the whole dataset. The training set is therefore most of the times enriched in rare classes. This seems to be beneficial as well for unsupervised learning approaches to large biological data mining.

Chapter 3

High-content screening data as a resource

Résumé - Les données de crible à haut débit en tant que ressources (see *infra* for English text)

Les cribles à haut contenu constituent des jeux de données riches et complexes, auxquels il est pour la plupart possible d'accéder sur le Web. Nous pensons qu'au-delà de cet accès, il serait encore plus bénéfique pour la communauté de généraliser les analyses secondaires de telles données. En effet, il y a généralement matière à plusieurs études dans un seul ensemble d'expériences à haut contenu, qu'il s'agisse d'approfondir celle du même processus biologique ou d'en étudier un autre.

Les analyses secondaires sont probablement encore peu nombreuses du fait des problématiques générales liées au partage de données (qualité des données et méta-données, partage de vocabulaire, persistance des données), comme de problématiques spécifiques : la taille des jeux de données complique leur partage, dont l'utilité n'a pas encore été démontrée.

C'est ce à quoi nous nous attachons dans ce chapitre, à travers trois exemples de ré-analyse des données Mitocheck [Neumann et al., 2010]. Dans la section 3.2, l'application de MotIW [Schoenauer Sebag et al., 2015] permet d'aboutir à une liste de gènes potentiellement impliqués dans la motilité cellulaire, comme à une ontologie des trajectoires cellulaires individuelles. La section 3.3 se concentre sur l'étude des gènes impliqués dans le cycle cellulaire. Enfin, la section 3.4 décrit le développement d'une nouvelle distance pour l'inférence de cibles thérapeutiques, appliquée aux expériences de Mitocheck ainsi qu'à un crible pharmacologique non publié.

3.1 Data re-use in Bioimage Informatics

Modern biology has many features of big data science: it is characterized by systematic large-scale studies, generating enormous amounts of biological data, and it relies on advanced data mining methods in order to infer the biological information from these large and heterogeneous data sets.

Imaging experiments are no exception to this. The advent of HCS has indeed led to the generation of extremely large and complex image datasets. Computational analysis of such data, and of genetic screens in particular, has helped to identify the genes required for cellular processes as diverse and as fundamental as protein secretion [Simpson et al., 2012], endocytosis [Collinet et al., 2010] and cell division [Neumann et al., 2010], and allowed for the phenotypic annotation of many genes.

On the other hand, producing such comprehensive and high quality data sets is a major investment in terms of time, manpower and money. The resulting data sets are extremely rich in information and usually, their information content goes beyond what is typically published in articles. It seems therefore more than reasonable to provide access to these data sets. Indeed, this has been recognized already by the scientific community: some of these large scale projects have made the initial image data publicly available as a scientific resource. For instance, [Neumann et al., 2010] generated a data set published on mitocheck.org. There, scientists can check the loss-of-function phenotype for genes they are interested in, both in terms of raw data and analysis results. On the same platform, several other phenotypic screens are published, which makes it possible to search across different phenotypic aspects.

While the utility as a resource for consultation has thus been shown, we believe that the usefulness of such data sets goes beyond single gene queries and visual inspection of the recorded image data. Indeed, although such data sets are mostly generated by a laboratory or a consortium in order to answer a specific question, the acquired data is usually informative about many different aspects of cellular phenotypes. Therefore, one single analysis - typically performed by the groups that performed the screen - does not exploit all information contained in these rich datasets.

Here, we advocate the re-use of HCS data by raw image data re-mining. This can be aimed at performing more detailed analyses of the same biological process, or in view of finding answers to different biological questions. For example, cell nucleus images can be used to analyze mitotic phases, the number of nucleoli, or even to assess phenotypic heterogeneity for different biological processes. Second, genome-wide data sources provide us with reference datasets, to which other experimental data can be compared. Third, availability and re-use of easy-to-access biological data sets enable method comparison.

Finally, as pointed out by [Choudhury et al., 2014], data accessibility and reuse is well in harmony with the current trend of general transparency in science. Indeed, there are several examples in particular in the gene expression literature, where re-analysis of data has led to a debate on the stability and validity of reported results.

While the use of annotated data sets for method comparison have become increasingly important to the computer vision community (see [Choudhury et al., 2014] for a list of benchmark data sets) and are indeed used frequently, re-mining data studies are still very sparse in this field. Regarding the Mitocheck data set for example, four articles only have been re-using it since the original paper [Neumann et al., 2010] was published: [Ostaszewski et al., 2012] is interested by linking phenotype and genotype in the context of cell cycle, [Suratanee et al., 2014] searches for evidence of protein-protein interactions in siRNA experiment similarity, and [Pau et al., 2013] dynamically models nuclear phenotypes.

I see various reasons for the little data re-use which is currently experienced in *Bioimage informatics*. First, the general issues associated to data sharing and re-use do apply to the case at hand: data and metadata should be of optimal quality, in order to ease re-analysis and make the data structure understandable. Laboratories should also share controlled vocabularies [Peng, 2008], as well as formats, for data to be easily accessed and understood by all parties. Data re-use should be born in mind while acquiring it [Wruck et al., 2014], since post-experimental data re-organization is harder and likely to never happen. Data should also be persistent, which demands time, fundings, and incentive. Second, HCS data set tend to be large. While many institutes have the IT infrastructure to both store these data and to computationally analyze them using compute clusters, it is still a challenge to actually transfer the data. The only working solution today is to actually send hard disks by mail. More appealing solutions such as a common computing cloud where people could perform their analysis directly without transferring entire data sets, would require an important funding effort, which might not be straightforward to obtain. The third reason is that so far, the usefulness of re-mining existing image data sets has not yet been shown sufficiently. Once the scientific community is aware of the potential data mining of HCS data sets provides us with, the technical challenges might be properly addressed.

Here we will show different cases of data reuse. Starting from the data set published in [Neumann et al., 2010], we will show an in-depth analysis of nuclear motility by applying MotIW to this data set. This led to the discovery of genes which are likely to be involved in cell motility, as well as an ontology of cell trajectories (section 3.2). Furthermore, combining tracking with nuclear phenotypic classification enables to detect

genes which have an impact on cell cycle length, as described in section 3.3. Finally, in the context of drug development, the comparison between Mitocheck phenotypic profiles and those following drug exposure enables to identify a few possible drug target pathways for each investigated substance, as is exposed in section 3.4. We hope that this work will make a case for the remining of HCS data and the usefulness of HCS as a scientific resource in terms of systematic and comprehensive analysis.

3.2 MotIW reveals modes of movement and genes involved in nuclear motility

Part of this section was published in [Schoenauer Sebag et al., 2015].

Analysis of the Mitocheck dataset with our workflow allows the identification of genes with putative role in nuclear and/or cellular motility (section 3.2.1). Furthermore, it reveals the existence of a cell trajectory ontology in the dataset. Without any prior assumption on cell motion, we are able to identify eight types of cell trajectories (section 3.2.2). Finally, we also observed that all motility behaviours exist in negative controls: the only effect of gene silencing is a change in the measured proportion of distinct motility modes.

3.2.1 Hit list

After evaluating MotIW on simulated data, we then apply it to the whole genome-wide screen Mitocheck [Neumann et al., 2010]. In the context of the Mitocheck dataset, the identification of an experiment in which cell motility is significantly different from negative controls leads to the identification of siRNAs which significantly and reproducibly alter cell motility. A gene was selected as possibly involved in motility mechanisms if it was targeted by at least one hit siRNA. Indeed, it is well known that only a proportion of the siRNAs targeting a particular gene will effectively lead to a significant down-regulation. The reasons which could explain this are still not completely understood. Therefore, requiring that more than one siRNA related to a given gene are selected for the gene to be selected would have led to too many false negatives.

The application of MotIW to the Mitocheck dataset enabled the identification of the experiments which significantly deviate from controls (5%; 7,153 out of 144,909). It amounts to 1,180 genes (out of 17,816), which are available as a supplementary to this thesis (see Supp. table 1).

3.2.1.1 Functional analysis

Some of these genes are well known to be involved in cellular motility, such as RhoA (Ras homolog family, member A) or CDK5 (cyclin-dependent kinase 5). However, the list is not overall significantly enriched in genes which are linked to cell motility according to the Gene Ontology database and DAVID online analysis tools [Jiao et al., 2012].

The three first functional annotation clusters which were found on MotIW's hit list are respectively related to protein kinases and ATP-binding proteins, G-protein coupled receptors, and neurotransmitter receptor activity (respective enrichment scores: 6.69, 4.86 and 4.56). Nevertheless, the fourth functional annotation cluster which was found using DAVID tools is composed of three "Cellular components" GO terms: *integral to plasma membrane* (GO:0005887), *intrinsic to plasma membrane* (GO:0031226) and *plasma membrane part* (GO:0044459) (enrichment score: 3.33). This is consistent with what has already been observed [Naffar-Abu-Amara et al., 2008]: changes in cell adhesion to the extra-cellular matrix and to neighbouring cells are bound to have major impacts on cell motility. As an example, cell motility processes are involved when cells divide as they detach from the plate surface, or as they probe their close environment. Perturbing cell adhesion or cell protrusion mechanisms through the silencing of one of the numerous membrane proteins which these processes involve, will therefore modify cell motility. This explains that our hit list is enriched in genes which are related to cell membrane.

Comparing this finding with results from other recent screens on cell migration, we observe that a lack of enrichment seems to be the rule rather than the exception. [Simpson et al., 2008] describes a study of 1,081 genes regarding cell motility in human breast cells (MCF-10A cell line). Using wound healing, they identify 66 high confidence genes of which only 24 were previously associated with cell motility. Similarly, [Lara et al., 2011] is focused on the involvement of kinases (779 genes) in cell motility of human lung cancer cells (A549 cell line). Using single cell tracking, they identify 70 hit genes, of which only 13 were previously linked to cell motility. Finally, [van Roosmalen et al., 2015] study 1,429 genes using phago-kinetic tracks in human lung cancer cells (H1299 cell line), finding 136 hits. Thanks to a personal communication of the authors, we could access their hit list and see that only 13 of their hit genes are functionally linked to cell motility (Gene Ontology biological process GO:0048870).

Hence it seems that medium- to HT approaches to cell motility study tend to complement the older ones, rather than abide by them. In fact, this seems to be a more general trend of genetic screen hit lists. For example, Mitocheck mitotic hit list contains more than 50% genes which were not functionally linked to cell division before [Neumann et al., 2010]; [Eggert et al., 2004]'s cytokinesis-linked gene list contains only 20% genes

which were previously known to be involved in this cellular process. Modern systematic and automatic approaches to gene functional inference are indeed likely to detect genes whose involvement in the cellular process at hand was too subtle to be detected by lower-throughput and more ancient methods.

3.2.1.2 Intersection with other published motility gene lists

The next step is to analyse the intersections between different published medium- to high-throughput studies about cell motility. They are detailed in table 3.1.

TABLE 3.1: Existing medium- to high-throughput studies of cell motility

Study	Assay	Cell line	Gene list	Hit gene list
[Simpson et al., 2008]	Wound healing	MCF-10A	1,081	66
[Lara et al., 2011]	Single cell tracking	A549	779	70
[Yang et al., 2013]	Matrigel invasion chambers	U87	1,954	25
[Zhang et al., 2013]	Cell area growth	HeLa	710	81
Us	Single cell tracking	HeLa	17,816	1,180
[van Roosmalen et al., 2015]	Phago-kinetic tracks	H1299	1,429	136

TABLE 3.2: Hit list intersections

	Simpson	Lara	Yang	Zhang	Us	van Roosmalen
[Simpson et al., 2008]	66	4	0	19	10	4
[Lara et al., 2011]	4	70	0	4	4	2
[Yang et al., 2013]	0	0	25	0	4	0
[Zhang et al., 2013]	19	4	0	81	13	5
Us	10	4	4	13	1,180	6
[van Roosmalen et al., 2015]	4	2	0	5	6	136

Following the information which is presented in table 3.2, there is little intersection between the hit genes which were obtained by the existing medium- to high-throughput studies of cell motility. There are different explanations: first, some of the assays investigate collective migration, some others, single cell migration. It might be that there is little overlap between the machineries that govern migration in these two situations. Second, each cellular model (such as normal breast epithelial cells, lung cancer epithelial cells, cervix cancer epithelial cells) is depending on specific and partly distinct machineries for cell motility. Third, all cell lines do not have identical basal speed and motility behaviours. As a consequence, genes that change motility parameters for slow cells such as HeLa cells, might have an opposite or even no effect on faster cells, such as H1299 cells. Indeed, the effects of a perturbation on a given cell depend both on the perturbation and its cellular state at the moment of the perturbation. This makes these comparisons difficult.

Nevertheless, we also observe that some "stars" of cell motility as RhoA are identified by all studies. Finally, as for hit list enrichment in previously known genes, little overlap is generally observed between medium- to HT screen outputs in other fields as well [Neumann et al., 2010].

3.2.2 Cell trajectory ontology

A question which is related to motility gene discovery is to know whether there exists an ontology of cell trajectories. The approach to answer it would be to apply unsupervised clustering methods on the whole trajectory dataset and try to identify a number of motility patterns for which the clustering is of good quality. This is measured by cluster quality indices, which depend on the clustering method (see e.g. [Tan et al., 2005, Chapter 8], or [Halkidi et al., 2001]). As an example, two common indices to evaluate the output of k-means are the intra-cluster cohesion $C(K)$ and the silhouette score $S(K)$. They both compare intra-cluster distances to inter-cluster distances, if N is the number of data points centered in c_{data} :

$$C(K) = \sum_k \sum_{x \in \mathbb{C}_k} \frac{d(x, c_k)^2}{d(x, c_{data})^2}$$

$$S(K) = \frac{1}{N} \sum_x \frac{b_x - a_x}{\max(a_x, b_x)}$$

where

$$a_x = \text{mean}\{d(x, y) | y \in \mathbb{C}_{k_x}\}, \quad b_x = \min_{k \neq k_x} \text{mean}\{d(x, y) | y \in \mathbb{C}_k\}$$

A slope change in $C(k)$ and a maximum in $S(k)$ are expected at the appropriate number of clusters, if it exists. Other interesting measures of clustering quality encompasses for example the Bayesian Information Criterion or clustering stability [Ben-Hur et al., 2002].

This approach did not prove to be successful when applied to pooled trajectories from all experiments, for a wide range of clustering techniques (k-means, Gaussian mixtures models, spectral clustering, fuzzy c-means, kernel k-means with a radial basis function - data not shown). Concretely, no combination of clustering algorithm and cluster number could be found, whose quality was clearly over the quality obtained by the same clustering algorithm and other cluster numbers. It did not seem possible to find any structure in the data.

Nevertheless, clustering succeeded when only trajectories from MotIW **hit** experiments were pooled together. Indeed, this small subset contains only experiments which have

been selected for being significantly different of controls in terms of single cell motility: it is enriched in rare trajectories. Our interpretation is the following. Let us assume that we have a number of k clusters in the whole trajectory dataset. Due to biological variability, each trajectory is at a certain random distance (in the feature space) to its latent trajectory cluster center. Given the dataset size (approximately 50 million trajectories), this produces a continuous dataset in the feature space, preventing the identification of any cluster. Furthermore, the different clusters are unbalanced. Uniformly subsampling the whole dataset is therefore inefficient as well (data not shown). Applying MotIW to find experiments which are enriched in rarer trajectories enabled us to identify the underlying cluster structure of the dataset, since it performed a stratified subsampling with respect to the latter.

After retaining the first seven principal components (explaining 95% of the variance), k-means was applied to the resulting dataset of approximately 1.1 million pooled trajectories. Fig. 3.1 shows the evolution of intra-cluster cohesion and silhouette score with respect to the number of clusters. It points to $k = 8$ as being both the best and a good quality clustering on this dataset. Indeed, a break and a maximum are respectively expected in the cluster cohesion and the silhouette score curves at the correct cluster number, if it exists. It is compared to the evolution of those indices, if k-means is applied to a uniformly random dataset and a normal dataset (in \mathbf{R}^8).

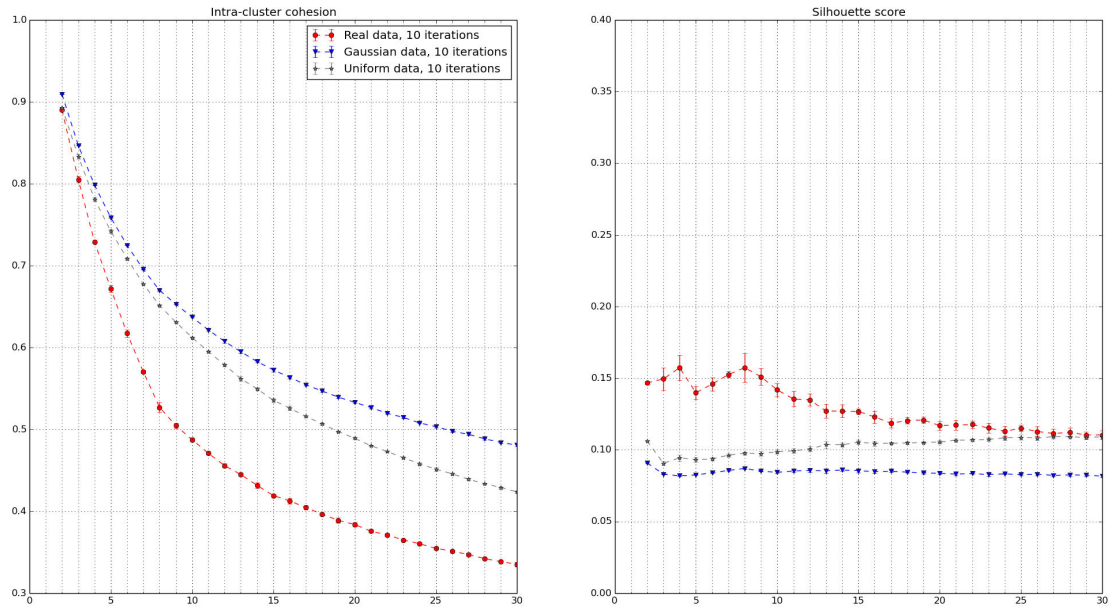


FIGURE 3.1: Evaluation of k-means clustering quality as a function of the number of clusters (average and standard deviation on 10 algorithm initializations). The same protocol was applied to a subset of the Mitocheck dataset, and two samples of the same dimensions, respectively drawn from the Uniform and the Normal distributions.

A first insight on cell motility from this clustering is presented figure 3.2. It presents the distribution of trajectory distributions in the eight identified clusters. One can observe

that there are no cluster which is specific to either controls or hit experiments. From this we see that rather than creating new modes of movement, gene silencing rather leads to shifting the probability of entering a certain mode of movement. This can be explained by the fact that a certain number of motility behaviours are possible for the cell to adopt at any time. Its molecular and cellular state, as well as the stochasticity of gene expression dictates which one it chooses. Hence some behaviours will be rarer in control videos, but become more frequent as siRNA exposure modifies cell molecular state.

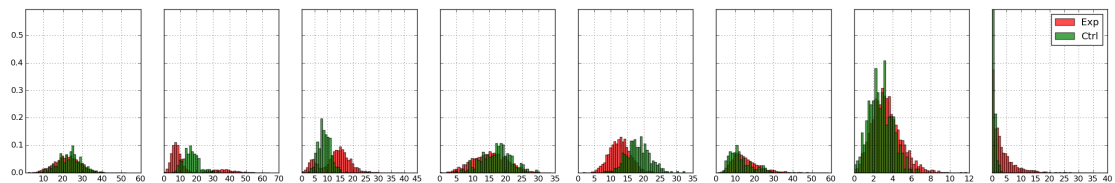


FIGURE 3.2: Comparison of cluster distributions between controls (Ctrl) and experiments (Exp) for the eight trajectory clusters which were identified in the Mitochek dataset. The clusters are in the same order as in figure 3.3.

The cluster characteristics are illustrated in fig. 3.3. Each column in the heatmap corresponds to one cell trajectory, for which the rows show the standard scores of a subset of features. The single cell trajectories are arranged according to the cluster to which they belong. For each cluster, we randomly selected the same number (1000) of trajectories.

A result about single cell motility patterns is obtained from experiments which were selected on the basis of their trajectory feature distributions. This shows that meaningful single cell information can be retrieved by our statistical procedure, which works at the experiment level.

In the second place, it shows that there is more than speed for differentiating trajectory types. For example, clusters 2 and 3 present very similar MSDs and *Effective space length*. However, trajectory curvatures are different: the features *Mean curvature* and *Straightness index* are quite distinct between the two clusters. This can be observed in the Supplementary movie, where cells whose trajectory belongs to cluster 2 (green) are much straighter than those belonging to cluster 3 (red). In this video, cells whose trajectory passed the trajectory quality control have a dot, whose colour corresponds to its cluster as indicated in fig. 3.3.

3.3 Cell cycle length study

The combination of the Mitochek dataset and of our methodological workflow is also very well suited to study cell cycle genes. Indeed, one only needs to combine tracking and

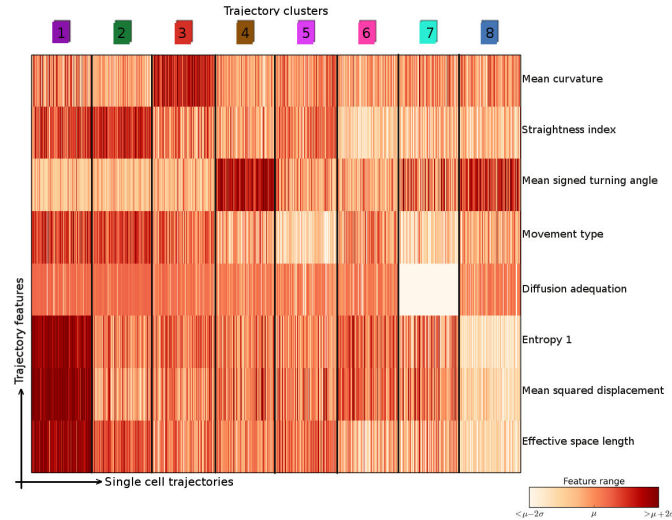


FIGURE 3.3: Characterization of our ontology of trajectories. Each column is a single cell trajectory ; trajectories are grouped by cluster label. 1,000 trajectories were randomly selected per trajectory cluster.

nucleus classification to recover a set of complete trajectories in each experiment, that is, trajectories which start with a mitosis and end with a mitosis. Using the methodological procedure described in 2.2.3, it is then possible to detect the experiments in which cell cycle length distribution is significantly different from that in control experiments. This results in a list of genes.

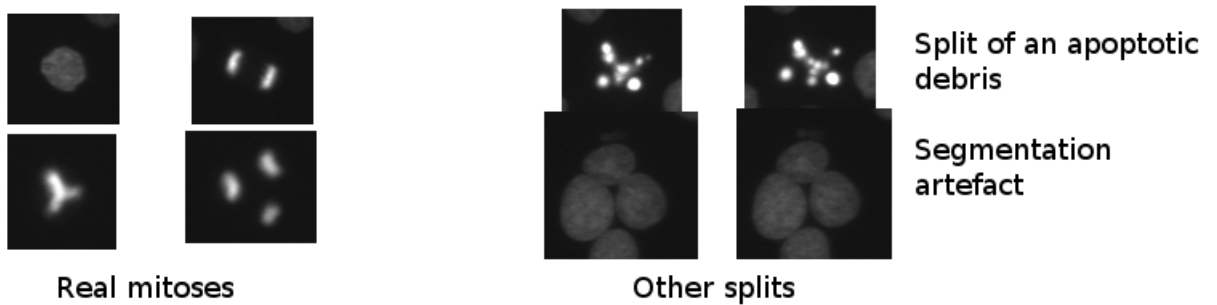


FIGURE 3.4: Examples of object divisions from the Mitocheck dataset

3.3.1 Complete cell cycle detection

The first step to study cell cycle length distribution according to siRNA exposure is to filter out complete trajectories from the others. By complete, we mean trajectory which start with a mitosis rather than the end of a *merge* event, an apparition, the split of an apoptotic debris or the beginning of the film, and which end with a mitosis rather than a *merge*, a disappearance or the end of the film. The distinctions are easily made by a human eye, as can be seen on figure 3.4. To automatize this filtering, one can rely on the

nucleus classification as described in section 3.4.1.3 and [Neumann et al., 2010]: prior to a mitosis, the nucleus will likely be observed in $M_{-1} = \{prometaphase, metaphase, metaphase\ alignment\ problem\}$. Similarly, following a mitosis, the nucleus will likely be observed in $M_{+1} = \{anaphase\}$.

However, given that the classifier is not 100% correct, it is not fully certain. Furthermore, there could be cases of accelerated mitoses, in which it would not be possible to observe both the mother cell in M_{-1} and the daughter cell in M_{+1} . Hence we developed a scoring approach with respect to each track of interest τ going from T_0 to T_f , as detailed in figure 3.5:

$$score_{1,\tau} = \mathbf{1}(Mother_\tau \in M_{-1}) + \mathbf{1}(\tau_{T_0} \in M_{+1})$$

$$score_{2,\tau} = \mathbf{1}(\tau_{T_f} \in M_{-1}) + \sum_{\tau' \text{ s children}} \mathbf{1}(c \in M_{+1})$$

Briefly, the scores compute the number of *right* classifications in the starting and ending splits. To evaluate where to threshold those scores in order to filter out the unwanted tracks, 20 movies were randomly sampled from the Mitocheck dataset, all splits scored and all 2,100 tracks manually divided into complete and other trajectories. Selecting tracks longer than 1 frame with $score_1 \geq 1$ and $score_2 \geq 1$ enables to select more than 87% complete trajectories.

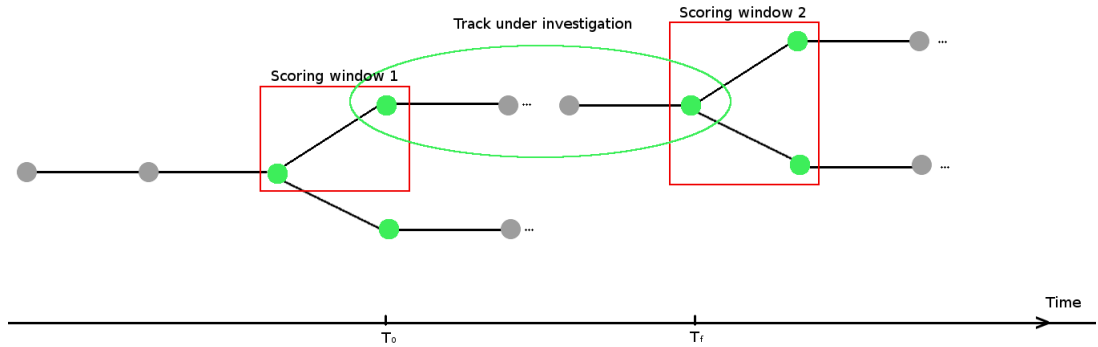


FIGURE 3.5: Approach to cell cycle study

3.3.2 Cell cycle length hit list

The principle of MotIW's statistical procedure (see 2.2.3 and [Schoenauer Sebag et al., 2015]) was then applied to the distributions of complete trajectory length: 2 sample Kolmogorov-Smirnov tests were realized between cell cycle length distributions of each experiment and the controls of the same batch. It was compared to the empirical null distribution in a second time. The latter is the distribution of 2 sample Kolmogorov-Smirnov tests comparing controls to controls. Finally, the empirical p-values which were

obtained were adjusted for multiple testing, following the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995].

Setting a threshold of 0.05 enabled the identification of 66 genes whose cell cycle length distribution differs from that of controls from the same batch. The list is fully provided in appendix, see section A.1. Interestingly, the down-regulation of only three genes produces a *longer* cell cycle length: APOA1, RPS20, SFMBT2 (coding respectively for the apolipoprotein A1, ribosomal protein S20 and Scm-like with four mbt domains 2). SFMBT2 silencing has already been found to decrease cell growth in human prostate cancer cells [Lee et al., 2013].

Gene Ontology analysis reveals that there are three annotation clusters which are highly enriched in this list. The first one is related to protein kinase activity. It contains such genes as BMPR-IB which encodes the bone morphogenetic protein receptor, type IB, and whose reduced expression is correlated to poor prognosis in breast cancer [Bokobza et al., 2009], and tumor grade in prostate cancer [Kim et al., 2000]. The second cluster is related to nucleotide binding. It contains genes such as the integrin-linked kinase whose silencing has interestingly been found to slow cell cycle in human gastric carcinoma cells [Song et al., 2013], whereas we have found its silencing to speed cell cycle in HeLa cells. Finally, the third cluster contains genes which encode proteins which are intrinsic to plasma membrane, such as the melatonin receptor 1A whose absence has been found to be correlated with bad prognosis in triple-negative breast cancer [Oprea-Ilie et al., 2013].

3.3.3 Discussion

Our approach for studying cell cycle length enabled us to obtain a list of 66 genes which may be involved in cell cycle regulation. Gene Ontology analysis revealed that a certain number of these genes has already been found to be linked to cell cycle duration regulation. This list contains 3 genes whose silencing lengthens cell cycle, and 63 whose silencing shortens it. It is possible that this more broadly reflects that there are more proteins which play a role of checkpoints rather than cell cycle enhancers, hence the fact that gene silencing experiments produce more experiments where cell proliferation is increased than decreased. This is supported by other studies such as [Moffat et al., 2006], which found 87 (resp. 15) genes whose silencing significantly increases (resp. decreases) the mitotic index of HT29 cells.

Method bias

Our method was applied to the whole Mitocheck dataset, enabling the obtention of

complete trajectory length distributions for all experiments. Experiments with less than 10 complete trajectories were not considered for further analysis. This explains why decreased proliferation genes as provided by [Neumann et al., 2010] could not be found: the proliferation is so low that no mitosis is observed, hence no complete trajectories can be found. Trying to diminish this bias in experiment selection, we also included trajectories which finish with the end of the experiment rather than a mitosis (hereafter called *incomplete* trajectories).

However, this approach was not successful. For some siRNAs, it seems that incomplete trajectory length distribution indeed has the same shift as that of *complete* ones (see the example of arylsulfatase F gene, ARSF, on fig. 3.6). However, for most of the siRNAs, incomplete trajectory length distributions seem to be more dependent on the batch than on the chemical exposure. Two examples are shown fig. 3.7 which concern the genes CACNA1D and DIMT1 (respectively coding for the calcium channel, voltage-dependent, L Type, Alpha 1D subunit and DIM1 dimethyladenosine transferase 1 homolog). The outcome of the statistical analysis consistently proved too noisy to enable the detection of any hit siRNA for incomplete trajectory length (up to the following threshold for adjusted p-values: 0.1).

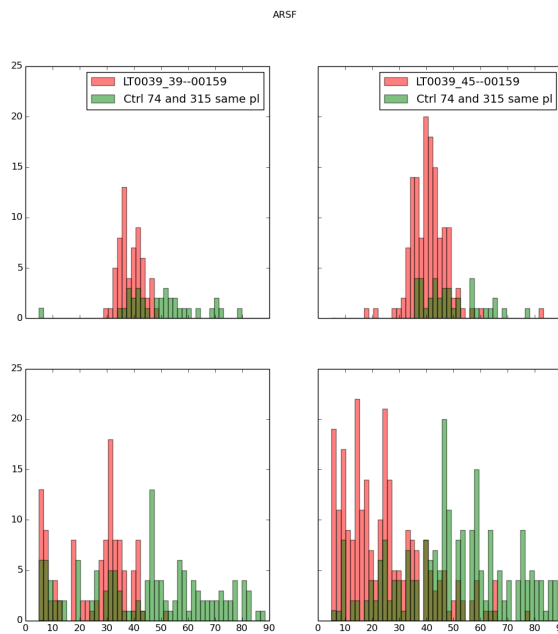


FIGURE 3.6: Histograms showing cell cycle length for complete (top) and incomplete (bottom) trajectories, for two experiments of the Mitocheck dataset concerning ARSF which were detected as significantly different from controls for cell cycle length.

Perspective: cell cycle phase detection

Cell cycle can furthermore be split in four sequential phases: G_1 , S , G_2 , and M . S is

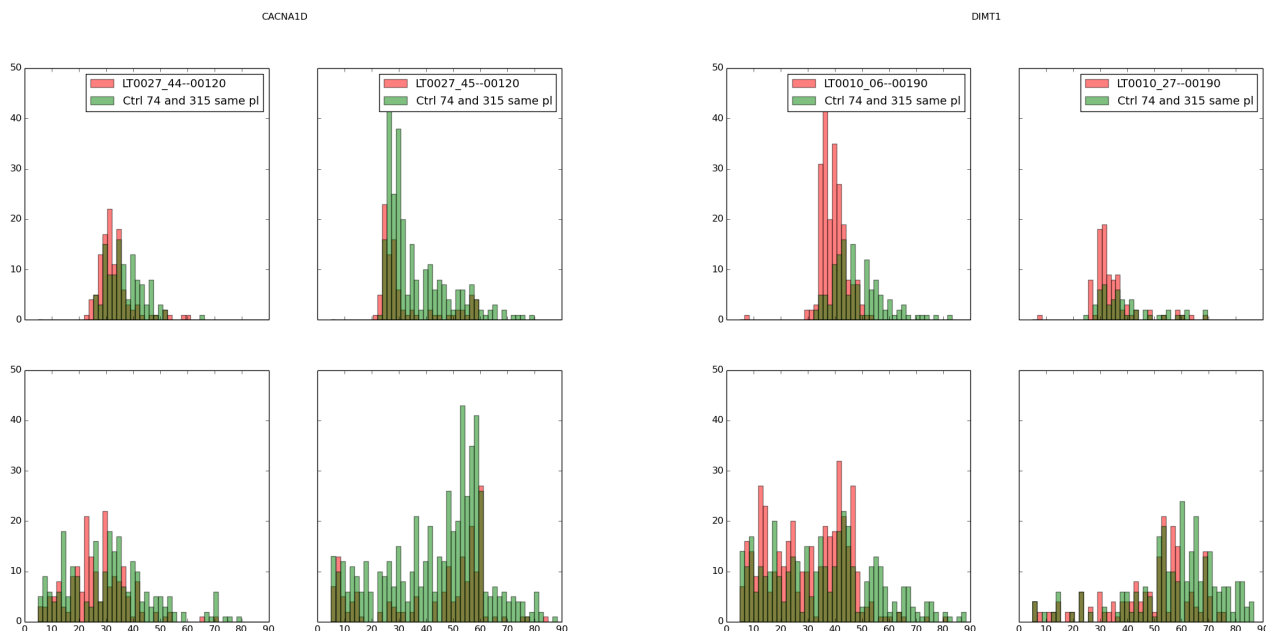


FIGURE 3.7: Histograms showing cell cycle length for complete (top) and incomplete (bottom) trajectories, for two experiments of the Mitochondrial dataset concerning CACNA1D (left) and DIMT1 (right), which were detected as significantly different from controls for cell cycle length.

the DNA replication phase and M stands for mitosis. To distinguish S from G_1 and G_2 , PCNA (proliferating cell nuclear antigen) is usually used. It forms sparkling nuclear dots during DNA replication, as it is recruited to replication foci [Leonhardt et al., 2000].

However, the only marker available in the Mitochondrial dataset is the core histone 2B (H2B). The question is then to know whether it is possible to use this marker for S detection. [Coquelle et al., 2006], using cell size and H2B-GFP fluorescence, managed to FACS-purify RKO cells in G_1 , S and G_2 . Furthermore, [Loo et al., 2007] uses a "cell cycle heuristic" to separate between the different phases of the cell cycle using DNA size and intensity as provided by Hoechst 33342 as a DNA marker (Supplementary figure 3,b-c). They use it to infer links between small molecule exposure and cell cycle modification; however they do not prove the accuracy of their heuristic.

This led us to the hypothesis that it might be possible to use the information contained in the H2B-GFP fluorescence signal for *in silico* sorting of HeLa cells. As expression levels vary between cells, we cannot expect that intensity features will be directly informative, but we hypothesized that changes in nuclear intensity and size should give us cues on potential G_1-S transitions. However, the problem turned out to be less straightforward. As shown on fig. 3.8, nuclei grow linearly during cell cycle without showing any clear slope break. This is confirmed on fig. 3.9. This latter plot has the same configuration as that of Supp. fig. 3 from [Loo et al., 2007]. However, although this study uses a heuristic

model that defines a clear separation between nuclei in different cell cycle phases, we do not find it in our data. Nevertheless, nuclear intensity and size are clearly linked to cell cycle phases.

Hence to answer this question, we had the idea to use a published dataset of HeLa cells which were stained for both H2B and PCNA [Held et al., 2010]. Cell cycle phase annotations of this dataset are available on the Cell Cognition website¹. This training set was created using the information on the PCNA channel. It makes it possible to test the following hypothesis: can a classifier be trained, which uses H2B information for cell cycle phase identification? Preliminary work shows that all tested methods have an accuracy below 70% (random forest, gradient boosting, support vector machine, logistic regression). However, using this information, cell tracking, and Hidden Markov Models for error correction can significantly improve these results.

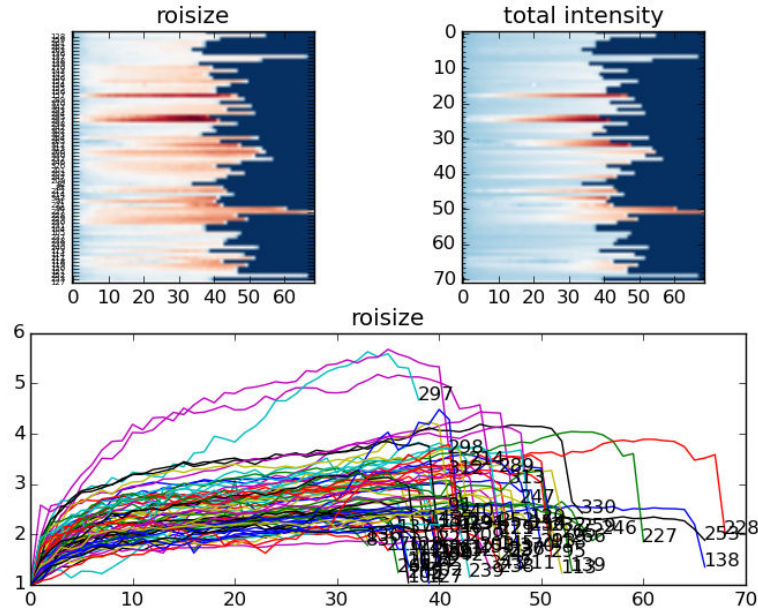


FIGURE 3.8: Example of the time evolutions of nuclear size ("roisize", top left and bottom) and nuclear intensity ("total intensity", top right) for all complete trajectories of a control experiment from the Mitocheck dataset. As discussed in the text, no clear slope break is seen for most trajectories for any of the two indicators, hence preventing the delimitation of cell cycle phases using only this information.

¹<http://www.cellcognition.org/downloads/data>

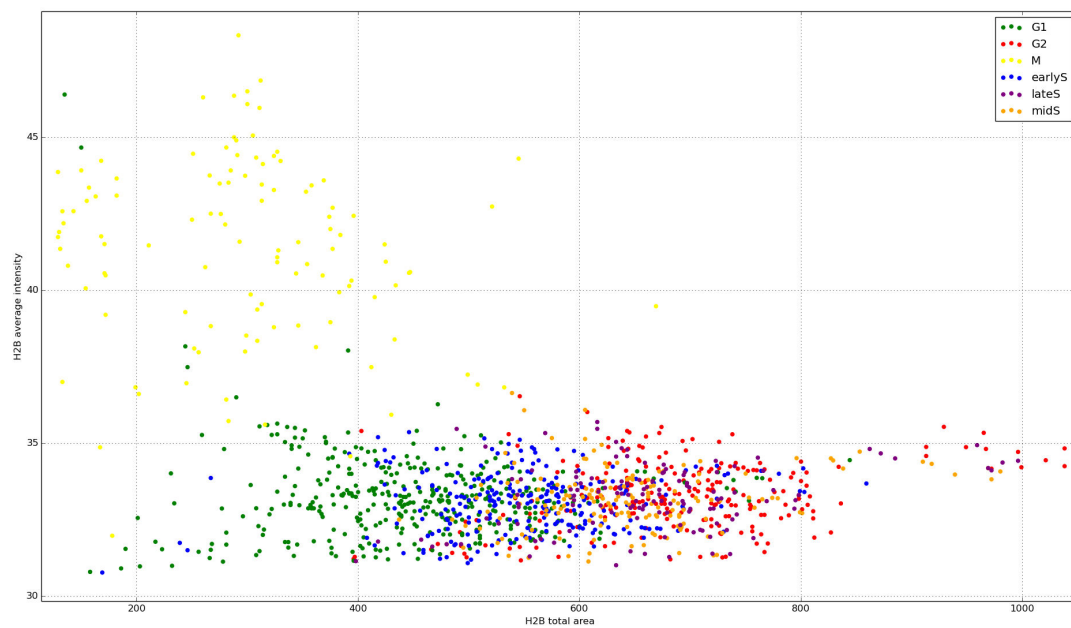


FIGURE 3.9: DNA intensity and size as provided by H2B-GFP information is not sufficient to differentiate between the different cell cycle phases. Data and labelling come from the PCNA dataset.

3.4 Functional inference by in silico comparison of small-molecule and siRNA screens

Finally, on top of enabling the discovery of cell cycle and motility genes, and identifying an ontology of cell trajectories, the Mitocheck dataset can be re-used for drug target inference².

Indeed, genome screens contain very rich information about how selective gene knock-down perturbs normal cell processes such as cell division, cytoplasmic morphology or motility. If cells are screened against drugs or putative drugs in the same experimental context (e.g. the same seeding density, the same markers, the same cell line), the comparison between drug exposure profile and genetic profiles can point at the biological processes which the drug targets [Parsons et al., 2006]. The comparison between gene knockdown and drug phenotypic profiles is therefore more and more considered as an efficient tool for drug discovery [Feng et al., 2009]. In principle, it should allow for the systematic and direct identification of drug target, as opposed to targeted bioassays which demand specific prior knowledge or an extreme amount of time and means.

This view has been applied under different forms in the recent years, either in the comparison of yeast deletion mutants to drug exposed yeast [Ohnuki et al., 2010], or the comparison of parallel RNAi and small molecule screens in *Drosophila* [Parsons et al., 2006] or HeLa cells [Young et al., 2008].

However, when applied, it rapidly appears that these comparisons cannot directly point at *the* drug target or mode-of-action. This is mainly due to two reasons, which are biological rather than technical. On the one hand, drug action can often not be summarized by a unique "entry point" into an organism's cellular processes [Schirle and Jenkins, 2015]. On the other hand, two drugs which are targeted at different molecules albeit occurring in the same pathway will very probably produce the same phenotypic profiles. Inferred targets remain to be confirmed and/or refined [Schenone et al., 2013].

A difficulty of another order is the complexity of HCS data and therefore the difficulty of exploiting it. For the same reason as the preference for univariate measures of cell motility, targeted bioassays can be preferred because they are easier to understand and exploit. Multivariate analysis methods for HCS profile comparison are however more and more developed [Loo et al., 2007], [Young et al., 2008]. Nevertheless, drug target inference from HCS **time-lapse** data has to our knowledge never been done.

This section therefore presents different methods for drug target inference by phenotypic profile comparison between siRNA and drug screen experiments in HeLa H2B-GFP

²Manuscript in preparation

cells. After introducing the dataset and the image analysis methods which were used (section 3.4.1), six distances between phenotypic profiles are introduced and compared (section 3.4.2) before being applied to drug similarity evaluation and target inference in section 3.4.3.

3.4.1 Materials and methods

3.4.1.1 Experimental work

This data set was not produced in the context of this PhD. Rather, the experiments were conducted at the EMBL (Heidelberg, Germany) by Beate Neumann, Jutta Bulkescher and Thomas Walter. See table 3.3 for a list of the drugs and doses which were tested.

TABLE 3.3: Selected drugs and dose ranges. All drugs were tested for 11 doses.

Drug	Dose 0 (μM)	Dose 10 (μM)
Acyclovir	0.036	36.768
(-)-Adenosine 3	0.025	25.152
Aminopurine, 6-benzyl	0.036	36.76
Anisomycin	0.03	31.209
Azacytidine-5	0.033	33.907
Camptothecine(S,+)	0.023	23.77
Daunorubicinhydrochloride	0.014	14.682
Dexamethasoneacetate	0.019	19.057
Doxorubicinhydrochloride	0.014	14.277
Epiandrosterone	0.028	28.509
Etoposide	0.014	14.068
Hesperidin	0.013	13.562
Idoxuridine	0.023	23.384
JNJ7706621	0.008	8.28
MLN8054	0.008	8.28
Methotrexate	0.018	18.221
Nocodazole	0.027	27.48
Paclitaxel	0.009	9.697
R763	0.008	8.28
Ribavirin	0.033	33.907
Sulfaguanidine	0.038	38.647
Sulfathiazole	0.032	32.43
Thalidomide	0.031	32.065
VX680	0.008	8.28
Zidovudine, AZT	0.03	30.984

3.4.1.2 Object segmentation

Since we are interested in nuclear morphologies in this context, the original segmentation of the Mitocheck project was used, as previously described [Walter et al., 2010].

3.4.1.3 Object classification and phenotypic scores

The use of the original segmentation from the Mitocheck project made it possible to re-use its training set, albeit strengthened for classes which were previously slightly under-represented. This can be seen on fig. 3.10 in comparison with figure 3 from [Walter et al., 2010]. A visual inspection of the dataset enabled us to verify that the drug screen experiments did not contain new morphological classes (that is, absent in the Mitocheck experiments). This would have made it necessary to include nuclei from drug screen experiments into our training set.

Cell Cognition [Held et al., 2010] was used for learning an RBF (Radial Base Function) kernel SVM classifier, whose precision and recall are also indicated on fig. 3.10. Its parameters were optimised by grid-search ($\gamma = 2^{-7}$, $C = 8$).

	Name	Samples	Color	%PR	%SE
1	Interphase	420		90.4	89.5
2	Large	80		86.0	92.5
3	Elongated	110		92.9	94.5
4	Shape1	346		91.5	93.4
5	Shape3	473		89.9	84.4
6	Grape	99		71.2	74.7
7	Metaphase	74		84.0	85.1
8	Anaphase	85		88.6	91.8
9	MetaphaseAlignment	176		86.7	81.8
10	Prometaphase	345		85.7	85.2
11	ADCCM	99		85.6	83.8
12	Apoptosis	308		90.7	88.6
13	Hole	114		89.3	80.7
14	Folded	58		56.9	63.8
15	SmallIrregular	165		71.2	82.4
16	Artefact	112		80.7	85.7
17	UndefinedCondensed	47		57.9	70.2
18	OutOfFocus	325		99.4	96.6
#	overall	3436		87.5	87.1

FIGURE 3.10: Precision and recall per class as provided by Cell Cognition. Compared with the original classifier as published in [Walter et al., 2010], classes *ADCCM* (Asymmetric Distribution of Condensed Chromosome Masses) and *Out of focus* were added. More nuclei were furthermore included for training in most classes. *Shape1* (resp. *Shape3*, *MetaphaseAlignment*) corresponds to binucleated (resp. polylobed, metaphase alignment problem) nuclei.

This provides a representation of each video as a set of time-series, which are the evolution of the percentage of nuclei in each phenotypic class over time. To evaluate how an experiment i diverges from control experiments from the same batch C_i for its temporal evolution of class k , we used phenotypic scores as previously described [Walter et al.,

2010]. Briefly, temporal evolutions of the percentages of nuclei in class k for experiment i , $(\%_{k,i,t})_t$ and its controls $(\%_{k,C_i,t})_t$ are regularized using a locally weighted scatterplot smoothing as implemented in the Python package statsmodels. The fraction of data points which is used for smoothing was manually chosen to be $f = 50\%$. The maximum deviation between the two regularized time series is then computed, where ps stands for "phenotypic score":

$$ps_{k,i} = \max_{0 \dots T} (\%_{k,i,t}^{reg} - \%_{k,C_i,t}^{reg})$$

3.4.1.4 Quality control

One plate had to be eliminated due to an issue during image acquisition. An abrupt increase in fluorescence intensity around the 80th frame of all experiments from this plate causes a sudden increase in the number of detected object which prevents any time-consistent analysis.

For the other experiments, a threshold of c cells at the beginning of the movie, and maximum $p\%$ out-of-focus objects were used to remove unexploitable movies. $c = 50$ and $p = 40$ were selected. Out-of-focus objects and cells that were neither artefacts nor out-of-focus objects were identified following segmentation, feature extraction and classification as described supra.

Out of 1,232 experiments on four plates, 904 experiments from three plates passed the quality control, among which 98 control experiments.

3.4.1.5 Selection of Mitochondria experiments for target inference

The Mitochondria project led to the identification of 1,249 mitotic hits in the primary screen. 1,042 were identified by manual thresholds on phenotypic scores from the following phenotypic classes: *Prometaphase*, *Metaphase Alignment Problem*, *Binucleated*, *Polylobed*, *Grape*. 207 genes were further identified by manual annotations. 1,128 of these genes were screened again in a validation screen.

Lists of hit genes according to the following measures were also published:

- phenotypic score of *Large* nuclei,
- cell death, as measured by phenotypic score of *Apoptosis* nuclei,
- nuclear dynamic changes, as measured by the sum of phenotypic scores for *Hole*, *Folded* and *Small irregular*,

- and cell proliferation.

Finally, a list of hit experiments for *Elongated* nuclei was computed, which can be found as a supplementary to this thesis (see Supp. table 2).

Given the evolution of the reference sequence of the genome, not all those genes were in fact targeted in the Mitocheck experiments. An updated mapping of the siRNAs which were used in the primary and validation Mitocheck screens to the present reference sequence of the genome was graciously provided by Jean-Karim Hériché (EMBL, Heidelberg, Germany). Once this and the quality control are taken into account, the final list of hits in at least one of the listed categories amounts to 2,614 genes (cf fig. 3.11), which are covered by 4,847 siRNAs.

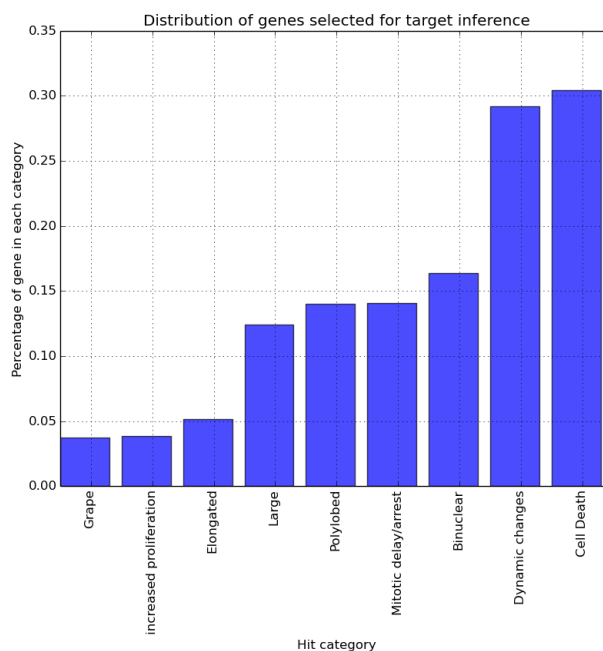


FIGURE 3.11: Number of hit genes per category. As hit detection is univariate, a gene can be in more than one category.

Given the variations in siRNA coverage between genes which were, for example, included or not in the validation screen, each gene was chosen to be represented by the siRNA which showed the maximum effect. This was measured by the median of the phenotypic scores for *Interphase* nuclei of this siRNA experiments.

3.4.1.6 Detection of drug screen hit experiments

Phenotypic scores of the drug screen experiments were computed as described supra. Experiments whose *Interphase* phenotypic score was lower than $Q_1^{ctrl} - 1.5 \times IQR^{ctrl}$ were

selected as hit experiments, where Q_1^{ctrl} and IQR^{ctrl} are respectively the first quartile and the inter-quartile range of control *Interphase* scores. This is a robust one-sided way to select outliers, as the distribution of control *Interphase* scores cannot be assumed to be Gaussian.

It corresponds to the 197 experiments which are under the bottom whisker on the *Interphase* subplot of fig. 3.12. Supplementary plots represent the distribution of phenotypic scores as a function of dose and drug, see "Phenotypic score plots_type1" in "Supplementaries" folder.

Hit conditions are conditions for which strictly more than 50% of their replicates are hit experiments.

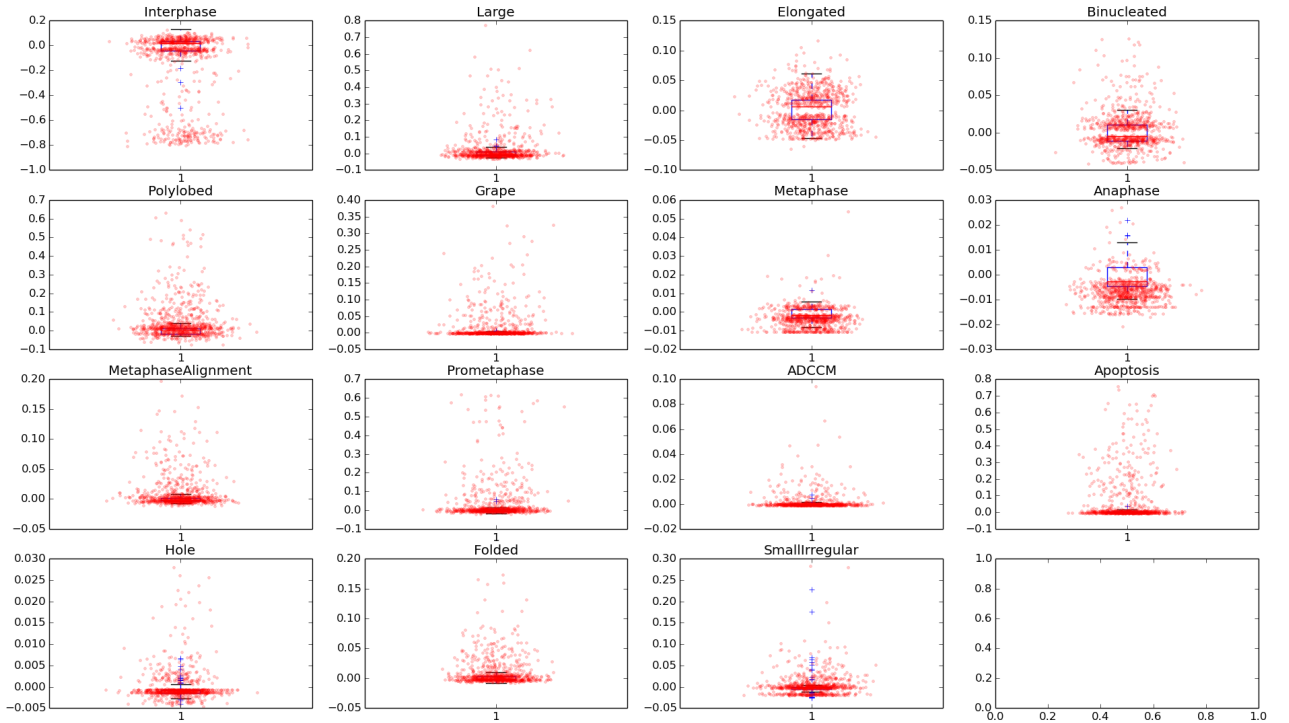


FIGURE 3.12: Distributions of phenotypic scores from the drug screen experiments. Each boxplot corresponds to the distribution of control phenotypic scores, whereas each red dot is an experiment in which cells were exposed to a drug.

3.4.1.7 Other analyses

Proliferation hit detection

Proliferation hit detection was realized similarly to hit experiment detection: experiments whose proliferation rate was higher than $Q_3^{ctrl} + 1.5 \times IQR^{ctrl}$ or lower than $Q_1^{ctrl} -$

$1.5 \times IQR^{ctrl}$ were considered proliferation hit experiments. Proliferation hit conditions were conditions for which strictly more than 50% of their replicates are proliferation hit experiments. Interestingly, only one condition (Azacytidin, dose 7) was a proliferation hit without being a hit.

Hierarchical clustering

Condition hierarchical clustering was performed using the median of condition replicates. The Python package fastcluster [Müllner, 2013] was used to this end. In each case, the clustering method which visually produced the best result was used, from Ward, centroid and single methods.

Target pathway inference and analysis

Target pathways were inferred for groups of selected conditions following hierarchical clustering. For each group, experiments from all group conditions were pooled together and considered as replicates of one artificial condition. This condition was characterized by all experiment distances to Mitocheck experiments. A way to consider each experiment nearest neighbours consistently accross group experiments is to use rank products [Breitling et al., 2004]. Briefly, the idea is that if a gene is consistently close to all experiments, it will consistently be in the top of each ranked experiment-gene list. The product of its ranks will therefore be small. This is the final variable which was used to compute the nearest neighbours for each condition group. The list of 200 nearest genes to each condition group was then analyzed using DAVID online tool [Jiao et al., 2012], and its enrichment in GO terms computed by comparison with the list of Mitocheck hit genes selected as described in section 3.4.1.5.

Data visualization

A Web-based user interface was designed and implemented for result visualization and result sharing among collaborators. It is described more in details infra, see section 4.1.2.

3.4.2 Phenotypic profile distances

Screening experiments provide us with temporal sequences of information. They can be seen as sequences of two-dimensional intensity distributions [Rajaram et al., 2012], sequences of object feature distributions or sequences of phenotypic class distributions. In our case, we chose to summarize each experiment by a set of temporal evolutions of phenotypic classes, that is, we chose to represent our experiments by their phenotypic profile. The question remains to know how to measure the similarity between two instances of this representation of the information.

We decided to test the following distances on the question to know whether we can apply phenotypic profiling for drug target inference from parallel drug and siRNA screens:

- a very simple approach, the Euclidean distance of phenotypic scores,
- a state-of-the-art approach, the phenotypic trajectory distance as defined in [Walter et al., 2010],
- a divergence which enables the use of biological prior knowledge, the Sinkhorn divergence [Cuturi, 2013].

Let i be a screening experiment, and p the number of phenotypic classes ($p = 15$ in our case).

3.4.2.1 Euclidean distance on phenotypic scores

i can be represented in \mathbb{R}^p by $(ps_{k,i})_k$ the vector of its phenotypic scores. The distance between two experiments is then the Euclidean distance of their vectors of phenotypic scores, excluding *Interphase* and *Anaphase* scores. *Interphase* score is excluded as its decrease is most of the time a summary of the increases of other scores. It is therefore the least specific measurement we can look at. *Anaphase* score is excluded because there is no single condition which leads to an accumulation in anaphases: this chromosome configuration is not stable and consequently, cells cannot remain in this phase. An observed accumulation of anaphase is therefore sure to correspond to an artifact (typically observed when anaphase is confounded with apoptosis).

These vectors can be normalized with respect to the mean and standard deviation of phenotypic scores in the dataset. This will correspond to the **Normalized phenotypic score** distance in the following, whereas the non-normalized version will simply be called **Phenotypic score** distance.

These distances are robust to time delay in the onset of phenotypic changes. Indeed, as controls basically show a constant percentage of *Interphase* nuclei, phenotypic scores will be identical for two experiments which show an increase in, e.g., *Apoptosis* nuclei respectively at the beginning and at the end of the experiments. Hence the strength of these distances is that even if one experiment is identical with a delay to another, their distance will be small. It will however still be small if they're distinctly ordered, e.g. if one experiments shows the same phenotypic events than the other, albeit in the opposite order.

3.4.2.2 Phenotypic trajectory distance

On the other hand, it is possible to use the phenotypic trajectory distance as published in [Walter et al., 2010]. Briefly, let us re-use the notations of section 3.4.1.3: i is seen as $(\%_{k,i,t})_{k=1\dots p, t=1\dots T}$, that is, a sequence in $[0; 1]^p$. This sequence is then approximated by two p -dimensional vectors. The phenotypic trajectory distance between two experiments is then a distance between their vectors, as defined in formula 7 of [Walter et al., 2010]. This distance will be called **phenotypic trajectory** distance in the following.

This distance does not take explicitly time into account, but it respects the order of phenotypic changes. Hence its strength is that even if one experiment is identical with a delay to another, their distance will be small. It will not if they're distinctly ordered, as opposed to the phenotypic score distances.

3.4.2.3 Sinkhorn divergence

Motivations

Finally, we wanted to test a distance which would enable us to use some prior biological knowledge of phenotypic class relationships. If we consider the two previous distances, they implicitly consider each phenotypic class to be independent of the others, and equally biologically far away from all. Indeed, the phenotypic score distances operate in \mathbb{R}^p to sum the squared differences of phenotypic scores, hence treating the different phenotypes independently of each other. However, different morphological classes do not all point to entirely different phenotypic situations. For example, *Hole*, *Folded* and *Small Irregular* all point to problems in nuclear stability. Accumulations in *Metaphase Alignment Problems (MAP)* and *Prometaphase* are observed if the mitotic spindle is not capable of aligning chromosomes in the metaphase plate. *Binucleated*, *Polylobed* and *Grape* nuclei arise as secondary consequence of mitotic failures that were not detrimental to the cell. Hence the biological intuition is that a chemical causing a great increase in *Polylobed* nuclei has probably a closer mode of action to that of another drug causing an increase in *Grape* nuclei than to another drug causing a strict increase in *Apoptosis*.

Ideally, the idea is then that the distance of $a\%$ *Grape* nuclei to $b\%$ *Polylobed* nuclei is smaller than that to $b\%$ *Apoptosis* nuclei, or that it "costs" less to go from $a\%$ *Grape* nuclei to $b\%$ *Polylobed* nuclei than to $b\%$ *Apoptosis*. This is precisely the idea behind the Earth Mover's distance (or transportation distance, or Wasserstein distance). This distance was developed in the first place to compute the cost to move a certain number of piles of dirt into a certain number of holes. To optimally do so, one needs to take into account the distance between piles and holes.

Definitions

Let us formalize this intuition and briefly introduce transportation distance³. We note $\Sigma_d = \{x \in \mathbb{R}_+^d | x^T \mathbf{1}_d = 1\}$ the probability simplex. In our case, $d = 13$: we can consider either the distributions of phenotypes in a given experiment over all time-points, or this distribution in a specific frame.

Given r and c in Σ_d , the transport polytope $U(r, c)$ is the set of matrices such that

$$U(r, c) = \{P \in \mathbb{R}_+^{d \times d} | P\mathbf{1}_d = r, P^T\mathbf{1}_d = c\}$$

If X and Y are two discrete random variables with values in $\{1, \dots, d\}$ whose distributions are r and c , the elements of $U(r, c)$ are in fact the possible joint probabilities of (X, Y) . Given a cost matrix M in $\mathbb{R}^{d \times d}$, the optimal transportation distance between r and c is the solution of the following optimization program, where $\langle \cdot, \cdot \rangle$ is the Frobenius matrix norm:

$$d_M(r, c) = \min_{P \in U(r, c)} \langle M, P \rangle \quad (3.1)$$

Optimal solutions P^* of 3.1 can be obtained. Furthermore, if M is a metric matrix, this quantity is a distance [Villani, 2009]. This optimization program's complexity is in $O(d^3 \log d)$ in theory and in practice, which makes it less applicable to high-dimensionality problems.

In our case however, $d = 13$, hence complexity is not a serious issue. An issue which is more relevant is that optimal solutions P^* will lie on the vertices of $U(r, c)$. This is due to the linear quality of the optimization problem. It will produce almost deterministic joint probabilities [Cuturi, 2013]. The idea is therefore to solve a regularized version of this program, placing ourselves in the following convex subset of $U(r, c)$, for $\alpha > 0$:

$$U_\alpha(r, c) = \{P \in U(r, c) | \mathbf{KL}(P || rc^T) \leq \alpha\}$$

The Sinkhorn divergence will be the following quantity, for $\alpha > 0$:

$$d_{M, \alpha}(r, c) = \min_{P \in U_\alpha(r, c)} \langle M, P \rangle \quad (3.2)$$

This will produce less deterministic optimal solutions, which will converge to P^* as α increases, while $d_{M, \alpha}(r, c)$ converges to $d_M(r, c)$. In practice, there exists an efficient method for solving the dual of this problem, Sinkhorn fixed-point algorithm [Sinkhorn and Knopp, 1967].

³References and proofs can be found in [Cuturi, 2013].

The solution for obtaining a faster computation of an approximate transportation distance is therefore to solve the dual problem of 3.2. Its solution will be used to compute the Sinkhorn divergence. For any $\alpha > 0$, there exists $\lambda > 0$ such that $d_{M,\alpha}(r, c) = d_{M,\lambda}(r, c)$, with

$$d_{M,\lambda}(r, c) = \langle M, P^\lambda \rangle, \quad P^\lambda = \arg \min_{P \in U(r, c)} \langle M, P \rangle - \frac{1}{\lambda} h(P)$$

The use of this divergence will enable us to take into account prior biological knowledge while computing distances between phenotypic distributions. This knowledge will be encoded in the cost matrix M . Its choice as well as λ 's is described infra. Practically, Sinkhorn fixed-point algorithm was implemented in Python to compute Sinkhorn divergences.

Finally, there are two ways to apply Sinkhorn divergence to the problem at hand. Let us consider two experiments i and j of duration T :

- one can pool all nuclei from all frames by representing the experiments in Σ_d . This distance will be called **global Sinkhorn divergence** in the following:

$$D_{M,\lambda}(i, j) = d_{M,\lambda}((\%_{k,i})_k, (\%_{k,j})_k) \quad (3.3)$$

- one can choose to keep the temporal information by representing the experiments in $(\Sigma_d)^T$. We will define two distances from this:

max time Sinkhorn divergence, which is the maximum of all timepoints Sinkhorn divergences:

$$D_{M,\lambda}^{max}(i, j) = \max_t d_{M,\lambda}((\%_{k,i,t})_k, (\%_{k,j,t})_k) \quad (3.4)$$

sum of time Sinkhorn divergence, which is the sum of all timepoints Sinkhorn divergences:

$$D_{M,\lambda}^{sum}(i, j) = \sum_t d_{M,\lambda}((\%_{k,i,t})_k, (\%_{k,j,t})_k) \quad (3.5)$$

Choice of phenotypic cost matrix

The phenotypic cost matrix summarizes our biological knowledge about the phenotypes which were observed in the Mitocheck dataset (they include those which were observed in the drug screen). We made the choice to set inter-phenotypic costs according to the cellular process which is perturbed when they appear, or which they represent. The phenotypic cost matrix which we chose is illustrated fig. 3.13.

To resume our previous example, *Polylobed*, *Grape* and *Binucleated* nuclei were considered to be closer to each other than to any other phenotype, as they are all non-detrimental failures of division defects. *Binucleated* nuclei were set slightly further apart from *Polylobed* and *Grape* nuclei, because they correspond to cytokinesis defects, whereas the latter correspond to segregation defects. Hence we set a cost of 1 between *Grape* and *Polylobed*, and 2 between *Polylobed* and *Binucleated*, and *Grape* and *Binucleated*. The three classes then have a cost of 3 between them and the different versions of interphases (*interphase*, *Elongated*, *Large*) and different mitotic classes (*Prometaphase*, *Metaphase*, *Anaphase*, *Metaphase Alignment Problems (MAP)*, *ADCCM*), and 5 to *Apoptosis*. Finally, they are at a cost 4 of nuclear stability phenotypes which do not result from cell division defects (such as *Hole* or *Folded*).

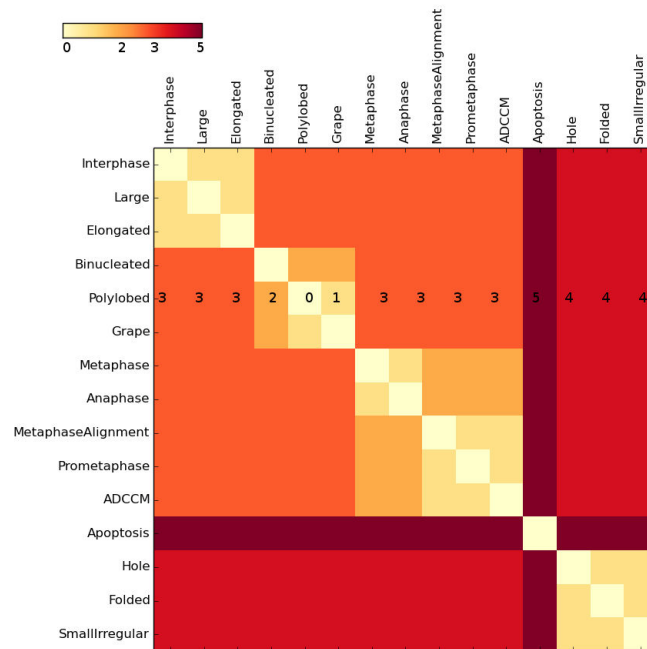


FIGURE 3.13: Cost matrix for phenotypic Sinkhorn divergence

Choice of λ

This choice will determine how close the Sinkhorn divergence is to the transportation distance. As expected, when λ increases, the Sinkhorn divergence converges. This is illustrated fig. 3.14. This figure also enables us to see that in the range of λ which we investigated, there seems to be mainly two different behaviours: one which is shown at $\lambda = 0.01$ and $\lambda = 0.1$, and one which is shown at $\lambda = 1$ and $\lambda = 10$.

Our choice of λ was driven by the ability to differentiate between Mitocheck hit experiments. As detailed in section 3.4.1.5, these experiments are grouped according to

the phenotype(s) of which they present a strikingly high percentage. We therefore visually compared the ability of λ 's two different value ranges to separate Mitocheck hit experiment from different phenotypic hit lists.

This is shown on figures 3.15 and 3.16: Mitocheck hit experiments are represented following the use of multi-dimensional scaling in 2 dimensions of their **global Sinkhorn divergences**. We clearly see that $\lambda = 10$ seems to distinguish - to a certain extent - between distinct phenotypes, as opposed to $\lambda = 0.1$. This is striking if we consider the example of **Binuclear** and **Cell death** hits (see also figures in Appendix, section A.2.1). $\lambda = 10$ was hence chosen.

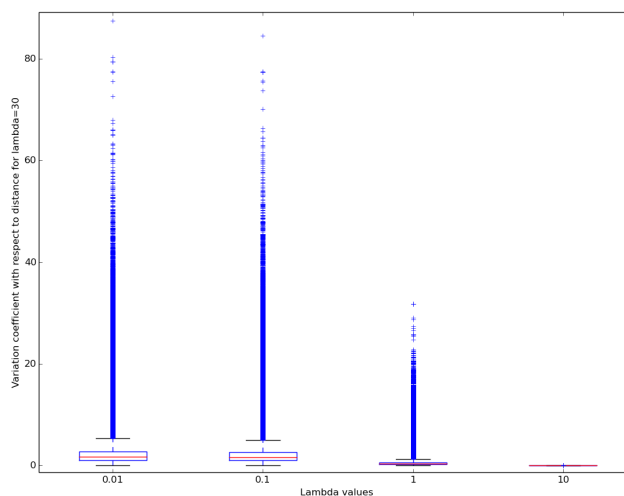


FIGURE 3.14: Convergence of Sinkhorn divergence as a function of lambda. Divergences were computed between drug screen experiments and Mitocheck hit experiments for different values of lambda, and the distribution of their relative variation to the divergences computed for $\lambda = 30$ are showed here.

3.4.2.4 Distance quality evaluation

Six distances were selected to compare phenotypic profiles following drug/siRNA exposure. We then wanted to evaluate their ability to distinguish between different conditions without distinguishing between condition replicates. For this purpose, we computed for each distance d and condition C a separability score $S_d(C)$ as defined in formula 3.6 and a replicability score $R_d(C)$ as defined in formula 3.7. Separability compares the distance between replicates of the same condition to the distance to other conditions, whereas replicability measures the correlation between condition replicates to Mitocheck hit experiments.

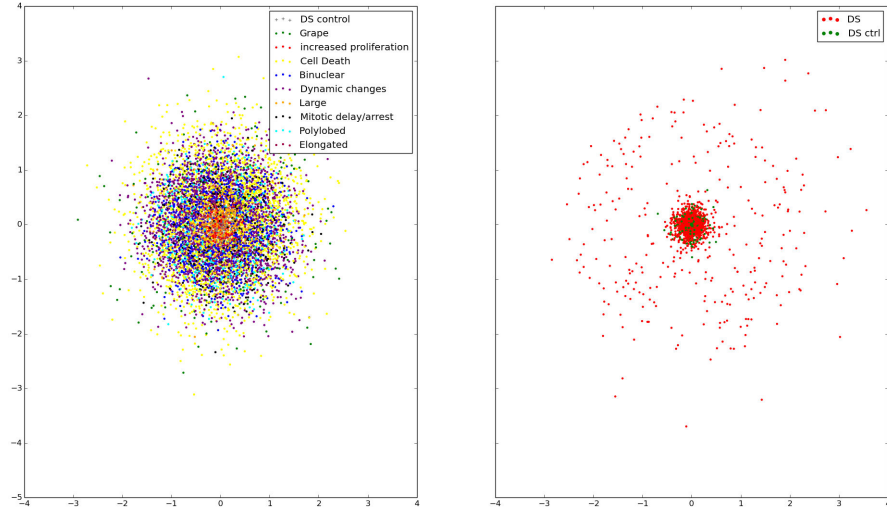


FIGURE 3.15: Separation between Mitocheck hit categories (left) for $\lambda = 0.1$. Global Sinkhorn divergences between Mitocheck hit experiments were computed for $\lambda = 0.1$, and multi-dimensional scaling was used for representing them in two dimensions in the first two lines. Divergences between these experiments and the drug screen were included and their multi-dimension scaling is shown on the right plot.

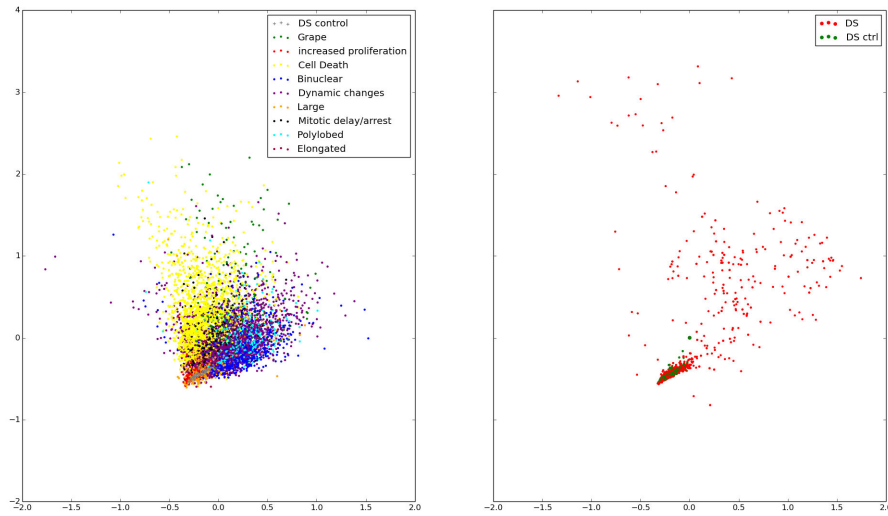


FIGURE 3.16: Separation between Mitocheck hit categories (left) for $\lambda = 10$. Global Sinkhorn divergences between Mitocheck hit experiments were computed for $\lambda = 10$, and multi-dimensional scaling was used for representing them in two dimensions in the first two lines. Divergences between these experiments and the drug screen were included and their multi-dimension scaling is shown on the right plot.

Notations: for each experiment i we note C_i its condition, $d(i, M)$ the vector of distances between i and all Mitocheck hit experiments and $corr$ the Pearson correlation.

$$S_d(C) = \frac{\sum_{i|C_i=C} \sum_{k|C_k \neq C} d(i, k)}{\sum_{i|C_i=C} \sum_{j \neq i|C_j=C} d(i, j)} \quad (3.6)$$

$$Ra(C) = \frac{2}{(n-1)(n-2)} \sum_{C_i=C}^i \sum_{C_j=C}^{j \neq i} corr(d(i, M), d(j, M)) \quad (3.7)$$

The results are presented fig. 3.17. We can observe that all investigated distances score the same on average in terms of replicability and separability on drug screen hit conditions ; they are more different on all drug screen conditions. In the latter case, Sinkhorn divergences and the simple phenotypic score distance seem to better separate conditions than normalized phenotypic score and phenotypic trajectory distance, although not significantly.

In both cases, there is a high standard deviation, as some conditions were visually observed to have lower reproducibility levels than the others. One can for example consider the reproducibility of the 10th dose of JNJ7706621 (see fig. A.3 in appendix, section A.2.2). This high standard deviation is therefore at least partly experimental, which means that none of the investigated distances is robust enough to cover it.

Based on these results, we chose to restrict ourselves to the phenotypic score distance, the phenotypic trajectory distance, the sum of time and global Sinkhorn divergences.

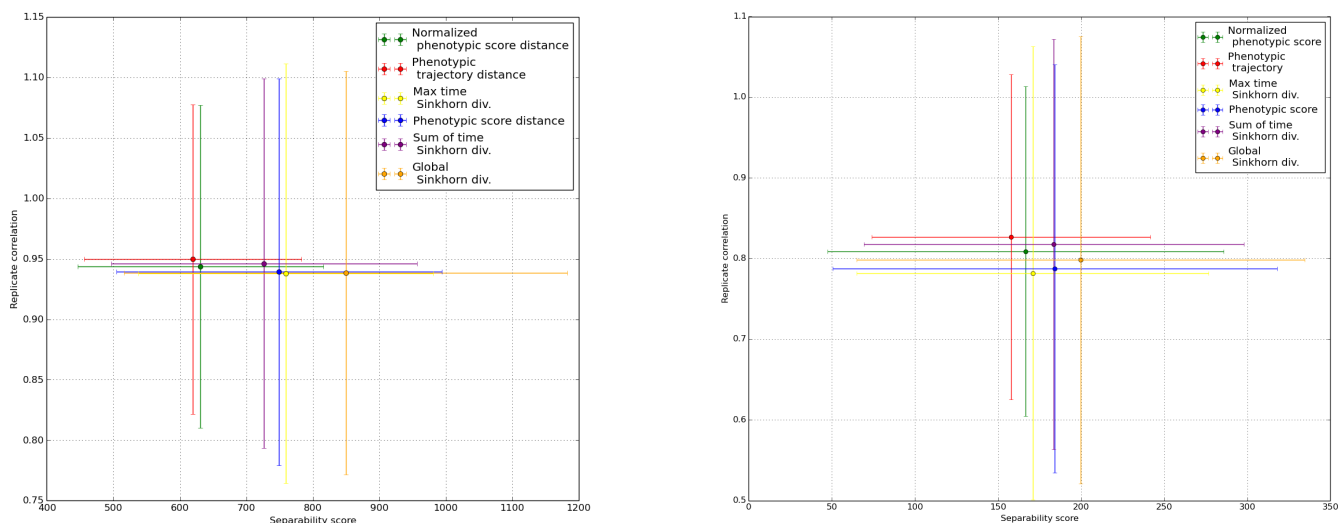


FIGURE 3.17: Mean separability and replicability scores of investigated distances on all conditions (left) and hit conditions only (right - bars represent standard deviations).

3.4.3 Applications

There are two main applications of phenotypic profiling to drug or small molecule screens [Loo et al., 2007]. On the one hand, one might be interested in studying the

similarity between the different conditions cells were exposed to. If there are drugs with known targets, we can then hypothesize that drugs with similar phenotypic effects will target the same pathway. On the other hand, if there exists a parallel genetic screen, that is, a genetic screen which was performed in the same experimental conditions, one might be interested in comparing phenotypic profiles resulting from gene silencing or over-expression to phenotypic profiles resulting from drug exposure. These two applications could lead to the inference of possible targets for unknown chemicals, and might even give a hint as to a possible mode of action.

3.4.3.1 Small molecule similarity evaluation

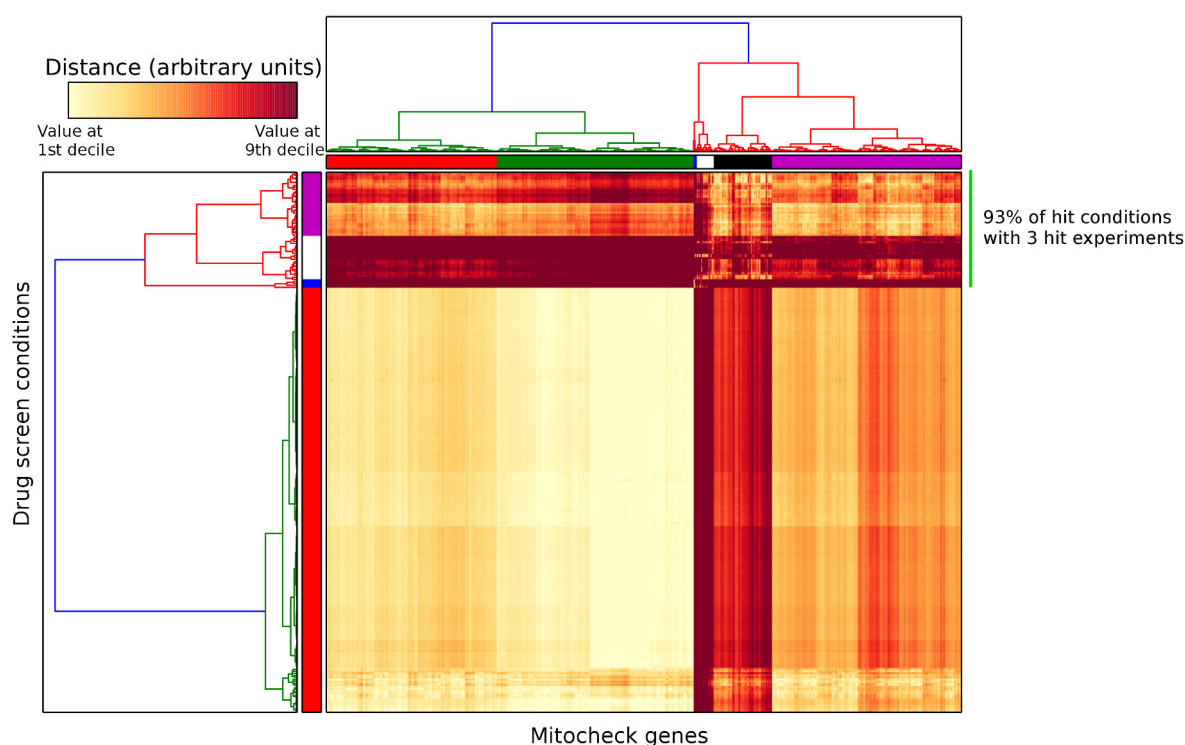


FIGURE 3.18: Drug screen condition - Mitochondrial siRNA two-dimensional hierarchical clustering using global Sinkhorn divergence. Ward method was used in combination with the Euclidean distance.

Condition clustering

The median distance of each condition replicates to Mitochondrial hit experiments was used to perform condition hierarchical clustering, for each investigated distance. The output for the global Sinkhorn divergence is illustrated fig. 3.18, and outputs for the other distances can be seen in appendix (see section A.2.3).

From 89% to 100% of (drug screen) hit conditions are clustering together, depending on the used distance. Furthermore, the clusters they belong to are composed of hit

conditions at purity levels ranging from 85% to 91%. The dendrogram shows that these clusters present subclusters, which are at a certain distance from each other (i.e. the dendrogram is not flat - see purple, white and blue clusters as indicated on the y-absciss colorbar on fig. 3.18). On the other hand, non-hit conditions are grouped in one large and flat cluster (see the red cluster on fig. 3.18). This means that non-hit conditions as described by their distance to Mitocheck hit experiments are not distinguishable, whereas hit conditions are.

We have chosen to summarize experiments by their nuclear phenotypic profiles, that is, the temporal sequence of nuclear phenotype distributions. This also means that any perturbation which do not result in a change in percentage of any of these class distributions cannot be detected. This might concern perturbations that alter nuclear morphologies without changing their class assignment. For example, if the number of nucleoli was changed by a perturbation, the resulting interphase would still be classified as interphase and the change would remain unnoticed. It can also concern perturbations which affect an aspect of cell life that is not measured at all by our assay. For instance thalidomide, whose known teratogenic effects might be linked to an inhibition of ubiquitin ligase [Ito et al., 2010], might simply have no effect on nuclear morphology. Some drugs might also not have any effect on HeLa cells, or demand a longer exposure time for an effect to be detected.

The conclusion of these clusterings is that our method (combining our choice for information representation and similarity evaluation) is suitable for detecting and inferring knowledge regarding mitotic hit conditions and conditions which modify nuclear morphology as described by the phenotypic classes which were chosen, whereas it is not for other conditions. We therefore restrict ourselves to the hit conditions which we have detected as described in section 3.4.1.6.

Hit condition clustering

Hit condition clustering was realized in the same way. The output for the global Sinkhorn divergence can be seen fig. 3.19 ; other distance outputs are displayed in appendix section A.2.4.

Existing knowledge about hit condition targets, as displayed in table 3.4, can be used to evaluate the consistency of clustering results⁴. Ideally, drugs with similar targets and similar doses from the same drug should cluster together. A way to visualize how the different distances perform with regard to that, is to look at the binary maps which are

⁴Unfortunately, our dose ranges are too different from that of [Perlman et al., 2004] to compare our clustering results to those of [Loo et al., 2007].

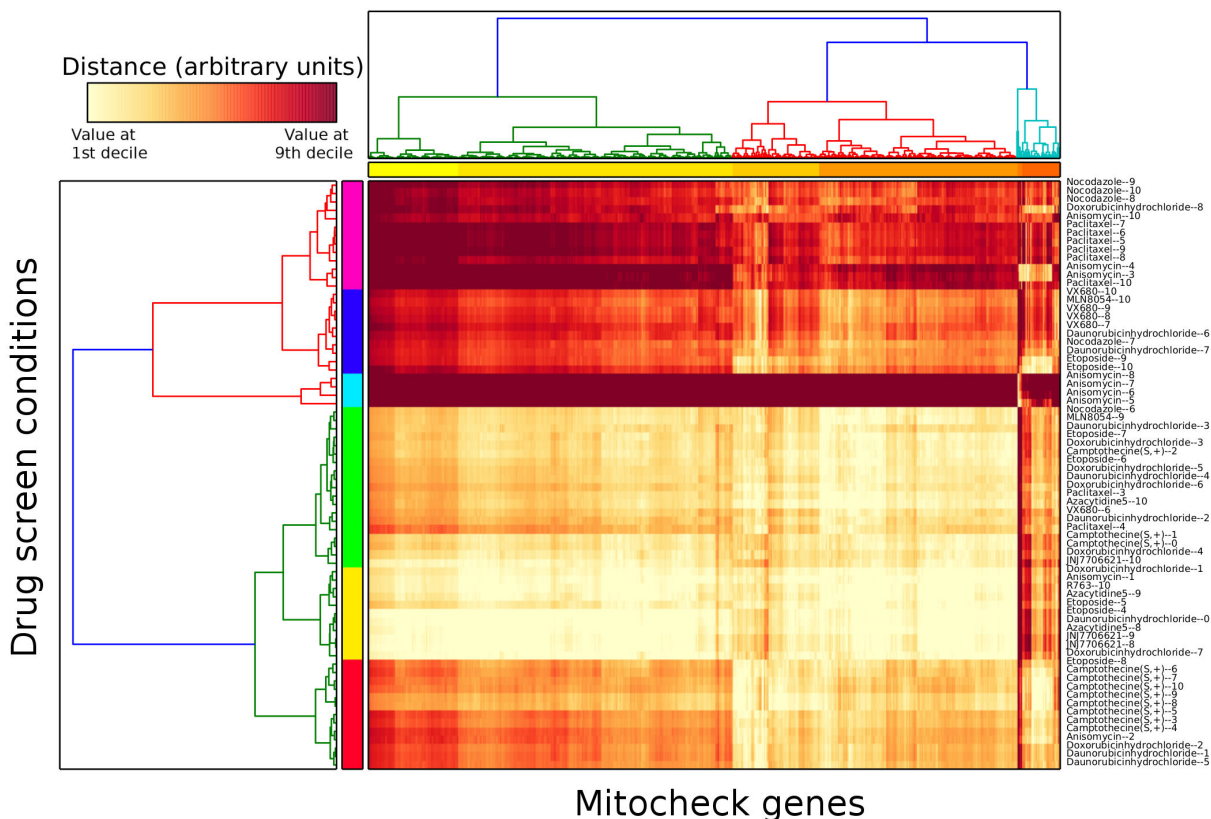


FIGURE 3.19: Drug screen **hit** condition - Mitochondrial siRNA two-dimensional hierarchical clustering using global Sinkhorn divergence. Ward method was used in combination with the Euclidean distance.

presented on fig. 3.20, 3.21 and in appendix section A.2.4. On these binary maps, drugs have been grouped by target similarity as in table 3.4.

The binary map from phenotypic score distance (fig. 3.20) shows little agreement with the literature: conditions belonging to the same clusters seem to be randomly distributed and in particular not to reflect the structure suggested by the literature. On the other hand, the binary map from global Sinkhorn divergence produces a structure which is more in agreement with the current knowledge (fig. 3.21). For example, doses 5 to 10 of Paclitaxel are clustered together with doses 8 to 10 of Nocodazole, which is consistent with their common target β -tubulin. Another interesting example is that of Anisomycin. Its doses 5 to 8 constitute a single cluster according to global Sinkhorn divergence. This is consistent with Anisomycin being the only drug in the drug screen to inhibit protein synthesis. The same effect can be observed for doses 3 to 10 of Camptothecin (S,+): they cluster together in an almost pure cluster, which is consistent with Camptothecin (S,+) being the only drug inhibiting topoisomerase (DNA) I (TOP1).

Furthermore, the hierarchy of these clusters is also understandable in the light of the existing literature. Indeed, the highest doses of drugs which are targeted at Aurora kinases

TABLE 3.4: Known protein targets of hit drugs (bold: present in Mitocheck hit experiments). Drugs are grouped by target similarity. Source: DrugBank [Wishart et al., 2008] unless specified.

Drug	Protein target (HUGO gene symbol)	Other targets
Anisomycin	RPL10L, RPL13A, RPL23, RPL15, RPL19 , RPL23A, RSL24D1, RPL26L1, RPL8, RPL37, RPL3, RPL11, NHP2L1	
Azacytidine 5	DNMT1	DNA and RNA hypomethylation
Camptothecin (S,+)	TOP1	
Daunorubicin hydrochloride	TOP2A, TOP2B	ds-DNA intercalating agent
Doxorubicin hydrochloride	TOP2A	ds-DNA intercalating agent
Etoposide	TOP2A, TOP2B	
JNJ7706621	AURKA , AURKB , CDK1 , CDK2, CDK3, CDK4 , CDK6 [Emanuel et al., 2005]	
MLN8054	AURKA [Huck et al., 2010]	
R763	AURKA , AURKB , FLT3, VEGFR2 [McLaughlin et al., 2010]	
VX680	AURKA , AURKB	
Nocodazole	HPGDS , TUBB (B2A , B4A , B4B , B6)	Microtubule destabilizer
Paclitaxel	BCL2 ,TUBB1, NR1H2, MAPT, MAP4, MAP2	Microtubule stabilizer

(JNJ7706621, MLN8054 and VX680 which are in the dark blue cluster on fig. 3.19) is the closest cluster to the one which contains high Paclitaxel and Nocodazole doses (pink cluster). The latter drugs are targeted at tubulin, when Aurora kinases are linked to microtubules as well (NCBI Gene webpages [Pruitt et al., 2014]). The cluster hierarchy is therefore consistent, since closer clusters point to a common biological process.

The binary map from time Sinkhorn divergence (fig. A.8) shows little difference with that from global Sinkhorn divergence. Finally, the binary map from phenotypic trajectory distance (fig. A.10) shows more agreement with the literature than the phenotypic score distance's (e.g. a high number of conditions of Etoposide, Doxorubicin hydrochloride and Daunorubicin hydrochloride cluster together), but seems less refined than any of the Sinkhorn's maps (e.g. most doses of Camptothecin (S,+) are included in the aforementioned cluster).

Another observation can be made from these clustering results. All three "sophisticated" distances, the phenotypic trajectory distance, the global and time Sinkhorn divergences, present a cluster which can be named a small dose cluster (respectively green, yellow and

red as indicated by the y-absciss colorbar on figures A.9, 3.19 and A.7). It contains mostly the smallest doses for each drug hit: Anisomycin dose 1, Daunorubicin hydrochloride dose 0, Azacytidin 5 doses 8 and 9, etc. Interestingly, these conditions have in common very unspecific phenotypic consequences: slight increases in *Elongated* and/or *Large* and/or *Polylobed* and/or *Binucleated* nuclei, without any increase in *Apoptosis* nuclei. This can be seen in the phenotypic plots which are in the supplementaries, folder "Phenotypic score plots_type2".

Global and time Sinkhorn divergence results present a second unspecific effect cluster (respectively green and yellow-green on figures 3.19 and A.7), which contains mostly small to medium doses for each drug hit: Daunorubicin hydrochloride doses 2, 3, 4, Azacytidin 5 dose 10, etc. Similarly to the small dose cluster, these conditions have very unspecific effects in common. Contrary to the small dose clusters, they present a non-negligeable level of *Apoptosis* nuclei. This cluster is not present in the phenotypic trajectory distance results.

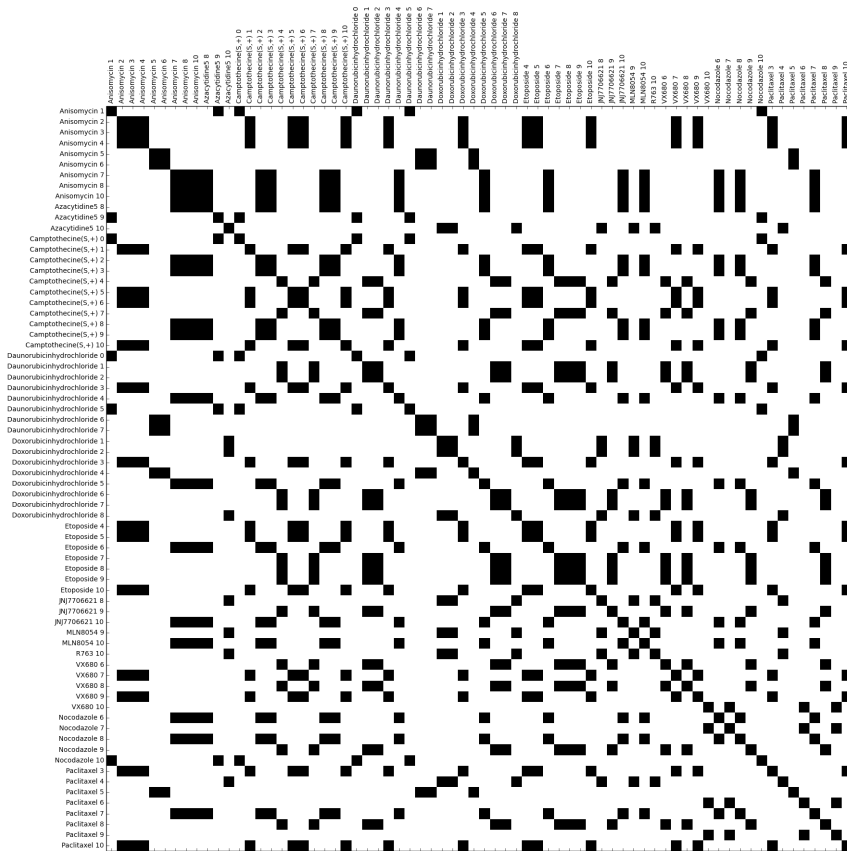


FIGURE 3.20: Visualization of condition clustering for phenotypic score distance. A black dot means that the conditions belong to the same cluster, a white dot that they do not.



FIGURE 3.21: Visualization of condition clustering for global Sinkhorn divergence. A black dot means that the conditions belong to the same cluster, a white dot that they do not.

From these observations we can conclude that both Sinkhorn divergences (and especially global Sinkhorn divergence) produce drug similarities which better correspond to the existing knowledge. Furthermore, these similarities are easily understandable by going back to the phenotypic scores and the cost matrix, as opposed to the phenotypic trajectory distance. As an example, Anisomycin doses 5-8 cluster is very likely a result of the exclusive high increase in *Apoptosis* nuclei which is observed in these experiments, and of *Apoptosis* being at a cost of 5 from all other phenotypes. Finally, these results also show that for almost all drugs, small doses produce unspecific effects which seem difficult to be robustly linked to any gene silencing experiment. Our focus for target pathway inference will therefore be condition clusters which are not small/medium dose clusters, using global Sinkhorn divergence.

3.4.3.2 Target pathway inference

Meaningful condition groups were established in the previous section. They enable us to circumscribe the search for drug target to dose ranges which are most likely to show on-target specific effects. We will consider the following condition groups (corresponding colors in fig. 3.18 in brackets):

1. Nocodazole, doses 8 to 10 (pink, top)
2. Paclitaxel, doses 5 to 9 (pink, bottom)
3. VX680, doses 8 to 10 and MLN8054, dose 10 (dark blue, top)
4. Anisomycin, doses 5 to 8 (light blue)
5. Camptothecine (S,+), doses 6 to 10 (red, top)

For each condition group, Mitocheck hit genes were ordered according to their global distance to group experiments, as described in section 3.4.1.7. The list of ordered genes for each condition group can be found in the Supplementaries, folder "Gene ranks".

TABLE 3.5: Rank of known drug targets, when applicable. Condition group are identical to that in the text.

Condition group	Protein target (HUGO gene symbol)	Rank
1 Nocodazole	HPGDS	577
	TUBB2A	548
	TUBB4A	over 1,000
	TUBB4B	416
	TUBB6	987
2 Paclitaxel	BCL2	389
	TUBB2A	466
	TUBB4A	over 1,000
	TUBB4B	317
	TUBB6	over 1,000
3 VX680	AURKA	39
	AURKB	21
4 Anisomycin	RPL19	over 1,000
	NHP2L1	15
5	na	

A first result from these lists is the index of known drug targets in each, as can be seen in table 3.5. This result does not enable to assess the predictions for any drug whose known target is not in Mitocheck hit list, such as Camptothecin(S,+). It furthermore only relies on two or three Mitocheck experiments in each case. A solution to this drawback is to

consider the closest genes to condition groups, and analyze significantly represented gene functions using Gene Ontology and DAVID online tools [Jiao et al., 2012]. A summary of the results from such an analysis are presented in table 3.6 ; the whole results can be found in the Supplementaries folder "Gene ranks".

From both tables, it appears that results from the different condition groups are unequal. On the one hand, functional analysis of the closest genes to the Anisomycin group does not show any terms which are related to ribosomal proteins, whom Anisomycin is mainly targeted at. Functional analysis identifies as the three most important functional clusters for the Anisomycin group one cluster related to cell cycle and microtubule cytoskeleton, one related to chemical homeostasis, and one related to serine/threonine protein kinase - whose enrichment scores range from 1.6 to 1.4. Similarly, functional analysis of Camptothecin(S,+) group shows relation neither with DNA binding nor with DNA topological change. They rather point at cell-cell junctions, cyclic nucleotide binding proteins and ion channel activity. This does not prevent one of Anisomycin targets, NHP2L1 (SNU13 homolog, small nuclear ribonucleoprotein (U4/U6.U5)), of being its 15th closest gene.

On the other hand, functional analysis of the closest genes to the VX680-MLN8054 group is quite consistent with them being targeted at Aurora kinases. Indeed, these cell-cycle regulated proteins are thought to be involved in the formation of microtubules and their stabilization at the spindle pole, which can be retrieved in the two first functional clusters. Furthermore, AURKA and AURKB (Aurora kinases A and B) are respectively the 39th and 21st closest genes to this condition group.

In between these cases, although the main targets of Nocodazole and Paclitaxel are not extremely close to these condition groups, functional analysis does retrieve their impact on microtubule-based process. In conclusion, our method provides a list of genes and their similarity to drug exposure, for each group of drug conditions it identifies. Functional analysis of the top 10% of these lists reveals that our method performs unequally well for retrieving target gene function in all cases. It works very well on drugs which directly act on mitosis-related proteins (e.g. VX680 or Paclitaxel), whereas it is not as helpful on drugs which are targeted at other biological processes (e.g. Anisomycin).

3.4.4 Discussion

In this section, we were interested in evaluating drug similarity and drug target from time-lapse HC-HT drug screening data. To this end, we introduced a new divergence, the Sinkhorn divergence, from which we derived three different flavours, and which we compared to a simple Euclidean distance on phenotypic scores, and a state-of-the-art measure, the phenotypic trajectory distance.

TABLE 3.6: Rank of known drug targets, when applicable. Condition group are identical to that in the text.

Condition group	Cluster rank	Terms from relevant functional annotation clusters	Enrichment score
1	5/66	Microtubule cytoskeleton organization Microtubule-based process Cytoskeleton organization	0.97
2	1/51	Microtubule cytoskeleton Cell division Cell cycle Centromere Chromosome segregation Spindle	2.24
	2/51	Microtubule cytoskeleton Microtubule cytoskeleton organization Microtubule organizing center organization Centrosome cycle Microtubule-based process	1.89
3	1/66	Regulation of cell cycle Mitotic cell cycle checkpoint Kinetochore	2.49
	2/66	Regulation of cell cycle process Cell cycle Microtubule cytoskeleton Spindle M phase Centrosome	2.19
	3/66	Chromosome, centromeric region Metaphase	2.13
4	-/62		
5	-/63		

Distance quality evaluation on our drug screening dataset showed that all distances were affected by a medium experimental reproducibility to the same extent. Interestingly enough, this is also the case for global Sinkhorn divergence, which should be more robust as it pools a temporal sequence of phenotype distributions into a single set of phenotype distributions. In order to reduce the impact of batch effect, one could think of normalizing all experiment phenotype distributions from one batch by the median of control

distributions from this batch. This should compensate batch effects which modify basal cell state. However, it could also be the case that batch effects impact how cells respond to drug exposure. If response timing is different, one could think of applying dynamic time warping when evaluating time Sinkhorn divergence, and possibly phenotypic trajectory distance. However, if the phenotypic response itself varies from batch to batch, this might be an interesting information about the drug: cell response might be stochastic in the same measure as is gene expression [Elowitz et al., 2002].

Drug similarity evaluation showed that our information representation is only suitable for drugs which show an impact on nuclear phenotypes as measured by chosen classes in less than 48 hours. Hence, drugs which have an impact on nuclear texture such as the number of nucleoli, or on other biological processes with no nuclear consequences within 48 hours, will not be detected. In the first case, new phenotypic classes could be added to the existing set, or hit experiment detection should be done directly from object feature distributions (e.g. nuclear perimeter, nuclear excentricity, etc.) rather than phenotypic distributions. [Young et al., 2008] gives an example of such a strategy, which is nevertheless challenging in high dimensional feature spaces. In the second case, one would need more fluorescent markers in order to quantify other phenotypic aspects (markers for the Golgi apparatus, plasma membrane, cytoskeleton, etc.). However, as the number of markers is necessarily limited, effects of all drugs can never be identified in the same assay. One strategy might therefore be to perform drug screens with respect to one specific biological process, and to chose cell markers such that they are informative about all aspects of this particular process.

Drug similarity evaluation on hit conditions performed well, with Sinkhorn divergences performing better than all other distances. This was measured by the grouping of drugs with similar targets and mode-of-actions, as could be found in the literature. Furthermore, not only were similar drugs grouped together, but the hierarchical clustering performed using global Sinkhorn divergences also proved to be biologically meaningful. Indeed, it grouped microtubule-related drugs JNJ7706621, MLN8054 and VX680, and Paclitaxel and Nocodazole in the same cluster, which was further refined in two different clusters as they are respectively targeted at Aurora kinases and tubulin. Finally, hierarchical clustering also identified conditions which produced non-specific effects (with and without apoptosis), enabling to restrict drug target inference to drug doses which showed specific effects. This restriction has already been searched for [Loo et al., 2007], as results from drug target inference on small doses will probably not be robust and/or meaningful.

Finally, we used global Sinkhorn divergence between Mitocheck hit experiment and meaningful condition groups to perform drug target inference. Although it should theoretically

be possible to identify drug targets from drug and genetic screen comparison, it is generally only providing a list of putative targets, which remain to be tested [Schenone et al., 2013]. This is due to the fact that drugs generally do not have a single target, and that many genes share the same phenotype upon knockdown. It is also explained by the fact that drug exposure and siRNA exposure do not strictly have the same impact on the targeted protein. Indeed, siRNA phenotypic onset can be rather different than that of a drug, because drugs directly act on their target, whereas siRNAs cause the destruction of target mRNAs. Furthermore, phenotype penetrance in drug screen experiments is always higher than that in siRNA experiments.

This is precisely what was experienced in this analysis, with known drug targets never being *the* closest gene to drug experiments. Furthermore, one can observe that due to the fact that we used a marker which was optimized to observe chromosome segregation, drug target inference performed well on mitotic hits *per se*, such as VX680. For other drugs such as Anisomycin, our method identifies closest genes based on their impact on the nucleus when silenced, which could explain that they were not related to translation. In these cases, the workflow will indicate the downstream impact of the drug on nuclear morphologies. But even in cases where the combination of markers and classes are not optimal, our method can still identify hits and group them together. In some cases however, no effect at all was observed. They demand additional phenotypic classes and/or other markers.

Chapter 4

Xenobiotic screen

Résumé - Les cribles xénobiotiques (see *infra* for English text)

Environ 100,000 nouvelles molécules sont synthétisées chaque année. D'autre part, les réglementations européennes sont de plus en plus strictes en ce qui concerne l'expérimentation animale. L'utilisation de cribles biologiques à haut débit et haut contenu semble donc indiquée en toxicologie environnementale : ils constitueraient une procédure expérimentale ayant l'avantage d'être à la fois *in vitro* et très informative. Toutefois, la plupart des tests actuellement utilisés en toxicologie environnementale sont à faible contenu, ou analysés manuellement - ce qui empêche leur utilisation à haut débit.

Afin d'étudier la faisabilité d'une telle approche, nous avons réalisé un crible de 5 xénobiotiques connus tels que la dioxine (TCDD), produisant un jeu de données de vidéomicroscopie à épifluorescence. Ces données ont d'une part été analysées à l'aide de MotIW [Schoenauer Sebag et al., 2015] pour la motilité cellulaire individuelle (cf. section 4.2.2), d'autre part à l'aide de la procédure développée par [Walter et al., 2010] pour le cycle et la division cellulaires (cf. section 4.2.3). Une interface Web a également été conçue pour le partage des résultats entre les laboratoires, qui est présentée dans la section 4.1.2.1.

Ces expériences n'ont toutefois pas permis de conclure à l'utilité de l'approche pour l'étude de l'impact sub-toxique des xénobiotiques choisis. Plusieurs pistes quant à l'origine de ces résultats sont discutées dans la section 4.3. Il serait en premier lieu intéressant de réaliser une batterie d'expériences primaires à faible contenu avant le crible lui-même, afin de choisir un ensemble de doses resserré auxquelles un effet est attendu de manière certaine. Dans un second lieu, le choix de la lignée cellulaire pourrait être revu afin de choisir une lignée la plus homogène possible.

Environmental health consists in studying the impact of Man's environment on his health. This can be done following either one of two major paradigms: *Epidemiology* and *Toxicology*. The former deals with human populations, looking for significant link between past exposure and present pathologies. Toxicology can itself be *in vivo*, *in vitro* or *in silico*, depending whether one chooses to study the impact of a precise exposure on a population of living organisms, a population of cells, or based on chemical descriptors of the exposure.

Causal links can be quite delicate to establish in Epidemiology. There is a great number of variables which are involved (e.g. genetic or behavioral) ; the effects can have a weak penetration and are often delayed by a certain number of years with respect to the exposure. To mitigate these effects, Toxicology enables to select the chemical (or xenobiotic¹) of interest, as well as precisely control the experimental settings.

However, Toxicology is currently facing two major challenges:

- Around 100 000 new compounds are synthesized each year. There is therefore the need for high-throughput (HT) and safe toxicity tests.
- European regulations are stricter and stricter with animal testing, hence the need for novel *in vitro* toxicity tests.

This led us to think of importing the technique of drug screening from pharmacological Toxicology, replacing prospective drugs with xenobiotics. The basic idea of screening is to perform a given assay on hundreds of compounds in parallel.

Classical *in vitro* toxicological screening procedures have been in use for some time, such as the Comet assay for DNA damage, which is known since 1984 [Ostling and Johanson, 1984]. This type of test is now routinely done in a HT setup. However, it provides very crude information compared to what high content (HC) assays would indicate regarding for example, cell cycle modulation following BPA exposure. Subtler tests are being developed in a HT setup, either endpoint [Freitas et al., 2014], [Vecchio et al., 2014] or real-time assays [Wlodkowic et al., 2011], [Timm et al., 2013]. This was recently made possible by technical and computational progresses.

Nevertheless, the majority of new *in vitro* toxicological methods are not HC², i.e. they do not allow detailed observation of phenotypes at the single cell level. Such methods

¹A xenobiotic is, with regard to a species, any compound which was not produced by an individual of this species.

²3 papers dealing with new toxicological assays out of 10 present HC methods (PubMed searches: "environmental AND cellular AND (screening[Other Term] OR High throughput/high content assays[Other Term])", and "time-lapse AND toxicology" on the 3/19/2015).

are common practice in molecular biology and have been scaled up, so as to be presently applicable in a HC setup. More specifically, there are many biological processes which are best studied with time-lapse microscopy. Surprisingly, this type of experiment is still rarely used in Environmental Toxicology - regardless of the throughput. When it is, data is manually analyzed most of the time (e.g. [Fiorini et al., 2008], [Gatti et al., 2004], or [Costa, 1983]). However, time-lapse experiments would be a significant improvement over endpoint assays, as they enable for example to assess event sequences following exposure rather than record cell death.

Environmental Toxicology time-lapse data was therefore newly generated, in order to assess whether HC time-lapse screening is applicable in this context : is this approach relevant for toxicity detection and characterization of environmentally relevant compounds? Time-lapse HCS usability in this context would lead to the potential development of a time-lapse HC-HT assay for Environmental Toxicology.

Hence five well-known xenobiotics were selected and screened for their effects on nuclear motility and nuclear morphology. The whole pipeline is illustrated fig. 4.1 (and the experimental settings are detailed fig. 4.2). Briefly, cells were chemically exposed prior to image acquisition over time. Cells were segmented on each image of each experiment using the open-source software CellCognition [Held et al., 2010]. Object features were extracted using the same software, for two purposes: nuclear tracking and nuclear morphology classification.

Nuclear tracking was performed as described in the methodological article [Schoenauer Sebag et al., 2015]. This step was followed by trajectory feature extraction, and statistical hit detection, as described in the same article.

CellCognition was used to establish a training set of annotated nuclear morphologies, to train a classifier and apply this classifier to each nucleus in the data set. This allowed us to describe each experiment by a set of class percentage time series, whose comparison with control time series allow us to detect significant differences.

As this was a proof of concept experiment, the goal was slightly different than that when applying MotIW on the Mitocheck dataset. In the latter case the goal was to select relevant genes for nuclear motility with high confidence, which explains why an ad hoc statistical procedure had to be developed to make sure that p-values were not overestimated. In the case at hand, the goal was rather to select experimental conditions with mild-to-high confidence for performing confirmatory experiments. Our goal was therefore to obtain a ranking of all conditions with respect to the tested endpoint (motility or cell cycle) rather than computing absolute p-values.

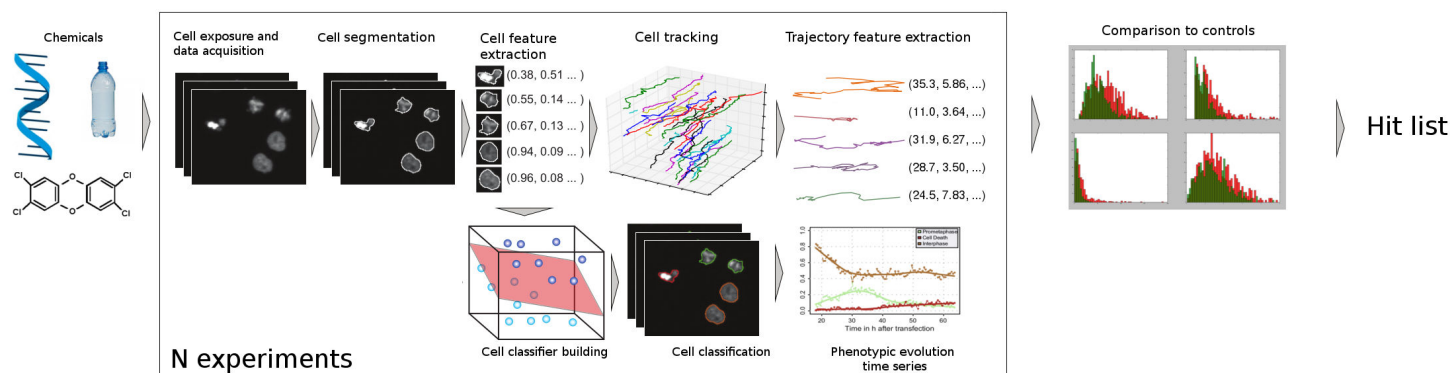


FIGURE 4.1: Xenobiotic screen complete workflow

4.1 Materials and methods

4.1.1 Experimental work

4.1.1.1 Chemicals

TCDD was bought from LGC Standards (Molsheim, France), TGF β 1 from R&D SystemsTM (Minneapolis, MN, USA), BPA, MeHg and PCB153 from Sigma Aldrich[®] (Saint-Louis, MO, USA). α -Endosulfan was bought from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA, USA) and DMSO from Merck Millipore (Billerica, MA, USA).

The following media were used:

- normal medium: DMEM (Gibco[®], Life TechnologiesTM, Puteaux, France) with phenol red and supplemented with 10% fetal calf serum (FCS), 200 units/ml penicillin, 500 μ g/ml streptomycin, 3g/ml glutamin, 10 μ g/mL insulin and 0.1nmol/mL of non-essential amino acids solution (all from Life TechnologiesTM), and 0.5 μ g/ml fungizone (Squibb, Princeton, NJ, USA),
- imaging medium: CO₂-independent medium (Gibco[®]) supplemented with 10% FCS, 200 units/ml penicillin, 500 μ g/ml streptomycin, 3g/ml glutamin, 10 μ g/mL insulin and 0.1nmol/mL of non-essential amino acids solution, and 0.5 μ g/ml fungizone.

4.1.1.2 Cell culture

The human mammary tumor cell line MCF-7 (ATCC[®] Catalog N^oHTB-22TM) was maintained in normal medium as defined above.

4.1.1.3 Cell transfection and clonal selection

MCF-7 cells were transfected with two plasmids : one containing human histone H2B fused to the gene encoding a red fluorescent protein (mCherry), isolated from *Discosoma* species (Addgene, plasmid #21045), one containing human membrane lipid Myr/Palm fused to the gene encoding the green fluorescent protein (GFP) of *Aequorea victoria* (Addgene, plasmid #21037). The aim was to generate a stable line constitutively expressing H2B-mCherry and Myr/Palm-GFP.

On the day before transfection, MCF-7 cells were seeded into 10cm dish with normal medium (2 millions per dish). On the day of transfection, cell medium was replaced with 2mL of normal medium and a mix composed of 72 μ L Lipofectamin[®] 2000 reagent and 24 μ g of total plasmid DNA (either H2B-mCherry alone, Myr/Palm-GFP alone or both), completed to a volume of 3mL with Optimem (Life technologiesTM). Cells were then incubated for 5 hours at 37°C, after what 5mL of normal medium was added. Antibiotic selection started 7 days after transfection, with 1mg/mL of neomycin and 1 μ g/mL of puromycin added to normal cell medium ; selection medium was replaced every other day.

Clonal selection was realized by infinite dilution: three weeks following the beginning of antibiotic selection, transfected cells were seeded in two 96-well plates with 64 wells at 0.3 cell/well, 64 wells at 1 cells/well and 64 wells at at 3 cells/well. Once clones had sufficiently grown, a few clones were selected based on their level of plasmid expression. They were tested for the expression of the following genes with and without exposure to 25nM TCDD for 48 hours: Aryl hydrocarbon receptor (AHR), Cytochrome P450 1A1 (CYP1A1) and E-cadherin genes. Most clones responded as expected (increased expression of AHR and CYP1A1 and decreased expression of E-cadherin - data not shown). One clone was discarded.

Following this step, modified cells were maintained in normal medium with 1mg/mL of neomycin and 1 μ g/mL of puromycin. Hereafter, "MCF-7 cells" refers to a selected clone of such modified cells.

4.1.1.4 Production of 96 well-plate for imaging

48h prior to imaging, MCF-7 cells were seeded in normal medium into one 96 well-plastic plate, at a density of 6,000 cells per well. The plate was placed back in the incubator at constant temperature (37°C) and CO₂ pressure (5%). 24 hours prior to imaging, dilutions (cf table 4.2) of selected compounds with constant solvent percentage were freshly prepared in imaging medium. Cell medium was changed with 198 μ L of imaging

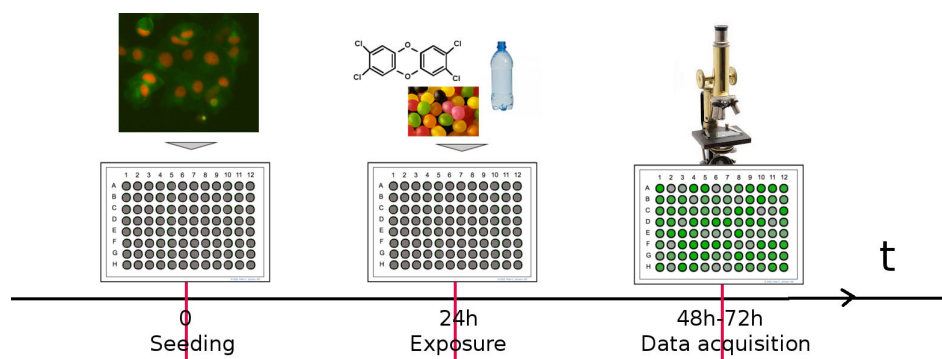


FIGURE 4.2: Illustration of the experimental settings

271214 plate

A	Cl1 PCB 10 Indtp 10	Cl1 PCB 9 Indtp 10	Cl1 DMSO 0 Indtp 10	Cl1 PCB 8 Indtp 10	Cl1 PCB 7 Indtp 10	Cl1 PCB 6 Indtp 10	Cl1 Nonane 0 Indtp 10	Cl1 PCB 5 Indtp 10	Cl1 PCB 4 Indtp 10	Cl1 PCB 3 Indtp 10	Cl1 PCB 2 Indtp 10	Cl1 PCB 1 Indtp 10
B	Cl1 Endo 10 Indtp 10	Cl1 Endo 9 Indtp 10	Cl1 Nonane 0 Indtp 10	Cl1 Endo 8 Indtp 10	Cl1 Endo 7 Indtp 10	Cl1 Endo 6 Indtp 10	Cl1 DMSO 0 Indtp 10	Cl1 Endo 5 Indtp 10	Cl1 Endo 4 Indtp 10	Cl1 Endo 3 Indtp 10	Cl1 Endo 2 Indtp 10	Cl1 Endo 1 Indtp 10
C	Cl1 TCDD 9 Indtp 10	Cl1 TCDD 8 Indtp 10	Cl1 DMSO 0 Indtp 10	Cl1 TCDD 7 Indtp 10	Cl1 TCDD 6 Indtp 10	Cl1 TCDD 5 Indtp 10	Cl1 Nonane 0 Indtp 10	Cl1 TCDD 4 Indtp 10	Cl1 TCDD 3 Indtp 10	Cl1 TCDD 2 Indtp 10	Cl1 TCDD 1 Indtp 10	Cl1 TGF 15 Indtp 10
D	Cl1 BPA 9 Indtp 10	Cl1 BPA 8 Indtp 10	Cl1 Nonane 0 Indtp 10	Cl1 BPA 7 Indtp 10	Cl1 BPA 6 Indtp 10	Cl1 BPA 5 Indtp 10	Cl1 DMSO 0 Indtp 10	Cl1 BPA 4 Indtp 10	Cl1 BPA 3 Indtp 10	Cl1 BPA 2 Indtp 10	Cl1 BPA 1 Indtp 10	Cl1 TCDD 1 Indtp 10
E	Cl1 MeHg 9 Indtp 10	Cl1 MeHg 8 Indtp 10	Cl1 DMSO 0 Indtp 10	Cl1 MeHg 7 Indtp 10	Cl1 MeHg 6 Indtp 10	Cl1 MeHg 5 Indtp 10	Cl1 Nonane 0 Indtp 10	Cl1 MeHg 4 Indtp 10	Cl1 MeHg 3 Indtp 10	Cl1 MeHg 2 Indtp 10	Cl1 MeHg 1 Indtp 10	Cl1 TGF 15 Indtp 10
F	Cl1 BPA 9 Indtp 10	Cl1 BPA 8 Indtp 10	Cl1 Nonane 0 Indtp 10	Cl1 BPA 7 Indtp 10	Cl1 BPA 6 Indtp 10	Cl1 BPA 5 Indtp 10	Cl1 DMSO 0 Indtp 10	Cl1 BPA 4 Indtp 10	Cl1 BPA 3 Indtp 10	Cl1 BPA 2 Indtp 10	Cl1 BPA 1 Indtp 10	Cl1 TCDD 1 Indtp 10
G	Cl1 PCB 10 Indtp 10	Cl1 PCB 9 Indtp 10	Cl1 TGF 15 Indtp 10	Cl1 PCB 8 Indtp 10	Cl1 PCB 7 Indtp 10	Cl1 PCB 6 Indtp 10	Cl1 Rien 0 Indtp 10	Cl1 PCB 5 Indtp 10	Cl1 PCB 4 Indtp 10	Cl1 PCB 3 Indtp 10	Cl1 PCB 2 Indtp 10	Cl1 PCB 1 Indtp 10
H	Cl1 Endo 10 Indtp 10	Cl1 Endo 9 Indtp 10	Cl1 Rien 0 Indtp 10	Cl1 Endo 8 Indtp 10	Cl1 Endo 7 Indtp 10	Cl1 Endo 6 Indtp 10	Cl1 Rien 0 Indtp 10	Cl1 Endo 5 Indtp 10	Cl1 Endo 4 Indtp 10	Cl1 Endo 3 Indtp 10	Cl1 Endo 2 Indtp 10	Cl1 Endo 1 Indtp 10
	1	2	3	4	5	6	7	8	9	10	11	12

FIGURE 4.3: Example of a plate setup. *Cl1*: clone number, *Indtp 10*: CO_2 -independent cell medium with 10% FCS.

medium and $2\mu L$ of chemical dilution or solvent dilution per well. Six control wells for each solvent and three wells with cell medium only were put on each plate. Plate design was not random, but control wells were not grouped (see fig 4.3 for an example of a plate setup).

The plate was placed back in the incubator. Immediately before image acquisition, the plate was sealed using an adhesive optical film. This procedure is illustrated on figure 4.2.

4.1.1.5 Time-lapse imaging

Images were acquired every 15 minutes in each well for forty-eight hours, with an automated epifluorescence microscope (Axio Observer Z1; Zeiss, Oberkochen, Germany)

with motorized objectives in z-axis (resolution 10nm) and using 10x objective (EC Plan-Neofluar;0.3 M27). Each well was imaged 800ms at lengthwave $\lambda = 555nm$ (mCherry) and 300 ms at $\lambda = 470nm$ (GFP).

The microscope is integrated into a microscope incubation chamber to provide constant temperature (+37°C), which was turned on at least one hour prior to the beginning of image acquisition. Zeiss software ZEN2011 was used for data recording. Focus was done manually, and was updated by definite focus.

4.1.1.6 Phototoxicity assays

One plate was prepared with 12 wells following the above-described procedure, except that no chemical was added prior to image acquisition. Images were acquired for 24h, with the four different imaging conditions :

- no imaging
- 500ms at $\lambda = 555nm$, 300 ms at $\lambda = 470nm$
- 1 000ms at $\lambda = 555nm$, 300 ms at $\lambda = 470nm$
- 1 500ms at $\lambda = 555nm$, 300 ms at $\lambda = 470nm$

2 μ L of *alarmarBlueTM Cell Viability Assay Reagent* (Thermo Scientific, Rockford, IL, USA) were then added to each well. Following 2h of incubation at 37°C, absorbance was measured at $\lambda = 570nm$ and 600nm using an EnSpire® (PerkinElmer, Waltham, MS, USA). The index of cell viability was computed following the manufacturer's formula.

4.1.1.7 Chemical dose choice

Doses were chosen so that the lower end is close to human exposure levels, and that the higher end is not cytotoxic.

The first goal was attained through a literature review, whose results can be seen in appendix, section A.3. The second goal was attained through cytotoxicity tests. One plate was prepared with 76 wells as described above, although with 5,000 cells per well. 24h after seeding, cells were exposed either to solvents, either to doses 6 to 10 of selected compounds. 24h after exposure, 2 μ L of *alarmarBlueTM Cell Viability Assay Reagent* were added to each well. After 24h of incubation, fluorescence was measured in each well at $\lambda = 555nm$, and cell viability index was computed following the manufacturer's formula. In the end, 9 to 10 doses for each of the five chosen xenobiotics were selected, the total summing to approximately 50 different conditions to which cells were exposed.

4.1.2 Bioinformatics methods

4.1.2.1 Web-based user interface for result visualization

A web-based user interface was designed for raw and quality control data visualization. It is accessible at the following address: <http://olympia.biomedicale.univ-paris5.fr/plates/> (login: *user*, password: *xbscreen*). It was implemented using Django³ web framework, the programming language Python 2.7⁴, and runs under Linux-Apache web server and mod_wsgi module. SQLite⁵ was used for storing experimental metadata (plate setups, experimental conditions such as cell medium or the percentage of FCS). The database was built as visible on fig 4.4.

Briefly, each well is linked to a unique plate, a unique condition (medium and percentage of FCS) and a unique treatment (xenobiotic and dose). As the data was (crudely) password-protected, a second database contains logins and passwords, as well as admin permissions.

After logging in, any user has access to the list of plates in anti-chronological order. A plate page displays the plate setup, as well as some overall features of all experiments, such as initial number of objects or the proliferation rate in each well. This enables to see a significant geographical bias, which could be due to chemical exposure or the experimental settings. In the latter case, the plate should not be used for analysis and the experimental protocol modified ; such visualization tools were useful for designing the protocol.

By clicking on a particular well on any image of the plate page, one accesses per-well information: well experimental conditions, well raw movies, and time evolutions of mean intensity, percentage of out-of-focus objects, number of cells, number of objects, and percentage of objects in all classes (as detailed below).

4.1.2.2 Quality control

Due to microscope hardware and/or software instability, images showed mean intensity variation on both channels. Hence, images t whose mean intensity on the mCherry channel I_t verified the following inequality were not considered for further analysis, where σ_I is the standard deviation of $(I_t)_{t=1\dots T}$:

$$\left| I_t - \frac{\sum_{i=0\dots 10} I_{t+i}}{10} \right| > 3\sigma_I$$

³<https://www.djangoproject.com/>

⁴<http://www.python.org>

⁵<https://sqlite.org/>

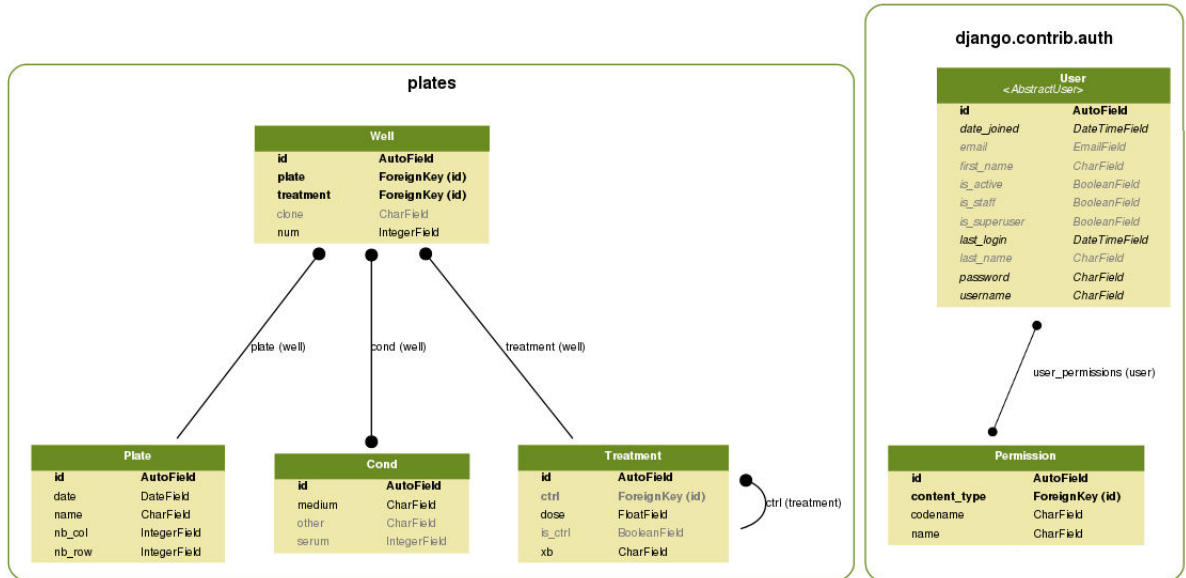


FIGURE 4.4: Diagram of the databases for experimental metadata storage

Furthermore, a threshold of c cells at the beginning of the movie, and maximum $p\%$ out-of-focus objects were used to further remove unexploitable movies. We fixed the threshold after visual inspection to $c = 23$ and $p = 37$. Out-of-focus objects and cells that were neither artefacts nor out-of-focus objects were identified following segmentation, feature extraction and classification as described below.

4.1.2.3 Object segmentation

Object segmentation was done using a newly-designed plugin implemented in the open-source software CellCognition [Held et al., 2010]. MCF-7 cells are smaller than HeLa cells and tended to form clusters in our experiments. While the method described in 2.2.1 and previously published in several papers, e.g. in [Held et al., 2010], is in principle capable of splitting clustered nuclei, we felt that the filtering of the distance transform, which ultimately influences the decision on whether to split or not, was suboptimal. Here, we used morphological dynamics to improve the splitting step of this segmentation method.

The first steps of the method are therefore identical to what we have presented in 2.2.1: images are prefiltered (here by a median filter) and we assign to each pixel the difference of its value to the average in a window centered in the pixel. This average value can be efficiently calculated with integral images. By applying a global threshold to the residue image, we obtain a first segmentation result which gives accurate results for isolated nuclei, but which tends to segment close nuclei together as a single object.

The last step is to calculate the Watershed transformation on the inverse distance map. This is a standard technique to divide close convex objects after segmentation [Lantuejoul, 1982, Chapter Geodesic segmentation]. This method splits the binary segmentation result in as many objects as there are local minima in the inverse distance map. As small irregularities in the contours can lead to such minima, it is often necessary to apply a filter on the distance map in order to avoid oversegmentation. One option is to use a simple Gaussian filter, as proposed in [Wählby et al., 2002] and [Held et al., 2010]. However, this does not permit an intuitive control over which minima are really kept, and even worse: it does not guarantee that larger filters always suppress more minima. Here, we propose to use morphological dynamics for this purpose [Soille, 2003]. While this technique is widely used in the morphology community, it is to our knowledge not used for the splitting of cells, even though it perfectly applies to this problem.

Morphological dynamics assign to each local minimum the value at which it fuses to a region coming from a minimum with lower value (the dynamic of the lowest minimum is set to ∞). As the watershed algorithm iterates through the values in an ordered way, this value is identical to the minimum height that has to be passed to reach a lower minimum. Let $p_{i,j}$ be a path that joins minima i and j , the dynamic of minimum i is the following:

$$dyn(i) = \min_{\substack{p_{i,j} \\ f(j) < f(i)}} \max_{x \in p_{i,j}} (f(x) - f(i))$$

Hence, to avoid the usual over-segmentation produced by the watershed algorithm, we only use the subset of minima with dynamic larger than a certain threshold. The number of objects decreases as this threshold increases, and the control is intuitive: small concavities produce local minima with relatively low dynamic (independently from their spatial arrangement or the size of the objects).

The computational cost of the morphological dynamics is the same as the watershed algorithm. Depending on the size of the data sets, this is not negligible, even though it compares favorably to smoothing filters, as they were used in this context. To further reduce the computational complexity, we implemented the dynamic filter in such a way that it is directly combined with the watershed algorithm. Indeed, the criterion can be checked each time a pixel is about to be assigned to the watershed line. Consequently no additional flooding step is necessary.

Finally, an object filter is applied to eliminate the objects that are (or whose mean intensity is) too small.

4.1.2.4 Object feature extraction

Object feature extraction was done using the open-source software CellCognition [Held et al., 2010], as previously described [Walter et al., 2010]. Briefly, for each object on each image, approximately 240 features are extracted, characterizing their shape and texture. These features enable to classify nuclei in user-defined nuclear phenotypic classes and to track them over time.

4.1.2.5 Object classification

Object classification was also performed using CellCognition. The class definitions are illustrated in table 4.1. The set of morphological classes is supposed to cover the morphological variability of the screen. The set therefore contains wildtype morphological classes, such as the morphologies corresponding to the different mitotic phases and aberrant morphologies indicating the presence of a phenotype. As our screen consisted in 50 conditions, we almost exhaustively inspected the dataset and are therefore confident that no aberrant nuclear phenotype was missed.

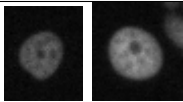
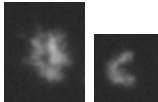
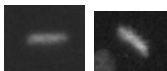
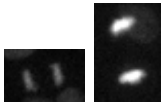
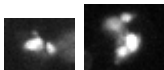
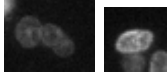
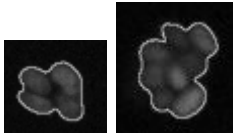
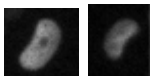
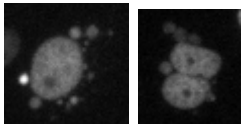
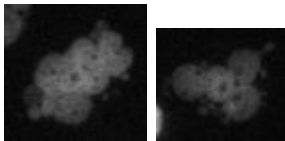

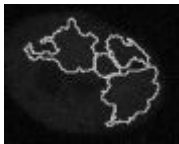
In detail, *Clusters* are clustered nuclei which the segmentation algorithm failed to split. *Folded* nuclei represent elongated or round nuclei with two shades of grey. *Frozen* nuclei are nuclei whose DNA shows a heterogeneous condensed pattern, persistent over time. Frozen nuclei either remain in the same class or lead to *apoptosis*. Biologically, they might correspond to dying nuclei and resemble nuclei experiencing phototoxicity (personal communication, Beate Neumann, EMBL Heidelberg, Germany). As these nuclei were observed following exposure to certain conditions only, we believe that this class does not translate a simple technological artifact. It would rather be the consequence of cell sensitization to phototoxicity which certain exposures could have produced.

A training set was annotated, containing 2,576 nuclei. Support Vector Machines (SVMs) were used for classification, as they work well for nucleus phenotypic classification ([Kovalev et al., 2006] for a comparison of classification algorithms in this context, and for application e.g. [Held et al., 2010], [Walter et al., 2010]). An RBF (Radial Base Function) kernel SVM was trained, whose parameters were obtained by grid search, using Cell Cognition's interface ($\gamma = 2^{-7}$, $C = 8$).

This step provides us with a representation of each video as a set of time-series, which are the evolution of the percentage of nuclei in each phenotypic class over time. The distance of an experiment i to its reference set j for class obj is the following:

$$d_{i,j}^{obj} = \int_0^T (\%obj_{i,t} - \%obj_{j,t}) dt$$

TABLE 4.1: Nuclear morphology classes with examples. The *artefact* and *cluster* examples are shown with segmentation contours.

Normal classes	
<i>Interphase</i>	
<i>Pro-metaphase</i>	
<i>Metaphase</i>	
<i>Anaphase</i>	
<i>Apoptosis</i>	
Aberrant morphology classes	
<i>Frozen</i>	
<i>Cluster</i>	
<i>Folded</i>	
<i>Micronucleated</i>	
<i>Polylobed</i>	
Technical problem classes	
<i>Out-of-focus</i>	
<i>Artefacts</i>	

Due to the visual similarities between *Micronucleated* and *Polylobed* nuclei in our dataset, these classes were pooled together for computing distances. The corresponding distance is named *Micronucleated* in the following.

4.1.2.6 Object tracking and trajectory feature extraction

Those steps were performed as described in the methodological article [Schoenauer Sebag et al., 2015].

TABLE 4.2: Chemical dilutions

Chemical	Solvent	Final solvent percentage (vol)	Doses (nM)
BPA	DMSO	$1.0 \cdot 10^{-1}$	0.1, 1, 10, 50, 100, 1 000, 5 000, 10 000, 50 000
Endo	DMSO	$2.0 \cdot 10^{-1}$	1, 10, 50, 100, 500, 1 000, 5 000, 10 000, 50 000, 100 000
MeHg	DMSO	$1.0 \cdot 10^{-3}$	0.01, 0.1, 1, 5, 10, 50, 100, 500, 1 000
PCB153	DMSO	$3.6 \cdot 10^{-1}$	0.1, 1, 10, 50, 100, 1 000, 5 000, 10 000, 50 000, 100 000
TCDD	Nonane	$3.2 \cdot 10^{-2}$	0.001, 0.01, 0.025, 0.1, 0.25, 1, 10, 25, 50

4.2 Results

4.2.1 Preliminary choices

After testing a few clones for the expected behaviour in response to TCDD exposure, one clone was selected and five plates were imaged following the described procedure, producing 415 videos. In the following, each plate is named after the day image acquisition was launched. The quality control eliminated 22% of the experiments, leaving 324 experiments for analysis with three biological replicates per condition minimum.

Due to the observed high variability in control trajectory statistics on the different plates (cf fig. 4.5), the reference was chosen to be the whole plate as opposed to control wells only. This is a valid choice as long as (1) there is no bias in the distribution of chemical conditions (and possibly results) on distinct plates, and (2) most wells do not show any effect [Birmingham et al., 2009].

4.2.2 Motility study

MotIW was applied to this dataset, which as described in [Schoenauer Sebag et al., 2015] enabled to go from a set of nuclear trajectories to a single statistic for each experiment. However, this statistic was not evenly distributed with respect to the plates. This was

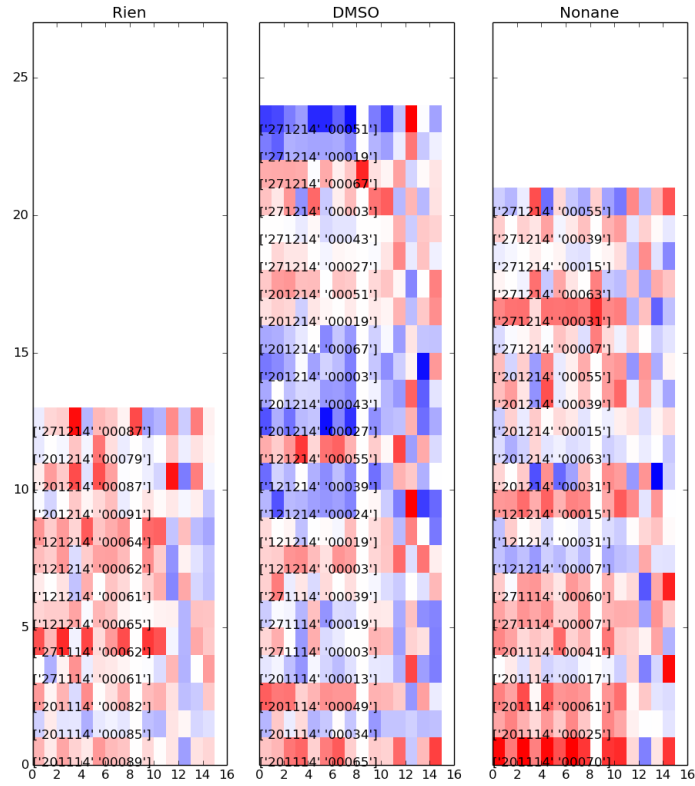


FIGURE 4.5: Trajectory feature heatmaps corresponding to control wells. Each line corresponds to an experiment, whose plate and name are indicated. Each column corresponds to a trajectory feature. A robust normalization (with median and inter-quartile range) was applied, using all plate, which still permits to see that control responses vary from well to well and plate to plate.

measured by a Mann-Whitney U test comparing the list of statistics for one plate with the list of statistics for all remaining plates. Indeed, plates 271214 and 271114 respectively output p-values of 0.06 and 0.02 at this test.

In the Mitochondria study, all experiments from all plates were ranked together ; a p-value threshold was set ; conditions which were more than 50% of the time under that threshold were considered as significantly modifying nuclear motility.

In the case at hand, statistics from all plates cannot be ranked together because of an important batch effect. Hence, the approach which was chosen is that of the *RankProduct* [Breitling et al., 2004]. Briefly, this approach consists in formalizing the following idea: we are interested in conditions which have consistently high statistics on the different plates. Our goal is therefore to compute their rank in the statistic list of each plate, and its variations depending on the plate. The *Rank Product* statistic of a condition c is the following, where $pl. i$ is the i^{th} plate and $card(i)$ the number of experiments

performed on the same plate⁶:

$$RP(c) = \prod_{pl. i} \frac{rg(c, i)}{card(i)}$$

Intuitively, the *Rank Product* of a condition which significantly alters nuclear motility is going to be small. Empirical p-values for the *Rank Product* statistics are computed by permutation, that is, each plate trajectory statistics are permuted N times and the *Rank Product* statistics for each condition computed each time. This produces the empirical null distribution of the *Rank Product* statistics, that is, the distribution of statistics under the hypothesis H_0 that no condition alters single nucleus motility more than the average response of all conditions (which was observed to be biologically non-interesting). Empirical p-values are then the proportion of permuted *Rank Product* statistics which are bigger.

Selecting conditions whose empirical p-values are smaller than 0.05 ($N = 10,000$ permutations) produces the following result:

Condition	Dose	P-value	Example
Endo,10	100 μM	0.0001	Supp. movie 1
BPA,9	50 μM	0.0001	Supp. movie 2
PCB,10	100 μM	0.0028	
Endo,9	50 μM	0.016	
MeHg,9	1 μM	0.042	
PCB,9	50 μM	0.048	

Rather than peculiar movement types, according to our approach, the consistent results are conditions which are so strong as to freeze any nuclear motion (cf. supplementary movies 1 and 2). We therefore conclude that motility is significantly altered, albeit not primarily: measured motility alterations result from potent effects on cell viability. In contrast, motility alterations in Mitocheck are not coupled to cell death or cell division phenotypes. Consequently, this confirms that the identified genes are candidates for cell motility regulators: they do not trigger motility alterations as a secondary effect of other alterations, as it seems to be the case here.

⁶If a condition was replicated on the same plate, we used $rg(c, i) = median(\{rg(c_j, i) | j \text{ technical replicate of } c \text{ on } i\})$.

4.2.3 Phenotypic study

4.2.3.1 Phenotypic class selection

Phenotypic classes were chosen after an almost exhaustive visual inspection of the dataset. The first result is that except *frozen* nuclei which were observed in medium and high dose experiments only, no other striking phenotypes were observed following xenobiotic exposure only. Indeed *micronucleated* and *polylobed* nuclei are present in all wells at a non-negligible base level, and do not significantly appear following xenobiotic exposure.

4.2.3.2 Results

As opposed to the motility case, no specific plate distance distribution for any of the classes is significantly different to that of all other plate distance distribution. Hence distinct plate distances can be directly compared.

When no effect is expected, most cells are in interphase. Therefore, strong effects will be visible in a decrease of *interphase* percentage. A preliminary step consists in looking at conditions for which *interphase* distance is especially low. The *interphase* distances are represented on fig. 4.6. One can observe that only high doses (plain red dots) are consistently under the rest of the scatter plot. This depletion of interphases is explained by an increase in *frozen* and *apoptotic* nuclei (cf. the *frozen* distance scatter plot on fig. 4.6).

The observation of the distance distributions of other classes does not provide other results as to possible consequences of xenobiotic exposure on cell division, and especially low dose exposure: there is no low dose condition which show a consistent effect over distinct plates for any phenotypic class. Graphs are shown in annex, section A.4.

These observations are supported by the computation of *Rank product* statistics for *Interphase*, *Frozen* and *Micronucleated* nuclei (cf. table 4.3). The significant increase in *Micronucleated* nuclei following exposure to PCB at the first dose could be visually confirmed in 1 out of 4 experiments only. Furthermore, the level of micronucleation in the latter experiment is comparable to the level that can be observed in other experiments with *nothing* or *Nonane*. Indeed, there is a significant base-level of aberrant cell divisions, as is stressed by the small p-value for wells containing nothing (*Rien*) and visual inspection of the dataset.

TABLE 4.3: *Rank product* p-values for different phenotypic distances (<0.05)

Condition	Dose	P-value	Example
Decreased <i>Interphase</i> distances			
BPA,9	50 μM	0.0001	
MeHg,9	1 μM	0.0048	
PCB,10	100 μM	0.0059	
Increased <i>Frozen</i> distances			
BPA,9	50 μM	0.0001	Supp. movie 2
PCB,10	100 μM	0.0003	Supp. movie 3
MeHg,9	1 μM	0.0054	
Endo,10	100 μM	0.0097	
PCB,9	50 μM	0.048	
Increased <i>Micronucleated</i> distances			
PCB,1	0.1nM	0.015	Supp. movie 4
<i>Rien</i>	0	0.072	Supp. movie 5

4.3 Discussion

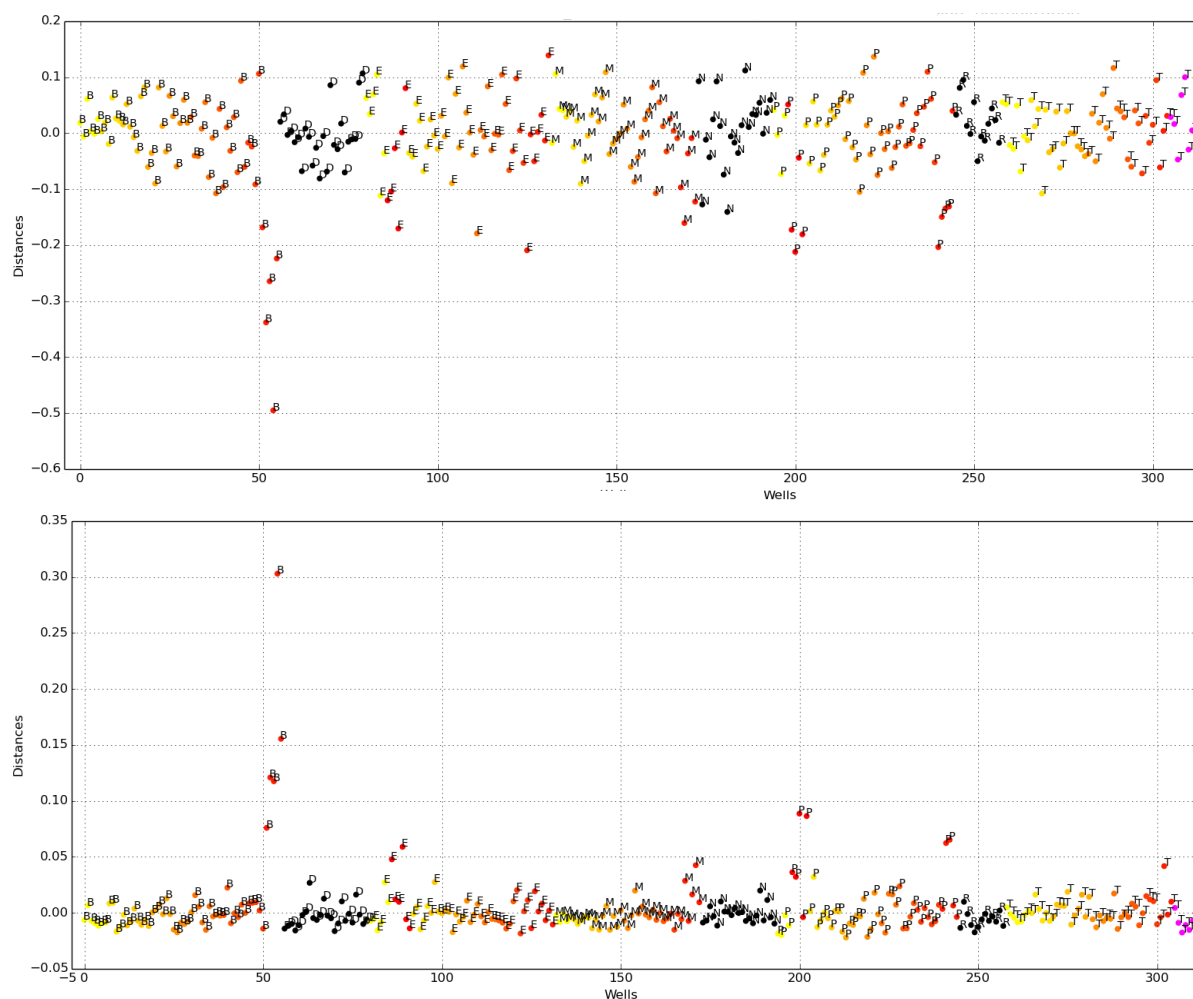
A workflow is established, which is accompanied by a Web interface, to enable the analysis of time-lapse xenobiotic screening experiments. However, on the panel of xenobiotics and doses which were chosen for our set of experiments, it was not possible to detect subtler effects than a toxic effect at the highest doses. This is probably due to one or a combination of the following reasons: we observed a relatively high level of experimental noise, even in negative controls, which might be due to (1) the biological variability of the used cell line, which showed many aberrant divisions, to (2) experimental noise due to microscope intensity variations and (3) low intensity of the fluorescence signal and (4) the sensitivity of our trajectory measurements. Even though our results are not conclusive on this project, we feel that it would be premature to conclude that the screening approach is not suited for Environmental Toxicology.

As a general remark, it should not be forgotten that it is very likely that a xenobiotic screen will never be as visually extraordinary as a siRNA or drug screen. Most xenobiotics are not - at sub-toxic doses - targeted at one single vital cellular process, as opposed to some siRNAs (especially those which were enlightened by the Mitocheck project, whose goal was to study cell division). Hence, they are less specific, and their effects can be slower. Nevertheless, previous results had let one hope that, e.g., significant cell division defects could be obtained following TCDD exposure ([Hutt et al., 2010], [Oikawa et al., 2008], [Oikawa et al., 2001]).

Dose choices

Ten doses were chosen for each xenobiotic, spanning from four to six orders of magnitude.

FIGURE 4.6: Interphase (up) and frozen(down) distances. Colors are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1.
 Legend: B: BPA, D: DMSO, E: Endo, M: MeHg, N: Nonane, P: PCB, R: nothing,
 T(red): TCDD, T(magenta): TGF- β 1



Dose ranges were selected in order to be non-toxic, as well as to include real human exposure dose (following a literature survey, see appendix A.3). This may not have been the surest way to detect an effect. Indeed, depending on their mechanisms of action, xenobiotics are going to be active in a smaller range, which we could easily have missed in the current setting.

In the future, dose choices could be improved in two ways. First of all, although human exposure doses are extremely relevant, they may be too small for an exploratory screen, where effects should be observed for the screen to be an efficient proof of concept. Rather, if previous similar studies exist, the doses which they find to be effective on cell division or cell motility should be used. However, given the novelty of the current approach in Environmental Toxicology, such previous similar studies are extremely rare. Previous studies using different cell lines or different measures for evaluating the same process

should be considered with caution⁷.

Hence a second and more robust way to improve dose choices would be to perform pre-screening (which also enables quality control setting and instrument and reagent validation - personal communications, Dr Wolfgang Huber and Dr Beate Neumann, EMBL, Heidelberg, Germany). The simplest way is to look for the lowest slightly toxic dose starting from 10 to 50 μM , and further dilute it of a factor 2 or 3 (rather than 10).

This is precisely the approach which was chosen by [Zimmer et al., 2014]. Willing to define a framework for developmental toxicity test battery, they had to select compounds for performing a proof of concept. They do note that human plasma concentrations are relevant when it comes to environmental contaminants such as PCB153. However given that they are around 1nM for the latter example, they also search for evidence of developmental toxicity in the literature, and do use higher concentrations while pre-screening and screening it. More elaborate approaches to pre-screening are also possible. As an example, [Peyre et al., 2014] used both cell viability assays and measures of cell area, mitochondrial activity and percentage of cells below 2N to define a "zone of interest" for further investigation.

In summary, dose choices should be chosen according to practical reasons (significant effect observed with local experimental parameters) instead of theoretical reasons (human exposure). Another parameter which should be set in this way is the time-lapse between cell exposure and screen start.

Cell line choice

A second parameter of importance is the cell line. Although MCF-7 cells are often chosen as a model for breast cancer, they exhibit a significant basal heterogeneity with regard to oestrogen receptor status and cell area [Palmari et al., 2000]. Screening is already subject to a certain amount of variance in its results, due to cell sensitivity to experimental parameters such as cell local confluence level [Snijder et al., 2009] or passage number, and gene expression stochasticity ([Elowitz et al., 2002], [Raj and van Oudenaarden, 2008]). For a proof of concept, it may therefore be relevant to choose a cell line which is inherently as homogeneous as possible.

In the current case, MCF-7 cells were genetically modified for incorporating H2B-mCherry and myrPalm-GFP. It appears that the chosen clone exhibits a high basal level of micronuclei. In the future, genetically modified clones should be checked both for a normal response to TCDD exposure and normal cell division.

⁷ As an example, although there exists several studies about measuring the impact of TCDD exposure on MCF-7 cell motility (e.g. [Diry et al., 2006], [Chen et al., 2012]), it is not clear at all that the same parameters are measured by MotIW (single cell motility versus cell population migration).

Chapter 5

Conclusion

Résumé - Conclusion (see *infra* for English text)

Grâce aux progrès dans les domaines de la robotique, de l'informatique, de la chimie organique et de la biologie moléculaire, les cribles biologiques à haut débit et haut contenu se sont multipliés ces dernières années. Parmi les techniques utilisées, la vidéomicroscopie a l'avantage de permettre une analyse plus fine des phénomènes rares et/ou dynamiques tels que la division ou la motilité cellulaires.

Cette approche produit de riches jeux de données, dont l'exploitation optimale reste une question ouverte. C'est ce à quoi nous nous sommes attachés à répondre dans cette thèse. Nous avons tout d'abord présenté le premier cadre méthodologique générale pour l'étude de la motilité cellulaire individuelle dans de telles données, MotIW. Le chapitre 3 démontre ensuite par trois exemples l'intérêt de la ré-utilisation des données de vidéomicroscopie produites dans le cadre de cribles à haut débit : l'application de MotIW aux données du projet Mitocheck [Neumann et al., 2010], l'étude du cycle cellulaire dans ces mêmes données, et enfin leur utilisation pour l'inférence de cibles thérapeutiques par leur comparaison avec un crible pharmaceutique non-publié. Le chapitre 4 présente en dernier lieu une approche méthodologique globale pour l'utilisation de la vidéomicroscopie en toxicologie environnementale.

Cette thèse a conduit à l'établissement de pistes sérieuses en ce qui concerne les gènes impliqués dans la motilité cellulaire, comme dans le cycle cellulaire. Elle a également abouti au développement d'une distance pour l'inférence de cibles thérapeutiques dans les données de vidéomicroscopie. Ces résultats gagneraient respectivement à être confirmés et utilisés dans d'autres systèmes. D'autre part, les expériences réalisées en toxicologie environnementale ont permis d'identifier les pierres d'achoppement de la procédure expérimentale. Une des perspectives de cette thèse serait par conséquent de reconduire les expériences en tenant compte des modifications suggérées.

In the last two decades, constant progress in the fields of molecular and cellular biology, laboratory hardware automation and computational methods for large scale data storage and data mining, have permitted high-throughput high-content experiments to presently be almost affordable and mainstream. As such, time-lapse microscopy has become more and more used. This has enabled a better understanding of complex dynamic biological processes such as cell division, which endpoint assays can more hardly help grasp. Not only are endpoint assays bound to miss rare and transient events, but they do not permit any assessment of the order in which displayed events happened.

Nevertheless, the question to know how to optimally develop computational methods for mining such large and complex datasets remains open. Indeed, time-lapse microscopy experiments produce three to five-dimensional datasets: 2 or 3 dimensions come from the images, time constitutes another one, and when a specific perturbation was studied (e.g. gene silencing, chemical exposure), it adds another dimension. The high-throughput quality of such experiments finally adds another difficulty: the size of the final dataset.

Main highlights

This is precisely the question which we have aimed at tackling in this thesis.

Given both that single cell motility is a particularly appropriate subject to be studied using time-lapse microscopy, and that there is currently no fully automated multivariate method for addressing such a question in this type of data, we have in the first place designed a generic methodological workflow for studying single cell motility in HT time-lapse microscopy experiments in **chapter 2**. This workflow was furthermore validated on a simulated screen, and applied to an existing dataset of approximately 150,000 videos.

This leads to the second main contribution of this thesis, namely the proof that HT HC time-lapse microscopy datasets constitute a rich and much valuable good. We think in particular that, should they be easily accessible and re-mined, their content could lead to more than one or two high-impact discoveries. In **chapter 3**, we therefore attached ourselves to re-discovering the Mitocheck dataset [Neumann et al., 2010] from the perspectives of single cell motility, cell cycle and drug target inference. This permitted to discover an ontology of single cell motility behaviours as well as a list of putative cell cycle genes. Furthermore, it enabled us to develop various other methods: for studying cell cycle in time-lapse experiments, and for performing drug target inference using phenotypic profiling on parallel siRNA and drug screens.

Finally, drawing on this methodological development, we exported the technique of HT HC time-lapse microscopy to *Environmental Toxicology* in **chapter 4**. Although the results we observed in our newly generated dataset were not up to our expectations, all

necessary methodological and practical tools have been developed, which are ready to be used on new data.

Perspectives

Application of our methodological workflow to the Mitocheck dataset produced a list of genes that might play a role in single cell motility. This list of genes was obtained in a specific model, HeLa cells, using a specific set of siRNAs. Therefore, confirmatory experiments in one (or more) different cell lines, using another set of siRNAs, should be performed in order to confirm our results. This would also permit to know if the ontology of single cell motility behaviour we obtained exists in other cell lines.

Similarly, we have developed a new distance for drug target inference by phenotypic profile comparison between parallel siRNA and drug screens. It would be interesting to confirm that this distance can apply to datasets using different markers and phenotypic classes. This distance could also benefit from further methodological development, which would take into account the temporal dimension of our data.

Finally, the reasons for the little success which was obtained in chapter 4 were at least partly identified. As the need for sophisticated and HT assays in Environmental Toxicology are rather increasing than diminishing, new data should be generated following the guidelines which this thesis permitted to identify. A simple and homogenous model organism should be used. It would be made fluorescent for one or two relevant markers, whose genetic insertion using modern techniques such as the CRISPR/Cas system would not alter any fundamental cellular processes. Last but not least, xenobiotics and doses should be chosen following robust preliminary experiments. HT use of such an assay in Environmental Toxicology would help us better understand our chemical environment.

Appendix A

Appendices

A.1 Cell cycle gene list

TABLE A.1: Cell cycle gene list. In bold are the three genes for which we found an extension of cell cycle length.

Hugo Gene Name	Ensembl gene id	Hugo Gene Name	Ensembl gene id
ADAMTS3	ENSG00000156140	MARK1	ENSG00000116141
APOA1	ENSG00000118137	MECR	ENSG00000116353
ARSF	ENSG00000062096	MGAT4A	ENSG00000071073
ATR	ENSG00000175054	MMP24	ENSG00000125966
B4GALT3	ENSG00000158850	MST1R	ENSG00000164078
BMPR1B	ENSG00000138696	MT-CO1	ENSG00000198804
BMPR2	ENSG00000204217	MTNR1A	ENSG00000168412
CACNA1D	ENSG00000157388	NEK10	ENSG00000163491
CDK15	ENSG00000138395	NOX1	ENSG00000007952
CDKN3	ENSG00000100526	NPC1	ENSG00000141458
CHRNA5	ENSG00000169684	NR1D1	ENSG00000126368
CIB3	ENSG00000141977	OR1F1	ENSG00000168124
CKB	ENSG00000166165	OSBP2	ENSG00000184792
CP	ENSG00000047457	OSMR	ENSG00000145623
DCK	ENSG00000156136	PAPD7	ENSG00000112941
DHPS	ENSG00000095059	PRPS2	ENSG00000101911
DIMT1	ENSG00000086189	PXDNL	ENSG00000147485
EIF2AK1	ENSG00000086232	PYGB	ENSG00000100994
F9	ENSG00000101981	RAB6B	ENSG00000154917
GDA	ENSG00000119125	RGL1	ENSG00000143344
GPR12	ENSG00000132975	RIPK2	ENSG00000104312
GPRC5C	ENSG00000170412	RNASEL	ENSG00000135828
HAS3	ENSG00000103044	RPS20	ENSG00000008988
ILK	ENSG00000166333	RPS6KA3	ENSG00000177189
ILVBL	ENSG00000105135	SERPINF1	ENSG00000132386
IP6K3	ENSG00000161896	SFMBT2	ENSG00000198879
KCNH6	ENSG00000173826	SGK1	ENSG00000118515
KCNMA1	ENSG00000156113	SPSB2	ENSG00000111671
KCNN1	ENSG00000105642	TAB1	ENSG00000100324
KIF20B	ENSG00000138182	TGM5	ENSG00000104055
KIFC2	ENSG00000167702	TRMT2A	ENSG00000099899
MAP4K4	ENSG00000071054	TTL	ENSG00000114999

A.2 Functional inference by in silico comparison of small-molecule and siRNA screens

A.2.1 Choice of λ parameter

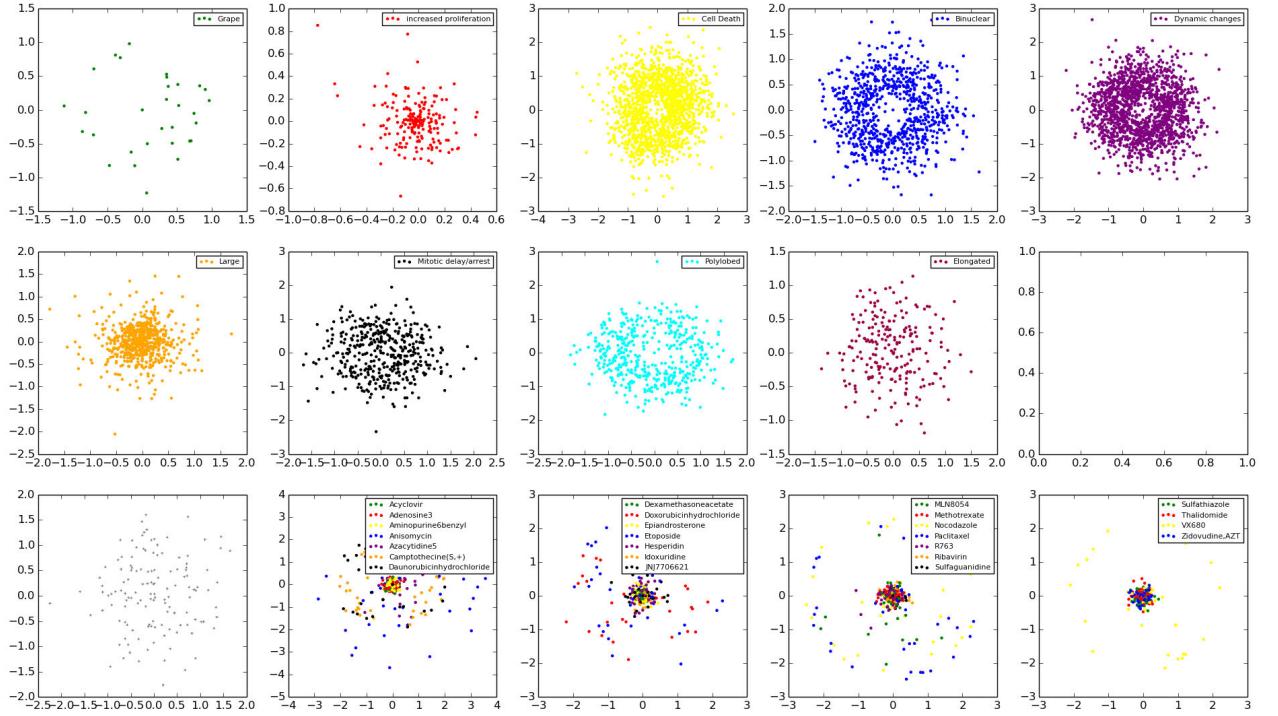


FIGURE A.1: Separation between Mitochek hit categories for $\lambda = 0.1$. Global Sinkhorn divergences between Mitochek hit experiments were computed for $\lambda = 0.1$, and multi-dimensional scaling was used for representing them in two dimensions in the first two lines. Divergences between theses experiments and the drug screen were included and their multi-dimension scaling is showed on the last line (grey: controls).

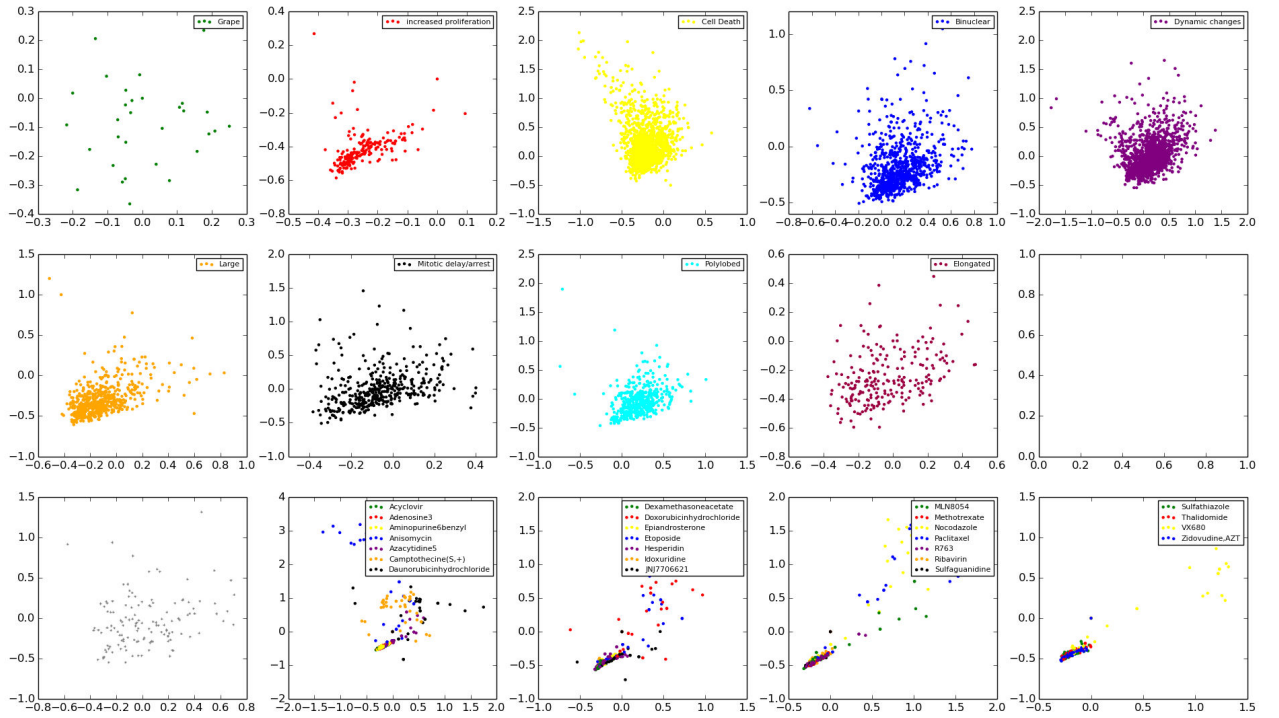


FIGURE A.2: Separation between Mitochek hit categories for $\lambda = 10$. Global Sinkhorn divergences between Mitochek hit experiments were computed for $\lambda = 10$, and multi-dimensional scaling was used for representing them in two dimensions in the first two lines. Divergences between these experiments and the drug screen were included and their multi-dimension scaling is showed on the last line (grey: controls).

A.2.2 Phenotypic scores of JNJ7706621

JNJ7706621

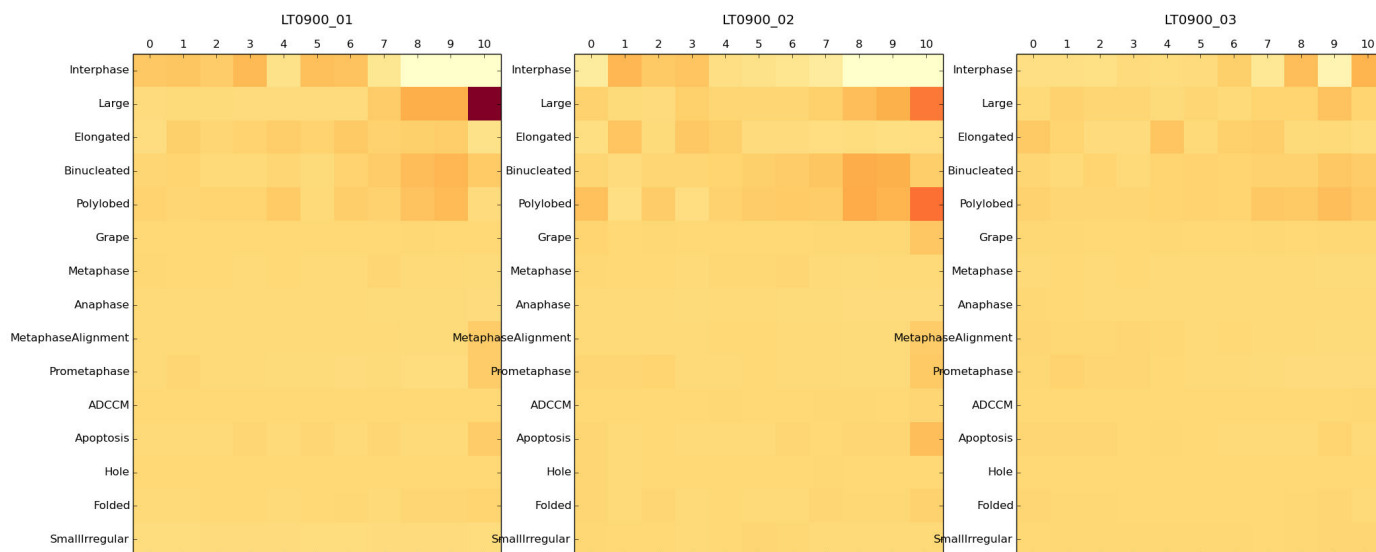


FIGURE A.3: Phenotypic scores of JNJ7706621 experiments, as a function of plate (left, middle, right) and dose (abscissa). The redder a square, the further away from control phenotypic scores.

A.2.3 Two-dimensional hierarchical clustering of drug screen condition distance to Mitochek siRNAs for different distances

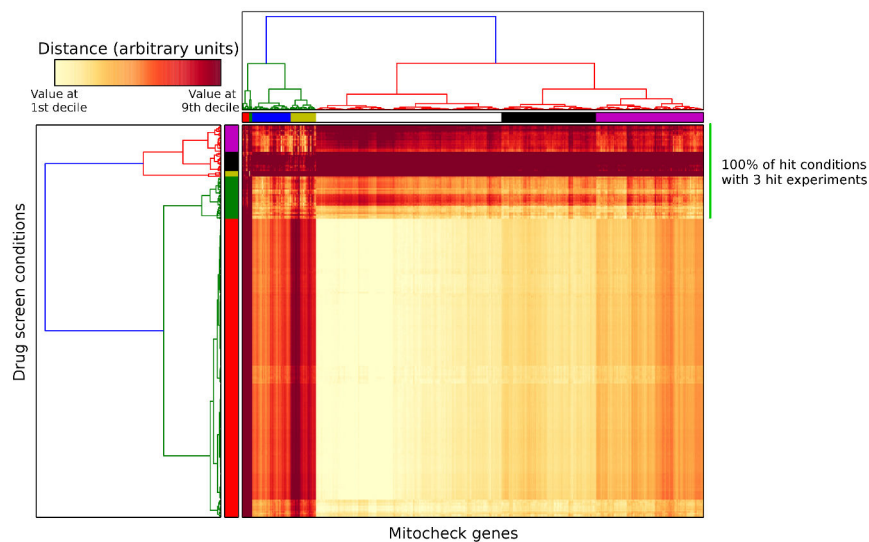


FIGURE A.4: Drug screen condition - Mitochek siRNA two-dimensional hierarchical clustering using sum of time Sinkhorn divergence. Ward method was used in combination with the Euclidean distance.

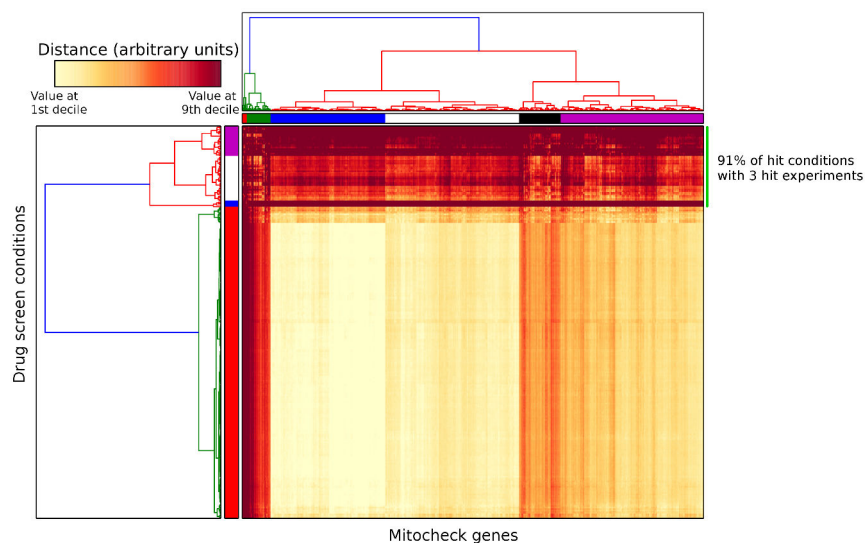


FIGURE A.5: Drug screen condition - Mitochek siRNA two-dimensional hierarchical clustering using phenotypic trajectory distance. Ward method was used in combination with the Euclidean distance.

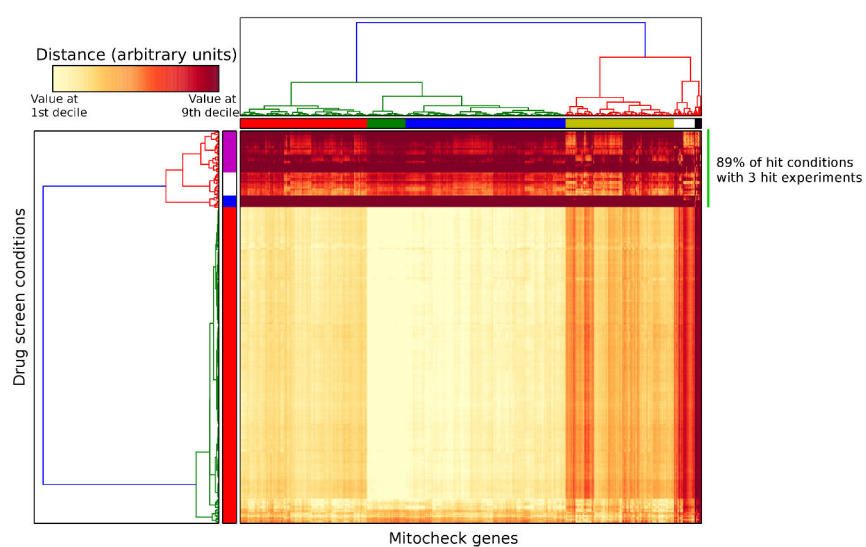


FIGURE A.6: Drug screen condition - Mitochondrial siRNA two-dimensional hierarchical clustering using Euclidean distance of phenotypic scores. Ward method was used in combination with the Euclidean distance.

A.2.4 Two-dimensional hierarchical clustering of drug screen hit condition distance to Mitochek siRNAs for different distances

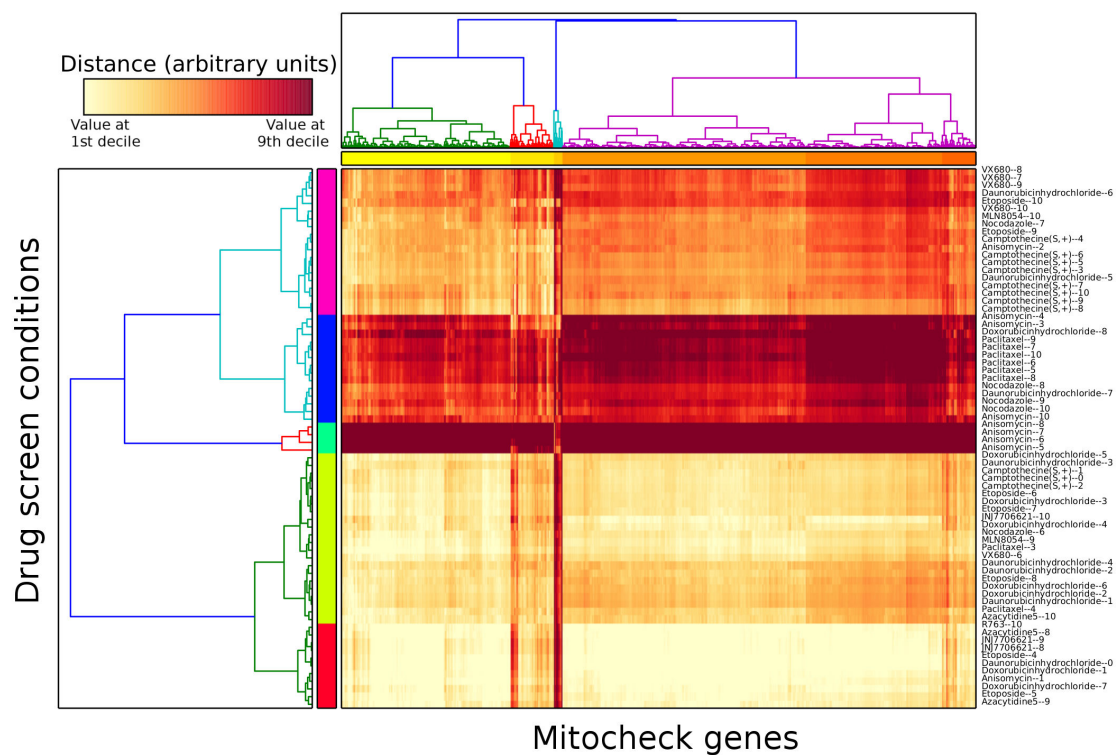


FIGURE A.7: Drug screen hit condition - Mitochek siRNA two-dimensional hierarchical clustering using sum of time Sinkhorn divergence. Ward method was used in combination with the Euclidean distance.

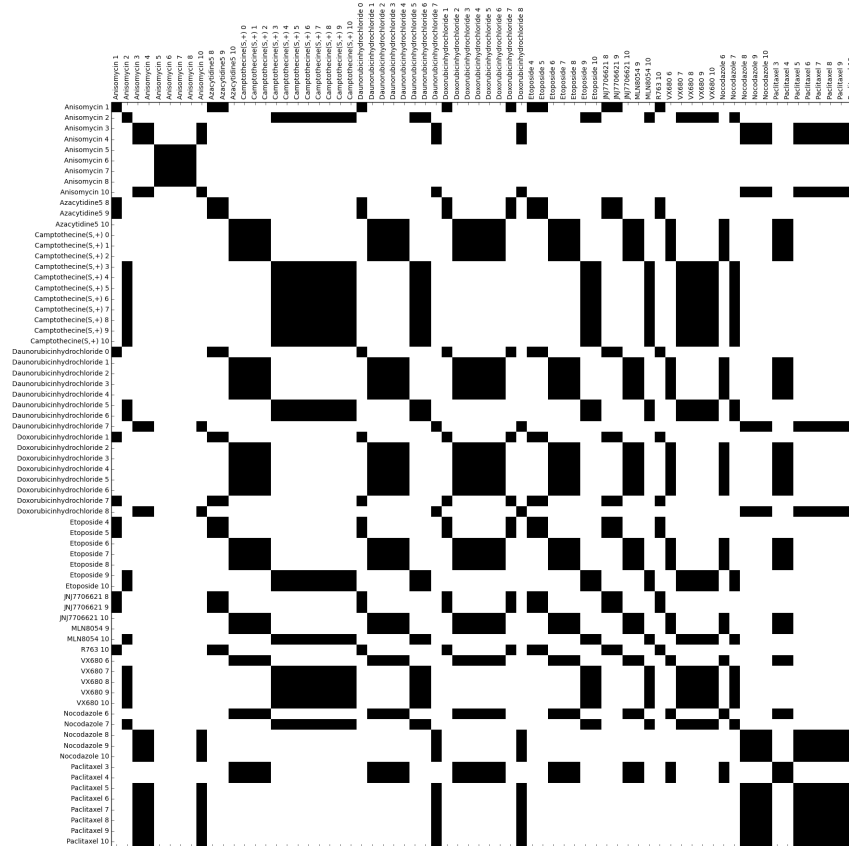


FIGURE A.8: Corresponding visualization of condition clustering for time Sinkhorn divergence. A black dot means that the conditions belong to the same cluster, a white dot that they do not.

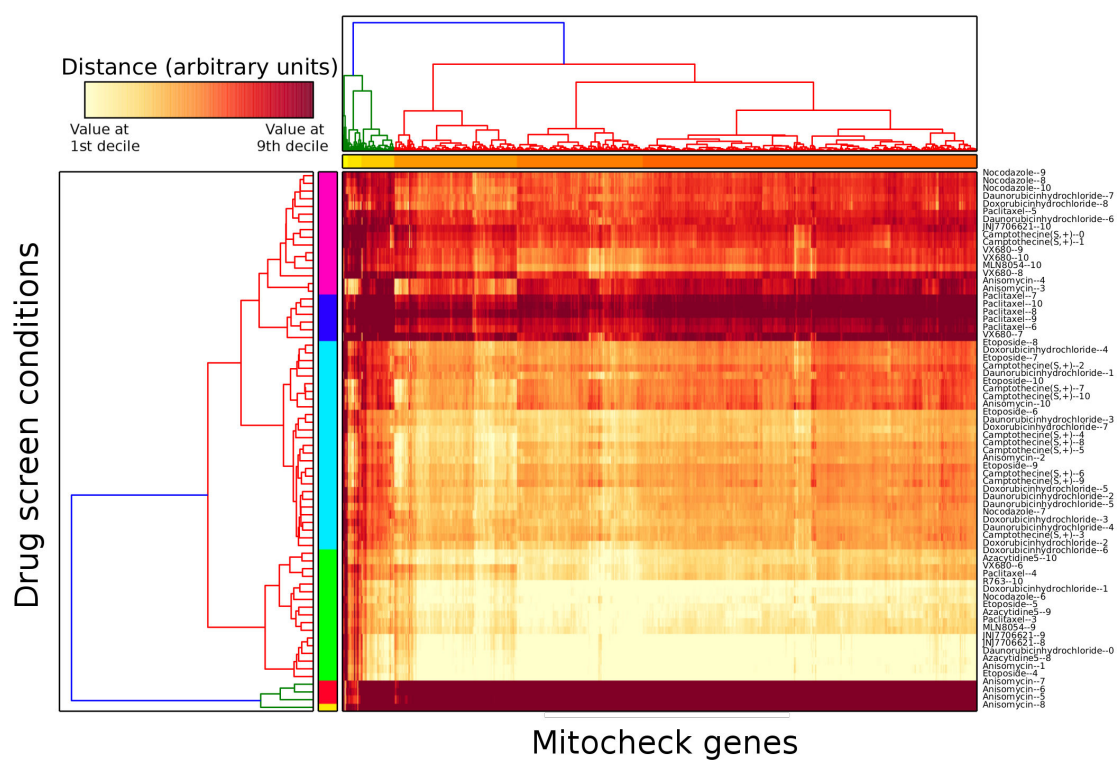


FIGURE A.9: Drug screen hit condition - Mitochondrial siRNA two-dimensional hierarchical clustering using phenotypic trajectory distance. Centroid method was used in combination with the Euclidean distance.

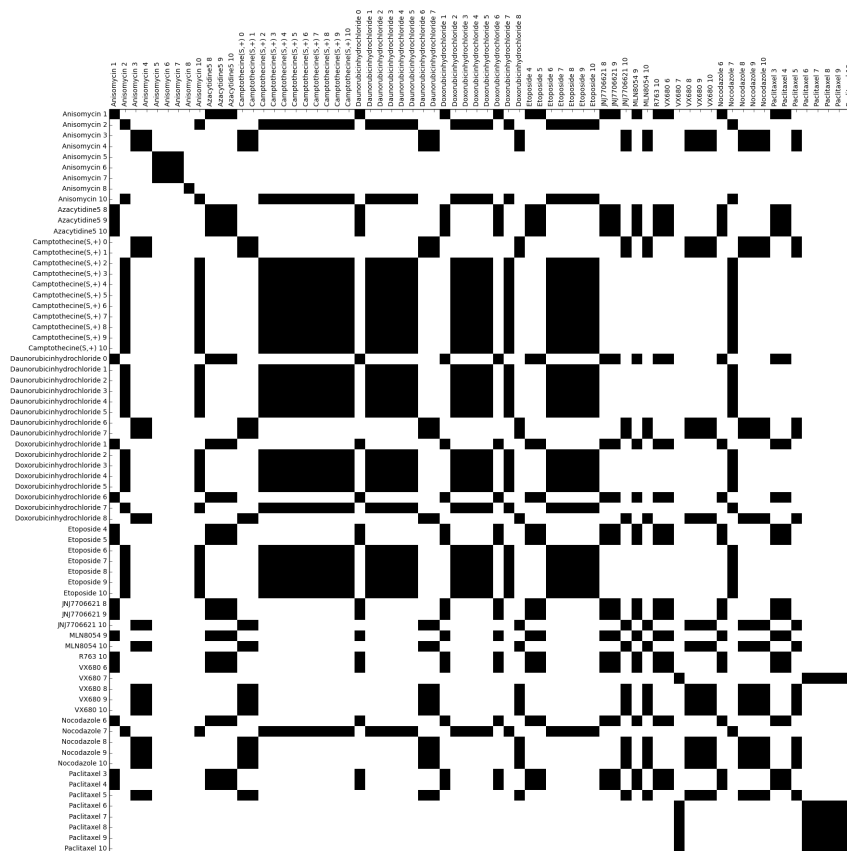


FIGURE A.10: Corresponding visualization of condition clustering for phenotypic trajectory distance. A black dot means that the conditions belong to the same cluster, a white dot that they do not.

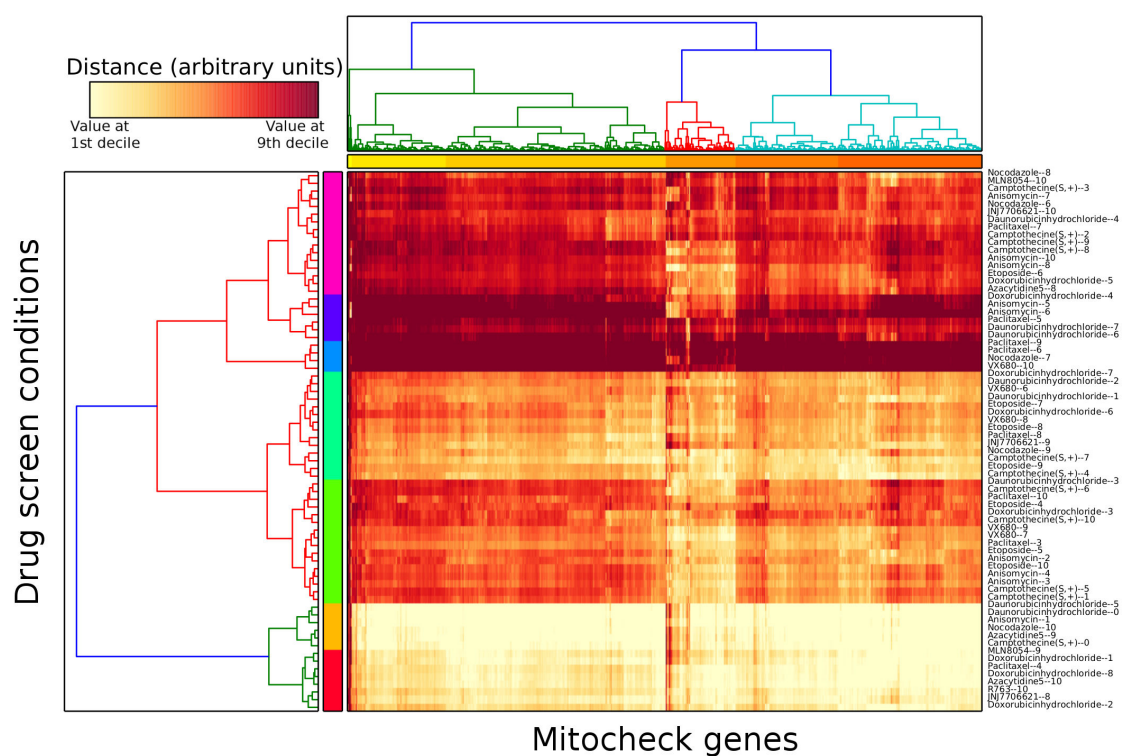


FIGURE A.11: Drug screen hit condition - Mitochondrial siRNA two-dimensional hierarchical clustering using Euclidean distance of phenotypic scores. Ward method was used in combination with the Euclidean distance.

A.3 Literature review

TABLE A.2: Human xenobiotic levels

Name	Human levels
BPA	Serum levels: order 0.9 to 87.6 nM(unconjugated, 0.2 to 20 ng/ml) [Vandenberg et al., 2009]
Dioxin	<p>Blood level: between $0.28 \cdot 10^{-3}$ pM and 0.27 pM</p> <ul style="list-style-type: none"> • Seveso, Italy: 12.4 pg/g lipid and 5.5 pg/g lipid (medians, plasma sampled between 1992 and 1994 vs accident in 1976, resp high contamination area and low contamination area, [Consonni et al., 2012]) • Japan: 1 pg/g lipid, 2 pg/g lipid (resp median, p75) [Arisawa et al., 2011]. Same order of magnitude in Canada [Rawn et al., 2012] • Germany: 0.020 pg/g lipid, 3.92 pg/g lipid (resp median, max) [Fromme et al., 2009] • Industrialized area, Germany: 1.3 pg/g lipid, 4.9 pg/g lipid (resp median, max) [Wittsiepe et al., 2007] <p>Adipose tissue: 2.05 pg/g lipid, 2.45 pg/g lipid (resp mean, P75 Spain [Lopez-Espinosa et al., 2008])</p> <p>Breast milk: 0.882 pg/g lipid, 3.58 pg/g lipid (resp median, max, China [Deng et al., 2012]), 1.5 pg/g lipid, 5.3 pg/g lipid (resp median, max, industrialized area Germany [Wittsiepe et al., 2007]). Same order of magnitude in France [Focant et al., 2013]</p>
Endosulfan	<p>Serum levels (1, 2):</p> <ul style="list-style-type: none"> • Contaminated Brazilian area: approx. 0.5, 0.6 nM (median), 1.1-1.2, 1.5-1.8 nM (p75) (resp. approx. 0.22, 0.25 ng/ml and 0.42-0.51, 0.62-0.75 ng/ml, [Freire et al., 2013]) • Baseline serum levels in farm workers: 1.30 μM (mean, 530 ng/ml, [Dalvie et al., 2009]) • Young male Spaniards: 3.61, 3.43 nM (median, 1.47, 1.00 ng/ml [Carreno et al., 2007])
MeHg	Blood plasma: 1.30 nM, 7.23 nM (resp mean, max, 0.28 μ g/L, 1.56 μ g/L, Hong-Kong residents [Liang et al., 2013])
Continued on next page	

Name	Human levels
	<p>Whole blood:</p> <ul style="list-style-type: none"> • 78.8 nM, 519.4 nM (resp median, max, 17.0 $\mu\text{g/L}$, 112 $\mu\text{g/L}$, Canadian inuits [Valera et al., 2013]) • 89.0 nM (P75 19.2 $\mu\text{g/L}$, contaminated environment, China [Chang et al., 2008]) • même ordre de grandeur (eg Sardaigne) ou un au-dessus (eg Brésil) dans différents endroits contaminés [Chang et al., 2008]
PCB153	<p>Serum levels: approx between 1nM and 500 nM</p> <ul style="list-style-type: none"> • Napoli: 42.3 ng/g lipid, 195.3 ng/g lipid (median, max [Esposito et al., 2014]) • Slovakia: 232/578 ng/g lipid, 5,193/25,089 ng/g lipid (background area/contaminated area, median, max [Esposito et al., 2014]) • >65 years old, Canada: 73.6 ng/g lipid, 208 ng/g lipid (median, P95 [Medehouenou et al., 2011]) • Inuits, Canada: 177 ng/g lipid, 6,020 ng/g lipid (geom mean, max [Medehouenou et al., 2010]) <p>Whole blood: 0.89 nM, 6.65 nM (0.32 $\mu\text{g/l}$ median, 2.4 $\mu\text{g/l}$ max, industrialized region, Germany [Wittsiepe et al., 2007])</p> <p>Breast milk: 20 ng/g lipid, 49 ng/g lipid (median, max, Philippines [Malarvannan et al., 2013]), 3.4 ng/g lipid, 8.0 ng/g lipid (mean, max, Turkey [Cok et al., 2012]), range 20-183 ng/g lipid (Northern Russia [Polder et al., 2008])</p>

A.4 Phenotypic study

FIGURE A.12: Apoptosis distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates.

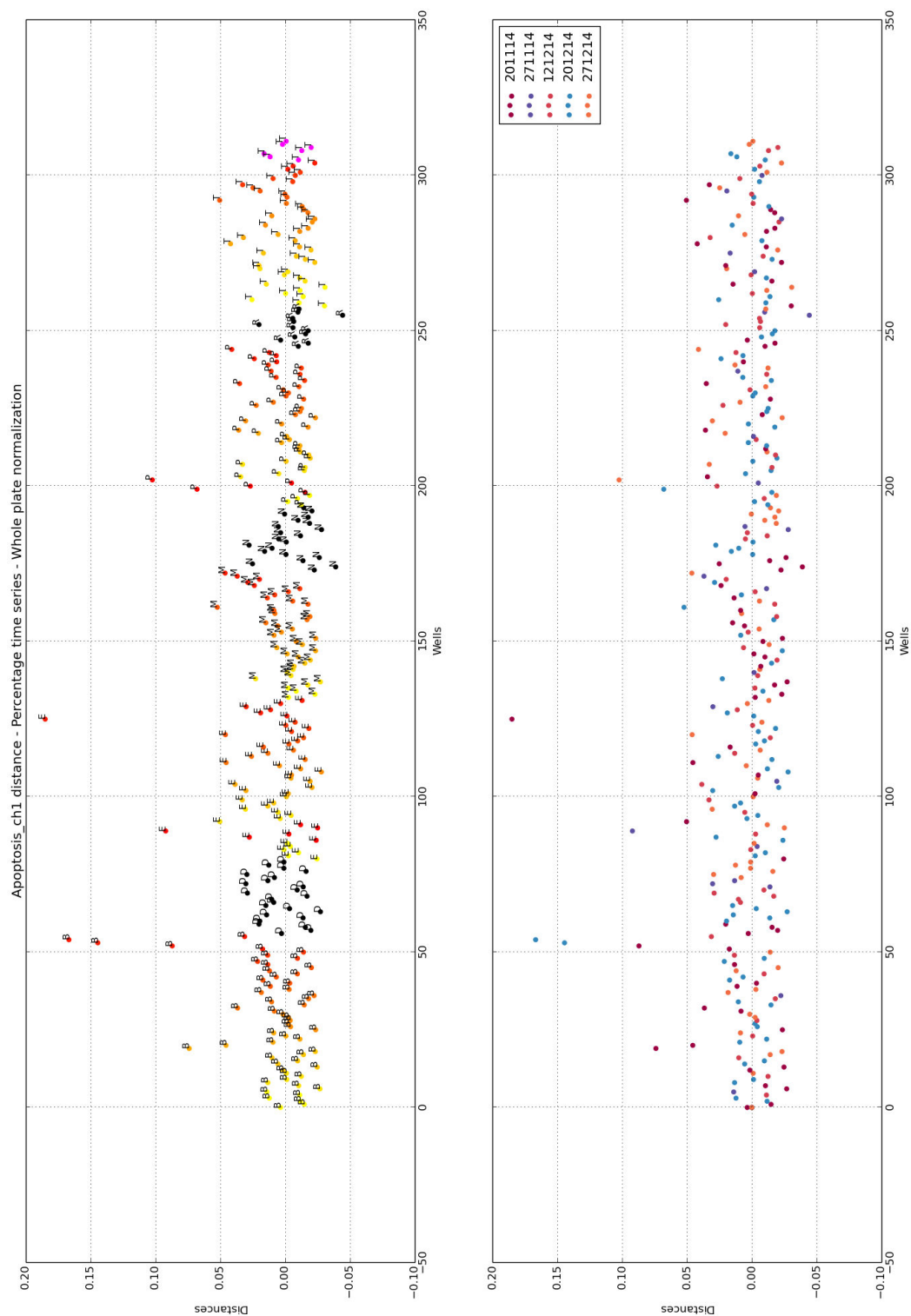


FIGURE A.13: Frozen distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates.

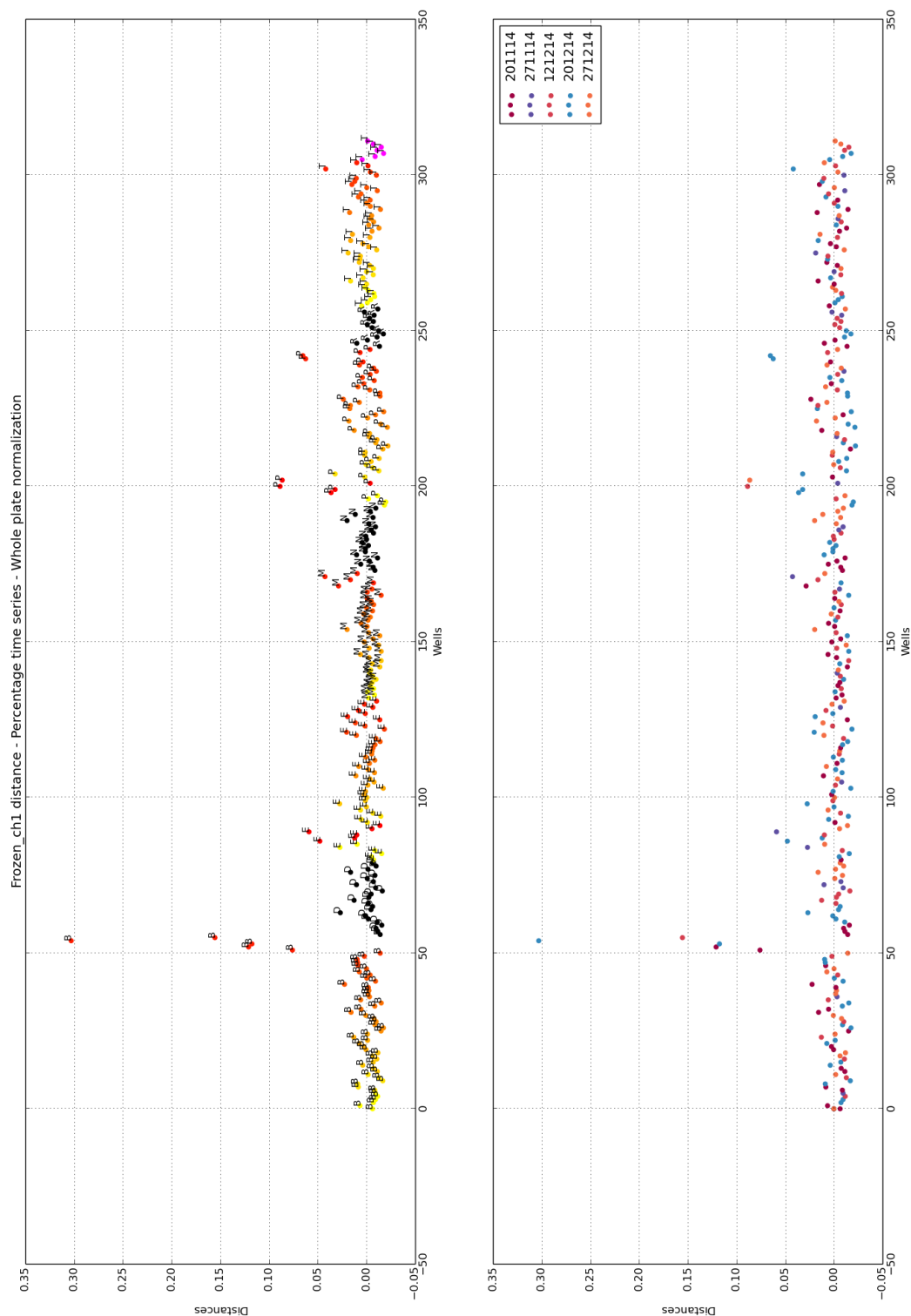


FIGURE A.14: Interphase distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates.

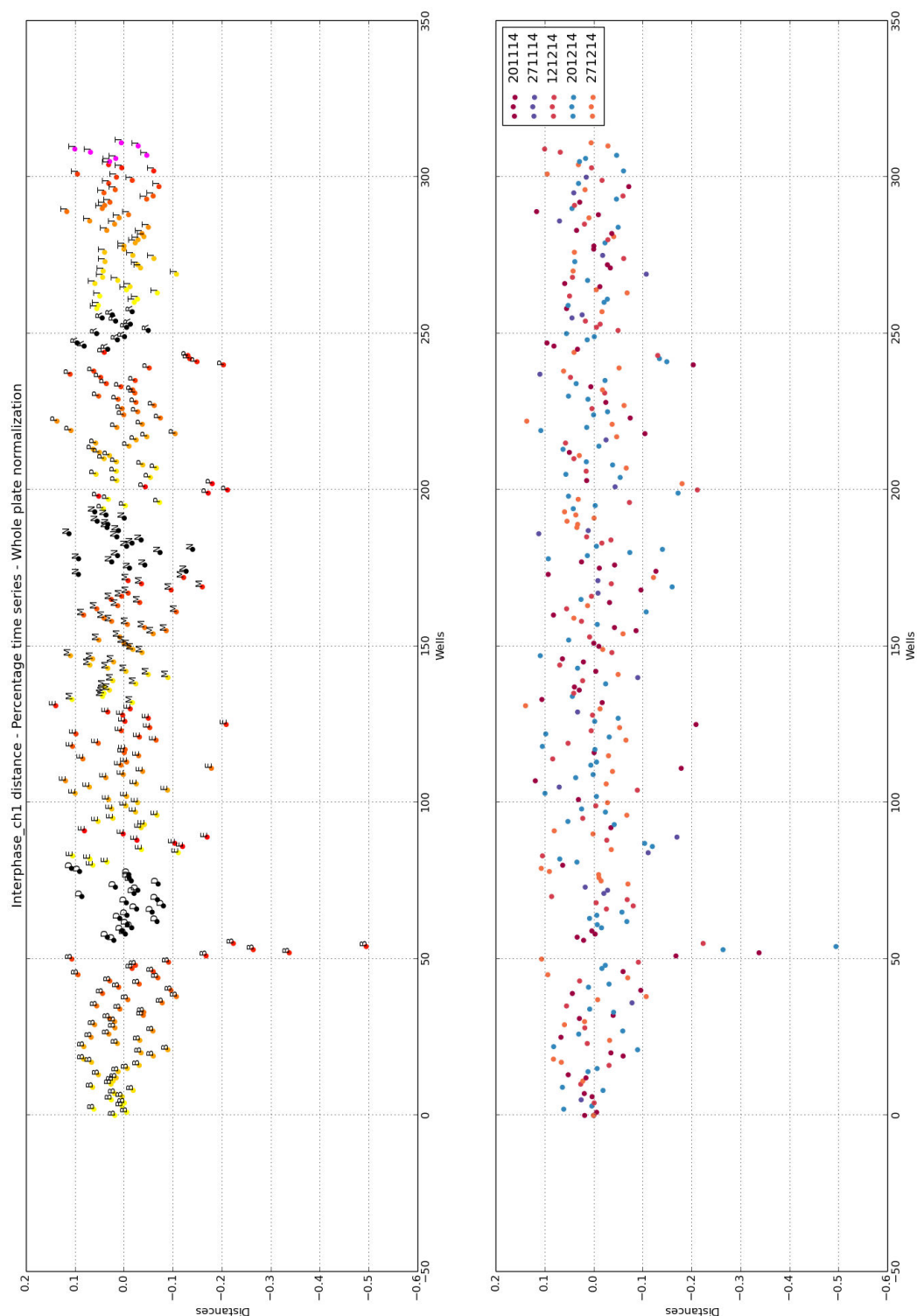


FIGURE A.15: Metaphase distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates.

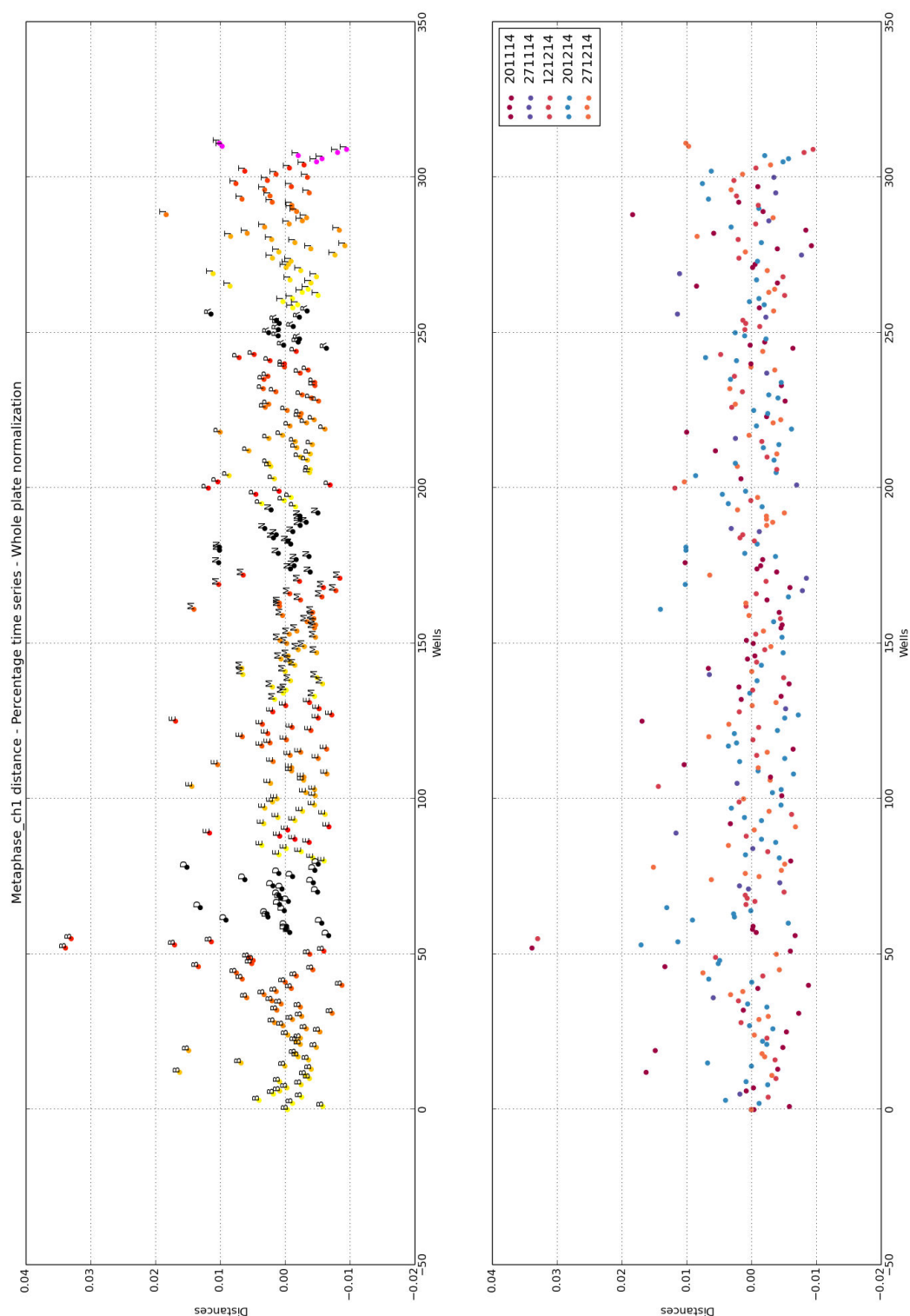


FIGURE A.16: Micronucleated distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates.

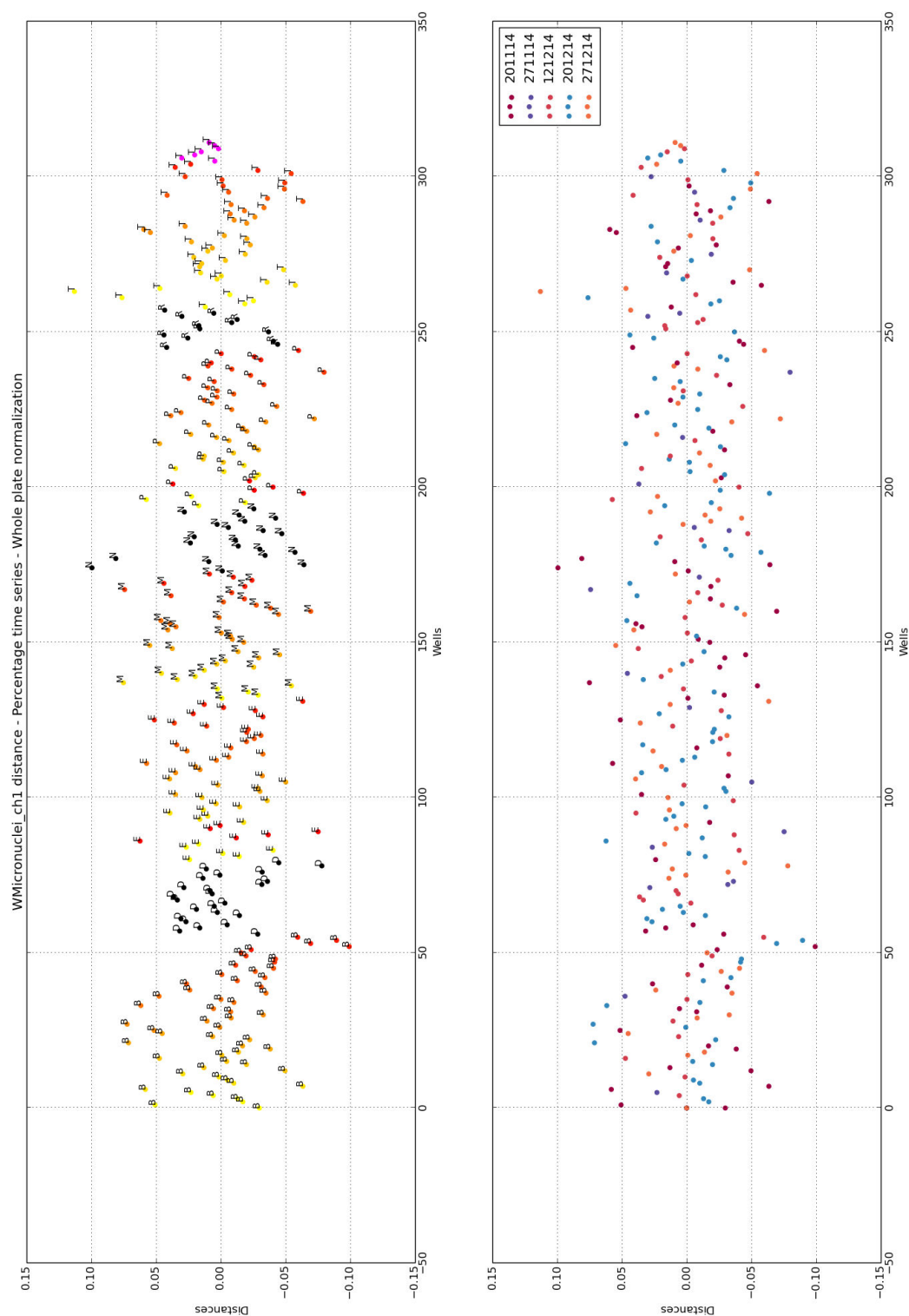
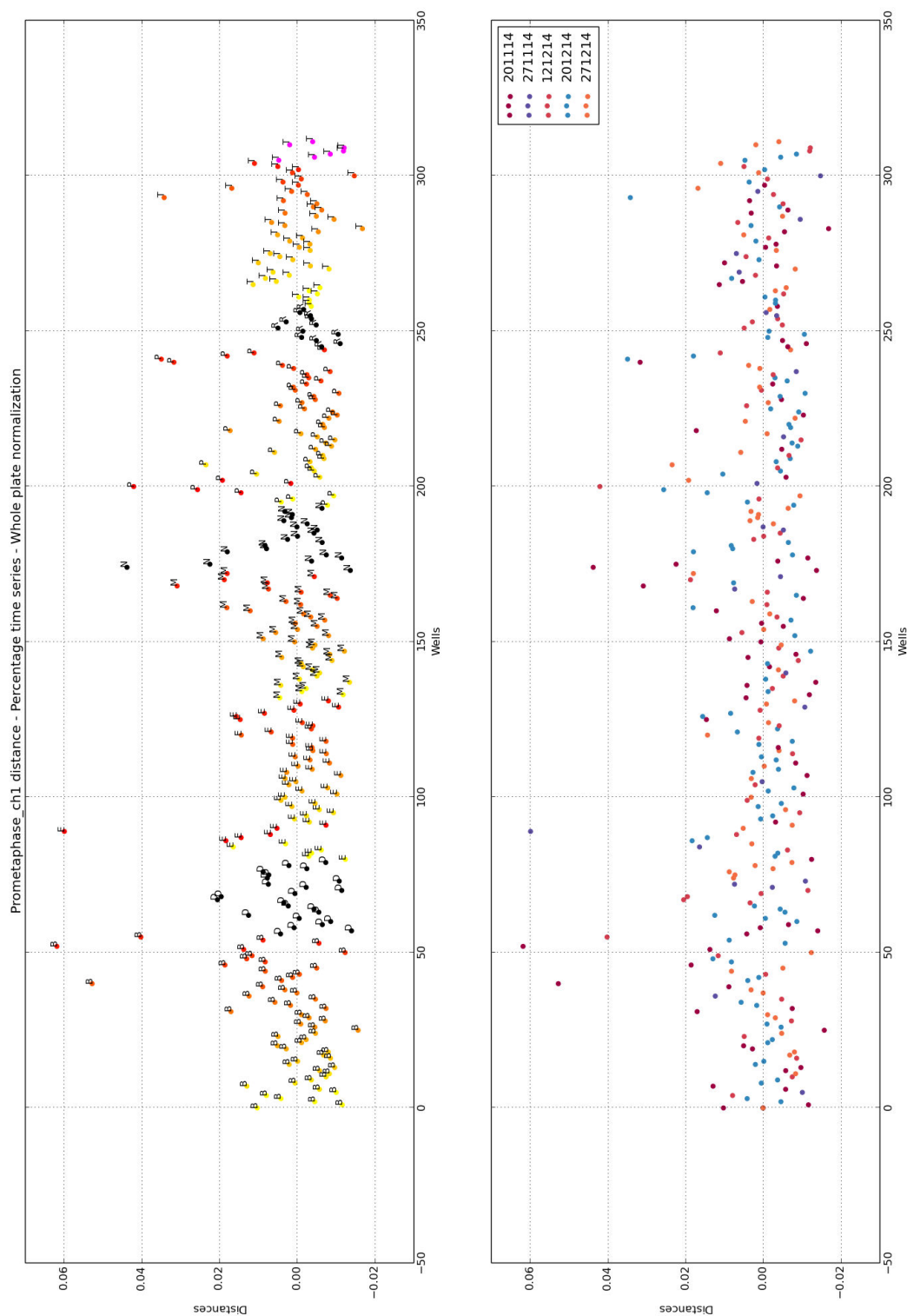


FIGURE A.17: Prometaphase distances. Colors (up) are black for control wells, yellow to red for xenobiotics ranked by increased dose, and magenta for TGF- β 1. B: BPA, D:DMSO, E: Endo, M: MeHg, N: Nonane, P:PCB, R: nothing, T(red):TCDD, T(magenta): TGF- β 1. Colors (down) are for plates.



Bibliography

- [Adey et al., 2013] Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., Qiu, R., Lee, C., and Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, 500(7461):207–211.
- [Albrecht-Buehler, 1977] Albrecht-Buehler, G. (1977). The phagokinetic tracks of 3T3 cells. *Cell*, 11(2):395–404.
- [Arisawa et al., 2011] Arisawa, K., Uemura, H., Hiyoshi, M., Kitayama, A., Takami, H., Sawachika, F., Nishioka, Y., Hasegawa, M., Tanto, M., Satoh, H., Shima, M., Sumiyoshi, Y., Morinaga, K., Kodama, K., Suzuki, T., and Nagai, M. (2011). Dietary patterns and blood levels of PCDDs, PCDFs, and dioxin-like PCBs in 1656 Japanese individuals. *Chemosphere*, 82(5):656–662.
- [Ben-Hur et al., 2002] Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pac Symp Biocomput*, pages 6–17.
- [Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300.
- [Benjamini and Yekutieli, 2001] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.
- [Biol., 2005] Biol., P. (2005). The Evolution of Self-Fertile Hermaphroditism: The Fog Is Clearing. *PLoS Biol.*, 3(1):e30.
- [Birmingham et al., 2009] Birmingham, A., Selfors, L. M., Forster, T., Wrobel, D., Kennedy, C. J., Shanks, E., Santoyo-Lopez, J., Dunican, D. J., Long, A., Kelleher, D., Smith, Q., Beijersbergen, R. L., Ghazal, P., and Shamu, C. E. (2009). Statistical methods for analysis of high-throughput RNA interference screens. *Nat. Methods*, 6(8):569–575.

- [Bokobza et al., 2009] Bokobza, S. M., Ye, L., Kynaston, H. E., Mansel, R. E., and Jiang, W. G. (2009). Reduced expression of BMPR-IB correlates with poor prognosis and increased proliferation of breast cancer cells. *Cancer Genomics Proteomics*, 6(2):101–108.
- [Boyden, 1962] Boyden, S. (1962). The chemotactic effect of mixtures of antibody and antigen on polymorphonuclear leucocytes. *J. Exp. Med.*, 115:453–466.
- [Breitling et al., 2004] Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, 573(1-3):83–92.
- [Carpenter et al., 2006] Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. a., Chang, J. H., Lindquist, R. a., Moffat, J., Golland, P., and Sabatini, D. M. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100.
- [Carreno et al., 2007] Carreno, J., Rivas, A., Granada, A., Jose Lopez-Espinosa, M., Mariscal, M., Olea, N., and Olea-Serrano, F. (2007). Exposure of young men to organochlorine pesticides in Southern Spain. *Environ. Res.*, 103(1):55–61.
- [Chalfie et al., 1994] Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., and Prasher, D. C. (1994). Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802–805.
- [Chang et al., 2008] Chang, J. W., Pai, M. C., Chen, H. L., Guo, H. R., Su, H. J., and Lee, C. C. (2008). Cognitive function and blood methylmercury in adults living near a deserted chloralkali factory. *Environ. Res.*, 108(3):334–339.
- [Chen et al., 2012] Chen, Y. J., Hung, C. M., Kay, N., Chen, C. C., Kao, Y. H., and Yuan, S. S. (2012). Progesterone receptor is involved in 2,3,7,8-tetrachlorodibenzo-p-dioxin-stimulated breast cancer cells proliferation. *Cancer Lett.*, 319(2):223–231.
- [Chenouard et al., 2014] Chenouard, N., Smal, I., de Chaumont, F., Maska, M., Sbalzarini, I. F., Gong, Y., Cardinale, J., Carthel, C., Coraluppi, S., Winter, M., Cohen, A. R., Godinez, W. J., Rohr, K., Kalaidzidis, Y., Liang, L., Duncan, J., Shen, H., Xu, Y., Magnusson, K. E. G., Jalden, J., Blau, H. M., Paul-Gilloteaux, P., Roudot, P., Kervrann, C., Waharte, F., Tinevez, J.-Y., Shorte, S. L., Willemse, J., Celler, K., van Wezel, G. P., Dan, H.-W., Tsai, Y.-S., de Solorzano, C. O., Olivo-Marin, J.-C., and Meijering, E. (2014). Objective comparison of particle tracking methods. *Nat Meth*, 11(3):281–289.

- [Choudhury et al., 2014] Choudhury, S., Fishman, J. R., McGowan, M. L., and Juengst, E. T. (2014). Big data, open science and the brain: lessons learned from genomics. *Front Hum Neurosci*, 8:239.
- [Cok et al., 2012] Cok, I., Mazmanci, B., Mazmanci, M. A., Turgut, C., Henkelmann, B., and Schramm, K. W. (2012). Analysis of human milk to assess exposure to PAHs, PCBs and organochlorine pesticides in the vicinity Mediterranean city Mersin, Turkey. *Environ Int*, 40:63–69.
- [Collinet et al., 2010] Collinet, C., Stöter, M., Bradshaw, C. R., Samusik, N., Rink, J. C., Kenski, D., Habermann, B., Buchholz, F., Henschel, R., Mueller, M. S., et al. (2010). Systems survey of endocytosis by multiparametric image analysis. *Nature*, 464(7286):243–249.
- [Consonni et al., 2012] Consonni, D., Sindaco, R., Agnello, L., Caporaso, N. E., Landi, M. T., Pesatori, A. C., and Bertazzi, P. A. (2012). Plasma levels of dioxins, furans, non-ortho-PCBs, and TEQs in the Seveso population 17 years after the accident. *Med Lav*, 103(4):259–267.
- [Coquelle et al., 2006] Coquelle, A., Mouhamad, S., Pequignot, M. O., Braun, T., Carvalho, G., Vivet, S., Metivier, D., Castedo, M., and Kroemer, G. (2006). Enrichment of non-synchronized cells in the G1, S and G2 phases of the cell cycle for the study of apoptosis. *Biochem. Pharmacol.*, 72(11):1396–1404.
- [Costa, 1983] Costa, M. (1983). Sequential events in the induction of transformation in cell culture by specific nickel compounds. *Biol Trace Elem Res*, 5(4-5):285–295.
- [Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2292–2300.
- [Dalvie et al., 2009] Dalvie, M. A., Africa, A., Solomons, A., London, L., Brouwer, D., and Kromhout, H. (2009). Pesticide exposure and blood endosulfan levels after first season spray amongst farm workers in the Western Cape, South Africa. *J Environ Sci Health B*, 44(3):271–277.
- [Decaestecker et al., 2007] Decaestecker, C., Debeir, O., Van Ham, P., and Kiss, R. (2007). Can anti-migratory drugs be screened in vitro? A review of 2D and 3D assays for the quantitative analysis of cell migration. *Med Res Rev*, 27(2):149–176.
- [Delattre and Roquain, 2015] Delattre, S. and Roquain, E. (2015). New procedures controlling the false discovery proportion via romano-wolf’s heuristic. *Annals of Statistics*, 43(3):1141–1177.

- [Deng et al., 2012] Deng, B., Zhang, J., Zhang, L., Jiang, Y., Zhou, J., Fang, D., Zhang, H., and Huang, H. (2012). Levels and profiles of PCDD/Fs, PCBs in mothers' milk in Shenzhen of China: estimation of breast-fed infants' intakes. *Environ Int*, 42:47–52.
- [Diry et al., 2006] Diry, M., Tomkiewicz, C., Koehle, C., Coumoul, X., Bock, K. W., Barouki, R., and Transy, C. (2006). Activation of the dioxin/aryl hydrocarbon receptor (AhR) modulates cell plasticity through a JNK-dependent mechanism. *Oncogene*, 25(40):5570–5574.
- [Eggert et al., 2004] Eggert, U. S., Kiger, A. A., Richter, C., Perlman, Z. E., Perrimon, N., Mitchison, T. J., and Field, C. M. (2004). Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol.*, 2(12):e379.
- [Elowitz et al., 2002] Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.
- [Emanuel et al., 2005] Emanuel, S., Rugg, C. A., Gruninger, R. H., Lin, R., Fuentes-Pesquera, A., Connolly, P. J., Wetter, S. K., Hollister, B., Kruger, W. W., Napier, C., Jolliffe, L., and Middleton, S. A. (2005). The in vitro and in vivo effects of JNJ-7706621: a dual inhibitor of cyclin-dependent kinases and aurora kinases. *Cancer Res.*, 65(19):9038–9046.
- [Esposito et al., 2014] Esposito, M., Serpe, F. P., Diletti, G., Messina, G., Scortichini, G., La Rocca, C., Baldi, L., Amorena, M., and Monda, M. (2014). Serum levels of polychlorinated dibenzo-p-dioxins, polychlorinated dibenzofurans and polychlorinated biphenyls in a population living in the Naples area, southern Italy. *Chemosphere*, 94:62–69.
- [Feng et al., 2009] Feng, Y., Mitchison, T. J., Bender, A., Young, D. W., and Tallarico, J. A. (2009). Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat Rev Drug Discov*, 8(7):567–578.
- [Ferrari et al., 2001] Ferrari, R., Manfro, A., and Young, W. (2001). Strongly and weakly self-similar diffusion. *Physica D*, 154:111–137.
- [Fiorini et al., 2008] Fiorini, C., Gilleron, J., Carette, D., Valette, A., Tilloy, A., Chevalier, S., Segretain, D., and Pointis, G. (2008). Accelerated internalization of junctional membrane proteins (connexin 43, N-cadherin and ZO-1) within endocytic vacuoles: an early event of DDT carcinogenicity. *Biochim. Biophys. Acta*, 1778:56–67.
- [Focant et al., 2013] Focant, J. F., Frery, N., Bidondo, M. L., Eppe, G., Scholl, G., Saoudi, A., Oleko, A., and Vandentorren, S. (2013). Levels of polychlorinated dibenzo-p-dioxins, polychlorinated dibenzofurans and polychlorinated biphenyls in human milk from different regions of France. *Sci. Total Environ.*, 452-453:155–162.

- [Freire et al., 2013] Freire, C., Koifman, R. J., Sarcinelli, P. N., Rosa, A. C., Clapauch, R., and Koifman, S. (2013). Association between serum levels of organochlorine pesticides and sex hormones in adults living in a heavily contaminated area in Brazil. *Int J Hyg Environ Health*.
- [Freitas et al., 2014] Freitas, J., Miller, N., Mengeling, B. J., Xia, M., Huang, R., Houck, K., Rietjens, I. M., Furlow, J. D., and Murk, A. J. (2014). Identification of thyroid hormone receptor active compounds using a quantitative high-throughput screening platform. *Curr Chem Genomics Transl Med*, 8:36–46.
- [Friedl and Weigelin, 2008] Friedl, P. and Weigelin, B. (2008). Interstitial leukocyte migration and immune function. *Nat. Immunol.*, 9(9):960–969.
- [Fromme et al., 2009] Fromme, H., Albrecht, M., Boehmer, S., Buchner, K., Mayer, R., Liebl, B., Wittschiepe, J., and Bolte, G. (2009). Intake and body burden of dioxin-like compounds in Germany: the INES study. *Chemosphere*, 76(11):1457–1463.
- [Gatti et al., 2004] Gatti, R., Belletti, S., Uggeri, J., Vettori, M. V., Mutti, A., Scandroglio, R., and Orlandini, G. (2004). Methylmercury cytotoxicity in PC12 cells is mediated by primary glutathione depletion independent of excess reactive oxygen species generation. *Toxicology*, 204:175–185.
- [Giaever et al., 2002] Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K. D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C. Y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., and Johnston, M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896):387–391.
- [Gwack et al., 2006] Gwack, Y., Sharma, S., Nardone, J., Tanasa, B., Iuga, A., Srikanth, S., Okamura, H., Bolton, D., Feske, S., Hogan, P. G., and Rao, A. (2006). A genome-wide *Drosophila* RNAi screen identifies DYRK-family kinases as regulators of NFAT. *Nature*, 441(7093):646–650.

- [Halkidi et al., 2001] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145.
- [Held et al., 2010] Held, M., Schmitz, M. H., Fischer, B., Walter, T., Neumann, B., Olma, M. H., Peter, M., Ellenberg, J., and Gerlich, D. W. (2010). CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nat. Methods*, 7(9):747–754.
- [Huck et al., 2010] Huck, J. J., Zhang, M., McDonald, A., Bowman, D., Hoar, K. M., Stringer, B., Ecsedy, J., Manfredi, M. G., and Hyer, M. L. (2010). MLN8054, an inhibitor of Aurora A kinase, induces senescence in human tumor cells both in vitro and in vivo. *Mol. Cancer Res.*, 8(3):373–384.
- [Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95.
- [Hutt et al., 2010] Hutt, K. J., Shi, Z., Petroff, B. K., and Albertini, D. F. (2010). The environmental toxicant 2,3,7,8-tetrachlorodibenzo-p-dioxin disturbs the establishment and maintenance of cell polarity in preimplantation rat embryos. *Biol. Reprod.*, 82(5):914–920.
- [Ito et al., 2010] Ito, T., Ando, H., Suzuki, T., Ogura, T., Hotta, K., Imamura, Y., Yamaguchi, Y., and Handa, H. (2010). Identification of a primary target of thalidomide teratogenicity. *Science*, 327(5971):1345–1350.
- [Jaqaman et al., 2008] Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S. L., and Danuser, G. (2008). Robust single-particle tracking in live-cell time-lapse sequences. *Nat. Methods*, 5:695–702.
- [Jiao et al., 2012] Jiao, X., Sherman, B. T., Huang, d. a. W., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806.
- [Jones et al., 2001] Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed 2015-01-07].
- [Kim et al., 2000] Kim, I. Y., Lee, D. H., Ahn, H. J., Tokunaga, H., Song, W., Devereaux, L. M., Jin, D., Sampath, T. K., and Morton, R. A. (2000). Expression of bone morphogenetic protein receptors type-IA, -IB and -II correlates with tumor grade in human prostate cancer tissues. *Cancer Res.*, 60(11):2840–2844.
- [Kovalev et al., 2006] Kovalev, V., Harder, N., Neumann, B., Held, M., Liebel, U., Erfle, H., Ellenberg, J., Eils, R., and Rohr, K. (2006). Feature selection for evaluating

- fluorescence microscopy images in genome-wide cell screens. In *CVPR (1)*, pages 276–283. IEEE Computer Society.
- [Kramer et al., 2013] Kramer, N., Walzl, A., Unger, C., Rosner, M., Krupitza, G., Hengstschlager, M., and Dolznig, H. (2013). In vitro cell migration and invasion assays. *Mutat. Res.*, 752(1):10–24.
- [Lantuejoul, 1982] Lantuejoul, C. (1982). *Multicomputers and Image Processing: Algorithms and Programs*. Academic Press, New-York, NY, USA.
- [Lara et al., 2011] Lara, R., Mauri, F. A., Taylor, H., Derua, R., Shia, A., Gray, C., Nicols, A., Shiner, R. J., Schofield, E., Bates, P. A., Waelkens, E., Dallman, M., Lamb, J., Zicha, D., Downward, J., Seckl, M. J., and Pardo, O. E. (2011). An siRNA screen identifies RSK1 as a key modulator of lung cancer metastasis. *Oncogene*, 30(32):3513–3521.
- [Lecuyer et al., 2007] Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., and Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 131(1):174–187.
- [Lee et al., 2013] Lee, K., Na, W., Maeng, J. H., Wu, H., and Ju, B. G. (2013). Regulation of DU145 prostate cancer cell growth by Scm-like with four mbt domains 2. *J. Biosci.*, 38(1):105–112.
- [Leonhardt et al., 2000] Leonhardt, H., Rahn, H. P., Weinzierl, P., Sporbert, A., Cremer, T., Zink, D., and Cardoso, M. C. (2000). Dynamics of DNA replication factories in living cells. *J. Cell Biol.*, 149(2):271–280.
- [Li et al., 2008] Li, K., Miller, E. D., Chen, M., Kanade, T., Weiss, L. E., and Campbell, P. G. (2008). Cell population tracking and lineage construction with spatiotemporal context. *Med Image Anal*, 12(5):546–566.
- [Liang et al., 2013] Liang, P., Qin, Y. Y., Zhang, C., Zhang, J., Cao, Y., Wu, S. C., Wong, C. K., and Wong, M. H. (2013). Plasma mercury levels in Hong Kong residents: in relation to fish consumption. *Sci. Total Environ.*, 463-464:1225–1229.
- [Loo et al., 2007] Loo, L. H., Wu, L. F., and Altschuler, S. J. (2007). Image-based multivariate profiling of drug responses from single cells. *Nat. Methods*, 4(5):445–453.
- [Lopez-Espinosa et al., 2008] Lopez-Espinosa, M. J., Kiviranta, H., Araque, P., Ruokojärvi, P., Molina-Molina, J. M., Fernandez, M. F., Vartiainen, T., and Olea, N. (2008). Dioxins in adipose tissue of women in Southern Spain. *Chemosphere*, 73(6):967–971.

- [Lou and Hamprecht, 2011] Lou, X. and Hamprecht, F. (2011). Structured Learning for Cell Tracking. In *NIPS 2011*.
- [Lou and Hamprecht, 2012] Lou, X. and Hamprecht, F. (2012). Structured Learning from Partial Annotations. In *ICML 2012*.
- [Maiuri et al., 2015] Maiuri, P., Rupprecht, J. F., Wieser, S., Ruprecht, V., Benichou, O., Carpi, N., Coppey, M., De Beco, S., Gov, N., Heisenberg, C. P., Lage Crespo, C., Lautenschlaeger, F., Le Berre, M., Lennon-Dumenil, A. M., Raab, M., Thiam, H. R., Piel, M., Sixt, M., and Voituriez, R. (2015). Actin flows mediate a universal coupling between cell speed and cell persistence. *Cell*, 161(2):374–386.
- [Maiuri et al., 2012] Maiuri, P., Terriac, E., Paul-Gilloteaux, P., Vignaud, T., McNally, K., Onuffer, J., Thorn, K., Nguyen, P. A., Georgoulia, N., Soong, D., Jayo, A., Beil, N., Beneke, J., Lim, J. C., Sim, C. P., Chu, Y. S., Jimenez-Dalmaroni, A., Joanny, J. F., Thiery, J. P., Erfle, H., Parsons, M., Mitchison, T. J., Lim, W. A., Lennon-Dumenil, A. M., Piel, M., and Thery, M. (2012). The first World Cell Race. *Curr. Biol.*, 22(17):R673–675.
- [Malarvannan et al., 2013] Malarvannan, G., Isobe, T., Covaci, A., Prudente, M., and Tanabe, S. (2013). Accumulation of brominated flame retardants and polychlorinated biphenyls in human breast milk and scalp hair from the Philippines: levels, distribution and profiles. *Sci. Total Environ.*, 442:366–379.
- [McLaughlin et al., 2010] McLaughlin, J., Markovtsov, V., Li, H., Wong, S., Gelman, M., Zhu, Y., Franci, C., Lang, D., Pali, E., Lasaga, J., Low, C., Zhao, F., Chang, B., Gururaja, T. L., Xu, W., Baluom, M., Sweeny, D., Carroll, D., Sran, A., Thota, S., Parmer, M., Romane, A., Clemens, G., Grossbard, E., Qu, K., Jenkins, Y., Kinoshita, T., Taylor, V., Holland, S. J., Argade, A., Singh, R., Pine, P., Payan, D. G., and Hitoshi, Y. (2010). Preclinical characterization of Aurora kinase inhibitor R763/AS703569 identified through an image-based phenotypic screen. *J. Cancer Res. Clin. Oncol.*, 136(1):99–113.
- [Medehouenou et al., 2011] Medehouenou, T. C., Ayotte, P., Carmichael, P. H., Kroger, E., Verreault, R., Lindsay, J., Dewailly, E., Tyas, S. L., Bureau, A., and Laurin, D. (2011). Polychlorinated biphenyls and organochlorine pesticides in plasma of older Canadians. *Environ. Res.*, 111(8):1313–1320.
- [Medehouenou et al., 2010] Medehouenou, T. C., Larochelle, C., Dumas, P., Dewailly, E., and Ayotte, P. (2010). Determinants of AhR-mediated transcriptional activity induced by plasma extracts from Nunavik Inuit adults. *Chemosphere*, 80(2):75–82.

- [Meijering et al., 2012] Meijering, E., Dzyubachyk, O., and Smal, I. (2012). *Chapter nine - Methods for Cell and Particle Tracking*, volume 504 of *Methods in Enzymology*, pages 183–200. Academic Press.
- [Moffat et al., 2006] Moffat, J., Grueneberg, D. A., Yang, X., Kim, S. Y., Kloepper, A. M., Hinkle, G., Piquani, B., Eisenhaure, T. M., Luo, B., Grenier, J. K., Carpenter, A. E., Foo, S. Y., Stewart, S. A., Stockwell, B. R., Hacohen, N., Hahn, W. C., Lander, E. S., Sabatini, D. M., and Root, D. E. (2006). A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, 124(6):1283–1298.
- [Mokhtari et al., 2013] Mokhtari, Z., Mech, F., Zitzmann, C., Hasenberg, M., Gunzer, M., and Figge, M. T. (2013). Automated characterization and parameter-free classification of cell tracks based on local migration behavior. *PLoS ONE*, 8(12):e80808.
- [Müllner, 2013] Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, 53(9):1–18.
- [Muzzey and van Oudenaarden, 2009] Muzzey, D. and van Oudenaarden, A. (2009). Quantitative time-lapse fluorescence microscopy in single cells. *Annu. Rev. Cell Dev. Biol.*, 25:301–327.
- [Naffar-Abu-Amara et al., 2008] Naffar-Abu-Amara, S., Shay, T., Galun, M., Cohen, N., Isakoff, S. J., Kam, Z., and Geiger, B. (2008). Identification of novel pro-migratory, cancer-associated genes using quantitative, microscopy-based screening. *PLoS ONE*, 3(1):e1457.
- [Neumann et al., 2010] Neumann, B., Walter, T., Heriche, J. K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., Cetin, C., Sieckmann, F., Pau, G., Kabbe, R., Wunsche, A., Satagopam, V., Schmitz, M. H., Chapuis, C., Gerlich, D. W., Schneider, R., Eils, R., Huber, W., Peters, J. M., Hyman, A. A., Durbin, R., Pepperkok, R., and Ellenberg, J. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–727.
- [Ohnuki et al., 2010] Ohnuki, S., Oka, S., Nogami, S., and Ohya, Y. (2010). High-content, image-based screening for drug targets in yeast. *PLoS ONE*, 5(4):e10177.
- [Oikawa et al., 2001] Oikawa, K., Ohbayashi, T., Mimura, J., Iwata, R., Kameta, A., Evine, K., Iwaya, K., Fujii-Kuriyama, Y., Kuroda, M., and Mukai, K. (2001). Dioxin suppresses the checkpoint protein, MAD2, by an aryl hydrocarbon receptor-independent pathway. *Cancer Res.*, 61(15):5707–5709.
- [Oikawa et al., 2008] Oikawa, K., Yoshida, K., Takanashi, M., Tanabe, H., Kiyuna, T., Ogura, M., Saito, A., Umezawa, A., and Kuroda, M. (2008). Dioxin interferes in

- chromosomal positioning through the aryl hydrocarbon receptor. *Biochem. Biophys. Res. Commun.*, 374(2):361–364.
- [Oprea-Ilies et al., 2013] Oprea-Ilies, G., Haus, E., Sackett-Lundeen, L., Liu, Y., McLendon, L., Busch, R., Adams, A., and Cohen, C. (2013). Expression of melatonin receptors in triple negative breast cancer (TNBC) in African American and Caucasian women: relation to survival. *Breast Cancer Res. Treat.*, 137(3):677–687.
- [Ostaszewski et al., 2012] Ostaszewski, M., Eifes, S., and del Sol, A. (2012). Evolutionary conservation and network structure characterize genes of phenotypic relevance for mitosis in human. *PloS one*, 7(5):e36488–e36488.
- [Ostling and Johanson, 1984] Ostling, O. and Johanson, K. J. (1984). Microelectrophoretic study of radiation-induced DNA damages in individual mammalian cells. *Biochem. Biophys. Res. Commun.*, 123(1):291–298.
- [Palmari et al., 2000] Palmari, J., Wallet, F., Berard, J., Berthois, Y., Martin, P. M., and Dussert, C. (2000). Morphological evidence for a subpopulation selection effect by estrogen and antiestrogen treatments in the heterogeneous MCF-7 cell line. *Anal Cell Pathol*, 20(2-3):99–113.
- [Parsons et al., 2006] Parsons, A. B., Lopez, A., Givoni, I. E., Williams, D. E., Gray, C. A., Porter, J., Chua, G., Sopko, R., Brost, R. L., Ho, C. H., Wang, J., Ketela, T., Brenner, C., Brill, J. A., Fernandez, G. E., Lorenz, T. C., Payne, G. S., Ishihara, S., Ohya, Y., Andrews, B., Hughes, T. R., Frey, B. J., Graham, T. R., Andersen, R. J., and Boone, C. (2006). Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell*, 126(3):611–625.
- [Pau et al., 2013] Pau, G., Walter, T., Neumann, B., Heriche, J. K., Ellenberg, J., and Huber, W. (2013). Dynamical modelling of phenotypes in a genome-wide RNAi live-cell imaging assay. *BMC Bioinformatics*, 14:308.
- [Peng, 2008] Peng, H. (2008). Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17):1827–1836.
- [Perlman et al., 2004] Perlman, Z. E., Slack, M. D., Feng, Y., Mitchison, T. J., Wu, L. F., and Altschuler, S. J. (2004). Multidimensional drug profiling by automated microscopy. *Science*, 306(5699):1194–1198.
- [Peyre et al., 2014] Peyre, L., Zucchini-Pascal, N., de Sousa, G., Luzy, A. P., and Rahmani, R. (2014). Potential involvement of chemicals in liver cancer progression: an alternative toxicological approach combining biomarkers and innovative technologies. *Toxicol In Vitro*, 28(8):1507–1520.

- [Polder et al., 2008] Polder, A., Gabrielsen, G. W., Odland, J. ., Savinova, T. N., Tkachev, A., Løken, K. B., and Skaare, J. U. (2008). Spatial and temporal changes of chlorinated pesticides, PCBs, dioxins (PCDDs/PCDFs) and brominated flame retardants in human breast milk from Northern Russia. *Sci. Total Environ.*, 391(1):41–54.
- [Pruitt et al., 2014] Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O’Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D., and Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, 42(Database issue):D756–763.
- [Raj and van Oudenaarden, 2008] Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226.
- [Rajaram et al., 2012] Rajaram, S., Pavie, B., Wu, L. F., and Altschuler, S. J. (2012). PhenoRipper: software for rapidly profiling microscopy images. *Nat. Methods*, 9(7):635–637.
- [Rawn et al., 2012] Rawn, D. F., Ryan, J. J., Sadler, A. R., Sun, W. F., Haines, D., Macey, K., and Van Oostdam, J. (2012). PCDD/F and PCB concentrations in sera from the Canadian Health Measures Survey (CHMS) from 2007 to 2009. *Environ Int*, 47:48–55.
- [Sbalzarini and Koumoutsakos, 2005] Sbalzarini, I. F. and Koumoutsakos, P. (2005). Feature point tracking and trajectory analysis for video imaging in cell biology. *J. Struct. Biol.*, 151(2):182–195.
- [Schenone et al., 2013] Schenone, M., Dan?ik, V., Wagner, B. K., and Clemons, P. A. (2013). Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.*, 9(4):232–240.
- [Schirle and Jenkins, 2015] Schirle, M. and Jenkins, J. L. (2015). Identifying compound efficacy targets in phenotypic drug discovery. *Drug Discov. Today*.
- [Schoenauer Sebag et al., 2015] Schoenauer Sebag, A., Plancade, S., Raulet-Tomkiewicz, C., Barouki, R., Vert, J. P., and Walter, T. (2015). A generic methodological framework for studying single cell motility in high-throughput time-lapse data. *Bioinformatics*, 31(12):i320–i328.
- [Simpson et al., 2012] Simpson, J. C., Joggerst, B., Laketa, V., Verissimo, F., Cetin, C., Erfle, H., Bexiga, M. G., Singan, V. R., Hériché, J.-K., Neumann, B., et al. (2012).

- Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. *Nature cell biology*, 14(7):764–774.
- [Simpson et al., 2008] Simpson, K. J., Selfors, L. M., Bui, J., Reynolds, A., Leake, D., Khvorova, A., and Brugge, J. S. (2008). Identification of genes that regulate epithelial cell migration using an siRNA screening approach. *Nat. Cell Biol.*, 10(9):1027–1038.
- [Singh et al., 2010] Singh, D. K., Ku, C. J., Wichaidit, C., Steininger, R. J., Wu, L. F., and Altschuler, S. J. (2010). Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol. Syst. Biol.*, 6:369.
- [Sinkhorn and Knopp, 1967] Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math*, 21(2):343–348.
- [Snijder et al., 2009] Snijder, B., Sacher, R., Ramo, P., Damm, E. M., Liberali, P., and Pelkmans, L. (2009). Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, 461(7263):520–523.
- [Soille, 2003] Soille, P. (2003). *Morphological Image Analysis: Principles and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2 edition.
- [Sommer et al., 2013] Sommer, C., Held, M., Fischer, B., Huber, W., and Gerlich, D. W. (2013). CellH5: a format for data exchange in high-content screening. *Bioinformatics*, 29(12):1580–1582.
- [Song et al., 2013] Song, W., Zhao, C., and Jiang, R. (2013). Integrin-linked kinase silencing induces a S/G2/M phases cell cycle slowing and modulates metastasis-related genes in SGC7901 human gastric carcinoma cells. *Tumori*, 99(2):249–256.
- [Sulston et al., 1983] Sulston, J. E., Schierenberg, E., White, J. G., and Thomson, J. N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.*, 100(1):64–119.
- [Suratane et al., 2014] Suratane, A., Schaefer, M. H., Betts, M. J., Soons, Z., Mannsperger, H., Harder, N., Oswald, M., Gipp, M., Ramminger, E., Marcus, G., Manner, R., Rohr, K., Wanker, E., Russell, R. B., Andrade-Navarro, M. A., Eils, R., and König, R. (2014). Characterizing protein interactions employing a genome-wide siRNA cellular phenotyping screen. *PLoS Comput. Biol.*, 10(9):e1003814.
- [Tan et al., 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

- [Timm et al., 2013] Timm, D. M., Chen, J., Sing, D., Gage, J. A., Haisler, W. L., Neeley, S. K., Raphael, R. M., Dehghani, M., Rosenblatt, K. P., Killian, T. C., Tseng, H., and Souza, G. R. (2013). A high-throughput three-dimensional cell migration assay for toxicity screening with mobile device-based macroscopic image analysis. *Sci Rep*, 3:3000.
- [Tsochantaridis et al., 2005] Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)*, 6:1453–1484.
- [Valera et al., 2013] Valera, B., Dewailly, E., and Poirier, P. (2013). Association between methylmercury and cardiovascular risk factors in a native population of Quebec (Canada): a retrospective evaluation. *Environ. Res.*, 120:102–108.
- [van Roosmalen et al., 2015] van Roosmalen, W., Le Devedec, S. E., Golani, O., Smid, M., Pulyakhina, I., Timmermans, A. M., Look, M. P., Zi, D., Pont, C., de Graauw, M., Naffar-Abu-Amara, S., Kirsanova, C., Rustici, G., Hoen, P. A., Martens, J. W., Foekens, J. A., Geiger, B., and van de Water, B. (2015). Tumor cell migration screen identifies SRPK1 as breast cancer metastasis determinant. *J. Clin. Invest.*, 125(4):1648–1664.
- [Vandenberg et al., 2009] Vandenberg, L. N., Maffini, M. V., Sonnenschein, C., Rubin, B. S., and Soto, A. M. (2009). Bisphenol-A and the great divide: a review of controversies in the field of endocrine disruption. *Endocr. Rev.*, 30(1):75–95.
- [Vecchio et al., 2014] Vecchio, G., Fenech, M., Pompa, P. P., and Voelcker, N. H. (2014). Lab-on-a-chip-based high-throughput screening of the genotoxicity of engineered nanomaterials. *Small*, 10(13):2721–2734.
- [Villani, 2009] Villani, C. (2009). *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin.
- [Wählby et al., 2002] Wählby, C., Lindblad, J., Vondrus, M., Bengtsson, E., and Björkesten, L. (2002). Algorithms for cytoplasm segmentation of fluorescence labelled cells. *Analytical cellular pathology : the journal of the European Society for Analytical Cellular Pathology*, 24(2-3):101–11.
- [Walter et al., 2010] Walter, T., Held, M., Neumann, B., Heriche, J. K., Conrad, C., Pepperkok, R., and Ellenberg, J. (2010). Automatic identification and clustering of chromosome phenotypes in a genome wide RNAi screen by time-lapse imaging. *J. Struct. Biol.*, 170(1):1–9.
- [Wei et al., 2012] Wei, X., Hoffman, A. F., Hamilton, S. M., Xiang, Q., He, Y., So, W. V., So, S. S., and Mark, D. (2012). A simple statistical test to infer the causality of

- target/phenotype correlation from small molecule phenotypic screens. *Bioinformatics*, 28(3):301–305.
- [Wishart et al., 2008] Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Database issue):D901–906.
- [Wittsiepe et al., 2007] Wittsiepe, J., Furst, P., Schrey, P., Lemm, F., Kraft, M., Eberwein, G., Winneke, G., and Wilhelm, M. (2007). PCDD/F and dioxin-like PCB in human blood and milk from German mothers. *Chemosphere*, 67(9):S286–294.
- [Wlodkowic et al., 2011] Wlodkowic, D., Faley, S., Darzynkiewicz, Z., and Cooper, J. M. (2011). Real-time cytotoxicity assays. *Methods Mol. Biol.*, 731:285–291.
- [Wong et al., 2014] Wong, I. Y., Javaid, S., Wong, E. A., Perk, S., Haber, D. A., Toner, M., and Irimia, D. (2014). Collective and individual migration following the epithelial-mesenchymal transition. *Nat Mater*, 13(11):1063–1071.
- [Wruck et al., 2014] Wruck, W., Peucker, M., and Regenbrecht, C. R. (2014). Data management strategies for multinational large-scale systems biology projects. *Brief. Bioinformatics*, 15(1):65–78.
- [Yang et al., 2013] Yang, J., Fan, J., Li, Y., Li, F., Chen, P., Fan, Y., Xia, X., and Wong, S. T. (2013). Genome-wide RNAi screening identifies genes inhibiting the migration of glioblastoma cells. *PLoS ONE*, 8(4):e61915.
- [Yilmaz and Christofori, 2010] Yilmaz, M. and Christofori, G. (2010). Mechanisms of motility in metastasizing cells. *Mol. Cancer Res.*, 8(5):629–642.
- [Young et al., 2008] Young, D. W., Bender, A., Hoyt, J., McWhinnie, E., Chirn, G. W., Tao, C. Y., Tallarico, J. A., Labow, M., Jenkins, J. L., Mitchison, T. J., and Feng, Y. (2008). Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, 4(1):59–68.
- [Zhang et al., 2013] Zhang, H., Wu, P. Y., Ma, M., Ye, Y., Hao, Y., Yang, J., Yin, S., Sun, C., Phan, J. H., Wang, M. D., and Xi, J. J. (2013). An integrative approach for the large-scale identification of human genome kinases regulating cancer metastasis. *Nanomedicine*, 9(6):732–736.
- [Zimmer et al., 2014] Zimmer, B., Pallocca, G., Dreser, N., Foerster, S., Waldmann, T., Westerhout, J., Julien, S., Krause, K. H., van Thriel, C., Hengstler, J. G., Sachinidis, A., Bosgra, S., and Leist, M. (2014). Profiling of drugs and environmental chemicals

for functional impairment of neural crest migration in a novel stem cell-based test battery. *Arch. Toxicol.*, 88(5):1109–1126.

- [Zimmer et al., 2002] Zimmer, C., Labruyere, E., Meas-Yedid, V., Guillen, N., and Olivo-Marin, J. C. (2002). Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: a tool for cell-based drug testing. *IEEE Trans Med Imaging*, 21(10):1212–1221.

Développements méthodologiques pour données de cribles temporels à haut contenu et haut débit

Résumé : Un crible biologique a pour objectif de tester en parallèle l'impact de nombreuses conditions expérimentales sur un processus biologique d'un organisme modèle. L'imagerie sur cellules vivantes est un excellent outil pour étudier en détail les conséquences d'une perturbation chimique sur un processus biologique. L'analyse des cribles sur cellules vivantes demande toutefois la combinaison de méthodes robustes d'imagerie par ordinateur et de contrôle qualité, et d'approches statistiques efficaces pour la détection des effets significatifs. La présente thèse répond à ces défis par le développement de méthodes analytiques pour les images de cribles temporels à haut débit. Les cadres qui y sont développés sont appliqués à des données publiées, démontrant par là leur applicabilité ainsi que les bénéfices d'une ré-analyse des données de cribles à haut contenu (HCS). Le premier workflow pour l'étude de la motilité cellulaire à l'échelle d'une cellule dans de telles données constitue le chapitre 2. Le chapitre 3 applique ce workflow à des données publiées et présente une nouvelle distance pour l'inférence de cible thérapeutique à partir d'images de cribles temporels. Enfin, le chapitre 4 présente une pipeline méthodologique complète pour la conduite de cribles temporels à haut débit en toxicologie environnementale.

Mots clefs : Fouille de données, Apprentissage statistique, Bioinformatique, Informatique de l'image biologique, Cribles à haut contenu, Toxicologie

The versatility of HC HT time-lapse screening data

Abstract: Biological screens test large sets of experimental conditions with respect to their specific biological effect on living systems. Live cell imaging is an excellent tool to study in detail the consequences of chemical perturbation on a given biological process. However, the analysis of live cell screens demands the combination of robust computer vision methods and quality control procedures, and efficient statistical approaches for the detection of significant effects. This thesis addresses these challenges by developing analytical methods for High Throughput time-lapse microscopy screening data. The developed frameworks are applied to publicly available HCS data, demonstrating their applicability and the benefits of HCS data remining. The first multivariate workflow for the study of single cell motility in such large-scale data is detailed in Chapter 2. Chapter 3 presents this workflow application to previously published data, and the development of a new distance for drug target inference by in silico comparisons of parallel siRNA and drug screens. Finally, chapter 4 presents a complete methodological pipeline for performing HT time-lapse screens in Environmental Toxicology.

Keywords: Data mining, Machine Learning, Bioinformatics, Bioimage informatics, High-content screening, Toxicology

