UNIVERSIDADE NOVA DE LISBOA

MScCBBi

MASTER IN
COMPUTATIONAL BIOLOGY
& BIOINFORMATICS

IAN LOPES TEIXEIRA
BSc in Cell and Molecular Biology

BIT by BIT: Associating Transcription to Ageing

September,2023

ITQB NOVA
NOVA SCHOOL OF SCIENCE & TECHNOLOGY
NOVA MEDICAL SCHOOL
NOVA IMS Information Management School
DESDE 1902 INSTITUTO DE HIGIENE E MEDICINA TROPICAL UNIVERSIDADE NOVA DE LISBOA

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
LIFE SCIENCES

BIT BY BIT: ASSOCIATING TRANSCRIPTION TO AGEING

IAN LOPES TEIXEIRA

BSc in Cell and Molecular Biology

Adviser: Sérgio Fernades de Almeida
Principal Investigator, iMM-JLA and Associate Professor with Habilitation, Medical School, University of Lisbon

Co-adviser: Ana Rita Grosso
Assistant Professor, NOVA University Lisbon

Examination Committee

Chair: Paula Maria Theriaga Mendes Bernardo Gonçalves
Associate Professor, FCT-NOVA

Rapporteur: Tomás Pires de Carvalho Gomes
Researcher, iMM-JLA

Member: Sérgio Alexandre Fernandes de Almeida
Principal Investigator, iMM-JLA and Associate Professor with Habilitation, Medical School, University of Lisbon

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

NOVA University Lisbon
September, 2023

BIT by BIT: Associating Transcription to Ageing

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.0.2) [1].

Acknowledgements

During the last 5 years, I called the School of Science and Technology of Nova University of Lisbon (FCT-UNL) my home far away from home. Besides all the hardships faced, I could not leave this place without thanking it and its professors and all working personnel for the incredible 3 years of my bachelor's, and finally, during this master's I kept enjoying its company while meeting many more of what UNL has to offer. Thank you so much for all you did for me, personally and professionally.

A special thanks to my supervisor, Dr. Sérgio F. Almeida, for accepting me in his lab and allowing me to face the reality of the world of science. Thank you for teaching me how to be a better professional, to express myself without stumbling on my own words, and for believing in me when I most needed it.

I would like to sincerely thank Dra. Ana Rita Grosso (forgive me for all the formalities). No words can express how much you have helped during the last few years. Thank you for listening to me, for helping me ease my confused and anxious thoughts, for making the hard science easy to understand, and for giving me the tools to grow at my own pace without slacking off. Thank you for always believing in me, and accepting that future Ian will solve present Ian problems and my lack of colour notion when it comes to plots.

This could not have been done without the help and company of my lab colleagues. Anne-Valerie, Robert, Bilal, Anna, Madalena, Inês, Margarida, and Cris, thank you for all your support, all the laughs and talks. Even in the most daunting moments of science, there's always someone to make it lighter. A special thanks to both, Balsinha and Rita, two colleagues who became some of my dearest friends in such a short time. Thank you for the laughs, the hugs, the coffee breaks, advice. I can't put into words this last year we spent together, even if the problem would be having faith I would still put all of mine in you both, my great scientist and friends.

To all my friends, thank you for giving my head the right to drift from the thesis, to have fun, and to enjoy many moments with you all. The biggest thanks are to Adão, Ema, Serra and Miguel, each of you, in its own way was here for me when I needed you the most. Thank you for being my friends, for all the moments out, the games, the laughs, and for all the care you have for me. Even if all my work goes wrong, I can always count

on you to make it the best day, even if I lose at snooker by putting the black in right after breaking the game.

To all my family, especially my grandma, mom and dad, thank you for always being here for me. I could never repay or thank you enough for all the support, and everything you did for me. I know my mood sometimes (a lot of times) is not the easiest, but you still are here for me no matter what, either to help me or just listen to me speaking about science, looking like the biggest gibberish ever. "The word of a man means more than its actions", thank you for showing the true meaning of this and for the best thing that I could ask for, even if we disagree sometimes. Thank you for loving me, your Yanko and Zé Fuinhas, even if I beat you at cards.

Não podia terminar esta tese sem te agradecer a ti, avô. Apesar de teres partido exatamente 1 semana antes de toda esta aventura começar, não passei um momento em que sentisse que esta tese não era fruto do teu trabalho. Obrigado por tudo, por sempre teres estado cá para mim, antes e durante este grande passo. Sei que independentemente de tudo, vais estar a ver-me crescer, a cair e levantar e continuar. Obrigado por me inspirares a não desistir mesmo quando parece a única opção viável. Obrigado por seres o melhor avô que podia pedir.

Finally, I would like to thank myself. Even if sometimes I hate past Ian, for the problems he brings to present and future Ian, I know he makes them with the hope everything goes amazing.

”

“O problema é ter fé.”

— Unknown Author

Abstract

A common and longstanding view of the ageing process poses that the onset of ageing hallmarks is driven by the gain of somatic mutations deriving from the inaccurate repair of DNA lesions, the most catastrophic of which is Double-strand Break (DSB). However, compelling evidence accumulated during the last decade demonstrated that, in addition to causing mutations, DSBs can initiate gene expression events. Using live-cell imaging, it was recently provided direct evidence of transcription initiation at DSB sites, a process dubbed Break-Induced Transcription (BIT), which was previously reported as an important step of DSB repair and gene expression regulation. A trademark of the ageing cell is the widespread rewiring of gene expression, including an increase in spurious transcription. This led us to propose the disruptive hypothesis that alternative gene expression events established through BIT at DSB contribute to the establishment of the ageing phenotype. By using public human data from genome-wide detection techniques of DSB (BLISS, DSBCapture, Break-seq) and high-throughput sequencing (RNA-seq), whose cells underwent no treatment, it was possible to identify natural BIT events at the whole genome level. We were able to develop a genome-wide detection tool of BIT independently of the DSB technique resolution. The available dataset provided no evidence of BIT events influencing ageing itself.

Keywords: Ageing, Double-strand break, Break-Induced Transcription, Spurious transcription

Resumo

Uma visão comum e duradoura do processo de envelhecimento sugere que o início das marcas de envelhecimento é uma consequência pela acumulação de mutações somáticas resultantes da reparação imprecisa de lesões no DNA, sendo a mais catastrófica delas a quebra de dupla hélice (DSB). No entanto, evidências convincentes acumuladas durante a última década demonstraram que, além de causar mutações, as DSBs podem levar a eventos de expressão genética. Através de microscopia em tempo real de células vivas, foi recentemente obtido provas da iniciação da transcrição nos locais de DSB, um processo denominado Indução de Transcrição por Quebra (BIT), que anteriormente foi relatado como uma etapa importante na reparação de DSB e na regulação da expressão genética. Uma característica da célula envelhecida é a reconfiguração generalizada da expressão genética, incluindo um aumento na transcrição espúria. Isso levou-nos propor a hipótese disruptiva de que eventos alternativos de expressão genética através de BIT em DSB contribuem para o estabelecimento do fenótipo de envelhecimento. Através do uso de dados públicos de humanos de técnicas de detecção de DSB em todo o genoma (BLISS, DSBCapture, Break-seq) e sequenciação de RNA de nova geração (RNA-seq), cujas células não foram submetidas a qualquer tratamento tratamento, foi possível identificar eventos naturais de BIT em todo o genoma. Fomos capazes de desenvolver uma ferramenta de detecção de BIT em todo o genoma, independentemente da resolução da técnica de DSB. Os conjuntos de dados disponíveis não forneceram evidências de que eventos de BIT influenciam o processo envelhecimento.

Palavras-chave: Envelhecimento, Quebra de dupla hélice, Indução de Transcrição por Quebra, Transcrição Espúria

Contents

List of Figures

List of Tables

Acronyms

LMNA	Lamin A/C *(p. 5)*
PALB2	Partner And Localizer of *BRCA2* *(pp. 9, 14, 18, 24, 26, 27, 29, 30)*
8-OHdG	8-hydroxy-2-deoxyguanosine *(p. 7)*
ATP	adenosine triphosphate *(p. 5)*
BIT	Break-Induced Transcription *(pp. 9, 10, 12, 15–21, 23–26, 29, 30)*
BLISS	Break Labeling In Situ and Sequencing *(pp. 14, 19, 46)*
ChrM	mitochondrial chromosome *(pp. 16, 20)*
DDR	DNA Damage Response *(pp. 7, 9, 10)*
DDRNA	DNA damage response RNA *(pp. 9, 10)*
dilncRNA	damage-induced long non-coding RNA *(pp. 9, 10)*
DNA	Deoxyribonucleic Acid *(pp. 3–11)*
DNA-PK	DNA-dependent protein kinase *(p. 8)*
DNA-PKcs	DNA-dependent protein kinase catalytic subunit *(p. 7)*
dNTP	deoxynucleotide triphosphate *(p. 5)*
DSB	Double-strand Break *(pp. 6–10, 12, 14–16, 19–23, 26, 29, 30)*
ELS	essential lifespan *(pp. 1, 5)*
FR	Forward Reverse *(p. 39)*
GEO	Gene Expression Omnibus *(pp. 14, 39–46)*
GTEx	Genotype-Tissue Expression *(pp. 14, 18, 27, 30)*
HGPS	Hutchinson-Gilford Progeria Syndrome *(pp. 5, 14)*
HR	Homologous Recombination *(pp. 7–9)*

IDE integrated development environment *(p. 13)*

LIG4 ligase IV *(p. 8)*
lncRNA long non-coding RNA *(pp. 24, 29)*

M Million *(pp. 15, 19, 20, 39–45)*

NA Not Applicable *(p. 46)*
NHEJ Non-Homologous End Joining *(pp. 7, 9)*
NHEK Normal Human Epidermal Keratinocytes *(pp. 14, 22)*
nt nucleotide *(pp. 16, 20)*

PCA Principal Components Analysis *(p. 17)*
Pol II RNA-polymerase II *(pp. 10, 16)*

RF Reverse Forward *(pp. 39–45)*
RNA Ribonucleic Acid *(pp. 9–11, 15, 19, 20)*
RPA replication protein A *(p. 9)*

SSB Single-strand Break *(pp. 7, 9)*
STAR Spliced Transcripts Alignment to a Reference *(pp. 16, 19)*

TBP TATA-binding protein *(pp. 11, 15)*
TPM Transcripts per Million *(p. 14)*
TSS Transcription Start Site *(pp. 10, 11, 16, 17, 20, 24, 29, 30)*

UV ultraviolet *(p. 7)*

XLF XRCC4-like factor *(p. 8)*
XRCC4 X-ray repair cross-complementing protein 4 *(p. 8)*

1 Introduction

1.1 Ageing phenotype

Giacomo Leopardi once said: *"Old age is the supreme evil, because it deprives us of all pleasures, leaving us only the appetite for them and it brings with it all sufferings. Nevertheless, we fear death, and we desire old age"*. The idea of ageing differs from person to person. While some consider it the process of getting older, others as the path to a person's demise. Many different definitions care to explain such a common phenomenon, either from a physiological and molecular perspective or by giving a more psychological and philosophic approach[2].

Even though Giacomo Leopardi could grasp one of many ageing definitions, biology does it differently. Homeodynamics refers to the ability of cells and organs to adjust to changes within the organism and maintain a state of equilibrium[3]. It can be said that ageing is a result of decreasing homeodynamism, leading to a higher likelihood of age-related changes. Throughout the individual's lifespan, those changes start to accumulate giving rise to the ageing phenotype, as explained by biogerontology, the biological field of ageing-related physiology and diseases[2,4,5].

It is noteworthy to mention that the process of ageing commences as essential lifespan (ELS) ends. ELS refers to the duration required for an individual to reproduce, thereby ensuring the continuity of the species, which aligns with the fundamental purpose of life as postulated by Charles Darwin. Notably, the ELS varies across different species and is a crucial determinant in the extension of life.[5,6]. This will depend on the speed of maturation, and the start of reproductive capacity. In the case of the *Homo sapiens* it was considered to be 50 years of ELS, however, each individual can present it differently since many factors can influence ageing and its starting moment[5,7].

As previously explained, ageing is a natural process that occurs due to a lack of homeodynamism in the system, which ultimately leads to the extension of an individual's lifespan. The factors responsible for ageing include genetics, epigenetics, and environmental factors. However, it is interesting to note that genetics only accounts for a mere 25% of determining an individual's longevity[6].

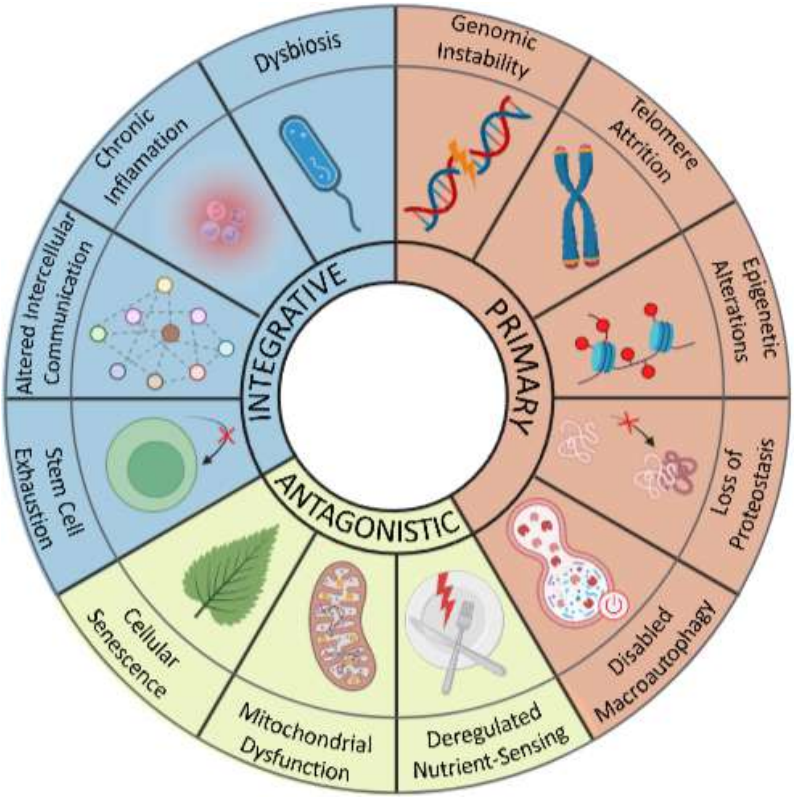


Figure 1.1: The 12 Hallmarks of Ageing organized by 3 categories, according to their effect on the phenotype.
The Hallmarks of Ageing correspond to the characteristics that characterize and drive this process, upon homeodynamics being impaired. The 12 hallmarks are divided into 3 groups: primary, being the cause of damage, promoting ageing; antagonistic, the response to damage itself; and integrative, as a manifestation of ageing itself, when the primary and antagonistic ones can no longer be compensated. Created with BioRender.com

1.1.1 Hallmarks of Ageing

There are many theories that attempt to explain the effects of homeodynamism loss in ageing. However, none of these theories can be attributed to a single characteristic. Many people believe that ageing can result from the accumulation of molecular damage, particularly in macromolecules, as well as defects in repairing or digesting mechanisms that can solve these issues. It is more accurate to say that ageing is influenced by its hallmarks[8,9].

The hallmarks of a disease or phenotype are the various aspects that can define such occurrences. While the hallmarks of ageing may seem separate, they rely on one another. *López-Otín* in 2013[10] proposed this idea of 9 driving forces to explain the ageing phenotype, nevertheless, 10 years later, concludes it to be composed of 12 hallmarks instead (Figure 1.1)[9]. Each of them has to follow three criteria:1) their age-association presence, and the ability to 2) accelerate the ageing process, but also work 3) as a therapeutic target to interrupt or even undo ageing itself[9].

To give a brief idea about each of the hallmarks, they have been categorized into three

major groups: 1) Primary hallmarks, those responsible for the causes of damage, promoting ageing; 2) Antagonistic hallmarks, the response to damage itself; and 3) Integrative hallmarks, as a manifestation of ageing itself, when the primary and antagonistic ones can no longer be compensated[9,11].

1.1.1.1 Primary Hallmarks

The Primary Hallmarks are composed of 5, which contribute to the ageing process by causing damage. These are genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, and disabled macroautophagy. The last one was added upon reviewing the hallmarks of ageing[9,10].

Genomic instability is the accumulation of Deoxyribonucleic Acid (DNA) damage, resulting in small-scale events, such as base pair changes, minor insertions, deletions, or inversions, as well as substantial changes like translocations, duplications, either of a segment or a chromosome, and chromosome loss. Further information on this topic is provided in Chapter 1.2. These alterations affect the transcriptomic profile and mitochondria function[9,12].

Telomere attrition is the consecutive shortening of telomeres, a repetitive part at the end of the chromosomes, as a result of successive cell divisions, which is a sign of a reduced lifespan of cells. Drugs that induce telomerase activity, a reverse transcriptase that maintains and elongates telomeres, can reduce telomeres attrition and treat age-related diseases, such as pulmonary fibrosis[9,13].

Epigenetic alterations are changes in the patterns of DNA methylation and acetylation, post-translational anomalies of histones, and atypical chromatin restructuring, more or less condensation of it, leading to different gene expression profiles from the expected. By looking into those epigenetic features, we can determine a person's biological age by using the epigenetic clock. The epigenetic clock is a molecular test that by measuring the DNA methylation levels and distribution, it is possible to determine a person's biological age. Multiple estimators have been developed, with different purposes and accuracies depending on the data available (e.g. some estimators are better suited for blood samples, but worse for other types of tissues) and goals[9,14,15].

When cells experience a deficit in proteome homeostasis, it can lead to the accumulation of protein damage, known as loss of proteostasis. This damage can manifest in different ways, including the misfolding of proteins or modifications that signal protein digestion, like ubiquitination. As a result, damaged proteins can form aggregates, which are linked to age-related diseases like Alzheimer's disease. This happens when the number of damaged proteins is too high for the repairing mechanisms (like proteasome, and lysosome degradation) to handle, or when such mechanisms aren't functioning. However, research has shown that pharmacological treatments can induce these mechanisms and pathways, and help ameliorate certain diseases, such as amyotrophic lateral sclerosis[9,16].

Macroautophagy is a type of autophagy in which the components to be degraded are

sequestred by a vesicle in the cytosol, the autophagosomes, of proteins, non-proteinaceous molecules (e.g. glycogen), organelles (e.g. mitochondria, peroxisomes), and even pathogens. One of the new hallmarks of ageing, disabled macroautophagy, refers to a reduction in autophagy events, increasing the amount of harmful non-digested components in the cytosol. By inducing autophagic events in specific cell types, was possible to revert the ageing phenotype, improving autophagic flux and immune response[9,17,18].

1.1.1.2 Antagonistic Hallmarks

The Antagonistic Hallmarks are associated with the biological response to damage, being composed of 3. These are deregulated nutrient-sensing, mitochondrial dysfunction, and cellular senescence.

When the pathways responsible for regulating cell metabolism in response to existing nutrients are damaged, as a result of cellular damage, it is known as deregulated nutrient-sensing. This can cause an increase in caloric intake and overnutrition, resulting in the inhibition of receptors that detect nutrient scarcity. Additionally, cellular responses to stress, like DNA repair, can be suppressed, and an uncontrolled inflammatory response may occur, along with excessive anabolic signalling[9,19].

Mitochondrial dysfunction results from multiple Primary Hallmarks taking effect - the accumulation of mutation on mitochondrial DNA, the loss of proteome homeostasis, the decline of macroautophagy of organelles, merging into a reduction in the homeodynamics of mitochondria. This dysfunction leads to a decrease in cellular energy production, compromised cellular fitness, and an increase in permeability of the mitochondria membrane. Moreover, it can trigger inflammatory responses, which can culminate in cell death. Fortunately, inducing the production of mitochondria proteins offers a viable solution to counteract age-related characteristics and improve metabolic and nutrient-sensing regulation[9,19].

Cellular senescence is a constant cell cycle arrest, caused by the presence of any of the Primary Hallmarks that leads to cellular damage, or by extracellular factors like inflammatory mediators. Although these cells cannot replicate, they can function as a signal for tissue repair by recruiting immune cells. However, in ageing, this recruitment process fails, resulting in the accumulation of senescent cells. By eliminating those cells - senolysis - it was possible to recover the immune response, leading to an extension of lifespan[9,20,21].

1.1.1.3 Integrative Hallmarks

The Integrative Hallmarks is composed of 4, resulting from the stress and accumulation of primary and antagonistic hallmarks, whose cell can't resolve. These are stem cell exhaustion, altered intercellular communication, chronic inflammation, and dysbiosis. Both chronic inflammation and dysbiosis were considered hallmarks in the most recent findings[9,10].

As a result of epigenetic changes, DNA damage, and cellular senescence, tissue self-renewal and repair become deficient, leading to stem cell exhaustion. It has been shown that inducing reprogramming of such tissues, can revert the epigenetic clock and telomere length, and also increase longevity, in mice[9,22,23].

The altered intercellular communication is a result of the increasing noise in the organism, harnessing cell-to-cell interaction by disrupting the hormone and neuroendocrine signalling pathways. By circulating young blood through old tissues, it was possible to rejuvenate and repair them[9,24].

Chronic inflammation results as a compilation of all other hallmarks being present and leading to an exacerbated inflammatory response. Nonetheless, just by influencing the immune response, it was possible to affect the ageing process and lifespan, without affecting other hallmarks, making it a hallmark of its own[9,25].

Dysbiosis corresponds to an alteration in the microbiome, converging to a less diverse microbiota, increasing the probability of disease as defence mechanisms are at fault. The microbiome trajectory of a healthy old individual tends to be unique, which confers an adaptive system to the organism's necessities, however, some are unable to drift towards the adequate microbiome. This can happen due to other hallmarks in place, such as altered intercellular communication and deregulated nutrient-sensing. It was also shown that by changing the microbiome composition, longevity and cognitive capacity were improved in mice[9,26–28].

1.1.2 Premature Ageing Syndromes

As previously explained, ageing starts as ELS reaches its end, however, some individuals experience the ageing process prematurely. Premature ageing syndromes, also known as progeria syndromes, are a condition in which the individual presents features of an ageing phenotype, both molecular and physiological, at an early age. Most of these syndromes start manifesting between neonatal and early childhood[29].

Studying one of these syndromes gives an insight into the ageing process itself and vice-versa. One of the most studied progeria syndromes is Hutchinson-Gilford Progeria Syndrome (HGPS). HGPS results of a mutation on the Lamin A/C (*LMNA*) gene, leading to the synthesis of progerin instead of lamin A. Progerin then aberrantly anchors to the nuclear envelope, leading to a function deficient nuclear lamina, which is involved in mitosis, transcription, nuclear organization and apoptosis[30,31].

As a result of progerin accumulation, HGPS patients suffer from mitochondria dysfunction and telomere attrition, both hallmarks of ageing (Chapter 1.1.1). The first is characterized by problems in mitochondria genesis and, consequently, decreased mitochondria adenosine triphosphate (ATP) production. The latter results from increased replication fork collapse and cell cycle arrest, and a decrease in the availability of deoxynucleotide triphosphate (dNTP). Nonetheless, mitophagy (i.e. autophagy of damaged mitochondria) is enhanced, contradicting a hallmark of ageing (disabled macroautophagy)[29,30].

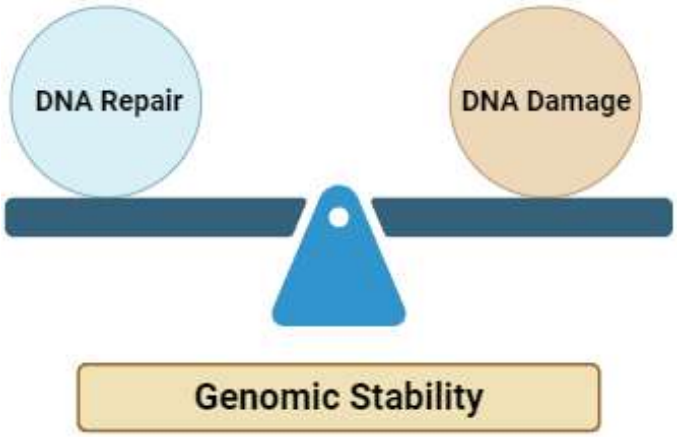


Figure 1.2: Genomic Instability results from more DNA Damage than the DNA Repair mechanisms can solve.
Genomic Stability is achieved when the DNA Repair mechanisms are enough to solve all DNA Damage, right after it occurs. When the repair mechanisms reach a plateau, DNA Damage accumulates, leading to Genomic Instability, corresponding to a higher probability of mutations, deletions in the DNA and chromosome rearrangements resulting from unsolved DNA Damage. Created with BioRender.com

Genomic instability is also associated with premature ageing. Lamin A is responsible for the maintenance of repair mechanisms of Double-strand Break (DSB), allowing for the maintenance of genomic stability. As for progerin production, those pathways are jeopardised by either delaying the recruitment of some repair factors or reducing their affinity to the nuclear matrix, leading to an accumulation of unresolved DSB. However, the cell can still identify them and signal those events for repair[29,32].

1.2 Genomic Instability

All organisms are composed of cells, which are the basic units of life. Each cell is composed of DNA, passed on to the next generation, by the process of replication. Cells have developed mechanisms to degrade malfunctioning and damaged proteins and organelles. In the case of damaged DNA, degrading it, even though possible, can compromise the cell cycle, gene expression, synthesis of proteins and organelles, and the cell function, especially in mononuclear cells[33]. To ensure DNA stability - an undamaged DNA molecule - the cells developed mechanisms to repair it[34,35].

Cells can reach a plateau where the amount of repair is insufficient to resolve all lesions. Besides, until being repaired, unsolved damage can lead to further DNA damage. As can be seen in Figure 1.2, Gnomic Stability is achieved when the DNA repair mechanisms can solve all DNA damage, although an increase in damage, when compared with the repair mechanisms, will lead to a state of Genomic Instability. This is characterized by a higher probability, during cell division, for mutations, deletions in the DNA and chromosome rearrangements[9].

1.2.1 DNA Damage and Responses

DNA damage is any alteration in the DNA structure that leads to cell injuries, by decreasing its fitness, proneness to divide or even organism survival. These structural changes can either be a base exchange, leading to different bonding between both strands, extension or deletion of the DNA, or chemical bonding alteration of bases, such as 8-hydroxy-2-deoxyguanosine (8-OHdG)[36].

Its source can either be exogenous or endogenous agents[8]. Exogenous sources or genotoxic elements can be ultraviolet (UV) light, ionizing radiation, and chemical compounds - aromatic amines, alkylating agents. In the case of endogenous, it can be due to metabolites and proteins that lead to alkylation, hydrolysis, oxidation, or endogenous processes such as replication fork collapses[8,37,38]. This results in one of the following outcomes: depurination and depyrimidination (breakage between a purine and pyrimidine from deoxyribose, respectively), 8-OHdG (guanine nucleoside affected by oxidative stress), cytosine deamination (cytosine turns into uracil), DNA adduction (covalent bond between the DNA and chemical compound), and DNA breakage (Single-strand Break (SSB) and DSB)[8,36].

To resolve any of these, cells have developed DNA Damage Response (DDR). DDR consists of 3 types of processes: repair mechanisms, damage tolerance, and cell-cycle checkpoints[39].

During cell-cycle, if the DNA is damaged, upon passing cell-cycle checkpoints, the cell is capable of signalling the damage for repair. The repair mechanisms aim to undo the lesion on the DNA, by reverting the process. Each type of DNA damage has its specific repair mechanism and machinery. For those lesions whose repair mechanisms are ineffective to solve, the cell uses damage tolerance processes to bypass replication-stalling upon encountering DNA damage. However, when the DNA is too damaged to either be repaired or pass through a damage tolerance process, those checkpoints signal apoptosis of the cell[37,39].

1.2.2 Double-strand Breaks

DSB is a break in both strands of the DNA molecule, being the most detrimental example of DNA damage[40]. They arise, spontaneously or programmed, from either exogenous agents, such as ionizing radiation, or endogenous like replication fork collapses, transposable elements, and endonucleases[34,41,42]. DDR starts a signalling cascade, upon checkpoint detection of it, leading to one of two repair mechanisms: Non-Homologous End Joining (NHEJ) or Homologous Recombination (HR) (Figure 1.3)[35,39,41,42].

NHEJ is the fastest and most used repair mechanism for DSB, however, it is error-prone. Besides being used during interphase, it is most important during G1. After the DSB is signalled and identified, both ends are processed to become blunt ends. Then, are bound by the heterodimer Ku70/80, recruiting factors for the repair machinery such as DNA-dependent protein kinase catalytic subunit (DNA-PKcs). DNA-PKcs binds to the

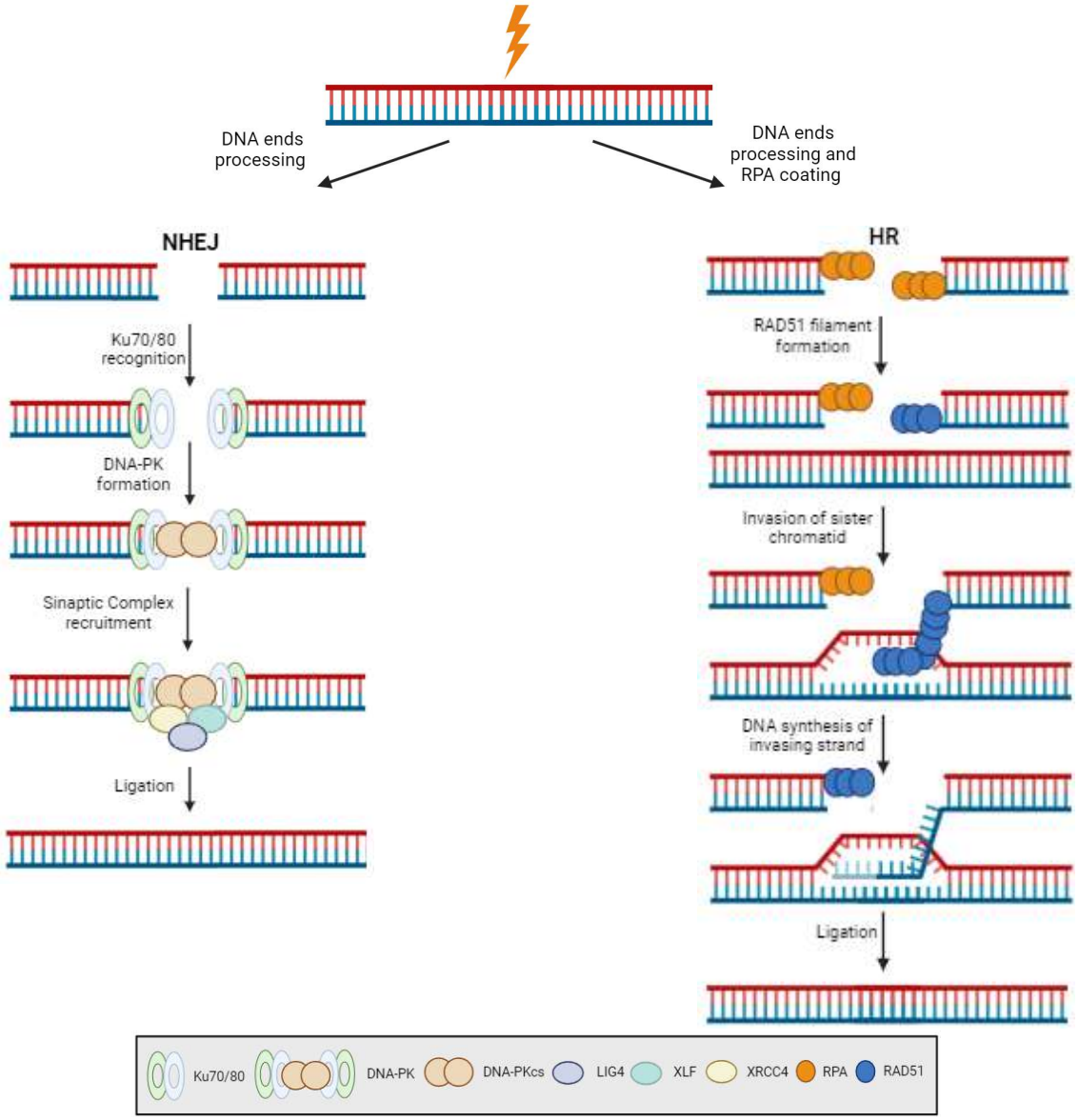


Figure 1.3: Double-strand breaks are repaired by Non-Homologous End Joining and Homologous Recombination.
In NHEJ after the DSB is signalled and identified, both ends are processed to become blunt ends. Then, are bound by the heterodimer Ku70/80, recruiting DNA-PKcs, binding to the heterodimer, forming the DNA-PK holoenzyme. To ligate the DNA a synaptic complex is required, bringing both ends from a long-range complex to a short-range complex. Finally, LIG4, XLF and XRCC4 repair the DSB. In the case of HR, both ends are hydrolysed by endonucleases, forming 3' end SSB, then bound by RPA. Aided by mediator proteins, is replaced by the RAD51 filament. Then the RAD51 filament pairs with a sister chromatid to finish the repair. Created by BioRender.com

Ku heterodimer, forming the DNA-dependent protein kinase (DNA-PK) holoenzyme, which is capable of phosphorylating other repair factors and itself. To ligate the DNA a synaptic complex is required, bringing both ends from a long-range complex to a short-range complex. Finally, ligase IV (LIG4), XRCC4-like factor (XLF) and X-ray repair cross-complementing protein 4 (XRCC4) repair the DSB (Figure 1.3)[41–43].

In the case of HR, which is only available during phases S and G2 of the cell cycle, is an

error-free method. After the DSB is signalled, both ends are hydrolysed by endonucleases, forming 3' end SSB, then bound by replication protein A (RPA). Aided by mediator proteins, such as Partner And Localizer of *BRCA2* (*PALB2*) and BRCA2, RPA is replaced by the RAD51 filament. Then the RAD51 filament pairs with a sister chromatid to finish the repair (Figure 1.3)[41,42].

Even though NHEJ is considered error-prone and HR error-free, both can result in mutations (error-prone) and accurate repair (error-free) depending on the DSB[44].

The formation of DSB can lead to mutations, chromosome rearrangements, and aneuploidy. Mutations are, mostly, a result of repair mechanisms. Chromosome rearrangements are represented as translocations, amplifications, insertions, inversions and deletions, which can either be from unsolved DSB or a consequence of DDR, namely NHEJ in the case of translocations. The accumulation of mutations and chromosome rearrangements lead to different gene expression and protein synthesis and function. Aneuploidy happens when DSB are not resolved, leading to chromosome loss and not to an increase in the number of chromosomes[34,35,42,43,45].

In Chapter 1.1, we discussed how genomic instability affects the ageing process and progeria syndromes. DSB being one type of DNA damaged, it influences both processes. Besides aged-tissues accumulate DSB when compared to younger ones, defects in the DSB repair and, consequently, its accumulation are common features of progeria syndromes and associated with induction of the ageing process[46].

1.3 Break-Induced Transcription

DSB have been shown to repress and induce transcription. In the latter case, also known as Break-Induced Transcription (BIT), the RNA-polymerase sits on the DSB and starts transcribing from there[45,47,48]. They can result from either programmed (i.e. DSB is a step of a biological procedure) or induced DSB (e.g. by endonucleases). The Ribonucleic Acid (RNA) resulting from a BIT (from now on referred to as BIT-RNA) were shown to synthesize different types of RNA, however, those BIT-RNA have regulatory functions. Whilst some regulate downstream genes in a pathway, others regulate the repair process of DSB[49].

Madabhushi R. et al shown that inducing DSB resulted in the overexpression of early-response genes, which consisted of transcription factors responsible for the activation of late-response genes. They proposed that the DSB of these BIT-RNA's, which localized in the promoter region, solved topological constraints, that would normally be solved by a topoisomerase[50].

Michelini F. et al presented a type of BIT-RNA, the damage-induced long non-coding RNA (dilncRNA). These are bidirectionally synthesized from the DSB, which can then form double-stranded RNA. Then, they are processed by Dicer and Drosha, two enzymes responsible for cleaving a double-strand RNA into smaller ones, the DNA damage response

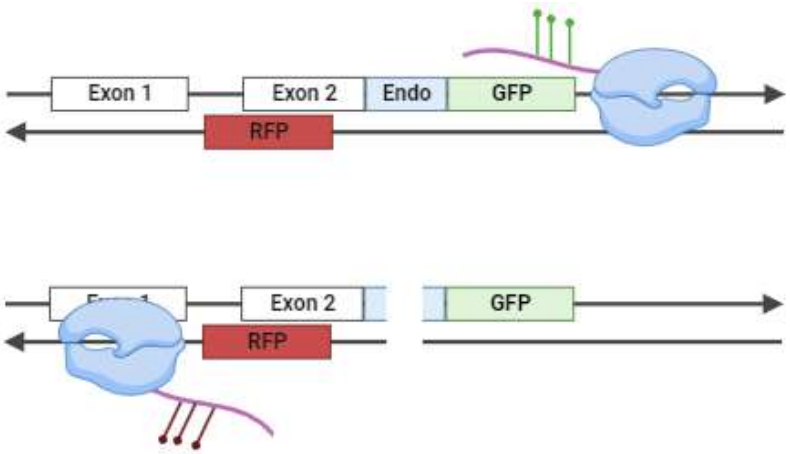


Figure 1.4: BIT detection at a single gene with live imaging. Schematic of single gene BIT detection, after DSB induction. Before the induction of DSB, the green fluorophore is expressed, showing transcription in the forward sense. When the DSB is present, it detects a red signal, corresponding to transcription in the reverse sense, where no transcription occurred prior to the DSB. Created with BioRender.com

RNA (DDRNA). DDRNA coupled with its precursor (dilncRNA) can then lead to a DDR focus on the DSB responsible for the dilncRNA[51,52].

Besides the previous reports of DSB inducing transcription, the term BIT was first introduced at *Vitor A. et al*[48], in which they were able to use the single-cell resolution to study this phenomenon. By using reporter genes they understood the sense of transcription. As presented in Figure 1.4, before the induction of DSB, the transcription was in the forward sense, since there was a green signal, however, when the DSB is present, the red fluorophore was expressed, showing transcription on the reverse sense, where no transcription occurred prior to the DSB. Using this technique, it was possible to witness BIT events and the change in RNA-polymerase II (Pol II) transcription sense, at a single-cell resolution.

1.4 Transcription Profile

Genes when expressed lead to the synthesis of RNA. However, cells present different gene expression, even though all cells in a healthy organism will have similar DNA (during cell division small mutations can occur). These differences result from the regulatory processes that influence gene expression on each cell[53].

Regulation happens on multiple levels: transcription, post-transcription, translation and post-translation. At the transcription level, one mechanism to alter gene expression is through the promoter[53,54]. The promoter is a region upstream of the Transcription Start Site (TSS), the standard start nucleotide of a transcript, with varying size and constitution. In eukaryotes, the promoter is composed of different sequence motifs that allow the transcription machinery to bind. TATA box and CpG islands are two of the most common promoter elements[54].

A TATA box is a canonical element of a promoter (i.e. only present in promoter and promoter-like regions) at position -25, relative to TSS, in humans, whose sequence is TATAWAWR (W is T or A, and R is A or G). These sequences are bound by TATA-binding protein (TBP), a protein required for transcription initiation in these promoters. The analyses presented in *Yella V. R. et al* in 2016 showed evidence that only 4% of human promoters contained a TATA box[55], while *Yang C. et al* in 2007 reported it as 24%[56]. Besides the low percentage of promoters containing a TATA box, it has been shown that some promoters contain TATA-like boxes, whose sequence besides being conserved (TTTCAA) is variable. Another important feature is that it can't be bound by TBP[54–57].

CpG islands are a common element found in the genome. These islands consist of CG dinucleotides that are highly methylated and have high mutation rates throughout the genome. However, CpG islands located within promoters are not methylated and have low mutation rates, being considered a noncanonical element of promoters. This type of promoter accounts for 70% of all human promoters, is common in housekeeping and tissue-specific genes, and typically lacks a TATA box[58].

While some genes only have one promoter for all transcripts, some regulate the transcription of different transcripts by having alternative promoters. An alternative promoter is a different region from which transcripts are synthesised, however, the term *"alternative promoters"* represents the presence of more than one promoter and does not give a preference when it comes to the choice of promoter. The usage of a promoter instead of others depends on two factors: a chromatin state that allows for transcription factor binding or the availability of transcription factors for binding on such promoters. This choice of different promoters for the same gene is an important step during the development and cell differentiation to a tissue-specific cell. The incorrect promoter choice was associated with cancer, neurological diseases and development disorders[59–61].

Besides promoters requiring specific sequences to start transcription, the constraints posed by it are not many, leading to many regions acting as promoters if the RNA-polymerase binds there. Spurious promoters refer to those regions of the genome which can unexpectedly act as promoters leading to unwanted transcription, also known as spurious transcription. Spurious transcription leads to non-functional RNA (protein synthesis, non-coding RNA functions), or transcripts whose function is different from the expected. This phenomenon increases transcription noise, which is a variation in gene expression as a result of stochasticity for supposedly identical cells, which can disrupt RNA-related pathways[62–64].

As a consequence of ageing and its hallmarks (Chapter 1.1.1), such as Genomic Instability, Epigenetic Alterations, and Loss of Proteostasis, the gene expression of many genes and transcripts changes[65]. Recently, spurious intragenic transcription, which can alter gene expression profiles, was associated with the ageing phenotype. In *Sen P. et al* they provided evidence of how enhancers can act as alternative promoters in intronic regions. These regions would have lower levels of methylation in both DNA and histones, than the remaining genome[63].

1.5 Aims of this Project

Our hypothesis relies on the possibility that the increase in DSB that is associated with ageing, could be responsible for the genomic profile alteration, as a result of BIT.

In order to confirm or reject such a proposal, it was required to construct a pipeline capable of identifying such events in a genome-wide approach and for not induced DSB. Coupled with transcriptional data from ageing datasets, it would open the doors to some candidates of interest for BIT-RNA's, which would then be confirmed *in vitro*.

This project aims to expand our knowledge in the fields of Ageing and their transcriptomic profile, spurious transcription, and the relevance of BITs in regular human physiology, by suggesting them as possible alternative promoters or spurious intragenic transcription activators.

2

Materials and Methods

2.1 Software and Tools

The analysis was performed using two programming languages: R and Shell. R (version 4.2.1) language and environment are routinely used for statistical data analysis and graphic visualization[66,67]. Rstudio (version 3.0.386) is an integrated development environment (IDE), which is used to facilitate the management of the R language[68]. Shell is, simultaneously, a command interpreter, due to its extensive GNU utilities, and also a programming language, since it allows the usage and combination of them. This project implemented Bash as the shell scripting interpreter[69].

Shell scripting mainly was used to run containers. A container image has a unique format, and is similar to a computer ready to run a specific application, without any root (i.e. administrative) privileges being needed. To allow it to run, an external tool capable of reading and executing that format is needed. For this project, containers were in a docker format and the tool was udocker[70].

Information about the R packages and the docker images that were utilized during this analysis, including a description of their functions and respective versions, are in Table 2.1 and Table 2.2, respectively.

The scripts developed for this project are available on the following link, as information about the pipeline, packages and datasets tested in: `https://github.com/IanLopesTeixeira/BIT_ageing`.

Table 2.1: R Packages

Package	Version	Description	Reference
ggplot2	3.4.2	Create elegant data visualization using grammar of graphics	[71]
dplyr	1.1.1	A grammar of data manipulation	[72]
rtracklayer	1.56.1	R interface to genome annotation files and the UCSC genome browser	[73]
DESeq2	1.36.0	Differential gene expression analysis based on the negative binomial distribution	[74]
edgeR	3.38.4	Empirical analysis of digital gene expression data in R	[75]
GEOquery	2.64.2	Get data from GEO	[76]
PCAtools	2.8.0	PCAtools: everything Principal Component Analysis	[77]

2.2 Large-scale profiles

For this project, it was used datasets for RNA-seq, DSB, genome and transcriptome sequence and annotation, TATA box, and CpG island information, from the following databases: Gene Expression Omnibus (GEO), GENCODE, Genome Browser and Genotype-Tissue Expression (GTEx). GEO is a free-access database in which researchers can submit datasets of their experiments, for others to use[78]. The GENCODE project aims to annotate the gene features of the human and mouse genome and provide it to the public[79]. UCSC Genome Browser Project is both a web-based tool and database, giving the chance to users to visualize multiple genomes, including the human and mouse, also retrieve data from those genomes, including the annotation and epigenetic markers[80]. GTEx is a public database whose goal is to understand tissue-specific gene expression and regulation[81].

Regarding RNA-seq, we used both raw and processed datasets. For the raw data two GEO datasets were selected: GSE26284 and GSE113957, corresponding to the MCF7 cell line (breast cancer cell line) and fibroblasts, respectively. The MCF7 dataset is composed of 2 samples, with no age relation. The fibroblasts dataset is an ageing dataset of 143 samples, from which 10 are of HGPS patients. The age range of this dataset goes from 1 to 96 years old. Further information about those samples is presented in Supplementary Table A.1. The processed transcriptomic dataset, which was from GTEx Analysis version 8, contained information from multiple tissues, with samples ranging from 20 to 79 years old. The data was in Transcripts per Million (TPM) (number of transcripts if there were a total of 1 million transcripts in the sample). From there was selected the information regarding *PALB2* isoforms, which were 6.

The DSB datasets used, GSE134798, GSE78172, GSE207714, and GSE124403, were available in GEO. However, each one was acquired by different techniques and was already processed, resulting in files with position for each DSB. GSE134798 data acquisition was with Break Labeling In Situ and Sequencing (BLISS), a high-resolution technique. GSE78172 was acquired with DSBcapture, and GSE207714 and GSE124403 with Break-seq, both techniques of lower resolution. Each dataset was from different cell lines: GSE134798 from MCF7, GSE78172 from Normal Human Epidermal Keratinocytes (NHEK), GSE207714 from MCF7 and MCF10A (breast tissue cell line), GSE124403 from GM06990 (lymphoblasts from an individual with Epstein-Barr virus). Further information about

Table 2.2: Docker Images

Container	Version	Alias	Software
ncbi/sra-tools	3.0.0	sra-tools	SRA-toolkit
biocontainers/fastqc	v0.11.9_cv8	fastqc	FastQC
ewels/multiqc	v1.14	multiqc	MultiQC
trinityctat/starfusion	1.10.1	star	STAR
clinicalgenomics/stringtie	2.2.1	stringtie	StringTie
argrosso/htstools	0.2.1	htstools	SAMtools, bedtools and Genome Browser Utilities
clinicalgenomics/gffcompare	0.11.2	gffcompare	GffCompare
nanozoo/pygenometracks	3.7–04632c7	pygenometracks	pyGenomeTracks

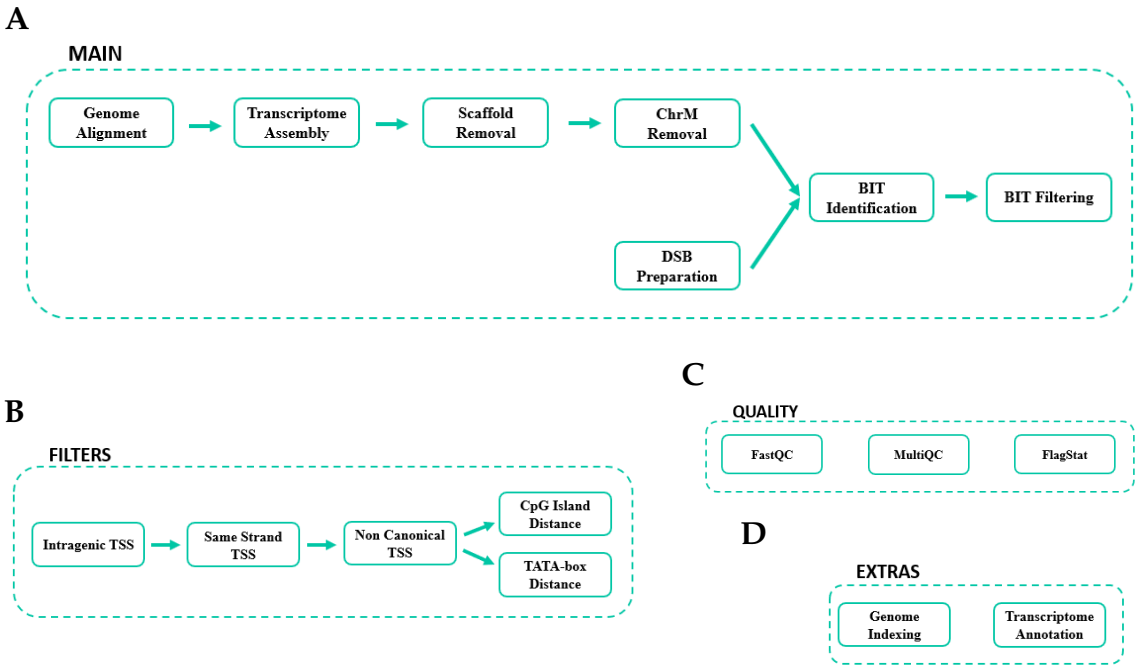


Figure 2.1: BIT identification Pipeline Schematics.
(**A**) Main pipeline. This scheme presents the major steps for the identification of BITs (**B**) Filters applied to further research of BITs of interest (**C**) All the tools used to ensure the quality of each file (**D**) Steps which were used through the pipeline to either complement with additional information or indexing files.

the samples is presented in Supplementary Table A.2.

The human genome, transcriptome and annotation were from build hg38 version 42, being provided by GENCODE.

The chromosome sizes, coordinate conversion from different human genome builds (hg19 to hg38), human CpG islands (32038 CpG Islands) and TBP locations (51 TBP), were available in Genome Browser. We considered the location of the TATA box as the same location of TBP.

2.3 Identification of BITs

To identify and select BITs genome-wide, we developed a pipeline as is presented in Figure 2.1. For the identification is required to associate a transcript to a DSB.

Regarding the transcripts, since BITs can produce novel transcripts, it was required to do transcriptome assembly. For the assembly, the RNA-seq data needed to fulfil the following requirements: 1) more than 60 Million (M) reads per sample, to deal with low coverage transcripts; 2) being strand-specific, for a more accurate assembly and to compare the BIT-RNA's to the known genome annotation; 3) polyadenylated RNA, to reduce the noise and focus on protein-coding RNA's. The transcriptome assembly was performed with StringTie, a tool designed for it when genome annotation is available and

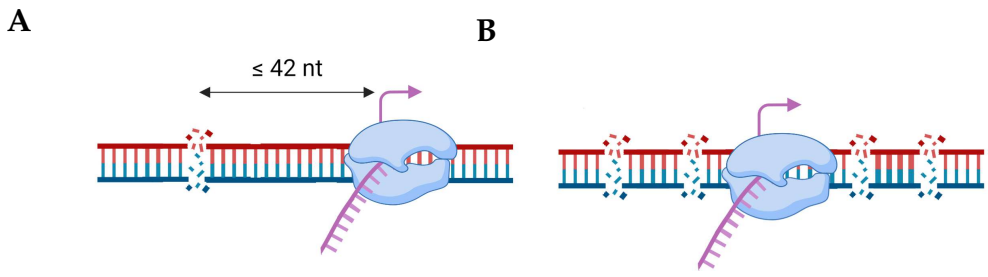


Figure 2.2: BIT identification models.
(**A**) Upstream Model. It is used for high-resolution DSB identification techniques, by identifying BITs as transcripts whose TSS is at 42 nt downstream of the DSB, upmost (**B**) DSBome Model. For low-resolution DSB identification techniques, considering BITs all transcripts whose TSS is inside the DSB hotspot region

of confidence, just like in model organisms as *Homo sapiens*, reducing time consumption while keeping accuracy[82]. The way it uses that information is by aligning the RNA-seq to the genome, which was aligned with Spliced Transcripts Alignment to a Reference (STAR) (Figure 2.1**A**)[83,84]. When running the alignment no multimapped reads were considered for a more accurate assembly. After the transcriptome assembly, it was trimmed by removing scaffolds and transcripts from the mitochondrial chromosome (ChrM), since we decided to not focus on BITs there.

The DSB data didn't go through any particular processing, since each dataset was in a particular format. So, each file was processed accordingly to output a bed format file. This would contain information of each DSB, independently of technique resolution, with coordinates, according to the hg38 build, for each DSB, with trimmed-out information of scaffolds and ChrM The hits considered are independent of expression level, although, they can't have been treated with the aim to induce those DSB.

The identification of BIT events was based on the relative position between transcripts in the assembled transcriptome, and the DSB. Depending on technique resolution, two models were developed (Figure 2.2): the Upstream model, for high resolution, and the DSBome model, for low resolution. By the first method, are BIT-RNA's those transcripts whose TSS have an upstream DSB and their distance is, at most, 42 nucleotide (nt), which is the Pol II footprint (Figure 2.2**A**). The DSBome model identifies transcripts as BIT-RNA's when their TSS is inside a DSB region (Figure 2.2**B**). Finally, those identified as BIT were dubbed BIT candidates (i.e. possible BIT of interest)

Throughout the pipeline, the results and input files underwent a quality checkup (Figure 2.1**C**).

2.4 Characterization of BITs

The BIT candidates were classified based on their relative position to the genome annotation, after which only one of the groups was selected for further filtering (Figure 2.1**B**).

First, they were classified into Intragenic or Intergenic, based on whether the BIT-RNA candidate TSS intersected an annotated gene or not, respectively.

After that, the Intragenic group of BIT candidates was filtered between the same strand or opposite strand, if the TSS intersected an annotated transcript with the same sense as the BIT-RNA candidate or not. In this case, belonging to the opposite strand group meant not intersecting any transcript with the same sense of transcription, while the other transcripts intersect in the same sense annotated transcripts, and simultaneously, intersect in the opposite sense.

Finally, those considered the same strand were classified as Canonical or Non-Canonical cases. The definition of canonical or not is through the Ensembl tag, *Ensembl Canonical*[85]. The first exon of the annotated transcripts tagged with it was selected and considered the Canonical set. If a TSS of a BIT-RNA would be located in any of these regions, would be considered Canonical, whilst the remaining Non Canonical.

To look for potential promoters in the region of the BIT candidates considered Intragenic, Same Strand, and Non-Canonical, simultaneously, it was looked for the closest CpG island and TATA box to the transcript.

2.5 Differential BIT Expression

The differential expression of BITs was performed for the fibroblast dataset. Using the initial alignment for each sample, and the merged transcriptome (prior to any characterization and filtering or BIT identification), StringTie estimates transcript expression and coverage for each sample. Next, a script designed for predicting the read counts based on the results from StringTie estimation was run for all samples.

Prior to the differential BIT expression, we did an unsupervised analysis, to understand if BIT gene counts could explain age differences. An unsupervised analysis aims to explore how data is divided by transcript counts, without sample labelling. Principal Components Analysis (PCA) organizes in a two-dimensional graphic, the multiple ways of data being divided based on two components at a time. In this case, each component obtained will show how specific transcripts can organize data. The higher the principal component, the more variability of the dataset it explains[86].

The differential analysis was performed using DESeq2. The significance value here corresponded to a p-value adjusted of 0.05 (5%). This value was established for statistical purposes. From the results obtained, the use of p-value was not the best choice, since the higher the number of analyses/samples/genes, the higher the number of false positives. By rectification of the p-values based on the number of analyses made, we chose to select them by p-value adjusted[87]. The differential analysis was between a young group and an old group.

Both the unsupervised analysis and differential BIT expression were performed for the healthy fibroblast samples (133 samples).

2.6 PALB2 Expression throughout Ageing

For analysis of *PALB2* BIT expression performed a Spearman test, with the fibroblast data. The Spearman test is a non-parametric correlation test, which is a test that tries to measure how is the relation between two variables and how they influence each other when the data doesn't have specific parameters or distribution or is unknown[88]. This test allowed us to understand if any of the *PALB2* BITs expression would have any association with ageing.

Regarding *PALB2* isoforms and GTEx data, it was performed an ANOVA and Tukey's Honestly Significant Difference test. The ANOVA test is used to test for statistically significant differences between three or more groups, however, it doesn't point to which groups are different from each other. The Tukey's Honestly Significant Difference test usually follows an ANOVA test whose groups have proven to be significantly different and aims to assess the statistical difference between all groups, two at a time, providing evidence of how each group contributes to the overall difference. In this project, we compared the expression of each isoform per tissue throughout ageing. Each ANOVA was aimed to understand if ageing and transcript expression were different for each age group. The age groups were in ranges of 10 years from 20 to 79 (i.e. 20-29, 30-39, and so on).[89]

3

Results

3.1 Genome Wide Identification of DSB

To understand molecular phenomena such as BIT, it would be required to use genome-wide detection techniques to identify both RNA and DSB.

In the first case, a pipeline is developed and established. RNA-seq or high-throughput transcriptome sequencing allows the detection of a transcription profile at a cell or group of cells at a specific time and conditions, allowing for further analysis.

However, for the detection of DSB, there is no consensus yet on a protocol to identify them and how to process raw data. Each developed technique aimed to identify them, such as BLISS, DSBCapture, Break-seq, and more, has its own particularities: number of cells, resolution, and cellular conditions to use. The resolution of a technique is defined as the size of an identified DSB, the higher the resolution the smaller DSB identified. When it's very low, the technique tends to identify enriched regions of DSB, instead of single hits. Nonetheless, each of those techniques has at least two common steps: the DSB ends need to be blunt, depend on an adaptor ligation for further identification[90–93].

3.2 Genome-wide detection of BITs is possible in MCF7

The method for identification of BIT events genome-wide (Figure 2.1) was developed and tested with MCF7 data. Detailed information regarding the MCF7 data is available in Supplementary Tables A.1 and A.2.

Unlike mentioned in Chapter 2.3, the RNA-seq samples of MCF7 had less than 60 M reads each. To solve it, they were aligned to the genome as a single sample instead of separately. This was possible because they were biological replicates and would have similar gene expression profiles. This aligned 152.4 M reads, instead of the 215.1 M, when considered the default multimapped feature of STAR (considers reads until a multimap alignment of 10 loci, at most). Next, the transcriptome was assembled and trimmed, leaving 46781 transcripts (Table 3.1).

The DSB data (GSE134798) was processed by concatenating all replicates into a single

file, removing duplicate hits between samples, and DSB of scaffolds and ChrM information, resulting in 476366 hits. Most of the remaining DSB had 1 nt of size (Supplementary Figure A.2).

Here, the BIT identification was performed with the Upstream Model (Figure 2.2**A**), which identified 2060 BIT-RNAs. This method allowed to point out such events independently of the strand (Figure 3.1) and even if multiple isoforms are present (Supplementary Figure A.1).

The characterization and following filtering of BIT candidates, as presented in Chapter 2.4 (Figure 2.1**B**), resulted in the values at Table 3.1. Examples of each category can be seen in Figure 3.2, for those of interest, and in Figure 3.3, for those removed (Supplementary Figures A.5 and A.6, for zoom on BIT location, respectively).

CpG islands and TATA box, two promoter elements, were evenly distributed around the TSS of the BIT-RNA candidates. Most of the transcripts intersected any of the promoter elements (Figure 3.4).

These results provided enough confidence to identify BITs in an ageing dataset.

3.3 Age-related fibroblasts show less BITs than MCF7

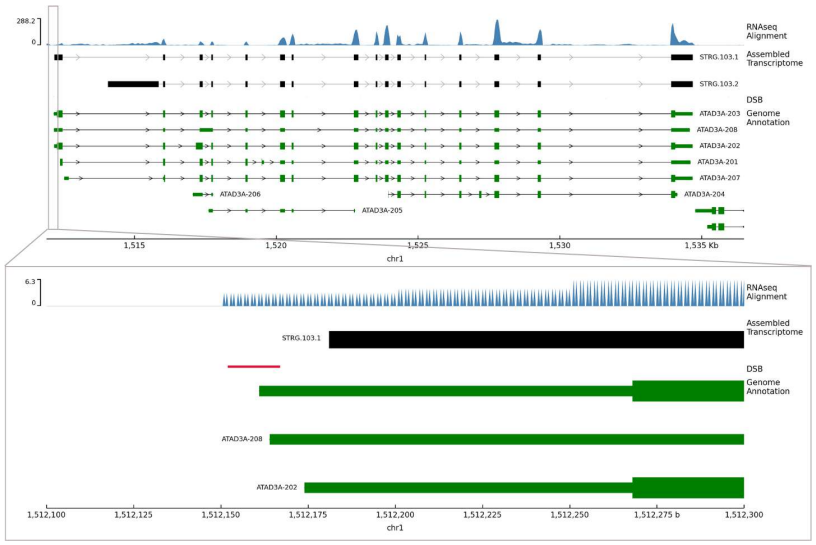
Using the pipeline tested in the previous chapter (Chapter 3.2), we identified BITs in age-related fibroblasts.

The fibroblast RNA-seq doesn't fulfil two of the established requirements. Instead of a polyadenylated dataset, it was total RNA and none of the samples had 60 M reads, rather most of them had less than 25 M reads. Unlike with the MCF7 RNA-seq, all samples were aligned separately, since they are not replicates (biological or technical). This led to 143 different transcriptomes, one per sample. All of them had both scaffolds and ChrM information removed. Next, they were merged, using the StringTie feature, into a single transcriptome. When performing the merging, the genome annotation needs to be provided, which adjusts the assembled transcripts to match annotated transcripts, in case

Table 3.1: MCF7 Transcripts

Step	Nº Transcripts
Transcriptome Assembly	46931
Scaffold Removal	46787
ChrM Removal	46781
BIT identification	2060
BIT-RNA intragenic TSS	1829
BIT-RNA intergenic TSS	231
BIT-RNA same strand	1702
BIT-RNA not same strand	127
BIT-RNA non-canonical	1636
BIT-RNA canonical	66

A



B

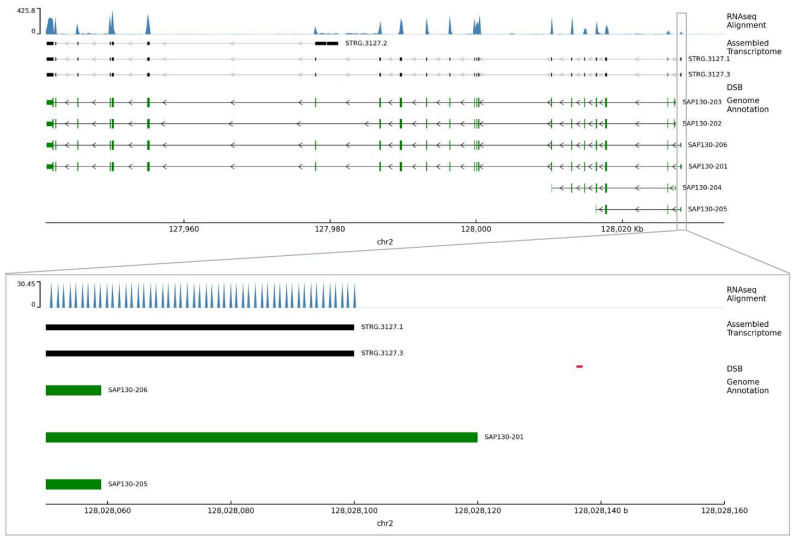


Figure 3.1: The pipeline was able to identify BIT-RNA candidates in MCF7 using the upstream model.
After assembled transcriptome trimming and DSB file preparation, the identification of BITs relied on a DSB being at 42 nt upstream of a TSS upmost. The representations highlight regions of BIT candidates and present the RNA-seq alignment of those samples, corresponding assembled transcripts, DSB, and annotated transcripts. Zoom of the BIT region is provided. Both (**A**) forward and (**B**) reverse strand cases were identified. Kb, Killobase; b, base

they are very similar.

While testing the pipeline, both DSB and RNA-seq were from the same cell line (MCF7), which was not the case with fibroblasts. It was decided to understand which DSB are common to multiple tissues to then be applied for BIT identification in fibroblasts. Those DSB constitute the DSBome. The DSBome assembly is by selecting which DSB are identified in two or more cell lines. Here, we did it with low-resolution techniques, which led to a DSBome of regions of DSB common to at least two cell lines (Figure 3.5). The

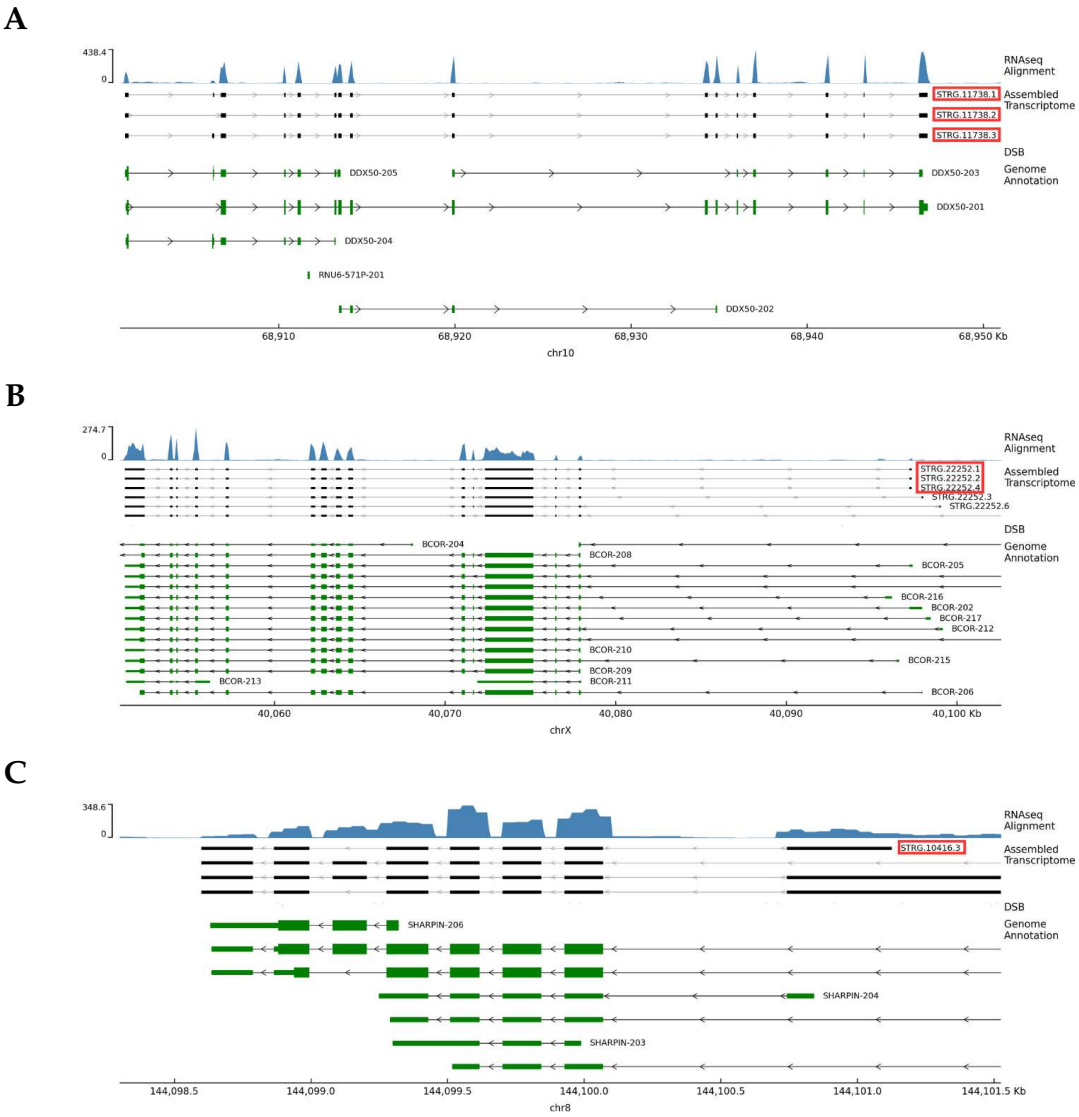


Figure 3.2: Filters select several BIT candidates identified based on the TSS position of the assembled transcripts and annotated transcripts.
Following BIT identification, certain filters were applied to choose the smaller group depending on the position of a BIT-RNA TSS and its relation to an annotated transcript. The representations highlight regions of BIT candidates and present the RNAseq alignment of those samples, corresponding assembled transcripts, DSB, and annotated transcripts. Their application was sequential, and their order is the one presented in Figure 2.1**B**, corresponding to **A**, **B** and **C**, respectively. This panel is a counterpart of Figure 3.2 and should be seen simultaneously. The red squares on the images highlight the ones who were selected in that filter. **(A)** the first selection was the BITs whose TSS was inside an annotated gene **(B)** then if the BIT-RNA sense of transcription was the same as the annotated transcript and **(C)** finally if the TSS location it's not in the first exon of a Canonical Transcript. Kb, Killobase; b, base

cell lines used were NHEK (GSE78172), MCF7 (GSE207714), MCF10A (GSE207714), and GM06990 (GSE124403) (Supplementary Figure A.3). Each of the datasets was processed prior to DSBome assembly, which was composed of 3715 DSB. The contribution of each cell line is presented in Supplementary Figure A.4.

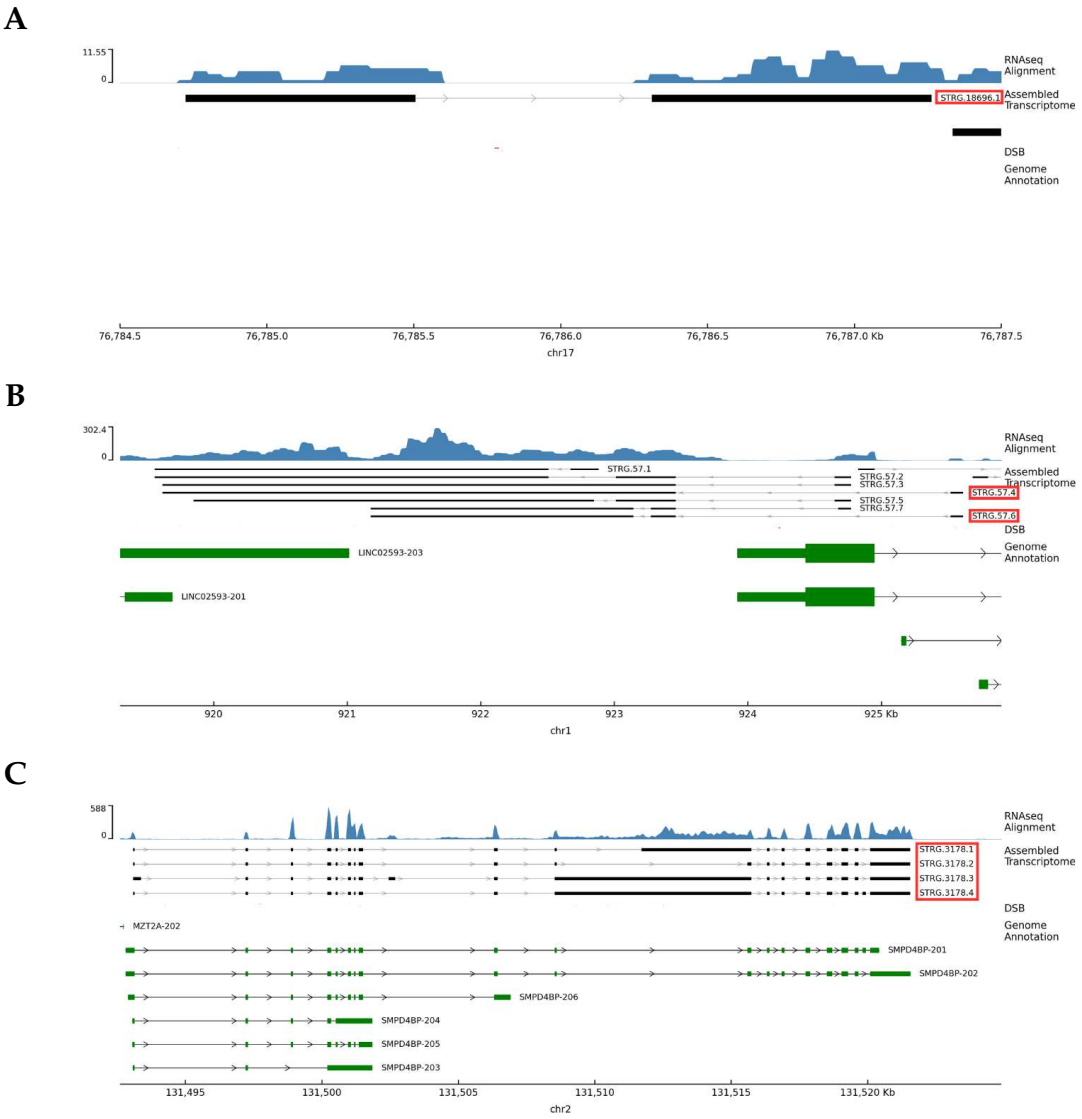


Figure 3.3: Filters remove several BIT candidates identified based on the TSS position of the assembled transcripts and annotated transcripts.
Following BIT identification, certain filters were applied to remove those whose characteristics were not of interest. The filtering depended on the position of a BIT-RNA TSS and its relation to an annotated transcript. The representations highlight regions of BIT candidates and present the RNAseq alignment of those samples, corresponding assembled transcripts, DSB, and annotated transcripts. Their application was to the ones which were not removed in the previous filter, corresponding to **A**, **B** and **C**, respectively. This panel is a counterpart of Figure 3.2 and should be seen simultaneously. The red squares on the images highlight the ones who were removed from that filter. **(A)** firstly, those whose TSS was outside an annotated gene **(B)** then if the BIT-RNA sense of transcription was not the same as the annotated transcript, which was different from selecting the ones on the opposite strand, since being the same strand and opposite strand simultaneously, and in those cases, they are not removed, and **(C)** finally if the TSS location is in the first exon of a Canonical Transcript. Kb, Killobase; b, base

Here, the identification of BIT candidates was with the DSBome model, since DSBome is of low resolution DSB (Figure 2.2). Even though the number of transcripts is higher than with MCF7, the total number of BIT candidates was much lower. The categorization

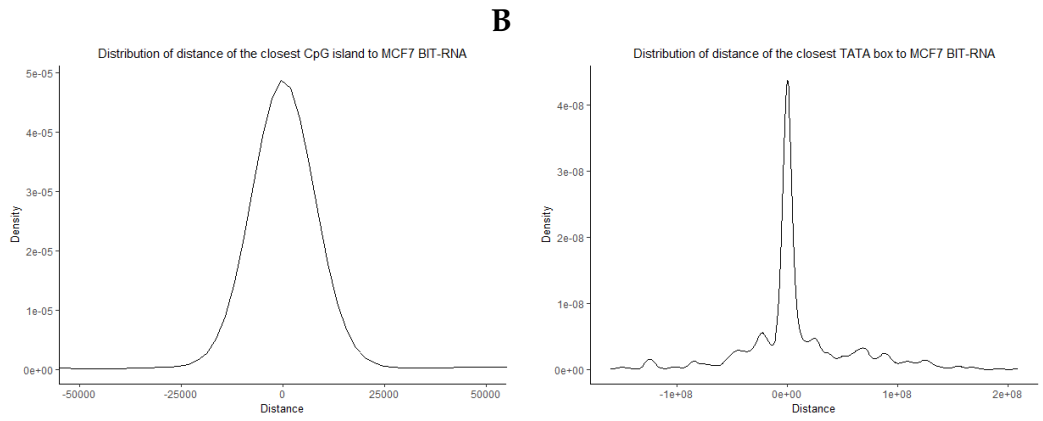


Figure 3.4: Most MCF7 Non-Canonical BIT-RNA's are transcribed from promoter-like regions. The majority of the Non-Canonical BIT-RNA have the closest **(A)** CpG Island and/or **(B)** TATA box near their TSS. The proximity of one of these elements and a TSS suggests the presence of a promoter-like region. The Distance axis in each plot is in nt, the negative values represent upstream and the positive downstream.

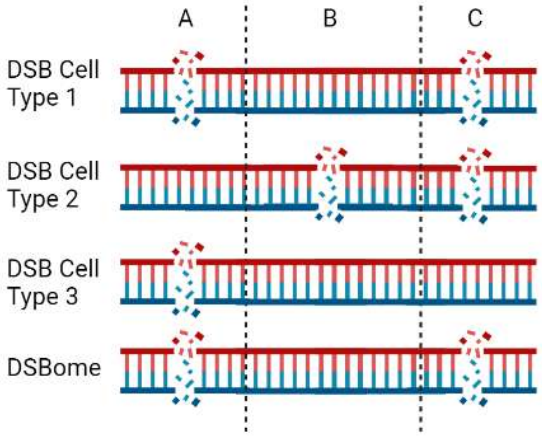


Figure 3.5: The formation of a DSBome is possible when a DSB region is present in two or more cell lines.
The DSBome is composed of DSB which are present in two or more cell lines. This process is done after all DSB files from multiple lines are processed

of them left some categories with 0 and 1 BIT-RNA, while the BIT-RNA non-canonical was about half of the initial number (Table 3.2).

Regarding the CpG island and TATA box to TSS, most of them were close to one of the two elements, just like with MCF7 results (Figure 3.6).

From the BITs identified as non-canonical, most of them were annotated, being either related to repair mechanisms, such as *PALB2*, or as long non-coding RNA (lncRNA), just like ENSG00000289277, ENSG00000283061, ENSG00000263594, LOC101928626, C21orf62-AS.

Table 3.2: Fibroblasts Transcripts

Step	Nº Transcripts
Transcriptome Assembly and Merge	2937306
BIT identification	62
BIT-RNA intragenic TSS	61
BIT-RNA intergenic TSS	1
BIT-RNA same strand	61
BIT-RNA not same strand	0
BIT-RNA non-canonical	32
BIT-RNA canonical	29

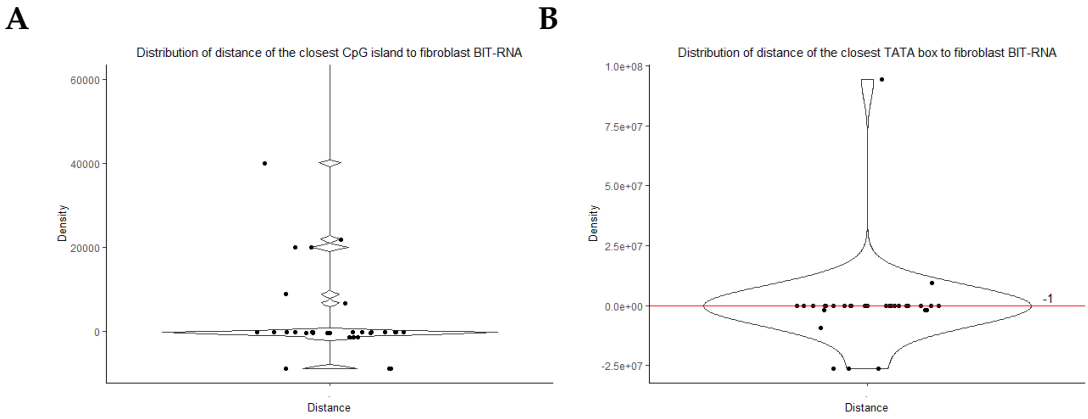


Figure 3.6: Non-Canonical BIT-RNA's from fibroblasts start at promoter-like regions. The majority of the Non-Canonical BIT-RNA have the closest **(A)** CpG Island and/or **(B)** TATA box near their TSS. The proximity of one of these elements and a TSS suggests the presence of a promoter-like region. The Distance axis in each plot is in nt, the negative values represent upstream and the positive downstream. The red horizontal line in **(B)** is where most BIT-RNA's have a TATA box, 1 nt upstream

3.4 BITs are not differentially expressed during ageing

Upon selection of the BIT-RNA non-canonical cases, we tried to understand how they could affect the ageing process, for unsupervised analysis and differential BIT expression. The non-canonical BIT-RNA couldn't distinguish younger individuals from older ones. We also tried to use all BIT candidates, and the conclusion was similar (Figure 3.7). After that, we looked if any of the BIT-RNA non-canonical cases were differentially expressed between the young and old groups. The young group is all individuals below 65 years old and the remaining in the old group.

No transcript showed differential expression between both groups. Then we rearranged the group composition to more extreme constitutions. The young group became individuals from 20 to 40 years old, and the old group from 70 to above, which led to the same outcome.

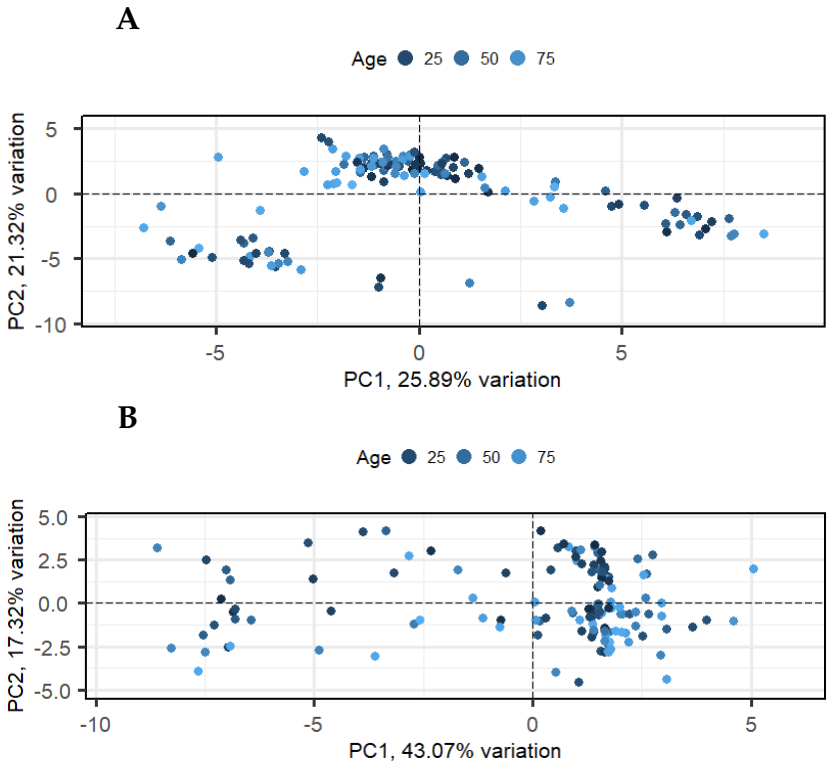


Figure 3.7: The BIT-RNA's can't distinguish younger from older individuals.
Both PCA performed with **(A)** all BIT-RNA candidates and **(B)** Non-Canonical BIT-RNA, could not show a clear distinction associated with ageing, since dark blue (younger) and light blue (older) points are in the same regions. The darker the blue the younger the individual

3.5 BIT located in PALB2 doesn't change with ageing

One of the BIT-RNA non-canonical cases corresponded to a *PALB2* transcript. *PALB2* is a gene associated with repair mechanisms of DSB (see Chapter 1.2.2). Based on previous findings, BIT have been associated with repair mechanisms of DSB (see Chapter 1.3).

The BIT-RNA's of *PALB2* started on the fourth exon (Figure 3.8), corresponding to existing isoforms -ENST00000697375 and ENST00000697378. Next, we tried to see if there was any correlation between ageing and any of the isoforms using a Spearman test, however, we only used one isoform, ENST00000697375, since the other had very low values of expression. Nevertheless, the isoform presented no correlation with the ageing process (Figure 3.9**A**).

3.6 Switch in PALB2 isoform from functional protein to non-functional with ageing in multiple tissues

After looking for all the different transcripts available for *PALB2*, it was seen that all those starting around the DSB region or downstream of it (since it is in the reverse strand) would be Non-sense mediated decays, with exception of one. PALB2-204 (ENST00000566069

Figure 3.8: PALB2 isoforms identified as BIT-RNA's.
The pipeline identifies PALB2 isoforms, with the DSBome model, as BIT-RNA's at the fourth exon of the PALB2 gene. The representation highlights the region of a BIT-RNA and corresponding assembled transcripts, DSB, CpG islands, TATA box, and annotated transcripts. Zoom of the BIT region is provided. Kb, Killobase; b, base

version 5) being one of the smaller isoforms of PALB2 transcripts seemed to be only able to produce a protein. Nonetheless, such protein seemed to not be functional.

To understand if this isoform would change also with ageing, we looked into GTEx data for the *PALB2* transcripts. From the seventeen transcripts associated with *PALB2*, only six isoforms were available. From those, and in multiple tissues, it was possible to see that in some tissues at least one of two phenomena was happening: 1) the Ensembl Canonical isoform, PALB2-201/ENST00000261584, decreases in expression with age; 2) PALB2-204 increases with ageing. In Figure 3.9**B**, both occur: the PALB2-201/ENST00000261584 decreases (ANOVA p-value: 1,27 x 10^{-5}), while PALB2-204/ENST00000566069 increases (ANOVA p-value: 8,35x10^{-10}). Both transcripts had Tukey test p-value significant when comparing one age group to either 60-69 or 70-79.

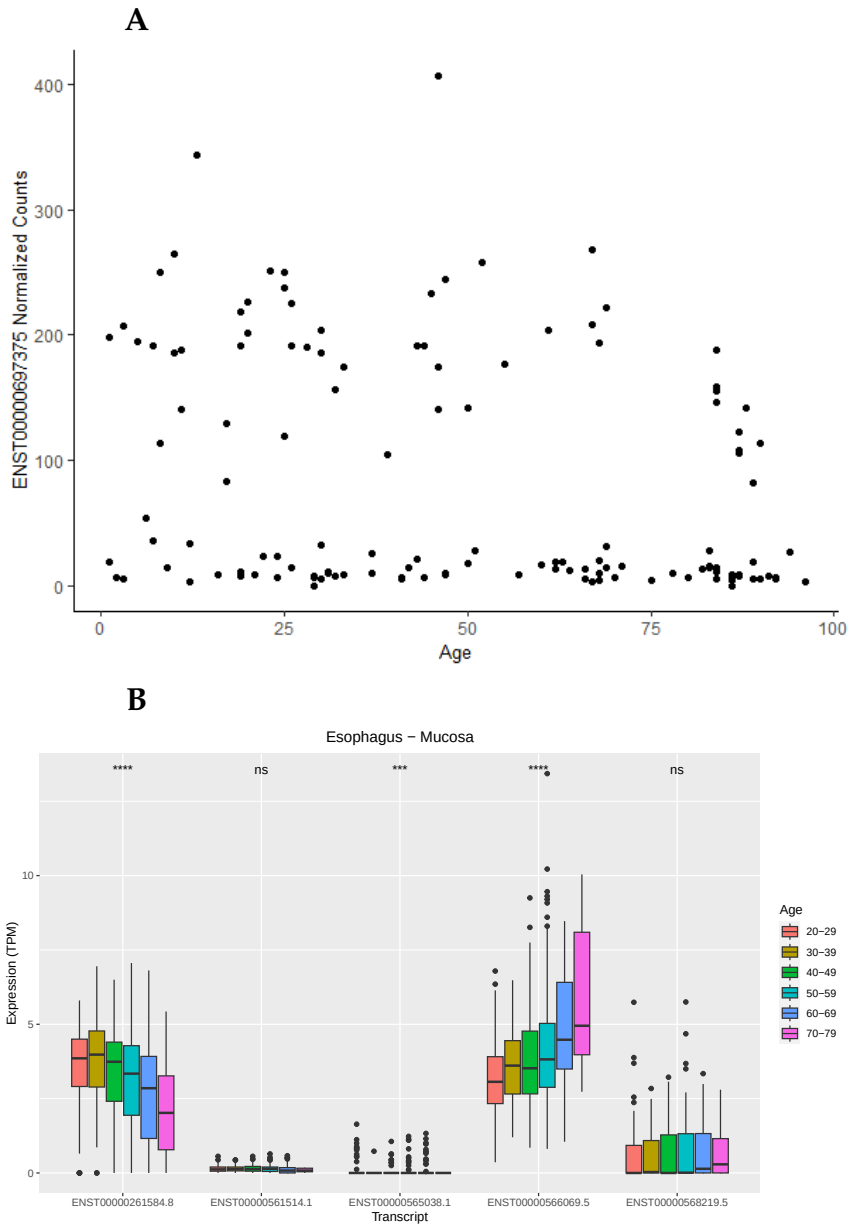


Figure 3.9: Different PALB2 isoforms have different correlation with ageing.
(A) The PALB2 BIT-RNA isoform has no correlation with aged-fibroblasts **(B)** The Canonical isoform is less expressed in older samples, while the isoform of the non-functional protein is more expressed

4

Discussion and Conclusion

Break-induced transcription occurs when the RNA-polymerase sits on the double-strand break and starts transcription from that point. As a consequence of ageing, the transcriptomic profile changes and there's an increase of spurious transcription. On the other hand, one of the hallmarks of ageing is the increase of genomic instability, which means a higher number of DSB. This raised the question of how BIT could be a source of spurious transcription, as a consequence of more DSB.

BITs were identified in a genome-wide approach, by two models (Figure 2.2), and characterized based on their relative position to the genome annotation, independently of the size of the DSB hits. The MCF7 BIT-RNA's were more specific to the cell line since we used the same cell line for RNA-seq and DSB datasets and the fibroblast ones would be more common, genome-wide, as a result of DSBome. Nonetheless, none of the BIT candidates was present simultaneously in MCF7 and fibroblast results. This could be a consequence of the MCF7 cell line not contributing much to the common BITs for multiple tissues, however, it is not likely, since we used the MCF7 dataset for the construction of the DSBome. Another explanation is the low number of reads per sample in the fibroblast dataset, which would reduce the low coverage transcripts identified, such as spurious transcripts, leading to fewer BIT candidates. This can also explain why the number of BIT in fibroblast is much lower than in MCF7. Although, both cell lines, MCF7 and fibroblasts, had most of their BITs of interest co-localize promoter elements (CpG Islands and TATA box) with their transcript TSS, supporting the hypothesis of BITs occurring at alternative promoter regions, which includes spurious promoters[63].

Even though BITs were identified in age-related fibroblasts, no BIT-RNA was differently expressed for different ages and age groups. Nonetheless, most of those BITs supported previous findings, while some were transcription factors[50], others lncRNA[51], and one a gene responsible for DSB repair mechanism (*PALB2*). There is still not enough information regarding the BIT and their transcripts to understand if they have a similar impact on the development and differentiation of cells, or even ageing in other species. However, we conclude that Break-Induced Transcription does not influence the Ageing process and phenotype.

Regarding the *PALB2*, an immune suppressor and mediator in the repair mechanism of DSB, although it didn't correlate in any way with Ageing and BITs, it was proven that in some tissues it has some association with ageing. The GTEx dataset shows how the canonical form (PALB2-201) was less expressed in older individuals, while the short isoform (PALB2-204) increased. The protein of PALB2-204 is smaller than the canonical one, having only a repeat domain, meaning that it loses functions in repair since it can't be bound and recognised by RAD51, BRAC2 and BRAC1, as provided by UniProt, a database with information regarding proteins, including its domains[94].

In summary, it was possible to develop methods for the identification of BIT events genome-wide, relying on public data. We also showed how promoter-like regions could be important for BIT events to happen, as BIT-RNA's TSS was close to promoter elements. Finally, we concluded that there is no evidence that BIT could affect the ageing process, proposing this phenomenon to be specific or certain events, such as DSB repair or pathologies.

4.1 Limitations and Future Perspectives

In this project, some limitations were present throughout the analysis regarding the methodology and dataset, which affected our conclusions.

In the case of the dataset, the DSB suffer from the lack of an established protocol for their identification. Besides, in the case of fibroblasts, the necessity of giving rise to a DSBome could hinder the results by lowering the number of matches between DSB and transcripts. Another problem was not having DSB datasets from aged samples coupled with correspondent RNA-seq. Regarding the transcriptomic information, the biggest issue was the fibroblast dataset not complying with our requirements, and having a low number of reads.

In future works, it should generate data, RNA-seq and DSB information, both for aged samples and coupled together. This would improve the identification of age-related BIT and if at any point they are important. Besides, BIT events should be studied for pathologies and see how they could affect them. It would also be important to create conditions in which those BITs could be of interest, to see if they are differentially expressed. In case unannotated or poorly annotated transcripts are identified, they should be then tested *in vitro* and *in vivo* to better understand their cellular function, and in which moments or pathways BITs could be more prevalent.

Another limitation was the *PALB2* information. As of July 2023, there was an update on the GENCODE annotation, in which the PALB2-204 isoform and protein are similar to the canonical ones. However, it still had a support level of 5, showing that there is not much evidence of the structure of this isoform. Further research should be in understand how is the structure of this isoform, and if it has any relation to ageing.

Bibliography

1. Lourenço, J. M. *The NOVAthesis LaTeX Template User's Manual* NOVA University Lisbon (2021). https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).
2. Da Costa, J. P. *et al.* A synopsis on aging-Theories, mechanisms and future prospects. *Ageing research reviews* **29,** 90–112 (2016-08) (cit. on p. 1).
3. Lloyd, D., Aon, M. A. & Cortassa, S. Why Homeodynamics, Not Homeostasis? *The Scientific World JOURNAL* **1,** 133–145. ISSN: 1537-744X (2001) (cit. on p. 1).
4. Harman, D. The aging process. *Proceedings of the National Academy of Sciences* **78,** 7124–7128. ISSN: 0027-8424 (11 1981-11) (cit. on p. 1).
5. RATTAN, S. I. Biogerontology: The Next Step. *Annals of the New York Academy of Sciences* **908,** 282–290. ISSN: 00778923 (1 2000-01) (cit. on p. 1).
6. Rattan, S. I. S. Biogerontology: research status, challenges and opportunities. *Acta Biomed* **89,** 291–301 (2 2018) (cit. on p. 1).
7. Rattan, S. & Clark, B. Understanding and Modulating Ageing. *IUBMB Life (International Union of Biochemistry and Molecular Biology: Life)* **57,** 297–304. ISSN: 1521-6543 (4-5 2005-05) (cit. on p. 1).
8. Yousefzadeh, M. *et al.* DNA damage—how and why we age? *eLife* **10.** ISSN: 2050-084X (2021-01) (cit. on pp. 2, 7).
9. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. Hallmarks of aging: An expanding universe. *Cell* **186,** 243–278. ISSN: 00928674 (2 2023-01) (cit. on pp. 2–6).
10. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153,** 1194–1217. ISSN: 00928674 (6 2013-06) (cit. on pp. 2–4).
11. Vujin, A. & Dick, K. The information theory of aging: Hacking immortality? *Health Science Inquiry* **11,** 148–154. ISSN: 2562-7791 (1 2020-08) (cit. on p. 3).

12. Weinert, T., Kaochar, S., Jones, H., Paek, A. & Clark, A. The replication fork's five degrees of freedom, their failure and genome rearrangements. *Current Opinion in Cell Biology* **21,** 778–784. ISSN: 09550674 (6 2009-12) (cit. on p. 3).

13. Povedano, J. M. *et al.* Therapeutic effects of telomerase in mice with pulmonary fibrosis induced by damage to the lungs and short telomeres. *eLife* **7.** ISSN: 2050-084X (2018-01) (cit. on p. 3).

14. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics* **19,** 371–384. ISSN: 1471-0056 (6 2018-06) (cit. on p. 3).

15. Duan, R., Fu, Q., Sun, Y. & Li, Q. Epigenetic clock: A promising biomarker and practical tool in aging. *Ageing Research Reviews* **81,** 101743. ISSN: 15681637 (2022-11) (cit. on p. 3).

16. Bella, E. D. *et al.* The unfolded protein response in amyotrophic later sclerosis: results of a phase 2 trial. *Brain* **144,** 2635–2647. ISSN: 0006-8950 (9 2021-10) (cit. on p. 3).

17. Feng, Y., He, D., Yao, Z. & Klionsky, D. J. The machinery of macroautophagy. *Cell Research* **24,** 24–41. ISSN: 1001-0602 (1 2014-01) (cit. on p. 4).

18. Alsaleh, G. *et al.* Autophagy in T cells from aged donors is maintained by spermidine and correlates with function and vaccine responses. *eLife* **9.** ISSN: 2050-084X (2020-12) (cit. on p. 4).

19. Mattison, J. A. *et al.* Caloric restriction improves health and survival of rhesus monkeys. *Nature Communications* **8,** 14063. ISSN: 2041-1723 (1 2017-01) (cit. on p. 4).

20. Xu, M. *et al.* Senolytics improve physical function and increase lifespan in old age. *Nature Medicine* **24,** 1246–1256. ISSN: 1078-8956 (8 2018-08) (cit. on p. 4).

21. Kumari, R. & Jat, P. Mechanisms of Cellular Senescence: Cell Cycle Arrest and Senescence Associated Secretory Phenotype. *Frontiers in Cell and Developmental Biology* **9.** ISSN: 2296-634X (2021-03) (cit. on p. 4).

22. Oh, J., Lee, Y. D. & Wagers, A. J. Stem cell aging: mechanisms, regulators and therapeutic opportunities. *Nature Medicine* **20,** 870–880. ISSN: 1078-8956 (8 2014-08) (cit. on p. 5).

23. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* **14,** R115. ISSN: 1465-6906 (10 2013) (cit. on p. 5).

24. Ma, S. *et al.* Heterochronic parabiosis induces stem cell revitalization and systemic rejuvenation across aged tissues. *Cell Stem Cell* **29,** 990–1005.e10. ISSN: 19345909 (6 2022-06) (cit. on p. 5).

25. Minhas, P. S. *et al.* Restoring metabolism of myeloid cells reverses cognitive decline in ageing. *Nature* **590,** 122–128. ISSN: 0028-0836 (7844 2021-02) (cit. on p. 5).

26. De Cedrón, M. G. & de Molina, A. R. in *Precision Medicine for Investigators, Practitioners and Providers* 291–299 (Elsevier, 2020) (cit. on p. 5).

27. Haran, J. P. & McCormick, B. A. Aging, Frailty, and the Microbiome—How Dysbiosis Influences Human Aging and Disease. *Gastroenterology* **160,** 507–523. ISSN: 00165085 (2 2021-01) (cit. on p. 5).

28. Lin, S.-W. *et al.* Lactobacillus plantarum GKM3 Promotes Longevity, Memory Retention, and Reduces Brain Oxidation Stress in SAMP8 Mice. *Nutrients* **13,** 2860. ISSN: 2072-6643 (8 2021-08) (cit. on p. 5).

29. Schnabel, F., Kornak, U. & Wollnik, B. Premature aging disorders: A clinical and genetic compendium. *Clinical Genetics* **99,** 3–28. ISSN: 0009-9163 (1 2021-01) (cit. on pp. 5, 6).

30. Cisneros, B., García-Aguirre, I., Ita, M. D., Arrieta-Cruz, I. & Rosas-Vargas, H. Hutchinson-Gilford Progeria Syndrome: Cellular Mechanisms and Therapeutic Perspectives. *Archives of Medical Research* **54,** 102837. ISSN: 01884409 (5 2023-07) (cit. on p. 5).

31. Shin, J.-Y. & Worman, H. J. Molecular Pathology of Laminopathies. *Annual Review of Pathology: Mechanisms of Disease* **17,** 159–180. ISSN: 1553-4006 (1 2022-01) (cit. on p. 5).

32. Zhang, H. & Cao, K. Mechanisms of genome instability in Hutchinson-Gilford progeria. *Frontiers in Biology* **12,** 49–62. ISSN: 1674-7984 (1 2017-02) (cit. on p. 6).

33. Travers, A. & Muskhelishvili, G. DNA structure and function. *FEBS Journal* **282,** 2279–2295. ISSN: 1742464X (12 2015-06) (cit. on p. 6).

34. Waterman, D. P., Haber, J. E. & Smolka, M. B. Checkpoint Responses to DNA Double-Strand Breaks. *Annual Review of Biochemistry* **89,** 103–133. ISSN: 0066-4154 (1 2020-06) (cit. on pp. 6, 7, 9).

35. Aguilera, A. & García-Muse, T. Causes of Genome Instability. *Annual Review of Genetics* **47,** 1–32. ISSN: 0066-4197 (1 2013-11) (cit. on pp. 6, 7, 9).

36. Kaufmann, W. K. & Paules, R. S. DNA damage and cell cycle checkpoints. *The FASEB Journal* **10,** 238–247. ISSN: 0892-6638 (2 1996-02) (cit. on p. 7).

37. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair, and mutagenesis. *Environmental and Molecular Mutagenesis* **58,** 235–263. ISSN: 08936692 (5 2017-06) (cit. on p. 7).

38. Hakem, R. DNA-damage repair; the good, the bad, and the ugly. *The EMBO Journal* **27,** 589–605. ISSN: 0261-4189 (4 2008-02) (cit. on p. 7).

39. Giglia-Mari, G., Zotter, A. & Vermeulen, W. DNA Damage Response. *Cold Spring Harbor Perspectives in Biology* **3,** a000745–a000745. ISSN: 1943-0264 (1 2011-01) (cit. on p. 7).

40. Huang, R. & Zhou, P.-K. DNA damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy. *Signal Transduction and Targeted Therapy* **6,** 254. ISSN: 2059-3635 (1 2021-07) (cit. on p. 7).

41. Vítor, A. C., Huertas, P., Legube, G. & de Almeida, S. F. Studying DNA Double-Strand Break Repair: An Ever-Growing Toolbox. *Frontiers in Molecular Biosciences* **7.** ISSN: 2296-889X (2020-02) (cit. on pp. 7–9).

42. Mehta, A. & Haber, J. E. Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. *Cold Spring Harbor Perspectives in Biology* **6,** a016428–a016428. ISSN: 1943-0264 (9 2014-09) (cit. on pp. 7–9).

43. Stinson, B. M. & Loparo, J. J. Repair of DNA Double-Strand Breaks by the Nonhomologous End Joining Pathway. *Annual Review of Biochemistry* **90,** 137–164. ISSN: 0066-4154 (1 2021-06) (cit. on pp. 8, 9).

44. Guirouilh-Barbat, J., Lambert, S., Bertrand, P. & Lopez, B. S. Is homologous recombination really an error-free process? *Frontiers in Genetics* **5.** ISSN: 1664-8021 (2014-06) (cit. on p. 9).

45. Ui, A., Chiba, N. & Yasui, A. Relationship among DNA double-strand break (DSB), DSB repair, and transcription prevents genome instability and cancer. *Cancer Science* **111,** 1443–1451. ISSN: 1347-9032 (5 2020-05) (cit. on p. 9).

46. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the ageing process. *Nature* **592,** 695–703. ISSN: 0028-0836 (7856 2021-04) (cit. on p. 9).

47. Gómez-González, B. & Aguilera, A. Transcription-mediated replication hindrance: a major driver of genome instability. *Genes & Development* **33,** 1008–1026. ISSN: 0890-9369 (15-16 2019-08) (cit. on p. 9).

48. Vítor, A. C. *et al.* Single-molecule imaging of transcription at damaged chromatin. *Science Advances* **5,** eaau1249. ISSN: 2375-2548 (1 2019-01) (cit. on pp. 9, 10).

49. Long, Q., Liu, Z. & Gullerova, M. Sweet Melody or Jazz? Transcription Around DNA Double-Strand Breaks. *Frontiers in Molecular Biosciences* **8.** ISSN: 2296-889X (2021-04) (cit. on p. 9).

50. Madabhushi, R. *et al.* Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. *Cell* **161,** 1592–1605. ISSN: 00928674 (7 2015-06) (cit. on pp. 9, 29).

51. Michelini, F. *et al.* Damage-induced lncRNAs control the DNA damage response through interaction with DDRNAs at individual double-strand breaks. *Nature Cell Biology* **19,** 1400–1411. ISSN: 1465-7392 (12 2017-12) (cit. on pp. 10, 29).

52. Francia, S., Cabrini, M., Matti, V., Oldani, A. & d'Adda di Fagagna, F. DICER, DROSHA and DNA damage-response RNAs are necessary for the secondary recruitment of DNA damage response factors. *Journal of Cell Science.* ISSN: 1477-9137 (2016-01) (cit. on p. 10).

53. Mitsis, T. *et al.* Transcription factors and evolution: An integral part of gene expression (Review). *World Academy of Sciences Journal* **2,** 3–8. ISSN: 2632-2900 (1 2020-01) (cit. on p. 10).

54. Brázda, V., Bartas, M. & Bowater, R. P. Evolution of Diverse Strategies for Promoter Regulation. *Trends in genetics : TIG* **37,** 730–744. ISSN: 0168-9525 (8 2021-08) (cit. on pp. 10, 11).

55. Yella, V. R. & Bansal, M. DNA structural features of eukaryotic TATA-containing and TATA-less promoters. *FEBS Open Bio* **7,** 324–334. ISSN: 22115463 (3 2017-03) (cit. on p. 11).

56. Yang, C., Bolotin, E., Jiang, T., Sladek, F. M. & Martinez, E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389,** 52–65. ISSN: 03781119 (1 2007-03) (cit. on p. 11).

57. Kramm, K., Engel, C. & Grohmann, D. Transcription initiation factor TBP: old friend new questions. *Biochemical Society Transactions* **47,** 411–423. ISSN: 0300-5127 (1 2019-02) (cit. on p. 11).

58. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & Development* **25,** 1010–1022. ISSN: 0890-9369 (10 2011-05) (cit. on p. 11).

59. Landry, J.-R., Mager, D. L. & Wilhelm, B. T. Complex controls: the role of alternative promoters in mammalian genomes. *Trends in Genetics* **19,** 640–648. ISSN: 01689525 (11 2003-11) (cit. on p. 11).

60. Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C. & Huang, T. H.-M. The functional consequences of alternative promoter use in mammalian genomes. *Trends in Genetics* **24,** 167–177. ISSN: 01689525 (4 2008-04) (cit. on p. 11).

61. De Klerk, E. & Hoen, P. A. t. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in Genetics* **31,** 128–139. ISSN: 01689525 (3 2015-03) (cit. on p. 11).

62. Wade, J. T. & Grainger, D. C. Spurious transcription and its impact on cell function. *Transcription* **9,** 182–189. ISSN: 2154-1264 (3 2018-05) (cit. on p. 11).

63. Sen, P. *et al.* Spurious intragenic transcription is a feature of mammalian cellular senescence and tissue aging. *Nature Aging* **3,** 402–417. ISSN: 2662-8465 (4 2023-03) (cit. on pp. 11, 29).

64. Raj, A. & van Oudenaarden, A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* **135,** 216–226. ISSN: 00928674 (2 2008-10) (cit. on p. 11).

65. Viñuela, A. *et al.* Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Human Molecular Genetics* **27,** 732–741. ISSN: 0964-6906 (4 2018-02) (cit. on p. 11).

66. *R: The R Project for Statistical Computing* https://www.r-project.org/ (2023-06-02) (cit. on p. 13).

67. *R: What is R?* https://www.r-project.org/about.html (2023-06-05) (cit. on p. 13).

68. team, P. *RStudio Integrated Development Environment for R* httpwww.posit.co (cit. on p. 13).

69. *Bash Reference Manual* (2022) (cit. on p. 13).

70. Gomes, J. *et al.* Enabling rootless Linux Containers in multi-user environments: the udocker tool. *Computer Physics Communications* **232,** 84–97. ISSN: 00104655 (2017-11) (cit. on p. 13).

71. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-3-319-24277-4. https://ggplot2.tidyverse.org (Springer-Verlag New York, 2016) (cit. on p. 13).

72. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *dplyr: A Grammar of Data Manipulation* https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr (2023) (cit. on p. 13).

73. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25,** 1841–1842. ISSN: 1367-4811 (14 2009-07) (cit. on p. 13).

74. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550. ISSN: 1474-760X (12 2014-12) (cit. on p. 13).

75. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. <tt>edgeR</tt> : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140. ISSN: 1367-4811 (1 2010-01) (cit. on p. 13).

76. Davis, S. & Meltzer, P. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **14,** 1846–1847 (2007) (cit. on p. 13).

77. Blighe, K. & Lun, A. *PCAtools: PCAtools: Everything Principal Components Analysis* R package version 2.8.0 (2022). https://github.com/kevinblighe/PCAtools (cit. on p. 13).

78. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research* **41,** D991–5. ISSN: 1362-4962 (Database issue 2013-01) (cit. on p. 14).

79. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* **47,** D766–D773. ISSN: 1362-4962 (D1 2019-01) (cit. on p. 14).

80. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research* **51,** D1188–D1195. ISSN: 0305-1048 (D1 2023-01) (cit. on p. 14).

81. *GTEx Portal* `https://gtexportal.org/home/` (2023-07-19) (cit. on p. 14).

82. Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS computational biology* **18** (ed Li, J.) e1009730. ISSN: 1553-7358 (6 2022-06) (cit. on p. 16).

83. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29,** 15–21. ISSN: 1367-4811 (1 2013-01) (cit. on p. 16).

84. Musich, R., Cadle-Davidson, L. & Osier, M. V. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Frontiers in Plant Science* **12,** 657240. ISSN: 1664-462X (2021-04) (cit. on p. 16).

85. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Research* **50,** D988–D995. ISSN: 0305-1048 (D1 2022-01) (cit. on p. 17).

86. Noel, L. in (InTech, 2012-02) (cit. on p. 17).

87. Hommel, G. A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test. *Biometrika* **75,** 383. ISSN: 00063444 (2 1988-06) (cit. on p. 17).

88. Zar, J. H. in (Wiley, 2005-02) (cit. on p. 18).

89. Nanda, A., Mohapatra, D. B. B., Mahapatra, A. P. K., Mahapatra, A. P. K. & Mahapatra, A. P. K. Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error. *International Journal of Statistics and Applied Mathematics* **6,** 59–65. ISSN: 24561452 (1 2021-01) (cit. on p. 18).

90. Yan, W. X. *et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications* **8,** 15058. ISSN: 2041-1723 (1 2017-05) (cit. on p. 19).

91. Lensing, S. V. *et al.* DSBCapture: in situ capture and sequencing of DNA breaks. *Nature Methods* **13,** 855–857. ISSN: 1548-7091 (10 2016-10) (cit. on p. 19).

92. Joshi, I., DeRycke, J., Palmowski, M., LeSuer, R. & Feng, W. Genome-wide mapping of DNA double-strand breaks from eukaryotic cell cultures using Break-seq. *STAR Protocols* **2,** 100554. ISSN: 26661667 (2 2021-06) (cit. on p. 19).

93. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature Methods* **10,** 361–365. ISSN: 1548-7091 (4 2013-04) (cit. on p. 19).

94. Bateman, A. *et al.* UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51,** D523–D531. ISSN: 0305-1048 (D1 2023-01) (cit. on p. 30).

A

Datasets Information

Figure A.1: Pipeline was able to identify BIT-RNA candidates in MCF7 using the upstream model even when isoforms of assembled transcripts are present. After assembled transcriptome trimming and DSB file preparation, the identification of BIT's relied on a DSB being at 42 nt upstream of a TSS upmost. Highlighted region of BIT candidates when isoforms of assembled transcripts are present. It presents the RNAseq alignment of those samples, corresponding assembled transcripts, DSB, and annotated transcripts. Zoom of the BIT region is provided. Kb, Killobase; b, base

Table A.1: RNAseq Sample Information. The number of reads (Nº of Reads column) values correspond to a single strand. Orientation refers to strand-specific sequencing orientation (Forward Reverse (FR) or Reverse Forward (RF))

Organism	Cell Line	GEO Series	SRX Code	SRR Code	Layout	Nº of Reads	Orientation
Human	MCF7	GSE26284	SRX084666	SRR315301	Paired-End	43 M	RF
Human	MCF7	GSE26284	SRX084666	SRR315302	Paired-End	44 M	RF
Human	Fibroblast	GSE113957	SRX4022456	SRR7093809	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022457	SRR7093810	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022458	SRR7093811	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022459	SRR7093812	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022460	SRR7093813	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022461	SRR7093814	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022462	SRR7093815	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022463	SRR7093816	Single-End	14 M	RF
Human	Fibroblast	GSE113957	SRX4022464	SRR7093817	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022465	SRR7093818	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022466	SRR7093819	Single-End	31 M	RF
Human	Fibroblast	GSE113957	SRX4022467	SRR7093820	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022468	SRR7093821	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022469	SRR7093822	Single-End	15 M	RF
Human	Fibroblast	GSE113957	SRX4022470	SRR7093823	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022471	SRR7093824	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022472	SRR7093825	Single-End	24 M	RF
Human	Fibroblast	GSE113957	SRX4022473	SRR7093826	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022474	SRR7093827	Single-End	20 M	RF

Organism	Cell Line	GEO Series	SRX Code	SRR Code	Layout	Nº of Reads	Orientation
Human	Fibroblast	GSE113957	SRX4022475	SRR7093828	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022476	SRR7093829	Single-End	27 M	RF
Human	Fibroblast	GSE113957	SRX4022477	SRR7093830	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022478	SRR7093831	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022479	SRR7093832	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022480	SRR7093833	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022481	SRR7093834	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022482	SRR7093835	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022483	SRR7093836	Single-End	14 M	RF
Human	Fibroblast	GSE113957	SRX4022484	SRR7093837	Single-End	37 M	RF
Human	Fibroblast	GSE113957	SRX4022485	SRR7093838	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022486	SRR7093839	Single-End	24 M	RF
Human	Fibroblast	GSE113957	SRX4022487	SRR7093840	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022488	SRR7093841	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022489	SRR7093842	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022490	SRR7093843	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022491	SRR7093844	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022492	SRR7093845	Single-End	15 M	RF
Human	Fibroblast	GSE113957	SRX4022493	SRR7093846	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022494	SRR7093847	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022495	SRR7093848	Single-End	25 M	RF
Human	Fibroblast	GSE113957	SRX4022496	SRR7093849	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022497	SRR7093850	Single-End	19 M	RF

Organism	Cell Line	GEO Series	SRX Code	SRR Code	Layout	Nº of Reads	Orientation
Human	Fibroblast	GSE113957	SRX4022498	SRR7093851	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022499	SRR7093852	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022500	SRR7093853	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022501	SRR7093854	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022502	SRR7093855	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022503	SRR7093856	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022504	SRR7093857	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022505	SRR7093858	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022506	SRR7093859	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022507	SRR7093860	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022508	SRR7093861	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022509	SRR7093862	Single-End	13 M	RF
Human	Fibroblast	GSE113957	SRX4022510	SRR7093863	Single-End	14 M	RF
Human	Fibroblast	GSE113957	SRX4022511	SRR7093864	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022512	SRR7093865	Single-End	28 M	RF
Human	Fibroblast	GSE113957	SRX4022513	SRR7093866	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022514	SRR7093867	Single-End	13 M	RF
Human	Fibroblast	GSE113957	SRX4022515	SRR7093868	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022516	SRR7093869	Single-End	15 M	RF
Human	Fibroblast	GSE113957	SRX4022517	SRR7093870	Single-End	14 M	RF
Human	Fibroblast	GSE113957	SRX4022518	SRR7093871	Single-End	24 M	RF
Human	Fibroblast	GSE113957	SRX4022519	SRR7093872	Single-End	13 M	RF
Human	Fibroblast	GSE113957	SRX4022520	SRR7093873	Single-End	14 M	RF

Organism	Cell Line	GEO Series	SRX Code	SRR Code	Layout	Nº of Reads	Orientation
Human	Fibroblast	GSE113957	SRX4022521	SRR7093874	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022522	SRR7093875	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022523	SRR7093876	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022524	SRR7093877	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022525	SRR7093878	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022526	SRR7093879	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022527	SRR7093880	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022528	SRR7093881	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022529	SRR7093882	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022530	SRR7093883	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022531	SRR7093884	Single-End	26 M	RF
Human	Fibroblast	GSE113957	SRX4022532	SRR7093885	Single-End	23 M	RF
Human	Fibroblast	GSE113957	SRX4022533	SRR7093886	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022534	SRR7093887	Single-End	13 M	RF
Human	Fibroblast	GSE113957	SRX4022535	SRR7093888	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022536	SRR7093889	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022537	SRR7093890	Single-End	24 M	RF
Human	Fibroblast	GSE113957	SRX4022538	SRR7093891	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022539	SRR7093892	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022540	SRR7093893	Single-End	11 M	RF
Human	Fibroblast	GSE113957	SRX4022541	SRR7093894	Single-End	25 M	RF
Human	Fibroblast	GSE113957	SRX4022542	SRR7093895	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022543	SRR7093896	Single-End	26 M	RF

Organism	Cell Line	GEO Series	SRX Code	SRR Code	Layout	Nº of Reads	Orientation
Human	Fibroblast	GSE113957	SRX4022544	SRR7093897	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022545	SRR7093898	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022546	SRR7093899	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022547	SRR7093900	Single-End	27 M	RF
Human	Fibroblast	GSE113957	SRX4022548	SRR7093901	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022549	SRR7093902	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022550	SRR7093903	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022551	SRR7093904	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022552	SRR7093905	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022553	SRR7093906	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022554	SRR7093907	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022555	SRR7093908	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022556	SRR7093909	Single-End	15 M	RF
Human	Fibroblast	GSE113957	SRX4022557	SRR7093910	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022558	SRR7093911	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022559	SRR7093912	Single-End	17 M	RF
Human	Fibroblast	GSE113957	SRX4022560	SRR7093913	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022561	SRR7093914	Single-End	25 M	RF
Human	Fibroblast	GSE113957	SRX4022562	SRR7093915	Single-End	24 M	RF
Human	Fibroblast	GSE113957	SRX4022563	SRR7093916	Single-End	16 M	RF
Human	Fibroblast	GSE113957	SRX4022564	SRR7093917	Single-End	24 M	RF
Human	Fibroblast	GSE113957	SRX4022565	SRR7093918	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022566	SRR7093919	Single-End	22 M	RF

Organism	Cell Line	GEO Series	SRX Code	SRR Code	Layout	Nº of Reads	Orientation
Human	Fibroblast	GSE113957	SRX4022567	SRR7093920	Single-End	25 M	RF
Human	Fibroblast	GSE113957	SRX4022568	SRR7093921	Single-End	21 M	RF
Human	Fibroblast	GSE113957	SRX4022569	SRR7093922	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022570	SRR7093923	Single-End	20 M	RF
Human	Fibroblast	GSE113957	SRX4022571	SRR7093924	Single-End	18 M	RF
Human	Fibroblast	GSE113957	SRX4022572	SRR7093925	Single-End	19 M	RF
Human	Fibroblast	GSE113957	SRX4022573	SRR7093926	Single-End	25 M	RF
Human	Fibroblast	GSE113957	SRX4022574	SRR7093927	Single-End	33 M	RF
Human	Fibroblast	GSE113957	SRX4022575	SRR7093928	Single-End	28 M	RF
Human	Fibroblast	GSE113957	SRX4022576	SRR7093929	Single-End	30 M	RF
Human	Fibroblast	GSE113957	SRX4022577	SRR7093930	Single-End	32 M	RF
Human	Fibroblast	GSE113957	SRX4022578	SRR7093931	Single-End	30 M	RF
Human	Fibroblast	GSE113957	SRX4022579	SRR7093932	Single-End	24 M	RF
Human	Fibroblast	GSE113957	SRX4022580	SRR7093933	Single-End	28 M	RF
Human	Fibroblast	GSE113957	SRX4022581	SRR7093934	Single-End	31 M	RF
Human	Fibroblast	GSE113957	SRX4022582	SRR7093935	Single-End	29 M	RF
Human	Fibroblast	GSE113957	SRX4022583	SRR7093936	Single-End	22 M	RF
Human	Fibroblast	GSE113957	SRX4022584	SRR7093937	Single-End	29 M	RF
Human	Fibroblast	GSE113957	SRX4022585	SRR7093938	Single-End	27 M	RF
Human	Fibroblast	GSE113957	SRX4022586	SRR7093939	Single-End	35 M	RF
Human	Fibroblast	GSE113957	SRX4022587	SRR7093940	Single-End	27 M	RF
Human	Fibroblast	GSE113957	SRX4022588	SRR7093941	Single-End	34 M	RF
Human	Fibroblast	GSE113957	SRX4022589	SRR7093942	Single-End	28 M	RF

Organism	Cell Line	GEO Series	SRX Code	SRR Code	Layout	Nº of Reads	Orientation
Human	Fibroblast	GSE113957	SRX4022590	SRR7093943	Single-End	29 M	RF
Human	Fibroblast	GSE113957	SRX4022591	SRR7093944	Single-End	35 M	RF
Human	Fibroblast	GSE113957	SRX4022592	SRR7093945	Single-End	33 M	RF
Human	Fibroblast	GSE113957	SRX4022593	SRR7093946	Single-End	40 M	RF
Human	Fibroblast	GSE113957	SRX4022594	SRR7093947	Single-End	36 M	RF
Human	Fibroblast	GSE113957	SRX4022595	SRR7093948	Single-End	38 M	RF
Human	Fibroblast	GSE113957	SRX4022596	SRR7093949	Single-End	35 M	RF
Human	Fibroblast	GSE113957	SRX4022597	SRR7093950	Single-End	30 M	RF
Human	Fibroblast	GSE113957	SRX4022598	SRR7093951	Single-End	36 M	RF

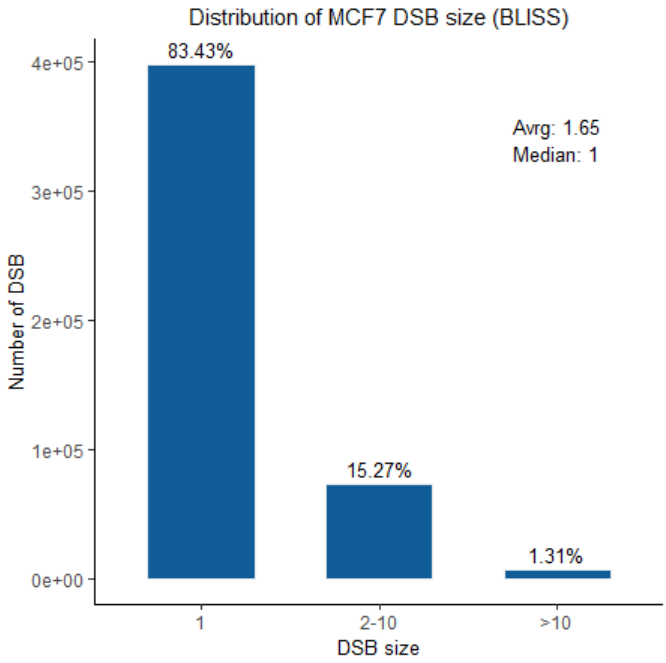


Figure A.2: MCF7 BLISS acquired DSB were mostly of 1 nt.
The barplot presents the number of DSB, distributed by three groups of DSB sizes: 1 nt, between 2 and 10 nt, and more than 10 nt. The average size of DSB and median value are also present. On top of the bars is the correspondent percentage

Table A.2: DSB Sample Information. The number of hits (Nº of hits) corresponds to the number of DSB present in a file

Organism	Cell Line	GEO Series	GSM code	Technique	Genome Build	Nº of hits
Human	MCF7	GSE134798	GSM4520693	BLISS	hg38	204388
Human	MCF7	GSE134798	GSM4520697	BLISS	hg38	153647
Human	MCF7	GSE134798	GSM4520701	BLISS	hg38	129103
Human	NHEK	GSE78172	Not Applicable (NA)	DSBCapture	hg19	87908
Human	MCF7	GSE207714	GSM6310749	Break-seq	hg19	2720
Human	MCF7	GSE207714	GSM6310750	Break-seq	hg19	3373
Human	MCF7	GSE207714	GSM6310751	Break-seq	hg19	1221
Human	MCF7	GSE207714	GSM6310752	Break-seq	hg19	4187
Human	MCF10A	GSE207714	GSM6310753	Break-seq	hg19	1640
Human	MCF10A	GSE207714	GSM6310754	Break-seq	hg19	2910
Human	MCF10A	GSE207714	GSM6310755	Break-seq	hg19	883
Human	MCF10A	GSE207714	GSM6310756	Break-seq	hg19	3496
Human	GM06990	GSE124403	NA	Break-seq	hg19	4149

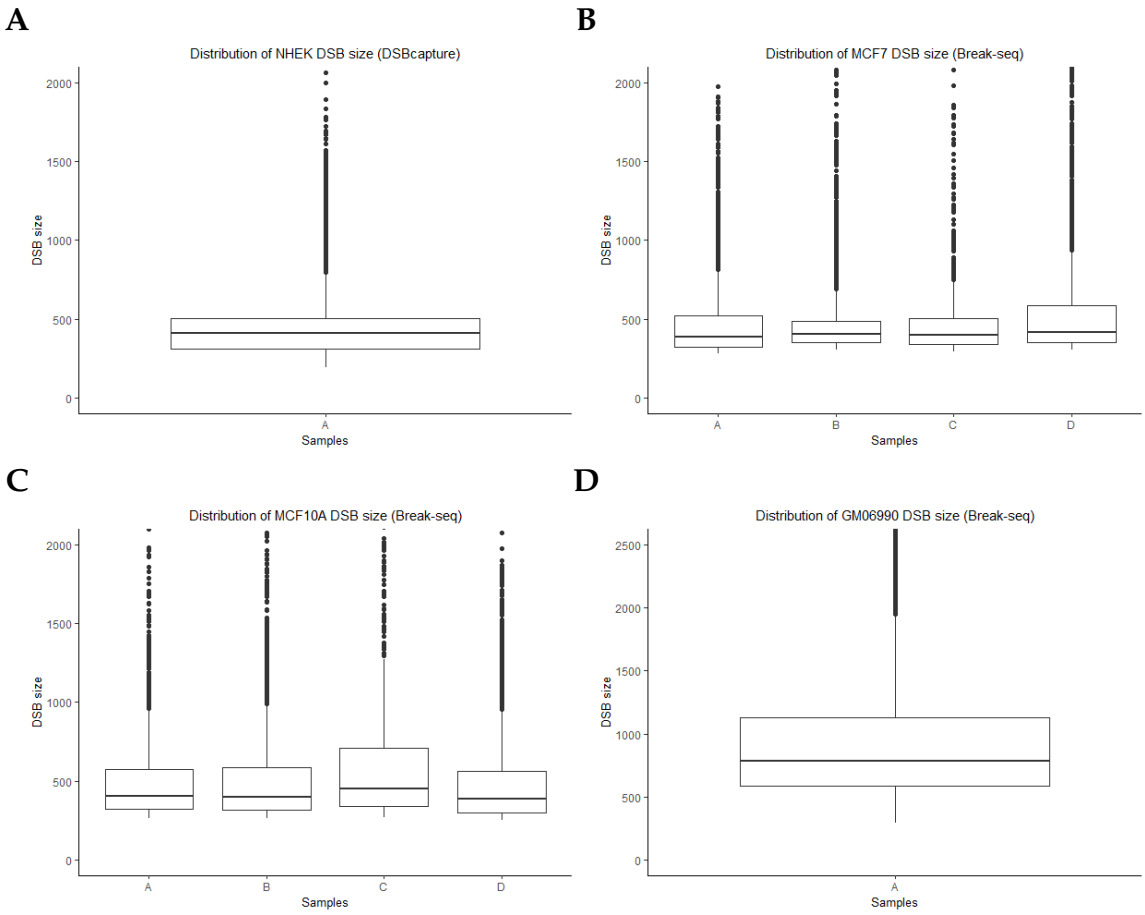


Figure A.3: Low resolution techniques of DSB vary in size.
The DSB acquisition techniques of low resolution identify hotspot regions of DSB. Each of the four cell lines (NHEK, MCF7, MCF10A, GM06990) was acquired with either (**A**) DSBcapture or **(B), (C), (D)** Break-seq, having hits varying in size. In **(A), (B)** and **(C)** outliers above 2000 were removed (i.e. DSB regions bigger than 2000), and **(D)** had outliers above 2500 removed (i.e. DSB regions bigger than 2500)

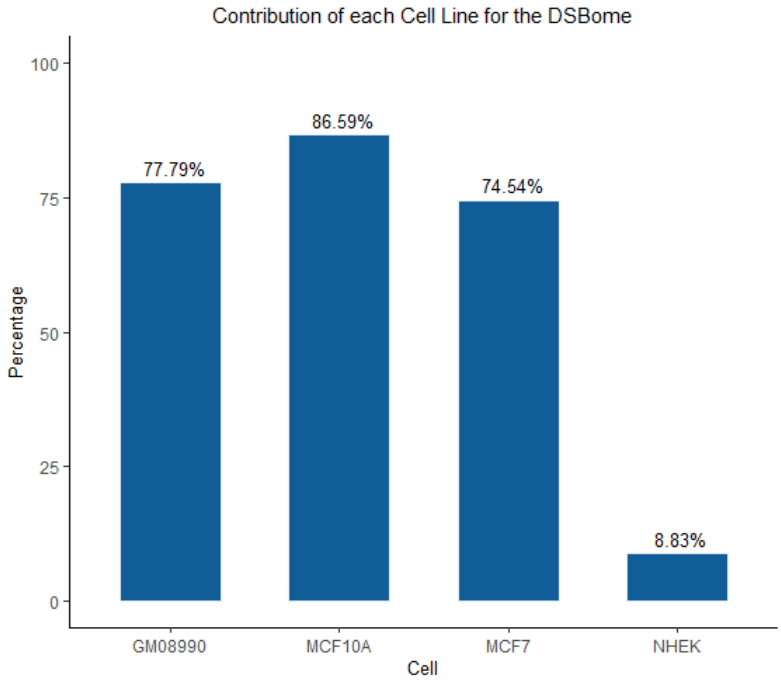


Figure A.4: DSBome does not present a preference in cell line. The DSBome is the DSBs common to 2 or more cell lines. Each bar represents the percentage of DSBs of DSBome that was from a cell line. Each DSB belongs to 2 or more bars, and besides NHEK, the remaining cell lines contribute similarly

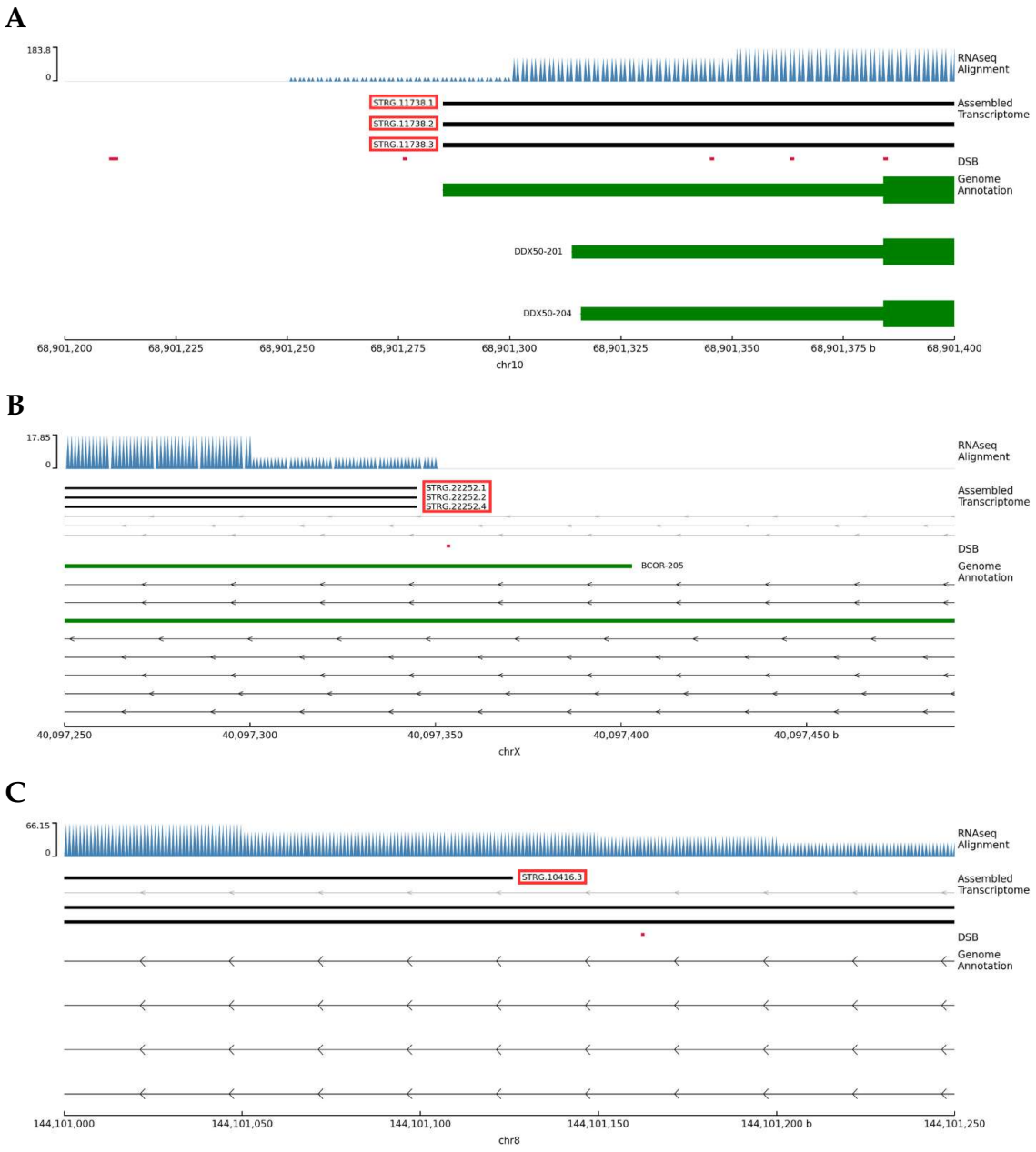


Figure A.5: Highlight BIT identification of those transcripts whose filters selected.
Following BIT identification, certain filters were applied to choose a smaller group of them depending on the position of a BIT-RNA TSS and their relation to an annotated transcript. The representations highlight regions of BIT candidates and present the RNAseq alignment of those samples, corresponding assembled transcripts, DSB, and annotated transcripts. Their application was sequential, and their order is the one presented in Figure 2.1**B**, corresponding to **A**, **B** and **C**, respectively. These images are a zoom of the BITs present in Figure 3.2 and should be seen simultaneously. The red squares on the images highlight the ones who were selected in that filter. This **(A)** the first selection was the BIT's whose TSS was inside an annotated gene **(B)** then if the BIT-RNA sense of transcription was the same as the annotated transcript and **(C)** finally if the TSS location it's not in the first exon of a Canonical Transcript. Kb, Killobase; b, base

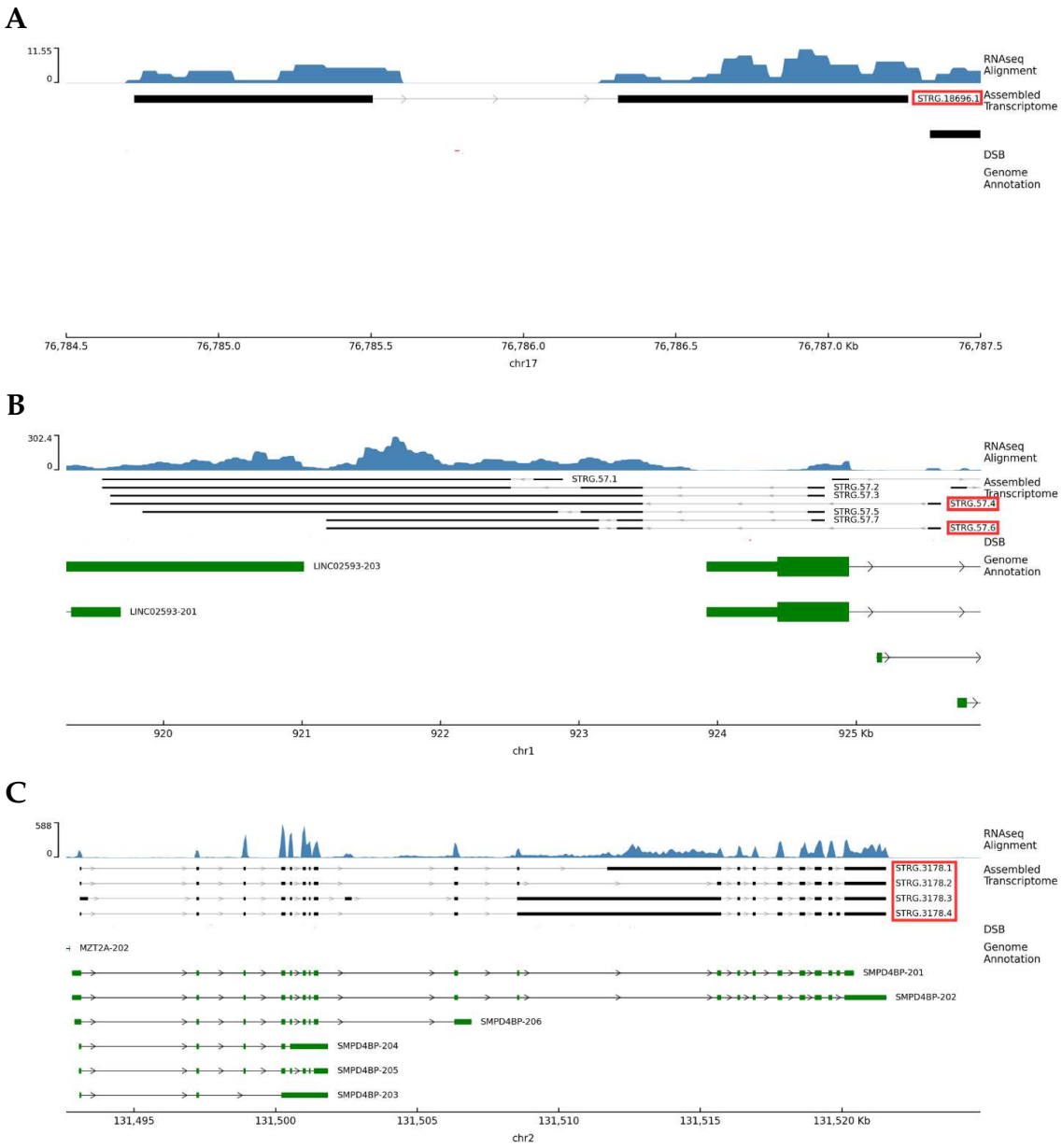


Figure A.6: Highlight of BIT identification of those transcripts whose filters were removed. Following BIT identification, certain filters were applied to remove those whose characteristics were not of interest. The filtering depended on the position of a BIT-RNA TSS and its relation to an annotated transcript. The representations highlight regions of BIT candidates and present the RNAseq alignment of those samples, corresponding assembled transcripts, DSB, and annotated transcripts. Their application was to the ones which were not removed in the previous filter, corresponding to **A**, **B** and **C**, respectively. These images are a zoom of the BITs present in Figure 3.3 and should be seen simultaneously. The red squares on the images highlight the ones who were removed from that filter. **(A)** firstly, those whose TSS was outside an annotated gene **(B)** then if the BIT-RNA sense of transcription was not the same as the annotated transcript, which was different from selecting the ones on the opposite strand, since being the same strand and opposite strand simultaneously, and in those cases, they are not removed, and **(C)** finally if the TSS location is in the first exon of a Canonical Transcript. Kb, Killobase; b, base

UNIVERSIDADE NOVA
DE LISBOA

2023

BIT by BIT: Associating Transcription to Ageing

IAN LOPES TEIXEIRA

MASTER IN **COMPUTATIONAL BIOLOGY & BIOINFORMATICS**

UNIVERSIDADE NOVA DE LISBOA