ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

Ανάπτυξη υπολογιστικών μεθόδων ανάλυσης της απόκρισης του κυττάρου στο γενοτοξικό στρες

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Δημήτρης Κωνσταντόπουλος

Βιοπληροφορικός, MSc

Κρήτη
Νοέμβριος 2020

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ανάπτυξη υπολογιστικών μεθόδων ανάλυσης της απόκρισης του κυττάρου στο γενοτοξικό στρες

Δημήτρης Γ Κωνσταντόπουλος

Α.Μ.: 802

ΕΠΙΒΛΕΠΩΝ: Γεώργιος Γαρίνης, Καθηγητής

ΤΡΙΜΕΛΗΣ	**Γεώργιος Γαρίνης**, Καθηγητής
ΣΥΜΒΟΥΛΕΥΤΙΚΗ	**Ιωάννης Ταλιανίδης**, Ερευνητής Α’
ΕΠΙΤΡΟΠΗ:	**Μαρία Φουστέρη**, Ερευνήτρια Β’

Νοέμβριος 2020

«Η έγκριση της Διδακτορικής Διατριβής από το Τμήμα Βιολογίας της Σχολής Θετικών Επιστημών του Πανεπιστημίου Κρήτης δεν υποδηλώνει αποδοχή των απόψεων του συγγραφέα (ν. 5343/1932, άρθρο 202)»

«Το κείμενο της Διδακτορικής Διατριβής δεν αποτελεί προιόν λογοκλοπής»

ABSTRACT

The purpose of this study is the development of a computational framework for studying the dynamic changes of active transcription, and its interaction with chromatin remodeling and chromatin alterations during cellular responses to genotoxic stress. For this purpose, ultraviolet light C (UVC) was used as a genotoxic stress factor, damaging skin cells, specifically skin fibroblasts (VH10, CSB and 1BR.3), while the activity of Nucletiode Excision Repair (NER) pathway and the repair products of Global Genome NER (GG-NER) and Transcription Coupled NER (TC-NER) sub-pathways were used to evaluate the examined mechanisms.
Various types of Next Generation Sequencing (NGS) experiments have been used to study the stages of the transcription cycle in normal conditions, and in response to Ultraviolet C irradiation (UVC) induced stress. Specifically, for studying the kinetics of RNA Polymerase 2 (RNAPII) molecules from the transcription initiation state, to promoter proximal pausing (PPP), and the transition to productive elongation, Chromatin immunoprecipitation sequencing (ChIP-seq) data of the hypophosphorylated RNAPII (RNAPII-hypo), the elongating isoform of RNAPII (RNAPII-ser2P), and the RNAPII-ser5P isoform (transcription initiation) was generated and analyzed.
To study the productivity of RNAPII molecules during the above stages, Capped Analysis of Gene expression sequencing (CAGE-seq) data and nascent RNA synthesis sequencing (nRNA-seq) data was used. To study the interactions of chromatin with active transcription and its alteration during the states of active transcription, Assay for Transposase-Accessible Chromatin (ATAC-seq) data was generated and analyzed, and ChIP-seq data of H3K27ac and H3K27me3 histone modifications.
To study the effectiveness and genomic landscape of NER repair-synthesis events, for both GG-NER and TC-NER sub-pathways, a novel assay called aniFOUND-seq was developed and analyzed, coupled with data of excised DNA during NER activity (XR-seq) and NER damage sequencing data (damage-seq). The functional assessment of TC-NER at active genes was carried out through the study of mutations in melanoma and lung adenocarcinoma cancer genomes, and XR-seq data meta-analysis respectively.

The results of these essays are divided into four sections:
(1) Development and application of algorithms for the analysis of NGS data related to human disease. (a) Implementation of stand-alone analysis pipelines for the analysis of ChIP-seq, nRNA-seq, and ATAC-seq datasets that include: Quality control (QC) assessment of sequenced short-reads, short-read preprocessing, short-read mapping against the under study reference genome/transcriptome, alignment processing, alignment summarization in genomic features and visualization via heatmaps and average profiles, generation of genomic tracks viewable in genome browsers (IGV, UCSC), NGS signal clustering upon functional genomic regions, correlation of biological and technical replicates, dimensionality reduction methods to identify technical/biological similarities/differences between samples, differential expression analysis, peak calling analysis, differential binding analysis, differential accessibility analysis and other statistical comparisons between biological conditions.
(b) Implementation of a “de novo” elongation wave identification algorithm using Hidden Markov Models (HMMs), and DRB-nRNA-seq datasets.

(2) Cellular responses under genotoxic stress conditions. (a) Development of a computational pipeline for the study of the reorganization of transcription and the chromatin rearrangements upon UV-induced stress that include: genome annotation reconstruction, and characterization of transcribed units' activity (promoters, genes, enhancers, PROMoter uPstream Transcripts - asPROMPs) along the human genome, the quantification of the RNAPII release from PPP sites, and the evaluation of the RNAPII elongation wave kinetics.
(b) A proposed model describing the 'safe' mode mechanism of transcription elongation; upon UVC-induced stress, steady-state transcription levels of virtually all actively transcribed genes are re-adjusted to fast and uniform release of RNAPII elongation waves from PPP sites that scan the transcribed genome for DNA lesions.
This mechanism maximizes the speed of lesion sensing, the probability that a damage will be identified by an elongating RNAPII molecule and removed by TC-NER along the actively transcribed elements. As a result, environmentally exposed genomes are characterized by a modest and homogeneous mutation prevalence across the actively transcribed genome in both strands, as opposed to the non-transcribed elements where higher mutation rates are observed. In case NER is unsuccessful or is not recruited efficiently during the stress recovery process, unrepaired DNA lesions can provoke error-prone DNA synthesis and result in mutagenesis during replication.

(3) Extending the previously described 'safe' mode mechanism of transcription elongation, the results of the particular dissertation also support a model of continuous transcription initiation that can fuel the widespread UV-triggered escape of RNAPII into transcription elongation, that safeguards the integrity of the actively transcribed genome. The particular mechanism is supported by a global increase of chromatin accessibility at all actively transcribed promoters serving as a platform that favors unrestrained transcription initiation, coupled by preservation of the active mark H3K27ac and repressive mark H3K27me3 mark during early response to genotoxic stress.

(4) A genome-wide analysis pipeline for the evaluation of aniFOUND-seq methodology. aniFOUND, is the first methodology (at the time of writing this thesis) that can exclusively label, capture and map the post-damage newly synthesized repaired chromatin in its native form (see materials and methods). Coupling of aniFOUND to NGS, allows the mapping and characterization of the NER efficacy of different chromosomal regions of the human genome. aniFOUND-seq was successfully applied to map the repair-synthesis activity along damaged skin fibroblasts (1BR.3 cells) with particular attention to promoter and enhancer sequences. Furthermore, aniFOUND-seq was applied for the assessment of NER-UDS activity in several chromosomal regions, including the fraction of repetitive DNA. Specifically, the repair efficacy during the first 4 hours after damage induction was clarified for rDNA and telomeres, for which contradictory explanatory models have been suggested. This is the first time that NGS-based approaches are adopted for shedding light in the above-mentioned inquiries regarding repair of telomeric DNA. Evidently, the cumulative nature of aniFOUND-seq (in terms of both damage types and repair assessment period) renders it applicable for the cases that require capturing of

the whole repair process, or the repair activity during moderately-to-considerably long-time windows.

ΠΕΡΙΛΗΨΗ

Ο σκοπός της συγκεκριμένης μελέτης, αποτέλεσε την δημιουργία ενός υπολογιστικού πλαισίου ανάλυσης για την μελέτη των δυναμικών αλλαγών της ενεργής μεταγραφής, καθώς και της αλληλεπίδρασής τους με την αναδιαμόρφωση της χρωματίνης, κατά την απόκριση στο γενοτοξικό στρες. Για τον σκοπό αυτό, η υπεριώδης ακτινοβολία C (UVC) χρησιμοποιήθηκε σαν στρεσογόνος παράγοντας για την δημιουργία βλαβών σε δερματικά κύτταρα, και συγκεκριμένα κυτταρικές σειρές ινοβλαστών δέρματος (VH10, CSB and 1BR.3), ενώ ο μηχανισμός εκτομής νουκλεοτιδίων (NER), και συγκεκριμένα τα επιδιορθωτικά παράγωγα των υπό-μονοπατιών Global Genome NER (GG-NER) και Transcription Coupled NER (TC-NER) χρησιμοποιήθηκαν για την αξιολόγηση των υπό μελέτη μηχανισμών.

Για την μελέτη των σταδίων του μεταγραφικού κύκλου σε κανονικές συνθήκες, καθώς και σε συνθήκες έκθεσης στην UVC ακτινοβολία, χρησιμοποιήθηκαν τεχνικές αλληλούχισης νέας γενιάς (Next Generation Sequencing - NGS). Συγκεκριμένα, για την μελέτη της κινητικής των μορίων της RNA πολυμεράσης 2 (RNAPII), από το στάδιο έναρξης της μεταγραφής, στο στάδιο παύσης (promoter proximal pausing - PPP), μέχρι και την μετάβαση στο στην παραγωγική επιμήκυνση, παράχθηκαν και αναλύθηκαν δεδομένα NGS ανοσοκατακρίμνησης της χρωματίνης (Chromatin immunoprecipitation sequencing - ChIP-seq) της υποφωσφορυλιωμένης RNAPII (RNAPII-hypo), της φωσφορυλιωμένης στην σερίνη 2 RNAPII (RNAPII-ser2P), και της φωσφορυλιωμένης στην σερίνη 5 RNAPII (RNAPII-ser5P)

Για την μελέτη της παραγωγικότητας των μορίων της RNAPII στα παραπάνω στάδια της μεταγραφής, χρησιμοποιήθηκαν δεδομένα CAGE-seq (Capped Analysis of Gene expression sequencing), καθώς και δεδομένα NGS αρτιγενούς έκφρασης RNA (nascent RNA synthesis sequencing - nRNA-seq)

Για την μελέτη της αλληλεπίδρασης της χρωματίνης και των αναδιαμορφώσεων της κατά το φαινόμενο της ενεργής μεταγραφής, ενεργοποιήθηκαν και αναλύθηκαν NGS δεδομένα καταγραφής της προσβασιμότητας της χρωματίνης ATAC-seq (Assay for Transposase-Accessible Chromatin), καθώς και δεδομένα ChIP-seq των επιγενετικών τροποποιήσεων της χρωματίνης H3K27ac και H3K27me3.

Για την αποσαφήνιση της αποτελεσματικότητας του μηχανισμού NER και των υπο-μονοπατιών GG-NER και TC-NER, αναπτύχθηκε μια νέα NGS τεχνολογία, το aniFOUND-seq, η οποία συνδυάστηκε με δεδομένα XR-seq (μεθοδολογία εντοπισμού εκτομής DNA κατά την δράση του μηχανισμού NER), καθώς και με NGS δεδομένα εντοπισμού βλαβών του DNA που προκαλούνται από την UVC ακτινοβολία (damage-seq). Η λειτουργική αποτίμηση του μηχανισμού TC-NER σε ενεργά μεταγραφικές μονάδες πραγματοποιήθηκε με την ανάλυση δεδομένων μεταλλαγών γονιδιώματων με καρκίνο του δέρματος (melanoma) καθώς και καρκίνο του πνεύμονα (lung adenocarcinoma), σε συνδυασμό με μετα-ανάλυση δεδομένων XR-seq.

Τα αποτελέσματα των παραπάνω συνοψίζονται σε 4 τμήματα:

(1) Ανάπτυξη και εφαρμογή αλγορίθμων για την ανάλυση δεδομένων NGS που σχετίζονται με την ανθρώπινη παθογένεια: (α) Ανάπτυξη ενός αυτόνομου πλαισίου ανάλυσης δεδομένων ChIP-seq, nRNA-seq, και ATAC-seq που περιέχει την ποιοτική αποτίμηση των δεδομένων (Quality Control - QC), την προ-επεξεργασία των μικρών διαβασμάτων (reads), την αντιστοίχιση των διαβασμάτων στο γονιδίωμα/ μεταγράφωμα αναφοράς, την επεξεργασία των

αντιστοιχίσεων, την σύνοψη των αντιστοιχίσεων σε γονιδιακές περιοχές αναφοράς και την οπτικοποίηση τους, την δημιουργία εγγραφών για πλοήγηση σε γονιδιακούς φυλλομετρητές (IGV, UCSC), την ομαδοποίηση του NGS σήματος σε λειτουργικά μετάγραφα, τις συσχετίσεις μεταξύ βιολογικών και τεχνικών επαναλήψεων των δεδομένων, την εφαρμογή μεθόδων μείωσης διαστάσεων για την αναγνώριση τεχνικών/βιολογικών ομοιοτήτων/διαφορών των δεδομένων, την ανάλυση διαφορικής έκφρασης, την ανάλυση εύρεσης περιοχών με ενισχυμένο ChIP-seq σήμα (peak calling), την ανάλυση διαφορικής πρόσδεσης, την ανάλυση διαφορικής προσβασιμότητας της χρωματίνης, και άλλες στατιστικές συγκρίσεις μεταξύ των υπο-μελέτη βιολογικών συνθηκών. (β) Ανάπτυξη ενός αλγορίθμου για τον εντοπισμό του “de novo” μεταγραφικού κύματος της RNAPII, χρησιμοποιώντας Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models - HMMs) και δεδομένα DRB-nRNA-seq.

(2) Κυτταρική απόκριση σε συνθήκες γενοτοξικού στρες: (α) Ανάπτυξη ενός πλαισίου ανάλυσης για την μελέτη της αναδιοργάνωσης της μεταγραφής και της χρωματίνης έπειτα από έκθεση σε UV-C. Στο συγκεκριμένο πλαίσιο ανάλυσης περιέχονται: Κατάλληλη προσαρμογή των μεταγράφων αναφοράς (υποκινητές, γονίδια, ενισχυτές, PROMoter uPstream Transcripts - asPROMPs) και χαρακτηρισμός της ενεργότητάς τους, ποσοτικοποίηση της εξόδου της RNAPII από το σημεία PPP και αποτίμηση της κινητικής της κατά την παραγωγική επιμήκυνση.

(b) Ένα προτεινόμενο μοντέλο που περιγράφει τον μηχανισμό “safe mode” της παραγωγικής επιμήκυνσης. Συγκεκριμένα, κατά την απόκριση στο γενοτοξικό στρες που προκαλείται από την UVC ακτινοβολία, παρατηρείται η γρήγορη και ομοιόμορφη απελευθέρωση της RNAPII από τα PPP σημεία των ενεργών μεταγράφων, προκαλώντας το ξέσπασμα ενός μεταγραφικού κύματος το οποίο με την σειρά του ανιχνεύει το μεταγραφώμενο γονιδίωμα για βλάβες.

Ο συγκεκριμένος μηχανισμός μεγιστοποιεί την ταχύτητα εντοπισμού DNA βλαβών, την πιθανότητα αναγνώρισης από τα επιμηκούμενα μόρια RNAPII, και αφαίρεσης τους από το TC-NER στις μεταγραφικές μονάδες των ενεργών γονιδίων. Σαν αποτέλεσμα, γονιδιώματα εκτεθημένα σε περιβαλλοντικούς παράγοντες χαρακτηρίζονται από περιορισμένο και ομοιόμορφο βαθμό μεταλλαγών, σε περιοχές ενεργών μεταγράφων, και στις 2 DNA αλυσίδες, σε αντίθεση με τις περιοχές που δεν μεταγράφονται από την RNAPII όπου παρατηρείται αυξημένος βαθμός μεταλλαγών. Σε περίπτωση που το NER είναι ανεπιτυχές, η δεν στρατολογηθεί επιτυχώς κατά την διαδικασία της ανάκαμψης από την έκθεση στις στρεσογόνες συνθήκες, μη διορθωμένες DNA βλάβες μπορούν να προκαλέσουν εσφαλμένη DNA σύνθεση η οποία θα έχει σαν αποτέλεσμα την μεταλλαξιγένεση.

(3) Επεκτείνοντας την περιγραφή του μηχανισμού “safe mode” της παραγωγικής επιμήκυνσης, τα αποτελέσματα της συγκεκριμένης διατριβής υποστηρίζουν ένα μοντέλο διαρκούς έναρξης της μεταγραφής, το οποίο τροφοδοτεί την εκτενή έξοδο των RNAPII μορίων από το PPP έπειτα από έκθεση στην UVC, διαφυλάσσοντας έτσι την ακεραιότητα του ενεργά μεταγραφώμενου γονιδιώματος. Ο μηχανισμός αυτός πλαισιώνεται από την καθολική αύξηση της προσβασιμότητας της χρωματίνης, σε όλες τις ενεργές μεταγραφικές μονάδες, παίζοντας τον ρόλο μιας πλατφόρμας η οποία ευνοεί την αδιάκοπη έναρξη της μεταγραφής, ενώ παράλληλα διατηρούνται οι επιγενετικές τροποποιήσεις των ιστονών H3K27ac και H3K27me3 κατά την διάρκεια των πρώτων σταδίων κυτταρικής ανάκαμψης έπειτα από την έκθεση στην UVC ακτινοβολια.

(4) Ανάπτυξη ενός πλαισίου ανάλυσης του ολόκληρου του γονιδιώματος για την αποτίμηση της aniFOUND-seq μεθόδου. Το aniFOUND, αποτελεί την πρώτη μέθοδο (κατά την διάρκεια της

συγγραφής της συγκεκριμένης μελέτης) που επιτρέπει τον αποκλειστικό χαρακτηρισμό και την ανάκτηση των νεοσυντιθέμενων τμημάτων της επιδιορθωμένη χρωματίνης, έπειτα από την απομάκρυνση των DNA βλαβών από τον μηχανισμό NER. Ο συνδυασμός της μεθόδου aniFOUND με την τεχνολογία NGS, επιτρέπει τον εντοπισμό και τον χαρακτηρισμό της αποτελεσματικότητας του μηχανισμού NER στο σε ολόκληρο το ανθρώπινο γονιδίωμα.
Το aniFOUND-seq εφαρμόστηκε επιτυχώς για την ανίχνευση επιδιορθωμένων περιοχών σε ινοβλάστες δέρματος (1BR.3 κυτταρική σειρά), με έμφαση στις περιοχές των υποκινητών και ενισχυτών. Επιπλέον, το aniFOUND-seq αξιοποιήθηκε για την αποτίμηση της δραστηριότητας του NER-UDS, σε διάφορες περιοχές του ανθρώπινου γονιδιώματος, όπως οι περιοχές επαναλήψεων DNA (DNA repeats). Συγκεκριμένα, αποσαφηνίστηκε η αποτελεσματικότητα της επιδιόρθωσης του DNA κατά τις 4 πρώτες ώρες έπειτα από την δημιουργία βλαβών, για τις περιοχές ριβοσωμικού DNA και των τελομερών, για τις οποίες μέχρι τώρα υπήρχαν αντικρουόμενα μοντέλα περιγραφής. Κατά συνέπεια, ο σωρευτικός χαρακτήρας του aniFOUND-seq (σε όρους τύπων βλαβών, καθώς και περιόδου επιδιόρθωσης) το καθιστά κατάλληλο για μελέτες που απαιτούν την ολική αξιολόγηση της διαδικασίας επιδιόρθωσης του DNA κατά την διάρκεια σχετικά μεγάλων χρονικών διαστημάτων.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τους γονείς μου, τους συναδέλφους μου και το άμεσο περιβάλλον μου που όλα αυτά τα χρόνια με στήριξε ώστε να ολοκληρωθεί η συγκεκριμένη διατριβή. Ιδιαίτερες ευχαριστίες θα ήθελα να δώσω στην Μαρία Φουστέρη, η οποία με καθοδήγησε και με στήριξε ηθικά και οικονομικά στην μέχρι στιγμής πορεία μου στον χώρο της έρευνας.

Contents

1 Introduction

1.1 Genotoxic Stress

Genomes are constantly exposed to DNA-damaging agents, which disrupt genome integrity by producing DNA lesions, altering DNA chemistry and structure. It has been estimated that every cell experiences up to 10^5 spontaneous or induced DNA lesions per day (De Bont & van Larebeke, 2004). Cells try to eliminate these alterations by either DNA repair or apoptosis, but lesions may not always be removed leading to mutagenesis and increasing the risk to develop cancer.

Genotoxic agents have long been associated with the development of human cancers. These include environmental agents such as the ultraviolet (UV) radiation that increase the risk of skin cancers (Pleasance, Cheetham, et al., 2010), cigarette smoke that increases the risk of lung

cancer (Pleasance, Stephens, et al., 2010), aflatoxins that are related with liver cancer (Alexandrov et al., 2013), amine dyes with bladder cancer (J. Kim et al., 2016), benzene with leukemia (Snyder, 2012), and vinyl chloride with hepatic cancer (Fedeli et al., 2019). Additional sources of genotoxic stress are several therapeutic agents, such as anticancer drugs cisplatin and Topoisomerase I and II inhibitors, but also some endogenous metabolic products or by-products such as reactive oxygen species (ROS) and errors generated during the replication procedure. A perplexing diversity of lesions arises in DNA by these genotoxins. ROS can cause DNA base lesions including hydrolysis (deamination, depurination, and depyrimidination) whereas exposure to alkylating agents (O^6-Methylguanine) or oxidation (8-oxoG). UV exposure is linked with formation of cycloboutene pyrimidine dimers (CPDs) and pyrimidine 6-4 pyrimidine photoproducts (6-4PPs). Ionizing radiation (IR) induces single and double DNA strand breaks, while chemotherapeutic drugs are responsible for inter- and intra-strand DNA crosslinks (Ciccia & Elledge, 2010). Over 100 oxidative modifications have been identified in DNA (Cadet et al., 2003).

The primary structure of the DNA double helix can be altered by such lesions, resulting in defects during the transcription and replication processes. Nevertheless, faulty repair of DNA lesions may lead to genomic mutations that can be inherited through cell division with deleterious consequences for human health. Since genotoxic stress effects can be (directly or indirectly) involved in both tumor initiation and tumor progression, or even be a prerequisite for tumorigenesis, studying and understanding the cellular responses to genotoxic insult is a vital step for the prevention and treatment of human disease.

Repair mechanism	Lesion feature	Genotoxic source (examples)
Base excision repair (BER)	Oxidative lesions	Reactive oxygen species (ROS)
Nucleotide excision repair (NER)	Helix-distorting lesions	UV radiation
Translesion synthesis	Various lesions	Various sources
Mismatch repair (MMR)	Replication errors	Replication
Single stand break repair (SSBR)	Single strand breaks	Ionizing radiation, ROS
Homologous recombination (HR)	Double-strand breaks	Ionizing radiation, ROS
Non-homologous end joining (NHEJ)	Double-strand breaks	Ionizing radiation, ROS
DNA interstrand crosslink repair pathway	Interstrand crosslinks	Chemotherapy

Table 1 DNA repair mechanisms are specialized to repair the different types of DNA damages. Adapted by (Torgovnick & Schumacher, 2015), TABLE1.

1.2 Cellular responses to genotoxic stress

Cells regularly respond to genotoxic insults using an intricate defense system. These responses involve various cellular factors that form an extensive signal transduction network. The specific network includes a complex signaling cascade termed as the DNA damage response (DDR) that bridges the DNA damage sensing (initial signal) with the activation of specific transcription factors, which successively regulate the expression of genes implicated in DNA repair pathways, cell cycle arrest to allow time for repair, and in some cases, initiation of senescence or apoptosis programs (Ciccia & Elledge, 2010).

Despite there is no single repair machinery that can handle all types of damage, evolution has molded a layer of complex and sophisticated DNA repair systems that altogether cover most of the genotoxic insults that affect cell's vital genetic information. These mechanisms are highly conserved across mammals and can be categorized to at least five distinct, partly overlapping pathways: Nucleotide-excision repair (NER), base-excision repair (BER), mismatch repair (MMR), homologous recombination (HR) and non-homologous end joining (NHEJ) (Friedberg et al., 2005; Lindahl & Wood, 1999). The main function of each mechanism can be depicted as follows (Table 1).

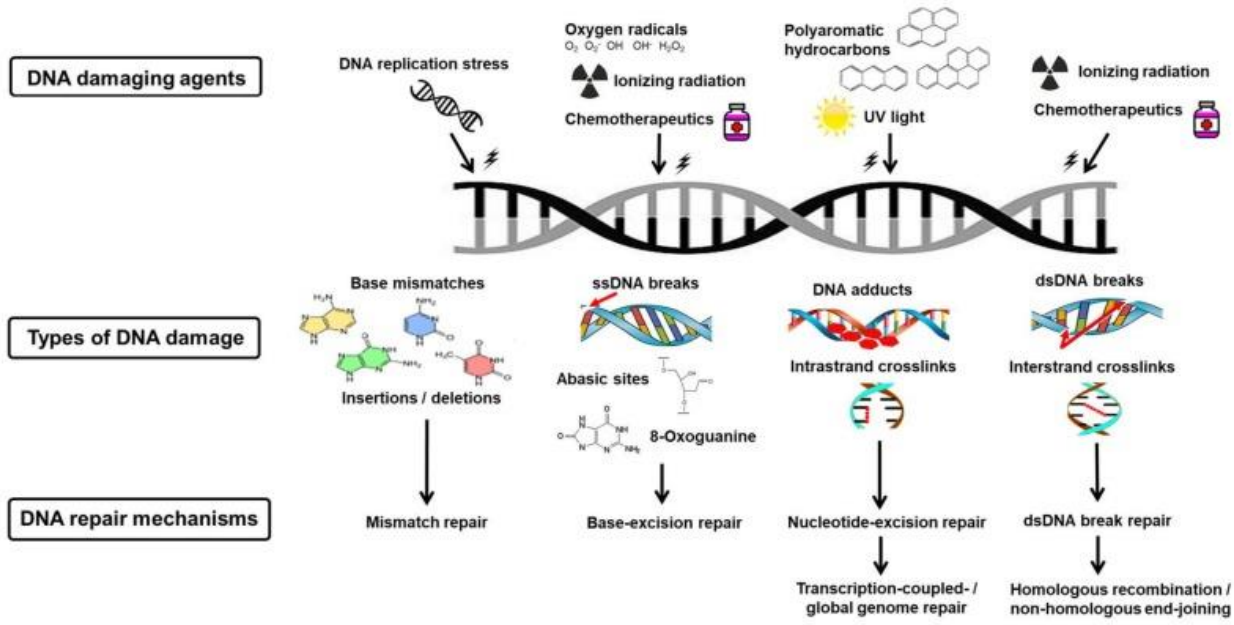


Figure 1 DNA damage and repair mechanisms. Adapted from figure 3 in (Helena et al., 2018).

NER is involved in the removal and replacement of bulky, helix distorting DNA adducts by sensing the distortion caused to the DNA double helix and by excising the oligonucleotide containing the lesion and replacing it with newly synthesized DNA (Figure 2). NER lesions mostly arise from exogenous agents (like UV), while exceptions include some kinds of oxidative lesions. This pathway is one of the main models used in this study and will be analyzed more precisely in the next sections.

BER removes and replaces small chemical alterations of DNA bases. This type of lesions is more frequently related to DNA miscoding that may be responsible for causing mutagenesis. BER is

mainly triggered by damages originating endogenously (like ROS). Lesions related with either NER or BER affect only one of the DNA strands where the injured part is removed and the resulting gap is replaced by using the intact complementary strand as a template.

MMR is activated when A-G and C-T do not pair correctly, but also when erroneous DNA replication or DNA polymerase misincorporation errors result in DNA insertions/deletions. During this process, mismatches are recognized, excised, and DNA resynthesis corrects the damaged sequence.

Double strand breaks (DSBs) seem to be more problematic, since both strands are affected, however HR and NHEJ are the specialized machineries that are dealing with such injuries. HR is activated during DNA replication, taking advantage of the original version of the sequence (copy of the sequence) for aligning the breaks. NHEJ is mostly activated during the G1 phase of the cell cycle and takes advantage of the DNA ligase IV that uses the overhanging pieces of DNA adjacent to the DSBs to join and fill in the ends.

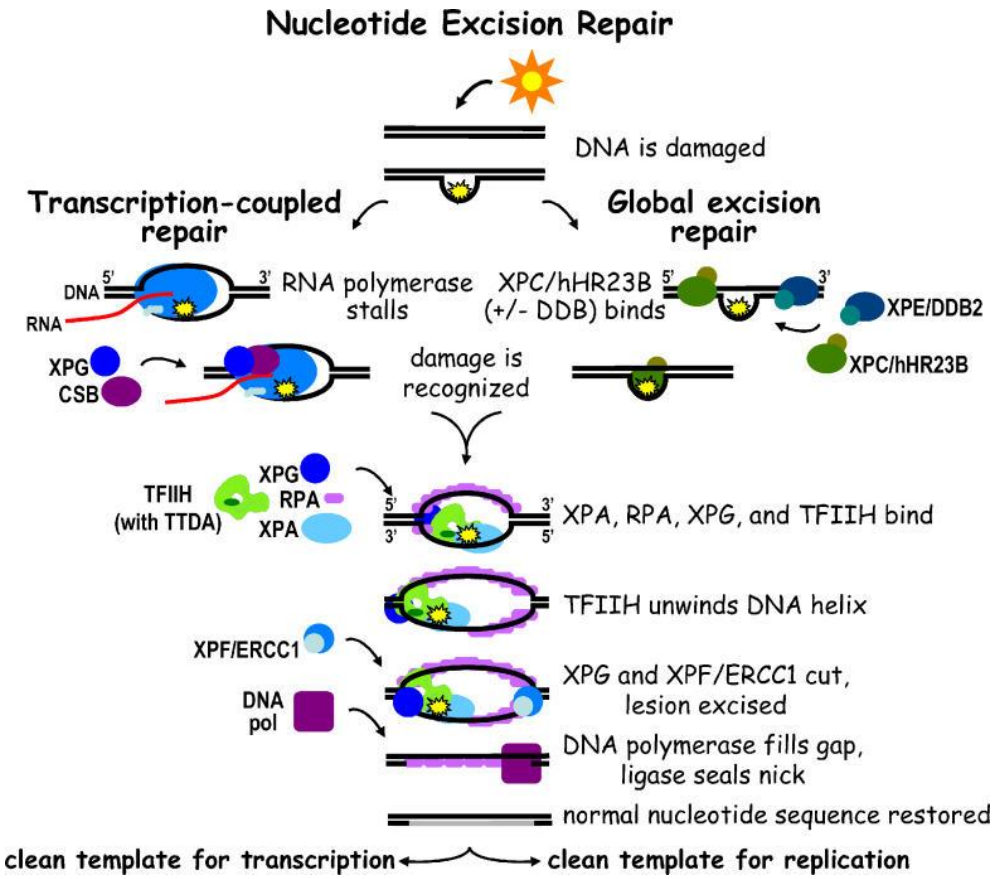


Figure 2 Nucleotide Excision Repair graphic. Image from (Fuss & Cooper, 2006).

1.3 NER

Nucleotide-excision repair (Figure 2) is responsible for the removal of the widest range of DNA lesions, including UV-derived photoproducts (6-4PPs and CPDs), numerous bulky chemical adducts, intrastrand crosslinks and ROS-generated lesions. NER is divided to two sub-pathways

(Figure 2): The Global Genome-Nucleotide Excision Repair (GG-NER), which is responsible for lesion removal throughout the whole genome, with a more stochastic activity, and Transcription Coupled-Nucleotide Excision Repair (TC-NER), which removes lesions from the transcribed strand of the actively transcribed regions.

NER is considered the most flexible repair mechanism in terms of damage recognition, since the sensing enzymes that participate in the recognition process do not focus on the lesions per se, but they rather recognize multiple formations of abnormal bulks in the DNA helix. In GG-NER, lesion sensing depends on the stochastic action of XPC--RAD23B or UV-DDB1/2 (XPE) complex. In TC-NER, damage recognition is performed by the stalling of RNA polymerase II (RNAPII) on DNA adducts and the impairing of RNA synthesis reaction during transcription. TC-NER is a faster procedure than GG-NER, but is exclusively limited to the template strand of the transcribed elements. Also, recruitment of the two TC-NER exclusive factors Cockayne Syndrome B (CSB) and Cockayne Syndrome A (CSA) at stalled RNAPII damage sites is crucial in humans and other eukaryotes for the activation and completion of the subsequent core NER reaction (Spivak, 2015; Vermeulen & Fousteri, 2013).
Following the damage detection, the two sub-pathways merge, and recruitment of the basal transcription factor TFIIH facilitates the opening/melting of DNA-containing lesion through the operation of Xeroderma pigmentosum (XP) XPB and XPD subunits.
Consequently, proteins XPA, RPA and XPG are recruited, with XPA checking the existence of any harmful DNA damage and transmitting signals to the 5′ DNA endonuclease XPF-ERCC1 complex, while RPA binds and secures the complementary to the damage single-stranded (ss) DNA, assisting the coordination of repair and the right orientation of the DNA endonucleases. The XPG DNA endonuclease associates with and grants stability to TFIIH (Ito et al., 2007), while incision 5′ to the damage by XPF-ERCC1 precedes the 3′ incision by XPG (Staresincic et al., 2009).
In turn, DNA repair synthesis factors use the undamaged strand as a template to fill the 25–30 nucleotide (nt) gap created by the excised damage-containing DNA, a procedure that is strongly affected by the cell cycle status (Lehmann, 2011). Particularly, gaps in non-cycling cells are filled by PCNA, RFC1 and DNA polymerases (DNA pol) delta and kappa, while gaps in dividing cells are filled by DNA pol epsilon and delta (Ogi et al., 2010). Accordingly, in non-cycling cells sealing of repaired DNA is performed by XRCC1-DNA ligase IIIa complex, while in dividing cells sealing of repaired DNA is performed by both DNA ligase I and XRCC1-DNA ligase IIIa (Moser et al., 2007).

Importantly, similar to all the repair machineries, NER acts on DNA in the context of chromatin. The presence of chromatin structure inhibits the repair process, thus chromatin remodeling and histones post-translational modifications (PTM) are essential (before and during the repair process) and serve as a primer of repair events, functioning further as the regulatory platform that guarantees that DNA repair is coordinated with other cellular events. After the DNA repair is completed, the prior chromatin structure must be faithfully restored. This procedure is described as the access/repair/restore model (ARR) (Green & Almouzni, 2002).
Regarding the NER sub-pathways, interaction between GG-NER and histone acetyltransferase (HAT) p300 has been reported, suggesting a functional role of p300 during the early stages of damage recognition (Datta et al., 2001; Rapić-Otrin et al., 2002). It has also been proposed that

p300 protein is recruited to UV-induced DNA lesions located in heterochromatin, contributing to the relaxation of chromatin structure in these loci (Q. E. Wang et al., 2013). DDB1 protein has also been reported to be associated with GCN5 acetyltransferase (Martinez et al., 2001), which in turn facilitates repair factors recruitment and NER induced repair through the H3K9 acetylation (H3K9ac)(Guo et al., 2011). Similarly, stabilization of the DDB2 protein by poly(ADP-ribosyl)ation leads to the recruitment of the chromatin remodeler enzyme ALC1, outlining a molecular mechanism for PARP1-mediated regulation of NER (Pines et al., 2012). Finally, chromatin assembly factor CAF-1 is believed to play a role in chromatin structure restoration after DNA repair is completed (Green & Almouzni, 2002)

Regarding the TC-NER mechanism, the CSB protein is a member of the SWI2 / SNF2 family of DNA-dependent ATPases, and has been linked to chromatin remodeling activity in vitro(Citterio et al., 2000). It has also been found to interact with acetyltransferase p300 and together with CSA are prerequisites of nucleosome binding protein HMGN1 recruitment, which enhances the rate of repair in chromatin (Birger et al., 2003; Vermeulen & Fousteri, 2013). Additionally, SPT16, a subunit of the histone chaperone FACT, facilitates H2A and H2B, which in turn are displaced at an accelerated pace from UV-induced DNA lesion sites. SPT16 is targeted to stalled RNAPII sites during TC-NER and is essential for efficiently restarting the RNA synthesis upon damage removal (Dinant et al., 2013). Finally, Histone methyltransferase DOT1L is a driver for gene expression recovery after a genotoxic insult (Oksenych et al., 2013).

Considering the above, it is clear that chromatin remodeling and histone PTMs are essential for the assembly of TC-NER at damage sites, but also for the subsequent restoration of active transcription.

1.4 Diseases related to defective repair mechanisms.

Defects in DNA repair mechanisms result in a very broad spectrum of human diseases including neurodevelopmental defects, premature ageing, neurodegeneration and cancer. (Table 2). Some characteristic examples will be described below, with an emphasis to NER-deficient related diseases (Figure 3).

Ataxia Telangiectasia (AT) is a neurodegenerative human disease, with a clinical outcome of radiation sensitivity, chromosomal instability and predisposition to cancer. AT is linked with homozygous mutations in the ATM gene (432 mutations have been reported, leading to protein instability), a protein kinase that initiates the DSB repair process (Torgovnick & Schumacher, 2015), with up to 30% of patients developing lymphoid cancer since ATM plays a critical role in T and B cells differentiation (Lumsden et al., 2004). Patients with heterozygous missense mutations have higher prevalence to develop breast, colorectal and stomach cancer (Paglia et al., 2010; Thompson et al., 2005), while hypomorphic mutations in ATR lead to Seckel syndrome, characterized by growth retardation, microcephaly and a characteristic “bird-headed” facial appearance (O’Driscoll et al., 2003).

In hereditary breast cancers, approximately 5–7% of mutations are related to BRCA1 and BRCA2, which in turn play a major role in different repair machineries. Particularly BRCA1 acts in HR and NHEJ and single-strand annealing (SSA) via its different interaction domains, while BRCA2 has the main role of mediating the recruitment of RAD51 protein to DSBs during HR.

Non-functional NER mechanism is a result of germ-line mutations in genes encoding for various factors that are involved in the different steps of the pathway. Although genetically distinct, the overlapping clinical features of these pathologies often create confusion to scientists regarding the correct classification and diagnosis of cases. Defects in GG-NER sub-pathway result in Xeroderma Pigmentosum (XP), which is an autosomal recessive and rare human disease, characterized by increased cancer risk (between 1000 to 10.000 times higher depending on the type) due to environmental stress sensitivity and an increased chance of developing tumors at internal organs. In addition, 25% of patients express progressive neurodegeneration.
Seven complementation groups with NER-deficiency have been genetically assigned in XP (XP-A to -G). An additional one, carries mutations in the POLH gene that encodes for DNA polymerase η (eta), that specializes in error--free replication of DNA damage-containing DNA, leading to XP variant (XPV) syndrome (Masutani et al., 1999).

Disorder	Main symptoms	DNA repair defect	Mode of inheritance
Xeroderma pigmentosum	Sensitivity to sunlight; slow neurodegeneration; skin cancer	NER (7 variants) pol η	Autosomal recessive
Cockayne's syndrome	Sensitivity to sunlight; growth retardation; neurological impairment; progeria	Defective NER and TCR	Autosomal recessive
Trichothiodystrophy	Sensitivity to sunlight; dystrophy; short brittle hair with low sulfur content; neurological and psychomotoric defects	Defective NER, particularly of ultraviolet-induced damage; closely related to ERCC2 and ERCC3 defects	Autosomal recessive
Down syndrome	Mental retardation; progeria	Defective repair of oxidative DNA damage (trisomy of chromosome 21)	No precise mode of inheritance
Ataxia-telangiectasia and ataxia-telangiectasia-like disorder	Progressive ataxia caused by cerebellar degeneration; progeria; wheelchair dependency	Defective DNA damage response and DSB repair	Autosomal recessive
Nijmegen breakage syndrome	Similar to ataxia-telangiectasia	Defective DNA damage response and DSB repair	Autosomal recessive
Alzheimer's disease	Progressive neurodegeneration leading to dementia, memory loss and cognitive decline	Increased oxidative stress and damage; defective repair of oxidative damage and DSB repair (nonhomologous end joining)	Autosomal dominant
Parkinson's disease	Tremor; bradykinesia; postural rigidity and postural instability; degeneration of dopaminergic neurons in substantia nigra area	Oxidative stress and DNA damage; mutations in α-synuclein and parkin variants	Autosomal dominant
Huntington's disease	Progressive chorea and dementia; severe neuronal loss in the striatum and cerebral cortex	CAG repeat expansion in *huntingtin* (*HD*) gene, and oxidative damage to DNA	Autosomal dominant
Several spinocerebellar ataxias	Various problems with bodily movements similar to those experienced with Huntington's disease; progressive loss of neurons in various loci	Expanded CAG repeats in various genes	Autosomal dominant
Friedreich's ataxia	Limb ataxia; cerebellar dysarthria; sensory loss; skeletal deformities	GAA expanded repeats in *frataxin* (*FXN*) gene	Autosomal recessive
Myotonic dystrophy types 1 and 2	Muscle weakness and wasting; cataracts; testicular atrophy; cognitive decline	CTG expansion (type 1); CCTG expansion (type 2)	Autosomal dominant
Spinocerebellar ataxia with axonal neuropathy-1	Progressive degeneration of postmitotic neurons	Mutated DNA tyrosyl phosphodiesterase 1 (*TDP1*) gene needed for SSB repair	Unknown
Triple-A syndrome	Adrenal insufficiency; achalasia; alacrima; neurodegeneration; autonomic dysfunction	Mutation in *AAAS* gene, which encodes ALADIN protein	Autosomal recessive
Amyotrophic lateral sclerosis	Progressive degeneration of motor neurons; muscle weakness and atrophy, leading to fatality	Defective Cu–Zn superoxide dismutase (SODC; SOD1); oxidative stress; defective DNA repair (BER?)	Autosomal recessive

Table 2 Disorders that arise by defective DNA repair mechanism. Table from .

Defective TC-NER results in Cockayne Syndrome, a progeroid disorder that is characterized by severe developmental abnormalities and mental retardation (Marteijn et al., 2014). CS is another rare human disease (2.7 per million live births) and it was first reported in 1936 (Cockayne, 1936) by the English physician Edward Alfred Cockayne (1880–1956), who made the first description of the features of the syndrome, based on the clinical characteristics of two siblings that expressed dwarfism, deafness and retinal atrophy (Cockayne, 1946). Mutations in CSA and CSB genes that encode for the homonymous indispensable TC-NER proteins is shown to be responsible for the classical CS pathology (Mayne & Lehmann, 1982; Tanaka et al., 1981). Approximately 60% of CS mutations were identified in the CSB gene, while the rest in CSA, but without a clear genotype / phenotype relationship (Laugel et al., 2010). CS patients live an average of 12 years and the clinical symptoms of the syndrome include (in addition to those aforementioned) cutaneous photosensitivity, deafness, cataracts, large cold extremities, growth and developmental abnormalities, microcephaly, dysmyelination, demyelination, increased brain calcification and vasculopathy, progressive neurodegeneration, and mental retardation (Karikkineth et al., 2017).
.

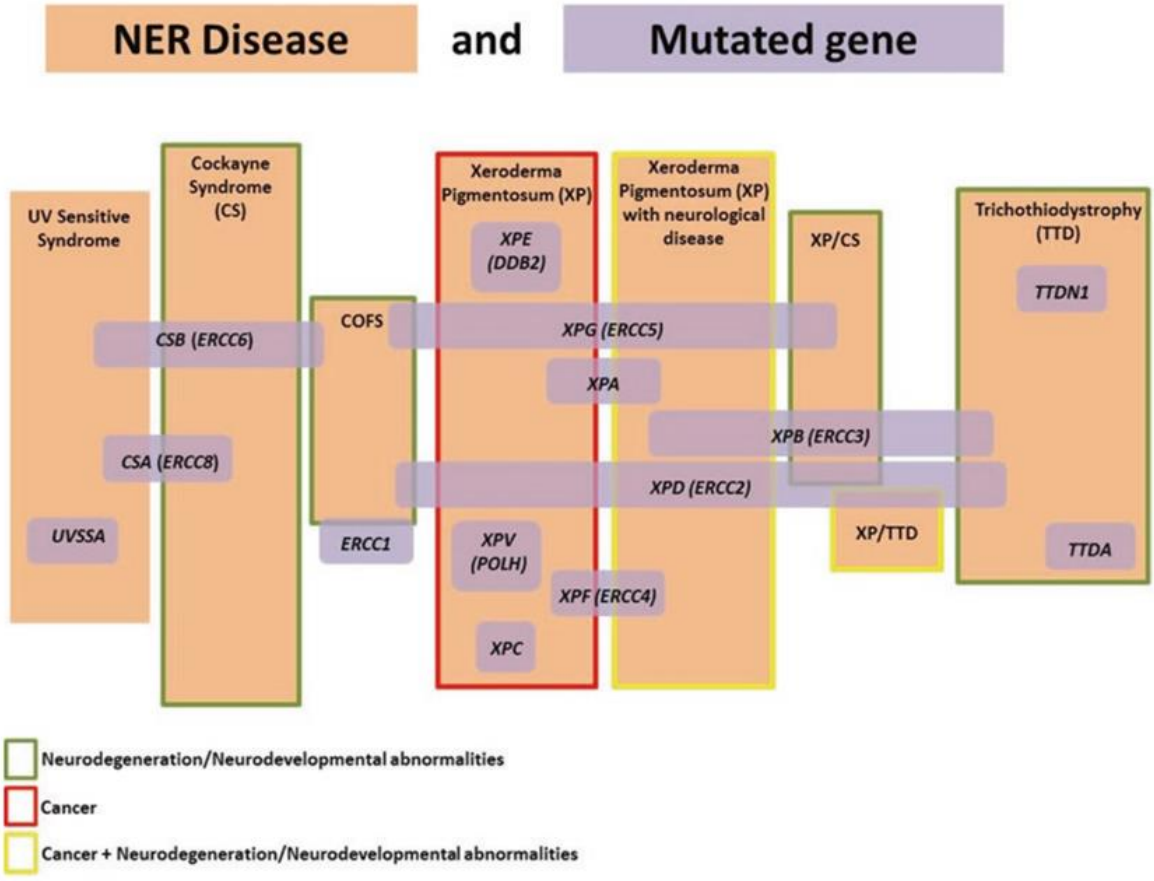


Figure 3 Genotype/phenotype relationships between NER disorders highlighting the overlap with observed neurological and cancer abnormalities. Adapted from (Liakos et al., 2017).

Additionally, some mutation in XPB, XPD and XPG encoding genes can lead to combined XP/CS phenotypes. The clinical characteristics of these patients include the skin disorders of the XP syndrome and the neurological abnormalities of the CS disease, as also severe developmental abnormalities, underdeveloped reproductive system and retinal atrophy. Moreover, patients with germ-line mutations in the genes XPA, XPB, XPD, XPF, XPF, XPG may express progressive neurodegeneration in combination with cancer depending on the genomic location of the mutation. The particular neurological disorders include progressive deafness, abnormal gait, mild microcephaly while in severe cases there is extensive neuronal death in various areas of brain, spinal cord and peripheral nervous system (Kraemer et al., 2007).

Cerebro-oculo-facio-skeletal Syndrome (COFS) is a rare human autosomal recessive syndrome characterized by microcephaly, cataracts and / or microphthalmia, severe developmental abnormalities, arthrogryposis, severe postnatal growth failure, facial dysmorphism and mental retardation. This syndrome is considered a NER-related disease, associated with a defective TC-NER mechanism. Mutations linked with COFS have been identified in the CSB, XPD, XPG and ERCC1 genes (Suzumura & Arisaka, 2010).

Mutations in the XPB, XPD, TTDN1 and TTDA genes can also cause an autosomal recessive disease called Trichothiodystrophy (TTD), that is characterized by brittle hair with a lack of sulfur, skin fading, developmental problems in the nervous system and demyelination (Kraemer et al., 2007). The lack of increased cancer susceptibility in TTD patients creates a partial overlap of symptoms with those of CS, however deafness, optic atrophy and cachexia are absent in TTD (Rapin, 2013).

UV Sensitive Syndrome (UVSS) is an autosomal recessive human disorder, with some common clinical characteristics with CS (photosensitivity, telangiectasia and freckles), but UVSS patients do not express the severe developmental and neurological abnormalities that CS patients express. This syndrome is linked with mutations found in CSB (Horibata et al., 2004), CSA (Nardo et al., 2009), and UVSSA. UVSSA protein has been found to interact with RNA polymerase II and other complexes of the TC-NER mechanism, and that it stabilizes CSB protein through interaction with USP7 protein (Nakazawa et al., 2012; Schwertman et al., 2012; X. Zhang et al., 2012).

Thus, mutations in one of the genes that are involved in the NER mechanism can lead to different diseases (Figure 3), pointing to the fact that mutation position and consequently the amount and stability of the produced protein could correlate with the resulting complexity. For example, different mutations in the CSB gene can lead to 3 different diseases related to the TC-NER sub pathway (UVSS, CS and COFS). Conversely, a disease may be the result of mutations in several genes involved in the NER mechanism.
Deficiencies in NER can directly result in increased mutation rates in affected cells that in turn may lead to carcinogenesis (Helleday et al., 2014; Marteijn et al., 2014). Most common NER-related mutations result by miss-replication of damaged and unrepaired DNA. In particular, UV related mutational signatures, which are associated to NER-deficient mutations (C > T), or smoking mutations (C > A) were identified in skin and lung cancers genomes (Hefferin & Tomkinson, 2005; Pleasance, Stephens, et al., 2010) respectively (Figure 4). Similar mutational

asymmetries have been reported to be associated with the TC-NER pathway and be related to mutagenesis in liver cancer (A > G) (Haradhvala et al., 2016). Notably, genome-wide quantification of mutation density, has uncovered a reduced mutation rate located at NER intact regions, such as DNA regulatory elements (Polak et al., 2014), and the complementary DNA strand of actively transcribed genes. Specifically, it was shown that in squamous cell carcinoma (SCC), lower mutation levels around DNAse1 hypersensitivity sites is related to XPC activity (Perera et al., 2016), while in melanoma and lung adenocarcinoma cancers, a lower mutation prevalence is observed in the complementary strand of all active genes, independently of the transcription levels, because of the homogenous activity of the TC-NER machinery (Alexandrov et al., 2013; Lavigne et al., 2017).

Therefore, it seems that the probability of developing a tumor strongly depends on the balance between the number of DNA damages accumulated in the cell and the capability of the repair machineries to handle them, in concert with the timely initiation of the appropriate cell-cycle check-points, or the programming of cell-death.

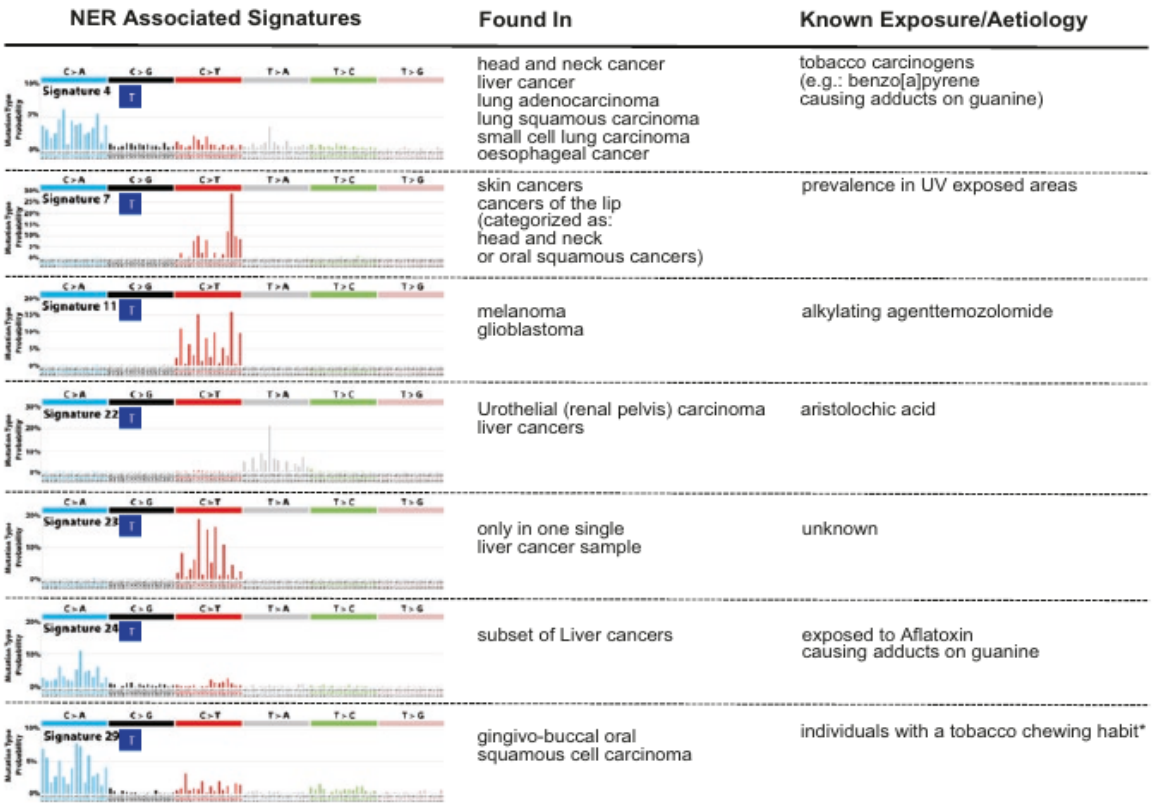


Figure 4 Mutational trinucleotide signatures identified from TCGA WES and WGS datasets, derived by cancer genomes, showing transcription strand bias (blue T boxes) and NER specific attributes. *: Different from mutational signature 4 (tobacco smoking). Adapted from (Liakos et al., 2017) Figure 2.3.

1.5 Repeats

Repetitive DNA is a major component of eukaryotic genomes. It has been estimated to comprise ~50% of the human genome, while there are computational studies reporting that this percentage might be even higher (66-69%) (de Koning et al., 2011).
There are two main groups of repeats in eukaryotes, tandemly repeated satellites, restricted to specific chromosomal regions, and repeats interspersed with genomic DNA. Interspersed repeats consist of mainly inactive copies of a large collection of currently and anciently active transposable elements (TEs) like DNA transposons and retroelements, which can be further classified into more distinct categories. Repetitive DNA sequences are considered to have played a major role in evolution of eukaryotic genomes (Garcia-Perez et al., 2016; Kidwell & Lisch, 1997). The particular sequences are considered to have a potential role in genetic variation and regulation, while their high tendency for co-localization within nuclear space, suggests that their genomic position may play a role in genome folding (Cournac et al., 2016; Shapiro & Von Sternberg, 2005).
Since repair mechanisms tend to prioritize their actions in functional regions in order to avoid critical cell dysregulation, it is yet unclear how damages in DNA repeats are treated during these processes. To shed more light in this question, and using NER products as a model (aniFOUND-seq, see materials and methods), a genome wide analysis methodology of NER-repair activity along these regions has been developed and is described in detail in the results section 4.4.

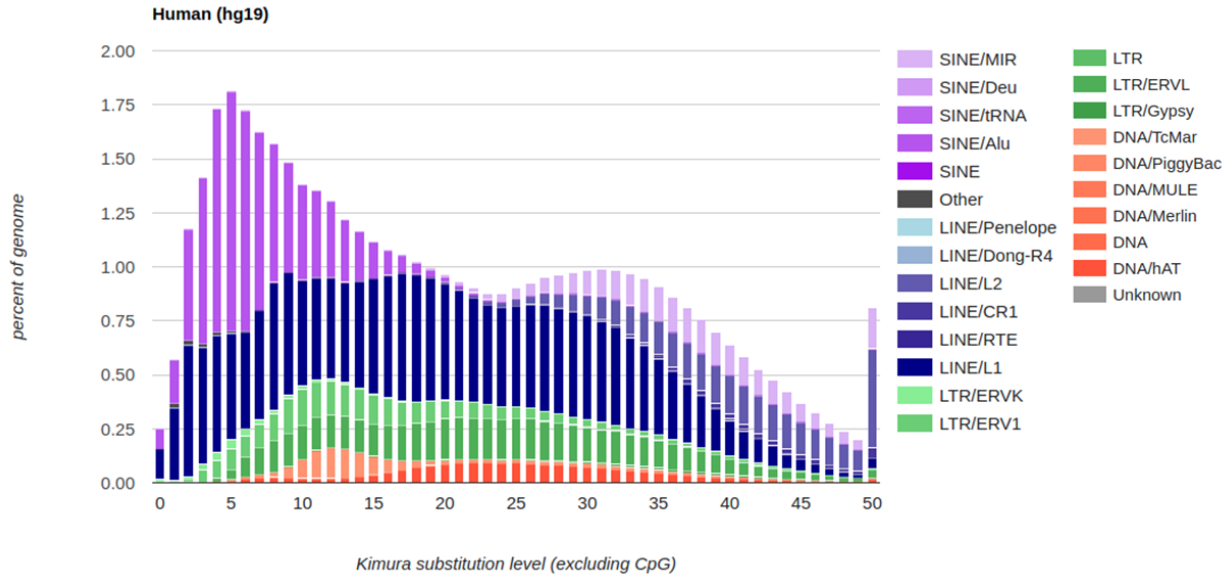


Figure 5 Repeats in human genome (UCSC hg19 buid). Figure from http://www.repeatmasker.org/

1.5.1 Telomeres

Telomeres are a special type of nucleoprotein complexes, located at the end of eukaryotic chromosomes, and they are considered to contain no valuable genetic information. Telomeres consist of multiple small repetitive nucleotide sequences (telomeric repeats) that form

heterochromatin and they participate in characteristic structures at the ends of chromosomes that are called “telomere caps”.

In humans, every telomere contains about 5,000 repetitions of the TTAGGG sequence. Telomeric sequences have a crucial role in cell viability, since they protect the ends of chromosomes from being identified by the cell as DSBs in need of repair. They protect the chromosome from euchromatin loss during DNA replication, and they also prevent chromosomes from binding one another (Blasco, 2005; Soediono, 1989). Telomere length is a determinant of cell reproductive age, and when it reaches a minimum "critical" length, "reproductive aging" is induced, which protects the organism against carcinogenesis. After each cell devision, telomeres’ length is reduces until a critical point is reached and DNA damage response is activated, leading cellular senescence or apoptosis. Decreased telomere length in healthy cells has been linked to diseases such as cancer, heart disease, diabetes, arteriosclerosis, pulmonary fibrosis, obesity.
To circumvent the limited number of possible cell divisions, tumors employ activation of telomerase or alternative lengthening of telomeres (ALT) as telomere maintenance mechanisms (TMMs) (Blasco, 2005). Telomerase is an enzyme that adds t-type repeats to the chromosome ends. In contrast, ALT is based on recombination of telomeric regions and results in several characteristics, including telomeres of heterogeneous length and sequence composition. These TMMs are crucial for tumorigenesis, making them valuable drug targets for cancer therapy. However, to precisely identify and interfere with these mechanisms in various tumor types, more insight into the different telomere structures is needed.

1.6 RNA polymerase II transcription machinery

RNA polymerase II (RNAPII) transcription is a fundamental and highly regulated cellular process, one of the most important steps in control of cell growth and differentiation. During transcription, encoded genetic information of DNA is transmitted to the messenger RNA (mRNA), in a process where the enzyme RNAPII uses coding DNA sequences as a template to synthesize RNA.
In eukaryotic organisms, synthesis of mRNA and some classes of non-coding RNAs like long non-coding RNA (lncRNA), microRNA (miRNA), some small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA), is performed by RNAPII. RNAPII enzyme, is a 550 kDa multiprotein complex consisting of 12 different subunits in humans (Hahn, 2004). RPB1 is the largest subunit of RNAPII, and its Carboxy-Terminal Domain (CTD) is composed by 52 heptapeptide repeats of the consensus Tyrosine-Serine-Proline-Threonine-Serin-Proline-Serine ($Y_1S_2P_3T_4S_5P_6S_7$) (Egloff, Dienstbier, et al., 2012). A special feature of the amino acid residues of YSPTSPS is that they can be post-translated modified independently, creating a great variety of combinations that characterize and specify the different stages of the transcription cycle (see below). CTD is a signaling and interaction platform between the transcription machinery and other factors that contribute in RNA splicing, mRNA modification (Bentley, 2014), and also with factors that modify the CTD, thus regulating the transcription process in a recursive manner (Egloff, Dienstbier, et al., 2012). The subunit complex RPB4/6 is required for the initiation of the transcription machinery, while the other 10-subunit catalytic core is capable of elongating the RNA transcripts.

1.6.1 Transcription cycle

As stated above, the CTD heptapeptide YSPTSPS repeats trigger a variety of post-translational modifications. Specifically, tyrosine, threonine and all three serines can be modified through phosphorylation. Even the two prolines can be divided between cis- and trans- conformation (Heidemann et al., 2013). Other CTD residue modifications include glycosylation and methylation (Kelly et al., 1993; Sims et al., 2011).
The phosphorylation status of RNAPII is regulated by a number of kinases and phosphatases, which act on the various stages of the transcription cycle and regulate its process. To clarify the functional role of phosphorylation in transcription, various monoclonal antibodies were developed to target the different isoforms of RNAPII (Heidemann et al., 2013) and used in Chromatin Immunoprecipitation (ChIP) techniques (Bataille et al., 2012; H. Kim et al., 2010; Mayer et al., 2010). The findings of these studies led to the concept of the "CTD code" and also the definition of transcription cycle (Komarnitsky et al., 2000).

1.6.1.1 Transcription Initiation

The first step of the transcription cycle is the formation of the Pre-Initiation Complex (PIC) at the promoters of transcribed elements (Rapić-Otrin et al., 2002; Sikorski & Buratowski, 2009; M. C. Thomas & Chiang, 2006). In summary, TFIID, a basal transcription factor, recognizes specific sequences in the promoter region (such as the TATA sequence and the DPE-Downstream Promoter Elements region) and binds to it. This is followed by the recruitment of the general transcription factors TFIIA, TFIIB, and TFIIF together with a hypophosphorylated CTD-containing RNAPII complex (RNAPII-hypo). This complex is initially unstable, as the double helix in the promoter region is not accessible. Subsequent binding of the transcription factor TFIIH, which contains protein subunits with helicase action (ATP-dependent), modifies the DNA by double helix strand dissociation. At the unwound DNA region of the promoter, the transcription bubble is formed and the PIC is stabilized (Figure 6).

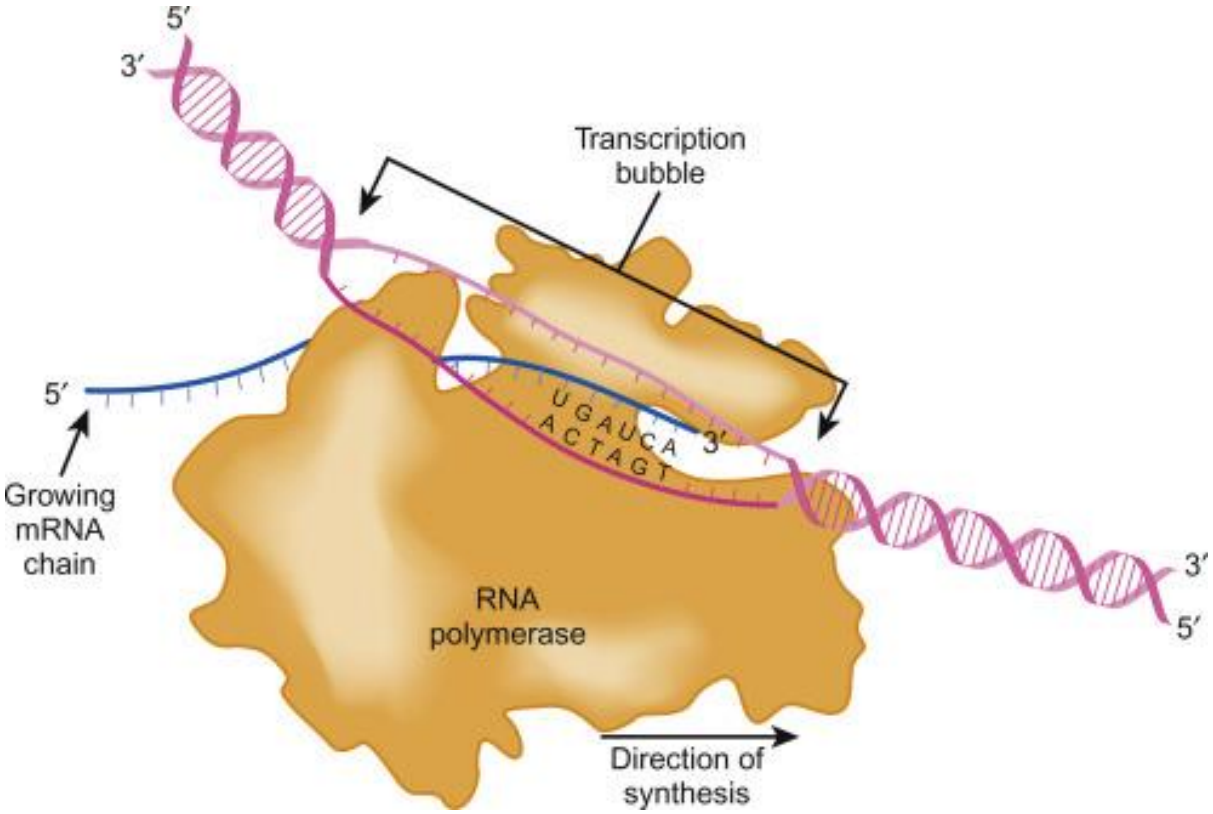


Figure 6 Transcription bubble during transcription elongation. The initiation of RNA synthesis is depicted. DNA strands are separated, and synthesis of complementary RNA takes place at the transcribed strand, while RNA polymerase remains bound at the promoter site. Figure from (Molecular Biology, 3rd Edition, 2019, Clark, Pazdernik and McGehee).

RNAPII starts composing RNA resulting in the production of small RNA molecules, smaller than 10 nucleotides, which are eventually eliminated. Whether RNA polymerase will progress to the early elongation step depends on TFIIH and other general transcription factors, such as the Mediator complex (Boeing et al., 2010; Hirose & Ohkuma, 2007; Levine, 2011). In particular, the CDK7 kinase, which is part of TFIIH protein complex, phosphorylates S_5 of $Y_1S_2P_3T_4S_5P_6S_7$ and RNAPII is able to escape from the promoter and slowly progress through the transcription start site (TSS) of the genes (Levine, 2011).

1.6.1.2 Promoter Proximal Pausing

When polymerase reaches +30 to +50 nucleotides from the (TSS), it pauses, in areas called Proximal Promoter Pausing (PPP) Sites (Gilmour & Fan, 2009). RNAPII pausing is mediated by DRB Sensitivity-Inducing Factor (DSIF) and Negative Elongation Factor (NELF), which bind to the newly synthesized RNA molecule. The release of RNAPII from PPP sites is considered a very important molecular switch of gene expression during development (Levine, 2011) (Figure 7).

1.6.1.3 Transcription Elongation

Transcription elongation begins with the release of RNAPII from PPP, a tightly regulated process, which depends also on various developmental and environmental signals. Initially, the Positive Transcription Elongation Factor b (P-TEFb) complex is recruited to PPP regions. The P-TEFb complex consists of the cyclin dependent kinase CDK9 and one of several cyclin subunits, cyclin T1, T2, and K (Fu et al., 1999). P-TEFb is essential for the regulation of RNAPII transcription elongation, as it phosphorylates S_2 of RNAPII CTD heptad repeats but also NELF (negative elongation factor) factor, which is subsequently removed, and the DSIF factor, which is converted to a positive elongation factor. After these steps, RNAPII is released into productive elongation and progresses fast towards the 3’ end of gene bodies synthesizing mRNA. (Lavigne et al., 2017; Sainsbury et al., 2015) (Figure 7).

1.6.1.4 Transcription Termination

When RNAPII passes through an active poly (A) site (PAS), while travelling through the transcribed element, cleavage and polyadenylation (CPA) factors bind to both the transcript and the RNAPII molecule. These factors include CPSF (cleavage and polyadenylation specificity factor), CstF (cleavage stimulatory factor), CGI (cleavage factor I) and CFII (cleavage factor II), and are responsible for the cleavage and the polyadenylation of the nascent RNA (nRNA) molecule. CPSF binds directly to the nRNA molecule, while CstF, CFI and CFII bind to the phosphorylated serine 2 of RNAPII CTD. CPSF and CstF also recognize specific patterns at the 3 'end of the newly formed RNA. Due to these interactions, transcription decelerates and pauses. Then, the RNA molecule is cleaved, and the 3 'end is polyadenylated, steps that facilitate the exit from the nucleus to the cytoplasm and the forthcoming translation. SETX (setaxin) is also reported to be involved in the transcription termination process of some genes, possibly by disassembling R-loops (M. Thomas et al., 1976), to allow the entry of 5'-3' exoribonuclease 2 (XRN2). Degradation of the 3’ region segment of the newly formed RNA by XRN2, results in the termination of transcription (tornado model) (Porrua & Libri, 2015) (Figure 7).

In some cases, termination might occur at several positions inside the transcribed element, for the prevention of aberrant transcript formation, but also the production of different transcripts (alternative transcripts). This can result in different mRNAs with altered regulatory properties or different encoded proteins. Finally, termination can be blocked/forced in response to particular cellular signals, as in cancer or virally infected cells. In such cases, unsuccessful transcriptional responses may have disastrous effects for the cell (Proudfoot, 2016).

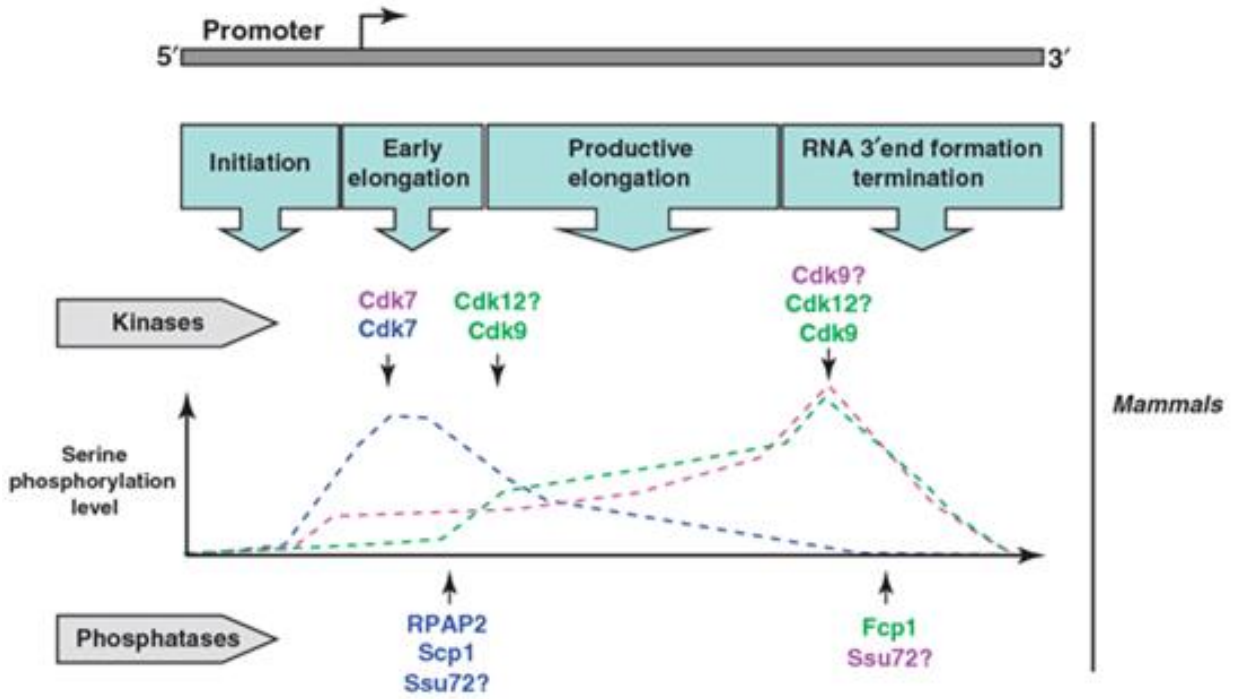


Figure 7 Transcription cycle (Egloff, Zaborowska, et al., 2012).

1.6.2 Non-coding transcription

In earlier decades, research on transcription had focused on protein coding genes due to their abundance and how easy it was for researchers to isolate their sequences, as also their associated transcripts. However, the development of new technologies with greater sensitivity and discretion revealed that only 2% of the genome corresponds to protein-encoding genes (Dunham et al., 2012), while it was also found that approximately 62-75% of the human genome is transcribed (Dunham et al., 2012).

Thus, in addition to non-coding (nc) transcripts such as transfer RNA (transfer RNA, tRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), research interest became also focused on other types of small non-coding transcripts like microRNAs (miRNAs), PIWI-associated RNAs (piRNAs), small interfering RNAs (siRNAs), but also long non-coding RNAs, lncRNAs like long intervening noncoding RNAs (lincRNAs), Natural Antisense Transcripts (NATs), enhancer RNAs (eRNAs), circular RNAs (circRNAs), competing endogenous RNAs (ceRNAs), PROMoter uPstream Transcripts (PROMPTs) and others (Figure 8).

ncRNAs are functional RNA molecules that lack protein coding capacity, but act through multiple mechanisms that regulate gene expression. These mechanisms include RNA-RNA base pairing, RNA-protein interactions and intrinsic RNA activity, gene splicing, nucleotide modification, protein transport, regulation of gene expression through degradation, regulation of diverse cellular functions such as RNA processing, mRNA stability, translation, protein stability and secretion (Szymański et al., 2003).

The distance of ncRNAs from their target protein-coding genes is more highly conserved than their RNA sequence, implying that position-specific cis effects are driving ncRNA evolution

(Kaikkonen & Adelman, 2018). ncRNAs are known to have a strong effect in epigenetic signaling, as they play an important role in genomic imprinting (Koerner et al., 2009), in chromatin remodeling and in defining DNA methylation patterns. Moreover, recent studies suggest that “the act” of transcription modulates chromatin accessibility, transcription factor occupancy, and epigenetic state, rather than the sequence or nature of the ncRNA product (Kaikkonen & Adelman, 2018). ncRNA activity occurs in a cell type (Qiu et al., 2017), tissue (Roadmap Epigenomics Consortium et al., 2015) and developmental stage specific manner, and their dis-regulation may result to pathogeny. In the particular thesis, the transcription activity at lncRNAs, PROMPTs and eRNAs, as well as of protein coding genes will be addressed.

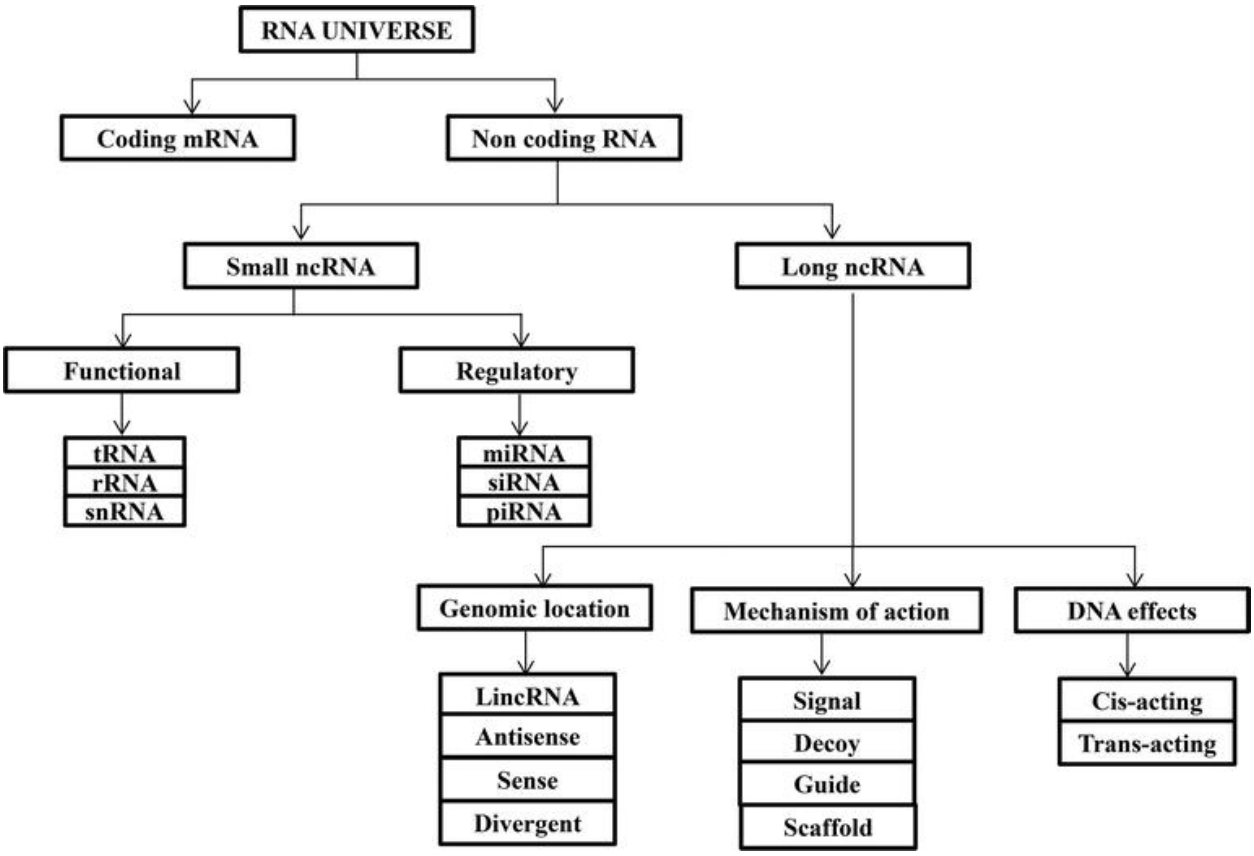


Figure 8 RNA Classification tree and the roles of non-coding RNAs in Transcriptional Regulation. Adopted by (Sriyothi et al, 2018).

1.6.2.1 eRNAs

Enhancer transcription units are very similar to protein coding genes promoters. In enhancers, transcription seems to start from a nucleosome depleted region surrounded by nucleosomes, at the edges of which independent PICs are formed that can trigger bidirectional transcription (sense and antisense in respect to the Watson strand). Additionally, it has been confirmed that binding of RNAPII is also present in enhancer elements’ promoters (Core et al., 2014; de Santa et al., 2010; Koch et al., 2011), while the typical steps of transcription cycle are also detectable,

including transcription initiation, transcription pausing in a distance of about 70 bases from the enhancer transcription start site (eTSS), and transcription elongation (Henriques et al., 2018). As mentioned above, enhancers are transcribed bidirectionally, producing similar amounts of eRNAs in both directions (Andersson et al., 2014). Enhancer transcription products are relatively short in length (~2kb), non-coding, unstable (Andersson et al., 2014), and sensitive to exosome degradation (Andersson et al., 2014; Core et al., 2014; Henriques et al., 2018), while the production levels are highly correlated with the enhancer's functional activity (W. Li et al., 2016).

Enhancers interact with DNA to upregulate gene transcription through enhancer-promoter looping and tracking of the transcriptional machinery (W. Li et al., 2016). This fact suggests that eRNAs favor enhancer activity and thus affect the protein-coding gene transcription, but there is also evidence that the act of transcription per se might also play a role in the activation of target genes (W. Li et al., 2016). This can be done either by creating a favorable chromatin environment for the activation of protein-coding genes located at distal regions, or in the case of some intragenic enhancers, by attenuating the host gene expression, thus regulating important cellular processes (Cinghu et al., 2017). Additionally, eRNAs could be bound to transcription repressors, to inhibit their function.

These mutually exclusive functions suggest that enhancers and their products may be functionally and mechanistically diverse, but further evidence is needed to fully understand their functions in gene regulation, development and disease.

1.6.2.2 PROMPTs

Recent studies have shown that the majority of transcriptionally active protein-coding genes show patterns of antisense transcription activity, initiating either upstream (divergent) or downstream (convergent) of the "host" TSS (Andersson et al., 2014; Core et al., 2008; Ntini et al., 2013). This phenomenon of bidirectional transcription occurs in a significant fraction of active promoters (Meng & Bartholomew, 2018) and is lately considered a general feature of protein-coding transcription (Andersson et al., 2015). These transcripts share some characteristics with eRNAs since they are relatively short, non-coding and unstable as they are degraded rapidly by the RNA exosome, but in contrast to eRNAs, they have a poly-A tail. Except from their antisense activity (asPROMPTs), some PROMPTs are transcribed in the sense direction of their host protein-coding promoter (assuming an extended promoter region +/- 2kb relative to TSS) (Ntini et al., 2013; Preker et al., 2008).

PROMPTs' transcript production seem to be positively correlated with the host gene's transcription activity, suggesting a possible role in the regulation of protein gene expression. It has been also shown that divergent asPROMPTs show a higher abundance of 3' poly(A) signals than divergent PROMPTs elements, resulting in a more rapid degradation, that in turn enables efficient elongation of downstream transcripts (Ntini et al., 2013). This imbalance might enforce the choice of promoter directionality. Finally, it has been shown that RNA levels of certain PROMPTs are altered in stress conditions, suggesting a possible regulatory role of this subset of elements that participate in some ways in the DDR process (Lloret-Llinares et al., 2016).

Nevertheless, the general function of PROMPTs remains obscure, and additional scientific efforts could unravel their potential role in transcription and regulation of gene expression.

1.6.3 Transcription during UV irradiation

UV-C induced stress affects transcription by interrupting the progression of an actively elongating RNAPII molecule. In particular, RNAPII complexes are stalled at DNA lesions, a phenomenon that triggers the recruitment and assembly of TC-NER factors.
Early in vitro experiments showed that RNAPII could remain stalled at CPD lesions for 20 hours (Selby et al., 1997), while in mouse CSB-deficient cells, it was found that it could remain stalled for more than 48 hours in vivo (Garinis et al., 2009). Additionally, the "footprint" of an RNAPII molecule that is stalled at a CPD lesion has been found to be “covering” the damage site 10 bases in front of the CPD and 25 bases behind it (Tornaletti et al., 1999). Since the TC-NER factors need access to the damage sites, the respective stalled RNAPII molecules should be removed after the damage recognition step.
Several models have been proposed for the fate of the stalled RNAPII; (Bregman et al., 1996) suggested that the damage-stalled RNAPII molecules are targeted by ubiquitination and then removed and degraded. In line with this model, a recent study suggests that the total damage-recovery of genes requires a continuous supply of RNAPII elongating molecules, as any molecule that encounters a damage will be removed through ubiquitination (Chiou et al., 2018). According to this model, the recognition of the next damage site (relative to the repaired damage site, in the direction of gene transcription) will be performed by the trailing RNAPII molecule. However, other studies suggest that this mechanism is acting only when the damage cannot be repaired, and the recovery of the transcriptional is impossible (Anindya et al., 2007; Woudstra et al., 2002).
The second model claims that RNAPII “backtracks” from the damage site, giving access to the repair factors to perform their function. This backtracking is followed by the activation of the nucleolytic activity of RNAPII, which cuts the overhang of the newly synthesized transcript, thus allowing the smooth recovery of transcription when the damages are repaired (Hanawalt & Spivak, 2008; Vermeulen & Fousteri, 2013). During this process, an important role is believed to be played by the TFIIS factor, which induces the nucleolytic activity of RNAPII (Donahue et al., 1994; Sigurdsson et al., 2010), and is colocalized with RNAPII at the regions of DNA lesions (Fousteri et al., 2006). Furthermore, a decrease in TFIIS levels has been found to lead to an abnormal recovery of cell transcription after UV irradiation (Jensen & Mullenders, 2010). The backtracking of RNAPII from the lesion site requires the relaxation of the chromatin structure behind the molecule, so that it can slide backwards. It has been suggested that proteins such as p300 and HMGN1 may modify the nucleosomes behind the stalled RNAPII, creating a looser structure to facilitate this process (Hanawalt & Spivak, 2008).

1.7 Chromatin and transcription

Eukaryotic genomic DNA coexists with proteins, forming a complex that is known as chromatin. The configuration of this complex and its relative flexibility regulate the overall function of the genome. Chromatin is a nucleoprotein complex that consists of repetitive histone octamers wrapped by DNA, forming special structures, called nucleosomes (Figure 9).

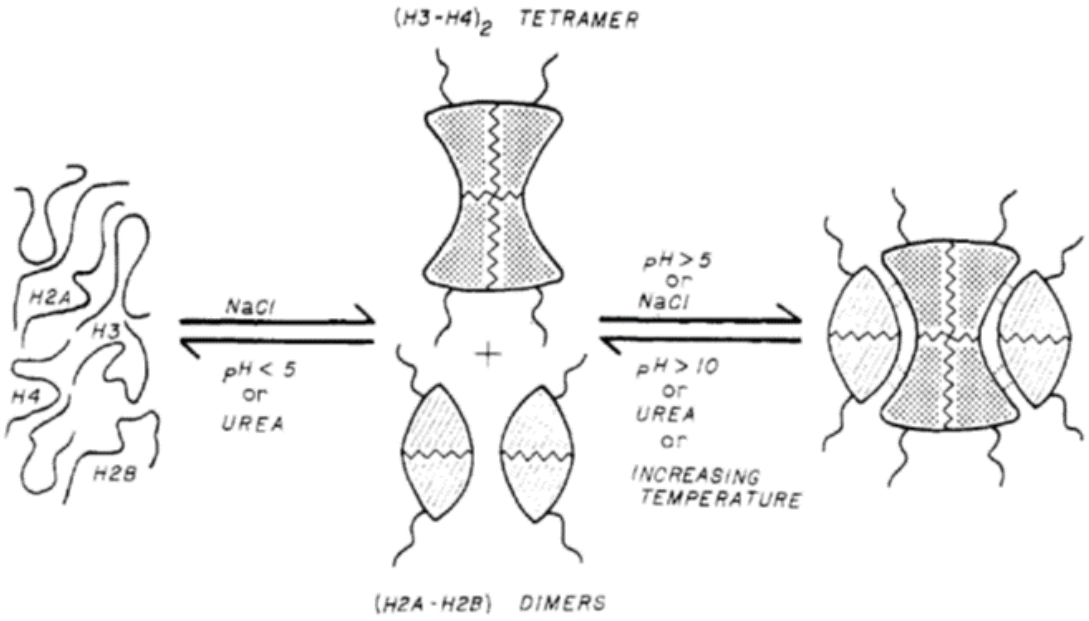


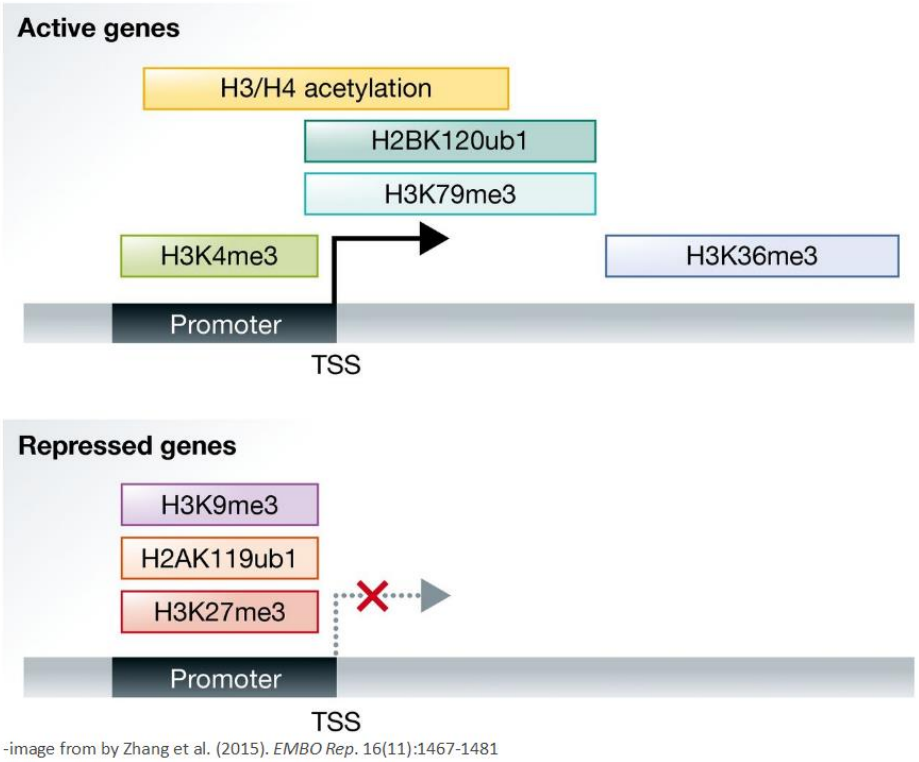
Figure 9 Histone octamer. Adopted from (Eickbush & Moudrianakis, 1978).

Histone octamers are composed by two histone 2A (H2A) - histone 2B (H2B) dimers, and a tetrameric core of histone 3 and histone 4, as depicted in Figure 9. In turn, nucleosomes are composed by 147 base pair (bp) length DNA, which wraps 1.7 times around the histone octamers, organizing very long DNA sequences into small structures. At the entry and exit sites located on the surface of the nucleosome core, DNA is bound by histone H1, known also as linker histone (Kowalski & Pałyga, 2012). This primary, and simple chromatin structure can be transformed to higher-order structures through interactions between histones and the linker histone. Through nucleosome forming, the DNA is compacted up to 20,000 times more so it can fit in the small volume of the nucleus.

Nucleosomes are constantly in a dynamic state and are flexible to alterations in order for cellular processes such as transcription, replication and DNA repair to take place in the context of chromatin. For this reason, several protein complexes are responsible for the rearrangement of nucleosome structure, reposition and redistribution (Zentner & Henikoff, 2013). Dysfunction of chromatin remodeling mechanisms has been associated with human disease, including cancer. Chromatin structure may be modified by several mechanisms. ATP-dependent remodeling complexes use ATP hydrolysis energy to shift the nucleosomes and swap or remove histones from the chromatin fiber. Histone variants create localized specific domains within the chromatin fiber, while histone chaperones control the delivery of free histones and act synergistically with chromatin remodelers during histone deposition and removal. Finally and most in focus in this study, post-translational modifications (PTMs) of histones, such as phosphorylation, ubiquitination, methylation and acetylation (Figure 10), directly or indirectly influence chromatin structure. These mechanism act cooperatively to regulate the chromatin structure and DNA accessibility (see next chapter) (Rossetto et al., 2012).

Figure 10 Schematic of the most common epigenetic modifications. Adopted from Abcam.

Histone PTMs affect gene expression, without changing the DNA nucleotide content. The histone N-termini that overhang from the packed octamer, are less structured and thus more exposed to PTMs and the enzymes (‘writers) that deposit them (Bannister & Kouzarides, 2011; Kouzarides, 2007). There are at least 9 different types of PTMs, summarized in Figure 10. Histone PTMs play an important role in processes such as replication, transcription, repair and packaging of DNA. Histone PTMs seem to act in two main ways. First, by affecting the link between nucleosomes and DNA, causing either local unwinding of the structure, or further condensation, and secondly, by inducing protein (PTMs ‘readers’) recruitment that further modifies chromatin through their enzymatic action. Furthermore, specific histone PTMs have been associated with cell cycle stages, as well as with regulatory genomic regions such as enhancers and promoters (Figure 11) (Smolle & Workman, 2013).



-image from by Zhang et al. (2015). *EMBO Rep*. 16(11):1467-1481

Figure 11 Histone PTMS associated with transcriptional activation and repression.

In regard to transcription, histone modifications can be divided into two major groups: those associated with actively transcribed chromatin that is called “euchromatin”, and those associated with inactive transcriptional chromatin, which is termed heterochromatin. Histone PTMs are functioning in activating and suppressing transcription, but their outcome depends on both the modified histone residue per se, as also their relative position in the genome. In this thesis, the focus will be centered in the acetylation of lysine 27 residue of the histone H3 protein (H3K27ac) modification, a major mark of associated to transcription activation, and the tri-methylation of the same lysine residue of histone H3 (H3K27me3), a PTM that is correlated with transcriptional repression of nearby genes

H3K27ac is a characteristic mark of active transcription in promoter and enhancer regions of mammalian genome, and a valuable tool for the identification of actively transcribed elements (Creyghton et al., 2010; ENCODE et al., 2012), while H3K27me3 is a characteristic modification of repressed elements. These two marks seem to exhibit mutual exclusive patterns of chromatin binding (Karlic et al., 2010; Shlyueva et al., 2014; Tie et al., 2009).

There are studies demonstrating that histone modifications turnover, and/or degradation around DNA lesions consist crucial steps in conserved pathways that assist the cell to cope with genotoxic stress (Misteli & Soutoglou, 2009; Polo & Almouzni, 2015).

1.7.1 Chromatin accessibility

The position of the nucleosomes play an important regulatory role in transcriptional activation as it regulates the “accessibility” of transcription binding sites to Transcription Factors (TFs) and other transcription complexes. Specifically, the structure of “open” (accessible) chromatin

defines a network of physical interactions through which promoters, enhancers, repressors and chromatin-binding factors simultaneously regulate gene expression. Thus, the accessible areas of chromatin are considered to be the main genomic regulatory regions (John et al., 2011) and are characterized by nuclease hypersensitivity (Gross, 1988). Consequently, chromatin accessibility plays a central role in several biological and pathological processes, such as development, differentiation (de la Torre-Ubieta et al., 2018; Maezawa et al., 2017; Murtha et al., 2015), tissue regeneration, aging, and cancer (Liu et al., 2019)(de la Torre-Ubieta et al., 2018; Maezawa et al., 2017; Murtha et al., 2015; Simon & Kingston, 2013; Tsompana & Buck, 2014). It must be noted that despite the importance of chromatin organization and post-translational histone modifications in the various cellular processes and responses, the way chromatin is reorganized in various regulatory areas of the genome, after exposure to genotoxic agents like UVC irradiation, have not yet been elucidated.

The Next Generation Sequencing revolution gave scientists the ability to develop sophisticated techniques and the opportunities to study the accessible regulatory regions of the chromatin in a genome-wide fashion. These techniques include MNase-seq, DNase-seq, FAIRE-seq, ATAC-seq, each of which has distinct characteristics, advantages and limitations (Tsompana & Buck, 2014)(Chang et al., 2018).

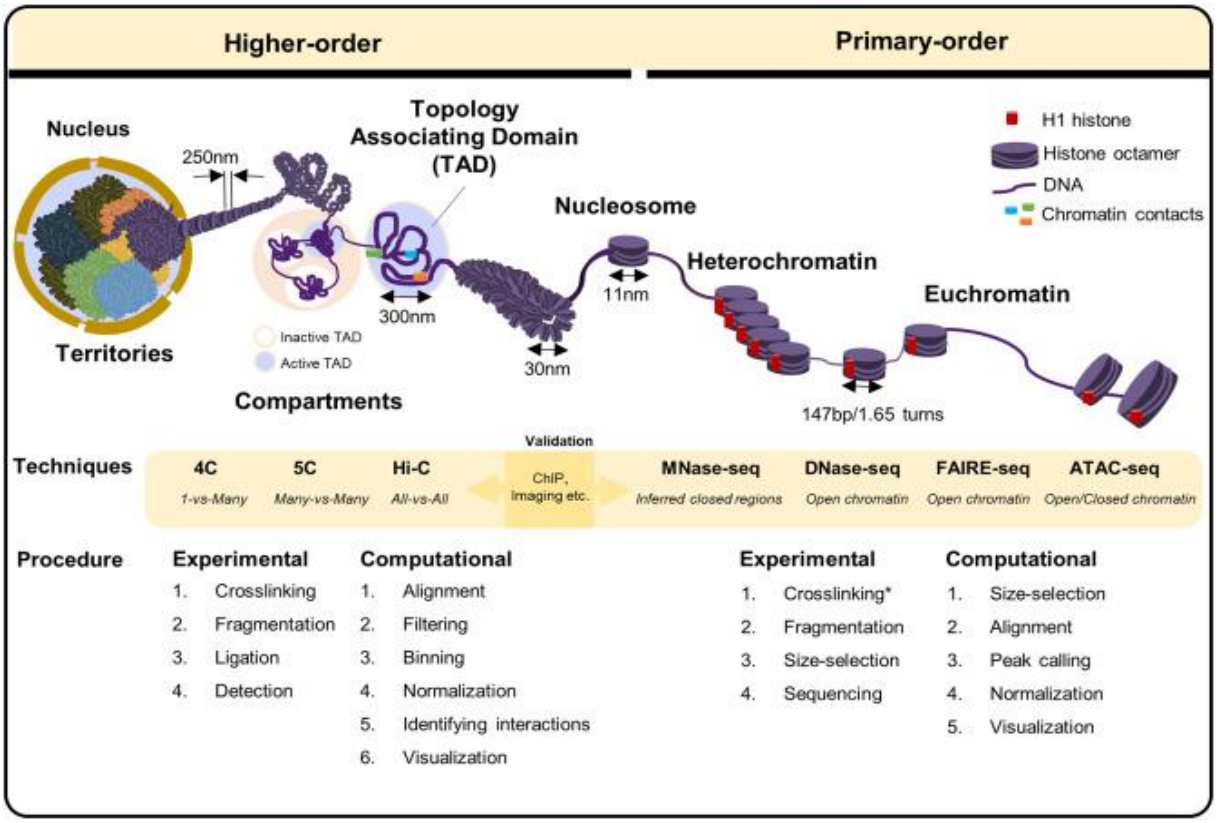
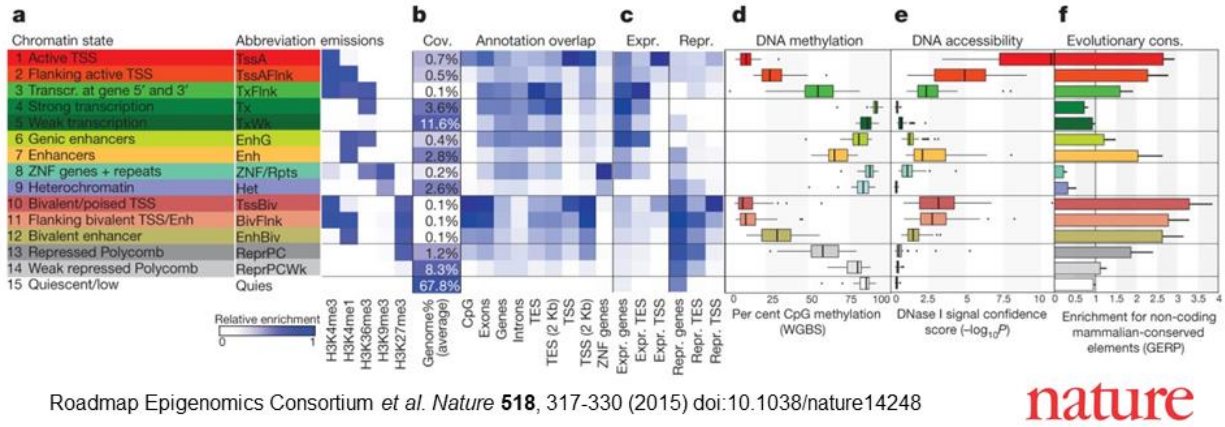


Figure 12 Genome organization in eukaryotes. From (Chang et al., 2018)

Furthermore, latest advances in single cell omics technologies allow scientists to study cell-to-cell heterogeneity, and draw rare variation within cell populations by applying single-cell ATAC-seq (scATAC-seq) and single-cell DNAse-seq techniques (scDNA-seq).
In this thesis, the focus will be on the ATAC-seq methodology as a tool to study chromatin accessibility dynamics in response to genotoxic factors, and particularly UV (Schick et al., 2015).

1.7.2 Roadmap chromatin states

The NIH Roadmap Epigenomics Mapping Consortium is a data repository of human epigenomic and transcriptomic data, as also a resource of genome-wide epigenetic information of over 100 human cell types and tissues, that assist basic-biological and disease-oriented research (Roadmap Epigenomics Consortium et al., 2015). The roadmap database includes processed data (alignment files, genome browser tracks, peak calling files, intergenic expression contigs, differentially methylated regions etc.) of multiple sequencing protocols, such as ChIP-seq of histone modifications (H3K27ac, H3K27me3, H3K4me1 etc.), chromatin accessibility (DNAseq-seq), mRNA-seq, and DNA methylation profiles, as also genome-wide chromatin state annotations using the chromHMM algorithm (Ernst & Kellis, 2017) coupled with processed ChIP-seq data of H3K4me3, H3K4me1, H3K36me3, H3K27me3, and H3K9me3 (core set - 15 chromatin states (see example Figure 13), supplemented by H3K27ac (18 chromatin state), or imputed data using H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H4K20me1, H3K79me2, H3K36me3, H3K9me3, H3K27me3, H2A.Z, and DNase (25 chromatin states) across multiple cell types. In this study, the stable 15-state annotation of primary Normal Human Dermal Fibroblasts (NHDF) cells is used.



Roadmap Epigenomics Consortium *et al. Nature* **518**, 317-330 (2015) doi:10.1038/nature14248

nature

Figure 13 Chromatin states of H1-Embryonic stem cells (H1-ES)

1.8 Illumina Sequencing

DNA (or complementary DNA in the case of RNA, cDNA) sequencing includes several methods and technologies that are used to determine the exact nucleotide sequence order in a DNA

molecule. The first attempts of DNA sequencing began in 1965 using 2-dimensional Chromatography, while nowadays next-generation sequencing (NGS) methodologies are the most widely used technologies.
In this study, all the analyzed NGS datasets are generated by the Illumina sequencing methodology, which is the most popular and widely used technology nowadays. The particular technology incorporates reversible dye-terminators that enable the identification of single nucleotides as they are washed over DNA strands. Illumina sequencing is widely used for ChIP-seq, RNA-seq, chromatin accessibility sequencing assays (ATAC-seq, DNAse-seq, FAIRE-seq, Mnase-seq), Exome sequencing, Whole genome sequencing, Methyl-seq, et al.
In Illumina sequencing, the use of adapters is a key step to a successful sequencing experiment, since they allow the fragment binding to the flow cell, enabling the PCR amplification of only the adapter-ligated DNA sequences, as also the indexing of each sample in order to perform multiplexed sequencing runs of multiple samples (see http://tucf-genomics.tufts.edu/documents/protocols/TUCF_Understanding_Illumina_TruSeq_Adapters.pdf) . Sequencing can be performed in single-end mode, where one stretch of each fragment is sequenced, or in paired-end sequencing mode, where both ends of each fragment are sequenced. That provides extra information, knowing exactly how far apart two reads are located in the genome. The diagram below (Figure 15) displays the difference.

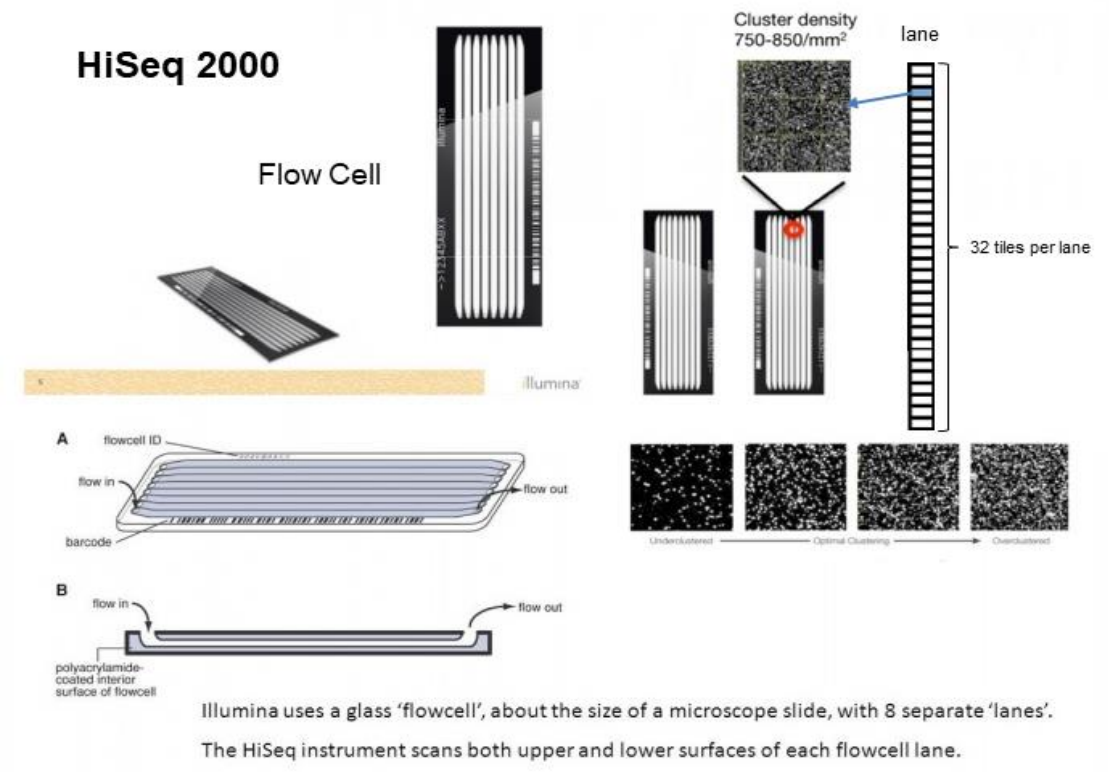


Figure 14 Illumina HiSeq 2000 sequencer parts. Adapted by https://hackteria.org/wiki/HiSeq2000_-_Next_Level_Hacking.

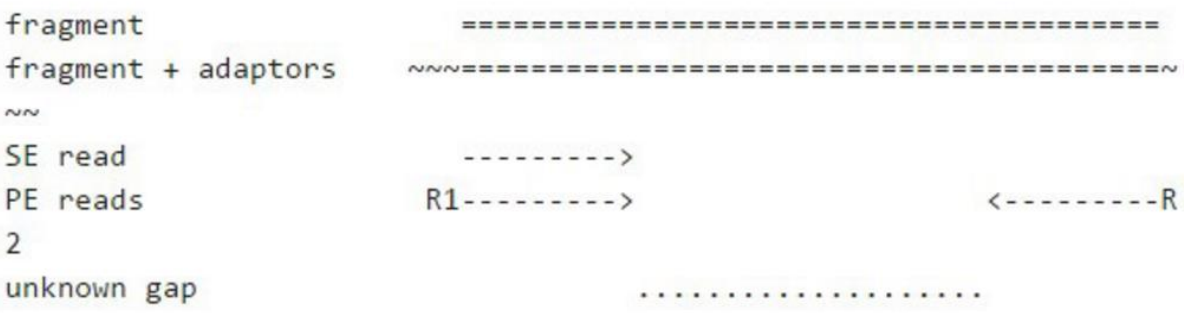


Figure 15 Difference between single and paired-end reads. From http://thegenomefactory.blogspot.com/2013/08/paired-end-read-confusion-library.html

1.9 Basic components of NGS data analysis

1.9.1 FASTA file

FASTA files are used to store nucleotide or amino acid sequence information. They may include at least one, or multiple sequences. Each FASTA record is characterized by two consecutive fields. The first field start with the ">" character and includes the name or/and id of the sequence record, as also comments and descriptions about the sequence that always should precede the sequence name. The second field includes the nucleotide or amino acid sequence string, either in one line or split into multiple lines. In Figure 16, an example of a FASTA record is illustrated, and in particular the first nucleotides of the human ribosomal DNA complete repeating unit with GenBank Id U13369.1.

```
>U13369.1
GCTGACACGCTGTCCTCTGGCGACCTGTCGTCGGAGAGGTTGGGCCTCCGGATGCGCGCGGGGCTCTGGC
CTCACGGTGACCGGCTAGCCGGCCGCGCTCCTGCCTTGAGCCGCCTGCCGCGGCCCGCGGGCCTGCTGTT
CTCTCGCGCGTCCGAGCGTCCCGACTCCCGGTGCCGGCCCGGGTCCGGGTCTCTGACCCACCCGGGGGCG
GCGGGGAAGGCGGCGAGGGCCACCGTGCCCCGTGCGCTCTCCGCTGCGGGCGCCCGGGGCGCCGCACAAC
CCCACCCGCTGGCTCCGTGCCGTGCGTGTCAGGCGTTCTCGTCTCCGCGGGGTTGTCCGCCGCCCCTTCC
CCGGAGTGGGGGGTGGCCGGAGCCGATCGGCTCGCTGGCCGGCCGGCCTCCGCTCCCGGGGGGCTCTTCG
ATCGATGTGGTGACGTCGTGCTCTCCCGGGCCGGGTCCGAGCCGCGACGGGCGAGGGGCGGACGTTCGTG
GCGAACGGGACCGTCCTTCTCGCTCCGCCCGCGCGGTCCCCTCGTCTGCTCCTCTCCCCGCCCGCCGGCC
GGCGTGTGGGAAGGCGTGGGGTGCGGACCCCGGCCCGACCTCGCCGTCCCGCCCGCCGCCTTCGCTTCGC
```

Figure 16 FASTA record of U13369.1 GenBank sequence

1.9.2 FASTQ files and quality control (QC)

During sequencing, for each sequenced nucleotide a quality score is assigned that reflects the possibility that the specific symbol is incorrectly reported. FASTQ files allow the storing of both the sequenced fragment fraction, and the corresponding quality of each nucleotide. Both strings are encoded with ASCII characters, since quality scores reach double digits. There are some discrepancies in the way that the quality scores are encoded between different platforms, but in this study only the Phred+33 system (Phred) will be considered, since nowadays this is the

default encoding method. Phred quality score was first used in the automated sequencing during the Human Genome Project (Adams, 2008). The sequence assembly program was called Phrap and the Phrap program used phred scores to help clear discrepancies in overlapping sequences. Quality scores remain of high importance, especially when short read technology is applied. Phred quality scores are defined by $Q = -10\,log_{10}P$, where Q is the actual Phred value, and P is the base-calling error probability. Indicative Phred values are depicted in the Table 3.

Phred-33 Q score	P of incorrect nt call	Nt call accuracy
10	10%	90%
20	1%	99%
30	0.1%	99.9%
40	0.01%	99.99%
50	0.001%	99.999%

Table 3 Indicative Phred quality scores

The FASTQ format is similar to FASTA, but the description of the sequence uses an "@" character instead of a ">" at the beginning. Immediately below the sequence is the description of the quality score, beginning with the + character. The next line contains the quality scores in ASCII format. In Figure 17, an example of two Illumina HiSeq 2000 FASTQ records is illustrated.

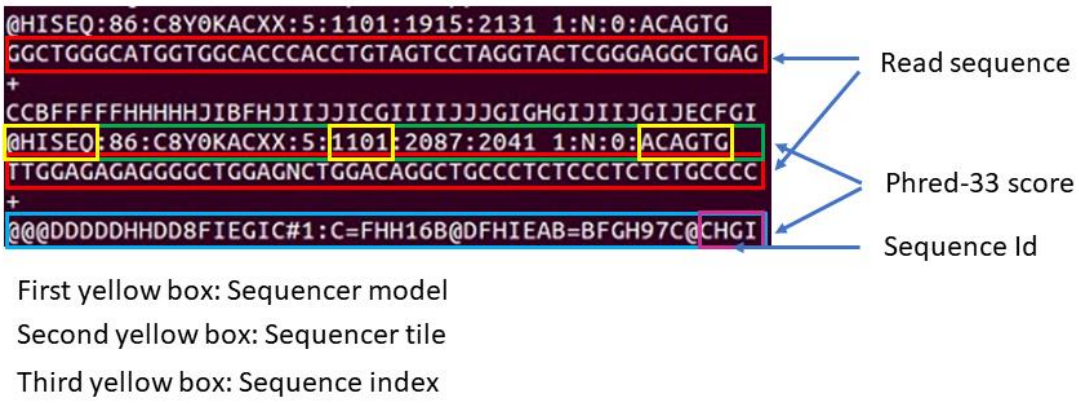


Figure 17 Example of two Illumina 2000 FASTQ records.

FASTQ sequence information and per nucleotide phred-33 quality scores gives the opportunity to generate quality control (QC) tests that are informative about the usability of the reads, or parts of the reads, or even the whole dataset. QC tests are also informative about basic statistics of the sequenced library, or about potential contamination of the libraries with unexpected endogenous or exogenous factors.
Examples of these tests include per base Phred quality check, per sequence Phred quality check, per sequence GC content check, read duplication rate check, read length distribution

check, overrepresented sequences reporting, adapter content reporting, check for external library contamination (blast search), nucleotide composition and others. These tests can be applied using the FASTQC suite (Andrews, 2015), FASTX-Toolkit (Gordon et al., 2014), RseQC (Liguo Wang et al., 2012), rnaqc (Zhou et al., 2018), or even custom scripts. Some indicative results of the above-mentioned QC tests are depicted in Figure 18.

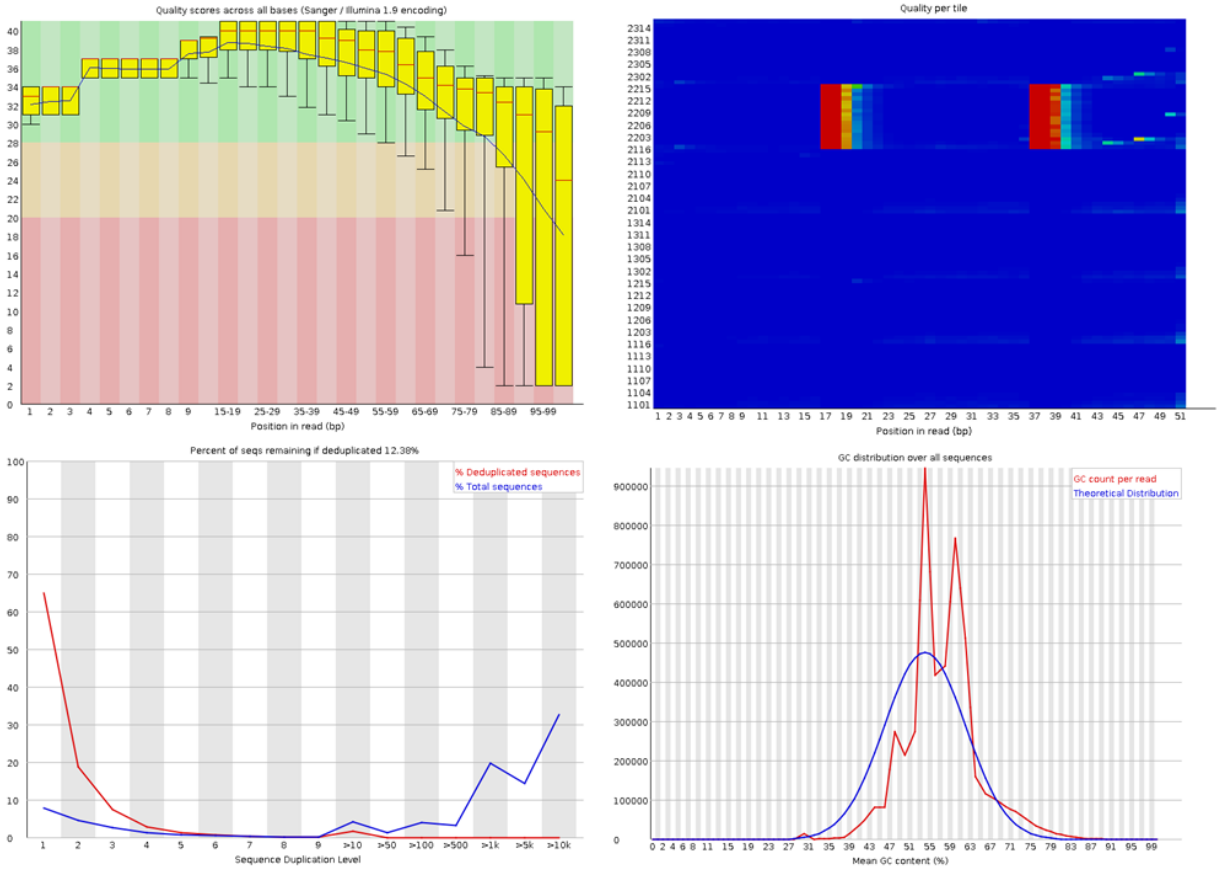


Figure 18 Indicative FASTQ QC tests using the FASTQC suite. (a) Per base sequence quality check. This quality check gives an overview of the Phred score distribution for each base pair. In many sequencing protocols, quality tends to be lower at the 3' end of the read, especially when the sequence length increases. If the mean of the Phred distributions are under 20, then if the particular nucleotide(s) is at one of the 5' or 3' ends of the reads, the trimming of that base pair should be considered for all the library reads. (b) Per tile sequence quality. The particular quality check refers only to Illumina experiments, and takes advantage of the information of the flowcell tile which each read came from (Figure panw me tiles). The particular heatmap visualization, illustrates clusters of low quality nucleotides in specific read positions and tiles, and can reveal transient problems such as bubbles or smudges through the flow cell, or debris inside the flow cell lane. If such problematic clusters occur in the inner body of the sequenced reads, it should be considered to remove all the reads coming from the problematic tiles. (c) Sequence duplication level reports. The specific quality test reveals the level of identical sequences coming from a sequenced library. If the sequence duplication is high (over 30%), then potential problems related with PCR overamplification or low complexity library material could be suspected. Especially, if the sequencing protocol is ChIP-seq or ATAC-seq, then after read mapping elimination of duplicated alignments should be considered. On the contrary, if the examined sequences

come from RNA-Seq libraries, some sequences belonging to highly expressed transcripts may be over-sequenced. (d) GC content distribution test. The particular QC module, tests the GC % along the examined sequenced reads, and compares it with the GC % distribution that comes from a random library. Deviations between the observed and the expected distribution may be a result of a specific library contamination, such as adapter sequence dimers.

Following QC, FASTQ files are often processed to eliminate low-quality nucleotides or/and sequences using appropriate software, such as cutadapt (Martin, 2011), trimmomatic (Bolger et al., 2014), seqtk (H. Li, 2012) et al.
Another source of FASTQ contamination are the adapter sequence that remains at the ends of the sequenced reads (see Figure 19).

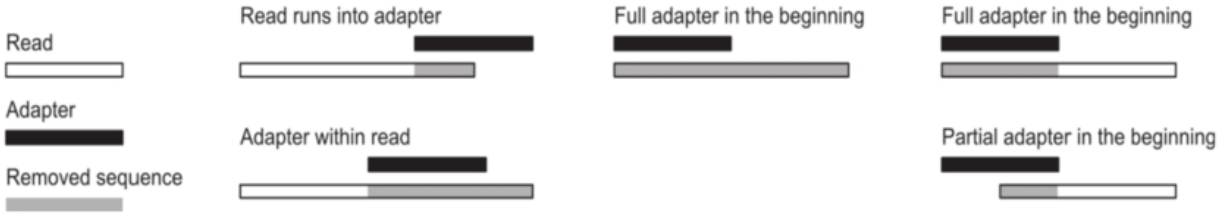


Figure 19 Adapter removal using cutadapt. Adapted by (Martin, 2011).

These contaminants can be removed by providing the respective nucleotide sequences to the afore-mentioned tools.
QC of short reads is also applied after the alignment of the sequenced reads against the reference genome/transcriptome of origin. This procedure includes visual inspection of the NGS signal using a genome browser as UCSC or IGV (Integrative Genome Browser), genome coverage calculation of the aligned library, estimation of the total number of actively transcribed genes, summarization of the total number of peaks in ChIP-seq experiments and chromatin accessibility assays (ATAC-seq, DNAse-seq), estimation of the noise-to-signal ratio et al (see next sections and results section).

1.9.3 Genome assembly

One of the major challenges in NGS analysis, and particularly in DNA sequencing, is to assemble the sequenced reads to their original order, to form the unified chromosomal sequences of the reference genome. The basic steps of this procedure include finding the overlapping regions between sequenced reads and form “contigs”, scaffolds, and finally chromosomes. Many genomes have only been assembled to the scaffold level (Hubbard et al., 2002).

Chromosome sequences are not identical between individuals of the same species. For example, each human has about 3-4 million single nucleotide polymorphisms (SNPs) with respect to the human “reference genome” (UCSC or Ensembl builds) which is an accepted,

standardized sequence. Since 2011, over 575,000 exonic sequences were annotated (Kent et al., 2002), while today there are over 20,000 annotated protein coding genes. Nowadays, hg19/GRCh37 and hg38/GRCh38 genome builds are the most commonly used for the analysis of NGS data.

De-novo assembly is applied when an organism is sequenced for the first time, and reads should be assembled from scratch. On the contrary, for individual genomes of a well-studied organism (like human), the respective reference genome serves as a template that guides the assembly procedure. If no reference is available, sometimes a closely related genome can be very helpful. Regarding the technology used for whole-genome sequencing, which is the appropriate methodology for genome assembly applications, different platforms produce different error rates and different read lengths. Long reads are very useful during the assembly process, but they are more prone to sequencing errors.

1.9.4 Short-read mapping

One of the most essential steps in NGS data analysis is the short-read mapping. After obtaining high-quality FASTQ files, reads must be assigned to their positional origin along the examined reference genome/transcriptome, to reconstruct biologically meaningful measurements, such as the level of mRNA produced by a gene (RNA-seq), the genomic locations of protein binding (ChIP-seq), the extent of chromatin accessibility (ATAC-seq, DNAseq-seq) et al. in order to gain valuable information about the biological outcome of the NGS experiment. For this reason, specialized and well-engineered software called short-read mappers or aligners is used. Because of sequencing errors and differences between the reference genome and the sequenced subject, the alignment process should allow nucleotide alterations such as mismatches, deletions and insertions.

Bowtie and BWA have been two of the most popular short-read aligners since 2009. Bowtie extends previous Burrows-Wheeler algorithm applications, by utilizing the Burrows-Wheeler indexing approach, with a quality-aware backtracking algorithm that permits mismatches. (Langmead et al., 2009) (Figure 20). BWA also utilizes Burrows-Wheeler transform algorithm, and for exact matches is very similar to Bowtie. For inexact matches, a backtracking approach is developed to seek for matches between genome segments and the read within constant distance (H. Li & Durbin, 2009) (Figure 21).

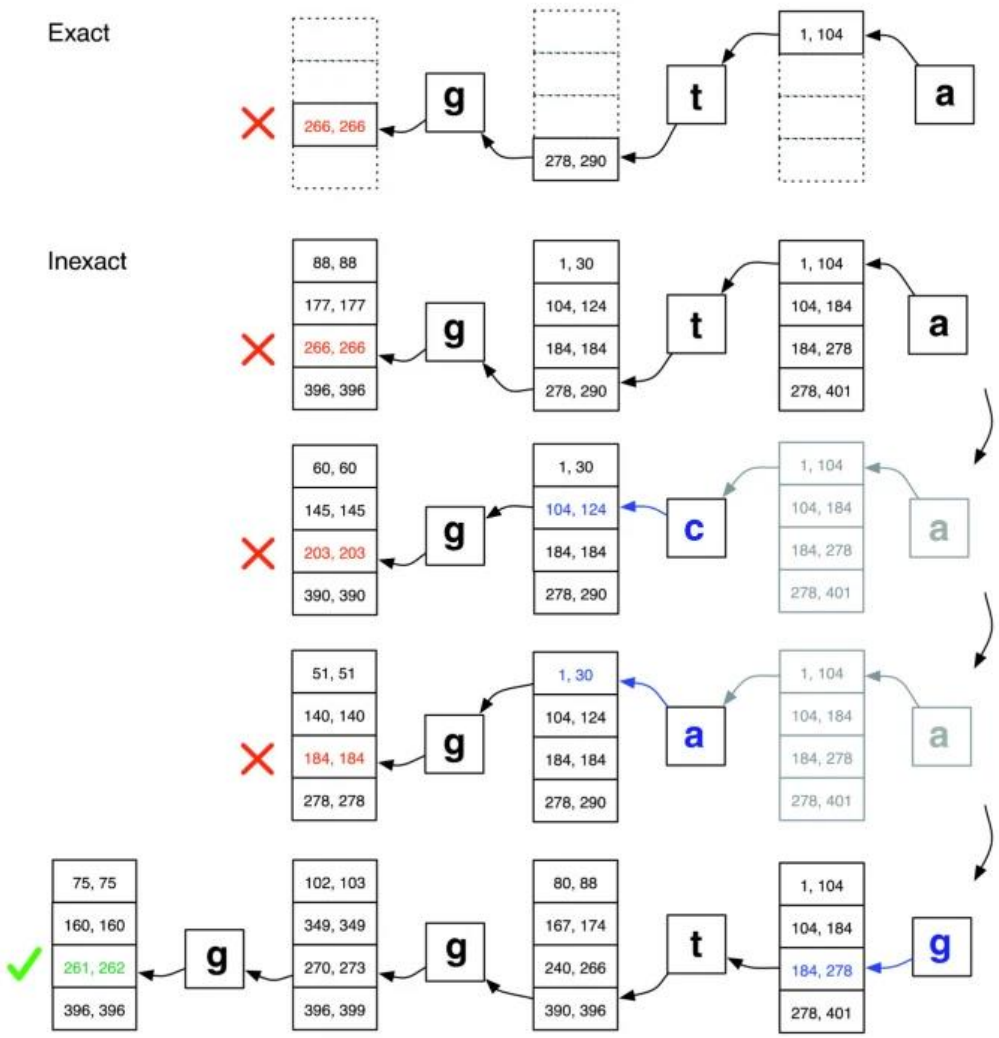


Figure 20 Exact and inexact alignment. Different approaches when there is no exact match for 'ggta' sequence (mismatch when 'a' is replaced by 'g'), exact alignment (top) and Bowtie (bottom) processes. Number pairs in boxes represent row matrix suffixes, X marks denotes an empty range and aborts aborts in the exact match algorithm, or backtracks in the inexact algorithm, and green ticks represents the finding of a non-empty range with one or more occurrences of a mapping for the read.

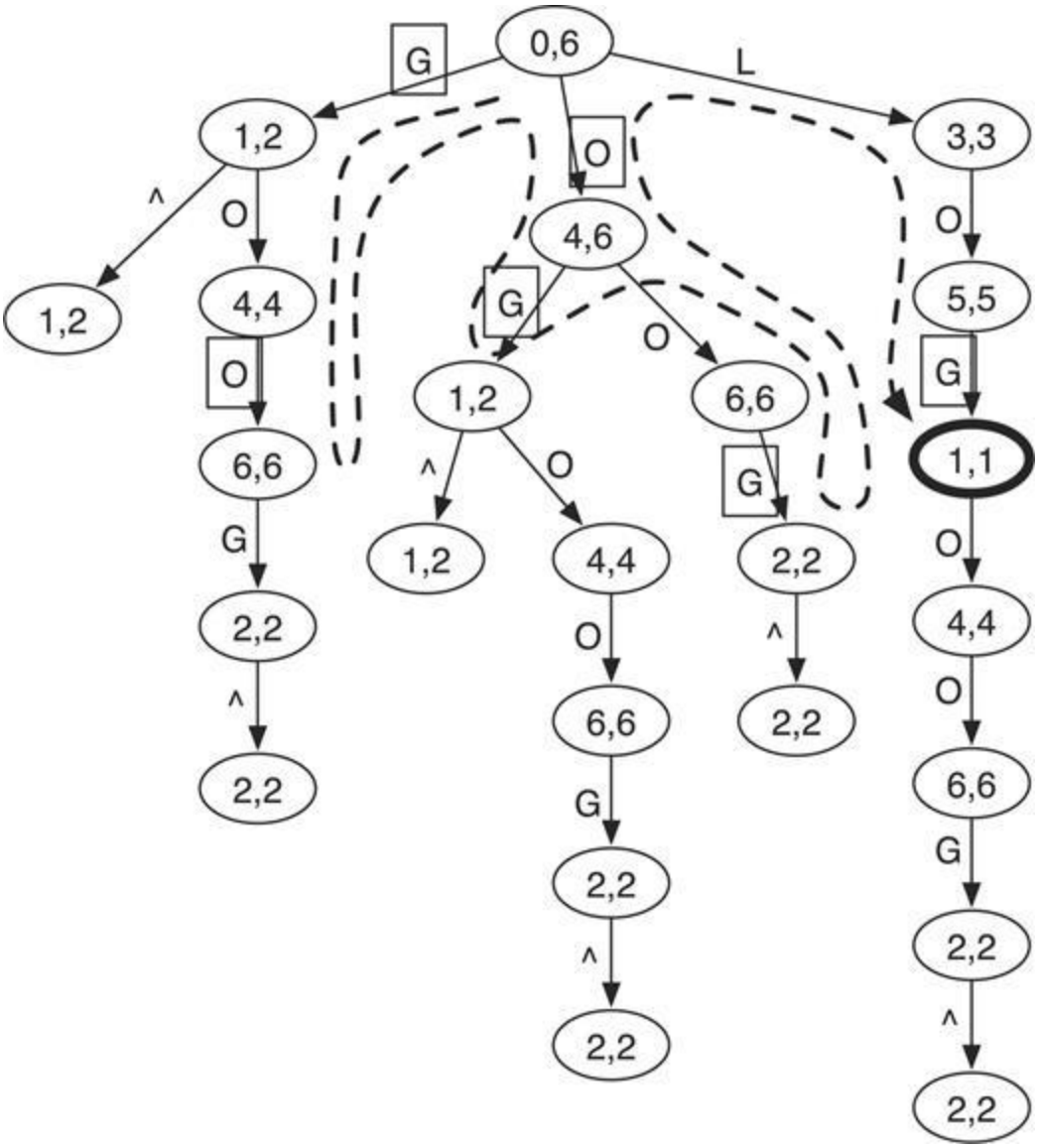


Figure 21 Digital tree of 'GOOGOL' string. "Λ" denotes the start of the sequence, while the number pairs give the SA interval of the string represented by the node (H. Li & Durbin, 2009). The dashed arrow illustrates the brute-force search route for the sequence 'LOL', allowing at most one mismatch, and labels in squares denote the mismatches. Finally, the only valid hit is the bold node [1, 1] representing the sequence 'GOL'.

1.9.5 SAM - BAM files

Short-read mappers report alignments of sequenced reads against the genome reference in several formats, but in the last years the golden standard has become the Sequence Alignment Map (SAM) format. This kind of files are usually processed by samtools (H. Li et al., 2009), a specialized toolkit that has been developed for this purpose. SAM files include a header section that informs the analyst if the alignments are sorted, reports the version of the software that generated the file, provides a list of chromosomes that were included in the genome reference,

as also their respective length, and lists all the operations that have been already applied on the file (alignment commands, samtools commands etc). An example of a SAM file header is illustrated in Figure 22.

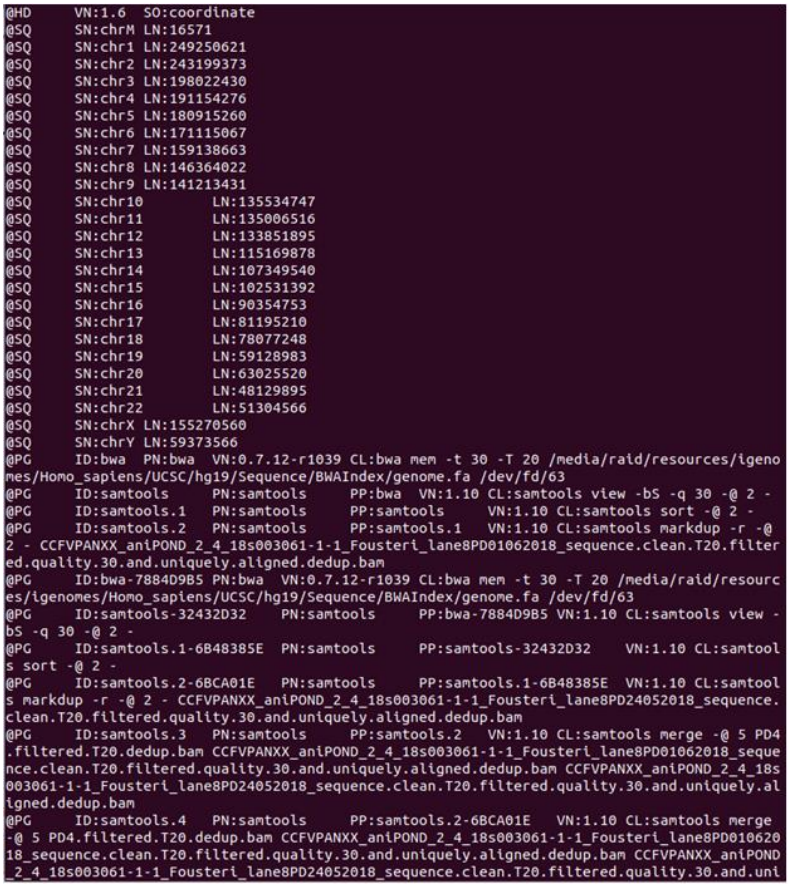


Figure 22 An indicative example of SAM file header

The rest of the SAM file includes the alignments, in a tab delimited format as illustrated in Figure 23.

Figure 23 An indicative example of two alignments stored in a SAM file

Sam records column-wise information includes (in the order reported as follows): (1) the read name/Id, (2) the SAM flag that indicates if the read is paired, mapped in a proper pair (paired-end), is unmapped, has its mate unmapped (paired-end), is mapped in the reverse strand, its mate is mapped in the reverse strand (paired-end), is first in a paired alignment (paired-end), is second in a paired alignment (paired-end), is not a primary alignment, is a low quality alignment, is a duplicated alignment, is a supplementary alignment, and also combinations of flags (table 4), (3) the chromosome name, (4) the chromosome position where the alignment between the

read and the chromosome starts, (5) the MAPQ mapping quality, (6) the CIGAR string that is informative about insertions or deletions (31M1D20M = 31 consecutive matches followed by one deletion and 20 consecutive matches), (7) the name of the read's mate (paired-end), (8) same as (4), but for the read's mate (paired-end), (9) the length of the template, (10) the read sequence, (11) the read quality (Phred score), and optional fields that include "tags" (may be aligner specific) and are informative about the number of mismatches (NM:i:2 = two mismatches), the position of the mismatches (MD:Z:4C3A42 = 4 consecutive bases with exact match with the reference, a "C" that does not match with the reference, 3 consecutive bases with exact match with the reference, an "A" that does not match with the reference, 42 consecutive bases with exact match with the reference), the aligner score (AS:i:42 = quality score equals to 42), and others.

MAPQ scores are similar to Phred scores (see section 1.9.2) and are informative about the quality of the alignment: $MAPQ = -10\ log_{10}P(mapping\ position\ is\ wrong)$. There are a lot of discrepancies regarding the definition of MAPQ values between different aligners, as each software generates different ranges of these values and consequently it's difficult to create a universal thresholding.

BAM file format is the binary form of SAM files.

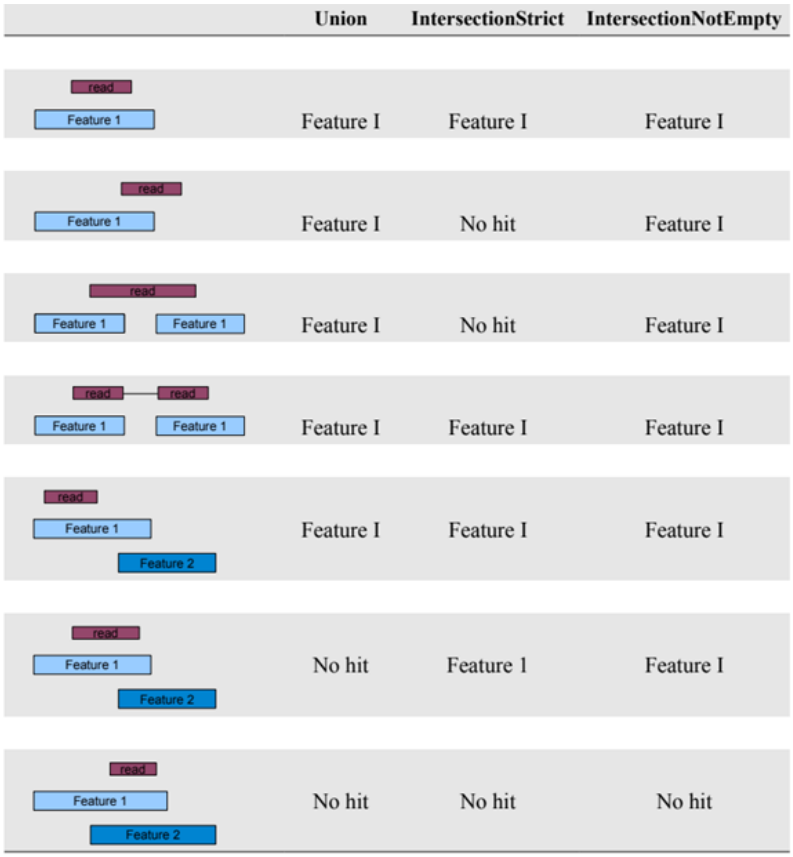
The SAM file structure enables the possibility to apply filters in order to discard low quality reads, uncertain alignments, potential PCR duplicates, but also to select alignments coming from specific genomic regions, such as particular chromosomes, intergenic regions, genes and enhancers, or even randomly sample mapped or/and unmapped reads. These operations are easily applied using samtools or custom scripts, and combinations of SAM flags ids (table 4).

Table 4 SAM flags with their decimal (fist column) and hexadecimal (second column) interpretation, and their description (third column). Valid combinations of flags are very common and provide valuable information about the alignment. For example flag = 1040 = 16 + 1024 means that the alignment is mapped in the reverse complement (16) and it is a potential PCR duplicate (1024). From (H. Li & Durbin, 2009).

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

1.9.6 Alignment counting

High quality alignments are further processed to create read-counts at genomic regions of interest, such as exons, promoters, genes, enhancers et al. Read-counts represent the total number of alignments coming from a specific genomic locus, and is reported as an integer. When overlapping genomic elements (for example overlapping genes) are present in an examined annotation, and reads are mapped in the overlapping region, uncertainties regarding the origin of the alignments occur. To avoid these ambiguities, read-counting can be performed using specialized tools that follow some specific rules (Figure 24) that aid the counting process (Anders et al., 2015; Liao et al., 2014).



* Picture reproduced from HTSeq web site :
http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

Figure 24 Different counting modes that provide different approaches to summarize alignments that overlap to multiple genomic features. From (Obenchain, 2013).

Other tools like bedtools (Quinlan & Hall, 2010), are simply reporting the total intersections between the alignments and the examined references, and are more useful for visualization purposes, such as average profiles of alignment density, heatmaps of read density, and others (see results section). Read-counting is also used to generate genome-wide profiles of NGS signal distribution, that in turn can be visualized by particular tools called “genome browsers”

(Kent et al., 2002; J. T. Robinson et al., 2011) (see examples in results section). The counting procedure is applied along the reference genome, by binning the chromosomes using a predefined segment size (for example 250 bp), counting reads in each bin, normalizing the bin-counts using a constant factor (optional), and generating bigWig files (see section 1.9.8) compatible with a genome browser interface. The counting procedure can be replaced by a similar process called “genome coverage” calculation, that summarizes the alignments in a per-base resolution (Figure 25).

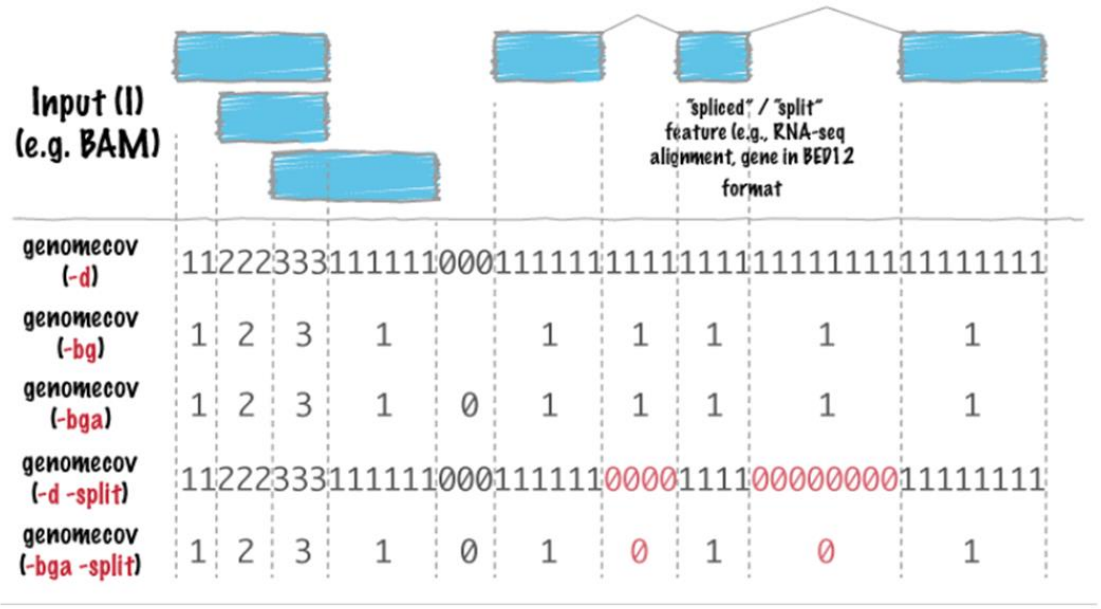


Figure 25 Bedtools genomecov [https://academic.oup.com/bioinformatics/article/26/6/841/244688] command for the creation of genome coverage profiles of an alignment file. Different options may generate different file types (BED, BEDGRAPH) as described in section 1.9.8.

1.9.7 RPKM, TPM and CPM normalization

Feature counts are not comparable between genomic elements of varying length, or between datasets of different alignment depth (different number of total alignments). This is because samples that are sequenced to a higher depth will naturally result in features that gain higher levels of counts, while longer features will also gain more mapped reads in their locus than smaller features. To tackle these biases, several normalization strategies are widely applied in the NGS analysis field. Reads per kilobase per million mapped reads (RPKM, applicable for single-end reads) or fragments per kilobase per million mapped reads (FPKM, applicable for paired-end reads) (Mortazavi et al., 2008) was one of the first approaches that addressed these issues, while transcripts per million (TPM) measurement came to make an improvement to the particular methodology (Wagner et al., 2012). The corresponding formulas that calculate RPKM/FPKM and TPM are described in Figure 27.

$$\text{RPKM/FPKM} = \frac{10^6 \times 10^3 \times C}{N \times L},$$

C = # of mappable reads/fragments of feature
N = # of mappable/sequenced reads of sample
L = gene length (bp)

Figure 26 RPKM/FPKM calculation formula

$$\text{TPM} = \frac{1}{\sum(A)} \times 10^6,$$

$$\text{A} = \frac{total\ reads\ in\ feature \times 10^3}{gene\ length\ (bp)}$$

Figure 27 RPKM/FPKM calculation formula.

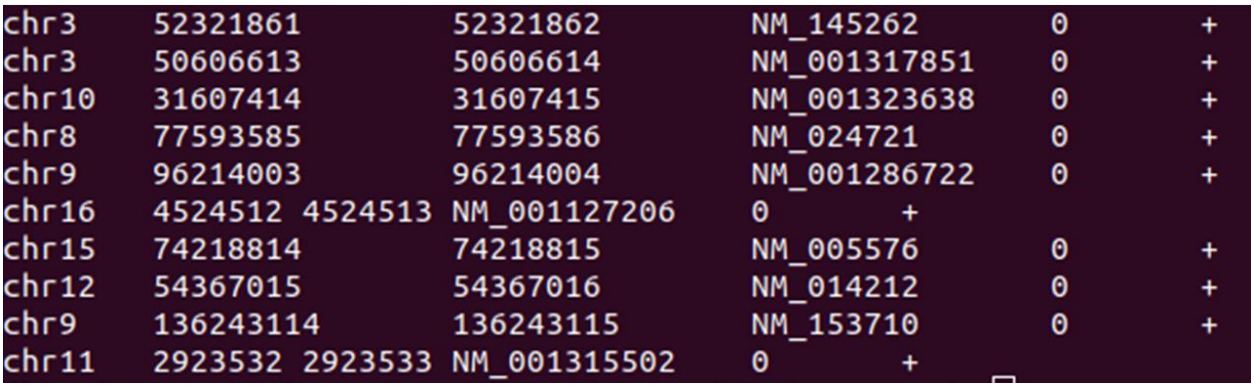
Both methods take into consideration the feature length and the library depth during the normalization process, but TPM is considered more consistent because after the calculation each examined sample is represented by the similar number of total normalized reads, while in RPKM not (Wagner et al., 2012).

While the particular normalization methods help to remove the gene length and sequencing depth bias, they should be used with caution when applying comparisons between datasets, since they don't take into account that different concentrations of NGS signal (RNA, total amount of binding of a specific protein, total nucleosome-free regions) might be very different between different biological conditions, tissues and treatment (library composition effect).
For that reason, some additional normalization methods have been implemented that take into consideration the so-called "library composition" effect. The two most widely used normalization methods that normalize for both library depth and library composition are the Median Ratio Normalization (MRN) method used by DESeq2 software (Love et al., 2014), and the trimmed mean of M-values (TMM) method used by edgeR software (M. D. Robinson et al., 2009), that are used for performing differential enrichment analysis of NGS datasets across a set of genomic features. The calculus behind these two methods will not be described in this section, since it's beyond the scope of this study. There are detailed descriptions of both approaches in the respective software publications (Love et al., 2014; M. D. Robinson et al., 2009).

1.9.8 BED, bedGraph and bigWig files

The Browser Extensible Data (BED) files are tab delimited files that are used to store alignments, or any type of genomic features that can be described by genomic coordinates. Each record is stored in one line, and contains 3-12 columns and one optional track definition line. The first three columns are mandatory since they describe the positional coordinates: (1) the chromosome name, (2) the starting position of the record in the respective chromosome, and (3) the ending position of the record in the respective chromosome. The starting and ending positions refer to the Watchon-Crick direction (plus strand) of the reference genome. The 9 additional lines include: (4) The record name, (5) a score value with a range between 0 and

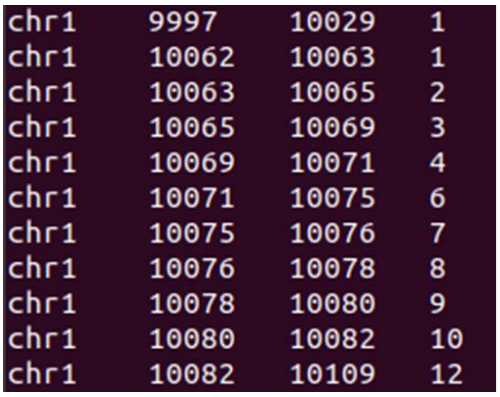
1000, (6) the feature strand orientation, (7) the thickstart, that may refer to the starting position of the starting codon of a gene, (8) the thickend, that may refer to the ending position of the stop codon of a gene, (9) the itemRgb is an RGB value that colors the record when it's displayed in a genome browser, (10) the blockCount which refers to the number of blocks (for example the total exons of a gene), (11) the blockSize, a comma separated list of block sizes and (12) the blockStart, a comma separated list of block starting positions. An example of 10 BED records is illustrated in Figure 28.



```
chr3    52321861        52321862        NM_145262       0       +
chr3    50606613        50606614        NM_001317851    0       +
chr10   31607414        31607415        NM_001323638    0       +
chr8    77593585        77593586        NM_024721       0       +
chr9    96214003        96214004        NM_001286722    0       +
chr16   4524512 4524513 NM_001127206    0       +
chr15   74218814        74218815        NM_005576       0       +
chr12   54367015        54367016        NM_014212       0       +
chr9    136243114       136243115       NM_153710       0       +
chr11   2923532 2923533 NM_001315502    0       +
```

Figure 28 BED records of protein coding and long non coding RNAs

The bedGraph format allows the interpretation of continuous values in a BED-like formal that is very useful for storing scores, such as normalized counts of NGS signal, or large blocks of genomic space with the same measurement. An example of such records is illustrated in Figure 29.



```
chr1    9997    10029   1
chr1    10062   10063   1
chr1    10063   10065   2
chr1    10065   10069   3
chr1    10069   10071   4
chr1    10071   10075   6
chr1    10075   10076   7
chr1    10076   10078   8
chr1    10078   10080   9
chr1    10080   10082   10
chr1    10082   10109   12
```

Figure 29 bedGraph records of BED alignments. Each record corresponds to a genomic block with the same number of alignments overlapping in each of the consecutive base pairs included in the particular block.

BigWig files are created using bedGraph files and are stored in an indexed binary format (Kent et al., 2010). They are very useful since they are of much smaller size than bedGraph files and they can be displayed in genome browsers as signal graphs (see results section).

1.9.9 RefSeq, UCSC and Ensembl human gene sets

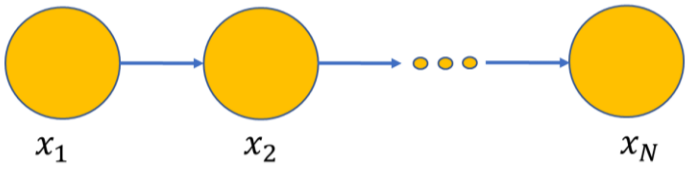
RefSeq genes is a comprehensive and non-redundant genome annotation supported by the National Center for Biotechnology Information (NCBI), and includes a set of curated and predicted gene models, transcripts, exons and UTRs. Annotation predictions use the accession prefixes XM_, XR_, and XP_, while the curated annotations (Genbank) start from NM_, NR_, and NP_ (O'Leary et al., 2016).
Ensembl (Flicek et al., 2011) uses both predicted and curated annotations for human, and the curation process is performed by the HAVANA project (Harrow et al., 2012). Automatically-annotated gene models include pseudogenes, non-coding RNAs, and alternative splicing events, while transcript annotation is based on experimental data coming from several data repositories like UniProt (Bateman et al., 2017) and RefSeq .
UCSC has been one of the many collaborators in Human Genome Project, and straight after the human genome assembly process was complete, the genome sequence was released in their genome browser site. UCSC gene annotation is constructed automatically, based on UniProt and Genbank (Hsu et al., 2006).
All human genome annotations can be downloaded from the respective database repositories, as also from specialized tools such as bioMart (Kinsella et al., 2011) and UCSC Table Browser (Karolchik et al., 2004).
At the date of this study, for the GRCh38/hg38 human genome build, RefSeq has annotated 26,671 gene models and 226,309 exons, Ensembl database includes 60,587 gene models and 510,285 exons, and UCSC includes 27,982 gene models and 236,062 exons.

1.10 Hidden Markov Models (HMMs)

Data sets in which data points are potentially interdependent, comprise a special data type that is known as "sequential data". The specific type of data points is displayed in a specific order, and its classification is important ($x_1, x_2, \ldots x_{N-1}, x_N,$). Some basic examples of sequential data are "time series" (weather data, stock data, audio data, etc.), and "spatially dependent data", such as DNA sequences and characters that form sentences in natural languages. This data type can be modeled by the following formula:

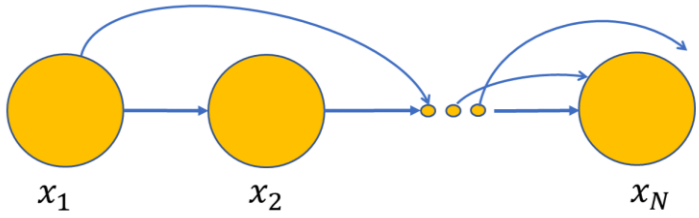
$$p(X) = p(x_1) \prod_{n=2}^{N} p(x_n | x_{n-1}, x_{n-2}, \ldots x_1)$$

In the particular notion, each conditional distribution depends on all the previous observations. In the special case where the above rules is relaxed and each conditional distribution depends only by the previous observation, the resulting model in known as "first-order Markov model":



$$p(X) = p(x_1) \prod_{n=2}^{N} p(x_n|x_{n-1})$$

Higher-order Markov models can also be defined, by letting more than one dependent observation to affect the model. For example, a second-order Markov model can be defined as follows:



$$p(X) = p(x_1)p(x_2|x_1) \prod_{n=3}^{N} p(x_n|x_{n-1}, x_{n-2})$$

Additionally, if the conditional distributions $p(x_n|x_{n-1})$ depend on adjustable parameters, such as those that might be inferred by training datasets, and all the distributions share the same parameter values, then the Markov model is considered “stationary” and “homogenous”.
In the special occasion where x_n latent values are discrete, then the model is called “Markov chain”. In the particular study, first-order stationary Markov models will be used.

The initial distribution of a first-order stationary Markov model, is a special latent variable since there is no parent observation:

$$\pi_k = p(x_{1k} = 1), \sum_{k=1}^{K} \pi_k = 1$$

The transition distribution $p(x_n|x_{n-1})$ is defined by a K x K matrix that is called transition matrix:

$$A_{jk} = p\left(x_{nk} = 1|x_{(n-1)j} = 1\right), \sum_{k=1}^{K} A_{jk} = 1\,, j = 1, \dots, K$$

where (π, A) are the parameters of the models.
To estimate (π, A), maximum likelihood estimation (MLE) is applied, for a given sequence of observations $\{x_1, x_2, \dots, x_N\}$:

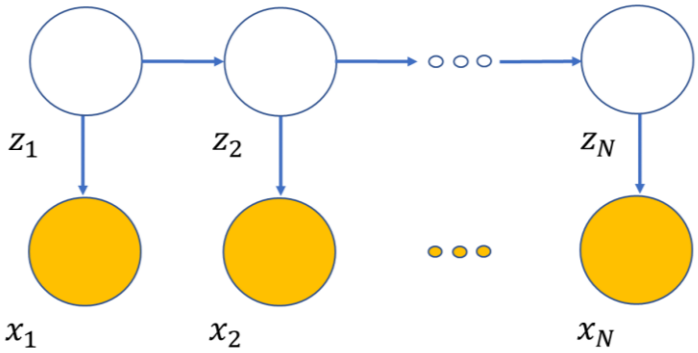
$$p(X|\pi, \mathrm{A}) = p(x_1)\prod_{n=2}^{N} p(x_n|x_{n-1})$$

$$= \left(\prod_{k=1}^{K} \pi_k^{x_{1k}}\right)\prod_{n=2}^{N}\left(\prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{X_{(n-1)j}X_{nk}}\right)$$

$$A_{jk} = \frac{\sum_{n=2}^{N} X_{(n-1)j}X_{nk}}{\sum_{l=1}^{K}\sum_{n=2}^{N} X_{(n-1)j}X_{nl}}$$

To train the transition matrix A of a first-order Markov model, K^2 parameters should be defined, while for a L-order Markov model, K^{L+1}parameters should be defined, making the parameter estimation procedure unfeasible for big values of L. This limitation is bypassed by Hidden Markov Models (HMMs) (Rabiner & Juang, 1986).

HMM is an extension of mixture models (schema below), where “hidden” latent variables $z_n, n = 1..N$ define the outcome of the observation, and hidden latent variables are described by a Markov chain:



$$P(X, Z) = \left(p(z_1)\prod_{n=2}^{N} p(z_n|z_{n-1})\right)\left(\prod_{n=1}^{N} p(x_n|z_n)\right)$$

$$= p(z_1)\; p(x_1|z_1)p(z_2|z_1)\; p(x_2|z_2)\ldots$$

where hidden variables are described by the Markov chain:

$$P(Z) = p(z_1) \prod_{n=2}^{N} p(z_n|z_{n-1})$$

with initial probability:

$$p(z_{1k} = 1) = \pi_k$$

with transition matrix:

$$p\left(z_{nk} = 1|z_{(n-1)j} = 1\right) = A_{jk}$$

and emission probability (observation probability):

$$p(x_n|z_n)$$

that may follow one of the discrete or continuous distributions. HMMs can be represented graphically as illustrated in Figure 30.

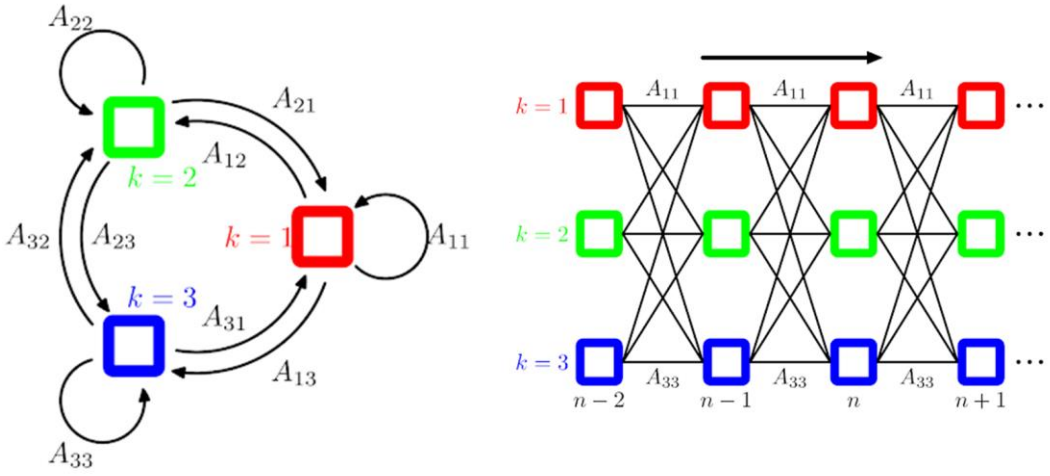


Figure 30 Lattice (left) and trellis (right) representations of the same HMM transition matrix. HMMs have a "sticky" before meaning that they stay in the same hidden state for multiple observation periods. From ("Pattern Recognition and Machine Learning," 2007).

HMMs are widely used in speech recognition (Rabiner & Juang, 1986; Sun & Jelinek, 1999), natural language modelling (Manning et al., 2002), handwriting recognition (Nag et al., 1986),

for the gene and protein sequence predictions (Best, 2004; Cunningham, 1999; Krogh et al., 1994) and other applications.

Parameter estimation (θ) of the HMM can be accomplished by using Baum-Welch algorithm (Munro et al., 2011), a special case of the Expectation Maximization (EM) algorithm, which uses a set of observations as input:

$$b_j(y_i) = P(X_n = x_i | Z_n = j)$$

to obtain a N (observations) x K (hidden state) matrix:

$$B = \{b_j(y_i)\}$$

a transition probability matrix A with predefined transition probabilities, and a predefined initial state probability π:

$$\theta = (\pi, A, \mathrm{B})$$

Baum-Welch is defined by two processes called “forward” and “backward” algorithms, and their recursions can be run in parallel. The two processes are described below.

Forward process:

The probability of observing $x_1 \ldots x_n$ in state i at time n is set as:

$$a_i(n) = P(X_i = x_i, X_n = x_n, Z_n = i | \theta)$$

and can be calculated using the recursion:

$$(1)\ a_i(1) = \pi_i b_i(x_1)$$

$$(2)\ a_i(n+1) = b_i(y_{n+1}) \sum_{j=1}^{N} a_j(n) a_{ji}$$

Backward process:

The probability of observing the partial sequence $x_{n+1} \ldots x_N$ using that starting state i at time n is set as:

$$b_i(n) = P(X_{n+1} = x_{n+1}, \ldots, X_{\mathrm{N}} = x_N | Z_n = i, \theta)$$

and can be calculated using the recursion:

$$(1)\ b_i(\mathrm{N}) = 1$$

$$(2)\ b_i(n) = \sum_{j=1}^{N} b_j(n+1) a_{ij} b_j(x_{n+1})$$

The temporary variables can be calculated using the Bayes' theorem. The probability of being at state i at time n, given the observations X with parameters θ, is calculated as follows:

$$\gamma_i(n) = P(\mathrm{Z}_n = i|Y, \theta) = \frac{\mathrm{P}(Y_{n=i}, X|\theta)}{P(X|\theta)}$$

$$= \frac{a_i(n) b_i(n)}{\sum_{j=1}^{N} a_j(n) b_j(n)}$$

while the probability of being in state i at time n, and j at time n+1, given the observations X and parameters θ, is calculated as follows:

$$\xi_{ij}(n) = P(\mathrm{Z}_n = i, Z_{n+1} = j|\mathrm{X}, \theta)$$

$$= \frac{\mathrm{P}(Z_n = i, Z_{n+1} = j, X|\theta)}{\mathrm{P}(\mathrm{X}|\theta)} \quad = \frac{a_i(n) a_{ij} b_j(n+1) b_j(x_{n+1})}{\sum_{k=1}^{N} \sum_{w+1}^{N} a_k(n) a_{kw} b_w(n+1) b_w(x_{n+1})}$$

After the above calculations, the HMM parameters are updated as follows:

(1) The initial state probability:

$$\pi_i^* = \gamma_i(1)$$

(2) The expected number of transitions from state i to state j over the expected total number of transitions from state i:

$$\alpha_{ij}^* = \frac{\sum_{n=1}^{N-1} \xi_{ij}(n)}{\sum_{n=1}^{N-1} \gamma_i(n)}$$

(3) How many times is expected that the observations will be equal to υ_k in state i, compared to how many times is expected that the state i will be visited:

$$b_i^*(v_k) = \frac{\sum_{n=1}^{N} 1_{x_n=v_k}\gamma_i(n)}{\sum_{n=1}^{N}\gamma_i(n)}$$

$$1_{x_n=v_k} = \begin{cases} 1\ if\ x_t = v_k \\ 0\ otherwise \end{cases}$$

All the above calculations are repeated until a desired number of iterations, or until convergence.

After parameter estimation, the most probable hidden state sequence $Z^* = (z_1^*, \ldots, z_N^*)$ is predicted, using the Viterbi algorithm (Viterbi, 1967):

$$Z^* = \underset{z}{\operatorname{argmax}}\, p(Z|X) = \underset{z}{\operatorname{argmax}}\, p(X,Z) = \underset{z}{\operatorname{argmax}}\, logp(X,Z)$$

Viterbi makes use of messages of the form:

$$\omega(z_n) = \log p(x_n|z_n) + \min_{z_{n-1}}[\omega(z_{n-1}) + logp(z_n|z_{n-1})]$$

$$\delta(z_n) = z_{n-1}^*$$

where δ(z_n) stores the z_{n-1}value, and are initialized as follows:

$$\omega(z_1) = \log p(x_1|z_1) + \log p(z_1)$$

After all ω messages are calculated, the next step is executed:

$$z_N^* = \underset{Z_N}{\operatorname{argmax}}[\omega(Z_N)]$$

and starting from z_N^* backtracking is performed based at:

$$z_{n-1}^* \leftarrow \delta(z_n^*)$$

that finds the most probable path.

All the theory, formulas and algorithm descriptions in this section are referring to the lecture presentations of the Machine Learning Course of the Department of Informatics and Telecommunications of the University of Athens, by Michael Titsias, the book: (Bishop, 2006), and Wikipedia https://en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm.

2 Materials and methods

2.1 Human cell lines

In the present study the experiments were conducted using normal human skin fibroblasts (VH10) (Kolman & Bohušová, 1992) as well as skin fibroblasts of CS-B patients (CS1AN) (Arlett et al., 2008)(Nardo et al., 2009). Both cell lines were immortalized by the human telomerase reverse transcriptase (htert) method (Lee et al., 2004). Cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, Thermo Scientific) enriched with 10% v / v fetal bovine serum (Fetal Bovine Serum, FBS, Thermo Scientific) and 1% v / v penicillin-streptomycin (Thermo Scientific), preserved in an incubator at 37 ° C and 5% carbon dioxide (CO2), and cultured in a laminar flow hood.

2.2 Cell population synchronization

Previously established protocols to synchronize cells in G1 were applied to limit cell-cycle heterogeneity and achieve steady-state levels of RNAPII, histone modifications, chromatin accessibility and nascent transcripts across the transcribed region (promoters, enhancers, asPROMPTs and gene bodies) (Lavigne et al., 2017; Liakos et al., 2020). Briefly, serum-starvation for 72 h of cells at confluency enriched for cells in G0/G1. After release in complete medium for 3 h, rapid recovery of steady-state levels of transcription and the examined factors was allowed (F. Chen et al., 2015) to take place before the exposure to UV irradiation.

Cell synchronization was also achieved at certain cases using inhibitors of transcription. The drug 5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) is a P-TEFb kinase inhibitor that inhibits transcription. Specifically, it inhibits the phosphorylation of the CTD sequence of RNAPII through interaction with the factors P-TEFb and DSIF. This results in the stalling of RNAPII molecules at PPP, as the mechanisms for their release to transcription elongation are not functional. Inhibition of transcription is reversible; once the DRB factor is removed, the RNAPII molecules are phosphorylated and released from the PPP regions to enter transcription elongation.
DRB treatment also allows the uncoupling of the dynamics of two previously indistinguishable subclasses of elongating RNAPII molecules: the ones that are already engaged in elongation prior to stress (pri-elongating) and the "de novo" PPP-released polymerases (Ip et al., 2011; Jonkers et al., 2014; Levens et al., 2016; Yamaguchi et al., 2013).
Triptolide (TRP) inhibits transcription by binding to the XPB subunit of TFIIH, which is required during the first steps in transcription to open the double-stranded DNA and to create a "transcription bubble" (Figure 1.6). TRP inhibits the ATPase activity of XPB, preventing the formation of the transcription bubble, and therefore inhibits transcription initiation. TRP treatment is irreversible as it binds covalently to XPB and activates a rapid proteasome-dependent degradation of RNAPII (Bensaude, 2011; Titov et al., 2011; Vispé et al., 2009).

The above inhibitors were placed directly in the medium, at time points described in detail in the respective experimental schematics (see next sections).

2.3 UVC Cell irradiation

The cells were exposed to UVC radiation (254nm, TUV Lamp, Philips). Doses corresponding to $8,\ 15$ and $20\ J\ /\ m^2$ were applied. Prior to UV exposure, the medium was removed from the cells, and a PBS wash was performed to remove nutrient residues that could absorb the irradiation. The plates were then left to recover in normal medium (10%FCS) for a certain period of time at 37 ° C.

2.4 Acetic histone extraction

Cells were placed on ice and washed twice with cold PBS, collected in PBS 1x solution containing 1 mM EDTA, 0.5 mM EGTA (Egtazic Acid) and 1 mM PMSF (Phenylmethylsulfonyl Fluoride) followed by centrifugation at 2000 rpm for 5 minutes at 4 ° C. The supernatant was removed, the cell pellet was resuspended in PBS (10X pellet volume) and then centrifuged at 2000 rpm for 5 minutes at 4 ° C. Cell pellet was resuspended in 10 volumes of Lysis Buffer (10mM HEPES PH 7.9, 1.5mM MgCl2, 10mM KCl, 0.5M Dithiothreitol (DTT), 1.5mM PMSF) and then sulfuric acid was added to a final concentration of 0.2M. The samples were incubated on ice for 30 minutes and then centrifuged at 10,080 x g for 10 minutes at 4 ° C. The supernatant was collected and trichloroacetic acid (TCA) was added at a final concentration of 20%. Vortex was applied and incubated for one hour on ice. Centrifugation at 14,000 rpm for 15 minutes at 4 ° C was applied, the supernatant was removed and 1 ml of cold acetone (-20 ° C) was added to the residue. Then, centrifugation at 14,000 rpm for 5 minutes at 4 ° C was applied, the acetone supernatant was removed and speedvac was performed. Finally, the pellet was resuspended in a suitable volume of TE solution (10 mM Tris, 1 mM EDTA) and the samples were stored at -80 ° C.

2.5 In vivo crosslinking

Formaldehyde creates reversible protein-DNA, protein-protein and protein-RNA chemical bonds. Formaldehyde was added to the cell medium at a final concentration of 1% (from 37% stock solution). After incubation for 12 minutes, glycine (stock 2.5 M) was added to the medium to a final concentration of 0.125 M for 6 minutes to stop the above reaction. The cells were then washed twice with cold PBS 1x and collected in PBS 1x containing 1 mM EDTA, 0.5 mM EGTA and 1 mM PMSF. Finally, cells were divided into 2 * 10^7 cell/pellets, and either used directly for chromatin lysis as described in Lavigne et al., 2017, or they were frozen in liquid nitrogen and stored at -80 ° C.

2.6 Chromatin Immunoprecipitation sequencing, ChIP-seq

Chromatin immunoprecipitation (ChIP) of crosslinked UV or non-irradiated chromatin was carried out from at least two independent cultures of cells per condition as described in (Lavigne et al., 2015, 2017). Antibodies used in the ChIP experiments are listed in the table below.

Table 5 Antibodies used in the ChIP-seq experiments in this study

Antibody	Brand	Catalogue Number
anti-Pol II-hypo (8WG16)	Millipore	05-952
anti-Pol II-Ser5P	Millipore	04-1572-I
anti-Pol II-Ser2P	Abcam	ab5095
anti-H3K27ac	Abcam	ab4729

After precipitation, ChIPped DNA was quantified on a Qubit 2.0 Fluorometer (dsDNA HS Assay Kit, Thermo Scientific) and ChIP specificity was checked by qPCR analyses performed with 10–100 pg of ChIP and Input DNA in duplicate reactions with qPCRBIO SyGreen mix (PCR Biosystems) on a Roche Light Cycler 96 instrument. At least two independent ChIP replicates were validated by ChIP–qPCRs. If individual ChIPs showed sufficient enrichment in control genomic regions, respective ChIP and Input DNA (1-10 ng) were subjected to library prep for NGS. (Lavigne et al., 2017; Liakos et al., 2020).

2.7 Total RNA and nascent RNA (nRNA) extraction

Cells were grown on 55 cm surface plates to a confluency of about 80%. After the medium was removed, cells were harvested in 500μl of trizol (Trizol, Life technologies) on ice. Next, 100 μl of chloroform (A1935 chloroform-isoamyl 24: 1, Applichem) was added and stirred with a vortex apparatus for a few seconds. The samples were then centrifuged for 15 minutes at 12,000 rpm. After centrifugation, the upper phase was carefully collected in a new eppendorf vial. In order to precipitate the RNA, 20 μg of glycogen, 1/10 sample volume of sodium acetate 3M pH5.5, and 2.5 volumes of ice-cold 100% ethanol were added. The samples were then left for at least 12-16 hours at -80^{o} C and then centrifuged at 16,000 rpm at 4^{o} C, and the pellet was rinsed with 70% ethanol. Then, the amount of nucleic acids in the samples was measured with the nanodrop spectrometer.
At this stage, except from RNA, the samples contain a quantity of DNA. To remove DNA, 20μg were incubated at 37^{o}C for 30 minutes with DNase I, according to the manufacturer's instructions (Turbo DNase, Ambion, Life Technologies). This was followed by purification with

acid phenol (acid phenol)/ chloroform pH 4.5 (ThermoFisher Scientific), homogenization, and centrifugation for 15 minutes at 16,000 rpm.
The supernatant was collected in new eppendorf vials, and an equal volume of chloroform (A1935 chloroform-isoamyl 24: 1, Applichem) was added. This was followed by homogenization and centrifugation at 16,000 rpm and finally precipitation for at least 12-16 hours at -80° C. The samples were centrifuged at 16,000 rpm at 4º C. The pellet was next washed with 70% ethanol and subsequently dissolved in clean RNAse-free water. RNA concentration was measured on the nanodrop spectrometer and stored at -80° C.
For the newly synthesized RNA (nRNA) isolation experiments, EU (Ethylene Uridine) was used as a uridine analog to label the newly synthesized RNA as follows: 5-10 minutes before the extraction of total RNA with Trisol, EU-labeled uridine analogue, 100 μM Click-iT ™ (Nascent RNA Capture Kit, C10365, ThermoFisher Scientific) was added and purification of total RNA was performed as described above at the indicated time. An initial amount of 5-10 μg of total RNA was used for each nRNA sample. After DNA removal by DNAse I, only EU-labeled molecules were selected. This was achieved by the biotin-azide -EU chemistry using the Click-iT ™ package (Nascent RNA Capture Kit, C10365, ThermoFisher Scientific), which contains magnetic beads covered with biotin-azide. The manufacturer's protocol was applied, with some modifications during the cDNA synthesis stage.
For the synthesis of cDNA, nRNA was used as the starting material, without being released from the magnetic beads. In other words, cDNA was synthesized on magnetic beads (on beads cDNA synthesis).

2.8 Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq)

ATAC-seq is a method used to map and study accessible ("open") chromatin regions in a genome-wide fashion (Buenrostro et al., 2013, 2015).The basic principles of the protocol are described in Figure 31.

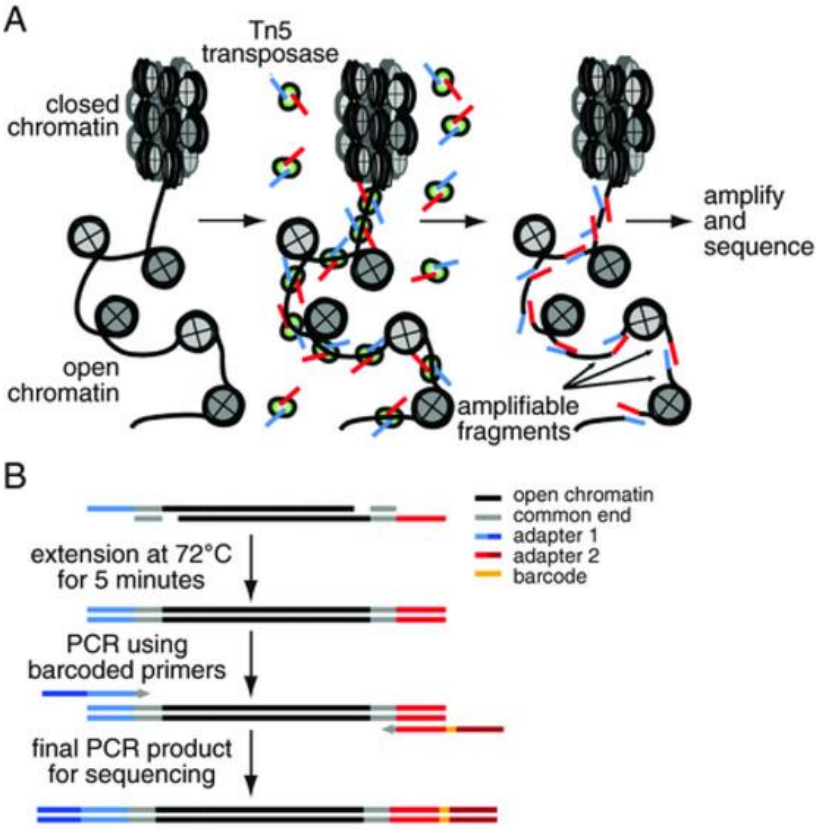
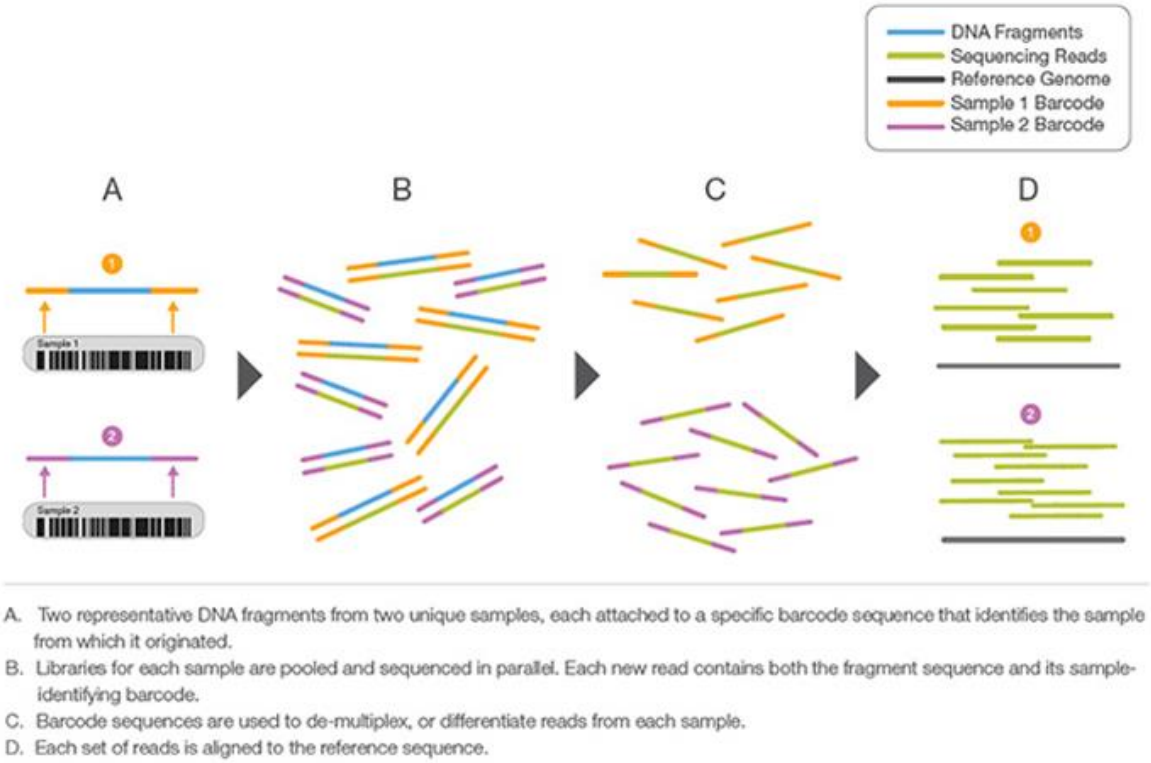


Figure 31 ATAC-seq methodology. (A) ATAC-seq is based on the activity of Tn5 transposase, which is attached with in vitro adapters, suitable and compatible with next generation sequencing (NGS) techniques. Tn5 transposase can cleave and integrate specific adapters at genomic regions where "open" chromatin is present, such as the regulatory regions of promoters and enhancers. (B) Following the transposase reaction, DNA is isolated and then amplified by PCR. Prior to amplification, adapters have to be ligated with a 72°C extension step. During the subsequent PCR, additional sequence is incorporated into the adapters, which include common sequencing ends and a sequencing barcode. From (Buenrostro et al., 2015).

For this study, an improved ATAC-seq protocol (omni-ATAC-seq) was used, which reduces mitochondrial DNA contamination and is characterized by a higher signal / noise ratio than the original method (Corces et al., 2017).

2.9 Construction of NGS compatible DNA libraries

Double-stranded DNA fragments derived from either chromatin immunoprecipitation or RNA isolation (cDNA) were modified to be compatible for NGS. The protocol that was used, results in the binding of a 6-nucleotide sequence (Illumina NEBNext adapters) to the cDNA, which acts as a molecular identity (index). This methodology allows for the parallel sequencing of multiple samples together (multiplexed samples), which can be separated after the sequencing procedure using bioinformatics techniques (demultiplexing) (Figure 32).



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

Figure 32 Adapted by Illumina

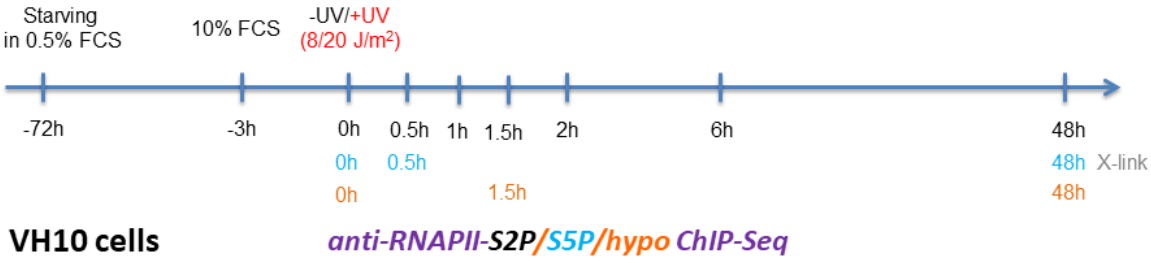
Briefly, the basic steps of this procedure are: (1) Blunt ending (end repair) addition at DNA ends, (2) A-base addition at 3' end of DNA, (3) Illumina NGS adapters attachment, (4) DNA fragmentation (150 bp - 500 bp), and (5) DNA fragment PCR amplification.

2.10 Next Generation Sequencing

All libraries in this study were sent to Genecore-EMBL and sequenced using the Illumina HiSeq 2000 platform for 50 sequencing cycles (maximum of 50 bp per sequenced read), resulting to one FASTQ file per sequenced library, including hundreds of millions of 50-nucleotide sequence reads, which were analyzed by the bioinformatics pipeline described in the results section.

2.10.1 ChIP-seq of RNAPII isoforms

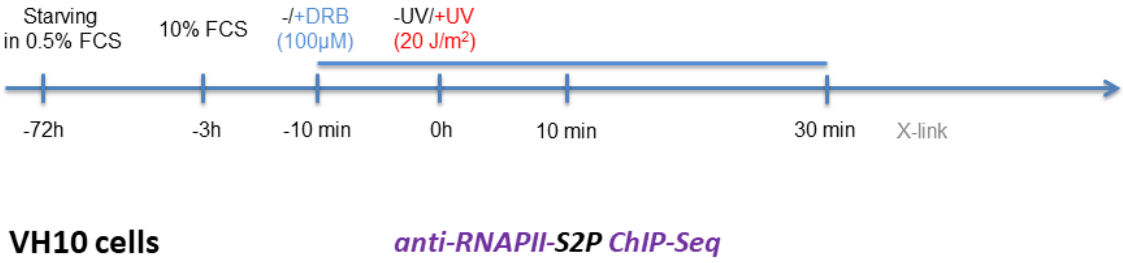
To study the genome wide binding kinetics of the different RNAPII isoforms (pre-initiating, initiating and elongating) during UVC stress recovery in normal human cells (VH10), a series of RNAPII ChIP-seq experiments were performed, as depicted in the schema below.



Specifically, for the RNAPII pre-initiating isoform “hypo” crosslinking was performed at 0 h (NO UV) and +UV 1.5 h after UVC irradiation, for the initiating isoform “ser5P” at 0 h (NO UV) and +UV 0.5 h, and for the elongating isoform “ser2P” at 0 h (NO UV), and +UV at 0.5, 1, 2, 6 and 48 h after UVC induction.

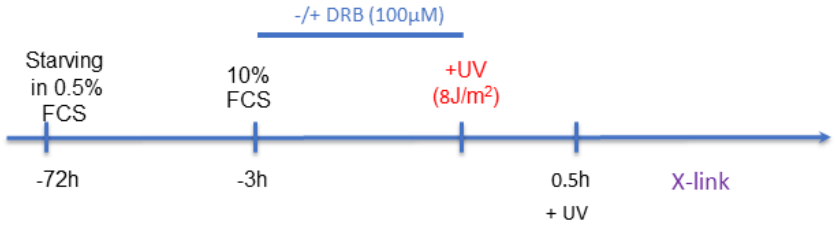
2.10.2 RNAPII-ser2P DRB ChIP-seq

Inhibition of RNAPII transition into transcription elongation, enables the unmasking of the kinetics of the already elongating -prior to UVC irradiation- RNAPII molecules (pri-elongating) from the ones that are released in response to UVC (de novo elongation). For studying the pri-elongating RNAPII molecules, transcription inhibition was performed by DRB, 10 min before UV irradiation. After irradiation, cells were allowed to recover for indicated times in the presence of DRB before crosslinking, chromatin isolation and ChIP for RNAPII-ser2P, as depicted in the schema below.



The generated samples by the above-mentioned experiments included NO UV -DRB, NO UV +0h +DRB, NO UV +10 min +DRB, NO UV +30 min +DRB, +UV +10 min +DRB, and +UV +30 min +DRB conditions.

Subsequntly, to study the genome-wide binding profile of “de-novo” elongating RNAPII molecules, the experimental set-up described in the schema below was followed.

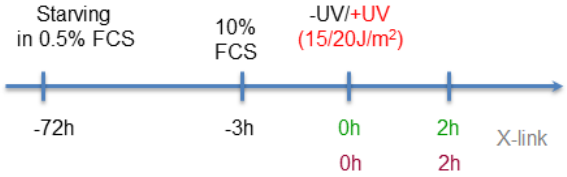


VH10 cells ***anti - RNAPII-Ser2P ChIP-Seq***

The generated samples will be referred as RNAPII-ser2 pre-DRB +UV -DRB and pre-DRB +UV +DRB.

2.10.3 Histone modifications- ChIP-seq

ChIP-seq experiments of histone modifications, and specifically H3K27ac and H3K27me3 are informative about the genome-wide transcriptional active or repressed status of chromatin, respectively, along all the functional genomic elements of interest in the particular study (genes, enhancers, asPROMPTs).

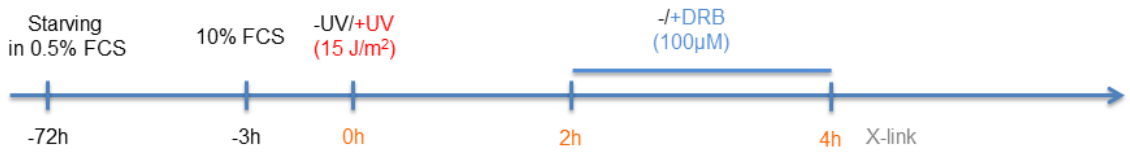


VH10 cells ***anti - H3K27ac/H3K27me3 ChIP-Seq***

In this study, to examine potential alterations of the particular histone modifications during early recovery from genotoxic stress in VH10 cells, H3K27ac NO UV and +UV 2 h, as also H3K27me3 NO UV and +UV 2 h were generated (schema above).

2.10.4 +DRB RNAPII-hypo ChIP-seq

To study the genome-wide profile of RNAPII-hypo binding in response to UVC stress, ChIP-seq experiments were performed using the RNAPII-hypo isoform specific antibody 8WG16 (table 5) in the experimental conditions described in the schema below.

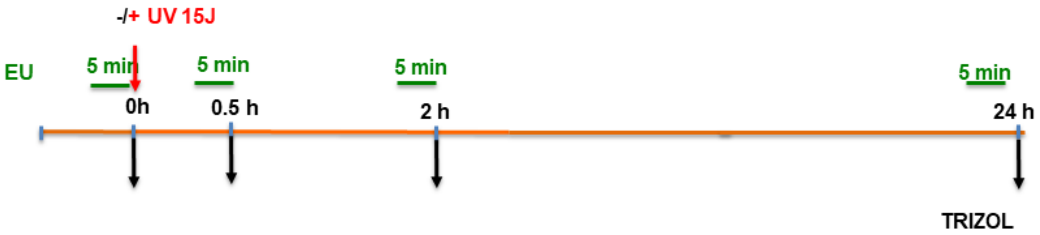


VH10 cells ***anti-RNAPII-hypo ChIP-Seq***

As depicted in the schema above, cells were UV- irradiated and left to recover for 2 h when the levels of RNAPII-hypo are known to be depleted (Heine et al., 2008; Lavigne et al., 2017; Rockx et al., 2000) (DMSO NO UV vs DMSO + UV +2 h). Consequently, DRB inhibitor was applied (or not) to block the release of RNAPII into productive elongation from PPP sites. Crosslinking was applied 2 h after the addition of DRB (or DMSO for the control cells).

2.10.5 VH10 and CSB nRNA-seq

To study the effect of UVC irradiation on nascent RNA synthesis in normal and TC-NER deficient cells, a set of nRNA-seq experiments was performed following the experimental set-up that is depicted in the schema below.

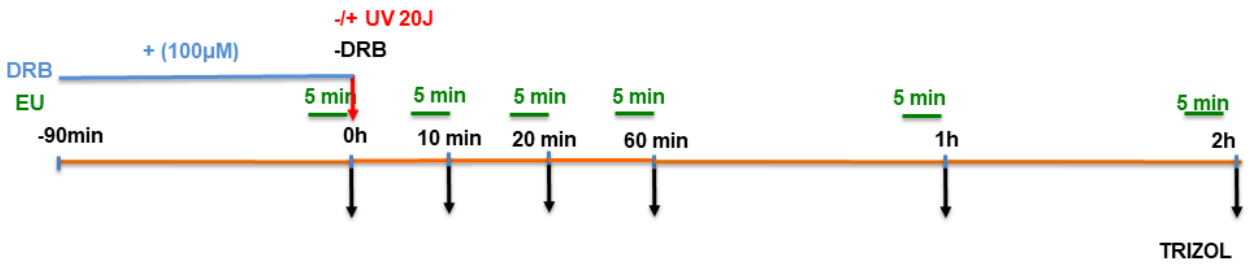


VH10 and CS-B cells

Chip-seq libraries from VH10 and CSB cells in NO UV +0 h, NO UV +24 h, +UV +0.5 h, +UV +2 h, and +UV +24 h conditions were generated.

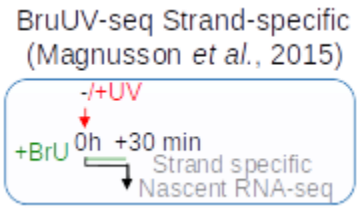
2.10.6 pre-DRB nRNA-seq

Transcription synchronization was achieved using 100 µM of the DRB inhibitor for 3 hours directly in the medium. At the end of this period, the medium was replaced with a fresh one without the inhibitor. This was followed by EU labeling and collection of total RNA, as depicted in the schema below.



VH10 and CS-B cells

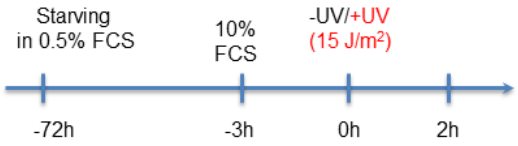
Pre-DRB nRNA-seq experiments according to the experimental set-up depicted in the schema above were performed for both VH10 and CSB cells at NO UV +0 min, +10 min, +1 h and +2 h conditions, and +UV 0 h, +10 min, +1 h and 2 h conditions.
Data generated with a variation of the nRNA-seq protocol, called BruUV-seq, were also analysed in this study. The particular methodology is strand-specific and incorporates Bromouridine (Bru) instead of EU for labeling the nascent transcripts (Magnuson et al., 2015) (schema below).



For the particular study, HF1 BruUV-seq datasets of NO UV and +UV 30 minutes were downloaded from GEO with accession number GSE75398.

2.10.7 ATAC-seq

To study the genome-wide landscape of chromatin accessibility in normal human skin cells (VH10) upon genotoxic stress, omniATAC-seq experiments were performed as described above, following the experimental schema below.

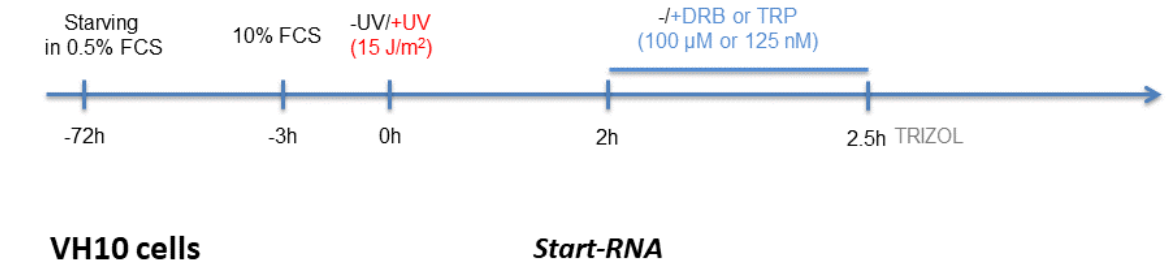


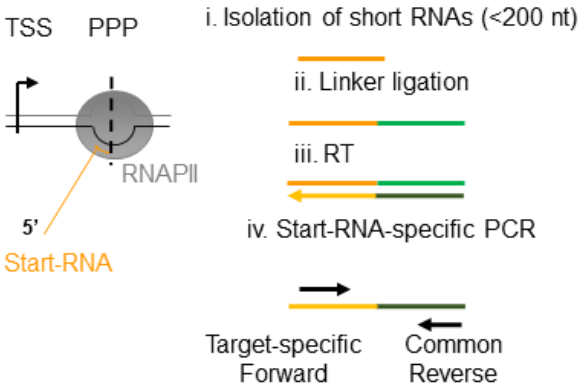
Specifically, as depicted above, VH10 omni-ATAC-seq libraries in NO UV and +UV +2 h conditions were generated.

2.10.8 Start-RNA synthesis

According to this protocol, VH10 cells were irradiated (or not) with UVC, allowed to recover for two hours, and then treated with the transcription elongation inhibitor DRB or the transcription

initiation inhibitor TRP by adding them to their medium (schema below). For each experimental condition, short RNAs less than 200 nucleotides (nt) were isolated and an RNA-DNA linker was attached to their 3 'end. Subsequently, a reverse transcription reaction was performed using a common primer complementary to the linker sequence. qPCR reactions were then conducted to quantitatively compare start-RNA levels in particular genetic regions where RNAPII-ser2P ChIP-seq or nRNA-seq signal was detected (Liakos et al., 2020).





According to the schema above, short RNAs at NO UV / + DRB / T 2.5h, + UV / - DRB / T 2.5h, + UV / + DRB / T 2.5h and + UV / + TRP / T 2.5h were isolated.

2.10.9 Cap analysis of gene expression sequencing (CAGE-seq)

Identification of transcription start sites (TSSs, eTSSs, PROMPT TSSs) and their associated promoters require 5′ end–specific signature sequences for annotating their transcription profiles. For this reason, techniques that perform cloning of short sequence tags from the 5′ end of cDNA, using cap analysis of gene expression (CAGE) (Shiraki et al., 2003) and 5′-SAGE (Hashimoto et al., 2004; Wan et al., 2004) were developed. In these protocols, DNA-linkers are attached to the 5′ end of cDNA to create a recognition site for the restriction endonuclease MmeI adjacent to the 5′ ends. cDNA cleavage is in turn performed by MmeI 20 and 18 nucleotides downstream of the recognition site, creating a two-base overhang. Finally, amplification is applied followed by concatenation of sequencing tags for NGS sequencing (Kodzius et al., 2006)(Figure 33). CAGE-seq accurately determines all kinds of transcription start sites, abundance and directionality of RNAPII transcription at TSSs (Andersson et al., 2014; Liakos et al., 2020; Noguchi et al., 2017).

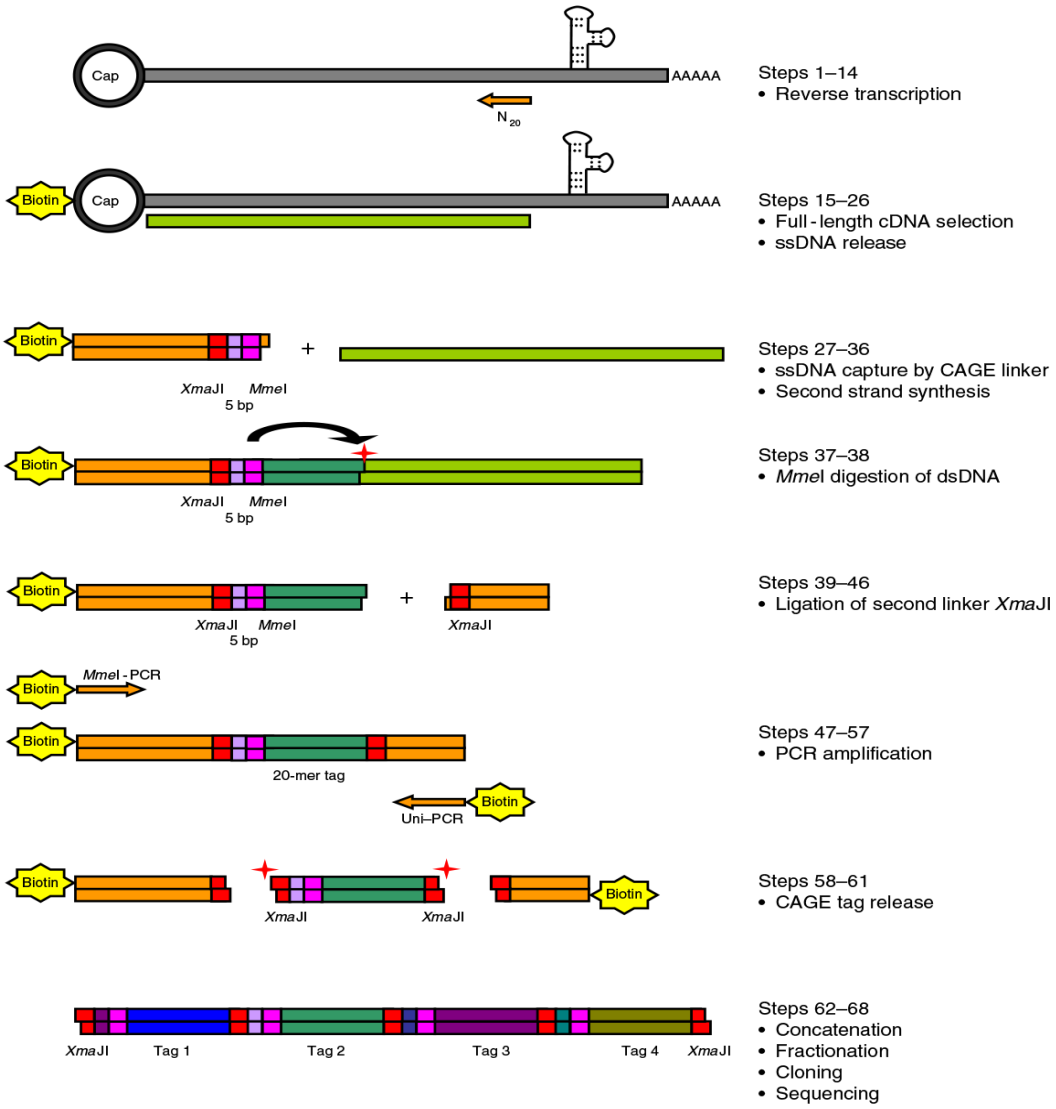


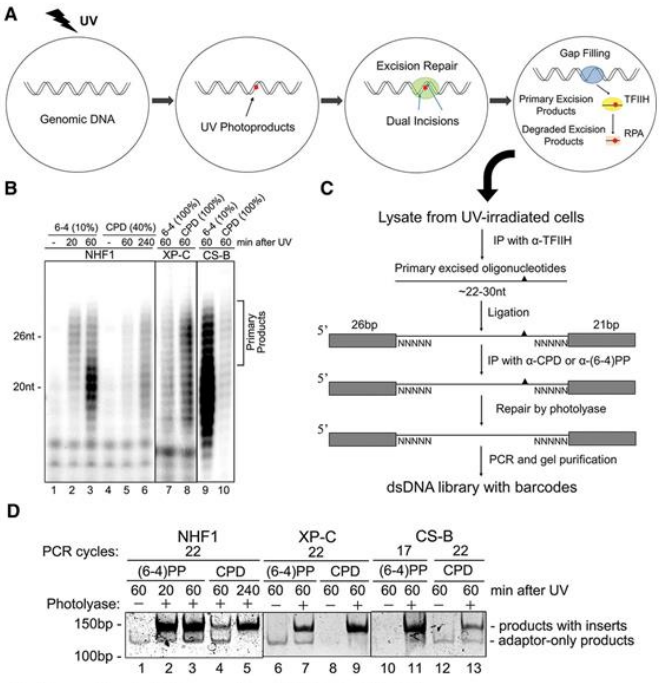
Figure 33 CAGE-seq protocol

For this study, FANTOM5 strand specific CAGE-seq alignment files of normal Dermal fibroblast primary cells (6 Donors with source codes: 11269-116G9, 11346-117G5, 11418-118F5, 11450-119A1, 11454-119A5 and 11458-119A9) and normal skin fibroblasts (2 Donors with source codes: 11553-120C5 and 11561-120D4) were downloaded from ftp://ftp.biosciencedbc.jp/archive/fantom5/datafiles/phase2.2/basic/human.primary_cell.hCAGE and were combined to generate a consensus BAM file. BAM files were further processed and separated into forward and reverse references, and saved as two separate BAM files.

2.10.10 EXcision repair sequencing (XR-seq)

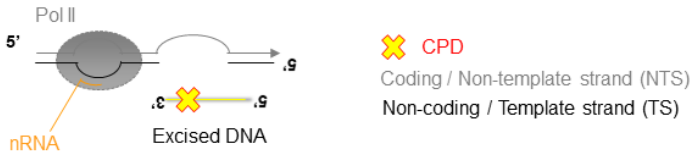
XR-seq methodology (Hu et al., 2015) provides a genome-wide map of excised-DNA sequences during NER repair activity (see section 1.2). Nucleotide excision repair in humans creates two cuts around the DNA-lesion site, resulting in a ~30 bp sequence. The particular fragments are

isolated and subjected to next-generation sequencing producing strand-specific, base-resolution maps of repair of the two classes of UVC-induced DNA lesions, cyclobutane pyrimidine dimers (CPDs) and (6-4) pyrimidine–pyrimidone photoproducts [(6-4) PPs]. Experiments were conducted in normal cells (NHF1 human skin fibroblasts), as also in cells defective in either transcription-coupled excision repair (CSB cells) or global genome excision repair (XP-C cells), addressing the contribution of each NER-pathway to the overall repair profile (Figures 34 and 35). Further analysis of XR-seq datasets enables the capturing TC-NER profile at promoters, enhancers and gene bodies of actively transcribed elements (Hu et al., 2015, 2017; Lavigne et al., 2017; Liakos et al., 2020).



Jinchuan Hu et al. Genes Dev. 2015;29:948-960

Figure 34 The XR-seq method. From (Hu et al., 2015).



TC-NER on TS around TSSs (XR-seq reads from XP-C cells 1 h after UV, data from Hu *et al.*, 2015)

Figure 35 XR-seq data after 1 h recovery from UVC irradiation pinpoints precisely and exclusively the location and levels of transcription-dependent repair (TC-NER pathway) when the assay is performed in GG-NER-deficient cells (xeroderma pigmentosum (XP)-C cells). From (Hu et al., 2015)

XR-seq data of CPD containing excised DNA fragments in wild-type (WT) NHF1 skin fibroblasts, XP-C, and CS-B mutant cells were downloaded by Gene Expression Omnibus with accession number GSE67941.

2.10.11 NHF1 time-course XR-seq

When XR-seq is applied over a time course, the kinetics of NER after UVC irradiation can be mapped (Adar et al., 2016)(schema below). Measurements of repair activity at UVC-induced CPDs at 1, 4, 8, 16, 24, and 48 h and 6-4 photoproducts at 5 and 20 min and 1, 2, and 4 h in normal human skin fibroblasts (NHF1) were generated using the protocol depicted in Figure 36, and were downloaded from GEO with accession number GSE76391.

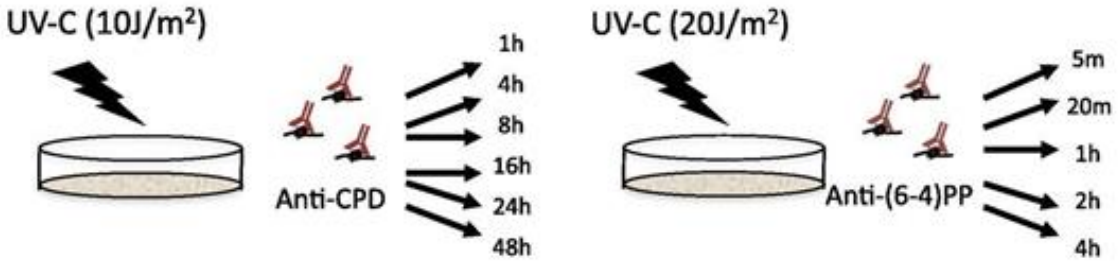


Figure 36 NHF1 time-course XR-seq

2.10.12 XPC XR-seq of CPD damages coupled with double DRB (DRB2) treatment (pulse–chase–pulse)

In the particular variation of the XR-seq protocol, cells were first incubated in DRB for 2 h, to block new molecules of RNAPII from entering transcription elongation and allowing the already elongating RNAPII complexes to complete and terminate transcription. After 2 h of DRB treatment, the inhibitor was washed off, and the cells were incubated for 10 min, and then DRB was added again (Figure 37).

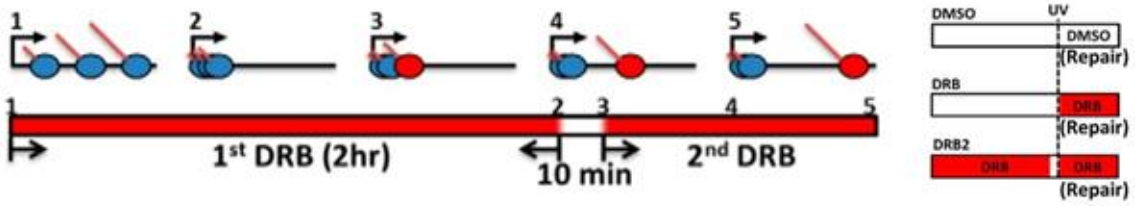


Figure 37 DRB2 XPC XR-seq of CPD experimental set-up. Adapted from (Chiou et al., 2018).

DRB2 procedure allowed the release of only a limited number of RNAPII molecules during the 10-min DRB-free chase period, enabling the repair of DNA lesions at the 5′ end of the gene. The three biological conditions denoted in Figure 37 are reported as CPD XPC +UV 1 h +DMSO, +UV 1 h +DRB and +UV 1 h +DRB2 XR-seq datasets in the rest of this study. The above datasets were downloaded from GEO with accession number GSE106823.

2.10.13 aniFOUND-seq

aniFOUND-seq, is a novel, antibody-free method that can capture the repaired chromatin after UVC irradiation. The particular methodology takes advantage of the unscheduled DNA synthesis (UDS), which occurs during the repair of the UVC-induced DNA lesions and in particular the DNA synthesis step that takes place after the incision of the damaged DNA fragment. The elimination of any DNA synthesis (i.e during replication) other than UDS is a key step in aniFOUND, so that the DNA resulting from UDS is solely and specifically labelled. This is achieved by arresting cells in G0/G1 by both contact inhibition and serum starvation. Additionally, to inhibit DNA synthesis in the small number of cells still escaping G0/G1, cultures were treated with hydroxyurea (HU) during the DNA labeling step. The chromatin associated with newly synthesized EDU- labeled DNA is next isolated (Click-IT chemistry) and this material can be subjected to high-throughput omics analyses like Protein Mass Spectrometry and DNA Next Generation Sequencing (NGS). The main steps of aniFOUND-seq protocol are depicted in Figure 38.

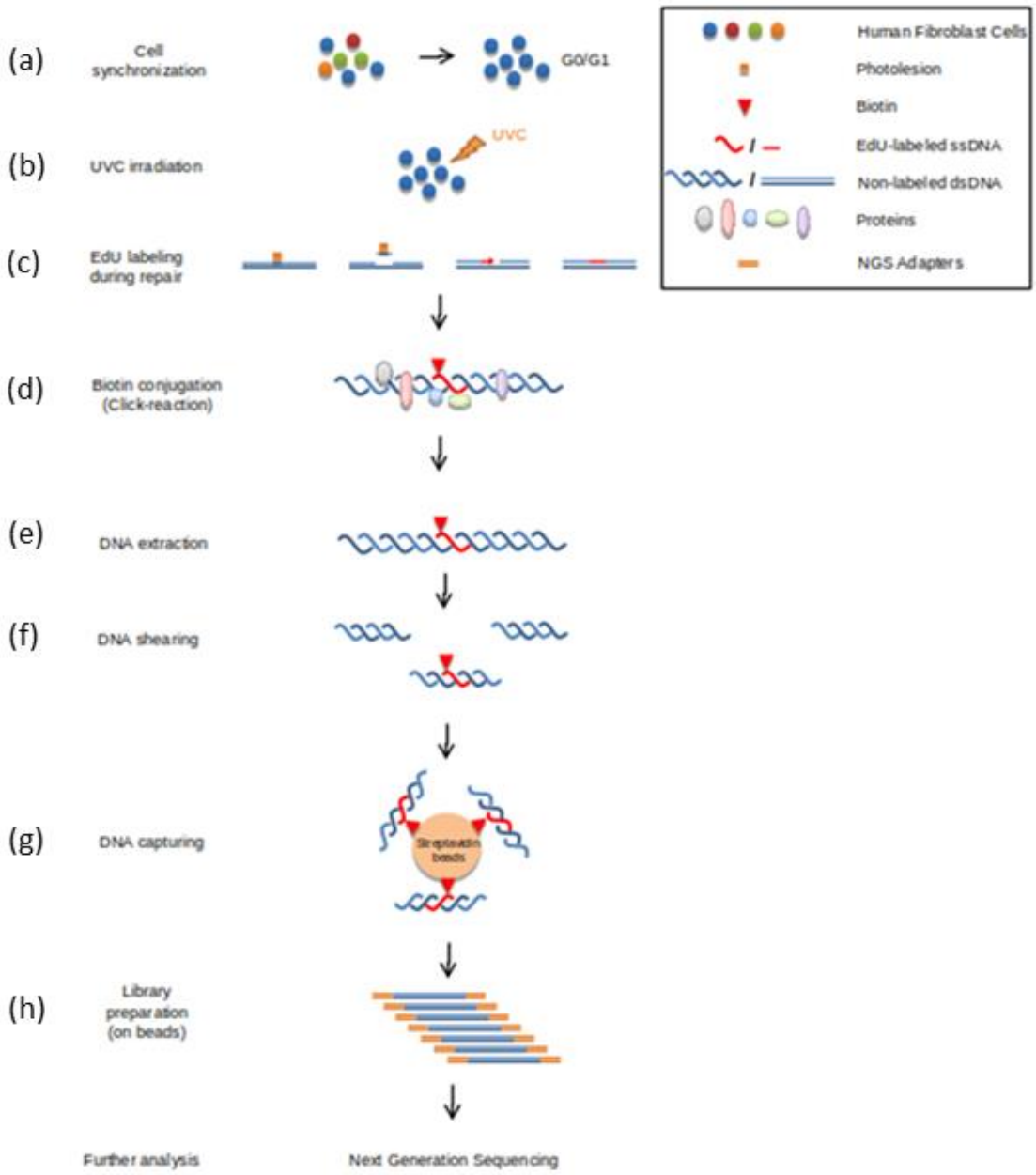


Figure 38 Illustration of aniFOUND method (a) An asynchronous population of fibroblasts is synchronized to G0/G1 phase by serum starvation and contact inhibition and (b) DNA lesions are induced by UVC-irradiation. (c) The lesions are left to be repaired in the presence of labelled nucleotides (EdU). HU is added in this step to eliminate the replication of any escapers. Potential fates of the UVC-derived lesions include repair by NER itself, procession by EXO1, repair of any formed DSBs. Since no replication occurs UDS will be the only source for DNA labelling. (d) Biotin molecules are conjugated on the labelled nucleotides, the chromatin is extracted (e) and sheared (f) and the labelled chromatin fragments are isolated with streptavidin beads. (g) On-beads library construction procedure is carried out that is followed by next generation sequencing (h).

The aniFOUND-seq protocol does not give strand-related information, since capturing of the whole repair related newly synthesized DNA is of particular interest. Even though, with particular experimental adjustments, strand-specificity can be applied to aniFOUND-seq, as only the repaired strand is labelled. This can be accomplished by a DNA denaturation step after the binding to the streptavidin beads followed by a strand-specific library protocol adaptation.

2.11 Peak Calling

One of the major analysis modules of ChIP-seq and ATAC-seq experiments is the identification of regions significantly enriched with NGS signal, that correspond to protein binding events (ChIP-seq) or nucleosome free regions (ATAC-seq). The identification procedure of these events is called “peak-calling”, and there are multiple choices of specialized software that can be applied, depending on the NGS protocol, the under-study factor, the data quality etc.
In this study, MACS2 (Y. Zhang et al., 2008) and epic2 (Stovner et al., 2019) were used, as described in the results section.
Briefly, MACS2 was designed for TF binding site identification (default algorithm behavior), but it is also applicable in chromatin accessibility assays (ATAC-seq, DNAse-seq) by using appropriate parameterization (the --shift -100 --extsize 200 parameters center a 200 bp window on the Tn5 binding site), in RNAPII ChIP-seq data (--nomodel option), or even in narrow and broad histone modifications (--nomodel, --broad parameters). Epic2 is an ultra-performant version of the SICER algorithm [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152844/] that is designed to identify peaks in ChIP-seq datasets with wide binding profiles, such as histone modifications (H3K27ac, H3K27me3, H3K4me1 etc.). During peak detection, the “wideness” can be controlled using appropriate window sizes, and setting the number of gaps between the window search.

2.12 Dimensionality reduction

Dimensionality reduction is a collection of methodologies where a set of dimensions, for example in a {gene by sample} matrix with genes, $(g_1, ..., g_n)$, $g_n \epsilon R^D$ is transformed into a smaller set of $(z_1, ..., z_n)$, $z_n \epsilon R^M$ dimensions, where M << D. These methodologies are considered a category of unsupervised learning systems.
The advantages of applying dimensionality reduction in high dimensional datasets include data compression (in terms of storage and processing speed), and data visualization, in a human readable interpretation of 2 or 3 dimensions.
Lower-dimensional spaces, such as principal component spaces, are often been treated as inputs to supervised learning algorithms (clustering methods) or non-linear dimensionality reduction methodologies (t-SNE, UMAP, diffusion maps) and contribute to better generalization (in the case of clustering) or human-readable visualization (in the case of non-linear dimensionality reduction).

Principal Component Analysis (PCA) (Pearson, 1901) is one of the most widely used methods of linear dimensionality reduction methodologies applied in the field of Bioinformatics. It is a mathematical algorithm that reduces the dimensional space, while maintaining most of the

variation in the data matrix, and reveals patterns of similarity/dissimilarity between the examined groups (in the case of {gene by sample} matrix example, groupings between the samples}) like clustering. In PCA, dimensions are reduced by data projection to lower dimensions (principal components - PCs) in a search of identifying the best data summary using the least number of PCs (at least less than the examined features - genes). Each PC is created with the requirement of minimizing the distance between the data points and their projection onto the particular PC, while consecutive PCs are created with the additional requirement of not being correlated with all previously created PCs (geometrically orthogonal) (Lever et al., 2017). Each PC explains a percentage of the variation of the dataset, starting from PC1 which depicts the higher variation value (see Figure 41b).
In the example of the {gene by sample} matrix, PCA can be applied in the whole, z-scored data matrix to reduce the gene dimensions, or by sub-setting the original matrix to the most variable rows (genes). By focusing on genes that exhibit the highest variability in the dataset is sometimes helpful in highlighting biological signals in NGS datasets and might favor the PCA procedure (Brennecke et al., 2013). The selection of these variable features can be done by using mean-by-variance analysis or dispersion analysis (Stuart et al., 2019).

2.13 Bootstrapping statistical analysis and effect sizes

To apply per sample comparisons between two read-count distributions, a permutation strategy was applied. For each distribution comparison and for each of the tested sets, 10,000 (or 1,000) samplings of 100 data points were randomly generated, and 95% confidence intervals of mean differences of log_2 counts between two groups were calculated. Effect sizes of log2 values between two distributions were calculated using Cohen's method (CES).

3 Summary

The purpose of this study is the development of a computational framework for studying the dynamic changes of active transcription, and its interaction with chromatin remodeling and chromatin alterations during cellular response to genotoxic stress. For this purpose, ultraviolet light C (UVC) was used as a genotoxic stress factor, damaging skin cells, specifically skin fibroblasts (VH10, CSB and 1BR.3), while the activity of Nucletiode Excision Repair (NER) pathway and the repair products of Global Genome NER (GG-NER) and Transcription Coupled NER (TC-NER) sub-pathways were used to evaluate the examined mechanisms.
Various types of Next Generation Sequencing (NGS) experiments have been used to study the stages of the transcription cycle in normal conditions, and in response to Ultraviolet C irradiation (UVC) induced stress. Specifically, for studying the kinetics of RNA Polymerase 2 (RNAPII) molecules from the transcription initiation state, to promoter proximal pausing (PPP), and the transition to productive elongation, Chromatin immunoprecipitation sequencing (ChIP-seq) data of the hypophosphorylated RNAPII (RNAPII-hypo), the elongating isoform of RNAPII (RNAPII-ser2P), and the RNAPII-ser5P isoform (transcription initiation) was generated and analyzed.
To study the productivity of RNAPII molecules during the above stages, Capped Analysis of Gene expression sequencing (CAGE-seq) data and nascent RNA synthesis sequencing (nRNA-seq) data was used. To study the interactions of chromatin with active transcription and its alteration during the states of active transcription, Assay for Transposase-Accessible Chromatin (ATAC-seq) data was generated and analyzed, and ChIP-seq data of H3K27ac and H3K27me3 histone modifications.
To study the effectiveness and genomic landscape of NER repair-synthesis events, for both GG-NER and TC-NER sub-pathways, a novel assay called aniFOUND-seq was developed and analyzed, coupled with data of excised DNA during NER activity (XR-seq) and NER damage sequencing data (damage-seq). The functional assessment of TC-NER at active genes was carried out through the study of mutations in melanoma and lung adenocarcinoma cancer genomes, and XR-seq data meta-analysis respectively.
The results of these essays are divided into four sections:
(1) Development and application of algorithms for the analysis of NGS data related to human disease. (a) Implementation of stand-alone analysis pipelines for the analysis of ChIP-seq, nRNA-seq, and ATAC-seq datasets that include: Quality control (QC) assessment of sequenced short-reads, short-read preprocessing, short-read mapping against the reference genome/transcriptome under study, alignment processing, alignment summarization in genomic features and visualization via heatmaps and average profiles, generation of genomic tracks viewable in genome browsers (IGV, UCSC), NGS signal clustering upon functional genomic regions, correlation of biological and technical replicates, dimensionality reduction methods to identify technical/biological similarities/differences between samples, differential expression analysis, peak calling analysis, differential binding analysis, differential accessibility analysis and other statistical comparisons between biological conditions.
(b) Implementation of a “de novo” elongation wave identification algorithm using Hidden Markov Models (HMMs), and DRB-nRNA-seq datasets.
(2) Cellular responses under genotoxic stress conditions. (a) Development of a computational

pipeline for the study of the reorganization of transcription and the chromatin rearrangements upon UV-induced stress that include: genome annotation reconstruction, and characterization of transcribed units' activity (promoters, genes, enhancers, PROMoter uPstream Transcripts - asPROMPs) along the human genome, the quantification of the RNAPII release from PPP sites, and the evaluation of the RNAPII elongation wave kinetics.
(b) A proposed model describing the 'safe' mode mechanism of transcription elongation; upon UVC-induced stress, steady-state transcription levels of virtually all actively transcribed genes are re-adjust to fast and uniform release of RNAPII elongation waves (green triangles, Figure 39) from PPP sites that scan the transcribed genome for DNA lesions.

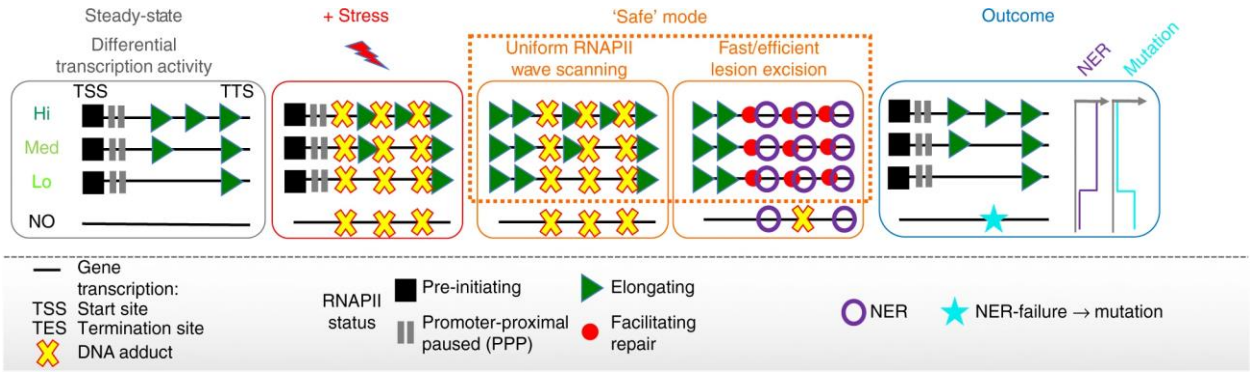


Figure 39 'Safe' mode mechanism of transcription elongation (Lavigne et al., 2017).

This mechanism maximizes the speed of lesion sensing, the probability that a damage will be identified by an elongating RNAPII molecule (red dots, figure 39) and removed (purple rings, figure 39) by the TC-NER (crosses, figure 39) along the actively transcribed elements. As a result, environmentally exposed genomes are characterized by a modest and homogeneous mutation prevalence across the actively transcribed genome in both strands, as opposed to the non-transcribed elements where higher mutation rates are observed (Alexandrov et al., 2013). In case NER is unsuccessful or is not recruited efficiently during the stress recovery process, unrepaired DNA lesions can provoke error-prone DNA synthesis and result in mutagenesis (turquoise star, figure 39) during replication (Lavigne et al., 2017).

(3) Extending the previously described 'safe' mode mechanism of transcription elongation, the results of the particular dissertation also support a model of continuous transcription initiation that can fuel the widespread UV-triggered escape of RNAPII into the transcription elongation, and safeguarding the integrity of the actively transcribed genome. The particular mechanism is supported by a global increase of chromatin accessibility at all actively transcribed promoters serving as a platform that favors unrestrained transcription initiation, coupled by preservation of the active mark H3K27ac and repressive mark H3K27me3 mark during early response to genotoxic stress.

(4)A genome-wide analysis pipeline for the evaluation of aniFOUND-seq methodology
aniFOUND, is the first methodology (at the time of writing this thesis) that can exclusively label, capture and map the post-damage newly synthesized repaired chromatin in its native form (see

materials and methods) (Stefos and Szentai, under revision]. Coupling of aniFOUND to NGS, allows the mapping and characterization of the NER efficacy of different chromosomal regions of the human genome. aniFOUND-seq was successfully applied to map the repair-synthesis activity along damaged skin fibroblasts (1BR.3 cells) with particular attention to promoter and enhancer sequences. Furthermore, aniFOUND-seq was applied for the assessment of NER-UDS activity in several chromosomal regions, including the fraction of repetitive DNA. Specifically, the repair efficacy during the first 4 hours after damage induction was clarified for rDNA and telomeres, for which contradictory explanatory models have been suggested. This is the first time that NGS-based approaches are adopted for shedding light in the above-mentioned inquiries regarding repair of telomeric DNA. Evidently, the cumulative nature of aniFOUND-seq (in terms of both damage types and repair assessment period) renders it applicable for the cases that require capturing of the whole repair process, or the repair activity during moderately-to-considerably long-time windows (Stefos and Szentai, under revision).

4 Results

4.1 Automated analysis of NGS data

The NGS data analysis process is performed automatically as a unified pipeline for each type of NGS data (ChIP-seq, nRNA-seq, ATAC-seq), or step by step, for more efficient inspection and revision of the intermediate results. Specific reports enable the possibility to apply changes in the initial configuration of the parameters of each analysis module, or even bypass (whenever applicable) one of the analysis steps that does not fit to the analysis plan. Most, but not all, modules can be run using multiple cores, while parallelization is also applied in the level of total processed files. All modules have been run and tested only in UNIX-based systems (Ubuntu and Kubuntu). The pipelines were applied in a Unix terminal by following a default set-up, or adjusted appropriately based on the analysis requirements.

4.1.1 Quality control (QC) of raw FASTQ files

QC of raw NGS files (FASTQ) is an essential step for the evaluation of the sequencing and library preparation quality (see introduction). For every FASTQ file, a quality control assessment is applied using the FASTQC toolkit. As described in the introduction, the resulting html report contains valuable sequencing quality metrics and statistics. This module can be run separately for inspection of QC reports that can result in adjustments of the analysis preferences.
An additional analysis option using a “blast search” can be applied in search of potential contaminant sequences based on the FASTQC “Overrepresented Sequences” results (Andrews, 2015), using the BLAST search engine (Altschul et al., 1990). In that case, a FASTA file containing the sequences of the candidate contaminants should be provided, and an additional blast report will be generated as an output (Figure 40 left panel). By default, this option is not enabled.

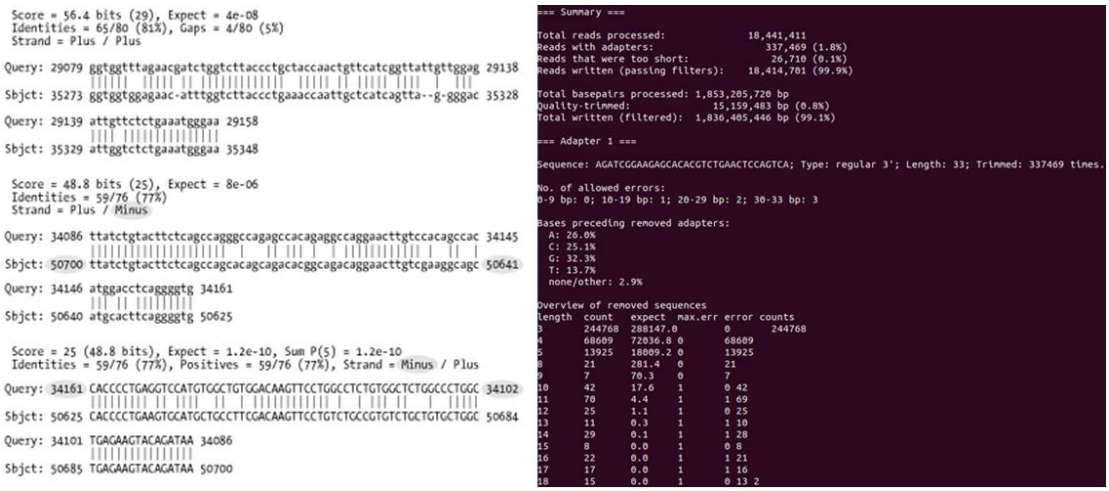


Figure 40 Blastn report (left panel) and cutadapt report (right panel).

4.1.2 Adapter clipping and quality trimming of raw FASTQ files

Any known or observed (based on the QC reports) abnormalities present in the FASTQ files should be eliminated before proceeding to the mapping step. For every FASTQ file, adapter clipping and quality trimming is performed using the python tool cutadapt(Martin, 2011). By default, the tool searches for occurrences of “Ns” (according to the IUPAC nucleotide code, any of the nucleotides A,T,C,G) in the 5’ and 3’ ends of each read and trim them, then searches for bases in the 5’ and 3’ for bases with Phred-33 quality score (see introduction) less than 20 and trims them, and finally using the adapter information included in FASTQC “Adapter Content” and “Overrepresented Sequences” it clips any contaminants present in the examined library. Filtered sequences of length less the 20 nucleotides are discarded. The output of this module is (a) a filtered FASTQ file, and (b) the cutadapt report (Figure 40 right panel), while it is also possible to output the filtered sequences in a separate FASTQ file. If the adapter sequences used during the generation of each examined library are known a-priori, a comma-separated list of the respective sequences are provided.

4.1.3 ChIP-seq analysis pipeline

4.1.3.1 Short-read mapping and alignment filtering

The most common step during NGS data analysis is the mapping of the filtered FASTQ reads against the under-study reference genome. For every filtered FASTQ file, bwa-mem (H. Li, 2013) with default parameters is applied, with a provided reference genome index generated by the “bwa mem index” command. To define a “uniquely” aligned set, hits with a MAPQ score (see introduction) less than 30 are filtered out using samtools, while chimeric and secondary alignments are filtered out using the ‘XA’ and ‘SA’ tags. Additionally, only alignments with at most 2 mismatches between the subject and the reference sequences are kept, in order to account for sequencing errors and SNPs between the reference cell line and the sequenced genome. In the case of paired-end reads, only proper paired-mates are kept, using the

command *samtools view -f 0x2* | *samtools sort -n -* | *samtool fixmate -m - -* | *samtools sort -* before applying deduplication with *samtools markdup*. If technical replicates are present in the dataset, files are concatenated using *samtools merge*, while if biological replicates are present, all files are first down-sampled to the lower alignment depth between replicates using samtools view -s, and concatenated using *samtools merge*.

4.1.3.2 Peak calling analysis

To identify genomic regions significantly enriched with ChIP-seq alignments that represent DNA binding events, peak-calling is applied. Peak-calling is performed at the merged datasets, if replicates are present. Several peak-calling approaches are available, depending on the type of the ChIP-seq protocols used in the study. In the particular study, 5 different procedures have been used, namely ”TF”, “Pol2”, “Histone_narrow”, ”Histone_broad”, and “ATAC”. ChIP-seq peak calling is commonly applied using a control library to model the background signal, but if this is not applicable, the background signal distribution can be formed by using the examined sample. “ATAC” mode, which is suitable for ATAC-seq datasets, is an exception since a typical ATAC-seq experimental design does not include generation of control libraries. Precomputed effective genome sizes (defined as the length of the “mappable” genome) are set accordingly, based on the examined organism and the respective genome build.
TF: This mode is appropriate for transcription factors and similar types of ChIP-seq experiments, where a relatively “narrow” binding profile is expected. In this mode, MACS2 peak-caller is applied, with “--keep-dup all” option enabled, and qvalue threshold set to 0.05, and log2FC > 1. If the reads are paired-end, “-f BAMPE” and “--shift 0” are also applied.
PolII: This mode is suitable for different types of RNAPII ChIP-seq experiments, like hypo-RNAPII, RNAPII-ser5P, and RNAPII-ser2P (see materials and methods). In this mode MACS2 peak-caller is applied, with “--keep-dup all” and “--nomodel” options enabled, and qvalue threshold equals 0.05, and log2FC > 1. If the reads are paired-end, “-f BAMPE” and “--shift 0” are also applied.
Histone_narrow: This mode is appropriate for ChIP-seq experiments of histone modifications, with a relatively narrow binding profile, like H3K27ac, H3K4m32 and H3K4me3. In this mode epic2 peak-caller is applied, with “--keep-duplicates” option enabled, window size set to 200, gap set to 1, and FDR threshold set to 0.05. An external threshold of log2FC (signal over background) is set to 1.
Histone_broad: This mode is appropriate for ChIP-seq experiments of histone modifications, with a relatively broad binding profile, like H3K27me3, H3K4me1 and H3K9me1. In this mode epic2 peak-caller is applied, with “--keep-duplicates” option enabled, window size set to 400, gap set to 3, and FDR threshold set to 0.05. An external threshold of log2FC is set to 1.
ATAC: This mode is suitable for ATAC-seq protocols, like classic ATAC-seq and omni-ATAC-seq (see materials and methods). MACS2 peak-caller is applied, with “--keep-dup” all, --nomodel, --shift -100 --extsize 200 and --call-summits options enabled, and qvalue threshold set to 0.05, and log2FC > 1. If the reads are paired-end, “-f BAMPE”, “--nolambda”, “--shift 0” are also applied.

Direct comparisons between the total number of peaks between samples cannot be applied without downsampling the BAM files to the same level of mapped reads, and without the same peak-calling mode applied.

4.1.3.3 Sample similarity analysis

Sample similarity assessment is a quality control step, essential for any NGS data analysis pipeline. The purpose of this analysis module is to provide visualizations that will reveal expected/unexpected similarities or differences between samples that will help the analyst to draw conclusions about the quality of the datasets. Identification of problematic datasets can help avoiding the creation of biases that could affect the subsequent analysis steps, and prevent the drawing of inaccurate biological conclusions.
High quality and non-merged BAM files produced during the 'Short-read mapping and alignment filtering' step (section 4.1.3.2) are summarized in 4 sets of genomic regions used as references, and are further processed to produce individual visualizations. The specific annotations are: (1) An "extended promoter" set, that includes the 4 kb regions centered at RefSeq transcription start sites (TSSs) (see introduction), (2) a "gene bodies" set, that includes all RefSeq genic regions, (3) all 3 kb annotation-agnostic genomic windows of the examined genome build, with a sliding window of 500 bp, using the BAM headers of the aligned samples, and (4) a consensus peak set from section 4.1.3.2. For the creation of the latter region set, all peak-sets are concatenated, sorted based on their genomic coordinate, and merged using the *bedtools merge -d 0* command.

4.1.3.3.1 Heatmaps of sample-to-sample correlations

Data correlation is a very common procedure that aids the identification of similarities or/and differences between the under-study datasets. For every library, filtered alignments are examined for overlaps using the aforementioned genomic sets, and are reported using the R summarizeOverlaps function, with "Union" mode enabled (Figure 19), resulting in a {region by sample} count matrix, for each feature set. Rows with a total sum less than 10 are discarded, and each count matrix is processed using the R *cor* function, producing a {sample by sample} Pearson correlation matrix. Euclidean distances of pairwise sample correlations are computed using the R *dist* function, to produce a distance dissimilarity matrix. Finally, hierarchical clustering is applied on the provided distance matrix using the R *hclust* function with the "average" method enabled.
Heatmaps of pairwise sample correlations are generated using R *pheatmap* function, with rows and columns clustered based on the aforementioned methodology. An example correlation heatmap of 4 ATAC-seq samples (see materials and methods) are depicted in Figure 41.

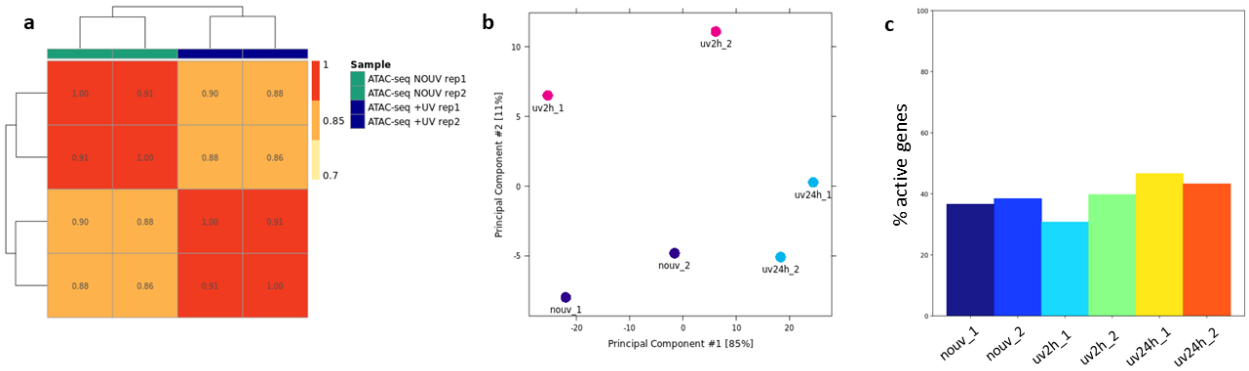


Figure 41 Sample similarity plots. (a) Correlation heatmap of ATAC-seq samples described in materials and methods. Hierarchical clustering groups the replicates of each condition together. (b) PCA plot visualization for a set of Hela H2Bub ChIP-seq datasets. The experimental set up includes 3 biological conditions: (1) nouv, where cell are in normal conditions, (2) uv2h, where cell are recovering for 2 hours after irradiation with UVC (20 $\boldsymbol{J/m^2}$) and (3) uv24h, where cell are recovering for 24 hours after irradiation with UVC (20 $\boldsymbol{J/m^2}$). The first replicates of nouv and uv2h were run in different time periods than the rest of the samples. Read counts are generated using the RefSeq human gene set as a reference, and PCA is calculated using the most variable features using a mean to variance strategy (see materials and methods). In this example, a batch effect is captured in the reduced dimensional space. Specifically, the first batch of nouv and uv2h conditions (nouv_1 and uv2h_1) are very similar based on the first PC. The same is true with the second batch of nouv and uv2h (nouv_2 and uv2h_2), while the second PC seems to capture the biological condition differences between both batches. On the contrary, the uv24h replicates which were generated in a single batch are grouped together using the PC1 and PC2 dimensions. (c) FRIP plot visualization for a set of Hela H2Bub ChIP-seq datasets described in (b). As a reference feature set, RefSeq transcripts were used.

4.1.3.3.2 Principal Component Analysis sample comparison

Principal component analysis (PCA) can be applied to visualize differences or similarities between NGS datasets. PCA is specifically useful for identifying problems with experimental designs, mislabeled samples, batch effects and other unexpected flaws.
Count matrices are created and filtered as described above, while the most variable genomic features (most variable rows) are selected using the R *mean.var.plot* method (Stuart et al., 2019). The new {most variable features x samples} matrix is z-transformed and used as an input for the R *prcomp* function, which in turn calculates the principal components.To visualize the new dimensional space, the first two principal components are used to generate scatter plots using R ggplot (Figure 41 (b)).

4.1.3.3.3 Fraction of Reads In Peaks (FRIP)

Calculating the percentage of mapped reads that fall into enriched regions (peaks or other functional genomic sets that are highly correlated with the under-study factor) is a good

indication of the quality of immunoprecipitation and the experiment per se (Ji et al., 2008). Typically, a small fraction of the alignments in ChIP-seq overlaps with significantly enriched genomic regions (peaks, genes, enhancers etc), as the majority of the mapped reads represents background. In most cases FRiP values show a high correlation with the magnitude of enriched regions. FRiP is a useful statistic for comparing ChIP-seq datasets generated by the same antibody across different cell types, as also for comparisons between antibodies using the same binding factor. Comparisons between biological/technical replicates or biological conditions within the same experimental setup can also be applied.
For this analysis, all the BAM files are indexed using samtools, and are processed using *deeptools plotEnrichment* (Ramírez et al., 2014), to produce signal enrichment fractions across the provided regions, relative to the total genome alignments. An example of the resulting bar graph is shown in Figure 41 (c).

4.1.3.4 Differential binding analysis

Differential binding analysis is performed to identify genomic regions with statistically significant differences in ChIP-seq enrichment, between different biological conditions and treatments. These differences may not be apparent through general visualization strategies such as average profiles (see below), because of the complexity of the datasets. In some cases where particular biological treatments may cause global alterations in the binding profile of the under-study factor (like UVC induced stress), the most significant changes can be captured using the appropriate methodologies such as diffBind with DESeq2 analysis enabled (Stark & Brown, 2011). There are two main types of differential binding tools: (1) “Peak-based” methods that perform the whole analysis in a predefined genomic region set (peaks, genes, HMM chromatin states) like diffBind and Manorm (Shao et al., 2012), and region-based systems, that perform the analysis in genomic windows using a predefined size, a gap size for performing concatenations between consecutive windows of similar binding profiles and a sliding window. This category of tools includes csaw (Lun & Smyth, 2015) and diffReps (Shen et al., 2013). In this study, the particular analysis module is performed by using diffBind software, and by applying pairwise comparisons “ConditionA_vs_ConditionB”, where ConditionB should represent the denominator of the underline comparison. Analysis for differential binding is applied on the filtered and merged peak-sets generated in the 4.1.3.2 section. Peaks of both conditions are concatenated and merged to create a consensus peak-set. If batch effects or any additional confounding factor are present in the experimental set-up, they can be modeled in the experimental design formula of the algorithm (see the diffBind vignette) by enabling the blocking factor parameter in the *dba.contrast* function. The analysis can be performed either by using DESeq2 (default) or/and edgeR. The particular analysis module generates a set of outputs depicted below:

(1) A tab-delimited text report of all the examined regions with column-wise information structure as follows:

1st column: Chromosome of the examined genomic region.
2nd column: Starting base position of the examined genomic region.
3rd column: End base position of the examined genomic region.

4th column: Length of the examined genomic region.
5th column: Strand orientation of the examined genomic region (if present).
6th column: “ConditionA” average RPKM of normalized read counts.
7th column: “ConditionB” average RPKM of normalized read counts.
8th column: Concentration - mean (log) reads across all replicates in both groups (normalization using the respective analysis algorithm).
9th column: "ConditionA" Concentration - mean (log) reads across all replicates of "ConditionA" condition (normalization using the respective analysis algorithm).
10th column: "ConditionB" Concentration - mean (log) reads across all replicates of "ConditionB" condition. (normalization using the respective analysis algorithm)
11th column: Fold difference - mean fold difference of binding affinity of group 1 over group 2 (Concentration ConditionA - Concentration ConditionB). Absolute value indicates magnitude of the difference, and sign indicates which one is bound with higher affinity, with a positive value indicating higher affinity in the first group
12th column: p-value calculation - statistic indicating the significance of the difference.
13th column: FDR (False Discovery Rate): adjusted p-value calculation - p-value subjected to multiple-testing.
14th column: Closest TSS (gene id) to the genomic region center.
15th column: Distance of the closest TSS to the closest genomic region center.

(2) Based on (1), a tab-delimited text report including all the significantly altered binding events present in the examined genomic regions. A threshold of FDR < 0.05 is applied.
(3) The {regions by samples} raw count matrix.
(4) The {regions by samples} normalized count matrix. Normalization is performed by using either DESeq2, or edgeR, or no normalization at all (custom normalization).
(5) A volcano plot (scatter plot), summarizing the significant differentially bound regions based on the aforementioned FDR threshold, expressed as -log10 FDR, and the magnitude of difference, expressed as log2 Fold difference. An example of such visualization is presented in Figure 42 (a).
(6) An MA plot (scatter plot), summarizing the significant differentially bound regions based on the mean (log) reads across all samples in both groups, expressed as log10 normalized counts, and the magnitude of difference, expressed as log2 Fold difference. An example of such visualization is presented in Figure 42 (b).
(7) A heat-density scatter plot, comparing ConditionA (y-axis) and ConditionB (x-axis) normalized counts on each examined genomic region, transformed into log10 space. Normalization is performed using the total alignment depth.

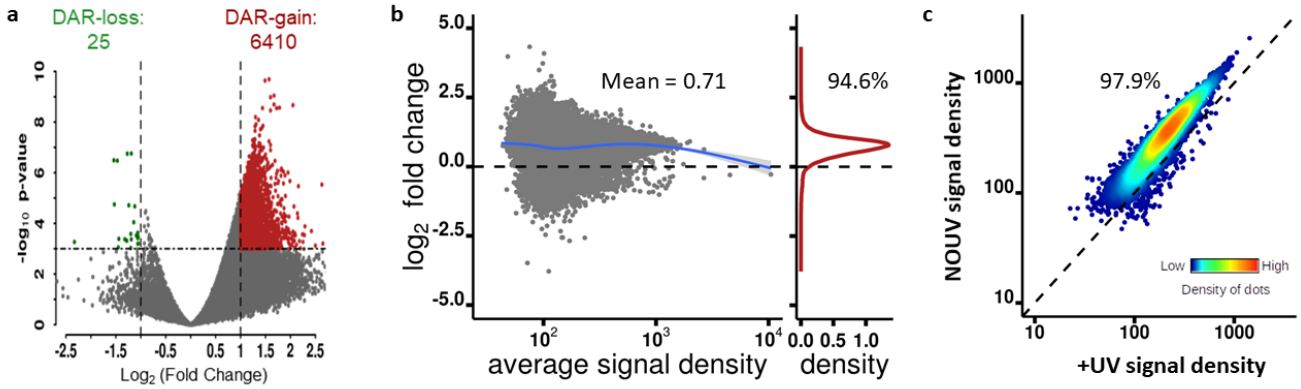


Figure 42 Differential binding/ accessibility visualization. (a) Volcano plot representing differentially accessible regions (DARs) between irradiated and non-irradiated cells. Regions with significantly increased (DAR-gain) or decreased (DAR-loss) accessibility are depicted in red and green, respectively. (b) Left panel: MA plot showing the individual (grey dots) and average (blue line) FC (Log2 FC) in ATAC read density at ATAC-seq peaks, between +UV and NO UV, as a function of the average (from replicates) ATAC-seq read density in NO UV. Right panel: Percentage of peaks with increased FC (Log2 FC > 0) is indicated on a kernel density plot. (c) Heat-density scatter plot comparing ATAC-seq read density before and after UV at all accessible regions (ARs - ATAC-seq peaks).

4.1.3.5 Peak annotation analysis

After the completion of the peak-calling and differential binding analysis modules, the regional distribution of these loci based on genome annotations is of particular interest. This information is very important, as biological hypotheses can be declared about persistent occurrences of DNA binding factors in specific regulatory areas. For this reason, genomic annotations of binding events are created. Annotations are generated using RefSeq gene-bodies and promoter annotations, as well as FANTOM5 enhancers. This creates the following categories:
(1) Extended promoter regions: 4 kb regions, centered to RefSeq TSSs.
(2) Genic enhancers: FANTOM5 enhancers, localized in a RefSeq gene region, without overlapping the respective extended promoter region.
(3) Genic regions: regions included in RefSeq genes, but not in an extended promoter region or a genic enhancer.
(4) Intergenic enhancers: FANTOM5 enhancers, localized between RefSeq genes, without overlapping an extended promoter region.
(5) Intergenic: All the regions not included in the categories (1) - (4).
Peak annotations are also performed with respect to roadmap chromHMM chromatin states (see introduction). Peaks are centered and examined for overlap with the chromatin state regions, and each peak is assigned to a unique state. The annotations are summarized as a percentage of the total annotations, using (a) a pie chart and (b) a radar plot of annotation fractions (Figure 43).

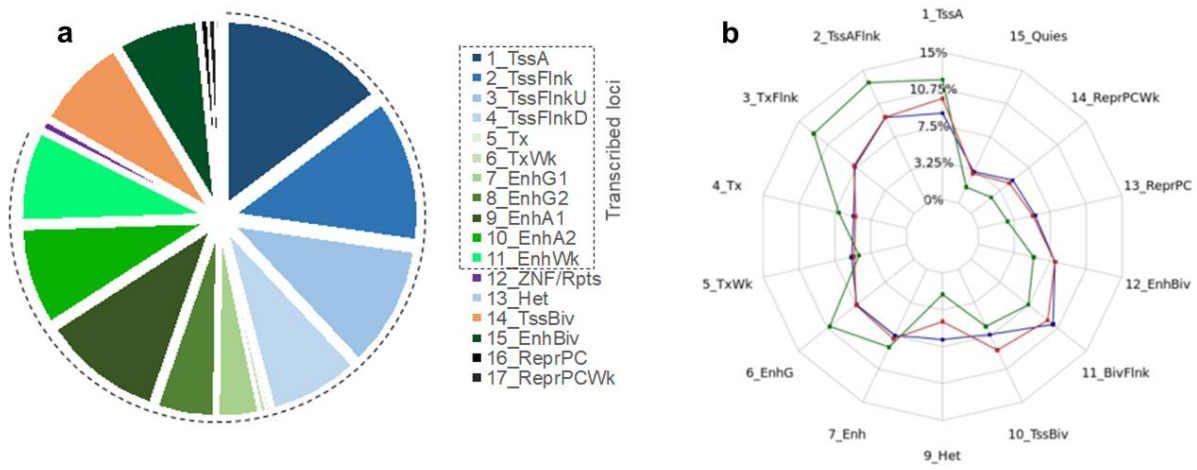


Figure 43 Classification of (a) ATAC-seq peaks (ARs) and (b) aniFOUND-seq and XR-seq repair signal (materials and methods) according to NHDF roadmap chromHMM annotation. The dashed line represents active regulatory loci.

4.1.3.6 Motif enrichment analysis

Peak regions contain valuable regional information that defines its functional dynamics, as they represent a snapshot of DNA binding events that potentially regulate transcription through promoter/enhancer interactions. Another layer of functional information is the sequence content included in these regions, and in particular motif sequences that correspond to transcription factors and repressors, and indicate potential binding in these genomic loci. Motif enrichment analysis is a way to validate the efficiency of a particular TF ChIP-seq, by identifying its corresponding binding motif in the called peaks, but also discovering multiple motifs that imply factor colocalization in potential protein complexes. The particular analysis is been applied using HOMER (Heinz et al., 2010)(Figure 44) or/and i-cisTarget (Herrmann et al., 2012; Imrichová et al., 2015) tools.

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-1015	-2.338e+03	44.05%	10.63%	118.8bp (155.0bp)	BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer(0.997) More Information | Similar Motifs Found	motif file (matrix)
2		1e-90	-2.079e+02	31.26%	20.51%	136.0bp (150.3bp)	ERG(ETS)/VCaP-ERG-ChIP-Seq(GSE14097)/Homer(0.931) More Information | Similar Motifs Found	motif file (matrix)
3		1e-89	-2.064e+02	33.99%	22.92%	133.0bp (147.3bp)	Egr2(Zf)/Thymocytes-Egr2-ChIP-Seq(GSE34254)/Homer(0.821) More Information | Similar Motifs Found	motif file (matrix)
4		1e-88	-2.042e+02	12.54%	5.84%	131.1bp (148.4bp)	CREB5(bZIP)/LNCaP-CREB5.V5-ChIP-Seq(GSE137775)/Homer(0.976) More Information | Similar Motifs Found	motif file (matrix)
5		1e-81	-1.875e+02	46.05%	34.43%	139.3bp (155.3bp)	POL002.1_INR/Jaspar(0.737) More Information | Similar Motifs Found	motif file (matrix)

Figure 44 Homer motif enrichment analysis results on VH10 ATAC-seq (see materials and methods) differentially accessible regions, with significant gains in accessibility (p-value<0.001) upon UVC stress (+UV 2 h).

4.1.3.7 Heatmaps, average profiles, boxplots and genomic tracks generation

Visualization of the alignment signal across the genome annotation is particularly useful, as it enables the examination of the binding profiles across the datasets in multiple genomic regions, allows the detection of global or partial differences between biological states, and reveals the quality of the ChIP between replicates, as also of the total experimental design. For all the types of visualization, read counting is performed using the filtered alignments generated in section 4.1.3.1 coupled with several genomic region sets as references. For this step, summarizeOverlaps is used with the “inter.feature=TRUE” and “ignore.strand=TRUE” options enabled. Also, reads are centered before applying read counting.

Four main types of visualizations are generated by this analysis module:

(1) Average profiles of read density: This type of visualization allows the examination of the average distribution of the NGS signal along a genomic region set, and reveals the “shape” of the binding, as also the enrichment level along this shape. For this purpose, RefSeq TSSs, TTSs, gene bodies, as well as peak summits (in case of MACS2 peaks) or peak centers (in case of epic2) are used:

(a) In the case of gene bodies, all regions are initially extended to a predefined length (2kb). The inner part of each region is divided to a total of 160 genomic segments (bins) of the same length, while the flanking regions to a total of 20 bins each, creating a 200-bin vector for each gene ($b_{i1}, b_{i2}, \dots, b_{i200}$), where i is the i-th element.

For each examined BAM file, read overlaps are generated for each bin, reverse-strand references are flipped, read depth normalization is applied (multiplication by 1,000,000/alignment depth), and the mean of counts of each bin position is calculated to generate a 200-length vector of average counts for each dataset.

Additional plots of gene bodies are also generated, using a gene length limit. In particular, genes with length over 10 kb, 20 kb, 40 kb, 60 kb, and 100 kb are extracted and limited to a total length of 10 kb, 20 kb, 40 kb, 60 kb, and 100 kb respectively, in order to create constant length references. This set-up can result in more realistic illustrations of the signal distribution, since the variable gene length effect is eliminated. Genes are then treated as described above to create one plot per gene set.

(b) In the case of TSSs, TTSs, peak summits and peak centers, regions are extended to 1 kb, 2 kb, and 5 kb, and binned to a total of 200 segments, and counting and averaging are performed as described in (a).

Representative examples of such visualizations are included in Figure 45.

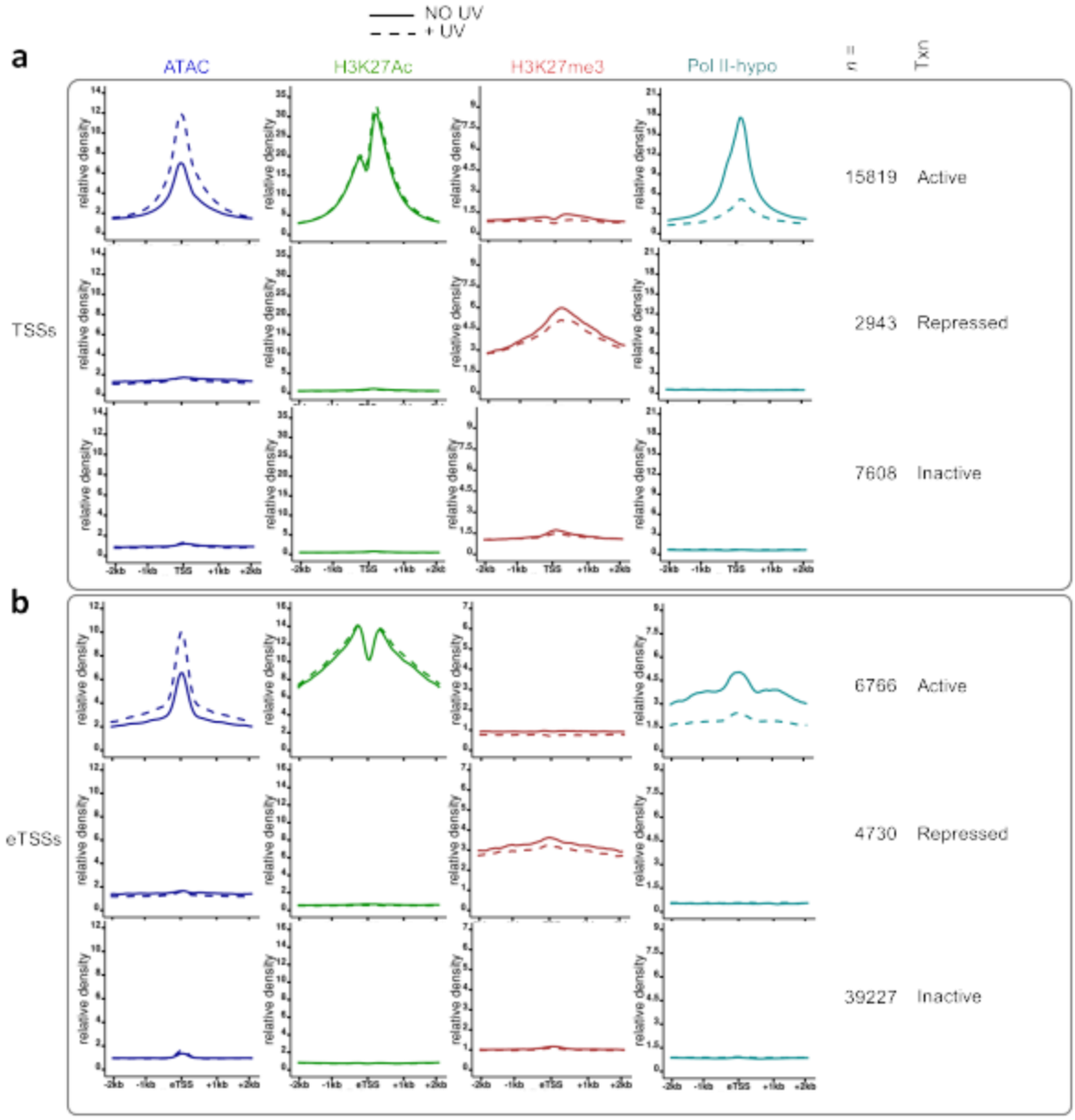


Figure 45 Average profile plots illustrating the read densities of ATAC-seq, H3K27ac, H3K27me3, and RNAPII-hypo ChIP-seq datasets before (NO UV, solid line), and after UV (+UV, dashed line), along active, inactive, and repressed transcription start sites (TSSs, a) and enhancer RNAs (eTSSs, b).

(2) Heatmaps of read counts: This type of visualization allows the examination of the global, or per-cluster distribution of the NGS signal in a set of genomic regions of interest, in a region-per-region resolution. Regions are sorted in ascending order based on their RPKM value, and bin-count vectors are generated as described in the above. The resulting {regions x bins} count matrix is used to generate heatmaps of read densities using the R *pheatmap* function and complexHeatmap R package (Gu et al., 2016). Bin-counts are also clustered using k-means clustering with a predefined k equals to 5, and/or hierarchical clustering based on euclidean distances that rearrange the {regions x bins} count matrix before generating additional heatmaps plots.

Representative examples of such visualizations are included in Figure 50.

(3) Boxplots of total read density: Total reads per examined region-set can define per sample read-count distributions that are informative about global differences/ similarities between signal enrichment in particular genomic clusters, between biological conditions and samples. For each genomic region-set described in 4.1.3.7 section, per sample read-count vectors are generated using summarizeOverlaps, counts are converted to RPKM values and boxplots are generated using R ggplot.
Representative examples of such visualizations are illustrated in Figure 46.

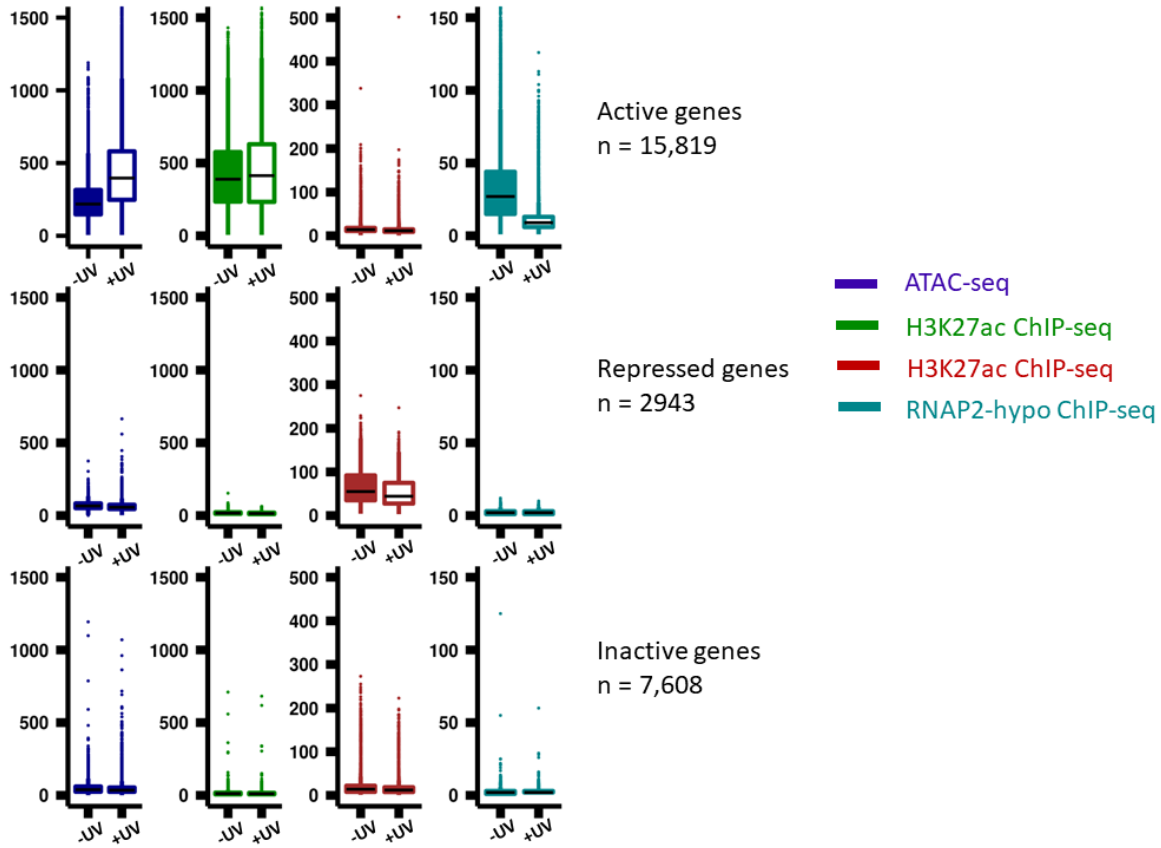


Figure 46 Boxplots of NGS signal read-counts for different assays at active, repressed and inactive regions, comparing non-irradiated (-UV) and irradiated (+UV) cells.

(4) Genome browser tracks: Genome browsers are powerful genomic exploration tools that can host NGS data using several file-types (BAM, BED, bedGraph, bigWig), that allow the identification of interesting signal patterns along the genome, and comparison with publicly available tracks. Exploration can be applied either in the context of specific genomic locations such as promoters, genes, enhancers, super-enhancer or even in very large genomic areas (chromosomes), as another quality control of the NGS signal distribution and the experimental set-up. Although genome browsers are compatible with several file types, the more efficient file types are bigWig for alignment files and bigBed for region files, because of their relatively small size.

In this analysis step, all filtered, sorted and indexed BAM files are processed with *deeptools bamCoverage* with RPKM value transformation enabled, to generate genome browser compatible bigWig files. Additionally, for every BED merged peak-set file, a bigBed file is also generated. The resulting UCSC compatible track lines have the following structure (bigWig and bigBed respectively):

track type=bigWig name=$Sample_name description="$Sample_name" color=$R,$G,$B maxHeightPixels=128:64:16 visibility=full autoScale=on bigDataUrl=$URL/$Sample_name.merged_reps.dedup.bw

track type=bigBed name=$Sample_name_peaks description="$Sample_name peaks" bigDataUrl=$URL/$Sample_name.merged_reps.dedup.TF_peaks.bb

where $Sample_name refers to the sample name, $URL is specified as the output folder, and $R,$G and $B correspond to Red, Green and Blue in RGB color code. Colors are generated using Colorbrewer (Brewer et al., 2003) and RGB transformations, using R *col2rgb* function. ColorBrewer colors are generated using the R diverging color palette “Dark2”. If replicates are present, they are assigned the same color. Regarding the generated file names, “merged” refers to merged replicates, while “dedup” refers to deduplicated alignments (see previous analysis steps).

4.1.4 Nascent RNA-seq (nRNA-seq) analysis pipeline

4.1.4.1 Short-read mapping and alignment filtering

In nascent RNA sequencing protocols like nRNA-seq, TT-seq (Schwalb et al., 2016), PRO-seq (Mahat et al., 2016), NET-seq (Mayer et al., 2015) et al, unlike ChIP-seq, libraries are generated using cDNA, and therefore different analysis methodologies are applied. In the alignment step, for each filtered FASTQ file, a first alignment run is applied against the ribosomal DNA repeat unit of the reference organism, in order to filter out ribosomal reads which is the main source of contamination in nRNA-seq protocols. In this analysis module, hisat2 (D. Kim et al., 2015) with default parameters is applied by setting a ribosomal DNA repeat unit reference index with disabled splicing-aware mapping. Unmapped reads are extracted from the BAM files and converted to FASTQ files using *samtools view -f 0x4 -b* | *samtools fastq* command, and a second round of alignment is performed. This time, to eliminate a fraction of mRNA reads, which is the second most common source of contamination in nRNA-seq protocols, splicing-aware mapping is performed. Hisat2 is run using predefined splice sites and a reference genome index. Using the sixth column of the SAM file and the flag marker “N”, all spliced alignments are excluded and reported in a separate BAM file, and uniquely aligned reads are detected using the ZS:i: SAM flag. Alignment filtering and replicate merging is applied as described in the ChIP-seq analysis module, omitting the deduplication step (only duplicate marking is performed), unless if the reads are paired-end. In the case of strand-specific protocols, strand-aware mapping is performed, and additional strand-aware alignment files are generated.

4.1.4.2 Transcription unit identification

This mode is equivalent to ChIP-seq peak-calling (see section 4.1.3.2). The purpose of this analysis module is to identify transcribed units across the genome, based on nascent RNA signal enrichment. These elements include actively transcribed genes, enhancers and super enhancers. This analysis module is applied on each filtered BAM file separately using an HMM-based algorithm which is described in detail in chapter 4.2.

4.1.4.3 Alignment similarity analysis

This analysis module is applied as described in the ChIP-seq analysis pipeline in section 4.1.3.3.

4.1.4.4 Fraction of reads in peak (FRIP)

This analysis module is applied as described in the ChIP-seq analysis pipeline (section 4.1.3.3). Comparing the ChIP-seq FRIP results (Figure 41) with the corresponding nRNA-seq results (Figure 47), it is obvious that in the nRNA-seq datasets, unlike the ChIP-seq datasets, the vast majority of the NGS signal is located within specific areas (genes in the particular example) indicating that in nRNA-seq the background signal is minimal.

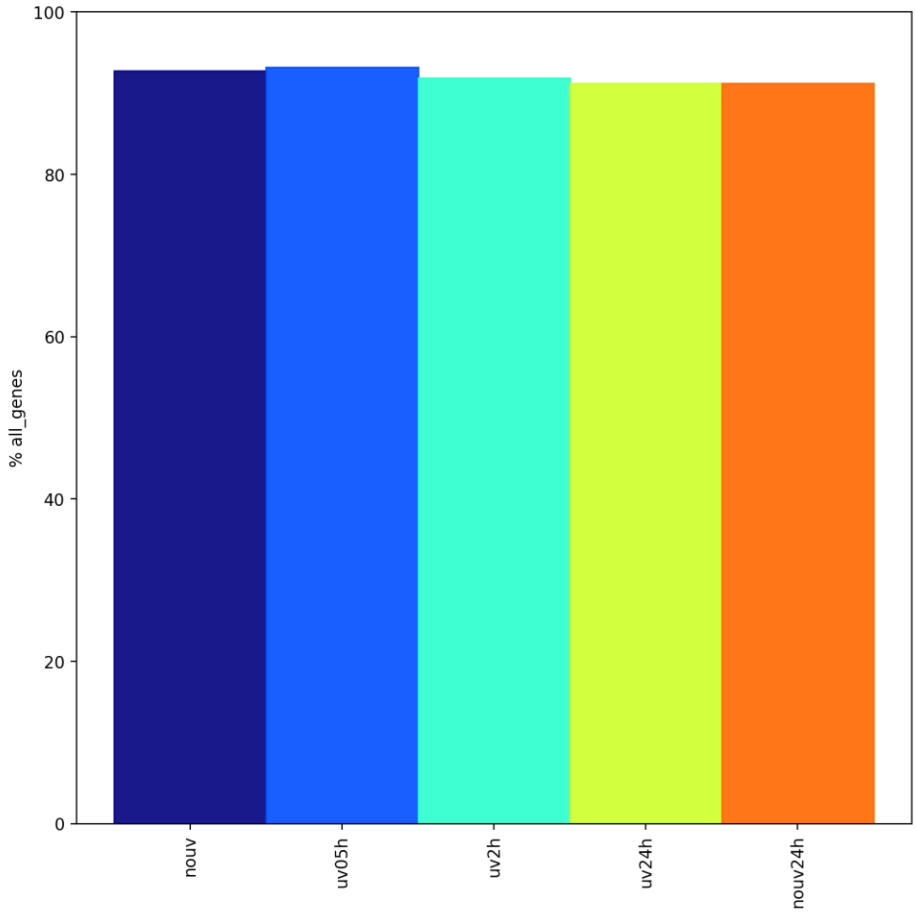


Figure 47 FRIP plot visualization for a set of VH10 nRNA-seq datasets described in section 2.10.5. (1) nouv, where cell are in normal conditions, (2) uv05h, where cell are recovering for 30 minutes after irradiation with UVC (15 $\boldsymbol{J/m^2}$), (3) uv2h, where cell are recovering for 2 hours after irradiation with UVC (15 $\boldsymbol{J/m^2}$), (4) uv24h, where cell are recovering for 24 hours after irradiation with UVC (15 $\boldsymbol{J/m^2}$), and (5) nouv24h, 24 hours after the nouv pull-down. As a reference feature set, RefSeq transcripts were used.

4.1.4.5 Differential expression analysis

Differential expression analysis is applied when seeking quantitative changes in gene expression levels between different biological conditions and samples, measured by RNA-seq. In the case of nRNA-seq methods, changes between nascent transcription levels. The particular analysis is performed using pairwise comparisons "ConditionA_vs_ConditionB", where ConditionB should represent the denominator of each comparison. Analysis for differential nascent RNA expression is applied using DESeq2 or/and edgeR packages, with the RefSeq gene-set as a reference, and counting reads only in the intron sequences to avoid overcounting processed RNA contaminants. If batch effects or any additional confounding factors are present in the experimental set-up, they can be modeled in the respective design matrix included in the comparison formula of each applied tool (Love et al., 2014; M. D. Robinson et al., 2009). Notably, confounding factors can be “removed” by the count matrix using the R function

removebatcheffect from limma package (Ritchie et al., 2015), and the corrected count matrix can be used for further analysis and visualization purposes (clustering, heatmaps, average profiles etc). The particular analysis module generates a set of outputs depicted below:

(1) A tab-delimited text report of all the examined regions with column-wise information structure as follows:

1st column: Chromosome of the examined genomic region.
2nd column: Starting base position of the examined genomic region.
3rd column: End base position of the examined genomic region.
4th column: Genomic region id (gene id as defined in the fourth column of the analyzed reference. If not present “chr:start-end” will be assigned, where “chr”,”start” and “end” refer to the 1st, 2nd, and 3rd column respectively).
5th column: GC content of the sequence content of the examined genomic region (in the case of a custom annotation, 0 is assigned).
6th column: Strand orientation of the examined genomic region (if it is absent in the custom annotation, * will be assigned).
7th column: Genomic region name (in the case of a custom annotation, gene id will be repeated).
8th column: Biotype of the examined genomic region (in the case of a custom annotation, “custom” will be assigned).
9th column: Length of the examined genomic region (in the case of a custom annotation, length will be calculated by the gene coordinates).
10th column: “ConditionA” average RPKM of normalized read counts.
11th column: “ConditionB” average RPKM of normalized read counts.
12th column: Concentration - mean (log) reads across all replicates in both groups (normalization using the respective analysis algorithm).
13th column: "ConditionA" Concentration - mean (log) normalized reads across all samples of "ConditionA" condition (normalization using the respective analysis algorithm).
14th column: "ConditionB" Concentration - mean (log) normalized reads across all samples of "ConditionB" condition (normalization using the respective analysis algorithm).
15th column: Fold difference - mean fold difference of expression enrichment of group 1 over group 2 (Concentration ConditionA - Concentration ConditionB). Absolute value indicates magnitude of the difference, and sign indicates which one is expressed higher, with a positive value indicating higher expression in the first group.
16th column: p-value calculation - statistic indicating the significance of the difference.
17th column: FDR (False Discovery Rate): adjusted p-value calculation - p-value subjected to multiple-testing correction.

(2) Based on (1), a tab-delimited text report including all the significantly altered differential expressed genes. A threshold of FDR < 0.05 is applied.
(3) The {genes by samples} raw count matrix.
(4) The {genes by samples} normalized count matrix. Normalization is performed by using either DESeq2, or edgeR.

(5) A volcano plot, summarizing the significant differentially expressed regions based on the significance level, expressed as -log10 FDR, and the magnitude of difference, expressed as log2 Fold difference.
(6) An MA plot, summarizing the significant differentially expressed genes based on the mean (log) reads across all samples in both groups, expressed as log10 normalized counts, and the magnitude of difference, expressed as log2 Fold difference.

4.1.4.6 Heatmaps, average profiles, boxplots and genomic tracks generation

This analysis module is applied as described in the ChIP-seq analysis pipeline in section 4.1.3.7.

4.1.5 ATAC-seq analysis pipeline

4.1.5.1 Short-read mapping and alignment filtering

For every filtered FASTQ file, mapping, alignment filtering, alignment deduplication and replicate merging is applied as described in the ChIP-seq analysis module (section 4.1.3.1). The output of this analysis module includes the individual and merged (if replicates are present) BAM files of each analyzed sample.

4.1.5.2 Peak calling analysis

This analysis module is applied as described in the ChIP-seq analysis pipeline in section 4.1.3.2.

4.1.5.3 Alignment similarity analysis

This analysis module is applied as described in the ChIP-seq analysis pipeline in section 4.1.3.3.

Differential accessibility analysis is applied when a scientist wants to identify alterations in chromatin accessibility between different biological conditions, as measured by ATAC-seq, DNAseq-seq, MNAse-seq or FAIRE-seq. The particular analysis module is designed for ATAC-seq protocols, and is applied as described in the ChIP-seq analysis pipeline in section 4.1.3.4.

4.1.5.4 Peak annotation analysis

This analysis module is applied as described in the ChIP-seq analysis pipeline in section 4.1.3.5.

4.1.5.5 Motif enrichment analysis

This analysis module is applied as described in the ChIP-seq analysis pipeline in section 4.1.3.6.

4.1.5.6 Heatmaps, average profiles and genomic tracks generation

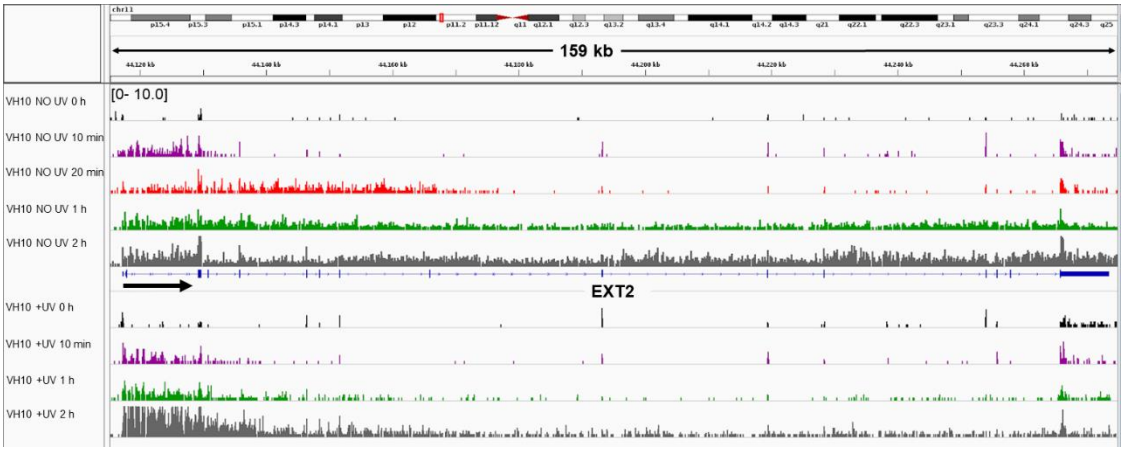
This analysis module is applied as described in the ChIP-seq analysis pipeline in section 4.1.3.7.

4.2 Genome-wide identification of de novo elongation waves

The particular algorithm is designed for estimating the transcription elongation wave-end of actively transcribed RNAP2 molecules during normal cellular conditions or during early genotoxic stress conditions, using primarily pre-DRB nRNA-seq datasets (see materials and methods). As an example dataset, human VH10 pre-DRB nRNA-seq experiments were used in NO UV +0 min, +10 min, +1 h and +2 h conditions, and +UV 0 h, +10 min, +1 h and 2 h, as also CSB pre-DRB nRNA-seq experiments in NO UV +0 h, +10 min, +20 min, +1 h and +2 h conditions, and +UV +0 h, +10 min, +20 min, and 2 h (materials and methods, section 2.10.6). For the prediction of the transcription wave front in each examined dataset, a Hidden Markov Model (HMM, see section 1.10) was implemented and applied as described below.

4.2.1 Quality control, prefiltering and read mapping

FASTQ files were processed for quality trimming and adapter clipping as described in the section 4.1.2. In order to exclude all the rRNA reads that comprise the major source of RNA contamination in these kind of data, high-quality FASTQ files were first aligned against the human ribosomal DNA complete repeating unit (U13369.1), keeping all the unmapped reads that were in turn aligned to the UCSC hg19 reference genome, using the module described in section 4.1.4.1. Only primary, and high-quality alignments were retained, and duplicated alignments were discarded, and genome browser tracks were generated accordingly (Figure 48).



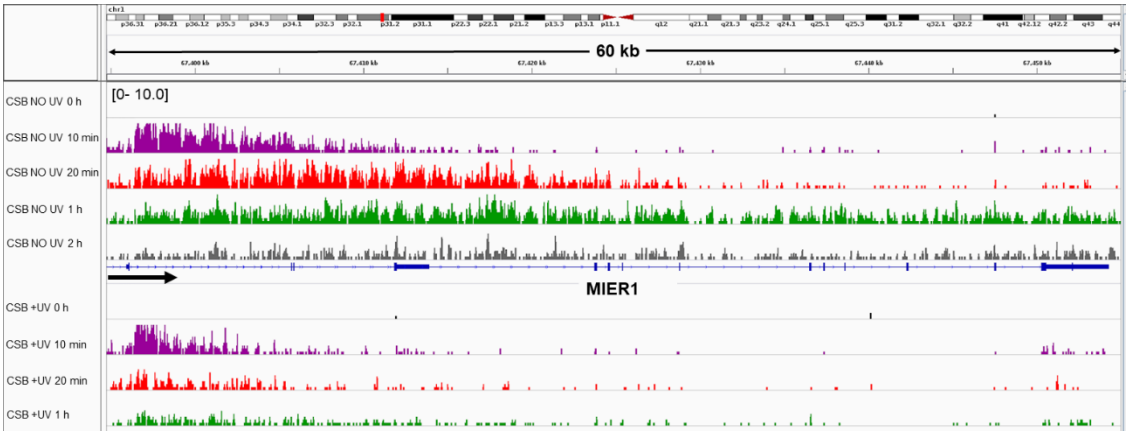


Figure 48 IGV genome browser tracks of VH10 and CSB pre-DRB and nRNA-seq datasets. Patterns of transcription elongation wave progression are observed in both cell types during transcription recovery from DRB

4.2.2 Genome annotation reconstruction

To define a reference gene set, 54,669 transcripts from RefSeq v.72 were downloaded. Initially, only protein coding and lncRNA genes were selected, two gene categories that are both transcribed by RNAPII (Bunch, 2017). These two biotypes will be referred to as “mRNAs” during the rest of this study. TSSs were clustered using a radius of 500 bp, and the longest transcript was kept for further analysis. Furthermore, TSS pairs with a distance less than 1kb were eliminated to avoid overlapping TSS flanking regions during read counting. This resulted in a total of 19,775 genes.

4.2.3 Transcriptional activity determination

To determine the activity status of each transcript, VH10 and CSB NO UV 2 h filtered BAM files were summarized at each transcript, excluding annotated exonic regions in order to minimize the effect of mRNA contamination. The first kilobase of each transcript was also omitted, in order to better gauge the density of polymerase that actively elongates through the gene-body, by avoiding the over-counting from PPP (Jonkers et al., 2014). Gene-counts were transformed to RPKM values and kernel density plots of $log_2 RPKM$ values for each dataset were plotted (Figure 49).

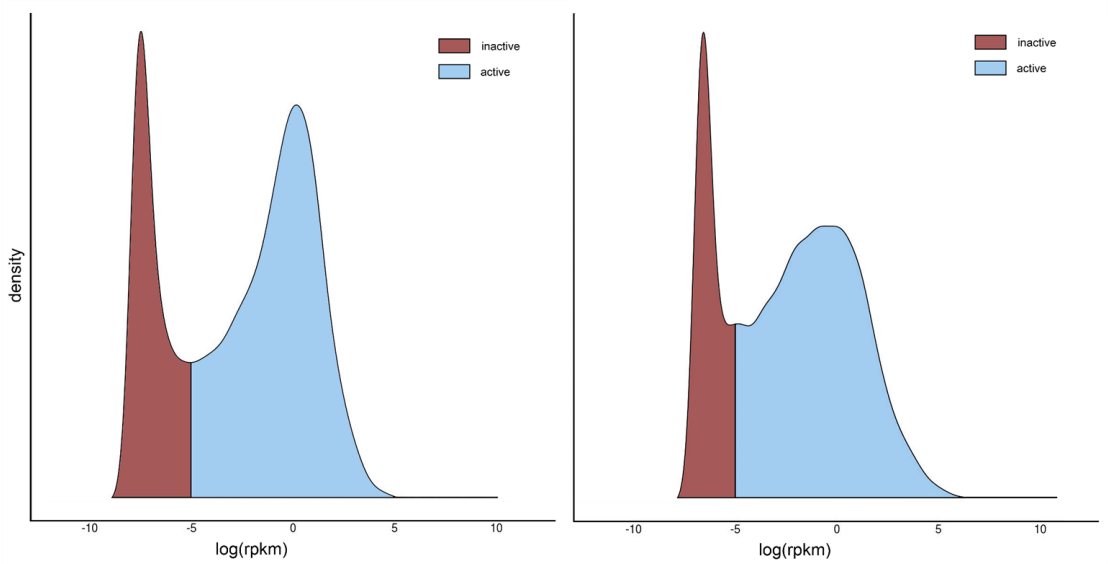


Figure 49 Kernel density plots of VH10 (left) and CSB (right) nRNA-seq NO UV 2h $\boldsymbol{log_2 RPKM}$counts at RefSeq transcripts reveal two main transcript populations for each cell line.

The resulting bimodal RPKM distributions denoted two main transcript populations for each cell line: active and inactive elements. The bisection point was commonly set to $RPKM = 0.03$ resulting in 12,435 active transcripts for VH10 cells, and 12,846 active transcripts for CSB cells. Taking into consideration that a wide range of transcription elongation rates are reported in different studies, from 1–6 kilobases per minute (kb/min) (Ardehali & Lis, 2009; Darzacq et al., 2007; Singh & Padgett, 2009), in order to gain robust results at all the processed datasets, only transcripts over 60 kb were considered for the rest of this analysis, as also the intersection of active genes between cell lines, resulting in a total of 3,048 commonly transcribed elements. As a negative gene set, 2,004 commonly inactive genes with a length over 60 kb were selected.

4.2.4 Data visualization

To generate heatmaps and average profiles of nRNA-seq signal the nRNA-seq analysis module was applied (section 4.1.4.6), using the 3,048 commonly transcribed genes (VH10 and CSB, see above) as a reference set (Figures 50 and 51).

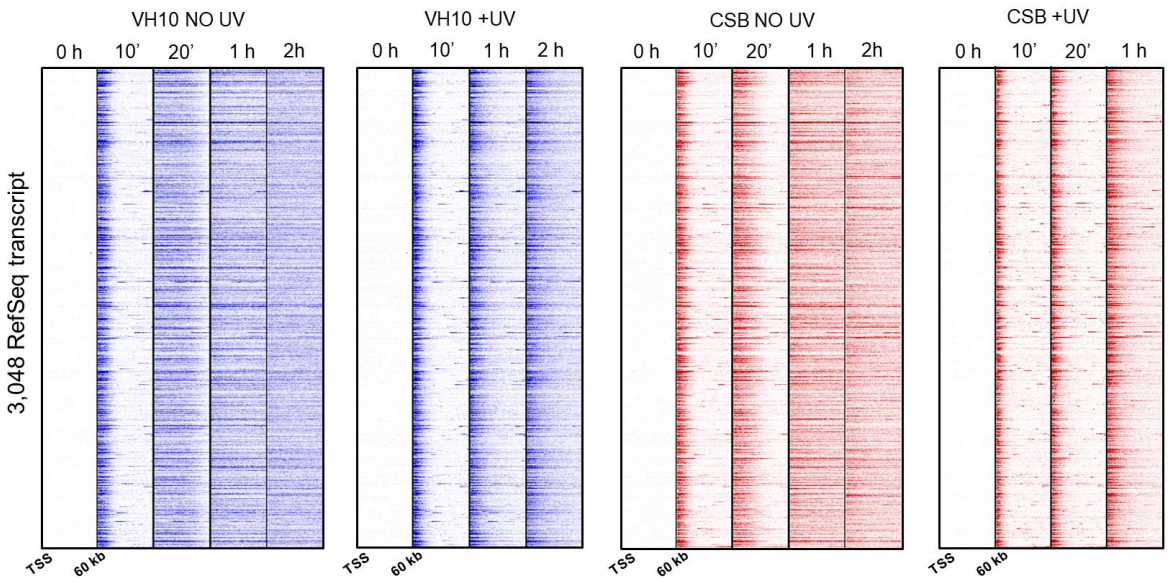


Figure 50 Heatmaps of nRNA-seq read-counts along actively transcribed RefSeq transcripts with length over 60 kb upon DRB removal. White color scale denotes low read density, while blue (VH10) and red (CSB) color scales denote high read density.

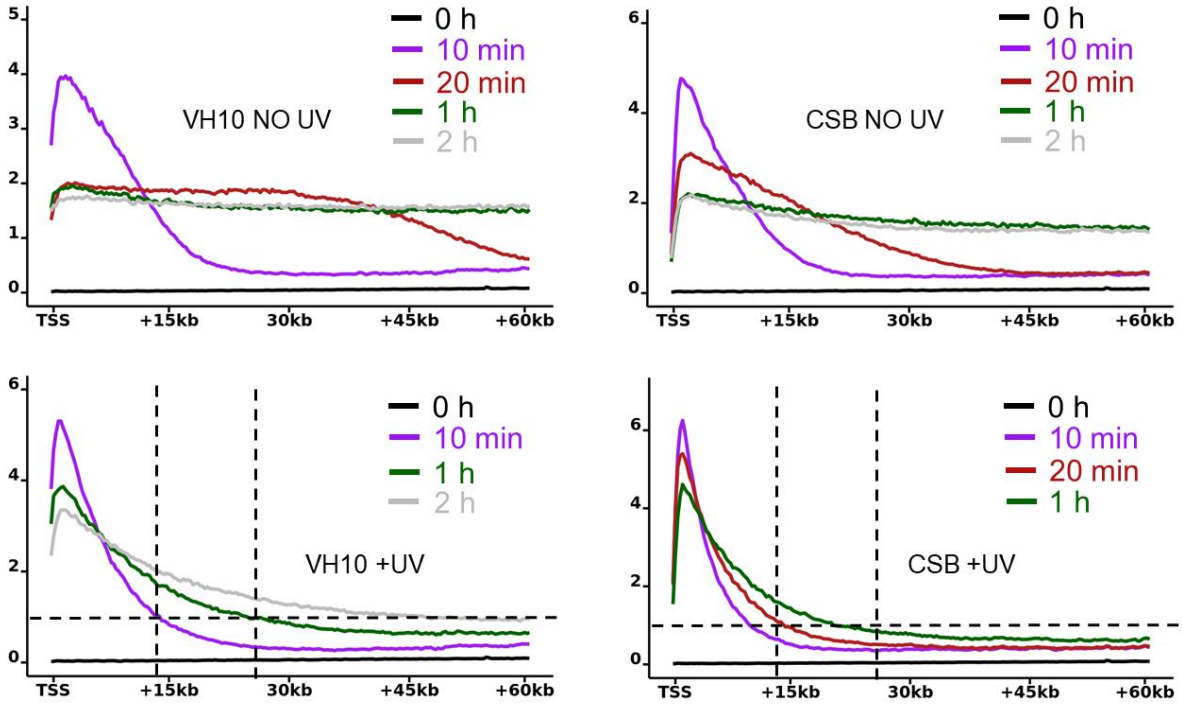


Figure 51 Average profiles of nRNA-seq read-counts along actively transcribed RefSeq transcripts with length over 60 kb after removal of DRB. Comparison of VH10 and CSB No UV and +UV conditions by setting an arbitrary threshold (horizontal dashed line, average normalized counts = 1) reveals patterns of elongation wave inhibition to a greater extent at TC-NER deficient cells (compare vertical dashed lines between VH10 and CSB).

Visual inspection of the generated graphs revealed patterns of elongation wave progression in all the analyzed datasets after removal of DRB (see materials and methods 2.10.6). In NO UV conditions, RNAPII elongates at a much higher rate than the +UV conditions for both datasets. This UV-induced deceleration of RNAPII progression is propably due to the stalling of RNAPII upon encountering DNA-damages. Additionally, in VH10 +UV condition, patterns of faster transcription recovery are gradually detected in contrast to CSB +UV where RNAPII molecules appear to remain stalled at DNA lesions for a longer time period due to the non-functional TC-NER mechanism (Figure 50, 10 minutes and 60 minutes +UV, and Figure 51 10 minutes and 60 minutes +UV).

4.2.5 Data preparation

Both active and inactive annotations were split into genomic bins, in order to create a bin-vector for each transcript, for each examined dataset, that will be used for NGS signal counting, data transformation, and normalization (section 4.1.4). For all the NO UV 10 min datasets, a bin size of 250 bp was used, while for NO UV 20 min, NO UV 1 h and NO UV 2 h, bin sizes of 500 bp, 1 kb and 2 kb were applied respectively. For +UV 10 min datasets, a bin size of 250 bp was used, while for the remaining +UV datasets, a bin size of 500 bp was applied. For the NO UV and +UV 0 h samples, bin-vectors of 250 bp, 500 bp, 1 kb, and 2 kb were also generated accordingly.

Read alignments in each examined dataset were extended to a 200 bp fragment length (in order to reach the average fragment length of the libraries), and only the 3' ends were considered for counting. For each examined dataset, the resulting 3' end points were examined for genomic overlap with the respective active and inactive bin-vectors to generate bin-count vectors.
For each bin-count feature, a pseudocount was added, and all bin-count vectors were divided by the corresponding 0 h bin-count vector, to eliminate the effect of the nRNA-seq background signal.
In order to eliminate the effect of mRNA contamination in each of the examined nRNA-seq experiments, for each genomic bin with an exon coverage larger than 20%, the corresponding normalized count was considered as a missing value, and all missing values were replaced by the outputs of a cubic splines interpolation, which is applied along the entire bin-count vector (R *smooth.spline* function). Consequently, to remove the PPP enrichment bias of the specific NGS protocol, for each bin-count vector, all bin-counts were divided by the average of the first five bin-counts, as an internal normalization.
Finally, all normalized bin-count ratios were discretized, using a range from 0.0 to 1.0, with a step size of 0.05, resulting in a maximum of 20 possible values for each vector, that represent the emission states of each HMM (see section 1.10). All elements that were annotated at the reverse strand were flipped in order to keep the same count-vector structure for all the examined elements.

<u>Training set</u>

To generate a robust training set for HMM parameter estimation, that includes all the instances of transcriptional activity, active transcripts of varying expression levels were included, as also

non-transcribed elements. Specifically, all active transcripts were grouped to 3 expression clusters based on their RPKM value, and a random choice of 450 highly expressed, 175 mediumly expressed, and 175 lowly expressed transcripts was applied. Also, a random choice of 250 inactive transcripts was added to the active training list, resulting in a training set of 950 bin-counts.

4.2.6 HMM set-up and training

For each examined dataset, an individual HMM was implemented, in order to predict which of the examined bins are engaged by the de novo RNAPII elongation wave, and which bins are not reached yet by RNAPII molecules (most probable hidden state path). To design these models, the hmm.discnp R package was used (https://cran.r-project.org/package=hmm.discnp), where each $HMM = (\pi, A, B)$ consists of a set of hidden states called "RNAPII engaged" (E) and "RNAP2 free" (F), $H = \{E, F\}$, with initial state probabilities $\pi = [0.8, 0.2]$ and hidden state transition probabilities $A = [0.95, 0.05;\ 0.05\ 0.95]$, and a total of 20 observed states $O = \{0.00, 0.05, \ldots, 1.00\}$ for each training bin-count feature, following any of the finite discrete distributions depending on the state of the Markov chain (https://cran.r-project.org/web/packages/hmm.discnp/index.html), specified non-parametrically, $Rho = [rho_{ij}], rho_{ij} = P(Y = y \mid S = y)$. Figure 52 summarizes the aforementioned design.

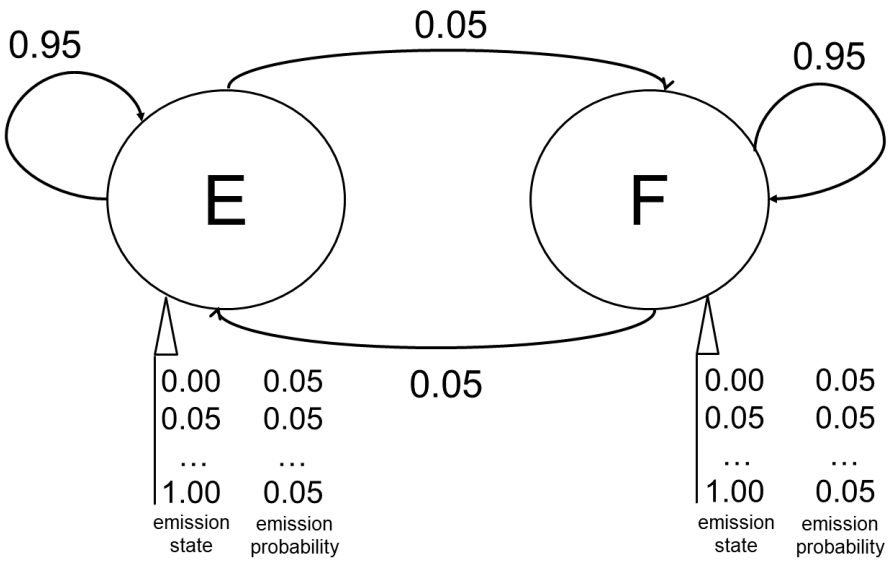
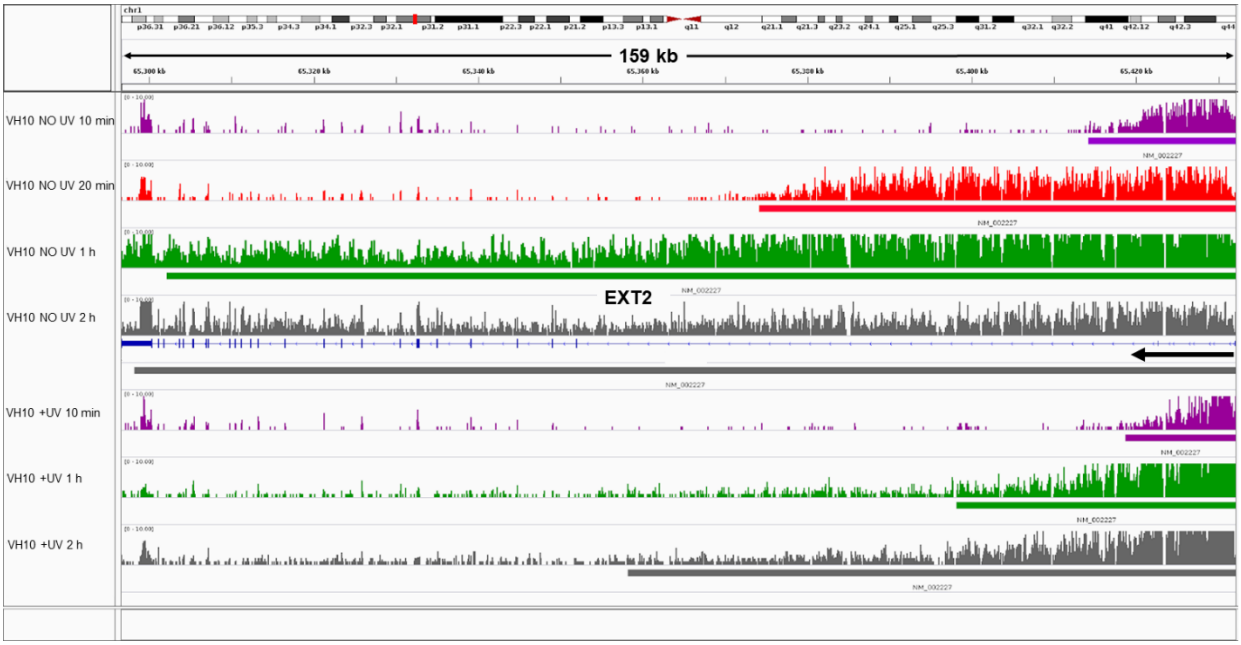


Figure 52 Automaton representation of the initially defined HMMs, with two hidden states: E (RNAPII engaged) and F (RNAP2 free). Before parameter estimation, emission probabilities were determined using the uniform distribution.

Emission probabilities B were estimated using the Baum-Welch algorithm (section 1.10).

4.2.7 HMM predictions

Predictions were applied at all 3,048 VH10-CSB commonly transcribed bin-counts, for each dataset, using the HMMs described in section 4.2.6. Particularly, the Viterbi algorithm (section 1.10) was used to predict the most probable path of hidden states underlying each of the observation sequences (https://cran.r-project.org/web/packages/hmm.discnp/index.html). The generated Viterbi paths were then translated into genomic coordinates (BED files), by merging consecutive predictions of the “E” state to a single BED record. All transcripts including more than one “E” blocks, were discarded from the rest of the analysis as instances of unreliable predictions. Additionally, for NO UV 10 min samples, all predictions less than 5 kb or greater than 20 kb were excluded, for NO UV 20 min all predictions less than 10 kb were excluded, for VH10 NO UV 60 min all predictions less than 40 kb were excluded, for +UV 10 min all predictions over 15 kb were excluded, for CSB +UV 20 min all predictions over 40 kb were excluded, for VH10 +UV 1 h all predictions over 60 kb were excluded, and for +UV 2 h all predictions over 80 kb were excluded. Moreover, only transcripts with valid predictions for both cell lines were considered for further analysis.
Finally, BED files of valid bin predictions were converted to BIGBED files to generate genome browser compatible track lines (Figure 53).



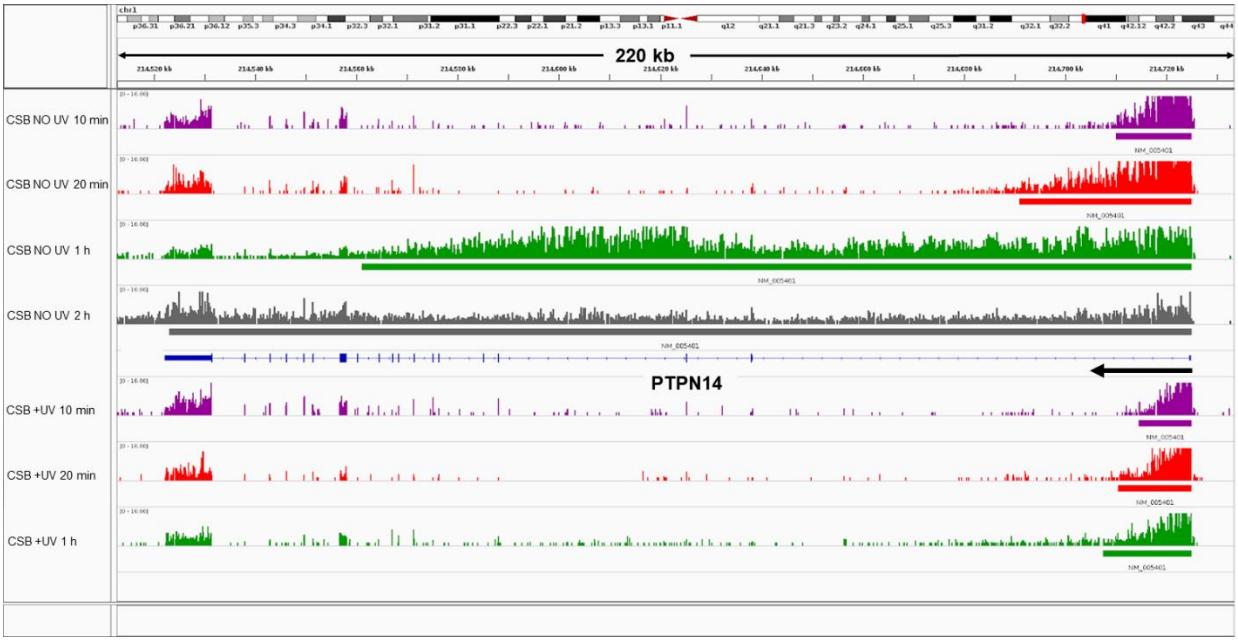


Figure 53 IGV genome browser tracks of two indicative genes, EXT2 and PTPN14 depicting the viterbi predictions for VH10 and CSB datasets respectively**.**

4.2.8 Wave front comparisons and elongation rate estimation

For each examined dataset, the average wave front position was reported. For datasets where the average prediction is over 60 kb, due to variable gene lengths in the reference transcript set, the wave front is reported to be over 60 kb (>60). For all the consecutive time point pairs where a fixed prediction was available, elongation rates were estimated as the average covered kilobases per minute for the particular time period (Figure 54).

	Time	Position (kb)	Elongation Rate (kb/min)
VH10 NOUV	0 h	NA	
	10 min	15.5	1.55
	20 min	45.7	3.02
	1 h	>60	
	2 h	>60	
CSB NOUV	0 h	NA	
	10 min	13.7	1.37
	20 min	38.5	2.48
	1 h	>60	
	2 h	>60	

	Time	Position (kb)	Elongation Rate (kb/min)
VH10 +UV	0 h	NA	
	10 min	9.3	0.9
	1 h	25.6	0.33
	2 h	33.8	0.13
CSB +UV	0 h	NA	
	10 min	6.1	0.61
	20 min	13	0.69
	1 h	20	0.18

Figure 54 Elongation wave progress prediction by HMMs. For each cell type and for each time point (whenever applicable) the average position of elongation waves in active genes of length over 60 kb are summarized. Elongation rates are calculated between consecutive time points (whenever applicable).

As expected, the HMM predictions in +UV conditions confirmed the initial observations regarding the average profiles and heatmaps of nascent RNA-seq signal (Figures 50 and 51), that in CSB cells where TC-NER is non-functional, DNA-lesions block the progression of

transcription elongation in a higher extent than in VH10 cells, since they remain essentially unrepaired. On the contrary, in VH10 cells where DNA damages are repaired at a reasonable rate by TC-NER, lesion repair allows the faster progression of the elongation wave along the recovery period (Figures 54 and 55).

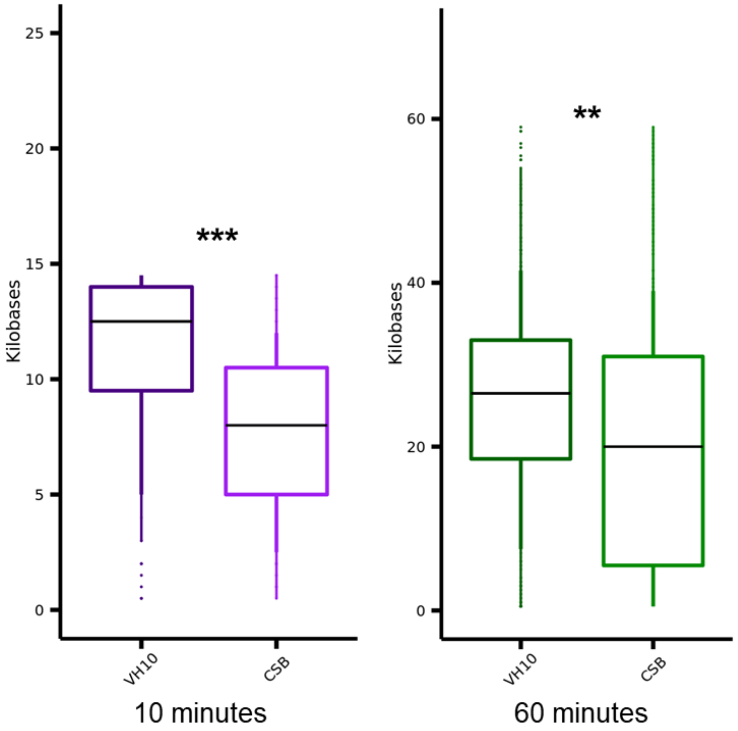


Figure 55 Box plots of elongation wave front positions as predicted by the HMMs. Comparisons between VH10 and CSB were applied using the Wilcoxon rank sum test, with p-value < 0.001 for the +UV 10 minutes comparison, and p-value = 0.0086 for the +UV 10 minutes comparison.

4.3 A computational pipeline for the study of the reorganization of transcription and chromatin alterations upon UV-induced stress.

High throughput profiles of RNAPII binding (ChIP-seq), histone modifications (ChIP-seq), nascent transcription activity (nRNA-seq) and chromatin accessibility in hTERT immortalized VH10 skin fibroblasts [see sections 2.1 and 2.2] were generated to obtain an accurate and global view into the molecular events regulating transcription re-organization and chromatin alterations in response to UVC induced stress.

4.3.1 Gene transcripts and exons annotation

RefSeq mRNAs defined in section 4.2.2 were used as a gene annotation set. To extract a consensus exon set, all overlapping transcripts were concatenated, and overlapping exons of the same strand were merged, while those encoded in opposite strands were excluded. This resulted in 164,896 RefSeq exonic regions

4.3.2 Transcript activity status determination

To study the process of transcription reorganization using the aforementioned mRNA set as a reference, transcripts were classified to 3 categories based on their transcriptional profile. These categories consist of active, poised, and inactive genes, and were classified as follows:
ChIP-seq datasets of RNAPII-ser5P, RNAPII-ser2P and RNAPII-hypo in NO UV condition (see materials and methods) were mapped against the UCSC hg19 reference genome (see section 4.1.3.1), and peak calling was performed using the "Pol2" mode described in section 4.1.3.2. Gene promoters (-250 bp to +100bp around TSS) with $\mathrm{NO\ UV\ RNAPII} - \mathrm{ser2P\ RPM}\ (\mathrm{read\ counts\ in\ the\ specified\ region} * 1{,}000{,}000/\mathrm{alignment\ depth}) > 0.7$, which overlap with any NO UV RNAPII-ser2P significant peak were characterized as active (8,954 transcripts). Active transcripts are considered affected by the transcription machinery showing patterns of transcriptional elongation. Promoters with $\mathrm{NO\ UV\ RNAPII} - \mathrm{ser5P\ RPM} > 0$ which overlap with any NO UV RNAPII-ser5P significant peak or NO UV RNAPII-hypo significant peak, but without overlapping any NO UV RNAPII-ser2P significant peak were characterized as poised (953 transcripts), while the rest of the annotations were characterized as inactive (genes that are not transcribed in the particular cell line and condition). The particular annotation characterization is summarized in Figure 56.

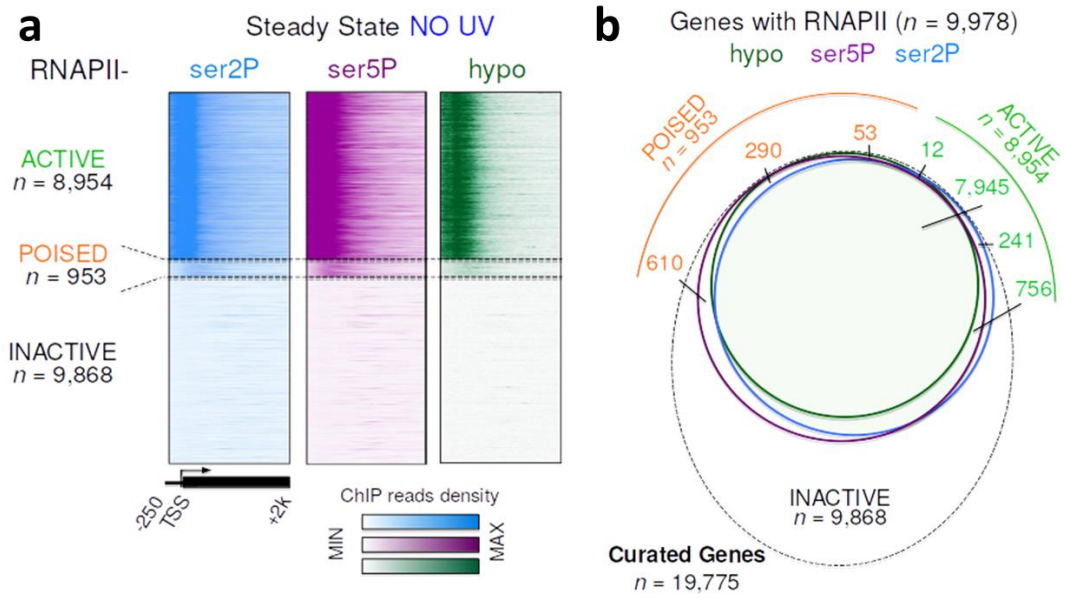


Figure 56 (a) Refseq mRNAs activity status defined by intersecting promoters with significant peaks RNAPII-hypo (green), -ser5p (purple) and -ser2P (blue) ChIP-seq. RNAPII-ser2P list was further filtered by selecting only genes with Dp > 0.7 RPM. RNAPII- ser2P containing genes were considered as active (solid green), while the union of RNAPII-ser5P and -hypo genes that did not overlap -ser2P genes were defined as poised. The rest of the transcripts were considered as inactive. (b) Heatmaps illustrating the distribution of RNAPII ChIP-seq signal (RNAPII isoforms as indicated) in NO UV condition at promoter regions (-250 bp to +2 kb relative to TSS) and gene bodies (+100 bp to +2 kb relative to TSS), categorized by transcript activity status (defined in a).

4.3.3 Transcription start site (TSS) annotation of mRNAs, enhancers and asPROMPTs

The TSS annotation was based on all known protein coding and non-coding RNA RefSeq transcripts release 86, that were retrieved from UCSC table browser, using the hg19 genome build (http://genome-euro.ucsc.edu/cgi-bin/hgTables). To classify the TSSs based on their biotypes (Table 6), the BioMart database was used (Kinsella et al., 2011), and all small non-coding RNAs and pseudogenes were excluded.

Table 6 Biotype classification of RefSeq TSSs.

antisense	1,084	processed_transcript	536
IG_V_gene	1	protein_coding	56,262
IG_V_pseudogene	1	pseudogene	904
lincRNA	3,442	sense_intronic	86
miRNA	1,478	sense_overlapping	16
misc_RNA	7	snoRNA	770
polymorphic_pseudogene	31	snRNA	22

For any Refseq gene model that contained more than one transcript, all elements were clustered together using a 50 bp TSS radius, and the longest transcript was finally reported, resulting in 30,473 TSSs.

4.3.4 mRNA TSS activity determination

All TSSs were divided into 3 categories, based on their transcriptional activity. Each element was extended to 2 kb in each direction, and the extended genomic elements were intersected with RNAPII-ser2P NO UV, H3K27ac NO UV and H3K27me3 NO UV peak sets. Regions that overlapped with RNAPII-ser2P and H3K27ac peaks, were categorized as active. Regions overlapping with H3K27me3 peaks, but not with RNAPII-ser2P and H3K27ac peaks were categorized as repressed. Finally, regions with no overlap with any of the aforementioned peak sets were categorized as inactive. The Figure 57 depicts the categorization procedure of the TSS references.

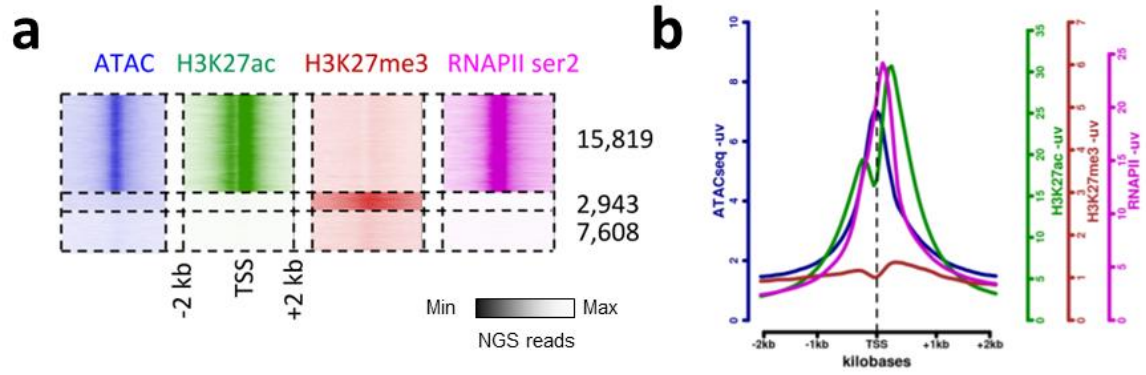


Figure 57 (a) Heatmaps of NO UV ATAC-seq and NO UV H3K27ac, H3K27me3 and RNAPII-ser2P ChIP-seq signal at RefSeq TSSs. TSS activity status was defined by intersecting TSS flanking regions (2 kb in each direction) with H3K27ac ChIP-seq (green), H3K27me3 ChIP-seq (red), and RNAPII-ser2P ChIP-seq (pink). H3K27ac and RNAPII- ser2P containing TSSs were considered as active, H3K27me3 TSSs that did not overlap H3K27ac or RNAPII-ser2P were defined as repressed. The rest of the TSSs were considered as inactive. (b) Average profiles of NO UV ATAC-seq and NO UV H3K27ac, H3K27me3 and RNAPII-ser2P ChIP-seq signal at active TSSs as defined in (a).

All elements that overlapped with both H3K27ac and H3K27me3 peaks were considered as dubious, and were excluded from the annotation. The activity categorization resulted to 15,819 active, 2,943 repressed and 7,608 inactive TSSs (Figure 57).

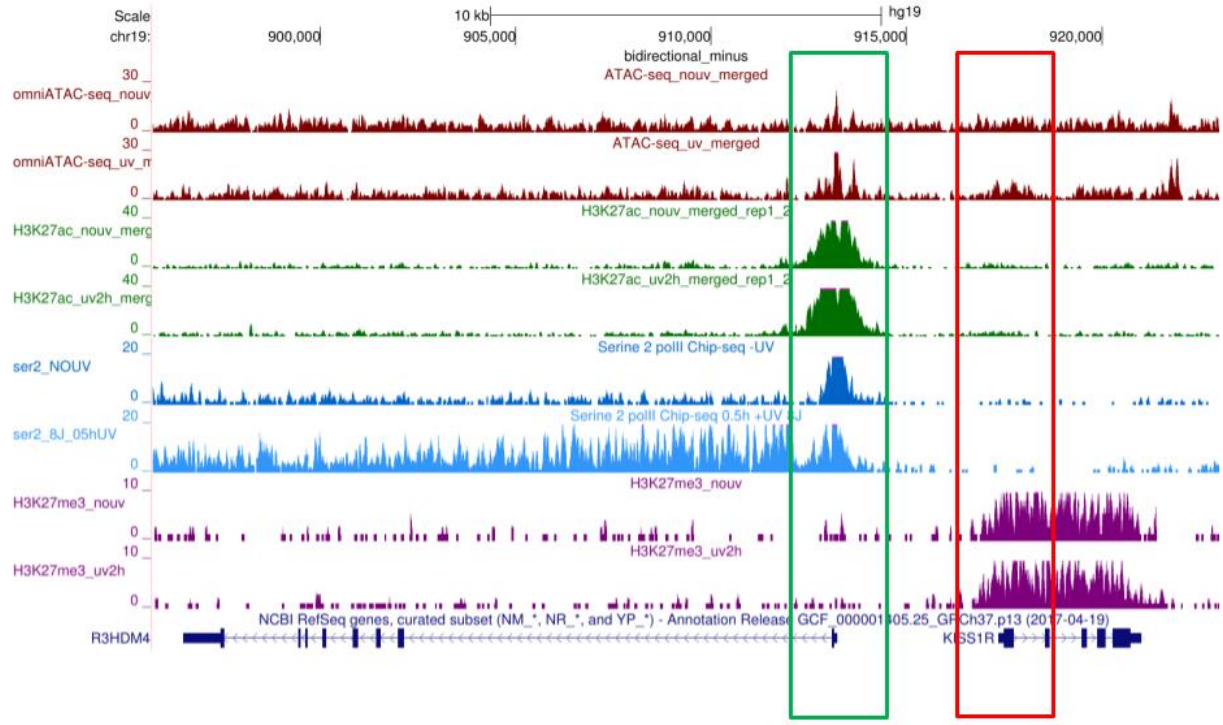


Figure 58 An indicative example of an active (R3HDMA TSS, green box) and a repressed TSS (KSS1R TSS, red box), showing mutually exclusive patterns of H3K27ac (NO UV and +UV, green track lines) and

H3K27me3 (NO UV and +UV, purple track lines) binding profiles, illustrated as a UCSC Genome Browser snapshot. R3HDMA TSS (active TSS) is also enriched with ATAC-seq signal (NO UV and +UV, dark-red track lines) and RNAPII-ser2P (N OU and +UV, blue track lines).

4.3.5 Transcriptional directionality of actively transcribed TSSs and actively transcribed enhancers determination

All active TSSs defined in section 4.3.4 were further analyzed in order to be classified with respect to their transcription directionality. Initially, all elements were split into unidirectional and bidirectional references. Active TSS pairs annotated in opposite strand orientation, with $TSS_{distance} \geq -2\ kb$ and $TSS_{distance} \leq +2\ kb$, where $TSS_{distance} = TSS_{forward\ strand\ coordinate} - TSS_{reverse\ strand\ coordinate}$ ("inter-TSS distance") were categorized as "bidirectional" TSSs, while the rest of the references were categorized as "unidirectional" TSSs (Figure 59).

Bidirectional TSS pairs were further grouped into two categories, convergent bidirectional pairs with $TSS_{distance} \leq 100\ bp$, and divergent bidirectional pairs with $TSS_{distance} > 100\ bp$. Convergent and divergent TSS coordinates were further adjusted using strand-specific CAGE-seq data of primary skin and dermal fibroblasts (see materials and methods, section 2.10.9) as follows: TSS regions were extended to 2 kb in each strand direction, and per-base CAGE coverage was calculated in order to detect the nucleotide occupied by the maximum sense CAGE signal (CAGE summit). Using the transcriptional inactive TSS pairs as a control reference set, CAGE summits were also detected using the same procedure to form the CAGE summit background distribution. Any active CAGE summit with a value over the mean of the background distribution, was set as the new TSS, while all the bidirectional pairs with a non-significant CAGE summit in the aforementioned 500bp region were excluded from the annotation. This procedure resulted in 1,410 active bidirectional TSS pairs, of which 905 pairs were characterized as divergent and 505 pairs as convergent. An example of an active bidirectional TSS-pair id is illustrated in Figure 59.

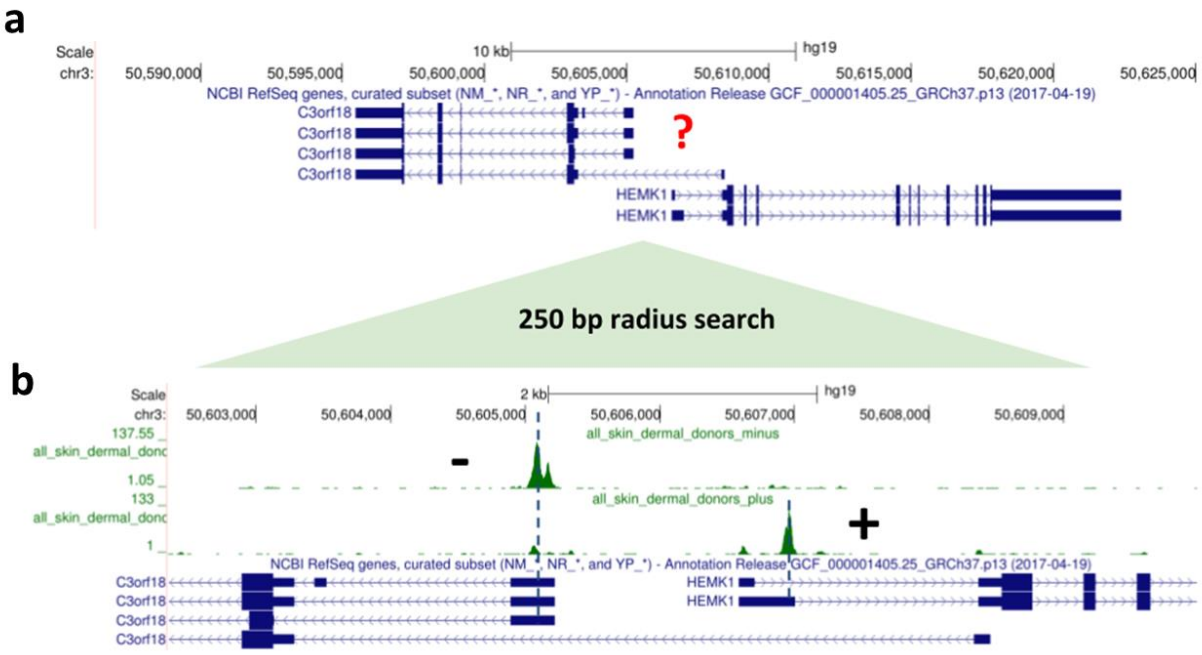


Figure 59 UCSC Genome Browser track of bidirectional TSSs (a) Bidirectional promoters with multiple annotated TSSs are further examined by searching CAGE-seq signal peaks in both directions (250 bp radius). (b) A unique sense CAGE-seq summit (green track lines) defines the new TSS of each TSS-pair member.

The remaining 12,859 unidirectional TSSs were further analyzed to detect asPROMPTs (see section 1.6.2.2) in order to gain a complete overview of the non-coding antisense transcription events occurring around active TSSs. To identify upstream antisense (uaRNA) and downstream antisense (daRNA) transcripts, any active gene model containing more than one mRNA transcripts, was processed in order to keep only the leftmost annotated TSS for the forward strand annotated genes, and the rightmost TSS for the reverse strand annotated genes. The antisense CAGE summit was detected as described above, using a search space of -2 kb upstream up to +1 kb downstream of each active unidirectional TSS. Using the inactive unidirectional references, and the same search space, an antisense CAGE summit background distribution was created as described above. All the antisense CAGE summits linked with an active unidirectional TSS with a with a higher summit than the average value of the respective background distribution, were considered as asPROMPT TSSs. This procedure resulted in 5,366 pairs of active unidirectional - asPROMPT TSSs, which were further subdivided to 1,444 divergent and 3,922 convergent pairs, as described above. Two examples of mRNA TSS - asPROMPT pairs (convergent and divergent) are depicted in igure 60.

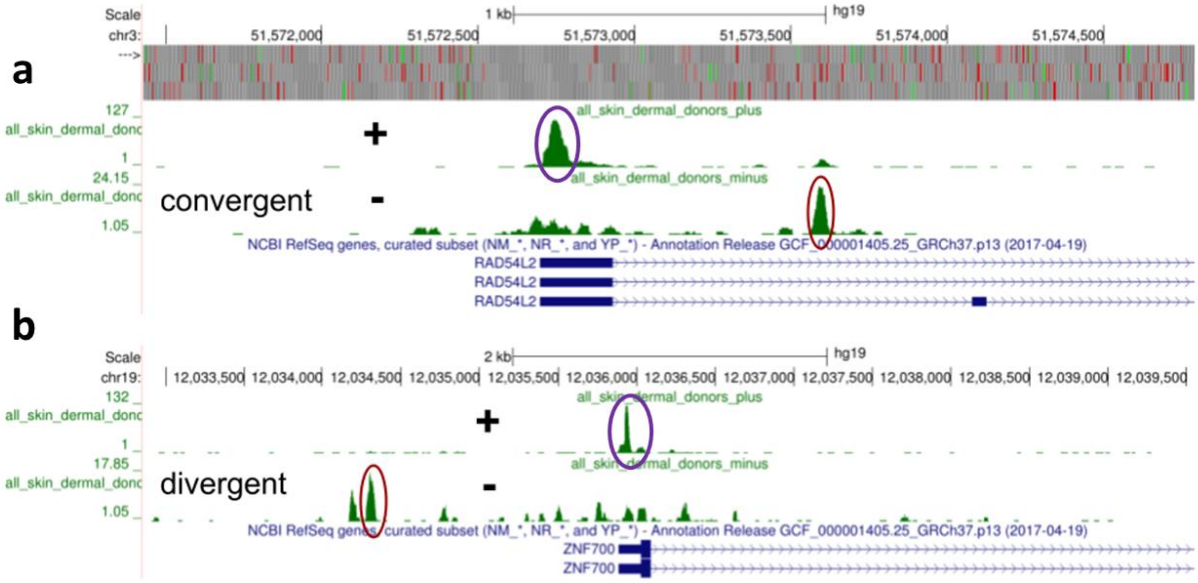


Figure 60 UCSC Genome Browser track of unidirectional TSS - asPROMPT pairs (a) Convergent TSS pairs. Cage signal summits define the mRNA TSS (purple circle) and the asPROMPT TSS (red circle). The asPROMPT TSS is located downstream of the mRNA TSS. (b) DivergentTSS pairs. Cage signal summits define the mRNA TSS (purple circle) and the asPROMPT TSS (red circle).The asPROMPT TSS is located upstream of the mRNA TSS.

Finally, to annotate enhancer TSSs (eTSSs), 65,423 human FANTOM5 enhancers were downloaded by FANTOM5 site (https://fantom.gsc.riken.jp/5/datafiles/latest/), and categorized to 6,766 active, 4,730 repressed and 39,227 inactive following the same described in section 4.3.4 (see Figure 61 for a summary).

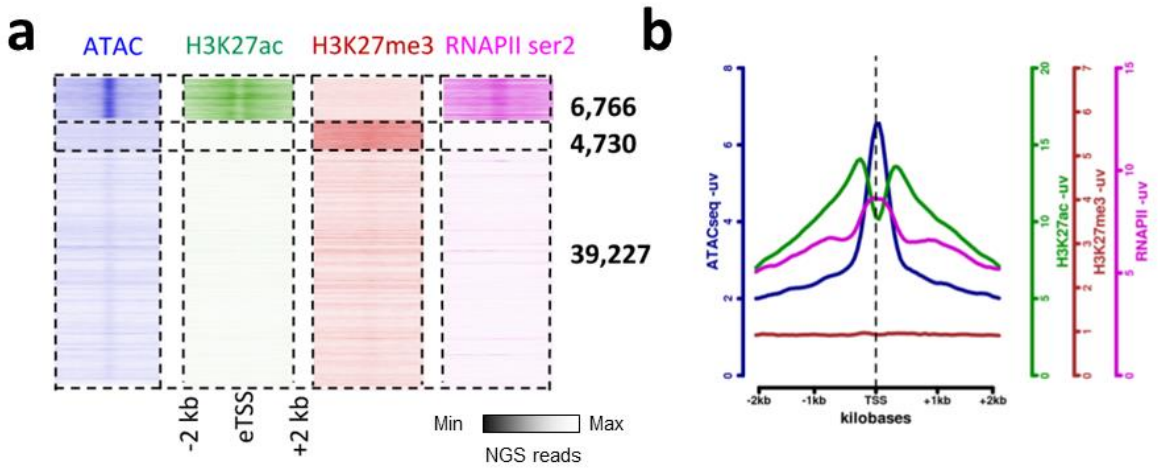


Figure 61 (a) Heatmaps of NO UV ATAC-seq and NO UV H3K27ac, H3K27me3 and RNAPII-ser2P ChIP-seq signal at FANTOM5 eTSSs. TSS activity status was defined by intersecting TSS flanking regions (2 kb in each direction) with H3K27ac ChIP-seq (green), H3K27me3 ChIP-seq (red), and RNAPII-ser2P ChIP-seq (pink). H3K27ac and RNAPII- ser2P containing TSSs were considered as active, H3K27me3 TSSs that did not overlap H3K27ac or RNAPII-ser2P were defined as repressed. The rest of the TSSs were considered as inactive. (b) Average profiles of NO UV ATAC-seq and NO UV H3K27ac, H3K27me3 and RNAPII-ser2P ChIP-seq signal at active eTSSs as defined in (a).

All active intergenic enhancers that don't overlap with actively transcribed promoters (2 kb around TSS) and their respective gene bodies (intergenic enhancers) were further processed, to keep only eTSSs (annotation mid-point) with a distance over 10 kb from the annotation borders of active transcripts, as also a distance over 2 kb from neighboring eTSSs. All active eTSSs were extended to 1 kb sideways, to detect sense and antisense CAGE summits as described above. The same strategy was repeated for all the inactive eTSSs in order to create the sense and antisense CAGE summit background distributions as described above. Consequently, only active intergenic sense and antisense CAGE summits with a height greater than the respective mean of the background distributions were considered, resulting in 1,228 active references. Summarizing the TSS and eTSS annotation, active bidirectional TSS pairs, active unidirectional TSSs paired with an asPROMPT, and active intergenic enhancers are illustrated in Figure 62 as graphical schemas.

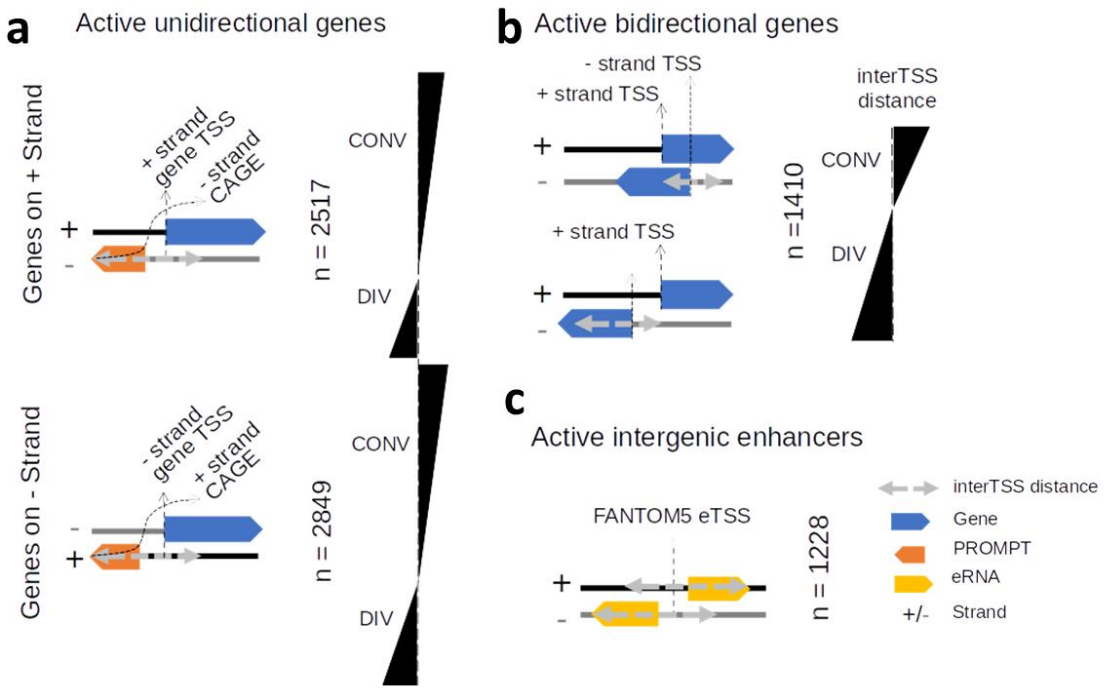


Figure 62 (a) active unidirectional mRNA TSSs (+ or - strand) for which an associated active PROMPT is transcribed in the antisense direction, (b) active bidirectional TSSs of mRNA-mRNA pairs transcribed in opposite directions, (c) active FANTOM5 intergenic enhancer TSSs (eTSS).

4.3.6 ChIP-seq read density analysis reveals patterns of extensive reorganisation of transcription

To examine the effect of low dose ($8\,J/m^2$) of UVC induced stress on transcription, RNAPII-ser2P ChIP-seq datasets in NO UV, +UV +0.5 h, +UV +1 h, +UV +2 h, +UV +6 h, +UV +48 h, RNAPII-ser5P ChiP-seq datasets in NO UV, +UV +0.5 h and +UV +48 h, and RNAPII-hypo ChiP-seq datasets in NO UV, +UV +0.5 h, +UV +1.5 h (see materials and methods section 2.10.2) were analyzed as described in the section 4.1.3. Read density heatmaps were generated using seqMINER (Ye et al., 2011) and revealed an extensive reorganization of RNAPII binding distribution, across all the active mRNAs defined in section 4.3.2. In particular, a global increase of RNAPII-ser2P elongating signal in the gene bodies of all long (over 60 kb) active genes was detected followed by a parallel RNAPII-ser5P and RNAPII-hypo signal decrease in the promoter regions (Figure 63). Interestingly, this phenomenon was not present in the poised and inactive regions (Figure 63 b).

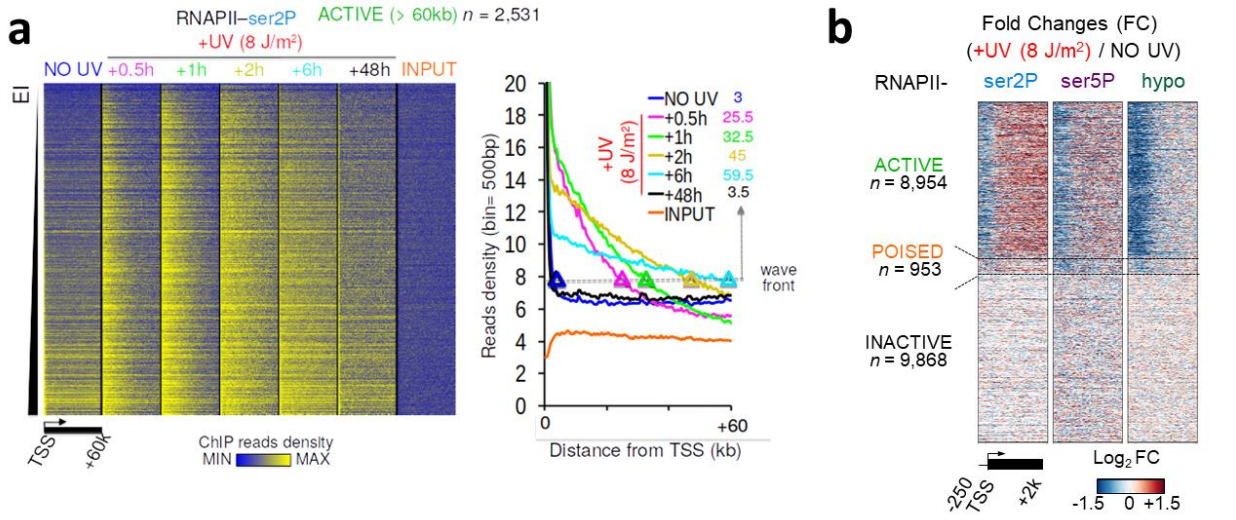


Figure 63 Transcription elongation wave is triggered at all active mRNAs, in response to UV irradiation. (a) Heatmaps of RNAPII-ser2P and INPUT signal at all active genes of length > 60 kb (visualized from TSS up to TSS+60 kb) before and after UV irradiation, ranked by increasing escape index (EI, see below) (b) Heatmaps illustrating the Log2 Fold Change (FC) of RNAPII-ser2P, -ser5P, and -hypo isoforms, comparing normalized read-counts between irradiated and non-irradiated cells at promoter (-250 bp to +100 bp relative to TSS) and gene bodies (+101 bp to +2 kb relative to TSS), separated by gene activity status.

Interestingly, RNAPII-ser2P ChIP-seq average gene density analysis revealed a gradual decrease of RNAPII molecule progression throughout the gene bodies (Figure 63 (a)). This decrease was determined by estimating the average transcription elongation wave front for each time point. Particularly, an arbitrary threshold was initially set (average per 500 bp - bin read density equals to 8) and the intersection point with each average count vector line was considered the transition point for each dataset. These positions were estimated with respect to the TSS and were expressed in kb. Average elongation rates (kb/min) were estimated by combining wave front information between pairs of datasets as previously described (Nicolai et al., 2015) (Lanfeng Wang et al., 2015; Zhong et al., 2011).

To quantify the observed RNAPII signal redistribution in response to UVC, the ratio of average RNAPII RPM at gene bodies (gene regions from +101 bp up to +2 kb relative to TSS, for genes over 2 kb length) over average RNAPII RPM of promoters (gene regions from -200 bp to +100 bp relative to TSS, for genes over 2 kb length) was calculated for each biological condition, for every gene, and every RNAPII ChIP-seq dataset (Promoter Escape Index - EI, see Figure 64).

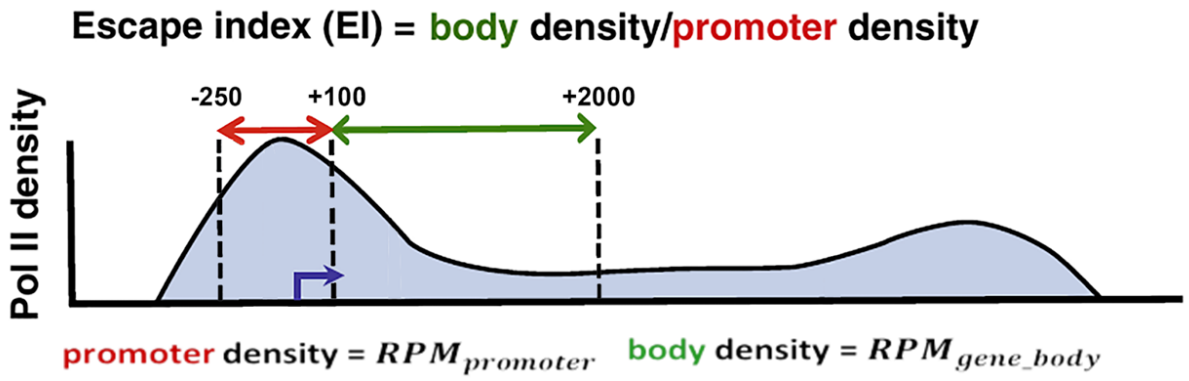
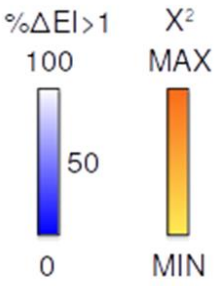


Figure 64 Escape Index calculation. The ratio of average RNAPII RPM at gene bodies (gene regions from +101 bp to +2 kb relative to TSS for genes over 2 kb length) is divided by the average RNAPII RPM at promoters (gene regions from -200 bp to +100 bp relative to TSS, for genes over 2 kb length). The figure is adapted by (Brannan et al., 2012) and edited accordingly.

The particular metric represents the degree of elongating RNAPII molecule escape from PPPs into transcription units.
As a result, a significant and time-resolved increase of EI was observed at the early time points (from +UV +0.5 h up to +UV +2 h) of RNAPII-ser2P and RNAPII-ser5P (Table 7), as compared to the NO UV condition ($\Delta\mathrm{EI},\ +\mathrm{UV\ EI/NO\ UV\ EI}$). In contrast, genes that are not regulated by the transcription machinery in steady state (inactive mRNAs, not significant RNAPII binding at promoters in NO UV condition), were still not affected by the reorganization process. These differences were tested for statistical significance using Chi-square tests (x^2), validating that the observed number of active genes with ΔEI>1 differ from the equivalent expected values in poised genes (Table 7).

Table 7 Summary of RNAPII ΔEIs > 1 for time series analysis of -hypo, -ser5P, and -ser2P isoforms and for gene activity categories defined in section 4.3.2. Chi-square test ($\boldsymbol{x^2}$) applied to validate if active genes' ΔEIs > 1 differ from expected value (poised genes' ΔEIs) for each Δ condition.

		%ΔEI>1			A vs P	
		ACTIVE	POISED	INACTIVE	Chi-Square	p-val
ser2P	0.5h 8J/NO UV	89.6	61.8	45.8	584.13	<.0001
	1 h 8J/NO UV	90.6	62.7	46.2	625.21	<.0001
	2 h 8J/NO UV	89.2	61.4	48.1	569.17	<.0001
	6 h 8J/NO UV	87.0	57.0	47.8	577.97	<.0001
	48 h 8J/NO UV	70.1	50.9	52.1	146.6	<.0001
hypo ser5P	0.5h 8J/NO UV	82.7	66.3	50.4	149.12	<.0001
	48h 8J/NO UV	40.9	48.9	48.8	22.35	<.0001
	1.5h 8J/NO UV	81.0	57.1	49.2	291.46	<.0001
	48h 8J/NO UV	46.0	45.9	47.3	0.01	0.9203



To examine if the aforementioned changes were explained by concurrent increase in gene-body signal density and decrease in promoter signal density, the average differences of RNAPII binding between +UV and NO UV conditions were summarized as an average profile of read density (Figure 65). The particular visualization was generated by a similar strategy as described in section 4.1.3.7 section with an intermediate step of dividing the +UV average count vector by the equivalent NO UV average count vector and plotting it as a Log2 FC average profile.

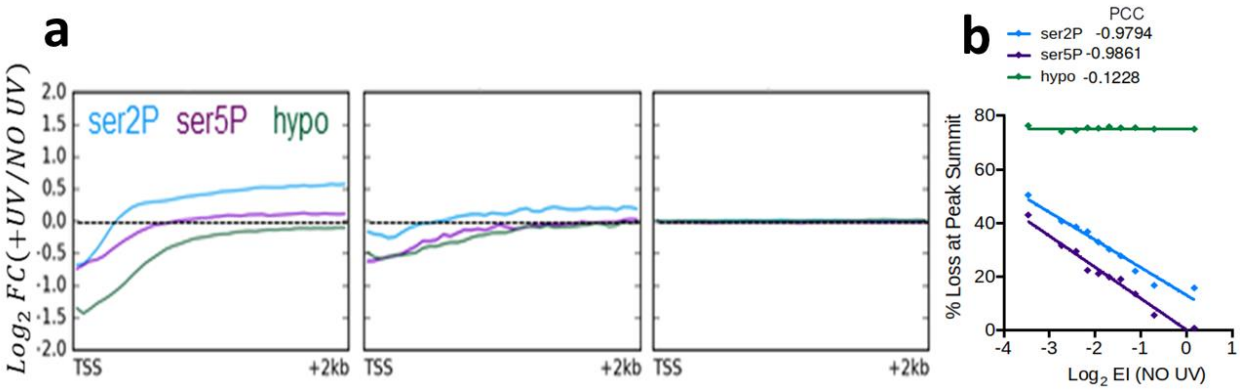


Figure 65 RNAPII reorganization during early UV stress recovery. (a) Average plots of read densities for RNAPII-ser2P, - ser5P and -hypo on active, poised and inactive genes from TSS to TSS + 2kb, before (NO UV) and after (+UV) irradiation, as differential binding profiles (Log2 FC= (normalized read-count vector +UV)/ (normalized read-count vector NO UV)). (b) Correlation plot between RNAPII EI (NO UV) and proportion of read-count loss at peak summits after UV irradiation for -ser2P, -ser5P, and -hypo isoforms, at NO UV condition.

The heatmap analysis of RNAPII-ser2P ChIP-seq (Figure 65 (a)) also revealed a homogenous pattern of transcription reorganization upon UVC induced stress, across the different gene activity levels in NO UV condition, as depicted by the EI values. To further explore this observation, correlation analysis between constitutive NO UV EI of RNAPII-ser2P or RNAPII-ser5P and EI changes after UV (ΔEI(+UV vs NO UV)) was performed for all active genes, demonstrating that EIs of lowly-expressed/ lowly-escaped genes are increased significantly compared to EIs of highly-expressed/ highly-escaped genes (anti-correlation, Figure 66 (a)). Furthermore, differential analysis of peak summit height between NO UV and +UV conditions revealed a more pronounced loss of early-elongating RNAPII reads around PPP sites for less expressed genes (Figure 65 (c)). Also, taking into consideration that the entry of RNAPII into gene bodies is not correlated to gene size (Figure 66 (b)), this mechanism facilitates the release of RNAPII elongation waves from PPP regions even at lowly-expressed genes, providing critical functionalities to the cell.

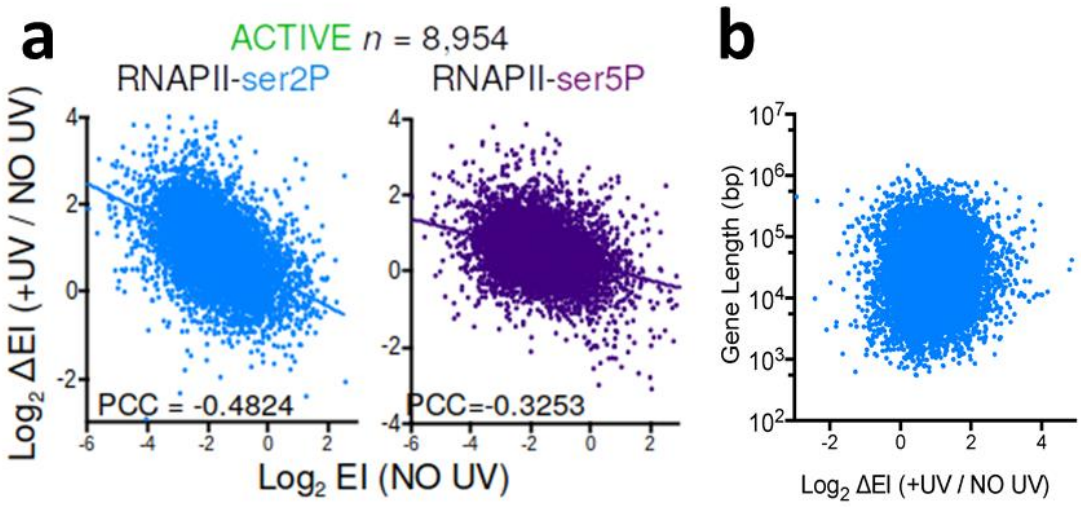


Figure 66 Correlation plots between EI (NO UV) for RNAPII-ser2P (a, left panel) and RNAPII-ser5P (a, right panel) and EI change after UV (ΔEI(+UV/NO UV)) for all active mRNAs. Pearson Correlation Coefficient (PCC) scores are reported. (b) same as in (a) for RNAPII-ser2P ΔEI (+UV / NO UV) vs gene length (bp).

4.3.7 nRNA-seq read density analysis reveals patterns of nascent RNA production asymmetries between proximal and distal gene regions

To explore the effect of UVC induced stress in nascent RNA production in actively transcribed genes, nRNA-seq libraries in VH10 and CSB cells, in NO UV +0 h, NO UV +24 h, +UV +0.5 h, +UV +2 h, and +UV +24 h (see materials and methods, section 2.10.5) conditions were analyzed as described in section 4.1.4. Gene activity was determined using the methodology described in section 4.2.3 and common active genes over 100 kb between VH10 and CSB cells were considered for further analysis. Average density vectors of genes with length greater than 40 kb were generated, and +UV +0.5 h, +UV +2 h vectors were divided by the NO UV +0 h vector, while the +UV +24 h by the NO UV +24 h vector. Log2 ratios were illustrated as average profiles, to reveal that regions directly downstream of PPP show considerable increase in nRNA-seq signal at +UV 0.5 h, and +UV +2 h conditions (Figure 67). This was followed by a global decrease of nRNA-seq signal in the more distal gene regions, while for the +UV +24 h ratio in VH10, a homogenous pattern of nRNA-seq was observed across the whole gene lengths. On the contrary, in CSB cells +UV +24 h / NO UV +24 h ratio the nRNA-seq asymmetry pattern persists.

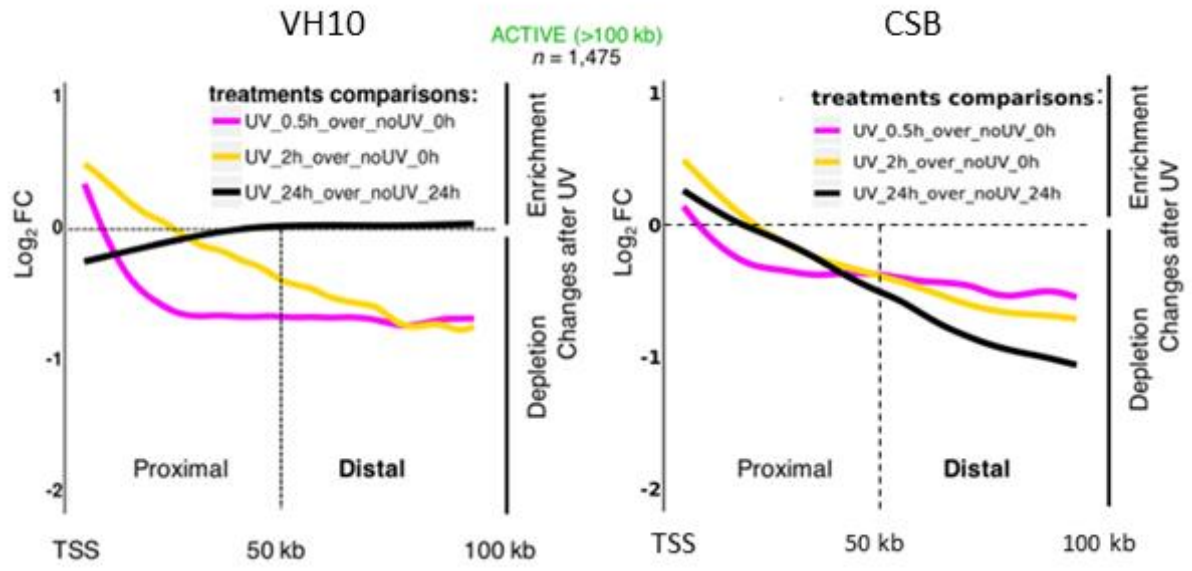


Figure 67 Characterization of the UV-dependent global and transient changes of nascent RNA synthesis. a, Average plots showing the difference in nRNA-seq read densities from transcription start sites (TSS) to transcription termination sites (TTS) (gene lengths are normalized to 100 bins and only bins 5 to 95 are displayed for clarity) for genes larger than 40 kb before and after low doses of UV irradiation (15 J/m^2) (Log2 FC between compared treatments are indicated). nRNA were labeled with EU for 10 min before indicated time, see also Methods.

Strikingly, although the overall nRNA level in CSB cells is reducedl, nascent RNA synthesis is still detectable at the beginning (5 'end) of active genes even at later time points when transcription initiation is expected to be inhibited (Figure 67, right panel). Comparing the nRNA signal between the two cell lines, 2 hours after UV exposure, we conclude that nRNA molecules are continued to be synthesized in both CS-B and VH10 cells
Reasonably, while in normal cells transcription recovery is completed after 24 h of UVC exposure, CS-B cells cannot complete the transcription of active genes, although new RNA molecules continue to be transcribed at the beginning of the genes.
The above quantifications suggest that the previously reported decrease of nascent RNA is due to the fact that the overall level of transcription elongation decreases throughout the gene bodies during early cellular response in UVC induced stress (Figure 63, right panel), possibly because of the increased stalling of RNAPII molecules upon encountering DNA lesions.
Additionally, regarding the observations in late response (+UV 24 h), it seems that transcription recovery is driven by a functional TC-NER machinery in VH10, as opposed by a non-functional TC-NER machinery in CSB cells.

4.3.8 Analysis of RNAPII-ser2P DRB ChIP-seq and pre-DRB nRNA-seq delineates the RNAPII elongation wave release in normal skin fibroblasts

To verify that the release of RNAPII from PPP is performed “de novo” upon UVC-stress, pre-DRB nascent RNA-seq and RNAPII-ser2P ChIP-seq experiments were generated (see materials and methods, sections 2.10.2 and 2.10.6).
VH10 VH10 RNAPII-ser2 pre-DRB +UV -DRB and pre-DRB +UV +DRB were analyzed as described in section 4.1.3. Heatmaps visualization of +UV conditions on active genes over 10 kb, revealed that DRB treatment cancels the aforementioned stress-dependent RNAPII wave generation and propagation (Figure 68 (a)). Additionally, escape Index analysis (EI, figure 64) shows that RNAPII escape dramatically decreases in promoter-proximal regions (Figure 68 (b)).

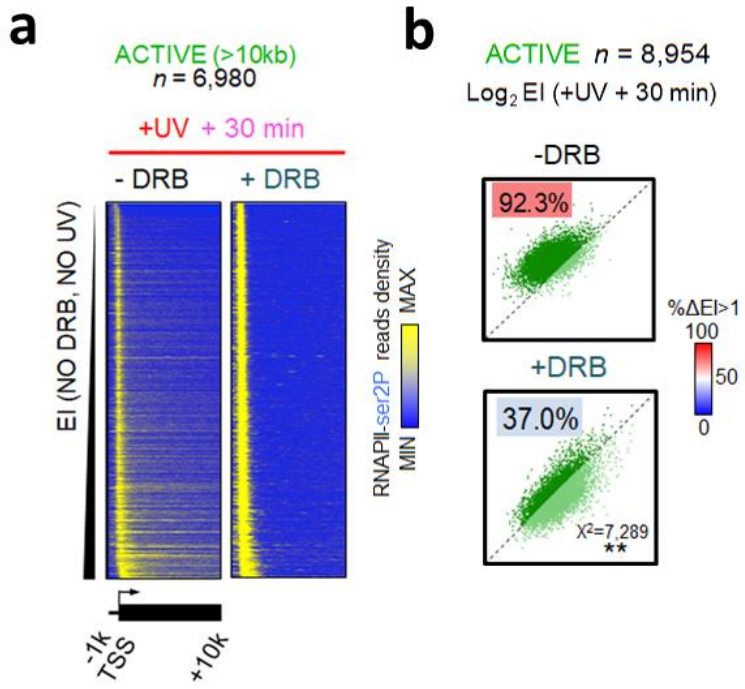


Figure 68 (a) Heatmaps illustrating the distribution of ser2P-RNAPII reads at active genes over 10kb (regions plotted: TSS -1 kb to TSS +10 kb, ranked by increasing EI as in Figure 63), after UV induction (+UV), in the presence (+) or not (-) of DRB. (b) RNAPII-ser2P NO UV EI comparison with +UV EI (+ and – DRB) at all active mRNAs. Percentage of transcribed elements with increased escape after UV (ΔEI > 1, dark green dots) is illustrated in a color scale (blue-white-red, from minimum to maximum values). $\boldsymbol{X^2}$ tests determine if the genes with ΔEI > 1 significantly differ between treatments (**P < 0.0001).

To focus on pri-elongation RNAPII molecules (polymerases that were engaged on elongation before UV irradiation), VH10 DRB ChIP-seq experiments of RNAPII-ser2P in NO UV -DRB, NO UV +0 h +DRB, NO UV +10 min +DRB, NO UV +30 min +DRB, +UV +10 min +DRB, +UV +30 min +DRB conditions (see materials and methods, 2.10.4) were analyzed using the methodology described in section 4.1.3. Heatmaps and average profile visualization showed a substantial retain of RNAPII molecules upon UV in the distal parts of long active genes over 60 kb that were traveling before the time of DRB as summarized in Figure 69.

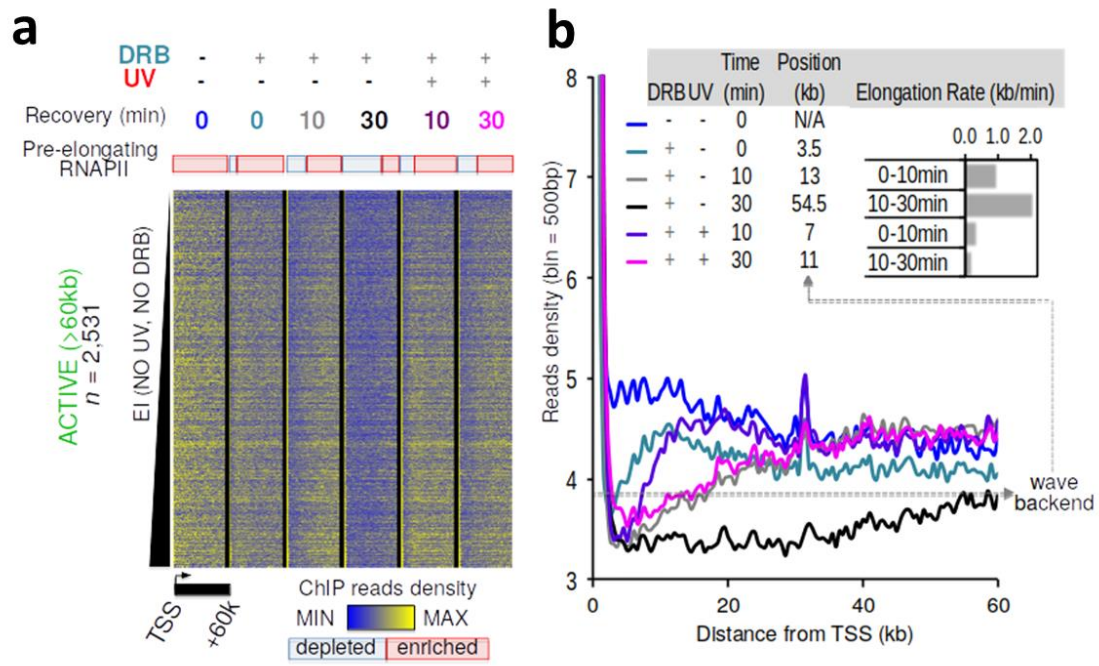


Figure 69 RNAPII stalling at PPP uncovers the kinetics of elongating RNAPII molecules prior to UV stress (pri-elongating). (a) Heatmaps of RNAPII-ser2P signal at active genes over 60 kb (region plotted: TSS to TSS+60 kb), ranked by increasing NO UV EI (Figure 63). (b) Average profile of RNAPII-ser2P signal derived from (a) highlighting the pri-elongating RNAPII wave backend positions, based on an arbitrary threshold (dashed line) representing the transition state. Elongation rates are calculated from differences between wave backend positions from consecutive time points.

To quantify the kinetics of pri-elongating RNAPII molecules based on RNAPII-ser2P DRB-ChIP-seq data, elongation rates in both NO UV and +UV conditions were estimated as described in section 4.3.6, showing a substantially decreased rate (Figure 69 (b)) but not a total loss of ongoing elongation.

Additionally, VH10 pre-DRB nascent RNA-seq libraries were analyzed in NO UV +0 h, NO UV +10 min, NO UV +20 min, NO UV +1 h and +UV +0 h, +UV +10 min, +UV +1 h , and +UV +2 h (see materials and methods, section 2.10.6). All samples were analyzed using the methodologies described in sections 4.1.4 and 4.2. The generated heatmaps and average profiles of nascent RNA signal revealed that DRB treatment had efficiently eliminated all the elongating RNAPII molecules from gene bodies at the time of UVC stress (Fig. 70, time = 0 h), while after the removal of the drug, UV- or non-irradiated cells resumed elongation with different kinetics (Figure 70).

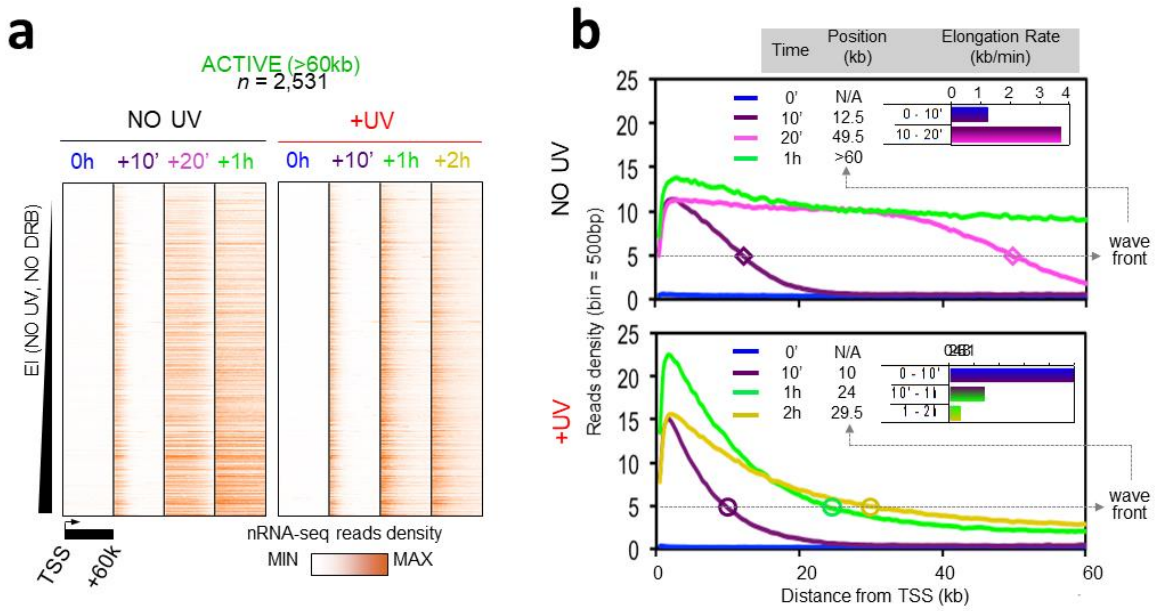


Figure 70 Released de novo elongation wave decelerates progressively in response to UVC induced stress. (a) Heatmaps of nascent RNA signal at long active genes (60 kb, plotted from TSS to TSS+60 kb), ranked by increasing NO UV EI (see Figure 63). (b) Average profiles of nascent RNA signal derived from (a), highlighting de novo elongation wave release of RNAPII. Differences in wave front positions at an arbitrary threshold (dashed line) were used to calculate average (n = 2,531) elongation rates in consecutive time points (when applicable).

Strikingly, UV irradiation did not suppress the nascent transcription recovery after the removal of DRB (Figure 70), but it triggered a wave of productively elongating RNAPII molecules, released in all the active gene bodies, replicating the UV-triggered phenomenon described above (Figure 63). Transcription elongation rates of de novo released RNAPII molecules, calculated as described above (section 4.3.6), revealed a decreased pattern of elongation recovery in early UVC response (Figure 70 (b)), similar to the pri-elongating (Figure 69 (b)) molecules, confirming the overall changes measured in ser2P-RNAPII molecules illustrated in Figure 63. To analyze these observations in higher resolution, the HMM algorithm described in section 4.2 was applied in each of the examined datasets (Figure 71, and section 4.2).

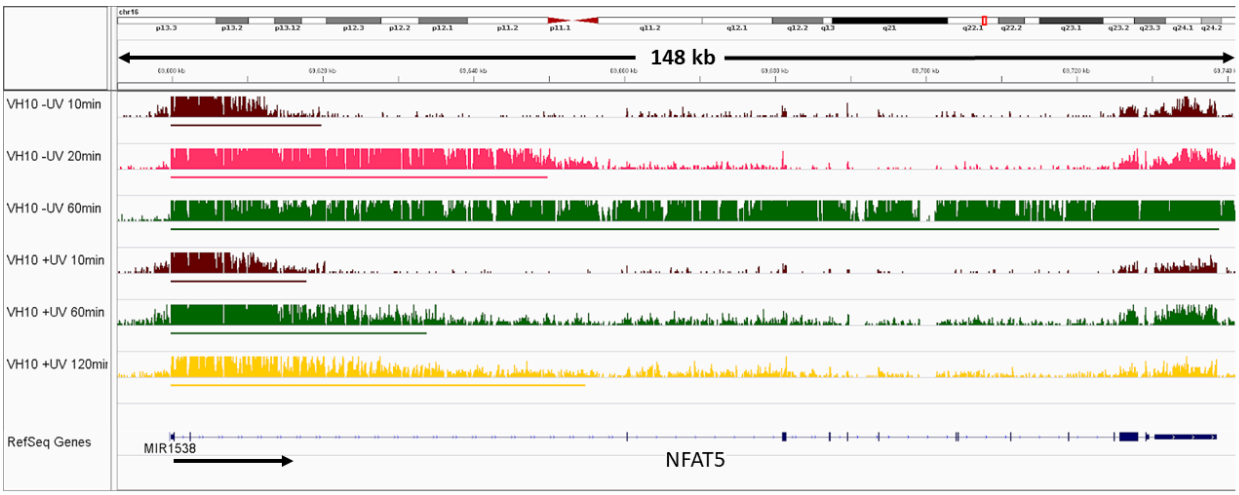


Figure 71 HMM wave-front prediction in VH10 pre-DRB nRNA-seq datasets. NFAT5 gene is illustrated as an indicative example. Solid lines under each dataset signal corresponds to the elongation wave annotation generated by the procedure described in section 4.2.

Interestingly, in distal regions of long active gene bodies, de novo nascent RNA-seq signal was still detectable, but in lower levels (Figures 51, 53, 55, and 70 (b)), showing that de novo released RNAPII molecules still elongate, but at slower rates. This phenomenon was also detectable at later time points during damage recovery, where RNAPII molecules progressed cumulatively between +2 h and +6 h (Figure 63).

Summarizing the nRNA-seq analysis, it's shown that in response to genotoxic threats, there is a significant increase of nascent RNA signal in PPP regions of all active genes, as an increasing trail of RNAPII molecules switches to a damage-sensing productive elongation state throughout the active gene-bodies. Consequently, nascent RNA synthesis rate is promptly and constantly affected along active genes, as a result of the deceleration of both de novo released and already transcribing RNAPII molecules during the DNA damage-sensing procedure. The particular model explains the recent findings, suggesting that transcription initiation and elongation still take place in PPP regions upon UV-irradiation, even though progress into gene bodies is significantly delayed (Bugai et al., 2019a; Liakos et al., 2020; Williamson et al., 2017a).

4.3.9 omni-ATAC-seq read density analysis reveals patterns of global chromatin accessibility increase along transcriptional regulatory regions upon UV

To examine if, and to what extent, the widespread UVC triggered release of elongating RNAPII molecules, and the increase in nascent RNA production downstream of the TSS of active genes (Figures 63 and 70) (Borisova et al., 2018; Williamson et al., 2017b) are linked with putative alterations in chromatin accessibility, a set of omni-ATAC-seq experiments were designed (materials and methods, section 2.10.7). In particular, VH10 omni-ATAC-seq experiments in NO UV and +UV +2 h conditions were generated, and analyzed using the methodology described in section 4.1.5.
Peak calling analysis of VH10 samples resulted in a set of 106,052 Accessible Regions (ARs)

across all conditions, and correlation analysis across the whole genome showed patterns of reproducibility among biological replicates of each condition (Figure 72 (a)).

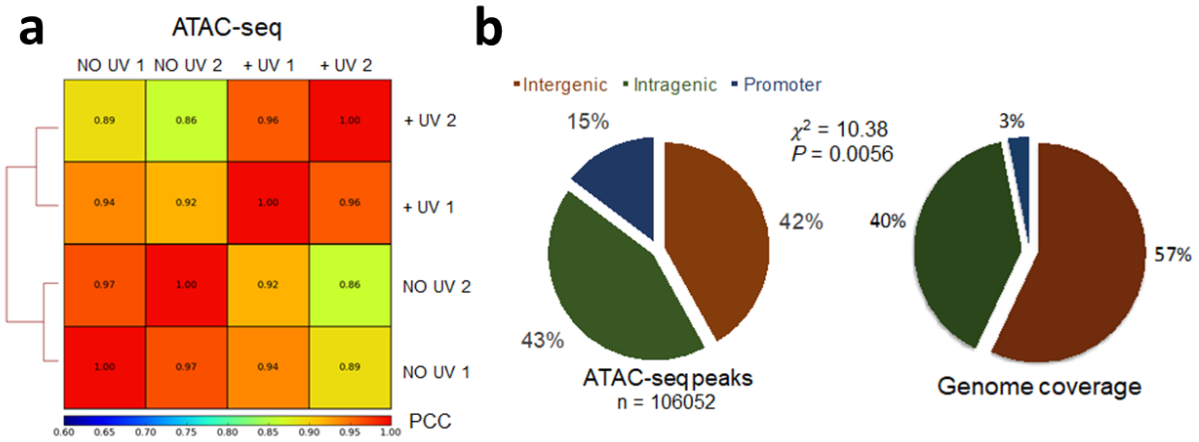


Figure 72 Quality Control (QC) of VH10 omni-ATAC-seq libraries (materials and methods). (a) Correlation heatmap of ATAC-seq signal along the datasets revealed biological replicate reproducibility for both NO UV and +UV conditions. Grouping of rows and columns was performed by using hierarchical clustering between Pearson's Correlation Coefficients (PCCs). Bin-count vectors of genome-wide segments (3 kb windows) of ATAC-seq signal were generated for each dataset, low-density windows among all datasets were excluded and PCCs were calculated for each pair of vectors.

AR annotation showed an enrichment of accessible loci at promoters, intragenic, and intergenic regions with transcriptional regulatory function (TSSs, TSS flanks and enhancers according to NHDF 15-state roadmap annotation, Figure 43).
ATAC-seq signal was summarized at all ARs to generate heatmaps and average profiles of chromatin accessibility signal along the consensus AR set (Figure 73 (a)). The particular visualizations revealed a pattern of global gain of chromatin accessibility in response to UVC-stress. Consequently, log_2 ratios of +UV over NO UV signal (FC) were generated and visualized as scatter plots, to report a global increase of chromatin accessibility after UVC stress induction at 97.9% of promoter, 94.6% of intragenic, and 94.4% of intergenic ARs (Figure 73 (c)). Differential accessibility analysis was performed using the DESeq2 (Love et al., 2014) mode of diffBind (Stark & Brown, 2011)(see section 4.1.3), avoiding edgeR, since the TMM normalization (M. D. Robinson et al., 2009) relies on a core of sites that don't systematically change their accessibility affinities, an assumption that is violated in the design of the particular experiment. Differentially Accessible Regions (DARs) were defined by applying $abs(log_2(FC)) > 1$ and $p - value < 0.001$ to extract 6,410 loci, with significant increase in chromatin accessibility upon UV (DAR-gain) (Figure 73 (d)).

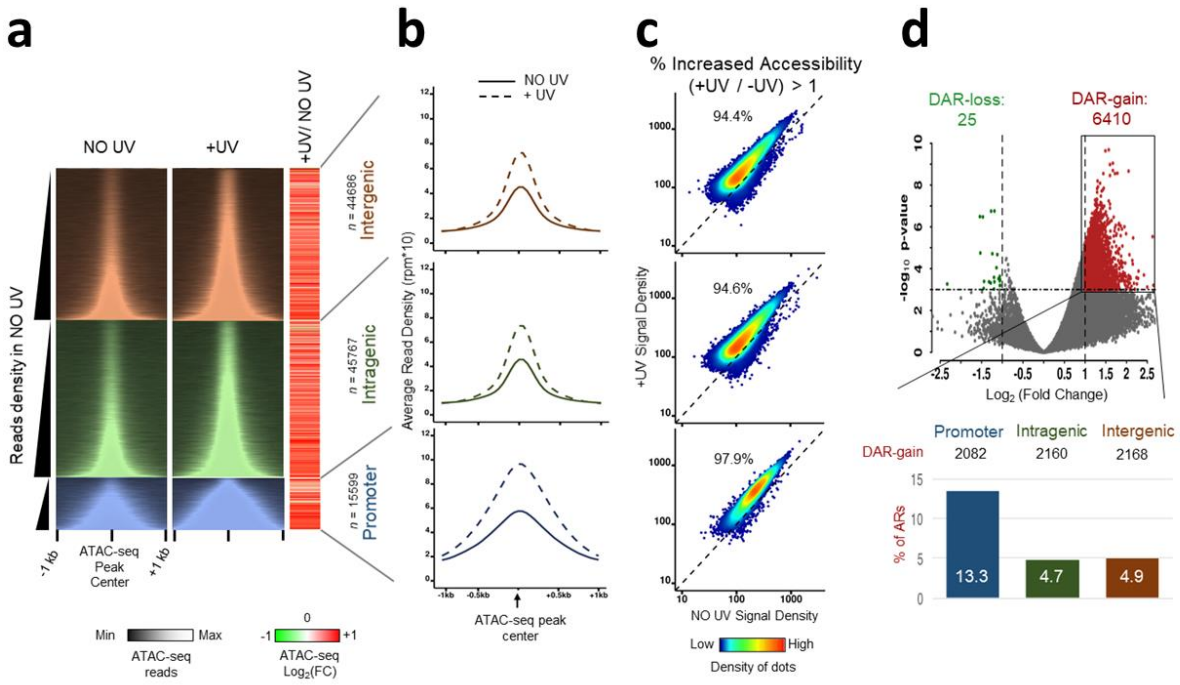


Figure 73 Global increase of chromatin accessibility during early recovery from UVC-stress induction. (a) Left panel: Heatmap illustration of ATAC-seq signal around ATAC-seq peak centers (1 kb flanks) in NO UV and +UV conditions, separated based on their genomic position relative to RefSeq transcripts (intergenic, intragenic and promoter peaks), sorted by increasing NO UV signal. Right panel: Heatmap illustrations depicting the $\boldsymbol{log_2}$ Fold Change ($\boldsymbol{log_2}$ FC) between +UV and UV ATAC-seq signal in regions described in the "Left panel". (b) Average profile of NOUV (solid curve) and +UV (dashed curve) ATAC-seq signal at regions described in (a). (c) Heat density scatter plots of ratios of +UV ATAC-seq signal over NO UV ATAC-seq signal, in regions described in (a). (d) Upper panel: Scatter plot (Volcano plot) summarizing the differential accessibility analysis results between ATAC-seq +UV and NO UV conditions. Differential accessible regions (DARs) with significantly increased (DAR-gain) or decreased (DAR-loss) accessibility, are visualized in red and green, respectively. Bottom panel: Proportion of DAR-gain loci in intergenic, intragenic and promoter ARs.

DAR-gain loci were localized at promoter regions representing 13,3% of all promoter ARs (Figure 73 (d)), defining a potentially functional relevant chromatin opening at TSS regions. DAR-gain loci located at intragenic regions or within active FANTOM5 enhancers, were examined for potential links to "targeted" promoters. All direct (promoter) and indirect (intergenic and intragenic enhancers) gene links were functionally analyzed using REACTOME pathway analysis (Fabregat et al., 2017) to identify ($adjusted\ p - value\ <\ 0.05$) a number of biological pathways previously associated with DDR processes, including cellular response to stress, DNA repair, transcription regulation by TP53 and cell cycle checkpoints, and a broad range of significant Gene Ontology (GO) terms (Figure 74). The particular results are in alignment with the widespread PPP release of elongating RNAPII molecules at all active genes upon UV irradiation (Figures 63, 67 and 70).

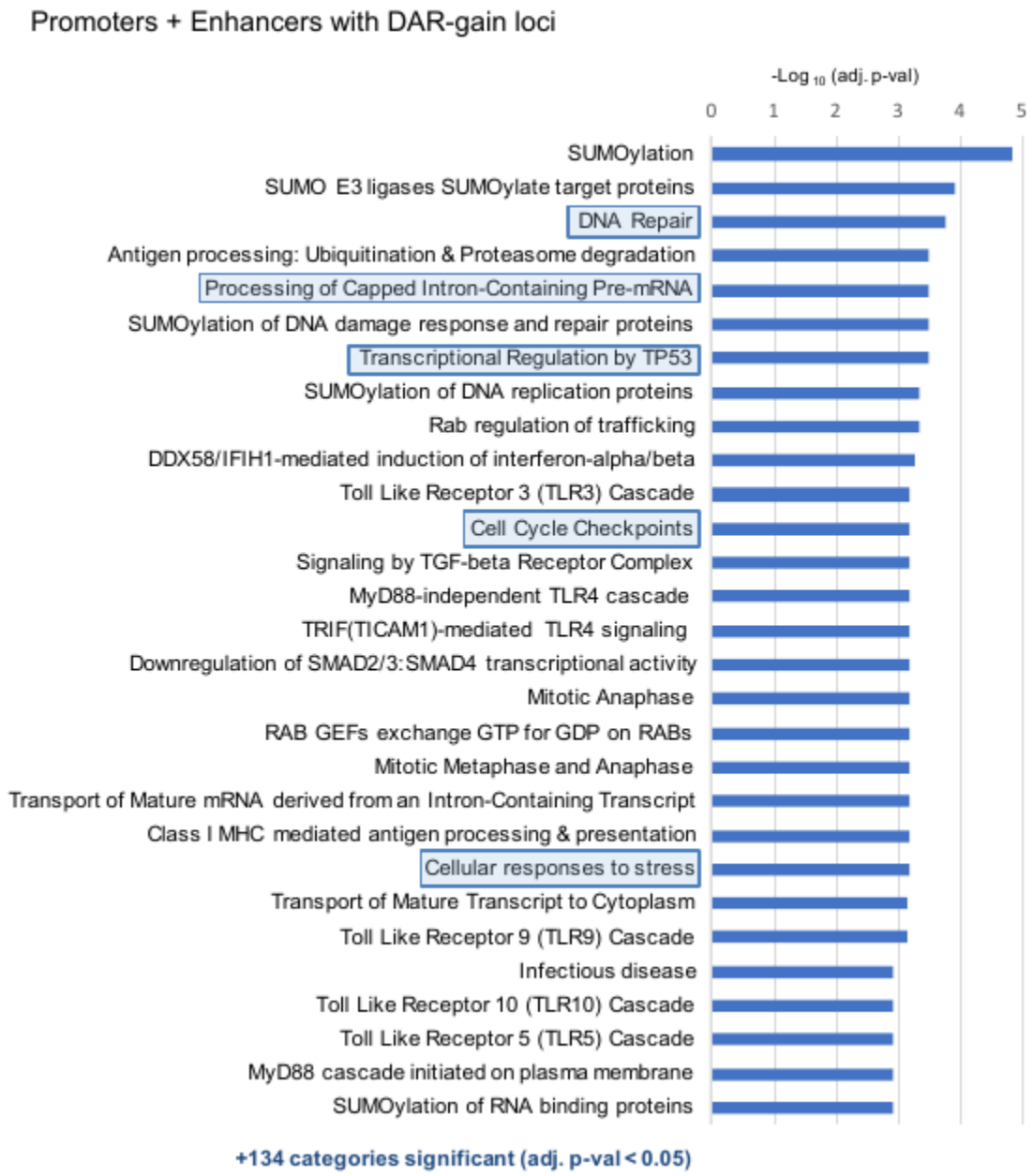


Figure 74 Reactome analysis of DAR-gain regions. The transcript list associated to DAR-gain regions were tested for biological pathway enrichment using the Reactome database. Significant results (p-adjusted value < 0.05) are ordered by decreasing significance. Blue boxes highlight the pathways that are relevant to the processes involved in UV-response or transcription.

4.3.10 H3K27ac and H3K27me3 marks remain stable after UV

To examine if the global increase in chromatin accessibility is coupled with changes in post-translational modifications (PTMs) of histones around transcriptional regulatory regions during the recovery period from UVC irradiation, ChIP-seq experiments of the silencing chromatin mark

H3K27me3 and the activation mark H3K27ac were conducted. In particular, VH10 H3K27ac and H3K27me3 ChIP-seq experiments in NO UV and +UV +2 h were designed and generated (see materials and methods, section 2.10.3). To focus on TSSs of mRNAs and enhancers, the NGS data analysis was performed using the genome annotation depicted in figures 57 and 61 as references. In particular, average profiles, heatmaps and boxplots of the H3K27ac, H3Kme3 and RNAPII-hypo ChIP-seq signal, as also the ATAC-seq signal were generated using the pipeline described in section 4.1.3, using 2 kb extended TSS and eTSS references (Figure 75).

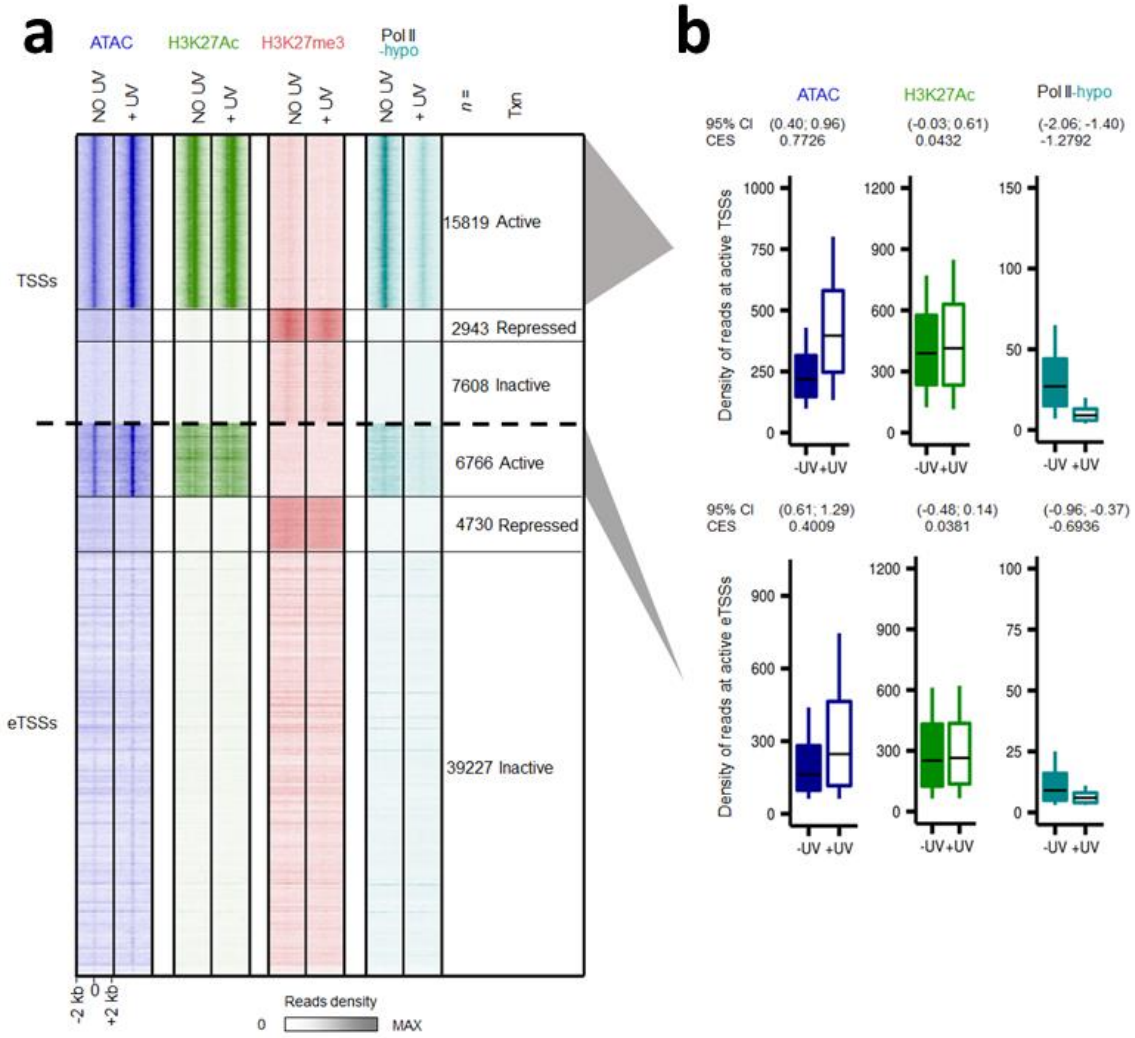


Figure 75 H3K27ac and H3K27me3 histone modifications remain essentially stable during early recovery from UV-stress induction. (a) Heatmap illustrating ATAC-seq, H3K27ac, H3K27me3 and Pol II-hypo ChIP-seq signal at NO UV and +UV conditions at genomic regions 2 kb around active, inactive and repressed TSSs and eTSSs, respectively. (b) Box plot summarization of ChIP-seq signal at genomic regions 2 kb around active TSSs and eTSSs, respectively. Each signal distribution contains the 25th–75th percentiles, while error bars represent the higher/ lower values included in 1.5 * IQR (inter-quartile range, or distance between the first and third quartiles). 95 % confidence intervals (CI) of mean differences between + UV and NO UV of $\boldsymbol{log_2}$ counts were calculated as described in materials and methods,

section 2.13. Effect sizes of $\boldsymbol{log_2}$ counts between NOUV and +UV samples were calculated using Cohen's method (CES).

The particular visualizations (Figure 75) confirmed that the genome-wide significant increase of chromatin accessibility that was detected in the accessible genome of VH10 cells during the early response to UV-induced damage (Figure 73), was also detectable in all actively transcribed mRNA and enhancer promoters. Markedly, the particular phenomenon was associated with preservation (slight but not significant increase) of H3K27ac +UV signal levels (Fig. 75 (a) and (b), 95% CI includes 0), but also no increase of H3K27me3 signal in response to UV at actively transcribed regions. Accordingly, no gain of H3K27ac, or RNAPII-hypo at repressed promoters was detected, and H3K27me3 showed a relatively stable pattern across all the repressed references.
Interestingly, RNAPII-hypo ChIP-seq signal level was significantly decreased at actively transcribed promoters upon UV-induced stress (Figure 75, 95% CI includes 0), a result that is in sharp contrast with the global increase of the ATAC-seq signal in the same references and cellular conditions.

4.3.11 Release of de novo elongation waves promote sensing of DNA damages

Since DNA lesions that are formed in the transcribed strand of actively expressed genes are detected by elongating molecules of RNAPII, the aforementioned UVC dependent trigger of elongation waves (sections 4.3.6, 4.3.7, and 4.3.8) could result in increased DNA lesion-sensing that will in turn enhance the assembly of TC-NER machinery for faster and more frequent damage repair.
To investigate this hypothesis, a functional link between the probability of RNAPII stalling and the detection of DNA lesions was examined. In particular, higher doses of UVC (20 J/m^2) were applied to VH10 cells in order to induce a larger number of DNA lesions, and RNAPII-ser2P ChiP-seq datasets in NO UV, +UV +1 h, +UV +2 h and +UV +48 h were generated (see materials and methods, section 2.10.1) and analyzed as described in section 4.3.6. Heatmaps analysis of ChIP-seq signal revealed that the widespread release of RNAPII wave at all active gene bodies (Figure 76) is also reproduced in higher lesion density, while average profile analysis of ChIP-seq signal and elongation rate estimation (see scetion 4.3.6) showed that the higher lesion rate was adequate to cause increased staling of RNAPII molecules at different time points during recovery (Figure 76).

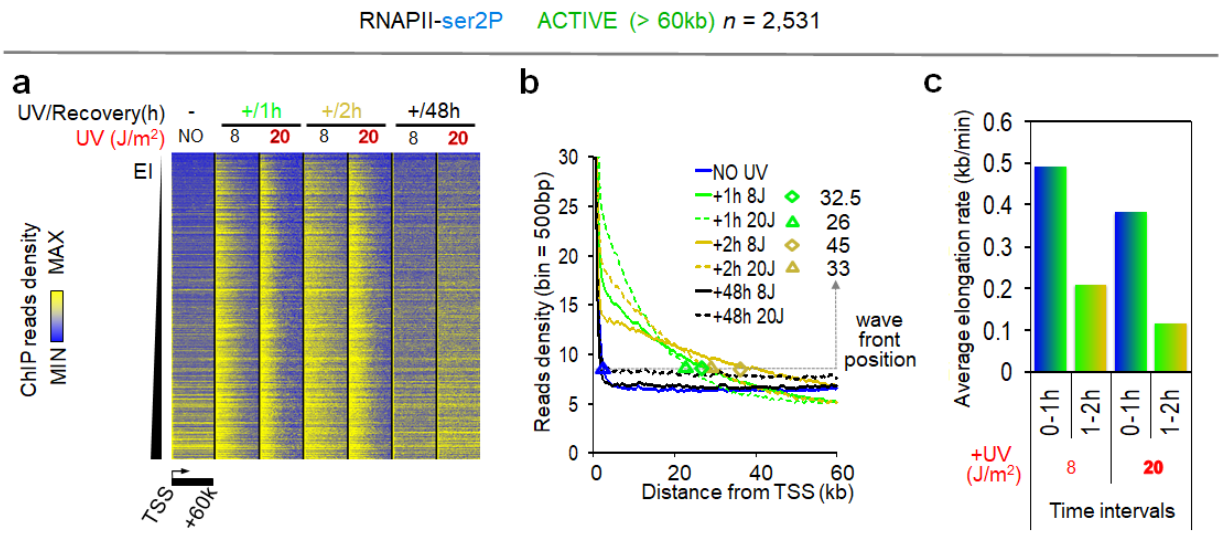


Figure 76 RNAPII elongation wave deceleration upon UVC-stress depends on UV dosage. (a) Heatmaps illustrating NO UV and +UV (1 h: bright green, 2 h: gold, and 48 h: black) RNAPII-ser2P signal distributions at genes over 60 kb (plotted from TSS to TSS +60 kb), ranked by increasing NO UV EI (see Figure 63), for 8 $\boldsymbol{J/m^2}$ and 20 $\boldsymbol{J/m^2}$ UVC doses. (b) Average profiles of RNAPII signal as described in (a). Wave front positions are estimated using an arbitrary threshold, representing the transition state. (c) Average (n = 2,531) elongation rates (kb/min) were estimated using consecutive time points for conditions defined in (a), and average wave front positions determined in (b).

Seeking for more insights regarding the functional consequences of the stress-dependent transcriptional wave release, an analysis of ChIP-seq signal at regions of actively transcribed genes, prone to lesion induction was performed. This analysis was based on regions that are considered potential DNA adducts, and in particular di-pyrimidine TpTs (TTs) since they are the most frequently dimerized pyrimidines after UV exposure (Ramanathan & Smerdon, 1986). Additionally, UV-induced CPDs and 64s are governed by TT abundance (Adar et al., 2016; Mao et al., 2016; Teng et al., 2011).

To efficiently annotate TT dinucleotide loci, all active genes were scanned for XTTX motif occurrences in the non-template strand, where X = {A, C, G}. All neighboring dinucleotides of a distance less than 70 bp were filtered out, and their distance from their corresponding TSS was recorded. This resulted in a final list of n = 29,612 active genic TTs.

RNAPII-ser2P ChIP-seq alignments were summarized to generate read density profiles at extended TT genomic regions (−400 bp to +400 bp relative to TT center). TTs were clustered in 6 categories, relative to their distance from their corresponding TSS (table 8), while the clusters were further annotated for each RNAPII-ser2P +UV condition (8 J/m^2 +UV +1 h, +UV +2 h, +UV +6 h) as “upstream” or “downstream”, based on their relative topology with respect to the wave-front, as estimated in section 4.3.6. PPP-specific TTs (TSS up to 3 kb) were not considered for this analysis.

Table 8 TT-cluster annotation, with respect to the wave-front positions as summarized in Figure 63**.**

		TT Distance to TSS (bp)		Wavefront		
CLUSTER	**Number of TTs**	**Start**	**End**	**+1h**	**+2h**	**+6h**
I	4218	3000	10000	Upstream	Upstream	Upstream
II	7916	10001	32500	Upstream	Upstream	Upstream
III	2242	32501	45000	Downsteam	Upstream	Upstream
IV	1954	45001	59500	Downsteam	Downsteam	Upstream
V	3254	59501	100000	Downsteam	Downsteam	Downsteam
VI	5267	100001	1150817	Downsteam	Downsteam	Downsteam

Heatmap analysis of RNAPII-ser2P signal at extended TT regions showed that RNAPII accumulation at potential damaged sites was maximized in clusters annotated as “upstream” at +UV +2 h condition (Figure 77).

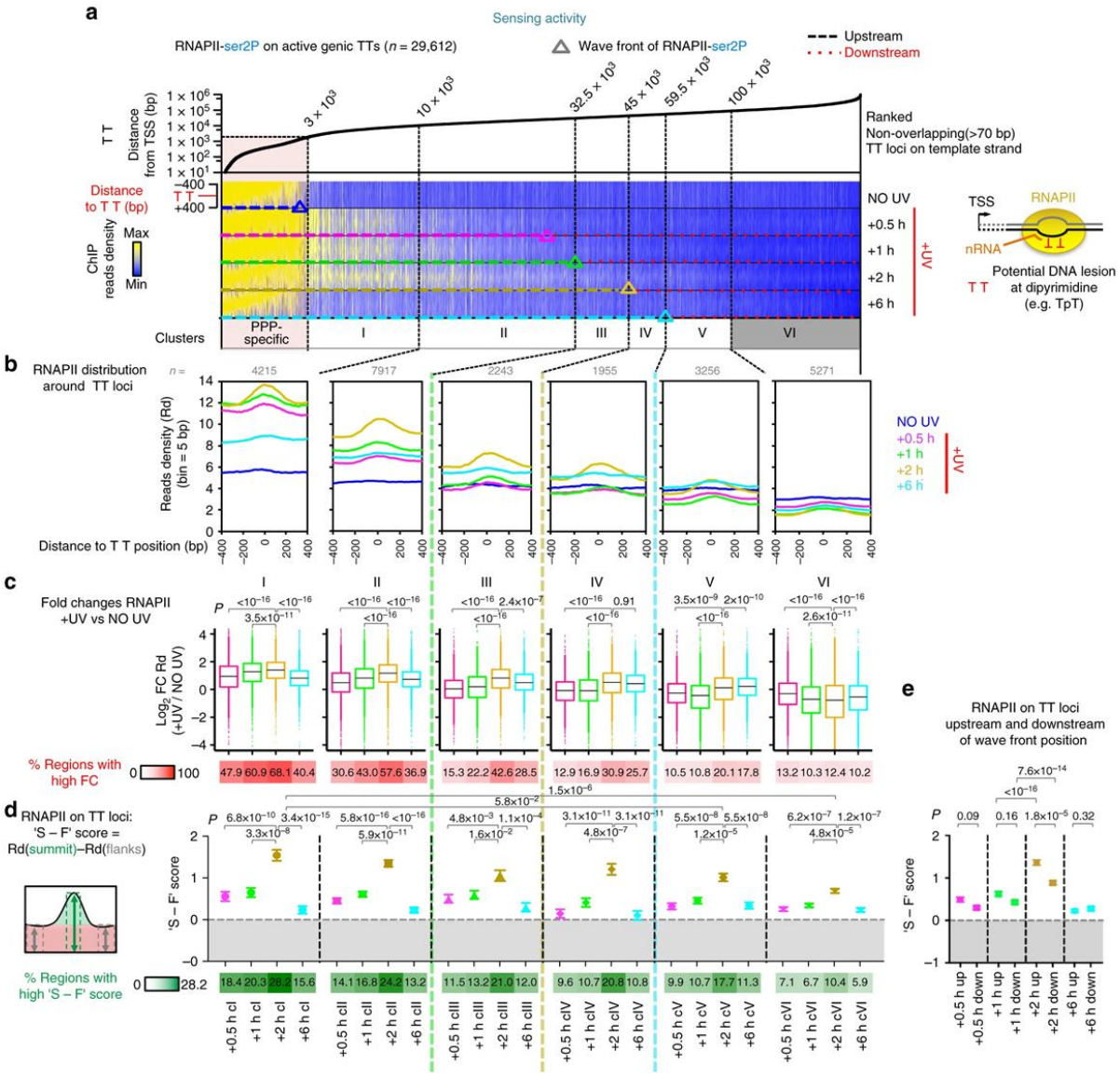


Figure 77 Sensing of DNA lesions by RNAPII is significantly increased in regions affected by the UVC-dependent elongation wave. (a) Heatmap visualization of NO UV and +UV RNAPII-ser2P signal along extended TT regions located at active mRNAs, sorted by increasing distance relative to TSS. TT loci were classified as upstream or downstream with respect to RNAPII-ser2P wave front positions, which pinpoints the border between de novo and pri-elongating RNAPII molecule populations. (b) Average profiles of RNAPII signal as described in (a). (c) Top panel: Box plots of log2 ratios of +UV RNAPII signal over NO UV RNAPII signal, as described in (a). Pairwise two-sided t-tests using Benjamini-Hochberg (BH) adjustment were performed, and corrected p-values are reported accordingly. Boxes refer to the first quartile, median and third quartile. Whiskers refer to the 10-90% interquartile range. Bottom panel: Percentages of boxplot data points with a value > 1 (for each boxplot). (d) Top panel: Average S-F scores (RNAPII read counts at the read density summit subtracted by the average RNAPII read counts at TT flanking regions) of all regions described in (a). Standard errors of the mean (SEM) are illustrated. Pairwise Wilcoxon rank-sum tests using Benjamini-Hochberg (BH) adjustment were performed, and corrected p-values are reported accordingly. Bottom panel: Percentages of high 'S-F' scores with a value > $\mathbf{average}[\mathbf{S}-\mathbf{F}]_{exon\ start}$TT + 3*$\boldsymbol{SD}[\boldsymbol{S}-\boldsymbol{F}]_{exon\ start}$. (e) Plot showing the comparison of average S-F

scores of all regions upstream (Up) and downstream (Down) of the respective wave-front position, for each +UV time point. Standard errors of the mean (SEM) are illustrated. Pairwise Wilcoxon rank-sum tests using Benjamini-Hochberg (BH) adjustment were performed, and corrected p-values are reported accordingly.

To quantify the RNAPII signal density around TT loci, the log_2 ratios between +UV and NO UV alignments were calculated (fold change (FC)) across all regions and plotted as boxplots (Figure 77). As depicted in Figure 77, in the +UV 2 h , up to 68.1 % (Cluster I, Figure 77 (c)) of the analyzed loci showed a higher enrichment of RNAPII signal, while RNAPII molecules exhibited a decreased stalling at TT regions at 6 hours after UV exposure. To precisely evaluate the distribution of RNAPII signal on TT dinucleotides, the difference between TT-counts and flanking region-counts for all the analyzed loci were calculated (S-F score) and plotted as a mean - standard error of the mean plots (Figure 77 (d)). S-F scores confirmed that the average RNAPII lesion stalling increases significantly in the genomic regions affected by the do novo elongation wave (Figure 77 (e)). Indeed, the fraction of damages detected by RNAPII gained the highest value of 28.2 % of the total analyzed loci in +UV 2 h - cluster I (Figure 72 (d)), thus validating that the UVC induced elongation wave release, increases the damage detection probability along active gene bodies. As a control region set, a set of Ensembl exon start positions were retrieved (see section 4.3.1), and analyzed as the TT loci (see above), to reveal that RNAPII signal is not specifically accumulated around exon start positions (because they are not preferentially enriched with di-pyrimidines), proving that the signal detected at exon start sites corresponds to the elongation wave that passes by, without specific stalling.

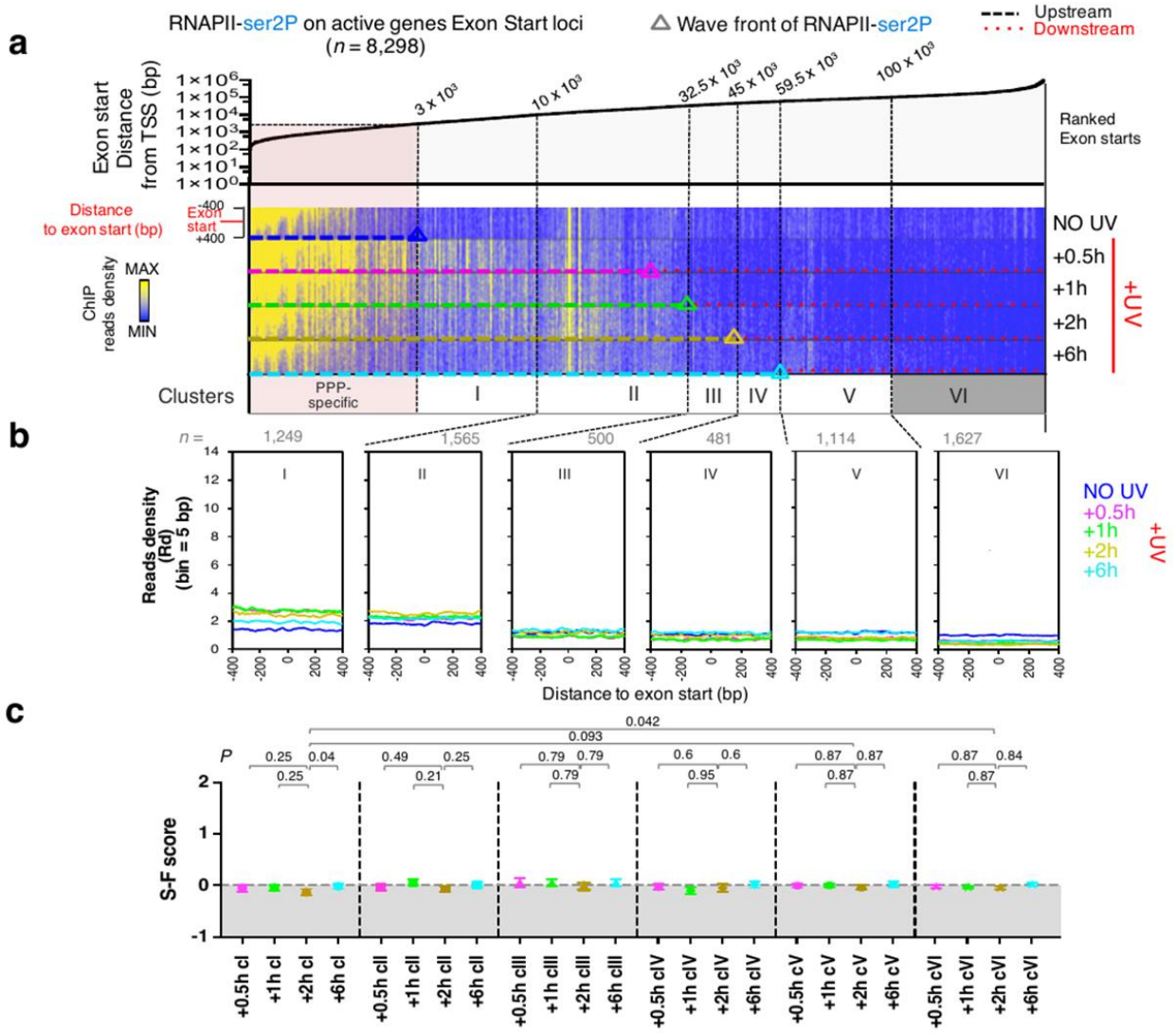


Figure 78 UVC-dependent elongation wave does not enhance exon specific stalling of RNAPII. (a) Heatmap visualization of NO UV and +UV RNAPII-ser2P signal at extended exon regions located at active mRNAs, sorted by increasing distance relative to TSS. Exons were classified as upstream or downstream with respect to RNAPII-ser2P wave-front positions, which pinpoints the border between de novo and pri-elongating RNAPII molecule populations. (b) Average profiles of RNAPII signal as described in (a). (c) Top panel: Average S-F scores (RNAPII read counts at the exon read density summit subtracted by the average RNAPII read counts at exonic flanking regions) of all regions described in (a). Standard errors of the mean (SEM) are illustrated. Pairwise Wilcoxon rank-sum tests using Benjamini-Hochberg (BH) adjustment were performed, and corrected p-values are reported accordingly. Bottom panel: Percentages of high 'S-F' scores with a value > $\mathbf{average}[\mathbf{S}-\mathbf{F}]_{exon\ start}$TT + 3*$\boldsymbol{SD}[\boldsymbol{S}-\boldsymbol{F}]_{exon\ start}$. Wilcoxon rank-sum tests using Benjamini-Hochberg (BH) adjustment were performed, and corrected p-values are reported accordingly.

To evaluate the importance of the de novo UVC-triggered elongation wave, as opposed to the already travelling RNAPII molecules (pri-elongating molecules) during the DNA-lesion

identification procedure, DRB VH10 RNAPII-ser2P in NO UV +30 min +DRB and +UV +30 min -DRB conditions (Figure 69) were analyzed as described above (Figure 77).

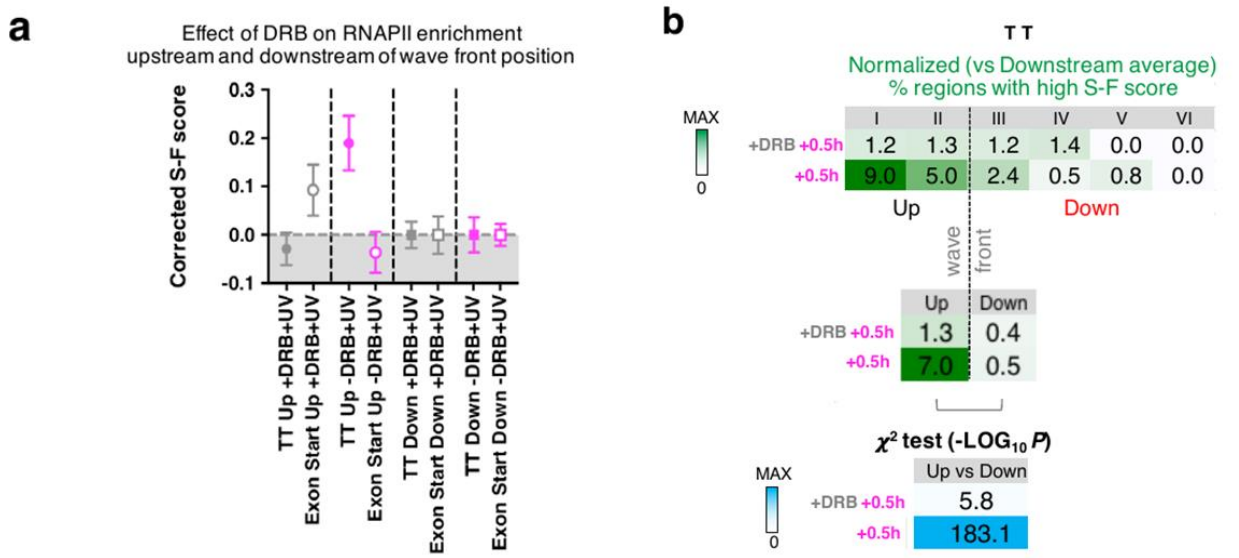


Figure 79 Inhibition of the de novo elongating RNAPII molecules, drastically reduces the RNAPII damage recognition process. (a) Average DRB-ChIP-seq S-F scores (pri-elongating) compared to no-DRB-ChIP-seq S-F scores (pri-elongating and de-novo elongating) for regions Upstream (Up) or Downstream (Down) of the theoretical wave front (Figure 63). S-F scores were corrected by inferring the average S-F score calculated for all Down loci. (b) Summary of the effect of DRB inhibition on differences in proportion of regions displaying high S-F scores (see Figure 77). Chi-square test ($\boldsymbol{X^2}$) compares the number of genes in Up and Down categories for each condition and determines if observed number of regions with high S-F scores differs from the expected values.

S-F scores were calculated for all the clusters described in table 8 and showed that in the absence of de novo elongated RNAPII molecules, the pri-elongating molecules were uniformly engaged in lesion detection along almost all the TT loci clusters.

4.3.12 De novo release of RNAPII elongation wave promotes DNA repair

To examine if the elongation wave-driven accumulation of RNAPII molecules in putative CPDs is also linked with preferential repair of those damages, a meta-analysis of XR-seq data (see materials and methods, section 2.10.10) was conducted. Specifically, XR-seq data of CPD damages at wild-type (WT) NHF1 skin fibroblasts, XP-C mutant cells (Xeroderma Pigmentosum) and CS-B mutant cells (Cockayne Syndrome) (see introduction, section 1.4) were retrieved by Gene Expression Omnibus (GEO), accession number GSE67941 using the sra toolkit (Leinonen et al., 2011). FASTQ files were generated using fastq-dump (sra-toolkit), and analyzed using the procedure described in section 4.1.3. An additional step of strand-specific alignment separation was performed to create forward and reverse alignment files. Heatmaps and average density profiles of strand-specific repair signal were generated to reveal patterns of preferential repair at specific loci, according to the examined dataset (see Figure 80).

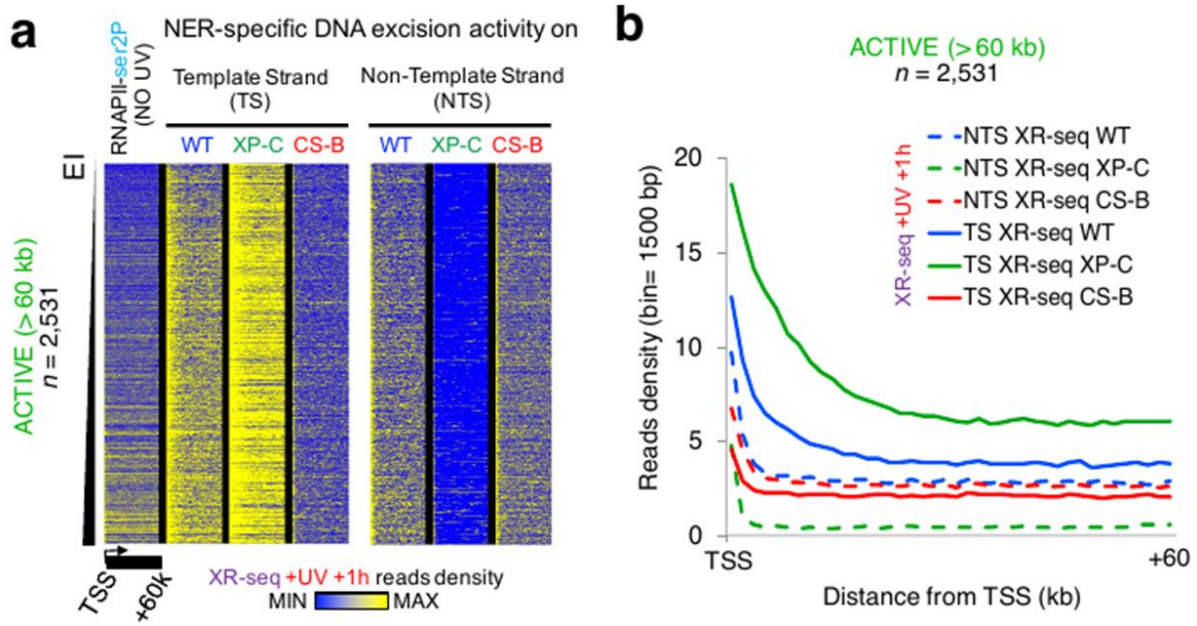


Figure 80 UVC-triggered wave release is coupled with increased repair activity in active mRNAs on both strands. (a) Strand specific heatmaps of XR-seq signal 1 h after irradiation at active transcripts over 60 kb (plotted from TSS to TSS +60 kb), ranked by increasing NO UV RNAPII-ser2P EI (see Figure 63), in WT cells (both TC-NER and GG-NER pathways are functional), in XP-C (GG-NER deficient cells), and in CS-B (TC-NER deficient cells). Heatmaps of RNAPII-ser2P ChIP-seq in VH10 cells in NO UV condition are shown (left). (b) Average density plots of XR-seq signal, as defined in (a).

In XP-C cells with non-functional GG-NER, TC-NER is favored, so the excision signal is preferentially accumulated at the transcribed strand of actively expressed genes (Figure 80). On the other hand, in CSB cells with non-functional TC-NER, GG-NER is favored, and the excision signal was shared between both transcribed and non-transcribed strands of active genes, depicting the stochastic function and global profile of the particular repair pathway. In WT cells, where both TC-NER and GG-NER are functional, excision signal is present in both transcribed and non-transcribed strands, with a higher prevalence at the transcribed strand of active genes, since most of the CPDs are repaired by the TC-NER mechanism in the first hour after UVC damage induction. Additionally, the repair signal is accumulated in an homogenous fashion at all active gene bodies, regardless of the steady state of RNAPII activity (EI values, Figure 63).

Next, using the merged strand alignments, the excision signal was summarized at TTs, using a new flanking space from −30 bp up to +30 bp relative to TT center. The analysis was performed as described in section 4.3.11, and the respective visualizations are depicted in Figure 81.

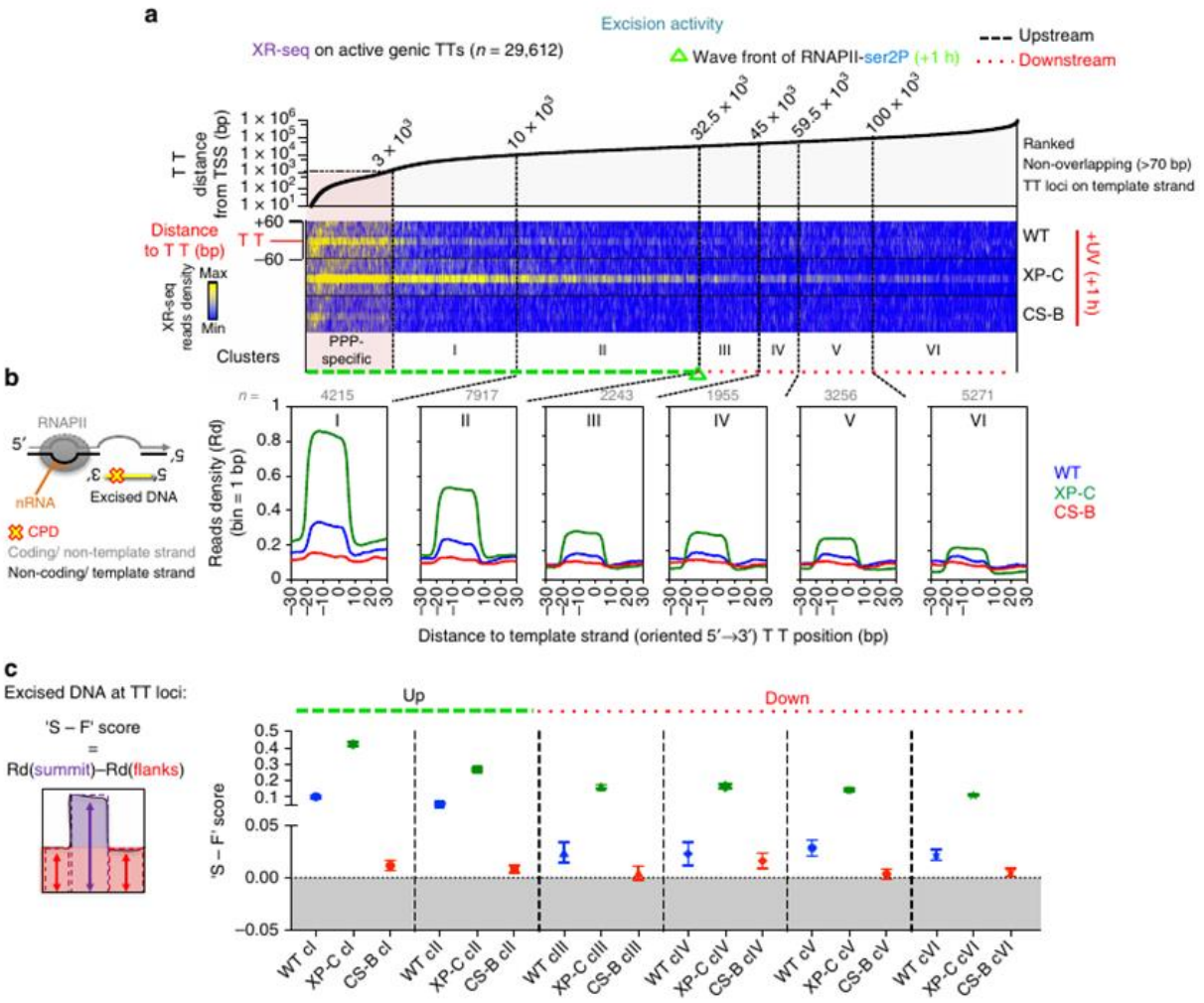


Figure 81 UVC-triggered de novo elongation wave, promotes NER repair of DNA lesions. (a) Heatmaps illustrating the XR-seq repair signal in NHDF (normal cell, both NER pathways functional), XPC (Xeroderma Pigmentosum cells, GG-NER deficient), and CSB (Cockayne Syndrome cells, TC-NER deficient) cell lines at extended TT loci (from -70 bp up to +70 bp relative to TT center) located in active genes, sorted by increasing distances relative to TSS. TT loci were classified as upstream or downstream with respect to RNAPII-ser2P wave front positions, which pinpoint the border between de novo and pri-elongating RNAPII molecule populations. (b) Average profiles of XR-seq signal as described in (a). (c) Top panel: Average S-F scores (XR-seq read counts at the TT read density summit subtracted by the average XR-seq read counts at TT flanking regions) of all regions described in (a). Standard errors of the mean (SEM) are illustrated. Pairwise Wilcoxon rank-sum tests using Benjamini-Hochberg (BH) adjustment were performed, and corrected p-values are reported accordingly.

The generated heatmaps of WT and XP-C XR-seq alignments showed an expected enrichment of repair activity in the clusters I and II, since these loci are affected by the released elongation wave during the first hour of UVC recovery (see section 4.3.12). Indeed, after the theoretical de novo wave front (clusters III, IV, V and VI), the WT and XP-C XR-seq signal accumulation in putative CPDs was uniformly lower (Figure 81), while in the CSB heatmap, all clusters showed a low, but stable pattern of GG-NER repair. Average density profiles of excision signal and S-F scores were also calculated as described in section 4.3.12, confirming the observations about

the wave-enhanced TC-NER repair on putative CPDs as opposed to the unaffected GG-NER pathway. Interestingly, even in the distal TTs, that are not affected by the wave-release, TC-NER activity was significantly detectable, because of the slowly travelling pri-elongating RNAPII molecules (Figures 77 and 79) that still detect downstream DNA-lesions.
Interestingly, given that XR-seq detects DNA excised damages at the time of the assay (Hu et al., 2015), and RNAPII molecules still stall at TT loci 2 hours after UV treatment (Figure 77), it seems that just a small proportion of CPDs are being repaired at 1 h post UV (time of excision), which implies that full recovery of all CPD lesions may last for several hours.

4.3.13 De novo release of RNAPII elongation wave restricts the mutation prevalence in the transcribed strand of all active genes

To examine if there is a causative effect between the UVC induced wave release and the mutation prevalence at actively transcribed genes, an analysis was conducted in datasets of clinical-relevant genotoxin-exposed tissues that have developed cancer, which were previously linked to NER activity (Alexandrov et al., 2013). These data revealed mutational asymmetries between the transcribed (TS) and non-transcribed strand (NTS), with lower mutational prevalence in the TS, implying a TC-NER dependent reduction of single nucleotide polymorphisms (SNPs). In particular, the analysis included skin melanoma, which is linked with high probability of UV (C(G) > T(A)) mutation, the hallmark of UV-exposed genomes (Helleday et al., 2014; Lehmann, 2000; Pleasance, Cheetham, et al., 2010; You et al., 2001), and smoke (G(C) > T(A)) mutation, the most frequent smoking adduct-generated mutation, repaired by TC-NER (Alexandrov et al., 2013; Haradhvala et al., 2016; Pleasance, Stephens, et al., 2010) (Figure 82).

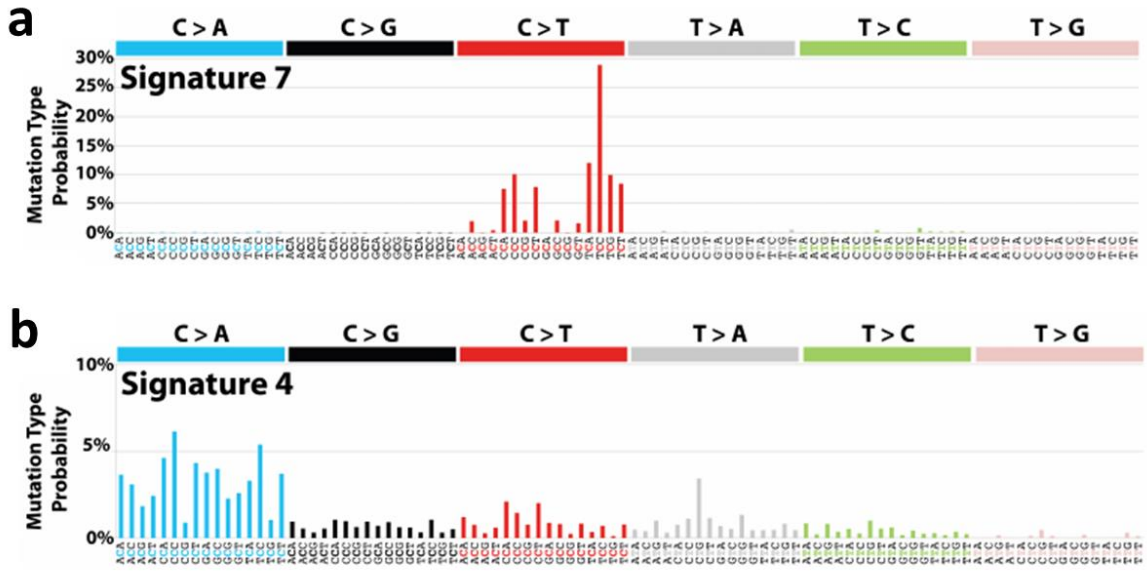


Figure 82 Patterns of substitutions for Signature 7 (a) and 4 (b) described in (Alexandrov et al., 2013). Mutational signatures are based in the trinucleotide frequencies of the human genome. The figure is adopted by the supplementary information of (Alexandrov et al., 2013).

Human melanoma and lung adenocarcinoma merged genome-wide maps of validated mutations were downloaded from ftp://ftp.sanger.ac.uk/pub/ cancer/AlexandrovEtAl (table 9).

Table 9 Summary of the mutation datasets of human melanoma and lung adenocarcinoma from (Alexandrov et al., 2013).

Cancer Type	Total Samples	Whole Genomes	Whole Exomes	Original Mutations	Removed Mutations	Filtered Mutations	Mutations Analyzed
Lung	660	24	636	1,963,661	117,685	1,845,976	1,797,343
Melanoma	396	-	396	340,592	47,986	292,606	292,046

Whole genome sequencing (WGS) and whole exome sequencing (WES) samples were analyzed separately. Melanoma was scanned for the UV-specific C > T (or the reverse complement G > A) substitutions, while the lung adenocarcinoma dataset was scanned for the TC-NER-specific G > T (or the reverse complement C > A) substitution. Next, using *bedtools* and *getfasta* (Quinlan & Hall, 2010), the respective trinucleotides were extracted and filtered in order to keep the most probable characterized trinucleotide events (Figure 82); for melanoma T(C)C > T(T)C and G(G)A > G(A)A, while for lung adenocarcinoma T(G)G > T(T)G and C(C)A > C(A)A, to generate trinucleotide BED-like files (Figure 83).

```
chr10   85984121    85984124    C->T    TCC
chr1    103345304   103345307   C->T    TCC
chr11   44286532    44286535    C->T    TCC
chr1    152281172   152281175   C->T    TCC
chr11   67203465    67203468    C->T    TCC
chr1    215853634   215853637   C->T    TCC
chr17   46868899    46868902    C->T    TCC
chr17   67041352    67041355    C->T    TCC
chr19   40354272    40354275    C->T    TCC
chr19   51920676    51920679    C->T    TCC
```

Figure 83 BED-like file of mutation trinucleotides for Melanoma datasets.

Mutations were further separated to template strand (TS) and non-template strand (NTS) trinucleotides based on the Watson-Crick strand reference and the strand orientation of the host transcript (Figure 84).

a

Select mutations according to (a)

↓

Generate BED files:
Aggregate for whole exome (WES) or whole genome (WGS) samples
→for each group
→ for each cancer type

	Melanoma		Lung Adeno			
	WES		WES		WGS	
	C > T	G > T	C > T	G > T	C > T	G > T
TS	26,048	876	1,958	4,632	3,000	6,311
NTS	30,470	805	1,674	5,973	3,122	10,671

Retrieve nRNA expression levels from normal cells →

↓

Map on annotated genes ranked by expression levels

b

Mutation name (substitution)	C > T		G > T	
Trinucleotide context	T(C)C > T(T)C		T(G)G > T(T)G	
Maximum frequency in	Melanoma		Lung Adenocarcinoma	
Reference Genome	C:G	G:C	C:G	G:C
Cancer Genome	T:A	A:T	A:T	T:A
Gene orientation	+ -	+ -	- +	- +
Mutation strand TS=Template strand NTS=Non-template strand	NTS TS	TS NTS	NTS TS	TS NTS
Select from annotated substitutions on Watson strand (from Alexandrov et al.): TS	C > T on – genes G > A on + genes		G > T on – genes C > A on + genes	
Select from annotated substitutions on Watson strand (from Alexandrov et al.): NTS	C > T on + genes G > A on - genes		G > T on + genes C > A on - genes	

Figure 84 (a) Procedure used to generate mutational profiles on VH10 active genes over 60 kb (from TSS to TSS +60 kb) using human melanoma and lung adenocarcinoma datasets [ref]. (b) Table summarizing the pipeline used to extract the most common UV-specific and smoking-specific mutation trinucleotides of human melanoma and lung adenocarcinoma tumors.

To analyze the mutation trinucleotides based on the nascent expression activity gene-status, non-stressed BRU-seq data from HF1 cells (human skin fibroblasts) (Andrade-Lima et al., 2015) and GRO-seq data from MRC5VA cells (Williamson et al., 2017b) were download by GEO with accession number GSE65985 and GSE91010 respectively, as sra files, and were analyzed using the methodology described in section 4.1.4. RefSeq mRNA expression counts were converted to $log_2 RPKM$ values, and gene activity was estimated by setting a vertical line that dichotomizes the RPKM bimodal distribution (see Figure 49 for example), separating genes to inactive and active references, for each examined cell line.

Active genes were further divided in three categories of the same size where Hi, Med, and Lo denote high, medium and low nascent expression levels respectively, and only the annotated references over 60 kb were retained, and sorted in a descending expression order. Melanoma trinucleotide signal was summarized at HF1-sorted transcripts, while lung adenocarcinoma trinucleotide signal was summarized at MRC5VA-sorted transcripts, by averaging the number of mutations detected in all analyzed samples (for WES or WGS datasets) over the examined references, and by considering a region of 1 Mb of DNA (Mutation Prevalence = number of mutations counted per Mb per sample). As WES data is not linear because of the heterogenous exon density across transcripts, mutation prevalence values were further corrected for each examined genomic window, as a function of the relative exon density measured in each region (see formula in Figure 85).

$$Mutational\ prevalence\ (P)(per\ Mb\ per\ Sample)$$

$$P(WES) = \frac{C \times Mb\_Norm\ \times WES\_Corr}{S},$$

$$P(WGS) = \frac{C \times Mb\_Norm}{S},$$

$$C = Count\ all\ hits\ for\ each\ mutation\ category\ in\ the\ defined\ genic\ regions\ (R),$$

$$Mb_{Norm} = Length\ normalization\ of\ the\ defined\ regions\ in\ 1\ Mbp\ genomic\ space\ \left(\frac{1{,}000{,}000}{length(R)}\right),$$

$$WES_{Corr} = Average\ coverage\ correction\ of\ exons\ in\ R\ \left(\frac{100}{\%coverage\ of\ exons\ in\ R}\right),$$

$$S = Number\ of\ samples$$

Figure 85 Mutational prevalence calculation.

Heatmaps and average density profiles of mutation prevalence were generated in a strand-specific fashion depicted below (Figures 86 and 87).

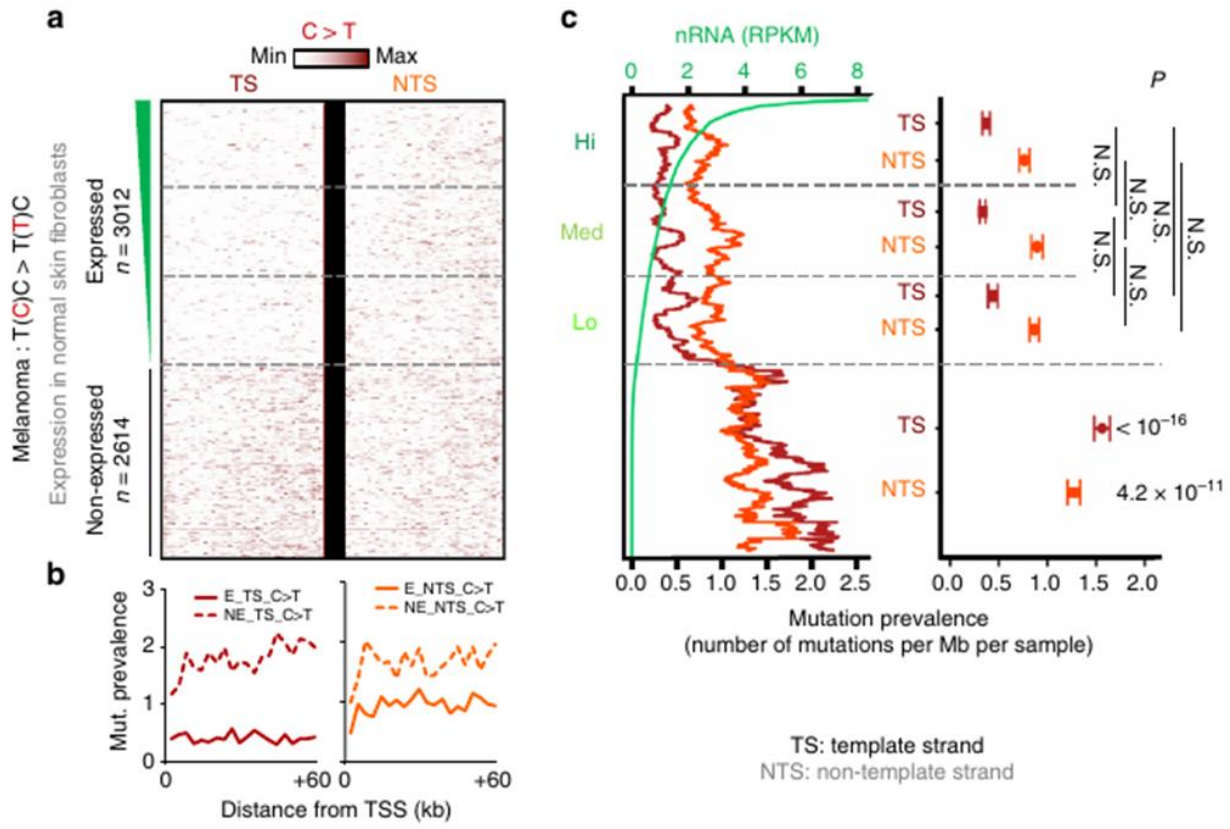


Figure 86 Low levels of mutation prevalence in actively transcribed genes in Melanoma samples. (a) Heatmaps illustrating the mutational density of UV (C > T) at active genes over 60 kb (plotted from TSS to TSS+60 kb), in a strand specific manner (template (TS) and non-template (NTS) strand separately). mRNAs are classified by activity levels, based on the respective normal skin fibroblasts nascent RNA levels (E for expressed, and NE for non-expressed). (b) Average mutation prevalence profiles across gene-bodies over 60 kb (plotted from TSS to TSS+60 kb), for "E" (solid line) and "NE" (dashed line) transcripts. (c) Left panel: Per gene average mutation prevalence profiles in the same order as described in (a). Curve correction was performed using a moving average with n=200. RPKM expression levels are depicted by the green curve. Right panel: Pairwise comparisons of average mutation prevalence

between TS and NTS, for Hi, Med, and Lo expression categories and NE transcripts. For each comparison a two-sided Wilcoxon rank-sum test using the BH adjustment was applied. N.S indicates Non-Significant P-value (> 0.05).

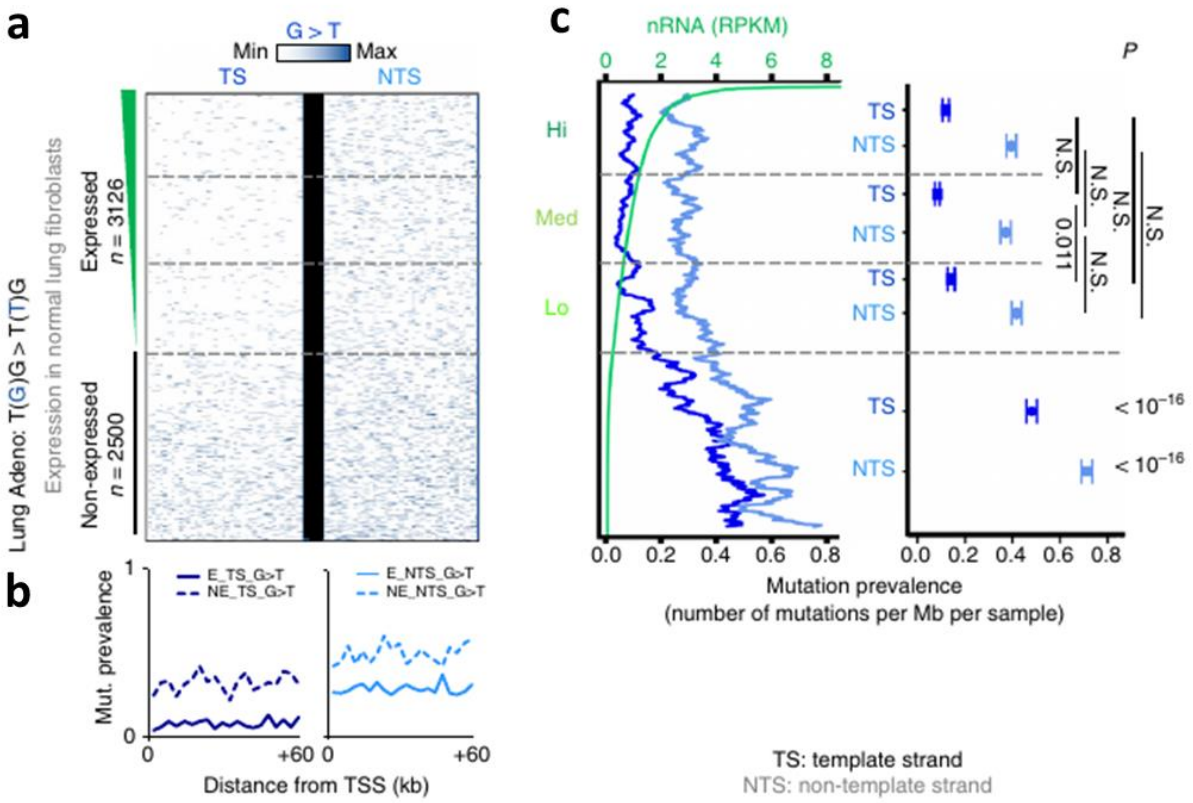


Figure 87 Low levels of mutation prevalence in actively transcribed genes in Lung Adenocarcinoma samples. (a) Heatmaps illustrating the mutational density of cigarette smoking (G > T) at active genes over 60 kb (plotted from TSS to TSS+60 kb), in a strand specific manner (template (TS) and non-template (NTS) strand separately). mRNAs are classified by activity levels, based on the respective normal lung fibroblasts nascent RNA levels (E for expressed, and NE for non-expressed). (b) Average mutation prevalence profiles across gene bodies over 60 kb (plotted from TSS to TSS+60 kb), for "E" (solid line) and "NE" (dashed line) transcripts. (c) Left panel: Per gene average mutation prevalence profiles illustrated in the same order as described in (a). Curve correction was performed using a moving average with n=200. RPKM expression levels are depicted by the green curve. Right panel: Pairwise comparisons of average mutation prevalence between TS and NTS, for Hi, Med, and Lo expression categories and NE transcripts. For each comparison a two-sided Wilcoxon rank-sum test using the BH adjustment was applied. N.S indicates Non-Significant P-value (> 0.05).

To apply statistical comparisons between prevalence score distributions, and to avoid inclusion of multiple zero data points (mutations are rare genomic events), the sparse mutation data matrices were aggregated over row groups of 15 genes within each expression cluster.
The particular visualizations (Figures 86 and 87) revealed detailed insights of the tumors' genomic mutation landscape. Specifically, the localization of mutations in gene-bodies was determined precisely to uncover a uniform pattern, even in the more distal parts of long genes (Figures 86 and 87).

As expected, both heatmaps and average profiles confirmed that NTS of actively transcribed genes are more prone to mutation forming than the TC-NER protected TS (Alexandrov et al., 2013; Haradhvala et al., 2016; Pleasance, Stephens, et al., 2010)(TS < NTS), while in both strands, lower mutation rates were observed in expressed genes (E) in contrast to the non-expressed genes (NE) (Figure 86 and 87). Strikingly, analysis of the mutational prevalence along the different expression groups in the two transcriptomes (skin and lung fibroblasts) revealed homogenous levels of genetic alterations for both DNA strands across the whole gene bodies, suggesting that the widespread and uniform release of RNAPII upon genotoxic stress impacts significantly on the mutation landscape of the active transcriptome.

4.3.14 UV-dependent increase of chromatin accessibility is paralleled by RNAPII transition into transcription elongation

To demonstrate the functional advantages that are linked with the phenomenon of the chromatin accessibility expansion during the early recovery after UV-induced stress, an integrative analysis of CAGE-seq data of normal dermal and skin fibroblast primary cells (materials and methods, section 2.10.9) and VH10 ATAC-seq NO UV and +UV +2 h, and VH10 RNAPII-ser2P NO UV and +UV +1 h was performed. The specific analysis was conducted using the annotation described in section 4.3.5, in order to examine the patterns of NGS signal along the transcriptional directionality of bidirectional genes, asPROMPTs and enhancer elements. Bidirectional genes and unidirectional genes-asPROMPT pairs were sorted by their inter-TSS distance, defined as the distance separating the significant CAGE summits detected on each strand (section 4.3.5). The heatmap and average profile visualizations of NGS signal were generated using the methodologies described in section 4.1.4.

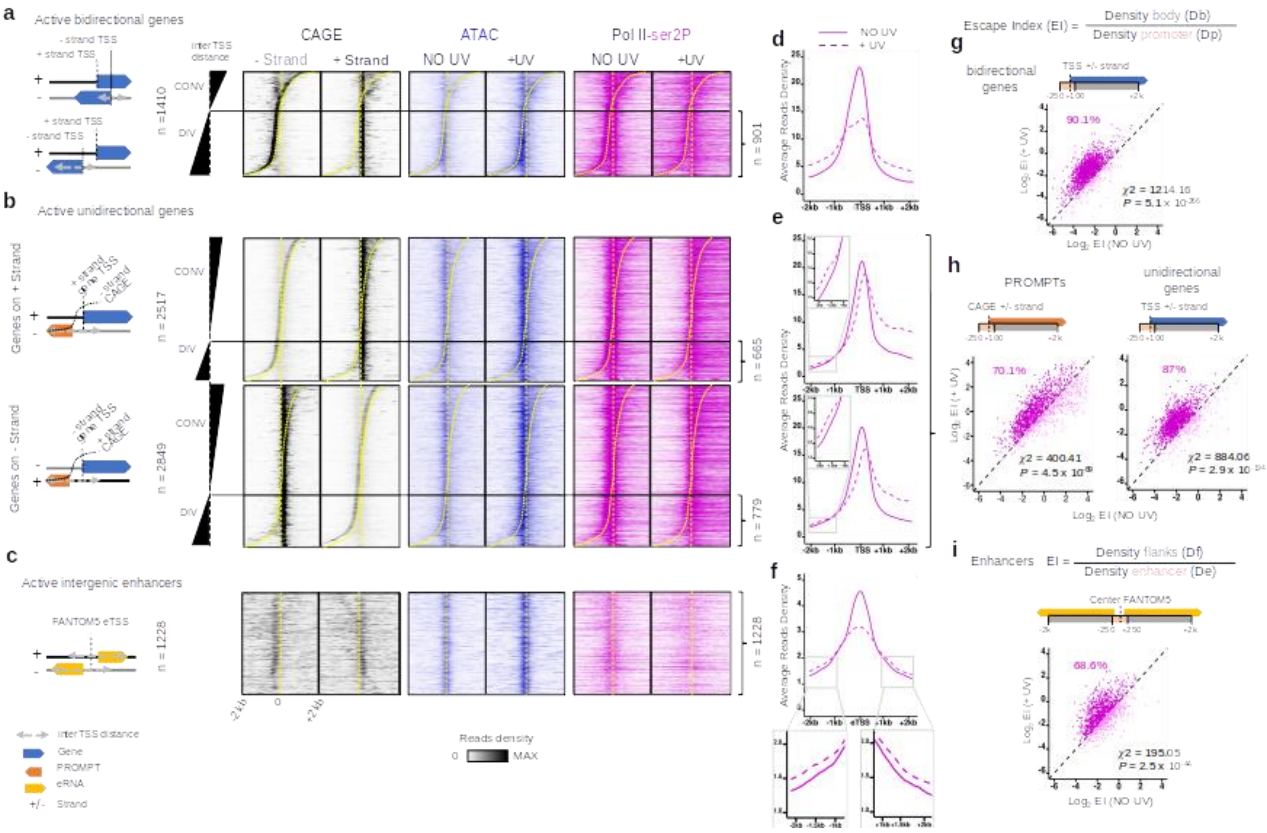


Figure 88 PPP release of RNAPII upon genotoxic stress. (a) Heatmap illustration of CAGE signal (black, strand specific analysis, two strands separated), ATAC-seq signal (blue, NO UV and +UV 2 h), and RNAPII-ser2P signal (purple, NO UV and +UV 2 h) at extended TSS regions of active bidirectional TSS pairs (+/- 2kb). TSS pairs are sorted by interTSS distance from the most convergent to the most divergent pair. (b) Heatmap illustration same as in (a) but for active convergent and divergent unidirectional TSS/ asPROMTs pairs. PROMPTs (orange arrow) are transcribed in the antisense direction relative to mRNAs, from either the minus strand (Upper panel, see CAGE minus strand signal) or the plus strand (Bottom pane, see CAGE plus strand signal). The straight dashed lines denote the mRNA TSS position, while the sigmoidal dashed line indicate the asPROMPT position. TSS pairs are ordered by interTSS distance from the most convergent to the most divergent pair. (c) Heatmap illustration same as in (a) but for active intergenic enhancers. (d), (e) Average profiles of NO UV and +UV RNAPII-Ser2P categories defined in (a) and (b), but only for divergent TSS-pairs. Zoomed illustrations are provided accordingly. (f) Average profiles same as in (d) and (e), but for all active intergenic enhancers. (g), (h), and (i) Scatter plots of pairwise Escape Index (EI) comparisons between NO UV and +UV RNAPII-ser2P, for categories indicated in (d), (e), and (f). Proportion of elements with higher escape in +UV condition are reported. Chi-square tests (X^2) between active and inactive elements of each annotation category were performed to determine if the observed number of elements (active elements) with ΔEI > 1 differs from the expected values (inactive elements ΔEIs).

To study the transcriptional dynamics at play in both directions (strands), avoiding the signal interference between overlapping references, the focus was addressed on the non-overlapping pairs of TSSs (divergent bidirectional TSSs, and divergent unidirectional TSS/ asPROMPT TSS pairs). Using the particular annotation set-up, the escape index of each annotated transcript was calculated as described in section 4.3.6, including the asPROMPTs, while the enhancer

elements' EI was calculated in a similar fashion: The read density of enhancer bodies is calculated as the average of two region flanks ranging from 2 kb up to 100 bp upstream of eTSS and from 100 bp up to 2 kb downstream of eTSS, while the read density of enhancer promoters is calculated at the region ranging from 100 bp upstream to 100 bp downstream of eTSS. The calculated EIs were visualized using scatter plots as described in section 4.3.6 (Figure 88 (g-i)). The particular illustrations revealed that the UV-dependent increase in chromatin accessibility depicted in figures 73 and 75 is corroborated by the transition of RNAPII into transcription elongation, at all the examined actively transcribed elements, as depicted by the reduction of RNAPII signal at promoters and the parallel increase of RNAPII signal at gene bodies. Quantifications using EI confirmed that the RNAPII elongation increases in response to UV-induced stress for the majority of actively transcribed elements (90.1% of bidirectional promoters, Chi-square test P = 5.1×10^{-266}, 70.1 % of asPROMPTs, Chi-square test P = 4.5×10^{-89}, and 68.6 % of enhancers, Chi-square test P = 2.5×10^{-44}).

4.3.15 Genome coverage analysis of nRNA-seq data reveals global inhibition of transcription upon early recovery from UVC-stress induction

To examine the percentage of the repressed transcription activity along the human genome during UVC recovery, all hg19 canonical chromosomes were split to 50 bp segments, and alignment-depth normalized counts were calculated per bin using VH10 nRNA-seq NO UV and +UV 2h datasets (see materials and methods, section 2.10.5). Summarization of all the bins where $log_2 FC$ (+UV / NO UV)) < 0, relative to the total number of bins, revealed that the 63.65 % of the transcribed genome shows inhibition of transcription, a result that is in agreement with other studies(Andrade-Lima et al., 2015; Bugai et al., 2019b; Lavigne et al., 2017; Magnuson et al., 2015; Williams et al., 2015), while a local increase in nRNA signal downstream of all active TSSs is detected during UV-recovery (Figures 67 and 70).

4.3.16 Treatment with DRB retains the RNAPII signal in PICs during early recovery from UVC-induced stress

To discover if RNAPII is able to be recruited at active promoters upon UV, +DRB RNAPII-hypo ChIP-seq experiments with conditions NO UV +DRB 2 h, +UV 2 h DMSO , +UV 4 h DMSO and +UV 4 h +DRB 2 h, (see materials and methods, section 2.10.4) were generated and analyzed using the methodology described in section 4.1.4, generating heatmaps and average profiles of NGS signal (Figure 89) using the extended TSS-pairs reference described in section 4.3.5.

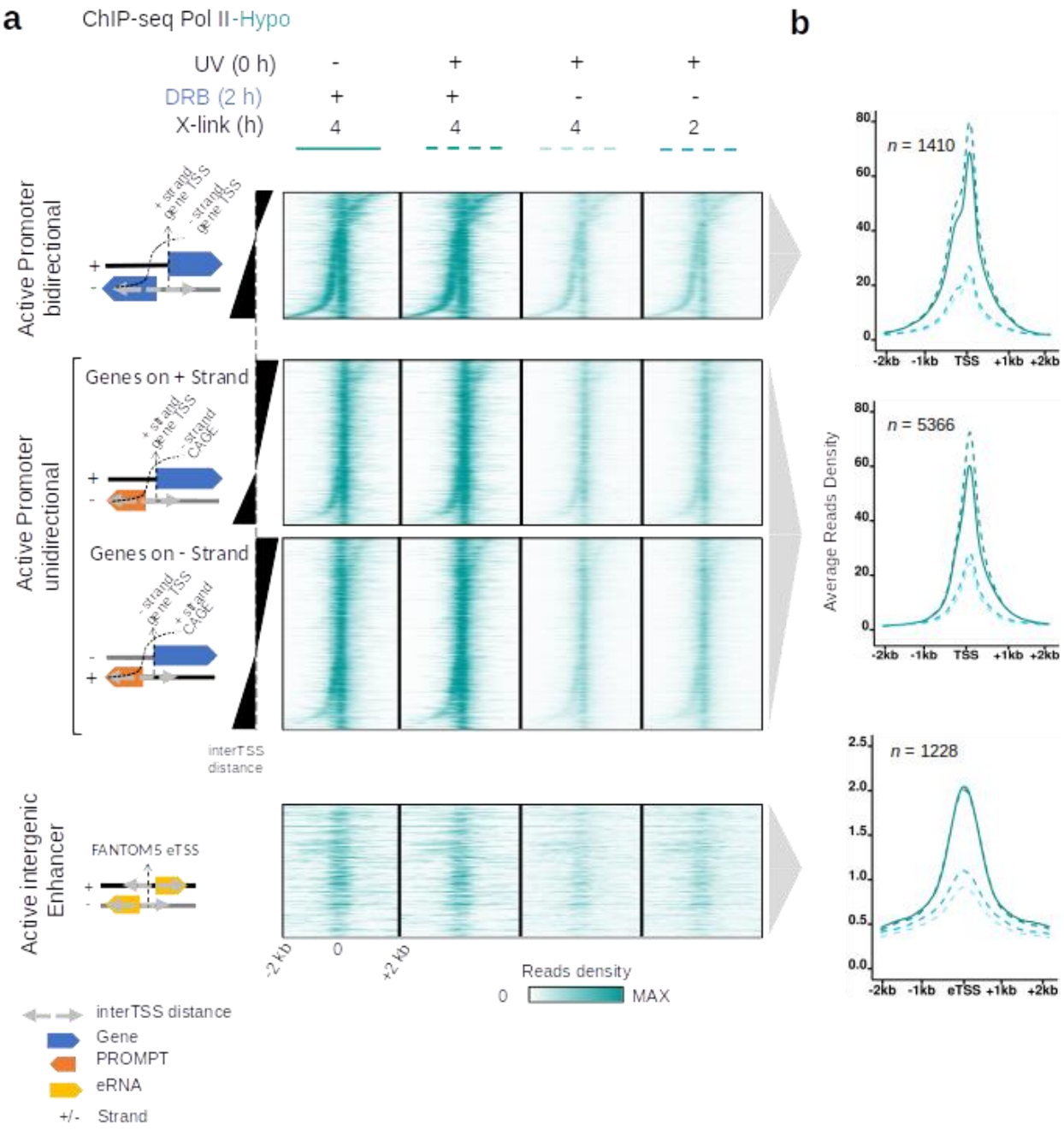


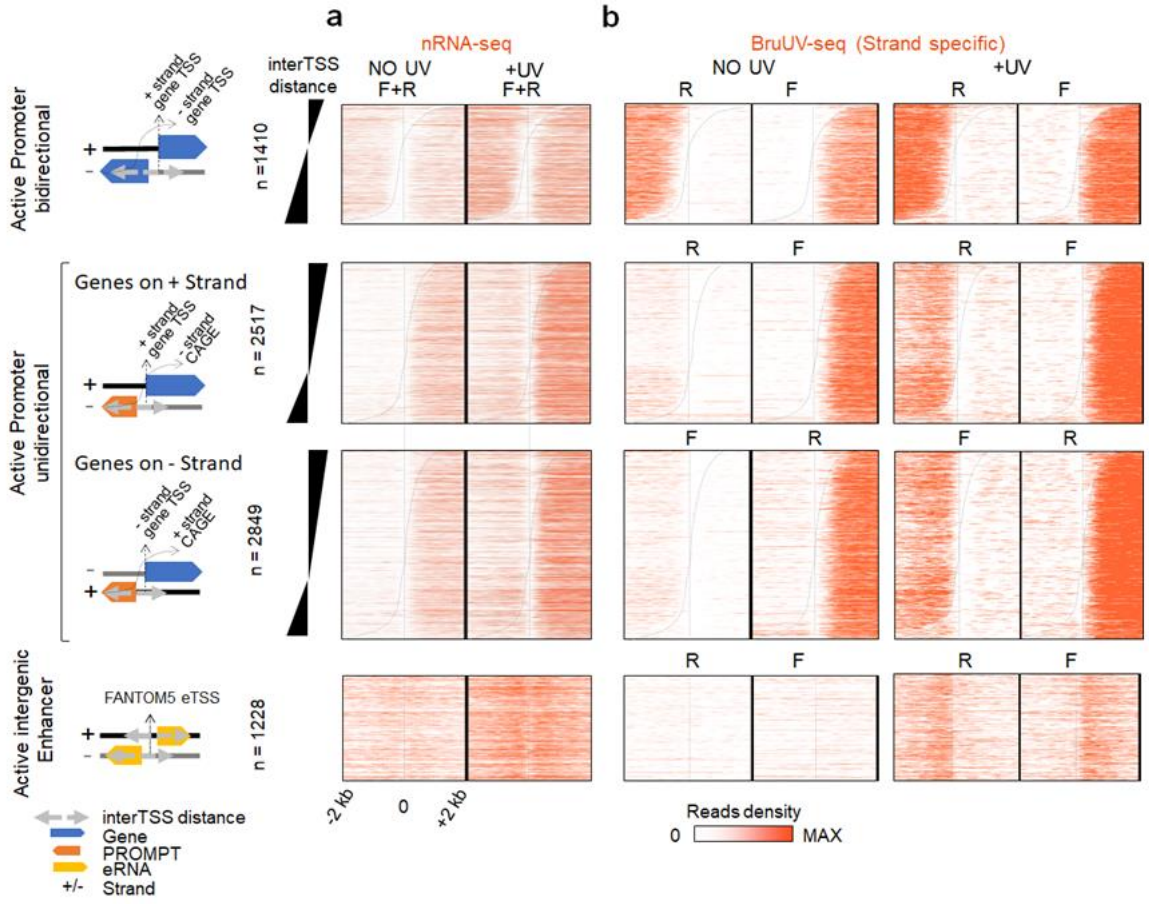
Figure 89 Inhibition of RNAPII PPP release retains pre-initiating RNAPII binding at active promoters after UVC-damage induction. (a) Heatmap illustration of +DRB RNAPII-hypo ChIP-seq experiments as described in Figure 88 (a-c). (b) Average profiles of +DRB RNAPII-hypo ChIP-seq experiments as analyzed in (a).

This analysis revealed that in the +UV 2 h DMSO and +UV 4 h DMSO conditions, only a minimal level of RNAPII-hypo signal is detected at actively transcribed promoters, as opposed to +UV 4 h +DRB 2 h condition, in which a significant retain of RNAPII-hypo signal is detected at all active elements (Figure 89). The recovery of RNAPII-hypo signal was even more evident when comparing the +UV 4 h +DRB 2 h with the NO UV +DRB 2 h condition, where the level of

RNAPII-hypo signal is rescued at all the analyzed TSS categories (Figure 89). Consequently, by preventing the UVC triggered transition of RNAPII molecules from PPP sites into active elongation at 2 h after UV induction, a time-point where the RNAPII-hypo level was almost non-detectable (Figure 75), a latent and continuous de novo recruitment of RNAPII-hypo molecules in PICs is revealed. This result clarifies the previously detected depletion of RNAPII-hypo (Figure 75, figure 89 -DRB samples), suggesting that upon early recovery from UVC-stress induction, new molecules of RNAPII are recruited at PPPs, and by the time of recruitment, they are released to gene bodies in order to increase the damage-scanning activity of the cell.

4.3.17 Increased nascent RNA synthesis from active promoters during early recovery from UVC-induced stress

Since UVC stress does not inhibit the initiation of transcription, nor the escape of RNAPII from PPP into productive elongation, the next step was to examine whether these phenomena are coupled by increased production of newly synthesized RNA around the genomic regions of active TSSs. To address this hypothesis, VH10 nRNA-seq datasets at NO UV and +UV 60 minutes conditions, as also the strand specific HF1 BruUV-seq datasets at NO UV and +UV 30 minutes from (see materials and methods, section 2.10.6) were analyzed using the methodology described in section 4.1.4, to produce heatmaps and average profiles of NGS signal using the TSS references described in section 4.3.5.


a
b
nRNA-seq
BruUV-seq (Strand specific)
interTSS distance
NO UV
+UV
F+R
R
F
Active Promoter bidirectional
+ strand gene TSS
- strand gene TSS
n =1410
Genes on + Strand
- strand CAGE
n = 2517
Genes on - Strand
n = 2849
Active intergenic Enhancer
FANTOM5 eTSS
n = 1228
-2 kb
0
+2 kb
Reads density
MAX
interTSS distance
Gene
PROMPT
eRNA
Strand

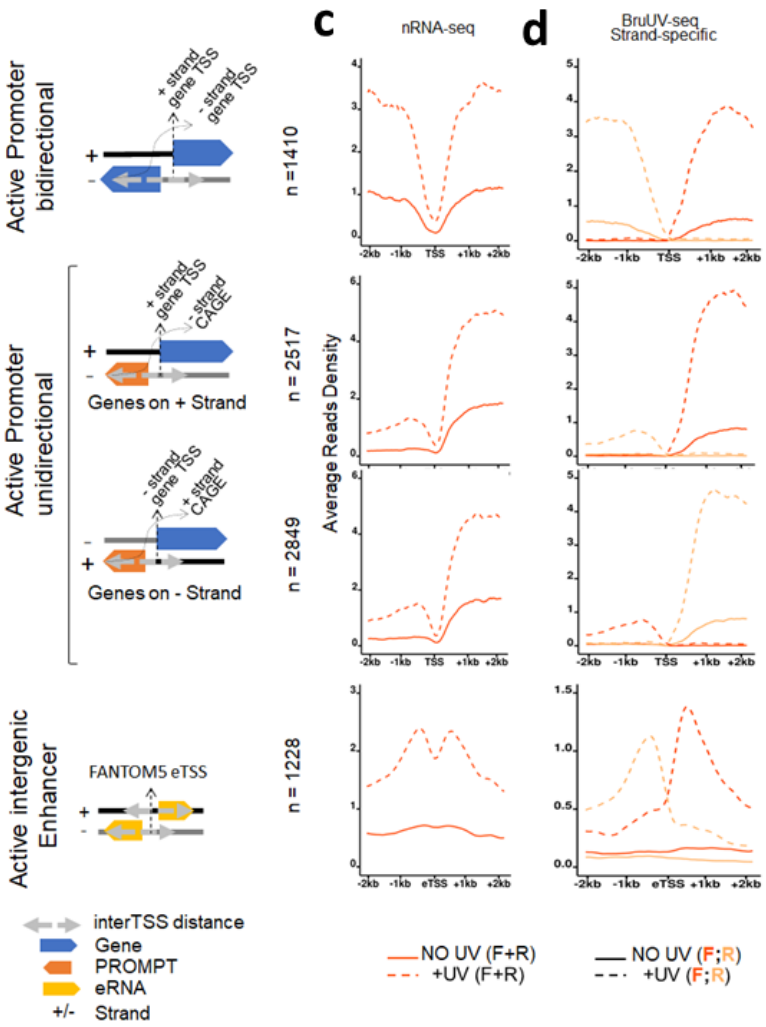


Figure 90 De novo UVC-derived nascent RNA synthesis at all active promoters. (a) Heatmap illustration of NO UV 1 h and +UV 1 h pre-DRB nRNA-seq signal as depicted in Figure 88 (a-c). (b) Strand specific heatmap illustration of BruUV-seq NO UV 30 min and +UV 30 min BruUV-seq as depicted in Figure 88 (a-c). Forward (F, +) and reverse (R, -) strands are visualized separately. (c), (d) Average profiles of nRNA-seq signal corresponding to (a) and (b) respectively.

The resulting visualization (Figure 90) replicates the previously shown global increase of EU-labelled and Bru-labelled RNA signal in the first kilobases of active genes (see section 4.3.7) and supports the hypothesis that this increase could arise by the elevated RNAPII initiation at active TSSs (Figure 89), as previously proposed (Magnuson et al., 2015). Specifically, at TSSs corresponding to unidirectional and bidirectional elements, nRNA level is significantly increased towards the mRNA direction, but also towards the antisense direction, due to the asPROMPT transcription activity. In the same fashion, active enhancers show a global increase in nRNA synthesis, towards both directions relative to the enhancer TSS (Figure 90). The later observations regarding the short-transcribed elements (asPROMPTS and enhancers), combined with the similar findings at active mRNAs (Figure 63), support the hypothesis that active promoter regions are transcribed “de-novo” during the early UVC-recovery process.

4.3.18 Continuous transcription initiation during UVC recovery is coupled to nascent RNA synthesis

To further verify that the initiation of transcription is productive and uninterrupted after UV radiation in the genomic regions proximal to the different classes of TSSs, localization and quantification of start-RNAs was performed. The particular procedure is informative about the magnitude of the engaged RNAPII production within the initially transcribed sequence (~ 100 first nucleotides) (Williams et al., 2015). For this purpose, VH10 start-RNA synthesis experiments were conducted using NO UV / + DRB / T 2.5h, + UV / - DRB / T 2.5h, + UV / + DRB / T 2.5h and + UV / + TRP / T 2.5h conditions (see materials and methods, section 2.10.8).

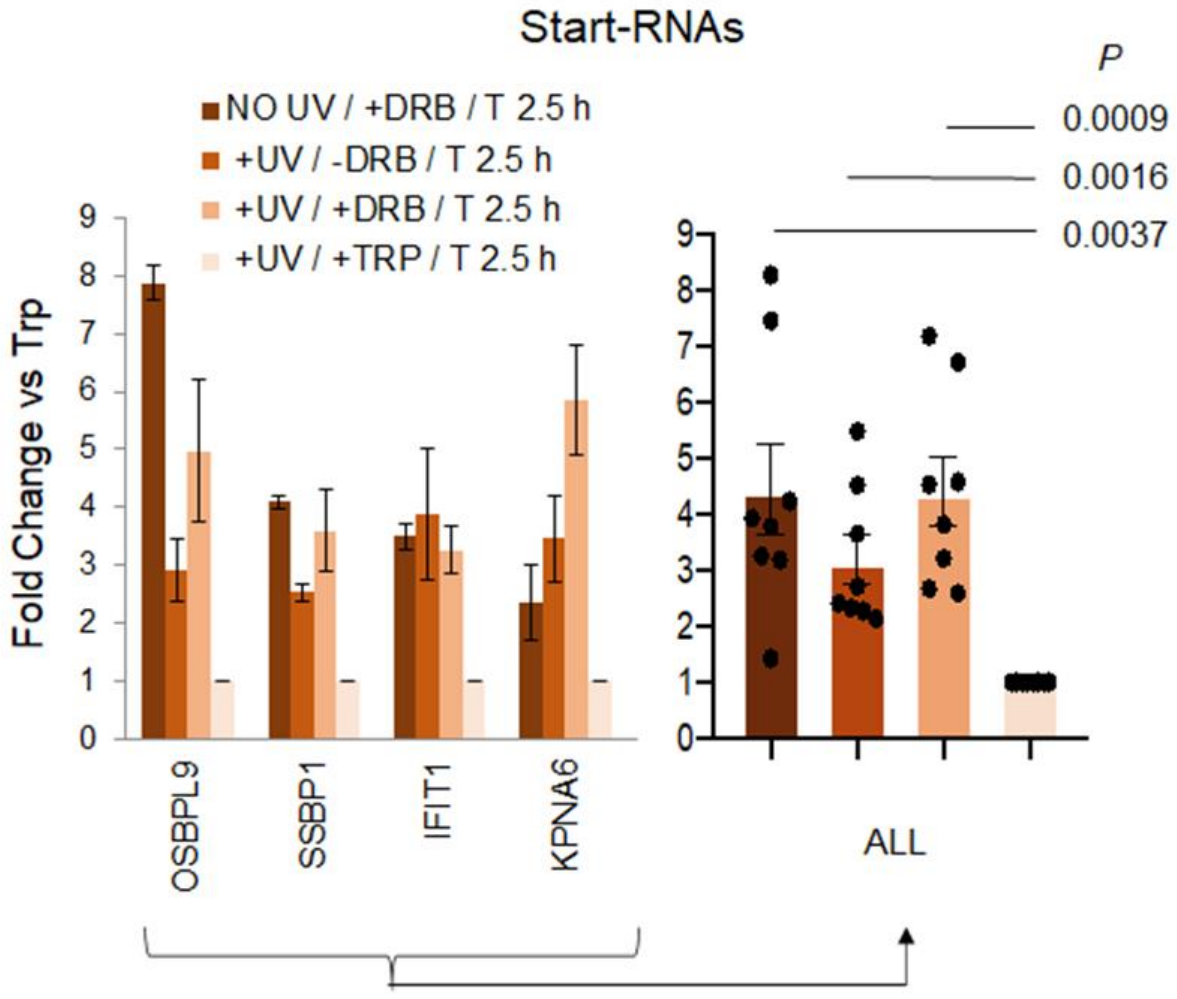


Figure 91 Start-RNA quantification using qPCR. Bar graphs depicting the Fold Change (FC), for each tested gene (left panel), and for all genes together (average of all genes, right panel). Standard Errors of

the Mean (SEM) are illustrated accordingly. Two-sided Student’s t test are applied as indicated (right panel) and p-values are reported

The resulting quantification of the qPCR analysis showed that the levels of start-RNAs after UV exposure were similar to those of non-irradiated cells (Figure 91). More precisely, DRB-dependent inhibition of RNAPII release from PPP sites did not prevent the detection of significant levels of start-RNAs after UV exposure, as opposed to the clear reduction in start-RNA levels following the TRP-dependent inhibition of transcription initiation (Figure 91, two sided Student's t test p-value = 0.0037 compared to "NO UV / + DRB / T 2.5h", p-value = 0.0016 compared to "+ UV / - DRB / T 2.5h" and p-value = 0.0009 compared to "+ UV / + DRB / T 2.5h ”), thus showing that after UV exposure, both transcription initiation and the corresponding RNA synthesis take place in the respective genomic regions.

4.3.19 Balanced level of RNAPII-hypo at PICs favors homogeneous TC-NER function

To clarify the functional implications of continuous transcription initiation during UV recovery, XR-seq data (analyzed in section 4.3.12), and specifically CPD XP-C datasets that precisely and exclusively pinpoint the location and levels of transcription-dependent repair (TC-NER pathway) were reanalyzed in the concept of the active TSS pairs and eTSSs. The alignments were analyzed in a strand-specific manner, considering only the excision of CPD-lesions from the transcribed strand (TS) of mRNAs, asPROMPTs, and enhancers, which corresponded to the forward “+” (blue) or the reverse “-” (red) genomic strands, (Figure 92) depending on the element annotation. Strand specific heatmaps and average profiles of XR-seq signal were generated as described in section 4.1.3.

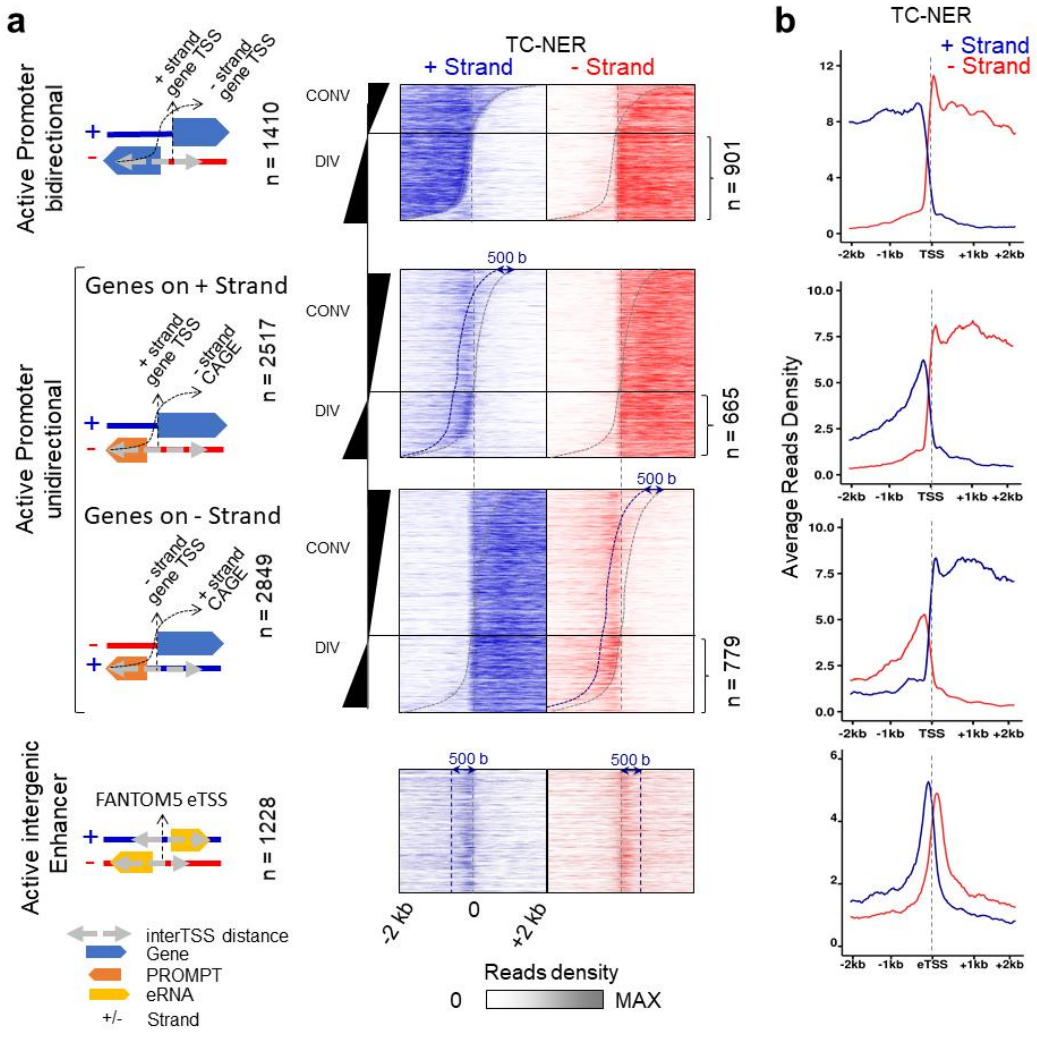


Figure 92 DNA-damages at transcribed strands of active elements are repaired homogeneously. (a) Strand-specific heatmaps illustrating the XPC XR-seq repair signal on template strand (TS, blue or red accordingly) of actively transcribed elements as depicted in Figure 88. Blue dashed lines set the border to 500 bases downstream of CAGE summits for each strand. (c) Strand-specific average profiles of XPC XR-seq signal as indicated in (a). Only divergent elements are included in this visualization.

The particular visualization revealed an expected balance in repair activity between transcription directions in active bidirectional promoters and enhancers, and a mild imbalance between mRNA-asPROMPT promoter pairs (Figure 92).

To further examine the patterns of TC-NER repair efficacy along transcription directionality, and in comparison with transcription initiation activity at the same regions, an analysis of XPC XR-seq, CAGE-seq and RNAPII-hypo signal (see materials and methods, sections 2.10.1, 2.10.9,

and 2.10.10) was conducted at divergent bidirectional TSS and unidirectional TSS/ asPROMTs pairs.

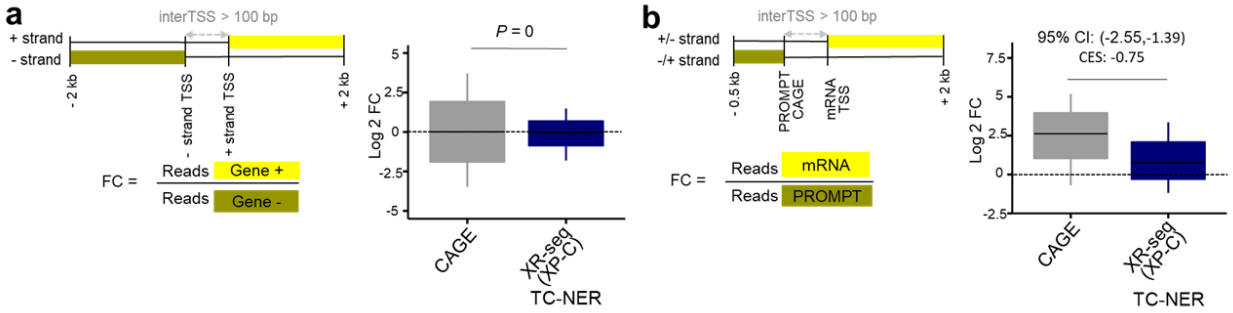


Figure 93 Comparisons between TC-NER repair activity at transcribed promoter regions and transcription initiation activity (a) Left panel: Representation of the genomic intervals used for calculating Log2 Fold Change (FC) ratios of sense XPC XR-seq and CAGE-seq signal. Right panel: Box plots of Log2 FC of CAGE-seq and XPC XR-seq sense reads between bidirectional promoter pairs. Box plots depict the 25th–75th percentiles and error bars depict the 1.5 * IQR (inter-quartile range). Two sample F-tests were applied for each of 10,000 sampling pairs of 100 data points with replacement from each population to test for significant difference between sample variance. The calculated P expresses the percentage of the non-significant F-tests (F-test P >= 0.05) out of all tests (b) Same as (a), but for unidirectional mRNA-PROMPT pairs. 95% confidence intervals (CI) of mean differences between log2 counts was applied as described in materials and methods tade. Effect sizes of log2 counts between datasets were calculated using Cohen's method (CES).

As depicted in Figure 93, sense CAGE-seq, sense XR-seq, and RNAPII-hypo (NO UV and +UV +1.5 h) sense alignments were counted at regions starting from mRNA TSS up to 2 kb (to the direction of the mRNA transcript), while for asPROMPTs, from CAGE summit up to 500 bp (to the direction of the asPROMPTs transcript). Counts were normalized by the element length and sample size and summarized as log_2 fold change (log_2FC) ratios between forward (+) and reverse (-) mRNA counts for bidirectional pairs, and log_2FC ratios between mRNA and asPROMPT counts for mRNA-asPROMPT pairs (Figure 93). XPC XR-seq and CAGE-seq count ratios were visualized using boxplots and coupled by a bootstrapping F-test approach (materials and methods 2.13) to support a balance of TC-NER repair efficacy in each direction of bidirectional active promoters (Figure 93, F-tests p-value = 0.). This result is also in agreement with RNAPII-hypo ChIP-seq data showing equal amount of RNAPII molecules recruitment at PICs (Figure 94) and equivalent production of capped mRNAs (CAGE-seq, Figures 93 and 94, median Log2 FC = 0, F-tests p-value = 0).

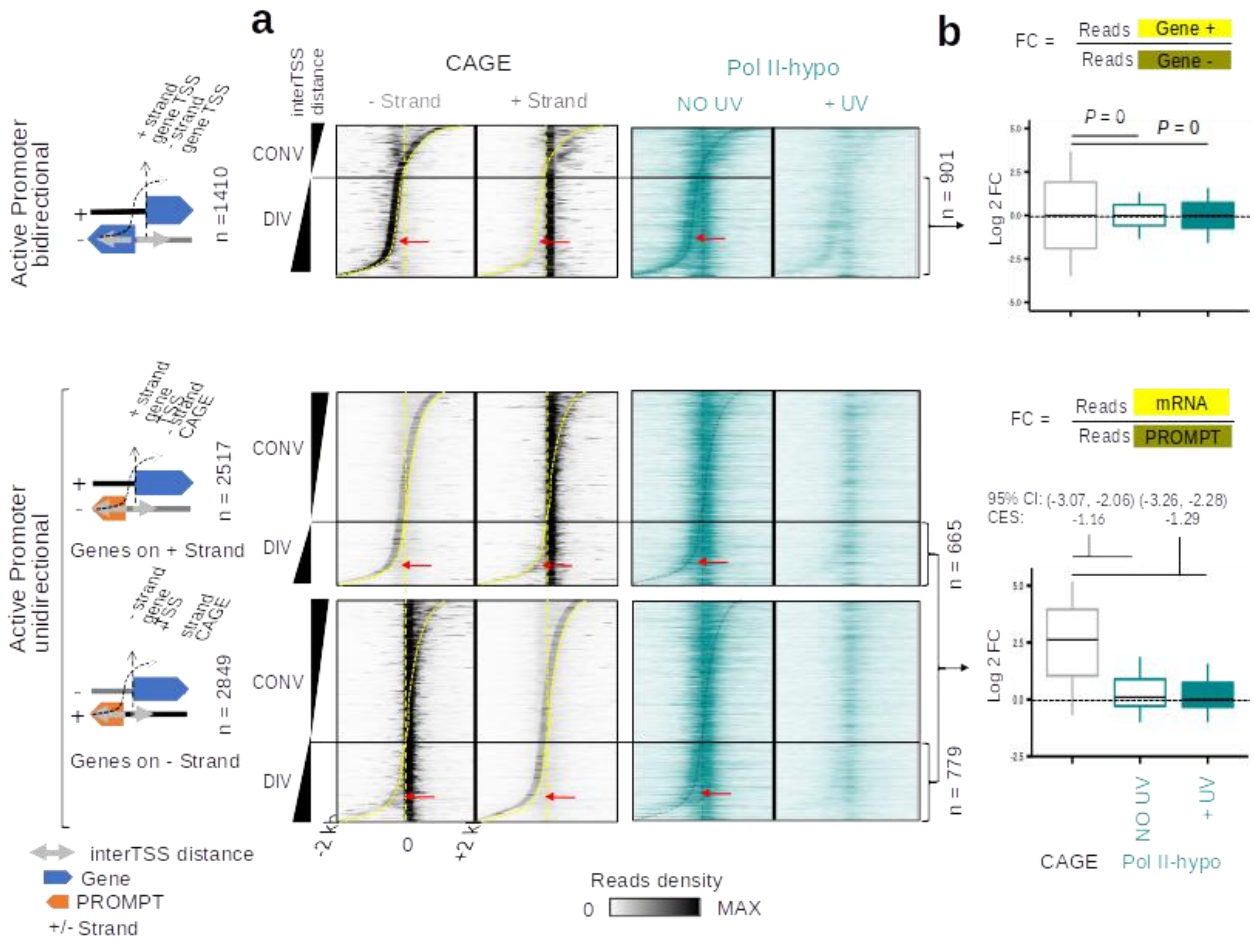


Figure 94 Pre-initiating RNAPII is bound homogeneously between pairs of transcribed elements, during early recovery from UV-stress. (a) Heatmaps of CAGE-seq (reads separated per strand) and RNAPII-hypo ChIP-seq signal (NO UV and +UV 1.5 h, see methods), at regions described in Figure 88. (b) Box plots of CAGE-seq and Pol II-hypo ChIP-seq signal ratios between forward (+ strand) and reverse (- strand) for divergent bidirectional promoters (upper panel) and mRNA over PROMPT (lower panel) for divergent unidirectional promoters. Box plots depict the 25th–75th percentiles and error bars depict the 1.5 * IQR (inter-quartile range). Upper panel: Two sample F-tests were applied as described in Figure 93. P-value denotes the proportion of the non-significant F-tests (F-test p-value >= 0.05) out of the 10,000 total tests. (Bottom) 95 % confidence intervals (CI) of mean differences between log2 counts of tested samples were calculated as described in materials and methods, section 2.13. Effect sizes of log2 counts between datasets were calculated using Cohen's method (CES).

Notably, in transcriptional pairs with a large variability in CAGE-seq signal levels between strands (mRNAs-asPROMPTs, Figure 94), signal density between strands was balanced for TC-NER (XR-seq (XP-C)) and RNAPII-hypo (Figure 94, F-Tests: P = 0).
While this phenomenon was previously observed, it was hardly explained (Adar et al., 2016; Hu et al., 2015). The particular quantification showed that TC-NER is not correlated with the steady state levels of CAGE at asPROMPTs (PCC = 0.1343). Additionally,the fact that the log_2FC of XPC XR-seq signal between mRNAs and PROMPTs is significantly smaller than the CAGE-seq signal (Figure 94, 95 % CI excludes 0) also matches with the UV-independent RNAPII-hypo uniformity (Figure 94).

The same analysis was also applied at enhancer regions, resulting in a balanced pattern of TC-NER repair between the bidirectionally transcribed enhancer units.

4.3.20 Uninterrupted transcription initiation drives the cell' transcriptome to DNA-damage recovery via TC-NER

To evaluate the biological importance of the uninterrupted transcription initiation at all active regulatory regions during the early UV-stress recovery, a strand-aware meta-analysis of XPC XR-seq of CPD damages in +UV 1 h +DMSO, +UV 1 h +DRB and +UV 1 h +DRB2 XR-seq conditions (see materials and methods, section 2.10.12) was conducted using the methodology described in section 3.1.3, to produce heatmaps and average profiles of CPD XR-seq signal around potential pyrimidine dimers (TTs, section 4.3.11) and all the classes of TSS-pairs defined in section 4.3.5. TT regions overlapping with enhancers were filtered out to avoid counting repair signal that arise from eRNAs, and 'S-F' scores for different TT-clusters were calculated as described in section 4.3.11.

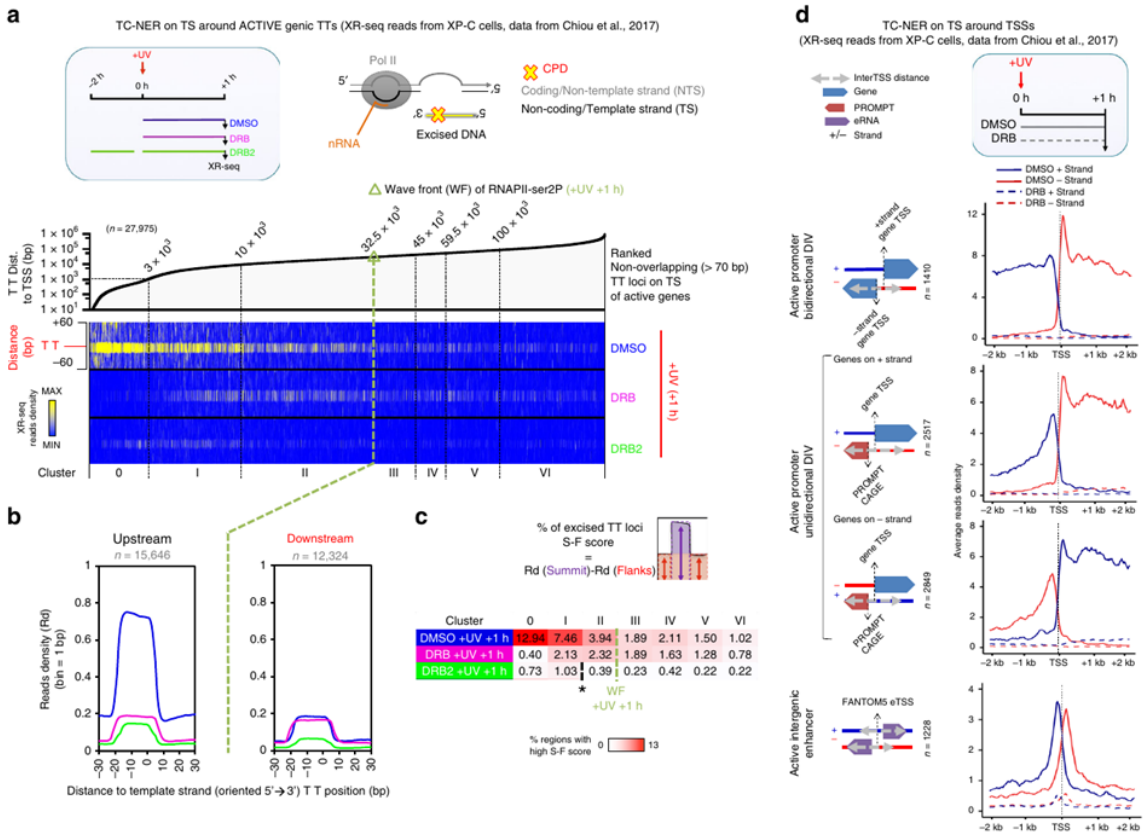
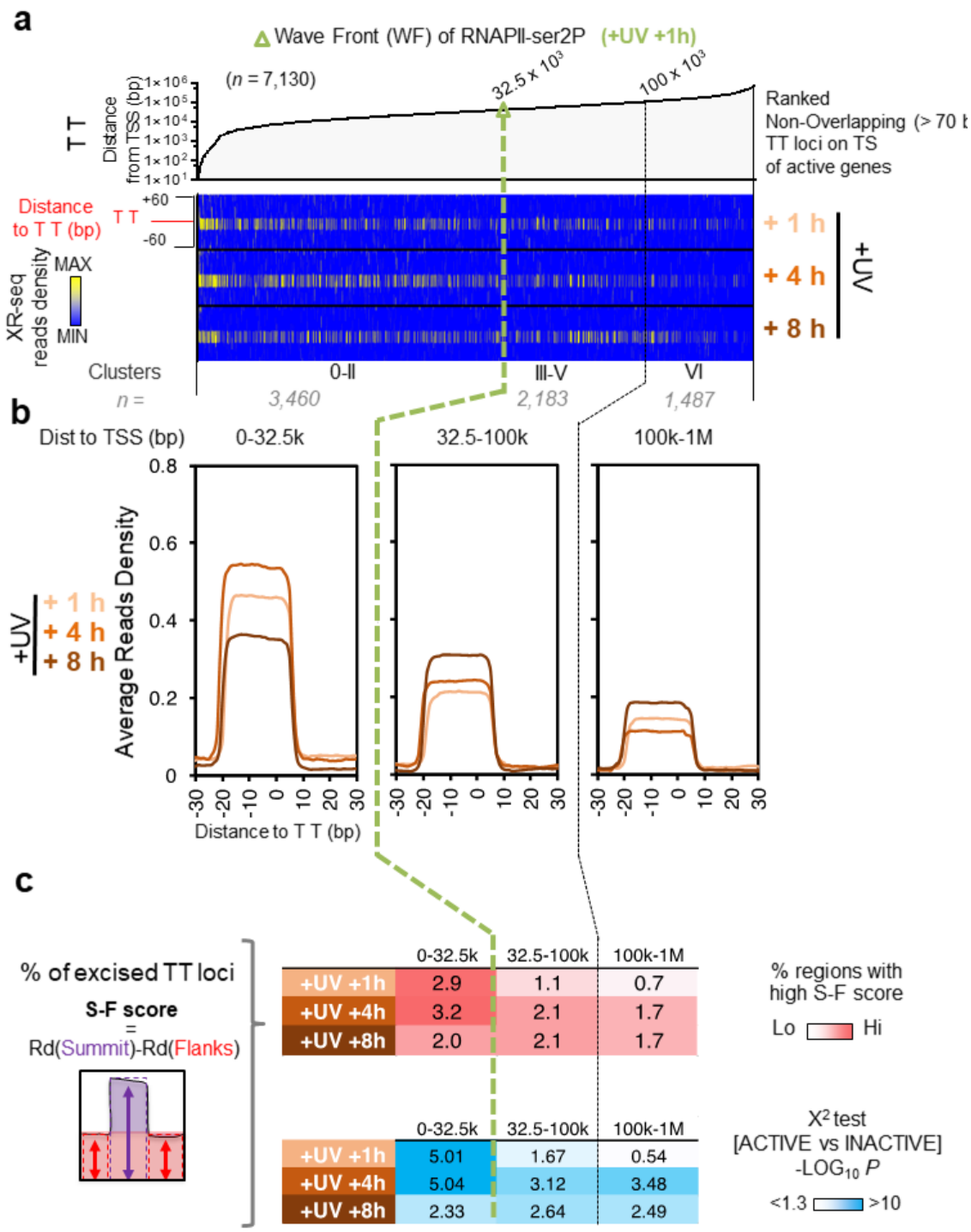


Figure 95 TC-NER activity is heavily dependent by transcription initiation. (a) Upper panel: Illustration of XPC DRB experimental timeline. Lower panel: Heatmaps of XPC XR-seq signal at TT regions located in the transcribed strand of active genes, at timepoints indicated in the experimental timeline illustration. (b) Average profiles of XPC XR-seq repair signal as indicated in (a) for TT-clusters defined in section 4.3.11 (+UV 1 h clustering). (c) Visualization of the percentage (%) of high S-F scores (see section 4.3.11) at all clusters presented in (a) and (b). Wave front is illustrated as a light green dashed line, while the asterisk

denotes high decrease of XR-seq signal in DRB2 condition. (d) Strand-specific average profiles of XPC XR-seq signal at TSS-pairs. Conditions analyzed are indicated accordingly.

Visualizations and quantifications depicted in Figure 95 outline the fact that when DRB is applied directly after UV-exposure, TC-NER activity at pyrimidine dimers localized between active TSSs and the +UV 1 h wave front of the stress-released RNAPII (as defined in section 4.3.11) is affected drastically (Figure 95, DRB +UV +1h, clusters 0-II, and (d)).
Subsequently, when only a restricted amount of pri-elongating RNAPII is allowed to be fired immediately before the UVC induction, and a parallel blockage of de novo RNAPII release after UVC irradiation is applied (Figure 95, DRB2 experiments), an inadequate delivery of RNAPII molecules impairs TC-NER activity at all transcribed loci (compare signal before and after asterisk positions in Figure 95).
To further evaluate whether the continuity of RNAPII initiation results to a high extent of ongoing repair activity as depicted in +UV +DMSO condition in Figure 95, NHF1 time-course XR-seq experiments of CPD damages in +UV +1 h, +UV +4 h and +UV +8 (see materials and methods, section 2.10.11) were analyzed as described above with minor adjustments. TT dinucleotides located in the reverse strand of active mRNAs, or between the TSS and 2 kb downstream of TSS of transcript-pairs with inter-TSS distance < 100 bp, were excluded. Heatmaps, average profiles and S-F scores of XR-seq signal at TT loci were generated as described in section 4.3.11.


a
Wave Front (WF) of RNAPII-ser2P (+UV +1h)
(n = 7,130)
32.5 x 10³
100 x 10³
Ranked
Non-Overlapping (> 70 b
TT loci on TS
of active genes
TT
Distance from TSS (bp)
1×10⁶
1×10⁵
1×10⁴
1×10³
1×10²
1×10¹
Distance to T T (bp)
+60
T T
-60
XR-seq reads density
MAX
MIN
+ 1 h
+ 4 h
+ 8 h
+UV
Clusters
n =
0-II
III-V
VI
3,460
2,183
1,487
b
Dist to TSS (bp)
0-32.5k
32.5-100k
100k-1M
Average Reads Density
0.8
0.6
0.4
0.2
0
+UV
+ 1 h
+ 4 h
+ 8 h
-30 -20 -10 0 10 20 30
Distance to T T (bp)
c
% of excised TT loci
S-F score
=
Rd(Summit)-Rd(Flanks)
0-32.5k
32.5-100k
100k-1M
+UV +1h 2.9 1.1 0.7
+UV +4h 3.2 2.1 1.7
+UV +8h 2.0 2.1 1.7
% regions with high S-F score
Lo
Hi
X² test
[ACTIVE vs INACTIVE]
-LOG10 P
0-32.5k
32.5-100k
100k-1M
+UV +1h 5.01 1.67 0.54
+UV +4h 5.04 3.12 3.48
+UV +8h 2.33 2.64 2.49
<1.3
>10

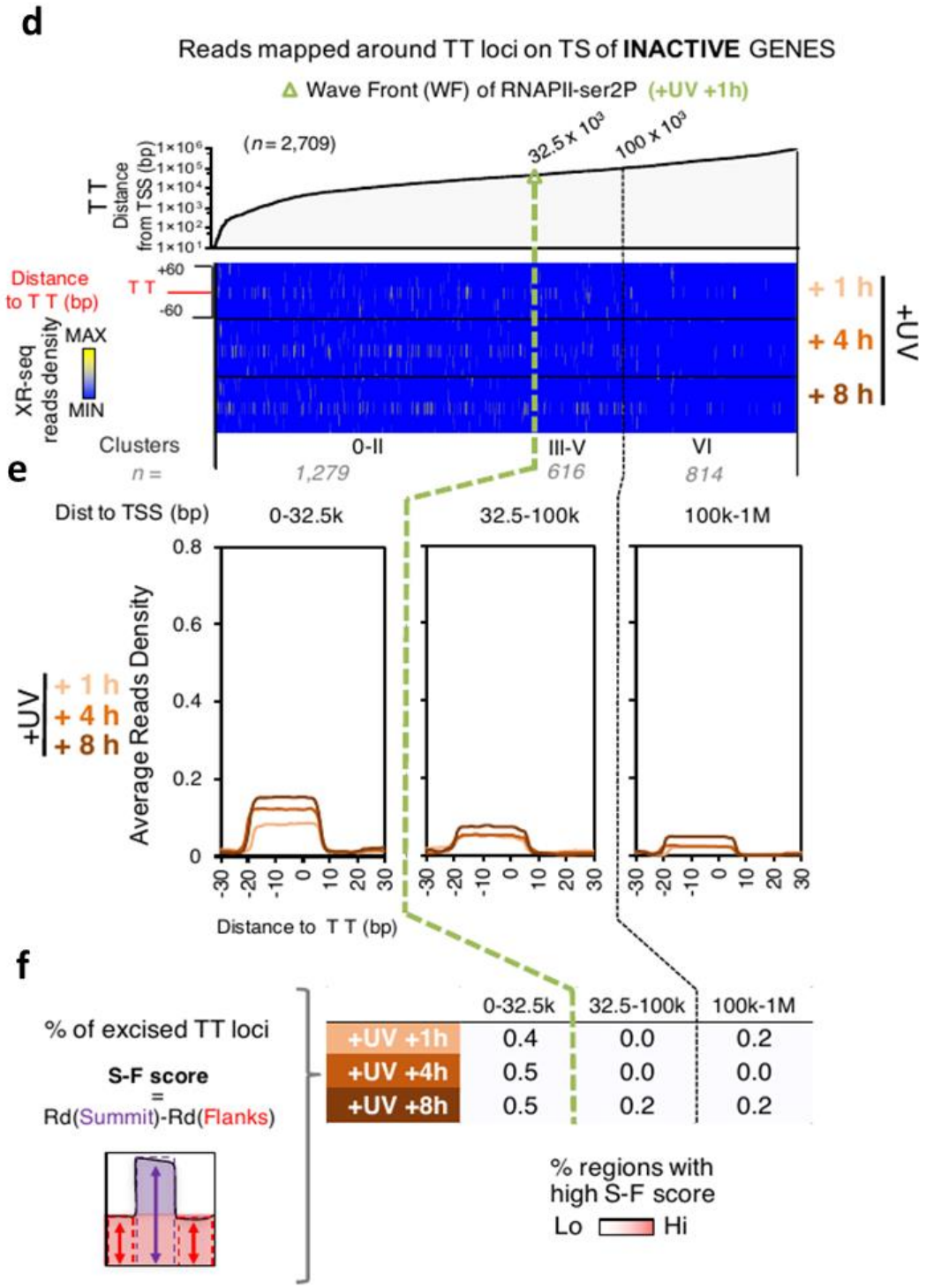


Figure 96 TC-NER activity is preserved along the recovery period due to the uninterrupted transcription initiation procedure. (a) Heatmaps of time-course NHF1 XR-seq signal at 1 h, 4 h and 8 h post UV irradiation at TT loci located in the transcribed strand (TS) of active mRNAs. TTs are separated to clusters as described in section 4.3.11 (+UV 1 h clustering) (b) Average profiles of XR-seq repair activity for the

regions illustrated in (a). (c) Percentages (%) of high S-F scores for clusters as indicated in (a) (see section 4.3.11 for details), (d), (e), and (f), Same as (a), (b), and (f) respectively, but for inactive genes.

This analysis showed that a considerable amount of TC-NER excision events was preserved at damage-sites localized immediately downstream of active TSSs at 4 h and 8 h during the damage recovery (compare Figure 95 with Figure 96). Importantly, the extent of excision events on the transcribed strand changes during damage recovery (+ 8 h) from the proximal to the distal part of long active genes (Figure 96 clusters III-VI, and Figure 95).

4.4 A genome-wide analysis pipeline for the evaluation of aniFOUND-seq methodology

To evaluate the specificity of aniFOUND-seq (see Material and Methods), a genome-wide comparison between XR-seq (see materials and methods) (Adar et al., 2016) and damage-seq (Adar et al., 2016) assays was performed. Since the particular variation of aniFOUND-seq (materials and methods tade) does not produce strand specific data, and XR-seq and damage-seq datasets are strand specific, the analysis was performed using a single-end set-up. Two replicates of 1BR.3 aniFOUND-seq +UV 4 h pull-down (PD) and aniFOUND-seq INPUT (INPUT) were analyzed using the methodology decribed above, and the resulting alignments were extended to an average of 200 bp fragments using the 5’ -> 3’ direction.

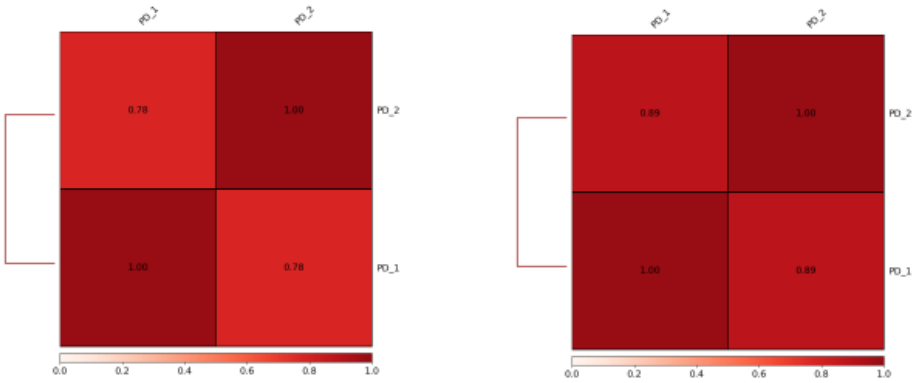


Figure 97 Correlation between aniFOUND-seq biological replicates. Left panel: Spearman correlation, calculated along the genome using 10 kb windows (as described in section 4.1.3.3.1). Right panel: Spearman correlation calculated at different chromatin states according to the NHDF 15-state ChromHMM annotation (see introduction, section 1.7).

The center of each read was used as described in section 4.1.3.5, in order to annotate each read uniquely based on the NHDF-Ad_Adult_Dermal_Fibroblasts core 15-state model roadmap chromatin state annotation (Roadmap Epigenomics Consortium et al., 2015). The 8th chromatin state, “ZNF genes & repeats" was excluded, since it is analyzed more precisely in a separate analysis module (see below). The same procedure was also applied at two replicates of NHF1 CPD XR-seq +UV 1 h and +UV 4 h datasets, and at two replicates of NHF1 64 XR-seq +UV 5 min, 20 min, 1 h, 2 h and 4 h datasets (see materials and methods, section 2.10.11), which

were merged based on the photolesion category (CPDs and 64s), omitting the step of fragment length extension. The merging of the different time points after UV exposure was applied after taking into consideration the main differences between the two repair assays: (a) XR-seq captures the excised DNA fragments along the early steps of NER, while aniFOUND-seq captures the newly synthesized DNA at the lesion gaps, after the DNA cleavage is completed; (b) aniFOUND captures the repair-synthesis events in a cumulative fashion, while XR-seq captures a 10-minute-long excision activity; (c) aniFOUND captures total UDS activity, which is associated with the repair of both CPDs and 6-4 PPs, while XR-seq focuses on one type of photolesion per experiment.

Similarly, for damage-seq, two replicates of NHF1 CPD +UV 0 h and NHF1 64 damage-seq +UV 0 h, as also NHF1 damage-seq INPUT libraries were processed as described above.

All chromatin state counts were aggregated per category, normalized by the total genome coverage of each chromatin category, as also by a sample size factor ($1{,}000{,}000$ / $total\ alignments$). The resulting normalized values were summarized either as ratios normalized by their corresponding input dataset (Figure 98 (a)), or as percentages of the total counts (Figure 98 (e)) using a radar plot visualization.

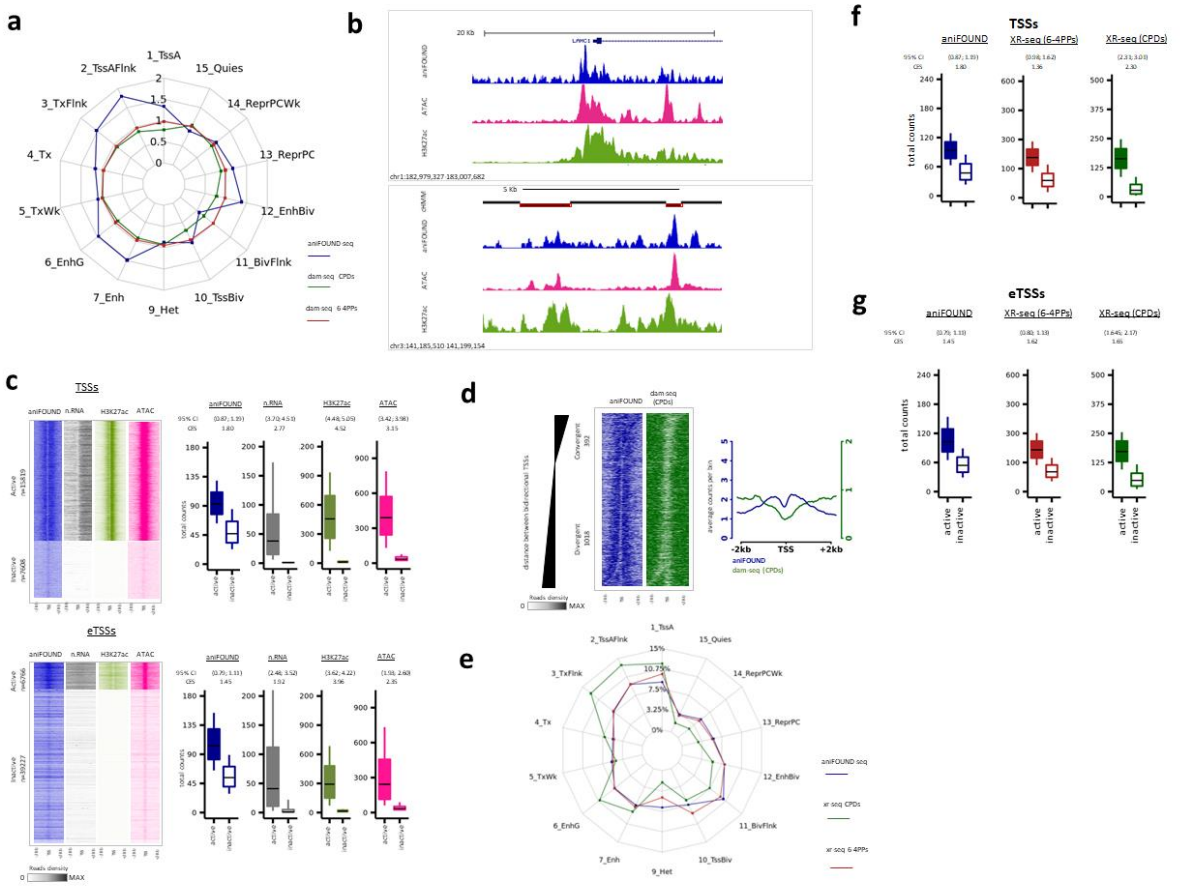


Figure 98 Genome-wide distribution of aniFOUND-seq signal. (a) Repair and damage ratios in different chromatin states. The chromatin states are defined according to the 15-state ChromHMM annotation (see materials and methods, section 2.10.11). Repair ratios are calculated by aniFOUND-seq reads,

normalized by their INPUT reads for each state. Similarly, damage ratios resulted from normalized damage-seq signal by their INPUT signal.
(b) Snapshots of UCSC Genome Browser. Upper panel: depiction of a gene and its flanking regions. The blue arrow indicates the direction of transcription. Lower panel: enhancers located in an area free of genes. The lower track (chromHMM) shows the ChromHMM states; yellow boxes with black outline correspond to enhancers. The enhancer regions in all tracks are shown in boxes.
(c) aniFOUND signal in active and inactive transcription start sites. Left panel: Heatmaps with the signal of aniFOUND, nRNA, H3K27ac and ATAC-seq 2 kb around the transcription start sites of active and inactive genes (TSSs), and active and inactive enhancers (eTSSs). Right panel: Box plots with the signal distributions of the gene sets shown in the corresponding heatmaps of the left panel. Boxes show the 25th - 75th percentiles and error bars show the data range to the larger and smaller values. For each active/inactive set, 10,000 samplings of 100 data points were randomly generated, and 95% confidence intervals of mean differences between active and inactive regions were calculated. Effect sizes of log2 counts between active and inactive sets were calculated using Cohen's method (CES).
(d) DNA damage and repair on bidirectional promoters. Left panel: Heatmaps of aniFOUND and damage-seq around the TSSs of bidirectional genes. The sorting was done based on the distance between the TSSs of the two bidirectional genes. Right panel: Aggregate plots of aniFOUND and damage-seq around the TSSs for the gene sets shown in the left panel.
(e) Distribution of aniFOUND and XR-seq signal along chromatin states. For XR-seq, all the available data sets up to 4 hours after irradiation were merged (5 min, 20 min, 1 h, 2 h and 4 h for 6-4PPs, and 1 h and 4 h for CPDs). The states are defined according to the 15-state ChromHMM annotation. For each library the number of reads that correspond to a chromatin state has been corrected by the total genomic length of the state. Y-axis shows the percentage of the corrected reads that fall in each state. For the hypothetical library in which all states were equally represented, a polygon with all its sides positioned at around 7 % (= 100 % / 14 states) would result.
(f), (g) Box plots of aniFOUND-seq and XR-seq signal in active and inactive TSSs and eTSSs, designated as in (c). Boxes show the 25th - 75th percentiles, and error bars show the data range to the larger and smaller values. For each active/inactive set, 10,000 samplings of 100 data points were randomly generated, and 95% confidence intervals of mean differences between active and inactive regions were calculated. Effect sizes of log2 counts between active and inactive sets were calculated using Cohen's method (CES).

Radar plots of damage-seq ratios revealed a rather expected (Adar et al., 2016) homogeneous formation of CPDs and 6-4PPs, since they are captured directly after irradiation by the protocol procedure (materials and methods, Figure 98 (a)). On the contrary, aniFOUND-seq radar plots showed that the UDS reads were unevenly distributed across the 14 chromatin states (Figure 98 (a), aniFOUND-seq). Notably, active TSSs and their corresponding flanking regions (states 1, 2, and 3), as well as enhancer-associated regions (states 6, 7 and 12) showed elevated repair-synthesis. These results suggest faster NER-activity during the 4-hours UVC recovery period in actively transcribed regions in comparison to repressed and quiescent regions. Heatmaps and boxplots of NGS signal on 2 kb-extended active TSSs and eTSSs (VH10 TSS/eTSS activity, see above) revealed that aniFOUND-seq repair signal is detected significantly around these regions (Figure 98 (b),(c)). Comparisons of the aniFOUND-seq repair signal with VH10 nRNA-seq +UV 2 h (see materials and methods), ATAC-seq +UV 2h (see materials and methods), and H3K27ac ChIP-seq +UV 2 h datasets (see materials and methods) displayed enhanced levels of UDS at highly accessible regions, and specifically around actively transcribed TSSs and eTSSs (Figure 98 b and c), as opposed to inactive elements (Figure 98

5c, 95 % Confidence Interval of log2 count differences does not include 0). Notably, a characteristic pattern of repair signal is observed at bidirectionally transcribed mRNA TSSs (Figure 98 (c) and (d)), equivalent to previously detected nascent-RNA NGS signal profiles in bidirectionally transcribed TSS-pairs (see Figure 90), confirming that NER takes place rapidly and effectively at all actively transcribed and accessible loci (Liakos et al., 2020). However, observing the UDS activity at non-transcribed regions, aniFOUND-seq signal is still detectable as a result of the GG-NER activity at these regions, captured by the assay.
To evaluate the potential effects of damage activity on UVC lesion repair at active bidirectional promoters, CPD damage-seq signal heatmaps and average profiles were also generated using the same annotation as reference, to reveal a complementary signal pattern between the two assays (Figure 98 (d)). Consequently, to validate the genome-wide UDS signal profile at actively transcribed promoters, NHF1 CPD and 64 XR-seq merged alignments were summarized at 14 roadmap chromatin states (as described above) as percentages of total counts, and also at actively transcribed TSSs and eTSSs, to generate boxplot quantifications (Figure 98 (g) and(f)). Radar plot visualization of aniFOUND-seq and XR-seq in figure 98 (e) demonstrates that the distribution of aniFOUND-seq repair signal across different chromatin states is analogous to the 64 XR-seq signal. This result is consistent with the fact that the majority of 6-4 photoproducts are repaired during the first 4 hours after damage induction. Further, the preferential enrichment of CPD XR-seq signal in chromatin state categories related to active transcription (TssA, TssAFlnk, TxFlnk, Tx, EnhG and Enh) was paralleled with reduced CPD signal in chromatin states related to repressed chromatin and heterochromatin (Het, TssBiv, BivFlnk, EnhBiv, ReprPC and ReprPCWk) and is in line with the fact that CPD repair is accomplished by TC-NER during the early UVC recovery (Adar et al., 2016; Hu et al., 2015).
Additionally, boxplots of repair signal distributions at actively transcribed TSSs and eTSSs (Figure 98 (c), (f), and (g)) demonstrate that the NER repair activity at the actively transcribed genome is elevated in comparison to the non-transcribed elements during the early recovery response, for all the repair datasets. Nonetheless, aniFOUND-seq and XR-seq activity is also detectable at inactive elements (Figure 98 (c), (f), and (g), white filled boxplots), showing that GG-NER is also present in the early UVC damage response.

4.4.1 An analysis pipeline for the estimation of NER activity on repeated genome using aniFOUND-seq

Repetitive DNA comprises a considerable part of the genome (~50%, see introduction) that is still “under-examined” in the field of DNA damage and repair. To study the UDS activity at these regions, aniFOUND-seq raw reads were analyzed as follows:
Raw FASTQ reads of both PD and INPUT conditions were processed using the methodologies described in section 4.1.1 and 4.1.2, but initially all sequences were trimmed at the 3’ end to a constant length of 50 bases in order to eliminate any effect of variable read length bias between the different datasets.
High quality FASTQ sequences were provided as an input to RepeatMasker software (Nishimura, 2000), by first converting them to FASTA files, and splitting them to 300,000 sequence chunks in order to run the algorithm more effectively. RepeatMasker was run with parameters: -e crossmatch -pa 30 -q -low -species human -a -inv -lcambig -html -source -gff -

excln -u -nopost to produce pairwise alignment files of repeat elements against the examined FASTA sequences, using RepBase (Jurka et al., 2005) and Dfam (Hubley et al., 2016) as a repeat species reference. The resulting alignments were further processed using ProcessRepeats, a RepeatMasker utility, to produce repeat specific annotation files, containing information about the alignment of every repeat species against each sequenced read. For each library, all annotation files were summarized to produce a count-like matrix with repeat species names as rows, sample ids as columns, and cells containing the number of total repeat species occurrences in each of the examined samples, resulting to a total of 1,279 unique repeat species identified in all datasets, that were further summarized to a total of 68 repeat families of origin. To determine potential differences of repair activity along the repeated DNA sequences, differential enrichment analysis between the aniFOUND-seq and the INPUT libraries was performed. The specific analysis was performed using the DESeq2 software (Love et al., 2014)] by providing the count matrix described above, and using the INPUT condition as a reference sample. Size factors and dispersion were estimated using the default settings of the program, and the statistical testing was performed using a negative binomial Generalized Linear Model (GLM), based on the estimated size factors. Only results with a p-adjusted value threshold lower than 0.05 were reported.

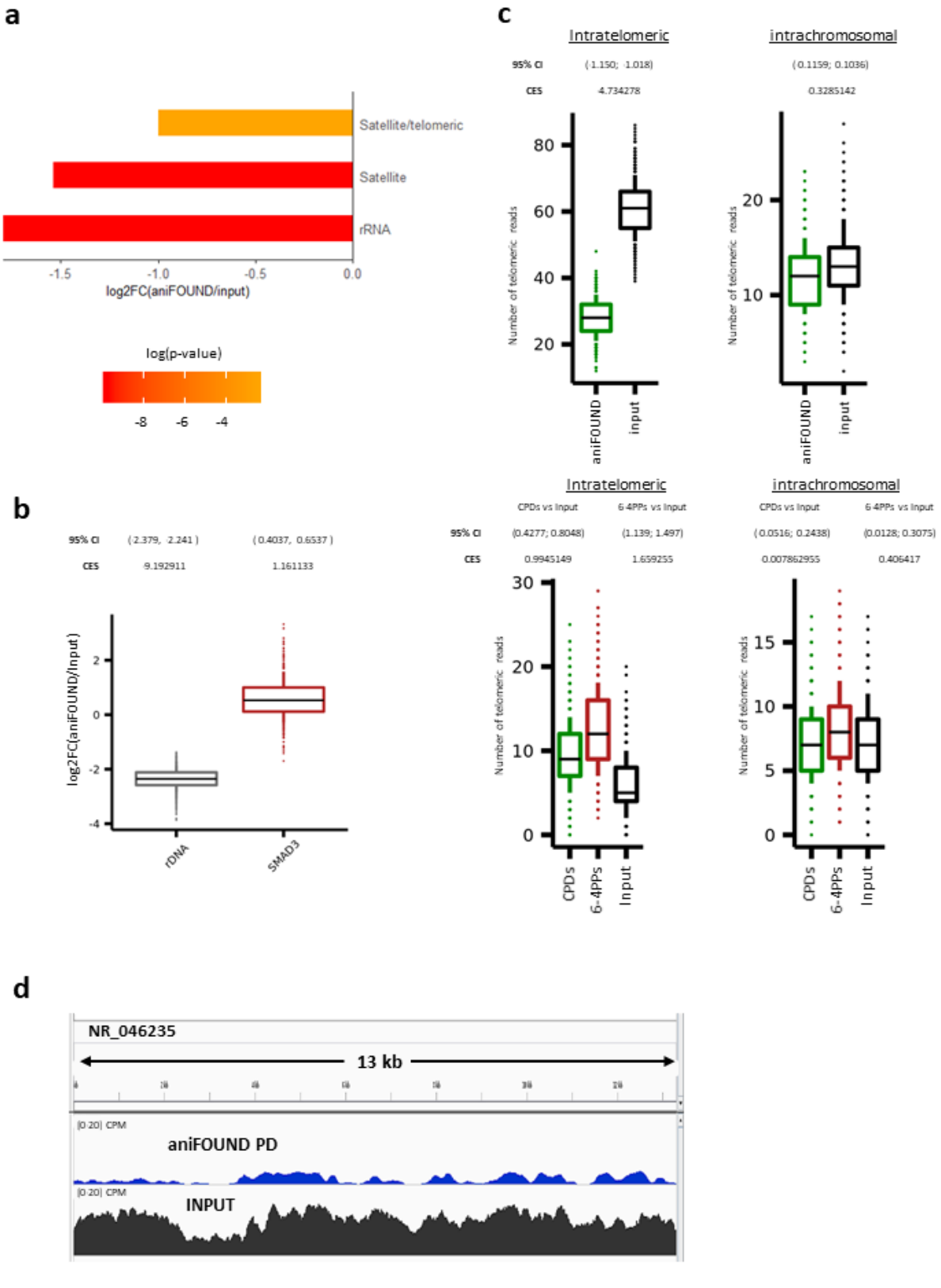

a
Satellite/telomeric
Satellite
rRNA
-1.5
-1.0
-0.5
0.0
log2FC(aniFOUND/input)
log(p-value)
-8
-6
-4
b
95% CI
(2.379, 2.241)
(0.4037, 0.6537)
CES
9.192911
1.161133
log2FC(aniFOUND/Input)
rDNA
SMAD3
c
Intratelomeric
intrachromosomal
95% CI
(1.150; 1.018)
(0.1159; 0.1036)
CES
4.734278
0.3285142
Number of telomeric reads
aniFOUND
input
Intratelomeric
intrachromosomal
CPDs vs Input
6-4PPs vs Input
95% CI
(0.4277; 0.8048)
(1.139; 1.497)
(0.0516; 0.2438)
(0.0128; 0.3075)
CES
0.9945149
1.659255
0.007862955
0.406417
Number of telomeric reads
CPDs
6-4PPs
Input
d
NR_046235
13 kb
[0-20] CPM
aniFOUND PD
[0-20] CPM
INPUT

Figure 99 Repeat enrichment on aniFOUND-seq reads (a) Differentially represented repeat families in aniFOUND-seq and input libraries. The bars show the log2 ratio of the aniFOUND-seq library reads over the input library reads that are annotated to the same repeat family. The repeat families are defined according to the classification system of Repbase. The color of each bar denotes its adjusted p-value. Only families with an adjusted p-value lower than 0.05 are shown. (b) Distributions of mapped read ratios on rDNA and SMAD3 gene between aniFOUND and input libraries. On the Y-axis, the logarithmized fold change of 1,000 random samples is shown. For each random sample the reads were aligned on an extended reference genome consisting of the UCSC hg19 and a single copy of the human rDNA (NR_046235) sequence (see Materials and Methods). Effect sizes refer to the difference from zero of the distributions depicted by the box plots and were calculated by using Cohen's method (CES). (c) Random samples of UDS (upper panel) and DNA damage (lower panel) signal on telomeres as estimated with aniFOUND-seq and damage-seq, respectively (see Materials and Methods). Y-axis shows the number of telomeric reads that resulted from 1,000 TelomerHunter runs on samples with 100,000 alignments each (see online methods). For both aniFOUND-seq and damage-seq, pull-down and input libraries have been plotted. 95% Confidence Intervals (95% C.I.) of log2 differences between pull-down and input libraries were calculated using 10,000 samples of 100 data points from each examined library. Effect sizes were calculated using Cohen's method (CES). (d) Custom IGV genome browser track of human NR_046235 repeat unit, illustrating aniFOUND-seq PD and aniFOUND-seq INPUT signal.

The results of this pipeline were used to evaluate the aniFOUND-seq repair prevalence on the repeated DNA. Figure 99 (a) summarizes that the most differentially enriched repeat family is rRNA, that seems to be less efficiently repaired compared to the INPUT background distribution, during the early UV-damage response, denoting that damages at the particular genomic sequences are repaired at a lower rate. Comparing the particular result with the literature, it seems that there are contradictory findings concerning the speed of rDNA repair, with some studies declaring that ribosomal repeats are repaired at a slower rate, likely because of inadequate repair factors accessibility at damage sites (34,35 Stefos). On the contrary, there is a study reporting significant TC-NER activity taking place at rDNA sequences during the early recovery process, attributing this phenomenon to the removal of damaged sequences to the nucleolar periphery that enables the repair machinery [36 Stef], while in another recent study, it was shown that rDNA is not subjected to TC-NER [37 Stef].
To further validate this finding, an additional analysis pipeline was conducted: Initially, the hg19 human reference genome (FASTA) was extended by adding the 45S pre-ribosomal N5 (RNA45SN5) NCBI sequence (https://www.ncbi.nlm.nih.gov/nuccore/NR_046235) as a new chromosome, using the >NR_046235.3 identifier. aniFOUND-seq PD and INPUT quality-filtered FASTQ files were aligned against the new genome build and analyzed using the methodology described in section 4.1.3.1 with some modifications: (a) The option -T 0 was added to bwa mem run, in order to allow low quality alignments, in order to maximize the number of the multiple alignments. (b) BAM files were not filtered using alignment quality or duplicated records information. Merged alignments were normalized to a similar read depth (19,000,000 reads) and sampled 1,000 times to produce 100,000 read chunks that were in turn summarized at NR_046235.3 chromosome and SMAD3, an indicative actively repaired gene to produce boxplots of log_2 PD normalized counts over INPUT normalized counts for each region (Figure 99 (b)). To apply a statistical comparison between the PD and INPUT count distributions for each element, 1,000 samplings of 100 data points were randomly generated, and 95%

confidence intervals of mean differences between PD and INPUT regions were calculated. Effect sizes of log_2 counts between PD and INPUT sets were calculated using Cohen's method (CES). The particular analysis confirmed that the rDNA is a region that is not preferentially repaired during the first hours of NER-repairs using the aniFOUND-seq set up (Figure 99 (b) , 95 % CI excludes 0).

Moreover, the differential repeat enrichment analysis showed that satellites were also repaired with a slower rate (Figure 99), a result that is in line with a previous study, reporting that satellite-rich regions are repaired slower by NER machinery than other regions (Sanders et al., 2004). To examine in more detail the UDS activity at human telomeres, a telomeric content enrichment analysis was performed. Telomeres are typical repetitive regions consisting of tandem 6 nucleotide-long sequences. Nevertheless, while their susceptibility to DNA damage and the cell's capability to repair them are tightly associated with aging and cancer, it is not yet clear whether they are prone to damaging factors and if they are repaired by the cell to the same extent as the rest of the genome. For this analysis, the same alignment set-up as in the rDNA analysis was used (see above). Both mapped and unmapped reads were scanned for TGAGGG repeat occurrence, but only the unmapped sequences are considered as telomeric content (Feuerbach et al., 2019). Merged alignments were down-sampled to a similar read depth, and each subsample file was scanned for TGAGGG enrichment, using TelomereHunter (Feuerbach et al., 2019) with default parameters. Candidate telomeric reads were classified into 3 categories: (1) "Intrachromosomal" reads, which comprise of telomeric repeats that are mapped to the chromosomal regions of the genome, except from the first and last band. These regions are considered "pseudo" telomeric ("pseudotelomeric") and were used as a control set. (2) "Subtelomeric" reads consist of telomeric reads aligned to the first or last band of a chromosome, while (3) all unmapped reads were categorized as "intratelomeric", which represent the actual telomeric content. The outputs of all the telomeric quantifications were summarized, to produce a telomeric content distribution for each region category, for both PD and INPUT, but also for CPD damage-seq, 64 damage-seq and INPUT damage-seq datasets. To compare the intratelomeric and intrachromosomal distributions between aniFOUND (or damage-seq) datasets and their corresponding INPUT libraries, a similar approach to calculate confidence intervals were applied as described in the rDNA sequence analysis pipeline (see above).

Boxplots of sampled counts revealed that true telomeric reads were under-represented in aniFOUND-seq, while pseudotelomeric reads were repaired to a baseline level (Figure 99 (c) upper panel), a finding that supports the hypothesis that telomeres are subjected to UVC-derived UDS at lower frequency compared to the rest of the genome. On the contrary, application of the same analysis pipeline at damage-seq samples showed that damage prevalence is higher in telomeric regions compared to the overall genome (Figure 99 (c) lower panel), showing that the observed, lower level of UDS activity at telomeres during the early response to UV irradiation is not an effect of reduced DNA damage occurrence. This result is in agreement with a proposed model suggesting that telomeres are vulnerable to UVC irradiation related lesions, but repair of these damages is almost absent (Rochette & Brash, 2010), opposing to another proposed model suggesting that telomeres are partly protected from UVC, and both categories of photolesions (CPDs and 64s) are removed fast and homogeneously, in comparison with other genomic sequences (Parikh et al., 2015).

5 Conclusions - Discussion

This study, describes a computational framework, developed for the study of transcription reorganization and chromatin alterations in response to UVC-induced stress, using primarily NGS data from human skin fibroblasts (Andrade-Lima et al., 2015; Lavigne et al., 2017; Liakos et al., 2020; Magnuson et al., 2015; Williamson et al., 2017b). The computational methodologies described above, provide genome-wide quantitative and qualitative illustrations of the NGS signals from a wide range of protocols (ChIP-seq, nRNA-seq, ATAC-seq, CAGE-seq, XR-seq, and aniFOUND-seq), regarding (1) the binding profiles of the three main RNAPII isoforms (from pre-initiation complex formation, to Promoter Proximal Pausing (PPP), and the entry into productive elongation), (2) the production of nascent RNA, (3) the chromatin accessibility, and (4) the histone modifications H3K27ac and H3K27me3, during the cellular responses to UVC-induced genotoxic stress.
The analysis pipelines included in the results section (see section 4) can serve as a guide for the analysis of the aforementioned NGS types, while the outputs of these modules can aid the research analyst with critical conclusions regarding the under-study biological phenomena.

Regarding the particular study, a novel metabolic function associated with active transcription is characterized, proposing that in response to UVC induced stress, damaged cells switch transiently to a ‘safe mode’ of RNAPII elongation (Figure 39). This mechanism promotes a global, accelerated and synchronous *de novo* escape of elongation waves of RNAPII molecules from PPP sites of active mRNAs into the gene bodies which cover the 50% of the transcribed genome [51 Lavigne]. The maximization of the entry of RNAPII molecules in gene bodies result to a rapid and homogenous DNA lesion identification at transcribed strands of mRNAs, regardless of the location of the DNA lesion, the mRNA length and prior to UVC levels of transcription. Complementarily, the expansion of NER activity is observed at damage sites overlapping the transcription elongation wave proximity.
In addition, detailed annotation of active regulatory regions revealed that the UV-induced release of RNAPII-Ser2P molecules from PPP sites is not limited to active genes, but is also detectable at PROMPTs and enhancers, as shown by the increase in RNAPII-Ser2P Escape Index (EI) in the respective genomic regions. In addition, the de novo binding of RNAPII-hypo molecules at PIC sites, and the detection of start-RNA molecules during cell recovery after UVC exposure, support a model where transcription initiation is not inhibited, but instead it supplies RNAPII molecules to the various transcription units (genes encoding proteins, long non-coding RNAs, PROMPTs, enhancers), in order to rapidly repair them via the TCR pathway. The experiments conducted in this study support that the continuous release of RNAPII molecules from PPP regions urges the molecules to shift to transcription elongation, thus reducing the NGS RNAPII-hypo signal in all actively transcribed TSSs. The particular defence mechanism affects the somatic mutation landscape of cancer genomes, such as melanoma and lung adenocarcinoma, by displaying low and homogenous mutation prevalence in all productively transcribed genes. Consequently, these results indicate that the widespread release of elongation waves boosts NER efficacy and can preserve genetic accuracy, while deficiencies in these mechanisms may hinder the genome-safeguarding effects. Interestingly, this mechanism

could potentially benefit the genotoxin-affected tissues and improve cancer therapeutics, by inhibiting the ability of tumour cells to boost transitioning of RNAPII into productive transcription elongation, while promoting genotoxic stress.

Comparing these results with two recent studies that investigate the effect of repair mechanisms in somatic mutations in cancer biopsies, it can be said that the described mechanism may be responsible for the fact that various point mutations appear to be significantly reduced in areas upstream of TSS, or around DNase hyper-sensitive sites (DHS) (Haradhvala et al., 2016; Perera et al., 2016; Sabarinathan et al., 2016). Maps of somatic mutations of genotoxins-exposed cancer genomes such as melanoma and lung adenocarcinoma [7 Lavigne] have previously been demonstrated to contain NER-specific signatures (see 6,14 Lavigne). The particular tumours arise from skin and lung tissues that may have been exposed to NER-related genotoxic stress, such as UV-irradiation and tobacco smoke. It's also shown (see Figure 87) that the mutation prevalence remains low throughout the TS of gene bodies of actively transcribed genes, independently of the level of expression, while reduced mutation prevalence is also observable in the NTS of actively transcribed genes, confirming better efficacy of both TC-NER and GG-NER.

Based on a mechanistic point of view, the widespread enhancement of productive elongating molecules into actively transcribed elements, is compatible with previous observations describing that, while in normal conditions P-TEFb function is restrained by the sequestering effect imposed by 7SK snRNP inactivating complexes (Nguyen et al., 2001), UV-irradiation favours an immediate increase in the totality of active P-TEFb molecules in the nucleus (R. Chen et al., 2008). The functional consequences of this activation are elucidated as follows: During UVC-stress recovery, the release of P-TEFb kinase activity (*via* cdk9) is followed by expanded hyper-phosphorylation of RNAPII CTD (Boeing et al., 2016; Heine et al., 2008), followed by a widespread and synchronous transition into productive elongation detectable in all actively transcribed elements (see Figures 76 and 88). The DRB absence further extends the outcome of the elongation wave-release in non-irradiated cells (see Figure 76), demonstrating that the magnitude of the wave-release depends on the amount of the engaged PPP loci by the paused molecules of RNAPII. Subsequently, a central role of P-TEFb in UVC DNA damage response is suggested, and is supported by a recently published study (Lavigne et al., 2015). Additionally, the determination of a global UVC-dependent elongation wave release of RNAPII molecules described in this dissertation, is in line with other finding regarding the detection of increased binding of RNAPII in most active gene bodies (see Figure 4 of (Gyenis et al., 2014)), and elevated levels of nascent transcription at the beginning of genes (Andrade-Lima et al., 2015; Williamson et al., 2017b). *De novo* RNAPII elongation wave release enables lesion-scanning at UVC damaged cells, and guarantees that damages located at the TSS proximity will be repaired. Cells seem to activate a program of 'safe' mode elongation that limits potential biases linked with the stochasticity of transcription initiation (Levine, 2011; Svejstrup, 2002), by transiently regulating gene expression at the level of PPP release. Additionally, the release of RNAPII molecules along the actively transcribed genome could enable the identification of the subsequent lesions by the trailing molecules, even in the case of the model that supports that

the RNAPII molecules are dissociated by the chromatin after the identification of a DNA lesion (Ratner et al., 1998)(Andrade-Lima et al., 2015)(Venema et al., 1992).

Functional assessment of the described defensive mechanisms with XR-seq data show that the global release of damage-sensing RNAPII molecules is paralleled by increased repair efficacy in all active genes, especially in genomic regions affected by the *de novo* wave propagation of RNAPII and to the substantial that increases the probability of transcription-dependent repair. It should be noted that XP-C cells demonstrate an increased excision activity at UV lesions as compared with WT cells, probably partially because of the lack of repair activity in NTSs of GG-NER deficient cells, that restricts the XR-seq signal coverage to a smaller part of the genome, and thus overestimates the read density enrichment in the TS regions of active genes. Additionally, the absence of GG-NER pathway might affect the probability of the available core NER factors to be recruited at transcription-blocking lesions.
While the proposed mechanisms promote TC-NER by rapid identification of DNA-lesions at the TS of active genes, the progression of the transcription elongation wave may result in a more accessible chromatin environment that could in turn enhance the repair rate of the NTS by GG-NER. Indeed, earlier studies have shown that repair in the NTS of active genes is faster than in inactive genes (Sabarinathan et al., 2016). Corroboratively, recent studies support the fact that chromatin accessibility promotes GG-NER repair along DNAse hypersensitive (DHS) regions (Adar et al., 2016; Jackson & Helleday, 2016; Perera et al., 2016), concluding that increased chromatin accessibility promotes both TC-NER and GG-NER accessibility to damaged DNA.

Of particular interest is the fact that the distribution of XR-seq signal in the sense and antisense strands of mRNAs and asPROMPTs of unidirectional TSSs and enhancers (Figure 92) is more homogeneous than it would be predicted by the corresponding CAGE signal. This finding reinforces the possibility of a replication process at the transcribed elements, promoted by the high levels of RNAPII-hypo binding at PIC positions in normal conditions, and the continuous transition of these molecules into transcription elongation in response to UVC. Regarding the role of antisense transcription, it has been suggested that the transcription process per se, and not the transcription products, utilizes and supports biological functions (Murray & Mellor, 2016). For example, asPROMPT sequences can function either as transcription factor binding platforms that regulate expression of their related genes (Scruggs et al., 2015), or various RNA-binding proteins that in turn regulate the expression of target genes (Seila et al., 2009). It is therefore understood that the successful repair of these loci is particularly important for maintaining the genomic expression programs, as also the processes necessary for the normal cell life.

Examining the accessibility of chromatin after exposure to UVC radiation using the ATAC-seq methodology, a global increase in accessibility at all active regulatory regions (promoters and enhancers) was observed, thus indicating that these regions remain “open” during the repair of damaged DNA. In the same context, ChIP-seq experiments showed that H3K27ac post-translational modification is preserved in the respective genomic regions.
These results are in agreement with similar studies which show that (i) in the case of rapid transcriptional induction, a significant increase in chromatin accessibility can be observed,

without changes in the degree of chromatin uptake by nucleosomes (Mueller et al., 2017), and (ii) increased gene expression (triggering transcription and promoting the productive stage of transcription elongation) is often coupled with increased chromatin accessibility (Gray et al., 2017; Ucar et al., 2017). In particular, this study supports that the increase in chromatin accessibility in actively transcribed regions is associated with the progression of RNAPII molecules from transcription initiation, to elongation, after cell exposure to UVC. The acting mechanism which carries out the increase of chromatin accessibility has not been further studied in this dissertation; however it is of particular interest to clarify, which chromatin remodeling molecules are involved in this process, and in what way.
Consequently, this study proposes a model in which RNAPII molecules continuously enter transcription initiation, and transit to transcription elongation through their release from PPP sites, to accelerate the processes of DNA lesion identification and repair in the entire transcribed genome. Overall, these results demonstrate a positive correlation between increased chromatin accessibility in active regulatory regions, transcriptional dynamics, and repair through the TC-NER repair pathway, revealing the high complexity of the cellular response during genotoxic stress.

Furthermore, the H3K27ac preservation at transcription initiation sites prevents the occurrence of H3K27me3, as these two post-translational modifications are mutually exclusive (Karlic et al., 2010). Indeed, the ChIP-seq data analysis showed that H3K27me3 modification was found to be located in a group of non-transcribed genes both before and after UVC exposure, and therefore did not occur in actively transcribed loci. This is consistent with the fact that H3K27me3 and PRC2 complex probably do not play a role in UVC-induced transcriptional response. Supporting the above, recent studies suggest that the presence of RNA inhibits the recruitment and further action of PRC2 at active genes (Beltran et al., 2016; Kaneko et al., 2014). The analyzed data also support that in the case of UVC exposure, nascent RNA production during activation and productive elongation of RNAPII molecules inhibit the binding of PRC2 to chromatin, and consequently the deposition of H3K27me3.
The stability of the binding levels and pattern of H3K27ac observed in active TSSs, is in agreement with previous studies claiming that during the early recovery period after UVC exposure (0-6 h), there is a dose-dependent increase (with higher exposure doses, smaller increase is observed) of histone acetylation (Ramanathan & Smerdon, 1986). In particular, the acetylation of histones H3 (Rubbi & Milner, 2003) and H4 (J. Wang et al., 2006) has been found to increase after UV exposure and these findings have been attributed to a more general process of chromatin structure "relaxation" after genotoxic stress induction. In fact, it is believed that DNA repair of the damaged sites requires relaxation of the chromatin structure, in order for the repair factors to have access at the DNA lesion sites. After the damage is repaired, the chromatin structure is restored (Polo & Almouzni, 2015; Soria et al., 2012). These results, regarding the increase in chromatin accessibility and stability of H3K27ac modification in active regulatory regions, show that the acquisition, or preservation of active chromatin, is essential for repairing the transcribed genome.

Nevertheless, it should be noted that the levels of chromatin reorganization and gene expression during cellular response to UVC, depends on the UVC exposure dose (Farrell et al.,

2011; G. Li & Ho, 1998). For example, a recent study showed that when mouse embryonic fibroblasts were exposed to a UV dose of $80\,J\,/\,m^2$, extensive chromatin reorganization was observed regarding both chromatin accessibility and histone modification levels (Schick et al., 2015). It seems that when cells are dealing with larger amounts of damages, they make drastic decisions related to the activation and of apoptosis programs, which reduce the risk of malignant cell transformation. Such cell fate decisions are accomplished through major alteration in the structure of chromatin and the pattern of gene expression. On the contrary, this study shows that low doses of UVC ($8-20\,J\,/\,m^2$), do not drive the cells to apoptosis, but triggers the mechanisms that promote the repair of DNA lesions.

Recent research in the field of transcriptional regulation, specifically focusing on the transcriptional response to heat stress, demonstrates that the activation of paused genes occurs through a transition from a state of premature transcription, to a state of elongation (Krebs et al., 2017). The above suggests that rapid induction of gene transcription requires a state of uninterrupted transcription initiation. In the case of transcriptional response to UVC exposure, we observe the release of transcriptional waves from PPP sites, into active genes. In addition, a recent study provides data suggesting that transcription pausing at PPP sites inhibits the initiation of transcription, as reduced RNAPII pausing leads to increased transcription initiation and nRNA production (Fitz et al., 2018). Consistent with the above, our findings show that RNAPII release from PPP sites, and the increase of nRNA signal at these regions are sufficient to lead to a de novo initiation of transcription and recruitment of RNAPII molecules at PIC regions, and in particular in active TSSs, PROMPTs and eTSSs. In a more general context, we can say that the results of this study extend the idea that continuous release of RNAPII molecules from PPP sites is an important element of regulation of gene expression (Steurer et al., 2018).

Moreover, the particular results suggest that bidirectional transcription starts from two distinct transcription initiation sites (PICs), corresponding to a Nucleosome Depleted Region (NDR) (Core et al., 2012; Ibrahim et al., 2018; Lai & Pugh, 2017). Indeed, in the bidirectionally transcribed genes, as also in the mRNA-PROMPTs pairs, the binding of RNAPII-hypo takes place at both ends of a highly accessible chromatin region (based on the ATAC-seq signal), surrounded by H3K27ac nucleosomes. The above is consistent with the fact that the mRNA PICs structure has a common architecture with non-coding PICs (Lai & Pugh, 2017). Consequently, it is arguable that differences in the level of transcription between different types of transcription elements (mRNAs, asPROMPTs, eRNAs) under normal conditions occur primarily because of the transcription initiation rate, the premature transcription termination, and the sensitivity that non-coding transcripts show in exosome degradation.

Regarding the newly developed aniFOUND-seq methodology, it can be considered as a very useful tool to complement XR-seq and subtractive Damage-seq (Adar et al., 2016; Hu et al., 2015, 2017), methods for providing all together a set of tools for the study of DNA damage and repair. The particular method is applied to map the repair-synthesis activity across the genome, with particular emphasis to promoters, enhancers and repeats. The newly developed analysis pipeline is specifically designed for the assessment of NER-UDS activity during the first 4 hours

after damage induction in particular chromosomal regions such as rDNA and telomeres, for which contradictory explanatory models have been suggested. Notably, this is the first time that NGS-based approaches have been adapted to shed light to these issues, especially regarding telomeric DNA. Thus, aniFOUND's unbiased (antibody-free) manner of detecting DNA repair-synthesis activity may offer advantages for refining the spatio-temporal understanding of genome maintenance requiring UDS after damage. The flexible nature of aniFOUND-seq (in terms of both damage types detected and the potential repair assessment period) renders it suitable for capturing of the whole repair process or repair activity during shorter or longer time windows thus allowing alternative perspectives of repaired-synthesized chromatin to be captured.

Importantly, aniFOUND-seq analysis results are in agreement with, and complete previous reports showing how NER activity is implemented with different speeds in different genomic areas/regions (Hu et al., 2015, 2017). Taken together these results confirm that aniFOUND can isolate and map in high resolution nascent chromatin loci that have undergone efficient NER of UV-lesions.

6 References

Adams, J. (2008). DNA Sequencing Technologies | Learn Science at Scitable. *Nature Education*.

Adar, S., Hu, J., Lieb, J. D., & Sancar, A. (2016). Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1603388113

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*. https://doi.org/10.1038/nature12477

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*. https://doi.org/10.1016/S0022-2836(05)80360-2

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu638

Andersson, R., Chen, Y., Core, L., Lis, J. T., Sandelin, A., & Jensen, T. H. (2015). Human Gene Promoters Are Intrinsically Bidirectional. In *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.10.015

Andersson, R., Refsing Andersen, P., Valen, E., Core, L. J., Bornholdt, J., Boyd, M., Heick Jensen, T., & Sandelin, A. (2014). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nature Communications*. https://doi.org/10.1038/ncomms6336

Andrade-Lima, L. C., Veloso, A., Paulsen, M. T., Menck, C. F. M., & Ljungman, M. (2015). DNA repair and recovery of RNA synthesis following exposure to ultraviolet light are delayed in long genes. *Nucleic Acids Research*, *43*(5), 2744–2756. https://doi.org/10.1093/nar/gkv148

Andrews, S. (2015). FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Institute*.

Anindya, R., Aygün, O., & Svejstrup, J. Q. (2007). Damage-Induced Ubiquitylation of Human RNA Polymerase II by the Ubiquitin Ligase Nedd4, but Not Cockayne Syndrome Proteins or BRCA1. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2007.10.008

Ardehali, M. B., & Lis, J. T. (2009). Tracking rates of transcription and splicing in vivo. In *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb1109-1123

Arlett, C. F., Green, M. H. L., Rogers, P. B., Lehmann, A. R., & Plowman, P. N. (2008). Minimal ionizing radiation sensitivity in a large cohort of xeroderma pigmentosum fibroblasts. *British Journal of Radiology*. https://doi.org/10.1259/bjr/27072321

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. In *Cell Research*. https://doi.org/10.1038/cr.2011.22

Bataille, A. R., Jeronimo, C., Jacques, P. É., Laramée, L., Fortin, M. È., Forest, A., Bergeron, M., Hanes, S. D., & Robert, F. (2012). A Universal RNA Polymerase II CTD Cycle Is Orchestrated by Complex Interplays between Kinase, Phosphatase, and Isomerase Enzymes along Genes. *Molecular Cell*, *45*(2), 158–170. https://doi.org/10.1016/j.molcel.2011.11.024

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., … Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1099

Beltran, M., Yates, C. M., Skalska, L., Dawson, M., Reis, F. P., Viiri, K., Fisher, C. L., Sibley, C. R., Foster, B. M., Bartke, T., Ule, J., & Jenner, R. G. (2016). The interaction of PRC2 with RNA or chromatin s mutually antagonistic. *Genome Research*. https://doi.org/10.1101/gr.197632.115

Bensaude, O. (2011). Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription*. https://doi.org/10.4161/trns.2.3.16172

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. In *Nature Reviews Genetics* (Vol. 15, Issue 3, pp. 163–175). https://doi.org/10.1038/nrg3662

Best, M. A. (2004). Bioinformatics: the Machine Learning Approach, 2nd edn. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. https://doi.org/10.1111/j.1467-985x.2004.298_2.x

Birger, Y., West, K. L., Postnikov, Y. V., Lim, J. H., Furusawa, T., Wagner, J. P., Laufer, C. S., Kraemer, K. H., & Bustin, M. (2003). Chromosomal protein HMGN1 enhances the rate of DNA repair in chromatin. *EMBO Journal*, *22*(7), 1665–1675. https://doi.org/10.1093/emboj/cdg142

Bishop, C. M. (2006). Pattern Recoginiton and Machine Learning. In *Information Science and Statistics*.

Blasco, M. A. (2005). Telomeres and human disease: Ageing, cancer and beyond. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg1656

Boeing, S., Rigault, C., Heidemann, M., Eick, D., & Meisterernst, M. (2010). RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M109.046565

Boeing, S., Williamson, L., Encheva, V., Gori, I., Saunders, R. E., Instrell, R., Aygün, O., Rodriguez-Martinez, M., Weems, J. C., Kelly, G. P., Conaway, J. W., Conaway, R. C., Stewart, A., Howell, M., Snijders, A. P., & Svejstrup, J. Q. (2016). Multiomic Analysis of the UV-Induced DNA Damage Response. *Cell Reports*. https://doi.org/10.1016/j.celrep.2016.04.047

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu170

Borisova, M. E., Voigt, A., Tollenaere, M. A. X., Sahu, S. K., Juretschke, T., Kreim, N., Mailand, N., Choudhary, C., Bekker-Jensen, S., Akutsu, M., Wagner, S. A., & Beli, P. (2018). P38-MK2 signaling axis regulates RNA metabolism after UV-light-induced DNA damage. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03417-3

Brannan, K., Kim, H., Erickson, B., Glover-Cutter, K., Kim, S., Fong, N., Kiemele, L., Hansen, K., Davis, R., Lykke-Andersen, J., & Bentley, D. L. (2012). MRNA Decapping Factors and the Exonuclease Xrn2 Function in Widespread Premature Termination of RNA Polymerase II Transcription. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2012.03.006

Bregman, D. B., Halaban, R., van Gool, A. J., Henning, K. A., Friedberg, E. C., & Warren, S. L. (1996). UV-induced ubiquitination of RNA polymerase II: a novel modification deficient in Cockayne syndrome cells. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.93.21.11586

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., & Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*. https://doi.org/10.1038/nmeth.2645

Brewer, C. A., Hatchard, G. W., & Harrower, M. A. (2003). ColorBrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science*. https://doi.org/10.1559/152304003100010929

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*. https://doi.org/10.1038/nmeth.2688

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*. https://doi.org/10.1002/0471142727.mb2129s109

Bugai, A., Quaresma, A. J. C., Friedel, C. C., Lenasi, T., Düster, R., Sibley, C. R., Fujinaga, K., Kukanja, P., Hennig, T., Blasius, M., Geyer, M., Ule, J., Dölken, L., & Barborič, M. (2019a). P-TEFb Activation by RBM7 Shapes a Pro-survival Transcriptional Response to Genotoxic Stress. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2019.01.033

Bugai, A., Quaresma, A. J. C., Friedel, C. C., Lenasi, T., Düster, R., Sibley, C. R., Fujinaga, K., Kukanja, P., Hennig, T., Blasius, M., Geyer, M., Ule, J., Dölken, L., & Barborič, M. (2019b). P-TEFb Activation by RBM7 Shapes a Pro-survival Transcriptional Response to Genotoxic Stress. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2019.01.033

Bunch, H. (2017). RNA polymerase II pausing and transcriptional regulation of the HSP70 expression. In *European Journal of Cell Biology*. https://doi.org/10.1016/j.ejcb.2017.09.003

Cadet, J., Douki, T., Gasparutto, D., & Ravanat, J. L. (2003). Oxidative damage to DNA: Formation, measurement and biochemical features. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*. https://doi.org/10.1016/j.mrfmmm.2003.09.001

Chang, P., Gohain, M., Yen, M. R., & Chen, P. Y. (2018). Computational Methods for Assessing Chromatin Hierarchy. In *Computational and Structural Biotechnology Journal*. https://doi.org/10.1016/j.csbj.2018.02.003

Chen, F., Gao, X., Shilatifard, A., & Shilatifard, A. (2015). Stably paused genes revealed through inhibition of transcription initiation by the TFIIH inhibitor triptolide. *Genes and Development*. https://doi.org/10.1101/gad.246173.114

Chen, R., Liu, M., Li, H., Xue, Y., Ramey, W. N., He, N., Ai, N., Luo, H., Zhu, Y., Zhou, N., & Zhou, Q. (2008). PP2B and PP1α cooperatively disrupt 7SK snRNP to release P-TEFb for transcription in response to Ca2+signaling. *Genes and Development*, *22*(10), 1356–1368. https://doi.org/10.1101/gad.1636008

Chiou, Y. Y., Hu, J., Sancar, A., & Selby, C. P. (2018). RNA polymerase II is released from the DNA template during transcription-coupled repair in mammalian cells. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.RA117.000971

Ciccia, A., & Elledge, S. J. (2010). The DNA Damage Response: Making It Safe to Play with Knives. In *Molecular Cell* (Vol. 40, Issue 2, pp. 179–204). https://doi.org/10.1016/j.molcel.2010.09.019

Cinghu, S., Yang, P., Kosak, J. P., Conway, A. E., Kumar, D., Oldfield, A. J., Adelman, K., & Jothi, R. (2017). Intragenic Enhancers Attenuate Host Gene Expression. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2017.09.010

Citterio, E., Van Den Boom, V., Schnitzler, G., Kanaar, R., Bonte, E., Kingston, R. E., Hoeijmakers, J. H. J., & Vermeulen, W. (2000). ATP-Dependent Chromatin Remodeling by the Cockayne Syndrome B DNA Repair-Transcription-Coupling Factor. *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.20.20.7643-7653.2000

Cockayne, E. A. (1936). Dwarfism with Retinal Atrophy and Deafness. *Archives of Disease in Childhood*, *11*(61), 1–8. https://doi.org/10.1136/adc.21.105.52

Cockayne, E. A. (1946). Dwarfism with retinal atrophy and deafness. *Archives of Disease in Childhood*. https://doi.org/10.1136/adc.21.105.52

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., … Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*. https://doi.org/10.1038/nmeth.4396

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*. https://doi.org/10.1038/ng.3142

Core, L. J., Waterfall, J. J., Gilchrist, D. A., Fargo, D. C., Kwak, H., Adelman, K., & Lis, J. T. (2012). Defining the Status of RNA Polymerase at Promoters. *Cell Reports*. https://doi.org/10.1016/j.celrep.2012.08.034

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*. https://doi.org/10.1126/science.1162228

Cournac, A., Koszul, R., & Mozziconacci, J. (2016). The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1292

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1016071107

Cunningham, P. (1999). Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. R. Durbin, S. Eddy, A. Krogh and G. Mitchison. *Cell Biochemistry and Function*. https://doi.org/10.1002/(sici)1099-0844(199903)17:1<73::aid-cbf799>3.3.co;2-#

Darzacq, X., Shav-Tal, Y., De Turris, V., Brody, Y., Shenoy, S. M., Phair, R. D., & Singer, R. H. (2007). In vivo dynamics of RNA polymerase II transcription. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb1280

Datta, A., Bagchi, S., Nag, A., Shiyanov, P., Adami, G. R., Yoon, T., & Raychaudhuri, P. (2001). The p48 subunit of the damaged-DNA binding protein DDB associates with the CBP/p300 family of histone acetyltransferase. *Mutation Research - DNA Repair*, *486*(2), 89–97. https://doi.org/10.1016/S0921-8777(01)00082-9

De Bont, R., & van Larebeke, N. (2004). Endogenous DNA damage in humans: A review of quantitative data. In *Mutagenesis*. https://doi.org/10.1093/mutage/geh025

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1002384

de la Torre-Ubieta, L., Stein, J. L., Won, H., Opland, C. K., Liang, D., Lu, D., & Geschwind, D. H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis.

Cell. https://doi.org/10.1016/j.cell.2017.12.014
de Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C. L., & Natoli, G. (2010). A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1000384
Dinant, C., Ampatziadis-Michailidis, G., Lans, H., Tresini, M., Lagarou, A., Grosbart, M., Theil, A. F., vanCappellen, W. A., Kimura, H., Bartek, J., Fousteri, M., Houtsmuller, A. B., Vermeulen, W., & Marteijn, J. A. (2013). Enhanced chromatin dynamics by fact promotes transcriptional restart after UV-induced DNA damage. *Molecular Cell*, *51*(4), 469–479. https://doi.org/10.1016/j.molcel.2013.08.007
Donahue, B. A., Yin, S., Taylor, J. S., Reines, D., & Hanawalt, P. C. (1994). Transcript cleavage by RNA polymerase II arrested by a cyclobutane pyrimidine dimer in the DNA template. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(18), 8502–8506. https://doi.org/10.1073/pnas.91.18.8502
Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*. https://doi.org/10.1038/nature11247
Egloff, S., Dienstbier, M., & Murphy, S. (2012). Updating the RNA polymerase CTD code: Adding gene-specific layers. In *Trends in Genetics* (Vol. 28, Issue 7, pp. 333–341). https://doi.org/10.1016/j.tig.2012.03.007
Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., & Murphy, S. (2012). Ser7 phosphorylation of the CTD recruits the RPAP2 ser5 phosphatase to snRNA genes. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2011.11.006
Eickbush, T. H., & Moudrianakis, E. N. (1978). The Histone Core Complex: An Octamer Assembled by Two Sets of Protein-Protein Interactions. *Biochemistry*. https://doi.org/10.1021/bi00616a016
ENCODE, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*. https://doi.org/nature11247 [pii]\n10.1038/nature11247
Ernst, J., & Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocol*s. https://doi.org/10.1038/nprot.2017.124
Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., D'Eustachio, P., Stein, L., & Hermjakob, H. (2017). Reactome pathway analysis: A high-performance in-memory approach. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-017-1559-2
Farrell, A. W., Halliday, G. M., & Lyons, J. G. (2011). Chromatin structure following UV-induced DNA damage-repair or death? In *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms12118063
Fedeli, U., Girardi, P., & Mastrangelo, G. (2019). Occupational exposure to vinyl chloride and liver diseases. In *World Journal of Gastroenterology*. https://doi.org/10.3748/wjg.v25.i33.4885
Feuerbach, L., Sieverling, L., Deeg, K. I., Ginsbach, P., Hutter, B., Buchhalter, I., Northcott, P. A., Mughal, S. S., Chudasama, P., Glimm, H., Scholl, C., Lichter, P., Fröhling, S., Pfister, S. M., Jones, D. T. W., Rippe, K., & Brors, B. (2019). TelomereHunter - In silico estimation of telomere content and composition from cancer genomes. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-019-2851-0
Fitz, J., Neumann, T., & Pavri, R. (2018). Regulation of RNA polymerase II processivity by Spt5 is restricted to a narrow window during elongation. *The EMBO Journal*. https://doi.org/10.15252/embj.201797965
Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kä hä ri, A.,

Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., … Searle, S. M. J. (2011). Ensembl 2011. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq1064

Fousteri, M., Vermeulen, W., van Zeeland, A. A., & Mullenders, L. H. F. (2006). Cockayne Syndrome A and B Proteins Differentially Regulate Recruitment of Chromatin Remodeling and Repair Factors to Stalled RNA Polymerase II In Vivo. *Molecular Cell*, *23*(4), 471–482. https://doi.org/10.1016/j.molcel.2006.06.029

Friedberg, E. C., Walker, G. C., Siede, W., Wood, R. D., Schultz, R. A., & Ellenberger, T. (2005). DNA Repair and Mutagenesis. In *DNA Repair and Mutagenesis*. https://doi.org/10.1128/9781555816704

Fu, T. J., Peng, J., Lee, G., Price, D. H., & Flores, O. (1999). Cyclin K functions as a CDK9 regulatory subunit and participates in RNA polymerase II transcription. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.274.49.34527

Fuss, J. O., & Cooper, P. K. (2006). DNA repair: Dynamic defenders against cancer and aging. In *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0040203

Garcia-Perez, J. L., Widmann, T. J., & Adams, I. R. (2016). The impact of transposable elements on mammalian development. In *Development (Cambridge)*. https://doi.org/10.1242/dev.132639

Garinis, G. A., Uittenboogaard, L. M., Stachelscheid, H., Fousteri, M., van Ijcken, W., Breit, T. M., van Steeg, H., Mullenders, L. H. F., van der Horst, G. T. J., Brüning, J. C., Niessen, C. M., Hoeijmakers, J. H. J., & Schumacher, B. (2009). Persistent transcription-blocking DNA lesions trigger somatic growth attenuation associated with longevity. *Nature Cell Biology*. https://doi.org/10.1038/ncb1866

Gilmour, D. S., & Fan, R. (2009). Detecting transcriptionally engaged RNA polymerase in eukaryotic cells with permanganate genomic footprinting. *Methods*. https://doi.org/10.1016/j.ymeth.2009.02.020

Gordon, A., Hannon, G. J., & Gordon. (2014). FASTX-Toolkit. In *[Online] http://hannonlab. cshl. edu/fastx_ toolkit http://hannonlab. cshl. edu/fastx_ toolkit*.

Gray, L. T., Yao, Z., Nguyen, T. N., Kim, T. K., Zeng, H., & Tasic, B. (2017). Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *ELife*. https://doi.org/10.7554/elife.21883

Green, C. M., & Almouzni, G. (2002). When repair meets chromatin. First in series on chromatin dynamics. In *EMBO Reports* (Vol. 3, Issue 1, pp. 28–33). https://doi.org/10.1093/embo-reports/kvf005

Gross, D. (1988). Nuclease Hypersensitive Sites In Chromatin. *Annual Review of Biochemistry*, *57*(1), 159–197. https://doi.org/10.1146/annurev.biochem.57.1.159

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw313

Guo, R., Chen, J., Mitchell, D. L., & Johnson, D. G. (2011). GCN5 and E2F1 stimulate nucleotide excision repair by promoting H3K9 acetylation at sites of damage. *Nucleic Acids Research*, *39*(4), 1390–1397. https://doi.org/10.1093/nar/gkq983

Gyenis, Á., Umlauf, D., Újfaludi, Z., Boros, I., Ye, T., & Tora, L. (2014). UVB Induces a Genome-Wide Acting Negative Regulatory Mechanism That Operates at the Level of Transcription Initiation in Human Cells. *PLoS Genetics*, *10*(7). https://doi.org/10.1371/journal.pgen.1004483

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. In *Nature Structural and Molecular Biology* (Vol. 11, Issue 5, pp. 394–403). https://doi.org/10.1038/nsmb763

Hanawalt, P. C., & Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. *Nature Reviews. Molecular Cell Biology*, *9*(12), 958–970. https://doi.org/10.1038/nrm2549

Haradhvala, N. J., Polak, P., Stojanov, P., Covington, K. R., Shinbrot, E., Hess, J. M., Rheinbay, E., Kim, J., Maruvka, Y. E., Braunstein, L. Z., Kamburov, A., Hanawalt, P. C., Wheeler, D. A., Koren, A., Lawrence, M. S., & Getz, G. (2016). Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell*. https://doi.org/10.1016/j.cell.2015.12.050

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*. https://doi.org/10.1101/gr.135350.111

Hashimoto, S. I., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S., & Matsushima, K. (2004). 5′-end SAGE for the analysis of transcriptional start sites. *Nature Biotechnology*. https://doi.org/10.1038/nbt998

Hefferin, M. L., & Tomkinson, A. E. (2005). Mechanism of DNA double-strand break repair by non-homologous end joining. In *DNA Repair*. https://doi.org/10.1016/j.dnarep.2004.12.005

Heidemann, M., Hintermair, C., Vo??, K., & Eick, D. (2013). Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. In *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* (Vol. 1829, Issue 1, pp. 55–62). https://doi.org/10.1016/j.bbagrm.2012.08.013

Heine, G. F., Horwitz, A. A., & Parvin, J. D. (2008). Multiple mechanisms contribute to inhibit transcription in response to DNA damage. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M707700200

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2010.05.004

Helena, J. M., Joubert, A. M., Grobbelaar, S., Nolte, E. M., Nel, M., Pepper, M. S., Coetzee, M., & Mercier, A. E. (2018). Deoxyribonucleic acid damage and repair: Capitalizing on our understanding of the mechanisms of maintaining genomic integrity for therapeutic purposes. In *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms19041148

Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3729

Henriques, T., Scruggs, B. S., Inouye, M. O., Muse, G. W., Williams, L. H., Burkholder, A. B., Lavender, C. A., Fargo, D. C., & Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. *Genes and Development*. https://doi.org/10.1101/gad.309351.117

Herrmann, C., Van De Sande, B., Potier, D., & Aerts, S. (2012). i-cisTarget: An integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks543

Hirose, Y., & Ohkuma, Y. (2007). Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. In *Journal of Biochemistry*. https://doi.org/10.1093/jb/mvm090

Horibata, K., Iwamoto, Y., Kuraoka, I., Jaspers, N. G. J., Kurimasa, A., Oshimura, M., Ichihashi, M., & Tanaka, K. (2004). From The Cover: Complete absence of Cockayne syndrome group B gene product gives rise to UV-sensitive syndrome but not Cockayne syndrome. *Proceedings of the National Academy of Sciences*, *101*(43), 15410–15415. https://doi.org/10.1073/pnas.0404587101

Hsu, F., Kent, J. W., Clawson, H., Kuhn, R. M., Diekhans, M., & Haussler, D. (2006). The UCSC known genes. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl048

Hu, J., Adar, S., Selby, C. P., Lieb, J. D., & Sancar, A. (2015). Genome-wide analysis of human

global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes and Development*. https://doi.org/10.1101/gad.261271.115

Hu, J., Adebali, O., Adar, S., & Sancar, A. (2017). Dynamic maps of UV damage formation and repair for the human genome. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1706522114

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., … Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research*. https://doi.org/10.1093/nar/30.1.38

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1272

Ibrahim, M. M., Karabacak, A., Glahs, A., Kolundzic, E., Hirsekorn, A., Carda, A., Tursun, B., Zinzen, R. P., Lacadie, S. A., & Ohler, U. (2018). Determinants of promoter and enhancer transcription directionality in metazoans. *Nature Communications*. https://doi.org/10.1038/s41467-018-06962-z

Imrichová, H., Hulselmans, G., Atak, Z. K., Potier, D., & Aerts, S. (2015). I-cisTarget 2015 update: Generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv395

Ip, J. Y., Schmidt, D., Pan, Q., Ramani, A. K., Fraser, A. G., Odom, D. T., & Blencowe, B. J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Research*. https://doi.org/10.1101/gr.111070.110

Ito, S., Kuraoka, I., Chymkowitch, P., Compe, E., Takedachi, A., Ishigami, C., Coin, F., Egly, J. M., & Tanaka, K. (2007). XPG Stabilizes TFIIH, Allowing Transactivation of Nuclear Receptors: Implications for Cockayne Syndrome in XP-G/CS Patients. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2007.03.013

Jackson, S. P., & Helleday, T. (2016). Drugging DNA repair. *Science*. https://doi.org/10.1126/science.aab0958

Jensen, A., & Mullenders, L. H. F. (2010). Transcription factor IIS impacts UV-inhibited transcription. *DNA Repair*. https://doi.org/10.1016/j.dnarep.2010.08.002

Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M., & Wong, W. H. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnology*. https://doi.org/10.1038/nbt.1505

John, S., Sabo, P. J., Thurman, R. E., Sung, M. H., Biddie, S. C., Johnson, T. A., Hager, G. L., & Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. In *Nature Genetics*. https://doi.org/10.1038/ng.759

Jonkers, I., Kwak, H., & Lis, J. T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *ELife*, *2014*(3). https://doi.org/10.7554/eLife.02407

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*. https://doi.org/10.1159/000084979

Kaikkonen, M. U., & Adelman, K. (2018). Emerging Roles of Non-Coding RNA Transcription. In *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2018.06.002

Kaneko, S., Son, J., Bonasio, R., Shen, S. S., & Reinberg, D. (2014). Nascent RNA interaction keeps PRC2 activity poised and in check. *Genes and Development*. https://doi.org/10.1101/gad.247940.114

Karikkineth, A. C., Scheibye-Knudsen, M., Fivenson, E., Croteau, D. L., & Bohr, V. A. (2017). Cockayne syndrome: Clinical features, model systems and pathways. In *Ageing Research Reviews* (Vol. 33, pp. 3–17). https://doi.org/10.1016/j.arr.2016.08.002

Karlic, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., & Vingron, M. (2010). Histone modification

levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0909344107
Karolchik, D., Hinricks, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkh103
Kelly, W. G., Dahmus, M. E., & Hart, G. W. (1993). RNA polymerase II is a glycoprotein. Modification of the COOH-terminal domain by O-GlcNAc. *Journal of Biological Chemistry*, *268*(14), 10416–10424.
Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*. https://doi.org/10.1101/gr.229102
Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq351
Kidwell, M. G., & Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.94.15.7704
Kim, D., Langmead, B., & Salzberg, S. L. (2015). Hisat2. *Nature Methods*. https://doi.org/10.1038/nmeth.3317
Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J. H., Pollock, D. D., Megee, P. C., & Bentley, D. L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nature Structural and Molecular Biology*, *17*(10), 1279–1286. https://doi.org/10.1038/nsmb.1913
Kim, J., Mouw, K. W., Polak, P., Braunstein, L. Z., Kamburov, A., Tiao, G., Kwiatkowski, D. J., Rosenberg, J. E., Van Allen, E. M., D'Andrea, A. D., & Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics*. https://doi.org/10.1038/ng.3557
Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., & Flicek, P. (2011). Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*. https://doi.org/10.1093/database/bar030
Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T. K., Zacarias-Cabeza, J., Spicuglia, S., De La Chapelle, A. L., Heidemann, M., Hintermair, C., Eick, D., Gut, I., Ferrier, P., & Andrau, J. C. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.2085
Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., & Carninci, P. (2006). Cage: Cap analysis of gene expression. *Nature Methods*. https://doi.org/10.1038/nmeth0306-211
Koerner, M. V., Pauler, F. M., Huang, R., & Barlow, D. P. (2009). The function of non-coding RNAs in genomic imprinting. In *Development*. https://doi.org/10.1242/dev.030403
Kolman, A., & Bohušová, T. (1992). Induction of 6-thioguanine-resistant mutants in human diploid fibroblats in vitro with ethylene oxide. *Mutation Research/Environmental Mutagenesis and Related Subjects*. https://doi.org/10.1016/0165-1161(92)91225-g
Komarnitsky, P., Cho, E. J., & Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes and Development*, *14*(19), 2452–2460. https://doi.org/10.1101/gad.824700
Kouzarides, T. (2007). Chromatin Modifications and Their Function. In *Cell* (Vol. 128, Issue 4, pp. 693–705). https://doi.org/10.1016/j.cell.2007.02.005
Kowalski, A., & Pałyga, J. (2012). Linker histone subtypes and their allelic variants. *Cell Biology International*. https://doi.org/10.1042/cbi20120133

Kraemer, K. H., Patronas, N. J., Schiffmann, R., Brooks, B. P., Tamura, D., & DiGiovanna, J. J. (2007). Xeroderma pigmentosum, trichothiodystrophy and Cockayne syndrome: A complex genotype-phenotype relationship. In *Neuroscience* (Vol. 145, Issue 4, pp. 1388–1396). https://doi.org/10.1016/j.neuroscience.2006.12.020

Krebs, A. R., Imanci, D., Hoerner, L., Gaidatzis, D., Burger, L., & Schübeler, D. (2017). Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2017.06.027

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., & Haussler, D. (1994). Hidden Markov Models in computational biology applications to protein modeling. *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1994.1104

Lai, W. K. M., & Pugh, B. F. (2017). Genome-wide uniformity of human “open” pre-initiation complexes. *Genome Research*. https://doi.org/10.1101/gr.210955.116

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. https://doi.org/10.1186/gb-2009-10-3-r25

Laugel, V., Dalloz, C., Durand, M., Sauvanaud, F., Kristensen, U., Vincent, M. C., Pasquier, L., Odent, S., Cormier-Daire, V., Gener, B., Tobias, E. S., Tolmie, J. L., Martin-Coignard, D., Drouin-Garraud, V., Heron, D., Journel, H., Raffo, E., Vigneron, J., Lyonnet, S., … Dollfus, H. (2010). Mutation update for the CSB/ERCC6 and CSA/ERCC8 genes involved in Cockayne syndrome. In *Human Mutation* (Vol. 31, Issue 2, pp. 113–126). https://doi.org/10.1002/humu.21154

Lavigne, M. D., Konstantopoulos, D., Ntakou-Zamplara, K. Z., Liakos, A., & Fousteri, M. (2017). Global unleashing of transcription elongation waves in response to genotoxic stress restricts somatic mutation rate. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-02145-4

Lavigne, M. D., Vatsellas, G., Polyzos, A., Mantouvalou, E., Sianidis, G., Maraziotis, I., Agelopoulos, M., & Thanos, D. (2015). Composite macroH2A/NRF-1 Nucleosomes Suppress Noise and Generate Robustness in Gene Expression. *Cell Reports*. https://doi.org/10.1016/j.celrep.2015.04.022

Lee, K. M., Choi, K. H., & Ouellette, M. M. (2004). Use of exogenous hTERT to immortalize primary human cells. *Cytotechnology*. https://doi.org/10.1007/10.1007/s10616-004-5123-3

Lehmann, A. R. (2000). Replication of UV-damaged DNA: New insights into links between DNA polymerases, mutagenesis and human disease. In *Gene*. https://doi.org/10.1016/S0378-1119(00)00250-X

Lehmann, A. R. (2011). DNA polymerases and repair synthesis in NER in human cells. In *DNA Repair*. https://doi.org/10.1016/j.dnarep.2011.04.023

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq1019

Levens, D., Baranello, L., & Kouzine, F. (2016). Controlling gene expression by DNA mechanics: emerging insights and challenges. In *Biophysical Reviews*. https://doi.org/10.1007/s12551-016-0243-5

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. In *Nature Methods*. https://doi.org/10.1038/nmeth.4346

Levine, M. (2011). Paused RNA polymerase II as a developmental checkpoint. In *Cell*. https://doi.org/10.1016/j.cell.2011.04.021

Li, G., & Ho, V. C. (1998). p53-Dependent DNA repair and apoptosis respond differently to high- and low-dose ultraviolet radiation. *British Journal of Dermatology*. https://doi.org/10.1046/j.1365-2133.1998.02305.x

Li, H. (2012). seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub*.

Li, H. (2013). [Heng Li - Compares BWA to other long read aligners like CUSHAW2] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint*

ArXiv. https://doi.org/arXiv:1303.3997 [q-bio.GN]
Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp324
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp352
Li, W., Notani, D., & Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2016.4
Liakos, A., Konstantopoulos, D., Lavigne, M. D., & Fousteri, M. (2020). Continuous transcription initiation guarantees robust repair of all transcribed genes and regulatory regions. *Nature Communications*. https://doi.org/10.1038/s41467-020-14566-9
Liakos, A., Lavigne, M. D., & Fousteri, M. (2017). Nucleotide excision repair: From neurodegeneration to cancer. In *Advances in Experimental Medicine and Biology* (Vol. 1007, pp. 17–39). https://doi.org/10.1007/978-3-319-60733-7_2
Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt656
Lindahl, T., & Wood, R. D. (1999). Quality control by DNA repair. In *Science*. https://doi.org/10.1126/science.286.5446.1897
Liu, L., Leng, L., Liu, C., Lu, C., Yuan, Y., Wu, L., Gong, F., Zhang, S., Wei, X., Wang, M., Zhao, L., Hu, L., Wang, J., Yang, H., Zhu, S., Chen, F., Lu, G., Shang, Z., & Lin, G. (2019). An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. *Nature Communications*. https://doi.org/10.1038/s41467-018-08244-0
Lloret-Llinares, M., Mapendano, C. K., Martlev, L. H., Lykke-Andersen, S., & Jensen, T. H. (2016). Relationships between PROMPT and gene expression. *RNA Biology*. https://doi.org/10.1080/15476286.2015.1109769
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. https://doi.org/10.1186/s13059-014-0550-8
Lumsden, J. M., McCarty, T., Petiniot, L. K., Shen, R., Barlow, C., Wynn, T. A., Morse, H. C., Gearhart, P. J., Wynshaw-Boris, A., Max, E. E., & Hodes, R. J. (2004). Immunoglobulin class switch recombination is impaired in Atm-deficient mice. *Journal of Experimental Medicine*. https://doi.org/10.1084/jem.20041074
Lun, A. T. L., & Smyth, G. K. (2015). Csaw: A Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1191
Maezawa, S., Yukawa, M., Alavattam, K. G., Barski, A., & Namekawa, S. H. (2017). Dynamic reorganization of open chromatin underlies diverse transcriptomes during spermatogenesis. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1052
Magnuson, B., Veloso, A., Kirkconnell, K. S., De Andrade Lima, L. C., Paulsen, M. T., Ljungman, E. A., Bedi, K., Prasad, J., Wilson, T. E., & Ljungman, M. (2015). Identifying transcription start sites and active enhancer elements using BruUV-seq. *Scientific Reports*, *5*, 1–12. https://doi.org/10.1038/srep17978
Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., Waters, C. T., Munson, K., Core, L. J., & Lis, J. T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nature Protocols*. https://doi.org/10.1038/nprot.2016.086
Manning, C. D., Schütze, H., & Weikurn, G. (2002). Foundations of Statistical Natural Language Processing. *SIGMOD Record*. https://doi.org/10.1145/601858.601867
Mao, P., Smerdon, M. J., Roberts, S. A., & Wyrick, J. J. (2016). Chromosomal landscape of UV

damage formation and repair at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1606667113

Marteijn, J. A., Lans, H., Vermeulen, W., & Hoeijmakers, J. H. J. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology*, *15*(7), 465–481. https://doi.org/10.1038/nrm3822

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*. https://doi.org/10.14806/ej.17.1.200

Martinez, E., Palhan, V. B., Tjernberg, a, Lymar, E. S., Gamper, a M., Kundu, T. K., Chait, B. T., & Roeder, R. G. (2001). Human STAGA complex is a chromatin-acetylating transcription coactivator that interacts with pre-mRNA splicing and DNA damage-binding factors in vivo. *Molecular and Cellular Biology*, *21*(20), 6782–6795. https://doi.org/10.1128/MCB.21.20.6782-6795.2001

Masutani, C., Kusumoto, R., Yamada, A., Dohmae, N., Yokol, M., Yuasa, M., Araki, M., Iwai, S., Takio, K., & Hanaoka, F. (1999). The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase η. *Nature*, *399*(6737), 700–704. https://doi.org/10.1038/21447

Mayer, A., Di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J. A., & Churchman, L. S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell*. https://doi.org/10.1016/j.cell.2015.03.010

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Söding, J., & Cramer, P. (2010). Uniform transitions of the general RNA polymerase II transcription complex. *Nature Structural & Molecular Biology*, *17*(10), 1272–1278. https://doi.org/10.1038/nsmb.1903

Mayne, L., & Lehmann, A. R. (1982). Failure of RNA synthesis to recover after UV irradiation: An early defect in cells from individuals with Cockayne syndrome and xeroderma pigmentosum. *Mutation Research*, *96*(1), 140. https://doi.org/10.1016/0027-5107(82)90047-1

Meng, H., & Bartholomew, B. (2018). Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II. In *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.R117.813485

Misteli, T., & Soutoglou, E. (2009). The emerging role of nuclear architecture in DNA repair and genome maintenance. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm2651

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. https://doi.org/10.1038/nmeth.1226

Moser, J., Kool, H., Giakzidis, I., Caldecott, K., Mullenders, L. H. F., & Fousteri, M. I. (2007). Sealing of Chromosomal DNA Nicks during Nucleotide Excision Repair Requires XRCC1 and DNA Ligase IIIα in a Cell-Cycle-Specific Manner. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2007.06.014

Mueller, B., Mieczkowski, J., Kundu, S., Wang, P., Sadreyev, R., Tolstorukov, M. Y., & Kingston, R. E. (2017). Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes and Development*. https://doi.org/10.1101/gad.293118.116

Munro, P., Toivonen, H., Webb, G. I., Buntine, W., Orbanz, P., Teh, Y. W., Poupart, P., Sammut, C., Sammut, C., Blockeel, H., Rajnarayan, D., Wolpert, D., Gerstner, W., Page, C. D., Natarajan, S., & Hinton, G. (2011). Baum-Welch Algorithm. In *Encyclopedia of Machine Learning*. https://doi.org/10.1007/978-0-387-30164-8_59

Murray, S. C., & Mellor, J. (2016). Using both strands: The fundamental nature of antisense transcription. In *BioArchitecture*. https://doi.org/10.1080/19490992.2015.1130779

Murtha, M., Strino, F., Tokcaer-Keskin, Z., Sumru Bayin, N., Shalabi, D., Xi, X., Kluger, Y., & Dailey, L. (2015). Comparative FAIRE-seq analysis reveals distinguishing features of the chromatin structure of ground state- and primed-pluripotent cells. *Stem Cells*, *33*(2), 378–391. https://doi.org/10.1002/stem.1871

Nag, R., Wong, K. H., & Fallside, F. (1986). SCRIPT RECOGNITION USING HIDDEN MARKOV MODELS. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Nakazawa, Y., Sasaki, K., Mitsutake, N., Matsuse, M., Shimada, M., Nardo, T., Takahashi, Y., Ohyama, K., Ito, K., Mishima, H., Nomura, M., Kinoshita, A., Ono, S., Takenaka, K., Masuyama, R., Kudo, T., Slor, H., Utani, A., Tateishi, S., … Ogi, T. (2012). Mutations in UVSSA cause UV-sensitive syndrome and impair RNA polymerase IIo processing in transcription-coupled nucleotide-excision repair. *Nature Genetics*, *44*(5), 586–592. https://doi.org/10.1038/ng.2229

Nardo, T., Oneda, R., Spivak, G., Vaz, B., Mortier, L., Thomas, P., Orioli, D., Laugel, V., Stary, A., Hanawalt, P. C., Sarasin, A., & Stefanini, M. (2009). A UV-sensitive syndrome patient with a specific CSA mutation reveals separable roles for CSA in response to UV and oxidative DNA damage. *Proceedings of the National Academy of Sciences*, *106*(15), 6209–6214. https://doi.org/10.1073/pnas.0902113106

Nguyen, V. T., Kiss, T., Michels, A. A., & Bensaude, O. (2001). 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature*. https://doi.org/10.1038/35104581

Nicolai, S., Filippi, S., Caputo, M., Cipak, L., Gregan, J., Ammerer, G., Frontini, M., Willems, D., Prantera, G., Balajee, A. S., & Proietti-De-Santis, L. (2015). Identification of Novel Proteins Co-Purifying with Cockayne Syndrome Group B (CSB) Reveals Potential Roles for CSB in RNA Metabolism and Chromatin Dynamics. *PloS One*, *10*(6), e0128558. https://doi.org/10.1371/journal.pone.0128558

Nishimura, D. (2000). RepeatMasker. *Biotech Software & Internet Report*. https://doi.org/10.1089/152791600319259

Noguchi, S., Arakawa, T., Fukuda, S., Furuno, M., Hasegawa, A., Hori, F., Ishikawa-Kato, S., Kaida, K., Kaiho, A., Kanamori-Katayama, M., Kawashima, T., Kojima, M., Kubosaki, A., Manabe, R. I., Murata, M., Nagao-Sato, S., Nakazato, K., Ninomiya, N., Nishiyori-Sueki, H., … Hayashizaki, Y. (2017). FANTOM5 CAGE profiles of human and mouse samples. *Scientific Data*. https://doi.org/10.1038/sdata.2017.112

Ntini, E., Järvelin, A. I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P. R., Andersen, P. K., Preker, P., Valen, E., Zhao, X., Pelechano, V., Steinmetz, L. M., Sandelin, A., & Jensen, T. H. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.2640

O'Driscoll, M., Ruiz-Perez, V. L., Woods, C. G., Jeggo, P. A., & Goodship, J. A. (2003). A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. *Nature Genetics*. https://doi.org/10.1038/ng1129

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1189

Obenchain, V. (2013). Counting reads with summarizeOverlaps. *Bioconductor*. https://doi.org/10.1002/aja.1001960208

Ogi, T., Limsirichaikul, S., Overmeer, R. M., Volker, M., Takenaka, K., Cloney, R., Nakazawa, Y., Niimi, A., Miki, Y., Jaspers, N. G., Mullenders, L. H. F., Yamashita, S., Fousteri, M. I., &

Lehmann, A. R. (2010). Three DNA Polymerases, Recruited by Different Mechanisms, Carry Out NER Repair Synthesis in Human Cells. *Molecular Cell*, *37*(5), 714–727. https://doi.org/10.1016/j.molcel.2010.02.009

Oksenych, V., Zhovmer, A., Ziani, S., Mari, P. O., Eberova, J., Nardo, T., Stefanini, M., Giglia-Mari, G., Egly, J. M., & Coin, F. (2013). Histone Methyltransferase DOT1L Drives Recovery of Gene Expression after a Genotoxic Attack. *PLoS Genetics*, *9*(7). https://doi.org/10.1371/journal.pgen.1003611

Paglia, L. La, Laugé, A., Weber, J., Champ, J., Cavaciuti, E., Russo, A., Viovy, J. L., & Stoppa-Lyonnet, D. (2010). ATM germline mutations in women with familial breast cancer and a relative with haematological malignancy. *Breast Cancer Research and Treatment*. https://doi.org/10.1007/s10549-009-0396-z

Parikh, D., Fouquerel, E., Murphy, C. T., Wang, H., & Opresko, P. L. (2015). Telomeres are partly shielded from ultraviolet-induced damage and proficient for nucleotide excision repair of photoproducts. *Nature Communications*. https://doi.org/10.1038/ncomms9214

Pattern Recognition and Machine Learning. (2007). *Journal of Electronic Imaging*. https://doi.org/10.1117/1.2819119

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space . *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. https://doi.org/10.1080/14786440109462720

Perera, D., Poulos, R. C., Shah, A., Beck, D., Pimanda, J. E., & Wong, J. W. H. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature*. https://doi.org/10.1038/nature17437

Pines, A., Vrouwe, M. G., Marteijn, J. A., Typas, D., Luijsterburg, M. S., Cansoy, M., Hensbergen, P., Deelder, A., de Groot, A., Matsumoto, S., Sugasawa, K., Thoma, N., Vermeulen, W., Vrieling, H., & Mullenders, L. (2012). PARP1 promotes nucleotide excision repair through DDB2 stabilization and recruitment of ALC1. *Journal of Cell Biology*, *199*(2), 235–249. https://doi.org/10.1083/jcb.201112132

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordõez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., … Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*. https://doi.org/10.1038/nature08658

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M. L., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordõez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., … Campbell, P. J. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*. https://doi.org/10.1038/nature08629

Polak, P., Lawrence, M. S., Haugen, E., Stoletzki, N., Stojanov, P., Thurman, R. E., Garraway, L. A., Mirkin, S., Getz, G., Stamatoyannopoulos, J. A., & Sunyaev, S. R. (2014). Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature Biotechnology*. https://doi.org/10.1038/nbt.2778

Polo, S. E., & Almouzni, G. (2015). Chromatin dynamics after DNA damage: The legacy of the access-repair-restore model. In *DNA Repair*. https://doi.org/10.1016/j.dnarep.2015.09.014

Porrua, O., & Libri, D. (2015). Transcription termination and the control of the transcriptome: Why, where and how to stop. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm3943

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., Schierup, M. H., & Jensen, T. H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. *Science*. https://doi.org/10.1126/science.1164096

Proudfoot, N. J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. In *Science*. https://doi.org/10.1126/science.aad9926

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*. https://doi.org/10.1038/nmeth.4402

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq033

Rabiner, L. R., & Juang, B. H. (1986). An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*. https://doi.org/10.1109/MASSP.1986.1165342

Ramanathan, B., & Smerdon, M. J. (1986). Changes in nuclear protein acetylation in u.v.-damaged human cells. *Carcinogenesis*. https://doi.org/10.1093/carcin/7.7.1087

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku365

Rapić-Otrin, V., McLenigan, M. P., Bisi, D. C., Gonzalez, M., & Levine, A. S. (2002). Sequential binding of UV DNA damage binding factor and degradation of the p48 subunit as early events after UV irradiation. *Nucleic Acids Research*, *30*(11), 2588–2598. https://doi.org/10.1093/nar/30.11.2588

Rapin, I. (2013). Disorders of nucleotide excision repair. *Pediatric Neurology, Part III*, *113*, 1637–1650. https://doi.org/10.1016/B978-0-444-59565-2.00032-0

Ratner, J. N., Balasubramanian, B., Corden, J., Warren, S. L., & Bregman, D. B. (1998). Ultraviolet radiation-induced ubiquitination and proteasomal degradation of the large subunit of RNA polymerase II: Implications for transcription- coupled DNA repair. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.273.9.5184

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv007

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*. https://doi.org/10.1038/nature14248

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. In *Nature Biotechnology*. https://doi.org/10.1038/nbt.1754

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp616

Rochette, P. J., & Brash, D. E. (2010). Human telomeres are hypersensitive to UV-induced DNA damage and refractory to repair. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1000926

Rockx, D. A., Mason, R., van Hoffen, A., Barton, M. C., Citterio, E., Bregman, D. B., van Zeeland, A. A., Vrieling, H., & Mullenders, L. H. (2000). UV-induced inhibition of transcription involves repression of transcription initiation and phosphorylation of RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(19), 10503–10508. https://doi.org/10.1073/pnas.180169797

Rossetto, D., Avvakumov, N., & Côté, J. (2012). Histone phosphorylation. *Epigenetics*. https://doi.org/10.4161/epi.21975

Rubbi, C. P., & Milner, J. (2003). p53 is a chromatin accessibility factor for nucleotide excision repair of DNA damage. *EMBO Journal*. https://doi.org/10.1093/emboj/cdg082

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., & Lopez-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*. https://doi.org/10.1038/nature17661

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm3952

Sanders, M. H., Bates, S. E., Wilbur, B. S., & Holmquist, G. P. (2004). Repair rates of R-band, G-band and C-band DNA in murine and human cultured cells. *Cytogenetic and Genome Research*. https://doi.org/10.1159/000077464

Schick, S., Fournier, D., Thakurela, S., Sahu, S. K., Garding, A., & Tiwari, V. K. (2015). Dynamics of chromatin accessibility and epigenetic state in response to UV damage. *Journal of Cell Science*. https://doi.org/10.1242/jcs.173633

Schwalb, B., Michel, M., Zacher, B., Hauf, K. F., Demel, C., Tresch, A., Gagneur, J., & Cramer, P. (2016). TT-seq maps the human transient transcriptome. *Science*. https://doi.org/10.1126/science.aad9841

Schwertman, P., Lagarou, A., Dekkers, D. H. W., Raams, A., van der Hoek, A. C., Laffeber, C., Hoeijmakers, J. H. J., Demmers, J. a a, Fousteri, M., Vermeulen, W., & Marteijn, J. a. (2012). UV-sensitive syndrome protein UVSSA recruits USP7 to regulate transcription-coupled repair. *Nature Genetics*, *44*(5), 598–602. https://doi.org/10.1038/ng.2230

Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C., & Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.04.006

Seila, A. C., Core, L. J., Lis, J. T., & Sharp, P. A. (2009). Divergent transcription: A new feature of active promoters. In *Cell Cycle*. https://doi.org/10.4161/cc.8.16.9305

Selby, C. P., Drapkin, R., Reinberg, D., & Sancar, A. (1997). RNA polymerase II stalled at a thymine dimer: Footprint and effect on excision repair. *Nucleic Acids Research*. https://doi.org/10.1093/nar/25.4.787

Shao, Z., Zhang, Y., Yuan, G. C., Orkin, S. H., & Waxman, D. J. (2012). MAnorm: A robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biology*. https://doi.org/10.1186/gb-2012-13-3-r16

Shapiro, J. A., & Von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. In *Biological Reviews of the Cambridge Philosophical Society*. https://doi.org/10.1017/S1464793104006657

Shen, L., Shao, N. Y., Liu, X., Maze, I., Feng, J., & Nestler, E. J. (2013). diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0065598

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., & Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.2136655100

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3682

Sigurdsson, S., Dirac-Svejstrup, A. B., & Svejstrup, J. Q. (2010). Evidence that Transcript Cleavage Is Essential for RNA Polymerase II Transcription and Cell Viability. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2010.02.026

Sikorski, T. W., & Buratowski, S. (2009). The basal initiation machinery: beyond the general transcription factors. In *Current Opinion in Cell Biology*. https://doi.org/10.1016/j.ceb.2009.03.006

Simon, J. A., & Kingston, R. E. (2013). Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. In *Molecular Cell*. https://doi.org/10.1016/j.molcel.2013.02.013

Sims, R. J., Rojas, L. A., Beck, D., Bonasio, R., Schüller, R., Drury, W. J., Eick, D., Reinberg, D., Egloff, S., Murphy, S., Fong, N., Bentley, D. L., Misteli, T., Spector, D. L., Ryan, K., Murthy, K. G. K., Kaneko, S., Manley, J. L., Cheng, D., … Yang, Y. Z. (2011). The C-terminal domain of RNA polymerase II is modified by site-specific methylation. *Science (New York, N.Y.)*, *332*(6025), 99–103. https://doi.org/10.1126/science.1202663

Singh, J., & Padgett, R. A. (2009). Rates of in situ transcription and splicing in large human genes. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.1666

Smolle, M., & Workman, J. L. (2013). Transcription-associated histone modifications and cryptic transcription. In *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* (Vol. 1829, Issue 1, pp. 84–97). https://doi.org/10.1016/j.bbagrm.2012.08.008

Snyder, R. (2012). Leukemia and benzene. In *International journal of environmental research and public health*. https://doi.org/10.3390/ijerph9082875

Soediono, B. (1989). Alberts - Molecular Biology Of The Cell 4th Ed. *Journal of Chemical Information and Modeling*. https://doi.org/10.1017/CBO9781107415324.004

Soria, G., Polo, S. E., & Almouzni, G. (2012). Prime, Repair, Restore: The Active Role of Chromatin in the DNA Damage Response. In *Molecular Cell*. https://doi.org/10.1016/j.molcel.2012.06.002

Spivak, G. (2015). Nucleotide excision repair in humans. In *DNA Repair*. https://doi.org/10.1016/j.dnarep.2015.09.003

Staresincic, L., Fagbemi, A. F., Enzlin, J. H., Gourdin, A. M., Wijgers, N., Dunand-Sauthier, I., Giglia-Mari, G., Clarkson, S. G., Vermeulen, W., & Schärer, O. D. (2009). Coordination of dual incision and repair synthesis in human nucleotide excision repair. *EMBO Journal*, *28*(8), 1111–1120. https://doi.org/10.1038/emboj.2009.49

Stark, R., & Brown, G. (2011). DiffBind : differential binding analysis of ChIP-Seq peak data. *Bioconductor*.

Steurer, B., Janssens, R. C., Geverts, B., Geijer, M. E., Wienholz, F., Theil, A. F., Chang, J., Dealy, S., Pothof, J., van Cappellen, W. A., Houtsmuller, A. B., & Marteijn, J. A. (2018). Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1717920115

Stovner, E. B., Sætrom, P., & Hancock, J. (2019). Epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz232

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*. https://doi.org/10.1016/j.cell.2019.05.031

Sun, D. X., & Jelinek, F. (1999). Statistical Methods for Speech Recognition. *Journal of the American Statistical Association*. https://doi.org/10.2307/2670189

Suzumura, H., & Arisaka, O. (2010). Cerebro-oculo-facio-skeletal syndrome. *Advances in Experimental Medicine and Biology*. https://doi.org/10.1007/978-1-4419-6448-9_19

Svejstrup, J. Q. (2002). Mechanisms of transcription-coupled DNA repair. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm703

Szymański, M., Barciszewska, M. Z., Zywicki, M., & Barciszewski, J. (2003). Noncoding RNA transcripts. In *Journal of Applied Genetics*.

Tanaka, K., Kawai, K., Kumahara, Y., Ikenaga, M., & Okada, Y. (1981). Genetic complementation groups in Cockayne syndrome. *Somatic Cell Genetics*, *7*(4), 445–455. https://doi.org/10.1007/BF01542989

Teng, Y., Bennett, M., Evans, K. E., Zhuang-Jackson, H., Higgs, A., Reed, S. H., & Waters, R. (2011). A novel method for the genome-wide high resolution analysis of DNA damage. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq1036

Thomas, M. C., & Chiang, C. M. (2006). The general transcription machinery and general cofactors. In *Critical reviews in biochemistry and molecular biology*.

https://doi.org/10.1080/10409230600648736
Thomas, M., White, R. L., & Davis, R. W. (1976). Hybridization of RNA to double stranded DNA: Formation of R loops. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.73.7.2294
Thompson, D., Duedal, S., Kirner, J., McGuffog, L., Last, J., Reiman, A., Byrd, P., Taylor, M., & Easton, D. F. (2005). Cancer risks and mortality in heterozygous ATM mutation carriers. *Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/dji141
Tie, F., Banerjee, R., Stratton, C. a, Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M. O., Scacheri, P. C., & Harte, P. J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development (Cambridge, England)*, *136*(18), 3131–3141. https://doi.org/10.1242/dev.037127
Titov, D. V., Gilman, B., He, Q. L., Bhat, S., Low, W. K., Dang, Y., Smeaton, M., Demain, A. L., Miller, P. S., Kugel, J. F., Goodrich, J. A., & Liu, J. O. (2011). XPB, a subunit of TFIIH, is a target of the natural product triptolide. *Nature Chemical Biology*. https://doi.org/10.1038/nchembio.522
Torgovnick, A., & Schumacher, B. (2015). DNA repair mechanisms in cancer development and therapy. In *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2015.00157
Tornaletti, S., Reines, D., & Hanawalt, P. C. (1999). Structural characterization of RNA polymerase II complexes arrested by a cyclobutane pyrimidine dimer in the transcribed strand of template DNA. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.274.34.24124
Tsompana, M., & Buck, M. J. (2014). Chromatin accessibility: A window into the genome. In *Epigenetics and Chromatin* (Vol. 7, Issue 1). https://doi.org/10.1186/1756-8935-7-33
Ucar, D., Márquez, E. J., Chung, C.-H., Marches, R., Rossi, R. J., Uyar, A., Wu, T.-C., George, J., Stitzel, M. L., Palucka, A. K., Kuchel, G. A., & Banchereau, J. (2017). The chromatin accessibility signature of human immune aging stems from CD8 + T cells . *The Journal of Experimental Medicine*. https://doi.org/10.1084/jem.20170416
Venema, J., Bartosova, Z., Natarajan, A. T., Van Zeeland, A. A., & Mullenders, L. H. F. (1992). Transcription affects the rate but not the extent of repair of cyclobutane pyrimidine dimers in the human adenosine deaminase gene. *Journal of Biological Chemistry*.
Vermeulen, W., & Fousteri, M. (2013). Mammalian transcription-coupled excision repair. *Cold Spring Harbor Perspectives in Biology*, *5*(8). https://doi.org/10.1101/cshperspect.a012625
Vispé, S., DeVries, L., Créancier, L., Besse, J., Bréand, S., Hobson, D. J., Svejstrup, J. Q., Annereau, J. P., Cussac, D., Dumontet, C., Guilbaud, N., Barret, J. M., & Bailly, C. (2009). Triptolide is an inhibitor of RNA polymerase I and II-dependent transcription leading predominantly to down-regulation of short-lived mRNA. *Molecular Cancer Therapeutics*. https://doi.org/10.1158/1535-7163.MCT-09-0549
Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*. https://doi.org/10.1109/TIT.1967.1054010
Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*. https://doi.org/10.1007/s12064-012-0162-3
Wan, D., Gong, Y., Qin, W., Zhang, P., Li, J., Wei, L., Zhou, X., Li, H., Qiu, X., Zhong, F., He, L., Yu, J., Yao, G., Jiang, H., Qian, L., Yu, Y., Shu, H., Chen, X., Xu, H., … Gu, J. (2004). Large-scale cDNA transfection screening for genes related to cancer development and progression. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0404089101
Wang, J., Chin, M. Y., & Li, G. (2006). The novel tumor suppressor p33ING2 enhances nucleotide excision repair via inducement of histone H4 acetylation and chromatin relaxation. *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-05-3444

Wang, Lanfeng, Zhou, Y., Xu, L., Xiao, R., Lu, X., Chen, L., Chong, J., Li, H., He, C., Fu, X. D., & Wang, D. (2015). Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature*. https://doi.org/10.1038/nature14482

Wang, Liguo, Wang, S., & Li, W. (2012). RSeQC: Quality control of RNA-seq experiments. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts356

Wang, Q. E., Han, C., Zhao, R., Wani, G., Zhu, Q., Gong, L., Battu, A., Racoma, I., Sharma, N., & Wani, A. A. (2013). P38 MAPK- and Akt-mediated p300 phosphorylation regulates its degradation to facilitate nucleotide excision repair. *Nucleic Acids Research*, *41*(3), 1722–1733. https://doi.org/10.1093/nar/gks1312

Williams, L. H., Fromm, G., Gokey, N. G., Henriques, T., Muse, G. W., Burkholder, A., Fargo, D. C., Hu, G., & Adelman, K. (2015). Pausing of RNA Polymerase II Regulates Mammalian Developmental Potential through Control of Signaling Networks. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.02.003

Williamson, L., Saponaro, M., Boeing, S., East, P., Mitter, R., Kantidakis, T., Kelly, G. P., Lobley, A., Walker, J., Spencer-Dene, B., Howell, M., Stewart, A., & Svejstrup, J. Q. (2017a). UV Irradiation Induces a Non-coding RNA that Functionally Opposes the Protein Encoded by the Same Gene. *Cell*. https://doi.org/10.1016/j.cell.2017.01.019

Williamson, L., Saponaro, M., Boeing, S., East, P., Mitter, R., Kantidakis, T., Kelly, G. P., Lobley, A., Walker, J., Spencer-Dene, B., Howell, M., Stewart, A., & Svejstrup, J. Q. (2017b). UV Irradiation Induces a Non-coding RNA that Functionally Opposes the Protein Encoded by the Same Gene. *Cell*, *168*(5), 843-855.e13. https://doi.org/10.1016/j.cell.2017.01.019

Woudstra, E. C., Gilbert, C., Fellows, J., Jansen, L., Brouwer, J., Erdjument-Bromage, H., Tempst, P., & Svejstrup, J. Q. (2002). A Rad26-Def1 complex coordinates repair and RNA pol II proteolysis in response to DNA damage. *Nature*. https://doi.org/10.1038/415929a

Yamaguchi, Y., Shibata, H., & Handa, H. (2013). Transcription elongation factors DSIF and NELF: Promoter-proximal pausing and beyond. In *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. https://doi.org/10.1016/j.bbagrm.2012.11.007

Ye, T., Krebs, A. R., Choukrallah, M. A., Keime, C., Plewniak, F., Davidson, I., & Tora, L. (2011). seqMINER: An integrated ChIP-seq data interpretation platform. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq1287

You, Y. H., Lee, D. H., Yoon, J. H., Nakajima, S., Yasui, A., & Pfeifer, G. P. (2001). Cyclobutane Pyrimidine Dimers Are Responsible for the Vast Majority of Mutations Induced by UVB Irradiation in Mammalian Cells. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M107696200

Zentner, G. E., & Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. In *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.2470

Zhang, X., Horibata, K., Saijo, M., Ishigami, C., Ukai, A., Kanno, S. I., Tahara, H., Neilan, E. G., Honma, M., Nohmi, T., Yasui, A., & Tanaka, K. (2012). Mutations in UVSSA cause UV-sensitive syndrome and destabilize ERCC6 in transcription-coupled DNA repair. *Nature Genetics*, *44*(5), 593–597. https://doi.org/10.1038/ng.2228

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., & Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*. https://doi.org/10.1186/gb-2008-9-9-r137

Zhong, S., Joung, J. G., Zheng, Y., Chen, Y. R., Liu, B., Shao, Y., Xiang, J. Z., Fei, Z., & Giovannoni, J. J. (2011). High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harbor Protocols*. https://doi.org/10.1101/pdb.prot5652

Zhou, Q., Su, X., Jing, G., Chen, S., & Ning, K. (2018). RNA-QC-chain: Comprehensive and fast quality control for RNA-Seq data. *BMC Genomics*. https://doi.org/10.1186/s12864-018-4503-6