Computational and Mathematical Modelling of DNA Repair in Budding Yeast

Leo Zeitler

université PARIS-SACLAY

Computational and Mathematical Modelling of DNA Repair in Budding Yeast

Intégration computationnelle et modélisation de la cinétique de réparation de l'ADN chez la levure bourgeonnante

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577 : structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat : Génétique
Graduate School : Life Sciences and Health. Référent : Faculté des Sciences d'Orsay

Thèse préparée dans l'unité de recherche **Institute for Integrative Biology of the Cell (I2BC) (Université Paris-Saclay, CEA, CNRS)**, sous la direction de **Julie SOUTOURINA**, Directrice de recherche, et le co-encadrement de **Cyril DENBY WILKES**, Chargé de recherche

Thèse soutenue à Paris-Saclay, le 21 septembre 2023, par

Leo ZEITLER

Composition du jury

Membres du jury avec voix délibérative

Karine DUBRANA Directrice de recherche, CEA - Université de Paris et Université Paris-Saclay, France	Présidente
Benjamin AUDIT Directeur de recherche, CNRS - ENS Lyon, Lyon, France	Rapporteur & Examinateur
Zhou XU Directeur de recherche, CNRS - Sorbonne Université, Paris, France	Rapporteur & Examinateur
Erik FRANSÉN Professeur, KTH, Stockholm, Sweden	Examinateur

THESE DE DOCTORAT
NNT : 2023UPASL077

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Intégration Computationnelle et Modélisation de la Cinétique de Réparation de l'ADN Chez la Levure Bourgeonnante

Mots clés : Réparation de l'ADN ; données NGS ; analyse pilotée par les données ; modèle computationnel

Résumé : L'interaction moléculaire entre les protéines et l'ADN régit le comportement de la cellule dans le contexte de son environnement. Le maintien de l'intégrité du génome est donc vital pour le fonctionnement normal et la survie des cellules. L'altération de la réparation de l'ADN a été associée à des maladies graves. Malgré des études intensives, la façon dont la réparation de l'ADN est orchestrée *in vivo* à l'échelle du génome reste mal comprise. Nous avons développé des outils informatiques pour une évaluation complète de la cinétique de réparation de l'ADN locus-spécifique après une irradiation UV en utilisant les données de séquençage à haut débit (NGS). Nous avons analysé les facteurs susceptibles d'influencer la réparation, tels que la disposition des nucléosomes, les niveaux de transcription et la taille des gènes. En tirant parti d'une interprétation différente des données, notre modèle minimal peut récupérer les informations manquantes pour étudier la réparation continue dans le temps à l'aide de points de données NGS peu nombreux. Contrairement à d'autres études qui considèrent les signaux de séquençage comme un comportement moyen, nous les prenons en compte comme la superposition d'interactions stochastiques ADN-protéine dans des cellules indépendantes. Cela a permis d'analyser la réparation de l'ADN dans le contexte d'autres processus nucléaires, tels que la transcription et le positionnement des nucléosomes. Cependant, pour une véritable compréhension du processus, il est nécessaire de combiner une analyse basée sur les données avec une modélisation mathématique. Nous avons développé et comparé deux approches—notamment une approximation *mean-field* et une méthode stochastique spécifique aux cellules—qui relient la dynamique de la cellule à des données NGS à l'échelle d'une population. Nos méthodes indiquent la cinétique de réparation spécifique aux gènes et permettent de comprendre le mécanisme de reconnaissance des dommages le long des régions codantes. Les deux modèles sont basés sur des interactions générales entre l'ADN et les protéines et peuvent être facilement appliqués à d'autres processus nucléaires. Ces travaux constituent un maillon manquant entre la dynamique temporelle interne aux cellules vivantes et le comportement à l'échelle de la population qui peut être mesuré.

Title : Computational and Mathematical Modelling of DNA Repair in Budding Yeast

Keywords : DNA repair ; NGS data ; data-driven analysis ; computational model

Abstract : The molecular interplay of proteins with the DNA governs cell behaviour in the context of its environment. Maintenance of genome integrity is therefore vital for normal cell functionality and survival. Impaired DNA repair has been associated with severe diseases. This includes cancer as well as neurological and premature-ageing disorders. Despite intensive studies, it remains poorly understood how DNA repair is orchestrated *in vivo* on a genomic scale. We developed computational tools for a comprehensive assessment of location-specific DNA repair kinetics after UV irradiation using Next Generation Sequencing (NGS) signals. We analysed possible repair influencing factors, such as nucleosome arrangement, transcription levels, and gene size. By leveraging a different interpretation of the data, our minimal model can recover missing information to study time-continuous repair using sparse NGS data points. In contrast to other studies that consider sequencing signals as an average behaviour, we understand them as the superposition of stochastic DNA-protein interactions in independent cells. This permitted the analysis of DNA repair in context of other nuclear processes, such as transcription and nucleosome positioning. However, for a true understanding of the process, it is necessary to combine a top-down data-driven analysis with bottom-up mathematical modelling. We developed and compared two approaches—namely a mean-field approximation and a cell-specific stochastic sampling method—that link single-cell dynamics with population-wide NGS data. Amazingly, both models indicate gene-specific repair kinetics, and they provide a mechanistic perspective of the damage recognition along coding regions. As they are based on general DNA-protein interactions, they can be readily applied to other nuclear processes. The work provides a missing link between cell-internal temporal dynamics in living cells and population-wide behaviour that can be measured.

To my parents and my sister.

Acknowledgements

La thèse dure trois ans [1]. And I was particularly lucky during these three years. I owe a debt of thanks to all the people who incredibly supported me during this time both professionally and privately. At the top of my list I want to thank Julie Soutourina, Cyril Denby Wilkes, and Arach Goldar. Their guidance, feedback, and patience gave me not only the necessary encouragement to develop my research projects, but also helped me to sort out the French bureaucracy and allowed me to frantically flee to England before the Covid lockdown. Their trust and support shaped me strongly. The work presented here is as much their achievement as mine. My thanks naturally extend to the rest of the team and their ex-members, in particular Adriana Alberti, Kévin André, Veronica Martinez-Fernandez, and Nathalie Giordanengo Aiach.

I also want to thank the jury and committee members Karine Dubrana, Benjamin Audit, Zhou Xu, and Erik Fransén for their time and commitment as well as for their comments and feedback.

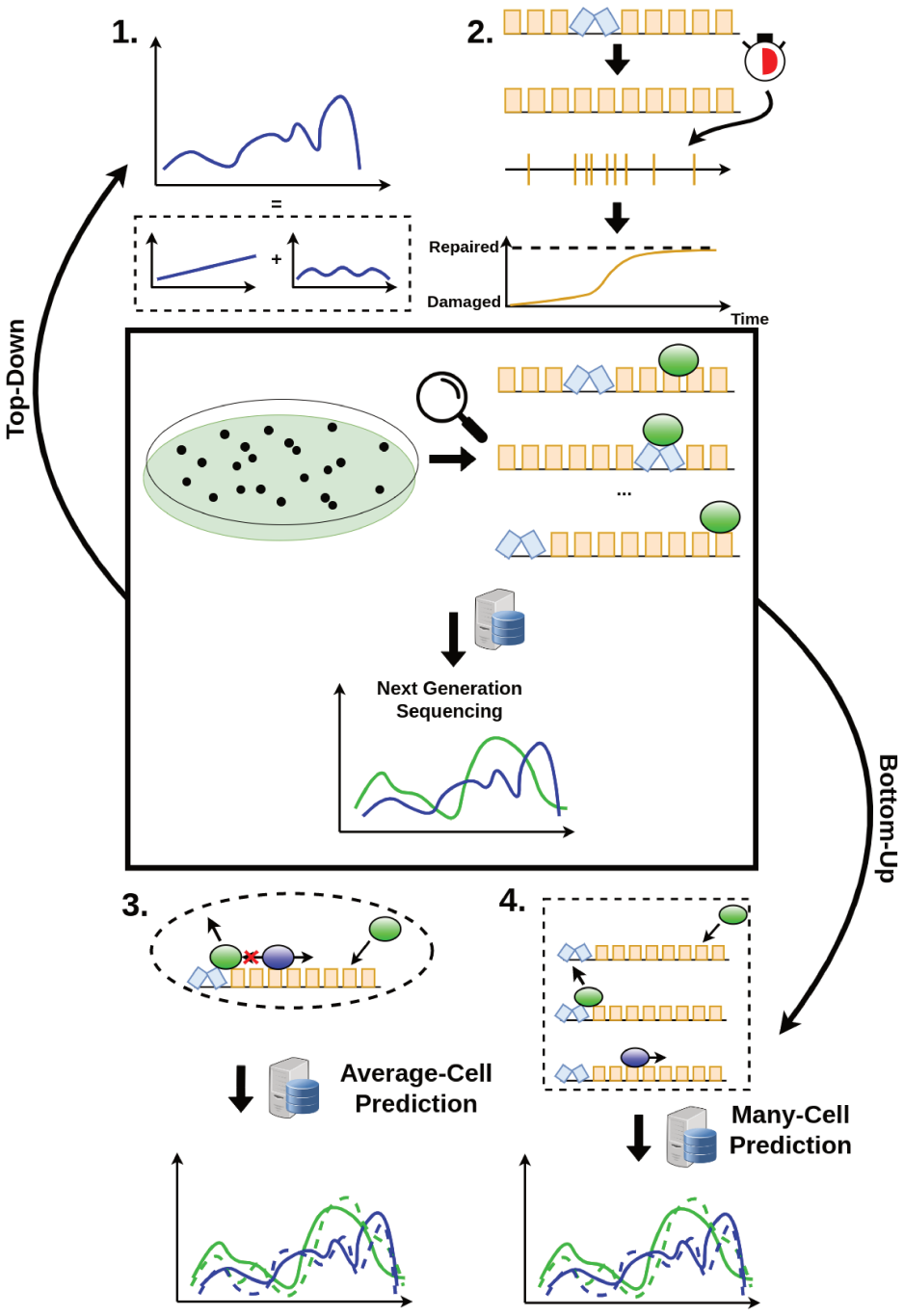
This work would not have been possible without the emotional support from Fiona Carter. The distractions during the long lockdowns, the trips together through all of Europe, and the late night calls allowed me not to worry about my personal life and to fully focus on my research work instead.

When I arrived in France, I was additionally faced with the challenge of finding friends to hang out with and to establish my social network. I owe an immense thank you to Julie Grandsire, Rhea Khoury Helou, and Andrea Porro who showed me around and welcomed me into their heartwarming friendship groups.

And, naturally, I would not be here without the unbelievable guidance and support from my family, Karin, Andreas, and Lena Zeitler. Thank you so much for making me who I am now. This PhD thesis is dedicated to you.

[1] Reference to *L'amour dure trois ans* by Frédéric Beigbeder, 1997

Graphical Abstract



Caption: The molecular interplay of proteins with the DNA governs cell behaviour in the context of its environment. Maintenance of genome integrity is therefore vital for normal cell functionality and survival. Impaired DNA repair has been associated with severe diseases. This includes cancer as well as neurological and premature-ageing disorders. Despite intensive studies, it remains poorly understood how DNA repair is orchestrated *in vivo* on a genomic scale. We developed computational tools for a comprehensive assessment of location-specific DNA repair kinetics after UV irradiation using Next Generation Sequencing (NGS) signals. In this PhD manuscript, we describe two top-down data analyses and two bottom-up mathematical modelling approaches. By evaluating the functional composition of data that explains the positioning of nucleosomes (1.) as well as analysing the repair transition over time in context of other genomic properties (2.), we determined repair-influencing factors that are supposedly important for lesion removal along protein-coding genes. In order to evaluate our hypotheses, we developed a mean-field approach (3.) and a cell-dependent stochastic sampling algorithm (4.) to link single-cell DNA-protein interactions to population-based data. Amazingly, both of our models predict gene-specific repair as well as an interaction between different damage recognition pathways. In this work, we present a new perspective on DNA repair by interfacing biological experiments with computational models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

What is our identity? What links us to other human beings or living organisms? It has been a central endeavour of western philosophy to address mankind's place in nature. With the development of a modern scientific methodology, research has found its own ways to contribute to the understanding of our own kind.

Deoxyribonucleic acid (DNA) is the molecule that encodes the hereditary information of all living organisms. Long before its discovery, Greek philosophers such as Aristotle and Pythagoras already hypothesised that parental information is transferred to the offspring, despite being far off the truth (Mukherjee (2016)). The methodological study of genes had its advent with the hereditary laws formulated by Gregor Mendel in 1865 (Mendel (1865, 1996)). However, this alone could not provide a mechanistic explanation. By purifying the DNA molecule itself, Oswald Avery showed that genes were in fact carried on a chemical (Avery et al. (1944)). Shortly after, in 1953, Watson and Crick discovered and characterised the actual molecular structure of the DNA (Watson and Crick (1953)).

Genetic and genomic research have shaped science and society alike, opening the doors to vast opportunities in medicine but also revealing their perils. The treatment development is foremost dependent on discoveries in fundamental research. For example, recombinant DNA technology (which was largely pioneered by Paul Berg for the tumor virus SV40 (Jackson et al. (1972))) came as remedy to produce clean concentrations of Factor VIII (FVIII)—a blood clotting factor that is dysfunctional in hemophilia patients—during the HIV crisis. In 1987, a hemophilia patient got successfully treated for the first time with synthesised FVIII from plasmids introduced into hamster ovary cells—without the risk of containing blood-borne pathogens (Mukherjee (2016)). A further development in medicine is giving hope to many patients with inheritable and often incurable diseases: gene therapy. A four-year-old girl with severe combined immunodeficiency (SCID) got treated with the first approved gene therapy in 1990 (Anderson (1990); Scheller and Krebsbach (2009)). Whilst the pro-

cedure was successful, other examples of premature applications led to prominent and tragic deaths. Jesse Gelsinger passed away in 1999 after the administration of an understudied gene therapy to replace the mutated ornithine-transcarbamylase (OTC) gene sequence (Sibbald (2001)). Despite many remaining unknowns, genetics bears undoubtedly a great potential to unlock so far unknown key functionalities that can be harnessed to develop novel treatments against various diseases. This trend has become increasingly clear over the last years. Eight gene therapies received approval by the Food and Drug Administration (FDA) in 2021, and there were more than 1300 under development in 2020 (Whittal et al. (2022)).

In order to harness biological mechanisms for the development of new treatments, it is indispensable to study the elemental molecular interactions that lead to their regulation in living cells. The interplay of proteins and RNA with the DNA regulate and affect all fundamental processes in different contexts (Cozzolino et al. (2021)). Due to the vast complexity, there remain uncountable questions to be answered. A matter that has been under intense study in recent years is DNA repair. It is a known fact that the physical composition of DNA is constantly changed by a variety of external and internal factors. Environmental agents like smoking (Swenberg et al. (2011); Yamaguchi (2019)), drinking (Brooks (1997)), and UltraViolet (UV) light (Rastogi et al. (2010); Mao et al. (2016); Hu et al. (2017)), but equally cell-internal metabolism can cause between 10,000 to 100,000 DNA distortions per cell per day in the human body (Marteijn et al. (2014); Swenberg et al. (2011)). It is therefore indispensable for cell survival to possess various mechanisms to repair molecular alterations and to maintain DNA integrity. The large number of genotoxic factors caused the development of several DNA repair pathways in nature, among others Nucleotide Excision Repair (NER). NER is an evolutionarily conserved pathway that can be found in almost all eukaryotes, including human cells (Reardon and Sancar (2005); Zhang et al. (2022)) and budding yeast (*Saccharomyces cerevisiae*). It is characterised by its exceptional ability to remove numerous lesion types—inter alia UV-induced damages such as Cyclobutane Pyrimidine Dimers (CPDs) and 6-4 Photoproducts (6-4PPs)—but also bulky chemical adducts, and cyclopurines that were generated by Reactive Oxygen Species (ROS) (Marteijn et al. (2014)). NER exhibits region-specific properties, which explains the conventional differentiation between Global-Genome Repair (GGR), which can be observed along the entire genome; and Transcription Coupled Repair (TCR), which is limited to genes that are actively transcribed by the multiprotein complex RNA Polymerase II (Pol II). The stalling of Pol II at Transcription Blocking Lesions (TBLs) initiates the recruitment of other NER proteins (Deaconescu et al. (2006)). The two different detection pathways converge subsequently to the same incision and replacement mechanism.

Despite laying the fundamental groundwork, biological experiments quickly reach their limits to

study an intricate process such as DNA repair. In fact, the genome-wide organisation of NER *in vivo* remains in the dark thus far, and it is unclear how repair dynamics are coordinated in context of other nuclear processes, such as transcription and chromatin folding. It is therefore necessary to combine location-specific DNA-protein interaction data—such as Next Generation Sequencing (NGS) data—with computational models to understand the nuclear kinetics in a fully controlled environment. Nevertheless, despite the clear need of interfacing computational and experimental methods, the number of modelling approaches for repair kinetics remain low. In this thesis work, we present top-down data analysis approaches and bottom-up mathematical models to explain DNA repair in the yeast species *Saccharomyces cerevisiae* as a model organism for lesion removal in human cells. The chapter starts with introducing some few conventions and abbreviations that are used throughout the manuscript. This is followed by presenting the biological and computational theory on which the thesis work is based. We close by motivating this work in a wider context of human diseases.

1.1 Conventions

We make use of the following conventions and notations. Standard laboratory strains or those whose phenotype are indifferent in a given context are called wildtype (WT) strains. Gene names in WT cells are written in *ITALIC* capitals. When referring to a mutated gene, we write the name in lower case italic, e.g. *rad7*. Gene deletions are given in lower case italic followed by a Delta symbol Δ, e.g. *rad7*Δ. Protein names are written in normal font, e.g. Rad7.

We make use of many mathematical notations. Bold symbols (e.g. $\mathbf{x}$ or $\boldsymbol{\mu}$) refer to vectors. Capital bold letters are vector functions (such as matrices), e.g. $\mathbf{W}$. If the same symbol appears not marked in bold, we refer to scalar values within the vector function or vector (e.g. W_{ij}). Throughout the manuscript, we make use of column vectors. We write sometimes $\partial_t x$ to represent $\frac{\partial x}{\partial t}$.

All abbreviations are introduced with its full explanation when used the first time. We provide a list of all acronyms in Tables 6.1 and 6.2.

1.2 Biological Background

All living beings need to react to changing environmental conditions. A plethora of cellular and nuclear mechanisms need to work in sync to allow an adequately adapted behaviour. It is therefore impossible to study NER dynamics isolated from its contextual setup. Various mechanisms—such as other DNA repair pathways, transcription, and DNA packaging—influence availability of proteins as well as accessibility to the lesion. To make matters more difficult: drastic changes in the environment—such

as an irradiation event that induces DNA damage—can evoke a stress response, during which many processes are strongly regulated. It remains largely unknown how stress response and DNA repair interact to ensure cell survival.

In the following, we will first present genome and gene organisation in *Saccharomyces cerevisiae* (Subsection 1.2.1). This allows the introduction of the location-dependent NER subpathways (Subsection 1.2.2). To grasp the complexity of repair in context of other nuclear processes, we explain the dynamics of other DNA repair mechanisms, transcription, and nucleosome positioning with respect to NER (Subsection 1.2.3). The section is closed with a description of the acquisition and treatment of NGS data (Subsection 1.2.4).

1.2.1 Genome and Gene Organisation

Budding yeast is a eukaryotic organism, which is characterised by the presence of membrane-bound organelles. This includes the cell nucleus, in which the DNA is spatially confined together with various proteins that orchestrate genomic integrity, maintenance, and usage. The limited space introduces a packaging problem which requires the folding and twisting of the DNA molecule (see Subsection 1.2.3). The DNA itself is a polymer structured into two separate strands which are coiled to a double-helix conformation. The opposing strands are connected by hydrogen bonds. Genetic information is represented by a sequence of the four nucleotides (nt) Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). Cytosine and thymine are pyrimidine bases, whereas adenine and guanine are purines. Pairing of the opposing strands through the hydrogen bonds follows the strict pattern A-T and C-G. The conformational integrity is provided by a sugar-phosphate backbone (Watson and Crick (1953)). The strand which has its 5'-end at the telomere of the shorter arm is called *Watson* (or plus) strand, and the other is called *Crick* (or minus) strand. Commonly, the word *genome* describes the linear information contained in the DNA.

In contrast to the genome, it is far more difficult to define a *gene*, especially in the context of human cells. In this work, we denote by the word *gene* a transcribed DNA region together with its regulatory sequences within the DNA. The information in the sequence itself determines the gene function. Stretches between genes are called intergenic regions. We distinguish between non-coding (which produce non-coding (nc)RNA when expressed); and coding genes (which result in messenger (m)RNA). The latter is subsequently translated to a protein to provide a specified cell functionality. The transcribed information is encoded on only one of the two strands in 5' to 3' direction (which is also referred to as Open Reading Frame (ORF))(Shafee and Lowe (2017)). A schematic representation of a gene is given in Figure 1.1.

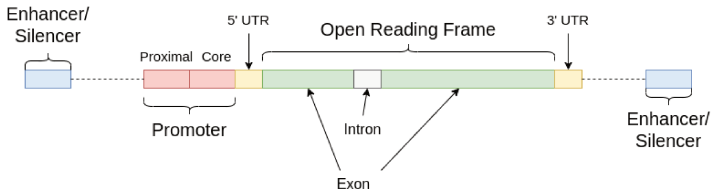


Figure 1.1: **Schematic representation of the gene organisation in a eukaryotic cell.** Boxes represented different sequences being part of gene expression, whereas black dotted lines indicate arbitrary stretches in between. The figure was taken and modified from Shafee and Lowe (2017).

The expression of genes is commonly governed by regulatory sequences. There is the distinction between promoters, enhancers and silencers. Promoters are subdivided into core promoter and proximal promoter sequences, which are located upstream of the transcribed region. The core promoter marks the starting position of gene expression by an RNA polymerase, whereas the proximal promoter region regulates transcription by binding Transcription Factors (TFs) that permit stable association of the RNA polymerase (Thomas and Chiang (2006)). Many core promoters contain a TATA-box, which is a *cis*-regulatory element. It is characterised by its repeating T and A base pairs, and it plays a fundamental role during transcription initiation (Watson (2014)). Enhancers and silencers are (commonly short) stretches of DNA that can increase and decrease gene activity, respectively. They can be positioned anywhere in the genome, independently of direction and location of the gene they are regulating. However, they are commonly found upstream in the yeast genome. One single enhancer or silencer can influence the expression of several genes. They can interact with promoters through TF loading. A DNA bending protein permits transient contact, a process which is also called DNA looping. Proteins that are responsible for regulation can compete in order to promote or inhibit expression. Transcription itself is introduced in Subsection 1.2.3.

Within the gene sequence, there are UnTranslated Regions (UTRs) on either side of the transcript with distinct responsibilities. The 3' UTR contains a termination codon which marks the dissociation of the polymerase. The 5' UTR, on the other hand, permits ribosome binding for commencing protein translation (Shafee and Lowe (2017)). Eukaryotic gene sequences can be divided into introns and exons, although intron sequences are rare in *Saccharomyces cerevisiae*. They are post-transcriptionally removed from the produced RNA and allow an evolutionarily accelerated way of recombining sequences to new genes (Gilbert (1978)).

Due to the rare occurrence of introns, we did not incorporate them specifically into our models. Similarly, enhancers and silencers were not considered as influential—as they contact the promoter region only transitively—and their effect on repair was not taken into account. Since UTRs are rather involved in post-transcriptional processes, they were not considered in this work. More sequence-dependent regions that are closely related to transcription are introduced in Subsection 1.2.3.

1.2.2 Nucleotide Excision Repair

Every cell possesses a number of different repair dynamics to undo molecular disruptions of the DNA. NER can remove various types of damages—such as UV-induced CPDs—and is evolutionarily conserved in all eukaryotes. The dynamics are considerably well understood *in vitro* (Mu et al. (1995) for human cells, Wang et al. (1995); Guzder et al. (1995) for yeast). However, the picture is less clear for studies *in vivo*, in particular in context of other nuclear processes such as transcription and DNA packaging. Whilst the *in vitro* assay indicates that CPDs can be repaired within 3 - 10 minutes after lesion recognition (Erixon and Ahnström (1979)), significantly elevated levels can be still observed after two hours post-irradiation *in vivo* (Mao et al. (2016)). It is hence pivotal to develop additional models that are adapted to the environment in a living cell.

NER kinetics are commonly divided into two recognition pathways—GGR and TCR—which subsequently converge to the same incision and replacement pathway. Involved protein components belong to the RAD3 epistasis group, which were revealed by UV-sensitivity screenings (Boiteux and Jinks-Robertson (2013)). Therefore, findings associated with NER functioning are chiefly related to UV-induced damage, although studies for other types of molecular disruptions exist as well. In the following, we present only the specifics and proteins concerning repair in budding yeast. Human homologues are separately introduced if necessary. An overview is given in Table 1.1 and Figure 1.2. We explain the the structure of CPDs before explaining known NER properties *in vivo* and *in vitro*.

Yeast Gene Name	Property	Mammalian Gene Name
RAD4	Forms a complex with Rad23 and Rad33 that binds damaged DNA.	XPC
RAD23	Forms a complex with Rad4 that binds damaged DNA.	HRAD23B
RAD33	Forms a complex with Rad4 that binds damaged DNA.	CEN2
RAD7	Forms a complex with Rad16.	DDB1
RAD16	Forms a complex with Rad7 that has ATP-dependent binding of damaged DNA, chromatin remodeling activity, and E3 ligase activity.	DDB2
RAD1	Forms a complex with Rad10 that has structure-dependent endonuclease activity; incises DNA on the 5'-side of lesions.	XPF
RAD10	Forms a complex with Rad1.	ERCC1
RAD2	Structure-dependent endonuclease; incises DNA on the 3'-side of lesions.	XPG
RAD14	Zinc-finger protein; binds damaged DNA.	XPA
RAD25	TFIIH subunit; DNA-dependent ATPase and X' to Y' helicase.	XPB
RAD3	TFIIH subunit; DNA dependent ATPase and helicase with Y' to X' polarity.	XPD
CDC9	DNA ligase 1	LIG1
RAD26	DNA-dependent ATPase required for transcriptional bypass of lesions and for TC-NER.	CSB
RAD28	WD40 repeat protein of unknown function.	CSA
RPB9	Nonessential RNA Pol II subunit required for Rad26-independent TC-NER.	POLE21

Table 1.1: **NER genes and their mammalian homologues.** The table gives the gene names of NER proteins with a short description and the mammalian counterpart. Table was taken from Boiteux and Jinks-Robertson (2013).

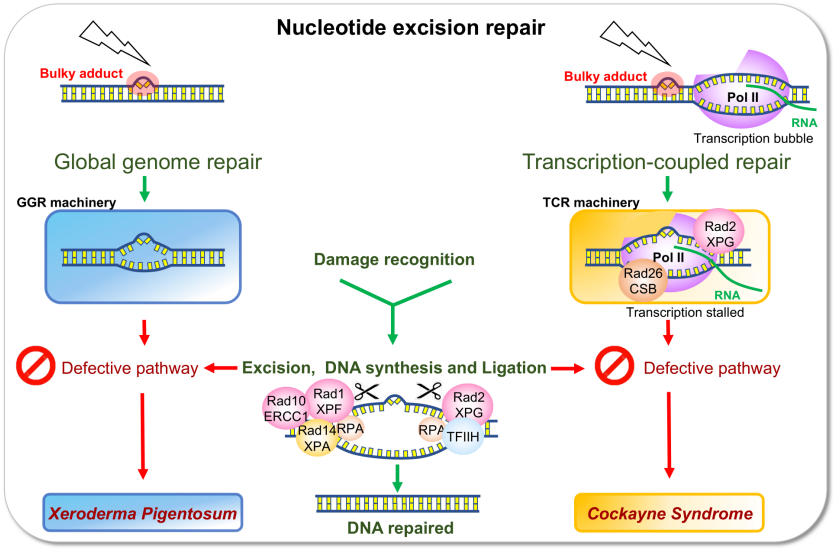


Figure 1.2: **Schematic overview of the NER pathway.** Lesions are either detected by GGR (left) or TCR (right). GGR lesion recognition is governed by the Rad4-Rad23-Rad33 complex. Protein loading is facilitated by Rad7-Rad16. This promotes the recruitment of TFIIH. During TCR, Pol II elongation is hindered by a TBL. The association of other NER proteins is predominantly triggered by Rad26 and Rad28, but also involves Rad2. Additionally, NER can be evoked independently of Rad26 by Pol II subunit Rpb9. Both recognition pathways subsequently follow the same incision and replacement mechanism. The helix is further opened by TFIIH and the presence of damage is verified through Rad14, Rad25, and RPA. Rad1-Rad10 as well as Rad2 incise the DNA strand on the 5' and 3' end of the lesion, respectively. The excised fragment is replaced by Pol δ or Pol ϵ and eventually sealed by DNA ligase 1. The figure was taken from André et al. (2021).

UV-Induced DNA Damage

The irradiation of cells with UV leads to the creation of different types of DNA disruptions. Most of them are CPDs and 6-4PP (both variants of pyrimidine dimers), the former of which accounts for up to 75-95% of all lesions (Bohm et al. (2023)). This motivates the focus of this work on CPDs.

The formation of CPDs are caused by a photochemical reaction during which UV is absorbed through a double bond between pyrimidine bases (Fig 1.3). By opening the hydrogen bond, the free nucleotide reacts with neighbouring molecules. If the adjacent nucleobase is another pyrimidine, they form new direct bonds (Goodsell (2001)). CPDs are therefore categorised as transition-type lesions—i.e. the succession of two bases—namely CT, TC, CC, or TT. They form so-called *bulky* DNA damage and TBLs that can be repaired by NER.

The UV irradiation emitted by the sun is commonly divided into UVA, UVB, and UVC, depending on their wavelength and consequently the transported energy. Nonetheless, all of them can induce damage to the DNA—including CPDs—although to varying levels. UVC has the shortest wavelength which corresponds to the highest amount of UV-transported energy. Whilst UVA and some UVB rays

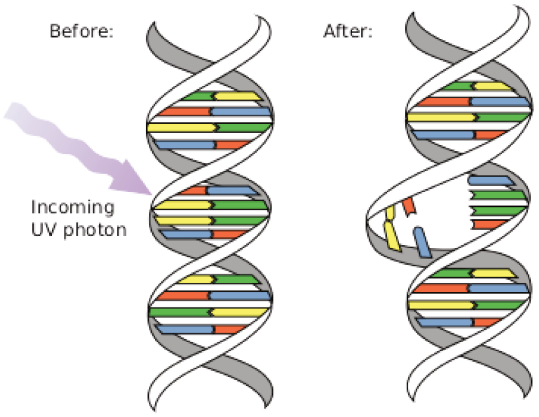


Figure 1.3: **Formation of CPDs.** The energy of an incoming photon is absorbed by the hydrogen bonds, so that they can freely react with neighbouring molecules. If the adjacent nucleotide is a pyrimidine, they create new bonds. The figure was taken from Herring (2010)

can penetrate the Earth's ozone layer, UVC (and partly UVB) can be absorbed. Despite the fact that most on earth-living organisms have never been in contact with UVC, it is commonly used in experimental setups to study CPD formation and repair. Its use results in a high number of lesions, and it therefore yields a more reliable data collection. All CPD measurements that were used in this work were obtained using UVC.

Damage Recognition by Global-Genome Repair

GGR finds and recognises DNA lesions by direct protein associations of Rad4-Rad23-Rad33 (van Eeuwen et al. (2021)) (Figure 1.2, left). The pathway can be observed along the entire genome. Recognition and following repair is independent of lesion site and chromatin structure, although it might be facilitated by interactions with chromatin remodelers such as SWI/SNF and Ino80 (Sarkar et al. (2010)). Whilst Rad4 is required for GGR, Rad23 or Rad33 deletions seem to solely decrease repair efficiency. However, double mutants result similarly in deficient lesion removal (den Dulk et al. (2006); Boiteux and Jinks-Robertson (2013)). Repair in Pol I-transcribed ribosomal genes additionally involves Rad34 (den Dulk et al. (2005)). Lesion recognition by Rad4 is driven by the detection of thermodynamically unstable base pairs (Min and Pavletich (2007)), yet CPDs—as a particular form of damage—do not support sufficiently stable binding. The Rad7-Rad16-Abf1 complex associates to damage sites to facilitate loading of Rad4 (Jones et al. (2010)). Indeed, gene knockout experiments proved Rad7-Rad16 to be essential for correct GGR (Verhage et al. (1994); Wang et al. (1997)). It has also been suggested that the protein complex plays multiple roles during repair. There is evidence that it is involved in post-incision steps (Reed et al. (1998)), and it fosters the UV-dependent ubiquitination of Rad4 (Gillette et al. (2006)). Rad4-Rad23-Rad33 itself permits damage verification

through recruiting the Transcription initiation Factor IIH (TFIIH). During this process, the distorted strand is further opened by the ATPase and helicase activity of Rad3 and Rad25 (Boiteux and Jinks-Robertson (2013)). The Rad4-Rad23-Rad33 recognition complex is subsequently released and the DNA is scanned in a 5'-3' direction for helicase-blocking lesions (Sugasawa et al. (2009)). If no DNA damage is found, the open strand is closed and the process reverted.

Damage Recognition by Transcription-Coupled Repair

There is a scientific consensus that lesions in transcribed regions exhibit quicker repair than silent downstream sequences. This promoted the idea that TCR is more efficient than GGR (Bohr et al. (1985); Mao et al. (2016); Li et al. (2018); Mao et al. (2020)). It is commonly assumed that damage removal from the Transcribed Strand (TS) is preferred over the Non-Transcribed Strand (NTS) (Mellon et al. (1987)), which was demonstrated on the *RPB2* gene (Sweder and Hanawalt (1992)). TBLs cause Pol II to be stalled at damage sites (Figure 1.2, right). The recruitment can be triggered either by Rad26 or Rpb9 (Duan et al. (2020)), the latter of which is a subunit of Pol II.

Rad26 is related to Pol II elongation, and it is therefore present during lesion detection as well (Malik et al. (2010)). Rad26-mediated NER is also associated to Rad28. In contrast to their homologues in human cells, Rad26 or Rad28 knockout mutants are not UV-sensitive (Boiteux and Jinks-Robertson (2013)). Blocked Pol II is assumed to stabilise protein interactions with Rad26, which might lead to lesion bypassing (Yan et al. (2021)). In this case, a repair cascade is not evoked, although the faulty site in the mRNA nucleotide sequence can produce erroneous proteins, which is also called *transcriptional mutagenesis* (Brégeon and Doetsch (2011)). TBLs that cause continued stalling trigger the execution of consecutive NER steps, which is facilitated through chromatin remodelling by Rad26 (Boiteux and Jinks-Robertson (2013)).

It has been reported that recruitment of other NER proteins can also be evoked independently of Rad26 by the non-essential Pol II subunit Rpb9 (Li and Smerdon (2002)), particularly at the Transcription Starting Site (TSS)-proximal half of the +1 nucleosome (Duan et al. (2020)). Rpb9's exact role during protein loading has not yet been characterised, although it is suggested that it promotes the association of TFIIH. Deletion of both Rad26 and Rpb9 renders cells TCR-deficient.

The further assembly of the repair machinery is impaired by the stalled Pol II complex, which covers around 35 nt of the transcribed strand, including the lesion (Tornaletti et al. (1999)). There have been various and non-excluding hypotheses about the fate of Pol II, among others dissociation, backtracking, or degradation. It is commonly conjectured that the most common mechanism is transcript cleavage followed by backtracking (Sigurdsson et al. (2010); Marteijn et al. (2014)), as it is also involved in other nuclear processes such as transcription proofreading. Nevertheless, the pre-

cise protein interactions remain still in the dark. It should be noted that deficient backtracking does only lead to a negligible phenotype in yeast; yet it results in severe disorders in human cells such as Cockayne Syndrome (CS). Thus, there might be different mechanisms in place in *Saccharomyces cerivisiae* versus human cells (Boiteux and Jinks-Robertson (2013)).

Incision and Replacement

In vitro screenings of the incision after lesion recognition on naked DNA identified six fundamental NER protein complexes: Rad4-Rad23, Rad14, TFIIH, Rad1-Rad10, Rad2, and replication protein A (RPA) (Guzder et al. (1995)). It should be emphasised, however, that there might be various other proteins involved *in vivo*. After damage detection through either GGR or TCR, the DNA is further opened by the ATPase / helicase interplay of the multiprotein complex TFIIH, in particular by Rad3 and Rad25. TFIIH components are also interacting directly with other NER factors (Compe and Egly (2012)). Although the TFIIH complex is primarily associated with transcription initiation, the catalytic activity of its submodule Rad3 is only required during NER, highlighting its multifunctional role (Feaver et al. (1993)). The pre-incision complex is stabilised through binding of Rad14 and RPA. Lesion presence is verified by TFIIH, Rad14, and RPA. If damage is absent, the DNA cleavage is not performed, and the proteins dissociate.

The lesion is removed by an incision on both sides of the distortion. This represents a *point of no return* (Marteijn et al. (2014)). Rad1-Rad10 and Rad2 are positioned on the 5' and 3' side of the lesion, respectively (Evans et al. (1997)). As they lack specificity to DNA damage, they are guided by interactions with other proteins (Tomkinson et al. (1993); Habraken et al. (1993)). Rad1-Rad10 and Rad2 incise the helix distortion on either side. The excised fragment is subsequently released together with the other NER components.

The dual incision is followed by DNA synthesis and ligation. Although this is poorly documented in yeast, data suggest that Pol δ or Pol ϵ (two DNA polymerases) perform the replacement of the missing oligonucleotide. This leaves an open nick, which is sealed by DNA ligase 1 (Budd and Campbell (1995)).

1.2.3 NER in Context of Other Nuclear Processes

DNA Repair

Virtually everything in the environment—and even cell-internal processes—can cause changes to the molecular structure of the DNA. It is therefore not surprising that several repair pathways developed over the course of evolution to remove the various types of damages. Particularly interesting for

the repair of UV-induced lesions in *Saccharomyces cerevisiae* are—next to NER—Photoreactivation (PR) and Base Excision Repair (BER).

PR is mediated by photolyase, a DNA repair enzyme that is activated by the energy of photons coming from (near-)visible light. When bound to the damage, it reverts the lesion by inserting it into the enzyme's active site (Sancar (2003)). Photolyase repairs 0.3 CPDs/kilo base (kb) around nucleosome positions in two hours, but it only needs 15 minutes in regions depleted of nucleosomes. Repair speed is therefore largely location-dependent. However, considering *in vitro* data, it can be up to six-fold faster than NER in NDRs and Autonomous Replication Sequences (ARS) (Suter et al. (1997, 2000b,a)). Prevalence of one or the other repair pathway might be hence position-specific. Interestingly, photolyase harnesses the energy carried by photons only during the enzymatic step, and it can associate to pyrimidine dimers in the absence of light. This raises the question whether NER components interact with photolyase during the repair process. Indeed, it was shown that cell survival was improved in the presence of photolyase, even though PR could not be carried out in the dark. This indicates that the enzyme promotes NER kinetics (Sancar and Smith (1989)). Due to its high (although region-dependent) efficiency, it is of paramount importance to control light exposure after UV irradiation during experiments to prevent a potential influence of PR on CPD repair. If overexpressed under strong light exposure, it can remove over 80% of all CPD lesions within 90 seconds (Bucceri et al. (2006)). It should be mentioned that PR is absent in human cells. It has been proposed that the loss of PR particularly in placental mammals can be explained by a lack of selection pressure and an increased mutagenesis rate induced by photolyase (Lucas-Lledó and Lynch (2009)). PR is hence not of interest for this work.

BER is—similar to NER—a multistep process involving several proteins. It is particularly responsible for repairing lesions with endogeneous cause or induced by ROS. Whilst NER can repair adducts with up to 30bp, damages removed by BER are typically smaller than 10bp, and they are commonly different types of lesions (Casal-Mouriño et al. (2020)). It is assumed that they fulfill distinct roles in maintaining the molecular integrity of the genome. However, it should be mentioned that a growing body of evidence suggests that they share common components and cooperate with each other (Kumar et al. (2020)). This becomes an important consideration when dealing with several types of lesions at the same time, such as ROS and CPDs. Nevertheless, the relatively strong DNA distortion that is produced by CPDs can be presumed to be solely repaired by NER if not specifically deleted. The influence of BER was not considered.

Nucleosome Positioning

The *Saccharomyces cerevisiae* DNA is—when fully unfolded—around five millimetres long and needs to be packaged into the spatially constraining nucleus with a two-micrometre diameter (Yanamoto et al. (2011)). This is accomplished by wrapping the DNA tightly around histone complexes named nucleosomes (Kornberg (1974); Luger et al. (1997)), which also neutralise the genome's negative charge (Jansen and Verstrepen (2011)). The condensed structure is commonly referred to as chromatin (Kornberg (1974)). Nucleosome positions are highly frequent and occur roughly every 200 base pairs (bp) in all eukaryotes. A nucleosome consists of $\approx$146 bp of DNA that is coiled 1.65 times around the histone octamer (Luger et al. (1997); Jansen and Verstrepen (2011)). Short stretches of linker DNA connect the nucleosomes along the genome (Figure 1.4).

The nucleosome core is composed of several histone units, namely the two H2A-H2B dimer and one H3-H4 tetramer (Figure 1.4(A, top)). H1 and H5 are linker histones that lock the nucleosome position by binding starting and ending sites. The histones' amino acids lysine and aginine establish salt and hydrogen bonds to the DNA, further stabilising its position. Histones also possess a net positive charge which increases binding stability with the negatively-charged DNA phosphor backbone (Figure 1.4(C)). All histones contain so-called tails, which are subject to chemical modifications. These post-translational histone marks allow the regulation of various nuclear processes (Allfrey et al. (1964)). As the histone N-terminal tail can make direct contact to adjacent nucleosomes, there is a scientific consensus that chemical modification can regulate chromatin conformation (Luger et al. (1997)). Moreover, they can evoke enzyme recruitment to remodel the nucleosomal position by utilising ATP (Bannister and Kouzarides (2011)). This is changing sequence accessibility, and they therefore influence other vital procedures, such as transcription and repair. The most influential modifications include acetylation, phosphorylation, and methylation (Figure 1.4(B)).

Histone acetylation is performed by acetyltransferases, which can—by catalysing the transfer of an acetyl group—neutralise the lysine's positive charge of the N-terminal tail. This weakens binding to the DNA molecule. The effect is reverted by histone deacetylase. Similarly, phosphorylation is governed by kinases, which add a phosphate group from ATP to one of the amino-acid residues, in particular serines, threonines and tyrosines. Phosphatases revert this process. This can occur at histone tails as well as core histones. Phosphorylation increases the negative charge, and it is therefore clearly changing DNA-protein interaction. It has known roles in DNA repair and regulation of transcriptional activity. Lastly, histone methylation is a chemical modification of the side chains—particularly lysines and arginines—that does not influence the protein charge. All histone modifications change directly the chromatin structure as well as regulating loading of other effector

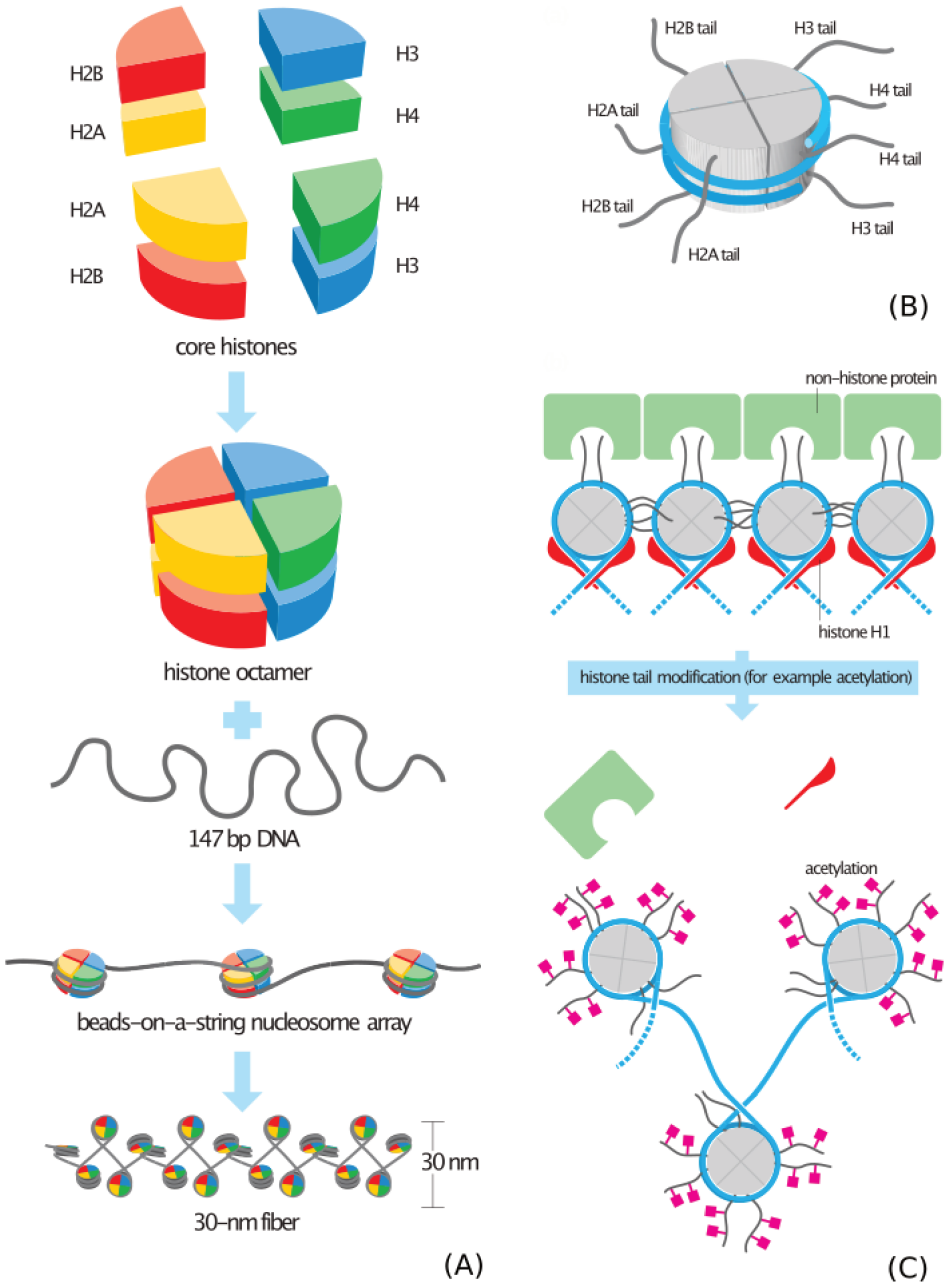


Figure 1.4: **Chromatin organisation and packaging.** Figure 1.4(A): DNA packaging occurs around histone complexes called nucleosomes. A single nucleosome core is composed of two H2A-H2B dimers and on one H3-H4 tetramer. Figure 1.4(B): Histones are subject to chemical modifications, which affects particularly the N-terminal tails. 1.4(C): The N-terminal interacts with other nucleosomes, non-histone proteins, as well as with the DNA directly. Linker histones (red) lock the DNA at a specific position around the nucleosome (figures were taken from Morgan (2007a,b)).

molecules (Bannister and Kouzarides (2011)).

It is intuitive that tightly wrapped DNA is less accessible than linker DNA. Hence, nucleosome positioning and kinetics play an active regulatory role in various nuclear pathways, including transcription, replication, and DNA repair. Despite the fact that positioning is partly sequence dependent (Tillo and Hughes (2009)), recent studies showed strong interactions with proteins and *trans* factors,

including nucleosome remodelers with an ATPase subunit such as RSC (André et al. (2023); Badis et al. (2008)). There is also evidence of interaction with other DNA-bound proteins, among others Pol II (Jansen and Verstrepen (2011)). As a matter of fact, nucleosomes are not only influenced by other proteins but also by the presence of their own kind. Their highly frequent positioning leads to preferred lengths of linker DNA ($\approx$ 18 bp) (Mavrich et al. (2008)). As neighbouring nucleosomes cannot overlap, it has been proposed that an array can be modelled as beads on a string (Jansen and Verstrepen (2011)) (Figure 1.4(A)).

Nucleosome organisation along the genes is highly structured and preserved along the entire genome. About 95% contain a Nucleosome Depleted Region (NDR) upstream of the TSS, which coincides with the gene promoter (Jiang and Pugh (2009); Jansen and Verstrepen (2011)). Nucleosomes are strongly positioned close to the TSS. Naming of the nucleosomes follows the positional order with respect to the NDR, i.e. the first downstream nucleosome is +1, the second +2, etc. Positions before are called -1, -2, etc. The +1 and -1 are adjacent to the NDR. Further downstream nucleosomes are phased with respect to the +1 position, which promoted the notion of the barrier model (Mavrich et al. (2008)).

A correct three-dimensional chromatin organisation is pivotal for yeast survival. It has been reported that the tight packaging in the nucleus prevents entanglement of the DNA molecules (Arsuaga et al. (2002)). The topological entanglement can have dramatic effects on the regulation of various processes, including gene expression (Portugal and Rodr1́guez-Campos (1996)). Therefore, it cannot be excluded that the 3D organisation influences DNA repair. Indeed, it has been found that chromatin mobility might play a crucial role to promote cell-cycle arrest and chromosome segregation during the removal of Double-Strand Breaks (DSB) (Strecker et al. (2016)). *Saccharomyces cerevisiae* chromosomes follow a Rabl-like conformation, which describes the localisation of centromeres and telomeres close to the nucleus membrane. It has not been fully resolved how genomic entanglement is minimised to permit correct functioning. However, it has been proposed that the Rabl configuration might be necessary to reduce entanglement incidence (Pouokam et al. (2019)). In this study, we do not consider the influence of the 3D chromatin folding. Incorporating it would require extensive study of various other data types such as microscopy data. As this study focus on the modelling of CPD repair using high-throughput sequencing data, any higher-order structure other than nucleosome positioning was chiefly ignored.

The arrangement of nucleosomes plays an important role during damage formation itself. It has been shown that outward-rotational DNA at strongly positioned nucleosomes is less protected against UV irradiation, leading to a so-called *photo-footprint* which persists during ongoing repair. Moreover, it could be demonstrated that there is a subtle but consistent effect of reduced repair speed close

to the nucleosomal dyad. This trend vanished at highly dynamic nucleosome positions (Mao et al. (2016)). Other studies also find that the presence of nucleosomes significantly inhibit CPD repair. Linker DNA as well the 5'-end of positioned nucleosomes exhibited faster repair than centre and core sites (Guintini et al. (2015)). Particularly repair at the NTS is seemingly influenced by the nucleosome organisation (Mao et al. (2020)). Surprisingly, though, there is evidence that the human Rad4-homologue XPC is more abundant at densely-packed heterochromatin. XPC mobility is significantly slowed down following UV treatment due to more stable binding at DNA lesions. This effect was drastically reduced after 2h, and mobility dynamics returned to pre-irradiation levels after $\approx$ 4h, far before the completion of CPD repair in human cells (Hoogstraten et al. (2008)). This could possibly indicate that damage recognition at heterochromatin is highly efficient, whilst access for other NER components is blocked by nucleosome packaging. Similar findings for Rad4 are lacking, although it should be emphasised that heterochromatin in *Saccharomyces cerevisiae* is limited to only some few regions (i.e. telomeres, the rDNA locus, and the silent mating-type cassettes) (Duina et al. (2014)).

Transcription

The genome comprises various expressed sequences, which are grouped depending on the process the transcript is involved in. Genes encoding ribosomal RNA (rRNA) and transfer (tRNA) are expressed by RNA Polymerase I (Pol I) and RNA Polymerase III (Pol III), respectively. Protein-coding sequences are expressed through Pol II, a process that produces mRNA. All RNA polymerases are part of the multisubunit RNA polymerase family. The overall transcriptional process for a single gene can be divided into three steps: initiation, elongation, and termination. Pol II is a 12 subunit multiprotein complex (which is the *complete* form) with a 10 subunit *core*. The ten subunits of the core complex can be categorised into several mobile modules. Firstly, the *trigger loop* opens and closes around newly added RNA bases, which support RNA sequence proofreading, a mechanism that has been kinetically described by Hopfield (1974). Secondly, the *cleft*—through which the DNA descends—synthesises RNA by passing the template between the *clamp* which consists of two modules named *jaws*. A *wall* with a magnesium ion separates the RNA-DNA hybrid, where the DNA is pushed 90° downwards and re-hybridises with the opposite strand (Cramer et al. (2001); Schier and Taatjes (2020)).

Although not all subunit-subunit interactions and functions are known, we want to highlight some few that we deem to be particularly important. All Pol II subunits are called Rpb followed by a number. The numbering order indicates subunit size from largest to smallest. Rpb1 (together with other subunits, in particular Rpb9) creates a groove where DNA is bound and transcribed to RNA. This

contact is particularly maintained during transcription by Rpb2. Rpb6 stabilises Pol II association during transcription. Rpb4 and Rpb7, which are not part of the core enzyme, can reversibly associate to the main complex. The core cannot initiate transcription without those subunits, although it is independent during elongation (Bushnell and Kornberg (2003)).

Transcription of all protein-coding genes is regulated by the sequence-specific binding of TFs either to Upstream Activating Sequences (UASs) or Upstream Repressing Sequences (URSs) in yeast, i.e. enhancers and silencers in multicellular eukaryotes. There is support that nucleosome presence prevents TF-independent transcription (Juan et al. (1993)). As suggested by the name, almost all UASs or URSs are positioned at the 5'-side of the promoter in yeast. In multicellular eukaryotes, however, they can be positioned at different distance and orientation with respect to the promoter. Although they are commonly close to the NDR next to the gene's TSS, they can be similarly located more than 1kb away (Hahn and Young (2011)). Instead of changing the activity of TFs, some pathways rather modulate the transcription levels themselves. The transcriptional program needs to be dynamically coordinated with the chromatin structure and nucleosome positioning. Indeed, bound activators recruit co-activators to modulate chromatin conformation to make it more accessible or to stimulate the assembly of the transcription machinery. One of these co-activators is the Mediator complex, which facilitates protein loading and stabilisation (Soutourina (2018)). Binding to UASs and URSs influences the behaviour of the transcription machinery assembly at the core promoter. This is commonly related to the assembly of the Pre-Initiation Complex (PIC) which is composed of Pol II and the general TFs TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, and the Mediator complex. PIC assembly commonly occurs at core promoters containing a TATA-box or TATA-like elements, although it has been reported that low levels of TATA-independent expression are possible at some genes (Pellman et al. (1990)). TFs do not bind directly to Pol II to regulate transcription, but instead rely on interaction with Mediator (Soutourina (2018); Schier and Taatjes (2020)). The cooperation between PIC and Pol II—both downstream and upstream of the TSS—are necessary for initiating transcription and stabilising the open complex. Transcription begins after scanning downstream for a suitable TSS (Hahn and Young (2011)) (Figure 1.5). The TSS is a distinguishable sequence composition, and is predominantly 40 to 120nt farther downstream of the TATA box (Struhl (1987)). Transcription itself is mediated by elongation factors, which can bind Pol II but do not necessarily constitute to the PIC.

It can take up to several minutes from transcription initiation to the translation and completion of a functional protein. Hence, there is a delay before a cell can react to a changing environment, and it might be important to take temporal aspects into account during modelling. Pol II moves at a limited and non-constant speed, stopping at several pausing sites along the transcript. There are various chemical changes that are caused by Pol II pausing and backtracking. For example,

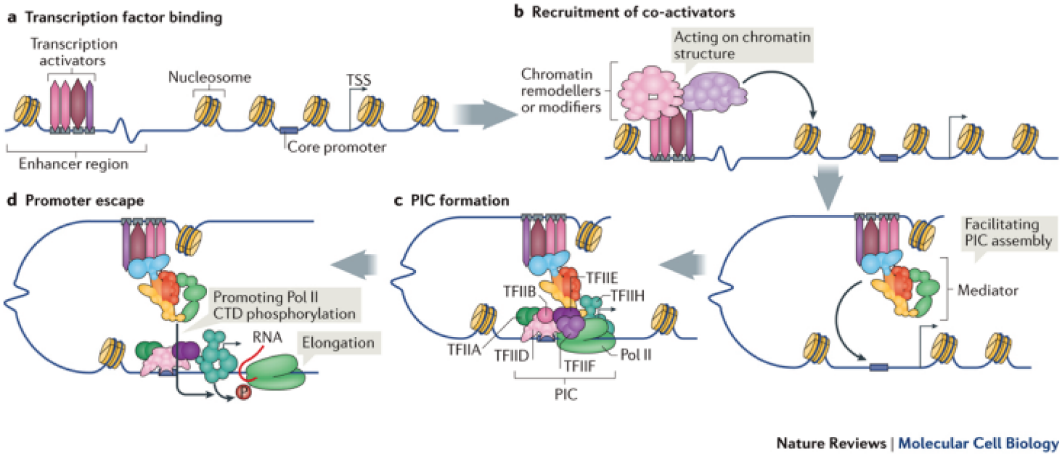


Figure 1.5: **Schematic description of the transcription initiation process.** Transient contact between transcription factors at enhancer and promoter regions—which is orchestrated by the Mediator complex—regulates PIC assembly and mediates promoter escape of Pol II. The figure was taken from Soutourina (2018).

the newly synthesised RNA strand is pushed back into the Pol II funnel (which is blocking further elongation), and the unhybridised DNA template is removed from the active site. TFIIS can stimulate the catalytic cleavage activity of stalled Pol II, making further elongation possible, even though Pol II's own enzymatic activity can largely remove the blocking RNA itself (Schier and Taatjes (2020)). On average, elongation might occur at 18-24 nt per second, resulting in 25-50 seconds for 1kb (Pérez-Ortín et al. (2007)).

It has long been assumed that transcription happens only at coding regions. However, an increasing number of studies reveal the importance of antisense and non-coding transcription. Indeed, antisense RNAs (asRNAs) can repress gene expression and therefore fulfill a pivotal regulatory role (Nevers et al. (2018)).

As aforementioned, TCR is a repair pathway that is linked to the transcriptional activity of Pol II. Therefore, gene expression itself naturally influences NER dynamics. Indeed, it can be shown that genes which were highly active prior to UV treatment exhibit quicker CPD removal (Mao et al. (2016); Li et al. (2018)). Nevertheless, the effect of TCR can be observed at all genes, independent of their transcription levels (Mao et al. (2020)). It should be emphasised that the link between transcription and repair is far from trivial, and there is no clear linear relationship (Li et al. (2018)).

A deep investigation is further complicated, as cells engage in a stress response upon UV-irradiation, which is manifested in a global transcription shutdown (Gregersen and Svejstrup (2018); Hauser et al. (2019)). Surprisingly, the few upregulated genes can be barely linked to processes necessary for cell survival and DNA repair. This indicates that essential proteins must be sufficiently present prior to stress exposure (Birrell et al. (2002)). Genes with high activity after UV treatment are

involved in metabolic, catabolic, and proteasotic activity, as well as transporters and iron regulation (Hauser et al. (2019)). Despite the identification of some few upregulated genes after irradiation, it remains poorly understood how the transcriptional stress response really unfolds over time.

1.2.4 Next Generation Sequencing Data

The development of high-throughput NGS drastically decreased costs and allows fast probing of genomic properties along the entire DNA molecule. Whilst the initial assembly of the human genome costed approximately 3 Billion USD, it is now possible to sequence an entire sample with some hundred dollars. It has been particularly used to study DNA-protein interaction, although it can be similarly applied to measure other features, including DNA damage and nucleosome positioning. It is based on the Sanger sequencing technique and largely replaced microarray methods, as it allows the genome-wide sequencing instead of being limit to predefined regions. Therefore, there is no prior knowledge of the probed process required. Furthermore, it is highly reproducible with a small error rate whilst requiring only a small quantity of nucleic acid. Despite being based on the first-generation Sanger sequencing, the chemical principle harnessed by NGS is fundamentally different. In particular, it performs the sequencing of many small fragments in parallel rather than using slow capillary electrophoresis (Behjati and Tarpey (2013)). In this section, we present the common steps of the Illumina NGS workflow that were used for this thesis (taken largely from Hu et al. (2021)). They can be divided into sample and library preparation, sequence determination, and data analysis. Other techniques—such as single-cell sequencing—are not considered here, as we do not work with tissues that analyse several cell types at the same time.

To describe the workflow in a few summarising words, cells are harvested, and the nucleic acid of interest—i.e. DNA or RNA—is extracted. They are partitioned into smaller fragments, which are ligated to platform-specific adaptor sequences. This permits binding to a hard surface in a device, therefore spatially structuring the sample. Short-read sequencing (which has been exclusively used in this work) is performed by Sequencing By Synthesis (SBS), a process during which sequence composition can be measured. All acquired data are computationally filtered and rectified using standard as well as customised processing pipelines. In the following, we will point out technical details of each step that we deem to be important for this work.

The library preparation consists of the gathering of the nucleic acid molecules, i.e. the DNA or RNA sample. Special caution should be spent to rRNAs, which make more than 90% of the total RNA. If they are not of interest, they must be depleted. DNA or RNA is commonly extracted from an entire cell culture at exponential growth. This corresponds to 8-12 million cells of *Saccharomyces*

cerevisiae (Puay Yen Yap (2017)). Consequently, NGS data represent a snapshot over an entire population.

Samples are usually amplified—for example using a Polymerase Chain Reaction (PCR)—and partitioned into short sequences of 250 - 800bp, with most fragments $\approx$ 300bp for Illumina sequencing. Fragmentation can be performed using different techniques, including sonication and enzymatic digestion (Head et al. (2014)). Sonication is commonly used for DNA-sequencing (DNA-seq) as well as Chromatin Immonuprecipitation sequencing (ChIP-seq). Fragmentation through digestion is preferred for MNase sequencing (MNase-seq), which probes the position of a nucleosomal dyad. In this case, MNase removes free linker DNA, conserving only DNA that was coiled around histone complexes. Fragmentation is followed by the ligation of a platform-specific adaptor to the sequences. They are used for the fragment recognition by the sequencing device. RNAs require an additional reverse-transcriptase step to produce complementary DNA (cDNA). To increase sequencing efficiency, adaptor ligation is followed by a size selection during which fragments outside a pre-defined range are removed.

Short-read sequencing, such as performed by Illumina, is based on the release of light through SBS of fluorescent-labelled nucleotides that are bound to reversible terminators (Goodwin et al. (2016)). At each cycle, the incorporated nucleotide emits a light signal that can be measured. The terminator is subsequently removed, which permits the continuation of the polymerase step. SBS is preceded by a cloning procedure during which samples are largely amplified using *bridging* PCR on a solid-phase called a flow cell. This improves the signal detection during sequencing. Illumina sequencing possess a relatively low error rate of 0.1% (Hu et al. (2021)).

An initial data analysis step is performed by the sequencing platform, which records and measures the quality of the read. Adaptor sequences are removed (which is called trimming), and reads are filtered based on their quality. This is followed by the read alignment to a reference genome. It determines the position of the read along the entire DNA. Many alignment algorithms today are based on a Burrows-Wheeler transform that can be compared to pre-computed values in a hash table. An optional variant calling step on the aligned reads permits the detection of Single Nucleotide Polymorphisms (SNPs) or larger structures that are different to the reference. The produced data files are then subjected to further downstream analysis and modelling techniques.

Technically, NGS allows single-nucleotide resolution. In practice, however, this can be difficult to achieve for the probed quantity. A CPD sequencing approach—which was adapted from a previous method for the assessment of ribonucleotide lesions (Ding et al. (2015)) using an additional enzymatic step—has been proposed to study UV-induced damages at a genome-wide single-nucleotide resolution (Fig 1.6) (Mao et al. (2016)). The free 3' hydroxyls (3'OHs) of the damaged and sonicated

DNA are ligated with an adapter sequence, which is followed by cleavage directly upstream of the CPD using the repair enzyme T4 endonuclease V and an apurinic/apyrimidinic endonuclease. The new free 3'OH end is subsequently ligated with another adapter sequence (adapter A), thus marking the exact CPD position. Purified fragments are then amplified and sequenced. The applied data analysis pipeline adds the location right downstream of adapter A to the sequencing signal. The CPD-seq data produced by Mao et al. (2016) were used extensively in this work to study and evaluate the DNA repair process on a population scale.

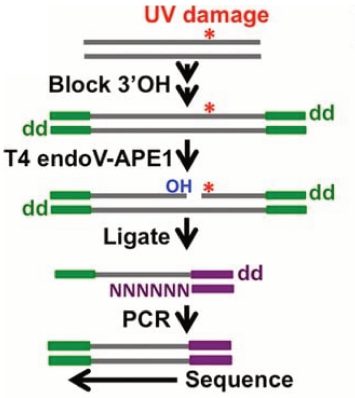


Figure 1.6: **Schematic explanation of the CPD-seq method.** The trP1 adapter is colored green, and the A adapter is given in purple. *OH* indicates a free 3'OH; *dd* indicates dideoxy (i.e., 3'H). The figure and caption was taken from Mao et al. (2016).

1.3 Computational Methods

Due to the sheer complexity of possible interactions in living systems, more and more biological projects include sophisticated computer models to understand and analyse their data. This led to the development of separate scientific branches such as system biology or computational biology. As introduced in Section 1.2, transcription, nucleosome phasing, and damage removal are deeply entangled through various molecular interactions, which is why we consider it as necessary to include computational approaches to study DNA repair. However, it is not straightforward to sensibly represent the plethora of entwined nuclear processes with mathematical formulas in a simplified model. To complicate matters, many of the protein-protein or DNA-protein interactions remain unknown, as explained above. It is therefore necessary to start with a very general description of molecular dynamics, which are then refined to the specifics of DNA lesion removal.

In this work, we presume that protein and DNA movements can be represented by stochastic particle dynamics. Interactions can be observed when they co-localise. All applied modelling techniques make use of processes that describe the motion of molecules. We utilised in particular the

mathematical framework of Brownian motion, which describes the probabilistic movement of particles suspended in a medium—in our case, the nucleoplasm. Fluctuations in the molecule's path come from interactions with other particles. Although we assume that every single molecule follows deterministic Newtonian dynamics, the formulation as a many-body problem makes it infeasible to account for all possible collisions during simulations. Consequently, it requires a statistical treatment. For many observations in nature, however, it is possible to find a phenomenological description of the average behaviour that ignores random fluctuations (see for example the model by Kolmogorov (1937), Johnson and Mehl (1939), and Avrami (1939, 1940, 1941) introduced below). It should be emphasised that the underlying process is nevertheless stochastic. First described by the botanist Robert Brown, the theory has been particularly developed by Bachelier (1900), Einstein (1905), and von Smoluchowski (1906).

We used the probabilistic framework of Brownian motion to describe the random and location-specific DNA-protein interactions along the genome. Indeed, it was already proposed by Schrödinger (1943)—before the discovery of the actual molecular structure of the DNA—that stochastic effects might be pivotal for genomic processes. It should be mentioned that we are oblivious of any three-dimensional movement in space—despite the fact that we incorporate spatial and position-specific NGS data. To model particle dynamics along the one-dimensional string, we presume that nearby interactions are more likely to happen in an infinitesimal time step than interactions that are farther away. We divide the one-dimensional sequence into segments where this holds reasonably true. This means that within these partitions, we conjecture that the effect of particle movements in three dimensions that appear as a jump in one dimension is negligible (Figure 1.7). We ignored the complicated three-dimensional DNA conformation in space. A more detailed assessment is given in Appendix A. Examples and consequences are stated and critically discussed further below.

Next to the formal mathematical description of the process, another major problem in computational biology is the incorporation of data into the model with the goal to find reasonable parameter estimates. Fortunately, different machine learning approaches can be remedially applied. To avoid any ambiguity, we distinguish between the mathematical description of the process; and the training procedure that changes the model parameters to find the best explanation of the available data. The actual choice of the learning method depends on the approach as well as the type of data available.

In this section, we introduce the fundamental modelling principles used in this work. We present first a very general introduction into stochastic processes and give two examples how to analyse them (Subsection 1.3.1). Thereafter, we introduce the formalism of Brownian motions and set our equations into context (Subsection 1.3.2). This is followed by a brief description of parameter estimation approaches (Subsection 1.3.3). In this work, the training procedures themselves are used merely as

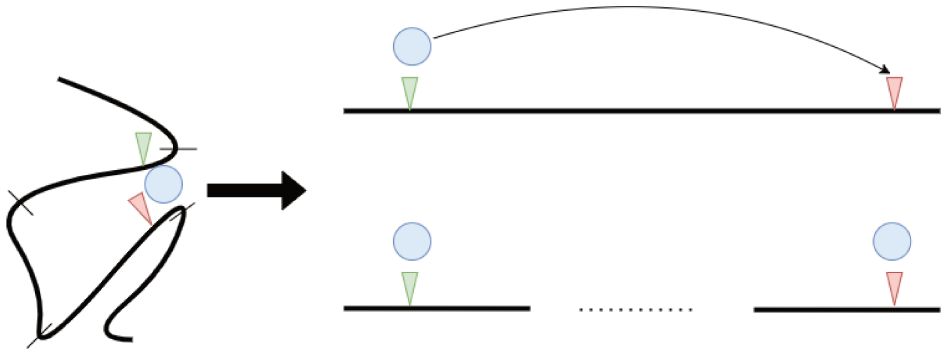


Figure 1.7: **Implications of a one-dimensional model.** The three-dimensional folding of the DNA polymer can place different positions next to each other that are far away in the one-dimensional sequence. This could allow a quick succession (first green triangle, then red triangle) of DNA-protein interactions that are not expected when considering only one-dimensional information, such as NGS data. We partitioned the DNA (e.g. at the small black lines) such that distances can be assumed to be sufficiently linear within that segment. Each partition is then analysed independently, therefore considering independent proteins per each fragment. The proteins are given in blue, and the DNA is displayed as a bold line.

a tool, and we surmise that they have little influence on the conclusions for NER. We refrain from a detailed description and refer instead to Bishop and Nasrabadi (2006). Lastly, we compare our work with existing computational DNA repair models (Subsection 1.3.4).

1.3.1 Stochastic Processes

A stochastic process is commonly defined as a collection of random variables X, which often vary in time t but can be equally dependent on other (multidimensional) properties such as space. If the nature of the process is known, they can be analysed with a given mathematical framework such as a Poisson point process.

A Poisson process is usually a counting process and is characterised by the Poisson distribution $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. The equation describes the probability of sampling an integer value k, e.g. the number of times an event is observed. λ is a shape parameter, and e is Euler's number. The temporal process (as a counting process) represents the succession of discrete and independent events (i.e. *points*), whose number follows the Poisson distribution. If the process is homogeneous in time, it is given by

$$P(X(t) = k) = \frac{(\lambda t)^k e^{-t\lambda}}{k!}. \tag{1.1}$$

Note that the shape of the distribution is now dependent on t. A Poisson point process can be equally applied in higher dimensions, where the intuitive representation over a one-dimensional time-line does not hold. More generally, a Poisson process is a collection of points randomly distributed in

a mathematical space.

If the underlying stochastic process is unknown, it is possible to determine the best functional descriptors by applying a functional Principal Component Analysis (fPCA). Related to the common PCA, it is a dimensionality reduction in Hilbert space which determines orthonormal eigenfunctions, given a number of basis functions that describe the functional data (e.g. B-spline or Wavelet). FPCA presumes a stochastic process $X(t)$ with mean $\mu(t)$ and noise $X(t) - \mu(t) = \eta(t)$. The latter can be represented by the sum over all orthonormal eigenfunctions $\phi_i(t)$, $i = 1, 2...$, which describe the maximal variance in X orthogonal to all $\phi_j(t)$, $j < i$. To be more precise, the Kosambi–Karhunen–Loève theorem states that every stochastic process can be represented as a linear combination of its eigenfunctions, i.e. $\sum_i \zeta_i \phi_i(t)$. We can therefore describe the noise by

$$\eta(t) = X(t) - \mu(t) = \sum_i \zeta_i \phi_i(t), \tag{1.2}$$

where ζ_i is the autocovariance operator

$$\zeta_i = \int (X(t) - \mu(t))\phi_i(t)dt. \tag{1.3}$$

By choosing only the first n eigenfunctions that explain most of the stochastic variance, we can approximate the process through

$$X(t) \approx X_n(t) = \mu(t) + \sum_k^n \zeta_i \phi_i(t). \tag{1.4}$$

Eq 1.4 reduces the functional dimensionality by combining the basis functions to their n major eigenfunctions.

1.3.2 Brownian Motion

Mathematical Description of Brownian Motion

Particle motion can be accurately described using Newtonian equations for every particle. However, this becomes computationally strenuous for even a small number of molecules. The Langevin equation (Lemons et al. (1908)) combines deterministic forces with noise to implement the apparently random particle collisions

$$m\frac{\partial^2 \mathbf{x}}{\partial t^2} = -\lambda \partial_t \mathbf{x} + \mathbf{F}(\mathbf{x}) + \boldsymbol{\eta}(t). \tag{1.5}$$

m denotes the particle mass, $\partial_t \mathbf{x} = \mathbf{v}$ is the velocity for a number of observed particles, $\mathbf{F}(\mathbf{x})$

describes the force field in which the motion occurs, and $\boldsymbol{\eta}(t)$ is the noise term. The introduction of $\boldsymbol{\eta}(t)$ reduces the number of molecules that need to be modelled for an accurate description. The stochasticity is also referred to as Brownian motion.

Eq 1.5 is a Stochastic Differential Equation (SDE). By integrating both sides, we find a description of the postional change over time

$$\frac{\partial \mathbf{x}}{\partial t} = \mathbf{A}(\mathbf{x}, t) + \mathbf{B}(\mathbf{x}, t)\boldsymbol{\sigma}(t), \tag{1.6}$$

where $\mathbf{x}$ and t denote position and time, respectively. The random variable $\boldsymbol{\sigma}(t)$ is the noise, incorporating uncertainty about the particle position and its interactions. $\mathbf{A}$ represents the deterministic component (called *drift term*), and $\mathbf{B}(\mathbf{x}, t)$ is the *noise term*. The latter describes amplitude and correlation of $\boldsymbol{\sigma}$. It is also called *diffusion term*, which we want to elaborate briefly.

In the following, we presume that there is no external force, i.e. $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. The left-hand side of Eq 1.5 represents the inertia, whereas $\lambda\mathbf{v}$ gives the friction. In the *limit of strong friction*, we suppose that $|\lambda\mathbf{v}| \gg |m\frac{\partial \mathbf{v}}{\partial t}|$. We can simplify Eq 1.5 to

$$\lambda\mathbf{v} = \boldsymbol{\eta}(t). \tag{1.7}$$

In other words, friction can be explained solely through random particle interactions. Similarly, Eq 1.6 becomes $\frac{\partial \mathbf{x}}{\partial t} = \mathbf{B}(\mathbf{x}, t)\boldsymbol{\sigma}(t)$. For the sake of simplicity, we only consider the one-dimensional case. Particles are distributed along the x-axis according to a distribution G, and they can move either to the left or right. We introduce the function $f(x, t)dx$ denoting the number of particles around a position x at time t wihtin the interval dx. In order to determine the temporal change of $f(x, t)$, we need to derive how many particles move into dx and how many move out. We define $\psi(\Delta_x, \tau)$ to be the probability that a particle moves the distance Δ_x within time τ (Figure 1.8). The number of particles at distance Δ_x that will move to x within τ are defined by $f(x + \Delta_x, t)\psi(\Delta_x, \tau)$. To calculate the total change of particles at x, we integrate over the entire spatial axis, i.e.

$$f(x, t + \tau) = \int f(x + \Delta_y, t)\psi(\Delta_y, \tau)d\Delta_y. \tag{1.8}$$

By applying the Taylor expansion for f to the first degree on the left-hand side and to the second degree on the right-hand side, we obtain

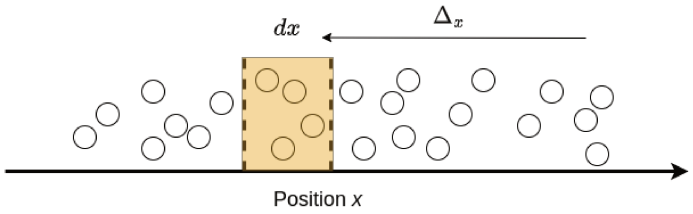


Figure 1.8: **Schematic description of diffusion with the motivation by Einstein.** Particles (circles) are distributed along the x-axis. Taking a container dx (orange) as a reference, we want to determine the change of the particle number $f(x,t)$ in dx. Consider the probability $\psi(\Delta_x, \tau)$ of a particle displacing by Δ_x (top arrow) within τ. By integrating over all possible distances Δ_x, we can determine the expected change $\partial_t f(x,t)$.

$$\begin{aligned} f(x,t) + \frac{f(x,t)}{dt}\tau &= \int \left(f(x,t) + \frac{\partial f(x,t)}{\partial x}\Delta_y + \frac{\partial^2 f(x,t)}{2\partial x^2}\Delta_y^2 \right) \psi(\Delta_y, \tau) d\Delta_y \\ &= f(x,t) \int \psi(\Delta_y, \tau) d\Delta_y + \frac{\partial f(x,t)}{\partial x} \int \Delta_y \psi(\Delta_y, \tau) d\Delta_y + \frac{\partial^2 f(x,t)}{2\partial x^2} \int \Delta_y^2 \psi(\Delta_y, \tau) d\Delta_y. \end{aligned} \tag{1.9}$$

Due to the fact that the probability of displacement is equally likely to the left and to the right (i.e. $\mathbf{F}(\mathbf{x}) = \mathbf{0}$), the term $\int \Delta_y \psi(\Delta_y, \tau) d\Delta_y = 0$. Moreover, the integral over a probability distribution is defined to be one. Therefore, we can simplify 1.9 to

$$f(x,t) + \frac{f(x,t)}{dt}\tau = f(x,t) + \frac{\partial^2 f(x,t)}{2\partial x^2} \int \Delta_y^2 \psi(\Delta_y, \tau) d\Delta_y, \tag{1.10}$$

which becomes

$$\frac{f(x,t)}{dt}\tau = \frac{\partial^2 f(x,t)}{2\partial x^2} \int \Delta_y^2 \psi(\Delta_y, \tau) d\Delta_y. \tag{1.11}$$

By setting $D = \frac{1}{2\tau} \int \Delta_y^2 \psi(\Delta_y, \tau) d\Delta_y$, we can write

$$\frac{f(x,t)}{dt} = D \frac{\partial^2 f(x,t)}{\partial x^2}. \tag{1.12}$$

The relationship in Eq 1.12 was revealed by Sutherland (1905), Einstein (1905), and von Smoluchowski (1906), and it is commonly known as Einstein diffusion equation. We want to provide some additional intuition. If we replace the derivative by the finite difference approximation, we obtain

$$\begin{aligned} \frac{f(x,t+\tau) - f(x,t)}{\tau} &= D \frac{f(x+\Delta_x, t) - 2f(x,t) + f(x-\Delta_x, t)}{\Delta_x^2} \\ &= \frac{2D}{\Delta_x^2} \left(\frac{f(x+\Delta_x, t) + f(x-\Delta_x, t)}{2} - f(x,t) \right) \end{aligned} \tag{1.13}$$

The right-hand side indicates that if the average number of particles around x is larger than at x, $f(x,t)$ increases, and vice versa. This refers to the observable diffusion phenomenon during which unequal particle distributions uniformise over time.

A general case of the Einstein diffusion equation describing the temporal change of the probability distribution over a particle position is given by the Fokker-Planck equation

$$\frac{\partial p(\mathbf{x},t|\mathbf{x}_0,t_0)}{\partial t} = \left(-\sum_i \partial_{x_i} A_i(\mathbf{x},t) + \frac{1}{2}\sum_{i,j} \partial_{x_i}\partial_{x_j}[BB^T]_{ij}(\mathbf{x},t) \right) p(\mathbf{x},t|\mathbf{x}_0,t_0). \tag{1.14}$$

Eq 1.14 can be derived similar to Eq 1.12. However, the Fokker-Planck equation allows $\mathbf{F}(\mathbf{x}) \neq \mathbf{0}$. This introduces a bias according to which the movement occurs. Consequently, the term $\int \Delta_y \psi(\Delta_y,\tau) d\Delta_y$ in Eq 1.9 can be unequal zero. By using Ito's calculus, Eq 1.9 can be transformed to 1.14. We refrain from presenting the derivation and instead refer to Schulten and Kosztin (2000). However, we want to emphasise some key assumptions. Most importantly, the noise $\boldsymbol{\sigma}$ in 1.6 is uncorrelated with zero mean (i.e. white noise). Moreover, Eq 1.14 is completely determined by the distribution $p(\mathbf{x},t|\mathbf{x}_0,t_0)$ at t. There is no temporal dependence on previous distributions. This is called a Markov process. It is presumed that the process is non-anticipative. This means that a random variable X can be adapted if and only if X_t is known at time t, which is why it is also known as *adapted process*. Intuitively, this means that there is no direct knowledge about X_t if not observed at t. Conveniently, there is no requirement that the system must be close to equilibrium. Eq 1.14 can be extended to include memory effects by convoluting over time:

$$\frac{\partial p(\mathbf{x},t|\mathbf{x}_0,t_0)}{\partial t} = \int_{-t_0}^{t} \left(-\sum_i \partial_{x_i} A_i(\mathbf{x},t-\tau) + \frac{1}{2}\sum_{i,j} \partial_{x_i}\partial_{x_j}[BB^T]_{ij}(\mathbf{x},t-\tau) \right) p(\mathbf{x},\tau|\mathbf{x}_0,t_0).d\tau. \tag{1.15}$$

It should be emphasised that the impact of memory effects is often negligible, and we only consider Markov processes in this work unless otherwise stated. A very general description of the evolution of probability distributions is given by the master equation. It assumes that the temporal change of the probabilistic combination of system states can be represented by a transition function between these states. This can be expressed for the discrete case in the following form

$$\frac{\partial p(x_n,t|\mathbf{x}_0,t_0)}{\partial t} = \sum_m \left(w(x_m \to x_n) p(x_m,t|\mathbf{x}_0,t_0) - w(x_n \to x_m) p(x_n,t|\mathbf{x}_0,t_0) \right), \tag{1.16}$$

where $w(x_i \rightarrow x_j)$ denotes the transition rate from x_i to x_j $(x_i, x_j \in \mathbf{x})$. This can be simplified to

$$\frac{\partial p(x_n, t|\mathbf{x}_0, t_0)}{\partial t} = \mathbf{W} p(\mathbf{x}, t|\mathbf{x}_0, t_0). \tag{1.17}$$

The Gillespie Algorithm

A stochastic simulation approach for a chemically reacting system based on the master equation was proposed by Gillespie (1977). Suppose a reactive system that is governed by M chemical reactions and N molecules. In order to stochastically simulate the temporal evolution, we introduce $p(\tau, \mu|\mathbf{x}, t)d\tau$, i.e. the probability of observing the reaction μ after time τ given the system state $\mathbf{x}$ at time t. The reaction probability of μ in an infinitesimal time step dt is given by θ_μ. By denoting the number of molecular combinations for μ with h_μ, we can define the sampling probability $a_\mu = h_\mu \theta_\mu$. Intuitively, if protein $D^{(1)}$ and $D^{(2)}$ participate in reaction μ, then the larger the number of $D^{(1)}$ and $D^{(2)}$, the more likely it is to observe μ. h_μ can be calculated for a bimolecular reaction by $h_\mu = [D^{(1)}][D^{(2)}]$, where the brackets denote the number of molecules. Gillespie (1977) defines $p(\tau, \mu|\mathbf{x}, t)d\tau$ as the joint probability of observing (or rather sampling) no reaction within time τ (i.e. $p_0(\tau|\mathbf{x}, t)$); and the probability that the subsequent reaction after τ within dt is μ (given by a_μ). By assuming independence, this can be formulated as the product

$$p(\tau, \mu|\mathbf{x}, t)d\tau = p_0(\tau|\mathbf{x}, t) a_\mu d\tau. \tag{1.18}$$

It is clear that the probability of no reaction within $d\tau$ is given through $p_0(\tau|\mathbf{x}, t)d\tau = \left(1 - \sum_\nu^M a_\nu\right) d\tau$. We define $a_0 = \sum_\nu^M a_\nu$. By comparing p_0 to a Poisson point process, we derive

$$p_0(\tau|\mathbf{x}, t) = \exp\left(-a_0 \tau\right). \tag{1.19}$$

When substituting Eq 1.19 in Eq 1.18, we obtain

$$p(\tau, \mu|\mathbf{x}, t) = \begin{cases} a_\mu \exp\left(-a_0 \tau\right) & \text{if } 0 \leq \tau < \infty \text{ and } \mu = \{1, ..., M\}; \\ 0 & \text{otherwise.} \end{cases} \tag{1.20}$$

Eq 1.20 describes the update probability of the chemically reacting system. τ and μ can be straightforwardly sampled, such that a computer simulation can be easily implemented. To be precise, given two random numbers r_1 and r_2 sampled over a unit-interval uniform distribution, we can

calculate

$$\tau = \frac{1}{a_0} \ln\left(\frac{1}{r_1}\right); \tag{1.21}$$

and μ is the integer which fulfills

$$\sum_{\nu}^{\mu-1} a_\nu < r_2 a_0 \leq \sum_{\mu}^{\mu} a_\nu. \tag{1.22}$$

The KJMA Model

In some cases, the impact of noise on a macroscopic scale can be largely ignored. It is then possible to find a phenomenological description of the process that is solely governed by Ordinary Differential Equations (ODEs). An example is the model for phase transitions in solids proposed by Kolmogorov (1937), Johnson and Mehl (1939), and Avrami (1939, 1940, 1941) (KJMA model). It is particularly interesting since the system has a solution for which the parameters can be conveniently estimated using linear regression (Subsection 1.3.3). Although it is a physical model, it has also been applied in biology to study DNA replication (Herrick et al. (2002)).

The KJMA model describes the phase transition from phase α to β by presuming random and uniform nucleation in untransformed material which is followed by isomorphic growth. Nucleation itself is a stochastic process by which material self-organises into structures (for example crystals). We presume that nucleation of new particles happens at a rate n, and growth occurs at speed G. The volume of transformed particles within the total volume V—by assuming that the entire sample is still untransformed (which is called *extended* volume)—is given by

$$dV_\beta^e = \omega G^m n V dt. \tag{1.23}$$

ω describes the space in which the transformation occurs, and m is the dimension of the space. For example, if growth can happen in all 3 dimensions, then $\omega = 4\pi/3$ and $m = 3$. However, if the processes has not just started (i.e. $t > 0$), only a fraction of Eq 1.23 can really occur, as material has already transformed to the new phase. The real transition can only happen within the volume $1 - V_\beta/V$, consequently $dV_\beta = dV_\beta^e\,(1 - V_\beta/V)$. With some straightforward algebra, we derive

$$f(t) = \frac{V_\beta}{V} = 1 - \exp\left(-kt^{m'}\right) \tag{1.24}$$

where k is the transformation rate, and $m' = m + 1$ is the Avrami exponent. Conveniently, the equation can be transformed to a linear regression problem to determine k and m'. By re-arranging Eq 1.24 and taking the logarithm twice, we obtain

$$\ln \ln \frac{1}{1-f(x)} = m' \ln t + \ln k. \tag{1.25}$$

Admittedly, any relationship with a biological process might appear far-fetched. However, the function generally describes the state transition of a substrate, e.g. from *damaged* to *repaired* DNA. We show below that by presuming particle movement in the nucleus, we can derive the same equation, which permits an interesting alternative interpretation of the CPD sequencing data (Chapter 3).

1.3.3 Parameter Estimation

Linear Regression

Suppose a function whose observed output y can be represented as a linear transformation of its input. i.e.

$$\mathbf{y}(\mathbf{X}) = \mathbf{w}^{\mathbf{T}}\mathbf{X} + b\mathbf{1} + \boldsymbol{\sigma}. \tag{1.26}$$

Here, $\mathbf{X} \in \mathbb{R}^{(m,n)}$ are m independent input variables of n measurements; $\mathbf{w} \in \mathbb{R}^m$ and b are weights and intercept, respectively; $\mathbf{1}$ is a vector only containing ones, i.e. $\{1\}^n$; and $\boldsymbol{\sigma}$ represents uncorrelated white noise in the data. We aim to find the parameter values for $\mathbf{w}$ and b that minimise the error of the model prediction $\hat{\mathbf{y}}$ to the observation $\mathbf{y}$. In particular, the parameters should minimise the mean squared error (MSE)

$$\begin{aligned} L(\hat{\mathbf{y}}, \mathbf{y}) &= \frac{1}{n}\sum_{i}^{n} (\hat{y}_i - y_i)^2 \\ &= \frac{1}{n}\sum_{i}^{n} \left(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b - y_i\right)^2 \end{aligned} \tag{1.27}$$

By minimising the error (also called *loss*), we find a solution for which the negative parameter derivative is 0, i.e $0 = -\partial_{\mathbf{w}} L$ and $0 = -\partial_b L$. In the following, we set $\mathbf{w'}^T = (w_0, w_1, ..., w_m, b)$ and $\mathbf{x'}_i^T = (x_0, x_1, ..., x_m, 1)$. We deduce for $\mathbf{w'}$

$$
\begin{aligned}
0 &= -\partial_{\mathbf{w}'} L \\
\iff 0 &= -\frac{2}{n}\left(\mathbf{w}'^T\mathbf{X}' - \mathbf{y}\right)\mathbf{X}'^T \\
\iff \mathbf{w}'^T\mathbf{X}'\mathbf{X}'^T &= \mathbf{y}\mathbf{X}'^T \\
\iff \mathbf{w}'^T &= \mathbf{y}\mathbf{X}'^T\left(\mathbf{X}'\mathbf{X}'^T\right)^{-1},
\end{aligned}
\tag{1.28}
$$

which gives a closed-form solution to determine $\mathbf{w}'$.

Stochastic Gradient Descent, Backpropagation and Other Parameter Learning Methods

Linear regression offers a way to derive parameters for a simple (i.e. linear) model. Unfortunately, most optimisation problems do not possess a closed-form solution, as variables of ODEs (i.e. $\partial_{\mathbf{w}'} L$) are commonly not separable. Therefore, other estimation methods are necessary. In the following, we introduce a class of learning algorithms that rely on iterative approaches, such as (stochastic) gradient descent (SGD). SGD is a training procedure that can be used to fit parameter values of a continuous function to a given set of data. The method presumes that: (a) data describing the function output are available; (b) the function itself is given and fixed (or a reasonable approximation); and (c) the function is continuous. SGD learning can be understood as follows. Although an equation does not have a closed-form solution, we know nevertheless that the negative loss derivative with respect to the parameters points towards the direction with the steepest error decline. By updating the parameters by a small increment towards the gradient (called *learning rate*, mostly denoted with α), the error is minimised over several iterations. Since the model is optimised over all data points—which contain noise—the gradient direction can change erratically. Statistically, however, the error is minimised with respect to the loss function L and the data. This is why it is commonly referred to as *stochastic* gradient descent. To reduce sensitivity of the parameter update to noise, it is possible to apply a *momentum*. Similar to the momentum in mechanics, it describes the dependency of the current update to previous updates, and it can be implemented as $\Delta\mathbf{w}(t_i) = \beta\Delta\mathbf{w}(t_{i-1}) + (1-\beta)\alpha\partial_{\mathbf{w}} L$. Here, β is a weight, $0 \leq \beta \leq 1$, governing the impact of previous parameter updates, and $\Delta\mathbf{w}(t_i)$ is the parameter update at iteration t_i. It should be emphasised that the stochasticity comes from the data and not from the method. However, a similar phenomenon (i.e. erratically updating model weights) can be also observed when parameters of the algorithm itself (called *hyperparameters*)—rather than parameters of the approximated function—are improperly set. For example, if the learning rate is set too large, the optimum can be easily missed. The gradient seemingly jumps around, possibly not converging at all (Figure 1.9(A)). Finding sensible choices is called hyperparameter optimisation.

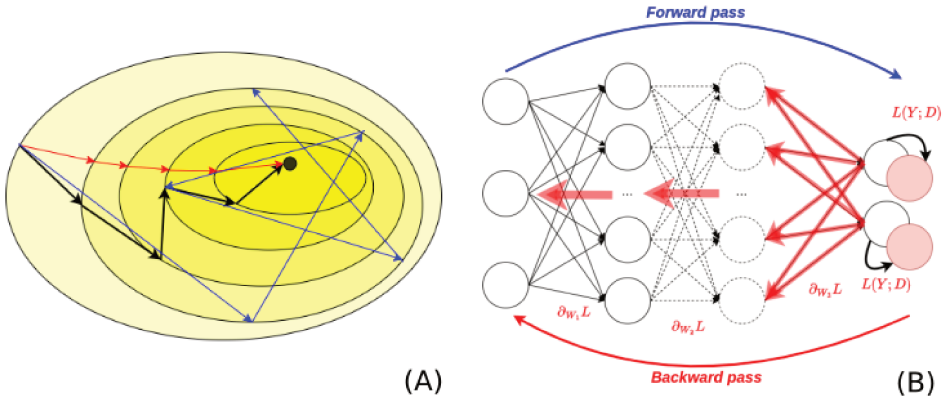


Figure 1.9: **Schematic explanation of concepts in machine learning.** (A) The yellow ellipses represent the error landscape. The darker the shade of yellow, the lower the value of the loss function L. Gradient descent iteratively minimises the error (red arrows). However, data contain commonly noise, which is why the gradients contain fluctuations (black arrows). This is why it is commonly referred to as stochastic gradient descent. If the learning rate is improperly chosen, the algorithm might not be able to converge (blue arrows). (B) Parameter update by backpropagation can be divided into forward and backward pass. Circles represent computational units (such as nodes in a neural network) including parameters that are not directly accessible by the input or output. Arrows represent connectivity. Dashed lines indicate that there might be varying numbers of these computational units. The forward pass calculates the current estimate based on an input and the learnt parameters. The difference is expressed by a loss function L. The gradient of the loss function with respect to the parameters W can then be passed back (propagated). This allows the update of parameters that are not directly connected to the functional input.

Calculating $\partial_{\mathbf{w}} L$ analytically can be difficult for complicated functions with many parameters. Fortunately, operations are often applied iteratively. This permits a straight forward implementation of the chain rule, during which the derivative is determined for a set of weights; their contribution is removed from the error; and subsequently, the updated error is passed *backwards* to the next set of operations. This approach is called backpropagation, and it is widely used, particularly for finding parameters in a neural network (Figure 1.9(B)). Adaptations include Backpropagation Through Time (BPTT), during which the parameters of a repeatedly applied function over several consecutive steps are determined. We refrain from showing a mathematical derivation and refer instead to common text books like Bishop and Nasrabadi (2006).

A downside of SGD approaches is that they assume differentiable equations and functions. Consequently, stochastic processes that rely on sampling cannot be estimated. Similarly, the algorithms fail when being positioned at non-differentiable points of the loss function. Other optimisation approaches such as Bayesian inference can be remedially applied. They are commonly implemented through parameter sampling. Well-known examples are Markov-Chain Monte-Carlo (MCMC) simulations that use Metropolis–Hastings sampling algorithms. They are commonly structured as follows. A large number of parameters are sampled based on an initial distribution which encodes our *prior* knowledge. By evaluating the function using the sampled parameters, it is possible to determine their

goodness with respect to an objective, e.g. minimising an error function. The parameter distribution is subsequently refined and the process is repeated. We refer for a more detailed description to Brooks et al. (2011).

1.3.4 DNA Repair Models of Other Studies

There have been several proposals to shine light onto DNA repair kinetics using a mathematical description of the process. One of the first computational models for DNA repair was based on BER dynamics in human cells (Sokhansanj et al. (2002)). By presuming stochastic effects to be negligible, they formulate a set of ODEs based on the Michaelis-Menten kinetics (Michaelis et al. (1913)). When comparing different hypotheses, their model predicted that cooperativity between repair complexes is necessary to describe the *in vivo* observation. This notion has been initially proposed by Hill et al. (2001), and it is today considered as an essential component of the BER pathway (Kladova et al. (2018)).

Politi et al. (2005) developed one of the earliest mathematical models specific for NER. By using imaging approaches with Green Fluorescence Protein (GFP) and gleaning data from the literature, they could find parameters that describe repair dynamics over time using a set of ODEs. Their model finds that a sequential protein assembly of the repair machinery is largely beneficial over a stochastic assembly for DNA repair speed.

Surprisingly, the results of Luijsterburg et al. (2010) oppose that point of view. By combining biological experiments with an ODE description, they analyse the timing of the repair protein assembly for the removal of 6-4PP in human cells. Through measuring the presence of fluorescence-tagged proteins as well as fluorescence loss though photobleaching, they determine protein dwell times and decline at damage sites. When combining them with CPD levels at various time points, they were able to deduce interaction rates by using an MCMC method. Their model favours a stochastic and reversible protein assembly that is guided through the repair program by irreversible enzymatic steps.

A comprehensive model that includes (possibly competing) NER and BER dynamics on a single-cell scale was introduced by Semenenko and Stewart (2005). Instead of relying on deterministic ODEs, they present a Monte Carlo model that reproduces the repair of ionising irradiation-induced damages in hamster cells as well as *Escherichia coli*.

Nevertheless, none of the models above incorporate the notion of space. Therefore, location-specific differences are chiefly ignored. Despite the progress, genome-wide computational descriptions for DNA repair are largely lacking. Unfortunately, it is notoriously difficult to retrieve information about dynamic interactions from static sequencing data. Microscopy images on a nucleotide resolu-

tion that would permit the measuring of ongoing location-specific dynamics are impossible to obtain with the currently available methods. Dion et al. (2007) proposed a workaround through using two different histone tags to measure competition for DNA association and nucleosome positioning. By fitting the parameters to a Poisson process, they were able to deduce nucleosome turnover rates. Similarly, Lickwar et al. (2013) applied competitive ChIP for determining binding dynamics for the TF Rap1. However, models that are not reliant on specifically adapted sequencing protocols have not been developed to our knowledge. It is therefore necessary to create new methods to establish the missing link between location-specific repair dynamics and static NGS data of nuclear processes.

1.4 Motivation

The accumulation of damages can lead to cell malfunctioning and premature cell death. Thus, deficient NER has been associated with several severe diseases, including a predisposition to cancer as well as neurological and ageing disorders (Sharma et al. (2020)). Some defects can be linked to a specific subpathway. For example, the GGR disorder Xeroderma Pigmentosum (XP) is characterised by UV hypersensitivity and sun-induced hypopigmentation and hyperpigmentation. The susceptibility to skin cancer is increased by more than a thousand-fold, and the risk of other tumour types is elevated as well (DiGiovanna and Kraemer (2012)). On the other hand, impaired TCR is associated with a great range of different symptoms, and the actual effects of TCR-specific diseases depend on various factors, such as accessibility to the lesion. In patients with the relatively mild UV-Sensitivity Syndrome (UVSS), other pathways such as GGR or BER remedially repair the lesion, as Pol II can be still removed from the damage site (Marteijn et al. (2014)). Severe forms of TCR deficiency include Cerebro-Oculo-Facio-Skeletal Syndrome (COFS) and CS, both of which are associated with premature aging, cession of growth, organ and neurodegeneration, as well as microcephaly and dysmyelination. The life expectancy of patients drops to between 2 and 12 years (Marteijn et al. (2014)). There is still an open debate whether the severe syndromes are a consequence of defective TCR (Vermeulen and Fousteri (2013)), dysregulated gene expression (Wang et al. (2014)), or both.

Despite the acknowledged associated disorders, the exact NER kinetics in living cells on the entire genome are not fully understood. This is especially pronounced with respect to other and possibly interacting nuclear processes such as transcription and nucleosome positioning. The advent of large-throughput NGS technology allowed the acquisition of many nuclear properties on a global scale, in particular DNA-protein interactions and damage distribution at various time points (Eyboulet et al. (2013); Mao et al. (2016); Li et al. (2018); Gopaul et al. (2022)). However, the actual analysis proves to be difficult due to the cellular complexity. This is even further complicated, as many data sets

contain only a few time points over several hours, making it strenuous to empirically derive NER dynamics and location-specific functioning. It is often necessary to glean heterogeneous data from different resources. Therefore, it is important to combine bottom-up mathematical modelling with data analysis frameworks in order to verify hypotheses and to fill-in missing information. The main objective of this PhD thesis is to develop different modelling techniques to assess holistically UV-induced CPD removal by NER in *Saccheromyces cerevisiae* as a model organism using NGS data.

Chapter 2

A Detailed Analysis of Nucleosome Coordination Along the Gene to Understand Implications for Sequence Accessibility

2.1 Context and Summary

Every scientific model needs to make assumptions and simplifications in order to describe the observed phenomena. A plethora of nuclear mechanisms interact with each other to permit cell survival and functionality. This is true under normal conditions as well as under stress. The complex interplay makes it difficult to identify factors that play key roles for lesion removal. The positioning of nucleosomes for example is pivotal to permit sequence accessibility. It is therefore conjectured to influence and to be influenced by various other processes in the nucleus, including Pol II presence and elongation (Koerber et al. (2009); Ocampo et al. (2016)) as well as DNA repair (Mao et al. (2016); van Eijk et al. (2019)). A clear picture of the dynamics to coordinate nucleosome phasing and gene-related processes is missing. There is a scientific consensus that arrangement largely relies on chromatin remodeler complexes, which can add, slide, or evict nucleosomes by using energy from ATP hydrolysis (Clapier et al. (2017)). However, it is not fully resolved how these remodeler complexes influence molecular processes along genes. This could have direct consequences for DNA repair. Indeed, as NER is a multistep process requiring DNA interactions with various proteins, it could be surmised

that coordinated nucleosome presence along a gene influences both TCR and GGR. By investigating positioning in non-irradiated cells, we can quantify a possible measurable effect of chromatin conformation on nuclear processes within the gene body; and consequently, whether nucleosome arrangement needs to be taken into account to explain lesion removal.

In this study, we combined classical Pearson correlation with position-specific fPCA to describe nucleosome dynamics along coding regions. By comparing MNase-seq data from chromatin remodeler-deficient strains (Ocampo et al. (2016, 2019)), we quantified their impact on phasing and spacing of multiple nucleosomes with respect to each other. FPCA permitted the identification of RSC as a key player to decouple arrangement between gene bodies, limiting the organisation strictly to the transcribed region. Correlating the distribution with other influencing factors in WT conditions suggested that chromatin remodelers render nucleosome positioning largely independent from sequence composition and presence of large protein complexes. However, interdependence with various properties—in particular Pol II occupancy—largely increased in *chd1*Δ strains, emphasising the important role of chromatin remodelers in WT cells. As our analysis indicates that remodeling complexes decouple arrangement from other genomic factors, we conclude that nucleosome phasing does not need to be taken specifically into account when investigating DNA repair in WT strains.

As lead author, I substantially contributed to the study conceptualisation and design. I implemented the analysis pipeline to determine and evaluate the fPCA as well as to measure correlation with other genomic factors. I guided paper writing and editing. I worked together with my colleagues to contact publishers.

A Genome-Wide Comprehensive Analysis of Nucleosome Positioning in Yeast

A Genome-Wide Comprehensive Analysis of Nucleosome Positioning in Yeast

Leo Zeitler[1], Kevin André[1], Adriana Alberti[1], Cyril Denby Wilkes[1 †], Julie Soutourina[1 ‡], Arach Goldar[1*]

1 Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

† **cyril.denby-wilkes@cea.fr**
‡ **julie.soutourina@cea.fr**
***** **arach.goldar@cea.fr**

Abstract

In eukaryotic cells, the one-dimensional DNA molecules need to be tightly packaged into the spatially constraining nucleus. Folding is achieved on its lowest level by wrapping the DNA around nucleosomes. Their arrangement regulates other nuclear processes, such as transcription and DNA repair. Despite strong efforts to study nucleosome positioning using Next Generation Sequencing (NGS) data, the mechanism of their collective arrangement along the gene body remains poorly understood. Here, we classify nucleosome distributions of protein-coding genes in *Saccharomyces cerevisiae* according to their profile similarity and analyse their differences using functional Principal Component Analysis. By decomposing the NGS signals into their main descriptive functions, we compared wild type and chromatin remodeler-deficient strains, keeping position-specific details preserved whilst considering the nucleosome arrangement as a whole. A correlation analysis with other genomic properties, such as gene size and length of the upstream Nucleosome Depleted Region (NDR), identified key factors that influence the nucleosome distribution. We reveal that the RSC chromatin remodeler—which is responsible for NDR maintenance—is indispensable for decoupling nucleosome arrangement within the gene from positioning outside, which interfere in *rsc8*-depleted conditions. Moreover, nucleosome profiles in *chd1*Δ strains displayed a clear correlation with RNA polymerase II presence, whereas wild type cells did not indicate a noticeable interdependence. We propose that RSC is pivotal for global nucleosome organisation, whilst Chd1 plays a key role for maintaining local arrangement.

Introduction

The eukaryotic DNA must be tightly wrapped into the spatially constraining nucleus. This is achieved in the form of chromatin, a DNA-protein complex within which the 1-dimensional DNA is condensed around histone octamers and folded to a 3-dimensional structure. To be more precise, these histone complexes are positively-charged multiprotein structures around which the DNA molecule is locally coiled, forming a linear organisation resembling the stringing together of beads. This is why the primary structure of chromatin is commonly represented by a so-called *beads-on-a-string* model. In yeast, a nucleosome refers to ≈147 base pairs (bp) of DNA that are wrapped around four histone units. Nucleosomes are closely spaced, with an averaged centre-to-centre distance of 165 bp, leaving roughly 15 bp of linker DNA between two adjacent histone complexes. There is a consensus that phasing is highly regular within coding regions, which is interrupted by Nucleosome Depleted Regions (NDRs) between two neighbouring genes. This

observation gave rise to the barrier model, which proposes that promoter-dependent properties (e.g. bound proteins or sequence composition) pose a limit for nucleosome assembly, and arrangement occurs with respect to this barrier [1, 2]. However, it is widely accepted that various factors establish and influence the genome-wide positional nucleosome landscape, including sequence composition, transcription, and chromatin remodelers [3–6]. Since the DNA molecule must bend to wrap around the histone octamer, the local nucleotide sequence naturally affects positioning. Generally speaking, GC-rich sequences are more flexible than AT-rich ones, and they are favorable to support the presence of a nucleosome [7, 8]. However, sequence-related properties might be dependent on specific motifs.

The condensed packaging also functions as regulator for various DNA-protein interactions. Most of these processes rely on chromatin remodeler complexes, which can—by consuming energy obtained from ATP hydrolysis—move, add, or evict the histone complexes to provide or inhibit direct access to the DNA sequence [9]. In yeast, chromatin organisation is maintained by four protein families, SWI/SNF, INO80, ISW, and CHD. The RSC remodeler complex of the SWI/SNF family is the only essential chromatin remodeler in *Saccharomyces cerevisiae*, and it is recruited to promoter regions where it is responsible for the maintenance of NDRs [10–12]. It has also been reported that the complex has an influence on nucleosome organisation in coding regions as well as supporting RNA Polymerase II (Pol II) elongation [13]. It is presumed to restore chromatin organisation after transcription [14]. However, RSC does not exhibit an impact on regular nucleosome spacing within the gene [14, 15]. Chd1—the only member of the CHD remodeler family in yeast—is associated with various transcription-regulating functions, including initiation, elongation, and termination [16]. It has been suggested that Chd1 stabilises perturbed nucleosomes during gene expression [17]. Isw1 and Chd1 are supposed to antagonise for nucleosome spacing within the gene, with Isw1 dominating profiles along genes with larger spacing, whereas Chd1 seems to control shorter spacing [12, 18]. It has been reported that deletion of Chd1 and Isw1 only disrupt inter-nucleosome distances and leave the +1 position unaffected [19]. Isw2 is similarly associated with regular spacing [20], and it is particularly affecting nucleosomes close to the NDR, which is presumed to regulate transcription [21]. However, the underlying mechanism for chromatin remodeling is still under debate, and a scientific consensus is missing [22–25].

Several studies showed an interdependence between nucleosome distribution and gene expression by using MNase-seq data, a Next Generation Sequencing (NGS) technique that allows the measurement of nucleosome profiles by using MNase digestion of purified chromatin [26, 27]. It has been suggested that high gene expression correlates with low nucleosome regularity [28] as well as extreme spacing (both short and long) [18]. There are contradicting results about the correlation between transcription and nucleosome phasing. Whilst [18], [29], and [30] report that transcription increases random positioning and weakened phasing, [28] show that nucleosome phasing of highly expressed genes is increased. The depletion of Pol II exhibited increased array regularity [31]. This phenomenon seems to be conserved across species, as indicated by studies using *Drosophila* [28] and mouse cell lines [32]. The outcomes indicate that gene expression can be partially explained by nucleosome positioning over the gene body. Nonetheless, the autocorrelation of MNase-seq profiles along genes revealed that nucleosomal organisation accounts for only ≈25% of the observed transcriptional variability, even though genes with similar regularity tend to have the same level of gene expression [33]. Surprisingly, many strains deficient for chromatin remodelers seem to show only a marginal effect on transcription [18, 19]. The only exception is *rsc8*-depleted cells, which exhibit a global decrease in gene expression [12]. A clear picture between nucleosome phasing and Pol II presence is still lacking.

Different approaches have been used to categorise collective nucleosome arrangement within transcribed regions using NGS data. However, many of them rely predominantly on measurements that describe only an average over the entire profile, such as Pearson [34] or autocorrelation measurements [33]. Another analysis that takes into account multiple nucleosomes upstream and downstream of the NDR was presented by [14]. However, the study focused on changes with respect to the NDR, and many phenomenological descriptions are based on the application of different

analysis techniques. In order to provide comparability of nucleosome positioning changes between various mutants, we aimed to use a single mathematical framework that can be applied to all strains. To our knowledge, a unifying approach assessing location-specific phasing properties along the entire nucleosome array over varying conditions has not been proposed, and a direct comparison of the effects in different remodeler-deficient strains is difficult.

In this work, we present a genome-wide analysis of collective nucleosome positioning along the gene. We define nucleosome positioning and phasing to be the positions of the MNase-seq signal peaks over an entire single nucleosome array. By clustering the MNase-seq signals of coding regions along 6-7 histone complexes into two groups using linear Pearson cross-correlation—which measures similarity of the entire nucleosome arrangement between each gene pair—we can categorise coding regions according to their likely phasing similarity imposed by chromatin remodelers. In order to interpret how profiles are classified into the two groups, we combined the clustering with an alternative data representation via functional Principal Component Analysis (fPCA). Whilst related to the conventional Principal Component Analysis (PCA), it assumes a functional relationship between positions along the profile, whereas PCA conjectures independence of every base pair along the gene. Therefore, fPCA implicitly considers spatial dependency, which is a fundamental assumption in common nucleosome phasing models like the barrier model, where nucleosomes phasing is coordinated with respect to a barrier and each other. FPCA is commonly used in time series and signal processing, and it has been used in biology for analysing crop yield [35], identifying child growth patterns [36], as well as studying genetic variation and the allelic spectrum [37]. However, it has never been applied to the spatial interdependence of nucleosome phasing to our knowledge.

The established Pearson clusters can be visually separated by considering only two fPCs, which are therefore sufficient to interpret the gene groups. Using our analysis, we can repeatedly investigate histone complex distributions of different chromatin remodeler-mutant strains using the same framework and interpret major differences along the entire nucleosome arrangement. By relating Pearson correlation with spatial properties along the profiles, our approach refines and complements other studies that focused either on a few individual nucleosomes close to the NDR or Transcription Starting Site (TSS); or which assessed only the average correlation of the entire array (e.g. via autocorrelation). Using MNase-seq data from yeast strains deficient for different chromatin remodelers [12, 18], we reveal that Rsc8 strongly limits coordinated nucleosome arrangement to the transcribed region. It might be therefore responsible for gene-specific phasing. By measuring how the Pearson cluster separation changes between mutants using a Support Vector Machine (SVM), we identified 5 combinations of gene deletions or protein depletions which have a notable impact on phasing properties compared to Wild Type (WT) conditions. Measuring correlation with other nuclear processes disclosed that none of the commonly assumed factors can easily explain long-reaching nucleosome arrangement in WT strains within the gene body. However, gene deletions—in particularly mutants that contained *chd1*Δ—caused a strong correlation with Pol II presence. Our results indicate a new mechanistic understanding of chromatin remodelers, where Rsc8 is responsible for long-range coordination and Chd1 for local positioning of nucleosomes. All customised source code was made available on GitHub (`https://github.com/leoTiez/nucleosome-fpca`) [38].

Results

Nucleosome Profiles Can Be Well Distinguished Based On Their Coordinated Positioning in WT

In order to compare nucleosome profiles over the gene body in WT conditions, we measured the pairwise Pearson cross-correlation of the MNase-seq data produced by [12, 18] for all protein-coding regions [39] using Eq 1. The Pearson correlation index is positive when the sequencing signals of both genes tend to change towards the same direction at the same position; and it is negative when

one profile is likely to increase whereas the other one decreases. Therefore, it compares similarity of the distributional shape—i.e. whether genes are apt to contain nucleosomes at similar positions—and it does not take the scaling of the sequencing data into account. The entire arrangement for each gene is treated as an entity. For both replicates, we considered 1000 bp after and 200 bp before the +1 position (= 1200 bp, approximately the average size of a gene in *Saccharomyces cerevisiae*), containing 6-7 nucleosome dyads.

Subsequently, Pearson coefficients were grouped into distinct partitions using k-mean clustering. In a nutshell, the algorithm divides a data set of m observations (here, pairwise Pearson indices over all genes) into k groups by minimising the variance within each cluster. Therefore, genes within a group tend to have nucleosomes at comparable positions, whereas profiles of different groups are likely to be less similar. Using a silhouette criterion measurement—which compares the similarity of an object to its own cluster with the similarity to other clusters—we determined that the Pearson coefficients are most distinctly divided when $k = 2$ (i.e. when having two groups, Fig 1(A)). By comparing the Jensen-Shannon (JS) distance of the Pearson clusters with the JS distance between 500 random group pairs using Eq 2, we proved their significance (outside the 95% prediction interval (PI) of a gamma distribution (Eq 3) estimated over the random partitions; SFig A.1). This shows that nucleosomal arrays can be significantly separated into two groups using linear correlation of MNase-seq data between genes (Figs 1(B, C)).

It is difficult to straightforwardly determine how the k-mean clustering algorithm distinguishes between these two groups; yet the interpretation of the discriminating boundary could reveal important insights about the nucleosome positioning that is presumably imposed by chromatin remodelers. As the data by Ocampo et al. [12, 18] contains several mutants, we want to identify this discriminator repeatedly with the same mathematical framework to make the results comparable. Due to the nature of the Pearson correlation index, we can make the following assumptions. As nucleosomes are commonly well positioned in budding yeast, the MNase-seq data resembles a wave-like function with one peak approximately every 200 bp. Moreover, single histone complexes cannot overlap in a single cell. The Pearson correlation measures therefore the average phasing similarity of the entire nucleosome array of two genes. Differences in similarity come either from shifts in exact positioning (i.e. well-defined peaks, Figure 1(D, left)) or from a change in the signal amplitude (i.e. increasing or decreasing MNase-seq magnitude over the profile or at particular locations, Figure 1(D, right)). The clusters must be separated based on either of these two trends, or possibly a combination of them.

In the following, we refer with *coordinated positioning* to the configuration of the entire nucleosome array, and consequently, to their behaviour with respect to the two separating trends of the k-mean clustering. Unfortunately, the Pearson coefficient measures only the average linear pairwise correlation over the entire profile, rather than taking position-dependent particularities into account. Therefore, simply extracting the boundary from the k-mean clusters does not explain whether the groups were established with respect to a shift or a change in amplitude (i.e. the previously determined discriminators). Instead, it is possible to investigate how the clusters distribute with respect to the data itself or a different description of it. By evaluating the major differences between the two groups of genes, we can interpret the separating clustering boundary and link it to particular properties along the nucleosome profile.

Conventional approaches apply dimensionality reductions like PCA to visually analyse clustering distributions. However, using PCA would implicitly mean that we assume independence between every position along the gene. By using the Pearson correlation measurement, we treat every profile as a single entity, which would be violated by the independence conjecture. This also contradicts the fundamental assumption of the barrier model where the positioning of earlier nucleosomes affect later phasing. Instead, we understand the arrangement as the result of a coordinated process. We assume that the MNase-seq signal along each gene can be described as a single (unknown) continuous function, which can be approximated by a mixture of a finite number of known simpler functions (so-called basis functions). In this study, we used 20 B-Splines to represent the MNase-seq data along each gene, which were subsequently averaged to a mean profile. This permitted the

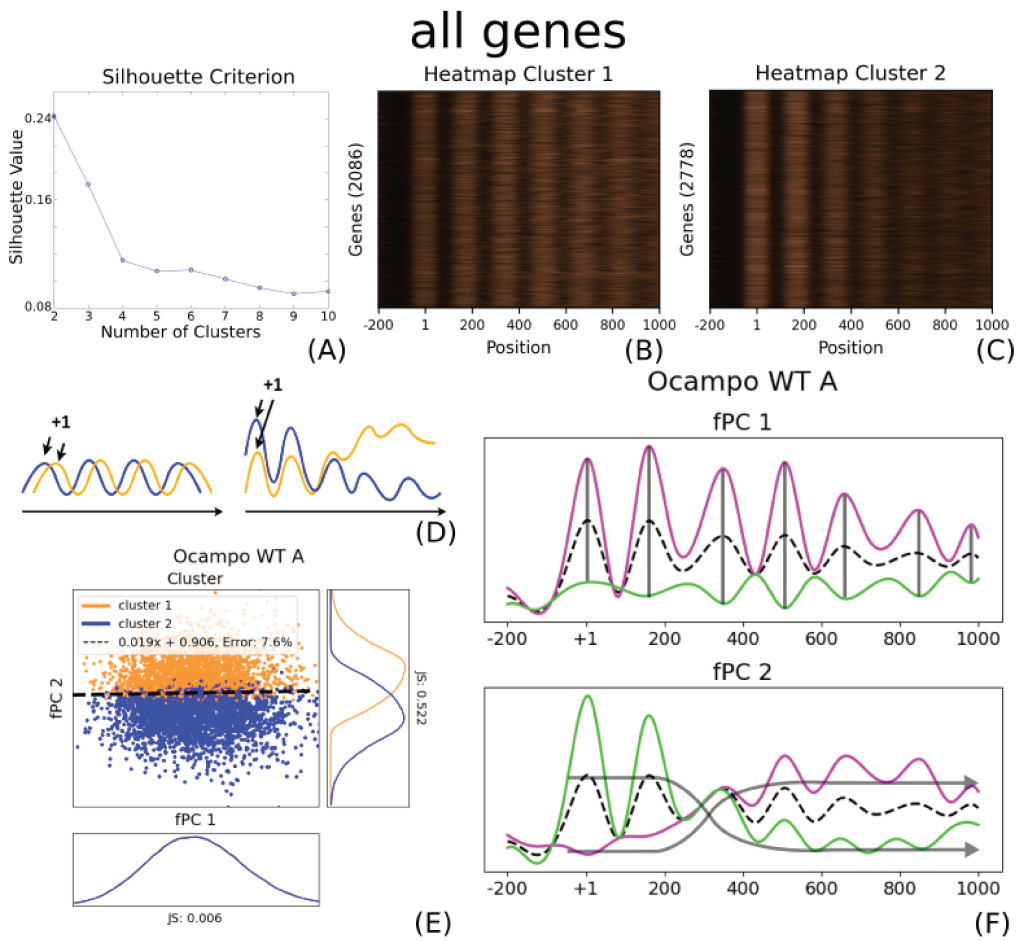


Figure 1: Pearson clusters and fPCA considering all protein-coding genes. (A) The silhouette plot clearly indicates that the data can be best divided into two clusters, and creating more groups would only decrease the difference between each cluster. (B) and (C) display the profiles for each cluster. However, it is difficult to quantify visually why these clusters were established. This is particularly true because the Pearson index measures only general trends in the profile, and it does not take the scaling into account. Each row represents a gene, and the x-axis shows the position along the coding region, with the +1 nucleosome defined to be at position 0 bp. The colour code represents MNase-seq amplitude, i.e. copper values show large MNase-seq signal values, whereas dark areas indicate a low amplitude. (D) The cartoon presents the hypothesised differences that could occur between the Pearson clusters. Due to the well-positioned nucleosomes and the wave-like structure of MNase-seq data, we presume that the Pearson correlation measures coordinated nucleosome positioning along the gene. If two profiles (orange and blue) are in two different clusters, this could indicate either a shift in the exact nucleosome positions (left); or a general trend in the MNase-seq signal amplitude, i.e. either increasing or decreasing (right). (E) Pearson clusters considering all genes are linearly separable with respect to their fPC scores. This indicates that two fPCs are sufficient to interpret the gene groups. The JS distance between the cluster distributions is much larger for fPC 2 than for fPC 1. Orange and blue indicate each one group, the dashed line symbolises the best linear separation using a SVM. The x-axis represents the score of the first fPC ζ^1, the y-axis gives the score for the second fPC ζ^2. Both axes are scaled to the same range. (F) When analysing the effect of the major fPCs, they describe predominantly position-dependent scaling (transparent black lines, fPC 1) and collective nucleosome phasing (transparent black arrows, fPC 2). The second fPC in WT indicate an increasing or decreasing signal magnitude as a function of distance from the TSS, suggesting stronger or weaker presence (corresponding to Figure 1(D, right)). The mean is given as a dashed black line, a positive contribution—i.e. adding the fPC to the mean—is displayed in magenta, and a negative contribution—i.e. subtracting from the mean—is shown in green. Trends over the entire array are indicated by grey arrows. When exact positions were seemingly not affected by the fPC, we marked the positions with a grey vertical bar. See Methods for more information about how the plots were produced.

application of fPCA to determine the two best-characterising functional Principal Components (fPCs) that describe each nucleosome arrangement. It incorporates specific assumptions about the spatial relationship in the distribution through the basis functions, which is the crucial difference between conventional PCA and fPCA. To be more precise, the establishment of the MNase-seq distribution is understood as a stochastic process with a mean behaviour. Each considered nucleosome array can be regarded as a realisation of this stochastic process with a deviance from the expected average distribution. Instead of defining a data representation for every gene individually, fPCA determines how the mean profile needs to be transformed to approximate a particular gene. This transformation is found by combining the basis functions over all coding regions to more complex functions that are orthonormal to each other and describe the most variance along the data (i.e. the fPCs, Eq 6). These functions transform the mean by adding them to the average profile with a gene-specific scaling factor (i.e. ζ_i^j for the j-th fPC of the i-th gene). Consequently, every nucleosome array can be also described exclusively by the factors ζ_i^j together with the respective fPCs, and we can evaluate how the two Pearson clusters distribute with respect to these factors.

Interestingly, the two clusters—which were independently obtained by classical hierarchical k-mean clustering of Pearson coefficients—are visually neatly separated by using only the first two fPCs, indicating that they are sufficient to quantify the difference between the two sets of genes (Fig 1(E)). In fact, the separating boundary is almost exclusively dependent on the second fPC, whilst it is seemingly independent of the first. This is slightly less clear for the B replicate, although still distinct (SFig A.2(B)). Using our previous considerations about how the algorithm establishes the two clusters, we intuited that the second fPC describes coordinated nucleosome phasing along the gene body. By analysing the effect of the second fPC on the function shape, we conclude that the clusters are determined based on the downstream presence of nucleosomes (corresponding to the right cartoon in Fig 1(D)). We found that the first fPC largely represents amplitude scaling at a given position, as it does not influence the location of the peak (Fig 1(F)). The analysis shows that position-dependent amplitude scaling and coordinated arrangement are the best two independent functional descriptors for the MNase-seq data. Despite the fact that the ratio of explained variance is not high (21.4% and 11.5% for fPC1 and fPC2, respectively), they are completely sufficient to distinguish between the Pearson correlation groups and permit an interpretation of the linear separating boundary between the clusters.

FPCA Reveals Size-Dependent Rsc8-Mediated Phasing of Nucleosome Positions

Since the smallest genes are ≈300 bp long, the 1000 bp window after the +1 position can contain much more than the actual length of the coding region. In order to analyse how nucleosome phasing is affected by the gene size, we repeated the fPCA considering exclusively small ($\leq$ 1000 bp, ≈26.7%) or large genes ($>$ 1000 bp, ≈73.3%). Consequently, the mean as well as the two fPCs changed, whilst we kept their allocation to the previously determined Pearson clusters the same (in the following also referred to as *all-gene* clusters). If coordinated positioning is substantially affected by the length of the transcribed region, we expected that the factors ζ_i^j of the two major fPCs should exhibit a changed behaviour with respect to the linear separability. We can confirm that the linear separation is preserved for large genes, although the boundary becomes slightly sloped (SFig A.2(C, D)). The fPCs for only large genes are almost identical to the *all-gene* fPCs (SFig A.4). We therefore presume that the clusters can be still largely separated by the second fPC. We also considered a possible impact of the downstream NDR by analysing exclusively very large genes ($\geq$ 3000 bp, ≈11.5%). Once again, the boundary was clearly visible (SFigs A.2(G, H)). We concluded that the MNase-seq distribution over the first 6-7 nucleosomes of all genes larger than 1000 bp can be best clustered by the collective positioning, and it can be surmised that phasing within the gene body is only negligibly affected by the downstream NDR or nucleosomes outside the 1000 bp window.

However, the neat separation between the two clusters fully vanished for small genes (Fig 2(A), for replicate B SFig A.2 (E)). Almost all data points belong to the same group, although both are

present. We want to remind that clusters were established using all coding regions, whereas the functional representation depends now exclusively on genes smaller than 1000 bp. The newly determined fPCs include overlapping positioning inside and outside the gene body due to their varying size (SFig A.3). The fact that the clusters are not separable indicates that coordinated nucleosome phasing disappears after the Transcription Termination Site (TTS), and we hypothesised that the arrangement is strictly limited to the gene body. Indeed, the second small-gene fPC indicates well-defined positioning only for up to the +2 nucleosome (≈300 bp), and the function loses quickly its frequent wave-like shape thereafter (Fig 2(D)). The two major fPCs for small genes are not sufficient anymore to separate the *all-gene* clusters, which are discriminated by the presence of downstream nucleosomes. To verify our hypothesis of gene-size dependent phasing, we divided the regions into small and large genes *before* performing the Pearson clustering. When considering exclusively small genes, the two Pearson groups become linearly separable again, which is—in accordance with our hypothesis—predominantly determined by the size (SFig A.5). This shows that the nucleosome arrangement is strictly limited to the gene body.

The data produced by [12,18] contain two replicates for *chd1*Δ, *isw1*Δ, and *isw2*Δ cells as well as *rsc8*-depleted strains, together with their combinations as double, triple, and quadruple mutants. In order to analyse how gene-size dependent nucleosome phasing alters in varying contexts, we compared the small-gene fPCs in mutant and WT conditions. Surprisingly, the separation of the *all-gene* clusters was clearly visible for the fPCs of small coding regions in *rsc8*-depleted strains (Fig 2(B)). Indeed, the average MNase-seq profile exhibits phased peaks along the entire 1000 bp-window (Fig 2(E)), and nucleosome positioning continued outside the gene boundaries (SFig A.3). The linear separability of the *all-gene* clusters using small-gene fPCs can be found in almost all mutants which are depleted of Rsc8 (SFig A.6), with the sole exception of Rsc8-depleted *chd1*Δ strains (2(C), replicate B SFig A.7(B)). Here, the groups cannot be visually separated by ζ^1 and ζ^2, and the determined fPCs resemble small-gene fPCs in WT conditions (Fig 2(F), replicate B SFig A.7(D)). This indicates that the gene-specific boundaries for nucleosome phasing can be re-established, and the second fPC loses its wave-like shape again after the +2 position (SFig A.3). Consequently, we hypothesise that Chd1 and Rsc8 have partially antagonistic roles for maintaining chromatin organisation that distinguishes transcribed from non-transcribed regions. Taken together, this analysis exhibits strictly constrained and Rsc8-mediated nucleosome organisation within coding regions.

Nucleosome Phasing Changes In Remodeler Mutants

We were particularly interested in how nucleosome remodeler complexes affect coordinated phasing. To remove the gene size-dependent bias from the clustering and the established fPCs, we applied the Pearson clustering to exclusively large genes (> 1000 bp) for all strains and determined their two major fPCs (SFigs A.2(C, D); A.4). We can confirm that the created groups for all mutants were again significant (outside the 95% PI of a gamma distribution for the JS distance over 500 random group pairs), with the exception of *isw1*Δ*rsc8* replicate *B* (SFig A.8). We consequently removed this value from the analysis. Interestingly, the Pearson clusters were always visually separated by using solely the first two fPCs, although some strains exhibited a larger overlap between the groups than others (SFigs A.9 and A.10 for replicate A and B, respectively). This suggests that coordinated phasing in all mutants can be represented by considering only the two fPCs that describe the most variance, and including more fPCs is not necessary in order to interpret the discriminating function.

The respective contribution of the two major fPCs to separate the clusters varied between the cell strains, suggesting that fPCA is sufficiently sensitive to capture strain-dependent consequences (SFigs A.9 and A.10 for replicate A and B, respectively). This caused the slope of the discriminating boundary to tilt. Therefore, the transformations of the mean distribution (i.e. fPCs and their factors ζ_i^j) changed for these strains. This indicates that they had not only a global effect on the average MNase-seq profile, but also caused a gene-specific disruption of the nucleosome positioning. We deemed those strains particularly important that altered the gene-specific collective behaviour of the nucleosome distribution with respect to the WT. We determined the slope for all strains using a

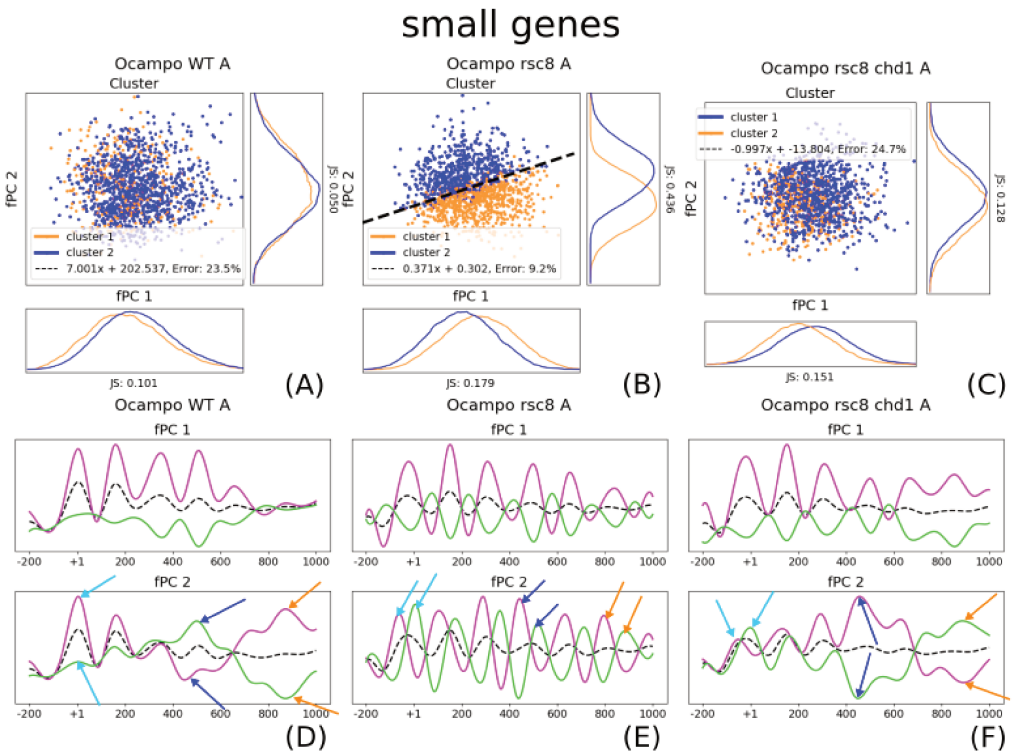


Figure 2: Nucleosome phasing is strictly limited to the gene body, which is maintained by Rsc8 but antagonised by Chd1. The cluster distribution plots in (A), (B), and (C) show the distribution of both gene groups with respect to the small-gene fPCs of WT, *rsc8*-depleted cells, and *chd1*Δ strains. Orange and blue indicate the two clusters, and the black dashed line shows the separating boundary determined by a linear SVM. The histograms present the cluster distribution with respect to each axis. Figures (D), (E), and (F) display the transformation of the average small-gene nucleosome profile by the two major fPCs for WT, *rsc8* depletion, and *chd1*Δ, respectively. The dashed black line as well as the solid lines in magenta and green display the mean, a positive contribution of the fPC, and a negative contribution. Turquoise arrows indicate the effect on the +1, dark blue arrows on the +4, and orange arrows on the +6 position. (A) When plotting the cluster distribution with respect to small-gene fCPs in WT, the linear separability is lost. (B) The fPCs of the *rsc8*-depleted strain maintain the linear separability, despite the fact that the groups were established for all genes. As we interpret the Pearson clusters as similarity in positioning between genes of 1000 bp mediated by chromatin remodelers, it possibly suggests that positioning outside coding regions influences nucleosomes inside and vice versa. (C) Whilst most mutants that were *rsc8* depleted could discriminate between the all-gene clusters using small-gene fPCs, this separability is lost again in *rsc8*-depleted *chd1*Δ, revealing partly antagonistic roles to maintain gene-specific phasing for Rsc8 and Chd1. (D) The effect of two fPCs sheds light on why the Pearson groups are not linearly separable in WT using small-gene fPCs. The the distribution of the second fPC loses its wave-like form after the +2 nucleosomes, which is approximately the size of the smallest genes in budding yeast. (E) Nucleosome positioning in *rsc8*-depleted conditions is clearly visible along the entire considered region, despite the included genes being smaller. This suggests that gene-specific nucleosome arrangement cannot be maintained. It is of note that the phasing also changes for the +1 nucleosome, and the NDR can be seemingly not conserved. (F) *rsc8*-depleted *chd1*Δ on the other hand loses the wave-like form of its second fPC after the +2 nucleosome, indicating the presence of gene-specific nucleosome profiles as in WT conditions. All axes are scaled to the same size for each strain; shapes and amplitudes are therefore comparable (see Methods for more details).

linear SVM. As aforementioned, the boundary is tilted when only considering large genes in WT conditions (Fig 3(A)), and the two available replicates differ slightly. The observed deviation

between replicates was used as a reference for the anticipated variability in the data. By using Eq 9, we determined chromatin remodeler-deficient strains that had a sufficiently different linear boundary with respect to the WT.

We provide three different perspectives on the data. Firstly, the cluster distribution with respect to the factors ζ_i^j together with the slope highlight mutants that particularly disrupt gene-specific collective nucleosome phasing. In the following, clusters are always indicated using the colours orange and blue. Secondly, analysing the transformation of the two major fPCs of the mean unlocks an additional understanding of the variance present in the data and allow quantifying the general impact of chromatin remodeler deficiency. Here, we plot a positive contribution to the mean in magenta and a negative contribution in green. Lastly, the location-specific effect of the discriminator links spatial properties to the Pearson gene clusters, which describe likely similarity of nucleosome positioning mediated by chromatin remodelers. The impact of the discriminator is in the following given as a grey area around the mean, indicating more important regions when the margin is larger. The median profile of each cluster using the determined fCPs will be given again in orange and blue. This allows a comprehensive analysis of the impact of gene deletions or *rsc8* depletion with respect to the WT.

We can determine the importance of particular positions to separate the clusters as follows. The slope of the SVM indicates the contribution of each fPC to separate the clusters. For example, a 0° angle shows that the descriminator can be solely described by the second fPC; 45° suggest an equal contribution of both fPCs to separate the clusters; and 90° indicate that collective nucleosome phasing is exclusively dependent on the first fPC. Consequently, by linearly combining both fPCs together as implied by the slope (Eq 11), we can evaluate which positions along the profile are particularly important for the classification. Indeed, understanding the separating boundary is not straightforward. Although the median profiles for each profile can differ substantially at some positions, this variance might be less important for separating and interpreting the clusters (e.g. the +2 nucleosome in WT conditions, Fig 3(C)). Reciprocally, whilst the median profiles for both groups can be very similar, the variance over all considered genes at this locus could be much larger and therefore play an important role for the classification (e.g. the +3 position in WT strains Fig 3(C)).

We identified 5 mutants—namely *chd1*Δ, *isw2*Δ*chd1*Δ, *rsc8*-depleted *chd1*Δ, *isw1*Δ*isw2*Δ, and *rsc8*-depleted *isw2*Δ*chd1*Δ—that evoked notable changes considering the experimental variability between replicates (Fig 3-5). For a correct interpretation of the results, it is crucial to highlight that this does not imply that other mutants had no effect on the profile. Rather, this suggests that the considered mutation caused a gene-specific change of nucleosome phasing regulated by chromatin remodelers, which we assume is represented by the deviance of the stochastic process (i.e. the variance to describe the MNase-seq profiles). Other gene deletions can have other impacts that do not disrupt the gene-specific collective positioning. All measurements are given in Table 1.

Most single mutants had only a small or negligible effect on the collective nucleosome phasing along transcribed regions, with the exception of *chd1*Δ (Fig 3(D-F)). Indeed, the boundary was notably tilted with respect to WT conditions (Fig 3(D)). This suggests that the functional composition of the MNase-seq signal changed. In fact, the amplitude of the second fPC decreases more quickly along the gene body in *chd1*Δ mutants, and the variance of the peak at the +2 position strongly diminished (Fig 3(E)). When interpreting the effect of the discriminating boundary, we observe that the +1 and +2 nucleosomes only exhibit a small importance for establishing the clusters, whereas the impact of the NDR and later nucleosomes increased (Fig 3(F)). Consequently, the +1 position remains largely unaffected. As Chd1 is responsible for nucleosome spacing along genes and is particularly involved in maintaining chromatin integrity during Pol II elongation, it is intuitive that the *chd1*-deletion influences phasing within the gene body. This outcome shows the clear effect of chromatin maintenance by Chd1 after the +2 nucleosome, whilst leaving the +1 position well preserved.

The double mutant *isw2*Δ*chd1*Δ exhibited also a noteworthy shift of the separating boundary (Fig 4(A)), yet with different results to the *chd1*Δ single mutant. The second fPC seemingly preserves its wave-like shape (Fig 4(B)). This indicates that nucleosome presence is less perturbed,

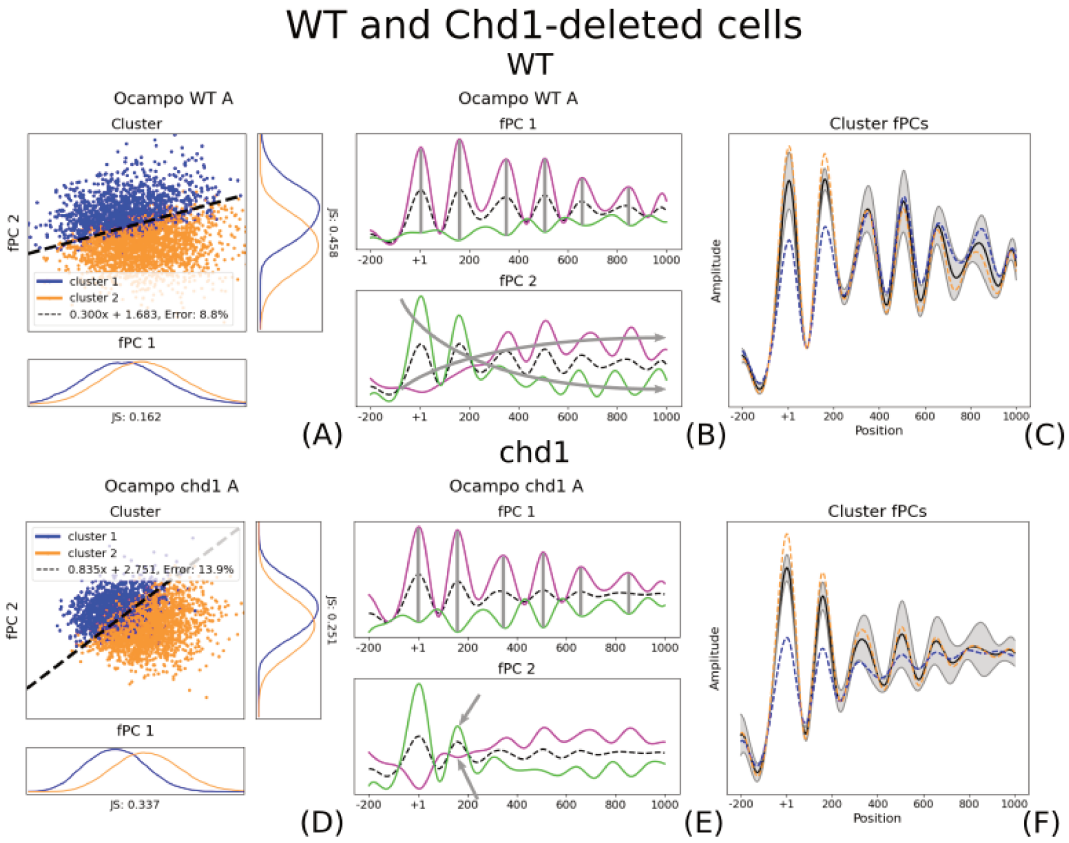


Figure 3: The fPCs, their gene specific scores, and the discriminating boundary explain collective phasing and how this changes in *chd1*Δ with respect to WT conditions. The figure shows the cluster distribution with respect to ζ_i^j, the impact of the determined fPCs, and the location-specific impact of the separating boundary for WT (i.e. (A), (B) and (C)) and *chd1*Δ conditions (i.e. (D), (E), (F)). (A) and (D) show the fPC scores ζ_i^j of WT and *chd1*Δ strains, respectively. For the latter, the boundary slope changed notably (black dashed line). As indicated by the fCPs in (B) and (E) for WT and *chd1*Δ, respectively, the functional description of the data changes. Indeed, the second fPC of *chd1*Δ abates quickly after the +1, with a strong effect on the effect of the +2 (grey arrows). The dashed black line as well as the solid lines in magenta and green indicate the mean, a positive contribution of the fPC, and a negative contribution, respectively. When exact positions were seemingly not affected by the fPC, we marked the positions with a grey vertical bar. General trends are given in grey arrows along the gene. The location-specific impact of the separating boundary is given in (C) for WT and (F) for *chd1*Δ strains. Interestingly, despite the median distributions of the clusters (blue and orange) are clearly different with respect to the +1 and +2 in WT conditions, later positions are much more important for allocating a profile to a particular group (grey areas, mean in black). Whilst this is also true for *chd1*Δ, the importance of later nucleosomes is even more accentuated, whereas the influence of the +1 and +2 positions are further decreased. All axes are scaled to the same size for each strain; shapes and amplitudes are therefore comparable (see Methods for more details).

and peaks are comparatively well positioned. Similar to the *chd1*Δ single mutant, both of the fPCs strongly contribute to distinguish between the Pearson clusters. The discriminating function exhibits similar local effects as the *chd1*Δ strain, but the positions after the +2 nucleosome clearly indicate an additional shift which contributes to the separation (Fig 4(C)). Interestingly, *rsc8*-depleted *chd1*Δ significantly decreases the slope tilting (instead of accentuating it), therefore making coordinated phasing almost exclusively dependent on the second fPC (Fig 4(D)). This can be better understood when analysing their respective effects (Fig 4(E)). The first fPC solely explains average signal

	WT	chd1	isw1	isw2	rsc8	isw1/chd1	isw2/chd1	chd1/rsc8
A	0.299	0.834	0.117	0.133	0.377	0.213	1.406	0.031
B	0.055	0.48	0.329	0.038	0.08	0.283	0.538	0.074
Mean μ	0.177	0.657	0.223	0.0855	0.2285	0.248	0.972	0.0525
s	0	**2.6674**	0.0409	0.3612	0.0366	0.2951	**2.9842**	**1.4773**
	isw1/isw2	**isw1/rsc8**	**isw2/rsc8**	**isw1/isw2 chd1**	**isw1/chd1 rsc8**	**isw2/chd1 rsc8**	**isw1/isw2 rsc8**	**isw1/isw2 chd1/rsc8**
A	1.452	0.347	0.153	0.112	0.216	0.057	0.466	0.066
B	1.074	~~0.072~~	0.567	0.295	0.207	0.068	0.245	0.174
Mean μ	1.263	0.2095	0.36	0.2035	0.2115	0.0625	0.3555	0.12
s	**12.7873**	0.0157	0.3315	0.0157	0.5420	**4.8846**	0.5909	0.1233

Table 1: SVM boundary slopes for both replicates. The first two rows give the boundary slope for replicate A and B, respectively. Mean μ is the mean slope for both. The **s** value represents our significance measurement defined in Eq 9. Noteworthy changes of the boundary slope are marked in green (bold), all others are red. The B replicate of *isw1*Δ*rsc8* was not significant, and the value was removed from the analysis (crossed out). The s-value in WT is per definition equal 0.

amplitude (which is not measured by the Pearson correlation index) and hence contains almost no information about coordinated positioning. As expected, the local effect of the discriminating boundary follows the trend described by fPC2 (Fig 4(F)). The second fPC also indicates that the NDR before the +1 cannot be maintained (see arrow in Fig 4(E) and the grey area in NDR and +1 position in Fig 4(F)), which is in line with other studies [12, 40]. Remarkably, all nucleosome positions along the entire array seem to be important for the classification—particularly the first two—which is not the case for the other two double mutants. It should be noted that not all double mutants that include *chd1*Δ show a similarly notable tilting of the slope as the single mutant. This could possibly mean that these double mutants have opposing effects, although it is difficult to give a clear indication with the variation between only two replicates. We found an interesting behaviour for *isw1*Δ*isw2*Δ (Fig 4(G)). The effect of the second fPC hints that the positioning of the +2 is strongly impacted, and following phasing becomes inharmonious (Fig 4(E)). The +1 is kept well positioned. The first fPC, on the other hand, resembles the first fPC of the *isw2*Δ*chd1*Δ mutant, with a minor difference at the +3 nucleosome (compare Fig 4(E) with Fig 4(B)). Indeed, when analysing the location-specific properties of the separating function (Fig 4(I)), nucleosome profiles in the *isw1*Δ*isw2*Δ strain seem to be clustered particularly with respect to a shift at the +3 and +4 position. This shift is apparently slightly corrected thereafter and becomes less important. Whilst seemingly similar, a shift in the *isw2*Δ*chd1*Δ strain after the +2 position remains important for the entire arrangement to determine the gene groups (compare Fig 4(I) with Fig 4(C)). This indicates that Chd1 and Isw1 contribute differently to nucleosome phasing in *isw2*Δ conditions, with the effect of Isw1 being possibly more confined. Taken together, these results show that double mutants can have varying and non-linear effects.

Among the triple and quadruple mutants, the only one that changed notably the clustering boundary is *isw2*Δ*chd1*Δ*rsc8* (Fig 5(A)). Once again, tilting is decreased. The effect of the fPCs and the separating boundary is almost identical to the *chd1*Δ*rsc8* mutant, suggesting that *isw2*Δ does not have a strong effect on the phenomenon (Figs 5(B, C)). However, it should be mentioned that the variability between the two replicates is considerably large, as the two clusters can be only neatly separated in replicate B, whereas replicate A exhibits a great overlap. Whilst the result in the latter replicate could suggest that more fPCs are necessary to interpret the gene groups, the results for replicate B indicate that sufficient information is preserved in the first two fPCs. More replicates would be needed to provide an answer. We also want to highlight that mutants with more than two gene deletions exhibited less clear nucleosome peaks, and a straightforward interpretation of the

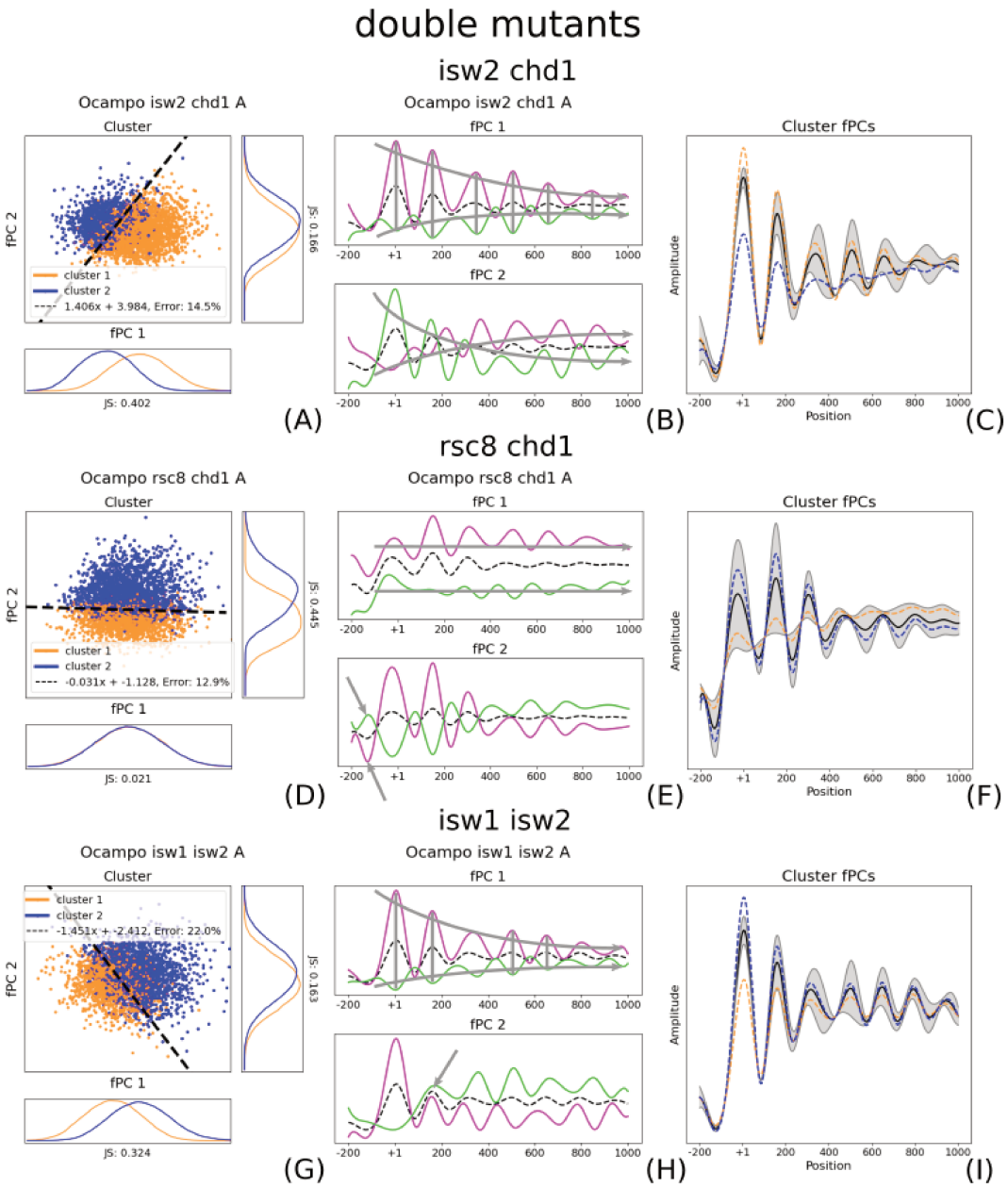


Figure 4: The fPCs, their gene specific scores, and the discriminating boundary explain changing collective phasing in double mutants. The figure shows the cluster distribution with respect to ζ_i^j, the impact of the determined fPCs, and the location-specific impact of the separating boundary for all double mutants, in particular *isw2*Δ*chd1*Δ (i.e. (A), (B) and (C)), *rsc8chd1*Δ (i.e. (D), (E), (F)), and *isw1*Δ*isw2*Δ (i.e. (G), (H), (I)). The linear separation of the cluster distribution with respect to factors ζ_i^j indicate a notable gene-specific change for the three mutants in (A), (D), and (G). The two clusters are given in orange and blue, and the SVM boundary is depicted by the black dashed line. Whilst *isw2*Δ*chd1*Δ and *isw1*Δ*isw2*Δ require both fPCs to linearly separate the Pearson clusters, *rsc8chd1*Δ is almost exclusively dependent on the second fPC, which means this mutant decreased the slope tilt. ↪

Figure 4: (continued) This can be better understood when analysing the two fPCs and their effect on the mean ((B) for *isw2*Δ*chd1*Δ, (E) for *rsc8chd1*Δ, and (H) for *isw1*Δ*isw1*Δ). The solid lines in magenta and green in these plots indicate a positive contribution of the fPC and a negative contribution, respectively, whereas the black dashed line depicts the mean. Grey arrows along the gene suggest general trends. Grey vertical bars suggest positions that remain largely unperturbed by the fPC. Grey arrows pointing to a single peak suggest remarkable properties. Interestingly, whilst the first fPC of the *isw2*Δ*chd1*Δ and *isw1*Δ*isw2*Δ strains shows a similar transformation of the mean, the second fPC indicates a different behaviour, particularly with respect to the +2 nucleosome. As suggested by the fact that clusters in the *rsc8chd1*Δ mutant are exclusively dependent on the second fPC, the first fPC explains only the average profile amplitude and does not contain any information about collective phasing. The location-specific effect of the linear separator for each mutant is given in (C), (F), and (I). The grey areas indicate the importance of each position to determine the clusters, whose median profile is shown as a blue and orange dashed line. The mean is depicted in black. Although the impact on the grouping of the +1 and +2 position in *isw2*Δ*chd1*Δ conditions is similar to the *isw1*Δ*isw2*Δ strain, the latter is seemingly particularly dependent on the +3 and +4 nucleosome. Positions thereafter become less important, which keep having a strong impact on clustering in *isw2*Δ*chd1*Δ. As expected *rsc8chd1*Δ is exclusively dependent on the second fPC. Interestingly, the entire profile seems to be influential for classifying genes, with the largest impact allocated to the first two nucleosomes. All axes are scaled to the same size for each strain; shapes and amplitudes are therefore comparable (see Methods for more details).

Pearson correlation with respect to the two discriminating trends (compare with cartoon 1(D)) could be difficult. The results for these strains should be taken with a pinch of salt.

Taken together, these outcomes show that remodeler mutants have varying effects on nucleosome positioning. Whilst most mutations do not notably alter the gene-specific nucleosome coordination with respect to the WT, we identified 5 mutants that exhibited a strong effect on phasing. Interestingly, most of them include *chd1*Δ, which indicates an important role of Chd1 for local arrangement within the gene body. Using fPCA to visualise the Pearson clusters permits the clear and position-specific quantification of the induced impact among varying strains.

Pol II Presence Correlates With Nucleosome Organisation in *chd1*Δ Mutants

In order to assess an interdependence of nucleosome organisation with other genomic properties, we compared the two Pearson clusters to Pol II levels, Sth1 occupancy, AT ratio over the entire gene, as well as upstream NDR length and orientation of the upstream NDR (i.e. tandem or divergent). We also included Mediator presence as a large protein complex with transient interactions predominantly at the NDR. All of these factors were clustered into two equally-sized groups where possible. For example, Pol II presence along the gene was evenly separated into transcribed regions with high and low Pol II occupancy. Interdependence was measured by training a simple neural network with no hidden neurons using Hebbian learning [41]. Consequently, we assessed which nuclear groups (e.g. high or low Pol II presence) corresponded to which Pearson clusters. We want to stress that we did not aim to find a predictive model. Rather, this approach allowed us to measure a direct correlation between similarity of nucleosome phasing and other genomics properties. The initial k-mean clustering did not impose a constraint on the group size, and they could therefore differ in the number of genes they contained. To remove any prediction bias, we forced the clusters to be of the same size. Genes in the larger group with closest Pearson coefficient to all distributions in the smaller group changed the cluster. We found the analysis for WT conditions non-conclusive, and correlations varied between A and B replicate (SFig A.11(top)). Whilst A was slightly correlated with the AT sequence content (Figs 6(A) and (B)), this trend disappeared for B, and it might in both replicates rather correspond to the fPC orthogonal to the cluster boundary (SFig A.13). Overall, we were surprised that none of the investigated properties could indicate a clear

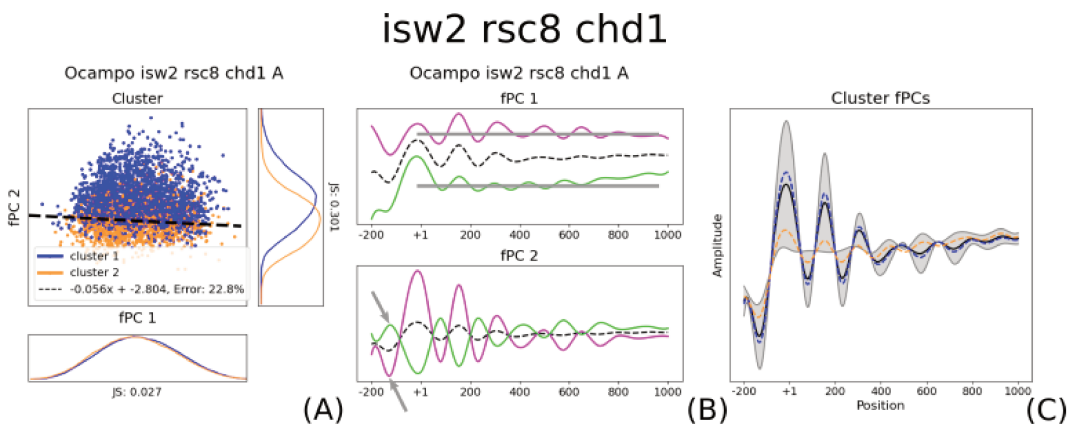


Figure 5: The fPCs, their gene specific scores, and the discriminating boundary explain changing collective phasing in *isw2*Δ*rsc8chd1*Δ. The two clusters are given in orange and blue. The figure shows the fPC scores ζ of the *isw2*Δ*chd1*Δ*rsc8* mutant and their separating boundary (black dashed line, A). The slope decreases with respect to the WT, making the gene groups almost solely dependent on the second fPC. Both fPCs transform the mean in a similar way as the double mutant *rsc8chd1*Δ (compare B with Fig 4(E)). The dashed black line as well as the solid lines in magenta and green indicate the mean, a positive contribution of the fPC, and a negative contribution, respectively. As expected, the separating boundary discriminate between the two clusters largely following the second fPC (C). The grey areas show the importance of each position to determine the clusters, whose median profile is shown as a blue and orange dashed line. The mean is depicted in black. All axes are scaled to the same size for each strain; shapes and amplitudes are therefore comparable.

interdependence with nucleosome phasing in WT (Fig 6(A)).

The correlation between positioning and other nuclear properties changed among the mutants (SFig A.11). The effect is particularly clear for *chd1*Δ (Fig 6(C)), as there is a strong interdependence between phasing and Pol II (Fig 6(D)), Mediator presence, and NDR size (SFig A.11). As aforementioned, Chd1 maintains, among others, chromatin integrity during Pol II elongation. The correlation is therefore in line with our previous conclusions and the function of Chd1. The established link between Pol II presence and nucleosome organisation remains conserved in all strains with a Chd1 gene deletion, except *isw1*Δ*chd1*Δ*rsc8*. This is similarly true for the correlation with Mediator occupancy and NDR length. There was also a slight correlation to Sth1 and AT ratio in cells containing *chd1*Δ, which were, despite being weak, still notably stronger than in WT. The results are in agreement with the effects of Chd1 on chromatin maintenance during gene expression.

Due to the Rsc8-mediated nucleosome organisation within transcribed regions, we were wondering whether there is an increased interdependence to NDR length or gene size. We can report that there is no correlation with NDR size in any *rsc8*-depleted strain (Fig 6(E)). This is in line with our hypothesis that Rsc8 decouples processes at different genes. However, *rsc8* mutant cells exhibited a slightly increased correlation with Sth1. By looking at the separation with respect to the Pearson cluster boundary, we find that there is no noticeable impact (Fig 6(F)). The results indicate that any correlation with region-specific properties is lost, which is likely due to the interference of nucleosome positioning of various regions.

*isw1*Δ single mutant did exhibit only a slightly increased correlation with Pol II, Sth1 and Mediator presence as well as AT ratio. *isw2*Δ might as well show a weak correlation with Pol II occupation. Each of the replicates of their double mutant indicates different correlations, and it is therefore difficult to tell whether transcription-related factors influence nucleosome phasing in the *isw1*Δ*isw2*Δ strain. However, none of them indicate any strong interaction, suggesting that—on a global scale—these effects might be negligible in comparison to the WT (SFig A.11).

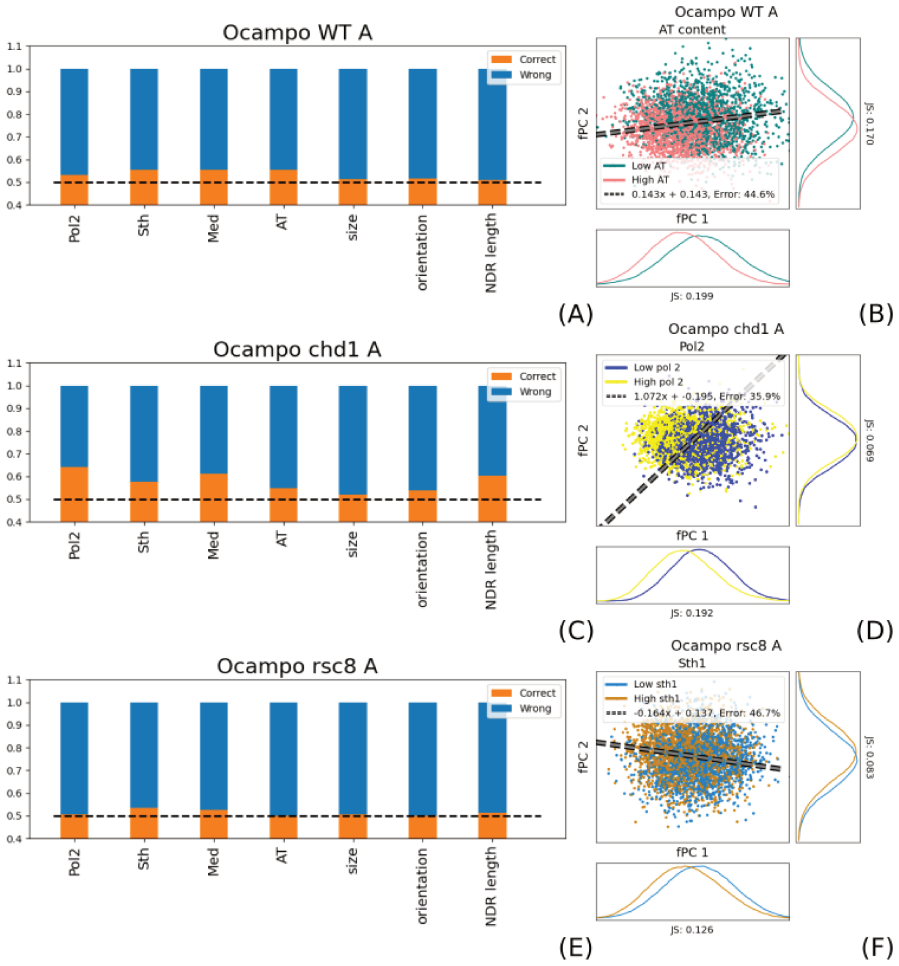


Figure 6: Remodeler deletions have varying effects on the interdependence with other genomic properties. The orange bars in Figs (A), (C), and (E) show the ratio of correct predictions, and blue bars are wrong guesses. As we distinguish between two clusters, the dashed black line at 0.5 indicates random guessing. The dashed grey line with black edging in Figs (B), (D), and (F) display the linear boundary for the Pearson clusters. (A) and (B): WT conditions are seemingly correlated with the sequence composition. However, the results are different for the *B* replicate, and therefore non-conclusive. All possible correlations are surprisingly low. (C) and (D): *chd1*Δ mutants increase particularly their dependence on Pol II and other transcription-coupled properties, such as Mediator presence. Surprisingly, the mutant showed also an increased interdependence on NDR length. (E) and (F): despite the Rsc8-mediated gene limits, there is no correlation with coordinated nucleosome phasing and the size of transcribed regions or NDR length. Although Sth1 indicates a slightly increased interdependence, this cannot be confirmed when plotting the groups with respect to the Pearson clusters. This is in line with our hypothesis that positioning in different regions interfere, and therefore, nucleosome localisation become increasingly independent from region-specific factors.

Interestingly, the *rsc8*-depleted *isw2*Δ indicated a strong correlation with Sth1 and Mediator presence as well as NDR length. The effect was observable in almost all strains that contained the double mutation with the exception of the quadruple mutant (SFig A.11). Taken together, this could indicate an impact along the gene body and the promoter region in strains that contain *rsc8*-depleted *isw2*Δ.

Surprisingly, combining two factors together (e.g. Pol II presence and AT ratio) to predict Pearson clustering did not increase accuracy. Instead, one factor dominated the correlation measurement, e.g. Pol II presence for *chd1*Δ strains. This could possibly suggest that—despite several factors showing increased interdependence—they can be reduced to a main influencing factor (which is not necessarily one of the tested properties).

Taken together, the results indicate a strong interdependence between local genomic properties—such as presence of large protein complexes or NDR length—and cell strains containing *chd1*Δ. This supports our hypothesis of Chd1 being responsible for local nucleosome coordination.

Discussion

In this work, we analysed the collective positioning of nucleosome arrangement within the gene body in WT and chromatin remodeler-deficient strains by combining clustering of Pearson coefficients with fPCA, the latter being an analysis framework for functional data. Although fPCA is well established in the assessment of time series, it has not been previously used to understand location-specific nucleosome profiles on a global scale. As we argue that the Pearson index measures similarity of nucleosome arrangement between genes, we interpreted the effect of chromatin remodelers on the positioning by visualising the distribution of two established significant Pearson clusters using fPCA. Indeed, we can show that the sets of genes for all mutants can be sensibly separated by the two fPCs that explain most variance in the data, and more fPCs are not necessary to describe the clusters. This allowed the quantification of the effect on coordinated phasing. The significant Pearson groups were compared with other nuclear properties—such as Pol II presence and NDR maintenance—and sequence-dependent characteristics. None of the commonly supposed influencing factors can easily explain coordinated nucleosome positioning in WT conditions. However, correlation between tested properties and phasing increases with some gene mutations. The analysis reveals the impact of different gene deletions of chromatin remodelers on nucleosome arrangement within the gene body. It shows Rsc8-defined boundaries for nucleosome positioning along the gene, suggesting a global impact over the entire array for each gene. On the other hand, the results for most strains that contained a Chd1 deletion indicated gene-specific local effects, which correlate largely with Pol II presence. In the following, we critically discuss the results and their significance.

We applied a pairwise Pearson cross-correlation index to measure profile similarity between genes. The linear correlation measurements evaluate the overall trend of the signal (i.e. increasing or decreasing distributions at similar positions), and it does not take signal scaling into account. Therefore, it assesses whether genes are apt to contain nucleosomes at similar positions. Indeed, similar nucleosome phasing could indicate similar but gene-specific chromatin remodeler dynamics, which justifies the rationale for measuring classical linear correlation. It also follows previous analyses using comparable measurements [33, 34].

We classified genes according to their Pearson coefficients by applying a k-mean clustering approach. k-mean was repeated over several random initialisations, therefore removing any prior bias. We used a silhouette criterion value to determine the best number of clusters, which was shown to be 2. It should be mentioned though that the cluster distribution according to the fPCA did not show a clear separation of the data points themselves (i.e. there were no distinct data accumulations). Therefore, this clustering is imposed by our assumptions using the Pearson index. Nonetheless, we argue that they reveal important information about nucleosome phasing linked to chromatin remodelers when compared with mutant strains. The validation using the silhouette criterion together with the shape-independent JS distance over 500 random clusters proved their significance. This shows that the data could not be better categorised using linear correlation. We acknowledge the fact that 500 random partitions for over ≈5000 transcribed regions is still comparatively low. However, as we approximate the distribution of JS distances over random clustering with a gamma distribution, we made our estimates independent of the actual number of samples. Gamma distributions are commonly used to represent unimodal strictly positive random variables, and it is therefore a sensible choice for JS distances over random partitions of the same data set.

As the Pearson correlation index only indicates average similarity over the entire nucleosome array, we aimed to compare the clusters to the data itself in order to interpret their differences. Dimensionality reductions are often used to visualise clusters, such as for single-cell sequencing analyses [42]. Common approaches include PCA, uniform manifold approximation and projection (UMAP) [43], and t-distributed stochastic neighborhood embedding (t-SNE) [44]. Whilst the latter two are non-linear dimensionality reductions, PCA and fPCA find a linear decomposition of the data into the axes (or functions) that explain most variance. It is challenging to retrieve the exact meaning of the discriminating boundary using non-linear approaches. Consequently, understanding the location-dependent differences in the profile between two clusters and interpreting their separating function is more straightforward for PCA or fPCA than for UMAP and t-SNE. Although PCA and fPCA are very similar, PCA assumes that every position in the MNase-seq data is independent, whereas fPCA conjectures that they were produced by a single stochastic process along the spatial axis. Therefore, positions are dependent on each other. This is inline with the barrier model for establishing nucleosome phasing, which makes fPCA preferrable over PCA. Moreover, as we treat each nucleosome profile as one entity by using the Pearson correlation, the independence assumption would violate the fundamental understanding in our analysis. Nonetheless, when comparing PCA and fPCA, we showed that the two clusters can be similarly separated (SFig A.12), although the two principal axes are slightly differently shaped due to the missing constraint of the spatial dependence.

FPCA assumes a stochastic process with a mean behaviour over the entire data set, and it characterises each data point with respect to their deviance from that mean (see Methods). The results therefore depend on the entire considered data set. Indeed, we find different results when including all genes or exclusively transcribed regions >1000 bp. However, these differences are not strong. Moreover, any possible bias was excluded by removing genes smaller than 1000 bp from a subsequent analysis. Due to the abundant and well-positioned nature of nucleosomes within the gene body in *Saccharomyces cerevisiae*, we find it justified to presume an average nucleosome distribution describing their wave-like profile. Nonetheless, we argue that the variance between genes contains important information about nucleosome phasing imposed by chromatin remodelers, which we roughly categorised into groups. We found that the two Pearson correlation clusters could be neatly separated by the fPC scores ζ_i^j, $j \in \{1, 2\}$. This indicates firstly that the Pearson index measures a trend that is explained by the largest variance in the data; and secondly, the two fPCs that describe most variance are sufficient to interpret the clusters.

Whilst linear-correlation measurements are limited to quantifying the average similarity, a combination with fPCA allows characterising location-specific differences and in which way gene deletions affect phasing from an average. Evaluating the effect of the linear boundary along positions within the gene body revealed detailed differences in the nucleosome profile that are important for establishing the groups. As our approach is largely dependent on general signal processing methods, we can repeatedly apply the same framework for all mutants and compare there results. Therefore, the combination of linear correlation with fPCA extends previous ways of analysing nucleosome distributions using only Pearson [34] or autocorrelation [33] by allowing a position-specific interpretation.

The analysis can clearly distinguish between mutant-specific effects on phasing. All mutants preserved the information of coordinated nucleosome arrangement in their first two fPCs, and the Pearson clusters could be separated by a neat line. Consequently, none of the chromatin remodeler gene deletions caused random positioning. Some mutants, however, showed an increased overlap between the two groups, which indicates increased independence between individual nucleosome locations, and positioning might be more random. Including more fPCs could help further separating the clusters. In all of those cases however, one of the two replicates always permitted a clearer separation by using only the first two fPCs. Considering the experimental variability in the data, it is not possible to draw direct conclusions without further replicates. In order to simplify the comparison between mutants, we restrained from including more fPCs.

Most strains did not alter notably their gene-specific collective arrangement (i.e. the slope), and

a linear separation of the Pearson clusters using the deviance from the mean did not change with respect to WT strains. Although they can nevertheless have an impact on the mean itself, coordination along the genes remains preserved in a similar way, at least as measured by the Pearson correlation index. Due to the focus of the study on coordinated nucleosome positioning along transcribed regions, we did not consider them as having notably changed their coordinated phasing.

Gene mutations of chromatin remodelers have been analysed previously in detail, including their influence on phasing [12–14, 18], NDR maintenance [45], and gene transcription [13]. RSC is the only essential chromatin remodeler complex in *Saccharomyces cerevisiae* [46], and it has been particularly associated with positioning of the +1 and -1 nucleosomes [12, 45, 47]. This mechanism has been proposed to be conserved among various yeast species [11]. It has also been reported that RSC regulates expression of Pol II and Pol III-transcribed genes [13, 48, 49]. Moreover, it has been found to impact Pol II elongation and termination [12]. All of these results imply that RSC is to some extent involved in limiting the transcribed region. However, this has been predominantly quantified with respect to changes at the core promoter. To our knowledge, a potential role for Rsc8 to decouple nucleosome phasing in independent genes has not been suggested. The presented functional analysis of MNase-seq profiles in *rsc8*-depleted strains clearly indicates a coordinated nucleosome arrangement that exceeds the limits of transcribed areas. This is further supported by our finding that correlation with other nuclear and sequence-dependent factors decreases. Furthermore, mutants that were *rsc8* depleted decreased notably the boundary slope between the two clusters, indicating that coordinated positioning becomes increasingly independent of other functional components. The strictly limited and Rsc8-mediated phasing barrier could have further implications for other processes—such as transcription—as nucleosome placing in one gene influences its neighbouring regions. The notion of gene-interfering positioning has been also proposed by [14]. The study shows that RSC could act as a bidirectional barrier, influencing upstream and downstream regions. Interestingly, they found that interference also plays a crucial role in WT strains, and that the same phenomenon remains preserved in *rsc8*-depleted cells. However, our fPCA reveals that the limiting role of the RSC remodeler complex is crucial in WT conditions, and that this behaviour is significantly altered when Rsc8 is depleted. Taking this into account, Rsc8 should fulfill the role of disentangling gene-related processes in WT strains, and it therefore allows for a flexible and uncorrelated transcriptional program. Indeed, *rsc8*-depleted cells exhibit significantly altered Pol II profiles [10, 12], which is in accordance with our hypothesis. We propose that the RSC chromatin remodeler globally disentangles nucleosome phasing, and it therefore plays a substantial role in long-range positioning.

Interestingly, our results indicate that positioning limited to the gene body can be re-established in *rsc8*-depleted *chd1*Δ mutants. We hypothesise that they have antagonistic effects in establishing gene size-dependent barriers for nucleosome arrangement. Indeed, it was reported that Rsc8 and Chd1 have opposing effects for Pol II termination. *rsc8*-depleted cells exhibit inhibition of Pol II dissociation at the TTS, whereas the double mutant *isw1*Δ*chd1*Δ increases release frequency, with seemingly *chd1*Δ dominating this effect [12]. The authors propose that this is related to the close packaging of nucleosomes at the TTS. Our outcomes suggest that they might have antagonistic effects in chromatin organisation that differs between transcribed and non-transcribed regions.

We found that *chd1*Δ mutants had a strong impact on coordinated positioning within the gene body. Indeed, Chd1 has been, among others, characterised with respect to its role in maintaining chromatin integrity during Pol II transcription [16, 50, 51], and it associates to both promoters and transcribed regions [52]. This is in line with our finding that correlation with Pol II presence and occupancy of Mediator increases in Chd1-deficient strains. With the exception of *isw1*Δ*isw2*Δ, all other noteworthy changes included deletion of *chd*, further emphasising its role for chromatin organisation within the gene. However, not all *chd1*Δ-containing mutants exhibit a notable effect. This can have various reasons, including experimental variability. However, particularly the mutant *chd1*Δ*isw1*Δ*isw2*Δ could indicate an interacting behaviour of the remodelers. Indeed, Chd1 has been reported to cooperate [16] as well as antagonise Isw1 [18], and therefore could have different effects depending on the context. With this being said, the behaviour of the triple mutant

*isw1*Δ*isw2*Δ*chd1*Δ is particularly interesting, as *chd1*Δ and *isw1*Δ*isw2*Δ each individually affect coordinated phasing, but not their triple mutant. This could suggest an antagonistic behaviour on nucleosome coordination. As Chd1 is highly conserved in all eukaryotes [53], this result could have consequences beyond *Saccharomyces cerevisiae*.

Analysing the MNase-seq data using fPCA allowed us to obtain a different view on the functionality of various remodelers to maintain chromatin organisation. We propose the following mechanism (Fig 7). The RSC remodeler complex is essential for allowing independent phasing in each single gene. It plays therefore a pivotal role in maintaining the barrier with respect to which nucleosome positioning is coordinated. This permits the decoupling of gene-specific processes such as transcription. Depletion of Rsc8 leads to the interference of different genomic regions, which therefore alters sequence accessibility on a global scale. Indeed, it has been reported that gene expression is dramatically changed in *rsc8* mutants [10, 49]. Chd1, on the other hand, maintains chromatin integrity during transcription [16, 50, 51], and it influences nucleosome phasing locally to permit Pol II-mediated expression. *chd1*Δ strains make positioning dependent on Pol II presence. Consequently, whilst RSC plays a global role, Chd1 is important for local nucleosome organisation.

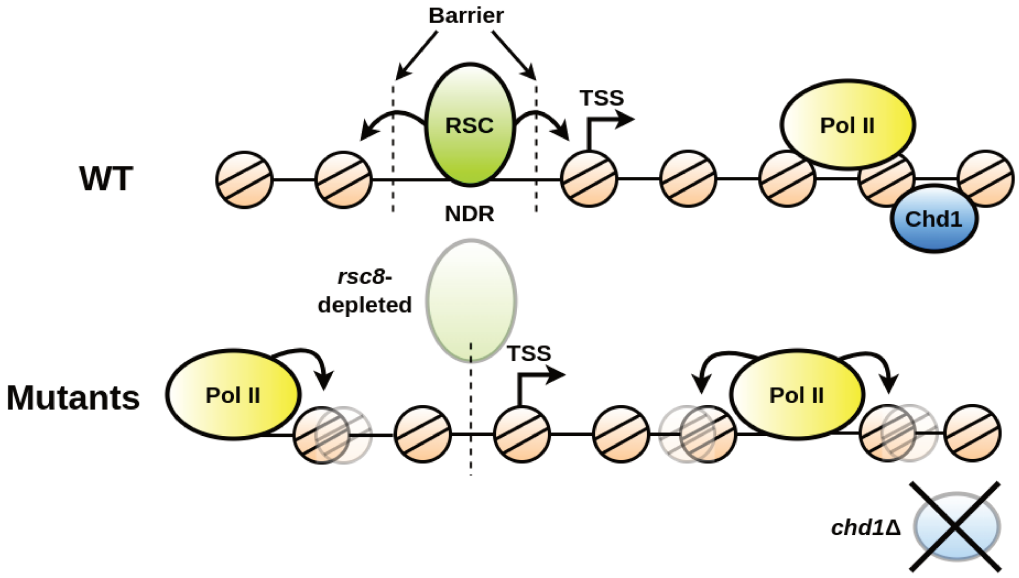


Figure 7: Chromatin remodelers maintain nucleosome organisation on a local and far-reaching scale. Top: RSC (green ellipse) establishes independent nucleosome phasing on each gene (two vertical dashed lines) by maintaining the NDR through positioning the +1 (cornered arrow) and -1 nucleosome. The ATP-dependent positioning is symbolised by black arrows pointing away from RSC. The local remodeling effect of Chd1 (blue ellipse) allows chromatin arrangement independent of Pol II transcription (yellow ellipse). Bottom: in *rsc8* strains, the NDR cannot be maintained anymore, and phasing in and outside a gene interfere with each other (single dashed line). We propose that this should equally lead to an increased interdependence of other nuclear processes such as transcription. If *chd1* is deleted, nucleosome arrangement is more sensitive to the presence of other large complexes, such as Pol II. During transcription, Pol II is affecting the local positioning (black arrows from Pol II).

Methods

Data Treatment

MNase sequencing reads were taken from [18] and [12] (GEO accession numbers GSE69400 and GSE73428, respectively) and treated as in our previous study [54]. To be precise, reads from Fastq files were trimmed with `trim_galore` (v0.6.5) [55] and `cutadapt` (v3.1) [56]. Subsequently, they were mapped on the *Saccharomyces cerevisiae* genome (University of California at Santa Cruz [UCSC] version `sacCer3`) using `bowtie2` (v2.3.4.3) [57]. Files were converted with `samtools` (v1.9) [58] and `deeptools` (v3.5.0) [59]. Read counts were normalised in Reads Per Million (RPM) of mapped reads. We used the option `--MNase` of `bamCoverage` so that only the mononucleosome fragments were kept. This means that fragments shorter than 130 bp and longer than 200 bp were removed from analysis. Mediator and Sth1 ChIP-seq were taken from [54] (ArrayExpress accession number E-MTAB-12198). We used Pol II ChIP-seq from our previous study [60].

Following [12, 18], we retrieved positioning profiles along the coding regions 200 bp before and 1000 bp after the +1 nucleosome. Genes on the Crick strand were inverted. Consequently, all data is calibrated such that the +1 position is at 200 bp. The profile of genes for which the +1 position is known were considered as in [18].

Measuring Profile Correlation and Clustering

The pairwise Pearson correlation of MNase-seq distributions for each gene was determined using equation

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{1}$$

Here, x and y denote two genes, $\bar{x}$ and $\bar{y}$ symbolise their respective average MNase-seq value along the coding region, and $n = 1200$ is the length of the considered region. Eq 1 ranges between -1 and 1, and indicates whether the two gene profiles tend to change into the same (positive Pearson correlation) or opposite directions (negative Pearson indices).

Genes were divided using the pairwise Pearson indices using the k-mean clustering implementation in MATLAB. Every gene is represented by a vector containing the cross-correlation values to all other profiles. To define the optimal number of k-mean clusters, we used the silhouette criterion measurement [61, 62]. For all analysed strains, the highest silhouette value occurs at 2 groups, suggesting that in order to divide the profiles into classes with respect to their Pearson indices, the optimal number of clusters is 2 (Fig 1(A)). Therefore data were grouped in two clusters (Figs 1(B, C)).

Cluster significance was validated by comparing the JS distance of the two determined groups with 500 random group pairs to which all genes were randomly assigned. The JS distance is bidirectional extension of the Kullback-Leibler (KL) divergence and can be calculated using

$$JS(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M), \tag{2}$$

where D and Q are two distributions (i.e. the MNase-seq profiles), and $M = 1/2(P + Q)$ is a mixture distribution.

As we compared a single value (i.e. the JS distance of the Pearson clusters) to a distribution (i.e. 500 random JS distances), standard significance tests are not applicable since they compare two distributions. We therefore approximated the distribution over all random JS distances with a gamma distribution and determined its 95% PI. The gamma distribution is defined by

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}e^{\beta x}\beta^{\alpha}}{\Gamma(\alpha)}. \tag{3}$$

Here, x denotes the value (i.e. JS distance), and α and β are parameters. $\Gamma(\cdot)$ is the Gamma function. If the value was outside the PI, we deemed it to be significant. Unfortunately, fits to the gamma distribution were sensitive to outliers, such that a single value far from the mode had a stronger weight in comparison to each value around it. We presumed that the distribution is best represented with the values close to the mode and removed therefore the 99% percentile of the random JS distances before fitting. All considered mutants were outside the 95% PI, with the exception of the *isw1*Δ*rsc8*, which was removed from the analysis.

Functional Principal Component Analysis

Functional clustering in a Hilbert space H can be achieved by fPCA. It applies—similar to PCA in Euclidean space—a functional dimensionality reduction in H to investigate the dominant mode in functional data. Instead of relying on values in discrete dimensions, fPCA uses a given number of basis functions (e.g. B-splines or Wavelet) to create the eigenfunction basis that accounts for most functional variation. Despite the fact that MNase-seq data is stored in a discrete array (i.e. one value per bp), we can nevertheless find a functional approximation over a range using a given choice of basis functions. It should be noted that this implicitly smooths out high frequencies in the signal. We presume that nearby values in MNase-seq data possess a strong interdependence, therefore justifying a smoothed and continuous functional representation of the high dimensional data. In this study, we apply B-splines as a basis to represent the nucleosome array (Fig 8). We use the Python library `scikit-fda` to determine the fPCs and the corresponding weights explaining the distribution [63]. Here, we describe briefly the underlying principles of the method.

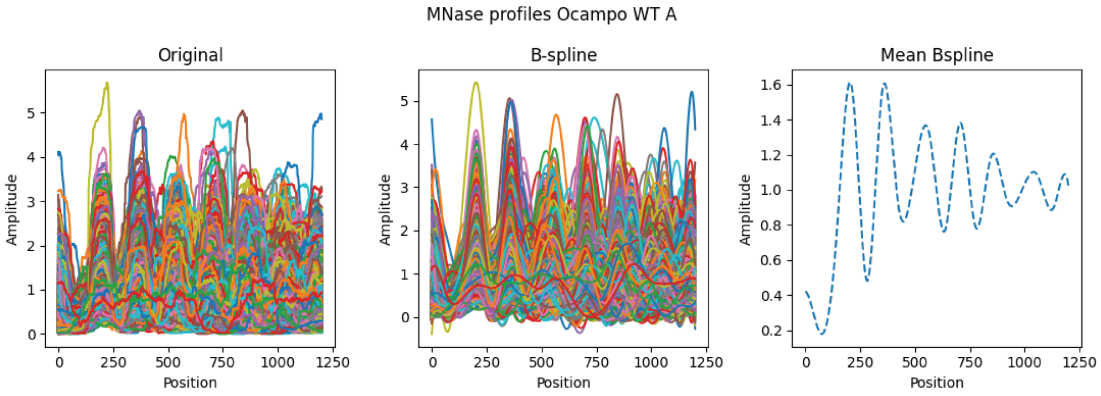


Figure 8: Representing the MNase data array as a composition of B-spline base functions in WT conditions. Left shows the raw data, each colour depicting one profile over a gene. Center gives the smoothed profiles after representing the data as B-splines. Right displays the average profile using the functional composition.

FPCA presumes that the functional data represents a stochastic process $X(t)$ with expected value $\mu(t) = \mathrm{E}[X(t)]$ and orthonormal eigenfunctions $\phi^i(t)$, $i = 1, 2....$ Intuitively, $\phi^i(t)$ describes the most variation in X orthogonal to all ϕ^j, $j < i$. This allows the iterative determination of the eigenfunctions in the functional data. It should be emphasised that in this study the process is defined in space rather than describing temporal data. We follow nevertheless the convention by denoting the independent variable as t. By using the Kosambi–Karhunen–Loève theorem, any stochastic process can be represented as an infinite linear combination $\phi^i(t)$. Consequently, we can describe the stochasticity in $X(t)$ via

$$X(t) - \mu(t) = \sum_k \zeta^k \phi^k(t). \tag{4}$$

ζ^k is the autocovariance operator

$$\zeta^k = \int (X(t) - \mu(t))\phi^k(t)dt. \tag{5}$$

To provide some intuition, it is presumed that the entire data set can be explained via an average behaviour $\mu(t)$. Variability to $\mu(t)$ for each gene is expressed by $\phi^k(t)$ together with a factor ζ^k. ζ^k can be loosely compared to a normal correlation measurement, i.e. ζ^k increases when $\phi^k(t)$ and $\int(X(t) - \mu(t))$ follow the same trend. If they describe opposing behaviours—for example $\phi^k(t)$ decreases when $\int(X(t) - \mu(t))$ increases—ζ^k becomes negative.

It is commonly justified to approximate Eq 4 as a finite sum

$$X(t) \approx X_n(t) = \mu(t) + \sum_k^n \zeta^k\phi^k(t). \tag{6}$$

It should be noted that $\phi^i(t)$, $i = 1, 2, ...$ is a basis of the functional space in H.

This understanding of the underlying process permits the application of fPCA. A smoothed representation with the basis functions (e.g. B-splines) fulfilling Eq 6 can be obtained using L^2 regularisation. To reduce the dimensionality to K, we keep only the first K components (i.e.$\phi^i(t)$) that represent the dominant mode of variation in X by setting the first component to

$$\phi^1 = \underset{\|\phi\|=1}{\arg\max} \left\{ \mathrm{Var}(\int_{\mathcal{T}} (X(t) - \mu(t))\phi(t)dt) \right\}, \tag{7}$$

and the following $K - 1$ components to

$$\phi^k = \underset{\|\phi\|=1, \langle\phi,\phi^j\rangle=0 \text{ for } j=1,...,k-1}{\arg\max} \left\{ \mathrm{Var}(\int_{\mathcal{T}} (X(t) - \mu(t))\phi(t)dt) \right\}. \tag{8}$$

$\|\phi\|$ is the square norm, i.e. $\|\phi\| = \sqrt{\left(\int \phi(t)^2\right)}$. It should be emphasised that ϕ^k can differ by a factor of -1 due to the square norm, and consequently, the operator ζ^k (Eq. 5) can be either positive or negative depending on ϕ^k. This means that the slope of the cluster-dividing boundary can be pointing upwards or downwards and still describe the same functional composition.

We exemplified the impact of the first two fPCs to analyse the consequences on nucleosome phasing in chromatin remodeler-deficient cells (see for example Figs 1, 2, and 4). It should be noted that the fPCs were amplified to highlight their functional contribution. We set the scaling factor to $\zeta^1 = \zeta^2 = 20$ in all figures that demonstrate their effects (i.e. magenta shows the effect of the fPC multiplied by 20 and added to the mean, and green depicts the fPC multiplied by 20 and subtracted from the mean). The determined factors were predominantly distributed in $\zeta^{1,2} \in [-20, 20]$ for all strains and replicates, and most of them were in fact much lower. Therefore, we limited the scaling of the axes for ζ^1 and ζ^2 to $[-20, 20]$ for all plots that show the cluster distribution with respect to the factors. Therefore, all figures and axes were directly comparable. The few outliers that were outside this range were incorporated into the analysis despite of being not shown in those plots.

Quantifying the Cluster Boundary

Long genes were linearly separable with respect to the Pearson coefficient clusters in all WT and mutant conditions. The boundary was determined using a linear SVM. We ignored the prediction error and the intercept of the linear boundary, and instead considered only the slope differences between the two replicates. As aforementioned, the sign of the slope m does not matter, and we consider therefore only $|m|$. To quantify the variability in the two replicates, we introduce the following measurement

$$s(i) = \frac{(\bar{m}_i - \bar{m}_{WT})^2}{(|m_i^A| - |m_i^B|)(|m_{WT}^A| - |m_{WT}^B|)}. \tag{9}$$

$\bar{m}$ denotes the average over the absolute slopes of both replicates. We defined a change as notable when $s(i) > 1$, which implies that the mean variability between WT and mutant is larger than the variability within the replicates, i.e.

$$(\bar{m}_i - \bar{m}_{WT})^2 > (|m_i^A| - |m_i^B|)(|m_{WT}^A| - |m_{WT}^B|). \tag{10}$$

As we consider only two replicates, we restrain from using the word *significant* as much as possible and use *noteworthy* or *notable* instead.

The slope of the boundary m indicates the contribution of each fPC to describe the discriminator between the clusters. As m shows the change of ζ^2 over one unit of ζ^1, we can determine the separating boundary by

$$\phi^{/} = \frac{m\phi^1 + \phi^2}{m+1}. \tag{11}$$

The impact of $\phi^{/}$ can be visualised by multiplying a scaling factor which is followed by addition to and subtraction from the mean. In this study, we used a factor of $\zeta^{/} = 5$ to create the grey bands in the plots that show the effect of the separating function.

Measuring Interdependence Between Nucleosome Phasing and Other Nuclear Properties

In order to analyse interdependence of nucleosome positioning with other nuclear properties, we divided all factors into two equally sized cluster using the median wherever possible. For example, the half with the smaller NDRs was assigned to group -1, whereas the larger half was group 1. This split was performed after filtering for the size (i.e. large or small genes). The analysis aimed to find a correlation between nuclear factor group and Pearson cluster. To remove any bias with respect to the group size, we forced both Pearson clusters to contain the same number of genes.

We used a simple feedforward network with no hidden neurons and a single output neuron whose activation indicated the predicted Pearson cluster. The number of input neurons varied between 1 and 2, depending on whether we considered a multivariate interdependence. The group of the nuclear factor (i.e. -1 or 1) was set as input neuron activation. This was weighted and summed together with all other input values. The activation function of the output was a modified sign function, which returned 0 when negative and 1 when positive. Therefore, if the weighted sum over the input was lower than or equal to 0, the output would be 0, and 1 otherwise.

Weights were trained using a Hebbian-like learning method [41]. In order to avoid any confusion, we name Pearson cluster 0 and nuclear factor group -1 *low* cluster, whereas we define group 1 in both cases to be the *high* cluster. The weight was defined to be the average number of genes where the nuclear factor group and Pearson coefficient cluster where both *low* or both *high*; minus the average number where one of them was *low* whilst the other *high*. The implementation as a neural network allowed the straightforward extension to compare interdependence with several factors at the same time using the same method.

Acknowledgments

This work was supported by Fondation ARC [PGA1 RF20170205342]; Comité Ile-de-France - La Ligue Nationale Contre le Cancer. K.A. was supported by a PhD training contact from the French Ministry of Higher Education and Research. L.Z. was supported by a PhD training contract from the CEA NUMERICS program, which has received funding from European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 800945. This project has received financial support from the CNRS through the MITI interdisciplinary programs.

References

1. Roger Kornberg. The location of nucleosomes in chromatin: specific or statistical? *Nature*, 292:579–580, 1981.

2. Travis N Mavrich, Ilya P Ioshikhes, Bryan J Venters, Cizhong Jiang, Lynn P Tomsho, Ji Qi, Stephan C Schuster, Istvan Albert, and B Franklin Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome research*, 18(7):1073–1083, 2008.

3. Kevin Struhl and Eran Segal. Determinants of nucleosome positioning. *Nature structural & molecular biology*, 20(3):267–273, 2013.

4. Yong Zhang, Zarmik Moqtaderi, Barbara P Rattner, Ghia Euskirchen, Michael Snyder, James T Kadonaga, X Shirley Liu, and Kevin Struhl. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature structural & molecular biology*, 16(8):847–852, 2009.

5. Pauline Vasseur, Saphia Tonazzini, Rahima Ziane, Alain Camasses, Oliver J Rando, and Marta Radman-Livaja. Dynamics of nucleosome positioning maturation following genomic replication. *Cell reports*, 16(10):2651–2665, 2016.

6. Amanda L Hughes, Yi Jin, Oliver J Rando, and Kevin Struhl. A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. *Molecular cell*, 48(1):5–15, 2012.

7. Job Dekker. GC-and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. *Genome biology*, 8:1–14, 2007.

8. Alberto Marin-Gonzalez, JG Vilhena, Ruben Perez, and Fernando Moreno-Herrero. A molecular view of DNA flexibility. *Quarterly Reviews of Biophysics*, 54:e8, 2021.

9. Cedric R Clapier, Janet Iwasa, Bradley R Cairns, and Craig L Peterson. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature reviews Molecular cell biology*, 18(7):407–422, 2017.

10. Timothy J Parnell, Jason T Huff, and Bradley R Cairns. RSC regulates nucleosome positioning at Pol II genes and density at Pol III genes. *The EMBO journal*, 27(1):100–110, 2008.

11. Carlo Yague-Sanz, Enrique Vázquez, Mar Sánchez, Francisco Antequera, and Damien Hermand. A conserved role of the RSC chromatin remodeler in the establishment of nucleosome-depleted regions. *Current Genetics*, 63:187–193, 2017.

12. Josefina Ocampo, Răzvan V Chereji, Peter R Eriksson, and David J Clark. Contrasting roles of the RSC and ISW1/CHD1 chromatin remodelers in RNA polymerase II elongation and termination. *Genome research*, 29(3):407–417, 2019.

13. Emily Biernat, Jeena Kinney, Kyle Dunlap, Christian Rizza, and Chhabi K Govind. The RSC complex remodels nucleosomes in transcribed coding sequences and promotes transcription in Saccharomyces cerevisiae. *Genetics*, 217(4):iyab021, 2021.

14. Dwaipayan Ganguli, Razvan V Chereji, James R Iben, Hope A Cole, and David J Clark. RSC-dependent constructive and destructive interference between opposing arrays of phased nucleosomes in yeast. *Genome research*, 24(10):1637–1649, 2014.

15. Manu Shubhdarshan Shukla, Sajad Hussain Syed, Fabien Montel, Cendrine Faivre-Moskalenko, Jan Bednar, Andrew Travers, Dimitar Angelov, and Stefan Dimitrov. Remosomes: RSC generated non-mobilized particles with approximately 180 bp DNA loosely associated with the histone octamer. *Proceedings of the National Academy of Sciences*, 107(5):1936–1941, 2010.

16. Claudia Alén, Nicholas A Kent, Hannah S Jones, Justin O'Sullivan, Agustın Aranda, and Nicholas J Proudfoot. A role for chromatin remodeling in transcriptional termination by RNA polymerase II. *Molecular cell*, 10(6):1441–1452, 2002.

17. Marta Radman-Livaja, Tiffani K Quan, Lourdes Valenzuela, Jennifer A Armstrong, Tibor Van Welsem, TaeSoo Kim, Laura J Lee, Stephen Buratowski, Fred Van Leeuwen, Oliver J Rando, et al. A key role for Chd1 in histone H3 dynamics at the 3' ends of long genes in yeast. *PLoS genetics*, 8(7):e1002811, 2012.

18. Josefina Ocampo, Răzvan V Chereji, Peter R Eriksson, and David J Clark. The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. *Nucleic acids research*, 44(10):4625–4635, 2016.

19. Triantaffyllos Gkikopoulos, Pieta Schofield, Vijender Singh, Marina Pinskaya, Jane Mellor, Michaela Smolle, Jerry L Workman, Geoffrey J Barton, and Tom Owen-Hughes. A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. *Science*, 333(6050):1758–1760, 2011.

20. Toshio Tsukiyama, Jeffrey Palmer, Carolyn C Landel, Joseph Shiloach, and Carl Wu. Characterization of the imitation switch subfamily of ATP-dependent chromatin-remodeling factors in Saccharomyces cerevisiae. *Genes & development*, 13(6):686–697, 1999.

21. Iestyn Whitehouse, Oliver J Rando, Jeff Delrow, and Toshio Tsukiyama. Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, 450(7172):1031–1035, 2007.

22. Janet G Yang, Tina Shahian Madrid, Elena Sevastopoulos, and Geeta J Narlikar. The chromatin-remodeling enzyme ACF is an ATP-dependent DNA length sensor that regulates nucleosome spacing. *Nature structural & molecular biology*, 13(12):1078–1083, 2006.

23. Kazuhiro Yamada, Timothy D Frouws, Brigitte Angst, Daniel J Fitzgerald, Carl DeLuca, Kyoko Schimmele, David F Sargent, and Timothy J Richmond. Structure and mechanism of the chromatin remodelling factor ISW1a. *Nature*, 472(7344):448–453, 2011.

24. Corinna Lieleg, Philip Ketterer, Johannes Nuebler, Johanna Ludwigsen, Ulrich Gerland, Hendrik Dietz, Felix Mueller-Planitz, and Philipp Korber. Nucleosome spacing generated by ISWI and CHD1 remodelers is constant regardless of nucleosome density. *Molecular and cellular biology*, 35(9):1588–1605, 2015.

25. Elisa Oberbeckmann, Vanessa Niebauer, Shinya Watanabe, Lucas Farnung, Manuela Moldt, Andrea Schmid, Patrick Cramer, Craig L Peterson, Sebastian Eustermann, Karl-Peter Hopfner, et al. Ruler elements in chromatin remodelers set nucleosome array spacing and phasing. *Nature Communications*, 12(1):3232, 2021.

26. Cizhong Jiang and B Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*, 10(3):161–172, 2009.

27. Assaf Weiner, Amanda Hughes, Moran Yassour, Oliver J Rando, and Nir Friedman. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome research*, 20(1):90–100, 2010.

28. Sandro Baldi, Stefan Krebs, Helmut Blum, and Peter B Becker. Genome-wide measurement of local nucleosome array regularity and spacing by nanopore sequencing. *Nature structural & molecular biology*, 25(9):894–901, 2018.

29. R Thomas Koerber, Ho Sung Rhee, Cizhong Jiang, and B Franklin Pugh. Interaction of transcriptional regulators with specific nucleosomes across the Saccharomyces genome. *Molecular cell*, 35(6):889–902, 2009.

30. Hope A Cole, Bruce H Howard, and David J Clark. Activation-induced disruption of nucleosome position clusters on the coding regions of Gcn4-dependent genes extends into neighbouring genes. *Nucleic acids research*, 39(22):9521–9535, 2011.

31. Ashish Kumar Singh, Tamás Schauer, Lena Pfaller, Tobias Straub, and Felix Mueller-Planitz. The biogenesis and function of nucleosome arrays. *Nature communications*, 12(1):7011, 2021.

32. Binbin Lai, Weiwu Gao, Kairong Cui, Wanli Xie, Qingsong Tang, Wenfei Jin, Gangqing Hu, Bing Ni, and Keji Zhao. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature*, 562(7726):281–285, 2018.

33. Jun Wan, Jimmy Lin, Donald J Zack, and Jiang Qian. Relating periodicity of nucleosome organization and gene regulation. *Bioinformatics*, 25(14):1782–1788, 2009.

34. Răzvan V Chereji and David J Clark. Major determinants of nucleosome positioning. *Biophysical journal*, 114(10):2279–2289, 2018.

35. Yanghui Liu, Yehua Li, Raymond J Carroll, and Naisyin Wang. Predictive functional linear models with diverging number of semiparametric single-index interactions. *Journal of Econometrics*, 230(2):221–239, 2022.

36. Reka Karuppusami, Belavendra Antonisamy, and Prasanna S Premkumar. Functional principal component analysis for identifying the child growth pattern using longitudinal birth cohort data. *BMC Medical Research Methodology*, 22(1):76, 2022.

37. Li Luo, Yun Zhu, and Momiao Xiong. Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation. *European Journal of Human Genetics*, 21(2):217–224, 2013.

38. Leo Zeitler. leoTiez/nucleosome-fpca: v1.0.0, September 2023.

39. Daechan Park, Adam R Morris, Anna Battenhouse, and Vishwanath R Iyer. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic acids research*, 42(6):3736–3749, 2014.

40. Avital Klein-Brill, Daphna Joseph-Strauss, Alon Appleboim, and Nir Friedman. Dynamics of chromatin and transcription during transient depletion of the RSC chromatin remodeling complex. *Cell reports*, 26(1):279–292, 2019.

41. Donald O Hebb. Organization of behavior. *New York: Wiley and Sons. J. Clin. Psychology*, 6(3):335–307, 1949.

42. Seungbyn Baek and Insuk Lee. Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. *Computational and structural biotechnology journal*, 18:1429–1439, 2020.

43. Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

44. Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

45. Slawomir Kubik, Maria Jessica Bruzzone, Philippe Jacquet, Jean-Luc Falcone, Jacques Rougemont, and David Shore. Nucleosome stability distinguishes two different promoter types at all protein-coding genes in yeast. *Molecular cell*, 60(3):422–434, 2015.

46. Bradley R Cairns, Yahli Lorch, Yang Li, Mincheng Zhang, Lynne Lacomis, Hediye Erdjument-Bromage, Paul Tempst, Jian Du, Brehon Laurent, and Roger D Kornberg. RSC, an essential, abundant chromatin-remodeling complex. *Cell*, 87(7):1249–1260, 1996.

47. Nils Krietenstein, Megha Wal, Shinya Watanabe, Bongsoo Park, Craig L Peterson, B Franklin Pugh, and Philipp Korber. Genomic nucleosome organization reconstituted with pure proteins. *Cell*, 167(3):709–721, 2016.

48. Marc Damelin, Itamar Simon, Terence I Moy, Boris Wilson, Suzanne Komili, Paul Tempst, Frederick P Roth, Richard A Young, Bradley R Cairns, and Pamela A Silver. The genome-wide localization of Rsc9, a component of the RSC chromatin-remodeling complex, changes in response to stress. *Molecular cell*, 9(3):563–573, 2002.

49. Huck Hui Ng, François Robert, Richard A Young, and Kevin Struhl. Genome-wide location and regulated recruitment of the RSC nucleosome-remodeling complex. *Genes & development*, 16(7):806–819, 2002.

50. Tiffani Kiyoko Quan and Grant Ashley Hartzog. Histone H3K4 and K36 methylation, Chd1 and Rpd3S oppose the functions of Saccharomyces cerevisiae Spt4–Spt5 in transcription. *Genetics*, 184(2):321–334, 2010.

51. Daechan Park, Haridha Shivram, and Vishwanath R Iyer. Chd1 co-localizes with early transcription elongation factors independently of H3K36 methylation and releases stalled RNA polymerase II at introns. *Epigenetics & chromatin*, 7(1):1–11, 2014.

52. Rajna Simic, Derek L Lindstrom, Hien G Tran, Kelli L Roinick, Patrick J Costa, Alexander D Johnson, Grant A Hartzog, and Karen M Arndt. Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes. *The EMBO journal*, 22(8):1846–1856, 2003.

53. Concetta GA Marfella and Anthony N Imbalzano. The Chd family of chromatin remodelers. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 618(1-2):30–40, 2007.

54. Kévin M André, Nathalie Giordanengo Aiach, Veronica Martinez-Fernandez, Leo Zeitler, Adriana Alberti, Arach Goldar, Michel Werner, Cyril Denby Wilkes, and Julie Soutourina. Functional interplay between Mediator and RSC chromatin remodeling complex controls nucleosome-depleted region maintenance at promoters. *Cell Reports*, 42(5), 2023.

55. Felix Krueger. Trim Galore. `https://github.com/FelixKrueger/TrimGalore/releases/tag/0.6.5`, 2019.

56. Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

57. Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012.

58. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *bioinformatics*, 25(16):2078–2079, 2009.

59. Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, 44(W1):W160–W165, 2016.

60. Adrien Georges, Diyavarshini Gopaul, Cyril Denby Wilkes, Nathalie Giordanengo Aiach, Elizaveta Novikova, Marie-Bénédicte Barrault, Olivier Alibert, and Julie Soutourina. Functional interplay between Mediator and RNA polymerase II in Rad2/XPG loading to the chromatin. *Nucleic acids research*, 47(17):8988–9004, 2019.

61. Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

62. Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

63. Carlos Ramos-Carreño, José Luis Torrecilla, Miguel Carbajo-Berrocal, Pablo Marcos, and Alberto Suárez. scikit-fda: a Python package for functional data analysis. *arXiv preprint arXiv:2211.02566*, 2022.

A Supplementary Figures

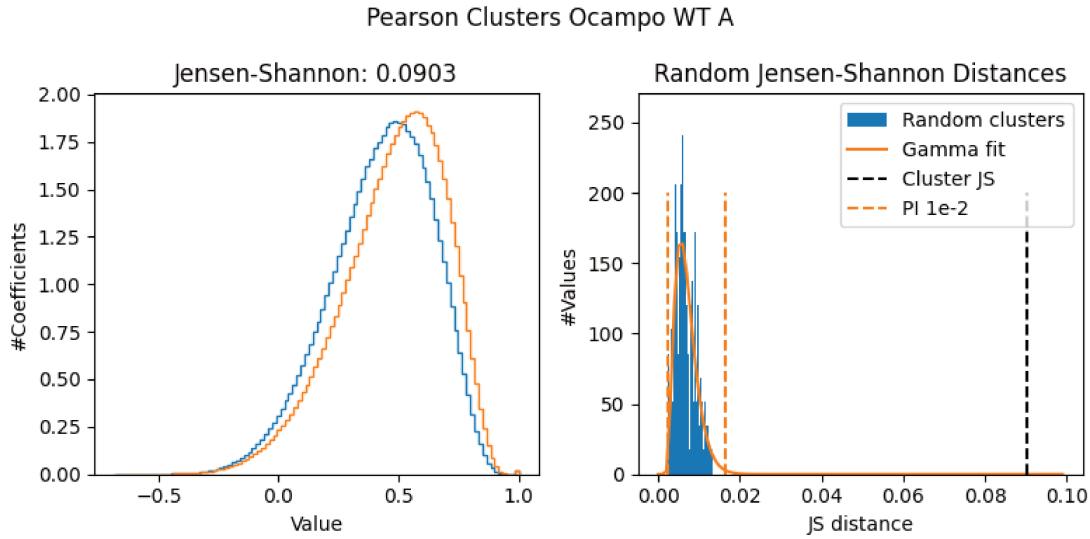


Figure A.1: Cluster significance test for all genes. Left: The Pearson correlation coefficients for each profile (blue and orange) in a cluster to all other distributions (independent of the cluster) is seemingly very similar for both groups, as indicated by shape-independent the JS distance. Right: By measuring the JS for 500 random and mutually distinct clusters (blue bars), we can approximate the expected distance over two random groupings of the Pearson coefficients using a gamma distribution (orange solid line). Indeed, the JS between our initially determined clusters (dashed black line) is outside the 99%-PI (dashed orange lines), proving that the separation is significant.

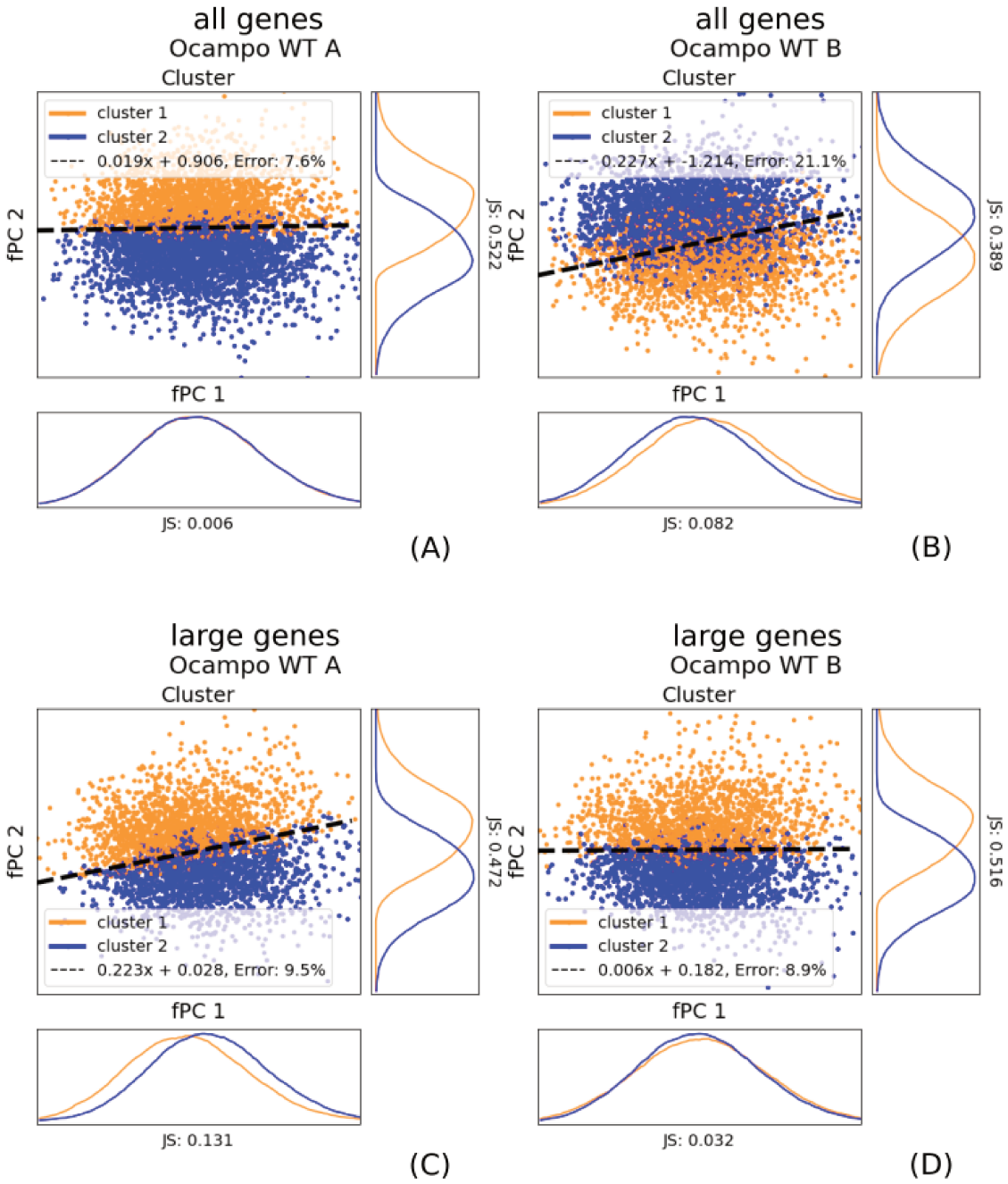


Figure A.2: The WT fPC scores ζ coloured with respect to the Pearson clustering using all genes (part 1). Blue and orange indicate each one group, the dashed line symbolises the best linear separation using a SVM. The x-axis represents the score of the first fPC ζ^1, the y-axis gives the score for the second fPC ζ^2. All axes are scaled to the same size; shapes are therefore comparable. (A) and (B) show all genes for replicate A and B. (C) and (D) display the fPC scores after filtering for large genes (> 1000 bp) for replicates A and B.

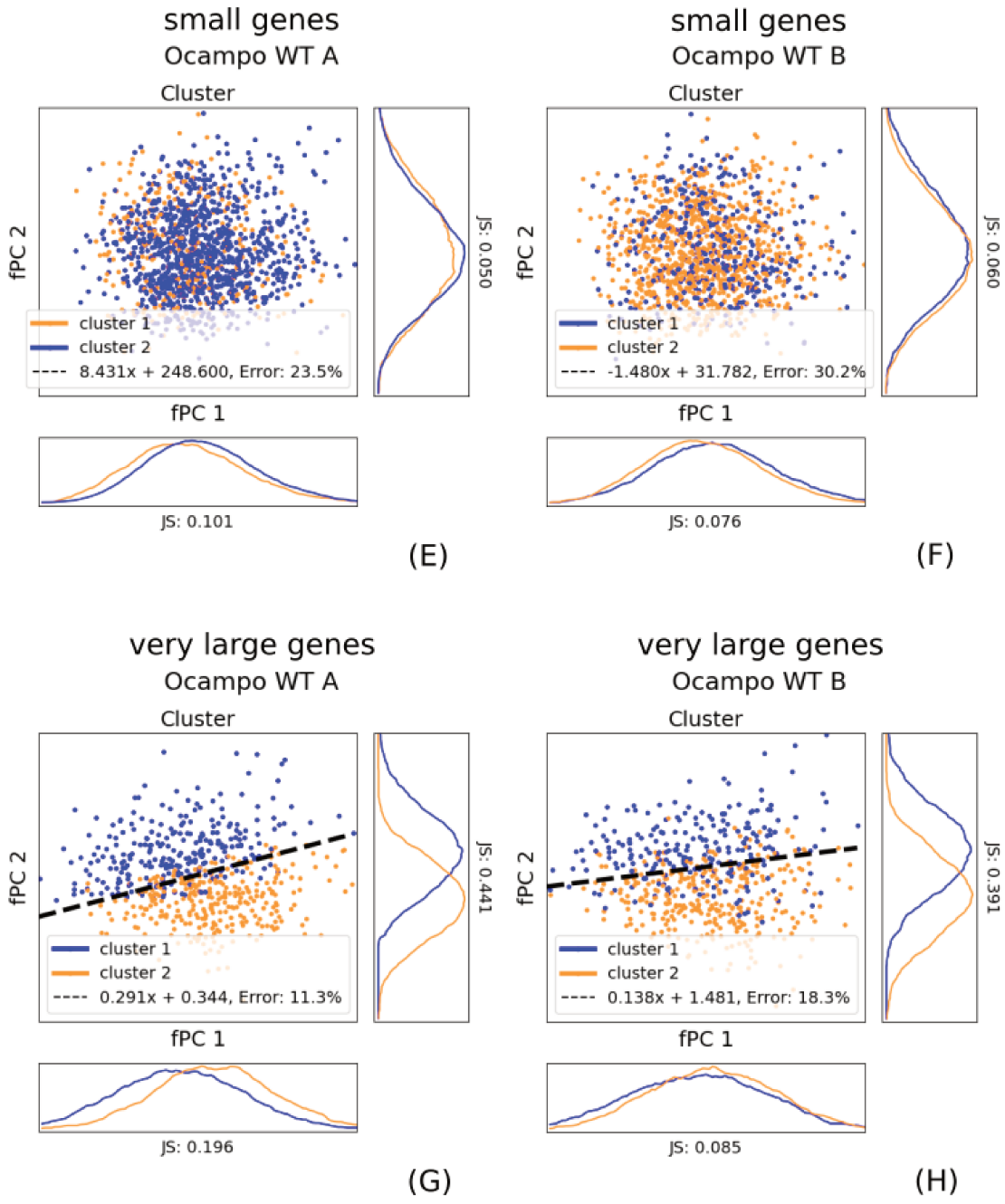


Figure A.2: The WT fPC scores ζ coloured with respect to the Pearson clustering using all genes (part 2). Blue and orange indicate each one group, the dashed line symbolises the best linear separation using a SVM. The x-axis represents the score of the first fPC ζ^1, the y-axis gives the score for the second fPC ζ^2. All axes are scaled to the same size; shapes are therefore comparable. (A) and (B) show small genes ($(\leq$ 1000 bp) for replicate A and B. We removed the separating boundary because it did not reasonably divide the clusters. Nevertheless, we kept the estimated linear function in the legend to allow a comparison with other boundaries. Of particular note is the bias, which can be even order of magnitudes different from large-gene clusters. (C) and (D) display the fPC scores after filtering for very large genes ($>$ 3000 bp) for replicates A and B.

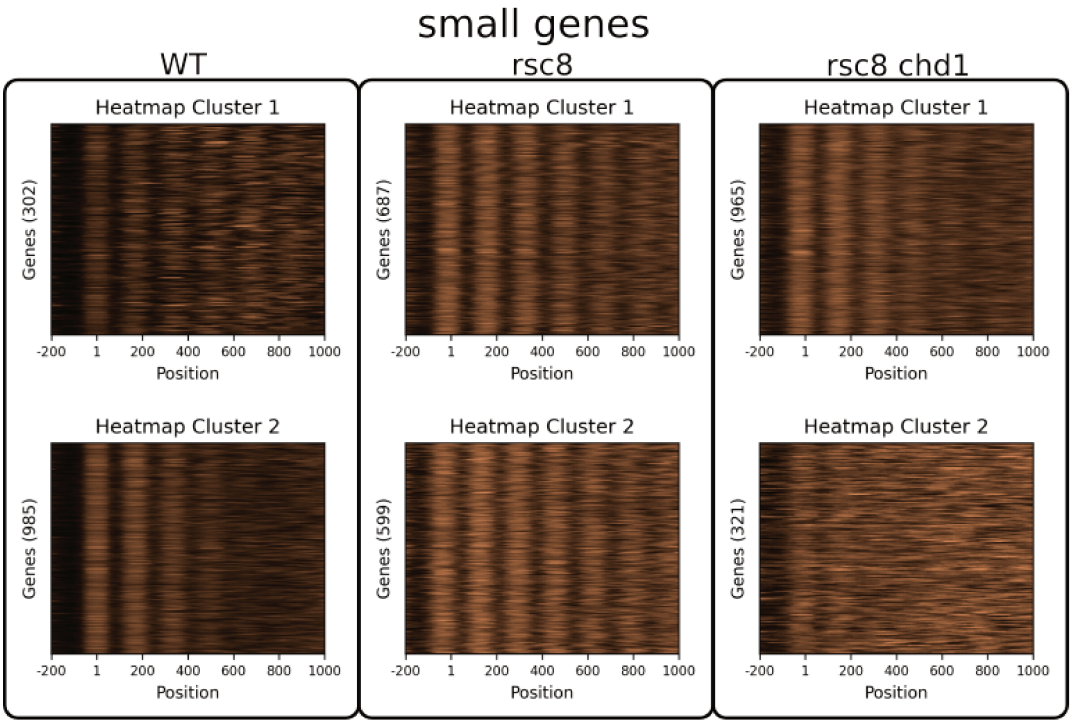


Figure A.3: Heatmaps for small-gene nucleosome profiles reveal antagonistic roles for Rsc8 and Chd1 to establish phasing boundaries. Cluster 1 and 2 for all genes in WT conditions were plotted only including small genes on the left. Indeed, correct positioning is either completely disrupted (Cluster 1), or clear phasing is lost after +3 or +4 position and individual peaks do not stand out thereafter (Cluster 2). However, both Pearson clusters for *rsc8*-depleted cells (centre) show clear phasing probabilities, despite all genes being smaller than the considered 1000 bp after the +1. The double mutant *chd1*Δ*rsc8* seems to re-establish the gene boundaries for nucleosome phasing, as positioning is either disrupted (Cluster 2, compare with Cluster 1 in WT) or does not exhibit clearly distinguishable peaks after the +3 or +4 nucleosome (Cluster 1, compare with Cluster 2 in WT). Defining a group as being 1 or 2 was arbitrary and has no significance. Copper values show large MNase-seq signal values, whereas dark segments indicate a low amplitude.

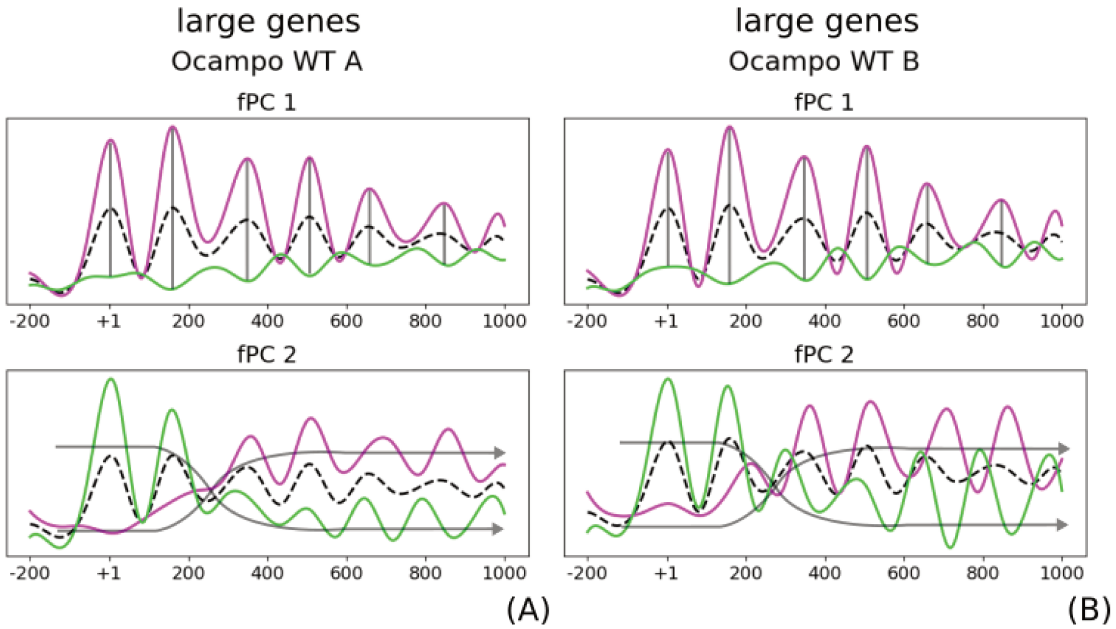


Figure A.4: The large-gene fPC effect in WT. Despite fact that the functions differ in the A and B replicate ((A) and (B)), they both describe the same properties as when considering all genes (Fig 1(F)). To be precise, the first fPC describes seemingly position-dependent scaling (grey vertical bars), and the second explains coordinated phasing (grey arrows). The mean is displayed as a black dashed line, whereas a positive and a negative functional contribution are given in magenta and green, respectively.

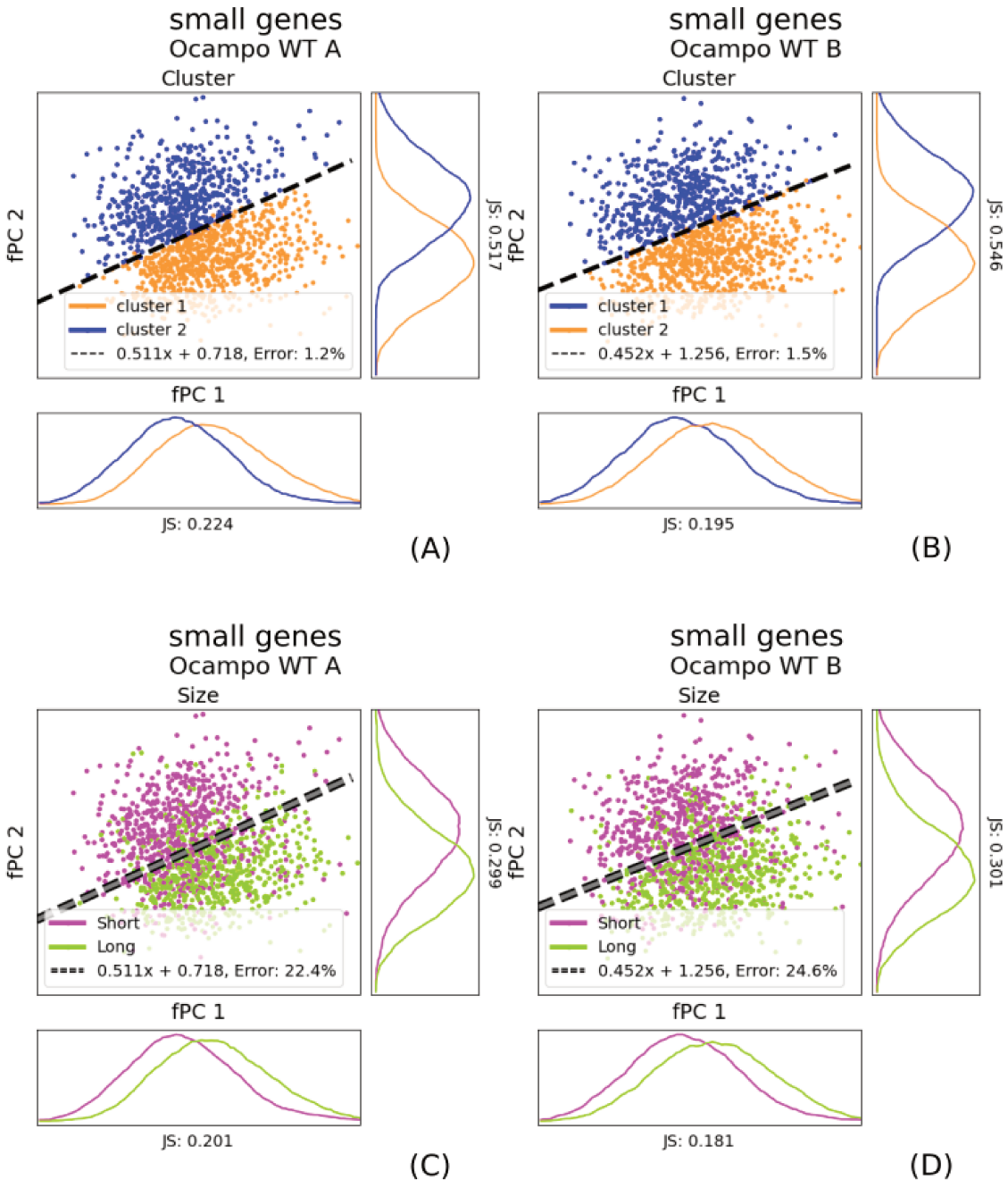


Figure A.5: The Pearson coefficient clusters for exclusively small genes correspond to the gene size. When we repeated the Pearson coefficient clustering considering exclusively small genes, we can linearly separate again the two groups (orange and blue). However, this is predominantly explained by the size of the gene (short pink, long green). This in line with the hypothesis that coordinated nucleosome phasing along the transcribed region is strictly limited within the gene body. The phase separating line was determined on the Pearson clusters (dashed black line) using an SVM. The same separating boundary was also plotted in right plot showing grouping with respect to the size. We plotted the original SVM boundary from the Pearson clusters with a dashed grey line to indicate that it was not determined using gene size. (A) and (B) give the Pearson clusters for replicate A and B. (C) and (D) show the size dependence of replicate A and B.

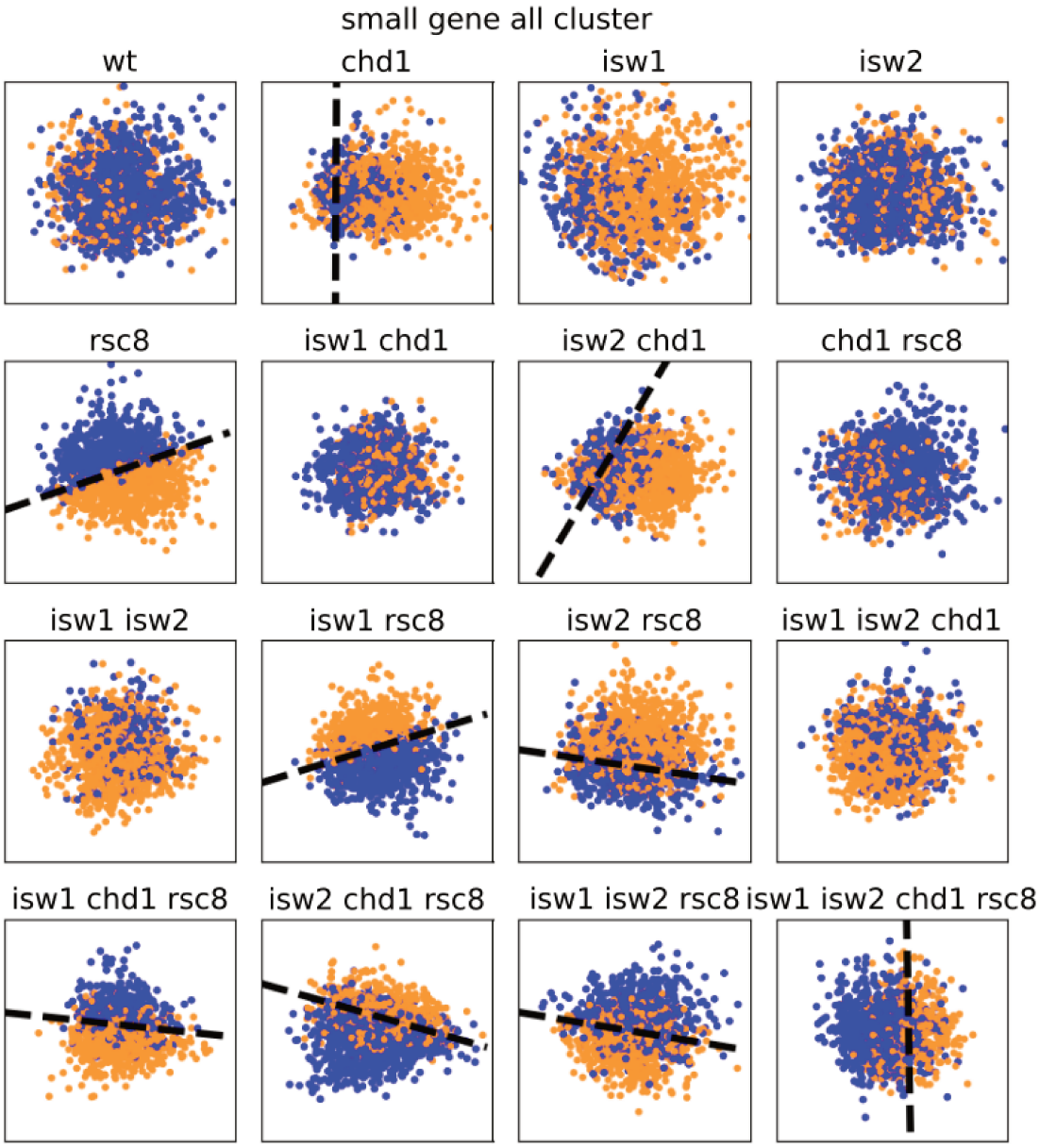


Figure A.6: Pearson clusters of small genes lose separability with respect to their fPC scores. The figure shows the fPC scores ζ of small genes (< 1000 bp) of all conditions coloured with respect to the *all-gene* Pearson clustering. Blue and orange indicate each one group, the dashed line symbolises the best linear separation using a SVM. We removed the linear boundary in plots where it went through the periphery instead of dividing the data points. The x-axis represents the score of the first fPC ζ^1, the y-axis gives the score for the second fPC ζ^2. All axes are scaled to the same size; shapes are therefore comparable.

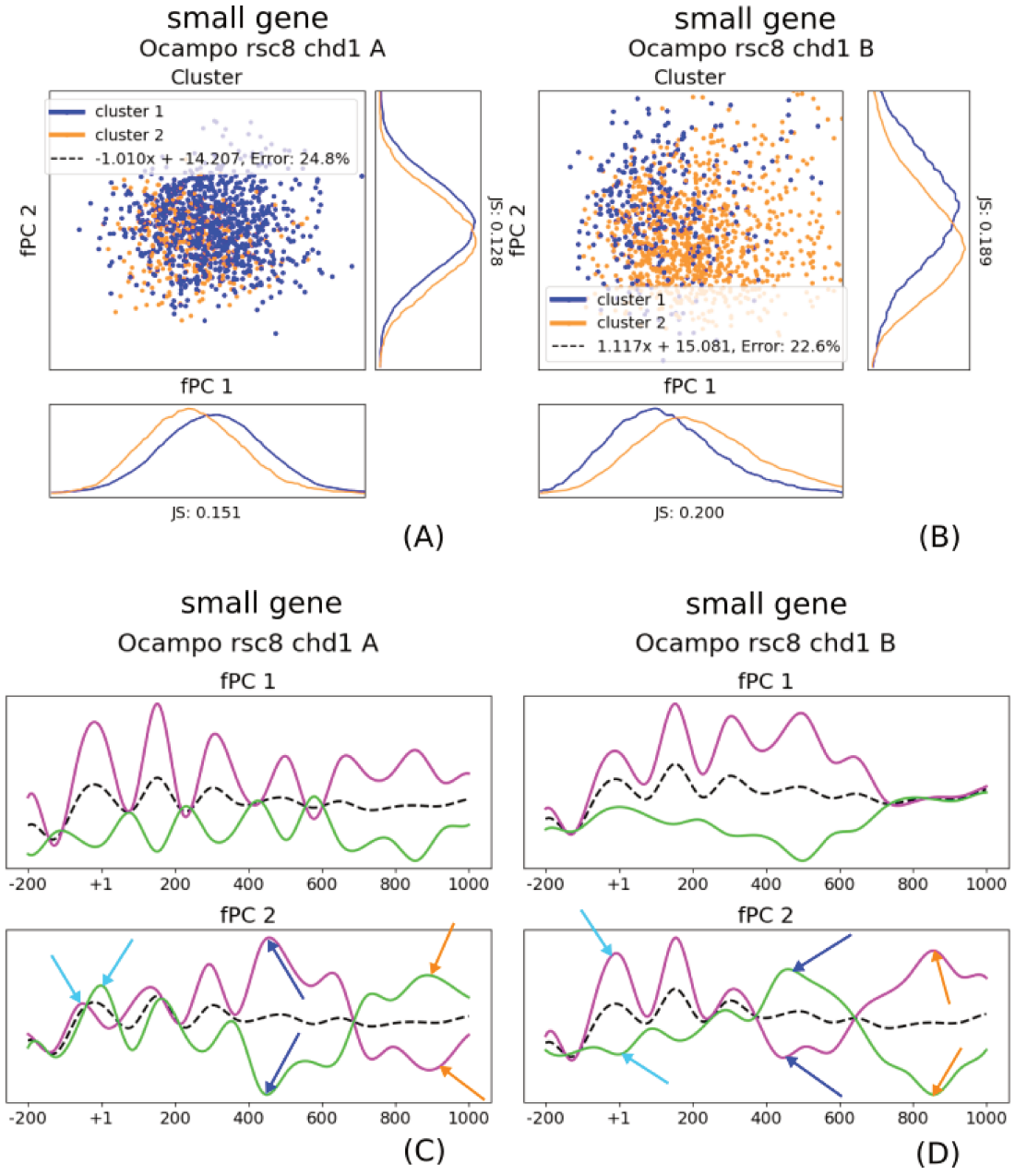


Figure A.7: The small-gene fPC effect in *chd1Δrsc8* strains. The double mutant seemingly re-establishes gene boundaries, and coordinated phasing is at least weakened after the +2 nucleosome (+1 in turquoise, +4 in blue, +6 in orange). This is true despite the fact that the A and B replicate differ. Figs (A) and (B) show the clusters for replicate *A* and *B*, and Figs (C) and (D) display their fPCs. We removed the separating boundaries in (A) and (B) because they did not reasonably divide the clusters. Nevertheless, we kept the estimated linear function in the legend to allow a comparison with other boundaries. Of particular note is the bias, which differs largely from large-gene clusters. The dashed black lines, the solid purple, and the solid green lines indicate the mean, a positive contribution, and a negative contribution, respectively.

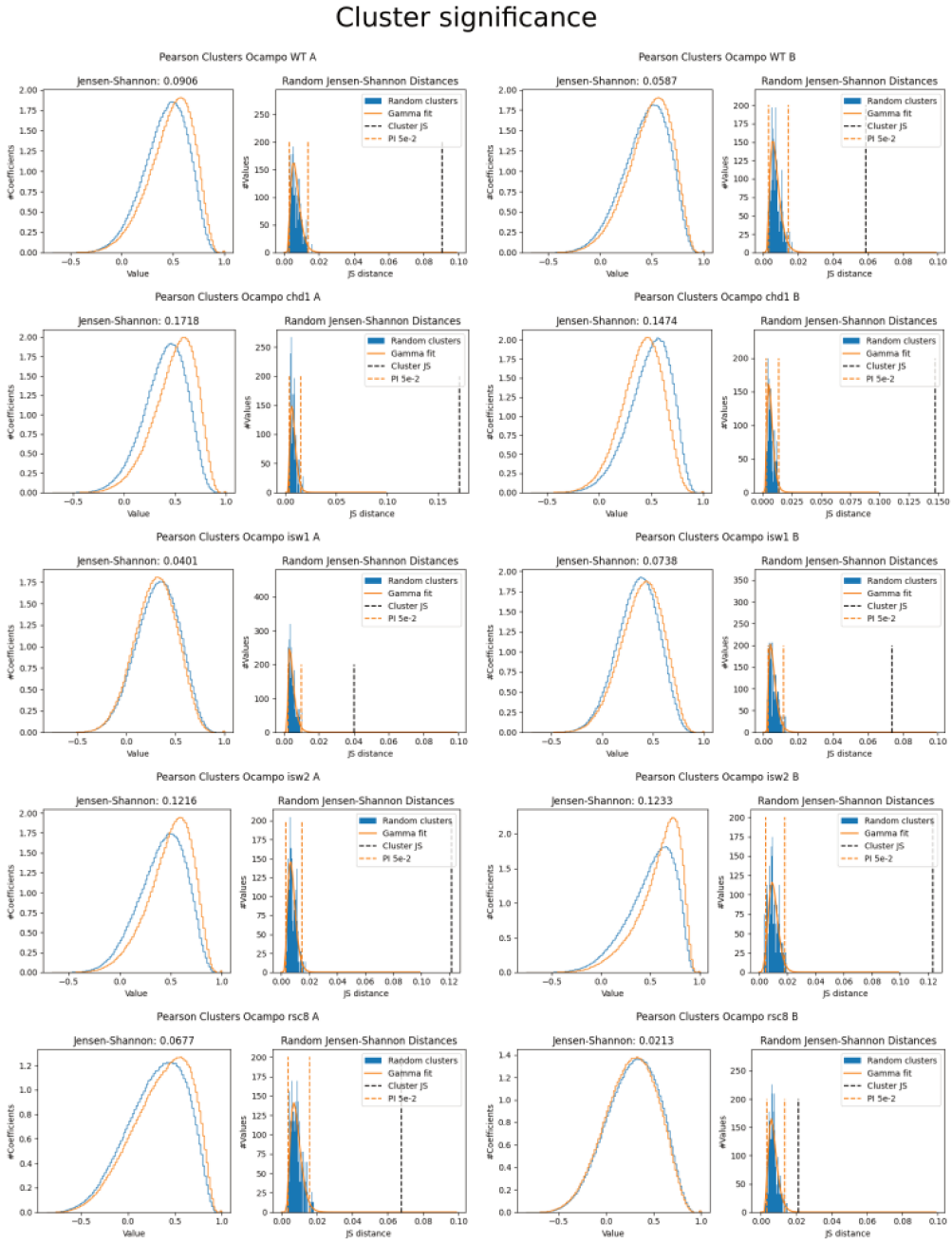


Figure A.8: Cluster significance test for large genes (part 1). Left: The Pearson correlation coefficients for each profile in a cluster to all other distributions (independent of the cluster) is seemingly very similar for both groups, as indicated by the shape-independent the JS distance. Right: By measuring the JS for 500 random and mutually distinct clusters, we can approximate the expected distance over two random groupings of the Pearson coefficients using a gamma distribution. Orange dashed lines indicate the 95% PI, the dashed black line display the JS of the Pearson clusters.

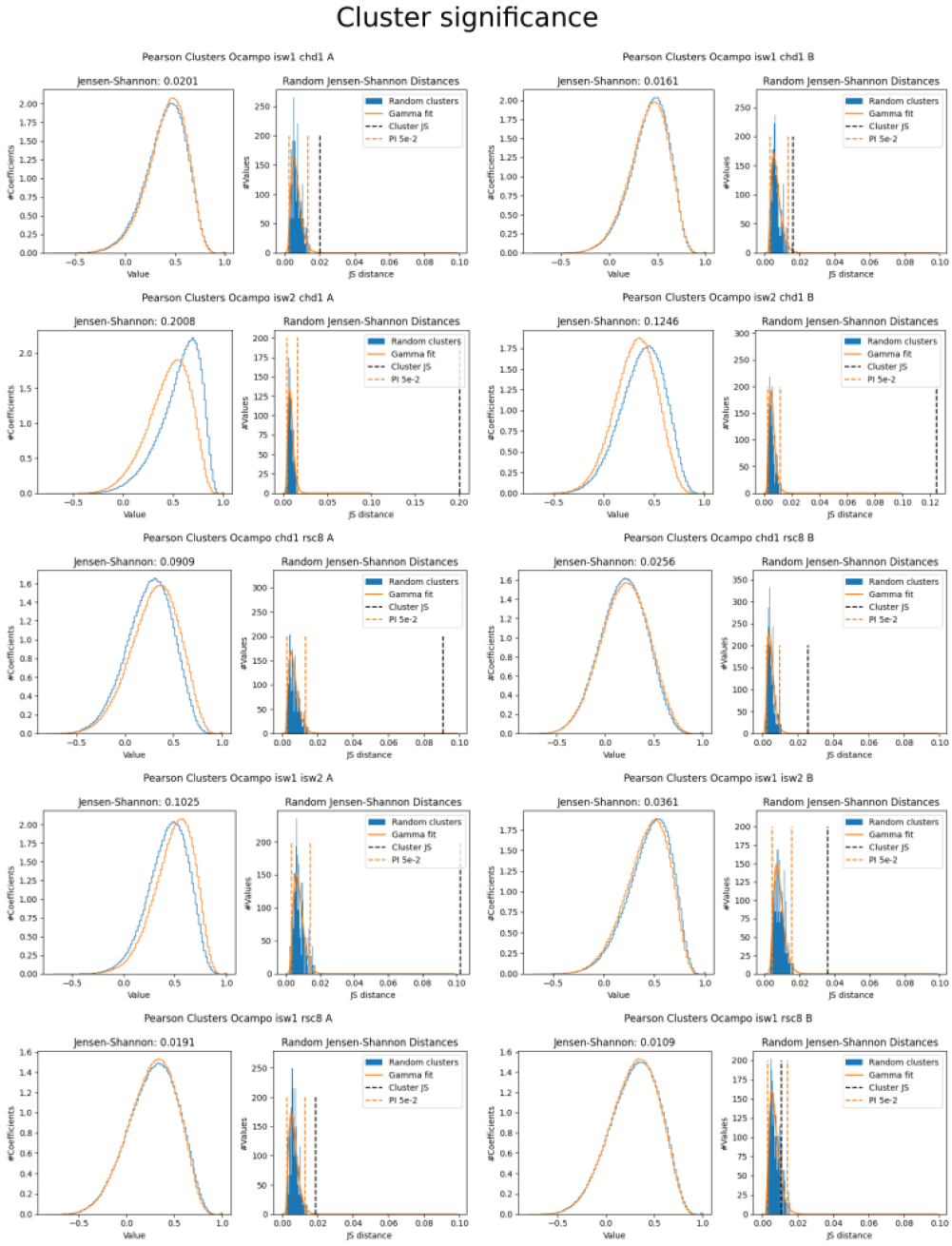


Figure A.8: Cluster significance test for large genes (part 2). Left: The Pearson correlation coefficients for each profile in a cluster to all other distributions (independent of the cluster) is seemingly very similar for both groups, as indicated by the shape-independent the JS distance. Right: By measuring the JS for 500 random and mutually distinct clusters, we can approximate the expected distance over two random groupings of the Pearson coefficients using a gamma distribution. Orange dashed lines indicate the 95% PI, the dashed black line display the JS of the Pearson clusters.

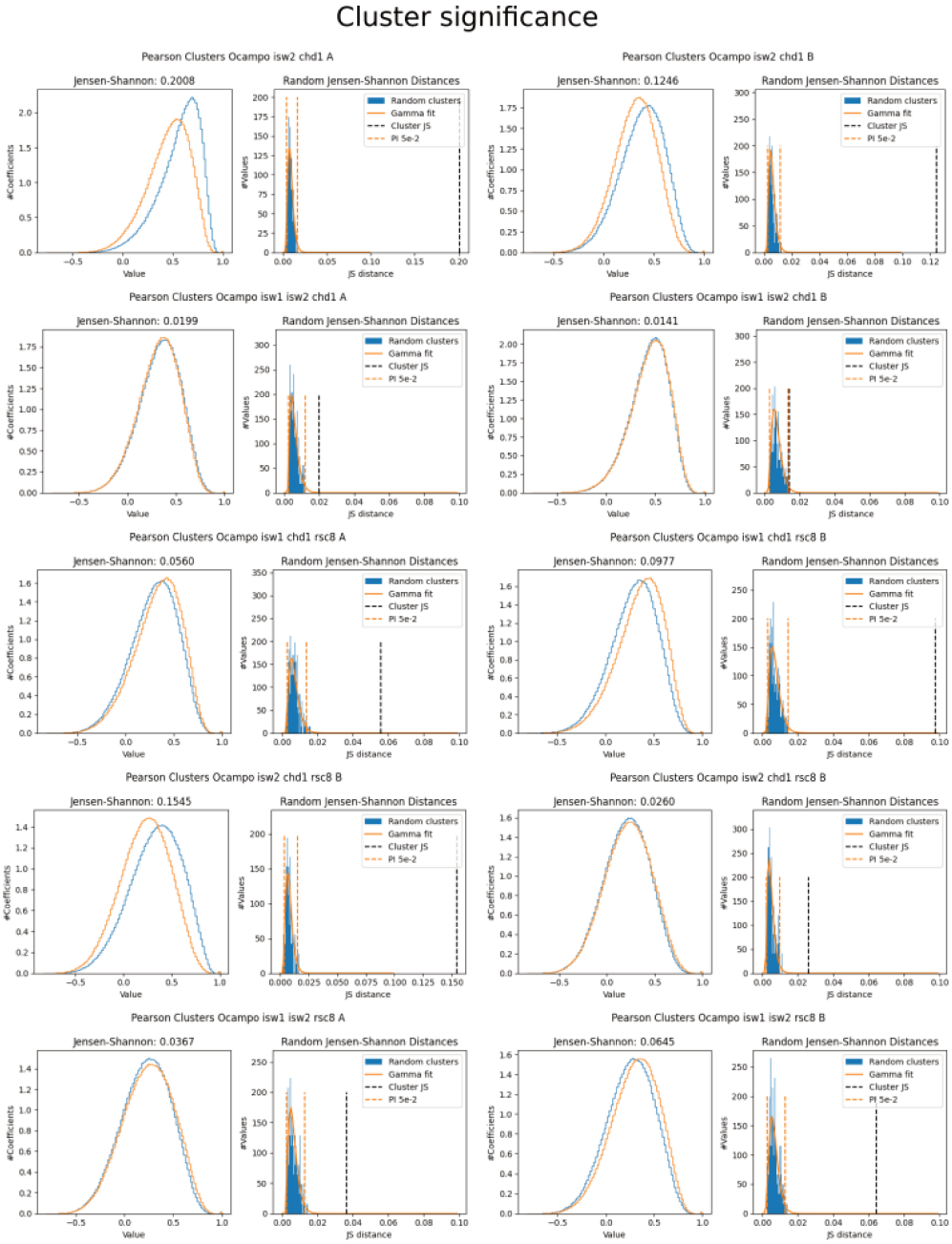


Figure A.8: Cluster significance test for large genes (part 3). Left: The Pearson correlation coefficients for each profile in a cluster to all other distributions (independent of the cluster) is seemingly very similar for both groups, as indicated by the shape-independent the JS distance. Right: By measuring the JS for 500 random and mutually distinct clusters, we can approximate the expected distance over two random groupings of the Pearson coefficients using a gamma distribution. Orange dashed lines indicate the 95% PI, the dashed black line display the JS of the Pearson clusters.

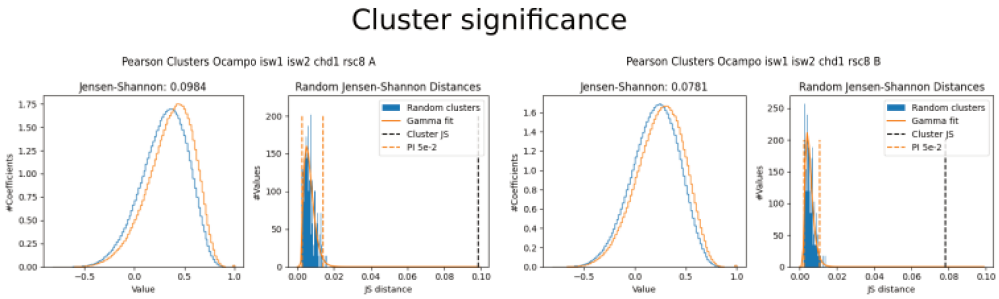


Figure A.8: Cluster significance test for large genes (part 4). Left: The Pearson correlation coefficients for each profile in a cluster to all other distributions (independent of the cluster) is seemingly very similar for both groups, as indicated by the shape-independent the JS distance. Right: By measuring the JS for 500 random and mutually distinct clusters, we can approximate the expected distance over two random groupings of the Pearson coefficients using a gamma distribution. Orange dashed lines indicate the 95% PI, the dashed black line display the JS of the Pearson clusters.

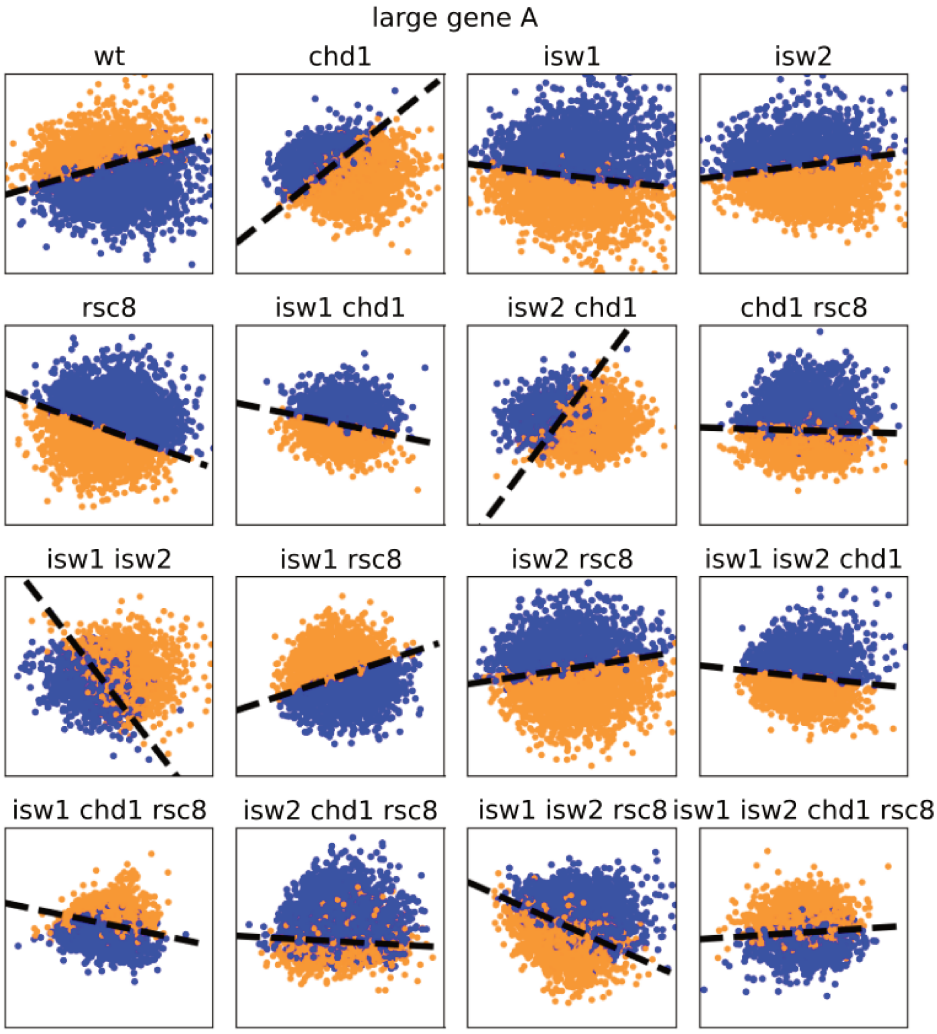


Figure A.9: Pearson clusters of large genes are linearly separable with respect to their fPC scores (replicate A). The figure shows the fPC scores ζ of all conditions coloured with respect to the Pearson clustering using only large genes ($\geq$ 1000 bp). Blue and orange indicate each one group, the dashed line symbolises the best linear separation using a SVM. The x-axis represents the score of the first fPC ζ^1, the y-axis gives the score for the second fPC ζ^2. All axes are scaled to the same size; shapes are therefore comparable. It should be emphasised that only the absolute slope value matters and not the sign (i.e. pointing upwards or downwards).

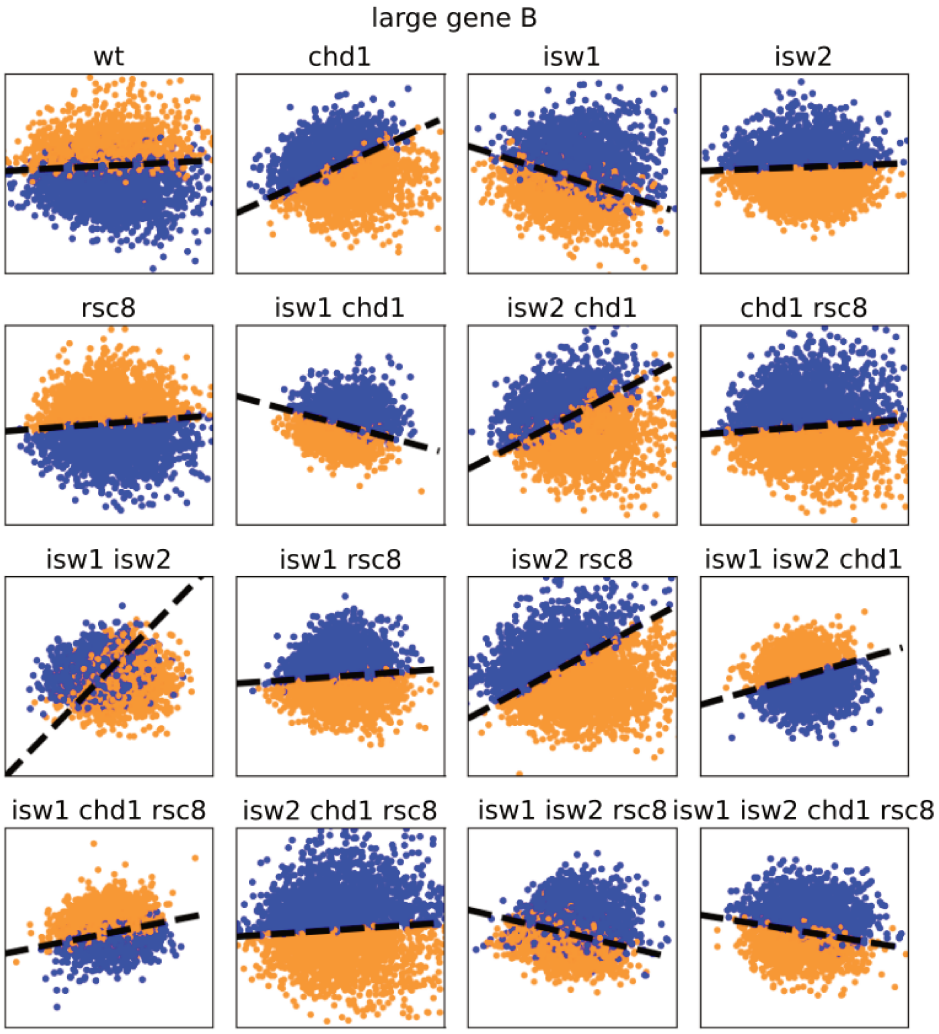


Figure A.10: Pearson clusters of large genes are linearly separable with respect to their fPC scores (replicate B). The figure shows the fPC scores ζ of all conditions coloured with respect to the Pearson clustering using only large genes ($\geq$ 1000 bp). Blue and orange indicate each one group, the dashed line symbolises the best linear separation using a SVM. The x-axis represents the score of the first fPC ζ^1, the y-axis gives the score for the second fPC ζ^2. All axes are scaled to the same size; shapes are therefore comparable. It should be emphasised that only the absolute slope value matters and not the sign (i.e. pointing upwards or downwards).

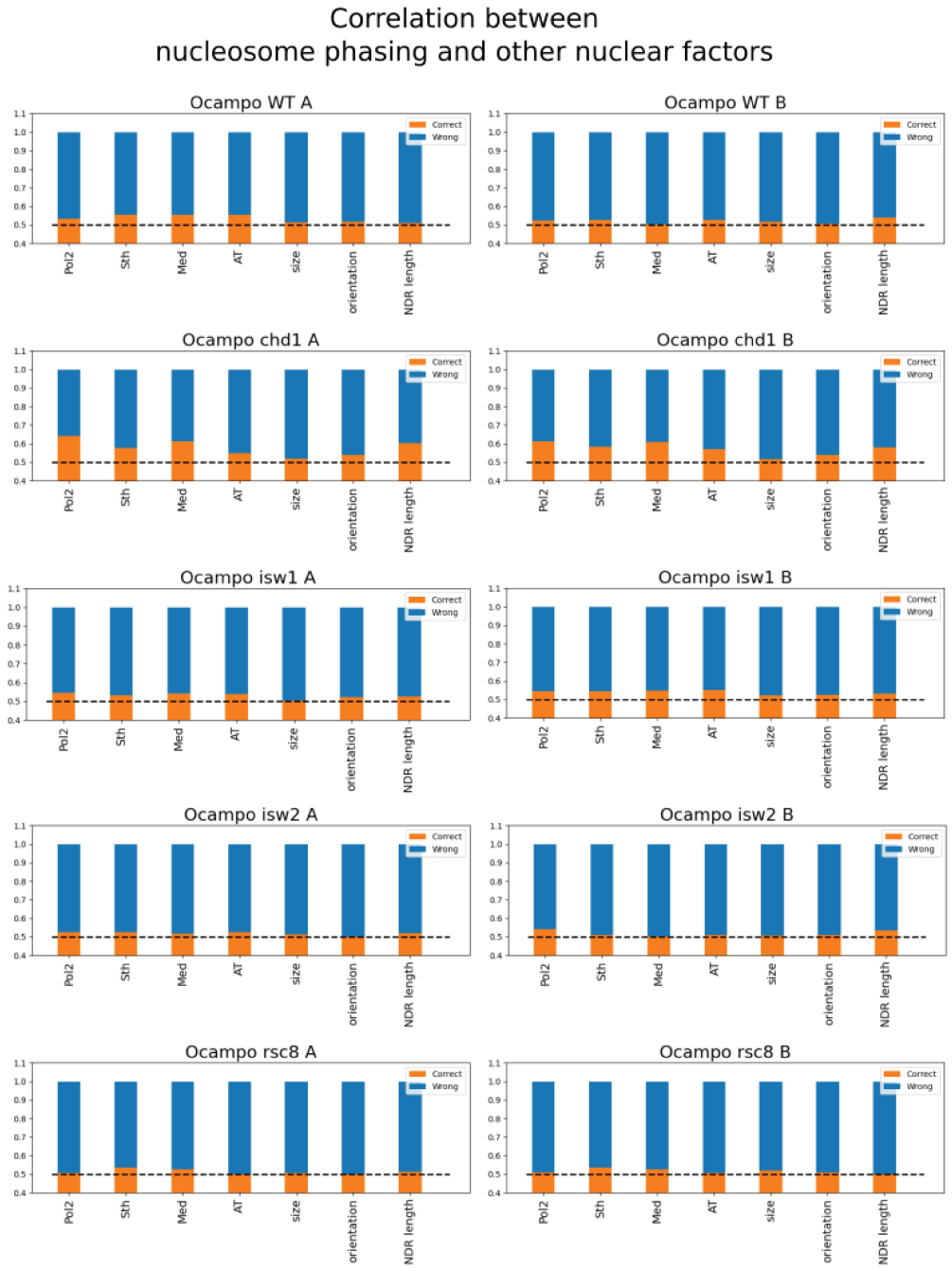


Figure A.11: Interdependence of Pearson clusters of MNase-seq profiles and other nuclear factors (part 1). The orange bar shows the ratio of cases where the nuclear factor could predict clustering, blue gives the wrongly classified ratio. Random guessing would be correct in 50% of the cases, which is given by the dashed black line. Consequently, the orange bar must exceed the dashed line to suggest interdependence.

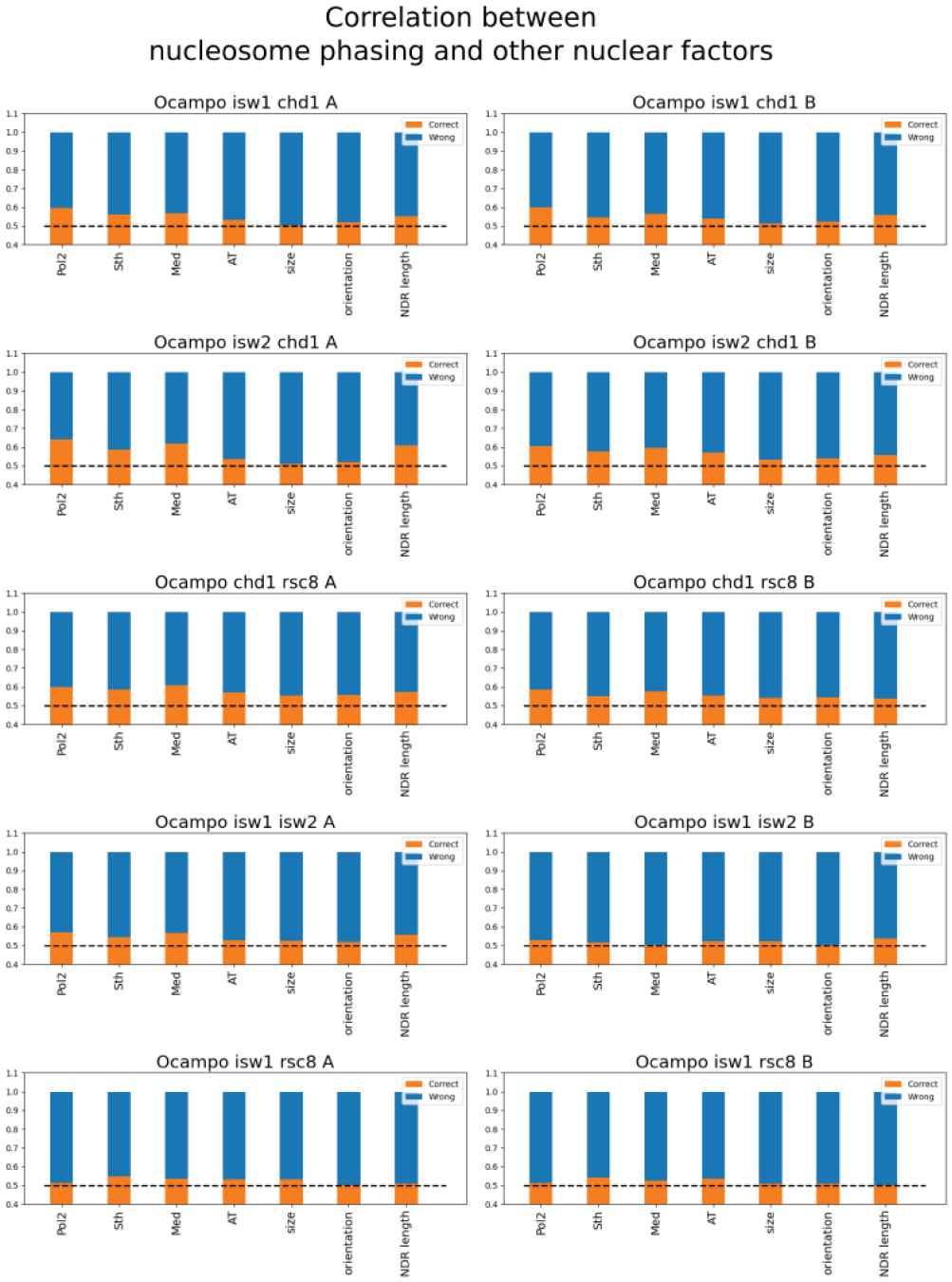


Figure A.11: Interdependence of Pearson clusters of MNase-seq profiles and other nuclear factors (part 2). The orange bar shows the ratio of cases where the nuclear factor could predict clustering, blue gives the wrongly classified ratio. Random guessing would be correct in 50% of the cases, which is given by the dashed black line. Consequently, the orange bar must exceed the dashed line to suggest interdependence.

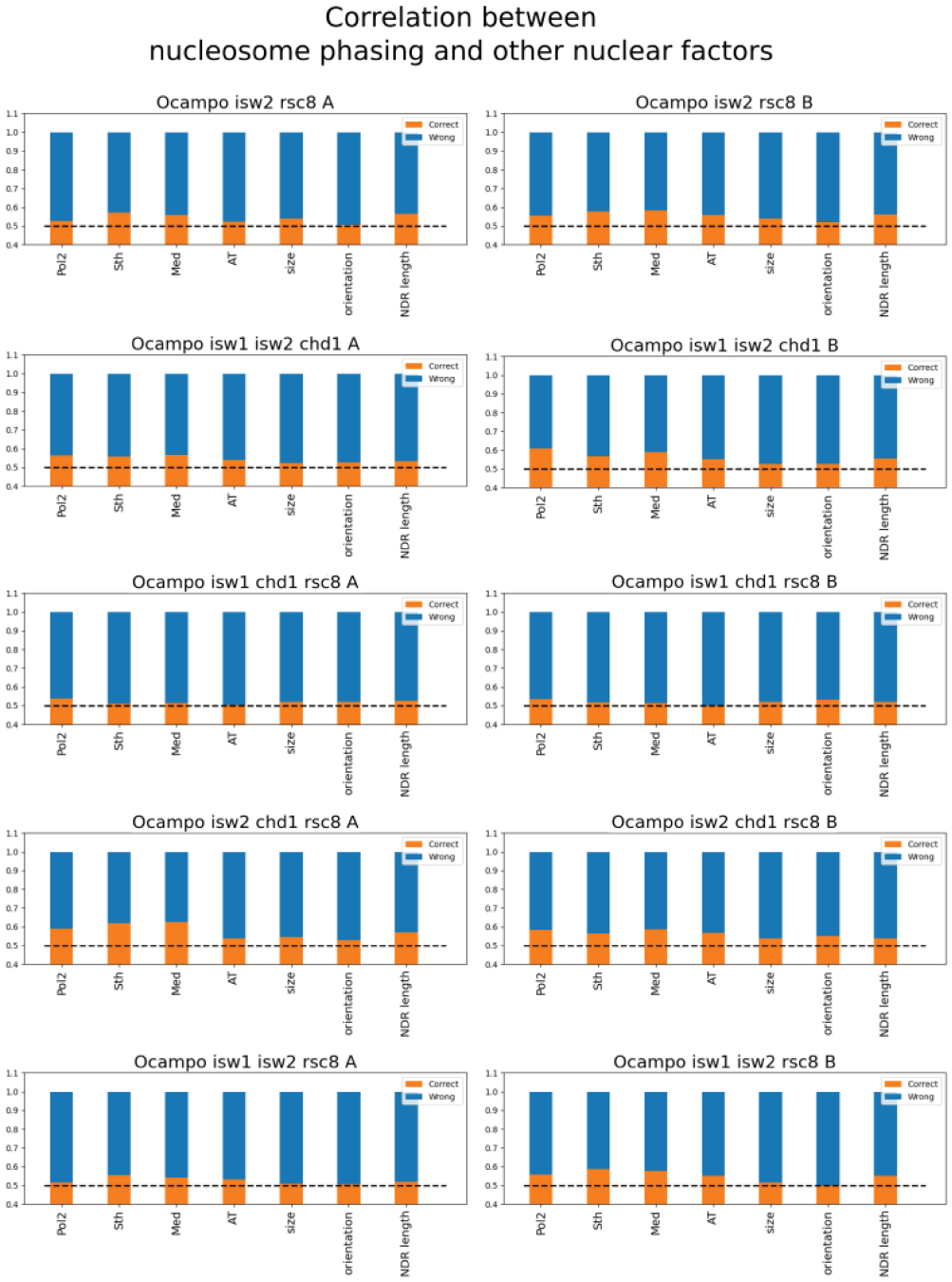


Figure A.11: Interdependence of Pearson clusters of MNase-seq profiles and other nuclear factors (part 3). The orange bar shows the ratio of cases where the nuclear factor could predict clustering, blue gives the wrongly classified ratio. Random guessing would be correct in 50% of the cases, which is given by the dashed black line. Consequently, the orange bar must exceed the dashed line to suggest interdependence.

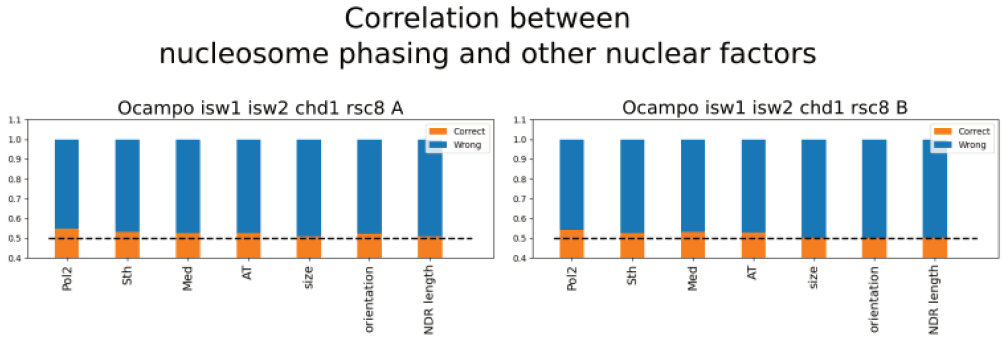


Figure A.11: Interdependence of Pearson clusters of MNase-seq profiles and other nuclear factors (part 4). The orange bar shows the ratio of cases where the nuclear factor could predict clustering, blue gives the wrongly classified ratio. Random guessing would be correct in 50% of the cases, which is given by the dashed black line. Consequently, the orange bar must exceed the dashed line to suggest interdependence.

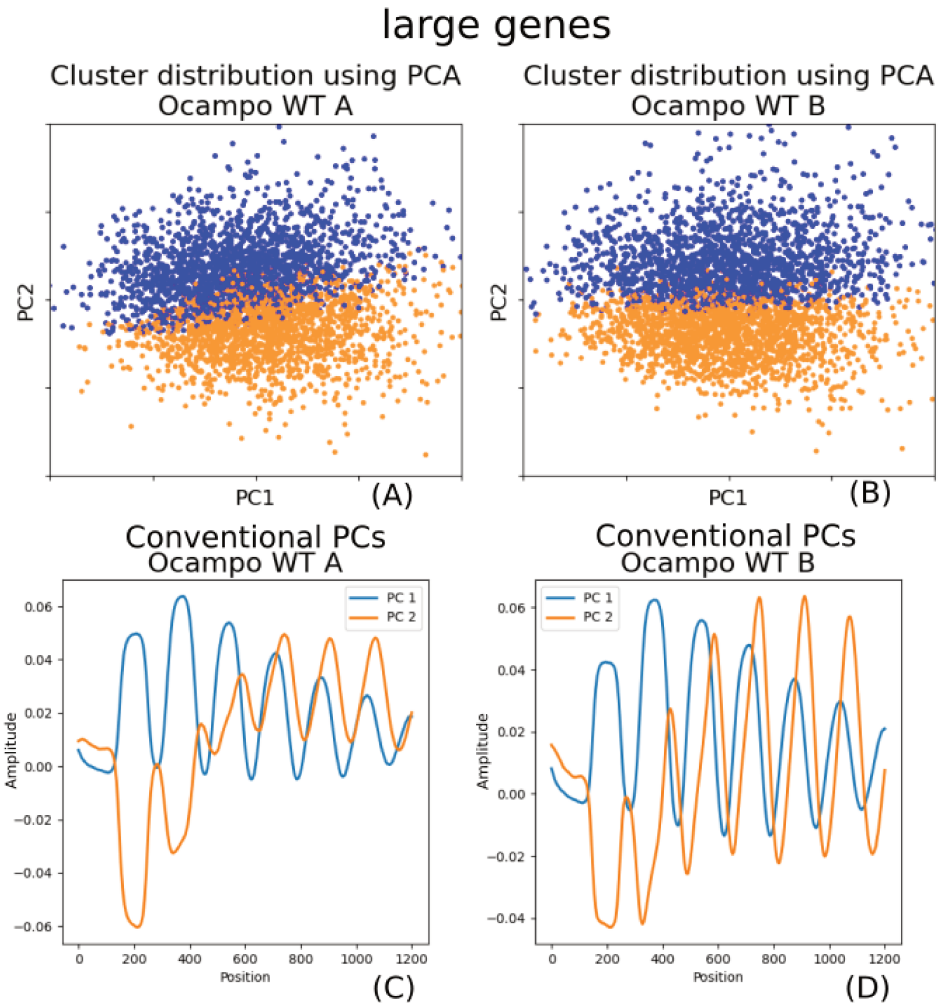


Figure A.12: Clustering distribution using PCA and their principal components. Indeed, conventional PCA can separate the clusters for all genes in WT conditions ((A) and (B) for replicate A and B) similarly to fPCA. The two clusters are given in blue and orange. However, the two determined PCs ((C) and (D) for replicate A and B) differ slightly with respect to the fPCA due to the independence assumption. Here, light blue and light orange indicate PC1 and PC2, respectively.

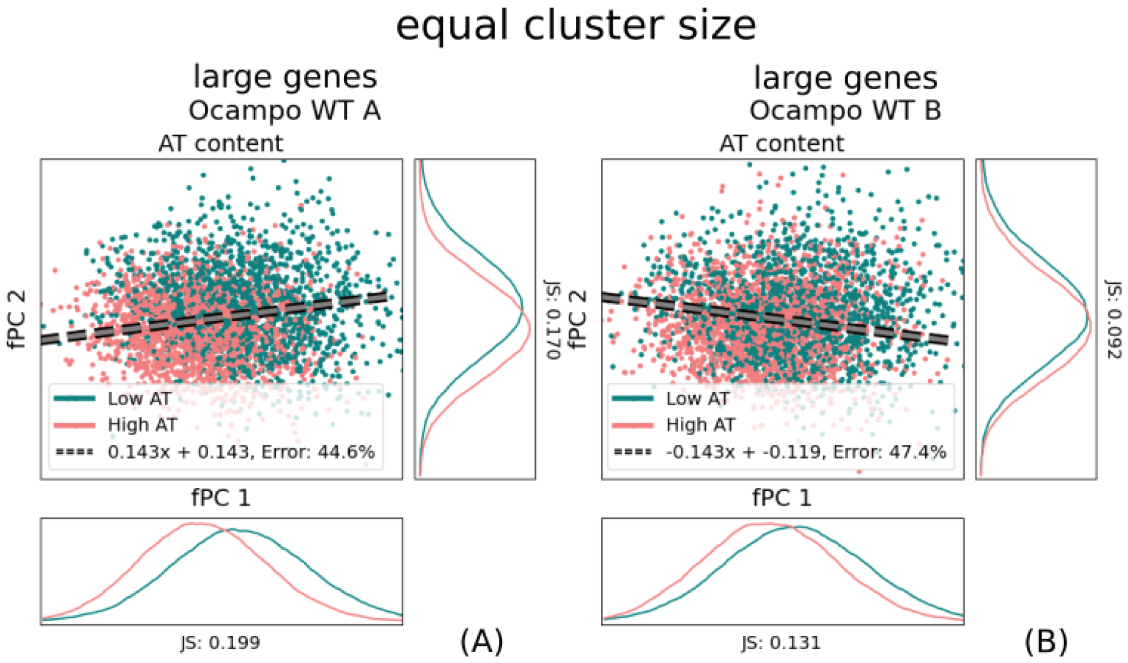


Figure A.13: AT-ratio distribution with respect to the fPC scores. Whilst there is seemingly a slight correlation between Pearson coefficient clusters and AT-ratio in the A replicate, this is trend vanishes for the B replicate. In fact, both replicates might rather distribute AT-rich and AT-poor genes orthogonal to the dividing boundary. We plotted the original SVM boundary from the Pearson clusters with a dashed grey line to indicate that it was not determined using the AT content.

Chapter 3

A Quantitative Modelling Approach for DNA repair on a Population Scale

3.1 Context and Summary

The intricate process of DNA repair requires the coordination of several steps, which need to be orchestrated together with other nuclear procedures. The number of possible influencing factors is large, and they include gene expression and nucleosome positioning. Although the previous chapter suggested that their arrangement in WT genes does not correlate with other genomic factors, we have not verified this conjecture with respect to DNA repair. NER dynamics are expected to change in different DNA regions, although it is not known how. A mechanistic explanation is still missing, and it is not clear how repair kinetics are organised on a global scale.

We developed a top-down data-driven modelling approach that avoids specific assumptions about the repair process itself. Instead, we solely presume that proteins follow particle dynamics as described in Section 1.3. This allows an unbiased evaluation of the temporal repair evolution in various DNA regions. By understanding CPD-seq data as the superposition of independent cells—rather than as an average over the entire cell culture—we derive a minimal model with only three parameters to describe repair on a population scale. Our predictions are in line with independently probed eXcision Repair sequencing (XR-seq) data (Li et al. (2018))—which measures ongoing repair at a given time point—validating our methods and data interpretation.

The model parameters—which describe the entire repair evolution rather than single time points—can be conveniently correlated with other biological properties, such as transcription and nucleosome density. Our outcomes are in line with other studies (Mao et al. (2016); Li et al. (2018); Yu et al. (2016);

van Eijk et al. (2019)), proving that the model can establish known links. As conjectured, the method does not suggest any correlation with nucleosome density along protein-coding genes, although it is highly important in intergenic regions. Surprisingly, the analysis indicates a strong link between DNA repair and gene size, which has never been proposed before to our knowledge. The system is therefore equally capable of finding new interrelationships.

As the first author, I was responsible for model development, mathematical formulation, and implementation, as well as leading on paper writing and editing. Together with my colleagues, I was involved in contacting journal editors and replying to reviewers.

A Quantitative Modelling Approach for DNA repair on a Population Scale

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A quantitative modelling approach for DNA repair on a population scale

Leo Zeitler, Cyril Denby Wilkes, Arach Goldar, Julie Soutourina*

Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France

* julie.soutourina@cea.fr

OPEN ACCESS

Citation: Zeitler L, Denby Wilkes C, Goldar A, Soutourina J (2022) A quantitative modelling approach for DNA repair on a population scale. PLoS Comput Biol 18(9): e1010488. https://doi.org/10.1371/journal.pcbi.1010488

Editor: Carl Herrmann, Heidelberg University, GERMANY

Received: March 30, 2022

Accepted: August 12, 2022

Published: September 12, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010488

Data Availability Statement: All code that was used for the study is available on https://github.com/leoTiez/jmak.

Funding: LZ was supported by the CEA NUMERICS program, which has received funding from

Abstract

The great advances of sequencing technologies allow the *in vivo* measurement of nuclear processes—such as DNA repair after UV exposure—over entire cell populations. However, data sets usually contain only a few samples over several hours, missing possibly important information in between time points. We developed a data-driven approach to analyse CPD repair kinetics over time in *Saccharomyces cerevisiae*. In contrast to other studies that consider sequencing signals as an average behaviour, we understand them as the superposition of signals from independent cells. By motivating repair as a stochastic process, we derive a minimal model for which the parameters can be conveniently estimated. We correlate repair parameters to a variety of genomic features that are assumed to influence repair, including transcription rate and nucleosome density. The clearest link was found for the transcription unit length, which has been unreported for budding yeast to our knowledge. The framework hence allows a comprehensive analysis of nuclear processes on a population scale.

Author summary

As DNA encodes our very identity, it has been subject to a plethora of studies over the last century. The advent of new technologies that permit rapid sequencing of large DNA and RNA samples opened doors to before unknown mechanisms and interactions on a genomic scale. This led to an in-depth analysis of several nuclear processes, including transcription of genes and lesion repair. However, the applied protocols do not allow a high temporal resolution. Quite the contrary, the experiments yield often only some few data signals over several hours. The details of the dynamics between time points are chiefly ignored, implicitly assuming that they straightforwardly transition from one to another. Here, we show that such an understanding can be flawed. We use the repair process of UV-induced DNA damage as an example to present a quantitative analysis framework that permits the representation of the entire temporal process. We subsequently describe how they can be linked to other heterogeneous data sets. Consequently, we evaluate a correlation to the whole kinetic process rather than to a single time point. Although the approach is exemplified using DNA repair, it can be readily applied to any other

European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 800945. The work was supported by the Fondation ARC (PGA1 RF20170205342) and Comité Ile-de-France - La Ligue Nationale Contre le Cancer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

mechanism and sequencing data that represent a transition between two states, such as *damaged* and *repaired*.

This is a *PLOS Computational Biology* Methods paper.

Introduction

As DNA represents the hereditary unit of life, maintaining its integrity is vital for every organism's survival. A large variety of different genotoxic factors have the potential to damage the molecular structure of DNA. Among others, it has been shown that UV light induces Cyclobutane Pyrimidine Dimers (CPDs). Nucleotide Excision Repair (NER) is an evolutionarily conserved repair mechanism in *Saccharomyces cerevisiae* that can remove a broad range of damage, including CPDs [1]. NER is conventionally divided into two subpathways. The first recognition mechanism is named Global-Genome Repair (GGR) and can be observed along the entire genome. DNA damage is recognised directly by protein association. There is evidence that protein loading is promoted through interactions with chromatin remodellers that change the nucleosome density or distribution [2, 3]. The second pathway is restricted to actively transcribed regions; hence the name Transcription-Coupled Repair (TCR). Expressed genes exhibit quicker repair than silent downstream regions [4]. This promoted the assumption that TCR is more efficient than GGR, although constrained to the transcribed strand (TS) [5, 6]. TCR is initiated by lesion-blocked RNA polymerase II (Pol II) which cannot continue elongation [7]. Thus, a potential link of TCR to transcription rate has been indicated by several studies [8, 9]. After recognition, TCR and GGR use the same incision and nucleotide replacement mechanism. DNA is incised to either side of the lesion leaving an approximately 30-nucleotide gap, which is subsequently replaced and ligated (for a comprehensive description and analysis, see the review by [10]).

Our understanding of such processes in living cells has been largely enhanced by Next Generation Sequencing (NGS). It allows the identification of enriched loci of a selected property on a genome-wide scale. Among others, it has been applied to investigate the CPD repair mechanisms *in vivo* through analysing temporal changes of the damage distribution. [8] obtained high-resolution CPD-seq data that are often used as a benchmark reference (see for example [9]). Their analysis indicates that single nucleosomes and DNA-bound transcription factors have an impact on the CPD formation. Moreover, they point out that repair is seemingly influenced by the CPD position with respect to the nucleosomal dyad as well as the transcription rate of genes. Another major contribution has been done by [9]. Their protocol for eXcision Repair sequencing (XR-seq) revealed strong TCR at early time points which is followed by repair in non-transcribed regions. Furthermore, [11] and [12] utilised CPD data to compute repair rates in different areas, which indicated that the process is highly organised into genomic regions. By using GGR-deficient strains, they show that repair is changing globally when the subpathway is repressed. This is compared to the distribution of repair proteins and histone modifications.

Unfortunately, due to costs and constraints in the experimental protocol, NGS data sets contain barely more than a few time points over several hours. Consequently, previous studies could only derive limited conclusions, e.g. the absolute change at different loci. We argue that

such an analysis ignores valuable information about the transitional process from one time point to another. Furthermore, it should be emphasised that sequencing signals are commonly understood as representing an average cell. We advocate an interpretation where the data is explained as the product of many independent cells. Thus, the repair dynamics are driven by non-interfering stochastic processes. Without assuming any specific molecular mechanism, we hypothesise that they are composed of two independent random variables namely accessibility to the lesion governing repair times and Brownian motions of proteins through the nucleus to find their target. It has been shown by several studies that proteins exhibit a range of different movements in the nucleus [13–18]. Diffusion has also been investigated and modelled in context of DNA repair for the Rad4-Rad23 complex [19]. Although protein movements have been used to understand specifics of NER kinetics, a framework to quantitatively describe population-based sequencing data is still lacking.

Our approach and main results can be summarised as follows. The sparse temporal resolution of NGS data sets makes it necessary to incorporate precise assumptions about the nature of the process in order to recover missing information. Here, we present a computational framework to analyse DNA repair kinetics. We derive a function to study CPD removal as a Poisson point process of independent cells. Since we do not impose any molecular mechanism, we obtain a simple and minimal representation. The parameters can be derived using the well-studied physical model for phase transitions, which is described in detail by Kolmogorov [20], Johnson and Mehl, [21], and Avrami [22–24] (KJMA model). It can be conveniently transformed to a linear regression problem and is therefore executable on almost any ordinary computer. A consequence that is implied by our repair model is that the observed change of CPDs is non-constant over time. To our understanding, this has not been explicitly incorporated in the analysis of NGS data. The model validity can be verified with independently probed XR-seq data [9]. We are able to recover particular aspects of the NER kinetics despite our broadly applicable assumptions. We ultimately use the framework to predict correlations with other nuclear processes. It is able to establish interrelationships that are supported by other studies such as nucleosome density [8, 12] and transcription rate [8]. It is most surprising, however, that we find the strongest correlation with transcription unit (TU) length, which is a new finding for budding yeast to our knowledge. Interestingly, our model allows also an alternative understanding of the data in which repair positions grow as patterns in a population. Although the analysis has been demonstrated for DNA repair, it can be applied to any process that can be modelled by an irreversible binary state transition. The source code is available on GitHub: https://github.com/leoTiez/jmak [25].

Results

Modelling DNA repair

In a single cell, CPD damage describes the mispairing of two adjacent pyrimidine nucleobases. Instead of establishing hydrogen bonds to the opposite strand, they cause two consecutive nucleotides to bind to each other. Consequently, there can be maximally one lesion per position. This results in a zero-one (i.e. *damaged-repaired*) state space per position and per cell. During ongoing repair, lesions are removed, and positions change subsequently their state to *repaired*. It can be assumed that this process is stochastic and involves to some extent unpredictable noise. If we could repeatedly and independently measure the repair times for a single position in a single cell, we could distribute the measurements over a timeline (Fig 1B). This type of data can be investigated by a Poisson point process, which allows the derivation of a predictive function that expresses the probability of repair over time. If we would temporally discretise the data over larger bins, all repair time points within a bin were aggregated together.

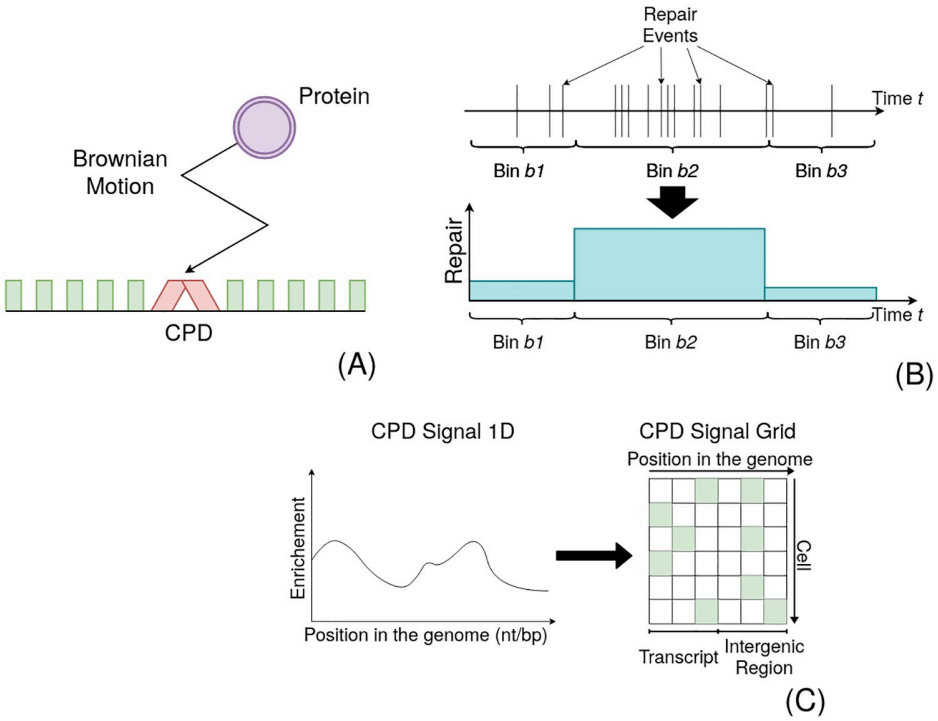


Fig 1. Schematic representation of the Process. (A) The repair proteins' search process (purple with arrow) to the lesion (red) can be understood as a Brownian motion. Repair happens after association with some delay, which we assume to be another random variable (waiting time). (B) If we could repeatedly measure the repair times for one position in a single cell individually, we could distribute the measurements over a timeline. This can be analysed by a stochastic point process. Binning the timeline should correspond to the observable change of CPDs in a given window. (C) We presume that this stochastic repair process happens independently in each cell. Therefore, CPD-seq data can be understood as an accumulation of several experiments. This allows an understanding of the signal as a two-dimensional grid. Since there can be only one lesion per position per cell, we understand the amplitude of the signal as a surrogate for the number of cells that contain a lesion at this locus (marked in green).

https://doi.org/10.1371/journal.pcbi.1010488.g001

In the following, we assume that this is given by the change of the CPD-seq signal, as the amplitude decrease at an arbitrary position must explain the number of cells that have repaired their lesions. Consequently, the data represent the process over the entire cell population. We conjecture that the dynamics are independent between cells. The CPD-seq data can be therefore alternatively interpreted as a two-dimensional grid: one axis representing the cells and the other the nucleotide positions (Fig 1C). We understand NGS signals not as representing an average cell but the mutual effect of multiple independent cells.

It is clear that the number of cells with a lesion at a given position is discrete. Therefore, the number of repair events in a given time window can be studied with a Poisson process. It expresses the probability for a number of events $N(t) = n$ in a given time t and takes the form $P(N(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda(t)t}$, where $\lambda(t)$ denotes the rate of events as a function of time. It is commonly assumed that CPD repair is irreversible if the irradiation source is removed. It follows that the probability that repair has happened (p_r) at time t can be expressed by the cumulative distribution function for the Poisson process, namely

$$p_r = 1 - e^{-\int \lambda(t) dt}. \quad (1)$$

To find a functional representation of $\lambda(t)$, we conjecture that repair proteins move through random Brownian motions (diffusion) to the repair sites, subsequently associate to the DNA and remove the lesion. The entirety of this mechanism can be understood as a mixture of two random processes: diffusion and waiting/repair time (Fig 1A). We surmise that the waiting time is determined by the accessibility to the lesion. The mutual effect of repair proteins removing DNA lesions becomes observable through the decrease of the CPD-seq signal. We thus assume that this change ΔC during time Δt is proportional to the searched volume by these proteins $D_{\hat{m}} t^{\hat{m}}$ (which is related to the mean squared displacement) and the average of the expected repair time $\hat{\beta}$

$$\lambda(\Delta t) = \Delta C \propto b\Delta t^{\hat{m}} + O(\Delta t^{\hat{m}}). \tag{2}$$

b denotes a scaling factor that accounts for the diffusion constant $D_{\hat{m}}$ and $\hat{\beta}$. $t^{\hat{m}}$ with the the anomalous coefficient $\hat{m}$ is the dominating term with the highest order. If $\hat{m} < 1$, the process is called subdiffusive; if $\hat{m} > 1$, the movement exhibits a superdiffusive behaviour. It is clear that the integral over Eq 2 also follows a power-law, i.e. $\int \lambda(t)dt \propto bt^m + O(t^m)$. Substituting in Eq 1 results in $p_r = 1 - e^{-bt^m}$. When setting $\sqrt[m]{b} = \beta = 1/\tau$ and assuming that only a fraction $\theta \in [0, 1]$ of cells have the ability to repair their lesions in a given time, we obtain

$$f(t) = \left(1 - \exp\left[-\left(\frac{t}{\tau}\right)^m\right]\right)\theta. \tag{3}$$

τ is the characteristic time until repair can be observed. An equation with a similar form to describe the phase transition in solids was derived independently by [20, 21], and [22–24] (KJMA model). As Eq 3 can be converted to a linear regression problem (Eq 5), it can be straightforwardly applied to find the necessary parameters. More interestingly, the KJMA model allows an alternative understanding of the data and the process, which is explained in Discussion.

Applying the model to the data

We formulated three expectations in order to prove the model validity. Firstly, we required that the estimated repair dynamics need to be in line with independently probed data; secondly, we thought it to be indispensable to recover NER-specific features that were not implicitly incorporated into the model; and lastly, Eq 3 needs to make verifiable predictions about other factors that influence repair. CPD data for the parameter estimation were taken from [8] (0, 20, 60, and 120 minutes after irradiation) and divided into different segments where $\lambda(t)$ was assumed to be spatially constant. We distinguish between TCR regions, which are the TS of genes that presumably exhibit TCR; the NTS of TCR regions; and non-TCR areas, which are composed of transcripts where the effect of TCR is not evident and intergenic regions (see S1 Appendix as well as S1 and S2 Figs). Moreover, TS and NTS of TCR regions were equally divided into start, centre, and end. Subsequently, CPD data was converted to represent repair using Eq 7 (in the following also called repair data). An example of the predicted repair dynamics is given in S3 Fig. We also compared the presented results with the analysis of a more traditional segmentation into TS and NTS of all genes as well as intergenic regions (S3 Appendix).

XR sequencing provides a snapshot of currently ongoing repair in the cell culture, and it therefore represents an independent angle on CPD removal. It should correspond to the derivative of Eq 3 (given in Eq 6). XR-seq signals were taken from [9] (5, 20, and 60 minutes after irradiation) and segmented as for the CPD-seq data. We assumed that a surrogate for ongoing

repair can be additionally derived from the CPD-seq data themselves by calculating the damage decrease per time (S4 Appendix Eq 4). This was used as a baseline value for the correlation between model and data. As we assumed a non-linear interrelationship, we used the distance correlation (DC) as a correlation measurement (S4 Appendix). Strikingly, the predicted repair rates correlate clearly better (DC = 0.441) than the actual data (DC = 0.209) (compare Fig 2A with 2B; see S2 Table). Moreover, the model predictions align fairly well with the XR data (exemplified in S4 Fig). We hence surmise that Eq 3 is in agreement with independently probed data.

Despite the fact that we model time-dependent repair, we nevertheless do not incorporate two (potentially competing) repair mechanisms, i.e. TCR and GGR. It is commonly presumed that CPD removal through TCR is quicker than by GGR [4, 26]. Moreover, as GGR acts genome-wide, they are spatially non-exclusive for genes. Indeed, we can recover the cumulative effect when averaging the repair evolution for a group of segments, e.g. the start of TCR regions. The beginning of TCR areas is almost solely repaired by a single mechanism (Fig 3A). The contribution of this pathway is decreasing as a function of distance from the transcription start site (TSS). Instead, a later acting mechanism becomes increasingly observable (Fig 3B and 3C). Lastly, lesions in non-TCR regions are only detected by the late-acting process (Fig 3D). This is even the case despite the fact that non-TCR areas also include transcripts. We deduce that these two distinct pathways show the effect of TCR and GGR along the gene. We were therefore able to separate the effect of two distinct NER processes without involving any particular mechanism. Parameter distributions (i.e. m, τ, and θ) that create these repair kinetics are exemplified in Fig 4. Surprisingly, the NTS possesses different dynamics whilst not exhibiting any difference between start, centre, and end (S5 Fig). The average repair fraction is much lower than for all other areas ($\theta \approx 0.6$ instead of ≈ 0.8). Moreover, we observe a subtle early increase of the derivative, indicating a larger presence of early repair in comparison to non-TCR regions. It is difficult to analyse this trend without additional experiments. It could simply be the impact of neighbouring overlapping regions. However, these results could equally point to different repair dynamics on the NTS.

Lastly, we extended the analysis to make predictions about influencing factors for CPD repair *in vivo*. Previous studies published various measurements of different nuclear properties that could possibly interact with lesion removal dynamics. To assess the predictive power of our model, we opted to analyse a link to transcription rate [8, 9] and nucleosome density as representing chromatin structure [8, 12, 27]. We also investigated a link to TU length and the relative distance to centromeres and telomeres as possible unreported affecting parameters. We used the NET-seq signal produced by [28] as a surrogate for transcription rate without UV irradiation. The TU length was measured by [29]. Nucleosome data after UV treatment were acquired by [12]. We excluded regions outside a reasonable parameter range from the subsequent analysis (see Methods and materials and S2 Appendix; the number of removed regions is given in S1 Table). Correlations with the model parameters were verified with a significance test, during which we compared a binary classification model with the performance of a random model (see Methods and materials and S5 Appendix for more information; the working of the classifier is explained in S6 Fig). Interrelationships to other sequencing data are elaborated and discussed in S6 Appendix. All data distributions are given in S7–S12 Figs. The results differed considerably depending on the genomic area and context (S3 Table). TCR has been repeatedly investigated with respect to transcription rate, and it is surmised to be positively correlated. Hence, higher transcription yields quicker repair, whereas low transcription results in slower lesion removal [8, 9]. Our model confirms these findings (Fig 4A and 4B). Non-TCR segments and the start of the TS seem to be starkly influenced by the nucleosome density, whereas all other areas do not show a strong correlation (Fig 4C and 4D). The clearest results,

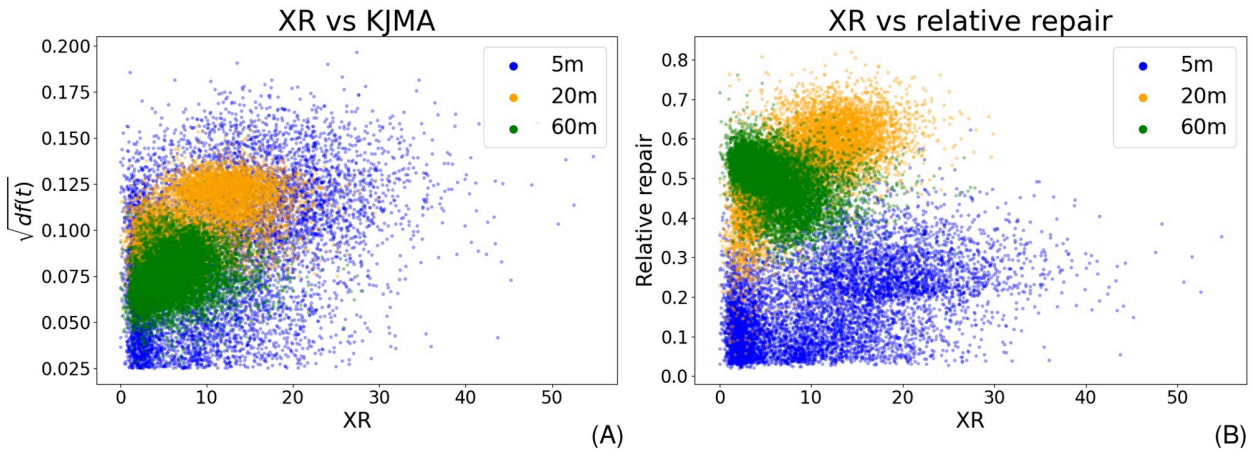


Fig 2. Comparing XR-seq data with model predictions. The values at 5 minutes are given in blue, 20 minutes are coloured yellow, and 60 minutes are green. The plots show that the distance correlation between prediction and XR-seq data is even higher than for the CPD-seq data. (A) Predicted repair rates with respect to the XR-seq data exhibit a considerably strong correlation (DC = 0.441). Predictions are given as the square root of the model prediction. This reduces the effect of increasing variance with larger derivatives. (B) The repair rates derive from the data as a function of XR-seq values shows a weaker correlation (DC = 0.209).

https://doi.org/10.1371/journal.pcbi.1010488.g002

however, were obtained by the TU length. Both, TS (Fig 4E) and NTS (Fig 4F) are clearly influenced. The TU length is therefore likely to be contributing to the lesion removal dynamics. This is an unreported finding for budding yeast to our knowledge. The developed quantitative framework has hence the potential to identify established as well as new interrelationships. Importantly, a correlation with the distance to telomeres and centromeres (Eq 9) did not indicate a significant link. This shows that the applied method is selective for certain correlations (S12 Fig).

Discussion

The few time points of NGS data sets require a temporal model to recover missing information between data samples. In this work, we developed a computational approach to describe the DNA repair kinetics on a population scale. We recover region-specific properties based on the genome-wide distribution of DNA damages. We assume a mixture of two stochastic processes (diffusion and lesion accessibility) that collectively explain the change of CPD data over time. Parameters of the derived equation can be estimated with the KJMA model, that is conveniently converted to a linear regression problem. This allowed the analysis of the temporal process as a whole rather than only comparing single time points. Importantly, it points out that the signal changes non-linearly over time. This is expected from a biological point of view, as TCR and GGR are commonly seen as acting within different time scales. However, it should be emphasised that it has not been incorporated in the analysis of temporal changes in sequencing data to our knowledge. The model therefore accounts specifically for dynamics on a population scale. Moreover, the derivative (Eq 6) provides key information about active ongoing repair. It thus permits linking CPD-seq data—showing the DNA damage distribution over the genome—and XR-seq data of excised DNA fragments generated by repair. This provides strong support for the validity of the model. Even though Eq 3 represents repair only with one mechanism per region, the combined effect of TCR shortly after UV treatment and GGR at later time points can be recovered when considering the average over several areas. The model can be readily used to uncover interrelationships between repair parameters and

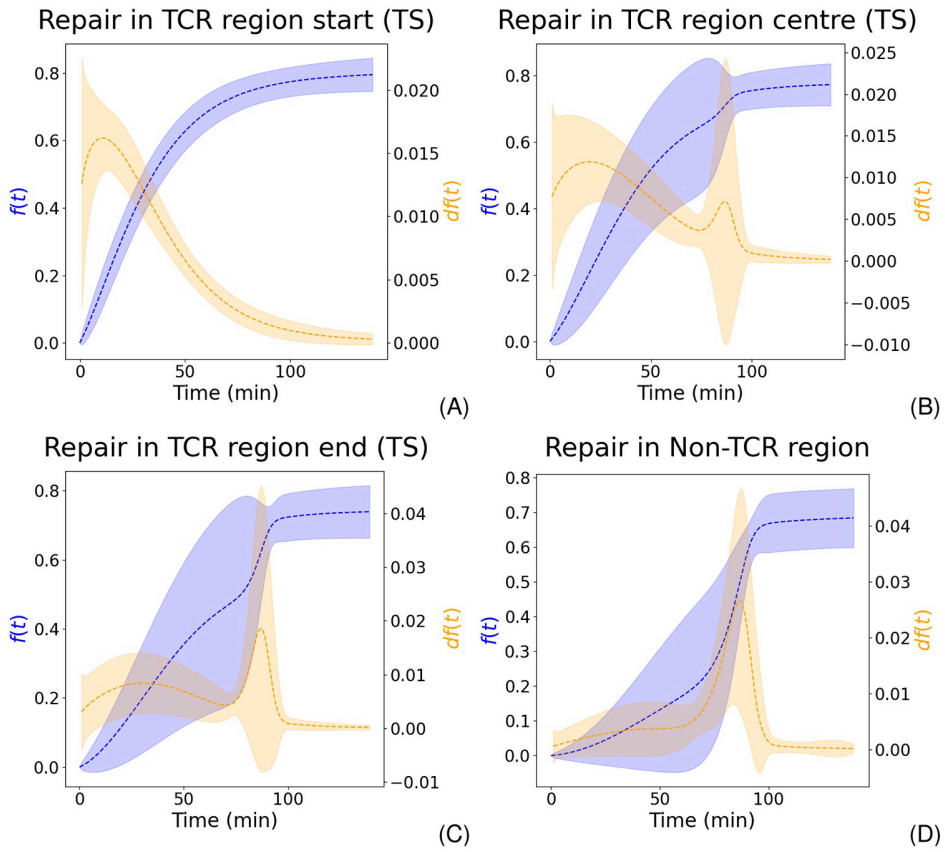


Fig 3. Collective behaviour of genomic regions can recover mutual effect of TCR and GGR. Dashed lines give the mean whereas the shaded areas show the standard deviation. Blue and orange represent the repair fraction and the repair rate (derivative of the repair fraction), respectively. (A) The start of TCR regions is repaired early after irradiation, demonstrating the effect of TCR. (B) At the centre of TCR areas, we can observe the mutual effect of TCR (first peak) and GGR (second peak). (C) GGR's contribution increases whilst the impact of TCR becomes less important towards the end of the gene. (D) Non-TCR regions are solely repaired by GGR. Therefore, repair is expected at later time points during the process.

https://doi.org/10.1371/journal.pcbi.1010488.g003

genomic contexts. Our outcomes are consistent with known influencing factors such as transcription rate and nucleosome density. Remarkably, the clearest link was established between the repair dynamics within genes and their length. To our knowledge, this is an unreported finding for budding yeast. In the following sections, we discuss the relevance of our approach and results within the context of previous publications.

Applying the CPD repair model

Several studies proposed temporal models for UV-induced lesion repair on different levels of detail. [30] represented NER kinetics in human cells using a Markov-Chain Monte-Carlo approach. It explains the removal of 6–4 photoproducts on a single-cell scale through the random and reversible assembly of repair complexes. A similar model was proposed by [31]. Interestingly, though, they derive very different conclusions, as they suggested that random or pre-assembly of repair proteins is unfavourable. Despite a great level of detail of both models, they are incapable to make region specific predictions. Moreover, as both models are based on microscopy data, they do not explain temporal changes in genome-wide sequencing data on a

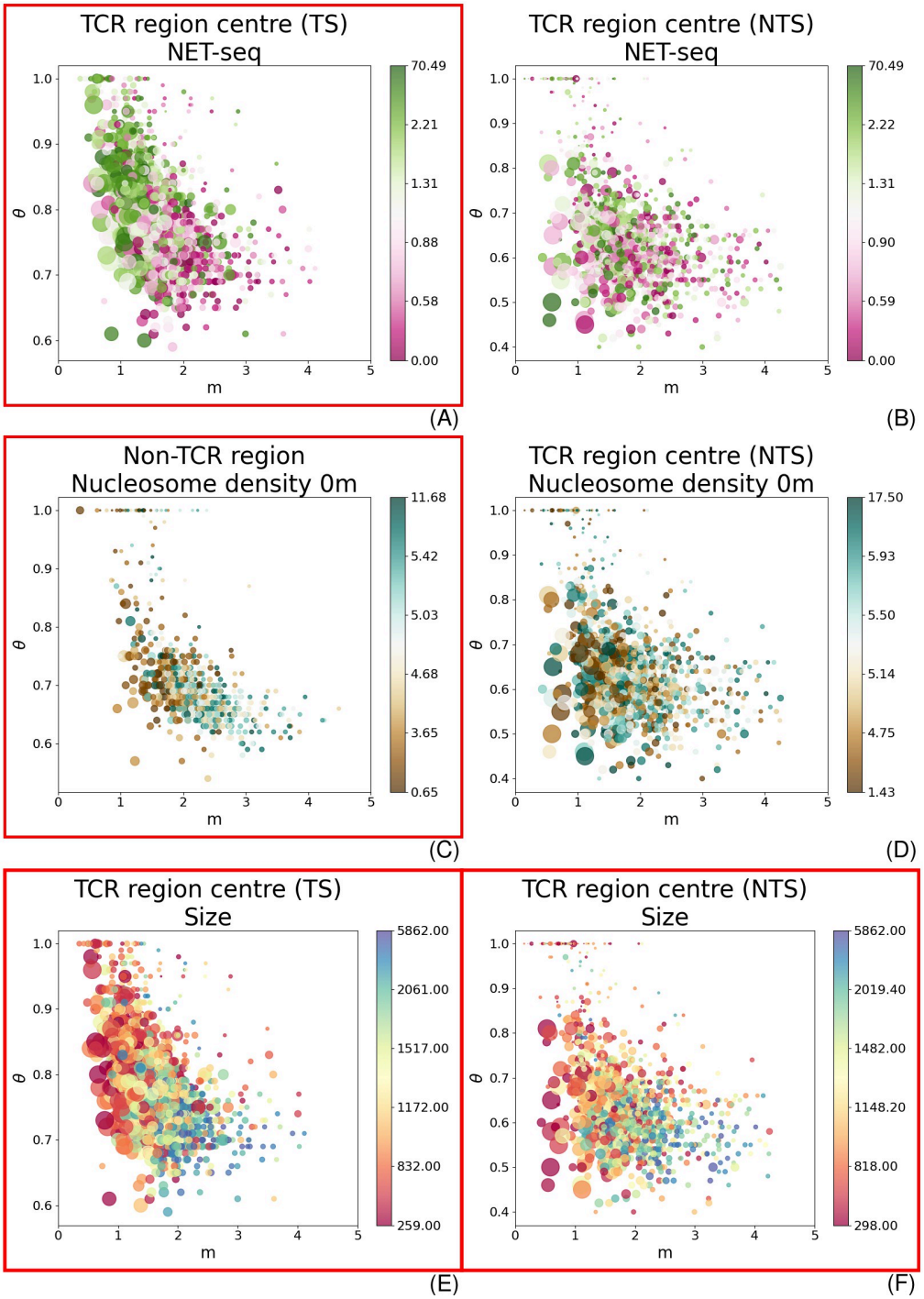


Fig 4. The parameter distribution is coloured with respect to different genomic properties. The x and y-axis give the values of m and θ, respectively. The size of the circles show $1/\tau$: the larger the circle, the shorter the characteristic time. Significant interrelationships are marked with a red frame. (A) and (B) are coloured with respect to NET-seq data (pink/low to green/high) for the centre of TS and NTS; (C) and (D) indicate the nucleosome density (turquoise/low to brown/high) for non-TCR regions and the NTS centre; (E) and (F) show the distribution with respect to the TU length (red/small to blue/big) for the centre of TS and NTS.

https://doi.org/10.1371/journal.pcbi.1010488.g004

population scale. A Monte Carlo approach to explain the damage distribution and subsequent repair induced by ionised irradiation in a single cell was proposed by [32]. It also incorporates the collective effect of NER and base excision repair (BER), therefore accounting for potentially competing mechanisms. It should be stressed that the lesion type is considerably different. Again, the predictions are location unspecific. [33] provided a different angle by presenting a protein-protein interaction landscape of NER components in yeast. Predictions about the repair efficiency in different regions were not established. To our knowledge, our model is the first that accounts for region specific changes in population-based data.

As there are only three time points, a data-driven machine learning model is prone to overfitting. In order to find a reasonable representation of the data, we incorporate explicit suppositions to derive Eq 3. To be precise, we presume that repair times follow a Poisson point process, and the non-constant rate $\lambda(t)$ can be described by a mixture of protein diffusion and repair times. This restricts the trajectory to an S-shape. Our understanding of the process makes two independence assumptions. Firstly, there is non-interfering DNA repair between cells; and secondly, protein diffusion is independent of lesion accessibility or repair time.

It is important that the points in a Poisson process are (sufficiently) independent from each other [34]. The repair times that are distributed along the timeline (e.g. as in Fig 1B) should hence symbolise values of independent stochastic variables. We find the presumption of non-interfering repair between cells trivial. *Saccharomyces cerivisiae* are single-cell organisms and should thus react independently to DNA damage. Moreover, yeast cultures were grown in rich medium after UV treatment, precluding any limitations for growth [8]. We conjecture the independence presumption to be reasonable.

The Poisson process is governed by the rate parameter $\lambda(t)$. It is presumed that nearby genomic positions posses similar rates. Therefore, $\lambda(t)$ should have a slow spatial variation, and a segmentation of the CPD-seq data into similar behaving areas is possible. A similar binning approach was applied by [8]. The simplest model sets λ to a constant and therefore independent of time. In such a setup, we would expect to observe the largest signal change right after irradiation which subsequently slows down (S13A Fig). This could indeed resemble the beginning of the TS of TCR regions (Fig 3A). However, the majority of areas exhibits the strongest repair rates between 20 and 60 minutes. We conclude that a non-constant $\lambda(t)$ provides a broader applicability for explaining CPD repair.

It remains to find a description of the function $\lambda(t)$. Repair happens through the collective working of several proteins, which need to move through the environment and find their target. Nuclear diffusion has been studied and modelled in detail in different contexts such as chromatin [15–17] and protein movements [13, 35], including the repair protein Rad4 [19]. It is clear that more DNA-protein interactions are possible during longer time windows, since proteins have more time to travel longer distances to reach their target. The distance is denoted by the random variable R which has the expected squared displacement $< R^2 >= D_{\hat{m}} t^{\hat{m}}$ with diffusion constant $D_{\hat{m}}$ [14, 16, 17, 19, 35]. Consequently, the searched volume is $< R^3 >= k'(D_{\hat{m}} t^{\hat{m}})^{\frac{3}{2}} = D_{\hat{m}} t^{\hat{m}}$, where k' is a scaling constant. We couple the Brownian motion through space with an independent random variable X which symbolises repair time or accessibility to the lesion. We define $< X >= \hat{b}$. As R and X are independent, we can write $< XR^3 >=< X >< R^3 >= bt^{\hat{m}}$, with $b = \hat{b} D_{\hat{m}}$. Substituting in Eq 1 results in Eq 3. We find it remarkable that β (representing a mixture of diffusion constant and expected waiting time) tends to be inversely proportional to m which is linked to the abnormality coefficient. Despite quick diffusion (m being large), we observe slow repair rates (β being low). We interpret this phenomenon as the repair time/accessibility X dominating the process, making diffusion negligible. This argument can be equally applied when diffusion is seemingly very slow.

The simplest representation of an irreversible transition between two states is given by an S-shaped function which contains at most one inflection point. It is important to emphasise that this function is not necessarily inversely symmetric around this point, meaning that the left side can be differently shaped than the right side. These requirements are fulfilled by the KJMA model, which is therefore a sensible choice (see S7 Appendix and S13 Fig).

Analysing genomic properties which influence repair kinetics

TCR has been identified as a rapid repair pathway on the TS. Intergenic regions and the NTS exhibit significantly slower lesion removal, which was demonstrated on the genomic scale in yeast and human cells [8, 9, 27, 36, 37]. It remains an unsolved quest to find an interrelationship between TCR efficiency and transcription rate. Whilst the two parameters are indeed assumed to be correlated [8, 9], some studies point out that TCR repairs CPDs efficiently at nearly all genes including those with a low transcription [27]. An in-depth analysis is still missing, and there is no clear consensus on how transcription rate is affecting repair. In this work, we compared the model predictions to gene expression. Our analysis clearly shows a significant correlation on the TS and is therefore supporting the common assumption (S7 Fig).

As repair proteins need to recognise and repair lesions on the DNA, it is conjectured that chromatin organisation can significantly modulate the efficiency of CPD repair [8, 12, 27]. However, previous studies were mostly scrutinising the positioning of damage at nucleosomes. CPD removal was shown to be less efficient at the dyad of strongly positioned nucleosomes in yeast [8]. Moreover, GGR on the NTS was asymmetrically inhibited in yeast and human cells with respect to the position within the nucleosome [27]. Even though nucleosome occupancy after UV treatment was already previously probed, the potential relationship of these data with CPD repair was not directly addressed [12]. Our results demonstrate a significant correlation between repair and nucleosome density in non-TCR regions (Fig 4C). We also discovered a clear influence on the beginning of TCR areas (S8 Fig).

Unexpectedly, our outcomes show a strong correlation between TU length and repair. Differences in transcription shutdown and restart after UV treatment relative to gene size were previously reported for human cells [38]. Both transcription regulation and efficient repair are necessary to orchestrate an effective cellular response to UV light. The restart of transcription to pre-irradiation levels is an important step at the final stages. However, a direct evaluation of lesion removal with respect to gene size was not performed. To our knowledge, this is a new finding for CPD repair in yeast. Due to our data pre-processing (Eq 7), we can rule out that the result derives only from the fact that larger areas have a greater potential to include more damage. This is true due to two reasons. Firstly, we normalised the CPD value in each bin (e.g. beginning of the TS) by the number of pyrimidine dimers in the sequence as described in [8]. Secondly, and more importantly, we want to point out that the quotient in Eq 7 lets any length dependence and normalisation of the binned data vanish. Therefore, the values become automatically comparable due to the design of Eq 7. It should be mentioned though that the regions of interest can become rather small when segmenting the gene into subareas. Influence or noise from neighbouring areas cannot be excluded. However, due to the fact that the same result can be obtained with a different segmentation (S3 Appendix and S9 Fig), we presume that it represents a genuine feature of the CPD removal mechanisms in yeast cells.

Lastly, we investigated a potential link to the distance relative to the centromere and telomere depending on which was closer to the region of interest (S12 Fig). A link to repair has not been proposed to our knowledge, which made it an interesting property to produce verifiable model predictions.

In conclusion, our work opens interesting perspectives for future research on DNA repair mechanisms and influencing genomic factors. New experimental data with increased temporal resolution will help to refine the model and analysis. The approach can be similarly used for other organisms including human cells. Moreover, it can be readily applied to sequencing data of any nuclear process that can be represented as a two-state system, and it is not restricted to repair.

Introducing the repair space—An alternative understanding of the data

We have discussed the model in detail with respect to a stochastic point process. We want to provide an additional interpretation that is motivated by the physical implications of the KJMA model. Next to assuming independence between cells, we conjecture in the following also independent repair within each cell (S8 Appendix). Moreover, we presume that CPD data were converted by Eq 7.

Considering the two-dimensional grid in Fig 1C, the independence assumptions above permit us to re-order repair positions to patterns. Nevertheless, we restrict the re-grouping to stay within areas of interest which are assumed to behave homogeneously. The growth of patterns in the virtual repair space reminds strongly of the phase transition in solids which is described by the KJMA model.

The creation of these repair patterns can be described by the nucleation rate n (which we link conceptually to the expected waiting/repair time) and a growth speed G (which is in this analogy linked to the diffusion process). In the following, we assume G to be constant in all directions. The transformed volume within Δt starting from a single nucleation site is therefore

$$v(\Delta t) = \sigma(G\Delta t)^{m-1}, \tag{4}$$

where σ denotes a parameter that describes the shape of the expanding pattern (which would become part of β). Interestingly, the parameters obtain a slightly different meaning from this point of view. m is the Avrami exponent and characterises the geometry of the area covered by repaired positions after their aggregation. For example, if $m - 1 = 2$, the area corresponds to regular disks in a two-dimensional space. Irregular forms can be expressed with non-integer values [39]. Nevertheless, a direct comparison of a physical shape with a virtual pattern might be difficult to imagine. We therefore advocate another interpretation. m can be understood to express time dependence of the repair process (compare with S14 Fig). A similar notion has been also proposed in the physical context [40]. We believe that such an understanding could possibly permit the inclusion of independent results from the realm of physics.

Methods and materials

Parameter estimation and derivative

Eq 3 explains CPD repair as an S-shaped transformation over time. It should be noted that the process has a defined starting point at $t = 0$. By applying the natural logarithm on both sides twice, we obtain

$$\ln \frac{1}{1 - f(t)/\theta} = m \ln t + m \ln 1/\tau. \tag{5}$$

Note that the expression is now continuous over $ln\ t$. Given the data points for repair and by assuming a value for θ, the parameters m and $1/\tau$ can be found by solving the linear regression problem defined in Eq 5 (compare with bottom plots in S14 Fig). θ was determined

through a systematic parameter search. We started with 0.5 as minimal value for transcribed/TCR regions and 0.4 for all other. We increased it thereafter by $\Delta\theta = 0.01$, until we reached $\theta = 1.0$. We chose the θ-value that can best describe the repair data together with the corresponding parameters m and $1/\tau$. This is determined by maximising the adjusted R^2. It represents the variance in the data that can be explained through the model and can be interpreted as goodness of fit. The derivative of Eq 3 is given by

$$df(t) = \frac{m\theta t^{m-1}}{\tau^m}\left(1 - \frac{1}{\theta}f(t)\right)dt. \tag{6}$$

Data processing

All experimental data that were analysed in this study comes from public databases (see overview in Table 1). CPD-seq data was taken from [8]. It contains two time courses with samples taken at $t_1 \in \{0, 60\}$min and $t_2 \in \{0, 20, 120\}$min, respectively. The location of transcribed areas was taken from [29]. Data signals were partitioned into different segments, i.e. the TS and NTS of TCR regions as well as non-TCR areas. For the latter, we combined both strands to one group. Consequently, the linear regression problem in non-TCR regions was required to find the best representation for both strands.

CPD-seq fragments were normalised by the number of available pyrimidine dimers, as explained in the supplementary material of [27]. The damage distribution was subsequently transformed into repair in area a through

$$R_a(t) = \frac{\sum_i^N CPD_{a_i}(0) - \sum_i^N CPD_{a_i}(t)}{\sum_i^N CPD_{a_i}(0)}, \tag{7}$$

where N denotes the size of a, $CPD_{a_i}(t)$ is the normalised CPD signal at time t and locus i in area a, and $t \in \{20, 60, 120\}$min. a is any of the previously described regions (e.g. the start of the NTS). We additionally take it for granted that no new CPD lesions can be induced during repair. Hence, data points were enforced to be greater than or equal to zero and monotonously increasing as a function of time. The rectification is defined by

$$\begin{aligned} R_a(20) &= \max\{R_a(20), 0\} \\ R_a(t_i) &= \max\{R_a(t_i), R_a(t_{i-1})\}, \end{aligned} \tag{8}$$

where $t_i \in \mathbf{t} = (20, 60, 120)$.

All other sequencing data that were used for the correlation analysis were averaged over the size of the area of interest. Start, centre, and end of TS and NTS were linked to the same value to smooth out the potential influence of noise. For example, all subregions of a TS were associated to the same transcription rate. Moreover, both strands were compared to the same data,

Table 1. Overview over the data sets that were used in this study.

Property	Strain	Data type	UV Dose	Reference
CPD	BY4741 (WT)	CPD-seq	125 J/m^2	[8]
CPD repair	Y452 (WT w.r.t. repair)	XR-seq	120 J/m^2	[9]
Abf1	BY4742 (WT)	ChIP-seq	100 J/m^2 (0min)	[12]
H2A.Z	BY4742 (WT)	ChIP-seq	100 J/m^2 (0min)	[12]
Nucleosome distr.	BY4742 (WT)	MNase-seq	100 J/m^2 (0min)	[12]
Transcription rate	YSC001	NET-seq	-	[28]

https://doi.org/10.1371/journal.pcbi.1010488.t001

e.g. the TS and the NTS were related to the same nucleosome density. We noticed that the NET-seq signal amplitude decreases as a function of distance from the TSS (S15 Fig). This could possibly induce a TU length-specific bias that is not removed by taking the average over the TU length. We could verify, however, that the NET-seq data strongly correlates with independently probed Pol2 ChIP-seq data [41] (S16 and S17 Figs). We therefore assume that it reasonably represents transcription rate, whilst allowing a direct comparison to the results obtained by [9] (S9 Appendix).

With the exception of nucleosome density, all biological data values possess a biased distribution. They strongly peak around a low value but contain large positive tails. To remove a potential bias introduced by outliers, we limited our analysis to the lower 95th percentile. As this procedure was applied to all data (except nucleosome density), we did not introduce a bias towards a certain model. Rather, we improved comparability. The only exception is the MNase-seq signal, as it is approximately normally distributed. We consider that trimming could introduce a bias rather than removing one.

The relative distance to centromeres (c') or telomeres (t') was measured as follows. Denoting the gene position by x, we can define

$$d_{mere} = \frac{2\ \min\{|x - c'|, |x - t'|\}}{|c' - t'|}. \tag{9}$$

We divide only by half of the length since the maximal distance ($d_{mere} = 1$) to both centromere and telomere should be the middle between them two.

Correlating repair dynamics to genomic contexts

Areas with parameter values outside a reasonable range were excluded from the subsequent analysis. We restricted $m \in [0.5, 5]$ and $\tau \in [20, 200]$. θ was constrained through the parameter search. Motivation and consequences are discussed in S2 Appendix. The repair parameters were investigated in context of other biological data. We opted for a nonparametric classification k-Nearest Neighbour (kNN) approach. We grouped biological data into high (class $c = 1$) and low values ($c = 0$), such that both classes contained the same number of samples to remove the difficulty posed by the biased data distribution. To train the machine learning model, the input values $\mathbf{x} = (m, \beta, \theta)$ where normalised so that every dimension was normally distributed with zero-mean and a standard deviation of one. We compared the results for several kNN models with $k \in \{5, 10, 20, 50, 100\}$ to remove any model specific bias. A trained model compares an unknown input $\hat{\mathbf{x}}$ to the k closest values of a known data set $\{\mathbf{X}, \mathbf{c}\}$ to predict class $\hat{c}$. We opted for the Euclidean distance as similarity measurement. Here, the i-th row of $\mathbf{X}$ is $\mathbf{x_i} = (m_i, \beta_i, \theta_i)$, and c_i is the associated class in $\mathbf{c}$. $\hat{c}$ is determined by a majority vote. For example, if more than 50% of the k neighbouring values are classified as group $c = 1$, then $\hat{c}$ is predicted to be group 1 as well. kNN is categorised as nonparametric model which permits the comparison of different results. The performance was measured through calculating the prediction error

$$E = \frac{\#\text{Incorrectly classified samples}}{\#\text{ All samples}}. \tag{10}$$

This was compared to a random baseline model, for which classes were randomly shuffled to a given parameter triple (m, β, θ) during training. Data $\{\mathbf{X}, \mathbf{c}\}$ were arbitrarily partitioned into learning and testing data sets. Every experiment was independently repeated 100 times to reduce the effect of any potential bias. We consider an interrelationship to be important if the prediction error of the true function is significantly lower than the error of the random model ($p < 0.00001\%$ of a one-sided t-test). Moreover, we require that 90% of the prediction errors

are below $E < 0.5$, which is the expected outcome of an unbiased coin-flipping experiment. This significance must be found in three out of the five evaluated k to indicate an interrelationship.

Supporting information

S1 Appendix. Determining TCR Regions.
(PDF)

S2 Appendix. Discussing the Effect of Data Transformation and Selection.
(PDF)

S3 Appendix. Dividing the Data into Genes versus Intergenic Regions.
(PDF)

S4 Appendix. Explaining the Correlation Between XR-seq Data and Repair Rate.
(PDF)

S5 Appendix. Discussing the *k*NN Approach.
(PDF)

S6 Appendix. Analysing Repair Kinetics in Context of Abf1 and H2A.Z Distribution.
(PDF)

S7 Appendix. Comparing the KJMA Model With Other Approximations.
(PDF)

S8 Appendix. Discussing the Model in Context of the Physical KJMA Model.
(PDF)

S9 Appendix. Investigating a Link Between Transcription Rate and TU length.
(PDF)

S1 Table. The number of models per region, before and after applying the requirements for parameter ranges. IGR abbreviates intergenic regions.
(PDF)

S2 Table. The DC between XR-seq and repair predictions / data for different experimental configurations.
(PDF)

S3 Table. Number of non-random interrelationships between model parameters and sequencing data over *k*. The table gives the number of *k*NN models that could find a correlation between model parameters and genomic context. $k \in \{5, 10, 20, 50, 100\}$. We defined a link to be significant if at least three out of five k find a non-random interrelationship.—means that data was not used in the given configuration. NET denotes NET-seq data, ND is nucleosome density, and meres give the relative distance to centromeres or telomeres. Suffixes S, C, and E denote start, centre and end of an area. NTCR are non-TCR areas. IGR are intergenic/non-transcribed regions.
(PDF)

S1 Fig. Scheme of the segmentation setup. The circles represent the number of cells with ongoing repair in the region. The arrows indicate the region and direction of transcription. The results in the paper follow the *TCR* setup. Here, only the first gene is considered as TCR area which shows more efficient repair than intergenic regions within the first 20 minutes after UV irradiation. All other parts are labelled as non-TCR region. Therefore, it spans from the

end of the first gene to the end of the second. The *gene* configuration (S3 Appendix) partitions the genome into the traditional notion of transcribed and intergenic regions. The TU positions were determined by [29].
(TIF)

S2 Fig. Relative repair distribution over genomic areas. Relative repair in non-transcribed regions is chiefly lower than 20% within the first 20 minutes (88.95%). Genic areas with stronger repair dynamics are thus likely supported by TCR. For all other transcripts, we cannot exclude the possibility that they are exclusively repaired by GGR.
(TIF)

S3 Fig. Example for model predictions. Data points are given by solid dots. The blue dashed line represents the repair fraction predicted by the model (left axis). The orange dashed line shows the derivative (right axis). (A) The *SNF6* gene can be well approximated. (B) However, *GEM1* exhibits no repair within the first 20 minutes, which results in a switch-like behaviour. (C) This is better understood when showing the data points after transformation according to Eq 5. A linear regression is difficult since they do not align.
(TIF)

S4 Fig. Example for model prediction and XR-seq data over time. XR-seq data (points) and the predicted repair rate (dashed lines) are exemplified for genes *BDH1* (orange) and *BDH2* (blue). When re-scaling XR data and repair rate prediction between 0 and 1, both follow clearly similar trends. *BDH2* has its largest XR-seq value at 20 minutes post-irradiation, whereas *BDH1* shows biggest repair rates after 5 minutes. This is indeed captured by the model.
(TIF)

S5 Fig. The evolution of repair along the NTS. The average repair evolution for the NTS (blue dashed line) shows a much lower repair fraction ($\approx$0.6) than the other areas. Moreover, the repair rate (orange dashed line) indicates repair at early time points. This could be caused by possible overlapping transcripts or by antisense-transcription-coupled repair. Shaded areas show the standard deviation. The repair trajectory is the same for (A) the beginning, (B) the centre, and (C) the end of the NTS.
(TIF)

S6 Fig. Example of the learnt function between model parameters and genomic context. (A) The learnt parameter distribution and the associated class for the true model after applying a principle component transformation. The x and y-axis give the first and second principle component, respectively. Red represent large genes, whereas blue shows low values. (B) The error distribution for the predictions follows the expected outline given by the learnt function in (A). The blue and red circles give values that were classified as short but were actually large and vice versa, respectively. White points are correctly classified. The right bar shows the error distribution along the colour axis, i.e. over estimated, correctly classified, and underestimated values from top to bottom. The lower histogram shows the distribution of overall correctly and incorrectly classified values. (C, D) The learnt parameter map and the error distribution of the random model.
(TIF)

S7 Fig. Model parameters with respect to transcription rate. Our results for the transcription rate support the hypothesis that it influences repair on the TS. The x and y-axis give the values of m and θ, respectively. The size of the circles show $1/\tau$: the larger the circle, the shorter the characteristic time. Significant interrelationships are marked with a red frame.
(TIF)

S8 Fig. Model parameters with respect to nucleosome density. Nucleosome density is seemingly influencing repair in non-transcribed/non-TCR regions as well as the beginning of the TCR TS and the TS in the *gene* setup. The x and y-axis give the values of m and θ, respectively. The size of the circles show $1/\tau$: the larger the circle, the shorter the characteristic time. Significant interrelationships are marked with a red frame.
(TIF)

S9 Fig. Model parameters with respect to size. The size is clearly influencing repair for both, TS and NTS in the *TCR* and *gene* configuration. The x and y-axis give the values of m and θ, respectively. The size of the circles show $1/\tau$: the larger the circle, the shorter the characteristic time. Significant interrelationships are marked with a red frame.
(TIF)

S10 Fig. Model parameters with respect to Abf1. The results for Abf1 are more ambiguous. Although we can find a significant correlation to non-TCR regions as expected in the *TCR* setup, the picture is less clear for the *gene* configuration. The x and y-axis give the values of m and θ, respectively. The size of the circles show $1/\tau$: the larger the circle, the shorter the characteristic time. Significant interrelationships are marked with a red frame.
(TIF)

S11 Fig. Model parameters with respect to H2A.Z. Similar to Abf1, the correlations with H2A.Z do not allow a straightforward interpretation. Whilst repair in all areas in the *TCR* configuration is seemingly linked to H2A.Z, this tends to be restricted to the TS and NTS in the *gene* setup. The x and y-axis give the values of m and θ, respectively. The size of the circles show $1/\tau$: the larger the circle, the shorter the characteristic time. Significant interrelationships are marked with a red frame.
(TIF)

S12 Fig. Model parameters with respect to cetromeres and telomeres. With the exception of the TS in the *gene* setup, the distance to telomeres or centromeres (shortened with *meres*) does not affect repair dynamics. The x and y-axis give the values of m and θ, respectively. The size of the circles show $1/\tau$: the larger the circle, the shorter the characteristic time. Significant interrelationships are marked with a red frame.
(TIF)

S13 Fig. Comparison of the KJMA model with other functional descriptions. (A) A homogeneous Poisson repair process with $\lambda(t) = c$ has the strongest change in the beginning which subsequently flattens out. In most investigated regions, such a behaviour is not observed. (B) We compared the performance of different models to describe the data, which is exemplified for gene *LDB16* (*YCL005W*). Black dots represent the repair data (converted CPD-seq data, see Eq 7), whereas the best fit of each model is given in dashed lines. (C) We applied the mean-squared error (MSE, S7 Appendix Eq 1) to compare the performance of the models with respect to the data. The KJMA model and the Hill equation perform undoubtedly better than simpler models like linear or logistic regression. Nevertheless, the Hill equation describes the data slightly yet significantly better. It should be emphasised though that the performance difference is marginal. Width of the shaded areas represents the number of genes that yielded the corresponding error, which is mirrored at the vertical line. The centre horizontal line with the corresponding numbers give the error median. The top and bottom horizontal lines show maximum and minimum, respectively. (D) Despite the fact that non-TCR regions are not expected to show an observable impact by TCR, the Hill equation indicates two mechanisms that act at different time points. Dashed lines give the mean whereas the shaded areas show the

standard deviation. Together with the fact that there is no straightforward interpretation of the Hill equation in context of repair evolution, we conclude that Eq 3 is a sensible choice.
(TIF)

S14 Fig. Example of the KJMA model. The KJMA model includes two governing parameters (as the original model does not involve θ), which are exemplified in (A) for m and (B) for τ. Eq 3 can be conviniently converted to a linear regression problem which is shown for the parameter settings of (A) in (C) and for the parameters of (B) in (D).
(TIF)

S15 Fig. Example of sequencing data representing transcription rate. The example of the NET-seq signal in comparison to the Pol2 ChIP-seq data probed by [41] shows that Pol2 exhibits a constant augmentation of the signal amplitude at transcribed regions, whereas NET-seq data decrease as a function of distance from the TSS. The Pol2 data is coloured in green, whereas NET-seq is given in blue (light blue represents the Watson, and dark blue is the Crick strand). The example is given for chromosome II around *CHS2* and *CHS3*.
(TIF)

S16 Fig. Correlation between size and transcription rate. The plots show the two-dimensional histogram distribution of TU length and different measurements of transcription rate. The number of genes per bin is given through the colour intensities and white numbers. The DC (S4 Appendix Eq 3) per measurement is given in the title. Size and transcription rate data was divided into 50 equally sized bins, which is given by the x and y axis. We use the 95th percentile to remove strong outliers. (A) The histogram distribution of NET-seq transcription with respect to size reveals that smaller genes tend to have higher transcription rates than larger genes. (B) This link is weakened when considering Pol2 ChIP-seq data.
(TIF)

S17 Fig. Correlation between NET-seq and Pol2 ChIP-seq data. The plot shows the two-dimensional histogram distribution of NET-seq and Pol2 data. The number of genes per bin is given through the colour intensities and white numbers. Size and transcription rate data was divided into 50 equally sized bins, which is given by the x and y axis. NET-seq data and Pol2 ChIP-seq signal are strongly related (DC = 0.75, S4 Appendix Eq 3). As we consider only two groups of genes with respect to transcription, i.e. genes with a low (blue) or high transcription rate (red), we can confirm that the majority of regions fall into the same category.
(TIF)

S18 Fig. Model predictions with respect to XR-seq data in the *gene* setup. (A) The model predictions in the *gene* configuration are less correlated with the XR-seq data than in the *TCR* setup (DC = 0.241, S4 Appendix Eq 3). (B) When correlating the relative repair rate and the XR-seq data in the *gene* setup, the DC is as low as for the model predictions (DC = 0.231). Therefore, we assume that the weak linkage is due to the data segmentation.
(TIF)

Acknowledgments

We thank Zoë Slattery for proofreading the manuscript.

Author Contributions

Conceptualization: Leo Zeitler, Cyril Denby Wilkes, Arach Goldar, Julie Soutourina.

Data curation: Leo Zeitler, Cyril Denby Wilkes.

Formal analysis: Leo Zeitler, Arach Goldar.

Funding acquisition: Julie Soutourina.

Investigation: Leo Zeitler, Cyril Denby Wilkes, Arach Goldar, Julie Soutourina.

Methodology: Leo Zeitler, Arach Goldar.

Project administration: Julie Soutourina.

Resources: Cyril Denby Wilkes, Arach Goldar, Julie Soutourina.

Software: Leo Zeitler.

Supervision: Cyril Denby Wilkes, Julie Soutourina.

Validation: Leo Zeitler, Cyril Denby Wilkes, Arach Goldar, Julie Soutourina.

Visualization: Leo Zeitler.

Writing – original draft: Leo Zeitler, Cyril Denby Wilkes, Arach Goldar, Julie Soutourina.

Writing – review & editing: Leo Zeitler, Cyril Denby Wilkes, Arach Goldar, Julie Soutourina.

References

1. Lehmann AR, Taylor EM. Conservation of eukaryotic DNA repair mechanisms. International journal of radiation biology. 1998; 74(3):277–286. https://doi.org/10.1080/095530098141429 PMID: 9737531
2. Gong F, Fahy D, Smerdon MJ. Rad4–Rad23 interaction with SWI/SNF links ATP-dependent chromatin remodeling with nucleotide excision repair. Nature structural & molecular biology. 2006; 13(10):902–907. https://doi.org/10.1038/nsmb1152 PMID: 17013386
3. Sarkar S, Kiely R, McHugh PJ. The Ino80 chromatin-remodeling complex restores chromatin structure during UV DNA damage repair. Journal of Cell Biology. 2010; 191(6):1061–1068. https://doi.org/10.1083/jcb.201006178 PMID: 21135142
4. Bohr VA, Smith CA, Okumoto DS, Hanawalt PC. DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. Cell. 1985; 40(2):359–369. https://doi.org/10.1016/0092-8674(85)90150-3 PMID: 3838150
5. Mellon I, Spivak G, Hanawalt PC. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. Cell. 1987; 51(2):241–249. https://doi.org/10.1016/0092-8674(87)90151-6 PMID: 3664636
6. Sweder KS, Hanawalt PC. Preferential repair of cyclobutane pyrimidine dimers in the transcribed strand of a gene in yeast chromosomes and plasmids is dependent on transcription. Proceedings of the National Academy of Sciences. 1992; 89(22):10696–10700. https://doi.org/10.1073/pnas.89.22.10696 PMID: 1438266
7. Deaconescu AM, Chambers AL, Smith AJ, Nickels BE, Hochschild A, Savery NJ, et al. Structural basis for bacterial transcription-coupled DNA repair. Cell. 2006; 124(3):507–520. https://doi.org/10.1016/j.cell.2005.11.045 PMID: 16469698
8. Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. Proceedings of the National Academy of Sciences. 2016; 113(32):9057–9062. https://doi.org/10.1073/pnas.1606667113 PMID: 27457959
9. Li W, Adebali O, Yang Y, Selby CP, Sancar A. Single-nucleotide resolution dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. Proceedings of the National Academy of Sciences. 2018; 115(15):E3408–E3415. https://doi.org/10.1073/pnas.1801687115 PMID: 29581276
10. Boiteux S, Jinks-Robertson S. DNA repair mechanisms and the bypass of DNA damage in Saccharomyces cerevisiae. Genetics. 2013; 193(4):1025–1064. https://doi.org/10.1534/genetics.112.145219 PMID: 23547164
11. Yu S, Evans K, Van Eijk P, Bennett M, Webster RM, Leadbitter M, et al. Global genome nucleotide excision repair is organized into domains that promote efficient DNA repair in chromatin. Genome research. 2016; 26(10):1376–1387. https://doi.org/10.1101/gr.209106.116 PMID: 27470111

12. van Eijk P, Nandi SP, Yu S, Bennett M, Leadbitter M, Teng Y, et al. Nucleosome remodeling at origins of global genome–nucleotide excision repair occurs at the boundaries of higher-order chromatin structure. Genome research. 2019; 29(1):74–84. https://doi.org/10.1101/gr.237198.118 PMID: 30552104
13. Shimamoto N. One-dimensional diffusion of proteins along DNA: its biological and chemical significance revealed by single-molecule measurements. Journal of Biological Chemistry. 1999; 274(22):15293–15296. https://doi.org/10.1074/jbc.274.22.15293 PMID: 10336412
14. Stauffer D, Schulze C, Heermann DW. Superdiffusion in a model for diffusion in a molecularly crowded environment. Journal of biological physics. 2007; 33(4):305–312. https://doi.org/10.1007/s10867-008-9075-2 PMID: 19669520
15. Bronstein I, Israel Y, Kepten E, Mai S, Shav-Tal Y, Barkai E, et al. Transient anomalous diffusion of telomeres in the nucleus of mammalian cells. Physical review letters. 2009; 103(1):018102. https://doi.org/10.1103/PhysRevLett.103.018102 PMID: 19659180
16. Tortora MM, Salari H, Jost D. Chromosome dynamics during interphase: a biophysical perspective. Current opinion in genetics & development. 2020; 61:37–43. https://doi.org/10.1016/j.gde.2020.03.001 PMID: 32304901
17. Oliveira GM, Oravecz A, Kobi D, Maroquenne M, Bystricky K, Sexton T, et al. Precise measurements of chromatin diffusion dynamics by modeling using Gaussian processes. Nature communications. 2021; 12(1):1–11. https://doi.org/10.1038/s41467-021-26466-7 PMID: 34702821
18. Diaz-Diaz F, Estrada E. Time and space generalized diffusion equation on graph/networks. Chaos, Solitons & Fractals. 2022; 156:111791. https://doi.org/10.1016/j.chaos.2022.111791
19. Kong M, Van Houten B. Rad4 recognition-at-a-distance: Physical basis of conformation-specific anomalous diffusion of DNA repair proteins. Progress in biophysics and molecular biology. 2017; 127:93–104. https://doi.org/10.1016/j.pbiomolbio.2016.12.004 PMID: 27939760
20. Kolmogorov AN. On the statistical theory of the crystallization of metals. Bull Acad Sci USSR, Math Ser. 1937; 1(3):355–359.
21. Johnson WA, Mehl RF. Reaction kinetics in processes of nucleation and growth. Am Inst Min Metal Petro Eng. 1939; 135:416–458.
22. Avrami M. Kinetics of phase change. I General theory. The Journal of chemical physics. 1939; 7 (12):1103–1112. https://doi.org/10.1063/1.1750380
23. Avrami M. Kinetics of phase change. II transformation-time relations for random distribution of nuclei. The Journal of chemical physics. 1940; 8(2):212–224. https://doi.org/10.1063/1.1750631
24. Avrami M. Granulation, phase change, and microstructure kinetics of phase change. III. The Journal of chemical physics. 1941; 9(2):177–184. https://doi.org/10.1063/1.1750872
25. Zeitler L. leoTiez/jmak: v1.0.0; 2022. Available from: https://doi.org/10.5281/zenodo.6669794.
26. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JH. Understanding nucleotide excision repair and its roles in cancer and ageing. Nature reviews Molecular cell biology. 2014; 15(7):465–481. https://doi.org/10.1038/nrm3822 PMID: 24954209
27. Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Asymmetric repair of UV damage in nucleosomes imposes a DNA strand polarity on somatic mutations in skin cancer. Genome research. 2020; 30(1):12–21. https://doi.org/10.1101/gr.253146.119 PMID: 31871068
28. Harlen KM, Trotta KL, Smith EE, Mosaheb MM, Fuchs SM, Churchman LS. Comprehensive RNA polymerase II interactomes reveal distinct and varied roles for each phospho-CTD residue. Cell reports. 2016; 15(10):2147–2158. https://doi.org/10.1016/j.celrep.2016.05.010 PMID: 27239037
29. Park D, Morris AR, Battenhouse A, Iyer VR. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. Nucleic acids research. 2014; 42(6):3736–3749. https://doi.org/10.1093/nar/gkt1366 PMID: 24413663
30. Luijsterburg MS, von Bornstaedt G, Gourdin AM, Politi AZ, Moné MJ, Warmerdam DO, et al. Stochastic and reversible assembly of a multiprotein DNA repair complex ensures accurate target site recognition and efficient repair. Journal of Cell Biology. 2010; 189(3):445–463. https://doi.org/10.1083/jcb.200909175 PMID: 20439997
31. Politi A, Moné MJ, Houtsmuller AB, Hoogstraten D, Vermeulen W, Heinrich R, et al. Mathematical modeling of nucleotide excision repair reveals efficiency of sequential assembly strategies. Molecular cell. 2005; 19(5):679–690. https://doi.org/10.1016/j.molcel.2005.06.036 PMID: 16137623
32. Semenenko V, Stewart RD, Ackerman EJ. Monte Carlo simulation of base and nucleotide excision repair of clustered DNA damage sites. I. Model properties and predicted trends. Radiation research. 2005; 164(2):180–193. https://doi.org/10.1667/RR3414 PMID: 16038589

33. Tran N, Qu PP, Simpson DA, Lindsey-Boltz L, Guan X, Schmitt CP, et al. In Silico Construction of a Protein Interaction Landscape for Nucleotide Excision Repair. Cell biochemistry and biophysics. 2009; 53 (2):101–114. https://doi.org/10.1007/s12013-009-9042-y PMID: 19156361

34. Daley DJ, Vere-Jones D, et al. An introduction to the theory of point processes: volume I: elementary theory and methods. Springer; 2003.

35. Schmidt HG, Sewitz S, Andrews SS, Lipkow K. An integrated model of transcription factor diffusion shows the importance of intersegmental transfer and quaternary protein structure for target site finding. PLOS one. 2014; 9(10):e108575. https://doi.org/10.1371/journal.pone.0108575 PMID: 25333780

36. Hu J, Lieb JD, Sancar A, Adar S. Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. Proceedings of the National Academy of Sciences. 2016; 113(41):11507–11512. https://doi.org/10.1073/pnas.1614430113 PMID: 27688757

37. Hu J, Adebali O, Adar S, Sancar A. Dynamic maps of UV damage formation and repair for the human genome. Proceedings of the National Academy of Sciences. 2017; 114(26):6758–6763. https://doi.org/10.1073/pnas.1706522114 PMID: 28607063

38. Vidaković AT, Mitter R, Kelly GP, Neumann M, Harreman M, Rodríguez-Martínez M, et al. Regulation of the RNAPII pool is integral to the DNA damage response. Cell. 2020; 180(6):1245–1261. https://doi.org/10.1016/j.cell.2020.02.009

39. Björner A. Some combinatorial and algebraic properties of Coxeter complexes and Tits buildings. Advances in Mathematics. 1984; 52(3):173–212. https://doi.org/10.1016/0001-8708(84)90021-5

40. Christian JW. The theory of transformations in metals and alloys. Newnes; 2002.

41. Georges A, Gopaul D, Denby Wilkes C, Giordanengo Aiach N, Novikova E, Barrault MB, et al. Functional interplay between Mediator and RNA polymerase II in Rad2/XPG loading to the chromatin. Nucleic acids research. 2019; 47(17):8988–9004. https://doi.org/10.1093/nar/gkz598 PMID: 31299084

S1 Appendix

Determining TCR Regions. Coordinates for transcribed regions were taken from [1]. This is set to be the TS. The area opposite of the TS is the NTS. All other segments are defined to be intergenic or non-transcribing. Here, we distinguish between Watson (positive) and Crick strand (negative). This allows us to show that there is no strand-specific bias. TCR-regions are defined to be transcripts that repair more or equal to 20% of their initial repair within 20 minutes. Approximately 89% of intergenic regions possess repair rates lower than 20% during the same time span (S2 Fig). Hence, we can have an increased confidence that genic regions with quicker repair are supported by TCR, leading to a more than 1%-point decrease of CPDs per minute. For all other genes, we cannot exclude the possibility that they are only repaired by GGR, as the pathway can function genome-wide. It is important to highlight that the observed CPD decrease is not uniform along the transcript. The end of the TS is seemingly less efficiently repaired. We require that only the first third of the gene after the TSS must possess more than a 20% decrease of damage within 20 minutes to be considered a TCR region.

References

1. Park D, Morris AR, Battenhouse A, Iyer VR. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. Nucleic acids research. 2014;42(6):3736–3749.

S2 Appendix

Discussing the Effect of Data Transformation and Selection. In order to make our results comparable, we followed the signal analysis described by [1]. However, we used three bins (beginning, centre, and end) instead of six. We also converted the data such that it represents repair instead of damage (Eq 7). This allowed a straightforward application of Eq 3. We additionally required that repair is greater than or equal to zero and monotonously increasing as a function of time (Eq 8). Some studies propose the notion of *dark* or *delayed* CPDs in human cells, which occur after UV treatment [2, 3]. However, to our knowledge there has not been a consensus over how *delayed* CPDs occur and influence repair dynamics. As we assume the biological process as well as the data probing itself to induce a considerable amount of noise, we prefer the interpretation that these data points should be rectified rather than representing damage created after irradiation.

In order to find potential groups that show similar repair dynamics, we compared the distribution of the model parameters against each other. Depending on the chosen segmentation and the type of genomic region, we found two to four clusters which were predominantly determined by the shape parameter m. When investigating the repair dynamics in detail, we found that all groups except one produced a switch-like behaviour (S3B Fig). This is due to the fact that data points cannot be brought into a line (S3C Fig). Whilst this could be a genuine property, we conjecture that this comes from the applied data normalisation. As discussed before, we require that no new lesions can be induced after irradiation. However, almost exclusively all regions with $m > 6.0$ originally possessed larger CPD signals after 20 minutes than directly after irradiation. During the data transformation, this data point was hence set to zero. Due to the form of Eq 3, lesion removal is seemingly acting exclusively between 20 and 60 minutes. Due to the lack of early repair and the data variability, we are convinced that these regions are not repaired by TCR and instead exhibit late acting GGR. They contribute significantly to the two distinguishable mechanisms in Fig 3. However, we have less confidence in the actual parameter values, as we gauge the data correction and consequent step-like behaviour to result from noise. A correlation with other nuclear processes could be therefore significantly biased. We excluded these regions from the

downstream analysis. It should be mentioned that we also tried an amendment to the algorithm to allow a larger flexibility for determining the repair kinetics in these regions. Here, we applied a weighted linear regression and required at least 1%-point difference in repair between two consecutive time steps. However, this solely brings the previously clearly separated groups closely together in parameter space. Consequently, the mutual effect of TCR and GGR becomes difficult to discern (despite being still detectable). There was no major change of the correlation analysis with respect to transcription rate, TU length, and nucleosome density in the *TCR* configuration. However, the relative distance to centromeres and telomeres changed drastically for the *gene* setup. As we have less confidence in the parameter values of regions with large variability, we prefer removing them from the correlation analysis while keeping two distinct repair mechanisms detectable.

In some cases, we could also find a grouping which was driven by the characteristic time τ. Large values only occurred in NTS or non-TCR/non-transcribed regions. We observed that these areas were all comparatively small, i.e. less than 300 base pairs (bp). Therefore, they are very susceptible to noise and processes from neighbouring regions. Instead of requiring a minimal length, we limited the range of τ to keep as many areas with potential useful information as possible. We assumed 200 minutes to be a sufficient time range for CPD repair to occur in yeast. All parameter ranges were set as follows: $m \in [0.5, 6.0]$; $\tau \in [20, 200]$; and $\theta \in [0.5, 1.0]$ for the TS of TCR regions (genes) and $\theta \in [0.4, 1.0]$ for all other areas. The number of remaining regions that fulfilled the set requirements changed considerably depending on the experimental setup. An overview is given in S1 Table. The TS was in almost all cases included in the subsequent analysis, although the end was more often outside the defined parameter ranges than the beginning and centre. Surprisingly, only around half of the NTSs met the requirements (both setups). The numbers are even worse for intergenic/non-TCR regions (both setups). Here, approximately a third of all non-transcribed/non-TCR areas were considered in the downstream computations. We hypothesise that repair at NTS and intergenic regions is dominated by accessinility to the lesion. Late repair times were also reported by [4] and [5]. Considering additionally the few time points, the data fitting of Eq 3 predicts no repair until 20 to 60 minutes, whereas all CPD decrease appears exclusively afterwards. This results in the aforementioned step-like behaviour. As

discussed before, we assume this to be rather unlikely. Analysing the repair dynamics in these regions could provide additional information. We hope that future research is inspired to repeat this analysis with a CPD-seq time course that has a finer temporal resolution.

References

1. Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Asymmetric repair of UV damage in nucleosomes imposes a DNA strand polarity on somatic mutations in skin cancer. Genome research. 2020;30(1):12–21.

2. Yim S, Lee J, Jo H, Scholten J, Willingham R, Nicoll J, et al. Chrysanthemum morifolium extract and ascorbic acid-2-glucoside (AA2G) blend inhibits UVA-induced delayed cyclobutane pyrimidine dimer (CPD) production in melanocytes. Clinical, cosmetic and investigational dermatology. 2019;12:823.

3. Fajuyigbe D, Douki T, van Dijk A, Sarkany RP, Young AR. Dark cyclobutane pyrimidine dimers are formed in the epidermis of Fitzpatrick skin types I/II and VI in vivo after exposure to solar-simulated radiation. Pigment cell & melanoma research. 2021;34(3):575–584.

4. Li W, Adebali O, Yang Y, Selby CP, Sancar A. Single-nucleotide resolution dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. Proceedings of the National Academy of Sciences. 2018;115(15):E3408–E3415.

5. Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. Proceedings of the National Academy of Sciences. 2016;113(32):9057–9062.

S3 Appendix

Dividing the Data into Genes versus Intergenic Regions. Instead of the commonly used partition into transcripts and intergenic regions, we opted to segment CPD-seq data into TCR (TS and NTS) and non-TCR regions (S1 Fig). In order to increase comparability with existing studies, we repeated the entire analysis for genes (TS and NTS, without sub-areas) and non-transcribed regions (Watson and Crick strand separately). Coordinates for transcripts were taken again from [1]. In the following, this is referred to as *gene* setup. The other aspects of the analysis remain the same.

We observed that the DC is much lower for the *gene* configuration (DC=0.241) (S18A Fig). Nevertheless, this is also true when comparing XR-seq with the repair rate derived from the data (DC=0.231) (S18B Fig). Thus, we judge this to be due to the data segmentation rather than the model assumptions (S4 Appendix) An overview over the correlation values in all setups is given in S2 Table.

When averaging the results over all instances of a region type (e.g. all genes), we can recover once again two distinct dynamics at genes. The early mechanism disappears on the NTS and both strands of intergenic regions, although NTS still shows higher repair rates at early time points than non-transcribed areas. As expected, Watson and Crick strand of intergenic regions follow identical trends. This is in line with our analysis at TCR regions.

Considering the parameter space, the clear pattern of early repair of TCR regions (low values for m) vanishes when considering all genes. The values are scattered much more broadly. As expected, the NTS tends to show low θ values. However, the distribution over m seems to be remarkably similar for all areas, with intergenic regions showing the largest dispersion.

The correlation with biological features differed sometimes considerably, which points out that the type of data segmentation is important. The transcription rate is seemingly weakly correlated with the NTS in the *gene* setup. Though, we should emphasise that both TS and NTS were compared to the same NET-seq value. Hence, we do not consider antisense transcription to be the reason. Instead, this could possibly indicate that accessibility to the lesion is influenced by transcription. Nucleosome density was clearly correlated to Crick and Watson strand in intergenic regions as well

as the TS. There was no link between nucleosome occupancy and repair on the NTS. TU length shows again a very clear interrelationship. Most interestingly, we report that the relative distance to centromeres and telomeres is seemingly important for the TS in the *gene* setup. We are unable to provide an intuition without further biological experiments. However, we could imagine a link between chromosome flexibility and repair, since it enables a larger number of DNA-protein interactions.

References

1. Park D, Morris AR, Battenhouse A, Iyer VR. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. Nucleic acids research. 2014;42(6):3736–3749.

S4 Appendix

Explaining the Correlation Between XR-seq Data and Repair Rate. As the derivative of Eq 3 represents the repair rate at a given time, we conjectured that it should correlate with XR-seq data. This is due to the fact that the signal shows the distribution of excised nucleotide sequences that were produced during DNA cleavage. It is surmised that they are quickly degraded, i.e. within five minutes [1]. However, this interrelationship is not necessarily linear, which speaks against the usage of Pearson's correlation. The distance correlation (DC) comes as a remedy by relating the distance of data points in a set to each other rather than the data itself (although it should be pointed out that it is not simply the Pearson's correlation of distances). It ranges from zero to one, with zero showing independence, whereas one indicates that the linear subspace between the data sets is equal. It is calculated as follows. Distance matrices A and B contain all pairwise distances, i.e. $\{A\}_{ij} = ||\mathbf{a_i} - \mathbf{a_j}||_2$ and $\{B\}_{ij} = ||\mathbf{b_i} - \mathbf{b_j}||_2$. Here, $\mathbf{a_x}$ and $\mathbf{b_x}$ ($\mathbf{x} \in \{\mathbf{i}, \mathbf{j}\}$) denote data points in sets $\mathcal{A}$ and $\mathcal{B}$, respectively. $||\dots||_2$ is the Euclidean distance. Each set contains n data points. Subsequently, A and B are double-centred. With the definition of the sample distance covariance

$$\mathrm{dCov}^2(A, B) = \frac{1}{n^2} \sum_j \sum_k A_{jk} B_{jk}, \tag{1}$$

as well as the sample distance variance

$$\mathrm{dVar}(A) = \mathrm{dCov}^2(A, A) = \frac{1}{n^2} \sum_j \sum_k A_{jk}^2, \tag{2}$$

,

we can introduce the DC:

$$DC(\mathcal{A}, \mathcal{B}) = \frac{\mathrm{dCov}^2(A, B)}{\sqrt{\mathrm{dVar}(A)\mathrm{dVar}(B)}}. \tag{3}$$

The DCs for all setups are given in S2 Table. In order to compare the values of Eq 6 to the CPD data, we transformed first the signal with respect to Eq 7. We calculated relative repair at three time points (i.e. 20, 60 and 120 minutes) by

$$r(t'_i) = \frac{R(t_i) - R(t_{i-1})}{(t_i - t_{i-1})/20}, \tag{4}$$

where $R(t)$ denotes repair determined by Eq 7. $t_i \in \mathbf{t} = (20, 60, 120)$ and $t'_i \in \mathbf{t'} = (5, 20, 60)$. The values must be re-scaled to the same time step to make them comparable. However, the CPD decrease within the first 20 minutes is relatively small for most areas. Calculating $r(t)$ per minute results in an almost flat line. All values represent therefore repair within 20 minutes. This makes the time points comparable whilst avoiding having too small values for $r(5)$.

Despite the fact that we determined the DC for all time points—i.e. 5, 20, and 60 minutes after repair—it is intuitive to see that $r(t)$ is more heterogeneous when taking all time points together. We therefore consider only the DC of the entire data set.

References

1. Li W, Adebali O, Yang Y, Selby CP, Sancar A. Single-nucleotide resolution dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. Proceedings of the National Academy of Sciences. 2018;115(15):E3408–E3415.

S5 Appendix

Discussing the kNN Approach. The parameters of Eq 3 can be used to set repair into the context of other biological and nuclear properties in the cell. As databases provide a large variety of data signals, it is reasonable to opt for a data driven approach. However, finding such a correlation is not straightforward. Many of the NGS histogram distributions peak sharply around a low value whilst also including far distant outliers. Moreover, it can be assumed that sequencing signals include a large amount of noise. Precise predictions for data values around the histogram peak are difficult. This excludes continuous regression approaches or widely used correlation indices, such as Pearson's correlation, DC, or mutual information. We circumvented these issues by transforming the mapping into a binary classification problem to analyse general trends. This reduces the impact of noise. Due to using equally many values for both classes, we removed any distribution specific bias.

Another requirement was comparability between results. It is a known fact that the performance of machine learning models can vary strongly depending on the number of parameters or used architecture [1]. Through using the nonparametric kNN approach, we could provide equivalent treatment for all setups. We also mitigated the impact of k by applying different values in a reasonable range, i.e. $k \in \{5, 10, 20, 50, 100\}$. Thus, we did not rely on any particular parameter setting or defined spline ranges. We are aware that parametric models can be efficiently implemented for a variety of tasks, as it has been recently shown with Alpha Fold 2 [2]. Moreover, architectural biases of parametric models could be possibly reduced by systematic parameter searches. Nevertheless, it should be emphasised that this study did not intend to find the best performing machine learning model. Rather, we aimed to show non-random correlations and therefore indicate potential repair influences. Some researchers even conjecture that nonparametric models could be generally better performing [3]. We conclude that the kNN approach is a sensible choice.

In the following, we want to provide some further intuition using the example of TU length. S6 Fig shows the learnt function and the prediction error distribution for the correct and random model, respectively. The correct mapping finds a distribution pattern with big genes predominantly distributed to the centre right, whereas small

genes are found in its periphery (S6A Fig). As expected, this pattern is destroyed in the random mapping (S6C Fig). The error distribution for the true kNN is equally large for values that are over and underestimated (right histogram in S6B Fig). This speaks for an unbiased mapping. We also want to point attention towards the spatial distribution of the prediction error. Red circles mark data values which were incorrectly classified as large genes, whereas the blue points are parameters that were wrongly associated with a small size. The distribution of the red and blue disks follows our expectations from the learnt map. The wrong classification could indicate noise in the data set or unknown information that cannot be represented. Compared with S6D Fig, it is clear that these trends vanish in the random model. However, the prediction error is surprisingly low (0.4). A repetition over 100 iterations is therefore indispensable to find significant links and to remove data specific biases.

References

1. Dyson F. A meeting with Enrico Fermi. Nature. 2004;427(6972):297–297.
2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589.
3. Perretti CT, Munch SB, Sugihara G. Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. Proceedings of the National Academy of Sciences. 2013;110(13):5253–5257.

S6 Appendix

Analysing Repair Kinetics in Context of Abf1 and H2A.Z Distribution. Different chromatin features including transcription factor binding sites or histone variants and modifications can affect CPD repair. An essential role in GGR recognition is allocated to the Rad7-Rad16-Abf1 protein complex. Yeast strains with respective gene deletions are incapable to repair lesions in non-transcribed regions and are inviable under genotoxic stress [1, 2, 3]. Abf1 binding was also proposed to inhibit CPD formation [4] and to influence GGR kinetics [2, 3]. Moreover, TCR and GGR are both reportedly influenced by multiple histone modifications and variants [5]. Binding sites for the hypothesised GGR-complex are flanked by H2A.Z histone variant-containing nucleosomes [6]. However, a direct relationship between lesion removal and Abf1 occupancy or H2A.Z distribution has not been investigated. Building up on previous work, we presumed particularly strong correlation in intergenic regions. Abf1 and H2A.Z distribution was probed after UV treatment by [6]. In the following, we include results from the *TCR* configuration (as presented in the main text) and the *gene* setup (see S3 Appendix).

The experiments with Abf1 yielded a mixed bag of results (S10 Fig). When considering the *TCR* setup, we found a non-random correlation with the repair dynamics in non-TCR areas regions for all k (S5 Appendix). A strong correlation is in line with the idea of Abf1 being part of the GGR complex, whose effect is presumed to be strong in non-TCR regions [6]. The end of TCR areas seemed to exhibit a slight correlation with Abf1 as well. However, the *gene* configuration found a link to almost all regions with the exception of the Watson strand in intergenic areas. This is surprising, as we would expect both strands to behave similarly. Whilst this could hint to a strand-specific bias, it is likely that the influence of Abf1 at intergenic regions in the *gene* setup is weaker than in the *TCR* configuration. A correct definition of genomic areas is hence clearly important to put the results into the right context. Abf1's role is associated with GGR [2, 3, 6] as well as transcription regulation and replication [7, 8]. It is intuitive that due to its multifunctional involvement, it is indirectly affecting a broad variety of regions.

The outcomes for H2A.Z were similarly ambiguous (S11 Fig). We found a

non-random correlation to all regions in the *TCR* configuration. On the other hand, we were unable to find a significant correlation in intergenic areas for the *gene* setup. Nonetheless, there was a definite interrelationship between the histone marker distribution and the repair dynamics at the TS and NTS. This indicates a non-negligible role for H2A.Z modification during lesion removal at active genes. This is unsurprising giving its regulatory role in gene expression [9]. However, a correlation with GGR might be less strong.

To put the results into context, it is important to mention that the histogram distribution of Abf1 and H2A.Z data was different in intergenic regions compared to transcribed areas. This was especially visible in the *gene* setup. It has been previously reported that Abf1 binding sites tend to colocalise with CID boundaries, which are usually found in intergenic regions. In the same paper, it was also proposed that they are flanked by H2A.Z-containing barrier nucleosomes [6]. Assuming that Abf1 is necessary for GGR [1, 3]—and therefore plays a specifically crucial role in intergenic regions—it is not surprising that the histogram exhibits different distributions for genes and non-transcribed areas. Since the distributions are still similar between strands, we do not consider this as having a strong influence on the final conclusions.

References

1. Boiteux S, Jinks-Robertson S. DNA repair mechanisms and the bypass of DNA damage in Saccharomyces cerevisiae. Genetics. 2013;193(4):1025–1064.

2. Yu S, Owen-Hughes T, Friedberg EC, Waters R, Reed SH. The yeast Rad7/Rad16/Abf1 complex generates superhelical torsion in DNA that is required for nucleotide excision repair. DNA repair. 2004;3(3):277–287.

3. Yu S, Evans K, Van Eijk P, Bennett M, Webster RM, Leadbitter M, et al. Global genome nucleotide excision repair is organized into domains that promote efficient DNA repair in chromatin. Genome research. 2016;26(10):1376–1387.

4. Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. Proceedings of the National Academy of Sciences. 2016;113(32):9057–9062.

5. Adam S, Dabin J, Polo SE. Chromatin plasticity in response to DNA damage: The shape of things to come. DNA repair. 2015;32:120–126.

6. van Eijk P, Nandi SP, Yu S, Bennett M, Leadbitter M, Teng Y, et al. Nucleosome remodeling at origins of global genome–nucleotide excision repair occurs at the boundaries of higher-order chromatin structure. Genome research. 2019;29(1):74–84.

7. Yarragudi A, Miyake T, Li R, Morse RH. Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast Saccharomyces cerevisiae. Molecular and cellular biology. 2004;24(20):9152–9164.

8. Kohzaki H, Murakami Y. A transcription factor Abf1 facilitates ORC binding onto the Saccharomyces cerevisiae replication origin via histone acetylase Gcn5. bioRxiv. 2019; p. 583310.

9. Giaimo BD, Ferrante F, Herchenröther A, Hake SB, Borggrefe T. The histone variant H2A. Z in gene regulation. Epigenetics & chromatin. 2019;12(1):1–22.

S7 Appendix

Comparing the KJMA Model With Other Approximations. The KJMA model is used as a tool to find the parameters. It remains to address that the representation is reasonable. We compared the accuracy of Eq 3 with a linear model as well as a logistic regression and the Hill equation, both of which produce an S-shaped function. Performance was measured using the mean-squared error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2. \tag{1}$$

An example is given in S13B Fig. We can confirm that the KJMA model achieves a lower error than the linear or the logistic regression model. However, the Hill equation seems to perform slightly yet significantly better (S13C Fig). Therefore, it seems likely that the desired function should follow an S-shape trajectory that is non-symmetric at the inflection point (in contrast to the logistic regression).

We want to provide some mathematical explanation why the Hill function performs similar yet slightly better. Assuming a ligand concentration [L], the fraction of bound receptor proteins can be described by

$$f(t) = \frac{1}{1 + \left(\frac{K_A}{[\mathrm{L}]}\right)^{m'}},$$

where K_A is the ligand concentration that is resulting in $f(t) = 0.5$, and m' represents the number of binding sites on the target protein. When assuming $\left(\frac{K_A}{[\mathrm{L}]}\right)^{m'}$ to be small—i.e. there is a sufficient surplus of the ligand L—we can approximate the Hill equation by $f(t) \approx 1 - \left(\frac{[\mathrm{L}]}{K_A}\right)^{m'}$. This is nothing else but the first order Taylor expansion of $1 - \exp\left[\left(-\frac{[\mathrm{L}]}{K_A}\right)^{m'}\right]$, which has the same form as the KJMA model. The slightly different behaviour is due to the shape of $1/x$ (which is the determining term in the Hill equation) and $1 - e^{-x}$.

It should be stressed that the Hill equation does not have a particular meaning to describe a temporal process such as DNA repair, since it was used to explain ligand occupancy with respect to the available quantity. There is no notion of time. We also noticed that when taking the average over regions that were supposed to have no TCR, the Hill equation still showed a double-repair behaviour (S13D Fig). Together with the

fact that both perform similarly, we think that the KJMA model provides a good representation.

S8 Appendix

Discussing the Model in Context of the Physical KJMA Model.

The KJMA model itself has been successfully applied in a biological context to analyse the dynamics of DNA replication in eukaryotes [1]. Nevertheless, the study did not include predictions with respect to independently probed data. More importantly, it is specific for DNA replication and cannot be applied to analyse CPD repair. In order to allow a physical interpretation and to re-group repair regions to patterns, we presumed independent repair kinetics between and within cells. Independence between cells has been discussed already above. The supposition of independent repair dynamics within a cell is based on two observations. Firstly, we assume that the spatial effect of lesion removal kinetics decreases as a function of distance. Hence, the farther away the CPD positions, the smaller the impact on each other. This is justified by the relatively small area of lesion removal ($\approx$ 30 nt [2]) and the notion of chromosome interaction domains (CIDs) [3]. Secondly, it has been reported for *Caenorhabditis elegans* that a UVC treatment of 100J/m^2 induces 0.4 to 0.5 CPDs per 10kb [4]. A similar UV dose (125 J/m^2) was used by [5]. Taking this as a reference, we presume that a comparable dose of UVC induces a corresponding number of CPDs in budding yeast. It is thus unreasonable to expect more than one lesion per CID per cell, as they are commonly less than 10kb in *Saccharomyces cerevisiae* [3]. It is true though that this could be species-dependent. Due to the lack of other studies, we take it as given that lesion removal does not affect each other within a single cell.

The independence assumptions permitted the application of the KJMA model. As we surmise that the sequencing data contains a hidden axis, we represent repair on a grid. Thus, it would be expected that the found shape parameter indicates a two-dimensional space, i.e. $m - 1 \approx 2$. However, as reported above, m exhibits a large range. We do not presume that such a deviation is only caused by noise. Instead, a similar behaviour can be observed when allowing the growth speed G to be larger during earlier time points rather than later in the process, and vice versa. Implicitly, this incorporates the possibility that $G(t)$ is non-constant in time. m can be interpreted to speak for the time-dependence of the process instead of a particular dimension [6]. Low values represent quicker repair in the beginning rather than in the end. A large m

indicates that G increases later on.

The nucleation rate n though is presumed to be constant. As explained above, this has as a consequence that the framework models repair with only one mechanism per region. There is a scientific consensus that intergenic regions and the NTS can be only repaired by GGR. For the TS of genes it is nevertheless surmised that TCR and GGR can act collectively. On a population scale, this would likely result in the repair rate to contain two peaks over time. TCR would be observable in the beginning and subsequently abate. GGR is supposedly acting later during the process. The collective effect of TCR and GGR in a genic area can be recovered by taking the average over an entire group, e.g. the beginning of TCR regions. Despite assuming similar kinetics, we presume that the noise in the process should lead to a representation by either TCR or GGR in a ratio comparable to their respective repair contribution. It should be highlighted that Eq 3 can be easily adapted to represent heterogeneous repair times by defining $n(t)$ (and therefore $\beta(t)$ since $n(t)$ is incorporated) as a function of time. However, any parameter estimation of such a function would be merely based on guesses due to the sparse temporal data resolution. We followed the principle of Occam's razor and opted for a simpler model. The production of CPD data with smaller time steps could permit such an estimation.

Finally, we also want address the analogy of the KJMA model to the stochastic point process. We linked the expansion of the pattern— and therefore $G(t)$—to the diffusion in our model. This can be explained by considering Eq 4, as it includes the Avrami exponent that we before linked to the time dependence of the process. From the perspective of the stochastic point process, the time dependence is incorporated by the diffusion, whose mean squared displacement is proportional to $D_{\hat{m}}t^{\hat{m}}$. The nucleation rate n was compared to the expected waiting time $\hat{\beta}$, both of which are time-independent. Consequently, the growth of repair patterns in the abstract repair space becomes an important property, since if we would only consider a constant nucleation rate, the resulting repair dynamics should follow the trajectory of a homogeneous Poisson point process with constant λ (S13 Fig).

We want to highlight that this alternative understanding cannot be taken literally and should be therefore used with some scepticism. However, we strongly believe that this interpretation could potentially unlock additional information, as it allows the

incorporation of well studied results from the physical model.

References

1. Jun S, Bechhoefer J. Nucleation and growth in one dimension. II. Application to DNA replication kinetics. Physical Review E. 2005;71(1):011909.

2. Boiteux S, Jinks-Robertson S. DNA repair mechanisms and the bypass of DNA damage in Saccharomyces cerevisiae. Genetics. 2013;193(4):1025–1064.

3. Hsieh THS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping nucleosome resolution chromosome folding in yeast by micro-C. Cell. 2015;162(1):108–119.

4. Meyer JN, Boyd WA, Azzam GA, Haugen AC, Freedman JH, Van Houten B. Decline of nucleotide excision repair capacity in aging Caenorhabditis elegans. Genome biology. 2007;8(5):1–17.

5. Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. Proceedings of the National Academy of Sciences. 2016;113(32):9057–9062.

6. Christian JW. The theory of transformations in metals and alloys. Newnes; 2002.

S9 Appendix

Investigating a Link Between Transcription Rate and TU length. It is usually assumed that the size of genes does not influence the frequency with which they are transcribed. Thus, both parameters are expected to be independent. During our analysis, we noticed that the NET-seq signal amplitude decreases as function of distance from the TSS (S15 Fig). This could possibly induce a size-specific bias if the decline occurs repeatedly within a specific distance, e.g. 500 bp from the TSS. In this case, it would affect small genes more strongly than large genes. In order to gauge the bias' impact, we compared the NET-seq signal with Pol2 ChIP-seq data [1], which we assume to represent transcription rate to a reasonable degree. We can verify that Pol2 ChIP-seq does not exhibit the same declining trend after the TSS. Indeed, NET-seq shows a larger correlation with respect to TU length (DC=0.321) than Pol2 occupancy (DC=0.224, S16 Fig). However, we can establish a rather strong interrelationship between NET-seq and Pol2 ChIP-seq data (DC=0.75, S17 Fig). When we scrutinised the link to repair, we divided all genes into two groups with high or low transcription rate. We can verify that the majority is still within the same group, independent of the use of NET-seq or Pol2 ChIP-seq data. Therefore, the conclusions about the relationship between transcription and repair remain nevertheless sensible. We opted to use NET-seq data to permit a direct comparison with the results from [2].

References

1. Georges A, Gopaul D, Denby Wilkes C, Giordanengo Aiach N, Novikova E, Barrault MB, et al. Functional interplay between Mediator and RNA polymerase II in Rad2/XPG loading to the chromatin. Nucleic acids research. 2019;47(17):8988–9004.

2. Li W, Adebali O, Yang Y, Selby CP, Sancar A. Single-nucleotide resolution dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. Proceedings of the National Academy of Sciences. 2018;115(15):E3408–E3415.

S1 Table

The number of models per region, before and after applying the requirements for parameter ranges. IGR abbreviates intergenic regions.

Experimental setup	Region name	#Total	#Filtered
Gene	TS	4973	4356
	NTS	4973	2294
	IGR +	4067	1591
	IGR -	4067	1583
TCR	TS start	1878	1865
	TS centre	1878	1703
	TS end	1878	1367
	NTS start	1878	840
	NTS centre	1878	1080
	NTS end	1878	1117
	IGR	1763	650

S2 Table

The DC between XR-seq and repair predictions / data for different experimental configurations.

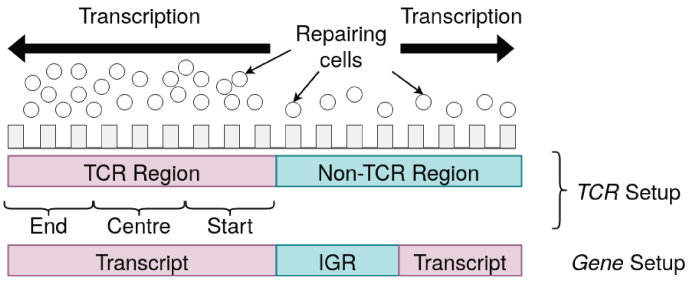
Experimental setup	5 min	20 min	60 min	Total
TCR setup: **model**	0.405	0.525	0.258	0.441
TCR setup: **data**	0.433	0.644	0.452	0.209
Gene setup: **model**	0.226	0.396	0.216	0.241
Gene setup: **data**	0.242	0.621	0.342	0.231

S3 Table

Number of non-random interrelationships between model parameters and sequencing data over k. The table gives the number of kNN models that could find a correlation between model parameters and genomic context. $k \in \{5, 10, 20, 50, 100\}$. We defined a link to be significant if at least three out of five k find a non-random interrelationship. - means that data was not used in the given configuration. NET denotes NET-seq data, ND is nucleosome density, and meres give the relative distance to centromeres or telomeres. Suffixes S, C, and E denote start, centre and end of an area. NTCR are non-TCR areas. IGR are intergenic/non-transcribed regions.

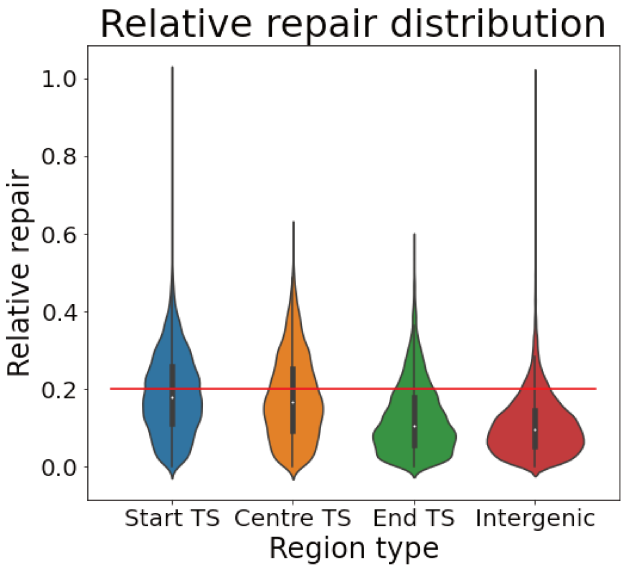
	***TCR* setup**							***Gene* setup**			
	TS S	TS C	TS E	NTS S	NTS C	NTS E	NTCR	TS	NTS	IGR+	IGR-
NET	5	5	5	0	1	1	-	5	3	-	-
Size	5	5	5	5	5	5	-	5	5	-	-
ND	3	2	0	0	0	0	5	5	0	5	5
Abf1	0	2	3	0	0	0	5	4	5	2	4
H2A.Z	5	5	5	5	5	4	3	5	5	2	0
Meres	0	0	0	0	0	0	0	5	0	0	0

S1 Fig



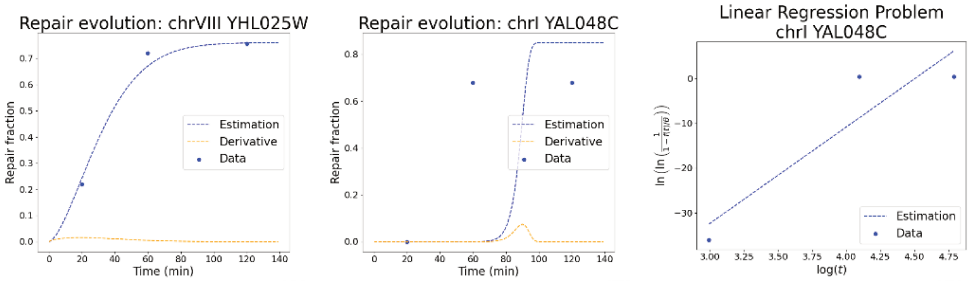
Scheme of the segmentation setup.

S2 Fig



Relative repair distribution over genomic areas.

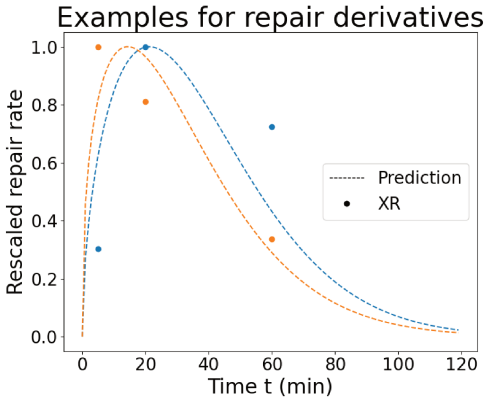
S3 Fig



(A) (B) (C)

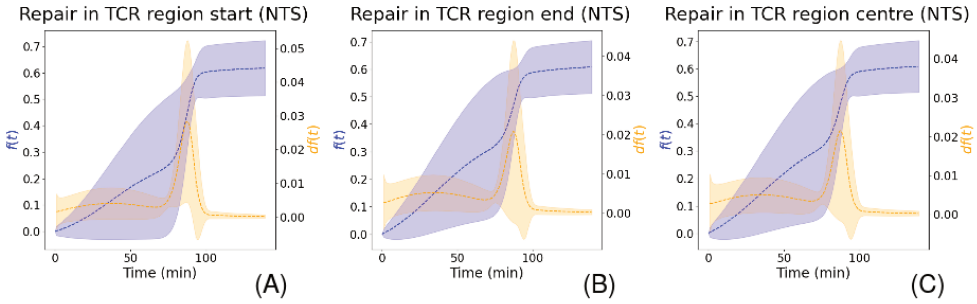
Example for model predictions.

S4 Fig



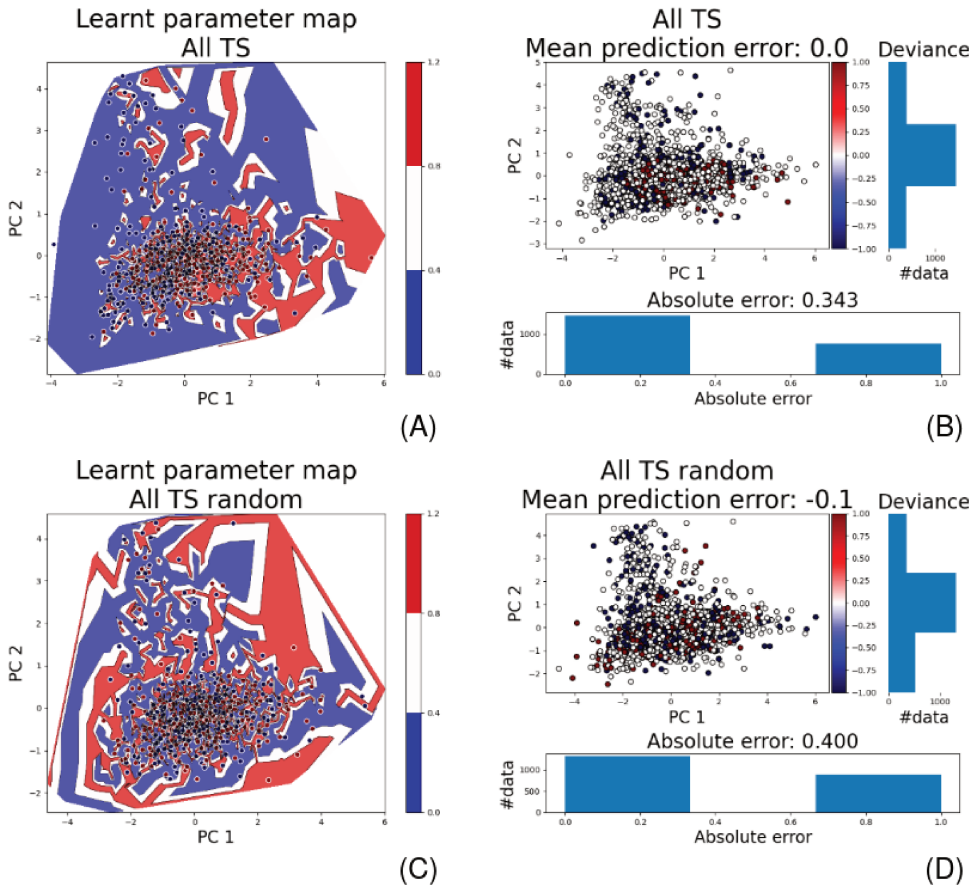
Example for model prediction and XR-seq data over time.

S5 Fig



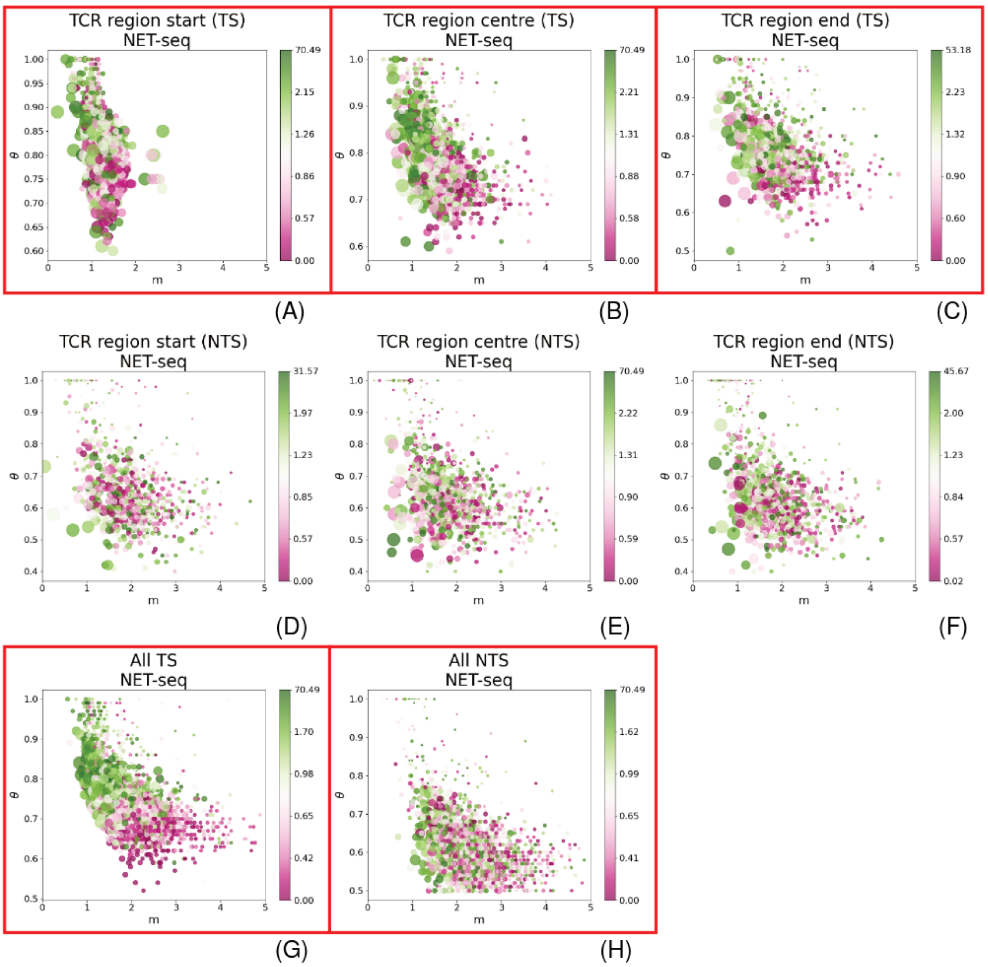
The evolution of repair along the NTS.

S6 Fig



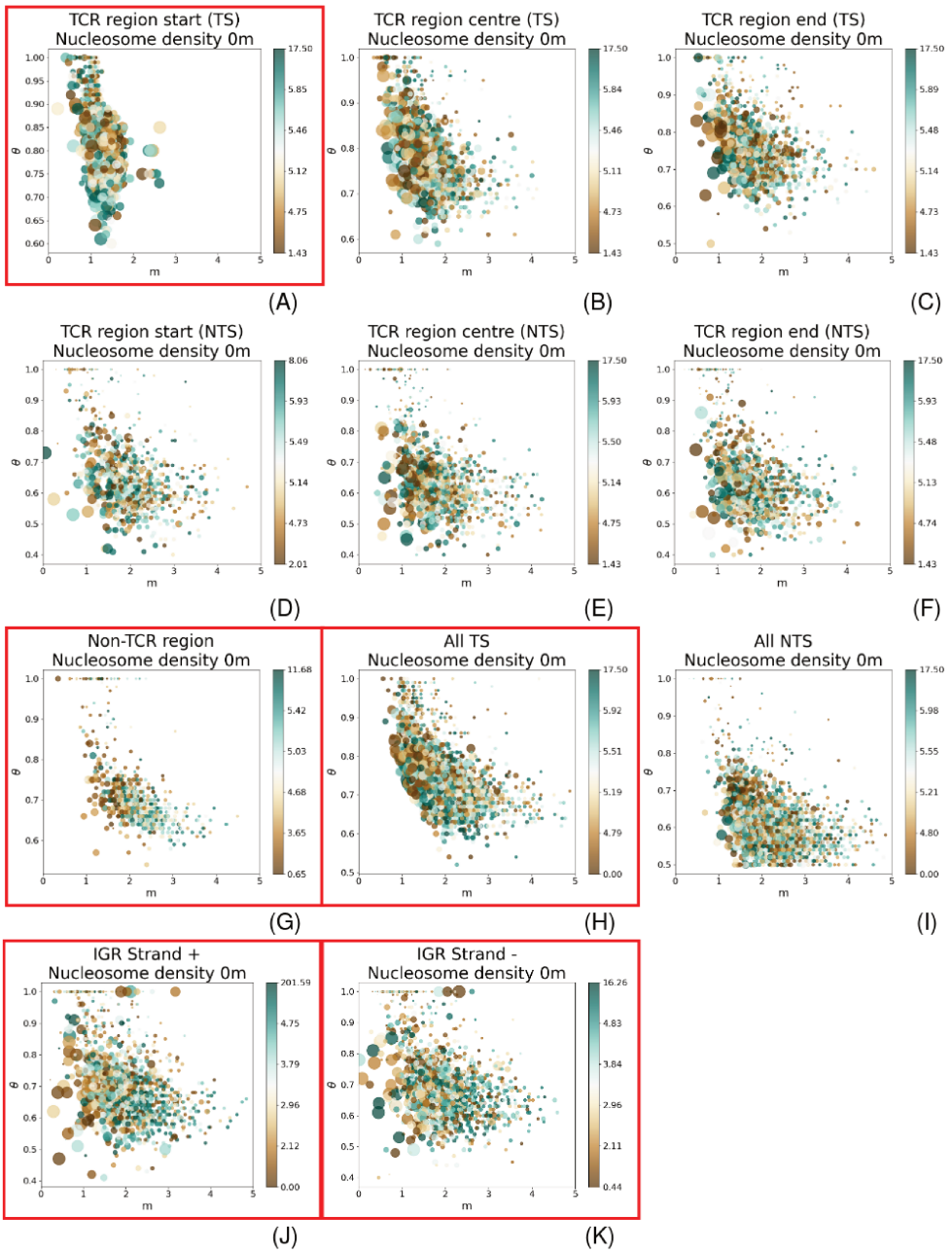
Example of the learnt function between model parameters and genomic context.

S7 Fig



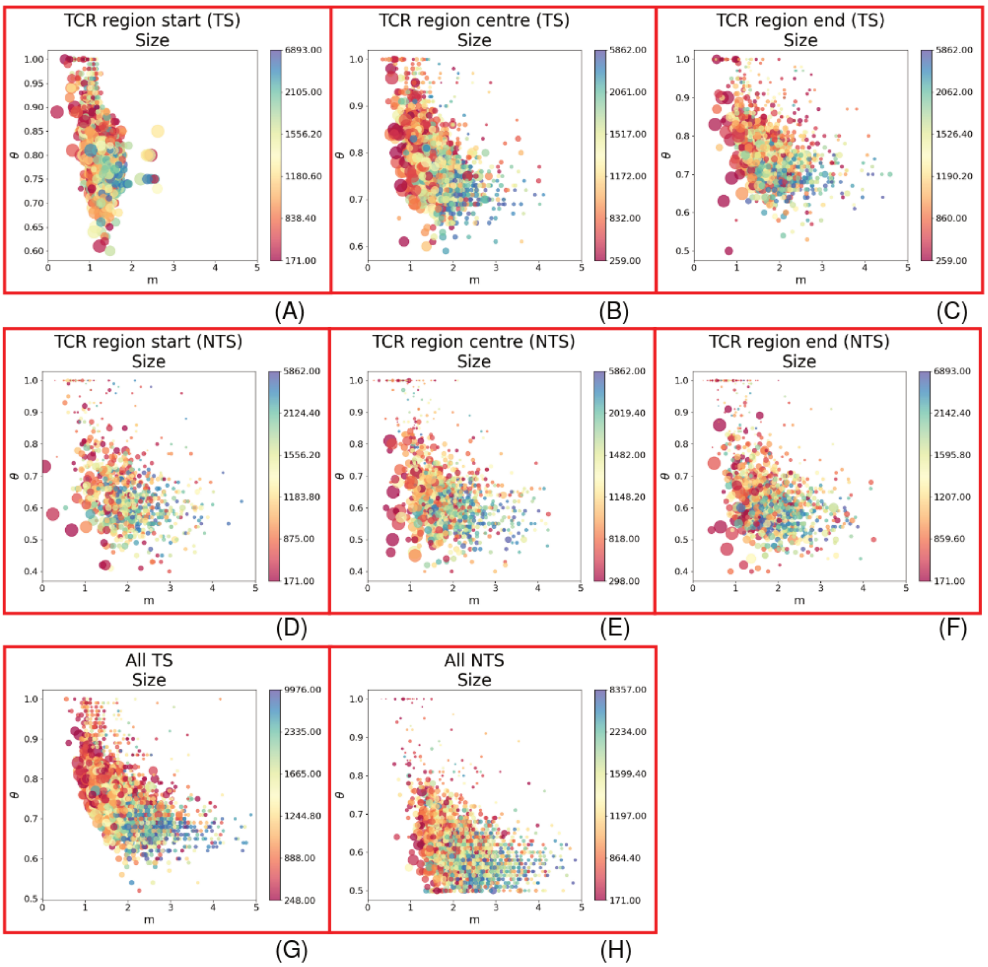
Model parameters with respect to transcription rate.

S8 Fig



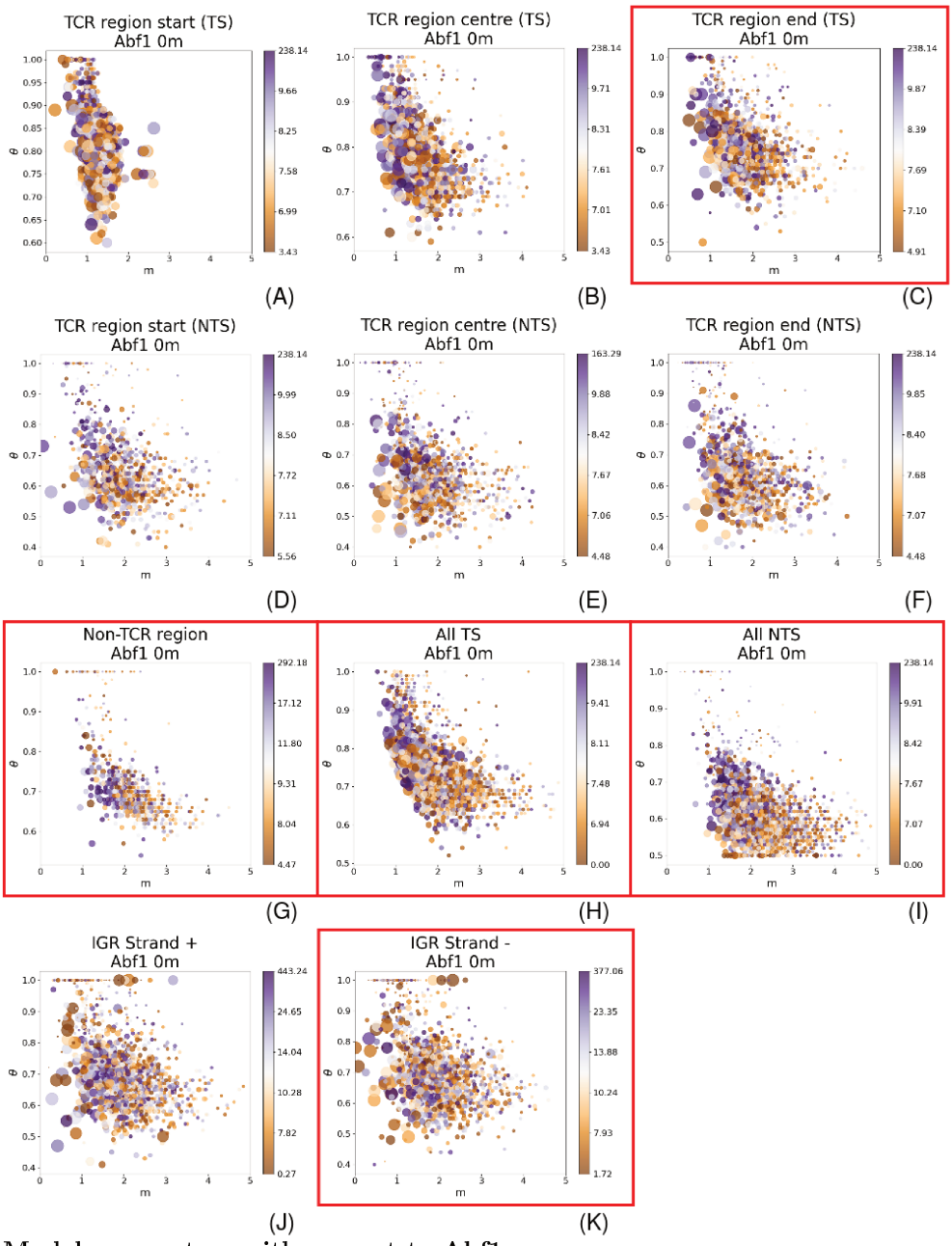
Model parameters with respect to nucleosome density.

S9 Fig



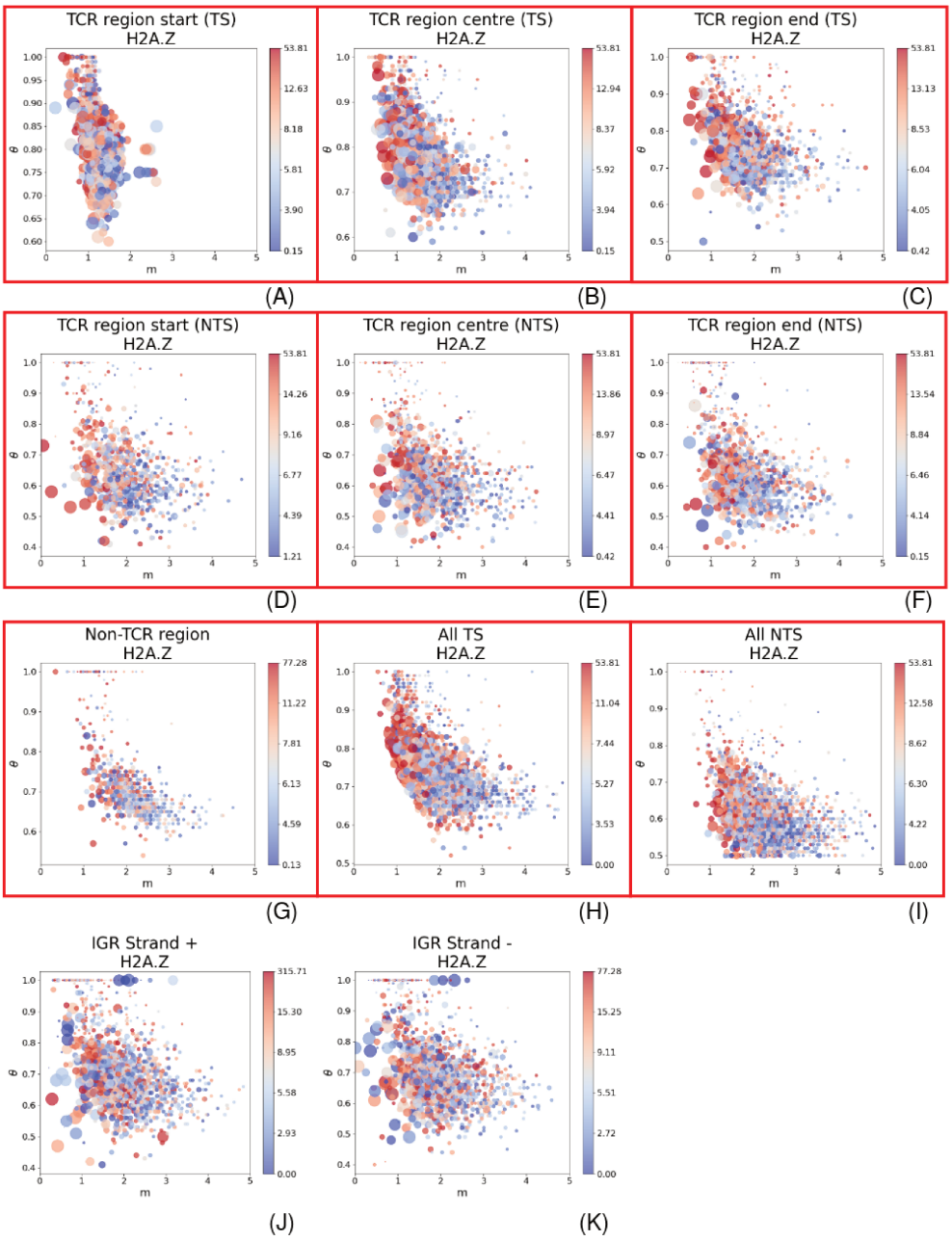
Model parameters with respect to size.

S10 Fig



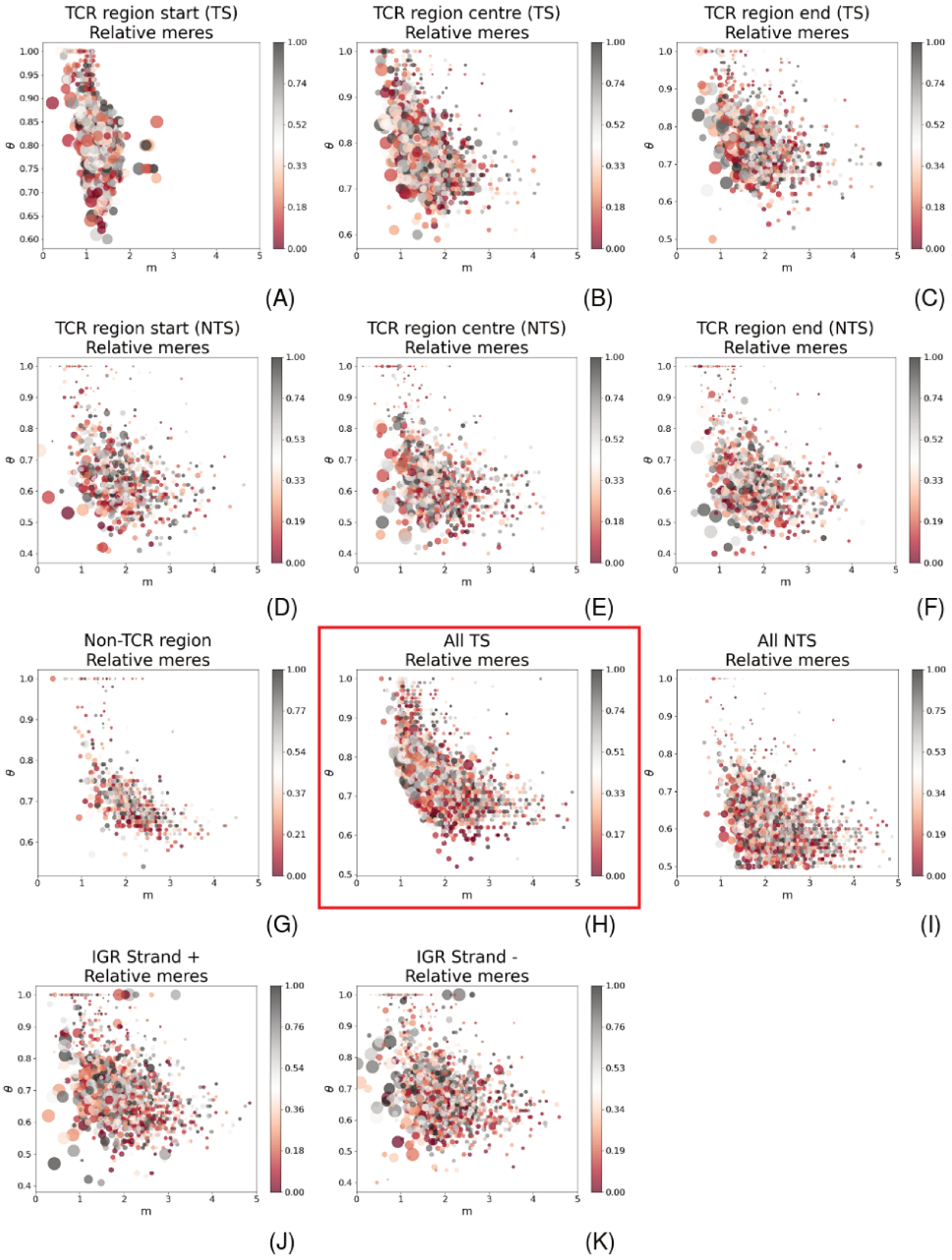
Model parameters with respect to Abf1.

S11 Fig



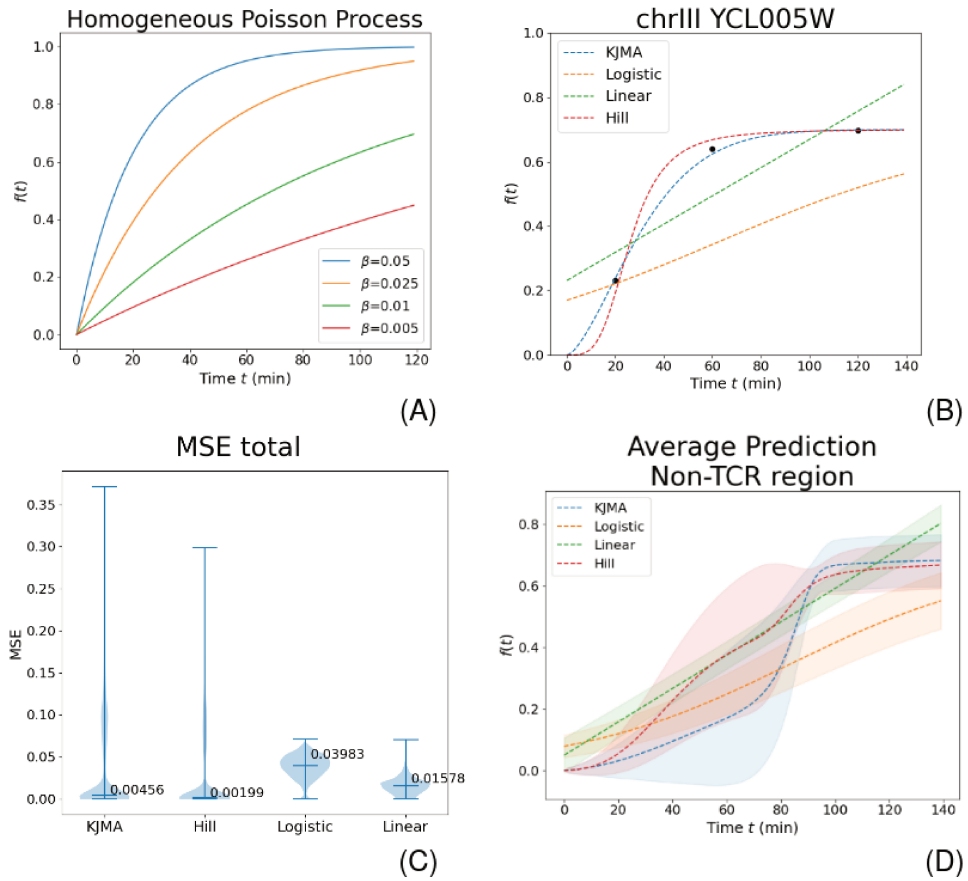
Model parameters with respect to H2A.Z.

S12 Fig



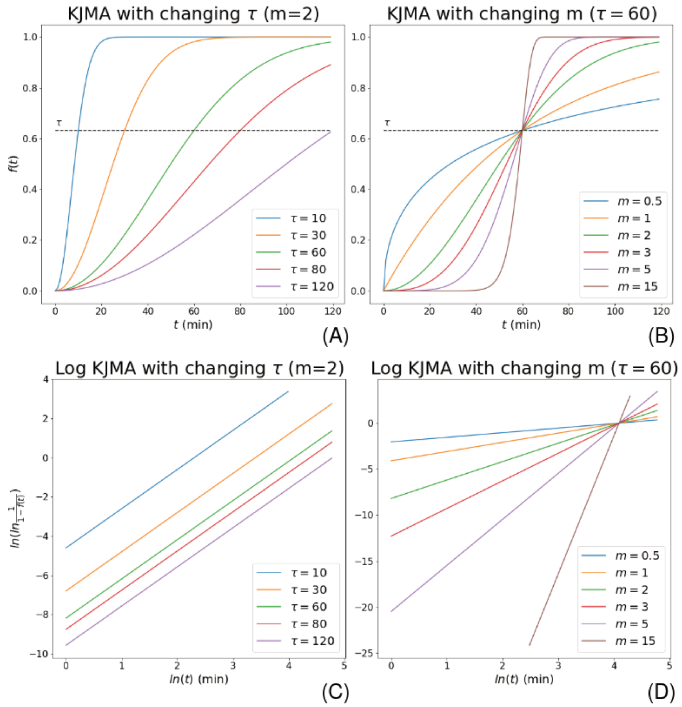
Model parameters with respect to cetromeres and telomeres.

S13 Fig



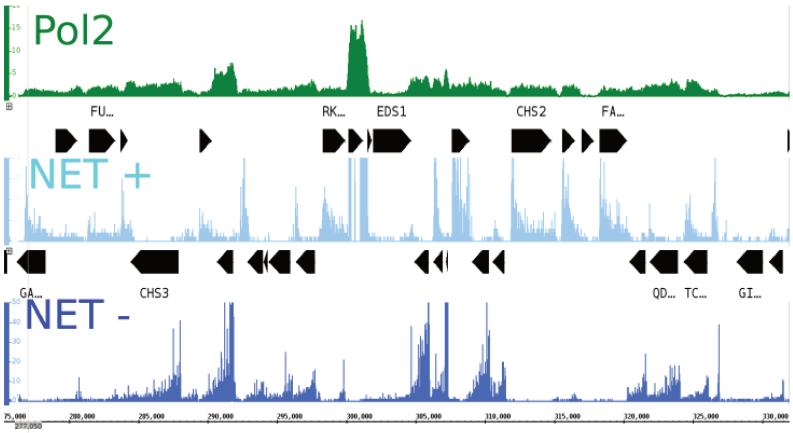
Comparison of the KJMA model with other functional descriptions.

S14 Fig



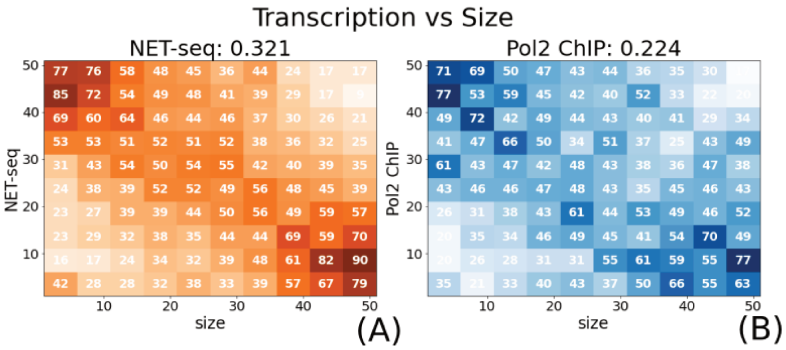
Example of the KJMA model.

S15 Fig



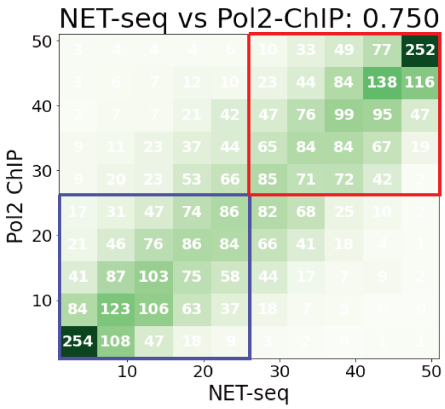
Example of sequencing data representing transcription rate.

S16 Fig



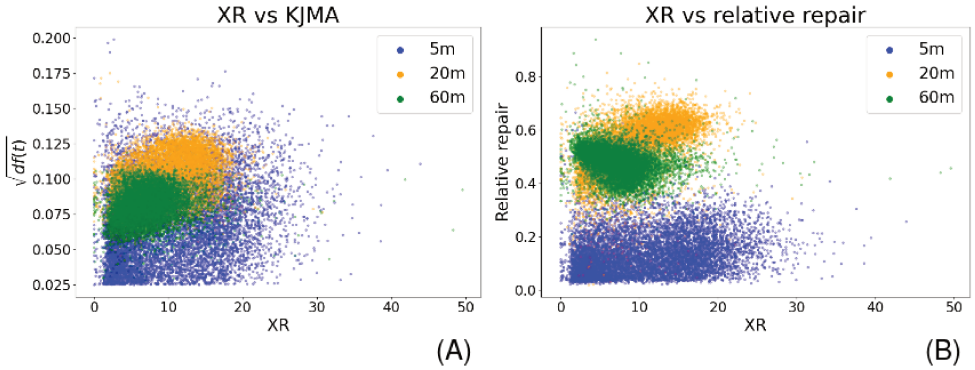
Correlation between size and transcription rate.

S17 Fig



Correlation between NET-seq and Pol2 ChIP-seq data.

S18 Fig



Model predictions with respect to XR-seq data in the *gene* setup.

Chapter 4

A Mean-Field Approach for Understanding DNA Repair

4.1 Introduction

The previous projects presented in Chapter 2 and 3 provided us with a top-down data-driven understanding about sequence accessibility and the temporal evolution of DNA repair. We can reasonably presume that WT strains—which possess Rsc8 and Chd1—exhibit nucleosome phasing independent of Pol II presence, at least for what can be measured using NGS data. We can therefore largely ignore the effects of chromatin accessibility during CPD removal in protein-coding regions that are repaired predominantly by TCR. This is supported by the KJMA model, which did not indicate a significant impact of nucleosome density along genes; yet repair parameters were correlated with transcription levels. Averaging the repair evolution in different coding regions revealed that lesion removal kinetics changed along the gene. Unfortunately, the KJMA model in Chapter 3 can only assess the temporal evolution for a given area, and it cannot describe the spatiotemporal dynamics. We hypothesise that the changing repair kinetics as a function of distance from the TSS is linked to protein interactions with the DNA, in particular Pol II movement. We aimed to assess this conjecture with a mechanistic bottom-up model. There have been various other studies that tried to recreate kinetics captured in imaging data of fluorescent-tagged proteins using ODEs (Politi et al. (2005); Luijsterburg et al. (2010)). However, a comprehensive representation of location-dependent TCR dynamics *in vivo* along the DNA is still lacking, and our understanding is predominantly based on *in vitro* studies.

In this chapter, we present the *traffic repair model*, which is derived from the simulation of vehicle movements. Similar approaches have been used to explain intracellular protein motion

(Hinsch et al. (2007)) and polymerisation during transcription and translation (Davis et al. (2014)). The model assumes that repair dynamics can be represented by an average behaviour—which is also referred to as mean-field approach—instead of taking stochastic specifics into account. By motivating repair in genes as a two step process during which the lesion is first found through protein movements and subsequently removed (i.e. combining excision an replacement in one step), we derive three ODEs that describe recognition by TCR and GGR as well as the repair of CPDs. The modelled TCR kinetics are driven by Pol II elongation, whereas GGR occurs through random association and dissociation without motion along the DNA. The formulas can be equally understood as master equations, which we introduced in Section 1.3.2. Parameters can be fitted to the sequencing data using Neural Ordinary Differential Equations (NODE) (Chen et al. (2018)), which apply a modified version of backpropagation using the adjoint sensitivity method (Pontryagin (1987)).

We show that a mean-field model can only make sensible predictions when property-dependent scaling of NGS signals is taken into account. By motivating a data normalisation based on single-cell estimates (Bucceri et al. (2006); Struhl (2007)), we demonstrate that repair dynamics measured over an entire cell culture are gene-specific. However, if not scaled appropriately, it is not possible to explain the observed CPD decrease, proving the importance of combining single-cell measurements with population-based data.

The method predicts a surprisingly strong and early influence of GGR along coding regions. Although the traffic model provides a good explanation for the average data evolution, it cannot provide a mechanistic cell-dependent explanation for why the presence of GGR is so important. Based on our presumption that lesions in single cells are rare, we hypothesised that it is pivotal to include cell-individual stochastic properties into our models to provide an in-depth understanding of the NER pathway (see Chapter 5).

In this project, I was leading in the model adaptation for NER as well as the implementation and validation. I was initially involved in the data production that included measurements for Rad4 as a representative of GGR. Unfortunately, Rad4-tagged strains exhibited a different repair behaviour to WT cells (Figure B.1). Due to the limited time and the lack of published data available online, we fell back to already produced but unpublished NGS signals from the laboratory.

4.2 Results

4.2.1 The Traffic Repair Model Explains DNA Repair in the Cell Population as an Average Cell Only With Appropriate Data Scaling

We presumed that molecular interactions with the DNA follow the pattern of association, dissociation, and one-dimensional movement (i.e. motion to the left or right) (Figure 4.1). The DNA itself is represented as a one-dimensional string with N discrete positions, and a protein can be present only at these locations. CPDs are removed either by TCR or GGR. Instead of modelling the intricate multistep process, we implemented the recognition dynamics through interactions with an abstract TCR or GGR protein (TCRP and GGRP, respectively). After detection, the lesion is removed with a TCR or GGR-specific rate. TCRP movements are motivated by Pol II elongation, i.e. association at the TSS, forward elongation along the coding region, and removal at the TTS. GGR interacts with the DNA substrate through random association or dissociation, yet without movement. When considering a single cell, we conjectured that there can be only one protein of the same kind at a given position. Consequently, association is blocked by already present proteins. Similarly, Pol II (or rather the abstract TCRP) temporarily stalls if blocked by another TCRP at the following position. When interacting with a CPD, TCRP stalls at the lesion site and cannot move forward. TCRP is removed after repair. We presume that TCRP and GGRP compete for lesion removal, and inhibit each other's presence at a given position. Both proteins repair CPDs with different rates, i.e. r_T and r_G for TCRP and GGRP, respectively. Motivated by our analysis of chromatin remodelers (Chapter 2) and correlation with other nuclear properties (Chapter 3), we presume that r_T and r_G are relatively independent of other processes and constant along the entire gene. Therefore, dynamics are exclusively driven by protein movements.

As mentioned before, the available NGS signals represent the state of an ensemble of cells. Therefore, single-cell dynamics are only observable by measurements over an entire population. We assume that sequencing data reasonably describe the state of an average cell. We extended the explained dynamics such that a state $s(P, x, t)$ of a property P (i.e. TCRP, GGRP, or CPD) symbolises the ratio of the cell culture possessing P at time t and position x. This can be also understood as the probability of presence when considering the frequentist point of view. Changes during a time step dt can only happen in the fraction of cells where such a transition is possible. For example, movement of TCRP from x to $x + 1$ can only occur in those cells that possess TCRP at x *and* a free position at $x + 1$. These dynamics can be summarised in three equations, which we present and explain in detail in the following.

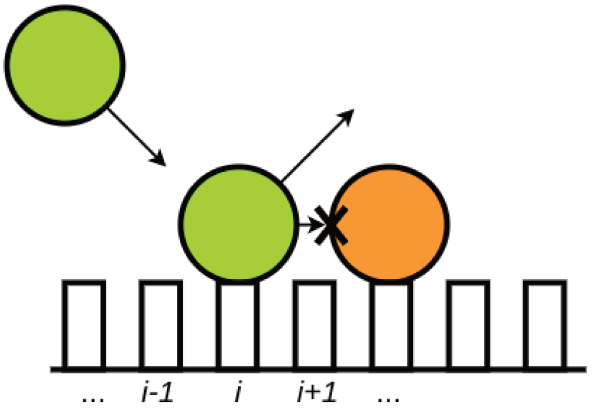


Figure 4.1: **Schematic explanation of the traffic repair model.** The DNA is structured into discrete compartments (e.g. bins $i-1$, i, $i+1$...). TCRP (green) can associate, dissociate, and move along the DNA in the direction of transcription. However, GGRP (orange) and TCRP inhibit each other (cross), such that DNA-protein interactions become less likely at positions where the other is already present.

When TCRP or GGRP are present at a lesion site, they remove the CPD at a rate r_T and r_G, respectively. When considering constant protein levels, more repair should be observable when more lesions are present, as it becomes more likely that both protein and damage can be found in the same cell. Similarly, the larger the TCRP or GGRP levels, the more likely it is to observe repair. The observed damage decrease in dt can be expressed as

$$\frac{\partial s(C,x,t)}{\partial t} = -s(C,x,t)\left(r_T s(T,x,t) + r_G s(G,x,t)\right). \tag{4.1}$$

$s(C,x,t)$, $s(T,x,t)$, and $s(G,x,t)$ are the levels at time t and position x of CPDs, TCRP, and GGRP, respectively.

Association of GGRP is blocked by already present GGRP and TCRP, and it can only occur at free positions with rate k_G^+. Dissociation happens at sites where GGRP is present at rate k_G^-. Dynamics are independent of CPDs, so that DNA-GGRP interactions can be described through

$$\frac{\partial s(G,x,t)}{\partial t} = k_G^+\left(1-s(G,x,t)\right)\left(1-s(T,x,t)\right) - k_G^- s(G,x,t). \tag{4.2}$$

TCRP dynamics are the most complex ones, as they involve movement along the DNA at a rate m_T, which we assumed to be constant. Due to the supposed competition between TCRP and GGRP, movement can only occur where the following position is free of either protein. TCRP is stalled once it colocalises with a DNA lesion. Moreover, to mimic nuclear Pol II dynamics, we presume that TCRP can only associate to the beginning of the gene with constant rate k_T^+ (i.e. $k_T^+(x) = k_T^+$ if $x < 0.5N$, and 0 otherwise), and it dissociates exclusively from the end of the gene at a constant rate k_T^- (i.e. $k_T^-(x) = k_T^-$ if $x > 0.5N$, and 0 otherwise). TCRP dissociates after lesion removal from the damage

site. This can be summarised as follows:

$$\frac{\partial s(T,x,t)}{\partial t} = (1-s(T,x,t))\,(1-s(G,x,t))\left(k_T^+(x)\,(1-s(C,x,t)) + m_T s(T,x-1,t)\,(1-s(C,x-1,t))\right)$$
$$- s(T,x,t)\left(m_T\,(1-s(T,x+1,t))\,(1-s(G,x+1,t))\,(1-s(C,x,t)) + r_T s(C,x,t) + k_T^-(x)\right) \tag{4.3}$$

The model behaviour was assessed for each parameter individually using the Sobol variance sensitivity analysis (Figure 4.2). It evaluates the effect of a parameter on the functional output (Sobol (1990)). First-order Sobol indices show that TCRP associates close to the TSS (k_T^+) and dissociates when approaching the TTS (k_T^-). GGRP can associate and dissociate freely along the entire gene (k_G^+ and k_G^-, respectively). Both TCRP and GGRP can repair uniformly along the entire transcribed region (r_T and r_G), making the dynamics dependent on TCRP motion and the interaction between TCRP and GGRP alone. The effect of TCRP movement (m_T) is strongest at the TSS where the protein also associates, and it loses importance towards the TTS. CPD levels are particularly effected at the centre. This is the most likely position where TCRP can encounter damage by motion, as TCRP dissociates thereafter. As TCRP is released after repairing a lesion, the effect of repair on TCRP occupancy is particularly strong close to the TSS, where it would not dissociate otherwise. Higher-order Sobol indices do not exhibit strong changes for transitive effects, with the exception of k_G^+ and k_G^-. As the presence of GGRP inhibits TCRP association and motion, it affects particularly TCRP levels closer to the TSS. The variance sensitivity analysis proves that the model works as intended, and that the dynamics are correctly represented by the formulas (Eq 4.1, 4.2, and 4.3).

Sequencing data (i.e. Pol II ChIP-seq and CPD-seq data, see Appendix B) was averaged into 5 bins along the gene body (i.e. $N = 5$), making it therefore independent of the actual gene size. We left GGRP levels as a hidden variable to account for missing repair that cannot be described by TCR alone. As in our previous publication (Zeitler et al. (2022)), we considered 1878 transcribed regions that exhibited a stark CPD decrease which presumably stems predominantly from TCR. Start and end sites were taken from Park et al. (2014). Unfortunately, publicly available sequencing signals that could be used to capture the dynamics described in the model could not be found. Consequently, data were produced lab-internally and are currently unpublished (no replicates available). They contain four time points for non-strand-specific Pol II ChIP-seq and CPD-seq before and right after irradiation ($t_\emptyset$ and t_0, respectively); 8 minutes after irradiation to measure immediate effects (t_8); and 38 minutes after irradiation (t_{38}) for assessing later changes. As lesion removal kinetics after recognition are supposed to take between 3-10 minutes *in vitro* (Erixon and Ahnström (1979)), we presumed that

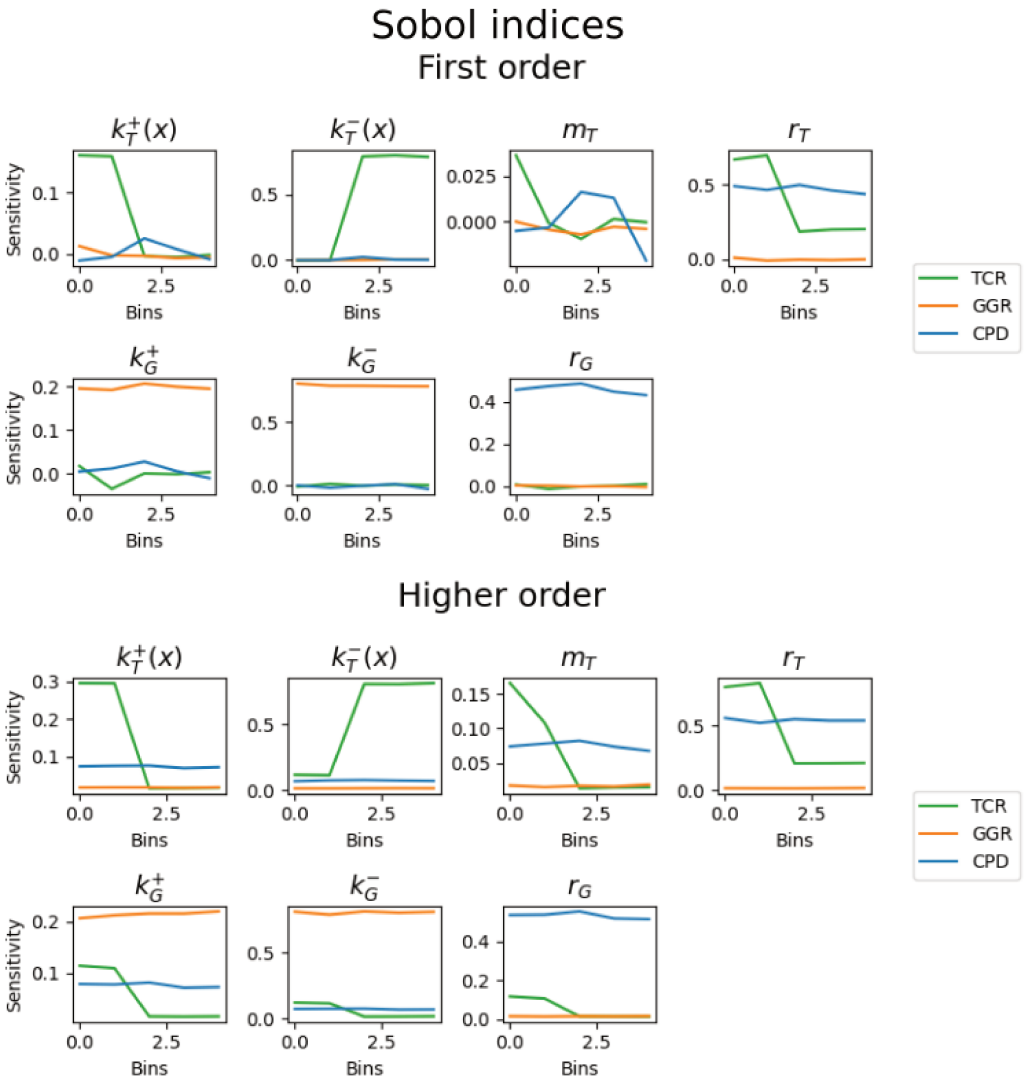


Figure 4.2: **Sobol sensitivity indices.** The variance sensitivity analysis indicates that the model is working as expected, and that the formulas Eq 4.1, 4.2, and 4.3 capture correctly the intended behaviour, as measured by the Sobol indices for first (upper plot) and higher order interactions (lower plot). First order indicates how the prediction is directly affected by the parameter, whereas higher order influences take into account transitive effects. The TSS is given at Bin 0 (left border of each plot), whereas the TTS is located after Bin 4 (right border of each plot). The y-axis displays the Sobol index, which is a measurement for the fraction of explained variance (Subsection 4.4.4). Green, orange, and blue lines show the Sobol indices per bin for TCRP, GGRP, and CPD levels, respectively. k_T^+, k_T^-, m_T, and r_T denote the association and dissociation rate, movement and repair speed of TCRP. k_G^+, k_G^-, and r_G are the association, dissociation and repair rate of GGRP.

CPD levels at t_8 follow repair dynamics of Pol II at t_0, and we made use of these data points as initial reference when the process is commencing. It is difficult to compare the actual NGS signal amplitude between CPDs and Pol II, as they might represent different numbers of cells. Initially, we intended to remove any bias by scaling both data sets individually such that the largest value of all time points was set to 1. With this approach, we aimed to take specifics of the repair procedure into account, whilst averaging out noise through a binning approach.

Starting from the initial distribution (i.e. the scaled Pol II ChIP-seq for TCRP at t_0, the scaled CPD-seq at t_8 for damage, and no associated GGRP), the goal was to predict the sequencing data at t_{38}. Parameters of Eq 4.1 - 4.3 were fitted to the sequencing data using NODE (Chen et al. (2018); Chen (2018)). We assumed that the CPD decrease should be fully captured by the model, but TCRP dynamics can slightly differ from the real Pol II distribution. We therefore implemented the loss computation as a weighted Mean Square Error (MSE)

$$L(Y;D) = \sum_{P\in\{C,T\}} w_P \frac{1}{N}\sum_{x=0}^{N}(Y(P,x) - D(P,x))^2. \tag{4.4}$$

Here, Y and D denote the prediction and data, respectively. $P \in \{C,T\}$ indicates CPD or TCRP prediction, or equivalently CPD or Pol II sequencing data. w_P is the error weight, which was set to $w_C = 2$ and $w_T = 1$ to encode our larger confidence into the captured repair kinetics. Indeed, L decreases over 200 training iterations (Fig 4.3(A)), and the model finds reasonable estimates given the data.

It is difficult to evaluate the model performance in the lack of a comparable baseline. Instead, we measured significance by comparing the predicted Pol II and CPD distributions to the results produced by random parameters. The latter were sampled within the range of the trained parameters over all genes. We hence assessed whether coding regions follow the same repair dynamics, or whether the mechanism captured by the model is gene-specific. In the former case, a significance test would not be able to discriminate between the prediction of the trained or random parameters. We therefore expected that model projections for CPDs and TCRP are either both significant or both insignificant. We estimated the prediction interval (PI) of a gamma distribution given the MSE between the data and 500 random predictions (Eq 4.4). The trained parameters were said to be significant if the prediction error was smaller than the lower bound of a 90% PI (which is equal to a 5% probability of yielding an error lower than a random model). Surprisingly, despite TCRP dynamics being highly significant, more than 90% of the CPD-seq data could be as well described using random parameters (Figure 4.3(B)). This indicates that whilst the movement and presence of TCRP changes substantially between genes, these kinetics do not translate into significant CPD repair. This is contrary to our assumption that spatiotemporal lesion removal kinetics are particularly linked to Pol II movement.

We were wondering why the TCRP predictions were gene-specific, whereas CPD removal was not. We proposed that inducing DNA damage is a rare event and significantly less present than Pol II, such that different NGS levels need to be taken into account. Indeed, it has been reported that irradiating living *Saccharomyces cerevisiae* yeast cells with 100 J/m^2 UVC induces ≈0.2 CPDs per

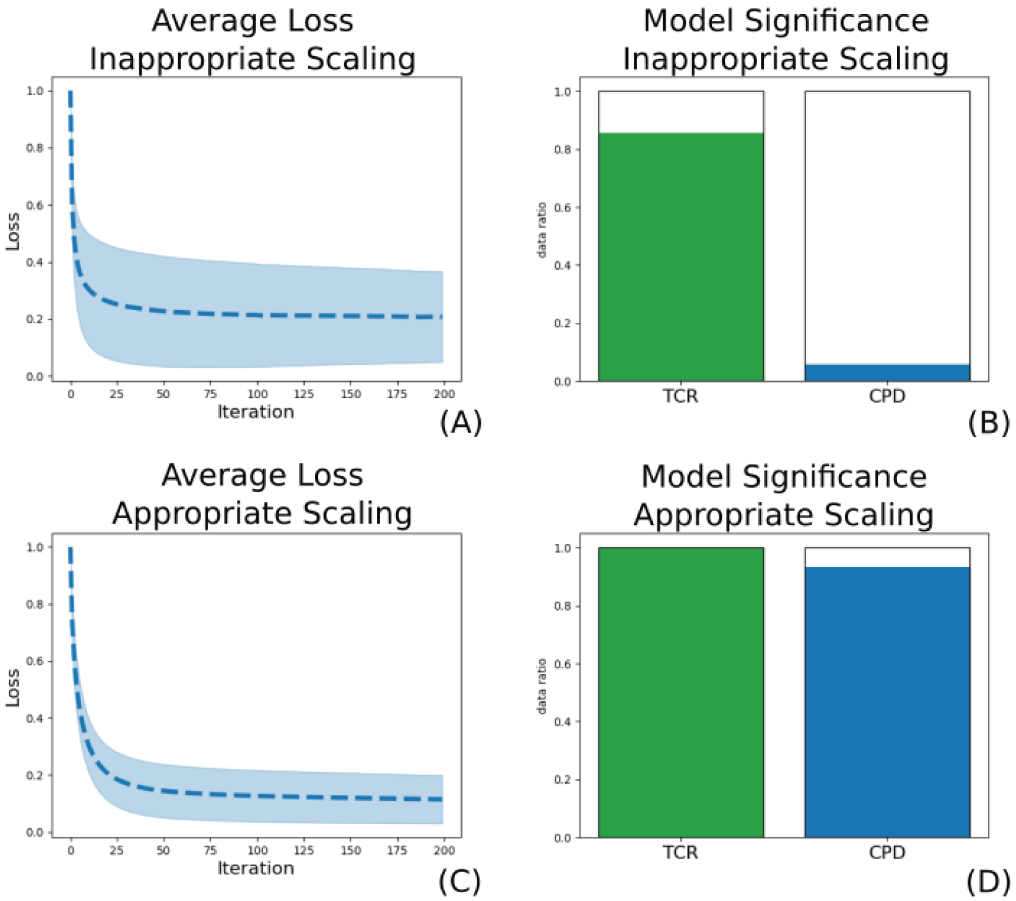


Figure 4.3: **Loss and significance of the traffic repair model.** (A) Using the NODE framework, we can indeed optimise the model parameters such that the average loss over all approximated regions decreases and eventually plateaus. This is even true for inappropriate data scaling. The dashed line gives the mean, the shaded area displays the standard deviation. The loss was rescaled such that the maximum error per estimation was set to 1. (B) When comparing the accuracy of the estimated parameters with the prediction using random parameters, we see that only the TCRP dynamics (green bar) can be significantly described (≈90%) when the data is scaled inadequately. However, the observed CPD decrease in the data could be equally well described with random parameters. Therefore, significant TCRP dynamics do not translate to DNA repair. This is contrary to our assumption that protein dynamics translate into repair. (C) When the data is appropriately scaled (Appendix B.9), the average error decreases to slightly lower values than for inadequate scaling, and the standard deviation is substantially reduced. This indicates that NODE finds parameters that approximate the data. (D) The model predictions become highly significant with respect to random parameter sampling over all genes for both CPDs and TCRP when considering appropriate scaling. This indicates gene-specific repair.

kilo base pair (kb) (Bucceri et al. (2006)). In other words, only 20% of the cells contain a single lesion in an average gene of 1000 bp, and the probability of sampling a cell that possesses damage at a given position is ≈0.01% (Appendix B.9). Within an entire cell culture in exponential growth phase with approximately 10 million cells, that results in around 1000 cells that may contain damage at a given location. On the other hand, presence of Pol II at a single gene might be larger, and varies depending on transcription levels and gene size. Zenklusen et al. (2008) demonstrated changing

presence depending on size, with some genes exhibiting a median of 2 transcribing Pol II. It is estimated that there are ≈20,000 Pol II molecules in *Saccharomyces cerevisiae*, of which roughly 60% are engaged in elongation (Struhl (2007)). By ignoring actual gene size and location-specific differences along the 12 Mega base pair (Mb) yeast genome, this results in approximately 1 Pol II complex per kb that is currently moving. We rescaled the data to account for Pol II and CPD-specific sequencing amplitudes (Appendix B.9) and repeated the parameter estimation. As the contribution of the damage prediction to the MSE (Eq 4.4) is now much lower (since the signal amplitude is weaker for CPD than for Pol II data), we increased the weight to $w_C = 10$. Once again, NODE improved model predictions and plateaued before training finished after 200 iterations (Figure 4.3(C)). Indeed, both predictions for TCR and CPD became significant when including appropriate scaling (Figure 4.3(D)). This result indicates that the observed average repair kinetics over the entire cell culture are location-dependent, and they are not solely conditional on the initial distribution of TCRP (i.e. Pol II) and CPDs. Similarly, the significance test suggests that spatiotemporal repair dynamics along the gene can be indeed explained through DNA-protein interactions, particularly Pol II movement. As this conclusion is not possible without accounting for adequate scaling (Appendix B.9), we demonstrate equally that presence of damage is substantially more sporadic than Pol II occupancy.

4.2.2 Understanding Genome-Wide Repair

To shine light onto the observation that population-wide repair dynamics are location-specific, we tested for functional relationships between parameter values, and how they are changing among genes. However, parameters might vary differently between transcribed regions in varying scales. To account for different correlation behaviours, ranges, and parameter influences, we reduced the dimensionality by applying a conventional Principal Component Analysis (PCA) (Section 4.4.5). We subsequently repeated the PCA when setting all parameter values to zero but one. We calculated the cosine similarity between fully transformed parameter set and single parameter transformation. The cosine similarity quantifies the angle between the two transformations and ranges between -1 and 1. By determining the similarity for every parameter to the full transformation, we can represent them within the same range and compare their functional relationship without any bias driven by the data (Figure 4.4).

Interestingly, many parameters exhibit a mutual interdependence along lines or ellipsoidal shapes (Figure 4.5). This is also clear for the correlation between r_G and r_T, which indicates a functional relationship between the two repair rates to describe CPD removal on a population scale. The parameter correlation explains why repair dynamics are predicted to be gene-specific, as the model

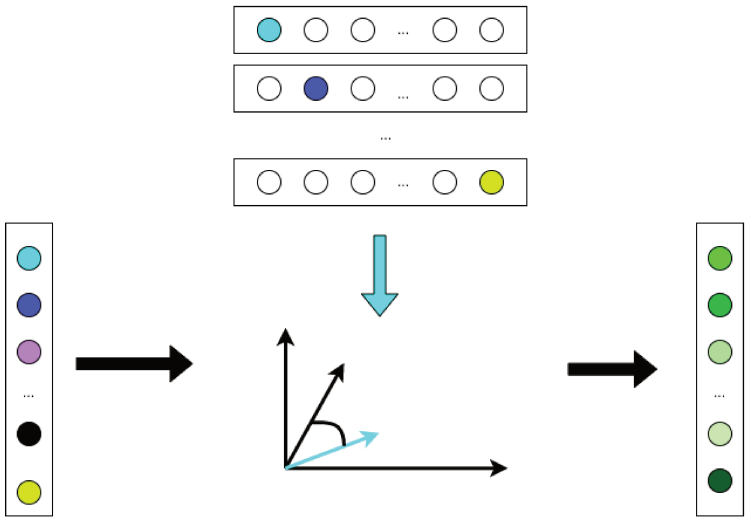


Figure 4.4: **Transforming model parameters to similar scales and behaviours.** All parameters have different values, ranges, and behaviours between genes. We determine a PCA transformation using the parameter values over all genes (left column and black arrow). We apply the same transformation to each parameter individually by setting all parameters per gene to 0 but one (top columns and turquoise arrow). This evaluates how each individual value is affecting the transformation. By calculating the cosine similarity between the two transformed vectors (i.e. using the full parameter vector in black and the vector containing only a single value in turquoise), we rescale every single parameter to similar ranges. The transformed representation (right in green) behaves similarly between genes, which simplifies the comparison.

expects that they change with respect to each other and cannot be drawn randomly. The histograms for TCR and GGR-specific lesion removal rates also reveal that they are similarly distributed with respect to their influence, making them equally important to explain the data. We grouped genes based on their PCA-dependent cosine similarities using a multivariate Gaussian classifier. Group 1 (green, $\approx 60\%$) is uniformly distributed or remains largely in the lower or left part for each parameter correlation plot, whereas Group 2 (yellow, $\approx 40\%$) seems to be predominantly present in the upper or right half. The restriction of Group 2 parameters to the top right corner is especially visible with respect to the GGR repair rate (r_G) and the TCR association rate (k_T^+). This could indicate a different GGR-related behaviour or distinct TCRP dynamics for a subset of genes. We can show that the observed phenomenon is particularly related to the temporal evolution of Pol II profiles (Figure 4.6). Whilst Group 1 remains fairly constant in Pol II occupancy levels and exhibits a shift towards the TSS at t_{38}, Group 2 shows a signal decrease over time, indicating dissociating Pol II. Therefore, the data in Group 2 is more in line with the implemented TCR dynamics during which Pol II dissociates, whilst the Pol II distribution in Group 1 might be more reliant on a large presence of GGRP with which TCRP must interact in order to explain the NGS data.

Since the two groups are seemingly linked to the Pol II profile along coding regions, we evaluated the transcription levels of the genes they contain. To provide a fair comparison, we randomly sampled

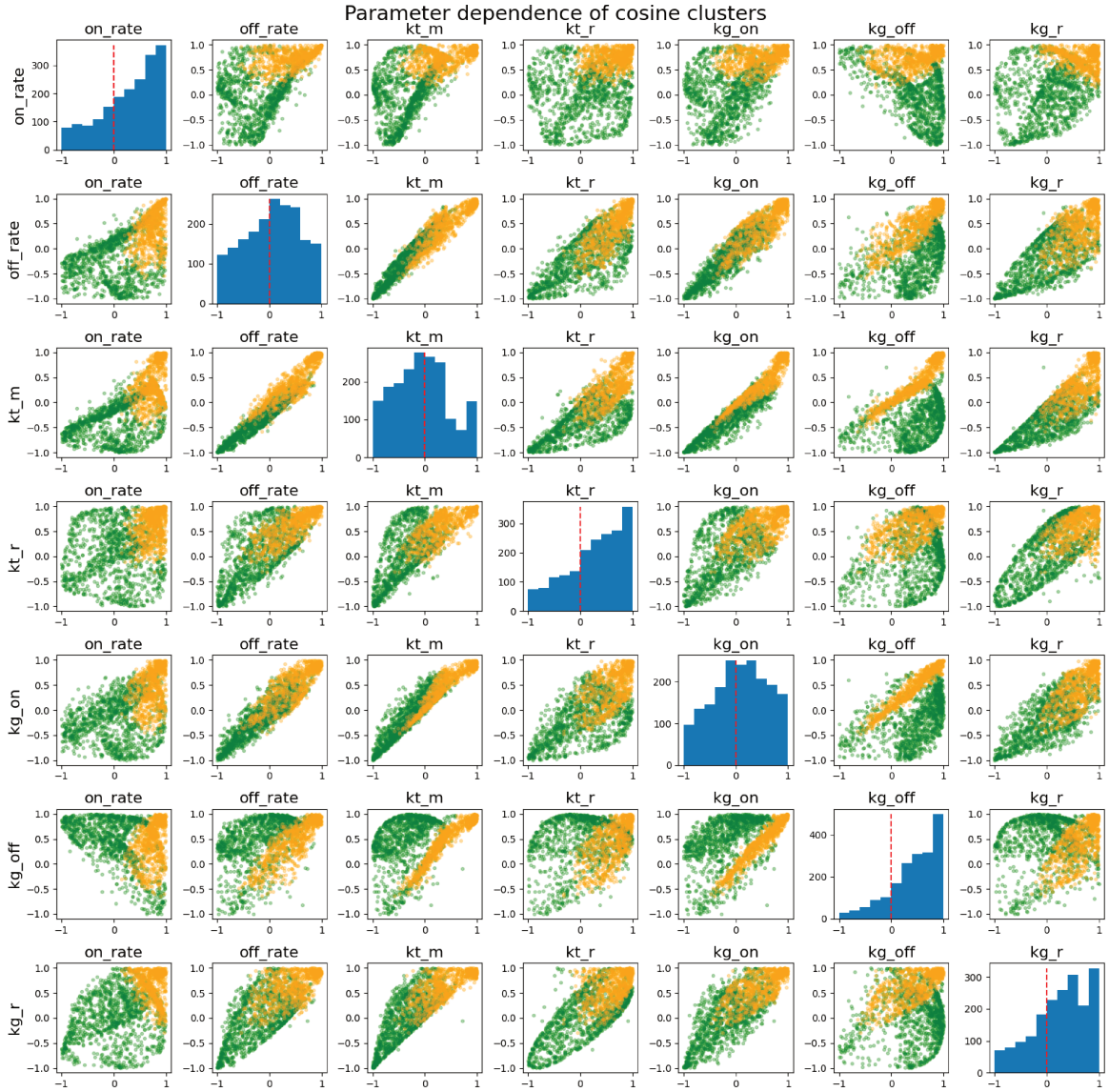


Figure 4.5: **Assessing the function relationship of the repair parameters.** We measured the cosine similarity of the transformed parameters as schematically explained in Figure 4.4. By comparing the similarity pairwise between parameters, we can investigate how parameters change over all genes with respect to each other. Each row and column represent a parameter, a single plot displays their pairwise comparison. The figures on the diagonal show the histogram distribution of the cosine similarities for a single parameter. The red dashed line in the histogram indicates a cosine value of zero, meaning that the transformation of a single parameter was orthogonal to the transformation of all parameters. The distribution of the pairwise cosine similarities along lines and ellipsoids indicates that most parameters are functionally dependent of each other. `on_rate` and `off_rate` are k_T^+ and k_T^-, respectively. `kt_m` represents m_T. `kg_on` and `kg_off` are the association and dissociation parameters for GGR (k_G^+ and k_G^-), whereas `kt_r` and `kg_r` represent the repair rates r_T and r_G. Genes were clustered based on the cosine values for each parameter. Group 1 is given in green dots, whereas Group 2 is displayed in yellow.

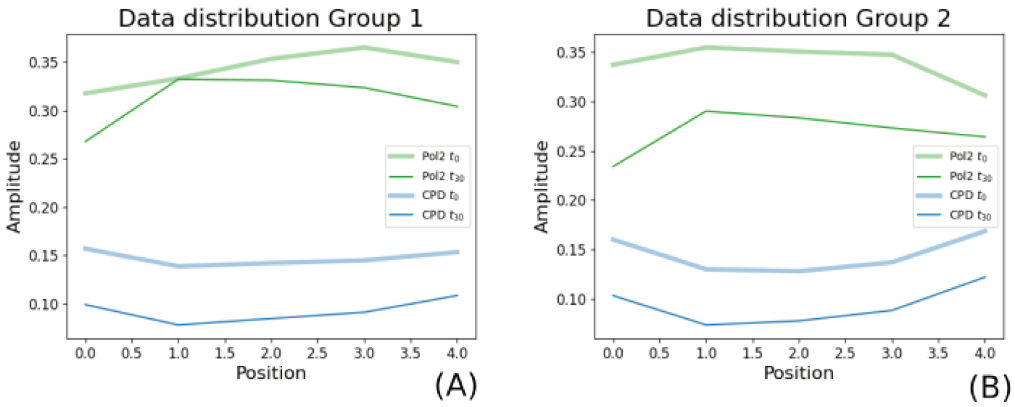


Figure 4.6: **Higher GGR rates indicate different Pol II distributions.** Green lines give Pol II data, whereas blue lines show the CPD-seq distributions. The bold lines are the initial condition, the thin lines display the distributions after 30 minutes of repair. (A) Pol II signals in Group 1 remain relatively constant over time with a slight shift towards the beginning of the gene. Implemented TCR kinetics cannot fully explain the CPD decrease, and genes might be more reliant on the presence of GGRP. (B) Pol II distributions decrease over time, which is more expected with the implemented TCR dynamics. Nonetheless, this does not exclude additional repair by GGR.

500 genes for each group to test their significance. Despite the fact that the histograms before UV treatment and 8 minutes after are significantly different (p-value of a Mann-Whitney-U test $5 \times 10^{-4}\%$, $2 \times 10^{-3}\%$, and 7% for before UV, 8 minutes, and 38 minutes, respectively), they show a large overlap at any time point (Figure 4.7(A)). We find it difficult to explain why two genes with similar expression levels should be in different repair groups. Moreover, the average sequencing data distribution along the coding region differs particularly with respect to t_{38}, whereas the average transcription levels between both groups are not significantly different at this time point. We obtain similar results for the gene size, where the group-dependent distributions are again significantly different (Mann-Whitney-U p-value $5 \times 10^{-4}\%$) but exhibit overall almost equal ranges for the transcription unit length (Figure 4.7(B)). Similarly, both groups show largely overlapping values when comparing initial CPD levels and the damage distribution after 30 minutes, where only the former is significant (Mann-Whitney-U p-value 0.3% and 9% for t_0 and t_{30}, Figure 4.7(C)). In our opinion, it is impossible to establish a clear determining link between the DNA repair group and another genomic property. Surprisingly, an ontology analysis with respect to the components the genes are involved in revealed that Group 2 is associated with ribosomal structure ($\approx 20\%$ of genes in Group 2 vs $2-3\%$ in Group 1). When evaluating the gene ontology with respect to the function, we find that Group 2 is predominantly involved in biosynthetic processes (e.g. transcription and translation, $\approx 50\%$), whereas Group 1 contains many genes related to protein localisation and protein transport ($\approx 30\%$). We find it difficult to identify their particular role during DNA repair or the UV stress response. Nonetheless, the fact that the different Pol II distributions are linked to non-identical repair dynamics that could be related

to distinct gene functions could indicate that CPD repair is globally orchestrated.

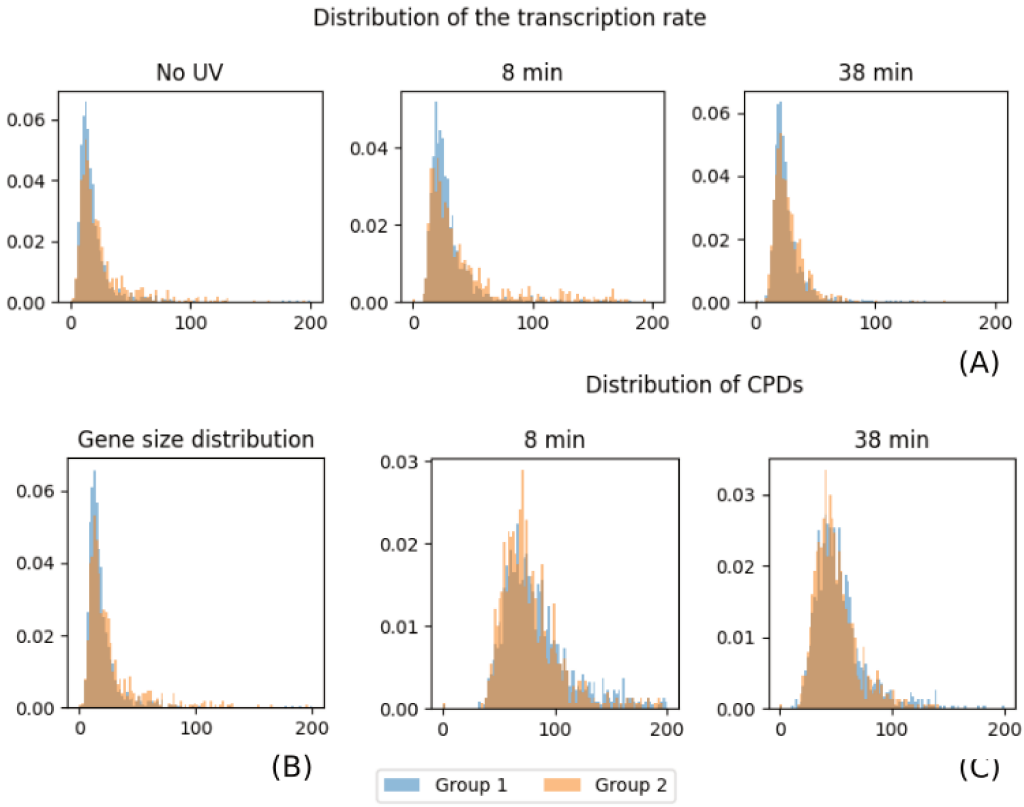


Figure 4.7: **Histograms of genomic properties with respect to the repair groups.** Figures (A), (B), and (C) show the distribution of transcription rate, transcription unit length, and CPD levels with respect to the two gene groups and time points if applicable. Blue bars display the histograms for Group 1, and orange bars show the distributions for Group 2. The differences in expression levels (No UV and after 8 minutes), gene size, and initial CPD presence are significant (measured using a Mann-Whitney-U test). However, there is a large overlap for all histogram pairs, and it becomes difficult to explain why genes that are comparable with respect to a genomic property are in different repair groups. The distributions are surprisingly similar in all cases.

4.2.3 Understanding Gene-Specific Repair

In order to obtain a better understanding of the different lesion removal kinetics between the parameter groups, we simulated the repair process per coding region with varying initial states. In all cases, CPD levels were forced to zero over time, which reflects the fact that no new lesions can be induced, and repair parameters can only be positive. Genes in Group 1 (exemplified for YAL020C in Figure 4.8), indicated high GGRP levels at early time points (e.g. after 15-20 minutes). By testing the repair kinetics over changing initial conditions which we randomly sampled, we can show that the important role for GGR dynamics remains conserved (Figure 4.8)(A)). The early presence of GGRP is surprising since we considered exclusively regions that exhibited strong early repair, which we presume to come from TCR dynamics. To improve our understanding of how the data are approximated during

training, we visualised predicted repair dynamics with the parameters at different learning iterations. Indeed, GGRP presence gained great importance during ongoing parameter fitting, and the best values were obtained for YAL020C when repair after 20 minutes was almost exclusively determined by GGR (Figures 4.8(B, C)). The protein presence equally translated into lesion removal, as TCRP and GGRP repair rates were identical. Nevertheless, the predicted CPD and TCRP levels at t_{38} were very close to the data, indicating that the fitting approach finds sensible parameter values (weighted MSE ≈ 0.0001, Figures 4.8(D, E)). It should be again emphasised that there were no data available to optimise GGR kinetics, and GGRP occupancy levels were treated as an unknown hidden variable that needs to account for the missing aspects in the data that are not explained by TCRP alone. The results could indicate that GGR dynamics might be much more important for repair in coding regions than previously appreciated.

Whilst the genes in Group 1 were largely dependent on GGR, repair in some coding regions of Group 2 could be achieved almost exclusively by TCR (exemplified for YAL053W in Figure 4.9). Varying initial conditions all led to a reduction of GGR, suggesting that the trained TCR kinetics are sufficient for repair, independently of the initial Pol II and CPD distribution (Figure 4.9(A)). Whilst the parameters would allow GGR to be happening during later time points when training commenced (Figure 4.9(B)), the algorithm forces GGRP association to zero during learning, such that it cannot interfere with Pol II (Figure 4.9(C)). Indeed, the approach can find once again a reasonably good fit, although the weighted MSE is admittedly higher than for YAL020C (≈ 0.0011, Figures 4.9(D, E)). The analysis reveals that the implemented TCR dynamics can better explain CPD removal in Group 2 (although it should be mentioned that this does not exclude an impact of GGR in other genes within the same group).

4.3 Discussion

In this work, we developed and implemented a mean-field approximation over the entire cell population to explain the damage decrease observable in CPD-seq data. We incorporated competing TCR and GGR by modelling damage recognition through an abstract TCR and GGR protein. TCRP dynamics are motivated by Pol II kinetics, and GGRP associates and dissociates randomly to and from the DNA. Both proteins repair CPDs with individual rates varying between genes but which are constant along the same transcribed region. We show that adequate data scaling is highly important when comparing the distribution of DNA lesions and repair proteins. Whilst TCRP dynamics were gene-specific using both data normalisations, repair represented in the CPD-seq signals could only be well explained when including appropriate scaling. By using reported CPD rates (Bucceri et al.

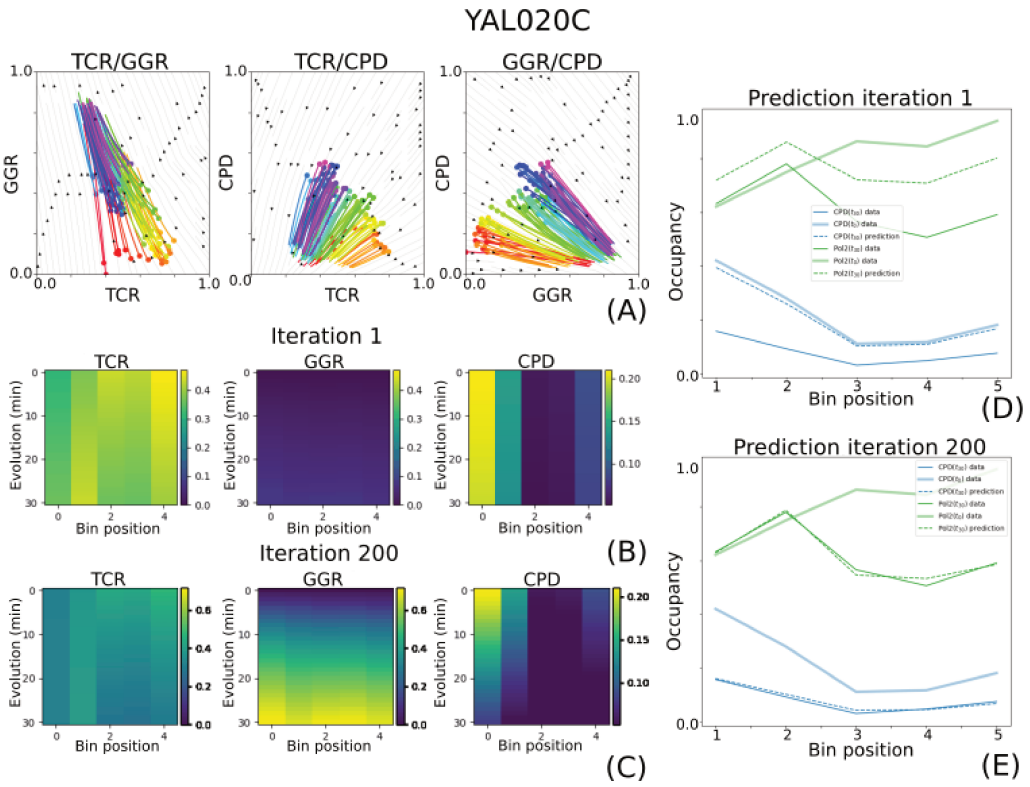


Figure 4.8: **Repair traffic dynamics exemplified at gene YAL020C (Group 1).** (A) The phase plane shows how variables in our system (i.e. TCRP, GGRP, and CPD presence) change with respect to each other over time and varying initial conditions. x and y-axis give the average occupancy levels over all bins. Each line represents a trajectory over 30 minutes, where the dot marks the initial state. From there, the line displays the change of TCRP, GGRP, and CPD values over time. The used colour-coding helps distinguishing different simulations. The phase plane of the parameters estimated for gene YAL020C suggests that GGR is essential to explain CPD removal over time. Despite the varying initial conditions in which GGRP presence is close to zero, occupancy levels rise to comparatively high values, whilst slightly decreasing TCRP levels or leaving them unaffected. (B) and (C) show the predicted dynamics at the first and last iteration of the training procedure. In each subplot, the x-axis represents the bin position, and the y-axis shows the evolution of TCRP, GGRP, or CPD levels over 30 minutes with the given parameters at the training iteration. Yellow indicates high levels, whereas dark blue mark low occupancy. Over time, the algorithm increases importance of GGR dynamics, as GGRP levels increase during parameter estimation. (D) and (E) display the prediction of TCRP and CPD values after 30 minutes at the first and last iteration of the training procedure. The bold shaded lines show the initial distribution. The dashed lines indicate the model prediction at t_{38}, and solid thin lines are the data at t_{38}. Green and blue show Pol II ChIP-seq and CPD-seq data, respectively. The learning approach finds sensible parameter values, such that the data distribution at t_{38} can be predicted given the initial occupancy levels.

(2006)) and Pol II occupancy values (Struhl (2007)) in *Saccharomyces cerevisiae* from the literature, we suggest that DNA damage might be much less present than Pol II along a gene, and we defined our normalisation factors accordingly. Indeed, when incorporating adequate scaling, the traffic repair model makes significant predictions for both TCRP presence and CPDs. This indicates that observed repair dynamics over an entire cell population are gene-specific and do not solely dependent on the initial distribution of lesions and repair proteins. Consequently, another and unknown factor

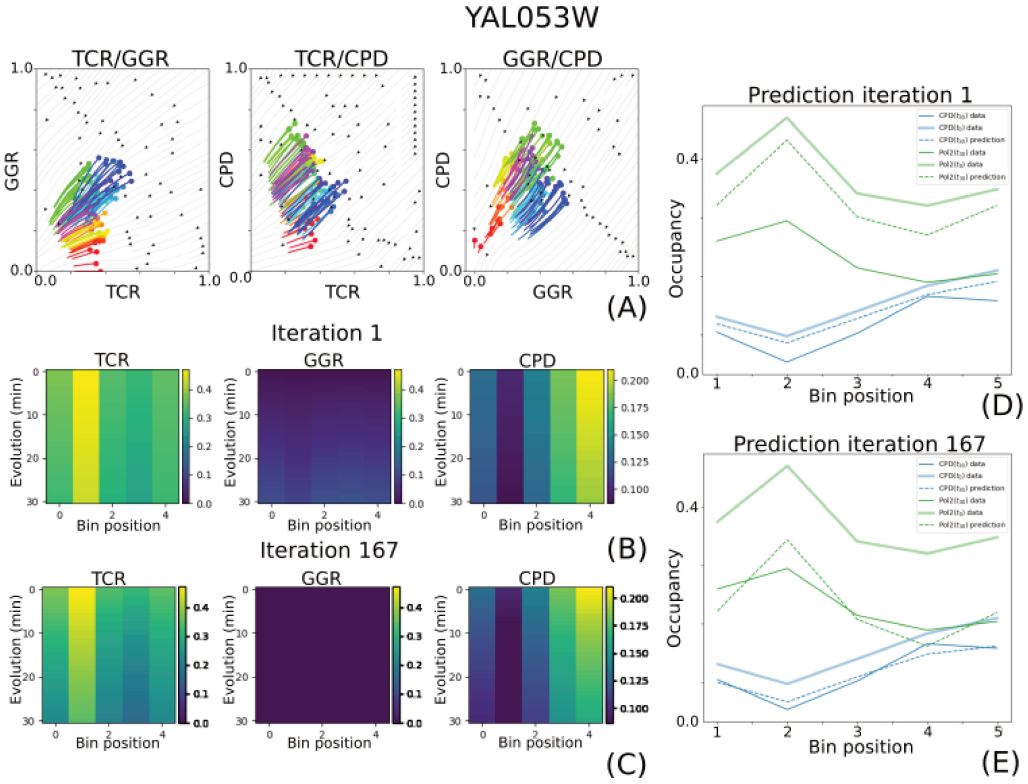


Figure 4.9: **Repair traffic dynamics exemplified at gene YAL053W (Group 2).** (A) The phase plane shows how variables in our system (i.e. TCRP, GGRP, and CPD presence) change with respect to each other over time and varying initial conditions. x and y-axis give the average occupancy levels over all bins. Each line represents a trajectory over 30 minutes, where the dot marks the initial state. From there, the line displays the change of TCRP, GGRP, and CPD values over time. The used colour-coding helps distinguishing different simulations. The plots indicate that the parameters of gene YAL053W force GGRP presence towards zero for various initial conditions. (B) and (C) shows the predicted dynamics at the first and last iteration of the training procedure. In each subplot, the x-axis represents the bin position, and the y-axis shows the evolution of TCRP, GGRP, or CPD levels over 30 minutes with the given parameters at the training iteration. Yellow indicates high levels, whereas dark blue mark low occupancy. The parameter approximation improves data fitting when forcing GGRP association to zero. (D) and (E) display the prediction of TCRP and CPD values after 30 minutes at the first and last iteration of the training procedure. The bold shaded lines show the initial distribution. The dashed lines indicate the model prediction at t_{38}, and solid thin lines are the data at t_{38}. Green and blue show Pol II ChIP-seq and CPD-seq data, respectively. The learning approach finds sensible parameter values, such that the data distribution at t_{38} can be predicted given the initial occupancy levels.

must influence lesion removal. We identified two gene groups with non-identical repair parameters. Group 1 ($\approx 60\%$) is predicted to be highly dependent on repair by GGRP whose dynamics are left as a hidden parameter. This is highly surprising, as we considered exclusively regions where we presume a strong impact of TCR. Group 2, on the other hand, can repair damages using exclusively TCR ($\approx 40\%$). A gene ontology analysis revealed that these coding regions are particularly involved in biosynthetic processes and ribosomal components. Unfortunately, the traffic repair model cannot provide a mechanistic understanding for why high GGRP levels might be necessary. Whilst this

could reflect solely the model's incapability to explain damage removal by TCR, it could equally indicate that GGR-related proteins might be earlier required than previously appreciated. In the following, we critically discuss the result and set it into context with existing studies.

Previously published repair models represent lesion removal using a set of ODEs. Luijsterburg et al. (2010) combines difference equations (i.e. time-discrete differential equations) for enzymatic steps with differential equations for repair intermediates. Indeed, their approach can explain microscopy data of fluorescent-tagged proteins and photobleaching. A similar method was proposed by Politi et al. (2005). The 6 equations that describe the evolution of transitional repair states were fitted using imaging data. By making them interdependent, they describe the repair steps as a sequence, where each stage is dependent on previous repair intermediates. Each differential equation contains a repair and a binding rate. We follow a similar method by modelling association and dissociation together with a damage-removal rate. Incorporating spatial information allows us to precisely define the recognition step mediated by Pol II. It needs to be said that we do not incorporate repair intermediates. However, as we do not possess any data for these compounds, we find incorporating more hidden variables unjustified. Politi et al. (2005) and Luijsterburg et al. (2010) report good agreement of their models with the data. We show that our training approach also strongly reduces the error, and predictions are significant. Surprisingly, GGR must be highly present in a subset of genes ($\approx 60\%$) to explain the observed CPD decrease in the sequencing signals. Whilst the mean-field dynamics find a good NGS approximation, they fall short when a mechanistic understanding is required, such as for explaining the necessary GGRP presence at some genes. It is therefore necessary to develop a model that takes cell-individual repair specifically into account.

Interaction between TCR and GGR was implemented as competition. This is justified by the fact that a single lesion can be detected only by one pathway but never both. Consequently, all cells whose lesions are repaired at a position x using GGR (represented by the presence of GGRP) do not remove the same damage using TCR (symbolised by the presence of TCRP). Although the real interaction in a biological cell might not be correctly incorporated, the observation over an entire cell culture should result seemingly in competition (presence of either protein but rarely both) due to the averaging and the mean-field approximation.

TCRP is stalled at the damage position after encountering a CPD site. Although Pol II must be physically backtracked to allow access to the lesion (Marteijn et al. (2014)), this is only considered to be a few base pairs, and it should remain within the same bin in our approach. After damage repair, our model removes TCRP from the DNA substrate. The actual fate of Pol II is admittedly not fully known, and it might be a combination of transcription continuation, removal, and degradation (Marteijn et al. (2014)). Nevertheless, some results propose a dominance of degradation or dis-

sociation, as this might be important for regulating the cellular stress response after UV exposure (Steurer and Marteijn (2017); Vidaković et al. (2020)). We find therefore that the modelling choice is in agreement with biological findings. Interestingly, we identify a subset of genes for which this requirement might not be fully met (Group 1), such that the observed repair can only be accounted for by a large presence of GGRP. This could possibly reveal that Pol II remains on the DNA in these regions, although it should be mentioned that it could equally indicate a large reliance on GGR. Without any further experimental validation, we find it justified to model Pol II dissociation after repair indiscriminately to all coding regions.

We applied the NODE framework (Chen (2018)) for finding reasonable parameter values of Eq 4.1 - 4.3. It was initially developed as a neural network architecture (Chen et al. (2018)), and therefore as a function approximation. However, we use it for the inverse problem, i.e. finding the parameters for a given set of ODEs that we suppose explain repair dynamics. We refrain from providing a detailed explanation of NODE and instead refer to Chen et al. (2018).

The traffic model itself was previously used to describe polymerisation of DNA and RNA (Davis et al. (2014)) and intracellular protein movements (Hinsch et al. (2007)). In particular, it has been applied for studying mRNA translation by ribosomes (Heinrich and Rapoport (1980)). The model was compared to non-spatial measurements for reaction rates. Chou (2003) used the traffic model to analyse mRNA-loop formation during translation. The study did not fit parameters to any data and instead compared the results qualitatively. Indeed, Pol II elongation movements—which are required for TCR—are similar to polymerisation and translation, which justifies the application of the method to explain CPD repair along genes. When accounting for appropriate scaling, we can find significant model predictions for almost all considered coding regions with a low error. The fact that random parameters—which were sampled over all estimated values—cannot describe the repair evolution suggests that parameters are substantially changing at different regions, making them dependent on another unknown factor. The model provides valuable insights into the repair evolution on a population scale.

To obtain a deeper understanding into the model behaviour, we evaluated the functional relationship between the parameters among genes. We applied a dimensionality reduction which was followed by a cosine similarity measurement between the transformed vectors of a single parameter and the full parameter set to rescale values into identical ranges in an unbiased way. This permits the comparison of different model parameters that behave heterogeneously over varying scales. PCA requires the data to be Gaussian distributed to maximise the mutual information between the transformation and the input. Nonetheless, it can be shown that PCA minimises the upper bound of the information loss if the noise in the data is more Gaussian than the underlying trend encoded in the

data (i.e. *signal* in information theory) (Geiger and Kubin (2013)). As we approximate all parameters to Gaussian distributions as closely as possible (Section 4.4.5), we presume a sufficient amount of information is preserved. Though, it should be emphasised that the revealed functional relationship is only indicative and cannot be seen as a proof due to the possible violations during the transformation.

It is difficult to explain the high importance of GGRP in our model (particularly for Group 1). Indeed, it could suggest that the hidden variable captures missing dynamics to explain observed lesion removal over the cell population; yet equally, it could point out that early GGR might be more important than previously appreciated. The significance test indicates gene-specific repair, and consequently, changing parameters for each coding region. Repair dynamics are therefore not exclusively dependent on the initial state (i.e. the distribution of damage and TCRP), and they change as a function of an unknown factor to explain repair in the cell culture. It should be stressed that the used data are non-strand-specific, and they represent TS and NTS of a transcribed region. The missing kinetics captured by the high importance of GGRP could stem from repair on the NTS. Nonetheless, the NTS does not exhibit spatially changing repair rates over the gene body, as indicated by Mao et al. (2016) and our own study (Zeitler et al. (2022)). Damage levels on the NTS should therefore uniformly raise non-strand-specific CPD-seq data at all time points without a spatial bias. Moreover, Group 1—for which GGRP is seemingly particularly important—does only differ from Group 2 with respect to the CPD distribution. As we assume Pol II presence on the NTS to be negligible, we surmise that considering both strands at the same time does not influence the overall conclusions.

Unfortunately, we are unable to suggest a biological property that renders repair gene-specific. It could be easily argued that the kinetics change with respect to the sequence accessibility, and lesion removal can be fully described when including the nucleosome distribution. However, we believe that the incorporation of MNase-seq data would not change this observation due to two reasons. Firstly, the results in Chapter 2 and 3 suggest that CPD repair in protein-coding regions is largely independent of nucleosome density and phasing. Secondly, as we divide the gene into 5 bins, each of them is assumed to contain at least one histone complex due to the highly regular positioning. The length of the coiled DNA together with the linker DNA accounts for $\approx$ 200bp. Each gene that is larger than 1000bp (i.e. approximately the average size of a gene in *Saccharomyces cerevisiae*) contains therefore at least one nucleosome per bin, and any possible effect related to sequence accessibility imposed by their positioning should be smoothed out. We can show that the majority of the considered genes is larger than 1000bp (Figure 4.10(A)). As we do not find a correlation between the two repair groups and gene size, we presume that the few genes that are smaller than 1000bp do not significantly impact the results.

Although we can phenomenologically explain why GGR might be more important in some genes,

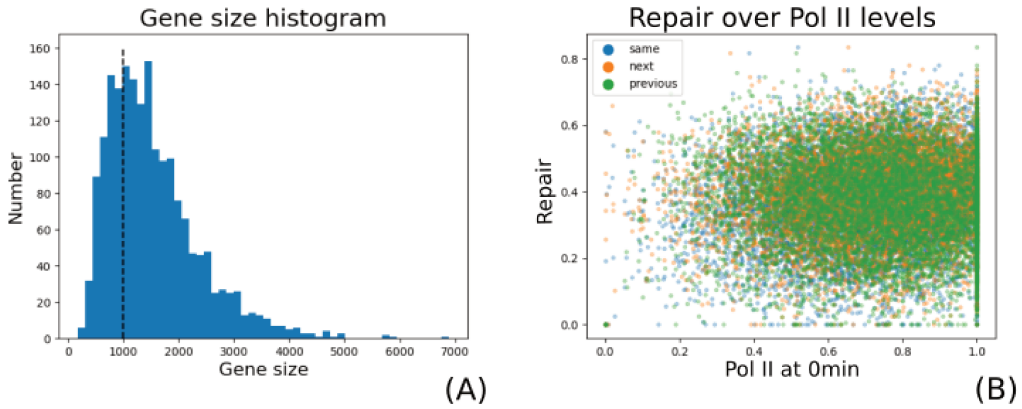


Figure 4.10: **Gene size and Pol II presence cannot explain repair rates.** (A) displays the histogram over all considered genes. The black dashed line shows the median. The considered genes are predominantly larger than 1000 bp (black dashed line). Therefore, each bin should contain in most cases at least one nucleosome, which should average out different dynamics due to chromatin packaging. (B) When plotting CPD repair over Pol II presence at t_0, there is clearly no correlation. This is even true when comparing the same bin (blue) as well as when considering movement by Pol II, since there is equally no link between repair at one bin x and the Pol II levels before (green) and after (orange).

the model fails to provide a mechanistic understanding. The mean-field model presumes that quantities collected over the entire cell culture are sufficiently large such that stochastic effects can be ignored. In other words, the sequencing signals symbolise an average cell, and the model can represent the mean behaviour in the population. As we presume that repair rates (i.e. r_T and r_G) are uniform along the entire gene, this implies that higher presence of TCRP results in more repair. This might be indeed the case for genes in Group 2. However, an analysis of the overall data suggests that this requirement is generally not fulfilled (Figure 4.10(B)). Nonetheless, the dynamics incorporated in Eq 4.1 - 4.3 are motivated by *in vitro* experiments, and it is reasonable to presume similar mechanisms in living cells. We hypothesise that cell-specific states need to be taken into account to explain repair mechanistically. The low presence of CPDs might require the incorporation of stochastic dynamics, as we derived that the probability of a damage site at a given position is $\approx$0.01%. Since yeast cells are single-cell organisms that repair lesions independently from each other, the CPD must be located after Pol II in the direction of transcription in order to be detected by TCR (Figure 4.11). However, if the presence of damage at a given position is a fluke, averaging ignores important information of positioning per cell. We believe that whilst mean-field dynamics provide a sensible phenomenological explanation, the sporadic presence of CPDs necessitates a stochastic and cell-specific model to make mechanistic conclusions about the process.

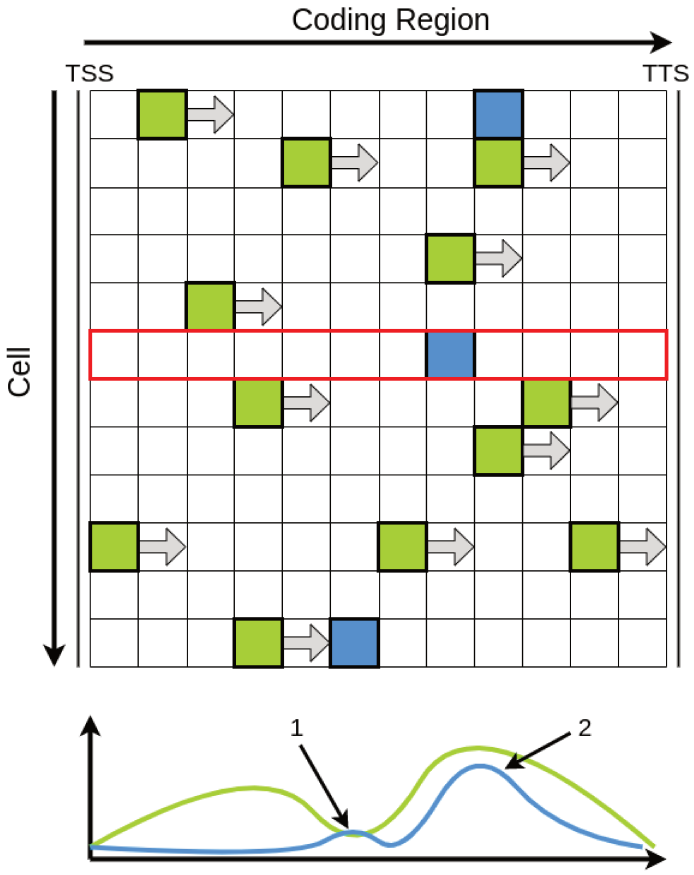


Figure 4.11: **Stochastic effects during repair need to be taken into account (cartoon).** Every cell needs to repair its lesions independently. Consequently, damage (blue) needs to be detected by Pol II (green) elongation (grey arrow) in order to be repaired by TCR, and both need to be present in the same cell. Whilst the NGS data (bottom) would suggest largest repair closer to the right side (as most Pol II is present there), some cells cannot remove their CPDs (red box). Computed repair rates for CPD location (1) would be 100%, whereas for location (2) would exhibit only 50% repair despite the fact that the Pol II signal at (2) suggests larger presence. The stochasticity induced by rare lesions need to be taken specifically into account when modelling the DNA repair process to provide a mechanistic explanation.

4.4 Methods

4.4.1 Data Normalisation

ChIP-seq and CPD-seq data (Appendix B) were grouped into 5 bins per coding region. We used the same positions as in our previous publication (Zeitler et al. (2022)). By using 5 bins, we can clearly distinguish between beginning, centre, and end of the gene with a transition in between. As replicates were not available, the data quality was evaluated using assumptions about the signal evolution. To be precise, we presume that no new CPDs can occur after UV exposure, and Pol II does not recover from the transcription shutdown within 38 minutes. Consequently, CPD levels can only decrease, and Pol II levels must be predominantly lower at t_{38} than shortly after irradiation, as the transcription

shutdown does not occur instantly. Surprisingly, the CPD signal increases at t_8 with a constant factor (Figure 4.12). Moreover, Pol II levels deviate substantially at t_8, and do not fit into the supposed dynamics of a global transcription shutdown (Figure 4.13). Therefore, we remove t_8 from the Pol II and t_0 from the CPD time course. Fortunately, this temporal delay is in line with the expected repair time after recognition, which is supposed to be between 3-10 minutes (Erixon and Ahnström (1979)). Thus, repair that has happened before t_8 must correspond to the Pol II distribution at t_0.

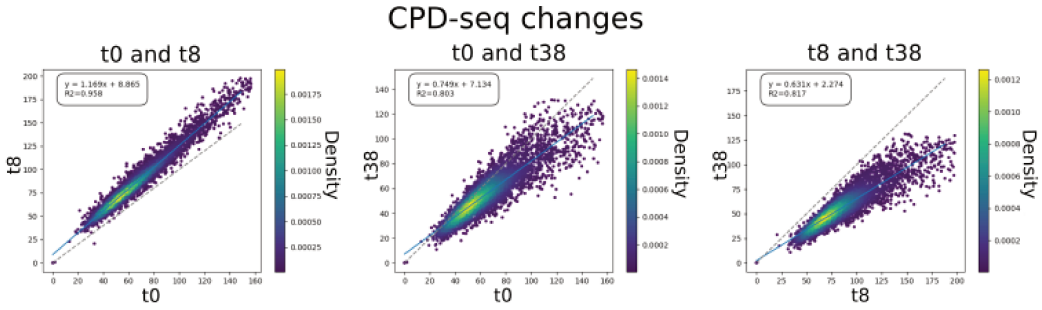


Figure 4.12: **CPD-seq data over time.** The plots show how the averaged CPD presence over each gene changes with respect to the measured time points. Surprisingly, the CPD-seq data at t_8 is larger than at t_0 at almost all transcribed regions. We opted therefore for the t_8-data point as initial condition. We removed values larger than the 97.5-percentile.

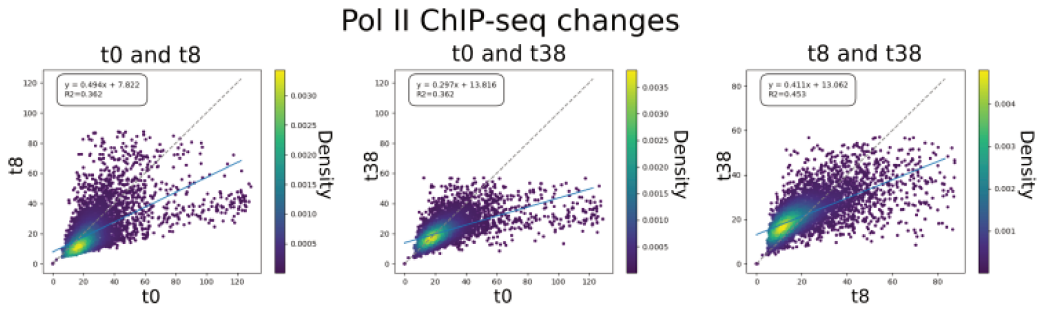


Figure 4.13: **Pol II ChIP-seq data over time.** The plots show how the averaged Pol II occupancy over each gene changes with respect to the measured time points. The data at t_8 is skewed. We chose t_0 as initial data point. We removed values larger than the 97.5-percentile.

4.4.2 Parameter Estimation

Parameters for ODEs 4.1 - 4.3 were estimated using the NODE framework (Chen (2018)). They were set to random initial values, which were improved over 200 iterations using a backpropagation algorithm modified with the adjoint sensitivity method (Chen et al. (2018)). Estimated values were required to be positive. The learning rate was set for all parameters to $\alpha = 0.01$. Time steps for updating a state using the ODEs were set arbitrarily to $\Delta t = 1/30$. This allowed a straightforward interpretation of the model predictions through time, i.e. a time unit was equal to the difference between data points. This came with the additional advantage that ODEs behave commonly better

when approximated over small time steps. Each gene was fitted independently.

4.4.3 Prediction Significance

We measured the prediction error using a classic MSE (Eq 4.4 with $w_P = 1$) for evaluating model significance. The distribution over 500 random model projections was approximated with a gamma distribution

$$f(x;\alpha) = \frac{x^{\alpha-1}\, e^{-x}}{\Gamma(\alpha)} \tag{4.5}$$

where $\Gamma(\cdot)$ is the gamma function. It is defined over the interval $[0, \infty)$ and is commonly used to model distributions over a range with a lower limit, such as the MSE. Random values for a given variable were drawn from a uniform distribution between the lowest and the largest estimated value, therefore taking parameter-specific ranges into account. By fitting Eq 4.5 to the MSE distribution of random model predictions, we determined a range for the possible performance presuming repair is independent of the location and when all genes exhibit similar kinetics. Estimated parameters were deemed to be significant—therefore specific to their region—if their prediction was smaller than the lower bound of the 90% PI, and therefore unlikely to be guessed by chance.

4.4.4 Sobol Sensitivity Analysis

In order to determine the magnitude of influence of a set of input variables on a functional output, the variance-based sensitivity analysis (or Sobol analysis) decomposes the variance into fractions that can be attributed to the input. Supposing an unknown function $Y = f(\mathbf{X})$, $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, it can be decomposed into subfunctions that are dependent on only a subset of the input, i.e.

$$Y = f_0 + \sum_{i}^{n} f_i(X_i) + \sum_{i<j}^{n} f_{ij}(X_i, X_j) + ... + f_{1,2,...,n}(X_1, X_2, ..., X_n). \tag{4.6}$$

Here, f_0 is a constant, f_i is a function of X_i, $f_{i,j}$ is a function of X_i and X_j etc. As f is unknown, the treatment is probabilistic and the function is decomposed with respect to the conditional expected values, i.e.

$$
\begin{aligned}
&f_0 = E[Y] \\
&f_i(X_i) = E[Y|X_i] - f_0 \\
&f_{i,j}(X_i, X_j) = E[Y|X_i, X_j] - f_0 - f_i(X_i) - f_j(X_j) \\
&...
\end{aligned}
\tag{4.7}
$$

Terms that include several variables encode higher-order interactions. The expression with respect to the expected value allows the decomposition of the variance

$$
\mathsf{Var}(Y) = \sum_{i}^{n} V_i + \sum_{i<j}^{n} V_{ij} + ... + V_{1,2,...,n}, \tag{4.8}
$$

with

$$
\begin{aligned}
&V_i = \mathsf{Var}\left(E[Y|X_i]\right) \\
&V_{i,j} = \mathsf{Var}\left(E[Y|X_i, X_j]\right) - V_i - V_j \\
&...
\end{aligned}
\tag{4.9}
$$

Sensitivity is measured by the index

$$
S_i = \frac{V_i}{\mathsf{Var}(Y)}, \tag{4.10}
$$

which can be straightforwardly extended for any higher-order interaction by using the appropriate V. Consequently, the sum over all sensitivity indices of all orders is equal to 1.

4.4.5 PCA and Cosine Analysis

We measured the functional relationship between parameters by performing a dimensionality reduction using PCA and comparing the cosine similarity after PCA transformation. PCA was applied in order to combine parameters according to their variance, which allows an unbiased comparison with respect to their feature importance. As it requires normally-distributed input values, parameters were transformed using a Box-Cox transformation with $\lambda = 0.5$ and normalised such that they had zero mean and unit variance. As many parameters were forced to zero during training, the Gaussian-transformed values were inflated with respect to their minimum. Nonetheless, the main body of the distributions fulfills seemingly the normality criterion (Figure 4.14(A)). A Pareto analysis over the num-

ber of dimensions indicated that we can explain more than 80% of the variance with 3 dimensions, which is why PCA weights were fitted for a reduction to a three-dimensional space (Figure 4.14(B)).

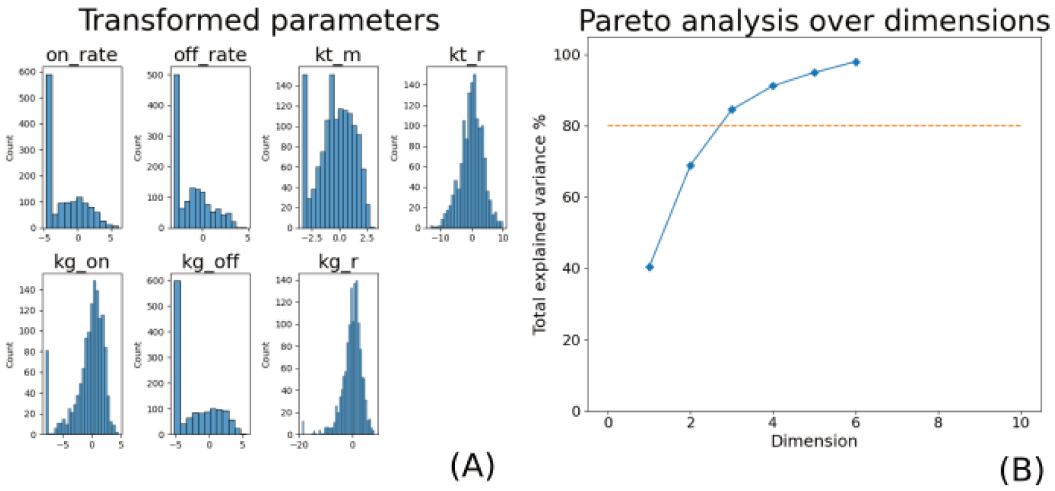


Figure 4.14: **Parameter transformation to near-Gaussian distributions and Pareto analysis.** (A) We show the parameter histograms after a Box-Cox transformation. The main body of the parameter distribution after centering and transforming to unit variance is seemingly normally distributed. However, due to the fact that some parameters were set to zero by the training algorithm, some of the distributions are inflated with respect to their lowest value. `on_rate` and `off_rate` are k_T^+ and k_T^-, respectively. `kt_m` represents m_T. `kg_on` and `kg_off` are the association and dissociation parameters for GGR (k_G^+ and k_G^-), whereas `kt_r` and `kg_r` represent the repair rates r_T and r_G. (B) A Pareto analysis over the reduced dimensions over the model parameters using PCA reveals that 3 dimensions can explain more than 80% (orange dashed line) of the parameter variance.

After determining the PCA weights $\mathbf{W}$ over the entire data set, we calculated the transformation when setting all parameters but one to 0. To be more precise, let $\mathbf{x}_i$ be the parameter vector of gene i, and $\bar{\mathbf{x}}_i^j$ denotes the vector when all parameters but j are set to zero. Both were transformed to $\mathbf{t}_i = \mathbf{W}\mathbf{x}_i$ and $\bar{\mathbf{t}}_i^j = \mathbf{W}\bar{\mathbf{x}}_i^j$, respectively. The cosine similarity c_{ij} between $\mathbf{t}_i$ and $\bar{\mathbf{t}}_i^j$ indicates how the parameter is influencing the transformation. By plotting c_{ij} for different parameters j over all genes i with respect to each other, we can establish a functional relationship between them. Clustering into two groups was performed on the cosine similarity per parameter and gene using a Gaussian mixture classifier. We did not find that using more clusters yielded better outcomes and instead made the interpretability of the clusters more difficult. This is why we restricted the analysis to two groups only.

Chapter 5

Providing a Mechanistic Understanding of Cell-Dependent Stochastic DNA Repair Using the *GillesPy* Algorithm

5.1 Introduction

DNA—as the vital hereditary unit of every living organism—is constantly damaged by internal and environmental factors. Fortunately, though, these lesion events are rare. We established in the previous chapter that—given the expected number of 0.2 induced CPDs/kb using 100 J/m^2 of UVC—we anticipate that less than 1000 cells possess UV-induced damage at the same position in a cell culture of 10,000,000 cells. As these values are low, we hypothesised that stochastic effects need to be taken into account to provide a mechanistic explanation on a single-cell scale. The traffic repair model (Chapter 4) revealed that repair dynamics are gene-specific, and therefore changing with respect to an unknown parameter. Moreover, some genes displayed a large dependence on repair by GGR, even though the regions have been selected because of their presumably strong TCR dynamics. Unfortunately, the mean-field approximation cannot provide a mechanistic understanding, i.e. linking the observations on a population scale to the necessary DNA-protein and protein-protein interactions in single cell. In order to shine light onto these observations, it is necessary to develop a probabilistic algorithm to explain location-specific DNA repair on a single-cell scale.

One of the best-known stochastic simulation approaches for chemical reactions was proposed by Gillespie (1977) (introduced in Section 1.3). By assuming that reactions in a well-mixed medium are sporadic, the Gillespie algorithm randomly samples a particular reaction μ and a time delay τ after which μ is observed based on the current state. A system state is defined as the number and types of particles present in the solution that can participate in a reaction. The reaction causes a state change, which can make other particle interactions more or less likely. A system can traverse different sequences of reactions when simulated repeatedly. The initial model proposed by Gillespie (1977) did not incorporate the notion of space. However, to represent DNA repair, it is important to include the position of a protein along the gene. Spatial versions of the Gillespie algorithm to model reaction-diffusion systems (Bernstein (2005)) and particle tracking in and around cells (Klann et al. (2012); Melunis and Hershberg (2017)) have used compartmentalisation (i.e. division of the space in smaller subregions) or lattice approaches. However, none of the proposed methods allow the modelisation of directed protein movements along the one-dimensional DNA. Moreover, methods to derive parameters within the stochastic algorithm to approximate the data have not been developed to our knowledge, although there have been studies that apply training approaches outside the simulations using neural networks (Arbona et al. (2021)). In this chapter, we propose a general simulation and training framework to represent particle interactions with and along a one-dimensional substrate. Single cells are mimicked independently, and their states are superimposed to reproduce NGS data. The framework fulfills two main purposes: firstly, it can link single-cell dynamics with static population-based sequencing data; and secondly, it can predict missing NGS distributions between sampled time points, showing how nuclear processes—such as DNA repair or gene transcription—really unfold over time.

The aim to include a general parameter estimation procedure for determining reaction rates to represent NGS data raises additional difficulties. The probabilistic treatment of the system evolution per cell renders the function discontinuous in time. Intuitively, if a protein A is sampled to associate to the DNA at time t in a given cell, there is a step-wise change in the presence of A at t. These sudden jumps make it impossible to directly apply efficient methods that are based on gradient optimisation to single-cell simulations (see Section 1.3.3). Other Bayesian inference approaches that rely on convergent sampling (such as MCMC) can be remedially applied. In short, they rely on the convergence of multiple tested parameter regimes to find a posterior distribution of likely parameters that match the provided data. Nonetheless, they require the simultaneous evaluation of several large parameter sets sampled from a given distribution, which comes with high computational costs. Alternatively, Reinforcement Learning (RL) is designed to learn a sequence of actions that transforms a system, such that a reward is maximised (e.g. a game like Go or chess

(Silver et al. (2018); Schrittwieser et al. (2020))). Therefore, it can deal inherently with discontinuities (for example the sudden change of moving a pawn in chess from one position to another). However, RL is largely reliant on the random exploration of possible actions that could improve the outcome; and only over time, the algorithm increasingly follows its current best policy. This exploration-exploitation trade-off requires commonly many training iterations, which can take a considerable amount of time on conventional hardware (e.g. 700.000 iterations over 9 hours on a server with specialised processing units for tensor operations (Silver et al. (2018))).

Therefore, we are faced with a multifaceted problem. Firstly, we need to model specific biological interactions that are necessary during DNA repair, and which we presume explain the observed decrease in the CPD-seq data; secondly, since we presume that damage (and potentially also Pol II presence) is rare in a single cell, we need to incorporate cell-specific and stochastic properties into the method; thirdly, it is important to develop a training algorithm to fit model parameters of the stochastic simulation to the available sequencing data; and lastly, parameter learning must be efficient such that it takes significantly less than several hundred thousand training iterations.

It is difficult to define a good approximation of DNA-related processes using a probabilistic algorithm. Within a reasonable degree, it is expected that the error between the stochastic simulation and the sequencing data will be larger than for a deterministic approximation approach that uses ODEs (such as the traffic repair model). We therefore introduce the *gateway problem* in order to quantify the liability and performance of our method as follows. Many DNA-protein interactions are dependent on the current state of chromatin conformation. Suppose a protein A must be present at position x, before another protein B can associate to position y and move to position z, with $y \neq z$. B cannot directly associate to z. This can be understood as A opening a gateway for the presence and motion of B, and without which binding is impossible. The temporal dependency of A and B as well as B's movement is a difficult problem that needs to be solved such that it happens repeatedly and statistically in each individual cell to match correctly the static NGS signal. We choose the initial parameters of a given set of reactions such that they are insufficient to allow binding and movement of B. We define that the algorithm can solve the gateway problem if the error trajectory shows a distinct decline over time.

In this chapter, we present the *GillesPy* framework—a portmanteau of Daniel Thomas *Gillespie*, who first proposed the non-spatial stochastic simulation algorithm for chemical reactions on which the framework is based, and the programming language *Python*. It is a general simulation and training module that can be used to represent any kind of interactions along a string or polymer, such as DNA, RNA, or amino acid chain. User-defined particles can associate, dissociate, and move along the one-dimensional substrate which is governed by a set of rules. Parameters can be conveniently

estimated independently of the implemented pathway by applying a gradient-descent approach over the time-continuous estimation of distributional data (i.e. NGS data), which is approximated by the single-cell simulations. We show that the method can efficiently reduce the error between simulation and NGS data by solving the gateway problem. Consequently, model parameters can be found without the need of sampling various parameter regimes, and it comes with reduced computational costs in comparison with variational Bayesian methods. The gateway problem itself was implemented as TFIIH-mediated transcription. TFIIH—of which Rad3 is a submodule—is a general transcription factor associated to the core promoter during PIC assembly and therefore governs Pol II association and transcription initiation (see Chapter 1). We apply the same parameter training technique to find interaction rates for CPD repair at genes that exhibit presumably a strong influence of TCR. Amazingly, the cell-dependent stochastic simulations are in line with the mean-field approximation, as both suggest gene-specific repair. The model indicates similarly that a high interaction rate of Rad4—which we presume represents GGR—with the DNA is necessary at some regions. By deconvoluting the simulated NGS signal into individual cells, we reveal that repair by GGR requires the probing of many positions, as random association and dissociation are less likely to find damage. Therefore, interactions of GGR-involved proteins with the DNA might be present at a much larger scale than previously appreciated, and the commonly associated later repair time comes solely from the comparatively ineffective way to scan for CPDs. This understanding of GGR at genes has not been previously proposed to our knowledge. The *GillesPy* model is thus capable of providing valuable cell-specific and temporal insights into nuclear processes.

For this project, I was leading in the mathematical formulation of the problem as well as the technical implementation. Initially, we aimed to use Rad4-tagged sequencing data to demonstrate functionality of the algorithm. As aforementioned, we could not construct a Rad4-tagged strain that followed repair dynamics similar to YPH (Figure B.1). Due to the time constraint and to provide comparability with the traffic repair model (Chapter 4), we used the same unpublished NGS data from the laboratory (Appendix B).

5.2 Results

5.2.1 Workings of the *GillesPy* Algorithm

In this section, we briefly introduce the fundamental principles of the *GillesPy* framework by focusing on DNA-protein interactions. Although referenced, we will not provide any precise mathematical formulation, and instead refer to Sections 5.4.1 and 5.4.2.

The algorithm makes use of the stochastic sampling approach proposed by Gillespie (1977), and it incorporates spatial information through compartmentalisation along one dimension. The notion of space can be implemented as probability distribution $P(\tau, \mu, x^-, x^+)$ by extending Eq 1.20 to Eq 5.1, where τ is the time delay, μ is the reaction, x^- denotes the reactant position, and x^+ gives the product position. The formula allows not only the simulation of particle interactions with a one-dimensional polymer, but it similarly permits a gradient-based parameter estimation. It is therefore the fundamental core of this project.

We surmise that the DNA molecule of length n is surrounded by a well-mixed solution containing proteins. The n positions are grouped into various predefined regions, which represent genomic areas with different DNA-protein interaction behaviour to model preferential binding. For example, TFIIH binds preferentially to the core promoter for its role in transcription initiation as part of the PIC, and it is less present within the gene body. Once proteins are bound to the DNA, they can transition through different states (e.g. ubiquitylated or phosphorylated) and move along the strand in a preferred direction. Similarly, they can influence the state of the DNA, e.g. evoke a change from damaged to repaired. Motion is implemented using the Smoluchowski equation, which is a specific form of the Fokker-Planck equation (Eq 1.14). Interactions between proteins and the DNA (or different associated proteins among each other) are implemented by user-defined rules. A rule μ includes the reacting DNA sites and proteins (together with their respective states) as well as the producing DNA sites and proteins (together with their respective states). A reaction can be optionally inhibited by a specific chromatin state, and a force value A_μ can be passed to modulate the direction and strength of protein movement. The rule-specific reaction rate defines the frequency at which we assume to observe a given reaction within a time unit when all reactants are present. Therefore, we slightly change the definition of θ_μ in Eq 1.20, which denotes a reaction *probability* and cannot exceed 1. However, this understanding does not influence the sampling approach, and we use in the following θ_μ to denote the reaction rate in the *GillesPy* algorithm.

The implementation keeps the state s_i of m cells simultaneously in memory; yet only a randomly-drawn fraction updates their state according to the Gillespie sampling during one time step. This avoids oscillations in the simulated sequencing data when starting from identical conditions in each cell. Consequently, instead of simulating an average behaviour as for the traffic repair model (Chapter 4), we sample the DNA-protein interaction in each cell individually (Figure 5.1). Sequencing itself is implemented as follows. We apply a modified maximum pooling function to each s_i that sets all proteins of a given type to the most occurring state within a moving window. To provide a general example which is unrelated to repair, all associated Pol II are set to the most occurring state (e.g. phosphorylated) within 100 bp in s_i, whereas the pooling of a simulated H3 histone tail is independent

and is set to the most occurring state of H3 (e.g. acetylated). This intends to represent many similar cells at the same time within one s_i. The cell states are subsequently averaged and smoothed along the DNA (Figure 5.2, top).

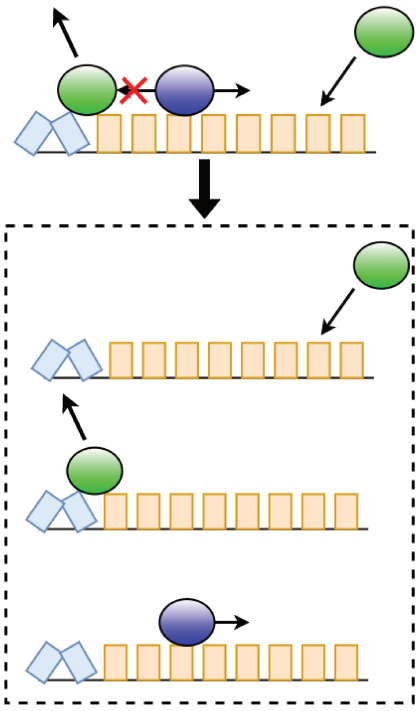


Figure 5.1: **The *GillesPy* algorithm simulates each cell individually.** The cartoon show the interactions (arrows) of two proteins (green and blue ellipses) with the DNA in possibly different position-dependent states (blue boxes indicates CPD damage, orange boxes are undamaged positions). Instead of simulating an average behaviour as for the traffic repair model (Chapter 4) (top row), the *GillesPy* algorithm represents each cell of an entire cell culture individually (dashed box). Therefore, every instance (a single cartoon within the dashed box) can traverse through different and independent sequences of DNA-protein and protein-protein interactions.

It is not functionally known how the single-cell dynamics alter the sequencing signals over time on a population scale. We assume we can approximate this behaviour by sampling and simulating the reactions in each individual cell. It should be noted that the equations presented by Gillespie (1977) describe the time-continuous change of a probability distribution. Such a distribution is indeed given by the data, as the NGS signal can be understood as indicating the likelihood of finding a measured property at a position. This allows us to link the parameters in Eq 5.1 (governing single-cell interactions) to the population-wide sequencing (Eq 5.3). By defining an error between simulation and NGS data (Eq 5.2), we can backpropagate the error gradient with respect to θ_μ in time (Figure 5.2, bottom).

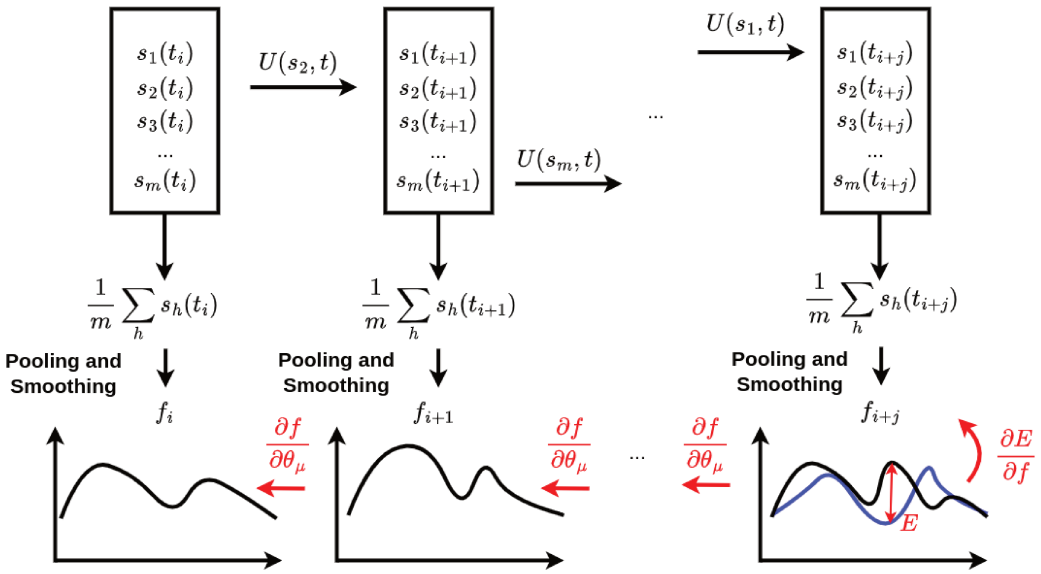


Figure 5.2: **Schematic representation of the *GillesPy* simulation flow.** In order to simulate the sequencing (or similar) data, the state s of m cells at time t (black boxes, top row) is averaged, smoothed, and pooled, which results in a distribution f along the one-dimensional substrate (bottom cartoon of a graph). The cell state s is updated using the update function U. Instead of sampling all cells at the same time, we sample cm random cells ($0 < c \leq 1$) which update their state. Repeating this over several steps simulates the temporal evolution of the sequencing signal. The prediction can be compared at several time points to the data (distribution in blue, bottom right), which allows an error computation E (red). When formulating the expected change in f as a function of U (Eq 5.3), we can backpass the error gradient in time (red arrows).

5.2.2 The *GillesPy* Algorithm Efficiently Solves the Gateway Problem

In order to evaluate whether the *GillesPy* model can efficiently solve the gateway problem, we estimated the parameters for a set of rules that represented TFIIH-governed transcription (Appendix C.1). In this setup, TFIIH needs to associate to the core promoter, which permits binding of Pol II to the TSS. Pol II can elongate into the coding region, after which it can dissociate (Figure 5.3). Sampling parameters θ_μ as well as movement strength of Pol II are unknown and need to be determined with respect to the data. We filtered for transcribed regions that contained no introns, are between 1-2kb, are Pol II-transcribed, and whose TATA-like element is 20-70bp before the TSS. We then removed the upper and the lower 5th-percentile, resulting in a selection 799 genes. Parameters were fitted to Pol II and Rad3 ChIP-seq, the latter representing the TFIIH distribution (Appendix B). The setup is presented in detail in Section 5.4.3.

We can report that the parameter estimation finds reasonable values and solves the gateway problem at almost all genes (i.e. ≈84% of the genes reduce the error at least by 30%) (Figure 5.4(A)). This result suggest that the approach can indeed find good approximations for sequence-dependent interactions (exemplified for gene YBR057C in Figures 5.4(B)-(D)). As we apply a gradient-descent

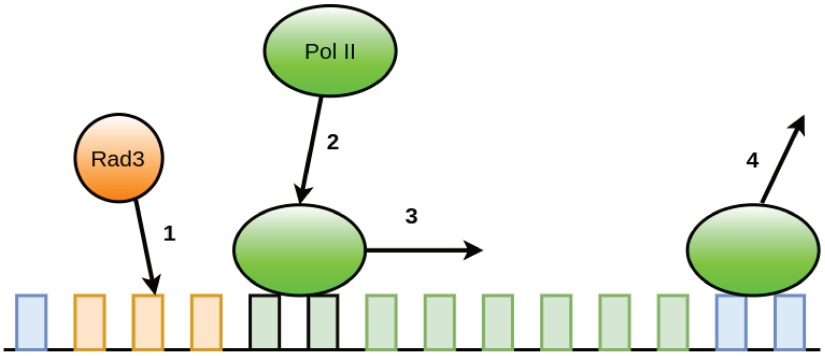


Figure 5.3: **Schematic representation of the gateway rule set.** Rad3 (orange circle)—one of the TFIIH submodules—needs to associate first (arrow 1) to the core promoter (orange rectangles). This permits Pol II (green ellipse) to associate (arrow 2) to the TSS (green rectangles, black border) and elongate (arrow 3) into the protein-coding region (green rectangles, green border). When it moves outside the gene (blue rectangles), Pol II dissociates from the DNA (arrow 4).

approach, parameters can improve without the need of evaluating many regimes at the same time. Therefore, we claim that the method is efficient with reduced computational costs in comparison to Bayesian inference approaches that rely on parameter sampling. We want to point out that the value θ_μ of a single reaction μ is dependent on all other reaction rates (Eq 1.20). Therefore, the algorithm needs to find a good balance between all of them. Moreover, the rates also determine how many reactions are likely to be sampled within a given time window (i.e. 35 simulated minutes). This means that the method needs to find good sampling rates to allow a sufficient number of drawn reactions that permit the completion of the temporal sequence. Despite these additional difficulties, the *GillesPy* model can drastically reduce the error in comparison to the initial parameters. By solving the gateway problem, we presume that the derived general training method is suitable for linking single-cell dynamics to NGS data.

5.2.3 Determining Repair Parameters

We repeated the parameter estimation for a defined CPD repair pathway that is comparable with the traffic repair model (Appendix C.2). Due to the limited time available, we sampled randomly 825 genes from the 1878 coding regions that we considered as being predominantly repaired by TCR (Zeitler et al. (2022)). Initial conditions were given by the double-strand CPD-seq and Pol II ChIP-seq data used in Chapter 4 (Appendix B). A detailed description of the setup is given in Section 5.4.4.

Once again, we can show that most genes improve their parameter values over time (Figure 5.5(A)). As the initialisation was already set to a sophisticated guess that comes close to the dynamics on average, some genes did not significantly improved their loss. Nevertheless, $\approx$74% of the coding regions reduced their error at least by 20%. The fact that we can reduce error for TFIIH-mediated

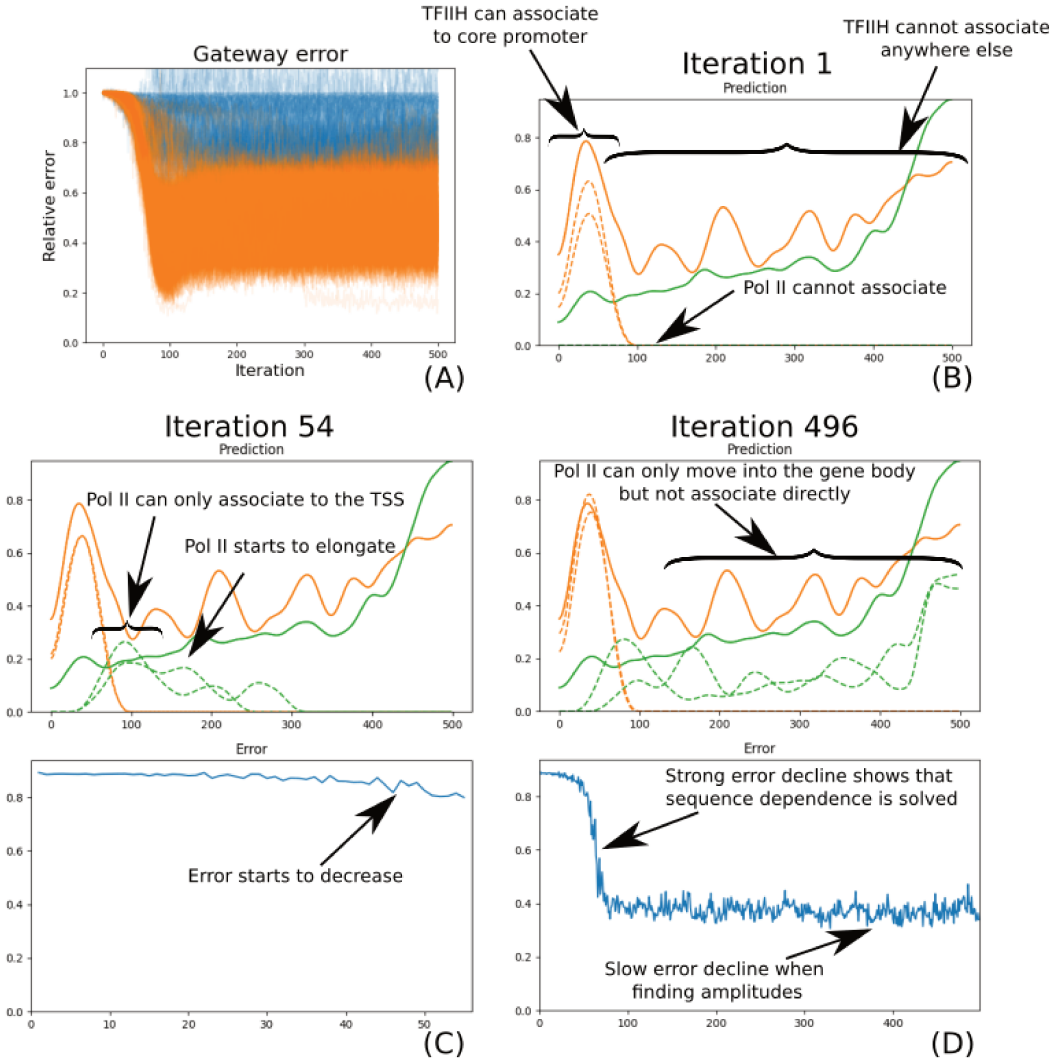


Figure 5.4: **The *GillesPy* algorithm can solve the gateway problem.**(A) displays the error function over 500 training epochs over all considered genes. It shows *GillesPy* can find improved parameter values at most genes, which is indicated by the strong error decline. Orange lines give the error decline for genes that successfully solved the gateway problem ($<$70% after 200 iterations for at least half of the remaining 300 iterations). This is $\approx$84% of all evaluated regions. Blue lines show the error for unsuccessful training. (B) - (D): The training evolution is exemplified over different iterations for gene YBR057C. Green lines give the Pol II distribution, and the TFIIH sequencing is shown in orange. The dashed line is the prediction by the *GillesPy* algorithm, solid lines are the NGS signals. As we suppose that the data represents an equilibrium state, we aim to match two time points (arbitrarily set to 28 and 35 minutes) to the NGS distribution. (B) Initially, no Pol II can associate to the gene. (C) After some iterations, the algorithm finds parameters that allow Pol II binding and initiation of elongation. The error starts to decline. (D) Pol II can now associate, move into the gene, and dissociate from the TTS. The error has dramatically decreased, although matching the exact amplitudes would take substantially more time. Note that TFIIH cannot associate outside the core promoter, and the Rad3 NGS signal within the gene is considered to be noise. Therefore, the error cannot go down to 0.

transcription and CPD repair shows that the *GillesPy* algorithm can be applied independently of the pathway and sequencing data.

As the traffic repair model predicted gene-specific repair, we evaluated the significance of the

GillesPy parameters. Simulations with the trained values was repeated over 20 iterations to determine an error distribution. This was compared using Mann–Whitney U significance test to the prediction of 20 random parameter regimes, which were sampled with respect to the trained values over all genes. The loss was measured using a standard MSE. Amazingly, repair predictions were once again highly significant, indicating location-specific repair dynamics (Figure 5.5(B)). The stochastic and cell-specific *GillesPy* algorithm therefore provides the same prediction as the mean-field traffic repair model, despite their substantial differences.

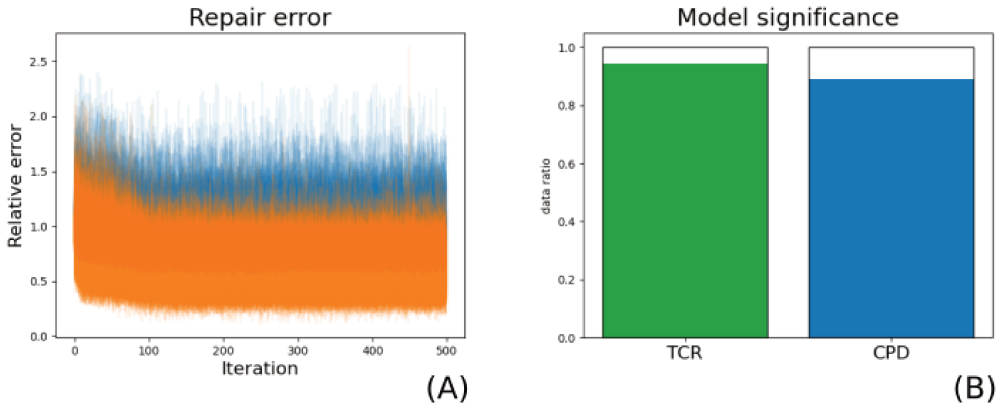


Figure 5.5: **Cell-dependent stochastic repair dynamics are gene-specific.** (A) displays the error function over 500 training epochs over all considered TCR regions. 74% of them genes can reduce the error at least by 20% after an initial training time of 200 iterations for at least 150 iterations over the remaining 300 (orange). Genes that did not reduce their error are given in blue. It should be stressed that this does not mean that training failed, since we started with good initial parameter guesses. (B) Model predictions for Pol II and repair are again highly significant for almost all genes, indicating location-specific repair dynamics.

The consequence that repair is gene-specific in *both* models implies that the observed damage removal cannot be solely explained based on initial conditions. The single-cell mechanisms are in line with the population-wide dynamics implemented using the traffic model. Indeed, when quantifying the effect of the model parameters using the cosine similarity (as presented in Chapter 4), we can show a strong functional interdependence of most parameters (Figure 5.6). The only parameter that exhibited more variance is the association of Pol II (parameter 2). This strongly suggests that repair kinetics change between genomic locations, and they must be functionally dependent on other external factors that had not been included. Similar to the procedure for the traffic repair model, we clustered the cosine similarities using a multivariate Gaussian classifier. The grouping is particularly linked to the Pol II association rate, although not exclusively. When comparing Pol II ChIP-seq and CPD-seq distributions, the two groups are seemingly dissimilar to the clusters determined using the traffic repair model (compare Figure 5.7 with Figure 4.6). Indeed, the ORFs changed drastically between the clusters, such that half of the genes from Group 1 for the traffic model parameters changed

to Group 2 for *GillesPy* parameters, and vice versa. This time, a gene ontology analysis could not reveal any appreciable functional differences. Interestingly, gene expression differs significantly for all time points (p-values of a Mann-Whitney-U test are 0.1%, $7 \times 10^{-5}\%$, and 0.6% for before UV, 8 min, and 38 min respectively) (Figure 5.8). Despite the dissimilarity of the overall distribution, they show largely overlapping values. It is difficult to explain why genes with similar transcription levels should be in different repair groups. We see it as unlikely that gene expression is the defining characteristic to which the two gene clusters are linked. It should be noted that we apply the classifier only to a subgroup of genes, as the time constraint limited the number of coding regions for which we could determine the parameters. However, to allow comparability, clustering was repeated for the traffic repair model parameters using only these genes. Hence, we do not expect that including all transcribed regions would change the overall result. Instead, we suppose that the different grouping is particularly linked to the different workings of the two approaches. For example, competition between TCR and GGR-related proteins is substantially weaker. Moreover, many simulated cells do not contain any damage, and Pol II elongation can continue unperturbed in most cases. We hypothesise that this is also the reason why Pol II association rate changes more drastically between genes. It should be nonetheless highlighted that both models predict gene-specific repair on a population scale.

5.2.4 Understanding Gene-specific Repair

In order to better understand the high presence of GGRP predicted by the traffic model for coding regions in Group 1, we simulated CPD removal at gene YAL020C with the estimated parameters. Amazingly, the model indicates a time scale and occupancy level for Rad4 comparable to GGRP dynamics in Chapter 4 (Figure 5.9). As the simulation of NGS data is achieved through interactions on a single-cell scale, we deconvoluted the sequencing signal into distinct cells, which displayed different repair scenarios. However, it should be mentioned that most cells did not exhibit damage due to the low rate. When a lesion was induced, the initial presence of Pol II could often swiftly detect CPDs (Figure 5.10). However, some simulated cells were more reliant on damage detection through GGR. As indicated by the population-wide simulation, a large number of cells displayed a high presence of Rad4. Despite its large occupancy levels, it took a surprisingly long time before Rad4 bound to the lesion and triggered repair (Figure 5.11). This suggests that the later observed lesion removal does not stem from less efficient DNA interactions, but instead points towards the protein-binding mechanism itself. Statistically, Rad4 needs to randomly associate and dissociate to a large number of positions before damage recognition. The Pol II-mediated TCR dynamics, on the other hand, provide a highly directed searching mechanism along the gene body. This is further supported by the

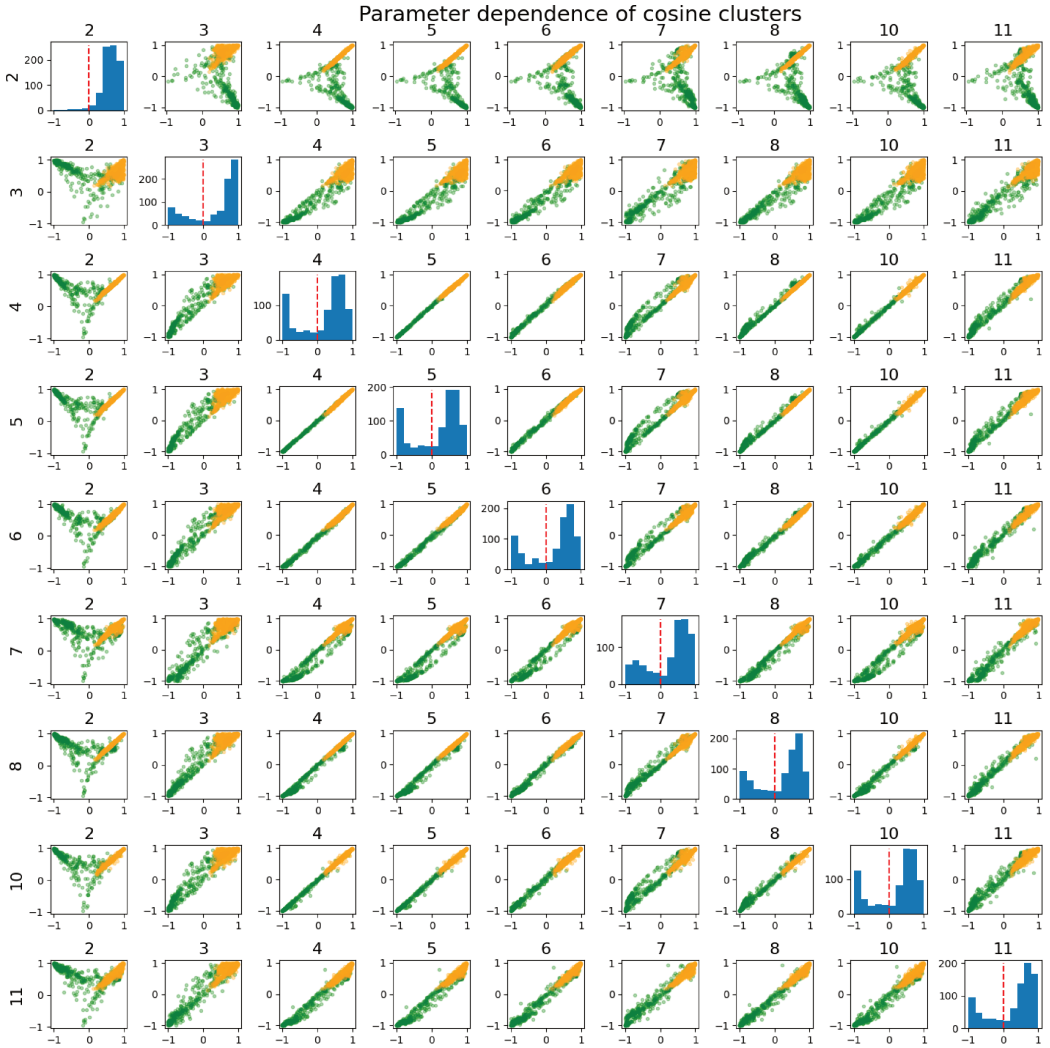


Figure 5.6: ***GillesPy* repair parameters are functionally interdependent.** We measured the cosine similarity of the transformed parameters as for the traffic repair model (Chapter 4). By comparing the similarity pairwise between parameters, we can investigate how parameters change over all genes with respect to each other. Each row and column represent a parameter, a single plot displays their pairwise comparison. Each point represents the parameter comparison for a single gene. The green dots show the distribution of Group 1, whereas the yellow dots display Group 2. The figures on the diagonal show the histogram distribution of the cosine similarities for a single parameter. The red dashed line in the histogram indicates a cosine value of zero, meaning that the transformation of a single parameter was orthogonal to the transformation of all parameters. The strong clustering along a line indicates a functional correlation between the parameters, which vary between genes. The numbers represent the parameters as follows. 2: Pol II association. 3: Pol II movement. 4: Pol II stalling. 5: Pol II dissociation. 6: Rad4 association. 7: Rad4 dissociation. 8: Pol II stalling at Rad4. 10: TCR repair rate. 11: GGR repair rate. We removed 0 (random association), 1 (random dissociation), and 9 (Pol II stalling at lesions) because these parameters were not trainable.

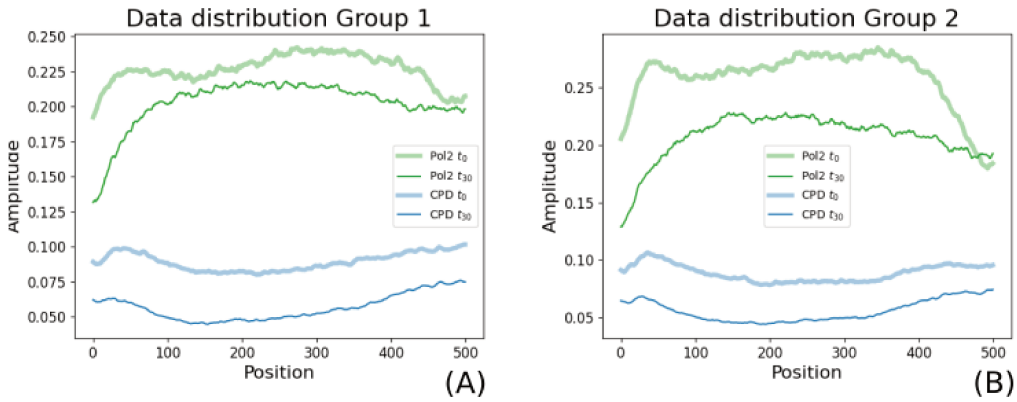


Figure 5.7: **Clustering for the *GillesPy* algorithm is different to the traffic repair model.** (A) and (B) show the average sequencing signal over all genes in Group 1 and 2. Green lines give the Pol II data, whereas blue lines display the CPD-seq distributions. The two groups determined by a multivariate Gaussian classifier on the cosine similarities change between the *GillesPy* algorithm and the traffic repair model.

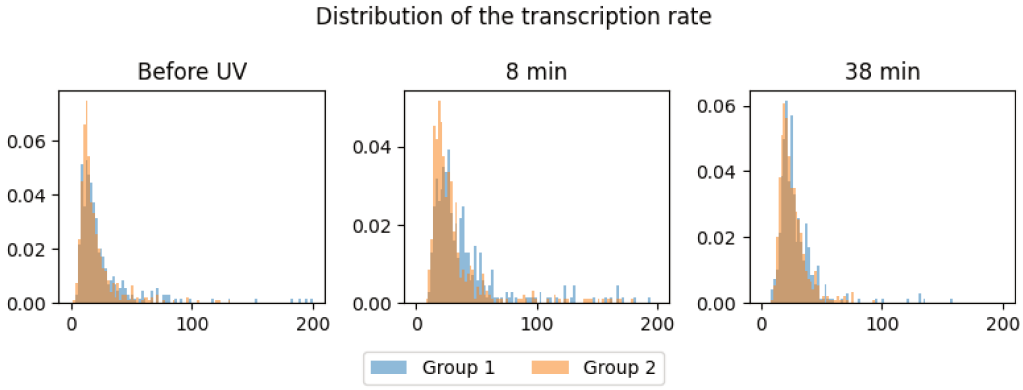


Figure 5.8: **The transcription rate distributions of the two repair groups exhibit a large overlap.** We compared the histogram distributions over Pol II presence before UV as well as at 8 min and 38 minutes after UV irradiation for gene group 1 and 2 (determined using the cosine similarities). As for the traffic repair model (Figure 4.7(A)), the differences in Pol II presence are significant, although the distributions are seemingly similar between Group 1 (blue) and Group 2 (orange). We find it unlikely that expression levels are the determining factor that defines the repair groups.

most observed repair scenario, during which Rad4 and Pol II were interacting synchronously with the DNA. Whilst Rad4 was bound at much larger quantities, it is Pol II that detects and removes the CPD (Figure 5.12). The model suggests that GGR-related proteins might be present at substantially higher levels—even at coding regions—than previously appreciated. Nonetheless, it should be highlighted that Rad4 was left as a hidden variable, and the algorithm had therefore no incentive find reasonable dissociation rates of Rad4 to match a specific NGS amplitude. Instead, the dynamics could rather be understood as first-visiting times, meaning that a position remains marked once Rad4 associated. Even with this interpretation, we find it remarkable that in ≈30% of the cells, GGR has scanned a substantial amount of the gene after 25 minutes (i.e. orange line in Figure 5.9).

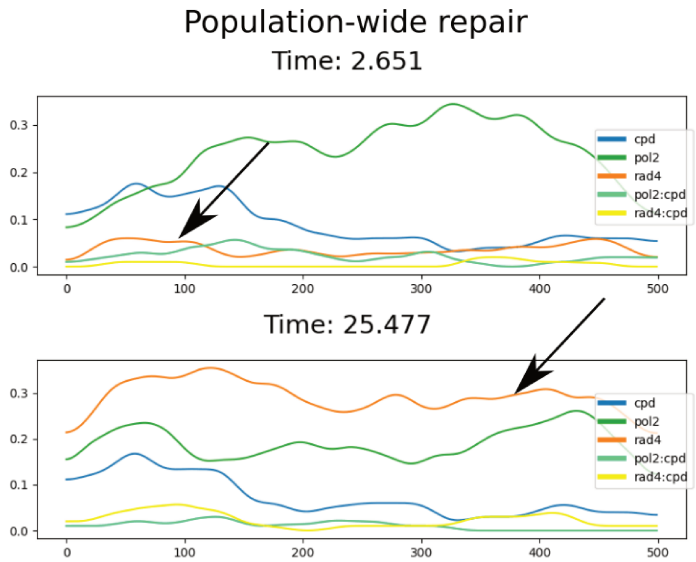


Figure 5.9: ***GillesPy* repair dynamics at gene YAL020C on a population scale.** The figure shows the simulated evolution of NGS data for repair-related properties with the trained parameters. The top row gives a time point early after the simulated induction of DNA damage, the bottom row displays the distribution of repair-related properties at a later time point. The x-axis indicates the position along the TCR region (over 500 bins), the y-axis shows the expected ratio of cells containing the property at this bin after maximum occurence pooling and smoothing. The importance of GGR strongly increases over time (orange line), which is predicted in a similar time scale as the traffic repair model (black arrows). Whilst Pol II levels only slightly decrease (green line), Pol II-CPD double sequencing is close to zero after 25 minutes (turquoise line). Rad4-CPD double ChIP-seq (yellow) approaches the damage distribution (blue) with increasing occupancy. The time is given in minutes.

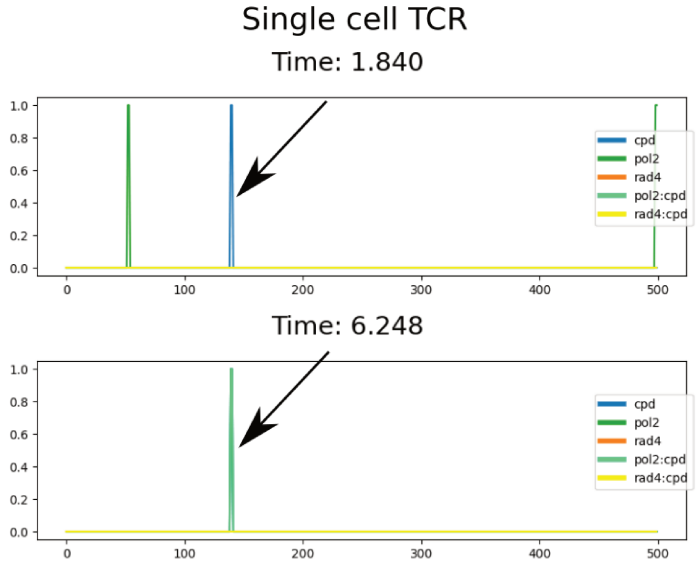


Figure 5.10: ***GillesPy* TCR dynamics at gene YAL020C on a single-cell scale.** The plot shows two snapshots during repair by TCR for a single simulated cell. The x axis indicates the position along the region (over 500 bins), and the y axis indicates presence or absence of a property. Presence of damage or proteins is indicated by a peak at this position. The elongation of Pol II (green) allows an efficient detection (turquoise) of the damage (blue). The time is given in minutes.

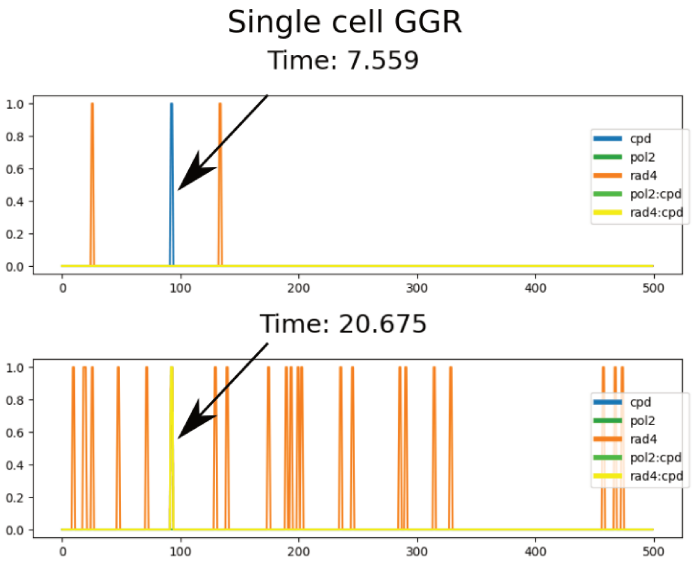


Figure 5.11: ***GillesPy* GGR dynamics at gene YAL020C on a single-cell scale.** The plot shows two snapshots during repair by GGR for a single simulated cell. The x axis indicates the position along the region (over 500 bins), and the y axis indicates presence or absence of a property. Presence of damage or proteins is indicated by a peak at this position. The random association and dissociation of Rad4 (orange) along the DNA requires the probing of more positions before the lesion (blue) can be found (yellow). The time is given in minutes.

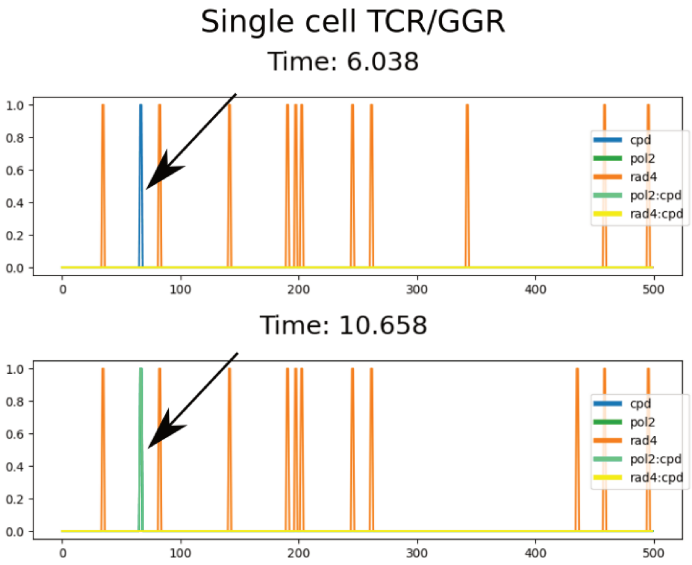


Figure 5.12: **Efficient TCR despite large presence of Rad4.** The plot shows two snapshots for a single simulated cell where TCR detects and repairs the induced CPD despite the large presence of Rad4. The x axis indicates the position along the region (over 500 bins), and the y axis indicates presence or absence of a property. Presence of damage or protein is indicated by a peak at this position. Despite the high presence of Rad4 (orange), TCR (turquoise) finds the CPD position (blue) more efficiently than GGR. The time is given in minutes.

The *GillesPy* model can equally predict other data types that the model was not fitted to. Naturally, Rad4 is one example, but the simulations can be extended such that two properties are sequenced at the same time. This can be understood as a double ChIP-seq experiment. The application of

the maximum pooling function during the simulated sequencing permits the spatial extension of a property in a single cell, and protein presence can be seemingly overlap within a given range. This is equivalent to the experimental procedure, during which the immunoprecipitation step is applied to, for example, a repair protein and CPD presence on a DNA fragment of a given length (e.g. 300 bp). Based on the estimated repair parameters, we predicted Pol II-CPD and Rad4-CPD double ChIP-seq data along the YAL020C gene (Figure 5.9). The Pol II-CPD distribution is lightly elevated during early time points, and then quickly abates. We found that it often did neither follow the Pol II nor CPD distribution alone. During ongoing repair, however, the Rad4-CPD double ChIP-seq signal approaches the damage distribution. The prediction suggests caution when analysing double-property sequencing data, as it could indicate a strong correlation between double-ChIP and CPD-seq signal that does not translate into a mechanistic interaction. Instead, it is solely produced by properties that are spatially close. If it turns out to be true that GGR-related proteins are present along the gene at larger quantities or earlier than previously thought, double-sequencing methods could allocate a bigger proportion of observed repair to GGR. Further experiments are necessary to verify the model prediction.

5.3 Discussion

In this chapter, we presented a general simulation and training framework—called the *GillesPy* algorithm—that can represent any particle interaction along a one-dimensional polymer. We demonstrated that it can link temporal sequence-dependent dynamics in single cells to static population-based NGS data by solving the gateway problem. Indeed, the approach finds reasonable parameters for TFIIH-mediated transcription at more than 84% of all tested genes. This shows that our efficient gradient-descent method—which does not require costly evaluation of several parameter regimes at the same time (i.e. Bayesian inference) or additional training for neural networks for action planning (i.e. deep RL)—can approximate the data based only on the current estimate. Amazingly, applying the method to the kinetics of repair data leads to the same predictions as for the traffic repair model (Chapter 4). These are in particular gene-specific lesion removal dynamics as well as high presence of GGR-related proteins at some coding regions. By disintegrating the signal into its cellular components, we can link the high presence of GGR to the inefficient DNA-interactions to scan for damage. This could indicate that they are substantially earlier present at genes (or at higher levels) than previously appreciated. Here, we discuss the applicability of the algorithm and set it into context with existing studies.

The Gillespie sampling approach (Gillespie (1977)) has been applied to model various chemical

systems, including mutual catalysis (Segré et al. (1998)) as well as cellular growth and division (Lu et al. (2004)). Other studies introduced the notion of space to represent reaction-diffusion interactions (Bernstein (2005)) and particle tracking (Klann et al. (2012); Melunis and Hershberg (2017)). It has also been used to simulate the spatial mean replication timing and replication fork directionality along the genome, therefore modelling nuclear processes along the DNA (Arbona et al. (2021)). Thus, it is an established and sensible approach for representing stochastic particle interactions with a one-dimensional polymer. However, previous implementations did not incorporate a pathway-unspecific training method to determine sampling parameters. Instead, they were commonly compared either qualitatively or used experiment-derived parameters (Bernstein (2005); Melunis and Hershberg (2017)), or alternatively, fitted independently of the stochastic simulation using a neural network (Arbona et al. (2021)). The latter requires careful and adequate transformation from one system to another (i.e. simulation to training and vice versa). Initially, we considered a training procedure that implemented the pathway as a general set of ODEs, for which parameters could be conveniently estimated through applying NODE (Chen (2018)). However, we were not able to find a general way of transferring parameters from one system to another independent of the use case. By approximating the data within the simulation framework, we successfully circumvented that problem, which makes learning progression monitorable and comparable with the actual behaviour of the system.

The sampling method used for the simulation of single chemical reactions randomly draws a reaction time τ. This value is dependent on the probability of observing any reaction (i.e. a_0) given a state s_i at time t. If many reactions are possible at t, τ is likely to be very small, and simulations can take a long time to complete. This becomes an important consideration for the approximation of NGS data, as the dynamics can be only derived from the static sequencing signal when provided with an explicit time frame T (e.g. repair within 30 minutes or reaching an equilibrium state within 25 minutes). It has been previously reported that Gillespie sampling becomes dramatically slower when fast reactions are considered, and alternatives have been presented (Cao et al. (2005)). However, we are not aware of ad-hoc solutions that can deal with different reaction time scales that are unknown *a priori*, and which would allow a straightforward parameter training. The time required to complete s system evolution within a given T depends on the number of possible interactions (e.g. proteins and length of the DNA) and the reaction rates of the system. To reduce temporal complexity, it is possible to reduce the length of the simulated polymer (for example through binning) and the number of interacting particles. Moreover, it is of paramount importance to set well-defined parameter ranges during training to prevent exploding values that can dramatically slow down forward simulations. Nevertheless, we want to stress this problem is inherent to the Gillespie sampling, and it is unrelated to our parameter training approach, which does not add a significant overhead.

Similar to slowing down forward simulations, τ can be sampled so large that only some few reactions are represented within a fixed T. Nonetheless, we can demonstrate through the gateway problem that the algorithm can find good approximations when starting with low parameter values. However, this can significantly increase the number of training iterations necessary. Providing reasonable lower parameter bounds can circumvent further complications.

To improve model interpretability, we redefined θ_μ to be the reaction rate rather than reaction probability. Consequently, a_μ represents the expected reaction rate given a state s_i. As mentioned before, a_μ is dependent on both θ_μ and the number of possible particle interactions h_μ. The latter value can vary largely between different rules. For example, random association and dissociation along the entire simulated DNA has more possible interactions than elongation of a single Pol II at a well-defined position. This means that parameters can differ greatly over several orders of magnitude, which should be carefully considered when choosing a learning rate α_μ. In our study, we apply a parameter-specific learning rate to correct for this bias. It should be said nonetheless that it is not possible to straightforwardly apply best-practices from standard machine learning approaches, as α_μ can similarly range over several orders of magnitude.

A well-known downside of gradient descent approaches over Bayesian inference is the sensitivity to the parameter initialisation. This is equally true for the *GillesPy* algorithm. However, the stochastic simulation allows the evaluation of different conditions and various temporal simulated sequencing progressions. The method is therefore less sensitive to initial parameter values if the boundaries are appropriately set, and a reasonable agreement with the data is likely to be found.

The parameter estimation finds interaction rates θ_μ which can recreate the NGS data amplitude in predefined regions. Appropriate data scaling is therefore essential. By gleaning repair rates and values for Pol II (Struhl (2007)) and CPD presence (Bucceri et al. (2006)) from the literature, we scale the NGS data such that the difference to the simulated sequencing signal is unbiased (Appendix B.9). This approach was also applied as a basis for the data scaling used in the traffic repair model (Chapter 4). It is essential to use single-cell values to provide a reasonable population-based data normalisation.

The implemented repair pathway is based on the supposed interactions that have been used in Chapter 4, and which were determined by *in vitro* experiments. Amazingly, the traffic repair model (a mean-field approximation) and the *GillesPy* algorithm (a stochastic model) both predict gene-specific repair, despite the fact that their method and level of detail is fundamentally different. This strongly supports our conclusions regarding repair. Both suggest gene-specific lesion removal and high presence of GGR-related proteins along coding regions that are presumed to be predominantly repaired by TCR. The *GillesPy* model allows to deconvolute the simulated sequencing data to observe CPD

repair unfolding on a single-cell scale. The observed scenarios indicate that GGR-involved proteins might be uniformly present at earlier time points (or at larger quantities) than currently appreciated. We hypothesise that the reason for later repair in non-transcribed regions—and which are consequently exclusively repaired by GGR—stems from less efficient lesion recognition kinetics provided by GGR, rather than lower protein presence or limited damage accessibility.

In conclusion, we presented a general simulation and training framework that can be applied to any particle interactions along a one-dimensional polymer. From a computer science perspective, parameters are determined such that a sequence of events represented in several cell instances can recreate a provided distribution over an entire population, such as ChIP-seq or CPD-seq data. Data approximation and simulations are linked within the same environment, and hence, they are not subject to any additional transformation between systems that would require careful consideration. We can prove its applicability to molecular DNA-protein interaction pathways through solving the gateway problem. By applying the algorithm to DNA repair, we can find a mechanistic explanation for the large importance of GGR in some regions. The method predicts gene-specific repair dynamics, which is in line with the traffic model. The *GillesPy* framework provides therefore a mechanistic understanding through the stochastic simulation of particle dynamics, which can be linked to static data distributions.

5.4 Methods

5.4.1 Sampling and Simulation

Interactions between N particles and the one-dimensional substrate of length n are governed by a set of M user-defined rules. Each rule defines the participating reactants (i.e. premises for a reaction to happen) and the reaction products. They contain a parameter for the sampling frequency θ_μ and optionally a force A_μ that encodes preferred particle movements into one direction. Reaction rules are implemented as *if-then* constructs to permit the straightforward formalisation using Boolean operators. This also permits the definition of several reactions with the same rule, e.g. the presence of particle A *or* B to enable loading of C. However, only one condition (e.g. either A or B) must be met. The polymer is surrounded by a well mixed solution. Particles can associate to the substrate; move along it, possibly with a direction-specific preference specified in A_μ; and dissociate again from the one-dimensional string back into the well mixed solution. Therefore, the notion of space is only important along the polymer. In the following, we consider exclusively DNA-protein interactions. We include the notion of space by compartmentalising the DNA string into n, discrete segments. They

can therefore represent either distinct base pairs or a larger region of the sequence. Proteins can exhibit a particular behaviour—such as loading preference—that is specific for certain DNA regions. Consequently, reaction μ can be dependent on a region R_μ with length ℓ. We incorporated the notion of space by extending Eq 1.18 as a joint probability distribution as follows

$$\begin{aligned} P(\tau, \mu, x^-, x^+ | s_i(t)) d\tau &= P(\tau, \mu | s_i(t))\ P(x^- | \tau, \mu, s_i(t))\ P(x^+ | x^-, \tau, \mu, s_i(t)) \\ &= \frac{1}{Z_{sim}} \underbrace{a_\mu \exp\left(-\sum_\nu a_\nu \tau\right)}_{\text{Gillespie Eq 1.18}} \underbrace{P(x^- | \tau, \mu, s_i(t))}_{\text{Sampling } x^-} \underbrace{\frac{1}{\sqrt{4\pi D_s \tau}} \exp\left(-\frac{(x^+ - x^- - A_\mu \tau)^2}{4 D_s \tau}\right)}_{\text{Sampling } x^+}. \end{aligned} \tag{5.1}$$

A_μ is the deterministic force that represents preferential protein movements. In the case of transcription, for example, Pol II movement can be explicitly modeled in the 3'-5' direction of the template DNA strand. D_s represents the fluctuation during the update step. The normalisation constant Z_{sim} accounts for the fact that we define θ_μ to be a reaction rate within a specified time unit, rather than a reaction probability. It is clear that the first term is identical to the probability distribution proposed by Gillespie (1977). Remember that $a_\mu = \theta_\mu h_\mu$. The second term represents the probability of sampling x^- given an already sampled reaction and the current state. The last term incorporates directed movement with uncertainty as described by von Smoluchowski (1906). Note that Eq 5.1 can also represent non-moving properties by setting D_s sufficiently small. In its limit, the Gaussian distribution becomes a Dirac function, therefore making any positional update that is other than x^- impossible. Consequently, properties of the DNA sequence itself or proteins that are not supposed to move can be adequately represented. Note that the positions x^- and x^+ also include the well-mixed solution around the DNA. However, in this case, the notion of segmentation is removed, and any application of a force A_μ or fluctuation D_s is meaningless.

In order to specify $P(x^- | \tau, \mu, s_i(t))$, we introduce $\ell(\mu, s_i(t))$ as the number of positions in R_μ where reaction μ is possible given state $s_i(t)$ at time t. Then, $P(x^- | \tau, \mu, s_i(t)) = \frac{1}{\ell(\mu, s_i(t))}$ if $\ell(\mu, s_i(t)) > 0$, and 0 otherwise.

Using Eq 5.1, we can sample μ and τ as in Gillespie (1977). $x^- \sim \mathcal{U}(R_\mu(s_i(t)))$ is a value randomly drawn from a uniform distribution within region R_μ which fulfills the reactant premise in state $s_i(t)$. Lastly, $x^+ \sim \mathcal{N}(x^- - A\tau, 2D\tau)$. In the following, we denote with $U(s_i, t)$ the update function that samples τ, μ, x^-, and x^+, and it changes state s_i at time t accordingly.

It is clear that when a state update is N' times independently drawn from U, it approaches Eq 5.1 when $N' \to \infty$. Consequently, we can approximate the probability distribution by simulating multiple

cell instances at the same time. In each update step, a fraction cm is updated, with $0 < c \leq 1$. We included uncertainty about the actual protein position by applying a modified max-pooling step to each simulated DNA, during which all values within a window of size l were set to the most occurring value. This can be equally understood as representing many similar cells in one state s_i. Motivated by the sonication step during the experimental data acquisition—which creates DNA fragments of similar size—we convoluted the mean particle presence with a smoothing function (Hann window of length 50 for all setups). This allowed the simulation of temporally evolving sequencing data (Figure 5.2). We calculate the update $\hat{\tau}$ of the cell sample by averaging over all sampled cell-specific τ.

5.4.2 Gradient Derivation

We aim to optimise the sampling parameters θ_μ and force value A_μ such that the averaged signal of the simulated cells at t_k has a minimal error to the NGS data at t

$$E(D, f; t_k, t) = \frac{1}{n} \sum_{x}^{n} \sum_{j}^{M'} w_j \left(D(j, x, t) - f(j, x, t_k) \right)^2 . \tag{5.2}$$

D is the sequencing data and M' are the number of probed properties (e.g. bound proteins or DNA damage). t_k is the first time value during simulation that is larger than or equal to t. j represents a probed property which sequencing data is available for. w_j is the property-specific error weight. $f(j, x, t_k)$ denotes the sequencing function of property j at update step k. By considering the *GillesPy* computational flow (Figure 5.2), it becomes clear that we can calculate the influence of a parameter θ_μ on the sequencing error back in time (red arrows) when finding an expression of $f(j, x, t_k)$ with respect to Eq 5.1. We define $f(j, x, t_k)$ at update step i such that its change over time is defined as follows

$$f(j,x,t_k)d\tau = \frac{1}{m}\sum_{i}^{m}\left(\underbrace{-\frac{1}{Z}\sum_{\nu}^{M} I(j,\nu^-)P(\tau,\nu|s_i(t_k))\; P(x^-|\tau,\nu,s_i(t_k))\; P(j|x^-,\tau,\nu,s_i(t_k))}_{\text{Particles that dissociate or move from } x \text{ during } d\tau}\right)$$

$$+\frac{1}{m}\sum_{i}^{m}\left(\underbrace{\frac{1}{Z}\sum_{\nu}^{M} I(j,\nu^+)P(\nu,\tau|s_i(t_k))\left(\int dy^- P(y^-|\tau,\nu,s_i(t_k))P(x^+|y^-,\tau,\nu,s_i(t_k))\sum_{j'}^{N} P(j'|y^-,\tau,\nu,s_i(t_k))\right)}_{\text{Particles that associate or move to } x \text{ during } d\tau \text{ from } y^-}\right). \tag{5.3}$$

$I(j, \nu^-)$ denotes the indicator function that returns 1 if j is a reactant of ν, and zero otherwise. $I(j, \nu^+)$ is defined equivalently for the list of products in ν. As reaction rules can be defined for several types of particles, the term $P(j|x^-, \tau, \nu, s_i(t_k))$ represents the probability of sampling j as reactant, given position x^-, time step τ, reaction ν, and state $s(t_i)$. It is simply defined as $1/\psi(\nu, s_i(x, t_k))$ if $j \in \nu^-$, and 0 otherwise, where ψ is a function that returns the number of possible reactants of ν in state s_i at x and t. Note that there is no such term for products, as it is assumed that they are fully defined by the sampled reactant. Z is a normalising value which accounts for the scaling between probability distribution and sequencing signal. Note that the probability is defined over all bound and unbound states. Z is therefore a constant, since *GillesPy* systems are simulated as closed system (no particles or DNA positions are added or removed). Due to the fact that we use discretised positions (e.g. bins), we reduce the integral to a sum. Let T_N denote the total number of update steps until a time t is reached. Then we can describe $f(j, x, t) = \sum_k^{T_N} f(j, x, t_k) d\tau$.

By applying the chain rule, we can write

$$\frac{\partial E}{\partial \theta_\mu} = \frac{\partial E}{\partial f} \frac{\partial f}{\partial a_\mu} \frac{\partial a_\mu}{\partial \theta_\mu}. \tag{5.4}$$

We can easily calculate $\partial_f E(j, x) = 2w_j \left(f(j, x, t_k) - D(j, x, t)\right)$ and $\partial_{\theta_\mu} a_\mu = h_\mu$. It is clear that $\partial_{a_\mu} f(j, x, t_k) = \partial_{a_\mu} \sum_{k'}^{k} f(j, x, t_{k'}) d\tau = \sum_{k'}^{k} \partial_{a_\mu} f(j, x, t_{k'}) d\tau$. When taking the derivative of Eq 5.3, we determine

$$\begin{aligned}
\frac{\partial f(j, x, t_{k'}) d\tau}{\partial a_\mu} \propto & -\frac{1}{m} \sum_i^m \Upsilon \left(\frac{I(j, \mu^-) \exp\left(-a_0 \tau\right)}{\ell(\mu^-, s_i(t_{k'})) \psi(j, \mu^-, s_i(t_{k'}))} \right) \\
& + \frac{1}{m} \sum_i^m \Upsilon \left(\frac{I(j, \mu^+)}{\sqrt{4\pi D_s \tau}} \frac{\exp\left(-a_0 \tau\right)}{\ell(\mu^+, s_i(t_{k'})) \psi(j, \mu^+, s_i(t_{k'}))} \right) \\
& \left(\sum_{y^-} \sum_{j'} \Upsilon \left(\frac{1}{\psi(\mu^-, s_i(t_{k'}), j', y^-)} \exp\left(-\frac{(x - y^+ - A_\mu \tau)}{4 D_s \tau} \right) \right) \right) \\
& + \frac{1}{m} \sum_i^m \sum_\nu^M \Upsilon \left(I(j, \nu^-) \frac{a_\nu \exp\left(-a_0 \tau\right) \tau}{\ell(\nu^+, s_i(t_{k'})) \psi(j, \nu^-, s_i(t_{k'}))} \right) \\
& - \frac{1}{m} \sum_i^m \sum_\nu^M \Upsilon \left(\frac{I(j, \nu^+)}{\sqrt{4\pi D_s \tau}} \frac{a_\nu \exp\left(-a_0 \tau\right) \tau}{\ell(\nu^+, s_i(t_{k'})) \psi(j, \nu^+, s_i(t_{k'}))} \right) \\
& \left(\sum_{y^-} \sum_{j'} \Upsilon \left(\frac{1}{\psi(\nu^-, s_i(t_{k'}), j', y^-)} \exp\left(-\frac{(x - y^- - A_\nu \tau)}{4 D_s \tau} \right) \right) \right).
\end{aligned} \tag{5.5}$$

$\Upsilon(x)$ denotes a modified identity function that returns x when $x \neq \infty$, and 0 otherwise. ∞ can occur when reaction is impossible in the given state, such that ℓ or ψ return 0. We can similarly find the

derivative with respect to A_μ:

$$\frac{\partial f(j,x,t_{k'})d\tau}{\partial A_\mu} \propto \Upsilon\left(\frac{I(j,\mu^+)}{\sqrt{4\pi D_s \tau}} \frac{a_\mu \exp\left(-a_0\tau\right)}{\ell(\mu^+, s_i(t_{k'}))\psi(j,\mu^+,s_i(t_{k'}))}\right) \left(\sum_{y^-}\sum_{j'} \Upsilon\left(\frac{1}{2\psi(j',\mu^-,s_i(t_{k'}),y^-)D_s} \exp\left(-\frac{(x-y^- - A_\mu\tau)^2}{4D_s\tau}\right)(x-y^- - A_\mu\tau)\right)\right). \tag{5.6}$$

We write that the gradients are proportional ($\propto$), since we remove the normalisation factor Z. It can be implicitly incorporated by a rule-specific learning rate α_μ.

The weighted MSE is minimised when updating the parameters θ_μ by $-\alpha_\mu \partial_{\theta_\mu} E$ and the force A_μ by $-\beta_\mu \partial_{A_\mu} E$, where α_μ and β_μ are rule-specific learning rates for reaction frequency θ_μ and force A_μ, respectively.

Gradient descent is repeated over several training iterations (in all setups set to 500). To improve performance and to keep values within bounded regions, we defined upper and lower limits for each parameters. Gradients were clipped to maximally 10% of their value to avoid exploding gradients. We applied a weak gradient momentum of 0.5. If less than 30% of the rules were sampled, we increased all θ_μ by 10% to increase number of reactions sampled in the given simulation time frame T.

5.4.3 The Gateway Problem

The ChIP-seq data for Pol II and the TFIIH subunit Rad3 were determined in Rad3-HA-tagged strains without UV exposure, and they were produced and treated as for the Rad4-tagged strains (Appendix B).

We included an extra 100bp window before the TATA-like element to allow for some padding. TFIIH could bind within a core promoter region of 100bp length, which started 25bp before the TATA-like element. Only a single TFIIH complex could be bound to the core promoter at the same time. The TSS was defined as an area of 120bp, which started 75bp after the core promoter and to which Pol II needed to bind before elongating into the coding region. We added another padding of 100bp after the coding region from which Pol II could dissociate. We improved reproducibility by binning all regions into 500 bins, making them therefore independent of the actual gene size. We simulated interactions with 250 proteins, 60% of which were Pol II and 40% were TFIIH. We implemented a slight bias due to the fact that more Pol II proteins could be present along the gene than the transcription factor. Initial reaction rate parameters were set to low values, such that it was impossible to sample any Pol II interactions. Elongation rate and speed were similarly chosen, such that once

Pol II was bound, it was impossible to elongate when no parameter update had been performed. All genes were trained over 500 iterations. At every iteration, the simulation was reset such that all positions along the DNA were free of bound proteins. The algorithm sampled reactions μ and update steps τ for 35 simulated minutes, after 28 of which an equilibrium needed to be found. Subsequently, parameters were updated based on the error between the prediction at both time points (i.e. 28 and 35 minutes) and the NGS data. We assumed a fixed rate of random association and dissociation for all proteins along the entire simulated DNA. As different values for association and dissociation can achieve the same ratio—and therefore, the same equilibrium amplitude in the NGS data—we fixed TFIIH dissociation rates before training. The aim was to find reaction rates that can describe the NGS distribution within the predefined genetic regions. This can only be achieved when the algorithm finds a way to make TFIIH binding sufficiently likely to allow Pol II association and elongation with appropriate rate and speed.

5.4.4 CPD Repair

As a baseline, all proteins could randomly associate and dissociate along the entire DNA with a fixed and low rate. To provide a fair comparison to the traffic repair model, we modelled Pol II-mediated TCR by association, directional movement, and dissociation without a transcription factor. Rad4—which represented repair by GGR—could associate and dissociate indiscriminately along the entire simulated region. Similar to the dynamics in Chapter 4, Pol II and Rad4 compete for lesion removal. We defined that Pol II stalls at Rad4-bound positions, and Rad4 cannot bind to the same region as Pol II. The implemented pathway included one more transitional step during repair through TCR than for GGR. To allow a fair parameter estimation for both pathways, we set the extra transition to an arbitrary high but fixed rate, such that the step was performed almost surely when possible. Pol II needed to associate to the TSS before it could move into the gene. 250 proteins could interact with the DNA, 60% of which were Pol II and 40% were Rad4. The slight bias was implemented to account for the different number of interactions possible, as Rad4 can interact with any position along the gene, whereas Pol II can only freely associate to the TSS, making it much less likely to be sampled. All genes were binned into 500 positions, therefore taking more details of the actual signal into account than for the traffic repair model.

Coding regions were selected as in Zeitler et al. (2022). Gene coordinates defined the TSS and Poly-A site, and they did not include the core promoter. The DNA layout was defined as for the gateway problem, with the exception that the core promoter was removed. To take cell-specific and stochastic repair into account, we randomly sampled 5 initial conditions based on the data (Appendix

B.9), which were selected by chance at every iteration during parameter estimation. Hence, the training considered varying initial cell states. Parameters were set according to sophisticated guesses before learning commenced. Thus, it differs in this regard from the gateway setup, where parameters were set to make training as difficult as possible. To allow equal and unbiased training conditions for TCR and GGR, learning rates as well as initial binding and repair parameters were set equally (Appendix C.2). Parameter estimation was repeated over 500 iterations.

5.4.5 PCA and Cosine Analysis

We measured the functional interdependence between the repair parameters as described previously in Chapter 4. However, the parameter distributions were much broader and more biased. We applied a Box-Cox log-transformation, and centred the mean at zero with unit variance. Unfortunately some of the transformed parameter distributions did not approximate a Gaussian bell curve (Figure 5.13(A)). As discussed in Chapter 4, the established functional relationship can be only seen as indicative due to possible violations in the normality assumption. Nonetheless, we assumed that they were reasonably close to apply a PCA transformation and to make the results comparable with the traffic repair model. The Pareto analysis revealed that 3 dimensions are necessary to explain at least 80% of the variance (Figure 5.13(B)). All other steps were identical to Chapter 4.

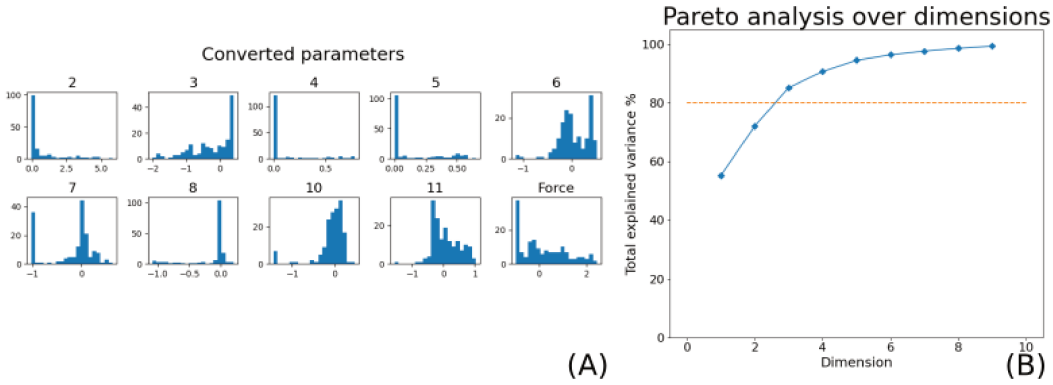


Figure 5.13: **Parameter transformation to near-Gaussian and Pareto analysis for *GillesPy* repair.** (A) Parameters are converted to a near-Gaussian distribution using a Box-Cox transform. However, some distributions are not adequately transformed, particularly parameters 3, 4, and 5 which govern Pol dynamics. We continued with the PCA transformation to allow comparability to the repair traffic model. Parameters are: 2: Pol II association. 3: Pol II movement. 4: Pol II stalling. 5: Pol II dissociation. 6: Rad4 association. 7: Rad4 dissociation. 8: Pol II stalling at Rad4. 10: TCR repair rate. 11: GGR repair rate. We removed 0 (random association), 1 (random dissociation) 9 (Pol II stalling at lesions) because these parameters were not trainable. (B) The Pareto analysis shows that 3 dimensions are necessary to describe at least 80% of the parameter variance.

Chapter 6

Discussion

In this work, we analysed CPD repair in *Saccharomyces cerevisiae* by combining top-down data evaluation approaches with bottom-up computational models. We determined first higher-level repair kinetics and influencing factors through a data-driven analysis. The mathematical modelling then permitted an improved explanation of the NGS signals, establishing the missing link between nuclear process and population-based data. By doing so, we aimed to provide a holistic understanding of spatiotemporal lesion removal kinetics, particularly at protein-coding genes. In this chapter, we want to summarise the results and methods. We close by discussing the research contribution to the work of the laboratory as well as to the wider field, and we explain how it can be extended in the future.

When analysing the coordination of nucleosome positioning along genes, we determined that chromatin remodeler WT strains should be able to largely decouple histone complex dynamics from the presence of multiunit proteins—such as Pol II—along coding regions. Such an understanding was further supported through comparing parameters of time-continuous repair with nucleosome density in transcribed regions, which did not indicate any significant impact. We therefore did not consider nucleosome positioning within the gene body in our mathematical models. However, transcription levels were significantly correlated with repair parameters, highlighting the role of damage detection by Pol II during TCR. The KJMA model also suggested a spatiotemporal change along TCR areas. We found a strong and early decline of CPD levels around the beginning of the gene, a trend which subsequently shifted towards later repair as a function of distance from the TSS. Whilst the method is unable to explain how and why the parameters vary, it nevertheless proves that repair kinetics tend to change in space. We hypothesised that the spatiotemporal dynamics are linked to DNA-protein interactions, in particular Pol II movement. In order to provide a mechanistic understanding of DNA damage removal in a single cell, we used two approaches—namely mean-field modelling and Gillespie sampling—to reproduce the data in TCR regions. Surprisingly, despite their fundamental

methodological differences, both models predict gene-specific repair dynamics. This means that parameters change with respect to an unknown repair-influencing factor that varies between coding regions. Moreover, both approaches suggest that a substantial subset of genes requires a large presence of GGR—or another repair factor that is not linked to Pol II-mediated TCR dynamics—to describe the CPD data evolution on a population-wide level. The presented stochastic *GillesPy* algorithm can link cell-specific mechanistic dynamics with static NGS data. The trained parameters indicated that GGR-related recognition proteins—such as Rad4—need to associate and dissociate statistically many times along the DNA before the lesion site is found. Whilst random interactions can sometimes lead to rapid CPD removal in single cells, it takes considerably longer to observe a GGR-driven change on a population scale. Pol II-mediated TCR, on the other hand, benefits from a systematic scanning approach through directed movement, and the effects of TCR become more quickly observable in NGS data. Although GGR-specific proteins might swiftly interact with the DNA and be present at high levels, the directed elongation dynamics do clearly allow on average a quicker damage recognition.

Surprisingly, the traffic repair model fails to link Pol II dynamics to the CPD decrease when data was not normalised adequately. We demonstrated that appropriate scaling needs to be specifically taken into account (Appendix B.9). When comparing with the number of induced CPDs per kb reported in the literature (Bucceri et al. (2006)), we concluded that the induction of damage is a rare event. In fact, one cannot expect more than a 1000 cells to contain a lesion at an arbitrary site in a cell culture of 10 million cells, and most of them do not possess damage at a given position. Similarly, we suppose that there is $\approx$1 Pol II complex per kb which is currently engaging in elongation (Struhl (2007)). As the values of presence in a single cell are low, we were wondering whether a stochastic and cell-specific treatment of nuclear processes might be necessary to explain CPD repair. Amazingly, we can report that the predictions of the probabilistic approach are in line with the traffic repair model.

Both presented bottom-up methods are highly general, although the traffic repair model was specifically implemented for repair. The *GillesPy* simulation and training framework, on the other hand, is independent of the actual implemented pathway, and it can be used for any representation of particle interactions along a one-dimensional polymer. This point is emphasised by the fact that it can solve the gateway problem, which we introduced to evaluate the method's behaviour to approximate the data. Although it takes undoubtedly more time than the traffic repair model, it allows to link single-cell dynamics with static population-wide sequencing data. The predictions of both methods are in agreement with each other. However, the mean-field approximation falls short when providing a mechanistic understanding. Therefore, both the traffic model and the *GillesPy* framework have their

own benefits and disadvantages. Future work on DNA repair—or other nuclear pathways—should consider both options depending on their use case.

With this study, we provided new insights into CPD repair in *Saccharomyces cerevisiae*, particularly in context of spatiotemporal changes along transcribed regions. The combination of top-down analysis with bottom-up modelling yields a phenomenological and mechanistic perspective on damage removal by NER. Our mathematical models can be generalised in many aspects. Consequently, the work does not only contribute to the group's projects but also to the wider research field. Indeed, we can distinguish between methodological and scientific contributions. Despite the focus on CPD repair at genes in this work, interpretations and algorithms can be applied beyond the scope. We want to emphasise that all models are species-independent, and they can be equally applied to other organisms, including human-cell data. Nonetheless, their treatment can be substantially more challenging. The repair process in human cells is, albeit similar, not identical to NER in yeast. Adaptations are undoubtedly necessary. Moreover, the human genome is two orders of magnitude larger than of *Saccharomyces cerevisiae*. Fortunately, binning or segmenting approaches can be remedially applied. With the adequate changes, the same models can be applied to analyse lesion removal in human cells. In the following, we want to elaborate in more detail possible extensions and applicability in future work.

6.1 Contribution to the Laboratory's Research

The research objective of the laboratory targets molecular mechanisms of eukaryotic transcriptional regulation and transcription-coupled processes in the context of chromatin conformation *in vivo* on a genome-wide level. By focusing on TCR kinetics, the project contributed to the agenda through providing a holistic understanding of the repair process in protein-coding regions.

By applying computational methods to publicly available and lab-internal data, we were able to disentangle repair-specific dynamics from other genomic processes, such as transcription and chromatin conformation. This has been proven to be especially valuable in context of the KJMA model (Zeitler et al. (2022)), where we could find known influencing factors as well as establish new links to other nuclear properties. We therefore investigated transcription-related lesion removal within its genomic context, and the presented work has contributed directly to the laboratory's research agenda.

The laboratory has always been engaged in NGS data analyses, and necessary evaluation frameworks were already set in place before the project commenced. Initial CPD-seq and Pol II ChIP-seq signals were available right from the beginning, which considerably supported an efficient start into the work. This research project builds up on the existing lab-internal infrastructure by appending to

the bio-informatics processing pipelines several sophisticated modelling approaches to deepen and widen the understanding of the repair kinetics on a genomic scale. We incorporated particular steps of the data acquisition protocol into our NGS analysis methods to account for technical details. This includes taking into account DNA fragment sizes during sonication and considering the entire cell culture during data collection. Due to our work on different levels of details, we deem the distinction between population-wide observable trends and mechanistic effects in single cells to be important. This promoted a better understanding of the data themselves.

Whilst the group worked already before on mathematical models of transcription—particularly on the PIC assembly—a combination of population-wide sequencing data with a bottom-up mechanistic description of DNA repair was largely lacking. Here, we provide two different dynamical models that explain the temporal data evolution via protein movements and DNA-protein interactions. Amazingly, both models yield the same conclusions, therefore significantly supporting the lesion removal kinetics hypothesised by the laboratory and linking single-cell mechanisms with population-wide data, which the group is specialised in.

6.2 Methodological Contributions to Computational Biology

Our findings indicate that an interpretation of NGS data as the superposition of individual cells might be convenient when measured quantities are supposedly low. By referencing reported values in the literature (Bucceri et al. (2006)), we can mathematically indicate that the expected number of cells containing a CPD at a given position in an entire population might be small in *Saccharomyces cerevisiae*. Viewing the data as cell superposition was fundamental for the development of the KJMA model (Chapter 3) and the *GillesPy* algorithm (Chapter 5). In fact, appropriate scaling for the traffic repair model could only be achieved using this understanding in the *GillesPy* algorithm (Appendix B.9). This might be even more important for the damage distribution in other species, as suggested by reported CPD rates in *Caenorhabditis elegans* (Meyer et al. (2007)). For modelling repair or similar cases, the data should be understood as the result of several cells stacked together, and stochastic effects might be essential to explain the data. Such a distinction for evaluating NGS data has not been proposed to our knowledge, and it could prove to be important beyond this work.

We combined classical Pearson correlation measurements with fPCA to evaluate nucleosome dynamics in WT and mutant conditions (Chapter 2). By doing so, we can characterise the two major functional descriptors of MNase-seq data, which is a location-specific evaluation of nucleosome phasing. The analysis indicated that local processes—such as presence of Pol II—might only negligibly affect histone complex positioning; at least for what can be sensibly measured using NGS data.

The application of a location-specific analysis was essential for quantifying the effect on coordinated nucleosome arrangement in mutants. This permitted the establishment of a mechanistic model over the local and long-range influences of chromatin remodelers on phasing. FPCA can be applied in general for any type of functional data independent of its form. Although predominantly used for the analysis of time series, we strongly encourage the use to study spatial correlation of nuclear properties along the genome. The fPCA scores offer a target for functional clustering. It is therefore a sensible choice for the grouping of genes based on the sequencing data of all kinds.

The KJMA model was motivated by the state transition from a damaged to a repaired position in the DNA (Chapter 3). We can recover missing temporal information by fitting an S-shaped function to the data. Although we only permit a constant repair rate per position (therefore ignoring the existence of two recognition pathways which possibly act with different rates), we can find two distinct dynamics when taking the average over all considered regions. We presume that the observed phenomena are early-acting TCR and late GGR, and hence, the method recovers mechanisms that were not explicitly incorporated. We are convinced that the KJMA model is a sensible choice for any process that can be understood as an irreversible binary state transition. This includes DNA replication and first-passage problems for protein binding. The straightforward parameter estimation using linear regression permits an efficient run time execution even on conventional personal computers, therefore allowing a quick evaluation and analysis. The applied methods are largely independent of the actual NER pathway. Moreover, the found parameters describe the entire temporal process, rather than a static time point captured by NGS. When correlating parameter values with other nuclear properties, we can find known repair-influencing factors as well as establish new correlations. The KJMA model allows therefore an unbiased embedding and comparison of a process with respect to its nuclear context, which could prove to be beneficial in any *in vivo* data analysis.

Although the traffic repair model was customised to represent NER (Chapter 4), the essential protein interactions—i.e. association, dissociation and movement—are independent of the process. Despite not being shown here, we evaluated equally the performance for nucleosome dynamics and Pol II transcription with promising preliminary results. The fact that it has been previously applied for various polymerisation processes in biology (Hinsch et al. (2007); Davis et al. (2014)) further encourages the development. Future projects could aim to extend the traffic model to a general simulation and training framework (similar to the *GillesPy* approach), which provides a user-friendly interface to model and evaluate NGS data independent of the actual DNA-protein interactions. In fact, we already considered a general traffic model that can be applied to all sequencing data using a Physics-Informed Neural Network (PINN) (Raissi et al. (2017a,b)). However, a sensible generalisation has not been achieved so far, as we could not sufficiently restrict the parameter approximation. Future work

could particularly address different approaches to link the parameters in a process-dependent manner. Alternatively, they could assess different machine learning architectures and their performance to describe protein movements through the equations in the traffic model.

We demonstrated pathway-unspecific functioning for the *GillesPy* algorithm by solving the gateway problem (Chapter 5). Other nuclear processes can be easily implemented and customised. The framework accepts the definition of the interactions and loading of the data in a separate file, which can be imported during run time. Consequently, it can be used with limited programming skills following a *GillesPy*-specific syntax to describe the particle interactions, for which we provide various examples in our repository. The algorithm converts the rules to the simulation system and finds the best parameters approximating the data automatically. This procedure can be equally applied to other species and theoretically even to any type of particle interaction along a one-dimensional substrate. However, as discussed in detail in Chapter 5, it is pivotal to provide reasonable hyperparameters and a sensible parameter initialisation. A basic knowledge of machine learning methods could help to find good values. Future work could target incorporating an exploration-exploitation trade-off during parameter search similar to RL approaches. This would allow a broad evaluation of different parameter regimes (similar to Bayesian inference) whilst applying a gradient-based approach which tends to require less resources. However, our own preliminary attempts did not result in a more robust parameter estimation and instead required only longer training time. The current implementation focused on model interpretability and user-friendliness for the trade-off of performance. The framework could significantly benefit from using a compiled programming language, such as C++. Similarly, the current Python code could gain performance by extensively leveraging tensor operations and parallelisation methods. However, it should be stressed that those adaptations should not impede a broad utilisation among researchers, for which we deem the ease of use to be important. We highly encourage future projects to use the *GillesPy* algorithm to evaluate their mechanistic hypotheses. This could not only provide a deepened understanding of the process in question itself, but it also would evaluate the framework in different contexts. This could further highlight necessary improvements of the approach, concerning e.g. forward simulation, parameter training techniques, and visualisation.

6.3 Scientific Contributions to Understanding Biological Processes

Indeed, the modelling approaches presented in this work did not only address technical issues and engineering problems, but they permitted scientific conclusions about the nuclear processes.

Through combining classical Pearson correlation indices with fPCA, we could provide a new mechanistic perspective on the functional role of the chromatin remodeling complexes RSC and

Chd1. Whilst RSC is a key player for globally decoupling gene-specific nucleosome phasing, Chd1 acts locally by allowing phasing independent of the presence of large proteins such as Pol II. Particularly the effect in *rsc8*-depleted cells on nucleosome positioning that exceeds the actual gene size could prove to be important for understanding coordinated arrangement with respect to a barrier in the core promoter. By using the same methodology, we analysed MNase-seq data in Mediator mutants which significantly impact interaction and colocalisation with RSC in the promoter region (not shown, data from André et al. (2023)). Our preliminary results indicate that coordinated nucleosome phasing is dramatically altered in Med17 point mutants, further supporting the notion of a RSC-related barrier model. Future projects could build up on our existing work to develop a mechanistic understanding of the establishment of such a barrier in the NDR. Similarly, the results can be used as a basis for developing models for the three-dimensional chromatin conformation by leveraging dynamic polymer simulations that take into account the nucleosome distribution in different mutants.

The KJMA model provides a description of the repair kinetics in time using only some few NGS data points. Linking the parameters—which describe now the entire temporal process—to other nuclear properties indicated necessary factors that need to be considered in our bottom-up models. It also suggested a new link between lesion removal dynamics and gene size, despite the fact that we normalised over the transcription unit length. Such a correlation had not been considered before to our knowledge. It remains an open question in what way the gene size is related to DNA repair. It should be mentioned that we described repair with respect to three equally-sized bins per gene. However, we obtained the same results when considering the entire coding region. As the KJMA model already suggested an increasing influence of later-acting repair as a function of distance from the TSS, a correlation with size could simply reflect exactly the same observation. If this is true, future projects could address why repair rates decrease when being further away from the beginning of the gene. However, it is also possible that there is a genuine link between NER and the size of a transcribed region. For example, it has been shown that shorter genes tend to be evolutionarily younger and are subject to a higher selection pressure (Vishnoi et al. (2010); Grishkevich and Yanai (2014)). It is conceivable that this also translates to DNA repair, which could be targeted by future work.

We developed two mathematical modelling approaches to explain DNA lesion removal kinetics through protein dynamics. By using the presented *GillesPy* framework, we can draw conclusions about single cell dynamics. We include data and findings from the literature to establish the missing link for a mechanistic understanding. However, the only observable to verify our models (i.e. sequencing data) comes from entire cell cultures, and a baseline model—which could support validity and compare performance of our methods—is missing. Our results are therefore exclusively based

on the dynamics in millions of cells at the same time, and translating these kinetics to single cells is not straightforward. Taking this into account, it is extraordinary that both of our models—i.e. the mean-field traffic repair model and the stochastic *GillesPy* algorithm—predict similar repair dynamics along genes. This even includes the behaviour of hidden parameters that govern GGR. It improves our confidence in the mechanistic conclusions in single cells. These are in particular gene-specific repair; and that GGR-related proteins might be present earlier or at higher rates than previously appreciated. We also established that there is a functional relationship between the repair parameters which indicates the presence of another external factor that we had not accounted for. Whilst nucleosome positioning—which we did not include in our bottom-up models—can indeed have an impact, we find it unlikely that they can explain exclusively the observed trend due to the findings presented in Chapter 2 and 3. Overall, the results indicate the existence of specific regulatory mechanisms to coordinate lesion removal on a genomic scale. We hypothesise that another parameter—such as interactions between Mediator and repair proteins (Eyboulet et al. (2013)), histone marks (Sun et al. (2020)), or similar—highly affect gene-specific repair. We strongly hope that future projects will address this point from a mathematical modelling as well as an experimental biological perspective.

We established that the different repair times for TCR and GGR are linked to the protein movement along the gene. The predicted early and uniform visiting times of GGR recognition proteins need ultimately experimental confirmation based on single-cell data, which could be produced by work building up on our results. Despite the fact that single-cell sequencing could indeed shine light onto the cell-dependent dynamics, we want to emphasise that population-based NGS has its own benefits. In fact, they combine multiple cell states together and are therefore ideally suited for time-continuous models due to the continuous overlap of state changes. NGS data were therefore a sensible choice for modelling a temporal multistep process such as NER. Nonetheless, we want to encourage future research projects to understand the interplay of different recognition pathways along coding regions on a single-cell scale.

Our models focused on representing DNA repair along genes. They can be equally used for damage removal in non-coding genes. Naturally, the DNA-protein dynamics need to be adapted to the specific region. For example, protein movements along the DNA must be removed if no polymerase is assumed to participate. However, the general training and simulation implementation for the traffic repair model and the *GillesPy* algorithm as well as the data do not need to be changed. Future work can target repair in non-coding regions by leveraging our methods.

Computational models can be harnessed to make predictions in varying contexts. By changing the parameters or the modelled protein interactions, it is possible to make projections about the repair process in mutant conditions. This can be straightforwardly done for the traffic repair model.

The implementation of the *GillesPy* framework, however, requires re-training the model parameters, as they are all linked to each other and determine the sampling rate. Nonetheless, the relatively easy modification of these methods can be particularly helpful for evaluating consequences related to human diseases. For example, our models allocate an important role to GGR in a subset of genes. Indeed, lab-internal data suggests that *rad7*Δ strains are more UV-sensitive. Mutations in the human homologue XPC of the GGR-related protein Rad4 can lead to Xeroderma Pigmentosum (XP). Together with the fact that XPC also acts as a Pol II co-factor for some genes (Bidon et al. (2018)), this could possibly suggest an altered TCR behaviour along protein-coding genes in XP patients. Consequently, the phenotype related to XP might be not only linked to damages in non-coding and regulatory sequences, but also to changing TCR dynamics. Future work could make use of our models for understanding repair in changing genetic contexts to better understand human disorders related to NER.

6.4 Concluding Remarks

In this work, we aimed to provide new perspectives on NER dynamics in living cells. We can establish known and unknown influencing factors (Chapter 3) as well as provide evidence that nucleosome arrangement might not dramatically alter repair along coding regions (Chapter 2). When aiming to reproduce the data with specific particle interactions implemented in different bottom-up models (Chapters 4 and 5), we reveal that repair parameters are functionally correlated and gene-specific. This is highly important as this implies that observed damage removal is not only dependent on the initial state at coding regions. Consequently, the *in vitro* dynamics—on which the approaches are based—are different to *in vivo* kinetics, and repair in living cells is influenced by additional parameters. It could be conjectured that the mechanism is influenced by the temporally-changing stress response, and parameters can only be reasonably understood as a function of time. However, we object such an interpretation. We believe that the strong functional correlation of the static repair parameters in both models reveals the importance of another influencing factor that we had not accounted for and which is gene-specific. The KJMA model indicated a dependence on transcription unit length; yet a correlation analysis with gene size did not reveal any appreciable interdependence (data not shown). It should be mentioned that the binning approach applied in Chapters 4 and 5 could have removed such a functional dependence. Future experimental work will be necessary to address gene-specific repair and to identify the factor (or even *several* factors) that regulate DNA repair *in vivo*.

The manuscript opened with a philosophical perspective, and we also want to close on a sim-

ilar tone. In fact, it is directly linked to the approaches we applied and the results we obtained. When analysing the influence of the *GillesPy* parameters over all coding-regions using PCA and cosine similarity measurements, we can once again show that they are functionally linked, further supporting the notion of gene-specific repair. However, the determined clusters differ fundamentally from the groups identified in Chapter 4. This can have indeed various reasons. Nonetheless, we want to emphasise that the two modelling approaches make different assumptions which have not only mathematical but also deeply philosophical implications. Since the traffic repair model is a mean-field approach, it considers that the lesion removal dynamics are identical in each cell. The differences come from independent and uncorrelated noise which we can average out. This implies that experimental procedures are reproducible and always yield the same result if the signal-to-noise ratio in the sequencing data allows a sensible analysis. The *GillesPy* algorithm, however, considers that every single cell repairs their lesions individually and independently by stochastic particle interactions. Therefore, the result (i.e. repair) can be achieved through varying sequences of DNA-protein interactions. Fluctuations in the population-based data are not only a result from white noise induced during the measurements, but they are also specifically linked to various cell states that traverse all through different state transformations. A particular cell state can be defined as the damaged DNA with the associated proteins. A rough calculation over a gene of 1000bp length, 2 proteins and one type of damage, all of which can only be either present or absent, yields $\approx 2^{3^{1000}}$ different configurations. To provide a comparison, this is more than 10^{476} orders of magnitude larger than the amount of atoms in the known universe. These are humbling numbers. The calculation is admittedly a strong simplification, as this assumes that damage can be present at every single position. Nonetheless, this large number has two important implications. Firstly, the *GillesPy* algorithm might find different repair parameters based on the initialisation—all yielding good agreement with the data—that would lead to changing gene groups after clustering. Secondly, this understanding of the process has direct implications for the reproducibility of results acquired over a cell population. Even when providing an extensive sequencing data set over one million experiments (each involving approximately 10 million cells), we would be still not able to reach a number that would allow for significant predictions when assuming all states are possible, at least technically speaking. The two models represent two fundamentally different views. Whilst on a macro-scale they are in good agreement, it is more than understandable that there are different on a micro-scale, i.e. the exact single-cell repair dynamics that lead to gene-specific damage removal.

Acronyms

Biological Acronyms

Table 6.1: **Biological Acronyms.** Capital letters in the acronym are also capitalised in the explanation. Protein names were not included.

Acronym	Explanation
6-4PP	6-4 Photoproducts
ARS	Autonomous Replication Sequences
asRNA	antisense RiboNucleic Acid
ATP	Adenosine TriphosPhate
BER	Base Excision Repair
bp	base pair
cDNA	complementary DeoxyRibonucleic Acid
ChIP-seq	Chromatin Immonuprecipitation sequencing
CID	Chromosome Interaction Domains
COFS	Cerebro-Oculo-Facio-Skeletal Syndrome
CPD	Cyclobutane Pyrimidine Dimers
CS	Cockayne Syndrome
DSB	Double-Strand Breaks

Continued on next page

Table 6.1: **Biological Acronyms.** Capital letters in the acronym are also capitalised in the explanation. Protein names were not included. (Continued)

Acronym	Explanation
DNA	DeoxyRibonucleic Acid
FDA	Food and Drug Administration
GFP	Green Fluorescence Protein
GGR	Global-Genome Repair
IGR	InterGenic Region (only abbreviated in Chapter 3)
kb	kilobase pair
Mb	Mega base pair
mRNA	messenger RiboNucleic Acid
ncRNA	non-coding RiboNucleic Acid
NDR (NFR)	Nucleosome Depleted Region (equiv. Nucleosome Free Region)
NER	Nucleotide Excision Repair
NGS	Next Generation Sequencing
nt	nucleotide
NTS	Non-Transcribed Strand
ORF	Open Reading Frame
OTC	Ornithine-TransCarbamylase
PCR	Polymerase Chain Reaction
PIC	Pre-Initiation Complex
PR	PhotoReactivation
ROS	Reactive Oxygen Species
rRNA	ribosomal RiboNucleic Acid
SBS	Sequencing By Synthesis
SCID	Severe Combined Immunodeficiency

Continued on next page

Table 6.1: **Biological Acronyms.** Capital letters in the acronym are also capitalised in the explanation. Protein names were not included. (Continued)

Acronym	Explanation
SNP	Single Nucleotide Polymorphisms
TBL	Transcription Blocking Lesions
TCR	Transcription Coupled Repair
TF	Transcription Factor
tRNA	transfer RiboNucleic Acid
TS	Transcribed Strand
TSS	Transcription Starting Site
TTS, (TES)	Transcription Termination Site (equiv. Transcription Ending Site)
TU	Transcription Unit
UAS	Upstream Activating Sequence
URS	Upstream Repressing Sequence
UTR	UnTranslated Regions
UV	UltraViolet
UVSS	UV-Sensitivity Syndrome
WT	WildType
XP	Xeroderma Pigmentosum
XR-seq	eXcision Repair sequencing

Mathematical and Computational Acronyms

Table 6.2: **Mathematical and Computational Acronyms.** Capital letters in the acronym are also capitalised in the explanation.

Acronym	Explanation
BPTT	BackPropagation Through Time
DC	Distance Correlation
fPC	functional Principal Component
fPCA	functional Principal Component Analysis
GGRP	abstract Global Genome Repair Protein
JS distance	Jensen-Shannon distance
KJMA model	Kolmogorov-Johnson-Mehl-Avrami model
KL divergence	Kullback-Leibler divergence
kNN	k-Nearest Neighbour
MCMC	Markov-Chain Monte-Carlo
MSE	Mean Squared Error
NODE	Neural Ordinary Differential Equations
ODE	Ordinary Differential Equation
PC	Principal Component Analysis
PCA	Principal Component Analysis
PI	Prediction Interval
RL	Reinforcement Learning
SDE	Stochastic Differential Equation
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine

Continued on next page

Table 6.2: **Mathematical and Computational Acronyms.** Capital letters in the acronym are also capitalised in the explanation. (Continued)

Acronym	Explanation
t-SNE	t-distributed stochastic neighborhood embedding
TCRP	abstract Transcription Coupled Repair Protein
UMAP	uniform manifold approximation and projection

Bibliography

Allfrey, V. G., Faulkner, R., and Mirsky, A. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences*, 51(5):786–794.

Anderson, W. F. (1990). September 14, 1990: the beginning. *Human gene therapy*, 1(4):371–372.

André, K. M., Aiach, N. G., Martinez-Fernandez, V., Zeitler, L., Alberti, A., Goldar, A., Werner, M., Wilkes, C. D., and Soutourina, J. (2023). Functional interplay between Mediator and RSC chromatin remodeling complex controls nucleosome-depleted region maintenance at promoters. *Cell Reports*, 42(5).

André, K. M., Sipos, E. H., and Soutourina, J. (2021). Mediator roles going beyond transcription. *Trends in Genetics*, 37(3):224–234.

Arbona, J.-M., Kabalane, H., Goldar, A., Hyrien, O., and Audit, B. (2021). Neural network and kinetic modelling of human genome replication reveal replication origin locations and strengths. *bioRxiv*, pages 2021–12.

Arsuaga, J., Vázquez, M., Trigueros, S., Sumners, D. W., and Roca, J. (2002). Knotting probability of DNA molecules confined in restricted volumes: DNA knotting in phage capsids. *Proceedings of the National Academy of Sciences*, 99(8):5373–5377.

Avery, O., MacLeod, C., and McCarty, M. (1944). Studies on the chemical nature of the substance causing transformation of the pneumococcal types. Induction by a desoxyribonucleic acid fraction isolated from pneumococcus Type III. *J. Exp. Med*, 79:137–158.

Avrami, M. (1939). Kinetics of phase change. I General theory. *The Journal of chemical physics*, 7(12):1103–1112.

Avrami, M. (1940). Kinetics of phase change. II transformation-time relations for random distribution of nuclei. *The Journal of chemical physics*, 8(2):212–224.

Avrami, M. (1941). Granulation, phase change, and microstructure kinetics of phase change. III. *The Journal of chemical physics*, 9(2):177–184.

Bachelier, L. (1900). Théorie de la spéculation. In *Annales scientifiques de l'École normale supérieure*, volume 17, pages 21–86.

Badis, G., Chan, E. T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C. D., Gossett, A. J., Hasinoff, M. J., Warren, C. L., et al. (2008). A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Molecular cell*, 32(6):878–887.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395.

Behjati, S. and Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, 98(6):236–238.

Bernstein, D. (2005). Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm. *Physical Review E*, 71(4):041103.

Bidon, B., Iltis, I., Semer, M., Nagy, Z., Larnicol, A., Cribier, A., Benkirane, M., Coin, F., Egly, J., and Le May, N. (2018). XPC is an RNA polymerase II cofactor recruiting ATAC to promoters by interacting with E2F1. *Nature communications*, 9(1):2610.

Birrell, G. W., Brown, J. A., Wu, H. I., Giaever, G., Chu, A. M., Davis, R. W., and Brown, J. M. (2002). Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. *Proceedings of the National Academy of Sciences*, 99(13):8778–8783.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Bohm, K. A., Morledge-Hampton, B., Stevison, S., Mao, P., Roberts, S. A., and Wyrick, J. J. (2023). Genome-wide maps of rare and atypical UV photoproducts reveal distinct patterns of damage formation and mutagenesis in yeast chromatin. *Proceedings of the National Academy of Sciences*, 120(10):e2216907120.

Bohr, V. A., Smith, C. A., Okumoto, D. S., and Hanawalt, P. C. (1985). DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. *Cell*, 40(2):359–369.

Boiteux, S. and Jinks-Robertson, S. (2013). DNA repair mechanisms and the bypass of DNA damage in Saccharomyces cerevisiae. *Genetics*, 193(4):1025–1064.

Borggrefe, T., Davis, R., Bareket-Samish, A., and Kornberg, R. D. (2001). Quantitation of the RNA polymerase II transcription machinery in yeast. *Journal of Biological Chemistry*, 276(50):47150–47153.

Brégeon, D. and Doetsch, P. W. (2011). Transcriptional mutagenesis: causes and involvement in tumour development. *Nature Reviews Cancer*, 11(3):218–227.

Brooks, P. J. (1997). DNA damage, DNA repair, and alcohol toxicity—a review. *Alcoholism: Clinical and Experimental Research*, 21(6):1073–1082.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Bucceri, A., Kapitza, K., and Thoma, F. (2006). Rapid accessibility of nucleosomal DNA in yeast on a second time scale. *The EMBO journal*, 25(13):3123–3132.

Budd, M. E. and Campbell, J. L. (1995). DNA polymerases required for repair of UV-induced damage in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 15(4):2173–2179.

Bushnell, D. A. and Kornberg, R. D. (2003). Complete, 12-subunit RNA polymerase II at 4.1-Å resolution: implications for the initiation of transcription. *Proceedings of the National Academy of Sciences*, 100(12):6969–6973.

Cao, Y., Gillespie, D. T., and Petzold, L. R. (2005). The slow-scale stochastic simulation algorithm. *The Journal of chemical physics*, 122(1):014116.

Casal-Mouriño, A., Ruano-Ravina, A., Torres-Durán, M., Parente-Lamelas, I., Provencio-Pulla, M., Castro-Añón, O., Vidal-García, I., Abal-Arca, J., Piñeiro-Lamas, M., Fernández-Villar, A., et al. (2020). Polymorphisms in the BER and NER pathways and their influence on survival and toxicity in never-smokers with lung cancer. *Scientific Reports*, 10(1):21147.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Chen, R. T. Q. (2018). torchdiffeq. `https://github.com/rtqichen/torchdiffeq`.

Chou, T. (2003). Ribosome recycling, diffusion, and mRNA loop formation in translational regulation. *Biophysical Journal*, 85(2):755–773.

Clapier, C. R., Iwasa, J., Cairns, B. R., and Peterson, C. L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature reviews Molecular cell biology*, 18(7):407–422.

Compe, E. and Egly, J.-M. (2012). TFIIH: when transcription met DNA repair. *Nature reviews Molecular cell biology*, 13(6):343–354.

Cozzolino, F., Iacobucci, I., Monaco, V., and Monti, M. (2021). Protein–DNA/RNA Interactions: An Overview of Investigation Methods in the-Omics Era. *Journal of Proteome Research*, 20(6):3018–3030.

Cramer, P., Bushnell, D. A., and Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 Ångstrom resolution. *Science*, 292(5523):1863–1876.

Davis, L., Gedeon, T., Gedeon, J., and Thorenson, J. (2014). A traffic flow model for biopolymerization processes. *Journal of mathematical biology*, 68:667–700.

Deaconescu, A. M., Chambers, A. L., Smith, A. J., Nickels, B. E., Hochschild, A., Savery, N. J., and Darst, S. A. (2006). Structural basis for bacterial transcription-coupled DNA repair. *Cell*, 124(3):507–520.

den Dulk, B., Brandsma, J. A., and Brouwer, J. (2005). The Rad4 homologue YDR314C is essential for strand-specific repair of RNA polymerase I-transcribed rDNA in Saccharomyces cerevisiae. *Molecular microbiology*, 56(6):1518–1526.

den Dulk, B., Sun, S. M., de Ruijter, M., Brandsma, J. A., and Brouwer, J. (2006). Rad33, a new factor involved in nucleotide excision repair in Saccharomyces cerevisae. *DNA repair*, 5(6):683–692.

DiGiovanna, J. J. and Kraemer, K. H. (2012). Shining a light on xeroderma pigmentosum. *Journal of investigative dermatology*, 132(3):785–796.

Ding, J., Taylor, M. S., Jackson, A. P., and Reijns, M. A. (2015). Genome-wide mapping of embedded ribonucleotides and other noncanonical nucleotides using emRiboSeq and EndoSeq. *Nature protocols*, 10(9):1433–1444.

Dion, M. F., Kaplan, T., Kim, M., Buratowski, S., Friedman, N., and Rando, O. J. (2007). Dynamics of replication-independent histone turnover in budding yeast. *Science*, 315(5817):1405–1408.

Duan, M., Selvam, K., Wyrick, J. J., and Mao, P. (2020). Genome-wide role of Rad26 in promoting transcription-coupled nucleotide excision repair in yeast chromatin. *Proceedings of the National Academy of Sciences*, 117(31):18608–18616.

Duina, A. A., Miller, M. E., and Keeney, J. B. (2014). Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. *Genetics*, 197(1):33–48.

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der physik*, 4.

Erixon, K. and Ahnström, G. (1979). Single-strand breaks in DNA during repair of UV-induced damage in normal human and xeroderma pigmentosum cells as determined by alkaline DNA unwinding and hydroxylapatite chromatography: Effects of hydroxyurea, 5-fluorodeoxyuridine and 1-β-d-arabinofuranosylcytosine on the kinetics of repair. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 59(2):257–271.

Evans, E., Moggs, J. G., Hwang, J. R., Egly, J.-M., and Wood, R. D. (1997). Mechanism of open complex and dual incision formation by human nucleotide excision repair factors. *The EMBO journal*, 16(21):6559–6573.

Eyboulet, F., Cibot, C., Eychenne, T., Neil, H., Alibert, O., Werner, M., and Soutourina, J. (2013). Mediator links transcription and DNA repair by facilitating Rad2/XPG recruitment. *Genes & development*, 27(23):2549–2562.

Feaver, W. J., Svejstrup, J. Q., Bardwell, L., Bardwell, A. J., Buratowski, S., Gulyas, K. D., Donahue, T. F., Friedberg, E. C., and Kornberg, R. D. (1993). Dual roles of a multiprotein complex from S. cerevisiae in transcription and DNA repair. *Cell*, 75(7):1379–1387.

Geiger, B. C. and Kubin, G. (2013). Signal enhancement as minimization of relevant information loss. In *SCC 2013; 9th International ITG Conference on Systems, Communication and Coding*, pages 1–6. VDE.

Gilbert, W. (1978). Why genes in pieces? *Nature*, 271:501–501.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361.

Gillette, T. G., Yu, S., Zhou, Z., Waters, R., Johnston, S. A., and Reed, S. H. (2006). Distinct functions of the ubiquitin–proteasome pathway influence nucleotide excision repair. *The EMBO journal*, 25(11):2529–2538.

Goodsell, D. S. (2001). The molecular perspective: ultraviolet light and pyrimidine dimers. *Stem cells*, 19(4):348–349.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

Gopaul, D., Wilkes, C. D., Goldar, A., Aiach, N. G., Barrault, M.-B., Novikova, E., and Soutourina, J. (2022). Genomic analysis of Rad26 and Rad1-Rad10 reveals differences in their dependence on Mediator and RNA polymerase II. *Genome Research*, pages gr–276371.

Gregersen, L. H. and Svejstrup, J. Q. (2018). The cellular response to transcription-blocking DNA damage. *Trends in Biochemical Sciences*, 43(5):327–341.

Grishkevich, V. and Yanai, I. (2014). Gene length and expression level shape genomic novelties. *Genome research*, 24(9):1497–1503.

Guintini, L., Charton, R., Peyresaubes, F., Thoma, F., and Conconi, A. (2015). Nucleosome positioning, nucleotide excision repair and photoreactivation in Saccharomyces cerevisiae. *DNA repair*, 36:98–104.

Guzder, S. N., Habraken, Y., Sung, P., Prakash, L., and Prakash, S. (1995). Reconstitution of Yeast Nucleotide Excision Repair with Purified Rad Proteins, Replication Protein A, and Transcription Factor TFIIH. *Journal of Biological Chemistry*, 270(22):12973–12976.

Habraken, Y., Sung, P., Prakash, L., and Prakash, S. (1993). Yeast excision repair gene RAD2 encodes a single-stranded DNA endonuclease. *Nature*, 366(6453):365–368.

Hahn, S. and Young, E. T. (2011). Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, 189(3):705–736.

Harrison, L. G. (1993). *Kinetic theory of living pattern*. Number 28. Cambridge University Press.

Hauser, M., Abraham, P. E., Barcelona, L., and Becker, J. M. (2019). UV laser-induced, time-resolved transcriptome responses of Saccharomyces cerevisiae. *G3: Genes, Genomes, Genetics*, 9(8):2549–2560.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., and Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, 56(2):61–77.

Heinrich, R. and Rapoport, T. A. (1980). Mathematical modelling of translation of mRNA in eucaryotes; steady states, time-dependent processes and application to reticulocytest. *Journal of theoretical biology*, 86(2):279–313.

Herrick, J., Jun, S., Bechhoefer, J., and Bensimon, A. (2002). Kinetic model of DNA replication in eukaryotic organisms. *Journal of molecular biology*, 320(4):741–750.

Herring, D. (2010). DNA UV mutation. `https://upload.wikimedia.org/wikipedia/commons/f/fd/DNA_UV_mutation.svg`. Online; last accessed: 11.04.23.

Hill, J. W., Hazra, T. K., Izumi, T., and Mitra, S. (2001). Stimulation of human 8-oxoguanine-DNA glycosylase by AP-endonuclease: potential coordination of the initial steps in base excision repair. *Nucleic acids research*, 29(2):430–438.

Hinsch, H., Kouyos, R., and Frey, E. (2007). From intracellular traffic to a novel class of driven lattice gas models. In *Traffic and Granular Flow'05*, pages 205–222. Springer.

Hoogstraten, D., Bergink, S., Ng, J. M., Verbiest, V. H., Luijsterburg, M. S., Geverts, B., Raams, A., Dinant, C., Hoeijmakers, J. H., Vermeulen, W., et al. (2008). Versatile DNA damage detection by the global genome nucleotide excision repair protein XPC. *Journal of cell science*, 121(17):2850–2859.

Hoogstraten, D., Nigg, A. L., Heath, H., Mullenders, L. H., van Driel, R., Hoeijmakers, J. H., Vermeulen, W., and Houtsmuller, A. B. (2002). Rapid switching of TFIIH between RNA polymerase I and II transcription and DNA repair in vivo. *Molecular cell*, 10(5):1163–1174.

Hopfield, J. J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proceedings of the National Academy of Sciences*, 71(10):4135–4139.

Hsieh, T.-H. S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O. J. (2015). Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell*, 162(1):108–119.

Hu, J., Adebali, O., Adar, S., and Sancar, A. (2017). Dynamic maps of UV damage formation and repair for the human genome. *Proceedings of the National Academy of Sciences*, 114(26):6758–6763.

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811.

Jackson, D. A., Symons, R. H., and Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences*, 69(10):2904–2909.

Jansen, A. and Verstrepen, K. J. (2011). Nucleosome positioning in Saccharomyces cerevisiae. *Microbiology and molecular biology reviews*, 75(2):301–320.

Jiang, C. and Pugh, B. F. (2009). A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genomes. *Genome biology*, 10(10):1–11.

Johnson, W. A. and Mehl, R. F. (1939). Reaction kinetics in processes of nucleation and growth. *Am. Inst. Min. Metal. Petro. Eng.*, 135:416–458.

Jones, K. L., Zhang, L., Seldeen, K. L., and Gong, F. (2010). Detection of bulky DNA lesions: DDB2 at the interface of chromatin and DNA repair in eukaryotes. *IUBMB life*, 62(11):803–811.

Juan, L.-J., Walter, P., Taylor, I., Kingston, R., and Workman, J. (1993). Nucleosome cores and histone H1 in the binding of GAL4 derivatives and the reactivation of transcription from nucleosome templates in vitro. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 58, pages 213–223. Cold Spring Harbor Laboratory Press.

Kladova, O. A., Bazlekowa-Karaban, M., Baconnais, S., Pietrement, O., Ishchenko, A. A., Matkarimov, B. T., Iakovlev, D. A., Vasenko, A., Fedorova, O. S., Le Cam, E., et al. (2018). The role of the N-terminal domain of human apurinic/apyrimidinic endonuclease 1, APE1, in DNA glycosylase stimulation. *DNA repair*, 64:10–25.

Klann, M., Ganguly, A., and Koeppl, H. (2012). Hybrid spatial Gillespie and particle tracking simulation. *Bioinformatics*, 28(18):i549–i555.

Koerber, R. T., Rhee, H. S., Jiang, C., and Pugh, B. F. (2009). Interaction of transcriptional regulators with specific nucleosomes across the Saccharomyces genome. *Molecular cell*, 35(6):889–902.

Kolmogorov, A. N. (1937). On the statistical theory of the crystallization of metals. *Bull. Acad. Sci. USSR, Math. Ser*, 1(3):355–359.

Kornberg, R. D. (1974). Chromatin Structure: A Repeating Unit of Histones and DNA: Chromatin structure is based on a repeating unit of eight histone molecules and about 200 DNA base pairs. *Science*, 184(4139):868–871.

Kumar, N., Moreno, N. C., Feltes, B. C., Menck, C. F., and Houten, B. V. (2020). Cooperation and interplay between base and nucleotide excision repair pathways: From DNA lesions to proteins. *Genetics and Molecular Biology*, 43.

Lemons, D., Gythiel, A., and Langevin's, P. (1908). Sur la théorie du mouvement brownien [On the theory of Brownian motion]. *CR Acad. Sci.(Paris)*, 146:530–533.

Li, S. and Smerdon, M. J. (2002). Rpb4 and Rpb9 mediate subpathways of transcription-coupled DNA repair in Saccharomyces cerevisiae. *The EMBO journal*, 21(21):5921–5929.

Li, W., Adebali, O., Yang, Y., Selby, C. P., and Sancar, A. (2018). Single-nucleotide resolution dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. *Proceedings of the National Academy of Sciences*, 115(15):E3408–E3415.

Lickwar, C. R., Mueller, F., and Lieb, J. D. (2013). Genome-wide measurement of protein-DNA binding dynamics using competition ChIP. *Nature protocols*, 8(7):1337–1353.

Lu, T., Volfson, D., Tsimring, L., and Hasty, J. (2004). Cellular growth and division in the Gillespie algorithm. *Systems biology*, 1(1):121–128.

Lucas-Lledó, J. I. and Lynch, M. (2009). Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. *Molecular biology and evolution*, 26(5):1143–1153.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260.

Luijsterburg, M. S., von Bornstaedt, G., Gourdin, A. M., Politi, A. Z., Moné, M. J., Warmerdam, D. O., Goedhart, J., Vermeulen, W., van Driel, R., and Höfer, T. (2010). Stochastic and reversible assembly of a multiprotein DNA repair complex ensures accurate target site recognition and efficient repair. *Journal of Cell Biology*, 189(3):445–463.

Malik, S., Chaurasia, P., Lahudkar, S., Durairaj, G., Shukla, A., and Bhaumik, S. R. (2010). Rad26p, a transcription-coupled repair factor, is recruited to the site of DNA lesion in an elongating RNA polymerase II-dependent manner in vivo. *Nucleic acids research*, 38(5):1461–1477.

Mao, P., Smerdon, M. J., Roberts, S. A., and Wyrick, J. J. (2016). Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 113(32):9057–9062.

Mao, P., Smerdon, M. J., Roberts, S. A., and Wyrick, J. J. (2020). Asymmetric repair of UV damage in nucleosomes imposes a DNA strand polarity on somatic mutations in skin cancer. *Genome research*, 30(1):12–21.

Marteijn, J. A., Lans, H., Vermeulen, W., and Hoeijmakers, J. H. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature reviews Molecular cell biology*, 15(7):465–481.

Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., Qi, J., Schuster, S. C., Albert, I., and Pugh, B. F. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome research*, 18(7):1073–1083.

Mellon, I., Spivak, G., and Hanawalt, P. C. (1987). Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. *Cell*, 51(2):241–249.

Melunis, J. and Hershberg, U. (2017). A spatially heterogeneous Gillespie algorithm modeling framework that enables individual molecule history and tracking. *Engineering Applications of Artificial Intelligence*, 62:304–311.

Mendel, G. (1865). Versuche uber pflanzen-hybriden. *Vorgelegt in den Sitzungen*.

Mendel, G. (1996). Experiments in plant hybridization (1865). *Verhandlungen des naturforschenden Vereins Brünn) Available online*.

Meyer, J. N., Boyd, W. A., Azzam, G. A., Haugen, A. C., Freedman, J. H., and Van Houten, B. (2007). Decline of nucleotide excision repair capacity in aging Caenorhabditis elegans. *Genome biology*, 8:1–17.

Michaelis, L., Menten, M. L., et al. (1913). Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369):352.

Min, J.-H. and Pavletich, N. P. (2007). Recognition of DNA damage by the Rad4 nucleotide excision repair protein. *Nature*, 449(7162):570–575.

Morgan, D. O. (2007a). Basic units of chromatin structure. `https://upload.wikimedia.org/wikipedia/commons/a/a9/Basic_units_of_chromatin_structure.svg`. [Online; last accessed: 13.03.23].

Morgan, D. O. (2007b). Basic units of chromatin structure. `https://upload.wikimedia.org/wikipedia/commons/f/f4/Histone_tails_and_their_function_in_chromatin_formation.svg`. [Online; last accessed: 13.03.23].

Mu, D., Park, C.-H., Matsunaga, T., Hsu, D. S., Reardon, J. T., and Sancar, A. (1995). Reconstitution of human DNA repair excision nuclease in a highly defined system. *Journal of Biological Chemistry*, 270(6):2415–2418.

Mukherjee, S. (2016). *The Gene: An Intimate History*. Scribner.

Nevers, A., Doyen, A., Malabat, C., Néron, B., Kergrohen, T., Jacquier, A., and Badis, G. (2018). Antisense transcriptional interference mediates condition-specific gene repression in budding yeast. *Nucleic Acids Research*, 46(12):6009–6025.

Ng, H. H., Dole, S., and Struhl, K. (2003). The Rtf1 component of the Paf1 transcriptional elongation complex is required for ubiquitination of histone H2B. *Journal of Biological Chemistry*, 278(36):33625–33628.

Ocampo, J., Chereji, R. V., Eriksson, P. R., and Clark, D. J. (2016). The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. *Nucleic acids research*, 44(10):4625–4635.

Ocampo, J., Chereji, R. V., Eriksson, P. R., and Clark, D. J. (2019). Contrasting roles of the RSC and ISW1/CHD1 chromatin remodelers in RNA polymerase II elongation and termination. *Genome research*, 29(3):407–417.

Park, D., Morris, A. R., Battenhouse, A., and Iyer, V. R. (2014). Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated noncoding RNA governed by TATA elements. *Nucleic acids research*, 42(6):3736–3749.

Pelechano, V., Chavez, S., and Perez-Ortin, J. E. (2010). A complete set of nascent transcription rates for yeast genes. *PloS one*, 5(11):e15442.

Pellman, D., McLaughlin, M. E., and Fink, G. R. (1990). TATA-dependent and TATA-independent transcription at the HIS4 gene of yeast. *Nature*, 348(6296):82–85.

Pérez-Ortín, J. E., Alepuz, P. M., and Moreno, J. (2007). Genomics and gene transcription kinetics in yeast. *TRENDS in Genetics*, 23(5):250–257.

Politi, A., Moné, M. J., Houtsmuller, A. B., Hoogstraten, D., Vermeulen, W., Heinrich, R., and van Driel, R. (2005). Mathematical modeling of nucleotide excision repair reveals efficiency of sequential assembly strategies. *Molecular cell*, 19(5):679–690.

Pontryagin, L. S. (1987). *Mathematical theory of optimal processes*. CRC press.

Portugal, J. and Rodríguez-Campos, A. (1996). T7 RNA polymerase cannot transcribe through a highly knotted DNA template. *Nucleic acids research*, 24(24):4890–4894.

Pouokam, M., Cruz, B., Burgess, S., Segal, M. R., Vazquez, M., and Arsuaga, J. (2019). The Rabl configuration limits topological entanglement of chromosomes in budding yeast. *Scientific reports*, 9(1):6795.

Prum, R. O. and Williamson, S. (2002). Reaction–diffusion models of within-feather pigmentation patterning. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1493):781–792.

Puay Yen Yap, D. T. (2017). DIRECT YEAST CELL COUNT AT OD600. `https://tipbiosystems.com/wp-content/uploads/2020/05/AN102-Yeast-Cell-Count_2019_03_17.pdf`. [Online; last accessed: 21.02.2023].

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2017a). Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations. *arXiv preprint arXiv:1711.10561*.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2017b). Physics Informed Deep Learning (Part II): Data-driven Discovery of Nonlinear Partial Differential Equations. *arXiv preprint arXiv:1711.10566*.

Rastogi, R. P., Kumar, A., Tyagi, M. B., Sinha, R. P., et al. (2010). Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair. *Journal of nucleic acids*, 2010.

Reardon, J. T. and Sancar, A. (2005). Nucleotide excision repair. *Progress in nucleic acid research and molecular biology*, 79:183–235.

Reed, S. H., You, Z., and Friedberg, E. C. (1998). The Yeast RAD7 and RAD16 Genes Are Required for Postincision Events during Nucleotide Excision Repair: IN VITRO AND IN VIVO STUDIES WITHrad7 AND rad16 MUTANTS AND PURIFICATION OF A Rad7/Rad16-CONTAINING PROTEIN COMPLEX. *Journal of Biological Chemistry*, 273(45):29481–29488.

Sancar, A. (2003). Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors. *Chemical reviews*, 103(6):2203–2238.

Sancar, G. B. and Smith, F. W. (1989). Interactions between yeast photolyase and nucleotide excision repair proteins in Saccharomyces cerevisiae and Escherichia coli. *Molecular and cellular biology*, 9(11):4767–4776.

Sarkar, S., Kiely, R., and McHugh, P. J. (2010). The Ino80 chromatin-remodeling complex restores chromatin structure during UV DNA damage repair. *Journal of Cell Biology*, 191(6):1061–1068.

Scheller, E. and Krebsbach, P. (2009). Gene therapy: design and prospects for craniofacial regeneration. *Journal of dental research*, 88(7):585–596.

Schier, A. C. and Taatjes, D. J. (2020). Structure and mechanism of the RNA polymerase II transcription machinery. *Genes & development*, 34(7-8):465–488.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

Schrödinger, E. (1943). What is Life? The Physical Aspect of the Living Cell. Based on lectures delivered under the auspices of the Institute at Trinity College, Dublin. *February 1943*.

Schulten, K. and Kosztin, I. (2000). Lectures in theoretical biophysics. *University of Illinois*, 117.

Segré, D., Pilpel, Y., and Lancet, D. (1998). Mutual catalysis in sets of prebiotic organic molecules: Evolution through computer simulated chemical kinetics. *Physica A: Statistical Mechanics and its Applications*, 249(1-4):558–564.

Semenenko, V. and Stewart, R. (2005). Monte Carlo simulation of base and nucleotide excision repair of clustered DNA damage sites. II. Comparisons of model predictions to measured data. *Radiation research*, 164(2):194–201.

Shafee, T. and Lowe, R. (2017). Eukaryotic and prokaryotic gene structure. *WikiJournal of Medicine*, 4(1):1–5.

Sharma, R., Lewis, S., and Wlodarski, M. W. (2020). DNA repair syndromes and cancer: insights into genetics and phenotype patterns. *Frontiers in Pediatrics*, 8:570084.

Sibbald, B. (2001). Death but one unintended consequence of gene-therapy trial. *Canadian Medical Association. Journal*, 164(11):1612.

Sigurdsson, S., Dirac-Svejstrup, A. B., and Svejstrup, J. Q. (2010). Evidence that transcript cleavage is essential for RNA polymerase II transcription and cell viability. *Molecular cell*, 38(2):202–210.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144.

Sobol, I. M. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118.

Sokhansanj, B. A., Rodrigue, G. R., Fitch, J. P., and III, D. M. W. (2002). A quantitative model of human DNA base excision repair. I. Mechanistic insights. *Nucleic acids research*, 30(8):1817–1825.

Soutourina, J. (2018). Transcription regulation by the Mediator complex. *Nature reviews Molecular cell biology*, 19(4):262–274.

Steurer, B. and Marteijn, J. A. (2017). Traveling rocky roads: the consequences of transcription-blocking DNA lesions on RNA polymerase II. *Journal of molecular biology*, 429(21):3146–3155.

Strecker, J., Gupta, G. D., Zhang, W., Bashkurov, M., Landry, M.-C., Pelletier, L., and Durocher, D. (2016). DNA damage signalling targets the kinetochore to promote chromatin mobility. *Nature cell biology*, 18(3):281–290.

Struhl, K. (1987). Promoters, activator proteins, and the mechanism of transcriptional initiation in yeast. *Cell*, 49(3):295–297.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature structural & molecular biology*, 14(2):103–105.

Sugasawa, K., Akagi, J.-i., Nishi, R., Iwai, S., and Hanaoka, F. (2009). Two-step recognition of DNA damage for mammalian nucleotide excision repair: Directional binding of the XPC complex and DNA strand scanning. *Molecular cell*, 36(4):642–653.

Sun, Z., Zhang, Y., Jia, J., Fang, Y., Tang, Y., Wu, H., and Fang, D. (2020). H3K36me3, message from chromatin to DNA damage repair. *Cell & bioscience*, 10(1):1–9.

Suter, B., Livingstone-Zatchej, M., and Thoma, F. (1997). Chromatin structure modulates DNA repair by photolyase in vivo. *The EMBO Journal*, 16(8):2150–2160.

Suter, B., Schnappauf, G., and Thoma, F. (2000a). Poly (dA· dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic acids research*, 28(21):4083–4089.

Suter, B., Wellinger, R.-E., and Thoma, F. (2000b). DNA repair in a yeast origin of replication: contributions of photolyase and nucleotide excision repair. *Nucleic acids research*, 28(10):2060–2068.

Sutherland, W. (1905). A dynamical theory of diffusion for non-electrolytes and the molecular mass of albumin. *London, Edinburgh Dublin Philos. Mag*, 9:781–785.

Svejstrup, J. Q., Li, Y., Fellows, J., Gnatt, A., Bjorklund, S., and Kornberg, R. D. (1997). Evidence for a mediator cycle at the initiation of transcription. *Proceedings of the National Academy of Sciences*, 94(12):6075–6078.

Sweder, K. S. and Hanawalt, P. C. (1992). Preferential repair of cyclobutane pyrimidine dimers in the transcribed strand of a gene in yeast chromosomes and plasmids is dependent on transcription. *Proceedings of the National Academy of Sciences*, 89(22):10696–10700.

Swenberg, J. A., Lu, K., Moeller, B. C., Gao, L., Upton, P. B., Nakamura, J., and Starr, T. B. (2011). Endogenous versus exogenous DNA adducts: their role in carcinogenesis, epidemiology, and risk assessment. *Toxicological sciences*, 120(suppl_1):S130–S145.

Thomas, M. C. and Chiang, C.-M. (2006). The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology*, 41(3):105–178.

Tillo, D. and Hughes, T. R. (2009). G+ C content dominates intrinsic nucleosome occupancy. *BMC bioinformatics*, 10(1):1–13.

Tomkinson, A. E., Bardwell, A. J., Bardwell, L., Tappe, N. J., and Friedberg, E. C. (1993). Yeast DNA repair and recombination proteins Rad1 and Rad1O constitute a single-stranded-DNA endonuclease. *Nature*, 362(6423):860–862.

Tornaletti, S., Reines, D., and Hanawalt, P. C. (1999). Structural characterization of RNA polymerase II complexes arrested by a cyclobutane pyrimidine dimer in the transcribed strand of template DNA. *Journal of Biological Chemistry*, 274(34):24124–24130.

van Eeuwen, T., Shim, Y., Kim, H. J., Zhao, T., Basu, S., Garcia, B. A., Kaplan, C. D., Min, J.-H., and Murakami, K. (2021). Cryo-EM structure of TFIIH/Rad4–Rad23–Rad33 in damaged DNA opening in nucleotide excision repair. *Nature communications*, 12(1):3338.

van Eijk, P., Nandi, S. P., Yu, S., Bennett, M., Leadbitter, M., Teng, Y., and Reed, S. H. (2019). Nucleosome remodeling at origins of global genome–nucleotide excision repair occurs at the boundaries of higher-order chromatin structure. *Genome research*, 29(1):74–84.

Verhage, R., Zeeman, A.-M., de Groot, N., Gleig, F., Bang, D. D., Van de Putte, P., and Brouwer, J. (1994). The RAD7 and RAD16 genes, which are essential for pyrimidine dimer removal from the silent mating type loci, are also required for repair of the nontranscribed strand of an active gene in Saccharomyces cerevisiae. *Molecular and cellular biology*, 14(9):6135–6142.

Vermeulen, W. (2011). Dynamics of mammalian NER proteins. *DNA repair*, 10(7):760–771.

Vermeulen, W. and Fousteri, M. (2013). Mammalian transcription-coupled excision repair. *Cold Spring Harbor perspectives in biology*, 5(8):a012625.

Vidaković, A. T., Mitter, R., Kelly, G. P., Neumann, M., Harreman, M., Rodríguez-Martínez, M., Herlihy, A., Weems, J. C., Boeing, S., Encheva, V., et al. (2020). Regulation of the RNAPII pool is integral to the DNA damage response. *Cell*, 180(6):1245–1261.

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannenhalli, S., and Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome research*, 20(11):1574–1581.

von Smoluchowski, M. (1906). Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen.

Wang, Y., Chakravarty, P., Ranes, M., Kelly, G., Brooks, P. J., Neilan, E., Stewart, A., Schiavo, G., and Svejstrup, J. Q. (2014). Dysregulation of gene expression as a cause of Cockayne syndrome neurological disease. *Proceedings of the National Academy of Sciences*, 111(40):14454–14459.

Wang, Z., Wei, S., Reed, S. H., Wu, X., Svejstrup, J. Q., Feaver, W. J., Kornberg, R. D., and Friedberg, E. C. (1997). The RAD7, RAD16, and RAD23 genes of Saccharomyces cerevisiae: requirement for transcription-independent nucleotide excision repair in vitro and interactions between the gene products. *Molecular and cellular biology*, 17(2):635–643.

Wang, Z., Wu, X., and Friedberg, E. C. (1995). The detection and measurement of base and nucleotide excision repair in cell-free extracts of the yeast Saccharomyces cerevisiae. *Methods*, 7(2):177–186.

Watson, J. (2014). Molecular biology of the gene. Watson, Watson.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Whittal, A., Idicula, E., and Hutchings, A. (2022). Exploring the economics of gene therapy innovation and price.

Yamaguchi, N. H. (2019). Smoking, immunity, and DNA damage. *Translational lung cancer research*, 8(Suppl 1):S3.

Yan, C., Dodd, T., Yu, J., Leung, B., Xu, J., Oh, J., Wang, D., and Ivanov, I. (2021). Mechanism of Rad26-assisted rescue of stalled RNA polymerase II in transcription-coupled repair. *Nature Communications*, 12(1):7001.

Yanamoto, T., Miyamoto, A., Ikeda, K., Hatano, T., and Matsuzaki, H. (2011). The relationship between chromosomal positioning within the nucleus and the SSD1 gene in Saccharomyces cerevisiae. *Bioscience, biotechnology, and biochemistry*, 75(9):1713–1721.

Yu, S., Evans, K., Van Eijk, P., Bennett, M., Webster, R. M., Leadbitter, M., Teng, Y., Waters, R., Jackson, S. P., and Reed, S. H. (2016). Global genome nucleotide excision repair is organized into domains that promote efficient DNA repair in chromatin. *Genome research*, 26(10):1376–1387.

Zeitler, L., Denby Wilkes, C., Goldar, A., and Soutourina, J. (2022). A quantitative modelling approach for DNA repair on a population scale. *PLoS Computational Biology*, 18(9):e1010488.

Zenklusen, D., Larson, D. R., and Singer, R. H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology*, 15(12):1263–1271.

Zhang, X., Yin, M., and Hu, J. (2022). Nucleotide excision repair: a versatile and smart toolkit. *Acta Biochimica et Biophysica Sinica*, 54(6):807–819.

Appendix A

Consequences of a Linear DNA String

As introduced in Section 1.2.3, the DNA is folded into a complex three-dimensional structure in order to fit into the spatially constraining nucleus. Our models—which implement repair along a one-dimensional polymer—are therefore highly simplified representations. It is difficult to quantify an impact of the three-dimensional structure on lesion removal using NGS data. We established that nucleosomes as the first-order compaction of the DNA are not expected to significantly influence CPD repair at protein-coding genes. However, it could be similarly surmised that the higher-order conformation in space affects protein distributions themselves, rendering protein recruitment to damage sites easier in some regions than others. Although we assume such an impact to be negligible along single genes, it is necessary to critically evaluate this conjecture.

Chromosomal Interaction Domains (CID) are regions within which DNA locations highly interact. They contain on average 2-3 genes in *Saccharomyces cerevisiae* (Hsieh et al. (2015)). Chromatin conformation within CIDs is commonly assumed to be fairly homogeneous along the population to allow the nucleosome contact. This further supports the hypothesis that we can suppose similar accessibility within a single gene on average. The higher-order structure might only play a negligible role on this scale.

Protein distributions themselves might differ greatly in space. Indeed, reaction-diffusion dynamics are highly important in nature to maintain flexible patterns with well-defined boundaries such as in skin and feather pigmentation (Prum and Williamson (2002)) as well as morphogenesis (Harrison (1993)). They could possibly govern protein distribution within the cell, too. However, studies in human cells suggest that most compounds involved in CPD repair are uniformly distributed in

the nucleus, both in absence and presence of DNA damage (Vermeulen (2011); Hoogstraten et al. (2008)). XPC and TFIIH seem to be the only exception. It is presumed that—despite changing local concentrations—TFIIH is sufficiently evenly scattered (Hoogstraten et al. (2002)). As XPC interacts directly with the DNA, it has been proposed that the uneven distribution might stem from heterogeneous chromatin packaging, in particular heterochromatin. In such a context, the role of reaction-diffusion dynamics might be essential. Generally speaking, however, it should be stressed that heterochromatin tends to be rare in *Saccharomyces cerevisiae*. Nonetheless, the distribution of the DNA quantity itself is—due to its polymer structure—non-uniform within the nucleus.

In order to quantify a possible effect of DNA-protein interactions on distribution and repair, we implemented a simple stochastic reaction-diffusion model in a two-dimensional plane. A sampled fraction of simulated particles updated their position to a random location within a certain radius. We want to remind that erratic protein motions come from particle collisions. Consequently, they are more likely to travel farther distances in less densely concentrated areas. By setting the update radius larger in regions with a lower particle density, we demonstrated that diffusion can be realistically represented (Figure A.1(top)). Chromatin conformation was modelled through a self-avoiding random walk, simulating different concentrations of DNA in space. The random walk was performed over 3600 iterations (i.e. 3600 positions), which corresponds to the size of an average CID in budding yeast. We measured visiting times of proteins along the DNA to evaluate a spatial effect on repair patterns based on either uniformly distributed particles; or high concentrations in regions with larger chromatin presence. Proteins associate randomly with rate k_{on}, and dissociate with a rate k_{off}. To follow the dynamics described by Hoogstraten et al. (2008) for XPC, we implemented a UV event after which dissociation rate was decreased to k'_{off}. Indeed, we found that when starting with a particle distribution that colocalises with DNA, some areas might exhibit quicker visiting times than when starting from a uniform distribution (Figure A.1). However, the effect was not strong, and they were clustered into larger regions within which the probability of association was fairly even. The biased distribution became quickly more uniform, and after a few iterations, there were no preferred areas distinguishable. We assume that when several cell states are superimposed (i.e. when producing NGS data), such an effect would become smoothed out. It should be mentioned that we do not consider protein-protein interactions which could lead to condensate formation that organise nuclear function. However, inclusion of those aspects would require a substantially different data set to compare with, which was not available to us. We conclude that with the best of our knowledge, any local fluctuations in protein concentrations do not significantly affect repair dynamics within genes, at least not on a population scale.

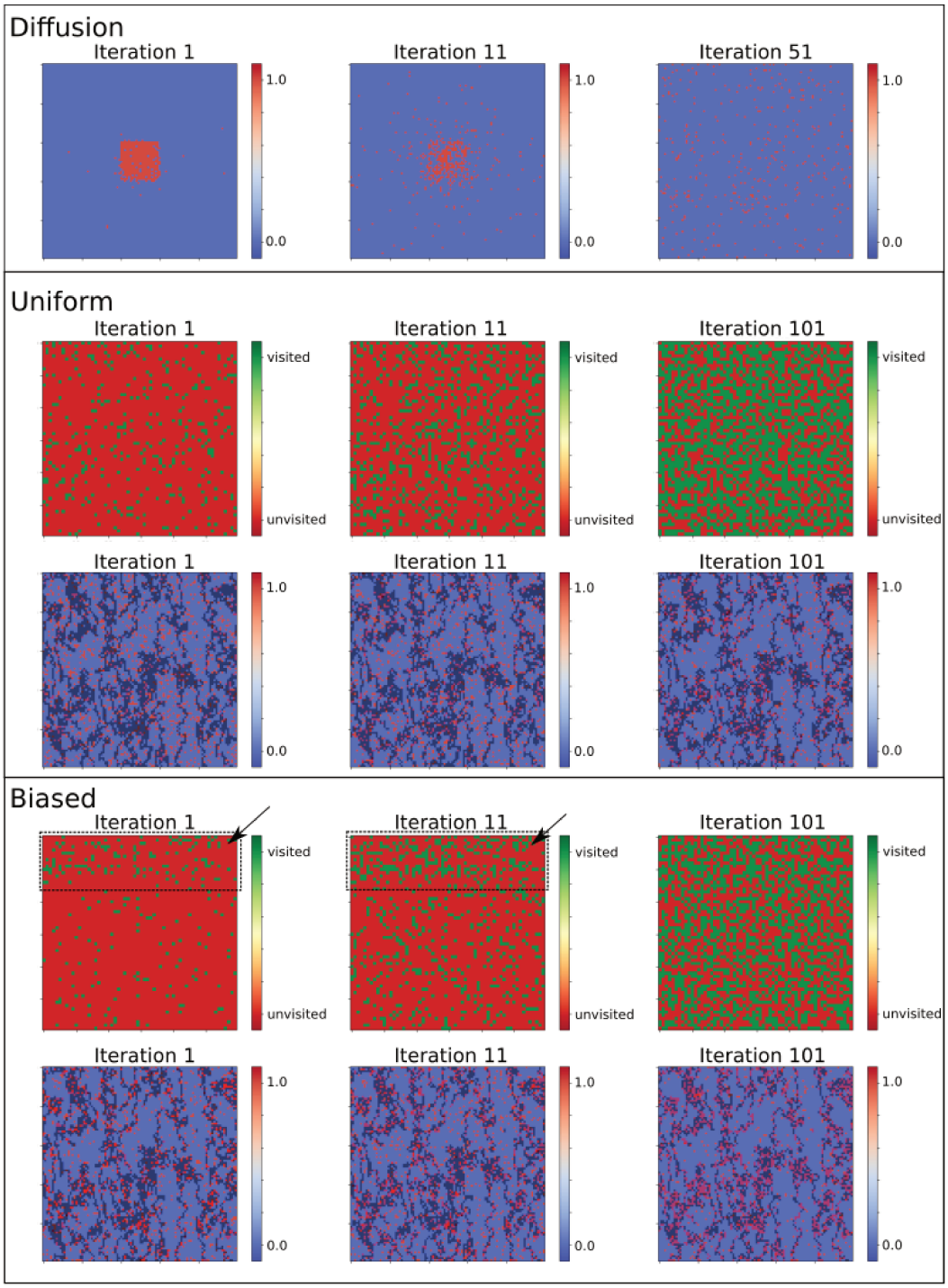


Figure A.1: **Reaction-diffusion dynamics of proteins associating to the DNA do not indicate a strong bias along a single gene.** Top: Diffusion dynamics are correctly captured by random particle movements (red dots). Centre: When starting with a uniform protein distribution in space, it is everywhere equally likely along the DNA to have been visited (green) by the simulated protein (top row, centre). Due to the reaction-diffusion dynamics, particles form colocalised clusters with the chromatin (dark shaded areas, bottom row, centre). Bottom: When proteins are initially more densely distributed at compacted DNA, we observe small patches that are earlier visited than other areas (dotted line with arrow, top row, bottom). However, within this region (approximately same size as a single gene), visiting times are homogeneous (green points). Colocalisation with chromatin remains preserved (bottom row, bottom).

Appendix B

Data Production and Treatment

B.1 Cell Culture

Rad3-HA-tagged cells were grown overnight in 200mL of YPD at 30℃ and harvested at exponential growth phase. Cells were transferred in PBS and irradiated with 100 J/m^2 UVC. Cells were cross-linked before and right after irradiation (t and t_0), 8 minutes after the first cross-link (t_8) and after 38 minutes (t_{38}). Cross-link was performed with 1% formaldehyde for 10 minutes, which was followed by a quenching step with 2.5M glycine for at least 5 minutes.

The data that was used for the gateway problem (TFIIH-mediated transcription) was produced independently without UV-treatment. In the following, both data types followed the same protocol before irradiation but when pointed out otherwise. The gateway data were initially produced for the measurements of CPDs; however, time points after UV were not used due to the discrepancy in the procedure, time points after UV exposure where not used for this study.

B.2 Cell Lysis and Chromatin Preparation

. Cell lysis was performed by bead-beating in FA/SDS buffer (50 mM Hepes-KOH pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton, 0.1% sodium deoxycholate, 0.1% SDS) supplemented with PMSF for 30 min at 4℃. Chromatin was recovered after a centrifugation step (13,400 rcf, 20 min, 4℃) and subjected to sonication on a S220 focused-ultrasonicator (Covaris) (180s ON – 30s OFF – 180s ON, 150W pulses, duty factor 10). Sonicated chromatin was recovered after centrifugation (9,300 rcf, 30 min, 4℃) and stored at -80℃.

B.3 Chromatin Immunoprecipitation

Immunoprecipitation was done using the IP-star SG8X (Diagenode) automated system. 10μL of magnetic beads coupled with protein A (Pierce) were washed twice (PBS, 0.05% Tween20, 0.1% BSA and 0.5M NaCl) and incubated for 1h at 4℃ with 3μL of anti-Pol2 (8WG16 antibody) in a total volume of 100μL (PBS, 0.05% Tween20, 0.1% BSA). Beads were then mixed with 100μL of chromatin during 2h at 21℃. The immunoprecipitation mix was supplemented with 10% (11μL) of chromatin from irradiated *Schizosaccharomyces pombe* that served as spike-in. The samples were then washed twice in FA/SDS/NaCl (50mM HEPES-KOH, pH 7.5, 500mM NaCl, 1mM EDTA, 0.1% sodium deoxycholate, 1% Triton X100, 0.1% SDS), once in IP buffer (Tris 10mM pH8, LiCl 0.25M, EDTA 1mM, NP40 0.5%, Na-Deoxycholate 0.5%), and once in Tris-EDTA (Tris 10mM pH 8, EDTA 1mM). All washes cycles lasted 5 minutes with fast mixing. Samples were ultimately eluted during 25min at 25℃ in the Diagenode iPure elution buffer. The data used for the gateway problem did not apply spike-in, and instead normalised the data using quantitative PCR (qPCR) over all produced time points.

B.4 DNA Purification

Immunoprecipitated chromatin and inputs (non-immunoprecipitated chromatin) were treated with pronase protease (1μg/µl final concentration) and incubated overnight at 65℃ to reverse the cross-linking with an addition of 4μL of 5M NaCl. Samples were subsequently incubated with RNAse (1μL, 1h at 37℃). Finally, DNA was purified with iPure-v2 kit from Diagenode and eluted in 25μL.

B.5 CPD Immunoprecipitation

500ng of DNA from the inputs were denaturated at 95℃ for 10 minutes and quickly cooled down on ice for at least 5 minutes and used for CPD immunoprecipitation. Using the IP-star SG8X (Diagenode) automated system, 10μL of magnetic beads coupled with protein A (Pierce) were washed twice (PBS, 0.05% Tween20, 0.1% BSA and 0.5M NaCl) and incubated for 2h at 4℃ with 2μL of anti-CPD (TDM2 antibody) in a total volume of 100μL (PBS, 0.05% Tween20, 0.1% BSA). Beads were then mixed with 500ng of single-stranded DNA during 4h at 4℃. The immunoprecipitation mix was supplemented with 10% (50ng) of DNA from irradiated *Schizosaccharomyces pombe* that served as spike-in. The samples were subsequently washed at 4℃ twice in FA/SDS/NaCl (50mM HEPES-KOH, pH 7.5, 500mM NaCl, 1mM EDTA, 0.1% sodium deoxycholate, 1% Triton X100, 0.1% SDS), once in IP

buffer (Tris 10mM pH8, LiCl 0.25M, EDTA 1mM, NP40 0.5%, Na-Deoxycholate 0.5%), and once in Tris-EDTA (Tris 10mM pH 8, EDTA 1mM). All wash cycles lasted 5 min with fast mixing. Samples were ultimately eluted during 25 minutes at 25℃ in the Diagenode iPure elution buffer. After elution, samples were purified with iPure (see DNA purification).

B.6 DNA Repair

Purified DNA samples were repaired with a CPD-photolyase purified from *Anacystis nidulans* (gift from Pavel Muller). Incubation was done during 90 minutes under yellow light (50mM Tris, 50mM NaCl, 10mM DTT). DNA was then purified using AMPure XP beads on the IP-star (30μL of samples and 54μL of Beads) and eluted in 20μL.

B.7 Quantitative PCR and Library Preparation

Samples were analyzed on a set of chosen regions by qPCR using the ABI7500 device. To get a genome-wide signal, we used ThruPLEX DNA-Seq Kit (Takarabio) for Pol II-ChIP, and Accel-NGS 1S Plus DNA Library Kit (Swift biosciences) for CPD immunoprecipitation. The former requires double-stranded DNA, whereas the latter uses single-stranded DNA and keeps the strand information up to the sequencing. Libraries were sequenced on a NexSeq550 at the IMAGIF plateform.

We repeated steps B.1-B.6 using Rad4-tagged cells and measured qPCR levels as for Rad3-tagged strains. When comparing the repair process in both cell lines, Rad4-tagged cells exhibited a clearly different CPD removal dynamics than the Rad3-tagged strain, which behaves like YPH (Figure B.1). Although we did not perform replicates, the observed difference was gauged to be very strong. We therefore did not continue with Rad4-tagged cells, and instead opted for representing GGR as a hidden variable in our models.

B.8 Data Treatment

Sequencing reads were first trimmed with `trim_galore`. In the case of single-stranded libraries, the first 10 bases from the second read were specifically removed as indicated by the manufacturer. *Saccharomyces cerevisiae* (version `SacCer3`) and *Schizosaccharomyces pombe* genome sequence (version `ASM294v2`) were concatenated to a hybrid reference genome for read mapping with `bowtie2`. Mapping files (`bam`) were filtered with `samtools` to keep only reads that mapped unambiguously to one or the other genome. Reads that mapped to *Schizosaccharomyces pombe* genome were counted

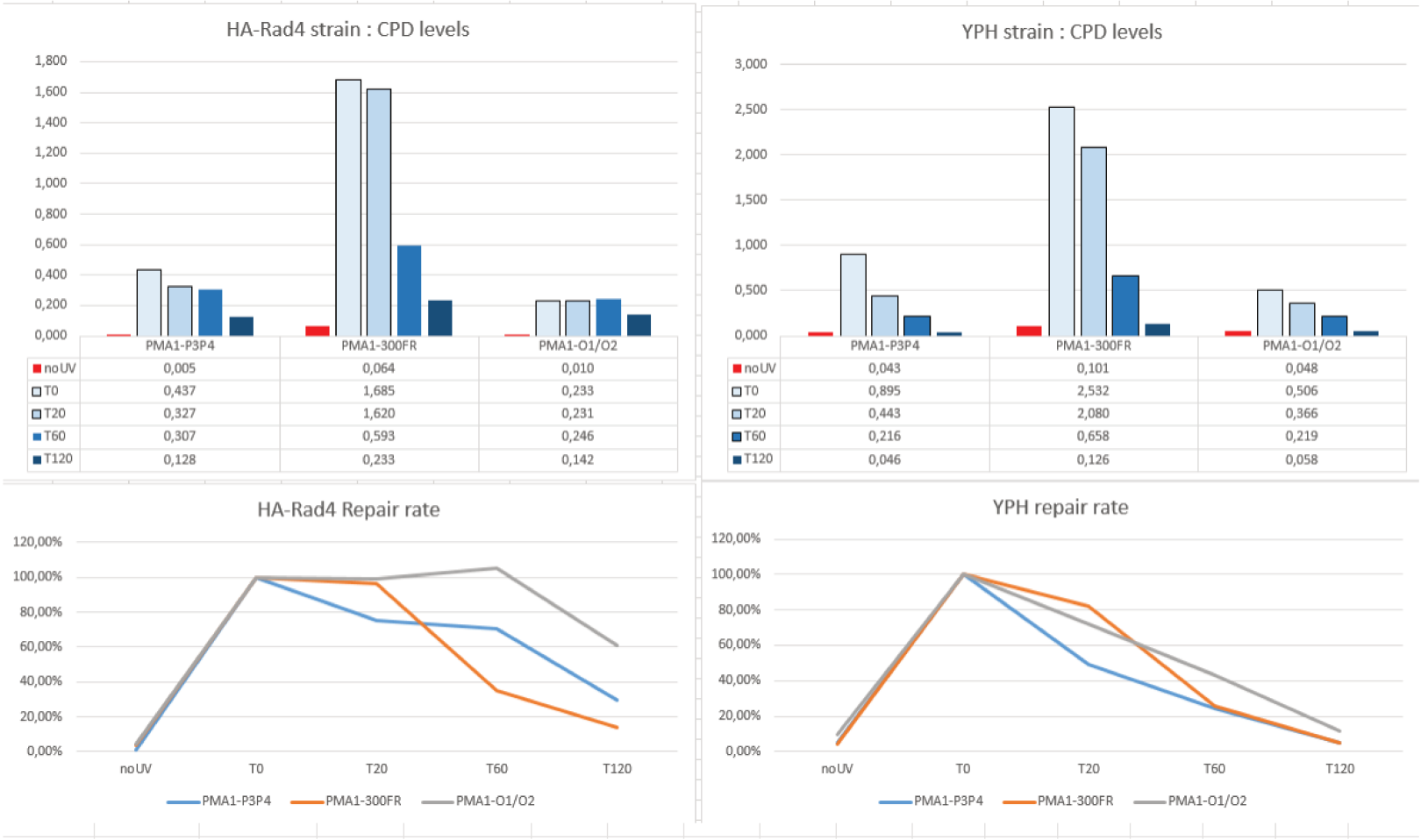


Figure B.1: **Rad4-tagged cells exhibit a different repair behaviour.** Rad4-tagged cells (left) show clearly a different repair behaviour than Rad3-tagged cells, which were verified to repair comparably to YPH strains (right).

and used as normalisation factor, such that the number of *Schizosaccharomyces pombe* reads were equal to 10% (spike-in factor). *Saccharomyces cerevisiae* reads were converted to `bigwig` files using `deeptools` and normalised with the spike-in factor. Data for the gateway problem were mapped only to the `SacCer 3` genome sequence and normalised using qPCR values.

B.9 Data Scaling

Data scaling proved to be pivotal for fitting the mathematical models to the NGS data. However, the sequencing signal amplitude of different properties, such as Pol II and CPDs, could refer to different quantities. This is due to the fact that library preparation for sequencing uses equal amounts of DNA. After the precipitation step through applying an antibody, it is impossible to determine to how many cells this quantity actually corresponds. Spike-in normalisation is useful for analysing changes of a single probed feature over time. However, it is insufficient to compare the amplitude of different properties, such as Pol II ChIP-seq and CPD-seq data. As indicated by the traffic repair model in Chapter 4, the damage distribution could represent significantly fewer cells than the signal for Pol II presence. Nevertheless, the model fitting—for both the traffic repair model and the *Gillespy algorithm* (Chapter 5)—relies on the inter-property comparability of the provided data. Furthermore, despite the fundamental different approaches of the *GillesPy* and the traffic repair model, we aimed to reasonably provide alikeness and similar conditions to avoid favouritism of one approach. In the following, we explain our methods to yield adequate data scaling.

Following Bucceri et al. (2006), we presume that 100 J/m^2 of UVC induces approximately 0.2 CPDs per kb. Consequently, there is a 20% chance that a cell possesses a lesion site in a general gene of size 1000 bp. Assuming a random sequence composition where each nucleotide is equally likely, the probability to find TT, TC, CT, or CC over all possible pairs is 25%. There are 500 pairs in the sequence, so therefore, the probability of having a lesion at a given position is $0.2 \times 1/500 \times 0.25 = 0.0001$ or 0.01%.

It has been reported that there are 20,000-30,0000 Pol II molecules in a single yeast cell (Borggrefe et al. (2001)), with 60% of it being hyperphosphorylated on the C-terminal domain and bound to chromatin, which indicates engagement in elongation (Svejstrup et al. (1997)). This results in 12,000 molecules being associated to the genome and able to encounter damage. By applying a similar nondiscriminatory approach as for CPDs and using the 12 Mb-*Saccharomyces cerevisiae* genome as a reference, we conclude that there is 1 Pol II per kb. However, this ignores a location-specific distribution of Pol II, as the complex can be found predominantly at coding regions and is dependent on the transcription rate. Indeed, other studies that take location-specific gene expression into ac-

count derive a much lower value of ≈0.078 molecules per kb (Pelechano et al. (2010)). However, the measurements rely on Pol II that conserves the nascent mRNA, and it is estimated that only 10% of the transcribing complexes are engaged in the production of mRNA (Struhl (2007)). When taking this into account, the average determined by Pelechano et al. (2010) is surprisingly similar to our approximation. Indeed, it has been shown that histone methylation—which is a marker of elongating Pol II—is also present at genes that are considered to be silent, indicating the presence of moving Pol II that is not producing mRNA (Ng et al. (2003)). The remaining 90% Pol II protein complexes are presumed to be involved in *junk* transcription, which does not result in the production of mRNA. In this work, we surmise that any elongating Pol II can be involved in the detection of CPDs, including those being engaged in junk transcription. This conjecture is in line with the observation that TCR can be observed at almost all genes ($n = 5205$), including those that are presumed to be inert (Duan et al. (2020)).

Since we simulate cell state-dependent repair in the *GillesPy* algorithm, we normalised NGS data first such that it matches the requirements during the stochastic approach; and only thereafter scaled the data accordingly for the traffic repair model to provide comparability. We hypothesise that the presence of CPDs is a rare event, and might only be present at a small fraction of cells. The expected number of damage sites was set to 0.5 CPDs per 500 bins, which is slightly higher than the 0.2 CPDs/kb and was presumed to reduce the number of necessary simulations. The number of Pol II protein complexes was set to 1.5 per 500 bins, similarly to a slightly larger value. We opted to apply the same rate indiscriminately to all coding regions. Although this prevents us from comparing location-specific differences, it nevertheless allows a correct representation within the same gene, whilst ensuring that the algorithm behaves similarly for all regions. For both Pol II and CPDs, the actual number per cell was Poisson distributed to allow variability (Figure B.2(A)). We subsequently performed maximum-occurrence pooling with a 100-bin window and smoothing using a 50-bin Hann window. The scaling was empirically determined such that the difference between data and sampling was normally distributed with zero mean (Figure B.2(B)). The same scaling was used for the data in the traffic model. Despite the fact that NGS signals were here only divided into five bins, we nevertheless increased comparability by using similar signal amplitudes during parameter estimation.

The gateway problem was designed to show that the *GillesPy* framework is able to learn inter-dependencies between associated proteins, such as Rad3 and Pol II. Exact scaling between protein signals was deemed to be less important. We normalised both data types such that the maximum value was equal to one after data binning along a coding region. Subsequently, data was smoothed, which further reduced the signal amplitude to ≈0.8 for the maximum value.

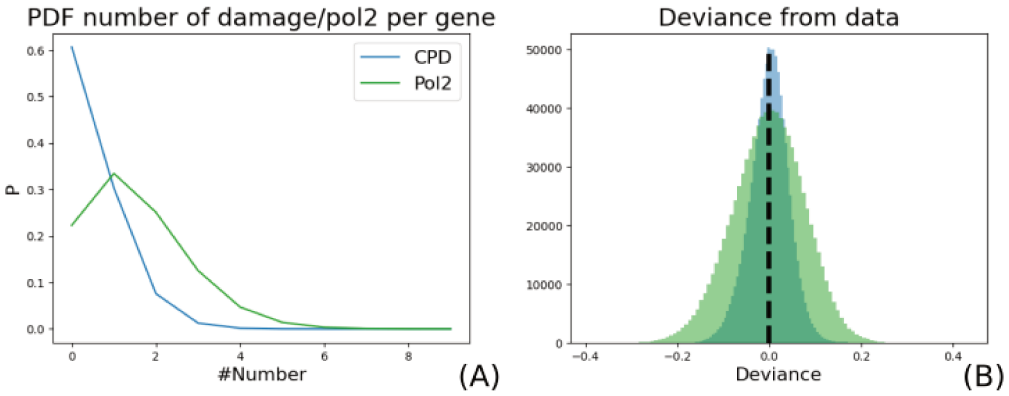


Figure B.2: **Data normalisation used for the computational models.** (A) The number of CPDs along the gene are randomly sampled following a Poisson distribution such that the expected number are either one damage site or 3 Pol II proteins over 1000 bins. (B) When CPD and Pol II presence are sampled, pooled, and smoothed, the difference between simulated distribution and real data is unbiased.

Appendix C

GillesPy Reaction Rules

C.1 Gateway Problem

All implemented proteins can associate and dissociate anywhere along the DNA at a low rate. TFIIH (measured using ChIP-seq for its subunit Rad3) can associate and dissociate from the core promoter. Association can only occur if no other TFIIH complex is already present at the core promoter. The binding event converts the transcription factor to an active state, which allows it to engage in transcription initiation. When TFIIH is associated and active at the core promoter, Pol II can associate to the TSS in an active state (meaning it can engage in transcription), which causes TFIIH to deactivate. This blocks any further association of Pol II, until TFIIH is present in an active state again. This can occur either through dissociation followed by association, or equivalently, by the conversion of a bound TFIIH from inactive to active. Any active Pol II can move along the entire simulated region, but can also randomly stall at any point. Once it reaches the TTS or areas outside the coding region, it dissociates from the DNA (Figure C.1).

C.2 Repair Dynamics

All implemented proteins can associate and dissociate anywhere along the DNA at a low rate. Pol II can associate to the TSS, at which point it becomes active and can move along the gene. Movement, stalling, and dissociation were implemented as for the gateway problem (Appendix C.1). To increase comparability with the traffic model, we implemented competition between TCR and GGR, such that Rad4 (as representative of GGR) can only associate when Pol II is not present. Rad4, on the other hand, can stall Pol II during elongation. Dissociation of Rad4 can happen anywhere independently. Once Pol II encounters a lesion during elongation, it is converted to a blocked state, and therefore, it

Present Reactants			Absent Reactants			Products			Force	Learning rate	Lower	Upper
Type	*Species*	*State*	*Type*	*Species*	*State*	*Type*	*Species*	*State*				
DNA	Unspecific	Unspecific				DNA	Unspecific	Unspecific				
Free Species	Unspecific	Unspecific				Bound Species	Unspecific	Associated	No	0	1.00E-10	1.00E-10
Bound Species	Unspecific	Free										
DNA	Unspecific	Unspecific				Free Species	Unspecific	Default	No	0	9.00E-03	9.00E-03
Bound Species	Unspecific	Unspecific										
DNA	Core Promoter	Default	DNA	Core Promoter	Default	DNA	Core Promoter	Default				
Free Species	Rad3	Unspecific	Bound Species	Rad3	Associated; Active	Bound Species	Rad3	Active	No	1.00E-05	5.00E-05	1
Bound Species	Rad3	Free										
DNA	Core Promoter	Default				DNA	Core Promoter	Default	No	1.00E-03	1.00E-05	1
Bound Species	Rad3	Associated				Bound Species	Rad3	Active				
DNA	Core Promoter	Default				DNA	Core Promoter	Default	No	0	5.00E-01	5.00E-01
Bound Species	Rad3	Associated; Active				Free Species	Rad3	Default				
DNA	TSS	Default				DNA	TSS	Default				
Free Species	Pol II	Default				Bound Species	Pol II	Active				
Bound Species	Pol II	Free							No	1.00E-03	1.00E-08	1
DNA	Core Promoter	Default				DNA	Core Promoter	Default				
Bound Species	Rad3	Active				Bound Species	Rad3	Associated				
DNA	Core Promoter; TSS; Coding Region; TES	Default				DNA	Core Promoter; TSS; Coding Region; TES	Default	Yes	1.00E+01	1.00E-01	2.00E+01
Bound Species	Pol II	Active				Bound Species	Pol II	Active				
DNA	Core Promoter; TSS; Coding Region; TES	Default				DNA	Core Promoter; TSS; Coding Region; TES	Default	No	1.00E-02	1.00E-10	1.00E-06
Bound Species	Pol II	Active				Bound Species	Pol II	Active				
DNA	TES; Default	Default				DNA	TES; Default	Default	No	1.00E-02	1.00E-10	1.00E+01
Bound Species	Pol II	Active				Free Species	Pol II	Default				

Figure C.1: ***GillesPy* rules for the gateway problem.** The table shows reactants that must be present or absent during a reaction as well as their products. Each reactant/product column is subdivided into the type of interacting species (i.e. DNA, free species, or bound species), the name of the species (e.g. core promoter for DNA, Pol II for proteins, or unspecific when the rule applies to all implemented species of a type), and the state (e.g. unspecific when the rule applies to all states or active when a protein must be in an active state). The dashed line indicates the conformation at two different sites along the DNA that must be simultaneously fulfilled. If several species or states are given separated by a semicolon, it indicates that any single one of these requirements is sufficient to participate in the reaction. If unspecific is given as a product (i.e. for species or state), then it is set to the same value as the participating reactant (e.g. if Pol II can participate in any state in a reaction, and an active one is sampled, then the product will be an active Pol II). The force column indicates whether a rule included protein movement. The last three columns give the learning rate as well as the lower and upper bound during parameter training. They were selected through evaluating the performance of different values on three to four randomly sampled genes.

cannot engage in transcription anymore. When Pol II is in a blocked state at a lesion site, the damage can be repaired. Similarly, if Rad4 is associated to a CPD, the lesion can be removed (Figure C.2).

Whilst the implemented pathway was related to *in vitro* dynamics and *in vivo* results, we could not find a reference for competition between Pol II and Rad4 on a single-cell level. Pol II stalling at Rad4 sites as well as inhibition of Rad4 loading when Pol II is present might not correspond to what is actually mechanistically happening. Nonetheless, it increased comparability with the traffic repair model presented in Chapter 4. In order to present modelling results side-by-side, we find the choice justified.

Present Reactants			Absent Reactants			Products			Force	Learning rate	Lower	Upper
Type	*Species*	*State*	*Type*	*Species*	*State*	*Type*	*Species*	*State*				
DNA	Unspecific	Unspecific				DNA	Unspecific	Unspecific				
Free Species	Unspecific	Unspecific				Bound Species	Unspecific	Associated	No	0	1.00E-10	1.00E-10
Bound Species	Unspecific	Free										
DNA	Unspecific	Unspecific				Free Species	Unspecific	Default	No	0	9.00E-03	9.00E-03
Bound Species	Unspecific	Unspecific										
DNA	TSS	Default				DNA	TSS	Default				
Free Species	Pol II	Default				Bound Species	Pol II	Active	No	1.00E-05	1.00E-05	1
Bound Species	Pol II	Free										
DNA	Core Promoter; TSS; Coding Region; TES	Default				DNA	Core Promoter; TSS; Coding Region; TES	Default	Yes	1.00E+03	1.00E-06	2.00E+01
Bound Species	Pol II	Active				Bound Species	Pol II	Active				
DNA	Core Promoter; TSS; Coding Region; TES	Default				DNA	Core Promoter; TSS; Coding Region; TES	Default	No	1.00E-02	1.00E-10	1.00E-02
Bound Species	Pol II	Active				Bound Species	Pol II	Active				
DNA	TES; Default	Default				DNA	TES; Default	Default	No	1.00E-02	1.00E-10	1.00E+01
Bound Species	Pol II	Active				Free Species	Pol II	Default				
DNA	Unspecific	Unspecific	DNA	Unspecific	Unspecific	DNA	Unspecific	Unspecific				
Free Species	Rad4	Default	Bound Species	Pol II	Unspecific	Bound Species	Rad4	Associated	No	1.00E-05	1.00E-10	1
Bound Species	Rad4	Free										
DNA	Unspecific	Unspecific				DNA	Unspecific	Unspecific	No	1.00E-02	1.00E-05	1
Bound Species	Rad4	Associated				Free Species	Rad4	Default				
DNA	Unspecific	Unspecific				DNA	Unspecific	Unspecific				
Bound Species	Pol II	Active				Bound Species	Pol II	Active	No	1.00E-02	1.00E-10	1
Bound Species	Rad4	Active				Bound Species	Rad4	Active				
DNA	Unspecific	Damaged				DNA	Unspecific	Damaged	No	0	1.00E+01	1.00E+01
Bound Species	Pol II	Active				Bound Species	Pol II	Blocked				
DNA	Unspecific	Damaged				DNA	Unspecific	Default	No	10	1.00E-06	1.00E+01
Bound Species	Pol II	Block				Free Species	Pol II	Default				
DNA	Unspecific	Damaged				DNA	Unspecific	Default	No	10	1.00E-06	1.00E+01
Bound Species	Rad4	Associated				Free Species	Rad4	Default				

Figure C.2: ***GillesPy* rules for repair.** The table shows reactants that must be present or absent during a reaction as well as their products. Each reactant/product column is subdivided into the type of interacting species (i.e. DNA, free species, or bound species), the name of the species (e.g. core promoter for DNA, Pol II for proteins, or unspecific when the rule applies to all implemented species of a type), and the state (e.g. unspecific when the rule applies to all states or active when a protein must be in an active state). If several species or states are given separated by a semicolon, it indicates that any single one of these requirements is sufficient to participate in the reaction. If unspecific is given as a product (i.e. for species or state), then it is set to the same value as the participating reactant (e.g. if Pol II can participate in any state in a reaction, and an active one is sampled, then the product will be an active Pol II). The force column indicates whether a rule included protein movement. The last three columns give the learning rate as well as the lower and upper bound during parameter training. They were selected through evaluating the performance of different values on three to four randomly sampled genes.

Intégration Computationnelle et Modélisation de la Cinétique de Réparation de l'ADN Chez la Levure

Introduction

L'acide désoxyribonucléique (ADN) est la molécule qui code pour l'information héréditaire de tous les organismes vivants. Depuis la découverte de sa structure moléculaire par Watson et Crick (Watson and Crick (1953)), il a façonné la science et la société, ouvrant la voie à de vastes possibilités en médecine. Il est d'une importance capitale pour la science médicale de comprendre les processus nucléaires liés à l'ADN pour permettre des traitements surs. Par exemple, l'application de la thérapie génique pourrait permettre de traiter des maladies considérées aujourd'hui comme incurables. En effet, huit thérapies géniques ont été approuvées par la *Food and Drug Administration* (FDA) en 2021, et plus de 1300 étaient en cours de développement en 2020 (Whittal et al. (2022)).

Réussir le développement d'un médicament implique l'étude des interactions moléculaires dans les cellules vivantes. L'interaction des protéines et de l'ARN avec l'ADN régule et influence tous les processus nucléaires dans différents contextes (Cozzolino et al. (2021)). En raison de cette vaste complexité, d'innombrables questions restent sans réponse. La réparation de l'ADN est un sujet qui a fait l'objet d'études intensives ces dernières années. On sait que la composition physico-chimique de l'ADN est constamment modifiée par divers facteurs externes et internes. Il est donc indispensable pour la survie des cellules de posséder divers mécanismes de réparation et de maintien de l'intégrité de l'ADN. La grande diversité d'altérations possibles a entraîné le développement de plusieurs voies de réparation de l'ADN dans la nature, notamment la réparation par excision de nucléotides (NER). La NER est une voie conservée au cours de l'évolution que l'on retrouve chez presque tous les eucaryotes, y compris les cellules humaines et la levure bourgeonnante (*Saccharomyces cerevisiae*).

Elle se caractérise par sa capacité exceptionnelle à éliminer de nombreux types de lésions, notamment les dommages induits par les UV tels que les dimères cyclobutylique de pyrimidines (*Cyclobutane Pyrimidine Dimers*, CPD) et les 6-4 photoproduits (*6-4 Photoproducts*, 6-4PP). La NER interagit différemment selon le type de région génomique, ce qui explique la différenciation conventionnelle entre la réparation globale du génome (GGR), qui peut être observée sur l'ensemble du génome, et la réparation couplée à la transcription (TCR), qui est limitée aux gènes activement transcrits par le complexe multiprotéique de l'ARN polymérase II (Pol II). Le blocage de la Pol II au niveau des lésions *Transcription-blocking DNA lesions*, TBL) déclenche le recrutement d'autres protéines de la NER (Deaconescu et al. (2006))). Les deux voies de détection différentes convergent ensuite vers le même mécanisme d'incision et de remplacement du brin endommagé.

Bien qu'elles aient posé les bases fondamentales de nos connaissances, les expériences biologiques atteignent rapidement leurs limites lorsqu'il s'agit d'étudier un processus complexe tel que la réparation de l'ADN. La manière dont la dynamique de réparation est coordonnée avec d'autres processus nucléaires—tels que la transcription et le repliement de la chromatine—n'est pas claire, et l'organisation de la NER à l'échelle du génome *in vivo* reste jusqu'à présent mal comprise. Il est donc nécessaire de combiner des données expérimentales génomiques, telles que les mesures d'interaction ADN-protéine couplées au séquençage à haut débit (NGS), avec des modèles informatiques pour comprendre les processus nucléaires dans un environnement entièrement contrôlé. Néanmoins, malgré le besoin évident d'interfacer les méthodes informatiques et expérimentales, le nombre d'approches de modélisation pour la cinétique de réparation reste faible. Dans ce travail de thèse, nous présentons des approches d'analyse de données et des modèles mathématiques pour expliquer la réparation de l'ADN dans la levure *Saccharomyces cerevisiae* en tant qu'organisme modèle pour l'élimination des lésions dans les cellules humaines.

La Théorie

L'irradiation des cellules par les UV entraîne la création de différents types de perturbations de l'ADN. La plupart d'entre elles sont des CPD et des 6-4PP (deux variantes de dimères de pyrimidine). Les premières représentent jusqu'à 75-95% de toutes les lésions (Bohm et al. (2023)). La formation des CPD est due à une réaction photochimique au cours de laquelle les UV sont absorbés par une double liaison entre les bases pyrimidines. En ouvrant la liaison hydrogène, la base libre réagit avec les molécules voisines. Si le nucléotide adjacent est une autre pyrimidine, ils forment de nouvelles liaisons directes (Goodsell (2001)). Elles forment des TBL qui peuvent être réparés par la NER.

Chaque cellule possède un certain nombre de voies de réparation différentes pour réparer les

perturbations moléculaires de l'ADN. La NER peut éliminer différents types de dommages—tels que les CPD induits par les UV—et elle est conservée au cours de l'évolution chez tous les eucaryotes. Le processus de la NER est généralement divisé en deux voies de reconnaissance—GGR et TCR—qui convergent ensuite vers la même voie d'incision et de remplacement.

La voie GGR trouve et reconnaît les lésions de l'ADN par des associations protéiques directes de Rad4-Rad23-Rad33. La voie peut agir sur l'ensemble du génome. La reconnaissance et la réparation qui s'ensuit sont indépendantes du site de la lésion et de la structure de la chromatine, bien qu'elles puissent être facilitées par des interactions avec des remodeleurs de la chromatine tels que SWI/SNF et Ino80 (Sarkar et al. (2010)). La reconnaissance des lésions par Rad4 est pilotée par la détection des paires de bases thermodynamiquement instables (Min and Pavletich (2007)). Rad4-Rad23-Rad33 permet lui-même la vérification des dommages en recrutant le facteur d'initiation de la transcription IIH (TFIIH). Le complexe de reconnaissance Rad4-Rad23-Rad33 est ensuite libéré et l'ADN est scanné dans le sens 5'-3' à la recherche de lésions bloquant l'hélicase (Sugasawa et al. (2009)). Si aucune lésion de l'ADN n'est trouvée, le brin ouvert est fermé et le processus recommence.

Il existe un consensus scientifique sur le fait que les lésions dans les régions transcrites sont réparées plus rapidement que les régions silencieuses en aval, ce qui a créé l'idée que la TCR est plus efficace (Bohr et al. (1985); Mao et al. (2016); Li et al. (2018); Mao et al. (2020)). Il est communément admis que l'élimination des lésions du brin transcrit (TS) est préférée à celle du brin non transcrit (NTS) (Mellon et al. (1987)), ce qui a été démontré sur le gène RPB2 (Sweder and Hanawalt (1992)). Les TBL empêchent la Pol II de poursuivre son élongation et si le blocage persiste, il déclenche le recrutement d'autres facteurs de la NER. Pol II couvre environ 35 nt du brin transcrit, y compris la lésion (Tornaletti et al. (1999)) et entrave donc la poursuite de la réparation. Des hypothèses diverses et non exclusives ont été émises sur le sort de Pol II, notamment sa dissociation, son retour en arrière (*backtracking*) ou sa dégradation. Il est communément admis que le mécanisme le plus courant est le clivage du transcrit et le retour en arrière (Sigurdsson et al. (2010); Marteijn et al. (2014)), car il est également impliqué dans d'autres processus nucléaires.

Après la détection des dommages par la GGR ou la TCR, l'ADN est ouvert par l'interaction AT-Pase / hélicase du complexe multiprotéique TFIIH, en particulier par Rad3 et Rad25. La présence de lésions est vérifiée par TFIIH, Rad14 et RPA. Si la lésion est absente, le clivage de l'ADN n'a pas lieu et les protéines se dissocient. Si elle est présente, la lésion est éliminée par une incision sur les deux côtés de la distorsion, réalisée par Rad1-Rad10 et Rad2. Le fragment excisé est ensuite libéré avec les autres composants de la NER. La double incision est suivie par la synthèse et la ligature de l'ADN.

La réparation de l'ADN dans les cellules vivantes doit être orchestrée dans le contexte d'autres processus nucléaires tels que la conformation de la chromatine et la transcription. Afin de confiner les molécules d'ADN dans le noyau, l'ADN est étroitement enroulé autour de complexes d'histones appelés nucléosomes (Kornberg (1974); Luger et al. (1997)). Les positions des nucléosomes sont très fréquentes et se produisent environ toutes les 200 paires de bases (pb) chez tous les eucaryotes. Bien que le positionnement soit en partie dépendant de la séquence (Tillo and Hughes (2009)), des études récentes ont montré de fortes interactions avec des protéines et des facteurs *trans*, y compris des remodeleurs de nucléosomes avec une sous-unité ATPase. Il existe également des preuves d'interaction avec d'autres protéines liées à l'ADN, notamment la Pol II. Comme les nucléosomes voisins ne peuvent pas se chevaucher, il a été proposé de modéliser l'ensemble comme des billes sur un fil (Jansen and Verstrepen (2011)). La disposition des nucléosomes joue un rôle important pendant la formation des lésions elles-mêmes. Il a été démontré que l'ADN en rotation vers l'extérieur au niveau des nucléosomes fortement positionnés est moins protégé contre l'irradiation UV, ce qui entraîne ce que l'on appelle une *empreinte photographique* qui persiste pendant la réparation en cours.

Les protéines sont vitales pour le fonctionnement des cellules et doivent être produites de manière dynamique pour réagir aux changements de l'environnement. Les séquences codant pour des protéines sont transcrites par la Pol II, un processus qui produit l'ARN messager (ARNm). La transcription de tous les gènes codant pour des protéines est régulée par la liaison spécifique des facteurs de transcription (*Transcription Factor*, TF) à des séquences activatrices en amont (*Upstream Activating Sequence*, UAS) ou à des séquences répressives en amont ((*Upstream Repressing Sequence*, URS) chez la levure. Le programme transcriptionnel doit être coordonné avec la structure de la chromatine et le positionnement des nucléosomes. En effet, les activateurs recrutent des co-activateurs qui modulent la conformation de la chromatine afin de la rendre plus accessible et/ou qui stimulent l'assemblage de la machinerie de transcription. L'un de ces co-activateurs est le complexe Médiateur, qui facilite le chargement et la stabilisation des protéines de la transcription (Soutourina (2018)). La TCR est une voie de réparation liée à l'activité transcriptionnelle de Pol II. Par conséquent, l'expression des gènes elle-même influence naturellement la dynamique de la NER. En effet, il peut être démontré que les gènes très actifs avant le traitement UV présentent une élimination plus rapide des CPD (Mao et al. (2016); Li et al. (2018)). Néanmoins, l'effet de la TCR peut être observé sur tous les gènes, indépendamment de leurs niveaux de transcription (Mao et al. (2020)). Il convient de souligner que le lien entre la transcription et la réparation est loin d'être trivial, et qu'il n'existe pas de relation linéaire claire (Li et al. (2018)).

Il existe depuis des décennies des techniques permettant de sonder les propriétés locales du

génome, sur un ou quelques loci. L'avènement des technologies de séquençage à haut débit, permet aujourd'hui d'utiliser ces mêmes techniques mais en accédant aux propriétés de l'ensemble du génome. Elles sont particulièrement utilisées pour étudier les interactions ADN-protéines, bien qu'elles puissent être appliquées de manière similaire pour mesurer d'autres caractéristiques, notamment les dommages à l'ADN et le positionnement des nucléosomes. En quelques mots, une culture de cellules est traitée pour extraire et sélectionner les acides nucléiques d'intérêt. Ils sont divisés en fragments plus petits, qui sont jointés à des séquences adaptatrices spécifiques à la technologie de séquençage et enfin liés à la surface en verre de la machine. Le séquençage à lecture courte, qui a été exclusivement utilisé dans ce travail, est effectué par séquençage par synthèse (SBS), un processus au cours duquel la composition de la séquence peut être déterminée. Toutes les données acquises sont filtrées et corrigées à l'aide de pipelines de traitement standard et personnalisés.

Dans ce travail, nous supposons que les mouvements des protéines et de l'ADN peuvent être représentés par une dynamique stochastique des particules. Les interactions peuvent être observées lorsqu'elles se co-localisent. Toutes les techniques de modélisation appliquées utilisent des processus qui décrivent le mouvement des molécules. Nous avons utilisé en particulier le cadre mathématique du mouvement brownien, qui décrit le mouvement probabiliste de particules en suspension dans un milieu—dans notre cas, le nucléoplasme. Les paramètres du modèle peuvent être estimés en utilisant une approche d'apprentissage, c'est-à-dire un algorithme spécifique au modèle et aux données qui permet de trouver des valeurs raisonnables qui expliquent les données.

Analyse du positionnement des nucléosomes à l'échelle du génome chez la levure

Le positionnement des nucléosomes est essentiel pour permettre l'accessibilité à la séquence, et il est donc supposé influencer et être influencé par divers autres processus dans le noyau, y compris la présence de Pol II et la transcription (Koerber et al. (2009); Ocampo et al. (2016)) ainsi que la réparation de l'ADN (Mao et al. (2016); van Eijk et al. (2019)). On pourrait supposer que la présence coordonnée de nucléosomes le long d'un gène influence à la fois la TCR et la GGR. Cependant, il manque une compréhension claire de la dynamique. Nous combinons la corrélation classique de Pearson avec l'analyse fonctionnelle en composantes principales (fPCA) pour décrire la dynamique des nucléosomes le long des régions codantes. En comparant les données MNase-seq de souches déficientes en facteurs de remodelage de la chromatine (Ocampo et al. (2016, 2019)), nous pouvons quantifier l'impact sur le phasage et l'espacement de plusieurs nucléosomes les uns par rapport aux

autres.

Nous avons commencé par mesurer la corrélation de Pearson des données MNase-seq produites par Ocampo et al. (2016, 2019) pour toutes les régions codant pour des protéines. L'ensemble des données contient deux réplicas pour les contextes *chd1*Δ, *isw1*Δ et *isw2*Δ ainsi que pour les souches appauvries en *rsc8*, de même que leurs combinaisons en tant que mutants doubles, triples et quadruples. Les coefficients de Pearson ont été regroupés en deux partitions distinctes à l'aide de la classification k-mean (dans la suite de l'article aussi *groupement de Pearson*). Cependant, il est difficile de caractériser mathématiquement les deux groupes en n'utilisant que l'indice de corrélation, en particulier celui qui permet d'identifier de manière cohérente les différences entre les mutants. Plus important encore, le coefficient de Pearson ne mesure que la corrélation linéaire moyenne sur l'ensemble du profil, plutôt que de prendre en compte les particularités liées à la position. Nous supposons que le signal MNase-seq peut être décrit comme une fonction continue et qu'il peut être approximé par un mélange d'un nombre fini de fonctions continues plus simples. Cela a permis l'application de la fPCA pour déterminer les deux composantes principales fonctionnelles (fPC) majeures qui expliquent chaque profil de nucléosome. Intuitivement, une fPC est une combinaison des fonctions plus simples avec un score ou poids ξ indiquant leur écart par rapport au profil moyen pour décrire une distribution particulière. Les deux fonctions s'étendent sur l'ensemble du gène (ou plutôt jusqu'à 7 nucléosomes) et prennent donc en compte les différences liées à la position. Les scores peuvent être utilisés pour regrouper les distributions des nucléosomes en fonction des deux principaux fPC. Il est étonnant de constater que les deux groupes en WT—qui ont été obtenus indépendamment par un regroupement hiérarchique classique des coefficients de Pearson—sont nettement séparés en ce qui concerne les scores le long du deuxième fPC, alors qu'ils sont apparemment indépendants du premier. Cela signifie que la deuxième fPC décrit le phasage coordonné des nucléosomes. Il était très surprenant que la séparation nette entre les deux groupes disparaisse complètement pour les petits gènes < 1000 pb, c'est-à-dire les gènes plus petits que la région considérée. Le fait que les clusters ne soient pas séparables implique que la mise en phase coordonnée des nucléosomes disparaît après le TTS et nous avons émis l'hypothèse que l'arrangement est strictement limité au corps du gène. Etonnamment, la séparation des deux groupes est clairement visible pour les petits gènes dans les souches appauvries en *rsc8*. En effet, le profil MNase-seq moyen présente des pics phasés tout au long de la fenêtre de 1000 pb, et le positionnement des nucléosomes se poursuit en dehors des limites du gène. Les résultats indiquent que Rsc8 est requis pour le découplage du phasage des nucléosomes entre les régions codantes leur régions flanquantes.

Le groupement de Pearson et la fPCA ont été répétés exclusivement pour les gènes de plus de

1000 pb pour tous les mutants. En effet, une limite de séparation linéaire des scores fPC peut être trouvée pour toutes les souches, indiquant que la mise en phase collective des nucléosomes—qui est représentée par le coefficient de corrélation de Pearson—reste préservée. Néanmoins, l'inclinaison de la pente a changé. Nous supposons que la mise en phase collective des nucléosomes a changé par rapport aux conditions WT si la pente est devenue significativement plus forte ou plus faible. Nous avons identifié 5 mutants - à savoir *chd1*Δ, *isw2*Δ*chd1*Δ, *chd1*Δ appauvri en *rsc8*, *isw1*Δ*isw2*Δ et *isw2*Δ*chd1*Δ appauvri en *rsc8*—qui ont provoqué des changements notables compte tenu de la variabilité expérimentale. En fait, les souches contenant la délétion du gène Chd1 étaient particulièrement affectées. Ceci est fortement corrélé avec la présence de Pol II et d'autres grands complexes protéiques tels que Médiateur.

Nous proposons le mécanisme suivant. Le complexe de remodelage de la chromatine RSC est essentiel pour permettre une mise en phase indépendante dans chaque gène. Il joue un rôle central dans le maintien de la barrière par rapport à laquelle le positionnement des nucléosomes est coordonné. Cela permet le découplage des processus spécifiques aux gènes, tels que la transcription. La déplétion de Rsc8 entraîne l'interférence de différentes régions génomiques, ce qui modifie l'accessibilité des séquences à l'échelle globale. Chd1, en revanche, maintient l'intégrité de la chromatine pendant la transcription et influence localement la mise en phase des nucléosomes pour permettre l'expression médiée par Pol II. Les souches *chd1*Δ rendent le positionnement dépendant de la présence de Pol II. Par conséquent, alors que RSC joue un rôle global, Chd1 est important pour l'organisation locale en nucléosomes.

Une approche de modélisation quantitative pour la réparation de l'ADN à l'échelle de la population

Le processus complexe de réparation de l'ADN nécessite la coordination de plusieurs étapes, qui doivent être orchestrées avec d'autres procédures nucléaires. Nous avons développé une approche de modélisation basée sur des données qui évite les hypothèses spécifiques sur le processus de réparation lui-même. Au lieu de cela, nous supposons uniquement que les protéines suivent la dynamique des particules. Cela peut être compris plus en détail comme suit.

Dans une cellule unique, les dommages CPD décrivent l'appariement erroné de deux nucléobases pyrimidiques adjacentes. Par conséquent, il peut y avoir au maximum une lésion par position. Il en résulte un espace d'état zéro-un (c'est-à-dire *endommagé-réparé*) par position et par cellule. Pendant la réparation, les lésions sont éliminées et les positions passent ainsi à l'état *réparé*. On

peut supposer que ce processus est stochastique et implique dans une certaine mesure un bruit imprévisible. Si nous pouvions mesurer de manière répétée et indépendante les temps de réparation pour une seule position dans une seule cellule, nous pourrions répartir les mesures sur une ligne de temps. Ce type de données peut être étudié par un processus de Poisson, qui permet de dériver une fonction prédictive exprimant la probabilité de réparation dans le temps. Nous supposons que cette fonction est donnée par le changement du signal CPD-seq (publié par Mao et al. (2016)), car la diminution de l'amplitude à une position arbitraire doit expliquer le nombre de cellules qui ont réparé leurs lésions. Par conséquent, les données représentent le processus sur l'ensemble de la population cellulaire. Nous supposons que la dynamique est indépendante entre les cellules. Cela nous permet de ne considérer pas les signaux NGS comme une cellule moyenne mais comme l'effet cumulé de plusieurs cellules indépendantes. Nous analysons la transition d'état, encodée dans les données, en combinant un processus de Poisson avec des mouvements aléatoires de protéines dans le noyau et le long de l'ADN. Les paramètres peuvent être facilement estimés par une régression linéaire. Étonnamment, la dynamique de réparation prédite est conforme aux données évaluées indépendamment qui mesurent la réparation active (XR-seq)(Li et al. (2018)).

Bien que nous modélisions la réparation en fonction du temps, nous n'intégrons pas deux mécanismes de réparation (potentiellement concurrents), c'est-à-dire la TCR et la GGR. Nous pouvons récupérer l'effet cumulatif en calculant la moyenne de l'évolution de la réparation pour un groupe de segments de gènes, c'est-à-dire le début, le centre et la fin du gène avec des niveaux élevés de TCR ainsi que des régions sans TCR. Nous montrons que l'influence d'un mécanisme agissant plus tardivement—dont nous supposons qu'il représente la GGR—augmente en fonction de la distance par rapport au TSS, jusqu'à ce que, dans les régions non TCR, nous observions presque exclusivement le processus de réparation le plus tardif. Cela suggère que la cinétique d'élimination des lésions change le long du génome.

Nous avons étendu l'analyse pour faire des prédictions sur les facteurs influençant la réparation des CPD *in vivo*. Pour évaluer le pouvoir prédictif de notre modèle, nous avons choisi d'analyser un lien avec le taux de transcription et la densité des nucléosomes, qui représentent la structure de la chromatine. Nous avons également étudié un lien avec la longueur de l'unité de transcription (TU) et la distance relative aux centromères et aux télomères en tant que paramètres d'influence possibles non signalés. Les corrélations prédites avec la transcription et la densité des nucléosomes sont conformes à la littérature. En particulier, la réparation dans les régions codantes est corrélée à l'activité transcriptionnelle mais pas à la densité des nucléosomes. Il est intéressant de noter que le lien le plus fort a été trouvé avec la longueur de la TU, ce qui, à notre connaissance, n'a jamais été montré. La TS et la NTS sont clairement influencées. Le modèle quantitatif développé a donc le

potentiel d'identifier des interrelations établies ainsi que de nouvelles interrelations. Il est important de noter qu'une corrélation avec la distance aux télomères et aux centromères n'a pas révélé de lien significatif, ce qui indique que la méthode appliquée est sélective.

En conclusion, notre travail indique une dynamique de réparation changeant dans l'espace. La corrélation avec d'autres processus nucléaires ouvre des perspectives intéressantes pour les recherches futures sur les mécanismes de réparation de l'ADN et les facteurs génomiques qui peuvent l'influencer. De nouvelles données expérimentales avec une meilleure résolution temporelle permettront d'affiner le modèle et l'analyse. Le modèle peut être facilement appliqué aux données de séquençage de tout processus nucléaire qui peut être représenté comme un système à deux états, et il n'est pas limité à la réparation.

Une approche *mean-field* pour comprendre la réparation de l'ADN

En utilisant les résultats des projets précédents, nous pouvons faire quelques hypothèses concrètes sur la dynamique de la cellule unique. Nous pouvons raisonnablement supposer que les souches WT—qui possèdent Rsc8 et Chd1—devraient présenter un phasage des nucléosomes indépendant de la présence de Pol II, du moins pour ce qui peut être mesuré à l'aide de NGS. Nous supposons que nous pouvons donc largement ignorer l'effet de l'accessibilité de la chromatine pendant la réparation dans les régions codant pour des protéines qui sont réparées principalement par la TCR. Cette hypothèse est étayée par le modèle KJMA, qui n'indique pas d'impact significatif de la densité des nucléosomes le long des gènes ; pourtant, les paramètres de réparation sont corrélés avec les niveaux de transcription. Plus intéressant encore, les résultats suggèrent que la dynamique de réparation change le long des régions codantes. Nous proposons que la cinétique spatio-temporelle de réparation des CPD résulte des interactions ADN-protéines, en particulier des protéines impliquées dans la détection des dommages par la GGR et la TCR, telles que Rad4 et Pol II respectivement. Nous visons à évaluer cette conjecture à l'aide d'un modèle mécanistique. Nous présentons le modèle de réparation par trafic, qui est dérivé de la simulation des mouvements de véhicules. Le modèle suppose que la dynamique de réparation peut être représentée par un comportement moyen—ce qui est également appelé l'approche *mean-field*—au lieu de prendre en compte les spécificités stochastiques. Nous considérons la réparation des gènes comme un processus en deux étapes au cours duquel la lésion est d'abord trouvée par les mouvements des protéines et ensuite réparée (c'est-à-dire en combinant l'excision et le remplacement en une seule étape). Ainsi, nous dérivons trois équations différentielles ordinaire (EDO) qui décrivent la reconnaissance par la TCR et la GGR ainsi que l'élimination des CPDs. La cinétique modélisée de la TCR est pilotée

par l'élongation de la Pol II, tandis que la GGR se produit par association et dissociation aléatoires sans mouvement le long de l'ADN.

Nous avons pris en compte 1878 régions transcrites qui présentaient une diminution rapide de CPD, que nous supposons provenir principalement de la TCR. Les paramètres des EDO développées ont été ajustés aux données de séquençage à l'aide d'équations différentielles ordinaires neurales (NODE) (Chen et al. (2018); Chen (2018)). Pour être précis, à partir d'une distribution initiale des données Pol II ChIP-seq et CPD-seq, nous avons cherché à prédire leur distribution respective à un point de temps ultérieur : après 30 minutes de réparation. Pour expliquer les aspects manquants de la cinétique NER, nous avons laissé les interactions ADN-protéine liées à GGR comme un paramètre caché pour lequel aucune donnée n'était disponible. Au départ, les distributions de séquençage ont été mises à l'échelle de manière à ce que la plus grande valeur soit égale à 1. L'erreur entre la prédiction et les données a diminué jusqu'à ce qu'elle atteigne un plateau. Toutefois, un test de significativité a révélé que bien que les distributions de Pol II étaient très significatives et ne pouvant être expliquées par l'utilisation de paramètres provenant d'autres gènes, la cinétique de réparation prédite pouvait être tout aussi bien expliquée en changeant les paramètres avec des gènes aléatoires. Cela indique que si le mouvement et la présence de Pol II changent de manière significative entre les gènes, cette dynamique ne se traduit pas par une réparation significative des CPD. Cela va à l'encontre de notre hypothèse selon laquelle la réparation spatio-temporelle est liée aux mouvements des protéines le long de l'ADN. De manière surprenante, lorsque l'on met à l'échelle les données par rapport à la présence de CPD et de Pol II dans une seule cellule rapportée dans la littérature, nous pouvons trouver des prédictions significatives à la fois pour la TCR et pour l'élimination des dommages. Ce faisant, nous démontrons que la présence des dommages est plus sporadique que l'occupation de Pol II, et qu'une mise à l'échelle adéquate doit être déterminée à l'aide d'hypothèses basés sur les cellules uniques. Plus important encore, le test de significativité indique que la dynamique de réparation est spécifique à la position et que les paramètres trouvés pour un gène ne peuvent pas être remplacés par les paramètres d'un autre gène. Nous avons évalué cela plus en profondeur en analysant l'influence des paramètres du modèle sur la prédiction dans toutes les régions évaluées. Cela a révélé non seulement une interdépendance fonctionnelle entre les valeurs des paramètres, mais aussi la présence de deux groupes de réparation distincts, qui diffèrent principalement en ce qui concerne l'association Pol II et le taux de réparation par GGR. Lorsque nous avons étudié la dynamique de réparation spécifique aux gènes, nous avons constaté que l'un des deux groupes nécessitait une présence importante de protéines liées à la GGR pour expliquer l'élimination des CPD, tandis que l'autre groupe présentait une dynamique de réparation qui pouvait être expliquée exclusivement par la TCR. Une analyse de l'ontologie des gènes suggère

qu'un groupe est principalement impliqué dans les processus de biosynthèse (par exemple, la transcription et la traduction, $\approx 50\%$), tandis que l'autre groupe contient de nombreux gènes liés à la localisation et au transport des protéines ($\approx 30\%$). Cela pourrait suggérer la présence d'un mécanisme de régulation de la réparation de l'ADN.

Comprendre le mécanisme de la réparation stochastique de l'ADN dépendante des cellules à l'aide de l'algorithme *GillesPy*

Alors que le modèle de réparation par trafic a révélé que la dynamique de réparation est spécifique à un gène—donc changeante par rapport à un paramètre inconnu—et qu'elle pourrait présenter une plus grande dépendance à l'égard de la GGR qu'on ne le pensait auparavant, l'approximation *mean-field* ne peut pas fournir une compréhension mécanistique, c'est-à-dire relier les observations à l'échelle d'une population, aux interactions ADN-protéine et protéine-protéine nécessaires dans une cellule unique. Afin d'éclairer ces observations, il est nécessaire de développer un algorithme probabiliste pour expliquer la réparation spécifique de l'ADN à l'échelle d'une cellule. L'une des approches de simulation stochastique les plus connues pour les réactions chimiques a été proposée par Gillespie (1977). En supposant que les réactions dans un milieu bien mélangé sont sporadiques, l'algorithme de Gillespie échantillonne au hasard une réaction particulière μ et un délai τ après lequel μ est observé en fonction de l'état actuel. Un état du système est défini comme le nombre et les types de particules présentes dans la solution qui peuvent participer à une réaction. La réaction entraîne un changement d'état, qui peut rendre plus ou moins probables d'autres interactions entre particules. Un système peut traverser différentes séquences de réactions lorsqu'il est simulé à plusieurs reprises. Cependant, le modèle initial n'incluait pas la notion d'espace. Pour décrire la dynamique de réparation spatio-temporelle le long des gènes codant pour les protéines, nous avons étendu le travail de Gillespie pour modéliser les interactions entre les particules le long d'un polymère unidimensionnel. Nous proposons un cadre général de simulation et d'apprentissage appelé *GillesPy* pour représenter les interactions des particules avec et le long d'un substrat unidimensionnel. Les cellules individuelles sont simulées indépendamment et leurs états sont superposés pour reproduire les données NGS. Le cadre remplit deux objectifs principaux : premièrement, il permet de relier la dynamique des cellules uniques aux données de séquençage statiques basées sur la population ; deuxièmement, il permet de prédire les distributions NGS manquantes entre les points temporels échantillonnés, en montrant comment les processus nucléaires, tels que la réparation de l'ADN ou la transcription des gènes, se déroulent réellement au fil du temps. L'algorithme est

implémenté indépendamment de toute voie métabolique, et il comprend une procédure d'estimation des paramètres spécifiquement développée pour relier les taux d'interaction des cellules uniques aux données de distribution basées sur la population. En effet, nous pouvons montrer que cet algorithme peut approximer de manière raisonnable les données CPD-seq et ChIP-seq de Pol II. Étonnamment, les simulations stochastiques dépendantes des cellules sont en accord avec l'approximation *mean-field*, et les deux suggérant une réparation spécifique aux gènes. Le modèle indique également qu'un taux d'interaction élevé entre Rad4—qui, nous le supposons, représente la GGR—et l'ADN pourrait être nécessaire dans certaines régions. En déconvoluant les signaux NGS simulé en cellules individuelles, nous révélons que la réparation par GGR nécessite le sondage de nombreuses positions, car l'association et la dissociation aléatoires sont peu susceptibles de trouver des dommages. Par conséquent, les interactions des protéines impliquées dans la GGR avec l'ADN pourraient être présentes à une échelle beaucoup plus grande que ce que l'on pensait jusqu'à présent, et le temps de réparation plus long généralement associé provient uniquement de la manière comparativement inefficace de rechercher les CPD. À notre connaissance, cette compréhension de la GGR au niveau des gènes n'a jamais été proposée auparavant. En évaluant le modèle *GillesPy* à l'aide d'autres processus liés à l'ADN, tels que la transcription médiée par TFIIH, nous pouvons démontrer que notre cadre de simulation et d'entraînement peut être appliqué à une grande variété de mécanismes indépendants de la réparation de l'ADN. Le modèle *GillesPy* est donc capable de fournir des informations temporelles et spécifiques aux cellules sur les processus nucléaires.

Discussion et conclusion

Dans ce travail, nous avons analysé la réparation des CPD chez *Saccharomyces cerevisiae* en combinant des approches d'évaluation des données avec des modèles informatiques. Nous avons d'abord effectué une analyse guidée par les données pour obtenir une vue d'ensemble de la cinétique de réparation et des facteurs pouvant l'influencer. La modélisation mathématique a ensuite permis une meilleure explication des signaux NGS, en établissant le lien manquant entre le processus de réparation et les données basées sur la population. Ce faisant, nous avons cherché à fournir une compréhension holistique de la cinétique d'élimination des lésions spatio-temporelles, en particulier au niveau des gènes codant pour des protéines.

En analysant la coordination du positionnement des nucléosomes le long des gènes, nous avons déterminé que les souches WT devraient être capables de découpler largement la dynamique des complexes d'histones de la présence de complexes multi-protéiques, telles que Pol II, le long des régions codantes. Cette hypothèse a été confirmée par la comparaison des paramètres de réparation

continue dans le temps avec la densité des nucléosomes dans les zones transcrites, qui n'a pas révélé d'impact significatif. Nous n'avons donc pas pris en compte le positionnement des nucléosomes dans le corps des gènes dans nos modèles mathématiques. Cependant, les niveaux de transcription étaient significativement corrélés avec les paramètres de réparation, soulignant le rôle de la détection des dommages par la Pol II au cours de la TCR. Le modèle KJMA a également suggéré un changement spatio-temporel le long des zones de TCR. Nous avons constaté deux profils temporel de réparation : un déclin rapide et précoce des niveaux de CPD autour du début du gène, et une influence croissante de la réparation tardive à mesure que l'on s'éloigne du TSS. Bien que la méthode ne permette pas d'expliquer comment et pourquoi les paramètres varient, elle prouve néanmoins que la cinétique de réparation a tendance à changer dans l'espace. Afin de fournir une compréhension mécanistique de l'élimination des dommages à l'ADN dans une cellule unique, nous avons utilisé deux approches, à savoir la modélisation *mean-field* et l'échantillonnage de Gillespie, pour reproduire les données dans les régions de la TCR. Il est surprenant de constater que, malgré leurs différences méthodologiques fondamentales, les deux modèles prédisent une dynamique de réparation spécifique aux gènes. Cela signifie que les paramètres changent en fonction d'un facteur d'influence de réparation inconnu qui varie entre les régions codantes. En outre, les deux approches suggèrent qu'un sous-ensemble substantiel de gènes nécessite une présence importante de GGR—ou d'un autre facteur de réparation qui n'est pas lié à la dynamique de la TCR médiée par la Pol II—pour décrire l'évolution des données CPD à l'échelle de la population. L'algorithme stochastique *GillesPy* peut relier la dynamique mécanistique spécifique d'une cellule aux données NGS statiques. Les paramètres entraînés ont indiqué que les protéines de reconnaissance liées à la GGR, telles que Rad4, doivent s'associer et se dissocier statistiquement de nombreuses fois le long de l'ADN avant que le site de la lésion ne soit trouvé. Alors que les interactions aléatoires peuvent parfois conduire à une élimination rapide des CPD dans les cellules individuelles, il faut beaucoup plus de temps pour observer un changement induit par la GGR à l'échelle d'une population. La TCR médié par la Pol II, en revanche, bénéficie d'une approche de balayage systématique grâce à un mouvement dirigé, et les effets de la TCR deviennent plus rapidement observables dans les données NGS. Bien que les protéines spécifiques à la GGR puissent interagir rapidement avec l'ADN et être présentes à des niveaux élevés, la dynamique d'élongation dirigée permet en moyenne une reconnaissance plus rapide des dommages.

Nous avons développé deux approches de modélisation mathématique pour expliquer la cinétique d'élimination des lésions de l'ADN par la dynamique des protéines. En utilisant le cadre *GillesPy* présenté, nous pouvons tirer des conclusions sur la dynamique des cellules uniques. Nous incluons des données et des résultats tirés de la littérature afin d'établir le lien manquant pour une

compréhension mécanistique. Cependant, le seul élément observable pour vérifier nos modèles (c'est-à-dire les données de séquençage) provient de cultures cellulaires entières, et un modèle de référence—qui pourrait soutenir la validité et comparer les performances de nos méthodes—est manquant. Nos résultats sont donc exclusivement basés sur la dynamique de millions de cellules en même temps, et la transposition de cette cinétique à des cellules uniques n'est pas simple. En tenant compte de cela, il est extraordinaire que nos deux modèles—c'est-à-dire le modèle de réparation par trafic en tant que l'approche *mean-field* et l'algorithme stochastique *GillesPy*—prédisent des dynamiques de réparation similaires le long des gènes. Cela inclut même le comportement des paramètres cachés qui régissent la GGR. Cela renforce notre confiance dans les conclusions mécanistiques concernant les cellules uniques. Il s'agit en particulier de la réparation spécifique aux gènes et du fait que les protéines liées à la GGR pourraient être présentes plus tôt ou à des taux plus élevés que ce que l'on pensait jusqu'à présent. Nous avons également établi qu'il existe une relation fonctionnelle entre les paramètres de réparation, ce qui indique la présence d'un autre facteur externe que nous n'avions pas pris en compte. Bien que le positionnement des nucléosomes—que nous n'avons pas inclus dans nos modèles—puisse effectivement avoir un impact, nous pensons qu'il est peu probable qu'il puisse expliquer exclusivement la tendance observée en raison des résultats présentés aux chapitres 2 et 3. Dans l'ensemble, les résultats indiquent l'existence de mécanismes de régulation spécifiques pour coordonner l'élimination des lésions à l'échelle génomique. Nous émettons l'hypothèse qu'un autre paramètre—tel que les interactions entre le Médiator et les protéines de réparation (Eyboulet et al. (2013)), les marques d'histones (Sun et al. (2020)) ou autres—affecte fortement la réparation spécifique des gènes. Nous espérons vivement que les projets futurs aborderont cet aspect du point de vue de la modélisation mathématique et de la biologie expérimentale. Nous avons établi que les différents temps de réparation pour la TCR et la GGR sont liés au mouvement des protéines le long du gène. Les temps de visite précoces et uniformes prédits pour les protéines de reconnaissance GGR nécessitent enfin une confirmation expérimentale basée sur des données de cellule unique, qui pourraient être produites par des travaux s'appuyant sur nos résultats. Bien que le séquençage d'une seule cellule puisse effectivement mettre en lumière la dynamique dépendante de la cellule, nous tenons à souligner que le séquençage basé sur la population présente ses propres avantages. En effet, il combine plusieurs états cellulaires et convient donc parfaitement aux modèles temporels continus en raison du chevauchement continu des changements d'état. Les données NGS constituaient donc un choix judicieux pour la modélisation d'un processus temporel à plusieurs étapes tel que la NER. Néanmoins, nous souhaitons encourager les futurs projets de recherche visant à comprendre l'interaction des différentes voies de reconnaissance le long des régions codantes à l'échelle de la cellule unique.

Les modèles informatiques peuvent être utilisés pour faire des prédictions dans différents contextes. En modifiant les paramètres ou les interactions protéiques modélisées, il est possible de faire des projections sur le processus de réparation dans des conditions mutantes. Cela peut être fait directement pour le modèle de réparation par trafic. La mise en œuvre du cadre *GillesPy* nécessite toutefois un nouvel entraînement des paramètres du modèle, car ils sont tous liés les uns aux autres et déterminent le taux d'échantillonnage. Néanmoins, la modification relativement facile de ces méthodes peut être particulièrement utile pour évaluer les conséquences liées aux maladies humaines. Par exemple, nos modèles attribuent un rôle important à GGR dans un sous-ensemble de gènes. En effet, des données internes au laboratoire suggèrent que les souches *rad7*Δ sont plus sensibles aux UV. Des mutations dans l'homologue humain XPC de la protéine Rad4 liée à GGR peuvent entraîner le Xeroderma Pigmentosum (XP). Avec le fait que XPC agit également comme cofacteur de Pol II (Bidon et al. (2018)), cela pourrait éventuellement suggérer un comportement altéré de la TCR le long des gènes codant pour les protéines. Par conséquent, le phénotype lié à XP pourrait être non seulement lié à des dommages dans les séquences non codantes et régulatrices, mais aussi à une dynamique changeante de la TCR. Les travaux futurs pourraient utiliser nos modèles pour comprendre la réparation dans des contextes génétiques changeants afin de mieux comprendre les troubles humains liés à la NER.