Genome-wide link between DNA replication and genome instability at the single cell level

Joseph Mark Josephides

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Institut Curie, UMR 3244

Genome-wide link between DNA replication and genome instability at the single cell level

Recherche pangénomique du lien entre la réplication de l'ADN et l'instabilité du génome au niveau de la cellule unique

Soutenue par
Joseph Mark JOSEPHIDES
Le 06 décembre 2023

École doctorale n° 515
Complexité du vivant

Spécialité
Génétique/Génomique

Composition du jury :

Leïla PERIÉ Directrice de recherche Institut Curie, PSL, CNRS, Sorbonne Univ.	*Présidente*
Aura CARREIRA Directrice de recherche CBMSO	*Rapporteuse*
Guillem RIGAILL Directeur de recherche INRAE	*Rapporteur*
Aurèle PIAZZA Chargé de recherche ENS de Lyon, UCBL1	*Examinateur*
Thierry VOET Professeur KU Leuven	*Examinateur*
Chun-Long CHEN Directeur de recherche Institut Curie, PSL, CNRS, Sorbonne Univ.	*Directeur de thèse*
Tatiana POPOVA Chargée de recherche Institut Curie, PSL, INSERM	*Codirectrice de thèse*

institutCurie | PSL

Abstract (French)

La réplication de l'ADN est essentielle pour les cellules, car elle permet de créer les quelque 30 000 milliards de cellules qui composent le corps humain à partir d'un seul zygote lors de l'embryogenèse. De plus, tout au long de la vie humaine, la réplication continue de l'ADN et la division cellulaire sont nécessaires pour remplacer les cellules âgées, mortes ou endommagées. Par conséquent, il est crucial que le programme de réplication de l'ADN fonctionne correctement à chaque division cellulaire. Cependant, de nombreux facteurs de stress, à la fois exogènes et endogènes, remettent régulièrement en question l'intégrité de l'ADN, ce qui entraîne une instabilité du génome. Cette instabilité est une cause majeure de cancers et d'autres maladies humaines.

Malgré l'importance du stress de réplication et de l'instabilité génomique dans les cancers, nous ne comprenons pas complètement les mécanismes sous-jacents ni leurs impacts sur le génome. Au cours de la dernière décennie, d'énormes progrès ont été réalisés dans l'analyse des cellules individuelles. L'étude des variants structuraux (VS) au niveau cellulaire est devenue cruciale pour comprendre l'instabilité génomique, en particulier dans des populations cellulaires hétérogènes telles que les échantillons de tumeurs, qui ne peuvent pas être facilement obtenus par des analyses de masse. Des études récentes ont révélé une corrélation importante entre le timing de réplication et l'apparition de VS dans les cancers, montrant que de nombreux VS résultent de mécanismes liés à la réplication. Cependant, il existe un manque d'études détaillées sur les mécanismes précis, en particulier sur les liens entre réplication, transcription et VS au niveau de la cellule unique. Comprendre ces mécanismes est crucial pour lutter contre les principales maladies humaines.

Pour répondre à cette question, ce projet développe et utilise de nouvelles méthodes informatiques basées sur l'intelligence artificielle. Il vise à (i) étudier directement le timing de réplication dans les cancers en analysant le nombre de copies au niveau de la cellule unique et (ii) examiner les interactions entre la réplication et les VS au niveau de la cellule unique. Les signatures des VS découvertes dans ce projet pourraient contribuer à améliorer le diagnostic et à définir de meilleures stratégies thérapeutiques. Dans l'ensemble, ce projet permet de mieux comprendre les mécanismes de la cancérogenèse et contribue à améliorer le diagnostic, le pronostic, le traitement et le suivi personnalisé des patients.

Mots clés :

Timing de réplication, Génomique, Cellule unique, Cancer, Intelligence artificielle, Variants structuraux.

Abstract (English)

DNA replication is a vital process of cells. Besides creating the ~30 trillion cells that comprise the human body from a single zygote during embryogenesis, continuous DNA replication and cell division is necessary during the entire human lifespan to replace the old, dead or damaged cells. It is therefore essential that the DNA replication program is correctly executed at each cell division. However, large numbers of exogenous and endogenous replication stresses routinely challenge DNA integrity and lead to genome instability, which is an important cause of cancers and many other human diseases.

Although replication stress and genomic instability are two important hallmarks of cancer, we lack full comprehension of the mechanisms that lead to these deregulations and the impacts they have on the genome. During the last decade, great progress has been made in analyses of individual cells. Determination of structure variations (SVs) in single cells has become an important approach to study genomic instability in heterogeneous cell populations, such as tumour samples, that cannot easily be obtained from bulk analyses. Recent studies have revealed that replication timing shows a strong association with the occurrence of SVs in cancers, and large amounts of SVs generated during tumorigenesis result from replication-associated mechanisms. However, studies addressing the direct mechanisms and, in particular, the links between replication, transcription and SVs at the single-cell level are missing. Investigating such mechanisms is critically important to address major human diseases.

To address this question, this project develops and uses novel computational methods, based on artificial intelligence, to: (i) directly investigate single cell replication timing (scRT) in cancers by single-cell copy number analysis, and (ii) examine the interactions of replication and SVs at the single cell level. The SV signatures in cancers revealed in this project might help to improve the diagnosis and better define therapeutic strategies. Altogether, this project provides further understanding of the mechanisms of carcinogenesis and contributes to improving the diagnosis, prognosis, treatment and/or personalised monitoring of patients.

Keywords:

Replication timing, Genomics, Single-cell, Cancer, Artificial intelligence, Structural variants.

Acknowledgements

I am thankful for the support I have received from a remarkable group of individuals throughout my PhD journey. It is with deep gratitude that I express my appreciation to my supervisor, Chunlong Chen, whose guidance, expertise, and support were instrumental and indispensable in completing this thesis. The opportunities he provided have been transformative, and I am genuinely thankful. I would also like to extend my warmest thanks to my dedicated colleagues, past and present, Manuela Spagnuolo, Ala Eddine Boudemia, Dalila Saulebekova, Minh Anh Vu, Nathan Alary, Win-Yan Bordes, Weitao Wang, Yaqun Liu, Stefano Gnan, and Christelle Lalanne. Their camaraderie and collaborative spirit made this academic journey truly fulfilling. Special recognition goes to Dalila, whose friendliness and kindness have been a source of motivation.

I would also like to express my appreciation to Eliana El Dawra and Jeanne Rakotopare, whose support and encouragement were unwavering throughout this process. I am also grateful to our assistants, Marie-France Lavigne and Win-Yee Liu Lefranc, for their tireless assistance and dedication.

I must also acknowledge the contributions of individuals outside my research unit, including Tatiana Popova, Marc-Henri Stern, Romain Koszul, Eric Letouzé, Josh Waterfall, Léa Wurges and the teaching staff at ENS. Furthermore, I would like to express my gratitude to my thesis jury members Aura Carreira, Guillem Rigaill, Leïla Perié, Aurèle Piazza and Thierry Voet, whose expertise and feedback have been important for the completion of this project.

Lastly, I want to thank my parents for their unwavering support and encouragement. To everyone mentioned, and to all those who contributed, no matter how small, to this endeavour, I extend my deepest gratitude. This thesis would not have been possible without your support and involvement.

Parts of this thesis were prepared using a limited access dataset obtained from BC Cancer and does not necessarily reflect the opinions or views BC Cancer.

Table of Contents

Abbreviations

A	Adenine
a-EJ	Alternative End-Joining
AI	Artificial Intelligence
BER	Base Excision Repair
BRCA1	BReast CAncer gene 1
BRCA2	BReast CAncer gene 2
BrdU	Bromodeoxyuridine
C	Cytosine
CDK	Cyclin-dependent kinase
CNAs	Copy-Number Alterations
CNVs	Copy-Number Variations
day-7	mouse embryonic stem cells after 7 days of differentiation into neurectoderm cells
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DKO1	Double Knock-Out of maintenance DNA methyltransferase DNMT1 and de novo DNA methyltransferase DNMT3B
DNA	Deoxyribonucleic acid
DSB	Dingle-Strand Breaks
EJ	End-Joining
EM	Expectation-Maximisation Algorithm
ERpos	Estrogen-Receptor positive
FACS	Fluorescence-Activated Cell Sorting
FISH	Fluorescence In Situ Hybridization
FNA	Fine-Needle Aspiration
G	Guanine
G0	Gap 0
G1	Gap 1
G2	Gap 2
HGP	Human Genome Project
HGSOC	High-Grade Serous Ovarian Cancer
HR	Homologous Recombination
kb	Kilobase
KNN	K-Nearest Neighbours
Mb	Megabase
MCM	MiniChromosome Maintenance
mESC	mouse Embryonic Stem Cells
MMEJ	Microhomology-Mediated End-Joining
MMR	MisMatch Repair
NER	Nucleotide Excision Repair
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End-Joining
ORM	Optical Replication Mapping

PDX	Patient-Derived Xenograft
RNA	RiboNucleic Acid
RT	Replication Timing
S-phase	Synthesis-Phase
sc	Single-Cell
scRT	Single-Cell Replication Timing
SNP	Single Nucleotide Polymorphism
SSA	Single-Strand Annealing
SSB	Single-Strand Breaks
SV	Structural Variant
T	Thymine
TNBC	Triple-Negative Breast Cancer
UMAP	Uniform Manifold Approximation and Projection
WGS	Whole-Genome Sequencing
WT	Wild-Type

1. Introduction

1.1. Timeline of genomics

1.1.1. Pre-genomics era

1.1.1.1. Foundations from antiquity to modern history

Throughout history, the concept of genetics has evolved and although it was not known to ancient civilisations, they laid the basis for our understanding of life. For instance, the ancient Greeks had made a fundamental contribution to the field of biology. This was done by founding the scientific method and employing systematic observation, experimentation, and hypothesis testing to explore nature. This approach is still used to guide modern-day research. In fact, a notable insight from antiquity came from the Greek philosopher Democritus. He proposed a theory where all matter consists of small particles called atoms. This concept was then revived in the 17th century through the work of the English chemist John Dalton who further suggested that atoms of different elements could combine to form molecules [1], a theory that changed our understanding of matter.

As we journey through the history of genetics, we stumble upon a significant milestone in 1869 when the Swiss chemist Friedrich Miescher identified a mysterious substance within the nuclei of white blood cells, which he named nuclein[2]. This substance, which is now known as Deoxyribonucleic acid (DNA), changed our understanding of heredity. The 20th century then saw a flurry of investigations in the relationship between DNA and heredity. In 1902, two biologists from different corners of the world, Walter Sutton from the United States and Theodor Boveri from Germany, independently introduced the chromosome theory of inheritance. This theory shaped the idea that chromosomes carry genes, the hereditary units of life, which allow to pass traits from parents to offspring[3]. Therefore, this discovery can be considered as a pivotal moment in our understanding of how traits are passed down through generations of living organisms.

1.1.1.2. The double helix

In 1953, another very important discovery was made. James Watson and Francis Crick published their ground-breaking paper describing the structure of DNA as a double helix[4]. This monumental finding earned them a Nobel prize in 1962 along with Maurice Wilkins[5]. However, it is important to note that a major contribution to this finding is also owed to Rosalind Franklin, a physical chemist who was working at King's College London. Despite being often overlooked both historically and in research papers, Franklin's efforts were significant in this work[5,6]. Nonetheless, this finding laid the foundation for a new field that we now know as genomics, the study of the entire genetic material in an organism, and thus, marked the birth of a novel branch in biology.

The DNA molecule is a marvel of nature, consisting of two intertwined strands of nucleotides, organic molecules that serve as monomeric units of DNA, which are complementary to each other. This means that the bases on one strand pair up with the corresponding bases on the other strand. Specifically, adenine (A) pairs with thymine (T), and guanine (G) pairs with cytosine (C). This complementary base pairing is a critical feature of DNA and it allows for the precise replication of the DNA molecule during cell division. When a cell divides, the double helix unwinds, and each strand operates as a template for the synthesis of a new complementary

strand. As a result, the two new DNA molecules produced are identical to the original template, ensuring the faithful transmission of genetic information throughout the future cell generations.

1.1.2. 1950s - 1970s: Sequencing

The origins of DNA sequencing which consists of decoding the molecular sequence of nucleotides, also called bases, of a DNA molecule date to the late 1950s and early 1960s. These first methods, which started to decrypt the genetic code, although slow and experimentally demanding, laid the foundation for the development of faster and more efficient sequencing technologies in the following years.

One of the pioneering DNA sequencing methods was developed by Frederick Sanger in 1997[7]. Sanger's technique involved using radioactive nucleotides to label fragments of DNA which were then separated by size using electrophoresis. The order of the nucleotides in the fragments could then be determined by reading the resulting autoradiograph. While Sanger's method marked a significant development in DNA sequencing, it was still relatively slow and expensive.

1.1.3. 1980s - 2000s: Genome projects and new technologies

1.1.3.1. Genome Sequencing Projects

In the following years, the launch of large-scale genome sequencing projects took advantage of new technologies. The most ambitious genome sequencing project launched during this period was the Human Genome Project (HGP), a partnership initiated in 1990 which had the goal of sequencing the entire human genome [8]. The HGP generated a draft sequence of the human genome in 2001 and two papers, one in Science[9] and one in Nature[10], were published describing the first complete sequence of a human genome, a major milestone in the history of genomics. The HGP sequence revealed that the human genome is comprised of approximately 3 billion base pairs of DNA and approximately 20,000 genes. As a result, this endeavour not only deepened our understanding of human genetics but also led the way for further discoveries, such as in medicine[8]. The HGP was not just a scientific endeavour; it was a journey into understanding what makes us human.

In parallel to the HGP, the French human genome project was another significant partnership launched in 1993 and which played a noteworthy role in the HGP by making data publicly accessible. The French project contributed to the publication of the draft sequence of the human genome, all while training a large cohort of scientists in the booming field of genomics[11].

Beyond the HGP, a plethora of other genome sequencing projects were completed between the 1980s and 2000s. These projects sequenced whole genomes of a wide range of organisms including plants[12–14], animals[15–19], fungi[20–22] and bacteria[23–26]. As a result of these collective efforts, the emergence of this flourishing field of genomics offered great opportunities for researchers. However, with these advancements, a multitude of questions on the genomes' organisations, functions and interactions were raised. These questions underlined the complexity of life from the molecular level to phenotypes and sparked a new wave of research aimed at unravelling these mysteries. It seemed that the field of genomics was just getting started.

1.1.3.2. Microarrays

Microarrays, a technology which grew in the 1980s, have been implemental in studying gene expression and various other molecular processes across a wide range of organisms and diseases[27]. This technology works by hybridising labelled samples, which can include DNA, RNA, or proteins, to a microchip containing specific probes. These probes are tailored to the target molecules. The degree of hybridisation to each probe is then quantified, which can be translated as the concentrations of the corresponding molecules by proportionality.

Microarrays have played an important role in making discoveries related to gene expression, molecular interactions, and various biological processes, particularly in the context of human diseases[27]. However, it would be important to note that microarrays have some major limitations[28]. Firstly, the number of molecules that could be studied simultaneously was often limited to a predefined set represented on the microarray chip. Secondly, the results obtained from microarray experiments could be influenced by environmental factors, such as humidity and temperature. Thirdly, the cost of microarrays is relatively high for a single experiment. Lastly, the typical microarray workflow involved a number of manual steps, which could add variability and complexity into the analysis process. Due to these limitations, microarray technology slowly phased out of fashion and is now considered outdated. While it was once a useful tool in the field of genomics, it has now been replaced with more advanced technologies that offer other advantages (explained below).

1.1.3.3. Next generation sequencing genomic technologies

In the late 1990s and early 2000s, a ground-breaking revolution occurred in genomics with the development of Next-Generation Sequencing (NGS) technologies. NGS involves the process of dividing DNA into small pieces and then sequencing these fragments in parallel. These new DNA sequencing methods represented a significant leap forward from previous laborious techniques. With commercial DNA sequencers offering high-throughput sequencing in the early 2000s, remarkable advantages were noticed as they proved to be faster and more cost-effective[29]. Effectively, the expense of sequencing decreased rapidly, with the cost per megabase (Mb) of DNA beating Moore's law[29,30.] Consequently, this high-throughput approach enables NGS technologies to sequence entire genomes at a much faster pace compared to previous methods and at a fraction of the price.

The impact of NGS technologies on genomics has been profound[31] and have facilitated numerous important discoveries, including the identification of new genes and genetic variants associated with human diseases[29] and the development of new medical diagnostic tools[29,31]. Furthermore, NGS has contributed to the study of evolutionary processes in organisms[32–34] and the creation of new crops with desirable traits[34–36] among many other applications[29,31]. While short-read sequencing (e.g. Illumina NovaSeq, HiSeq, NextSeq; BGI MGISEQ, BGISEQ; Thermo Fisher Ion Torrent sequencers) provides reads of up to 600 bp, this may not be sufficient for some applications[37]. Long-read sequencing offers several advantages with reads that can exceed 10 kb and can therefore improve de novo assembly, mapping certainty, transcript isoform identification, and detection of structural variants[37]. In essence, NGS technologies have transformed the way we study genomics while also opening new research interests.

1.1.3.4. Functional Genomics

In 2003, the initiative known as the ENCODE project was launched with the primary goal of studying the function of the elements in the human genome[38]. Functional genomics is the

branch of genomics that explores how genes and their other associated genomic elements function to express specific phenotypes and can encompass techniques ranging from differential gene expression analysis to proteomics and other omics data. For instance, it can be used to discover the role of genes in specific diseases, the interaction between genes and the environment or how genes control development. Importantly, one of the most notable outcomes of the ENCODE project was the debunking of the 'junk DNA' assumption. Prior to this project, many supported the assumption that the 98.5% of human DNA that does not contain genes does not have a regulatory function of the genome[39]. However, ENCODE revealed that many of these regions actually play an important role in regulating cellular processes. Overall, this marked another major moment in the history of genomics and shifted the understanding of human and, by extension, eukaryotic genomes.

1.1.4. 2010s: Genomics becomes bigger and cheaper

1.1.4.1. Synthetic biology

Synthetic biology – the field that involves engineering organisms to give new abilities – is being used to develop new diagnostic tools, vaccines and new treatments for cancer and other diseases[40,41]. The advent of CRISPR, a cutting-edge gene editing technology allowing to insert, delete, or replace the DNA at a desired site [42], has made gene modifications much easier for researchers. Consequently, synthetic biology became a rapidly growing field with potential to revolutionise many industries, most notably medicine and agriculture[41,43]. In 2015, it was estimated that genetically modified crops covered 70.9 million hectares of land in the United States[44], exemplifying a non-negligible impact of synthetic biology. Yet, one must bear in mind the various ethical concerns associated with genome editing, such as the potential for misuse or mismanagement of synthetic organisms[45]. As we continue to push the boundaries of the applications of synthetic biology, discussions and observations on the ethical implications of our advancements should also be considered.

1.1.4.2. Consumer genomics

In recent years, as the cost of DNA sequencing has declined, several companies have recognised an unprecedented opportunity to offer easily accessible and personalised solutions to individuals. Through the delivery of direct-to-consumer home kits, these companies collect, sequence, and analyse customers' DNA to provide what they claim to be information on the individuals' ancestry, health risks and physical traits.

While these DNA tests may seem appealing to the general public, they also raise important concerns[46–48]. One notable concern revolves around the accuracy of the information provided. The interpretation of genetic data can be complex, and errors or misunderstandings may occur. Additionally, the reliability of some health-related findings from these tests may not be well-established or validated through rigorous scientific research. Furthermore, it is widely acknowledged that this industry often monetises the genomic data it collects, either through research collaborations or marketing endeavours [49]. These data are also a target for malicious individuals who are able to steal genomic data. Recently, leaked credentials from 23andMe were used by hackers to steal genomic data of the platform's users and were sold on the dark web for as little as $1,000 USD for 100 or $100,000 USD for 100,000 profiles ($1 per genome)[50]. These concerns should be addressed and explained to platform users to mitigate ethical and scientific implications.

1.1.4.3. Single-cell sequencing

Classical sequencing methods, often referred to as bulk sequencing, have traditionally provided an averaged perspective of the spatial, temporal, and genetic variability within a biological sample[51]. With the advent of single-cell sequencing, researchers are finally able to dive deeper into the relationships between individual cells[51,52]. This technique allows the genomes of individual cells to be sequenced, enabling the identification of rare or abnormal cell types which is achieved by studying cell-to-cell variability, an important limitation of bulk sequencing.

Ever since single-cell sequencing has since become a powerful tool for investigating the heterogeneity of cell populations and tissues, it has played a crucial role in identifying distinct subpopulations of cells in cancer and other diseases [53–56]. This kind of information opens new doors for the development of more precise and effective treatments for cancer as it paves the way towards new therapeutic targets and pathways[51].

Despite its great potential, single-cell sequencing comes with its own challenges, most particularly in managing and exploring the large amount of data it generates. To address this, the single-cell community is constantly developing new tools and methods to unlock the full potential of single-cell sequencing. Already, a large number of single-cell computational tools have been developed and made accessible to the research community[57–77]. The ongoing development of new tools in this field testifies the power and promise of single-cell sequencing in the forceable future.

1.1.4.4. Bioinformatics

As the amount of biological data continues to expand, it becomes increasingly challenging to make use of this vast amount of information. Fortunately, bioinformatics, a field that is situated on the crossroads of biology and computer science, has taken a leading role in the process of managing and understanding the vast amount of biological data that is constantly generated. While the term 'bioinformatics' first appeared in 1970 by Paulien Hogeweg and Ben Hesper [78], its widespread adoption within biology-oriented research laboratories occurred in the aftermath of large-scale genomic projects. Nowadays, it is now common practice for new biologists to possess basic programming skills.

Bioinformatics-led research has already yielded a plethora of important discoveries. These include the identification of genes associated with diseases [79–81], the development of new diagnostic tools [82–84], and transformative impacts on agronomy[34,35]. As a powerful tool, bioinformatics holds the potential to address a multitude of pressing scientific questions spanning various omic levels of life and across all species. Consequently, this field represents an important bridge between big data and biology and will most likely continue to grow and contribute to the understanding of the vast and complex datasets appearing through modern biological, environmental and medical research.

1.1.5. 2020s: Current and future uses

1.1.5.1. Comprehensive genome projects

As our knowledge expands and genomic data becomes more accessible, researchers are able to explore increasingly more complex biological issues. Ambitious international projects have emerged seeking to provide more comprehensive answers to fundamental questions. These

consortiums generate and explore large amounts of data which are often analysed with advanced bioinformatic tools.

One such project is the draft human pangenome reference[85], a comprehensive genome programme that encompasses the genetic diversity found within the human population. By sequencing and assembling the genomes of numerous individuals originating from various populations, a more representative version of the human genome was created. This new reference could be beneficial to researchers studying human genetics, diseases, and evolution. This is because a reference for the human pangenome has the advantage of making it easier to identify rare or population-specific genetic variations, thereby offering a window into the rich diversity of human genetics.

Recently, the telomere-to-telomere consortium published a complete sequence of all the DNA in a human chromosome, from one telomere to the other [86]. Telomeres are the caps at the ends of chromosomes that safeguard them from damage. Sequencing the telomeres and centromeres has been difficult in the past due to their repetitive sequences they enclose. However, with new technologies, scientists have now been able to sequence the complete human genome, including the telomeres and other problematic regions, presenting new opportunities for genomics [87]. As a result, chromosome-specific studies, such as that of the complete sequence of a human Y chromosome[88] have also emerged.

Beyond solely human-oriented projects, the vertebrate genome project is an international effort to sequence the genomes of all known vertebrate species[89,90]. By using the genomes issued from this project, evolutionary scientists will be able to compare the genomes of various vertebrate species more accurately and gain new insights in the relationships of vertebrates. This project is expected to take many years to be completed, but it will undoubtedly help researchers better understand the evolutionary history of life.

Finally, the pan-cancer analysis of whole genomes consortium published a series of studies in early 2020 on the genetic basis of cancer across all types of cancer[91]. These investigations were made possible because of recent advances in whole-genome sequencing and high-performance computing. The pan-cancer studies have provided new insights into the development and progression of cancer. For example, novel findings were made by comparing mutational signatures[92], clonal evolution[93], and patterns of structural variants[94] of thousands of cancers. Although none of the published datasets from these studies contained single-cell data, possibly limiting the results in terms of intra-cancerous heterogeneity, these studies open new prospects for fundamental research in cancer.

In conclusion, these ambitious and large-scale projects represent the cutting edge of genomics research, providing advances in tackling some of the most pressing questions in biology. As these investigations, along with other projects, continue to generate new data, our understanding of the relationships, implications and functions of DNA will undoubtedly continue to change.

1.1.5.2. Artificial intelligence

Nowadays, datasets are larger than ever before, computational units are more powerful and complex questions continue to challenge biologists and medical practitioners. Therefore, it is only natural that bioinformaticians are turning to the latest wave of computational excellence – artificial intelligence (AI)[95]. AI, a branch of computer science, focuses on creating models capable of autonomous reasoning, learning, and action. AI research has achieved remarkable

success in developing effective techniques to solve a diverse array of problems, ranging from strategic game playing[96] to medical diagnosis[97] and autonomous vehicles.

AI algorithms can be categorised into two main branches: supervised learning and unsupervised learning. In supervised learning, models are trained on labelled data, such as providing images of animals with corresponding animal names. Unsupervised learning, on the other hand, comes into play when data lacks these labels, and elements are grouped based on their similarities. Beyond these two categories, there are also other approaches, such as semi-supervised learning, where not all the data is labelled, and reinforcement learning, where models are trained through a reward system, such as a self-driving car learning how and encouraged to stop at a red light. Finally, deep learning is an umbrella term that encompasses artificial neural networks, inspired by the structure and function of the human brain, and can be trained to perform a wide range of tasks, from image recognition to language processing.

Although machine learning, a branch of AI allowing software applications to become accurate in predicting outcomes, has existed since 1959 when IBM employee Arthur Samuel coined the term[96], it has increasingly gained momentum in genomics in recent times (Figure 1). Thus, these methods are on the rise and for good reason.

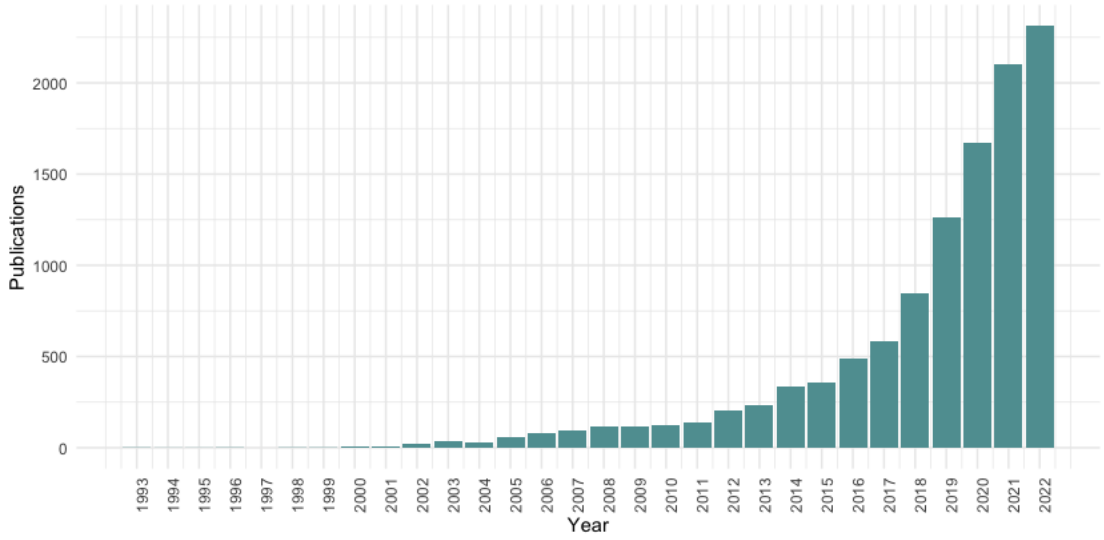


Figure 1: Annual number of publications listed on Pubmed containing the keywords 'genomics' and 'machine learning'.

Multiple machine leaning algorithms have been used in genomics (Figure 2) and are starting to become common practice. In fact, some recent publications have exemplified the great power harvested from AI algorithms applied on genomic datasets. These can range from simple visualisations such as dimension reduction on genomic datasets[62,68,98], to more insightful and predictive tools[95,97,99–101]. One example is DeepVariant, is a deep learning model that can identify genetic variants from NGS data[102]. This tool is more accurate than traditional methods for identifying genetic variants and identified specific variants in cancer and Alzheimer's disease. Another famous AI tool is AlphaFold, a deep learning-based protein structure prediction system developed by DeepMind, a subsidiary of Google. Its latest version was published in Nature in 2021[103] and has since become a powerful tool for genomics research. AlphaFold, also based on deep learning, can predict the 3D structure of proteins from their amino acid sequence with high accuracy, which is a task that has been challenging the field for decades. This knowledge can be used to develop new drugs and treatments for diseases, as well as to design new materials

and enzymes. Hence, it is safe to say that the power harnessed by AI, and especially deep learning, as demonstrated by these two singular examples alone, is transforming genomics as we know it.

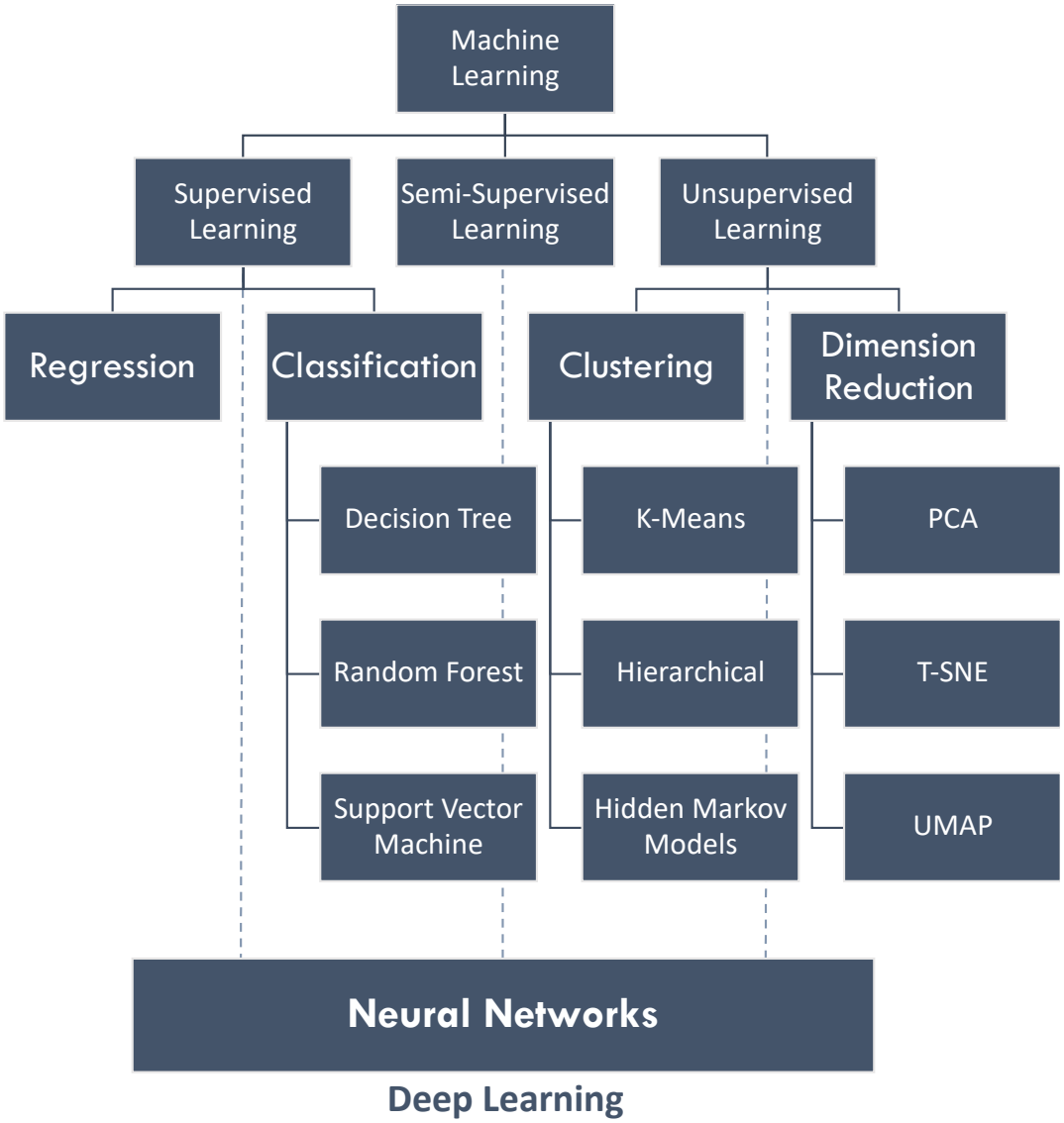


Figure 2: A non-exhaustive and simplified catalogue of the leading AI algorithms employed in genomics*. Machine learning approaches in genomics can be categorized according to the type of data used for model training. Labelled data can be used for supervised learning, unlabelled data can be used for unsupervised learning and partially-labelled data can be used for semi-supervised learning. Supervised learning can be useful for classification or regression while unsupervised learning can be useful for clustering or dimension reduction. Deep learning is based on neural networks and can be based on supervised, semi-supervised or unsupervised learning.*

1.1.5.3. Precision medicine

Precision medicine, also known as personalised medicine, is an emerging approach in healthcare that considers individual variability in genes, environment, and/or lifestyle to provide tailored prevention and treatment strategies. This approach represents a shift away from the traditional one-size-fits-all model of healthcare, where most people receive the same treatment regardless of their individual characteristics. For instance, two people with the same type

of cancer may respond differently to the same treatment[104]. This is because the cancer cells in each person may have different genetic mutations, or other characteristics which influence how they respond the therapeutic solutions.

Genomics and personalised medicine hold the potential to revolutionise the treatment of diseases, prospectively leading to a transformative change in healthcare management. Although significant progress has yet to be made, new opportunities appear to be on the horizon. It is projected that advancements in patient-tailored therapies will emerge and have the potential to change how we approach and manage health challenges. As our understanding of the genetic and environmental factors continue to grow, so will our ability to develop new targeted therapies that could improve patient outcomes.

1.2. The cell cycle

The cell cycle is an essential process in life that can be described as the sequence of events a cell goes through as it duplicates and divides. It is a highly regulated process which is crucial for the growth, development, and maintenance of all known organisms. The cell cycle is divided into two distinct phases – interphase and the mitotic phase – controlled by checkpoints to ensure accurate cell division.

1.2.1. Interphase

Interphase is divided into three subphases – gap 1 (G1), synthesis (S), and gap 2 (G2). The G1 and G2 phases represent the gaps between DNA duplication during S phase and mitosis. Gap 0 (G0) is a phase that cells can enter to leave the cell cycle, which is the general case of neurons.

1.2.1.1. G1 phase

The G1 phase of the cell cycle is the first and longest part of the interphase, accounting for approximately 60-80% of the total cell cycle duration[105]. During G1, the cell physically grows, synthesises proteins and other molecules in preparation for DNA replication, and repairs DNA damage. G1 is also a time when the cell makes the decision of whether to enter the S phase and divide, a decision influenced by a variety of factors, including growth factors, nutrient availability, cell density, and DNA damage.

The G1 phase is regulated by a complex network of signalling pathways and transcription factors which are explained in section 1.3.4. Once this network has completed its task, if the cell has the necessary resources and is not under any stress, it will then be able to progress into S phase. However, if the cell is lacking in resources or is under stress, then it may arrest in G1 phase or even enter a state of senescence, generally an irreversible cell cycle arrest. Disruptions to the G1 phase can lead to a variety of diseases. For example, mutations in some genes can lead to uncontrolled cell proliferation and cancer [105].

1.2.1.2. S phase

The S phase of the cell cycle is a crucial stage during which DNA replication takes place. The mechanisms responsible for DNA replication vary across the three domains of life with bacteria, archaea and eukaryotes exhibiting somewhat different molecular paths. Here we will focus on the eukaryotic cellular replicating machinery, also known as the replisome. DNA replication is the complex process of duplicating the cell's DNA to ensure that each daughter cell inherits a complete and identical copy of the genome. However, the DNA replication process is not

initiated at completely random points across the genome. Specific origins are regulated by proteins and cell cycle kinases, followed by the coordination of different proteins and enzymes[106].

Indeed, in eukaryotic cells, the progression of the S phase is strictly regulated by a complex network of signalling pathways and transcription factors[107]. At the heart of this process lies the origin recognition complex (ORC), a protein complex that selectively binds to specific DNA sequences known as origins of replication[108]. ORC serves as a critical initiation factor of DNA replication and organises the assembly of the replication machinery.

Another key player at this stage is the minichromosome maintenance (MCM) complex, which is loaded onto the DNA molecule with the recruitment of Cdc6 and Cdt1[108,109]. CDKs play an important role in the process as they phosphorylate specific proteins involved in origin firing, thereby promoting the activation of the helicase, an enzyme that breaks the hydrogen bonds between the complementary base pairs, which will initiate DNA synthesis. Thus, the firing of the origins is tightly regulated to ensure that individual origins are only activated once per cell cycle to prevent any re-replication of DNA segments.

As the replication origins are fired with the helicase unwinding the DNA strands, the elongation phase where DNA polymerases add complementary nucleotides to the exposed DNA templates can begin[106,108]. The dynamic structure of open DNA is called the replication fork and enzymes called primases will synthesize short RNA primers on the single-stranded DNA to provide a starting point for DNA polymerases. The two templates of the fork are replicated in opposite directions. This allows the distinction of leading and lagging strands. Leading strands are synthesised continuously in the 5'to 3'direction, in continuity with the replication fork, while lagging strands are synthesised in the opposite direction and discontinuously in shorter fragments, called Okazaki fragments. The latter are finally joined by a DNA ligase to form a continuous strand.

Meanwhile, various enzymes and proteins insure the processivity and accuracy of the process. The sliding clamp loaders will load and unload the sliding clamps which are responsible for fastening DNA polymerases to the DNA templates. Single-strand binding proteins will stabilise the single-stranded DNA strands to prevent them from reannealing (reforming a double-stranded DNA structure) or forming secondary structures. Topoisomerases enzymes prevent DNA from getting tangled, avoid supercoiling DNA while also relieving topological stress in the DNA molecule to allow it to unwind. Eventually, replication forks from neighbouring origins will meet leading to the collision of the DNA polymerases which will then be released.

DNA repair mechanisms can correct any errors or mismatches in the new molecules to ensure the integrity of the genetic information (see Section 1.3.5). DNA replication ends with the completion of DNA synthesis, the disassembly of the replication machinery and verification of accuracy. If needed, DNA synthesis could be reinitiated, or additional replication origins could be activated (see Section 1.4.1). This intricate orchestration of molecular events during the S phase ensures the accurate and efficient replication of the genetic material, a fundamental process in cell division and inheritance. The completion of DNA replication marks the end of the S phase, and the duplicated molecules can then be separated into the daughter cells during mitosis (see Section 1.2.2).

1.2.1.3. G2 Phase

The G2 phase, the final stage of interphase in the cell cycle, is an important period where the cell continues its growth and prepares for mitosis. During this phase, the cell synthesises the

essential proteins and molecules. The decision to advance into mitosis depends on the cell's growth status and the integrity of its DNA. If conditions are favourable, the cell progresses to mitosis[107]. However, if resources are lacking or the cell experiences stress, it may arrest in the G2 phase.

During this phase DNA repair may continue, ensuring that the genetic material remains intact and functional[105]. Chromosomes undergo condensation, forming compact structures that facilitate their orderly separation during mitosis[110]. Simultaneously, the spindle apparatus, responsible for chromosome segregation in mitosis, initiates its formation[110]. Protein synthesis is also a key activity during this phase as the cell produces tubulin, a component of the spindle apparatus, and cyclins[110]. Thus, G2 phase plays a pivotal role in ensuring the cells are prepared to accurately distribute genetic material to the daughter cells.

1.2.2. Mitosis

Mitosis is the next phase in the cell cycle, following interphase. This step is characterised by the careful distribution of replicated genetic material, along with the centrosomes, the organelles involved in cell polarity, equally among the sister cells[110]. The final step of mitosis is the completion of chromosome cohesion, which is achieved when the condensed chromosomes are aligned on a metaphase plate, and the faithful inheritance of a complete and accurate set of chromosomes for future generations of cells is attained[110].

Mitosis is composed of several distinct stages (i.e. prophase, metaphase, anaphase and telophase), which terminate at the step of cytokinesis. This final process involves pinching the cell membrane for the formation of the two daughter cells, splitting the cytoplasm evenly. However, it is important to be aware that mitosis is a highly regulated and precise process meaning that any errors of chromosome segregation during this phase can lead to disorders and diseases such as cancer[110].

1.2.3. Checkpoints

Cell cycle checkpoints are essential for ensuring the integrity of cell cycle progress and essentially safeguard genomic integrity. They are strategically placed throughout the cell cycle to ensure that essential conditions are met prior to cells enter the following phase. Accordingly, the G1, S phase and G2/M checkpoints act as molecular arbiters of the cell cycle, responsible for verifying that the preceding steps were correctly executed[111].

Under normal conditions, when irregularities or anomalies are identified, these checkpoints initiate a temporary interruption of the cell cycle. For example, when the G1 checkpoint detects DNA damage, it requires a period of time for the DNA lesions to be diligently repaired[110]. These checkpoints are regulated by a complex combination of signalling pathways, transcription factors, and key regulators, such checkpoint kinases and tumour suppressor proteins[110]. In cancer, however, these checkpoints can be inactivated. Therefore, they play a crucial role in keeping the fidelity of the whole cell cycle process intact to prevent diseases.

1.3. Cancer genomics

1.3.1. Cancer biology

Cancer is not a simple concept; it is a group of diseases involving abnormal cell proliferations and the result of a complex combination of genetic and environmental factors. Genetic

alterations can be inherited or even acquired throughout life[93]. Hereditary variations are passed down from one generation to the next and have the capacity to evolve through family generations, for better or worse.

The set of functional capabilities acquired by cells that move towards tumoral neoplastic growth (i.e. abnormal and excessive tissue growth) are dubbed hallmarks of cancer. Although not mutually exclusive, many cancers exhibit multiple hallmarks. The original six hallmarks of cancer were presented in 2000[112] with further additions suggested later on (Figure 3)[113–116]. These hallmarks can be used not only for cancer research but also as targets for the development of novel and more effective targeted cancer therapies.

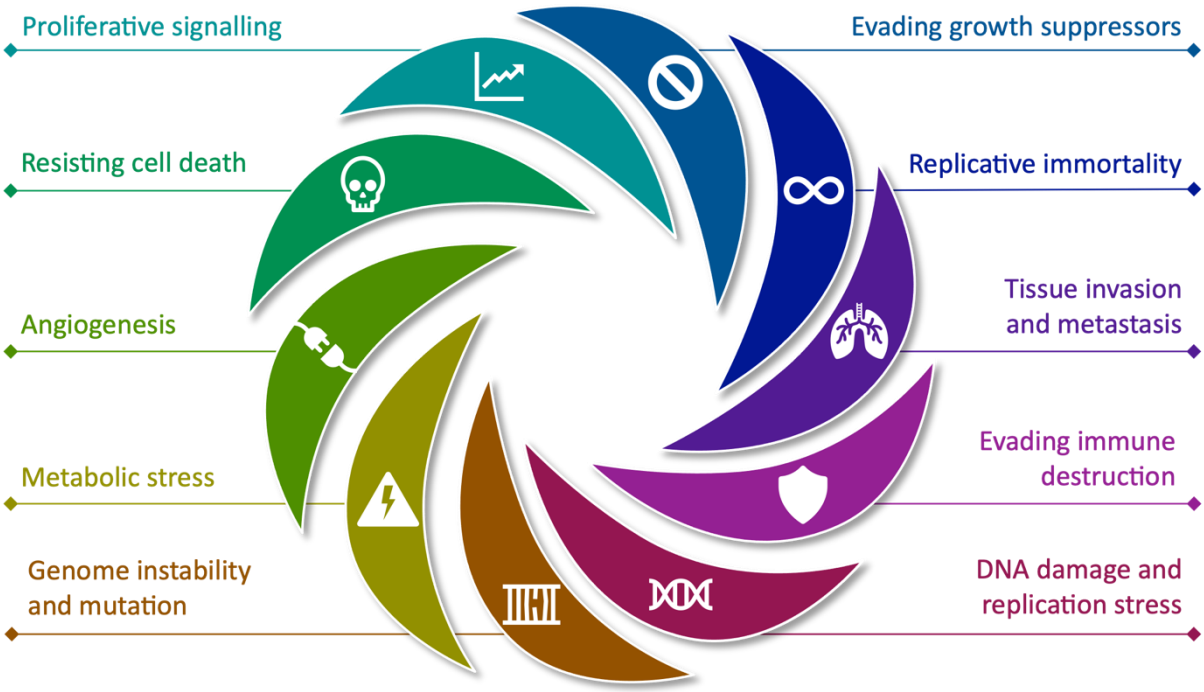


***Figure 3: A selection of the suggested hallmarks of cancer*[112–116].**

Intuitively, with the help of the previously mentioned advancements in the field of genomics, researchers have been able to uncover a vast collection of genetic changes associated with cancer. These alterations can range from single nucleotide polymorphisms (SNPs) – changes in a single letter of our genetic code – to insertions and deletions – changes in the number of copies of DNA in a genomic region. They can also include changes in the whole number of chromosome copies or rearrangements like jigsaw puzzles being rearranged. These genetic tweaks are grouped under the term structural variants (SVs)[94,117]. It is believed that they can initiate cancer by activating genes that drive uncontrolled cell growth, called oncogenes, silencing genes that normally keep cell growth in check, called tumour suppressor genes, or even obstructing the repair of damaged DNA[116].

Genome sequencing has demonstrated to be a powerful tool for understanding cancer, but also for diagnosing and treating it. By sequencing cancers, one can gain a better understanding of the genetic anomalies that are driving each cancer type. This allows selecting targeted therapies that are tailored to address the genetic abnormalities of the cancer cells[118]. Additionally, genomic sequencing opens the door to novel immunotherapy. By discovering the unique surface proteins that cancer cells present, treatments that mobilise the immune system to combat the disease can be created[104]. Hence, as the field of genomics continues to evolve, it is expected

that these new discoveries will have the potential to improve cancer treatments and thus, patient outcomes.

1.3.2. Structural variants in DNA

It was previously believed that the main source of genetic and phenotypic human variation was due to SNPs. However, the generation and analysis of data issued from new technologies, such as NGS, have uncovered a surprisingly large number of SVs, a term introduced in Section 1.3.1. In fact, it has been estimated that each individual has around 100 copy-number variations (CNVs) that are greater than 50 kilobases (kb) in size[117]. Thus, SVs is an umbrella term for types of genetic alteration which include:

i. Copy-number alterations/variations (CNAs/CNVs): These occur when a fragment of DNA that is present in an abnormal number of copies in comparison to the reference genome. Depending on the amount of these segments, they can be classified as insertions (addition of nucleotides), deletions (removal of nucleotides), or duplications (replicas of segments of DNA). Indels refer to both insertions and deletions.
ii. Inversions: This type of SV involves a fragment of DNA that is reversed in orientation. They can be pericentric (include the centromere) or paracentric (not including the centromere).
iii. Translocations: This occurs when a chromosomal fragment changes position within a genome without changing the total DNA content. These can be intra-chromosomal (within the same chromosome) or inter-chromosomal (onto another chromosome).

Other alterations that can also be considered as SVs include heteromorphisms (microscopically visible regions of a chromosome that vary in size or morphology), fragile sites (small breaks of chromosomes), marker chromosomes (additional chromosomes to the normal chromosome number), isochromosomes (a chromosome with two identical arms) and double minutes (small fragments of extrachromosomal DNA usually containing a particular locus; amplified genes as a result of chromothripsis)[117].

Additionally, aneuploidy is a term that refers to the presence of an irregular number of chromosomes in a cell. Although this has been reported to be a common feature of tumour genomes[94,119], its role in tumour development has been a matter of speculation. A recent report demonstrated that certain types of aneuploidies that are commonly found in cancer genomes play a crucial role in cancer progression[119]. Specifically, they found that 25% of cancers exhibit gains on the q arm of chromosome 1, and eliminating this abnormality increased the expression of TP53, an important tumour suppressor gene. Eventually, more drugs that demonstrate selective toxicity toward aneuploid cells could be used as anticancer agents.

1.3.2.1. Technologies for detecting structural variants

Initial evidence of human genetic variation was observed through microscopes[117], with karyotypes representing an individual's complete set of chromosomes. These consisted of condensed chromosomes that were mostly indistinguishable from one another, but aneuploidies, gross rearrangements and the Y chromosome were nonetheless identified[117]. As technology advanced, chromosome banding techniques and methods based on fluorescence in situ hybridisation (FISH) allowed for the detection of more subtle abnormalities such as large deletions, insertions, duplications, translocations, and inversions.

With the development of both experimental and computational methods, human SVs can now be analysed at a much higher resolution (Table 1). These methods can either be genome-wide or targeted, and include PCR, array, and computational approaches. Furthermore, recent advances in optical mapping enabled the identification of SVs across the genome by sequencing long stained DNA molecules[117].

Table 1: Selected methods for detecting SVs in the genome, adapted and updated from Feuk et al. 2006[117]. *Detection restrictions are in parentheses where applicable. NGS CNV methods are the only way to identify all SV types.*

	Method	Translocation	Inversion	CNV (>50 kb)	CNV indel (1–50 kb)	Small sequence variants (<1 kb)
Genome-wide	Karyotyping	Yes (>3 Mb)	Yes (>3 Mb)	Yes (>3 Mb)	No	No
	Oligonucleotide-based array-CGH	No	No	Yes (>35 kb)	Yes (>35 kb)	No
	SNP array	No	No	Yes	Yes	Yes (SNPs)
	Clone paired-end sequencing (fosmid)	Yes	Yes (break-points)	Yes (>8 kb of deletions)	Yes (>8 kb of deletions; <40 kb of insertions)	No
	NGS computational CNV detection[94]	Yes	Yes	Yes	Yes	Yes
	Optical mapping[120]	Yes (>50kb)	Yes (>30kb)	Yes	Yes	Yes (limited?)
Targeted	Microsatellite genotyping	No	No	Yes (deletions)	Yes (deletions)	Yes
	MAPH	No	No	Yes	Yes	Yes
	Real-time qPCR	No	No	Yes	Yes	Yes
	FISH	Yes	Yes	Yes	Yes	No
	Southern blotting	Yes	Yes	Yes	Yes	Yes

1.3.3. Tumour suppressor genes and oncogenes

The detection and analysis of genetic variation has come a long way since its initial observation. Despite significant differences between cancer types and individual cancers, they almost always share one point in common – they are genetically driven. Cancer-related genes can be classified in two distinct categories: oncogenes, which induce tumoral growth and development and tumour suppressor genes, which naturally inhibit cancerous development [116]. The human genome is believed to contain approximately 20,000 protein-coding genes, but some genes have gained attention in the cancer research community due to their mutation prevalence in cancer populations.

Oncogenes include PIK3CA, having single-nucleotide mutations in 17.8% of all cancers, KRAS, which can decrease the replication fork speed, and EGFR which stimulates cell proliferation[116]. Strikingly, TP53, a tumour suppressor gene that activates DNA repair proteins when DNA has sustained damage, is one of the most reoccurring genes in relative studies. It was estimated that this tumour suppressor gene has a single-nucleotide modification in over 40% of cancers, making it the most frequently mutated gene in human cancer[121]. While it is still not completely understood how this gene is so often targeted, its important role could create an ideal environment for strong evolutionary pressure leading to its inactivation and thus, allowing tumours to survive and proliferate. A previous study has shown that unlike other DNA repair genes, when mutated, TP53 does not induce specific types of SVs, making it a more general genome instability perpetrator[94].

1.3.3.1. BRCA1, BRCA2 and RAD51 genes

Prominently, two other tumour suppressor and protein-coding genes are also of great importance in cancer research: breast cancer genes 1 and 2 (BRCA1 and BRCA2), localised on chromosomes 17q21 and 13q12, respectively. These genes are implicated in many breast and ovarian cancers[122] as well as in DNA damage response pathways by interreacting with other genes, particularly RAD51. In the presence of a modification that makes either of these genes partially or completely defective, DNA damage may not always be correctly repaired, potentially leading to the rise and accumulation of serious SVs over time. When this kind of damage occurs in an important region of the genome, such as in gene bodies or regulatory elements, it can lead to the emergence or evolution of a cancerous cell.

It is well known that individuals carrying BRCA mutations on either allele have a significantly increased risk of developing breast and/or ovarian cancer in comparison with wild-type (WT) BRCA individuals[123]. The DNA repair mechanism in which BRCA genes are involved, called homologous recombination, detailed in Section 1.3.5, is essential for cell survival as it allows for the repair of any DNA damage caused by chemotherapy, radiation, and DNA replication stress, among other factors. The loss of function in either or both BRCA genes can lead to defects in the repair of DNA double-strand breaks[124]. BRCA1 and BRCA2 play a major role in the fidelity of DNA repair, and when defective, cells use backup strategies to attempt to repair these lesions by alternative but more error-prone mechanisms.

Nevertheless, the study of BRCA mechanisms has been fruitful as it is now known that tumours in which DNA repair pathways have been inefficient are most likely to respond to emerging targeted therapies, such as inhibitors of poly-ADP ribose polymerase (PARP), cancer drugs that target tumours with BRCA mutations[125,126]. RAD51 pathogenic mutations have also been identified in breast and ovarian cancers, showing similarities with BRCA1/2 mutations[124,127]. Furthermore, genetic testing for BRCA1 and BRCA2 pathogenic mutations has been valuable for defining eligibility for cancer screening and prevention programmes and methods for detecting these mutations are now widely accessible[128].

One of the reasons RAD51 has been linked to BRCA-less phenotypes is because RAD51 is also a key player in the homologous recombination pathway and interacts with BRCA2. This protein forms nucleoprotein filaments on single-stranded DNA which then search for and invade a homologous double-stranded DNA molecule[129]. Once this occurs, it forms a loop structure allowing the exchange of genetic information between the two DNA molecules. Besides its role in DNA repair, RAD51 has also been reported to be involved in replication fork processes. For instance, RAD51 allows DNA replication to restart when a replication fork encounters DNA damage[130].

1.3.4. Cyclin-Dependent Kinases (CDKs)

Some of the key genes that play a central role in regulating the eukaryotic cell cycle are cyclin-dependent kinases (CDKs)[131]. CDKs are a family of protein kinases that are activated by binding to regulatory proteins called cyclins (Table 2), proteins that accumulate and degrade during the cell cycle[132]. Besides regulating the cell cycle, CDKs also play a role in transcription and the differentiation of nerve cells[133].

The cell cycle can be deregulated in cancer cells due to genetic or epigenetic changes in CDKs. Recent studies have shown that while CDK1 is essential for embryonic cell division, CDK2, CDK4 and CDK6 are not essential for the mammalian cell cycle but remain important for the

proliferation of certain cell types[110]. CDK activity can be increased by DNA damage and changes in the cell cycle checkpoints, a combination that drives tumour cell cycles. Evidence suggests that inhibiting certain CDKs could offer a therapeutic prospect for certain cancers[110].

CDK12 has recently gained spotlight for its role regulating the cell cycle and impacting cancer. This kinase is associated with the elongation of RNA polymerase II and when depleted, it reduces the expression of homologous recombination (HR) DNA repair genes, among others, and therefore promotes genomic instability[131,134]. In fact, one study illustrated that CDK12 is indirectly required for G1/S progression because it is essential for the transcription of DNA repair genes (see Section 1.5.1)[131]. Furthermore, the absence of CDK12 results in premature cleavage and the loss of expression of genes larger than 45 kb[135] while inducing large tandem duplications across the genome[136]. Additionally, genomic alterations of this gene have been detected in a large number of cancer types with reports suggesting that up to 15% of cancers can be concerned by such mutations[137]. It can therefore be an important clinical biomarker for these cases and thus, a potential therapeutic target[137]. Indeed, CDK12 is a determinant of PARP inhibitor sensitivity[126]. Moreover, it has been shown that CDK12 is essential for the correct phosphorylation of RNA polymerase II, which enables the expression of the BRCA genes, and is often mutated In BRCA1- or BRCA2-defiecient cancers[126].

Table 2: A family portrait of the human CDK genes.

CDK	**Cyclin(-like) partners**[132–134]	**Functions**[132,133]	**Chromosome**[133]
CDK1	A, B	M	10
CDK2	A, B, D, E	G1/S, S, G2	12
CDK3	A, E, C, Cables1	G1	17
CDK4	C, D	G1	12
CDK5	p53, p59, Cables1		7
CDK6	D	G1	7
CDK7	H	transcription	5
CDK8	C	transcription	13
CDK9	K, T	transcription	9
CDK10		G2/M, transcription	16
CDK11A	D, L	transcription	1
CDK11B	D, L	transcription	1
CDK12	K, (L?)	transcription	17
CDK13	(L?)	transcription	7
CDK14	D, Y		7
CDK15			2
CDK16	p53, Cables1		X
CDK17	Cables1		12
CDK18	K		1
CDK19	C	transcription	6
CDK20			9

1.3.5. DNA damage response processes and genome stability

DNA damage is a constant occurrence in cells with an estimated minimum of 100,000 DNA polymerase errors occurring per cell cycle and around 20,000 potentially mutagenic lesions arising per diploid mammalian cell per day[138]. Considering that the average adult has between 28 and 36 trillion cells[139], many of which are cycling, this amounts to a colossal quantity of DNA damage incidents per day, highlighting the necessity to repair them to maintain cell functions and reduce cancer susceptibility. Therefore, the DNA damage response is required to be an efficient process to safeguard genome stability round the clock and evade the formation of SVs. Indeed, on average, only a mere ~10^{-10} mutations per bp per cell division escape the DNA repair radar[138], demonstrating the extraordinary fidelity of these pathways.

DNA breaks can occur on a single or on both strands, each requiring different ways to be repaired (Figure 4). Single-strand breaks (SSB) being the most common lesions occurring at a rate three times higher than double-strand breaks (DSB), and usually due to oxidative stress or abortive activity of DNA topoisomerase I[140]. These breaks are usually detected by PARP1 which acts as a first responder by initiating a decisive process that selects which DNA repair pathway can be used[140]. Base excision repair (BER) is the most common repair mechanism for SSB in mammalian cells, initiated with the recognition and removal of the erroneous base by a glycosylase enzyme, followed by a completion with a DNA polymerase[140]. Other SSB repair pathways include nucleotide excision repair (NER) and mismatch repair pathways (MMR).

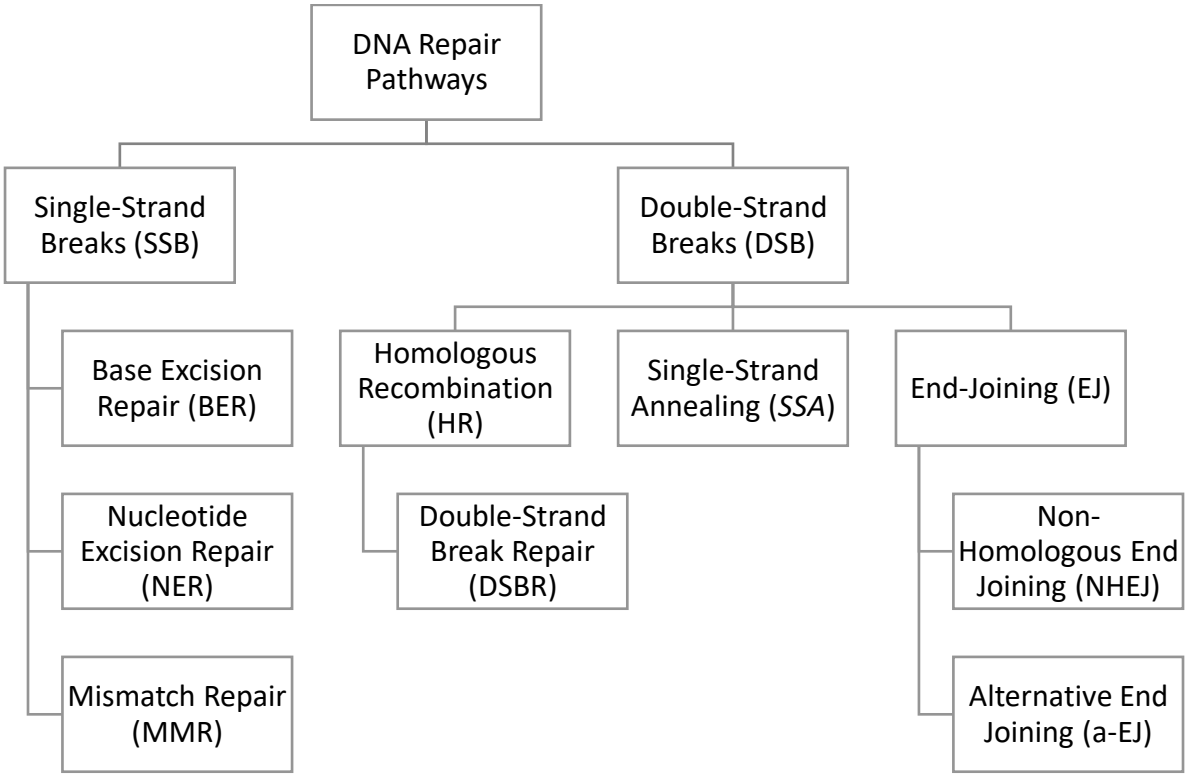


Figure 4: A classification of the main DNA repair pathways.

DSB can have a larger impact on the development of SVs because both strands need to be repaired. They can be mended by two main pathways: end-joining (EJ) and HR[141]. HR is considered to be a high-fidelity repair system that, with the help of BRCA1 and BRCA2, uses the undamaged sister chromatid as a template to restore the damaged DNA strand. During this process, RAD51, a key protein in HR that links to BRCA2, assists in the search and reparation

of damaged DNA. In various cancers, RAD51 can be linked to poorer patient survival[142], over-expressed[143], or under-expressed[144], leading to more DNA damage and potential mutations that could cause cancer. On the other hand, EJ pathways are only active during interphase and can be subdivided in non-homologous EJ (NHEJ) and alternative EJ (a-EL) pathways which are more error prone and therefore less preferred. NHEJ does not require a template and proceeds with direct ligation, making it prone to errors but probably accurate in most cases[141]. Microhomology-mediated EJ (MMEJ) uses sequences of microhomology as a template to repair the broken DNA while single-strand annealing (SSA) is able to mend homologous repeats between each other[145].

1.4. DNA replication timing

1.4.1. Introduction to DNA replication timing

As previously mentioned, DNA replication is a fundamental biological process in which, under normal circumstances, a cell creates an identical copy of its genome to ensure accurate transmission of genetic information to the daughter cells. However, one may wonder how and why the replication programme is regulated. A key metric in studying replication is replication timing (RT), which refers to the order in which different segments of the genome are copied during S phase. As previously mentioned, DNA replication is initiated at specific sites across the genome called replication origins[146], which each lead to the formation of a replication fork. The replication forks eventually merge when they meet to ensure that the whole genome is copied (Figure 5). Since this process is not simultaneously launched at all the possible replication origins at once, there is a selection and order which is specific to each cell type in which these origins are activated. The rate of replication fork elongation, which is a whole different parameter, is relatively stable and so the number of activated replication origins is the factor that will determine actual duration[146].

One critical question in this field is whether RT results from a stochastic or deterministic firing of replication origins. Early genome-wide RT studies found that budding yeast origin firing was better explained by stochastic firing whereas mammalian cells were firing origins in a deterministic manner[52]. Although stochastic origin firing has also been reported in mammalian cells[52], these observations could suggest that RT regulation was acquired over evolution. Moreover, it was suggested that this phenomenon is due to variability in mammalian cells seeming small due to longer S phases and the absence of gene-poor regions in yeast (gene significance discussed later). There has now been enough evidence supporting that the eukaryotic replication programme is globally deterministic (i.e. the firing probability of a given genomic region of a given cell type is pre-defined) but with individual replication initiation events displaying stochasticity[62,147–149].

Another interesting question is why replication patterns even change across cell types. In other words, why some regions replicate early in S phase for one cell type and late for others. It has been demonstrated that perturbating RT does not significantly change cell functions, possibly suggesting, at least at a first glance, that RT evolved to have specific patterns, without any major significance[146]. However, other studies have shown that RT is modified in cancer and other diseases, implying that it might play a larger role[62,146,150,151]. Replication stress can cause genomic instability, and promote tumorigenesis, making it a hallmark of cancer[116,152]. It therefore seems that RT is essential to maintain normal cell functions even though its biological significance has still not been fully elucidated.

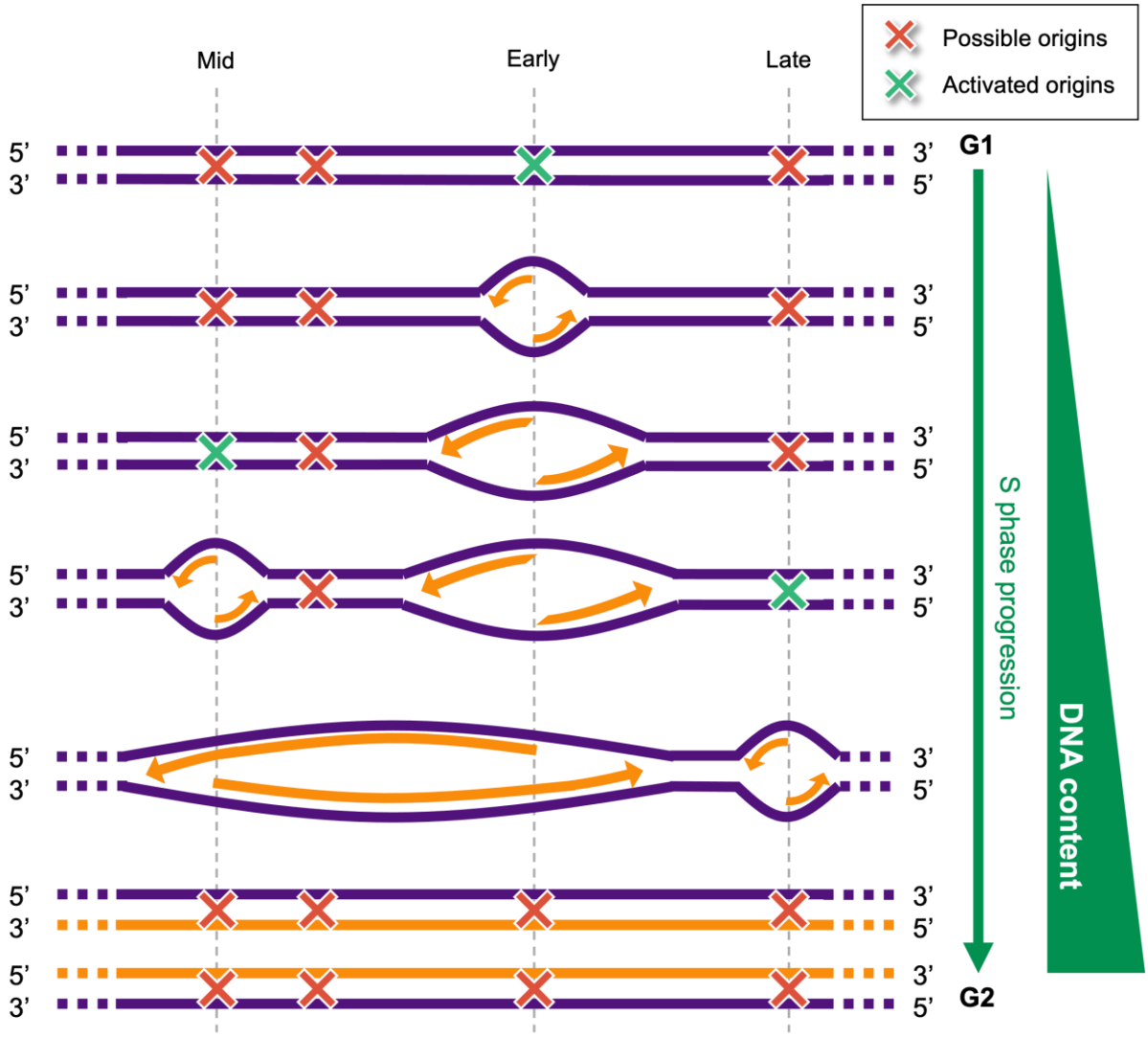


Figure 5: The chronological order of S phase events. *S phase is initiated with the activation of specific replication origins. This leads to the formation of a replication fork for each origin that elongates neosynthesised DNA in both directions. Later, more origins are fired across the genome and the resulting replication forks are eventually merged until there is an identical copy of the genome is generated and the cell can exit S phase. Replication origins can be activated at various moments of S phase which result in the classification of early, mid, and late replicating regions.*

1.4.2. RT, part of a multi-omic landscape

In order to further understand the role of replication patterns during S phase, a lot of research has been done in the past years. Resulting studies show that RT correlates with other cellular processes, including gene expression[153], DNA methylation[154], chromatin structure[155] and the 3D organisation of chromosomes[156] (Figure 6). Yet, many questions still remain unanswered around the exact relationship between these features and DNA replication.

The relationship between RT and transcription has long been a topic of interest[52]. Correlation between transcription and DNA replication was confirmed molecularly in the 1980s by analysing cell-type specific genes and then genome-wide with the emergence of microarrays on budding yeast and flies[153]. These findings demonstrated that budding yeast did not appear to show any significant RNA/RT correlation, in contrast to *Drosophila* which had transcribed regions generally replicate early. Since there appears to be some kind of link between transcription and

RT, it would be expected that changes in RT would impact transcription too, and vice versa. However, this is not always the case[157]. Consequently, the underlying reasons responsible for the correlation between transcription and RT remain a gap in our understanding.

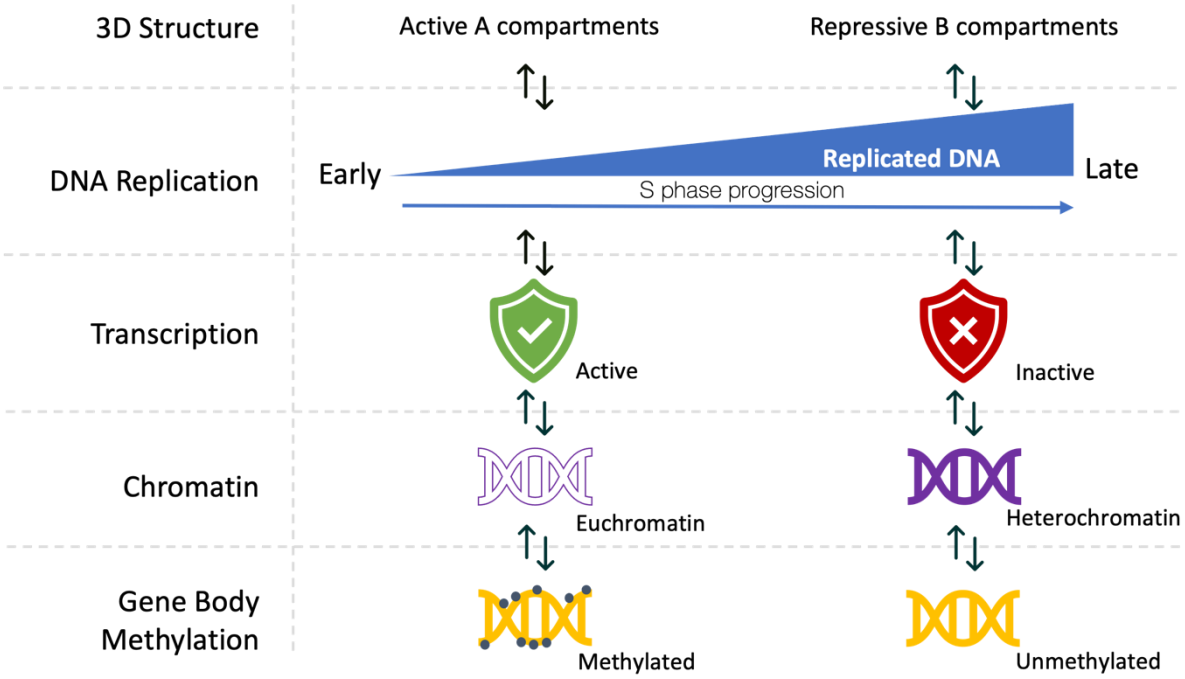


Figure 6: The multi-omic landscape related to RT. *Active A compartments, transcriptionally active regions, euchromatin and methylated gene bodies all correlate with early-replicating regions while B compartments, transcriptionally inactive regions, heterochromatin and unmethylated gene bodies correlate with late replicating regions.*

It is possible that these two features are linked by an intermediate factor and chromatin structure, DNA that is wrapped in histone proteins, could be this middle player. In eukaryotes, regions of DNA that are in an open chromatin state (euchromatin), and thus more accessible to the molecules surrounding them, are typically replicated earlier than those in a highly condensed state (heterochromatin)[155,158]. Chromatin, in hand, is linked to the 3D organisation of the genome with early and late regions, respectively, corresponding to A and B compartments in Hi-C data, an NGS chromatin conformation technology[52,154]. Finally, DNA hypomethylation leads to heterogeneity in RT[154].

1.4.3. DNA Replication: Methods of study

DNA combing experiments were among the first to study replication and they involved visualizing stretched Mb-sized DNA fibres on a microscope slide using fluorescently labelled nucleotides to visualise neosynthesised DNA. Thus, they could detect the density and origin of the replication. However, this technique did not allow to distinguish RT but was rather a method to study various DNA replication events (i.e. replication fork speed, replication initiation and termination, etc.). Nowadays, we have access to newer and efficient genome-wide methods to extract RT from cell populations (Figure 7).

Firstly, Repli-seq is a technique that reveals replication domains by immunoprecipitation of bromodeoxyuridine (BrdU)-labeled DNA. This classical method involves incubating cells with BrdU, a thymidine analogue that is incorporated into replicating DNA, for a certain amount of time ranging between 30 min to 2 h. This is followed by flow-sorting replicating cells into

different fractions of S phase, immunoprecipitation of BrdU-labeled DNA, and genomic analysis of the DNA by microarrays[142] (Repli-chip) and later with NGS[159]. Microarrays were used to quantify DNA in different regions but were phased out for the reasons previously mentioned (Section 1.1.3.2). To enable genome-wide RT, this approach requires cells to be sorted in S phase. Consequently, some studies compare late S phase cells to early S phase cells[160] while others have used between four and sixteen fractions of S phase[161–163]. The latter comes with a higher resolution because the use of multiple fractions of S phase can improve the spatial resolution and indicate loci that replicate asynchronously[164]. Repli-seq had established what was known as the "replication domain" concept, referring to regions with consistent RT appearing as plateaus on replication profiles, it is no longer an employed term. Moreover, the use of BrdU and cell sorting can come with technical restrictions in terms of the size of labelled DNA units and the precision of cell sorting respectively. Whether these limitations are largely reflected in RT measurements was unknown until we were able to obtain RT with higher resolutions.

Secondly, RT obtained via CNV, a technique based on DNA quantification, has been developed and provides a more detailed RT landscape. Counting the number of copies of DNA, based on the number of reads in each genomic region of both G1 and S phase cells, provides a more complete representation of the replicated domains. The number of copies of DNA of each S phase cell divided by the number of copies in the same region of the G1 cells acts as a genome-wide normalised method to correct any deviation from a diploid count (e.g. repetitive regions, polymorphism, etc…). Replicated regions are expected to have twice the amount of DNA in comparison to non-replicated regions [98,156,165]. While this technique provides a higher resolution across the genome compared with Repli-seq, the reliance on cell sorting (S/G1) meant that very early or very later regions would be lost since some cells lie on the G1 and G2 borderlines. More recently, the need for cell sorting has been eliminated thanks to the ability to detect copy-number changes by whole-genome sequencing. Thus, as little as 5-10% of S phase cells are sufficient to generate high-resolution RT profiles[166].

With single-cell whole-genome sequencing upgrading RT studies, more detailed investigations in mammalian replication dynamics at the genome-wide level have emerged[52,98,154,156,167]. The principle of single-cell RT (scRT) is the detection of copy numbers of DNA of every available genomic region in each cell, a method was first reported in 2018 by a genome-wide study of scRT achieved with the use of mouse embryonic stem cells (mESCs)[149]. This study found that the borders between replicated and non-replicated regions were highly conserved between individual cells originating from the same sample. This finding was further confirmed by another publication that followed, conducted by Ichiro Hiratani's lab, which suggested that besides the small degree of visually detectable cell-to-cell heterogeneity, replication organisation is conserved among mammalian cells[98]. Both these studies focused on mid-S phase cells and harvested various computational methods to unravel RT profiles. Moreover, when scRT profiles were averaged for each sample (pseudo-bulk RT), they showed remarkable similarity to the profiles obtain from whole cell populations (bulk RT) indicating that RT of the cell population is globally followed within each individual cell.

Another important approach to study DNA replication is with optical replication mapping (ORM). This technique is not a single-cell approach, but a single-molecule method. Fluorescently tagged nucleotides can be mapped to the genome by treating stretched DNA molecules with a nicking endonuclease. The molecules are then photographed and analysed computationally to be mapped on the genome and analysed[147]. Single-molecule data has allowed an extraordinarily large coverage of the genome, in comparison with single-cell data which is currently unable to achieve such a high coverage. With fibres measuring 300 kb on average, ORM

has been able to confidently identify individual replication initiation sites and has thus demonstrated that, although probability of firing is pre-defined, stochastic origin activation might be a feature of eukaryotic DNA replication.

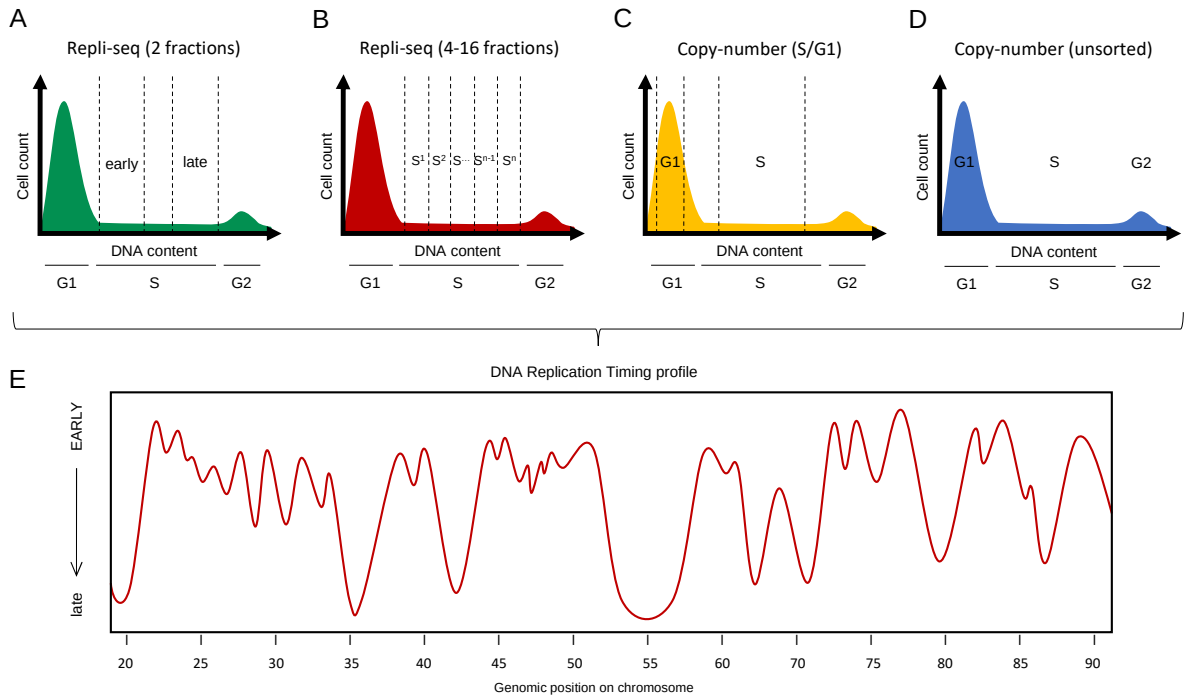


Figure 7: Methods for DNA replication study adapted from Hulke et al. (2020)[166]***.*** *Repli-seq can be carried out in 2 (A) or multiple fractions (B) while copy-number based methods can be carried out with sorted (C) or unsorted (D) cell populations. All methods will result in a similar replication timing profile (E).*

1.5. Interplay between replication timing, structural variants, and cancer

1.5.1. Current understanding

While genomic features such as RT, chromatin states and transcription seem to be linked between each other by some manner, they can all influence SVs, in particular copy-number changes across the genome[168]. Remarkably, one of the pan-cancer analysis of whole genomes consortium publications (mentioned in section 1.1.5.1) showed that out 38 different genomic features, RT had the strongest link with the appearance of SVs[94]. Precisely, tandem duplications along with translocations were mainly found to occur in early-replicating regions while deletions were enriched in late-replicating regions. When comparing individual tumours, they discovered that SVs, when not distributed heterogeneously, were clustered in either early- or late-replicating regions, adding evidence to this particularly strong relationship between SVs and RT. Furthermore, the authors described how different defective DNA repair genes induce different types of SVs (Table 3). Complementary findings from previous studies showed that mutation rates are higher in late-replicating regions, which could have evolutionary implications (e.g. deciding on which genes will evolve at a faster rate)[169,170].

Table 3: Association of different pathogenically mutated DNA repair genes with SVs in cancer; adapted from Li et al. (2020)[94]. *Coloured cells indicate an association between the SV type and muted gene. TD: Tandem Duplications; small: <50-kb; mid: 50-500kb; large: >500kb.*

	Small deletion	Mid deletion	Large deletion	Early small TD	Late small TD	Early mid TD	Late mid TD	Large TD	Fragile site	Fold-back	Unbalanced translocation	Reciprocal translocation
BRCA1	■			■	■						■	
BRCA2	■										■	■
CDK12							■	■				
FANC					■							
PALB2	■											■
TP53	■	■	■	■	■	■	■	■	■	■		

Aneuploidy can be found in most cancers[119,171] making it apparent that it is highly linked to cancer. Whether aneuploidy is the result from errors in the mitotic checkpoint or DNA repair errors can be verified by examining if the copy-number changes impact only segments of or an entire chromosome (arm). Indeed, it would not be expected that DSB alone could cause whole-chromosome duplications. In any case, it is clear that aneuploidy promotes tumorigenesis and that DNA repair genes play an important role in maintaining genome stability. Different mutations on these genes can have different outcomes and cancer is the result of a long process of replication stress and genomic instability as demonstrated by publications from the pan-cancer consortium[91–93]. Altogether, the arguments and known pathways in the previous sections are summarised in an over-simplified representation of the key mechanisms that link together RT and SVs in cancers (Figure 8). The interplay of these features is only a small, but important, part of the molecular orchestration of the cell cycle in cancers.

DNA replication stress can be induced (to kill cancerous cells) and regulated with chemotherapeutic drugs. One of them, oxyplatin, a DNA crosslinker, will arrest the cell cycle in G1 by repressing the expression of proteins involved in DNA replication. Moreover, in view of the strong link between SVs and RT, modifying the order in which the segments of DNA replicate in cancer cells to create catastrophic damage could induce apoptosis, and thus provide a new therapeutic method. However, more research is required to be able to find appropriate pathways for this. Regardless, RT could eventually be used a predictive metric to understand where SVs could arise across the genome, for each cancer type, or even each patient, providing insights in the possible genomic impacts. This might be an important tool that could predict cancer based on individualised mutational landscapes, years before it manifests. Thus, it is important to further conduct research in this field to build the knowledge surrounding genome instability and replication stress.

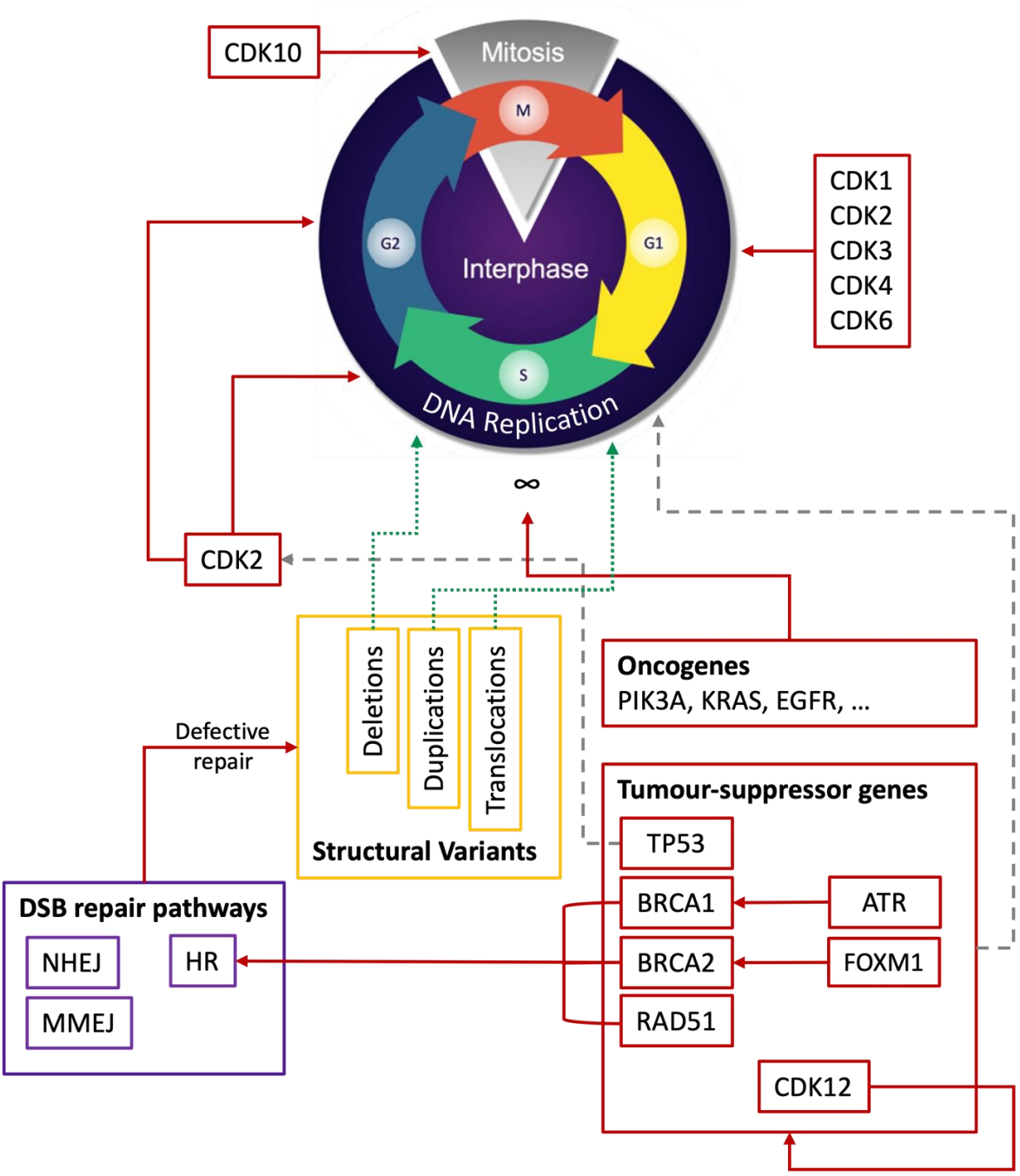


Figure 8: An extremely simplified yet complicated representation of factors of the cell cycle and its relationship to SVs in cancer. *The cell cycle is regulated by CDKs which concentrate at moments of the cell cycle. Oncogenes can lead to infinite proliferation of the cells. CDK12 phosphorylates RNA polymerase II which helps transcribe the tumour-suppressor genes which can indirectly regulate the entry of the cell in S phase if the DNA repair mechanism they regulate is not in place. Under normal conditions, if there is DNA damage, TP53 will block the activity of CDK2 through other intermediate mechanisms, blocking the termination of S phase and thus allowing the cell to enter apoptosis. ATR phosphorylates BRCA1 while FOXM1 upregulates BRCA2*[172]*. BRCA1 and BRCA2 create a complex with RAD51 enabling homologous recombination. If the DNA repair pathway is not executed correctly, or at all, SVs can arise. Duplications along with translocations are more likely to develop in early S phase and deletions in late S phase. Genes are in red boxes, DNA repair pathways in purple boxes and SVs in yellow boxes. Red lines indicate direct impact, pointed lines indicate an indirect impact of one factor on the other while green pointed lines indicate a positive correlation. HR: Homologous recombination; NHEJ: Non-Homologous End Joining; MMEJ: Microhomology-Mediated End Joining.*

1.5.2. Research gaps

Until recently, due to technical reasons, scRT investigations were limited by the number of cells (i.e. ~100) that could be studied in each sample. Some recent studies that have overcome these hurdles[62,154,165] have confirmed that there is some heterogeneity in RT between individual cells within a population. Yet, there has been an absence of unified frameworks for scRT analysis that unite all the individual computational methods starting from the alignment of reads to scRT extraction and further analyses.

We have previously shown that it is possible to distinguish subpopulations from CNAs and extract distinctive replication patterns issued from a single cell line sample (Section 6)[62]. However, this process was not automated and as a consequence, RT is still not routinely studied in tumours. The main reason behind this is because there is a lack of methods adapted to study DNA replication specifically in cancerous cell populations and the scarcity of available WGS data from patient tumours containing enough cells in S phase for scRT extraction. Although it was reported that replication patterns are largely conserved in cancer[162], heterogenous patient-derived replication profiles have still not been studied. Whether cell lines accurately portray the DNA replication timeline remains undetermined. Indeed, more realistic conditions, such as the tumour microenvironment, could play a role on regulating origin firing. Although, due to the lack of previous studies on this question, this remains a mere speculation. Since RT heterogeneity in cancer samples is not studied, new automated methods are required to create such genome-wide investigations.

Here, I attempted to address these issues. Firstly, with the aim of developing a computational tool for automatic scRT extraction and analysis directly from single-cell whole genome sequencing (scWGS) data obtained from asynchronous cells. Secondly, with the aim of developing new methods allowing the investigation of scRT in heterogenous patient-derived tumour samples. The automation of the discovery and extraction of copy-number heterogeneity from scWGS data was used in an attempt to answer the following questions: How can one efficiently discover and extract subpopulations from a single heterogenous sample? How heterogenous are different subpopulations between each other? What is the precise relationship between SVs and RT at the single-cell level?

2. Materials and methods

2.1. WGS data collection and preparation

2.1.1. 10X single-cell data

The BAM files of GM12878[165] that were aligned to hg19 were not available as fastq files (requiring realignment to hg38) and were obtained directly from the SRA website, sorted by read name with samtools[173] sort (v1.16.1; option -n) and converted to fastq files with samtools fastq (options -T CB --barcode-tag CB) to have barcodes transcribed in the fastq headers from the BAM headers. These files were demultiplexed with demultiplex[174] demux (v1.2.2; options -m 0 --format x). The other samples obtained by the 10X scCNV solution[53,154,165] (see Data Availability Section 6.1) were acquired as fastq files with the NCBI SRA toolkit (v3.0.6) under the fasterq-dump command and demultiplexed using demultiplex demux (options -r -e 16) with barcodes extracted directly from the first 16 bp of forward reads. All extracted barcodes were filtered based on the 10X barcode whitelist.

2.1.1.1. Identification of valid barcodes with the EM algorithm

Single cells were considered for further use if they originated from valid barcodes which were identified as follows. Data was prepared by counting the number of lines of each demultiplexed fastq file and then divided by 4 to reflect the number of total reads per single-cell. The resulting list containing the number of reads per barcode was then used to make a distinction between corrupted or low-read (invalid) barcodes from qualitative (valid) ones through a custom R[175] (v4.0.4) script. Barcodes containing less than 30,000 reads were considered to not be qualitative due to the very low number of reads and were systematically removed to eradicate any noise in the initial peak with the goal of only keeping a mixture of two distinguishable distributions (valid and invalid barcodes). The em command from the cutoff R library (v0.1.0) was used to identify the cut-off point of 2 log-normal distributions of the read counts from the Expectation-Maximisation (EM) algorithm for each demultiplexed file.

EM was comprised of two stages, the expectation (E-step) and maximisation (M-step) steps, which occurred after initialisation of the μ and σ parameters (see below) for the 2 log-normal distributions (D1 and D2). The probability density function (PDF) used during the E-step, which represented the probability of observing a particular read count per barcode (continuous random variable) given the following parameters, of the log-normal distribution can be described as:

$$f(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/(2\sigma^2)}$$

where:

- x represents the read count of the barcode.
- μ represents the mean (also called location parameter).
- σ represents the standard deviation (also called the scale parameter).

Using this, γ_i which represents the probability that barcode read count i belongs to the valid distribution is calculated as:

$$\gamma_i = \frac{f(x_i \mid \mu_{D2}, \sigma_{D2})}{f(x_i \mid \mu_{D2}, \sigma_{D2}) + f(x_i \mid \mu_{D1}, \sigma_{D1})}$$

where:

- $f(x_i \mid \mu_{D2}, \sigma_{D2})$ is the PDF of the log-normal distribution with parameters μ_{D2}, σ_{D2} evaluated at the read count x_i for the valid distribution (D2).
- $f(x_i \mid \mu_{D1}, \sigma_{D1})$ is PDF of the log-normal distribution with parameters μ_{D1}, σ_{D1} evaluated at the read count x_i for the invalid distribution (D1).

This E-step computed the expected value of any missing data points and calculated the probabilities of the missing or overlapping data given the estimates of μ and σ. The following M-step then updated the parameters of the log-normal distributions using the estimated probabilities as follows:

$$\mu_{new} = \frac{\sum_{i=1}^{N} \gamma_i \ln x_i}{\sum_{i=1}^{N} \gamma_i}$$

$$\sigma_{new} = \sqrt{\frac{\sum_{i=1}^{N} \gamma_i (\ln x_i - \mu_{new})^2}{\sum_{i=1}^{N} \gamma_i}}$$

The E- and M-steps were repeated iteratively until the estimated probabilities γ_i converged (when the parameters and probabilities stopped changing between iterations). The exact cut-off value between D1 and D2 was obtained with the cutoff command from the same package, having D1, the lower read count distribution, belonging to the Type-I error. Only the barcodes having a number of reads superior or equal to the EM cut-off value were considered to be valid and those with a lower number of reads were discarded. Histograms containing representations of the read counts and the cut-off values were systematically generated for visual inspection and validation. The valid barcodes were retained with their respective reads used to keep fastq files which corresponded to single-cells for further analysis.

2.1.2. Read alignments

MCF-7[62], JEFF[62], HeLa S3[62], hTERT-RPE1[98] and all mouse cells[98] were aligned as previously reported[62] using the Kronos FastqToBam module to the UCSC hg38 and mm10 reference genomes. Other single-cell fastq files had their reads trimmed and filtered by quality score with trim_galore[176] (v0.6.4; options –fastqc, –gzip, --paired when paired-end data were used or omitted otherwise, and --clip_R1 16 except for GM12878 data from BAM files) based on Cutadapt[177] (v3.7) and FastQC[178,179] (v0.11.9) and mapped onto the UCSC hg38 reference genome with BWA mem[180] (v0.7.17-r1188; option -M). Mate coordinates were corrected using samtools fixmate (option -O bam) whenever the data were issued from paired-end sequencing or skipped otherwise. All BAM files were then sorted by coordinates with samtools sort (-O bam) before read duplicates were removed with Picard[181] MarkDuplicates (v2.26.11; options ASSUME_SORT_ORDER=coordinate, METRICS_FILE) via java (v19; options -Xmx16g -jar). MultiQC[182] (v1.10.1) was used to visually inspect single-cell quality.

2.2. RT/Multi-omic comparisons

2.2.1. Functional analysis of theoretical transcriptomic activity

Pseudo-bulk mESC and day-7 RT profiles were obtained in 200 kb non-overlapping bins and evaluated with a two-sided Wilcoxon rank sum test with continuity correction via the wilcox.test R command. Regions with RT switching from early to late, late to early or remaining stable during differentiation were distinguished under the following conditions:

$$Early\ to\ Late = RT_{mESC} > 0.5 \cup RT_{day-7} < 0.5$$

$$Late\ to\ Early = RT_{mESC} < 0.5 \cup RT_{day-7} > 0.5$$

$$Early = RT_{mESC} > 0.5 \cup RT_{day-7} > 0.5$$

$$Late = RT_{mESC} < 0.5 \cup RT_{day-7} < 0.5$$

The list of known mouse genes from the UCSC mm10 database was loaded under the TxDb.Mmusculus.UCSC.mm10.knownGene R package (v3.10.0). The annotateBedFromDb command from the CompGO package (v1.26.0) was used to annotate the regions to provide a list of embodied genes by id. Gene duplicates were removed and gene ontologies were extracted from gene ids with lookUp from the annotate package (v1.68.0). For each gene, the biological processes were isolated from the ontologies and counted. Only the 25 most frequent processes were retained and the ones in common between the two cell types were tested for independence with Pearson's Chi-squared test on R (chisq.test command).

2.2.2. Chromatin accessibility

Pre-analysed GM12878, HeLa and MCF-7 scATAC fragments were obtained from GEO (see data availability Section 6.2) and converted to hg38 from hg19 with a custom R code based on the liftOver package (v1.14.0). Peaks were called in 500 bp tiles using another custom R code based on the ArchR[70] (v1.0.2) package with the random seed set to 20201125 for reproducibility (see code for detailed parameters). Clustering was achieved with LSI dimensionality reduction and Seurat's graph clustering[183] from ArchR. Significant marker genes were determined by multiple-hypothesis testing (binomial, Wilcoxon, two-sided t-testing) when the false discovery rate was lower or equal to 0.01 and the log2 fold-change was higher or equal to 1.25, organised in regions located 100 kb upstream and downstream of transcription start sites. Bulk ATAC of the same cell lines were obtained with the SRA toolkit (v3.0.6) under the prefetch and fastq-dump commands. Fastq files were merged per sample and passed through fastqc (v0.11.9). Alignment to the hg38 reference genome was made with Kronos fastqtoBAM and reads with a mapping score under 30 were removed. Peak calling was performed with the macs2[184] (v2.2.7.1) filterdup, predict and callpeak commands.

2.3. Copy-number matrix organisation

Copy-numbers from the resulting single-cell BAM files were estimated with the Kronos scRT Binning and CNV commands in either 20 or 25 kb windows (see code for parameters). Systematically problematic genomic regions were masked with the hg38 blacklist. The resulting BED files were regrouped by sample and used as an input for MnM. Marie Curie's date of birth in a YYYYMMDD format was used as a random seed when running MnM.

Genomic regions from all MnM input files were rearranged in 100 kb non-overlapping genomic windows (as a median of the copy-numbers from the input file each 100kb window included) delimited by the chromosome sizes of the hg38 reference genome provided by bedtools[185,186], and then in 25 kb and 500 kb for the replication state classifier models. Median copy-numbers

per bin were calculated when at least 50% of the bin was covered. MnM then automatically processed the data by temporarily removing windows containing no data and any remaining sporadic missing values were filled in with the integrated sklearn k-Nearest Neighbors (KNN) imputation algorithm[187] (options n_neighbors=5, weights='distance'). The nearest neighbours were defined as the 5 closest cells based on the Euclidean distance of the genome-wide copy-numbers (distances calculated in pairs for genomic regions that neither of the 2 cells were missing). A weighted average of copy-numbers from the region of the closest neighbours was used as the imputation value. The imputation method can be described as:

$$\hat{X}_{ij} = \frac{\sum_{k=1}^{n_{neighbours}} w_{ik} \cdot X_{kj}}{\sum_{k=1}^{n_{neighbours}} w_{ik}}$$

where:

- $\hat{X}_{ij}$ represents the imputed value for the copy-number of the region j in cell i.
- X_{kj} denotes the value of region j in the k-th neighbour.
- $n_{neighbours}$ is the number of nearest neighbours considered for imputation. Here n =5.
- w_{ik} represents the weight assigned to the k-th neighbour for cell i based on their Euclidean distance.

This imputation method was also used for the imputation of 5-55% in intervals of 5% of single-cell copy-number values that were randomly selected and removed after the elimination of any windows containing missing values of an S-phase enriched population of MCF-7 cells. A random imputation method where each missing copy-number value was substituted by a randomly selected non-missing value from the matrix, along with a median imputation method where the median of each genomic region was imputed, were implemented for comparison to the KNN imputation method under the same random seed (see code for details). Accuracy was calculated as the percentage of identity of the imputed values compared to the original values. Similarity was calculated as the percentage of values that differed less than ±1 copy number for KNN imputation compared to the original value. Invariance was calculated matrix-wide as the percentage of unchanged copy-numbers after imputation.

2.4. Replication state detection

To organise the data for the replication state classifier, cells phases were either extracted with Kronos (HeLa, MCF-7, JEFF)[62], solely from the FACS metadata (hTERT-RPE1)[98] or from the intersection of common replicating states from the FACS metadata and Kronos (HCT-116, GM12878)[165]. The resulting single-cell copy-number matrices were concatenated. Any partially or completely missing regions (i.e. any genomic region containing at least one missing copy-number value) were removed while only autosomal data were retained. 80% of the cells were used as training data and the remaining 20% were used as testing data. To prepare the replication state classifier to be able to distinguish noisy copy-number non-replicating profiles (e.g. from low-quality cells or technical noise) from replicating cells data, augmentation was performed. Half of the training cells were randomly selected and copied. For each of these copied cells, noise was induced by altering the copy-numbers by ±1 between 5-75% of the genomic regions which were selected from a uniform distribution.

The replication state classifier was built on a Sequential architecture which is a feed-forward neural network. The model was designed with the Keras[188] python library (v2.13.1) to facilitate the construction of a linear stack of neural network layers, each connected to the subsequent

one. As an input, the single-cell copy-number matrix of the training dataset containing the 6 cell types and the augmentation data was used. The sequence of layers aimed at hierarchical feature extraction and predictive modelling consisting of three hidden layers with 64, 32 and 16 units, respectively. These layers facilitated the extraction of increasingly complex and abstract representations of the input copy-number profiles. The model terminated in an output node having a single unit with a sigmoid activation. This configuration was suited for binary classification tasks, enabling the model to produce a probability estimation in a [0,1] range. Upon construction, the model was compiled with a binary cross-entropy loss function to optimise the network's performance concerning binary classification. An 'adam' optimiser, known to be efficient and adaptive on learning rates, was used to optimise the parameters throughout training. In order to avoid overfitting, an early stopping mechanism was implemented on an epoch-based patience of 15 iterations.

With the completion of training, the resulting neural network model along with the list of genomic windows comprised in the matrix were saved for further use. The model was then integrated and automatically loaded with MnM to predict the single-cell binary replication states (Replicating/S-Phase, Non-Replicating) of the scWGS data obtained from tumours or cell lines. In the case where any regions required by the model were not present, MnM compensated for these missing values by using linear interpolation from both directions. Compensating for these missing values ensured the continuity and integrity of replication state predictions.

2.5. Subpopulation discovery

Commencing with non-replicating cells, the number of variables were reduced from the number of autosomal regions to 2 dimensions with Uniform Manifold Approximation and Projection (UMAP)[189]. The Density-based spatial clustering of applications with noise (DBSCAN) algorithm[190,191] was then used to detect the number of groups on this reduced dataset (option min_samples= 10%). The epsilon parameter (ε) was calculated as:

$$\epsilon = \frac{\max(UMAP1) - \min(UMAP1)}{\max(UMAP2) - \min(UMAP1)} \times 1.25$$

where UMAP1 and UMAP2 correspond to UMAP's first and second output parameters respectively. Epsilon was always restricted between 1.25 and 2 while min_samples had a minimal requirement of at least 10 cells. UMAP was repeated with 6 randomly generated seeds and the most frequent number of subpopulations, as determined with DBSCAN, was retained. Subpopulations discovered with DBSCAN were redefined and merged iteratively in a descending similarity order if the median copy numbers per region were 98.5% identical. Copy-numbers of both S-phase and non-replicating were reduced to 10 UMAP dimensions (second round of UMAP) and then matched each S-phase cell to the closest non-replicating group with the sklearn nearest neighbour command (options n_neighbors=50% of cells, metric='euclidean'). The number of nearest neighbours was required to have a minimal value of 5. Both rounds of UMAP were performed on the single-cell copy-number matrices with the addition of 5 artificial cells stretching from complete haploid to pentaploid profiles for subpopulation calibration.

2.6. DNA replication timing

Kronos scRT was modified to work with R v4.0.5, ignore copy-number confidence during quality-control filtering and produce an extra metadata file containing cell diagnostic details. We used the diagnostic module at a first stage for quality control based on the number of reads per Mb under the customised developer mode (option -d) created for this purpose. The data

were filtered and then passed through MnM for replication state classification and subpopulation detection. The Kronos scRT WhoIsWho module was used to assign the cell phases from the replication state classifier or FACS data accordingly (see code for details) followed by the diagnostic module, which was used a second time to correct the early and late S-phase copy-numbers (option -C). For each subpopulation and biological replicate, the copy-number data were split into a different file with a custom python (v3.9.11) code. Kronos scRT was then used to calculate the replication timing profiles through the RT module in 200 kb windows. The resulting scRT binary values were used to produce scRT trajectories with the DRed module using the random seed '18671107' for reproducibility. Permutation tests on the trajectories were made using a custom python code under 1,000 permutations. The observed test statistic was calculated as the absolute mean of sum of differences in means between subpopulations for both UMAP coordinates:

$$Observed\ statistic = \sum_{groups} \left| \bar{x}_{group} - \bar{y}_{group} \right|$$

Where $\bar{x}_{group}$ and $\bar{y}_{group}$ are the means of the UMAP1 and UMAP2 coordinates for each subpopulation respectively. The permutation test was executed by randomly shuffling the group labels while keeping the same UMAP coordinates. The test statistic was computed on the shuffled data in the same way as the observed statistic. This allowed to calculate a p-value as follows:

$$Permutation\ test\ p-value = \frac{\sum_{permuted}(permuted\ statistic\ \geq observed\ statistic) + 1}{number\ of\ permutations + 1}$$

Pseudo-bulk and bulk RT correlations were calculated with the Spearman method and scRT correlation clustering for the scRT atlas was ordered with the Ward.D2 hierarchical clustering method. Bulk RT profiles were lifted over from hg19 to hg38 with the liftover command[192] after being converted to bed files with bigwigtobedgraph[193]. When applicable, the Kronos scRT compare TW module was used to compare the Twidth values of mouse scRT under 1,000 iterations.

3. Results

3.1. Single-cell replication timing in mammalian cells

3.1.1. Kronos scRT: a computational tool for scRT studies

Kronos scRT is a computational tool designed to extract and analyse single-cell replication timing (scRT) from single-cell whole-genome sequencing data (Figure 9). Developed in R[175], it acts as a unified framework to offer a comprehensive approach to analyse scRT from the post-sequencing to downstream analysis steps. Kronos scRT aligns the sequenced reads to the selected reference genome (human hg38 and mouse mm10 used here) with the FastqToBam module. The number of reads of each single-cell are then counted in non-overlapping bins which are 20 kb wide by default (modifiable parameter based on the coverage of the dataset) obtained from the Binning module. A human or mouse blacklist were systematically used in the analyses here to remove any problematic genomic regions[194]. Read counts were then corrected for GC content and mappability bias and were translated to raw copy-numbers. Late S phase cells systematically appeared to have a median raw copy-number inferior to that of G1 cells (see ref. [62]). The copy-numbers were corrected, and the cell-phases were distinguished based on bin-to-bin copy-number variability and median ploidy (median copy-number) with the diagnostic module. This method of *in silico* cell phase sorting could not distinguish G2 cells from G1 cells because, in principle, both have similar genome-wide copy-number profiles (copy-numbers are estimated in possible ploidy slots e.g. 1, 2, 3, etc..). We therefore refer to these cells as "G1" or "non-replicating" cells interchangeably hereafter. Finally, scRT was calculated with the RT module by dividing the copy-numbers of the bins of the S phase cells by the median copy-numbers of the respective G1 cells.

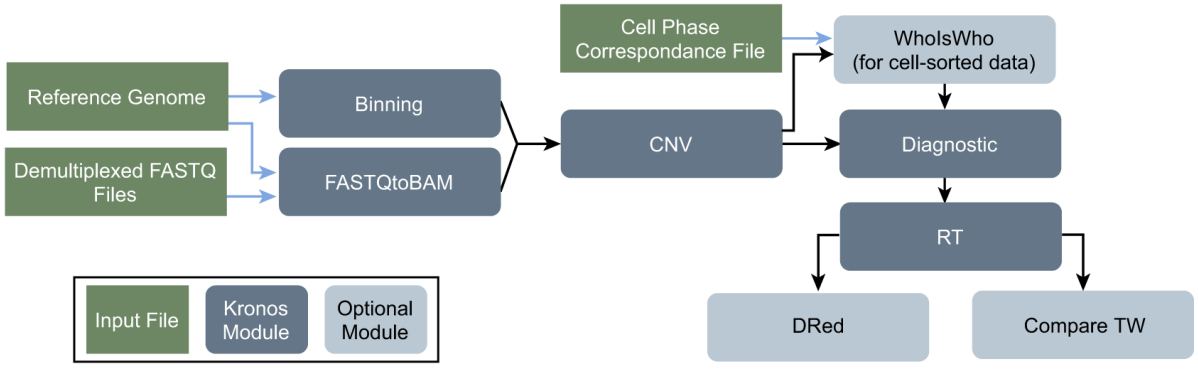


Figure 9: Kronos scRT, a unified pipeline for scRT analysis. *The obtained reads are aligned to the reference genome of choice which are then counted in fixed-sized genomic windows and translated to DNA copy-numbers. The copy-numbers are then used for an attempt to detect cell phases or alternatively imported from cell-sorting metadata. RT is then calculated from each individual cell by dividing the copy number of each genomic window of S phase cells by the median copy-number of the G1 cells for the corresponding window. Further analyses can include dimensionality reduction and Twidth comparison. Figure obtained from the Kronos scRT publication[62]. CNV: Copy-Number Variation; RT: Replication Timing; Dred: dimensionality reduction; TW: Twidth.*

We applied Kronos scRT to human droplet-based single-cell WGS data generated for this study[62] as well as published mouse data[98]. Specifically, the human data used were estrogen receptor-positive breast cancer MCF-7, cervical cancer HeLa (S3), and B-lymphoblastoid JEFF cell-lines. The mouse data used were mid-S phase mouse embryonic stem cells (mESCs) and

mouse neurectoderm cells at day 7 of differentiation (day-7). MCF-7 cells were sequenced under two repeats, the first being unsorted cells and the second being FACS-sorted cells to enrich the S-phase population. The cells were sorted because under normal circumstances, one would expect to only find ~20% of a population of cycling cells in S-phase. Kronos scRT was able to distinguish S-phase cells from non-replicating cells automatically in the unsorted population, and manually from the sorted population. This was because the non-replicating population was used as a target to be able to then detect the replicating cells. If the G1 cell population was not predominant, Kronos was not always able to detect the replicating states of the cells.

We obtained scRT profiles for each of the cell-types analysed and proceeded with analysis as described in the pipeline. A striking discovery was the in the analysis of MCF-7 cells, copy-number heterogeneity between cells was discovered. Upon further investigation, the discovery of two different aneuploid karyotypes was noticed. Although, MCF-7 cells are known to almost be tetraploid (ploidy ~3.8 here), we discovered that different genomic regions exhibited different copy-numbers which were both chromosome-wide and on the sub-chromosomal scale. Notably, chromosome 3 was present in either 4 or 5 copies depending on the subpopulation. In order to verify these *in silico* findings, a FISH experiment was carried out and indeed confirmed the presence of 4 or 5 copies of chromosome 3. Thus, the MCF-7 cells were split by subpopulation and analysed separately. When comparing the pseudo-bulk RT profiles of the two subpopulations, a 94.6% Spearman correlation was observed, highlighting that despite large SVs, the replication machinery is robust enough to reduce the effect on the replication patters. Nonetheless, differences were visible enough on the scRT trajectories to notice that the scRT patterns were divergent between the two subpopulations but, in comparison to the other human cell-types, not enough to characterise them as two separate cell-lines.

Additionally, we confirmed that variability in RT was higher in mid S phase compared to early and late S phase in all the cell types analysed. This observation was based on the T-width measurement, which will be explained later in Section 3.1.2, a measurement only made accessible with the arrival of single-cell technologies in RT studies. This remark suggested that the replication process is more asynchronous during this stage of S phase because different genomic regions do not replicate at the same time during mid S phase leading to this higher degree of variability. To further address cell-to-cell variability, regions were classified into late S (<30% genome replicated), mid S and late S (>70% of the genome replicated) based on the pseudo-bulk data. In all examined cell-lines, between 1-5% of the cells had late replicating regions already replicated in early S phase, which supported a certain degree of stochasticity in the RT programme.

Overall, this study showed that Kronos scRT is a scalable and comprehensive tool allowing the study of RT at the single-cell resolution in homogenous or heterogeneity-resolved populations. The detailed methods and results can be found in the publication (Section 6), Kronos scRT is available at https://github.com/CL-CHEN-Lab/Kronos_scRT and it is protected by the French Agency for the Protection of Programs (APP) under registration number IDDN.FR.001.370044.000.S.C.2022.000.20700. Personal input for the development of Kronos scRT and published scRT analyses[62] included:

i. Extension of scRT to non-human reference genomes (e.g. mm10).
ii. Inclusion of a blacklist, a list of problematic genomic regions that will be excluded from the analysis.
iii. Optional use of FACS metadata to define cell phases (WhoIsWho module).
iv. Subpopulation detection of MCF-7 cells.

v. Calculation of binary cell-to-cell RT distances with the simple matching coefficient (adapted distance metric for binary data).
vi. Dimensionality reduction and scRT trajectory generation (DRed module).
vii. Day-7 and mESC scRT extraction and bioinformatic analyses.
viii. Step-by-step case study tutorial on GitHub.

Although Kronos scRT enabled automatic scRT extraction and analysis from raw sequenced reads, it failed to address cell-to-cell copy-number heterogeneity to detect copy-number signatures resulting from CNAs. Furthermore, on smaller datasets (e.g. ref.[98,167]), Kronos scRT failed to detect replication states of the single-cells, or required manual cut-offs with larger ones that contained a disproportionate number of cells in S phase (e.g. ref.[165]). The following approaches (Sections 3.3, 3.4) were used to address this issue by creating new methods that would allow investigations in single-cell genomic heterogeneity.

3.1.2. Mammalian replication patterns differ between cell type

To compare RT of mouse samples, we had previously used previously published data[98] from mid-S phase mESCs as well as day-7 (n=46)[62] cells. Here, the data were enlarged by incorporating early and late S-phase mESCs (n=146) from the same source along with mouse embryonal carcinoma cells (Figure 10 a; n=25). The scRT profiles were extracted from the data, reconstructing the replication landscapes for each cell type (Figure 10c). Although the number of cells per sample was limited, there was a relatively high Spearman correlation with bulk RT profiles when compared in 200 kb genomic windows, amounting to 89%, 87.9% and 81.7% for mESCs, day-7 and carcinoma cells, respectively. These results indicate that even when there are a limited number of cells, the replication patterns can still be extracted successfully. As expected, the overall replication timing profiles of mouse cells differed between the 3 cell types. Specifically, there was a weaker relationship concerning the carcinoma cells when compared to mESCs and day-7 cells, correlating at 79.5% and 79.4%, respectively (Figure 10 b). On the other hand, mESCs and day-7 cells correlated higher between each other at 84.1%.

To quantify the variability of each part of S-phase, Twidth – the number of hours required for a genomic region to be replicated in 75% of the cells of the population, starting from 25% of the cells having that region replicated – was calculated (Figure 10 d) to reflect variability of the different parts of S phase. In accordance with previous studies[62,156,195], the most variable timing point was mid-S phase for both mESCs and day-7 cells. Contrariwise, the carcinoma cells were most variable in late-S phase, implying that there is a temporal deregulation of the replication timing programme in these cancerous cells. We found that there was a significant difference between the Twidths of the 3 samples which displayed bootstrap p-values of less than 1e-03 (Figure 10 d, Supplementary Section 8.3.1), indicating that the RT profiles can be ordered by variability in the following order: mESC, day-7, carcinoma cell-types.

3.1. Exploring the multi-omic landscape

3.1.1. Functional genomics from RT

Bearing in mind the multi-omic landscape RT is a part of, one would expect early-replicating regions of the genome to be transcriptionally active and late replicating regions inactive. Based on this hypothesis, a functional analysis was performed to discover gene functions located in regions that switch from early to late, or late to early during embryonic differentiation. The RT profiles between mESC and day-7 cells were proven to be significantly different by a Wilcoxon test which yielded a p-value less than 2.2e-16. 1,535 200 kb regions switched from early to

late, 548 from late to early, 5,438 remained early and 4,444 remained late during differentiation.

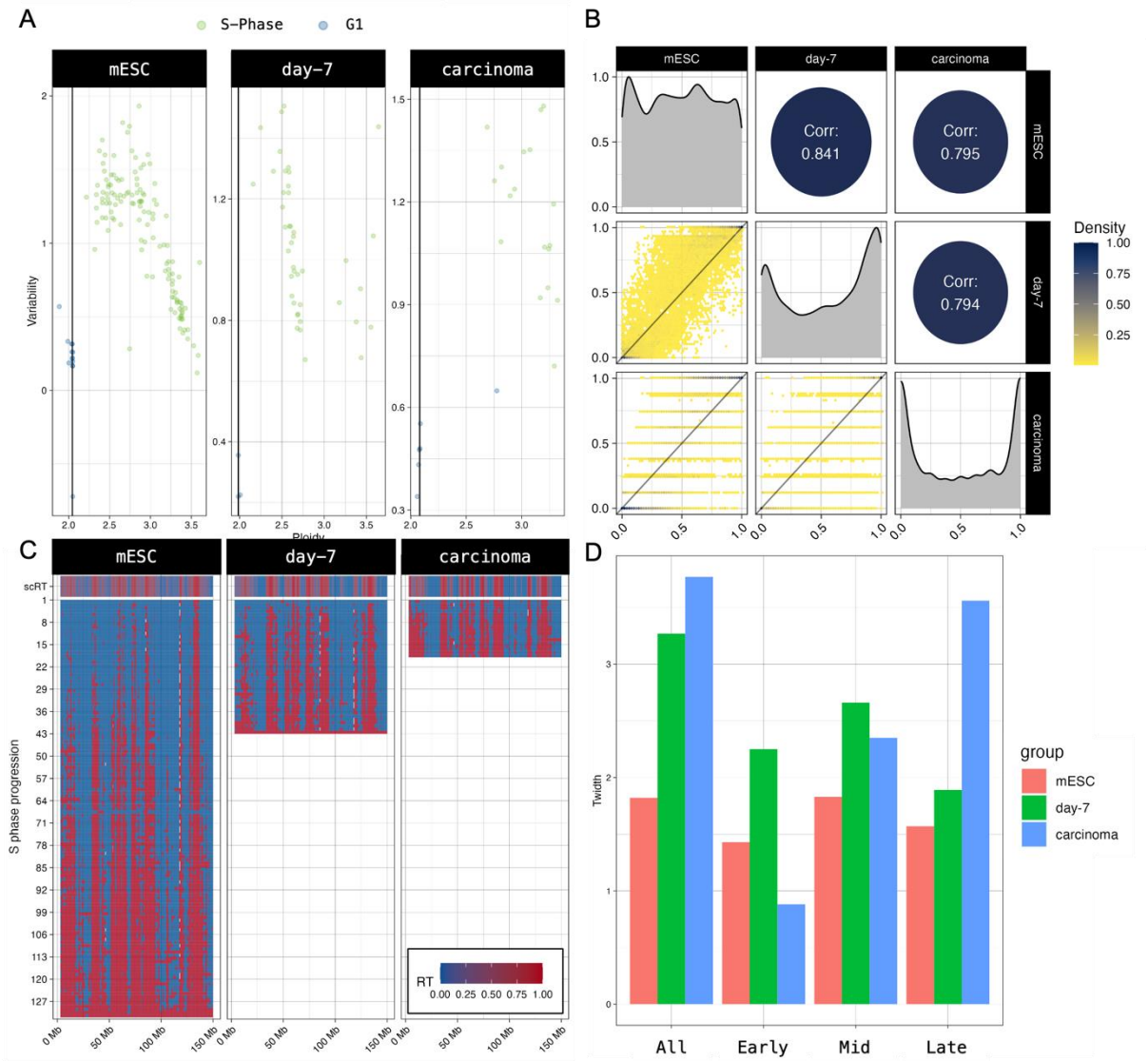


Figure 10: RT differs between mouse cell types. *A. G1 and S-phase ploidies and bin-to-bin variability of mESC, day-7 and embryonal carcinoma cells. B. Genome-wide Spearman correlations between mouse cell samples. C. Replication landscapes of the mouse cell types with single-cells sorted from early to late S-phase. D. Twidth values for early, mid and late S-phase indicating replication stress in mouse carcinoma cells, leading to higher cell-to-cell variability in late S-phase regions.*

The biological processes of genes present in the UCSC mm10 database that were located in regions that switched from late to early (activated during differentiation) were extracted and counted. A total of 15,805 biological processes were uncovered from these genes, of which 4,459 were unique occurrences. The highest 25 occurrences were extracted and corresponded to both general processes along with development- and differentiation-specific ones (Figure 11 a). Likewise, for early to late regions (deactivated during differentiation), 6,528 processes, of which 2,475 unique, were uncovered. Contrary to the late-to-early regions, development- and differentiation-related gene functions were less present (Figure 11 b). After grouping the 17 highest processes in common between the 2 temporal changes, a Chi-squared test was performed to evaluate whether they are independent. The p-value obtained was found to be less

than 2.2e-16 indicating that the distribution of functional ontology term counts was not the same between mESCs and day-7 cells.

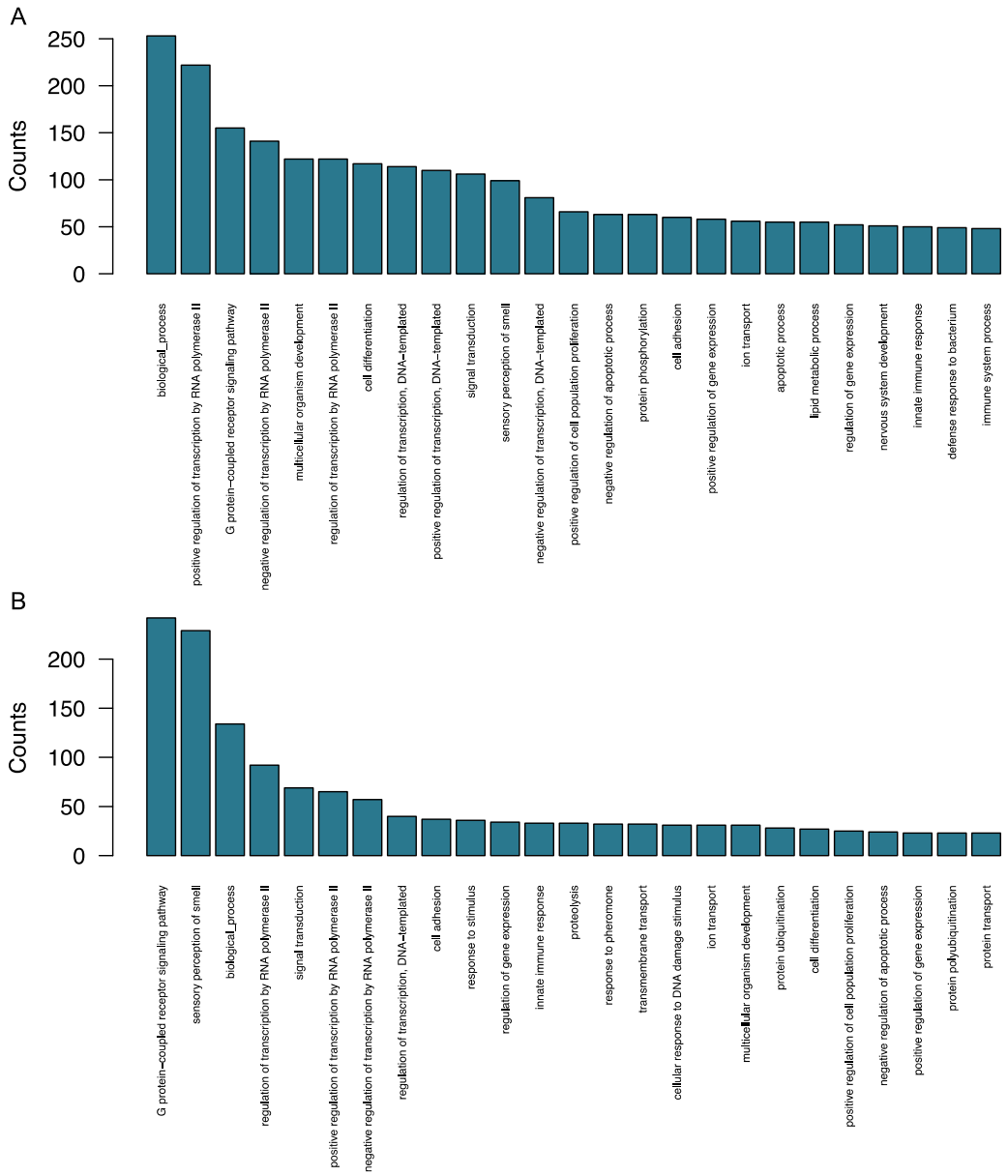


Figure 11: Theoretical transcriptomic activity based on RT changes. *A-B. The 25 highest occurring biological functions of genes embodied in regions switching from late to early, theoretically activated, (A) and from early to late, theoretically deactivated, (B) during differentiation.*

3.1.2. Chromatin accessibility and RT

While studies have shown that RT correlates well with transcription, they have also shown that there is an association with chromatin. In order to examine the extent of this association, bulk and single-cell ATAC (scATAC) data was obtained for HeLa (cervical cancer; n=2,610), MCF-7 (breast cancer; n=875) and GM12878 (lymphoblastoid; n=2,663) cell lines. The scATAC peaks were called with a revealed 3 individual clusters (Figure 12 A-B) corresponding to the 3 human cell types previously used for RT[62]. Bulk and single-cell ATAC peak counts and RT in 200 kb regions from these cell types were compared, with GM12878 acting as a similar lymphocyte cell line for JEFF cell RT comparisons used here.

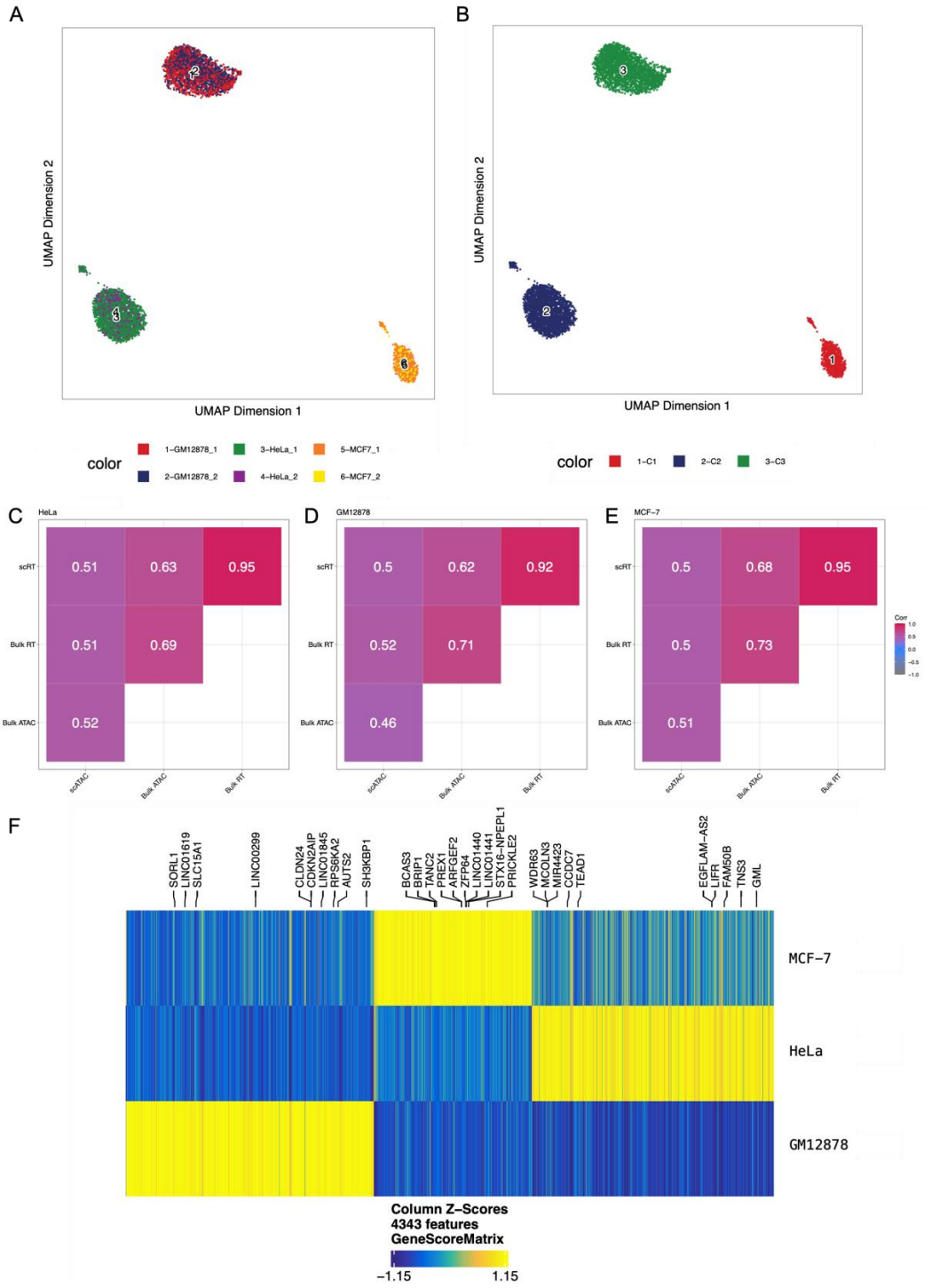


Figure 12: ATAC data reveals moderate links with RT. *A-B. scATAC peaks per replicate (A) and clustered to reflect the 3 cell-lines (B). C-E. Single-cell and bulk RT and ATAC comparisons only show moderate Spearman correlations between the two omic features for Hela (C), JEFF/GM12878 (D) and MCF-7 (E) samples. F. The 10 highest gene markers per cell-type based on multiple-hypothesis testing.*

scATAC peak counts only showed a moderate correlation with both bulk and pseudo-bulk RT (Figure 12 C-E). The 10 most significant marker genes per cell type were displayed (Figure 12 F) revealing a distinct gene signature in GM12878 cells, associated with the CDKN2A Interacting Protein (CDKN2AIP) gene which is central the cell cycle control, senescence, and DNA damage response through various signalling pathways, including the p53-HDM2-p21 (WAF1) pathway, consistent with inefficient mitotic progression and thus, the development of SVs [196].

This open-chromatin region suggests that this gene is transcribed and thus, in agreement with a functional DNA repair mechanism preventing the generation of CNAs as observed on both GM12878 and JEFF copy-number profiles (Supplementary Section 8.3.3, Figure 16).

3.2. Automatic single-cell data preparation for copy-number analysis

3.2.1. Single-cell barcode disentanglement

In single-cell technologies, distinguishing the reads by the individual cells they originated from is essential in order to be able to reconstruct cellular genomes and compare cells between each other. The use of unique molecular sequences that are attributed to each cell in order to do so are called barcodes. During the preparative pre-sequencing steps, each single cell has an individual barcode attributed to all its reads. Nonetheless, due to sequencing errors, some barcodes are not sequenced perfectly, or some cells do not have enough sequenced reads and should be removed as part of quality control. Although some methods do exist to identify valid barcodes, they are either owned by private companies (e.g. Cell Ranger from 10X Genomics), are applied to certain technologies, could be discontinued in the future, performed after unnecessary mapping which could lead to longer processing times[197] (e.g. Cell Ranger from 10X Genomics), manually curated by deciding on cut-offs (e.g. ref.[154]) or are made for other omic data and are not adapted to WGS data (e.g. ref.[198]). Furthermore, with carbon footprints becoming an aspect to consider in bioinformatics, it would be preferred to avoid unnecessary mapping of reads that will not be used in downstream analyses.

To overcome these issues, a machine learning method was implemented on GM12878 cells (detailed cell counts in Supplementary Table 3, Section 8.2.3, see Methods Section 2.1.1.1 for details). After removing the 13,404 barcodes that had a low read count (i.e. contained less than 30,000 reads), the remaining 7,163 barcodes were then clustered into valid and invalid groups. This removal allowed a clearer visual identification of the two distributions (i.e. valid and invalid) and avoided any noise in the identification of the cut-off value between valid and invalid barcodes. The Expectation Maximisation (EM) algorithm[199] was used to find the cut-off value between two log-normal distributions. The first distribution with a lower read count was considered to contain the invalid barcodes and the second one the valid ones (Figure 13). In total, 92.02% (n=6,591) GM12878 barcodes were retained (Section 8.3.4). Thus, a new method to extract single-cells out of a large number of invalid barcodes, that does not require mapping reads beforehand was developed and applied on all data originating from the 10X scCNV solution (see code in Section 6 for details).

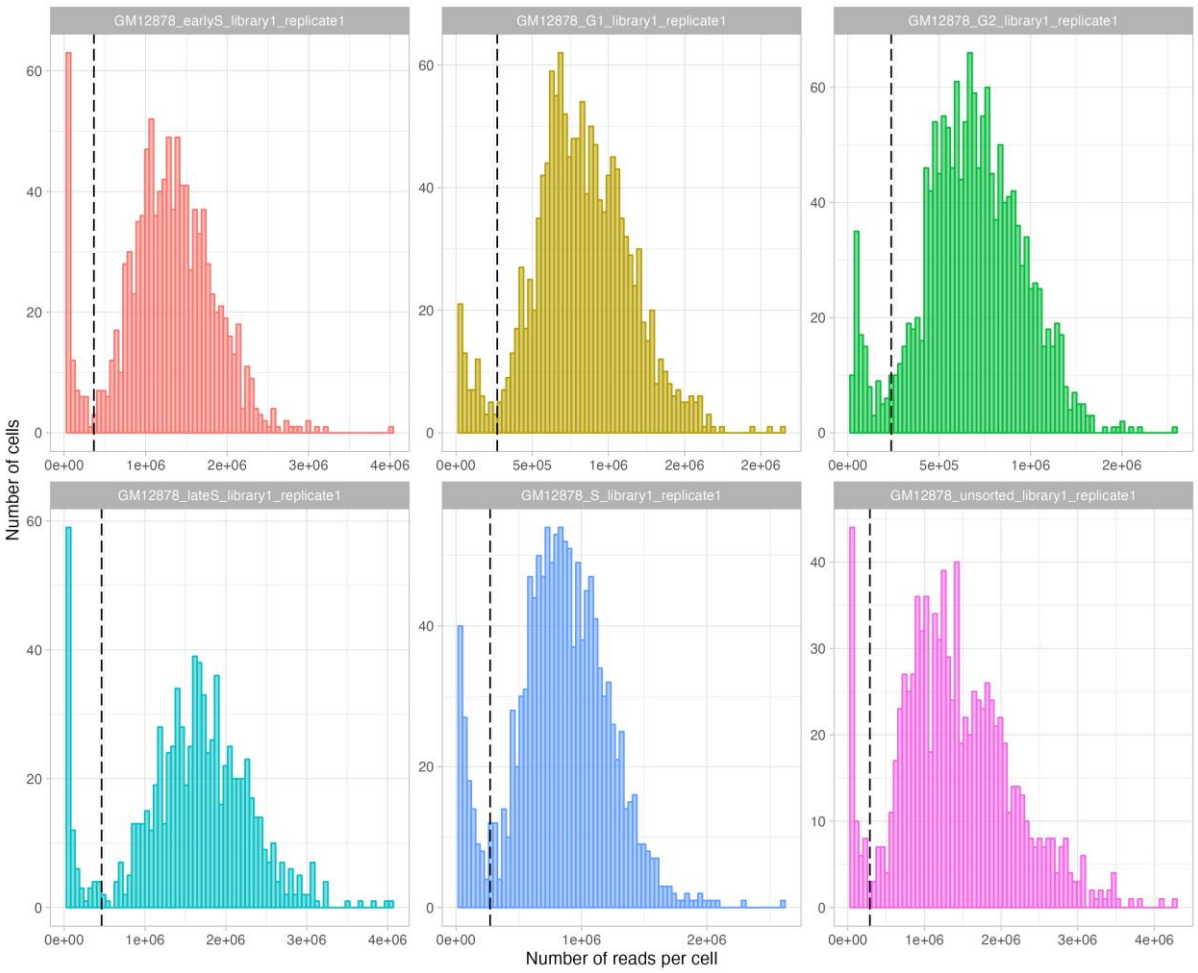


Figure 13: Separation of valid and invalid barcodes through read distributions. *Cut-off values appear as dashed lines on the number of reads per cell; splitting valid (right) and invalid (left) barcodes for GM12878 sorted (early S, G1, G2, late S, S) and unsorted cells.*

3.2.2. Highly accurate data completeness in single-cell analyses

Due to technical limitations, single-cell whole-genome sequencing data frequently has a lower read coverage across the genome (e.g. <1X) in comparison to bulk experiments, leading to sporadic missing copy-number values. To address this issue, we turned to the KNN imputation technique – a data completion method that takes into consideration the closest cells in terms of genome-wide copy-number profiles. To take into account rare copy-number aberrations, or replication events, we used a weighted genome-wide copy-number distance between single cells for the KNN imputation which generated an imputed value proportional to the closeness of the copy-number profiles between the cells based on their Euclidean distances. For each missing value, the existing copy-number for that same region from the closest 5 single cell profiles were used to fill in the missing data (See Methods Section 2.3 for details).

We proceeded to empirically validate this method by introducing random voids within the 100 kb bin single-cell matrix of MCF-7 cells (n=2,321; 1,288 genomic regions), a cell-line with a large number of CNAs [62], after removing any regions that already contained missing values. We removed between 5% to 55% random values in increments of 5%. We observed that KNN imputation predicted and integrated these missing values with an average accuracy of 83.959%, thereby reconstructing the single-cell copy-number landscape. Remarkably, our findings showed an invariance rate, defined as the total percentage of intact values of the whole matrix,

ranging between 99.209% and 90.911% for 5% to 55% of missing values respectively (Figure 14A), underscoring the robustness of the KNN approach in this context. Furthermore, the imputed values with an absolute difference no larger than 1, in comparison with the original values, ranged between 99.917% and 99.015%, illustrating that the vast majority of the errors introduced through this process were not radically inaccurate, even when more than half of the dataset contained missing values.

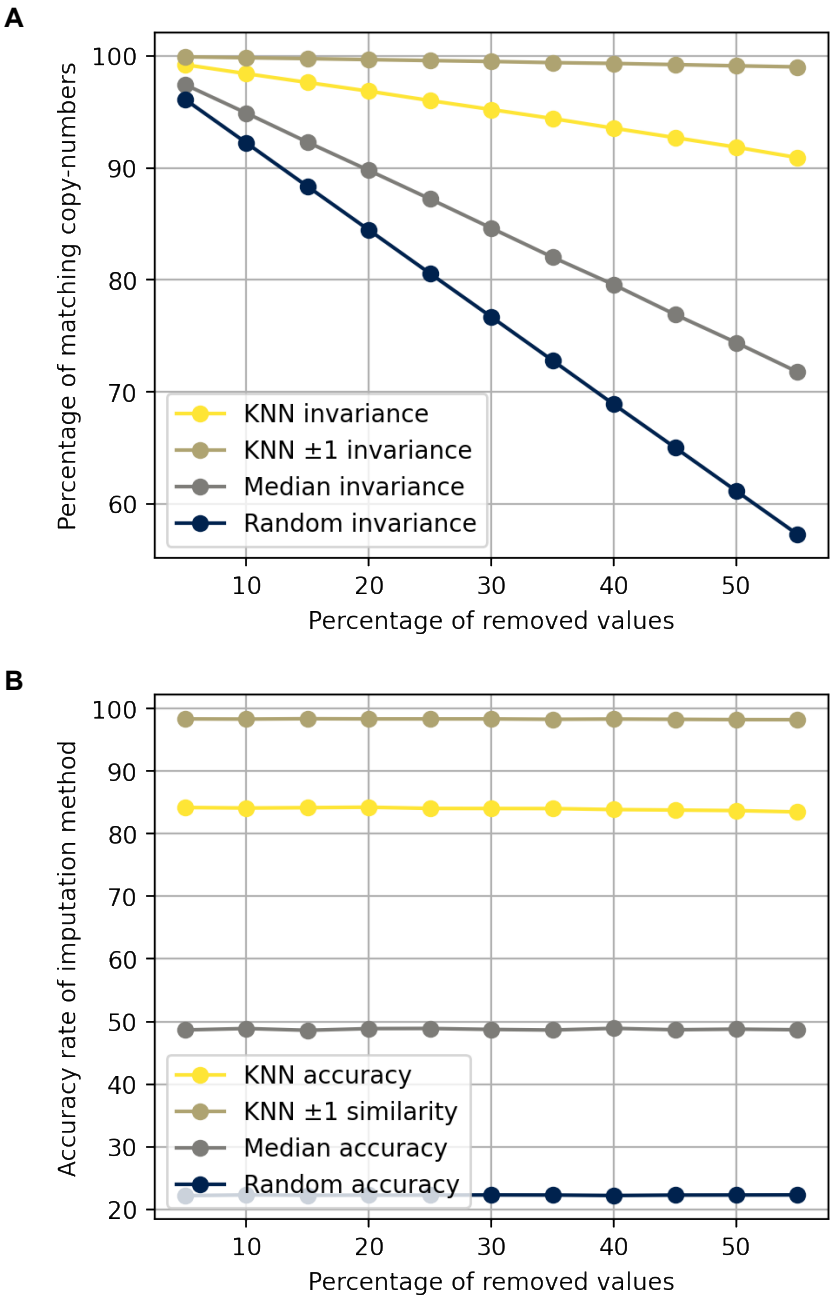


Figure 14: KNN imputation is an efficient method compensating for single-cell copy-number scarcity. *Missing values were simulated by using MCF-7 copy-numbers in 100 kb windows which underwent random value removal ranging from 5 to 55% of the total number of values in the single-cell copy-number matrix (regions/cells). KNN, median and random imputations were performed while KNN imputed values that varied by ±1 copy-number were calculated. These values were compared to the original values for copy-number matrix-wide invariance (A) and accuracy of imputed values (B).*

As a comparison, we used median and random imputation methods (see methods for details) on the same missing values as those used for the KNN imputation. These techniques provided accuracy rates that were lower and significantly different to the KNN method (paired t-test p-values of 5.786e-23 and 1.965e-25) which averaged at 48.758% and 22.269% respectively (Figure 14B). The invariance rates of the matrices were also lower for these methods compared to the KNN imputation as they ranged between 97.433% and 71.776% for median imputation and 96.11% and 57.27% for random imputation. To extract the most information possible from our data, we applied the KNN imputation method to all datasets in this study (Supplementary Section 8.2.3). Imputed values accounted for merely 0.84%, 0.68% and 1.11% missing values for HeLa, JEFF and MCF-7 cells, respectively, suggesting that the imputed value fidelity for these samples was high (>99%) based on the simulations (Supplementary Section 8.2.1).

We observed that KNN imputation accurately predicted and integrated these missing values, thereby reconstructing the single-cell copy-number landscape. Remarkably, our findings showed an accuracy rate range of 99.185 to 90.285% for 5% and 55% of missing values, respectively (Figure 14) underscoring the robustness of the KNN approach in this context. Furthermore, the imputed values with an absolute difference no larger than 1, in comparison with the original values, ranged from 99.915 to 99.034%, illustrating that the vast majority of the few errors introduced through this process were not drastically inaccurate, even when more than half of the dataset contained missing values. To extract the most information possible from out data, we applied this imputation method to all datasets in this study (Supplementary Table 1). Imputed values accounted for merely 0.84%, 0.68% and 1.11% missing values for HeLa, JEFF and MCF-7 cells, respectively, implying that the imputed value fidelity for these samples was high.

3.3. Unravelling cell-to-cell copy-number heterogeneity *in silico*

3.3.1. Deep learning single-cell replicating state classifier is superior to FACS sorting

In order to extract scRT profiles, it is essential to know which cells are replicating. One popular method to distinguish the replication state of cells is fluorescence-activated cell sorting (FACS). Considering that not all single-cell data undergo FACS sorting and that sorting cells, by replication state for example, can induce errors[200], it would therefore be interesting to further validate replication states by other means. Current computational tools that do so[62,201] require manually established thresholds or complementary information such as GC content and intra-cellular variability measurements, which are not always directly accessible. To create a method that can bypass any need for metadata, we amalgamated single-cell copy-numbers issued from datasets with replication states inferred from either FACS[98] (hTERT-RPE1: retinal pigment epithelial), Kronos scRT[62] (MCF-7, JEFF, HeLa), or the inter-section of both[154,165] (HCT-116: colon cancer; GM12878: lymphoblastoid) depending on appropriate extraction methods and harvested the labelled replicating states to create a deep learning model based solely on single-cell DNA copy-numbers (See Methods Section 2.4 for details). A total of 5,250 replicating and 2,273 non-replating cells spanning amongst these 6 cell-lines resulted from the amalgamation (Supplementary Table 2). We hypothesised that the diverse ploidy landscapes of the selected cell lines (Supplementary Section 8.3.3) would make this prediction tool universal and adapted to any ploidy state.

We split our dataset in an 80:20 proportion to create training and test datasets. The training dataset was augmented by replicating half of the cells it contained and artificially altering them to induce random noise of +/-1 copy sporadically. We trained the model for copy-numbers in

25, 100 and 500 kb bins which resulted in 97.94, 98.54 and 98.14% replicating state classification accuracy rates on the test datasets respectively. To quantify how well our 100 kb model performs in comparison with FACS sorting, we calculated the discordance percentage between our in-silico predictions and the FACS metadata of HCT-116 and GM12878 cells. We observed that FACS misclassifications accounted for 17.67% of wild-type (WT) HCT-116, 27.69% of double-knockout (DKO1) HCT-116, and 25.72% of GM12878 cells (Figure 15). These results demonstrate that even when taking into consideration the 1.56% error rate of our model, it generated results with accuracy superior to FACS for cell-phase sorting.

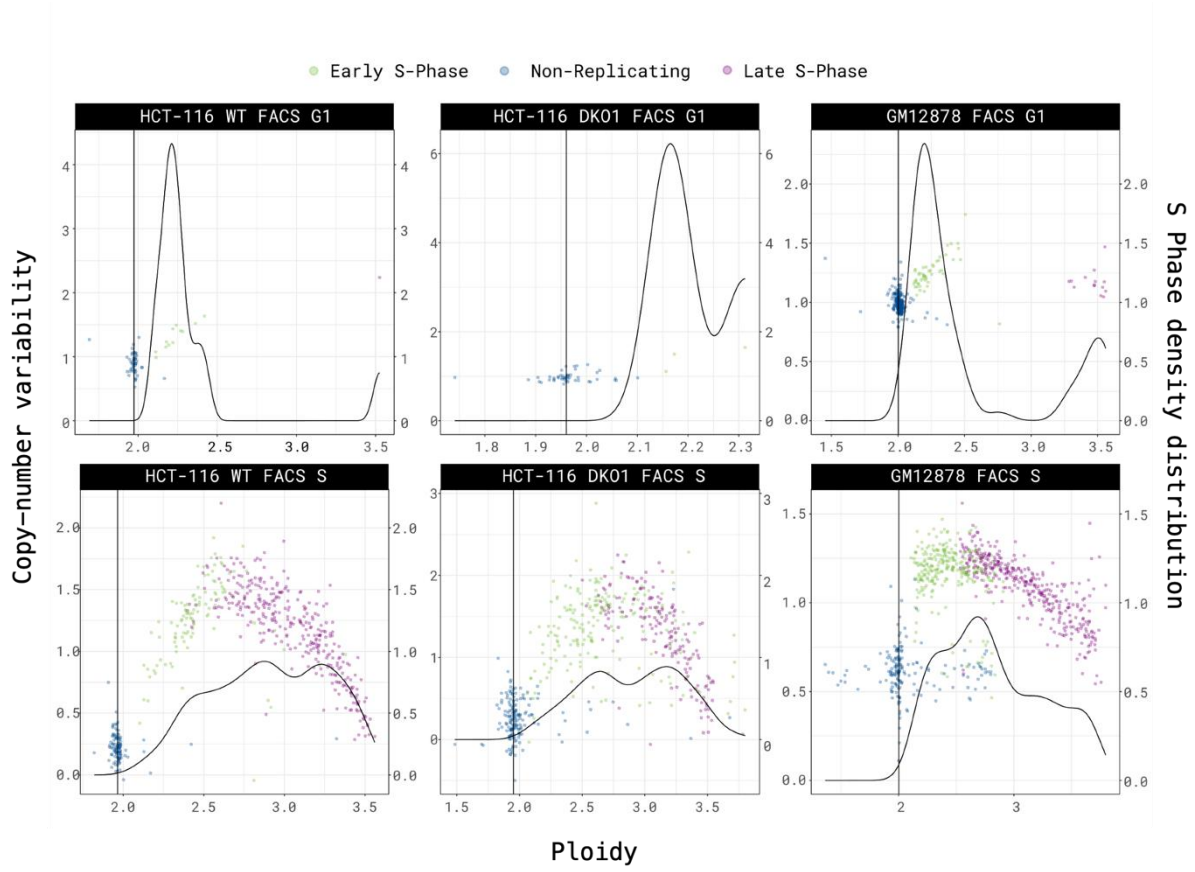


Figure 15: Discordance among FACS and the supervised deep learning method developed here on replication states of single-cells. *The upper row corresponds to FACS sorting of G1 phase cells and the lower row to S phase cells of HCT-116 wild-type (WT; left), double knock-out (DKO1; centre) and GM12878 (right) cells.*

3.3.2. Unsupervised machine learning automates subpopulation discovery of cancerous cells

Our next goal was to detect aneuploidy differences between cells, a crucial aspect of cancer emergence and evolution. To do so, we created a 3-step framework to detect genomic subpopulations – groups of cells that have distinct CNA signatures in comparison to other cells originating from the same sample. The autosomal copy-numbers of non-replicating cells, determined from our deep learning model, underwent dimensionality reduction to be represented in a two-dimensional pane. The 2D cell coordinates in these new representations would then be used to detect subpopulations with Density-Based Spatial Clustering of Applications with Noise (DBSCAN), an unsupervised spatial clustering algorithm. Although UMAP is relatively stable, due to it being a stochastic algorithm[189] that could generate non-representative distances

of high-dimensional data, the UMAP/DBSCAN steps were repeated another 6 times under random seeds ranging between 3 and 2^{30}. The number of subpopulations was counted with each seed. If the predominate number of clusters was not found in the original iteration, the seed would change to the first encountered of the 6 random seeds that would. Subpopulations were merged while they presented >98.5% median copy-number identity in a prioritised order. Finally, replicating cells would also be included and a second dimensionality reduction step in 10 dimensions allowed the KNN algorithm to match replicating cells to their corresponding non-replicating subpopulations.

To validate this method on genome-wide distinct CNA landscapes, we first mixed JEFF and HeLa copy-number data to be analysed as if they were a single sample, with the expectation that the two cell lines would be correctly distinguished. We first observed that the replicating cells in both cell lines were visually distinguishable in the 2D landscape (Figure 16 A-C). After running our 3-step subpopulation detector, we confirmed that, without providing any in-formation on the cell origins, they were matched back into 2 populations corresponding to JEFF and HeLa for both non-replicating (Figure 16 D) and replicating (Figure 16 E-F) cells. Furthermore, we exposed the existence of only one copy of chromosome X of JEFF (Figure 16 D) cells, instead of two which is typical for females, a phenomenon compatible with acquired monosomy X.

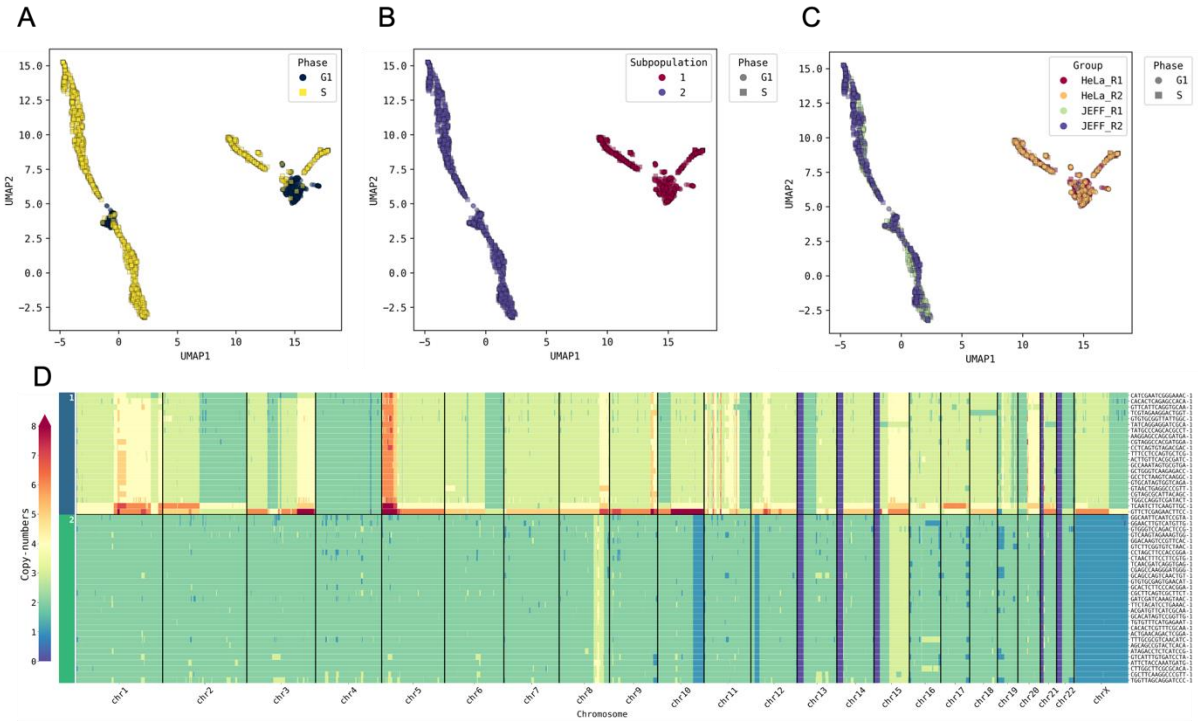


Figure 16: scCNV distinctions with unsupervised learning. *A-C. UMAP pane of JEFF and HeLa samples coloured by replication state (A), subpopulation (B) and replicate (C). D. 50 randomly selected cells and their genome-wide single-cell copy-numbers.*

We previously reported the revelation of 2 subpopulations of MCF-7 cells[62], a breast cancer cell-line known for unstable aneuploidy. We used our process to automatically detect subpopulations from the single, but heterogenous, MCF-7 sample. The 2 subpopulations could be distinguished by sub-chromosome (Figure 17 A) and whole-chromosome (Figure 17 B) copy-number differences and were divergent on UMAP's reduced dimension pane Figure 17 E). We then applied this same method to HCT-116 wild-type (WT) cells and discovered the existence

of 2 sub-populations (Figure 17 F), which was previously unreported[154]. Contrary to the MCF-7 cells, the observed local CNA changes (Figure 17 C-D), were likely due to DNA repair pathways rather than global genomic instability. This agrees with the fact that the HCT-116 cell line is known to be defective for the MMR pathway, containing a homozygous mutation of the MMR gene hMLH1 on chromosome 3, while also exhibiting microsatellite instability[202,203], which could be an explicable cause for this state.

To further validate our subpopulation discovery technique, published copy-number data from 42,759 single-cells issued from ref.[56] which were obtained from ref.[201]. These single-cell copy-numbers were based on the hg19 human genome and generated with HMMCopy (computational copy-number estimator based on a Hidden Markov Model), a reference genome and copy-number estimator that were different to the other data analysed in the study. Upon visual inspection of single-cell genome-wide copy-number heatmaps (Supplementary Figures from Section 8.3.3), we determined that there copy-number signatures specific to each subpopulation. Thus, we concluded that our subpopulation discovery approach is efficient on different reference genomes and with copy-numbers obtained from wither Kronos scRT or HMMCopy.

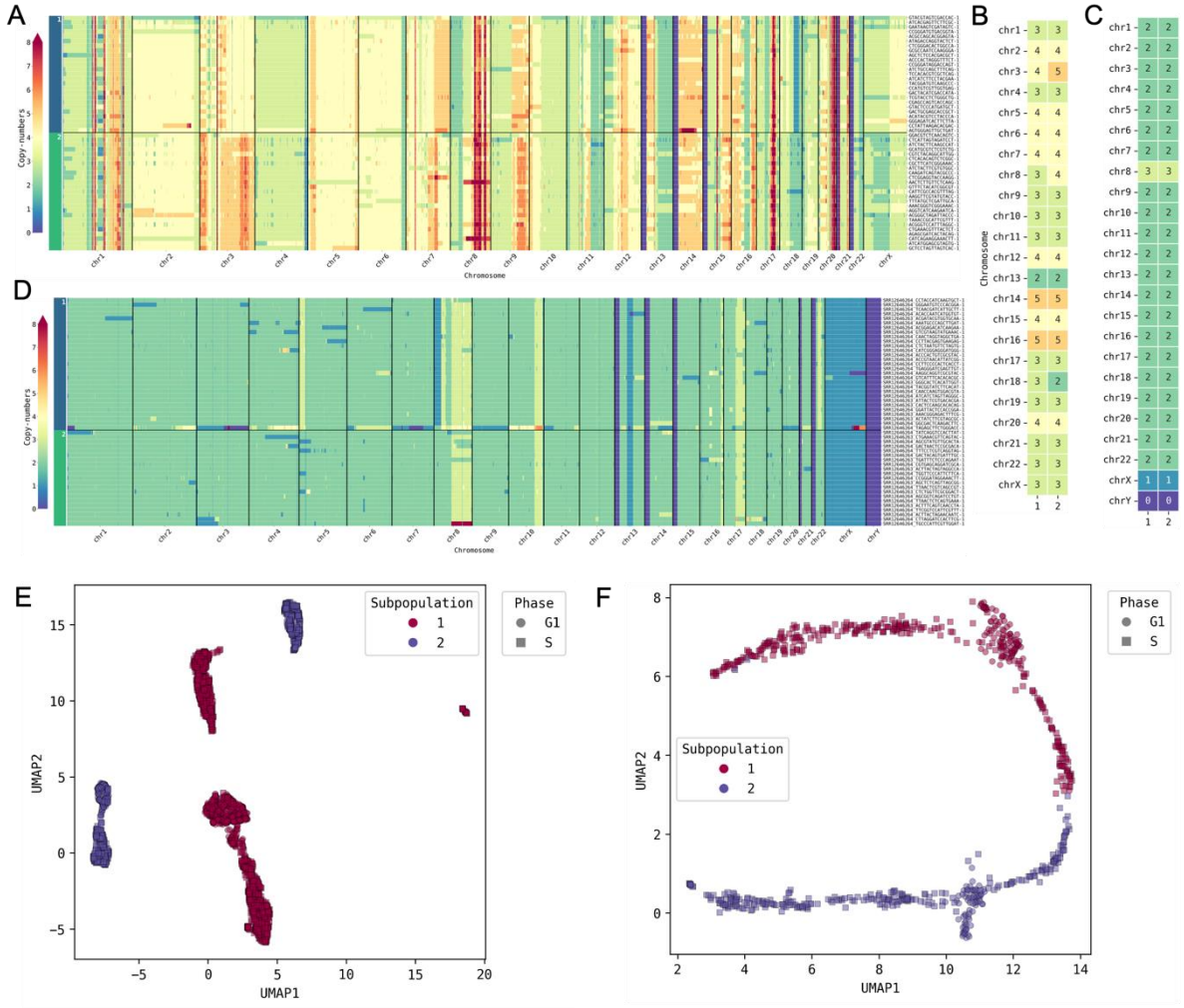


Figure 17: Genomic heterogeneity detected in individual samples of cancer cell-lines. *A-D: genome-wide copy-numbers (A,D) summarised by their median (B-C) of MCF-7 (A-B) and HCT-116 (C-D). E-F: reduced dimension planes by UMAP of MCF-7 (E) and HCT-116 (F).*

3.3.3. MnM: a fast and accurate tool integrating machine learning for replication states and subpopulation discoveries

The machine learning approaches from Sections 3.2.2, 3.3.1 and 3.3.2 were integrated to provide a single ready-to-use tool, MnM: Mix 'n' Match, that unifies these techniques under one program (Figure 18). Copy-number imputation, replicating state classification and subpopulation detection enabled scRT extraction from heterogenous cell populations and related downstream analyses, in vivo and in vitro. In addition to the reported accuracies, MnM is a fast tool, with a runtime of 7m:22s for 713 HCT-116 WT cells in 100kb bins running on a macOS v13.5.2 computer system with 6 intel i5 cores. MnM's source code is available on GitHub (https://github.com/CL-CHEN-Lab/MnM) and is protected by the French Agency for the Protection of Programs (APP) under the registration number IDDN.FR.001.340005.000.S.P.2023.000.31230. The simplified pseudocode can be found in the Supplementary Section 8.1.

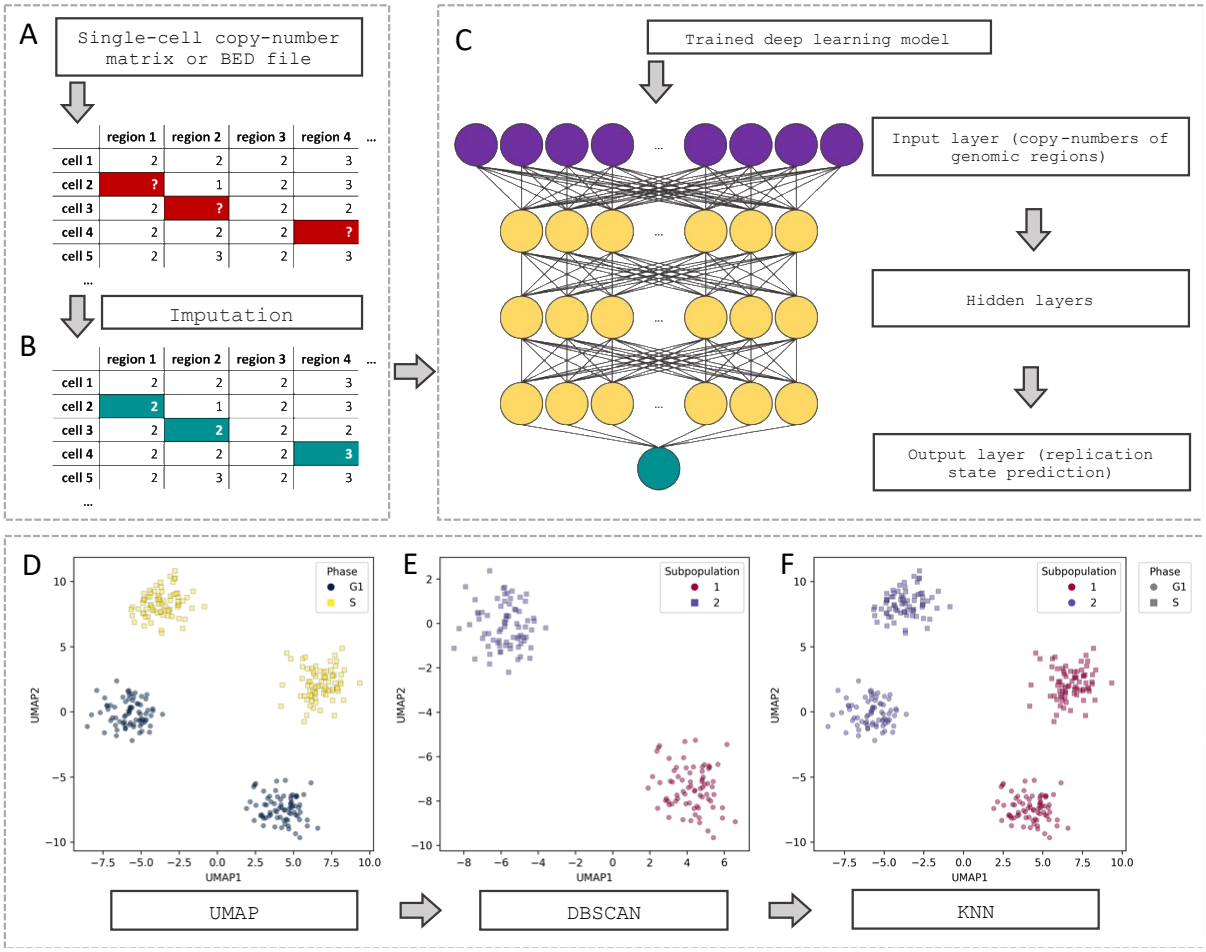


Figure 18: Machine learning techniques used throughout MnM. *A-B. Copy-number imputation with KNN. Single cell data is used as an input in either a matrix or BED file format with missing copy-number values (A) which are filled in with KNN imputation (B). C. Deep learning for a single-cell replication state classifier. The trained deep learning model consisting of 3 hidden and 1 output layer is then loaded and used to distinguish replication states of the single-cells. D-F. Subpopulation discovery in 3 steps. Dimensionality reduction is performed with UMAP on non-replicating cells under 2 dimensions to provide representative lower dimensions of the copy-number data (D). DBSCAN clusters the data based on the UMAP coordinates (E) which allows replicating cells to be matched to non-replicating subpopulations with KNN under 10-dimension UMAP coordinates (F).*

3.4. Unravelling scRT in heterogenous samples

3.4.1. DNA RT retains high fidelity in cell-lines despite CNAs

With a new tool, MnM, allowing to detect subpopulations from both replicating and non-replicating single-cells, it was therefore possible to extend the techniques developed to heterogenous cell populations issued from a single sample (single experiment). Copy number data by subpopulation could now be split and the detected cell phases could be provided to Kronos scRT to obtain the RT profiles. MCF-7 and HCT-116 cells were analysed with MnM to discover heterogeneity and extract scRT profiles. Because the copy-numbers calculated with Kronos scRT were relative, the first and second parts of S-phase copy-numbers were corrected for MCF-7 (Figure 19A-B) and HCT-116 (Figure 19 C-D) cells in 200 kb bins. This let to mid-S phase regions sometimes being miscategorised as illustrated on the region-to-region density plots (Figure 19 G-H) but did not disrupt global RT profiles, as seen when compared to bulk data (92.8-94.4% Spearman correlation; (Figure 19 G-H). We observed that the S/G1-phase borderline was non-linear on the variability scale (Figure 19 A-D), signifying that separation of the replicating states with previous computational methods using linear techniques with a unique cut-off value, as in previous studies[62,165], would have introduced a larger error rate. For each subpopulation, scRT profiles were inferred and visualised (Figure 19 E-F). Despite genome-wide CNAs, the pseudo-bulk RT profiles of the 2 MCF-7 subpopulations had a Spearman correlation of 93.6% (Figure 19 G). As expected, due to the smaller copy-number signatures, the HCT-116 profiles were also highly correlated at 97.5% (Figure 19 H).

Although these reported correlations between the subpopulations may seem strong, it is uncertain whether we can say with certainty if the replication timing profiles are significantly different. A previous attempt here, used the Wilcoxon test to compare RT profiles. However, this may not always be a suitable for the nature of RT (pseudo-)bulk data due to the large number of genomic regions and measurements with the same of very similar values[204]. We previously introduced the concept of scRT trajectories[62] which position the replicating single-cells in a representative manner, illustrating the progression of S phase (from early to late). This representation was also efficient for visual distinction of the scRT paths between subpopulations. In an attempt to assess whether there was a significant difference in the trajectories between the subpopulations, the scRT trajectory coordinates were collected, grouped by subpopulation and used for a permutation test (under 1,000 permutations), a non-parametric method that does not assume specific data distributions and allows for the randomisation of group labels, thus compatible with UMAP coordinates which can produce results that vary between datasets. In agreement with visual inspection (Supplementary Section 8.3.5), the trajectories of the two MCF-7 subpopulations were significantly different (9.99e-4 p-value) while those of the HCT-116 cells were not (0.26 p-value).

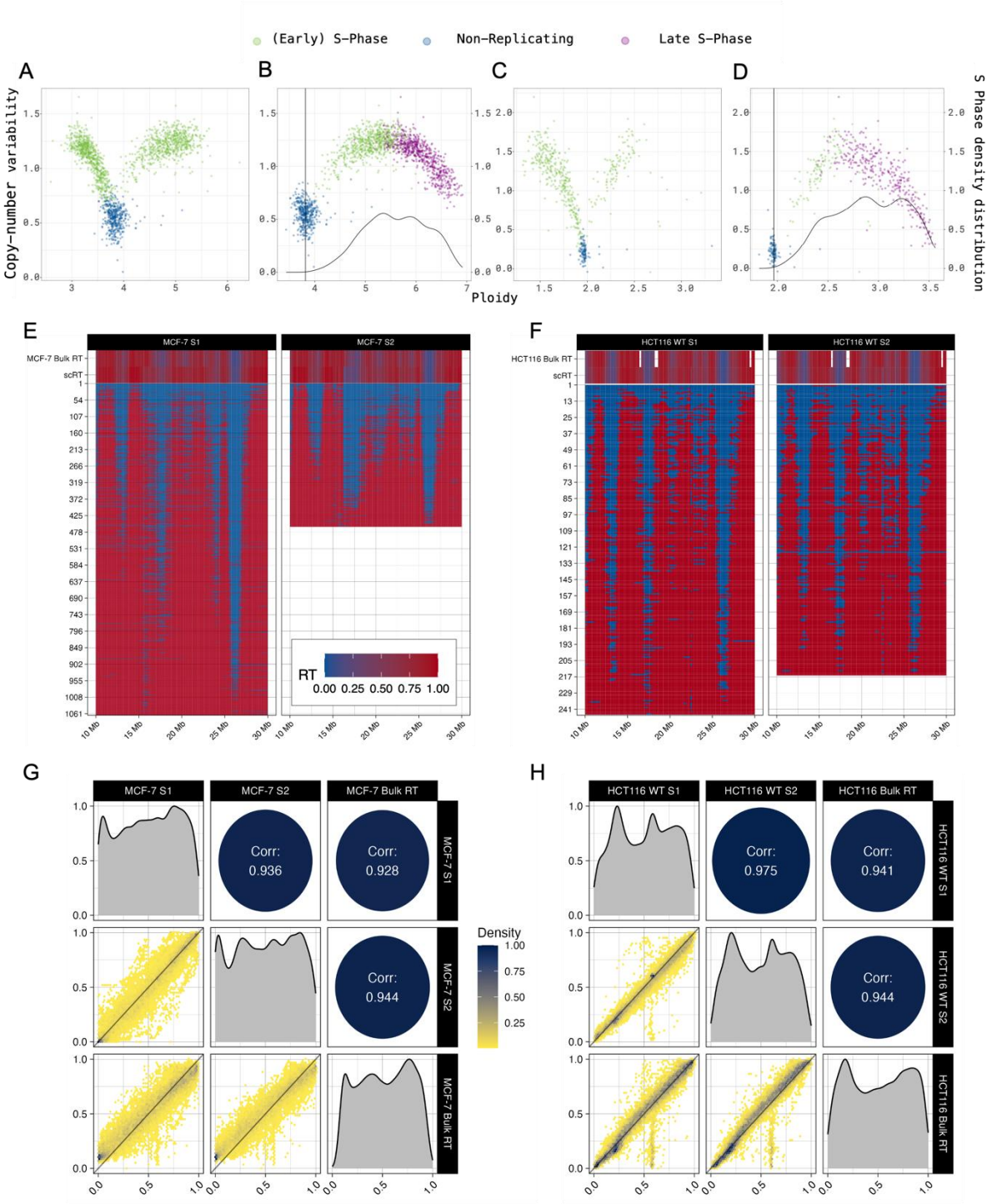


Figure 19: scRT of heterogenous cancer cell lines uncovered. *A-D: Early (green) and late (purple) S-phase cells are corrected (B,D) from raw scCNV data (A,C) displaying non-replicating (blue) and replicating (green) cells. E-F: scRT landscapes of chromosome 16 from MCF-7 (E) and HCT-116 (F) display minor differences between subpopulation with pseudo-bulk (scRT) and bulk RT are displayed in the upper window. G-H: correlations between pseudo-bulk subpopulation scRT and bulk RT for MCF-7 (G) and HCT-116 (H) cells.*

3.4.2. Replication timing changes in patient-derived breast cancer

Since replication timing in heterogenous tumours has not been studied, we used the methods we developed in the same manner as done with the cell-lines to discover cell phases and subpopulations from published data obtained from a triple-negative breast cancer (TNBC) tumour sample (SA1135). As in the original study[201], we discovered 1 diploid and 3 aneuploid subpopulations (Figure 20 A-B). Out of the 345 cells that passed quality control, 193 were not and 152 were replicating, showing a larger than expected proportion of replicating cells than those obtained from the cell lines models, which is concordant with persistent proliferation of cancer cells. We then calculated the scRT profiles for each subpopulation. We considered that subpopulations 1 (n=13 S phase cells) and 3 (n=29) did not have a representative S phase landscape and disregarded them in the analysis (Figure 20 D). Remarkably, we discovered that subpopulations 2 (n=36) and 4 (n=74) showed distinct RT programmes, indicating a deregulated replication programme in vivo. The two replication profiles from the same tumour correlated at 73.3%, and the permutation test on their trajectories yielded a 9.99e-4 p-value. Thus, these results demonstrate that RT can be modified in subpopulations of the same TNBC tumour in vivo (Figure 20 C).

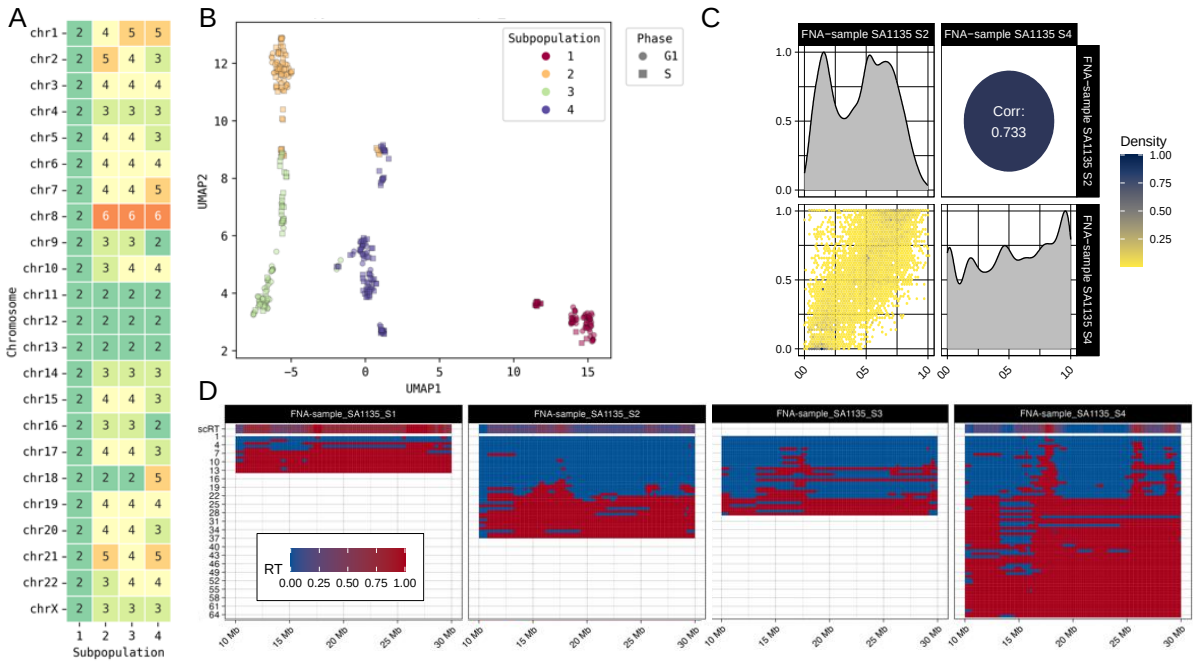


Figure 20: scRT extraction of the subpopulations from the human triple negative breast cancer sample SA1135. *A: median DNA copy-numbers per chromosome for the 4 subpopulations displaying major differences. B: Reduced dimension UMAP plane of the copy-numbers of the single-cells of the tumour. C: Spearman correlation between the subpopulation 3 and 4. D: scRT landscapes of chromosome 21 from the 4 subpopulations with the pseudo-bulk RT from the scRT profiles in the upper windows. FNA: Fine-Needle Aspiration (cell collection technique).*

3.4.3. RT cancer comparisons reveal cell-type relationships

We extended the use of our methods to more datasets. In total we analysed the copy-numbers of 119,991 quality-controlled cells originating from 92 different samples spanning across 21 different somatic/cancer cell-lines, 35 patient tumours and 19 patient-derived xenografts (PDX) samples. These cells originating from 60 individual samples (Supplementary Table 3). We calculated RT when we had enough cells to reconstruct a representative S phase land-scape.

This was determined either by software failure or visual inspection of the replication patterns and manual elimination. A total of 41 (sub)populations were used and the Spearman correlation for each pair was calculated (Figure 21).

In contrast to two subpopulations from the same sample, we noticed that MCF-7 samples from different laboratories[62,165] only presented an 84.5% correlation, on average. Knowing that this cell line is known to have variable karyotypes, we speculated that the RT differences could be caused by wide-spread copy-number differences. Indeed, we discovered that this cell line only shared between 11 and 13 common median chromosome copies be-tween the two sample origins (Figure 17 B, Supplementary Section 8.3.3). JEFF and GM- lymphoblastoid cell lines on the other hand, had extremely high RT correlations all ranging >91%, regardless of the sample origin, showing consistent RT in the same-cell type. Despite the fact that both H7 hESCs and GM12892 present a perfectly diploid karyotype (Supplementary Section 8.3.3), their replication tracks present a 79% correlation, illustrating that RT can be used as a cell-type specific biomarker.

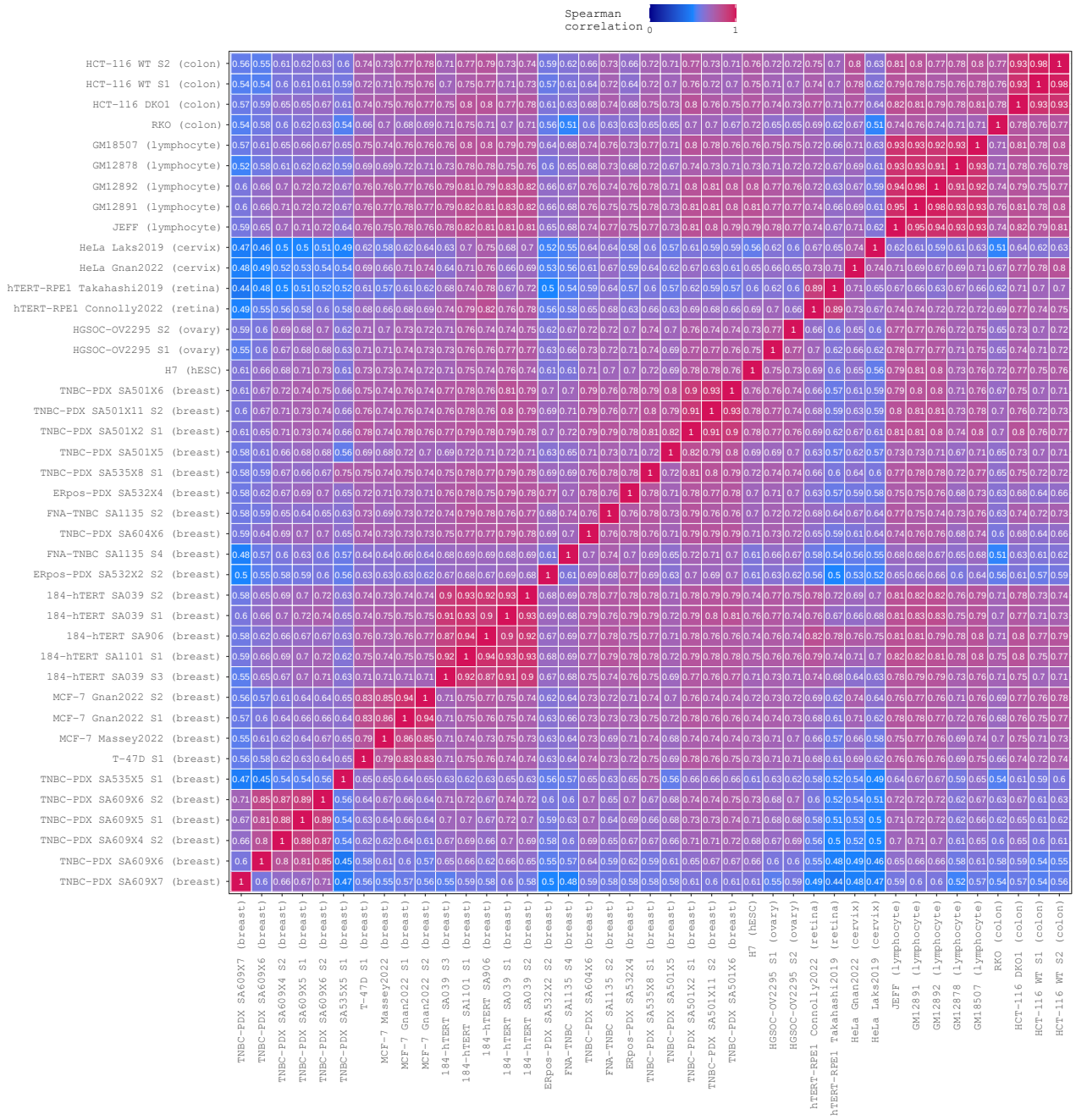


Figure 21: The scRT atlas represented as Spearman correlations between the 41 scRT pseudo-bulk profiles extracted from the human (sub)populations in a multi-sample approach reveal cell-type specific relationships of DNA replication. *Subpopulations were ordered by hierarchical clustering using the second version of Ward's minimum variance method (Ward.D2).* ***Subpopulation enumeration is given at the end of the sample name, preceded by S (i.e. S1, S2, etc…) Cell origins are given in brackets after the sample name. PDXs are encoded as [sample number]X[mouse passage]. RPE: retinal pigment epithelial cells; hESC: human embryonic stem cell; TNBC: triple-negative breast cancer; ERpos: estrogen-receptor positive; PDX: patient-derived xenograft; HGSOC: high-grade serous ovarian carcinoma; hTERT: human telomerase reverse transcriptase; FNA: fine-needle aspiration (cell collection technique).***

4. Discussion

In this PhD project two innovative computational tools were developed to advance our understanding of single-cell replication timing (scRT) and its association with genomic subpopulations: Kronos scRT, a unified pipeline for scRT analysis, and MnM, a fast and efficient tool to establish single-cell replicating states and reveal genomic subpopulations from a single heterogenous sample. MnM, a tool designed to detect single-cell replicating states and genomic heterogeneity, utilises single-cell copy-number data to eventually discover replication patterns with Kronos scRT or any other scRT method. Through a rigorous validation process of its methods, MnM demonstrated remarkable accuracy and speed in missing-value imputation and cell replicating state classification. By performing subpopulation clustering, MnM can discern different cell types and subpopulations within a heterogenous sample. These subpopulations are then used to extract scRT profiles using Kronos scRT, marking it the first automated method for conducting RT studies in heterogeneous cancers and patient-derived samples. MnM requires a minimum of 10 single-cells for subpopulation detection and the replication state classifier can work on a single cell.

By leveraging the genomic information from the single-cells analysed in this project, MnM adeptly disentangled heterogeneity, aligning replicating and non-replicating cells for each subpopulation. The single-cell RT atlas, a major outcome of this study, revealed an additional layer of heterogeneity in cancer progression, the existence of different RT profiles within single tumours. This finding highlights the dynamic nature of cancer biology and further underlines the importance of considering intra-tumoral heterogeneity when studying cancer samples or treating patients. By uncovering these different RT profiles within samples, these tools and data open new opportunities for a better understanding of the spatiotemporal dynamics throughout tumorigenesis and cancer progression. In the future, if RT signatures can be linked to therapeutic outcomes, this more comprehensive understanding of genomic heterogeneity could be added to the equation when designing combination therapies, which would improve how treatment resistance and recurrence are addressed.

Despite these advancements in better understanding heterogeneity, the precise relationships between RT and other genomic features remain incompletely understood. Exploring regions that transition between RT categories, such as late to early replication, has yielded limited insights into the connection between transcription and RT. Many of the biological functions were generic and did not give a complete explanation of the presumed activated genes during differentiation. Likewise, the ATAC peak counts only provided moderate correlations with RT data. In order to further study the relationship between transcription, chromatin organisation and RT, new multi-omic data needs to be made available. Techniques such as multi-omic single-cell sequencing[205–207] could be an ideal method to employ in order to study this. With the transcriptome and genome of individual cells obtained in a high-throughput fashion, a comprehensive investigation of this relationship would be achieved. Nonetheless, it remains unknown why RT is usually intact besides tweaks on other features. One reason could be that the exact replication patterns are not important, but rather ensure that the genome does not replicate too fast. The cells have a limited number of resources meaning that too many active replication forks at once, would inevitably lead to genome instability. Furthermore, too many simultaneous double strand breaks could lead to improper ends being reconnected and thus, translocations.

Besides the two main computer programs highlighted in this project, smaller scripts were also created and are provided to the scientific community in an open-source manner. Notably the valid barcode detector based on the EM algorithm could be useful for single-cell analyses.

Although not demonstrated here, there is reason to believe that this method could also be extended to other omic data such as scATAC and scRNA barcodes, especially those from the 10X Chromium single cell multiome ATAC and Gene Expression solution. Thus, the code used here could also be applied to other projects containing other single-cell omic data types.

An important finding was that FACS sorting is prone to a high error rate of up to 27.69% for cell phase sorting. This is a similar figure to a previously reported FACS sorting error rate for sorting single-cell phases obtained by using hidden Markov models[165]. Although in some cases there is a legitimate interest to sort single-cells before sequencing, the work performed here shows that it is important that cell-sorted metadata is verified computationally to avoid any erroneous conclusions from noise that might be induced from high rates of missorted cells. While the replicating state classifier of MnM was only trained on the hg38 reference genome, this method can easily be extended to other genomes and used routinely. Furthermore, in some cases, the FACS metadata may not exist (e.g. unsorted samples). This could be the case of precious samples such as embryonic stem cells and tumours which contain a limited number of cells one would not want to lose from a limited yield after sorting. It is known that technical and human errors during sorting can lead to cells being miscategorised[200]. Indeed, many factors can play a role in the efficacy of cell sorting such as the cell type, the condition of the cells (e.g. drug treatments), the pressure at which the cells are sorted as well as the type of buffer used (e.g. carbonate/phosphate) and will only provide a 75-90% yield of the initial cell population[208]. Seeing the results in this project, *in silico* predictions could be valuable for these cases and should be routinely performed.

Furthermore, we introduced the concept of scRT trajectories, which we compared with a permutation test to be able to determine whether the RT programmes differ or not between samples or subpopulations. Other studies have failed to determine with certainty whether the RT profiles are indeed different between different conditions[98,154]. Although we obtained scRT of PDXs, we did not infer the SVs to compare to single-cell lineage tracing to further understand the relationship between CNAs and RT. A related project undertaken in our team[209] has shown that various single-cell lineage tracing algorithms do not yield the same results, and thus, there is no such algorithm that has prevailed for such tasks. It is therefore necessary that such algorithms, possibly inspired by mitochondrial DNA from single-cells which provide an excellent coverage compared to the rest of the single-cell genome[66,210], are developed and made available in order to further comprehend this relationship. Moreover, this project focused on CNAs and therefore, any translocations or inversions were overlooked, limiting the results of the RT and SV relationship.

Another discovery made here was that JEFF cells had lost a copy of chromosome X, a phenomenon correlated with aggressive tumour growth or occurring from ageing[211]. Unlike the other cells lines present in this project which have well-documented histories, the JEFF cell line is less documented. Despite the scarcity of documentation, the absence of the Y chromosome, along with personal communications from colleagues, have allowed us to confirm that this cell-line was derived from a female patient. Important chromosomal aberrations found in the analysed samples, in cell-lines and patient tumours, further underline the importance of DNA copy-number screening. Aneuploidy is an omnipresent trait in the genomes of tumours[119]. Though the presence of genomic instability in cancer has been recognized for a long time[171], the exact role it plays in tumour development remains unclear. The belief that chromosome gains might amplify the expression of genes promoting tumours, called oncogenes, cushioned within the altered regions, has been proposed[212]. Yet, the generalisation of this theory remains disputable. Alternatively, it has been suggested that aneuploidy could stem from the

disruption of checkpoint control – a common occurrence in advanced malignancies[110,141]. What adds an intriguing layer to this discourse is the observation that individuals with Down syndrome, arising from the triplication of chromosome 21, exhibit a significantly diminished susceptibility to most solid tumours[213]. As discussed by others[119], this intriguing correlation suggests that aneuploidy might surprisingly exert tumour-suppressive effects in specific cases.

The robustness of the DNA replication machinery is a cornerstone of cellular integrity. Without it, genomic instability, a hallmark of cancer, can be triggered. In this study, we observed a noteworthy contrast between cell-line models and patient-derived samples in terms of DNA replication timing disruptions. Remarkably, cell-line models exhibited relatively modest distortions in DNA replication dynamics. These models, cultured under controlled conditions, often reflect simplified representations of cellular systems. However, a compelling finding emerged in our analysis of patient samples, where we identified substantial and impactful disruptions in DNA replication patterns. The mouse embryonal carcinoma analysed also exhibited temporal changes, but the number of cells was limited, and therefore the interpretation of the resulting temporal changes cannot be confident. Nonetheless, these observations resonate with the idea of the relevance of the tumour microenvironment and its intricate interplay with genomic stability. The disparities between cell-line and patient sample dynamics highlight the necessity of integrating complex, patient-specific factors into our understanding of DNA replication mechanisms in the context of cancer progression.

Furthermore, it is essential to acknowledge the persistent challenge in whole-genome single-cell studies which are still failing to overcome the low coverage of reads across the genome. As new methods are emerging with promising advancements, notably a recent report of long-read single-cell sequencing[214], future investigations will be able to dig further into the precise relationship between the mutational landscape, aneuploidy and the replication programme. If used with muti-omic sequencing, we will be able to have a much better image of the molecular processes in the cells during DNA replication. Eventually, with the imminent generation of higher resolution data, studies will be able to address the replication differences of different homologues with scRT. Thus, we underline the necessity for detailed analyses examining the replication synchronicity of alleles, an even more complex task for aneuploid cancers.

While the focus of this research primarily revolved around the development of computational tools for scRT analysis, it is essential to recognise the growing significance of single-cell spatial multi-omics in the broader context of genomics and cancer biology. Although not explored in this project, the integration of spatial multi-omics data with scRT could hold immense potential. Spatial multi-omics techniques, such as spatial transcriptomics and spatial proteomics, enable the concurrent analysis of multiple molecular layers within individual cells while preserving their spatial context. By combining scRT with spatial multi-omics, we could gain deeper insights into how replication timing impacts the spatial organisation of genomic features within individual cells. This integrated approach could uncover critical connections between replication timing and spatial genomic architecture, shedding light on how these factors contribute to the development and progression of heterogeneous cancers. As we move forward, exploring this uncharted territory in single-cell spatial multi-omics could pave the way for a more comprehensive understanding of the intricate interplay between replication dynamics and the spatial organisation of the genome, ultimately advancing our knowledge of cancer biology.

Furthermore, in the era of personalised medicine, genomic profiling has emerged as a powerful tool for tailoring treatment strategies to each individual patient. MnM, could play a role in genomic profiling by being included in comprehensive analyses of a patient's genetic makeup to identify specific copy-number DNA alterations that may drive their cancer. By

understanding the genetic underpinnings of a patient's tumour, oncologists could make more informed decisions about treatment options, based on availability, including targeted therapies and immunotherapies. The systematic integration of genomics into clinical practice would mark a significant step towards personalised medicine. Although for MnM to be included in such practices, it would need to be further tested, and tweaked to suit diagnosis indicators.

As the field of genomics is continuously evolving, artificial intelligence is now becoming an important tool for the analysis of such complex data. Machine learning has already provided tools that allow to address questions that were previously hard or impossible to answer with the speed and accuracy these models provide[102,103]. MnM adds on to this list of tools which harness the power of both these fields. While such tools could hold potential in enabling new means of medical diagnosis, inherent risks must be considered. One concern is algorithmic bias, where models may amplify disparities in the data. Indeed, overfitting, which occurs when the model provides accurate predictions for training data but not for new data, is a well-known risk[215,216]. As a result, machine learning solutions for medical diagnosis should still be interpreted with caution. Moreover, the interpretability of these algorithms could be difficult, which may further raise questions of trust in decision-making configurations. As the advantages of machine learning for research and health are being harnessed, these technological advancements might outpace legal and ethical frameworks. Thus, it is important to address these challenges proactively to ensure a responsible use of these new tools.

In conclusion, throughout this project two computational tools have been created, MnM and Kronos scRT. MnM, an AI based tool, was designed to democratise single-cell subpopulation detection from DNA copy-numbers. The outcomes of this tool can help contribute to our understanding of cancer emergence and progression. When used with Kronos scRT to compute replication profiles, it allows the study of scRT in heterogenous cancer samples. This project provides a large amount of single-cell copy-number and scRT data for the community, which could be an important resource for further research and discoveries. The data showed that large CNAs can modify the RT programme whereas smaller sub-chromosomal CNAs do not modify RT. This project regrouped various machine learning techniques from unsupervised learning, such as dimensionality reduction (UMAP, t-SNE) and clustering (expectation-maximisation, DBSCAN, hierarchical), as well as supervised leaning (deep learning sequential model). Thus, this work further demonstrates the importance of machine learning in genomics. Finally, our results underline the necessity to consider tumour samples in order to fully understand the mechanisms governing DNA replication in cancer as well as the generation of single-cell multiomic data to completely understand the relationship between RT and other factors. Although cell lines constitute an easier research model to study, they lack some critical environmental factors that interact with cancer. Therefore, this project paves the way for further detailed investigations in the RT programme with the new methods developed.

5. Citations

1. Rouvray, D. H. John Dalton: the world's first stereochemist. *Endeavour* **19**, 52–57 (1995).

2. Dahm, R. Friedrich Miescher and the discovery of DNA. *Dev Biol* **278**, 274–288 (2005).

3. Satzinger, H. Theodor and Marcella Boveri: chromosomes and cytoplasm in heredity and development. *Nat Rev Genet* **9**, 231–238 (2008).

4. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

5. Maddox, B. The double helix and the 'wronged heroine'. *Nature* **421**, 407–408 (2003).

6. Cobb, M. & Comfort, N. What Rosalind Franklin truly contributed to the discovery of DNA's structure. *Nature* **616**, 657–660 (2023).

7. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977).

8. Hood, L. & Rowen, L. The human genome project: big science transforms biology and medicine. *Genome Med* **5**, 79 (2013).

9. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (1979)* **291**, 1304–1351 (2001).

10. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

11. Catherine Vincent. Notre patrimoine génétique décrypté. *Le Monde* 16–17 (2000).

12. Derelle, E. *et al.* Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proceedings of the National Academy of Sciences* **103**, 11647–11652 (2006).

13. Goff, S. A. *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science (1979)* **296**, 92–100 (2002).

14. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).

15. Adams, M. D. *et al.* The Genome Sequence of Drosophila melanogaster. *Science (1979)* **287**, 2185–2195 (2000).

16. C. elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science (1979)* **282**, 2012–2018 (1998).

17. Aparicio, S. *et al.* Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. *Science (1979)* **297**, 1301–1310 (2002).

18. Holt, R. A. *et al.* The Genome Sequence of the Malaria Mosquito Anopheles gambiae. *Science (1979)* **298**, 129–149 (2002).

19. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

20. Goffeau, A. *et al.* Life with 6000 Genes. *Science (1979)* **274**, 546–567 (1996).

21. van den Berg, M. A. *et al.* Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum. *Nat Biotechnol* **26**, 1161–1168 (2008).

22. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* **414**, 450–453 (2001).

23. Dufresne, A. *et al.* Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proceedings of the National Academy of Sciences* **100**, 10020–10025 (2003).

24. Bolotin, A. *et al.* The Complete Genome Sequence of the Lactic Acid Bacterium Lactococcus lactis ssp. lactis IL1403. *Genome Res* **11**, 731–753 (2001).

25. Stephens, R. S. *et al.* Genome Sequence of an Obligate Intracellular Pathogen of Humans: Chlamydia trachomatis. *Science (1979)* **282**, 754–759 (1998).

26. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences* **99**, 17020–17024 (2002).

27. Southern, E. M. DNA Microarrays. in *DNA Arrays: Methods and Protocols* (ed. Rampal, J. B.) 1–15 (Humana Press, Totowa, NJ, 2001). doi:10.1385/1-59259-234-1:1.

28. Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J. & Kimmel, M. Microarray experiments and factors which affect their reliability. *Biol Direct* **10**, 46 (2015).

29. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**, 133–141 (2008).

30. Check Hayden, E. Technology: The $1,000 genome. *Nature* **507**, 294–295 (2014).

31. Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics* **38**, 95–109 (2011).

32. Lazarevic, V. *et al.* Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* **79**, 266–271 (2009).

33. Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat Genet* **40**, 987–993 (2008).

34. Hardigan, M. A. *et al.* Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proceedings of the National Academy of Sciences* **114**, (2017).

35. Yano, K. *et al.* Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat Genet* **48**, 927–934 (2016).

36. von Wettberg, E. J. B. *et al.* Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation. *Nat Commun* **9**, 649 (2018).

37. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 30 (2020).

38. THE ENCODE PROJECT CONSORTIUM. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (1979)* **306**, 636–640 (2004).

39. Pennisi, E. ENCODE Project Writes Eulogy for Junk DNA. *Science (1979)* **337**, 1159–1161 (2012).

40. Charlton Hume, H. K. *et al.* Synthetic biology for bioengineering virus-like particle vaccines. *Biotechnol Bioeng* **116**, 919–935 (2019).

41. Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology. *Nat Rev Microbiol* **12**, 381–390 (2014).

42. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (1979)* **337**, 816–821 (2012).

43. Lu, T. K., Khalil, A. S. & Collins, J. J. Next-generation synthetic gene networks. *Nat Biotechnol* **27**, 1139–1150 (2009).

44. James, C. *Global Status of Commercialized Biotech/GM Crops: 2015*. https://www.isaaa.org/resources/publications/briefs/51/executivesummary/pdf/b51-execsum-english.pdf (2015).

45. National Academies of Sciences, E. and M. *Human Genome Editing*. (National Academies Press, Washington, D.C., 2017). doi:10.17226/24623.

46. Jiang, S., Liberti, L. & Lebo, D. Direct-to-Consumer Genetic Testing: A Comprehensive Review. *Ther Innov Regul Sci* (2023) doi:10.1007/s43441-023-00567-5.

47. Oh, B. Direct-to-consumer genetic testing: advantages and pitfalls. *Genomics Inform* **17**, e33 (2019).

48. FDA. FDA allows marketing of first direct-to-consumer tests that provide genetic risk information for certain conditions. *FDA NEWS RELEASE* https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-direct-consumer-tests-provide-genetic-risk-information-certain-conditions (2017).

49. Adam, S. & Friedman, J. M. Individual DNA samples and health information sold by 23andMe. *Genetics in Medicine* **18**, 305–306 (2016).

50. Varanasi, L. User data stolen from genetic testing giant 23andMe is now for sale on the dark web. *Insider* (2023).

51. Wang, Y. & Navin, N. E. Advances and Applications of Single-Cell Sequencing Technologies. *Mol Cell* **58**, 598–609 (2015).

52. Hiratani, I. & Takahashi, S. DNA Replication Timing Enters the Single-Cell Era. *Genes (Basel)* **10**, 221 (2019).

53. Minussi, D. C. *et al.* Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* **592**, 302–308 (2021).

54. Kinker, G. S. *et al.* Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat Genet* **52**, 1208–1218 (2020).

55. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

56. Funnell, T. *et al.* Single-cell genomic variation induced by mutational processes in cancer. *Nature* **612**, 106–115 (2022).

57. Edrisi, M. *et al.* Phylovar: toward scalable phylogeny-aware inference of single-nucleotide variations from single-cell DNA sequencing data. *Bioinformatics* **38**, i195–i202 (2022).

58. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, (2018).

59. Kuipers, J., Tuncel, M. A., Ferreira, P., Jahn, K. & Beerenwinkel, N. Single-cell copy number calling and event history reconstruction. (2020) doi:10.1101/2020.04.28.065755.

60. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* **14**, 565–571 (2017).

61. Wang, F. *et al.* MEDALT: single-cell copy number lineage tracing enabling gene discovery. **22**, (2021).

62. Gnan, S. *et al.* Kronos scRT: a uniform framework for single-cell replication timing analysis. *Nat Commun* **13**, 2329 (2022).

63. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun* **10**, (2019).

64. Liu, L. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat Commun* **10**, (2019).

65. Zafar, H., Navin, N., Nakhleh, L. & Chen, K. Computational approaches for inferring tumor evolution from single-cell genomic data. *Curr Opin Syst Biol* **7**, 16–25 (2018).

66. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325–1339.e22 (2019).

67. Baker, S. M., Rogerson, C., Hayes, A., Sharrocks, A. D. & Rattray, M. Classifying cells with Scasat, a single-cell ATAC-seq analysis tool. *Nucleic Acids Res* **47**, e10–e10 (2018).

68. Do, V. H. & Canzar, S. A generalization of t-SNE and UMAP to single-cell multimodal omics. *Genome Biol* **22**, (2021).

69. Mallory, X. F., Edrisi, M., Navin, N. & Nakhleh, L. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol* **21**, (2020).

70. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* (2021) doi:10.1038/s41588-021-00790-6.

71. Markowska, M. *et al.* CONET: copy number event tree model of evolutionary tumor history for single-cell data. *Genome Biol* **23**, (2022).

72. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol* **20**, (2019).

73. Wang, R., Lin, D.-Y. & Jiang, Y. SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. *Cell Syst* **10**, 445–452.e6 (2020).

74. Kozlov, A., Alves, J. M., Stamatakis, A. & Posada, D. CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. *Genome Biol* **23**, (2022).

75. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput Biol* **14**, e1006245 (2018).

76. Mallory, X. F., Edrisi, M., Navin, N. & Nakhleh, L. Assessing the performance of methods for copy number aberration detection from single-cell DNA sequencing data. *PLoS Comput Biol* **16**, e1008012 (2020).

77. Zafar, H., Navin, N., Chen, K. & Nakhleh, L. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res* **29**, 1847–1859 (2019).

78. Ouzounis, C. A. & Valencia, A. Early bioinformatics: the birth of a discipline—a personal view. *Bioinformatics* **19**, 2176–2190 (2003).

79. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001–D1006 (2014).

80. Bossé, Y. & Amos, C. I. A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiology, Biomarkers & Prevention* **27**, 363–379 (2018).

81. Rietveld, C. A. *et al.* GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science (1979)* **340**, 1467–1471 (2013).

82. Tschandl, P. *et al.* Human–computer collaboration for skin cancer recognition. *Nat Med* **26**, 1229–1234 (2020).

83. Vilsker, M. *et al.* Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics* **35**, 871–873 (2019).

84. Alawi, M. *et al.* DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. *Sci Rep* **9**, 16841 (2019).

85. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

86. Nurk, S. *et al.* The complete sequence of a human genome. *Science (1979)* **376**, 44–53 (2022).

87. Mao, Y. & Zhang, G. A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics. *Nat Methods* **19**, 635–638 (2022).

88. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).

89. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).

90. Genome 10K Community of Scientists. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. *Journal of Heredity* **100**, 659–674 (2009).

91. Aaltonen, L. A. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

92. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

93. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

94. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

95. Li, Y. *et al.* Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* **166**, 4–21 (2019).

96. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J Res Dev* **3**, 210–229 (1959).

97. Randhawa, G. S. *et al.* Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS One* **15**, e0232391 (2020).

98. Takahashi, S. *et al.* Genome-wide stability of the DNA replication program in single mammalian cells. *Nat Genet* **51**, 529–540 (2019).

99. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* **20**, 389–403 (2019).

100. Hill, T. & Unckless, R. L. A Deep Learning Approach for Detecting Copy Number Variation in Next-Generation Sequencing Data. *G3 GenesGenomesGenetics* **9**, 3575–3582 (2019).

101. Pounraja, V. K., Jayakar, G., Jensen, M., Kelkar, N. & Girirajan, S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome Res* **29**, 1134–1143 (2019).

102. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983–987 (2018).

103. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

104. Middleton, G., Robbins, H., Andre, F. & Swanton, C. A state-of-the-art review of stratified medicine in cancer: towards a future precision medicine strategy in cancer. *Annals of Oncology* **33**, 143–157 (2022).

105. Israels, E. D. & Israels, L. G. The Cell Cycle. *Oncologist* **5**, 510–513 (2000).

106. Zhang, D. & O'Donnell, M. Chapter Six - The Eukaryotic Replication Machine. in *DNA Replication Across Taxa* (eds. Kaguni, L. S. & Oliveira, M. T.) vol. 39 191–229 (Academic Press, 2016).

107. Wu, L., Liu, Y. & Kong, D. Mechanism of chromosomal DNA replication initiation and replication fork stabilization in eukaryotes. *Sci China Life Sci* **57**, 482–487 (2014).

108. Bleichert, F., Botchan, M. R. & Berger, J. M. Mechanisms for initiating cellular DNA replication. *Science (1979)* **355**, (2017).

109. Remus, D. *et al.* Concerted Loading of Mcm2–7 Double Hexamers around DNA during DNA Replication Origin Licensing. *Cell* **139**, 719–730 (2009).

110. Malumbres, M. & Barbacid, M. Cell cycle, CDKs and cancer: a changing paradigm. *Nat Rev Cancer* **9**, 153–166 (2009).

111. Murray, A. Cell cycle checkpoints. *Curr Opin Cell Biol* **6**, 872–876 (1994).

112. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).

113. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

114. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31–46 (2022).

115. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* **11**, 220–228 (2010).

116. Macheret, M. & Halazonetis, T. D. DNA Replication Stress as a Hallmark of Cancer. *Annual Review of Pathology: Mechanisms of Disease* **10**, 425–448 (2015).

117. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85–97 (2006).

118. da Costa, A. A. B. A., Chowdhury, D., Shapiro, G. I., D'Andrea, A. D. & Konstantinopoulos, P. A. Targeting replication stress in cancer therapy. *Nat Rev Drug Discov* **22**, 38–58 (2022).

119. Girish, V. *et al.* Oncogene-like addiction to aneuploidy in human cancers. *Science (1979)* (2023) doi:10.1126/science.adg4521.

120. Bionano Genomics, Inc. Reveal More Genomic Variation That Matters With Optical Genome Mapping. https://bionano.com/wp-content/uploads/2022/12/Saphyr-System-Brochure-DIGITAL.pdf (2023).

121. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).

122. Zhang, H., Tombline, G. & Weber, B. L. BRCA1, BRCA2, and DNA Damage Response: Collision or Collusion? *Cell* **92**, 433–436 (1998).

123. Chen, S. & Parmigiani, G. Meta-Analysis of BRCA1 and BRCA2 Penetrance. *Journal of Clinical Oncology* **25**, 1329–1333 (2007).

124. O'Donovan, P. J. & Livingston, D. M. BRCA1 and BRCA2: breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair. *Carcinogenesis* **31**, 961–967 (2010).

125. Zatreanu, D. *et al.* Pol inhibitors elicit BRCA-gene synthetic lethality and target PARP inhibitor resistance. *Nat Commun* **12**, (2021).

126. Bajrami, I. *et al.* Genome-wide Profiling of Genetic Synthetic Lethality Identifies CDK12 as a Novel Determinant of PARP1/2 Inhibitor Sensitivity. *Cancer Res* **74**, 287–297 (2014).

127. Smith, P. *et al.* The copy number and mutational landscape of recurrent ovarian high-grade serous carcinoma. *Nat Commun* **14**, 4387 (2023).

128. Tung, N. M. & Garber, J. E. BRCA1/2 testing: therapeutic implications for breast cancer management. *Br J Cancer* **119**, 141–152 (2018).

129. Baumann, P. & West, S. C. Role of the human RAD51 protein in homologous recombination and double-stranded-break repair. *Trends Biochem Sci* **23**, 247–251 (1998).

130. Petermann, E., Orta, M. L., Issaeva, N., Schultz, N. & Helleday, T. Hydroxyurea-Stalled Replication Forks Become Progressively Inactivated and Require Two Different RAD51-Mediated Pathways for Restart and Repair. *Mol Cell* **37**, 492–502 (2010).

131. Manavalan, A. P. C. *et al.* CDK12 controls G1/S progression by regulating RNAPII processivity at core DNA replication genes. *EMBO Rep* **20**, (2019).

132. Morgan, D. O. CYCLIN-DEPENDENT KINASES: Engines, Clocks, and Microprocessors. *Annu Rev Cell Dev Biol* **13**, 261–291 (1997).

133. Malumbres, M. *et al.* Cyclin-dependent kinases: a family portrait. *Nat Cell Biol* **11**, 1275–1276 (2009).

134. Blazek, D. *et al.* The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev* **25**, 2158–2172 (2011).

135. Krajewska, M. *et al.* CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nat Commun* **10**, (2019).

136. Popova, T. *et al.* Ovarian Cancers Harboring Inactivating Mutations in CDK12 Display a Distinct Genomic Instability Pattern Characterized by Large Tandem Duplications. *Cancer Res* **76**, 1882–1891 (2016).

137. Lui, G. Y. L., Grandori, C. & Kemp, C. J. CDK12: an emerging therapeutic target for cancer. *J Clin Pathol* **71**, 957–962 (2018).

138. Preston, B. D., Albertson, T. M. & Herr, A. J. DNA replication fidelity and cancer. *Semin Cancer Biol* **20**, 281–293 (2010).

139. Hatton, I. A. *et al.* The human cell count and size distribution. *Proceedings of the National Academy of Sciences* **120**, (2023).

140. Caldecott, K. W. Single-strand break repair and genetic disease. *Nat Rev Genet* **9**, 619–631 (2008).

141. Hustedt, N. & Durocher, D. The control of DNA repair by the cell cycle. *Nat Cell Biol* **19**, 1–9 (2017).

142. Li, Y. *et al.* Elevated expression of Rad51 is correlated with decreased survival in resectable esophageal squamous cell carcinoma. *J Surg Oncol* **104**, 617–22 (2011).

143. Hannay, J. A. F. *et al.* Rad51 overexpression contributes to chemoresistance in human soft tissue sarcoma cells: a role for p53/activator protein 2 transcriptional regulation. *Mol Cancer Ther* **6**, 1650–1660 (2007).

144. Yoshikawa, K. *et al.* Abnormal expression of BRCA1 and BRCA1-interactive DNA-repair proteins in breast carcinomas. *Int J Cancer* **88**, 28–36 (2000).

145. Sallmyr, A. & Tomkinson, A. E. Repair of DNA double-strand breaks by mammalian alternative end-joining pathways. *Journal of Biological Chemistry* **293**, 10536–10546 (2018).

146. Rhind, N. DNA replication timing: Biochemical mechanisms and biological significance. *BioEssays* 2200097 (2022) doi:10.1002/bies.202200097.

147. Wang, W. *et al.* Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Mol Cell* (2021) doi:10.1016/j.molcel.2021.05.024.

148. Bechhoefer, J. & Rhind, N. Replication timing and its emergence from stochastic processes. *Trends in Genetics* **28**, 374–381 (2012).

149. Dileep, V. & Gilbert, D. M. Single-cell replication profiling to measure stochastic variation in mammalian replication timing. *Nat Commun* **9**, (2018).

150. Donley, N. & Thayer, M. J. DNA replication timing, genome stability and cancer. *Semin Cancer Biol* **23**, 80–89 (2013).

151. Ryba, T. *et al.* Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. *Genome Res* **22**, 1833–1844 (2012).

152. Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat Cell Biol* **16**, 2–9 (2014).

153. Schübeler, D. *et al.* Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. *Nat Genet* **32**, 438–442 (2002).

154. Du, Q. *et al.* DNA methylation is required to maintain both DNA replication timing precision and 3D genome organization integrity. *Cell Rep* **36**, 109722 (2021).

155. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20**, 761–770 (2010).

156. Miura, H. *et al.* Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nat Genet* **51**, 1356–1368 (2019).

157. Peycheva, M. *et al.* DNA replication timing directly regulates the frequency of oncogenic chromosomal translocations. *Science (1979)* **377**, eabj5502 (2022).

158. Stewart-Morgan, K. R., Reverón-Gómez, N. & Groth, A. Transcription Restart Establishes Chromatin Accessibility after DNA Replication. *Mol Cell* **75**, 284–297.e6 (2019).

159. Marchal, C. *et al.* Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat Protoc* **13**, 819–839 (2018).

160. Hiratani, I. *et al.* Global Reorganization of Replication Domains During Embryonic Stem Cell Differentiation. *PLoS Biol* **6**, e245 (2008).

161. Chen, C.-L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20**, 447–457 (2010).

162. Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat Commun* **10**, (2019).

163. Emerson, D. J. *et al.* Cohesin-mediated loop anchors confine the locations of human replication origins. *Nature* **606**, 812–819 (2022).

164. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* **107**, 139–144 (2010).

165. Massey, D. J. & Koren, A. High-throughput analysis of single human cells reveals the complex nature of DNA replication timing control. *Nat Commun* **13**, 2402 (2022).

166. Hulke, M. L., Massey, D. J. & Koren, A. Genomic methods for measuring DNA replication dynamics. *Chromosome Research* **28**, 49–67 (2020).

167. Connolly, C. *et al.* SAF-A promotes origin licensing and replication fork progression to ensure robust DNA replication. *J Cell Sci* **135**, jcs258991 (2022).

168. De, S. & Michor, F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol* **29**, 1103–1108 (2011).

169. Caballero, M., Boos, D. & Koren, A. Cell-type specificity of the human mutation landscape with respect to DNA replication dynamics. *Cell Genomics* 100315 (2023) doi:10.1016/j.xgen.2023.100315.

170. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat Genet* **41**, 393–395 (2009).

171. Weaver, B. A. & Cleveland, D. W. Does aneuploidy cause cancer? *Curr Opin Cell Biol* **18**, 658–667 (2006).

172. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).

173. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).

174. Laros, J. F. J. Demultiplex: FASTA/FASTQ demultiplexer. https://github.com/jfjlaros/demultiplex (2023) doi:10.5281/zenodo.8362958.

175. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/ (2021).

176. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. Trim Galore. https://github.com/FelixKrueger/TrimGalore doi:10.5281/zenodo.5127898.

177. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**, 10 (2011).

178. FastQC. Preprint at https://qubeshub.org/resources/fastqc (2015).

179. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

180. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint* (2013) doi:10.48550/arXiv.1303.3997.

181. Broad Institute. Picard toolkit. https://broadinstitute.github.io/picard/ (2019).

182. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

183. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

184. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

185. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).

186. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

187. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

188. Chollet François. Keras. https://keras.io (2015).

189. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint* (2018) doi:10.48550/arXiv.1802.03426.

190. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *kdd* vol. 96 226–231 (1996).

191. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN Revisited, Revisited. *ACM Transactions on Database Systems* **42**, 1–21 (2017).

192. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res* **12**, 996–1006 (2002).

193. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–7 (2010).

194. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).

195. Zhao, P. A., Sasaki, T. & Gilbert, D. M. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol* **21**, (2020).

196. Stott, F. J. The alternative product from the human CDKN2A locus, p14ARF, participates in a regulatory feedback loop with p53 and MDM2. *EMBO J* **17**, 5001–5014 (1998).

197. Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S. & John, D. Comparative analysis of common alignment tools for single-cell RNA sequencing. *Gigascience* **11**, (2022).

198. Avey, D. *et al.* Single-Cell RNA-Seq Uncovers a Robust Transcriptional Response to Morphine by Glia. *Cell Rep* **24**, 3619-3629.e4 (2018).

199. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–22 (1977).

200. Vlad, G. Sources of errors in flow cytometry. in *Accurate Results in the Clinical Laboratory* 401–422 (Elsevier, 2019). doi:10.1016/B978-0-12-813776-5.00027-3.

201. Laks, E. *et al.* Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* **179**, 1207-1221.e22 (2019).

202. Vikhanskaya, F. *et al.* Cooperation between p53 and hMLH1 in a human colocarcinoma cell line in response to DNA damage. *Clin Cancer Res* **5**, 937–41 (1999).

203. Koi, M. *et al.* Human chromosome 3 corrects mismatch repair deficiency and microsatellite instability and reduces N-methyl-N'-nitro-N-nitrosoguanidine tolerance in colon tumor cells with homozygous hMLH1 mutation. *Cancer Res* **54**, 4308–12 (1994).

204. United States Environmental Protection Agency. *Data Quality Assessment: Statistical Methods for Practitioners*. https://www.epa.gov/sites/default/files/2015-08/documents/g9s-final.pdf (2006).

205. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12**, 519–522 (2015).

206. Hou, Y. *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res* **26**, 304–319 (2016).

207. Han, K. Y. *et al.* SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res* **28**, 75–87 (2017).

208. University of Virginia School of Medicine. FAQs for Cell Sorting. https://med.virginia.edu/flow-cytometry-facility/equipment/faqs-for-cell-sorting/.

209. Lalanne, C. Analysis of DNA copy number variation for the study of cancer cell evolution. (Université de Rennes, Institut Curie, Paris, 2023).

210. Penter, L. *et al.* Mitochondrial DNA Mutations as Natural Barcodes for Lineage Tracing of Murine Tumor Models. *Cancer Res* **83**, 667–672 (2022).

211. Spatz, A., Borg, C. & Feunteun, J. X-Chromosome Genetics and Human Cancer. *Nat Rev Cancer* **4**, 617–629 (2004).

212. Ding, J. *et al.* Gain of miR-151 on chromosome 8q24.3 facilitates tumour cell migration and spreading through downregulating RhoGDIA. *Nat Cell Biol* **12**, 390–399 (2010).

213. Rethoré, M.-O., Rouëssé, J. & Satgé, D. Cancer screening in adults with down syndrome, a proposal. *Eur J Med Genet* **63**, 103783 (2020).

214. Hård, J. *et al.* Long-read whole-genome analysis of human single cells. *Nat Commun* **14**, 5164 (2023).

215. Ying, X. An Overview of Overfitting and its Solutions. *J Phys Conf Ser* **1168**, 022022 (2019).

216. Kernbach, J. M. & Staartjes, V. E. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II—Generalization and Overfitting. in 15–21 (2022). doi:10.1007/978-3-030-85292-4_3.

6. Data and code availability

6.1. Single-cell WGS/CNV data

Dataset	Accession code	Cell-type
Takahashi2019[98]	GSE108556	mESCs
		Day-7
		Embryonal carcinoma
		hTERT-RPE1 (retinal)
Laks2019[201]	EGAS00001003190	GM18507 (lymphocyte)
		T-47D (breast cancer)
		184-hTERT (breast epithelial)
		TOV2295 (HGSOC)
		OV2295 (HGSOC)
		HeLa (cervical cancer)
		PDX Breast Cancer
		FNA Breast Cancer (patient samples)
		Follicular lymphomas
Funnell2022[56]	ZENODO.6998936	CNV data on hg19 (partially from Laks2019)
Gnan2022[62]	GSE186173	MCF-7 (breast cancer)
		Hela-S3 (cervical cancer)
		JEFF (lymphocytes)
Massey2022[165]	PRJNA770772	GM12878 (lymphocyte)
		GM12891 (lymphocyte)
		GM12892 (lymphocyte)
		H1 (hESC)
		H7 (hESC)
		H9 (hESC)
		HCT-116 (colon cancer)
		RKO (colon cancer)
		MCF-7 (breast cancer)
Connolly2022[167]	E-MTAB-10234	hTERT-RPE1 (retinal)
Du2021[154]	GSE158009	HCT-116 (WT colorectal cancer)
		HCT-116 (DKO1 colorectal cancer)
Minussi2021[53]	PRJNA629885	TN[1-8] (8 TNBC patient samples)
		MDA-MB-231 (breast cancer)
		MDA-MB-453 (breast cancer)
		MDA-MB-157 (breast cancer)
		BT-20 (breast cancer)

6.2. ATAC data

Dataset	Accession code	Cell-type
Granja2021[70]	GSE162690	Bulk ATAC
		scATAC

6.3. Other files and code

File(s)	Link
10X barcode whitelist	https://github.com/TheKorenLab/Single-cell-replication-timing/blob/main/align/10x_barcode_whitelist.txt
Custom scripts and code	https://github.com/josephides/PhD-scripts https://github.com/CL-CHEN-Lab/MnM/tree/main/scripts_publication
hg38 and mm10 reference genomes	https://support.illumina.com/sequencing/sequencing_software/igenome.html
Modified version of Kronos scRT	https://github.com/josephides/Kronos_scRT
Blacklists	https://github.com/Boyle-Lab/Blacklist
Bulk MCF-7 RT	GSE34399
Bulk HCT-116 RT	GSE158011
Liftover chains	https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver

7. Manuscripts

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-30043-x OPEN

Kronos scRT: a uniform framework for single-cell replication timing analysis

Stefano Gnan[1], Joseph M. Josephides[1], Xia Wu[1,4], Manuela Spagnuolo[1], Dalila Saulebekova[1], Mylène Bohec[2], Marie Dumont[3], Laura G. Baudrin[2], Daniele Fachinetti[3], Sylvain Baulande[2] & Chun-Long Chen[1✉]

Mammalian genomes are replicated in a cell type-specific order and in coordination with transcription and chromatin organization. Currently, single-cell replication studies require individual processing of sorted cells, yielding a limited number (<100) of cells. Here, we develop Kronos scRT, a software for single-cell Replication Timing (scRT) analysis. Kronos scRT does not require a specific platform or cell sorting, which allows investigating large datasets obtained from asynchronous cells. By applying our tool to published data as well as droplet-based single-cell whole-genome sequencing data generated in this study, we exploit scRT from thousands of cells for different mouse and human cell lines. Our results demonstrate that although genomic regions are frequently replicated around their population average RT, replication can occur stochastically throughout S phase. Altogether, Kronos scRT allows fast and comprehensive investigations of the RT programme at the single-cell resolution for both homogeneous and heterogeneous cell populations.

[1] Institut Curie, PSL Research University, CNRS UMR3244, Dynamics of Genetic Information, Sorbonne Université, 75005 Paris, France. [2] Institut Curie, Genomics of Excellence (ICGex) Platform, PSL Research University, 75005 Paris, France. [3] Institut Curie, PSL Research University, CNRS UMR144, Cell Biology and Cancer, 75005 Paris, France. [4] Present address: Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong 510080, China. ✉email: chunlong.chen@curie.fr

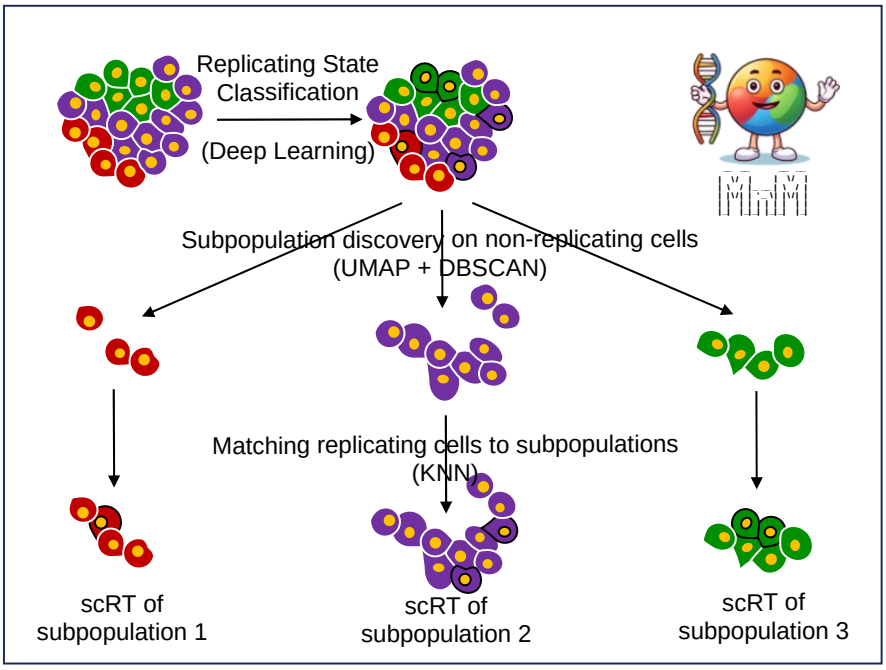
MnM: a machine learning approach to detect replication states and genomic subpopulations for single-cell DNA replication timing disentanglement

Joseph M. Josephides[1], Chun-Long Chen[1,*]

[1] Institut Curie, PSL Research University, CNRS UMR3244, Dynamics of Genetic Information, Sorbonne Université, 75005 Paris, France.

* Email: chunlong.chen@curie.fr

GRAPHICAL ABSTRACT



8. Supplementary data

8.1. Algorithm/Pseudocode of the main steps of MnM.

1. Read the reference genome chromosome sizes.
2. Divide the reference genome into bins.
3. Read the CNV file.
4. Divide the CNV data into bins in a BED format.
5. Convert CNV data to a 2D matrix (regions, single-cells).
6. Perform K-nearest neighbors (KNN) imputation on the binned CNV data to fill in missing values.
7. If user requests replication states and the reference genome is hg38:
    a. Load the deep learning model (25, 100 or 500kb accordingly).
    b. Create a copy of the matrix to be used until step 7e.
    c. Extract regions considered by the model.
    d. Use linear interpolation to fill in any missing regions from the data.
    e. Provide replication states from the model.
8. If user requests subpopulations detection:
    a. If replication states detected, retain only non-replicating cells for steps 7a-e.
    b. Ignore allosomes until step 7g.
    c. Perform UMAP with 2 dimensions.
    d. Apply DBSCAN to detect subpopulations.
    e. Repeat the UMAP/DBSCAN step 6 times with different random seeds.
    f. If the predominant number of subpopulations is not issued with the first iteration (7d), use the first random seed that does.
    g. If replication states detected:
        i. Perform UMAP of all regions of all cells with 10 dimensions.
        ii. Match S phase cells individually to their subpopulation with KNN.

8.2. Supplementary Tables

8.2.1. Supplementary Table 1: Percentage of missing values per scCNV matrix

Dataset	Cell-type	Missing Values	Percentage of missing values
Du2021	HCT-116 DKO1	82572	0.42%
Du2021	HCT-116 WT	77757	0.37%
Gnan2022	HeLa	139581	0.84%
Gnan2022	JEFF	232765	0.68%
Gnan2022	MCF-7	734532	1.11%
Laks2019	FNA-sample_SA1135	88581	0.91%
Massey2022	GM12878	106746	0.07%

8.2.2. Supplementary Table 2: scWGS data used for deep learning replication state classifier

Dataset	CellType	Phase	CellCount
Massey2022	GM12878	G1	1193
		S	1655
Du2021	HCT-116 WT	G1	49
		S	434
	HCT-116 DKO1	G1	40
		S	338
Gnan2022	HeLa	G1	224
		S	299
	JEFF	G1	132
		S	998
	MCF-7	G1	632
		S	1512
Takahashi2019	hTERT-RPE1	G1	3
		S	14

8.2.3. Supplementary Table 3: scWGS samples used in this project

Source	Sample	Pre-QC Cell Count	CellCount	% Cells QC Loss	Cell Type	Cell Type Description	Sample	Median Coverage (reads/Mb)
Du2021	HCT116 DKO1	669	668	0.15%	HCT-116	Colon cancer	Cell-line	1 882.02
Du2021	HCT116 WT	713	713	0.00%	HCT-116	Colon cancer	Cell-line	1 176.35
Gnan2022	HeLa	752	752	0.00%	HeLa	Cervical carcinoma	Cell-line	788.81
Gnan2022	JEFF	1 461	1 455	0.41%	JEFF	Lymphocyte	Cell-line	425.48
Gnan2022	MCF-7	2 768	2 768	0.00%	MCF-7	Breast cancer	Cell-line	720.07
Laks2019	184-hTERT_SA039	5 290	5 285	0.09%	184-hTERT	Mammary epithelial	Cell-line	344.11
Laks2019	184-hTERT_SA1101	3 358	2 788	16.97 %	184-hTERT	Mammary epithelial	Cell-line	272.89
Laks2019	184-hTERT_SA906	12 707	10 695	15.83 %	184-hTERT	Mammary epithelial	Cell-line	475.97
Laks2019	ERpos-PDX_SA532X2XB00147	755	441	41.59 %	PDX	Breast cancer	PDX	688.52
Laks2019	ERpos-PDX_SA532X4XB00273	635	498	21.57 %	PDX	Breast cancer	PDX	404.47
Laks2019	ERpos-PDX_SA532X8XB01398	589	371	37.01 %	PDX	Breast cancer	PDX	716.82
Laks2019	ERpos-PDX_SA611X3XB00821	531	436	17.89 %	PDX	Breast cancer	PDX	417.50
Laks2019	ERpos-PDX_SA995X5XB01910	465	152	67.31 %	PDX	Breast cancer	PDX	571.08
Laks2019	FNA-sample_SA1135	800	473	40.88 %	Tumour	Breast cancer	Patient tumour	982.93
Laks2019	FNA-sample_SA1137	88	37	57.95 %	Tumour	Breast cancer	Patient tumour	105.54
Laks2019	GM18507_SA928	8 218	7 461	9.21%	GM18507	Lymphocyte	Cell-line	678.99
Laks2019	HeLa_SA1087	656	601	8.38%	HeLa	Cervical carcinoma	Cell-line	416.94
Laks2019	HGSOC-OV2295_SA1090	741	696	6.07%	OV2295	HGSC	Cell-line	596.04
Laks2019	HGSOC-OV2295_SA922	1 085	368	66.08 %	OV2295	HGSC	Cell-line	849.26
Laks2019	HGSOC-TOV2295_SA921	1 118	371	66.82 %	TOV2295	HGSC	Cell-line	868.14
Laks2019	Lymphoma_SA1088	648	530	18.21 %	Lymphoma	Follicular lymphoma	Patient tumour	459.55
Laks2019	Lymphoma_SA1089	375	346	7.73%	Lymphoma	Follicular lymphoma	Patient tumour	620.18
Laks2019	T-47D_SA1044	1 436	1 332	7.24%	T-47	Breast cancer	Cell-line	1 021.92

Laks2019	TNBC-PDX_SA501X11XB00529	1 063	954	10.25 %	PDX	Breast cancer	PDX	233.80
Laks2019	TNBC-PDX_SA501X2XB00096	488	451	7.58%	PDX	Breast cancer	PDX	1 335.05
Laks2019	TNBC-PDX_SA501X2XB00097	492	37	92.48 %	PDX	Breast cancer	PDX	733.97
Laks2019	TNBC-PDX_SA501X5XB00877	615	270	56.10 %	PDX	Breast cancer	PDX	364.75
Laks2019	TNBC-PDX_SA501X6XB00969	636	355	44.18 %	PDX	Breast cancer	PDX	194.31
Laks2019	TNBC-PDX_SA535X5XB00517	928	444	52.16 %	PDX	Breast cancer	PDX	447.63
Laks2019	TNBC-PDX_SA535X8XB01043	1 072	341	68.19 %	PDX	Breast cancer	PDX	506.48
Laks2019	TNBC-PDX_SA604X6XB01979	968	476	50.83 %	PDX	Breast cancer	PDX	487.41
Laks2019	TNBC-PDX_SA609X3XB01584	480	212	55.83 %	PDX	Breast cancer	PDX	335.83
Laks2019	TNBC-PDX_SA609X4XB01721	606	392	35.31 %	PDX	Breast cancer	PDX	388.63
Laks2019	TNBC-PDX_SA609X5XB01844	561	396	29.41 %	PDX	Breast cancer	PDX	226.28
Laks2019	TNBC-PDX_SA609X6XB01898	626	410	34.50 %	PDX	Breast cancer	PDX	188.30
Laks2019	TNBC-PDX_SA609X6XB01899	635	499	21.42 %	PDX	Breast cancer	PDX	286.65
Laks2019	TNBC-PDX_SA609X7XB02184	844	634	24.88 %	PDX	Breast cancer	PDX	772.79
Massey2022	GM12878	8 947	7 942	11.23 %	GM12878	Lymphocyte	Cell-line	288.45
Massey2022	GM12891	2 742	2 621	4.41%	GM12891	Lymphocyte	Cell-line	170.15
Massey2022	GM12892	2 596	2 450	5.62%	GM12892	Lymphocyte	Cell-line	165.07
Massey2022	H1	2 370	1 216	48.69 %	H1	hESC	ESC	68.59
Massey2022	H7	1 923	1 780	7.44%	H7	hESC	ESC	234.20
Massey2022	H9	915	888	2.95%	H9	hESC	ESC	101.70
Massey2022	HCT-116	1 555	1 264	18.71 %	HCT-116	Colon cancer	Cell-line	73.84
Massey2022	MCF-7	1 337	982	26.55 %	MCF-7	Breast cancer	Cell-line	175.78
Massey2022	RKO	2 315	2 149	7.17%	RKO	Colon cancer	Cell-line	124.94
Minussi2021	BT20	1 229	1 228	0.08%	BT20	Breast cancer	Cell-line	215.55
Minussi2021	MDA-MB-157	1 210	1 210	0.00%	MDA-MB-157	Breast cancer	Cell-line	243.43
Minussi2021	MDA-MB-231	2 710	2 710	0.00%	MDA-MB-231	Breast cancer	Cell-line	252.10
Minussi2021	MDA-MB-453	1 260	1 260	0.00%	MDA-MB-453	Breast cancer	Cell-line	235.28

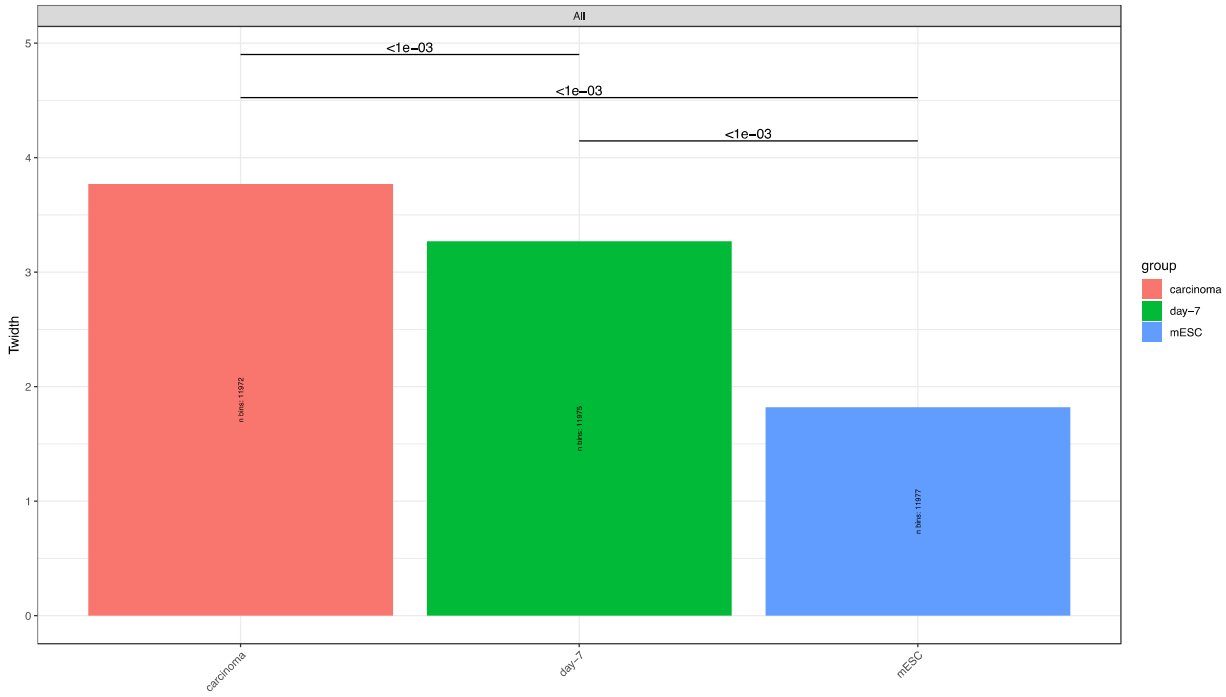
Minussi2021	TN1	1 978	1 977	0.05%	Tumour	Breast cancer	Patient tumour	362.75
Minussi2021	TN2	1 024	1 023	0.10%	Tumour	Breast cancer	Patient tumour	542.82
Minussi2021	TN3	2 192	2 190	0.09%	Tumour	Breast cancer	Patient tumour	260.54
Minussi2021	TN4	1 301	1 301	0.00%	Tumour	Breast cancer	Patient tumour	225.59
Minussi2021	TN5	1 238	1 238	0.00%	Tumour	Breast cancer	Patient tumour	229.75
Minussi2021	TN6	1 205	1 205	0.00%	Tumour	Breast cancer	Patient tumour	218.43
Minussi2021	TN7	907	907	0.00%	Tumour	Breast cancer	Patient tumour	214.02
Minussi2021	TN8	1 224	1 224	0.00%	Tumour	Breast cancer	Patient tumour	222.53
Connolly2022	hTERT-RPE1	63	63	0.00%	hTERT-RPE1	Retinal pigment epithelial	Cell-line	824.76
Takahashi2019	hTERT-RPE1	17	17	0.00%	hTERT-RPE1	Retinal pigment epithelial	Cell-line	683.32
Funnell2022*	Pre-processed_CNV_hg19_SA039	878	878	0.00%	184-hTert	Mammary epithelial	Cell-line	
Funnell2022*	Pre-processed_CNV_hg19_SA1054	382	382	0.00%	184-hTert	Mammary epithelial	Cell-line	
Funnell2022*	Pre-processed_CNV_hg19_SA1055	391	391	0.00%	184-hTert	Mammary epithelial	Cell-line	
Funnell2022*	Pre-processed_CNV_hg19_SA1056	496	496	0.00%	184-hTert	Mammary epithelial	Cell-line	
Funnell2022*	Pre-processed_CNV_hg19_SA1188	2 003	2 003	0.00%	184-hTert	Mammary epithelial	Cell-line	
Funnell2022*	Pre-processed_CNV_hg19_SA1292	404	404	0.00%	184-hTert	Mammary epithelial	Cell-line	
Funnell2022*	Pre-processed_CNV_hg19_SA906a	3 711	3 711	0.00%	184-hTert	Mammary epithelial	Cell-line	
Funnell2022*	Pre-processed_CNV_hg19_SA906b	5 716	5 716	0.00%	184-hTert	Mammary epithelial	Cell-line	
Funnell2022*	Pre-processed_CNV_hg19_DG1134	133	133	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_DG1197	115	115	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1049	1 283	1 283	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1091	506	506	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1096	802	802	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1162	254	254	0.00%	HGSC	HGSC	Patient tumour	

Funnell2022*	Pre-processed_CNV_hg19_SA1180	774	774	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1182	214	214	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1050	990	990	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1051	892	892	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1052	556	556	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1053	825	825	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1181	296	296	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1184	621	621	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1047	347	347	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA1093	346	346	0.00%	HGSC	HGSC	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA530	324	324	0.00%	TNBC	Breast cancer	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA604	2 139	2 139	0.00%	TNBC	Breast cancer	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA609	6 033	6 033	0.00%	TNBC	Breast cancer	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA610	268	268	0.00%	TNBC	Breast cancer	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA501	2 473	2 473	0.00%	TNBC	Breast cancer	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA535	1 801	1 801	0.00%	TNBC	Breast cancer	Patient tumour	
Funnell2022*	Pre-processed_CNV_hg19_SA605	65	65	0.00%	TNBC	Breast cancer	Patient tumour	
TOTAL		**134 168**	**119 991**					

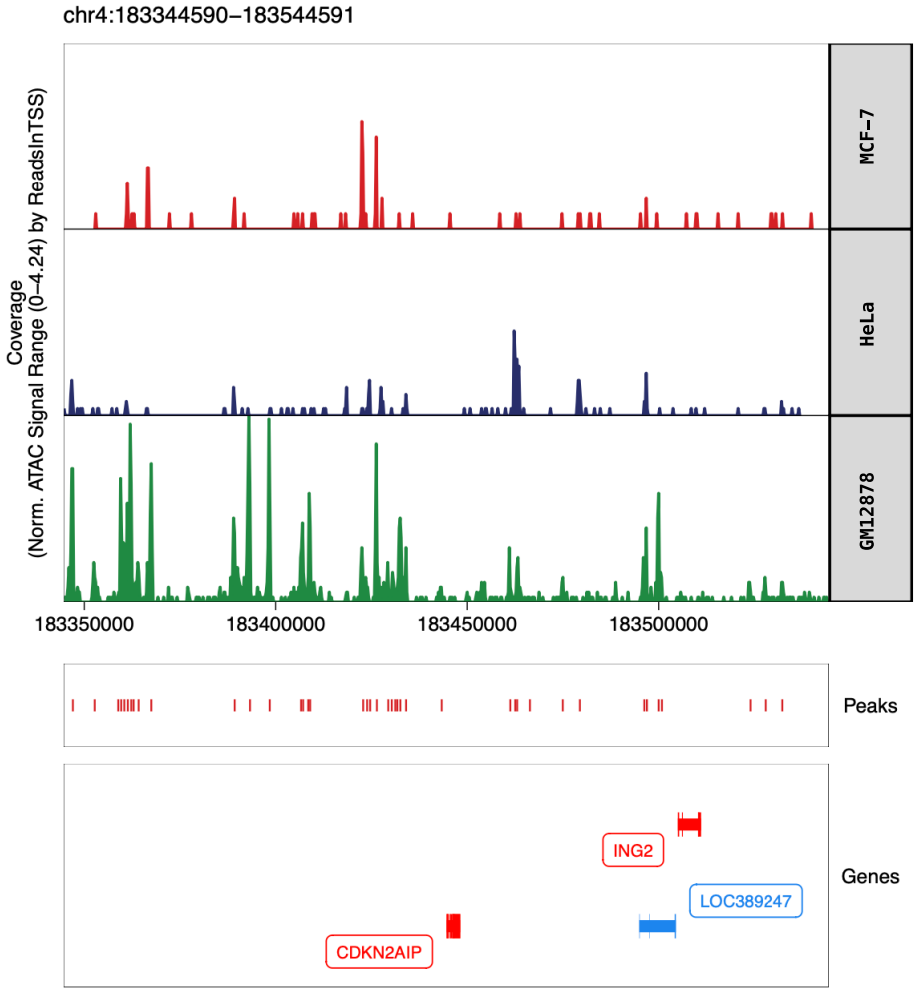
*Duplicates from Laks2019 are not included in the cell count.

8.3. Supplementary Figures

8.3.1. Mouse Twidths


All
<1e−03
<1e−03
<1e−03
Twidth
0
1
2
3
4
5
n bins: 11972
n bins: 11975
n bins: 11977
carcinoma
day−7
mESC
group
carcinoma
day−7
mESC

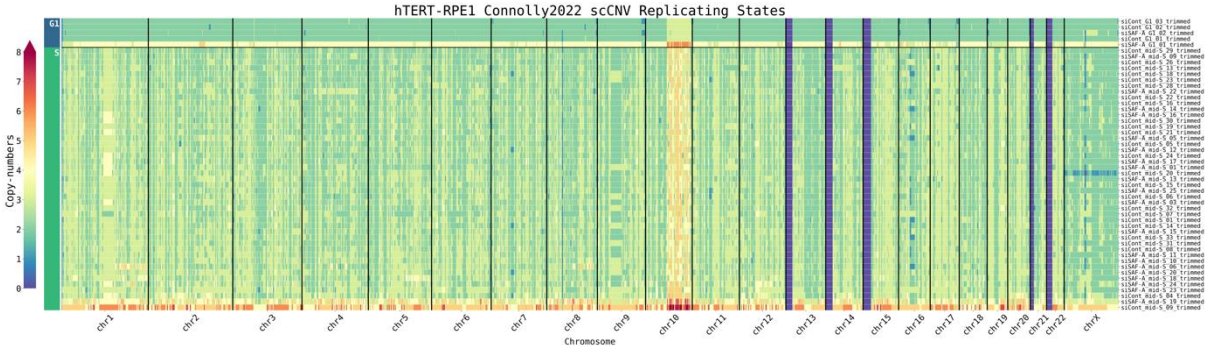
8.3.2. CDKN2AIP chromatin state from scATAC data



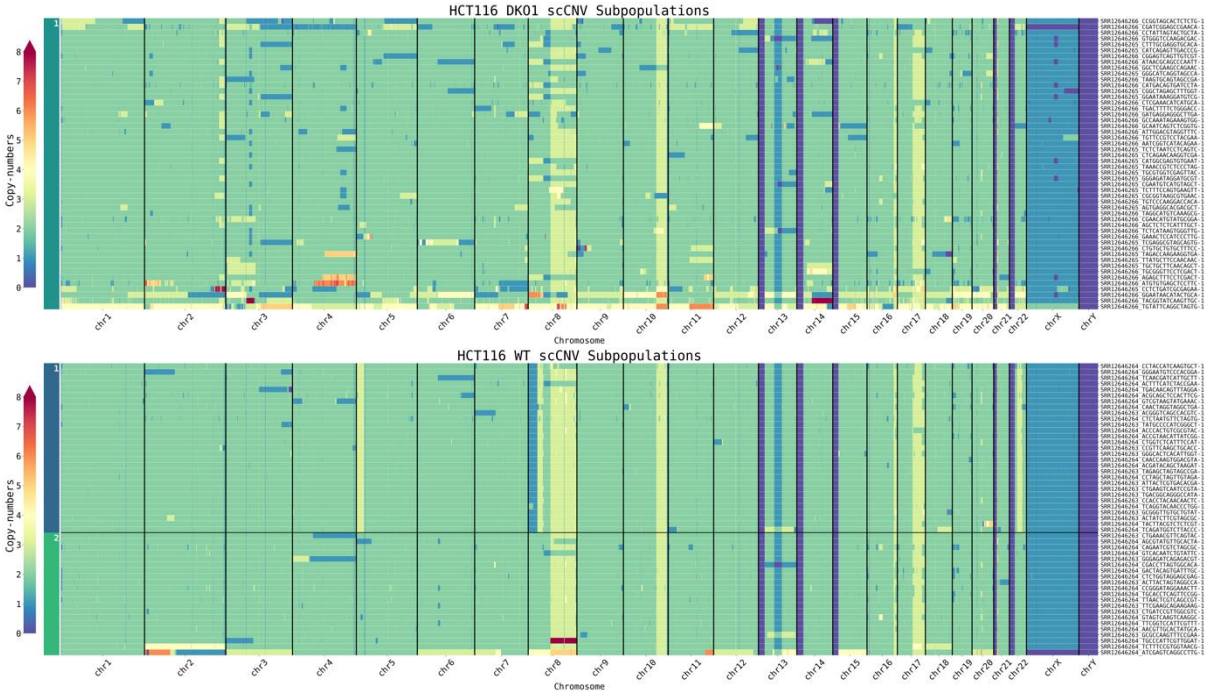
8.3.3. Genome-wide single-cell copy-numbers of non-replicating human samples

Genome-wide scCNV profiles of a maximum of 50 randomly selected cells for Connolly2022 (A), Du2021 (B), Gnan2022 (C), Laks2019 (D), Massey2022 (E), Minussi2021 (F), Takahashi2019 (G) and Funnell2022 (H) split by cell-phase (A,G) or subpopulation (B-F,H).

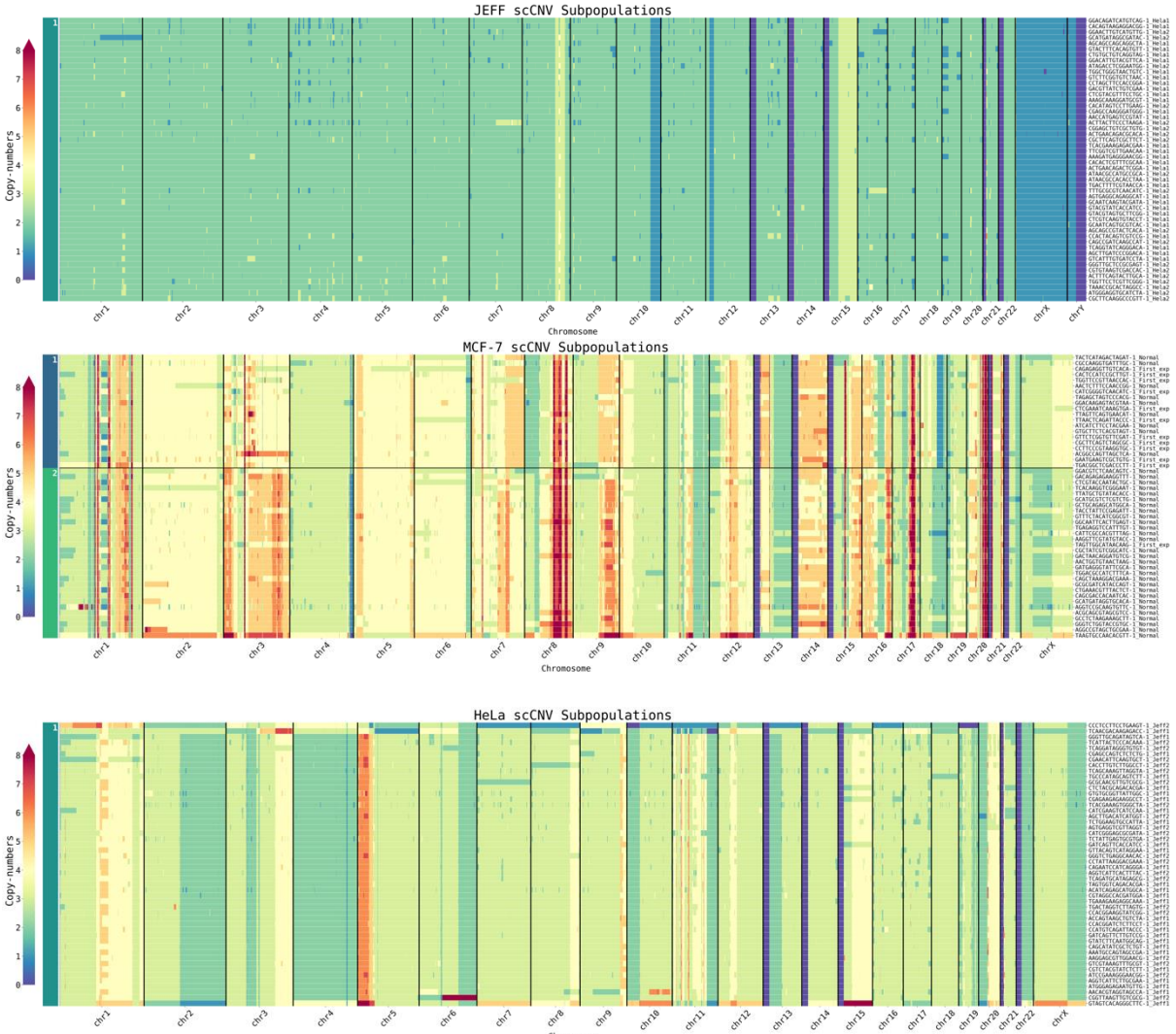
A



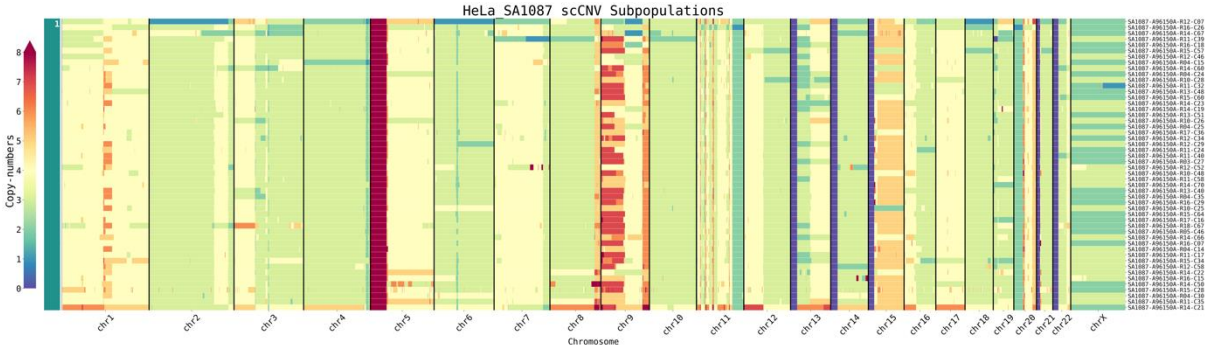
B

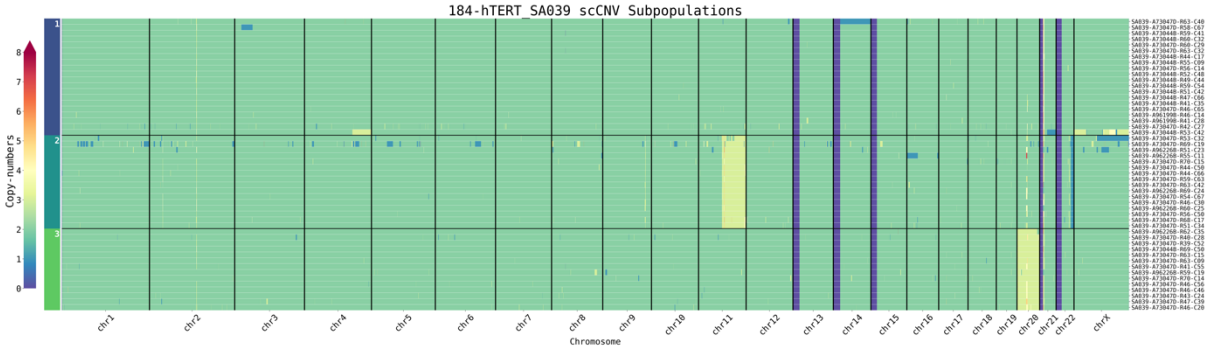


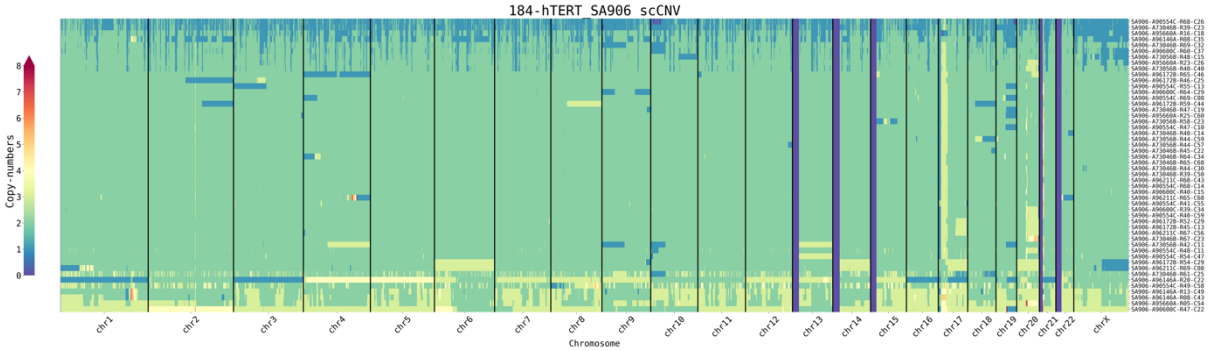
C

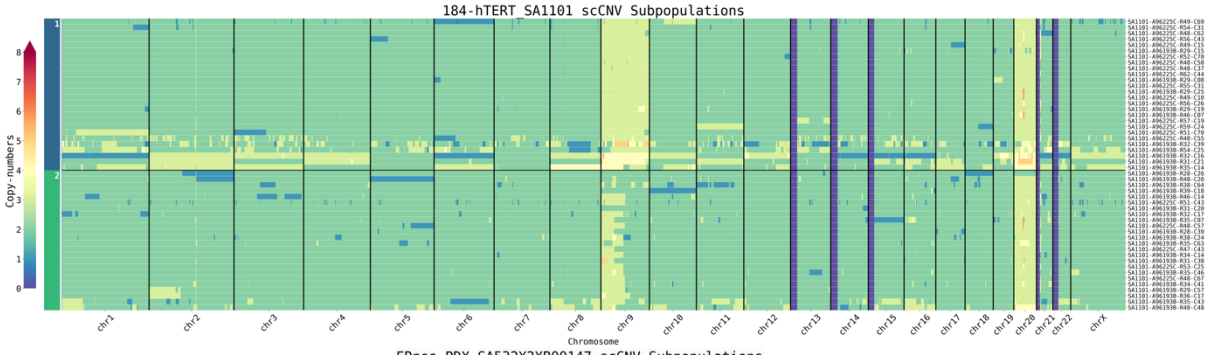

JEFF scCNV Subpopulations
MCF-7 scCNV Subpopulations
HeLa scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX

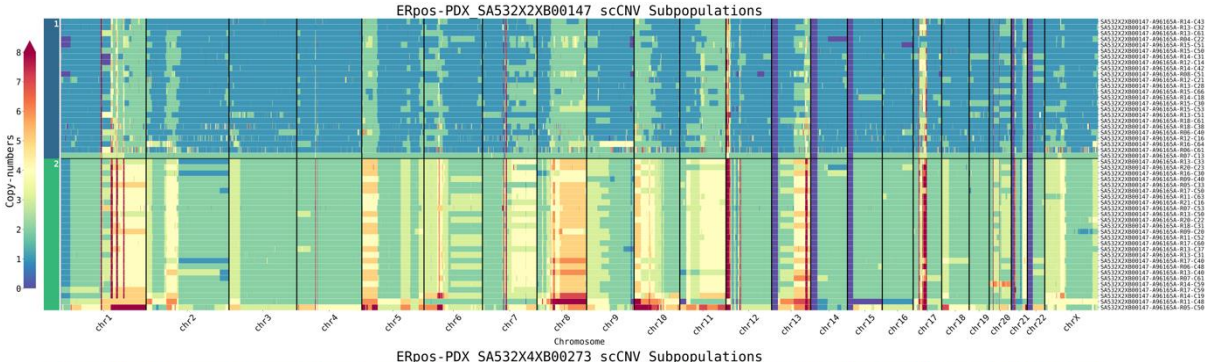
D

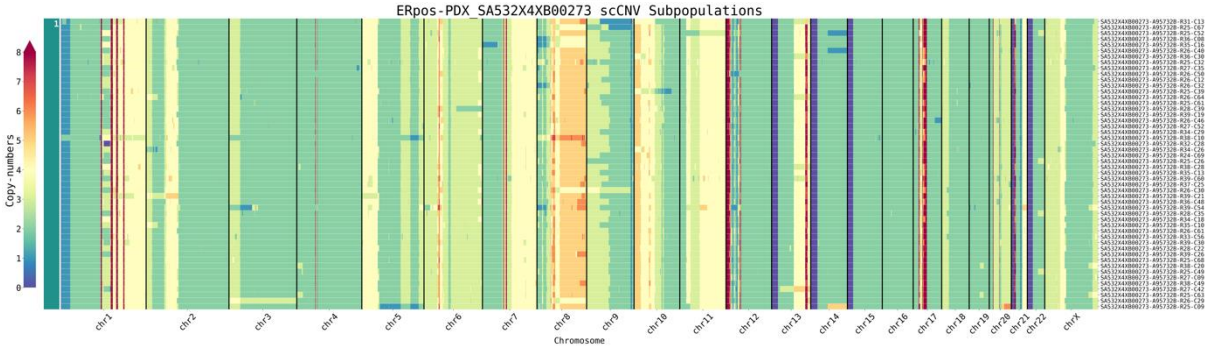

HeLa_SA1087 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX

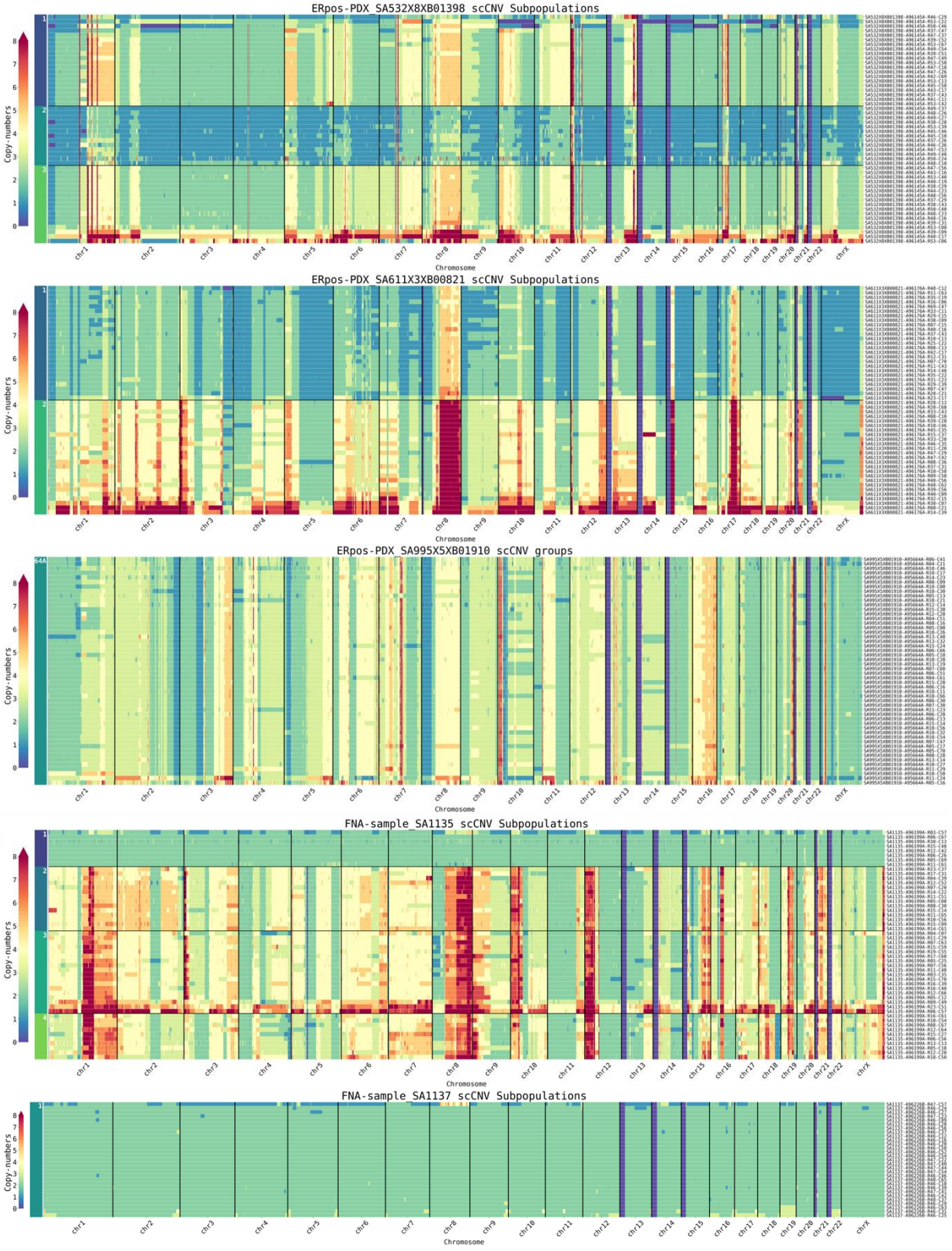

184-hTERT_SA039 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

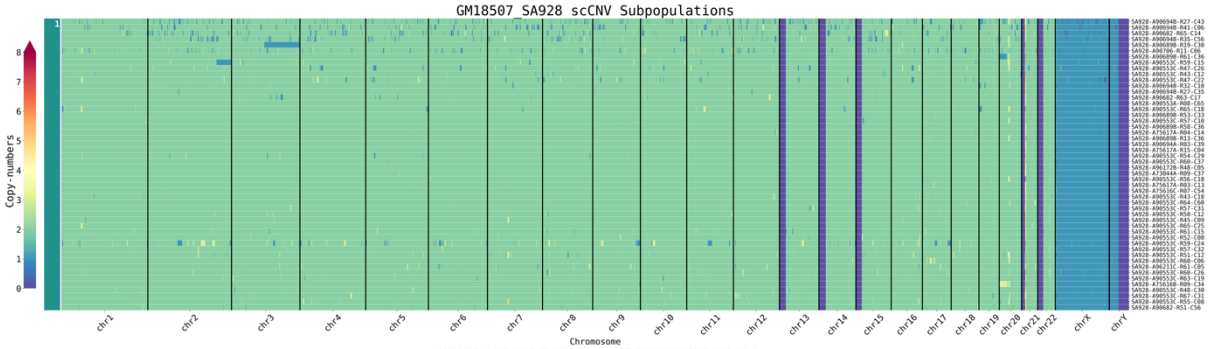

184-hTERT_SA906 scCNV
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

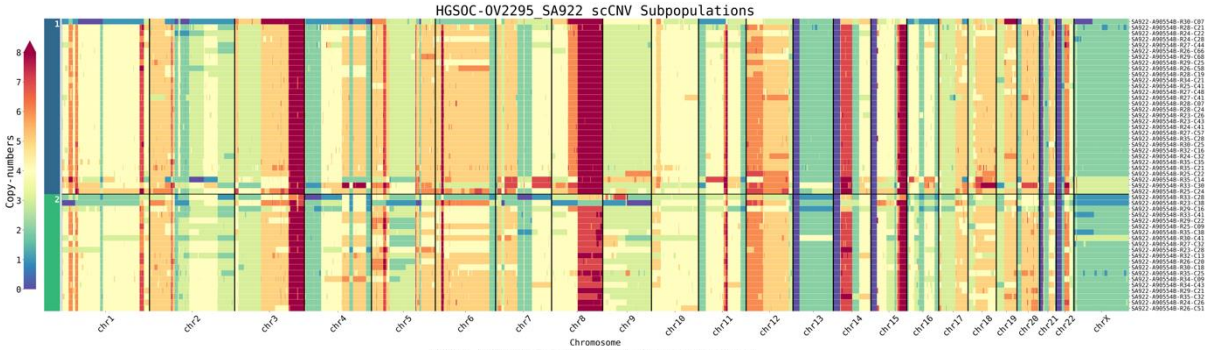

184-hTERT_SA1101 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

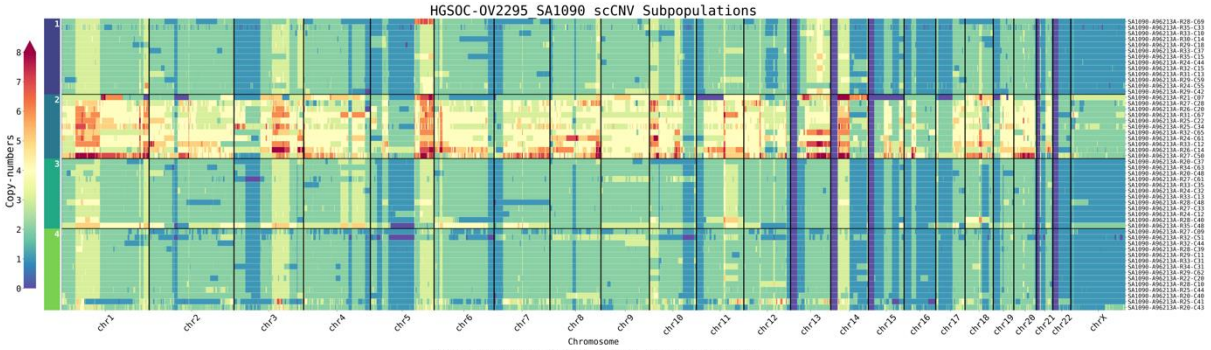

ERpos-PDX_SA532X2XB00147 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

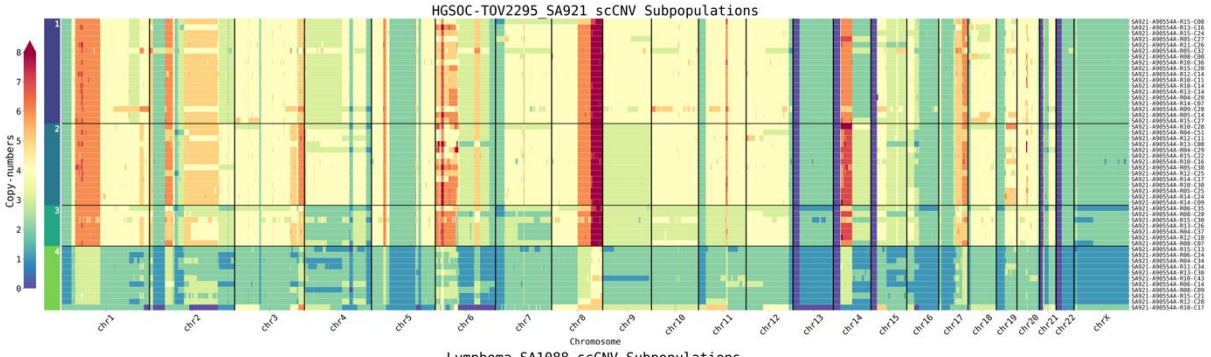

ERpos-PDX_SA532X4XB00273 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

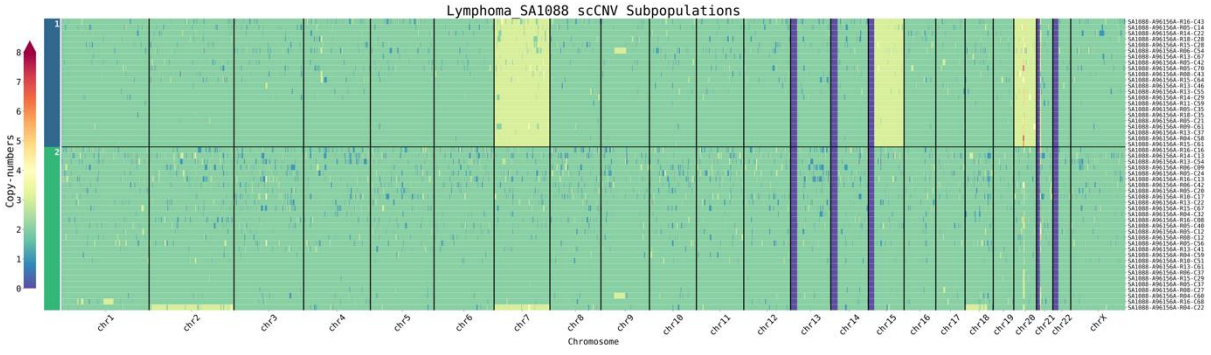

ERpos-PDX_SA532X8XB01398 scCNV Subpopulations
ERpos-PDX_SA611X3XB00821 scCNV Subpopulations
ERpos-PDX_SA995X5XB01910 scCNV groups
FNA-sample_SA1135 scCNV Subpopulations
FNA-sample_SA1137 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX

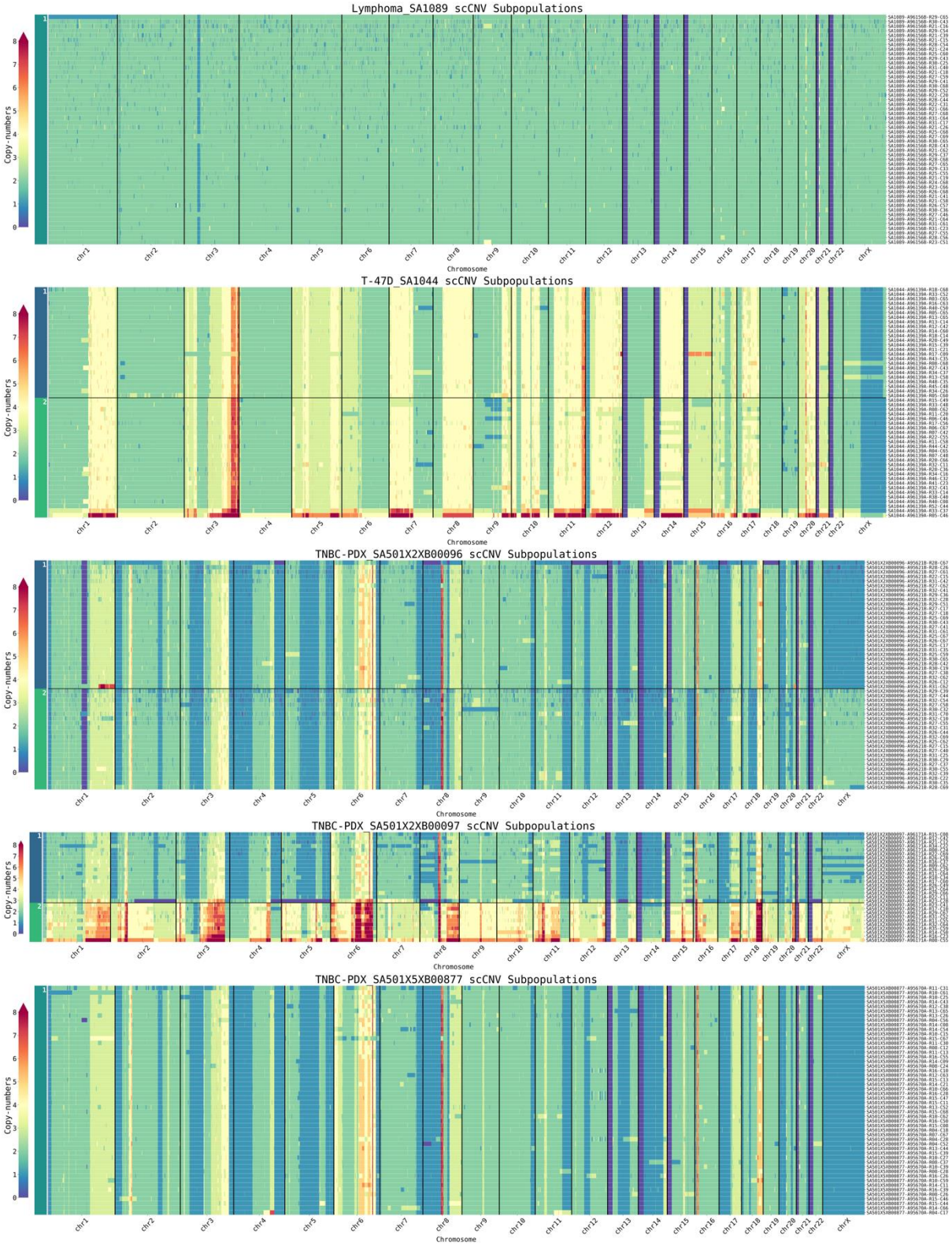

GM18507_SA928 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX chrY

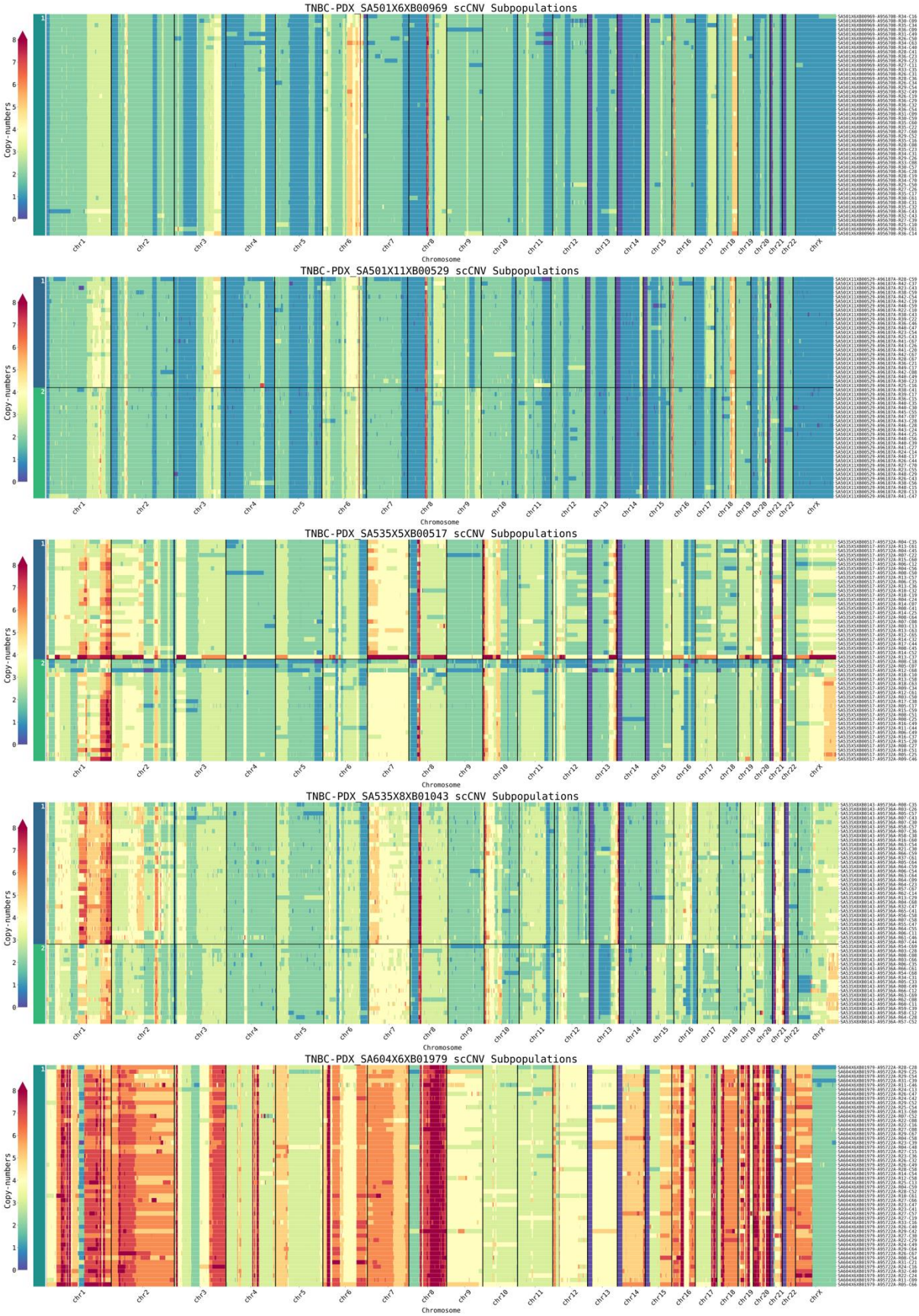

HGSOC-OV2295_SA922 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

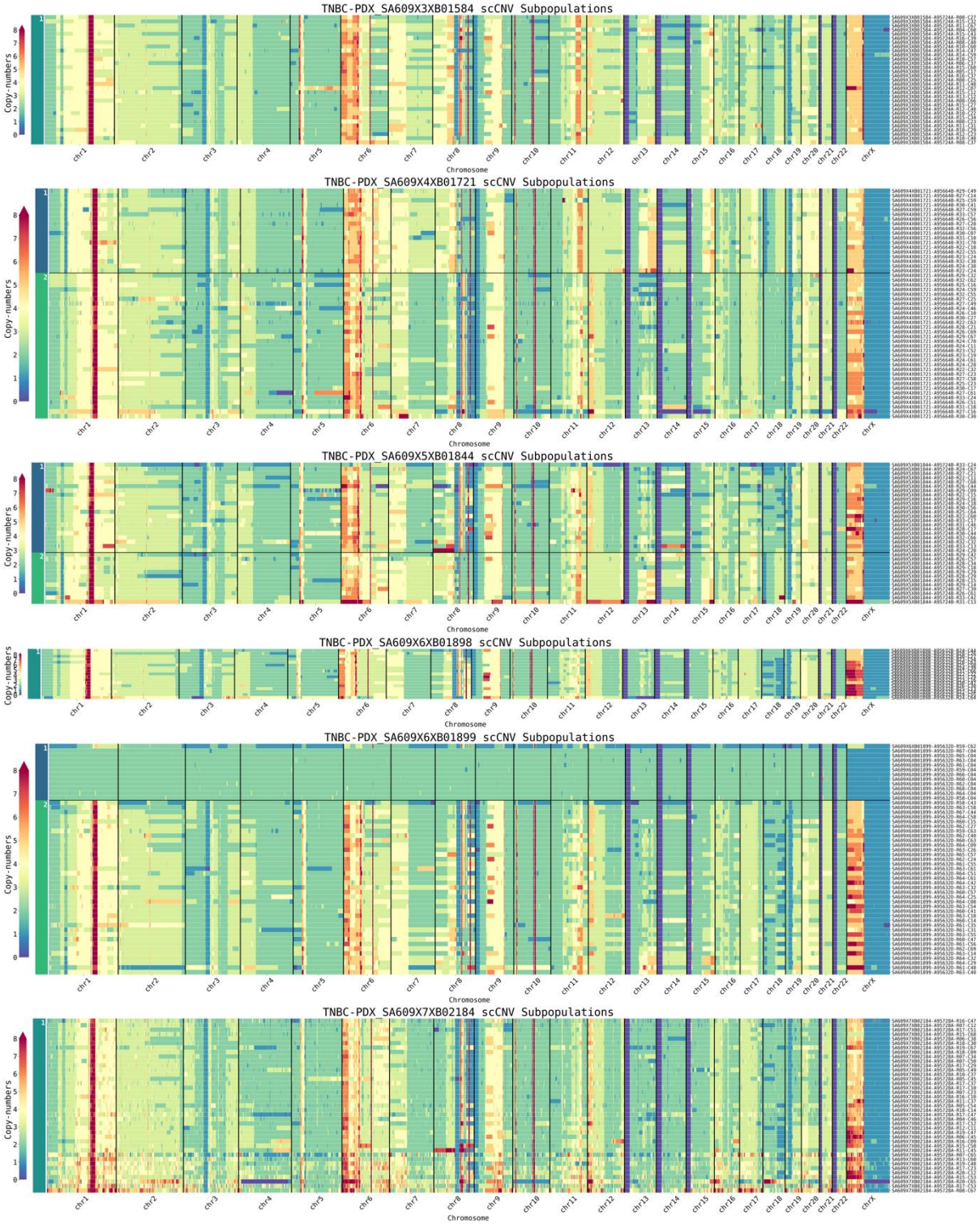

HGSOC-OV2295_SA1090 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX


HGSOC-TOV2295_SA921 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

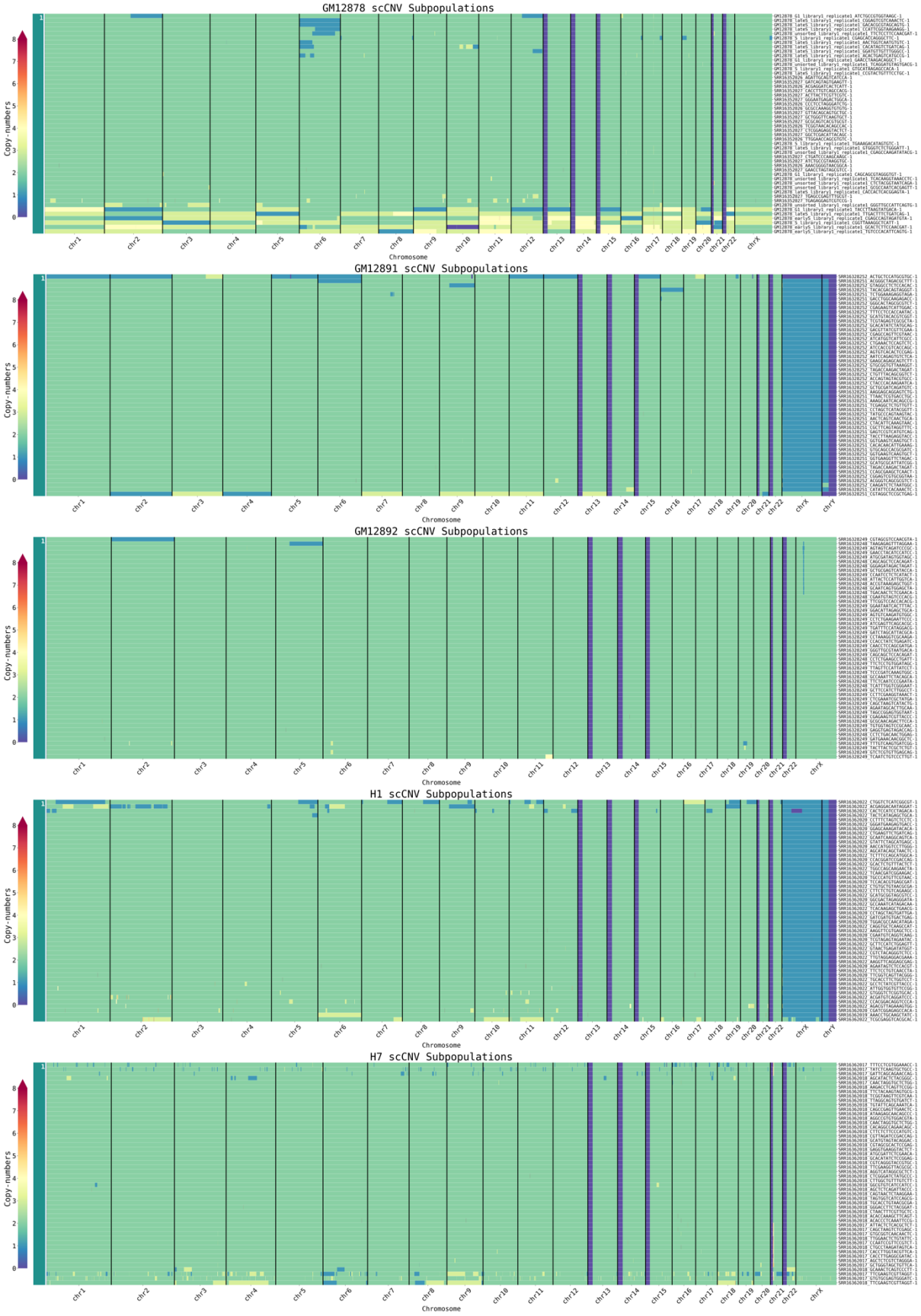

Lymphoma_SA1088 scCNV Subpopulations
Copy-numbers
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

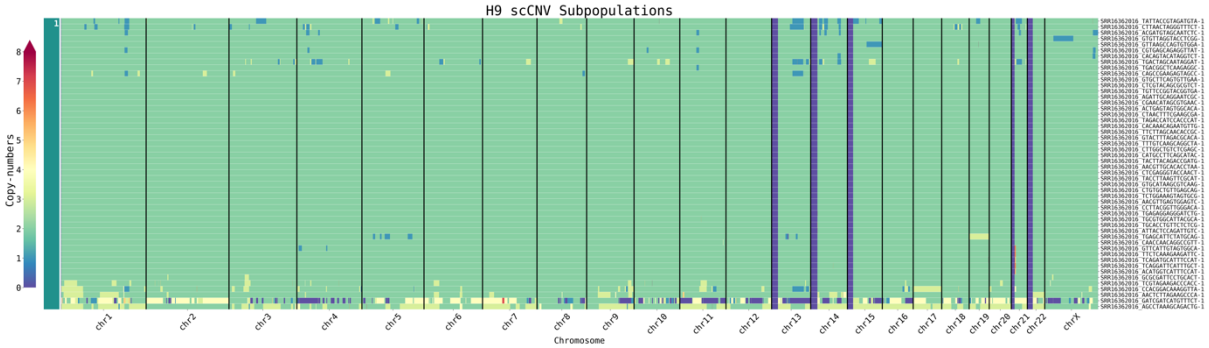

Lymphoma_SA1089 scCNV Subpopulations
T-47D_SA1044 scCNV Subpopulations
TNBC-PDX_SA501X2XB00096 scCNV Subpopulations
TNBC-PDX_SA501X2XB00097 scCNV Subpopulations
TNBC-PDX_SA501X5XB00877 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX

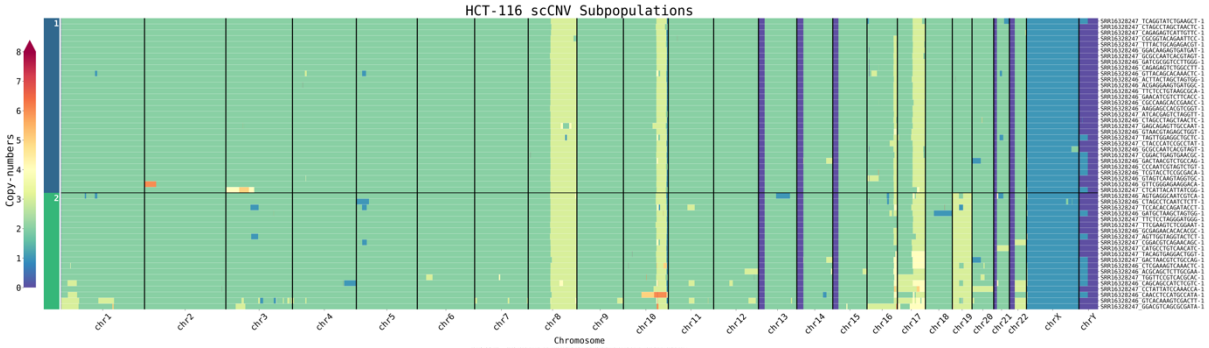

TNBC-PDX_SA501X6XB00969 scCNV Subpopulations
TNBC-PDX_SA501X11XB00529 scCNV Subpopulations
TNBC-PDX_SA535X5XB00517 scCNV Subpopulations
TNBC-PDX_SA535X8XB01043 scCNV Subpopulations
TNBC-PDX_SA604X6XB01979 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX

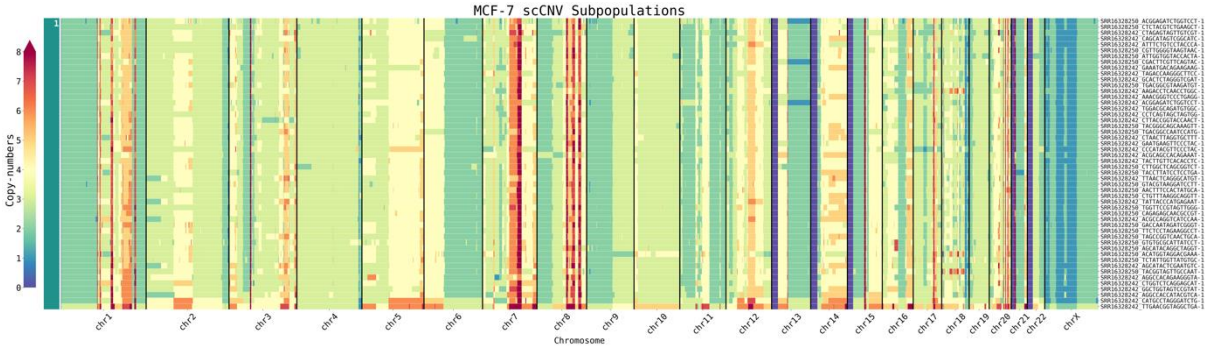

TNBC-PDX_SA609X3XB01584 scCNV Subpopulations
TNBC-PDX_SA609X4XB01721 scCNV Subpopulations
TNBC-PDX_SA609X5XB01844 scCNV Subpopulations
TNBC-PDX_SA609X6XB01898 scCNV Subpopulations
TNBC-PDX_SA609X6XB01899 scCNV Subpopulations
TNBC-PDX_SA609X7XB02184 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX

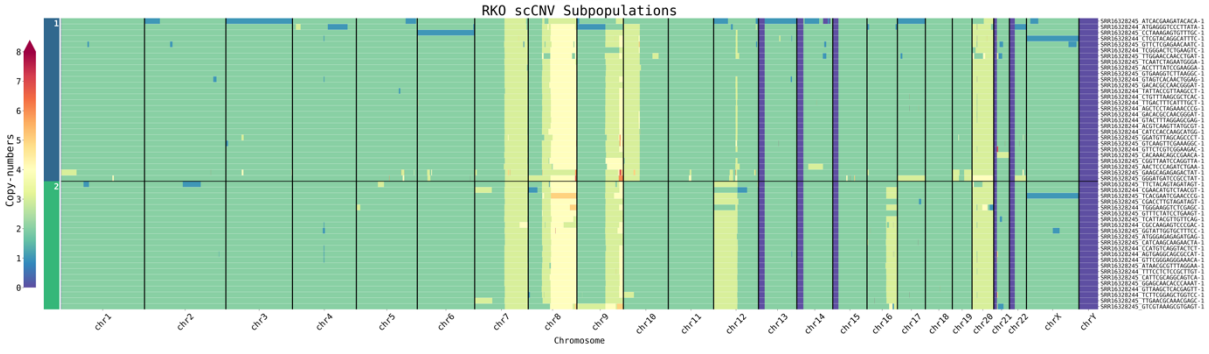
E


GM12878 scCNV Subpopulations
GM12891 scCNV Subpopulations
GM12892 scCNV Subpopulations
H1 scCNV Subpopulations
H7 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY

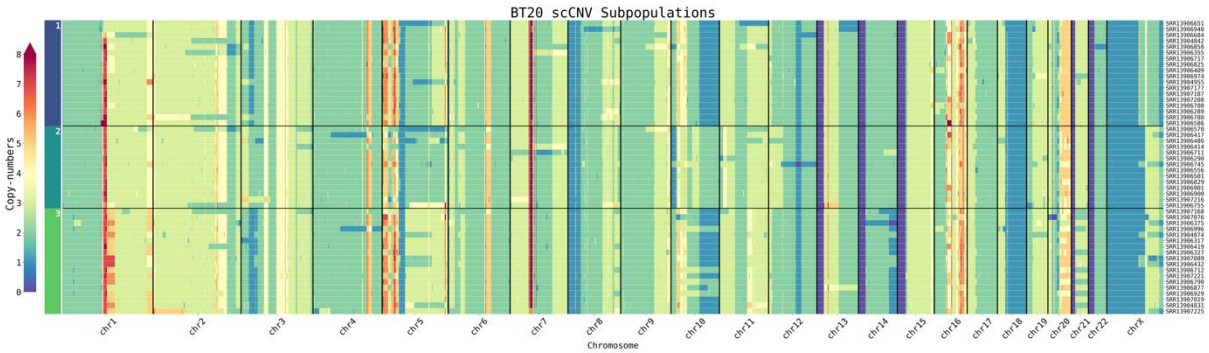

H9 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY

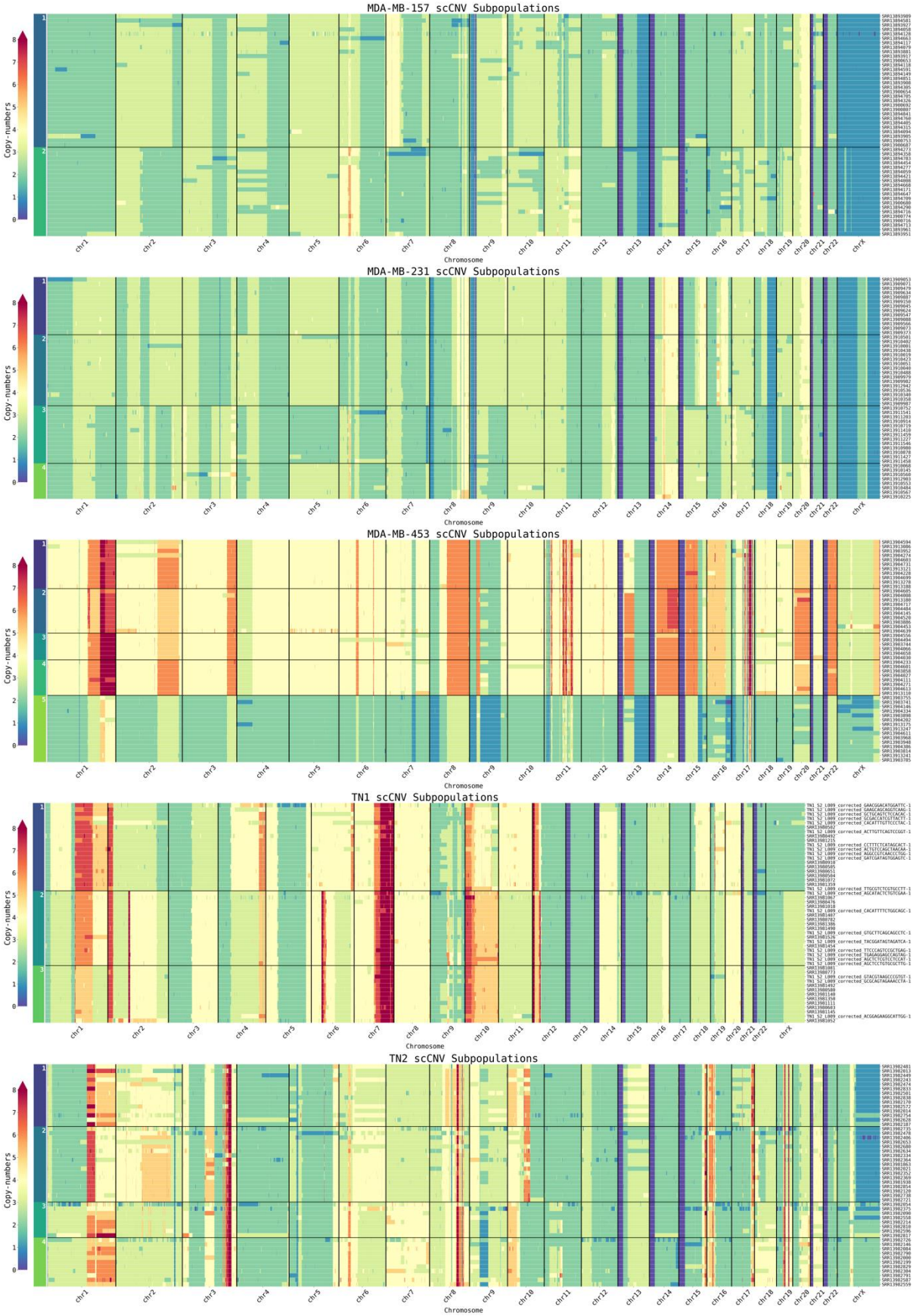

HCT-116 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY

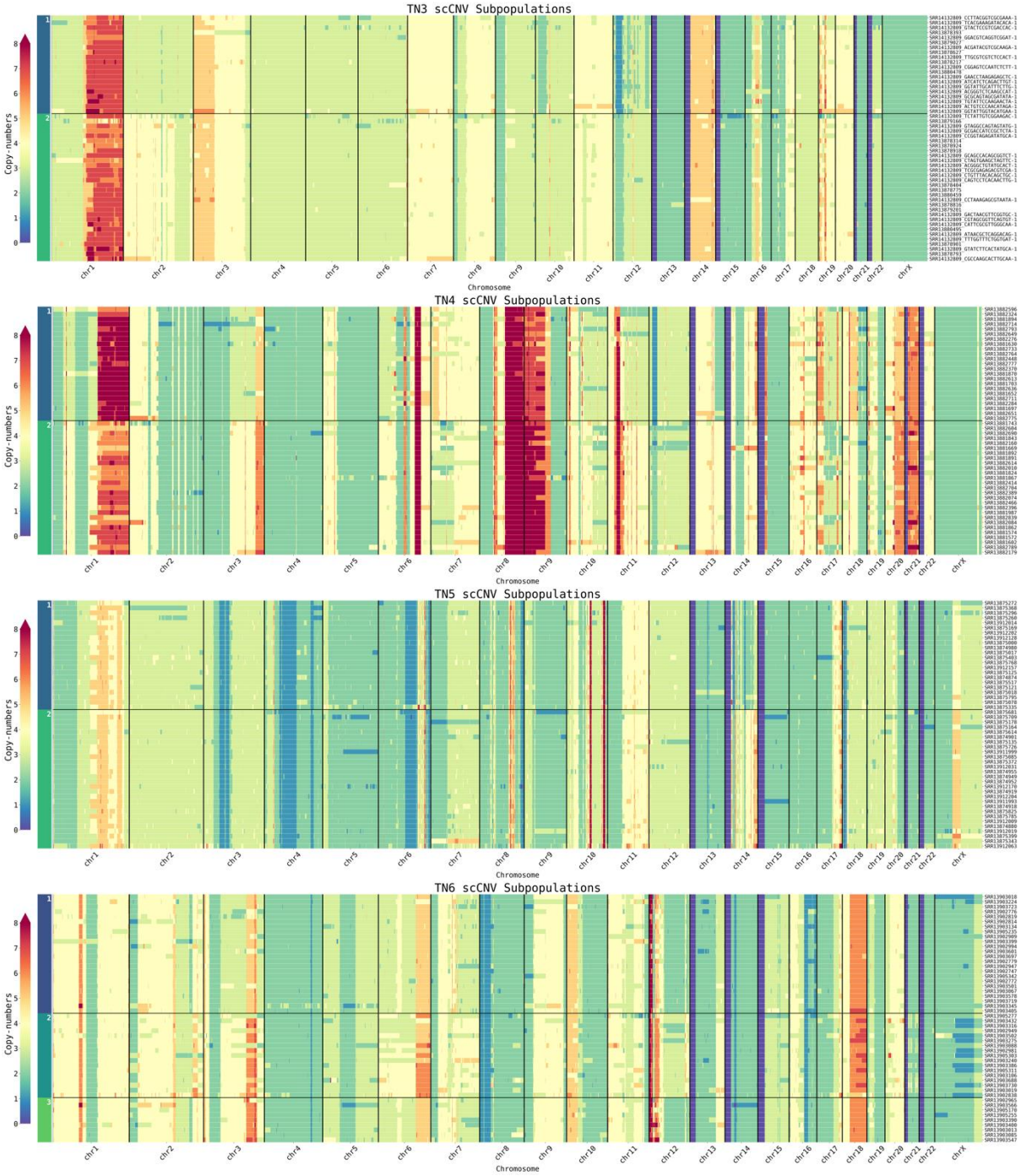

MCF-7 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY

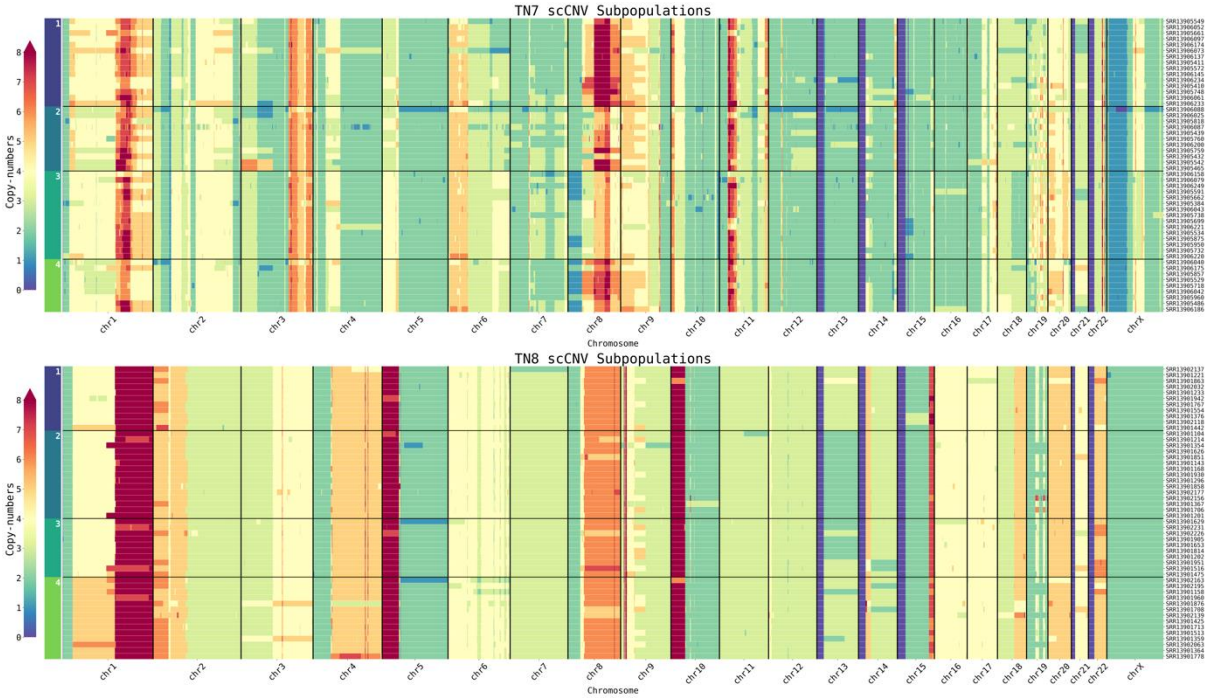

RKO scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY

F

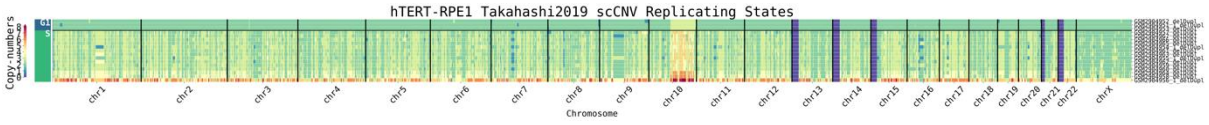

BT20 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX


MDA-MB-157 scCNV Subpopulations
MDA-MB-231 scCNV Subpopulations
MDA-MB-453 scCNV Subpopulations
TN1 scCNV Subpopulations
TN2 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX

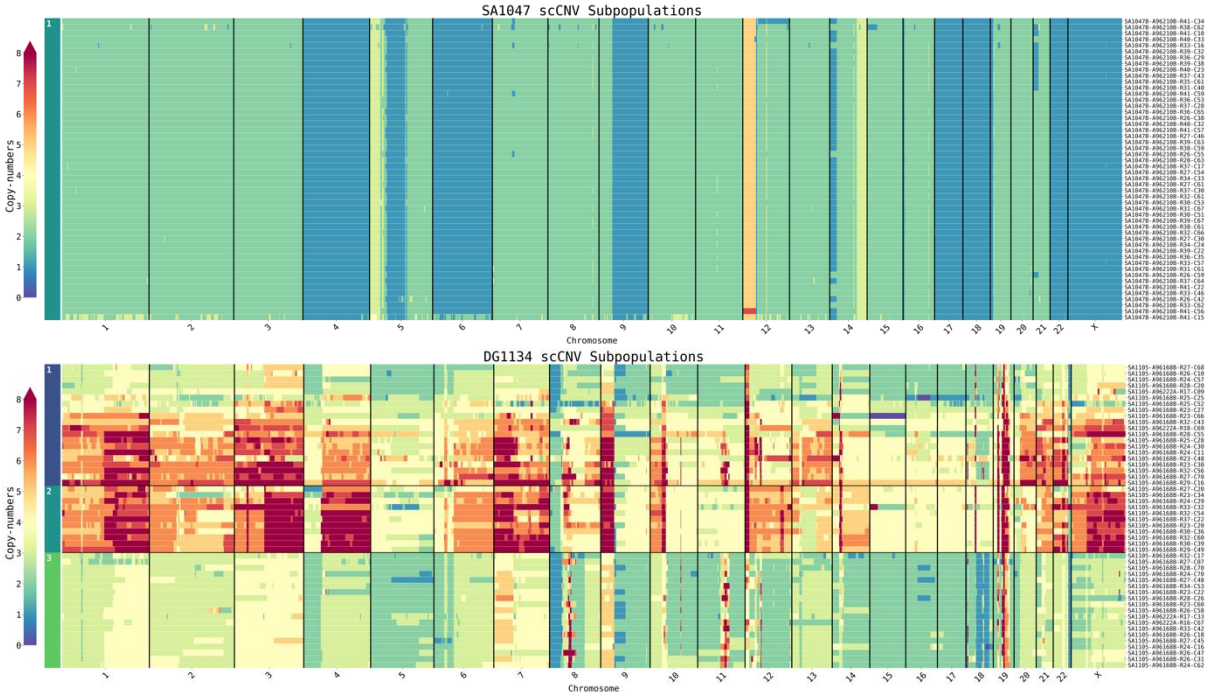

TN3 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
TN4 scCNV Subpopulations
Copy-numbers
Chromosome
TN5 scCNV Subpopulations
Copy-numbers
Chromosome
TN6 scCNV Subpopulations
Copy-numbers
Chromosome


TN7 scCNV Subpopulations
Copy-numbers
Chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
TN8 scCNV Subpopulations
Copy-numbers
Chromosome

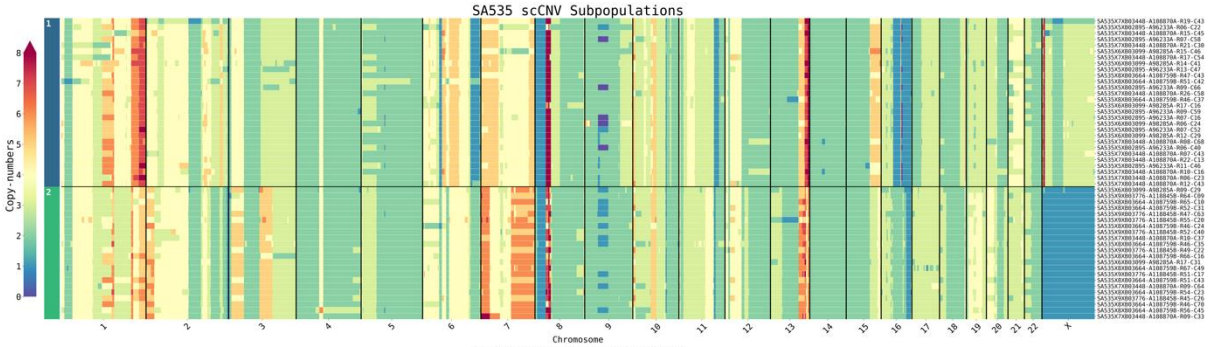
G

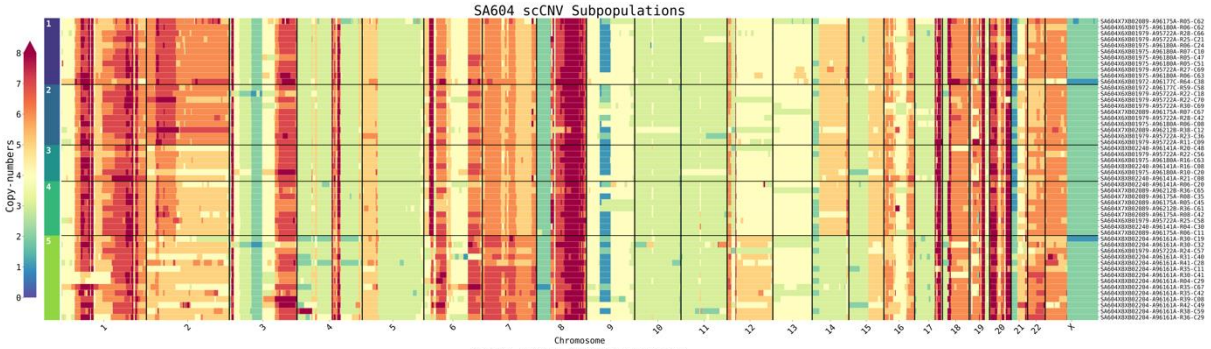

hTERT-RPE1 Takahashi2019 scCNV Replicating States
Copy-numbers
Chromosome

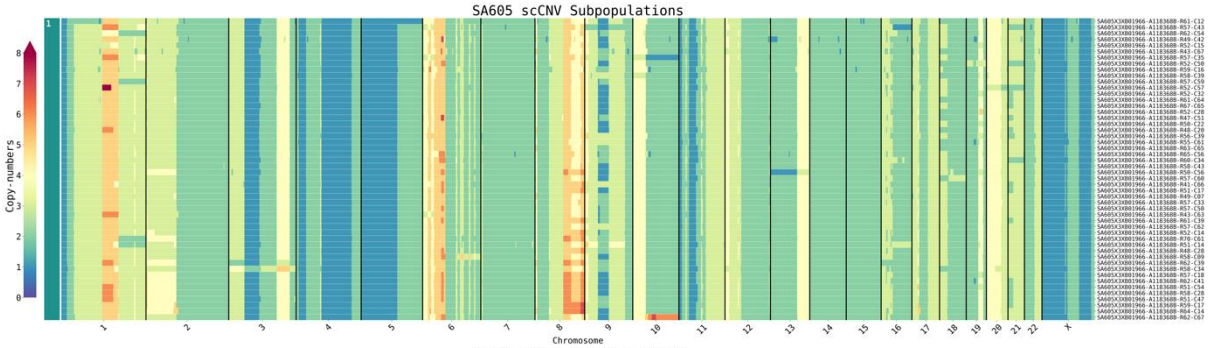
H

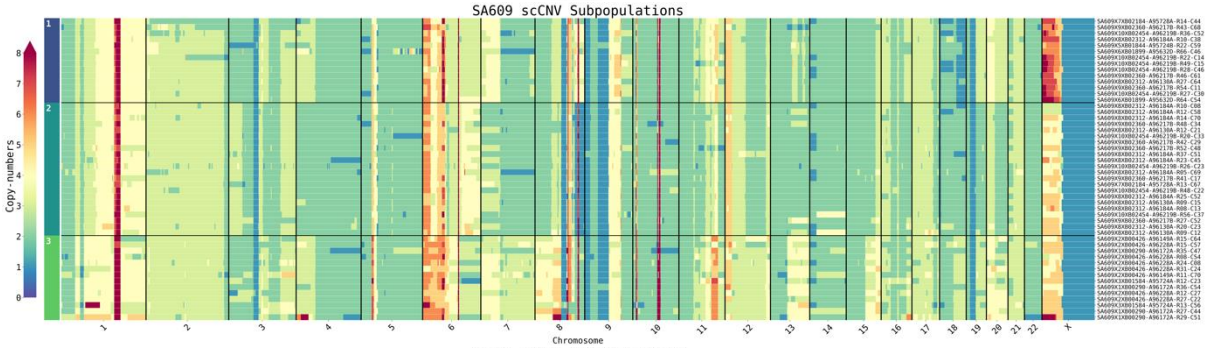

SA1047 scCNV Subpopulations
Copy-numbers
Chromosome
DG1134 scCNV Subpopulations
Copy-numbers
Chromosome

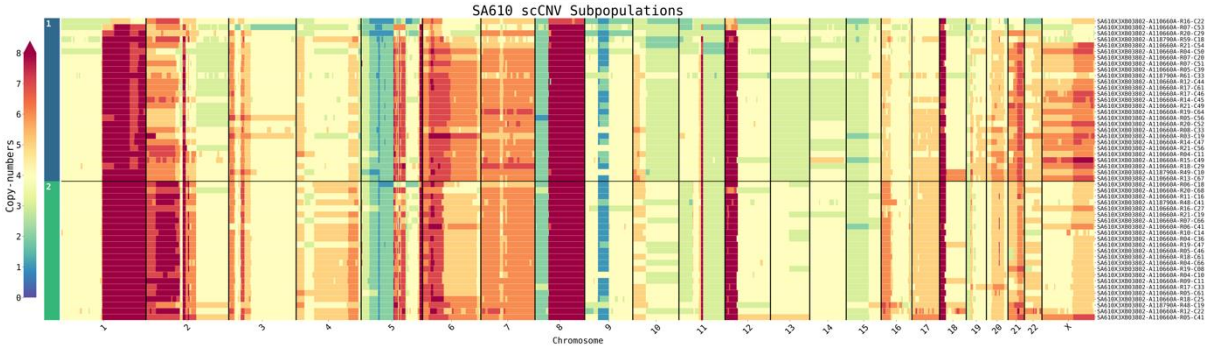

DG1197 scCNV Subpopulations
OV2295 scCNV Subpopulations
SA039 scCNV Subpopulations
SA501 scCNV Subpopulations
SA530 scCNV Subpopulations
Copy-numbers
Chromosome

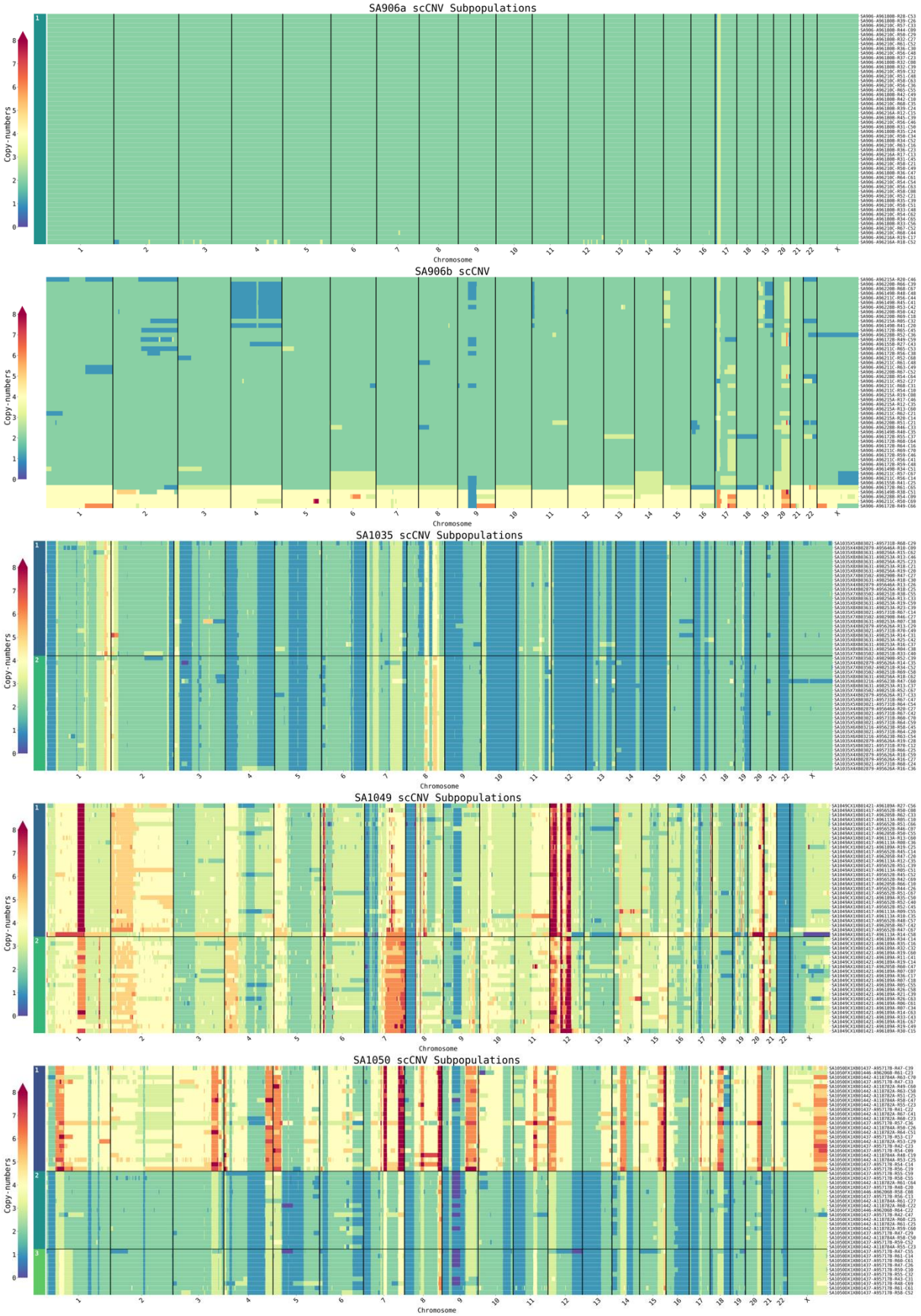

SA535 scCNV Subpopulations
Copy-numbers
Chromosome

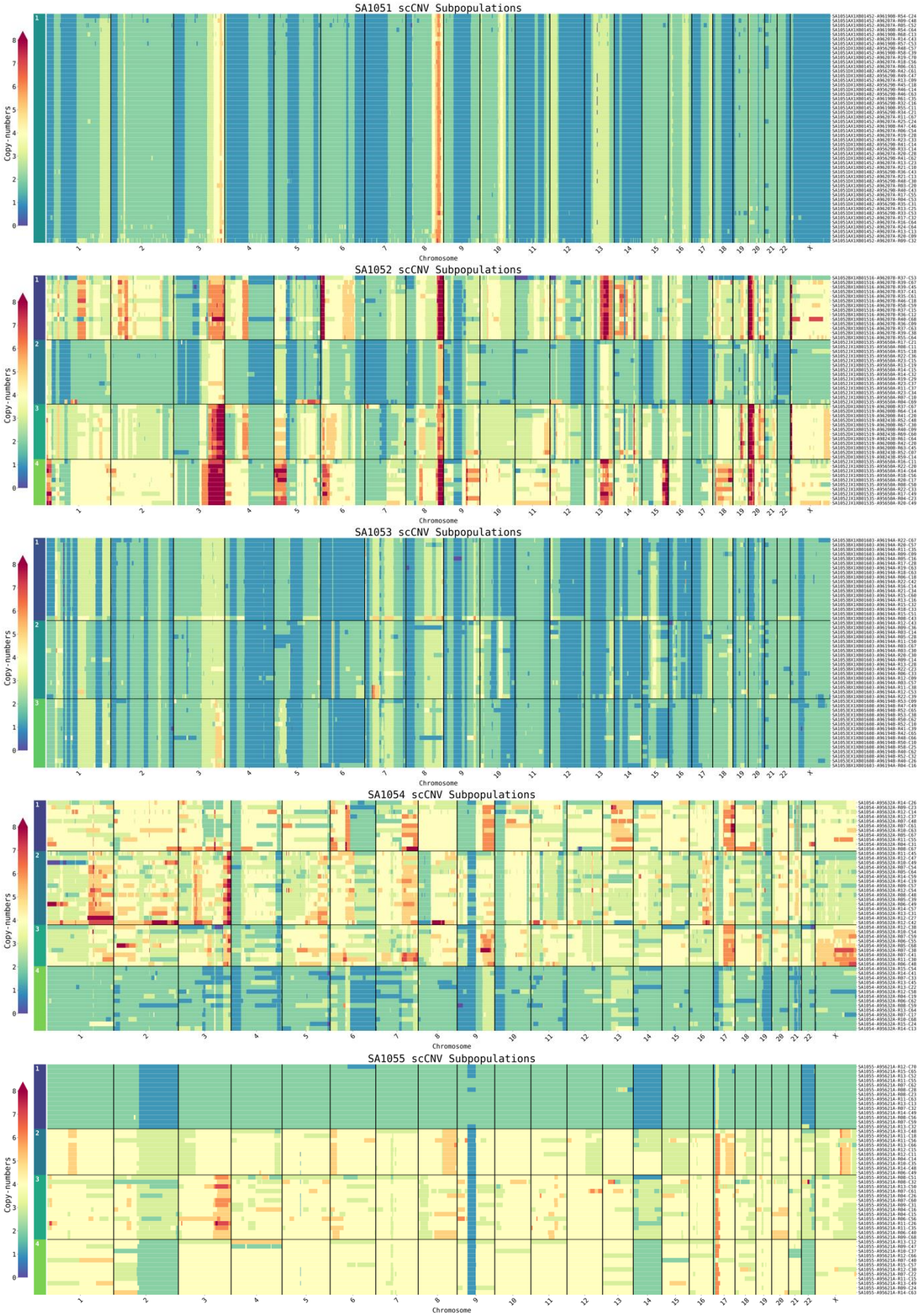

SA604 scCNV Subpopulations
Copy-numbers
Chromosome

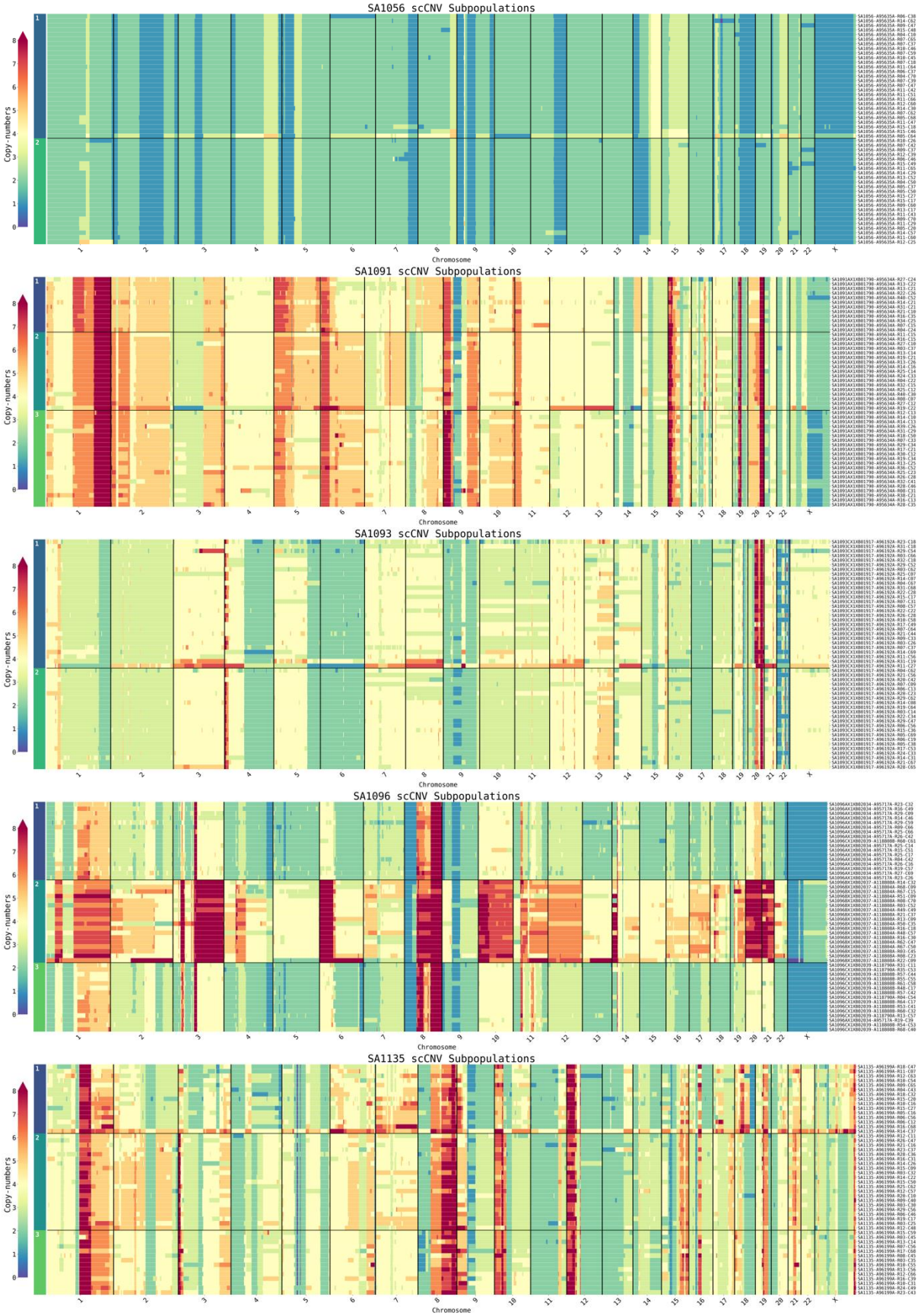

SA605 scCNV Subpopulations
Copy-numbers
Chromosome

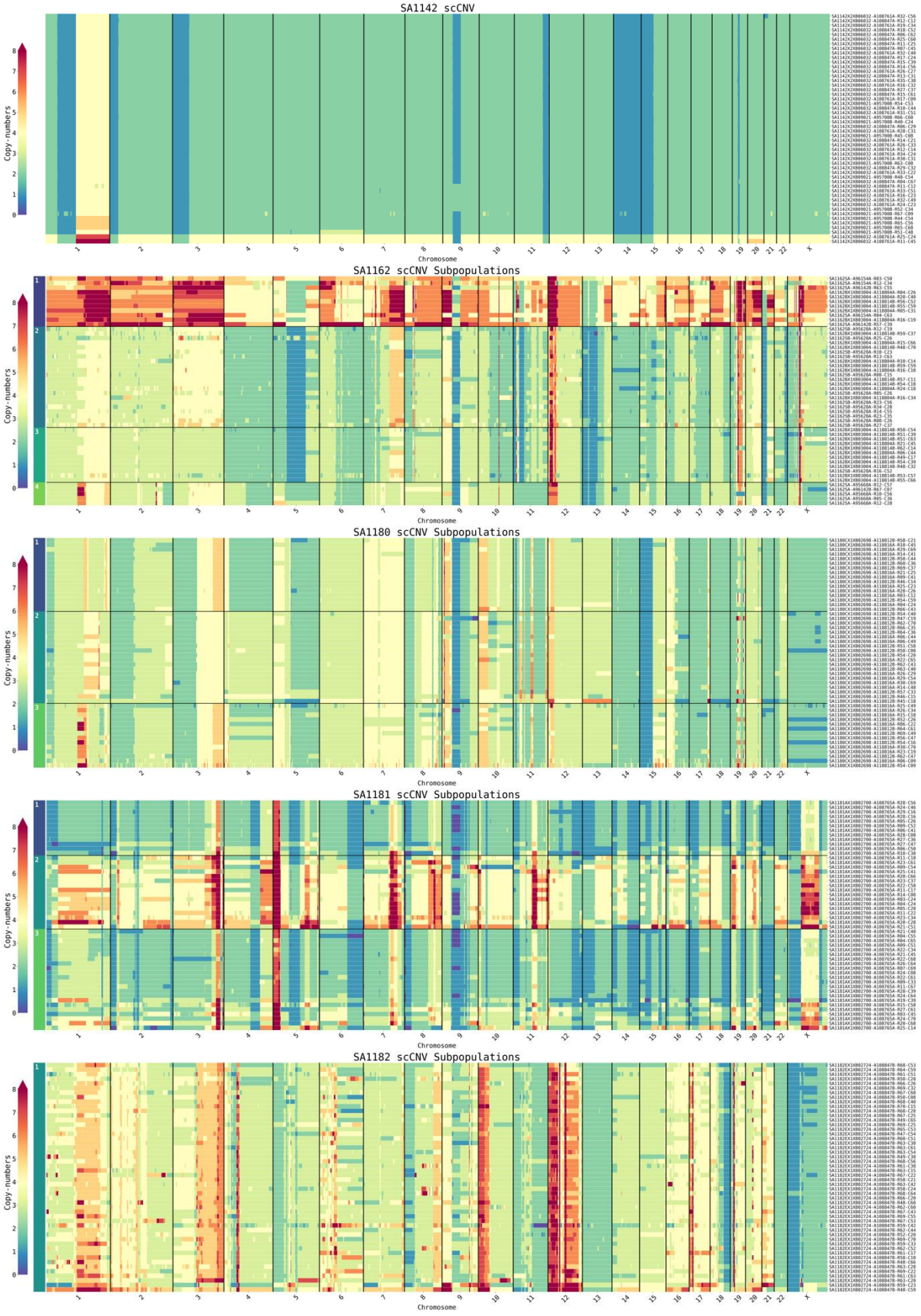

SA609 scCNV Subpopulations
Copy-numbers
Chromosome

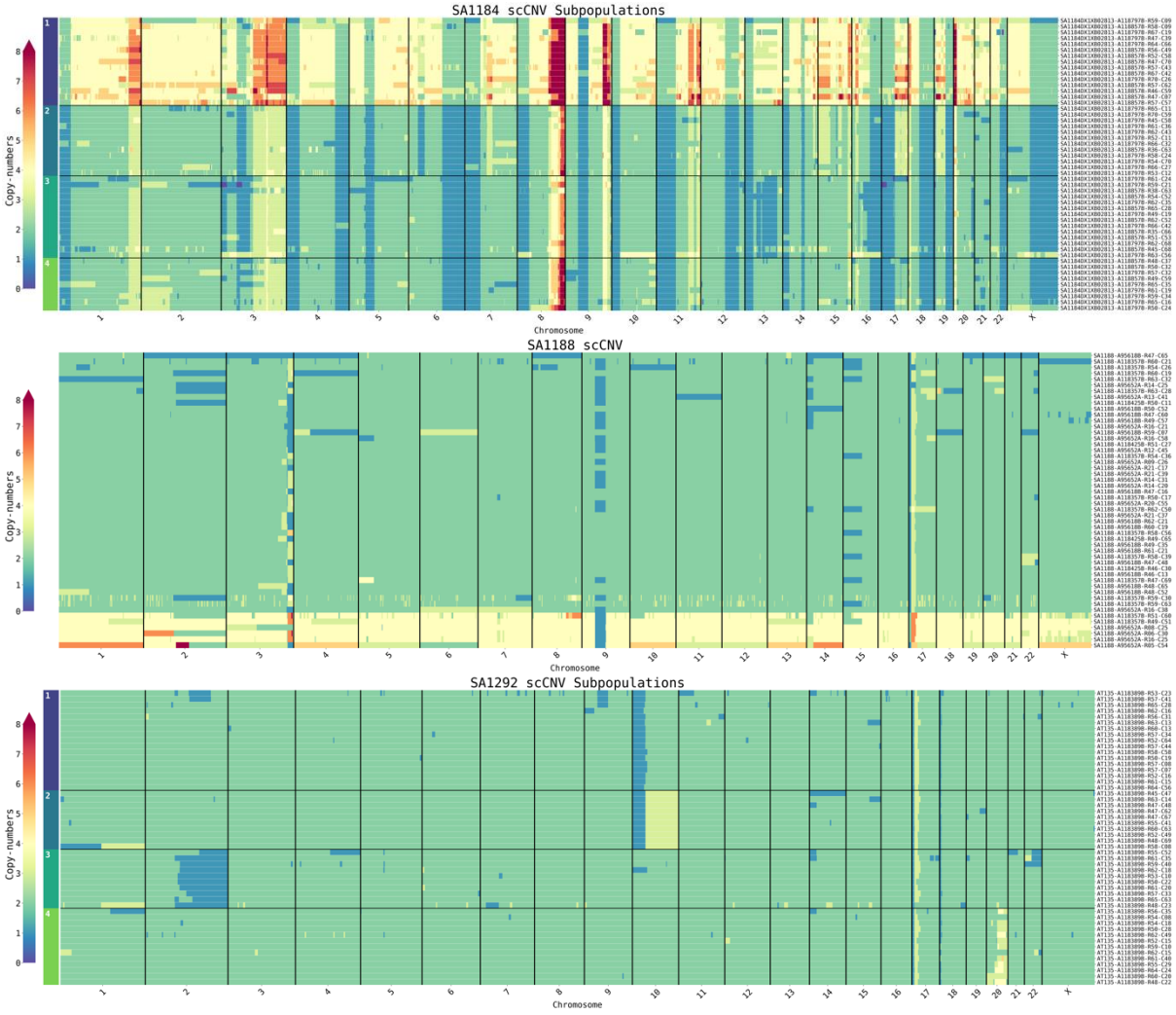

SA610 scCNV Subpopulations
Copy-numbers
Chromosome


SA906a scCNV Subpopulations
SA906b scCNV
SA1035 scCNV Subpopulations
SA1049 scCNV Subpopulations
SA1050 scCNV Subpopulations
Copy-numbers
Chromosome

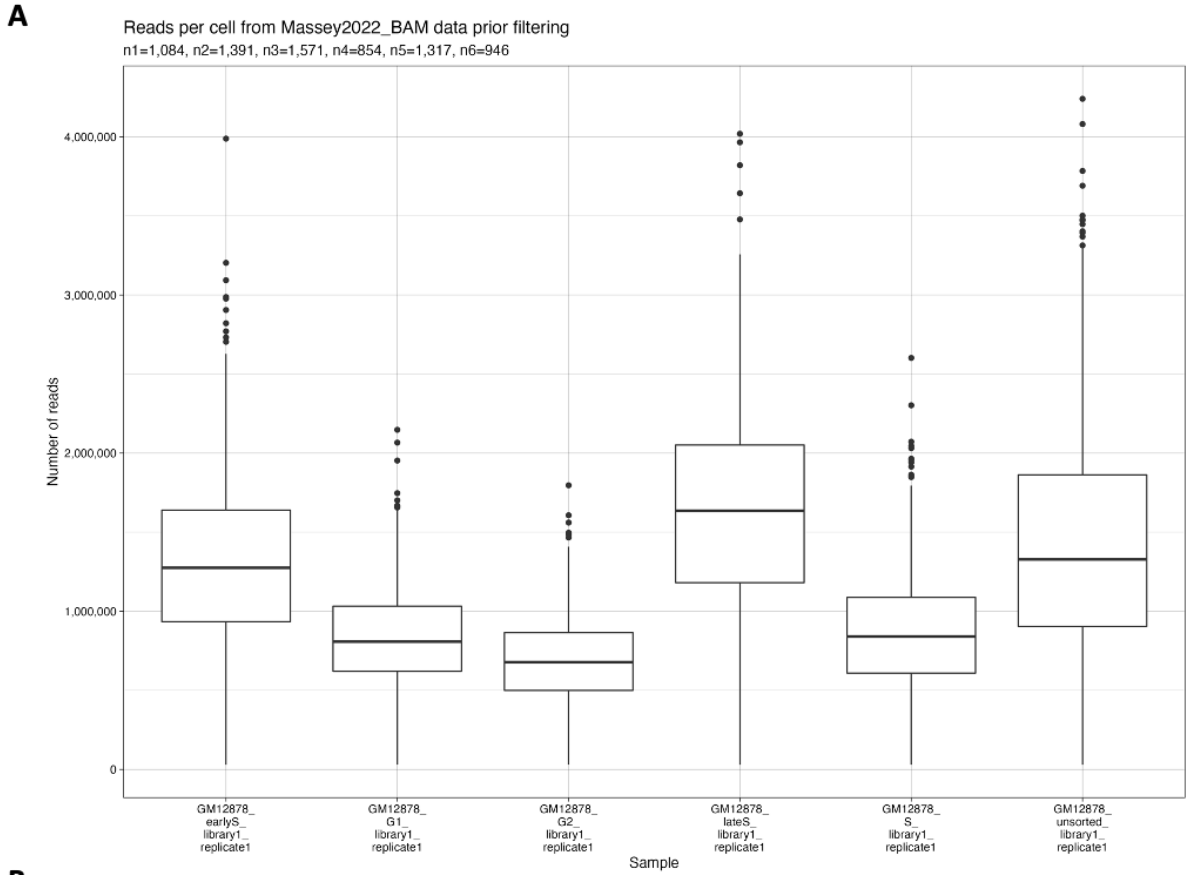

SA1051 scCNV Subpopulations
SA1052 scCNV Subpopulations
SA1053 scCNV Subpopulations
SA1054 scCNV Subpopulations
SA1055 scCNV Subpopulations
Copy-numbers
Chromosome

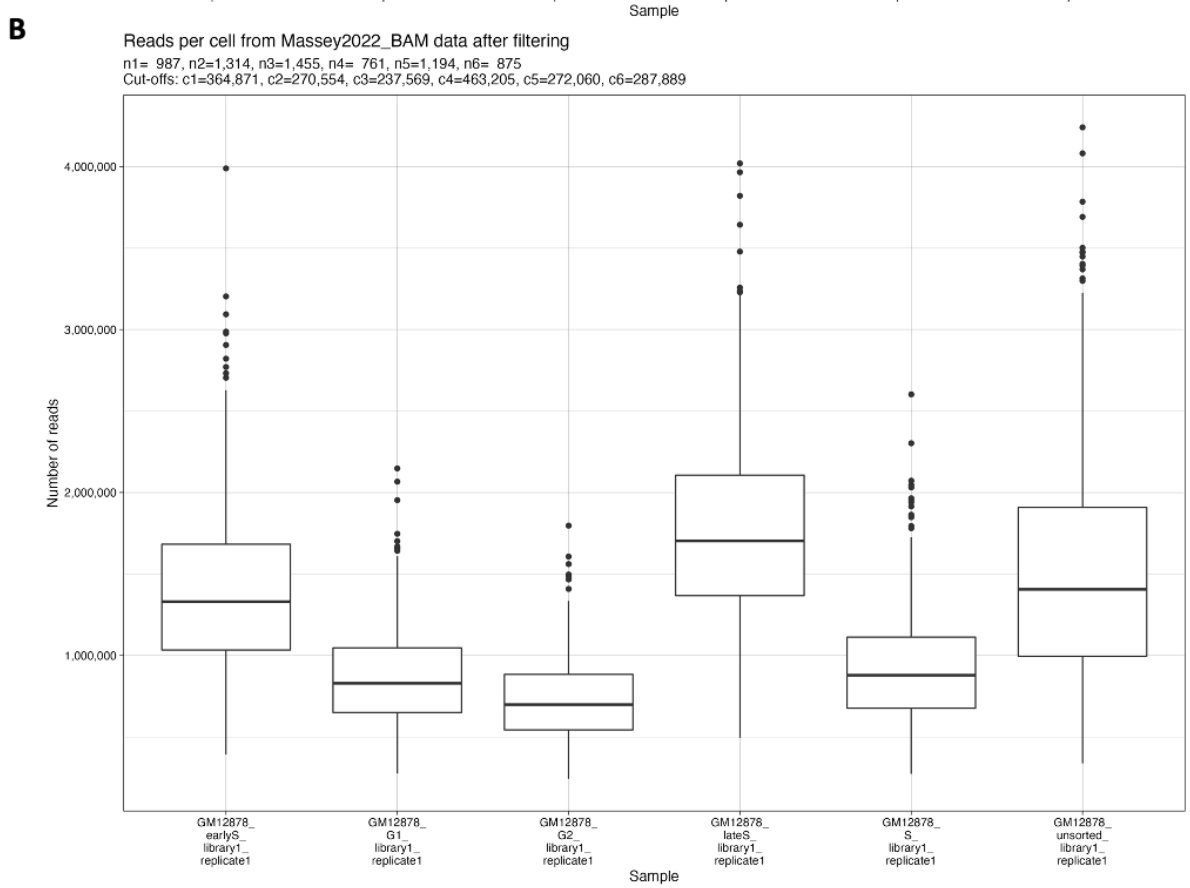

SA1056 scCNV Subpopulations
SA1091 scCNV Subpopulations
SA1093 scCNV Subpopulations
SA1096 scCNV Subpopulations
SA1135 scCNV Subpopulations
Copy-numbers
Chromosome


SA1142 scCNV
SA1162 scCNV Subpopulations
SA1180 scCNV Subpopulations
SA1181 scCNV Subpopulations
SA1182 scCNV Subpopulations
Copy-numbers
Chromosome

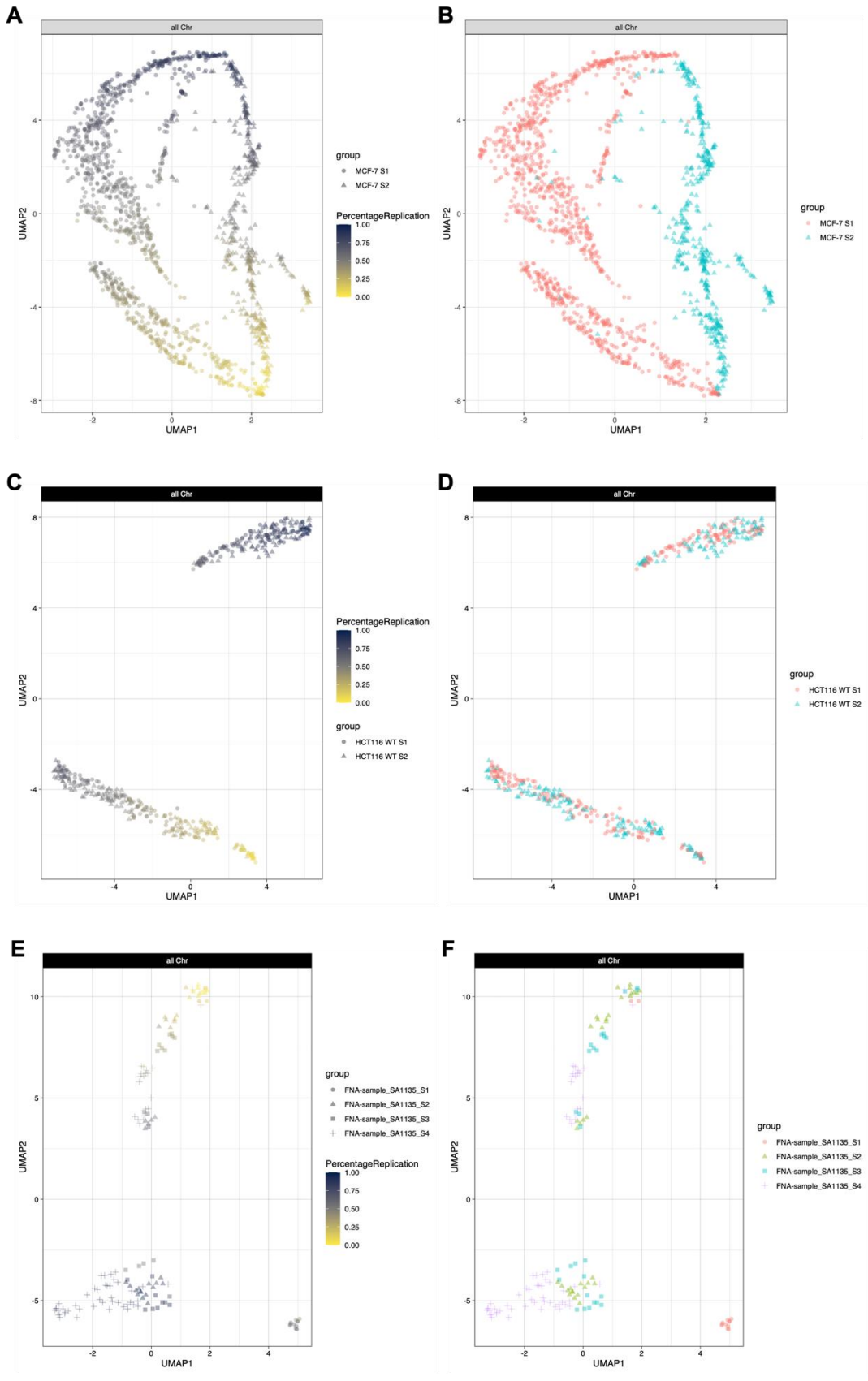

SA1184 scCNV Subpopulations
Copy-numbers
Chromosome
SA1188 scCNV
Copy-numbers
Chromosome
SA1292 scCNV Subpopulations
Copy-numbers
Chromosome

8.3.4. Single-cell cut-off reads with the EM algorithm





8.3.5. Single Cell RT trajectories



RÉSUMÉ

La réplication de l'ADN est essentielle pour les cellules, car elle permet de créer les quelque 30 000 milliards de cellules qui composent le corps humain à partir d'un seul zygote lors de l'embryogenèse. De plus, tout au long de la vie humaine, la réplication continue de l'ADN et la division cellulaire sont nécessaires pour remplacer les cellules âgées, mortes ou endommagées. Par conséquent, il est crucial que le programme de réplication de l'ADN fonctionne correctement à chaque division cellulaire. Cependant, de nombreux facteurs de stress, à la fois exogènes et endogènes, remettent régulièrement en question l'intégrité de l'ADN, ce qui entraîne une instabilité du génome. Cette instabilité est une cause majeure de cancers et d'autres maladies humaines.

Malgré l'importance du stress de réplication et de l'instabilité génomique dans les cancers, nous ne comprenons pas complètement les mécanismes sous-jacents ni leurs impacts sur le génome. Au cours de la dernière décennie, d'énormes progrès ont été réalisés dans l'analyse des cellules individuelles. L'étude des variants structuraux (VS) au niveau cellulaire est devenue cruciale pour comprendre l'instabilité génomique, en particulier dans des populations cellulaires hétérogènes telles que les échantillons de tumeurs, qui ne peuvent pas être facilement obtenus par des analyses de masse. Des études récentes ont révélé une corrélation importante entre le timing de réplication et l'apparition de VS dans les cancers, montrant que de nombreux VS résultent de mécanismes liés à la réplication. Cependant, il existe un manque d'études détaillées sur les mécanismes précis, en particulier sur les liens entre réplication, transcription et VS au niveau de la cellule unique. Comprendre ces mécanismes est crucial pour lutter contre les principales maladies humaines.

Pour répondre à cette question, ce projet développe et utilise de nouvelles méthodes informatiques basées sur l'intelligence artificielle. Il vise à (i) étudier directement le timing de réplication dans les cancers en analysant le nombre de copies au niveau de la cellule unique et (ii) examiner les interactions entre la réplication et les VS au niveau de la cellule unique. Les signatures des VS découvertes dans ce projet pourraient contribuer à améliorer le diagnostic et à définir de meilleures stratégies thérapeutiques. Dans l'ensemble, ce projet permet de mieux comprendre les mécanismes de la cancérogenèse et contribue à améliorer le diagnostic, le pronostic, le traitement et le suivi personnalisé des patients.

MOTS CLÉS

Timing de réplication, Génomique, Cellule unique, Cancer, Intelligence artificielle, Variants structuraux.

ABSTRACT

DNA replication is a vital process of cells. Besides creating the ~30 trillion cells that comprise the human body from a single zygote during embryogenesis, continuous DNA replication and cell division is necessary during the entire human lifespan to replace the old, dead or damaged cells. It is therefore essential that the DNA replication program is correctly executed at each cell division. However, large numbers of exogenous and endogenous replication stresses routinely challenge DNA integrity and lead to genome instability, which is an important cause of cancers and many other human diseases.

Although replication stress and genomic instability are two important hallmarks of cancer, we lack full comprehension of the mechanisms that lead to these deregulations and the impacts they have on the genome. During the last decade, great progress has been made in analyses of individual cells. Determination of structure variations (SVs) in single cells has become an important approach to study genomic instability in heterogeneous cell populations, such as tumour samples, that cannot easily be obtained from bulk analyses. Recent studies have revealed that replication timing shows a strong association with the occurrence of SVs in cancers, and large amounts of SVs generated during tumorigenesis result from replication-associated mechanisms. However, studies addressing the direct mechanisms and, in particular, the links between replication, transcription and SVs at the single-cell level are missing. Investigating such mechanisms is critically important to address major human diseases.

To address this question, this project develops and uses novel computational methods, based on artificial intelligence, to: (i) directly investigate single cell replication timing (scRT) in cancers by single-cell copy number analysis, and (ii) examine the interactions of replication and SVs at the single cell level. The SV signatures in cancers revealed in this project might help to improve the diagnosis and better define therapeutic strategies. Altogether, this project provides further understanding of the mechanisms of carcinogenesis and contributes to improving the diagnosis, prognosis, treatment and/or personalised monitoring of patients.

KEYWORDS

Replication timing, Genomics, Single-cell, Cancer, Artificial intelligence, Structural variants.