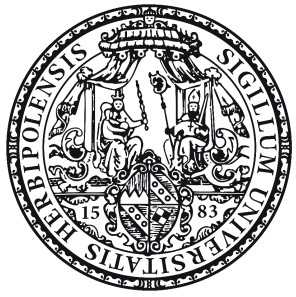


Aus dem Institut für Virologie und Immunobiologie
der Universität Würzburg
Vorstand: Professor Dr. med. Lars Dölken

Analyse der RNA-Landschaft und Chromatinorganisation in lytischer HSV-1 Infektion und Stress

Inaugural - Dissertation

zur Erlangung der Doktorwürde der
Medizinischen Fakultät
der
Julius-Maximilians-Universität Würzburg

vorgelegt von

Tobias Eberhard Haas
geboren am 20.11.1992 in Wiesbaden

Würzburg, August 2021

Erstprüfer:	Herr Prof. Dr. med. Lars Dölken
Zweitprüfer:	Frau Prof. Dr. Caroline Friedel

Referent: _____

Korreferent: _____

Dekan: Herr Prof. Dr. Matthias Frosch

Tag der mündlichen Prüfung: _____

Der Promovend ist Arzt

Inhaltsverzeichnis

I. Abkürzungsverzeichnis

4sU „4-thiouridine“ . . . 3

4sU-seq „4-thiouridine sequencing“ . . . 3

agg. Peaks „aggregierte Peaks“ . . . 38

ATAC-seq „Assay for Transposase Accessible Chromatin with high-throughput sequencing“ . . . 3

BAM „Binary Sequence Alignment Map“ . . . 13

Bash „Bourne-again shell“ . . . 13

BED „Browser Extensible Data“ . . . 17

bp Basenpaar(e) . . . 17

CTD C-terminale Domäne . . . 2

DNA „deoxyribonucleic acid“ . . . 1

DNAse-seq „DNase I hypersensitive sites sequencing“ . . . 14

dOCR „downstream Open Chromatin Region(s)“ . . . 9

DoG „Downstream of Genes“ . . . 7

DoTT „Disruption of Transcription Termination“ . . . 6

dsDNA „doppelsträngige DNA“ . . . 6

ENCODE „Encyclopedia of DNA Elements“ . . . 15

FACT „Facilitates Chromatin Transcription“ . . . 71

FDR „False Discovery Rate“ . . . 17

F-Seq „Feature Density Estimator for High-Throughput Sequence Tags“ . . . 16

GEO „Gene Expression Omnibus“ . . . 12

GRC „Genome Reference Consortium“ . . . 20

GRCh38 „Genome Reference Consortium Human Build 37“ . . . 13

GTF „General Transfer Format“ . . . 20

1. Einleitung

1.1. Physiologische Beeinflussung der RNA-Landschaft und Chromatinorganisation

Um auf die Analyse der RNA Landschaft und Chromatinorganisation in lytischer HSV-1 Infektion und Stress vorzubereiten, wird zunächst auf physiologische Prinzipien ebendieser eingegangen.

Die rund 2 Meter „deoxyribonucleic acid“ (DNA) einer menschlichen Zelle werden um Histone zu Nukleosomen verdrillt, welche weiterhin zu Chromatin komprimiert werden. Dadurch kann die DNA im Zellkern Platz finden. [1]
Dies hat erheblichen Einfluss auf die Aktivität bzw. Inaktivität verschiedener Gen-Regionen: Durch die Kompression von Promotoren, Enhancern oder sonstiger regulatorischer Elemente sind diese räumlich nicht für den Transkriptionsapparat zugänglich [2].
Der Transkriptionsapparat selbst hat weitreichende Folgen für die Chromatinorganisation, weshalb er im Weiteren detailliert besprochen wird. Splicing, „ribonucleic acid“ (RNA)-Editing und Translation werden im Rahmen dieser Arbeit nicht besprochen.

1.1.1. Transkription

Transkription ist die Synthese von RNA anhand einer DNA-Matrize. Diese kann entlang des Referenzstrangs ablaufen oder auf dem komplementären Strang in entgegengesetzter Richtung (Antisense-Transkription).

Initialisierung und Elongation

Unter physiologischen Bedingungen katalysiert die RNA-Polymerase II (Pol-II) im Zellkern die Bildung von prä-mRNA aus proteinkodierenden Genen sowie die Bildung von verschiedenen kleinen RNAs:

Nach Bildung eines Initiationskomplexes, welcher von diversen Promotoren abhängig ist, kommt es zu einer Phosphorylierung der C-terminale Domäne (CTD) der Pol-II an Ser 5 des Heptapeptides. Daraufhin werden Initiationsfaktoren an der Pol-II von Elongationsfaktoren abgelöst und die Elongation beginnt.
Damit die Elongation voranschreiten kann, muss die DNA vorübergehend von den Histonen abgelöst werden. Das Chromatin liegt also während der Transkription vorübergehend offen vor. [3]

Polyadenylierung und Termination

Trifft die Elongation auf die spezifische Polyadenylierungssequenz, bestehend aus den sechs Basen AAU-AAA, kommt es zur Polyadenylierung. Bei dieser werden noch weitere 50 bis mehr als 200 Adenylylreste an das 3'-Ende der entstehenden Prä-mRNA angeheftet. Dies ist der sogenannte Poly(A)-Schwanz, welcher die entstehende RNA vor dem Abbau durch Nukleasen schützt und damit die Halbwertszeit beeinflusst. Im Einzelnen löst dabei eine Änderung des Phosphorylierungsmusters der CTD der Pol-II folgende Vorgänge aus:
Die Prä-mRNA wird 10-30 Nukleotide downstream der Polyadenylierungssequenz durch den Endonukleasekomplex aus „cleavage stimulation factor“ (CstF) und „cleavage and polyadenylation specificity factor“ (CPSF) gespalten. CstF löst sich ab. Die Polyadenylatpolymerase (PAP) bindet am 3'-Ende und polyadenyliert, bis sich das Poly(A)-Bindeprotein PABP II (Poly(A) tail binding protein II) anlagert und PAP sowie CPSF freigesetzt werden. [3]
Die molekularen Einzelheiten der Termination sind nicht vollständig verstanden. Daher gibt es das allosterische Modell und das Torpedomodell, die sich jedoch nicht gegenseitig ausschließen (Hybridmodell) [4].

Das allosterische Modell besagt, dass es nach der Transkription des Poly(A)-Schwanzes zu einer Konformationsänderung der Pol-II durch den Verlust eines oder mehrerer Anti-Terminations-Faktoren kommt. Daraufhin dissoziiert die Pol-II von der DNA. [3, 5]

Das Torpedomodell besagt, dass die Pol-II durch die Anlagerung des Polyadenylierungskomplexes und dann durch die Formung einer R-loop Struktur ausgebremst wird. Während die mRNA durch die oben beschriebenen Prozesse nicht mehr an die Pol-II gebunden ist, fährt die Pol-II mit der Transkription der DNA-Matrize fort. Dieses Transkript (also nicht die mRNA) wird vom neu entstandenen 5'-Ende aus

durch „5'-3' Exoribonuclease 2“ (XRN2) abgebaut.
Da XRN2 schneller abbaut als die gebremste Pol-II transkribiert, trifft XRN2 wie ein molekularer Torpedo auf die Pol-II, wodurch diese von der DNA-Matrize freigesetzt wird und zur erneuten Initialisierung bereitsteht. [3, 5]

1.1.2. Histonmodifikationen

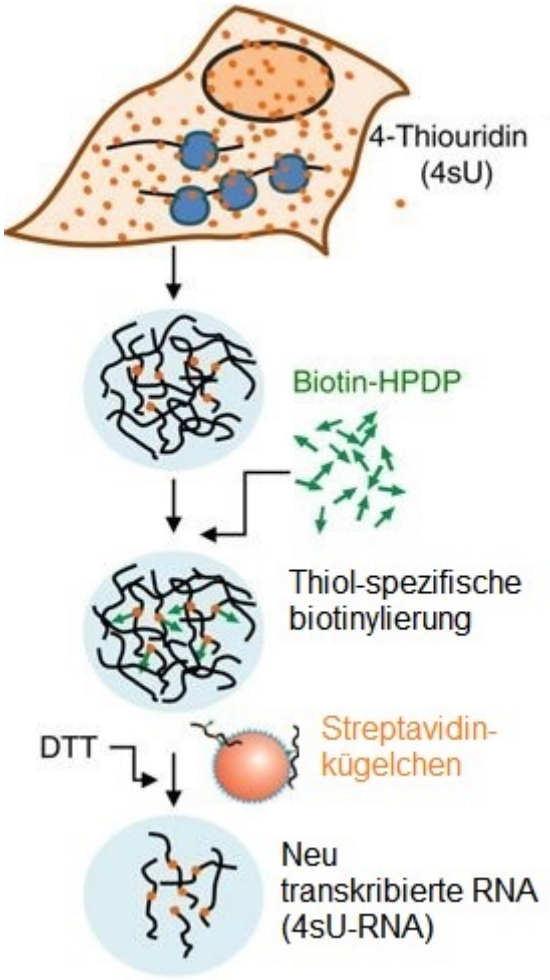
Die Histone bestimmen anhand ihrer zahlreichen Modifikationen den Offenheitsgrad des Chromatins. Dazu gehören die typischen Histonmarker ungestresster Zellen an transkribierten Genen H3K36me3 und H3K79me2 und an Enhancern H3K4me1 und H3K27ac sowie Histon-Acetylierung. [6]
Histonmodifikationen werden im Rahmen dieser Arbeit nicht weiter untersucht.

1.2. Untersuchung der RNA-Landschaft und Chromatinorganisation

Im Rahmen dieser Arbeit wurde die RNA-Landschaft mit „4-thiouridine sequencing“ (4sU-seq) [7–10] und die Chromatinorganisation mit „Assay for Transposase Accessible Chromatin with high-throughput sequencing“ (ATAC-seq) [11] untersucht.

1.2.1. „4sU-tagging“ und „4sU-seq“

4sU-tagging ist ein Verfahren, bei dem Zellen mittels „4-thiouridine“ (4sU) pulsmarkiert werden (s. Abb. 1.1). Es folgt eine Biotinylierung der neu transkribierten RNA am Schwefelatom mit 4sU. Nach Isolation der gesamten RNA aus Zellen kann das in neu synthetisierte RNA Moleküle eingebaute 4sU thiol-spezifisch biotinyliert werden. Dies erlaubt die Aufreinigung der im zeitlichen Markierungsfenster von der Zelle neu gebildeten RNA Moleküle über mit Streptavidin umhüllter magnetischer Partikel. Die Kombination von 4sU-tagging und Sequenzierung wird 4sU-seq genannt. Durch 4sU-seq zu unterschiedlichen Zeitpunkten nach einer Infektion lassen sich damit Veränderungen der RNA-Landschaft im zeitlichen Verlauf einer Infektion beschreiben. [12]



Quelle: Rutkowski 2015 [12]; Rechteinhaber: Springer Nature; Lizenz: CC BY 4.0

Abbildung 1.1.: Prinzip des „4sU-tagging"- schematisch
In einem ersten Schritt werden Fibroblasten mit 4sU pulsmarkiert. Anschließend erfolgt die Isolation der gesamten RNA. RNA, welche nach der Pulsmarkierung synthetisiert wird, ist am Schwefelatom des 4sU biotinyliert (4sU-RNA). Durch magnetische Streptavidin-beschichtete Plättchen wird die 4sU-RNA von der Gesamt-RNA getrennt und durch Zugabe des Reduktionsmittels Dithiothreitol (DTT) von den Plättchen zurückgewonnen. [12]
Damit kann neu transkribierte RNA von bereits vor der Markierung vorhandener RNA unterschieden werden.

1.2.2. ATAC-seq

„Assay for Transposase Accessible Chromatin with high-throughput sequencing" (ATAC-seq) ist eine Methode zur genomweiten Kartierung von offenem Chromatin. Nach Lyse einer etwa 50.000 Zellen umfassenden Probe verwendet ATAC-seq eine hyperaktive Tn5 Transposase, die offenes Chromatin gleichzeitig schneidet und mit

Adaptern für „high-throughput sequencing“ ligiert. Da die Tn5 Transposase Heterochromatin nicht effektiv schneidet, wird Heterochromatin nicht mit Adaptern ligiert und damit bei anschließender „Polymerase Chain Reaction“ (PCR) nicht amplifiziert. Daher entsprechen die durch ATAC-seq bestimmten Nukleotidabfolgen Regionen mit erhöhter Zugänglichkeit bzw. Offenheit des Chromatins (s. Abb. 1.2). [11]

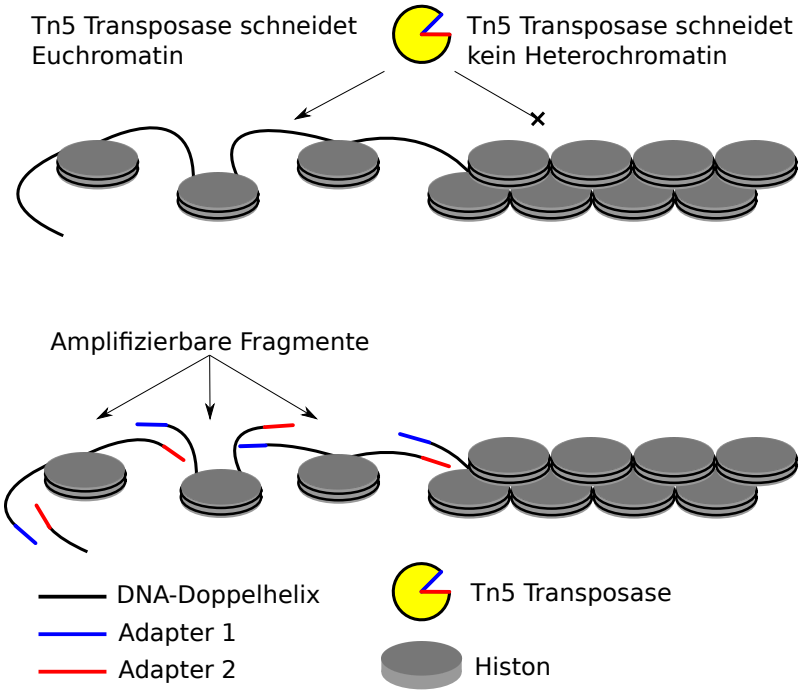


Abbildung 1.2.: Prinzip von ATAC-seq - schematisch
Die Tn5 Transposase schneidet Euchromatin, aber kein Heterochromatin. Beim Schneiden wird die DNA mit Adaptern ligiert, die nach weiterer Aufbereitung als Ausgangspunkt für eine PCR dienen. In der PCR werden „barcoded-primers“ verwendet, sodass anhand der anschließenden Sequenzierung auf Regionen mit erhöhter Zugänglichkeit bzw. Offenheit des Chromatins zurückgeschlossen werden kann. [11]

1.3. Pathophysiologische Beeinflussung der RNA-Landschaft und Chromatinorganisation

Die in dieser Arbeit untersuchten Einflussfaktoren der RNA-Landschaft und Chromatinorganisation sind die lytische Infektion mit Herpes simplex Virus 1 (HSV-1) und Zellstress in Form von Hitze und Salz.

1.3.1. HSV-1

Herpesviren sind große, umhüllte „doppelsträngige DNA“ (dsDNA) Viren, die global vorkommen und deren über 130 Spezies hauptsächlich Wirbeltiere, aber auch Wirbellose, infizieren können [13]. Darunter finden sich die 8 humanpathogenen Herpesviren, HHV 1-8. Das HSV-1 entspricht dem Humanen Herpesvirus 1 (HHV-1); beide Namen werden synonym für das gleiche Virus verwendet. Das HSV-1 hat ein 152 kb großes Genom [14], wobei Whisnant et al. (2020) 284 „open reading frames“ (ORFs) und 201 Transkripte identifizieren [15]. Es wird nicht nur in symptomatischen Intervallen oder sexuell übertragen, sondern auch durch Mutter-Kind-Kontakt im Kindesalter [16]. HSV-1 ist in Deutschland eine der Infektionskrankheiten mit der höchsten Seroprävalenz (78,4 %) [17].
Neben dem Herpes labialis oder genitalis kann HSV-1 auch zu schwerwiegenderen Komplikationen wie postherpetisches Erythema exsudativum multiforme, Herpeskeratokonjunktivitis und in seltenen Fällen auch HSV-Bronchopneumonie und HSV-Enzephalitis führen [18, S.424]. Zudem wurde unter den HSV-Infektionen eine erhöhte Suszeptibilität für HIV und Alzheimer beobachtet [19, 20].
HSV-1 ist durch Infiltration sensorischer Ganglien in der Lage, sich der Kontrolle des Immunsystems zu entziehen und damit lebenslang zu persistieren. In der latenten Phase der Infektion werden außer dem nicht kodierenden „Latency Associated Transcript“ (LAT) kaum virale Proteine synthetisiert [21].
Im Zuge der lytischen Infektion schaltet HSV-1 die zelluläre Genexpression praktisch vollständig aus (Englisch: Host shut-off). Hierbei spielen die beiden viralen Proteine „virion host shut-off“ (vhs) und „Infected Cell Polypeptide 27“ (ICP27) eine zentrale Rolle. Neben zahlreichen weiteren Funktionen im RNA Metabolismus stört ICP27 hierbei die Termination von Pol-II am Ende von Genen. Diese im Englischen als „Disruption of Transcription Termination“ (DoTT) bezeichnete Vorgang, führt zur Transkription weit (z.T. >100.000 Basen) über die Genenden hinaus in flussabwärts gelegene Genregionen [22]. Ein zentrales Ziel dieser Arbeit bestand darin, das Ausmaß der vom DoTT verursachten Transkription von Basenpaaren downstream der „transcription terminator site“ zu untersuchen.

Der „host shut-off“ ist dadurch gekennzeichnet, dass nur geringe Mengen an Host-RNA, hauptsächlich Haushaltsgene, translatiert werden. Host-RNA ist die RNA, welche dem Genom des infizierten Wirts zugehörig ist. Die viralen Gene sind hiervon jedoch nicht betroffen, sodass in der lytischen Phase 80 % der translatierten Genprodukte viralen Ursprungs sind, obwohl nur rund 20 % der Gesamt-RNA vira-

len Ursprungs sind [12].
Effektiv wird die Syntheseleistung der Zelle damit auf Reproduktion von weiteren Viren umprogrammiert. Durch die Unterdrückung der Apoptose wird die Freisetzung neu synthetisierten Viren durch Egress und Zytolyse gesichert.

1.3.2. Zellstress

Im Rahmen dieser Arbeit wird Zellstress in Form von Salz (Milieu mit 80mM KCl für 1 h) und Hitze (44 °C) untersucht. Zellstress beeinflusst die Chromatinorganisation über verschiedene Mechanismen. Dieser Einfluss wird nachfolgend erklärt.

1.3.3. „Disruption of Transcription Termination"(DoTT) und „Downstream of Genes"(DoG) Transkription

Wie in Kap. 1.3.1 beschrieben, wird durch die lytische HSV-1 Infektion ein „host shut-off" hervorgerufen. Zellstress in Form von Salz, Hitze oder Oxidation haben ähnliche Auswirkungen auf die Zellphysiologie [22, 23].
Rutkowski et al. (2015) konnten für HSV-1 Infektion folgendes zeigen: 8 h post interventionem (entspricht im Kontext mit Viren post infectionem) (p.i.) wurde eine transkriptionelle Herabregulation von 70 % der Gene und scheinbare transkriptionelle Hochregulation von 5,8 % der Gene gemessen, wobei eine Hochregulation nur in 0,34 % der Gene auch mit einer erhöhten Translation korreliert hat. Dagegen wurden 77 % der hochregulierten Gene 8 h p.i. nicht translatiert. Außerdem scheinen 44.5 % der long intergenic non-coding RNAs (lincRNAs) hochreguliert zu sein. [12]

Die computergestützte Analyse der Reads zeigt, dass dies auf DoTT zurückzuführen ist. Hierbei wird anstatt der in Kap. 1.1.1 beschriebenen Polyadenylierung die neu entstehende Prä-mRNA von der Pol-II über das Polyadenylierungssignal hinaus entsprechend der DNA-Matrize verlängert (Englisch: Read-through). Es wird also der Bereich transkribiert, der downstream des Bereiches liegt, der ein Gen enthält, weshalb man auch von „Downstream of Genes" (DoG) Transkription spricht. Zur besseren Unterscheidung wird hier bei lytischer HSV-1 Infektion von DoTT und bei Zellstress von DoG Transkription gesprochen.
Die Pol-II kann die Transkription über 100.000 Basenpaare hinaus und damit auch in andere Gene hinein („Read-in") oder sogar durch mehrere Gene hindurch fortsetzen. Dies erklärt die scheinbare Hochregulation mancher Gene und lincRNAs. Es wird diskutiert, inwieweit die dabei entstehende mRNA zum „host shut-off" beiträgt,

indem sie nicht richtig prozessiert, exportiert und translatiert werden kann. [12]

1.3.4. „Open Chromatin Regions (OCRs)“

Im Gegensatz zu Zellstress in Form eines Salz- oder Hitzeschocks, führt die lytische HSV-1 Infektion darüber hinaus vor allem bei stark exprimierten Genen zu offener vorliegendem Chromatin im Bereich des Read-throughs („Open Chromatin Regions“ (OCRs)).
OCRs kommen bei vereinzelnden Genen auch physiologisch vor. Bei HSV-1 Infektion sind sie ab 4 h p.i. stark positiv mit der Transkription von Genen und mit Read-through korreliert, wobei dies nicht DoTT verursacht, sondern von DoTT abhängig ist. Es gibt Hinweise, dass OCRs die unter HSV-1 beobachtete Antisense-Transkription begünstigen und eine Rolle in der Unterdrückung der Apoptose spielen. [22, 24]
Vilborg et al. hingegen berichten, dass OCRs auch im Zellstress vorkommen und ziehen in Betracht, dass das Offenhalten des Chromatins ein Zweck von DoG Transkription sein könnte [23].

1.4. Begrifflichkeiten der Informatik

1.4.1. Reads, „mapping“ und „aligned reads“

Reads stellen Basenpaarfolgen dar, die aus einer DNA- (oder RNA-)Sequenzierung abgeleitet wurden. Von einem „aligned read“ spricht man, wenn ein Read einer Position auf einem Referenzgenom zugeordnet wurde. Dieser Vorgang wird als „mapping“ bezeichnet.
Aus der Anzahl der „aligned reads“ können damit von der Art der Sequenzierung abhängige Rückschlüsse auf die Ausprägung eines Merkmals getroffen werden: Bei ATAC-seq sind Reads ein Maß für die Offenheit des Chromatins. Bei 4sU-seq sind Reads ein Maß für die Menge an transkriptioneller Aktivität an einer korrespondierenden DNA-Matrize.
Es ist zu beachten, dass die Anzahl an Reads aufgrund der stochastischen Natur der zugrunde liegenden Experimente einer statistischen Unsicherheit unterliegt. Aus diesem Grund werden von Reads abgeleitete Größen wie $\log_2$-fold-changes streng genommen geschätzt anstatt berechnet.

1.4.2. Peaks, Peak-Caller und Score

Ein Peak im Sinne dieser Arbeit ist ein Intervall, welches den Beginn und das Ende eines mit „aligned reads“ angereicherten Bereiches kennzeichnet. Peak-Caller sind Programme zur Findung von Peaks. Die Peak-Caller ordnen auf unterschiedliche Art und Weise jedem Peak einen Score zu. Dieser Score ist ein Maß für die Passgenauigkeit eines Peaks zu den Daten. Das heißt, dass einem Peak ein höherer Score zugeteilt wird, wenn er sich stärker von der Umgebung abhebt. Dies ist mit einer geringeren Rate an falsch-positiv annotierten Peaks assoziiert. Auf die Funktion der unterschiedlichen Peak-Caller wird in Kap. 2.2.1 eingegangen.
Durch die Verwendung von Peak-Callern auf ATAC-seq-Daten wird in dieser Arbeit genomweit auf die Positionen von OCRs geschlussfolgert.

1.4.3. Parameter

In der Informatik ist ein Parameter eine Variable, durch dessen Angabe sich Einstellungen an Programmen vornehmen lassen. Bei einer höher entwickelten Benutzeroberfläche (UI) entspricht dies einer Schaltfläche, einem Regler, oder dem Abhaken einer Checkbox. Parameter können jedoch auch durch reine Texteingabe gesetzt werden.
Damit ermöglichen Parameter die Feinadjustierung von Peak-Callern.

1.4.4. „Downstream Open Chromatin Regions“

Downstream Open Chromatin Regions (dOCR) ist eine von Hennig et al. (2018) eingeführte Hilfsgröße, die sich folgendermaßen berechnet [22]:
Die „downstream Open Chromatin Region(s)“ (dOCR) eines Gens ist die Summe der Längen von OCRs in Basenpaaren über einen gewissen Downstreambereich. Hierzu werden zuerst die Längen in Basenpaaren aller „Open Chromatin Regions“ (OCRs) im 10 kbp Downstreambereich summiert. Dann wird vom letzten mitgezählten OCR die Länge des folgenden OCR hinzu summiert, solange sich der folgende OCR nicht mehr als 5 kbp vom zuletzt mitgezählten OCR entfernt befindet. Das wird wiederholt, bis 5 kbp lang kein weiterer OCR zu verzeichnen ist [22]. Dies geschieht ohne Rücksichtnahme auf Inhalt (z. B. andere Gene) des Downstreams.

1.4.5. „Aggregierte Peak-Länge“

Dieser Parameter wird analog zu dOCR berechnet mit dem Unterschied, dass die Basenpaare in Peaks in einem definierten Abstand zum 3’ Ende des Gens aggregiert werden, ohne diesen Abstand in einem weiteren Schritt iterativ zu verlängern.

1.4.6. Genom Browser

Genom Browser sind Programme zur Visualisierung der „aligned reads“. Für die Darstellungen dieser Arbeit wurde „Genomic data integration platform“ (Gedi) verwendet [25].

1.4.7. Pseudocounts

Mit ATAC-seq und 4sU-seq lassen sich für jedes Basenpaar in einem untersuchten Genombereich eine Häufigkeit an Reads ableiten. Durch das ins Verhältnis setzen dieser Häufigkeiten in Form von Quotienten lassen sich unterschiedliche Genombereiche vergleichen, wie es bei der Berechnung bzw. Schätzung (s. 1.4.1) des prozentualen Read-through der Fall ist. Weiterhin kann ein Genombereich mit sich selbst zu einem anderen Zeitpunkt verglichen werden, wie es bei der Berechnung des $\log_2$-fold-change der Reads pre und post infektionem der Fall ist.
Da die Häufigkeiten der Reads einer statistischen Unsicherheit unterliegen und Null (oder annähernd Null) betragen können, kann eine Quotientenbildung keine (oder unsinnig hohe) Ergebnisse ergeben. Ein sinnvolles Dividieren ist erst möglich, wenn zu jeder Häufigkeit eine positive Konstante addiert wurde. Diese Konstante nennt man Pseudocounts.
Setzt man einen relativ hohen Wert für den Pseudocount ein, so werden Verhältnisse in Richtung 1 verzerrt, da der Wert sowohl in den Divisor als auch in den Dividenden einfließt.
Setzt man einen relativ niedrigen Wert ein, so können niedrige Werte im Divisor die Verhältnisse nach oben hin verzerren.
Eine willkürliche Wahl des Pseudocount ist daher problematisch. Stattdessen kann mit dem R-Paket „Log fold change distribution tools for working with ratios of counts“ (LFC) ein Wert für den Pseudocount geschätzt werden, der anhand der Reads auf die Minimierung der oben genannten Verzerrungen optimiert ist. [26]

1.5. Ziel der Arbeit

Neue Errungenschaften im „Next Generation Sequencing“ (NGS) wie 4sU-tagging und ATAC-seq ermöglichen es, den unter anderem durch HSV-1 induzierten „host shut-off“ genauer zu untersuchen. Gleiches gilt für die Unterdrückung der Transkriptionsterminierung unter Hitze- und Salzstress. Daher wird sich in dieser Arbeit folgender Fragestellung genähert:

Welche Gemeinsamkeiten und Unterschiede gibt es zwischen Zellstress in Form von Hitze-, Salzstress und lytischer HSV-1 Infektion gemessen an ATAC-seq und 4sU-seq als Indikator für OCRs und DoTT im Downstreambereich von Genen?

Hierzu sollte zuerst der optimale Peak-Caller ermittelt werden, um sowohl in uninfizierten Zellen als auch post infektionem anhand der Reads aus ATAC-seq genomweit die Offenheit des Chromatins zu beschreiben.
Dann sollten mit diesem Peak-Caller Unterschiede zwischen pre und post infektionem in Bezug auf die Offenheit des Chromatins beschrieben werden. Dazu sollten erstens Peak-Längen aggregiert, zweitens dOCR berechnet und drittens Reads in verschiedenen Downstreambereichen direkt gezählt werden.
Diese drei Hilfsgrößen für das Ausmaß des Read-throughs sollten verglichen und bewertet werden. Außerdem sollten sie den Ergebnissen aus 4sU-seq als Maß für die RNA-Landschaft gegenübergestellt werden. Dadurch sollten sich neue Erkenntnisse über die Zusammenhänge zwischen Veränderungen in der Chromatinorganisation und RNA-Landschaft erlangen lassen.
Die daraus gewonnenen Erkenntnisse sollten auf Hitze und Salzstress übertragen werden, um Unterschiede und Gemeinsamkeiten zwischen lytischer HSV-1-Infektion und Zellstress herauszuarbeiten.

2. Material und Methoden

2.1. Rohdaten

2.1.1. Bereitstellung der Rohdaten

Die Rohdaten wurden über die Datenbank „Gene Expression Omnibus“ (GEO) erhalten.
Im Detail handelt es sich um die Datensätze GSE59717 für 4sU-seq nach HSV-1 Infektion, GSE100469 für 4sU-seq nach Salz- und Hitzestress, GSE100611 für ATAC-seq nach HSV-1 Infektion und GSE101731 für ATAC-seq nach Salz- und Hitzestress.
Hennig et al. (2018) beschreiben unter anderem die labortechnische Gewinnung dieser Daten. [22]

Virus und Zellen

Es wurde mit dem HSV-1 Stamm 17 gearbeitet. Bei den Wirtszellen handelt es sich um humane fetale Fibroblasten aus der Vorhaut (Human fetal foreskin fibroblasts (HFF)).

2.1.2. Read mapping

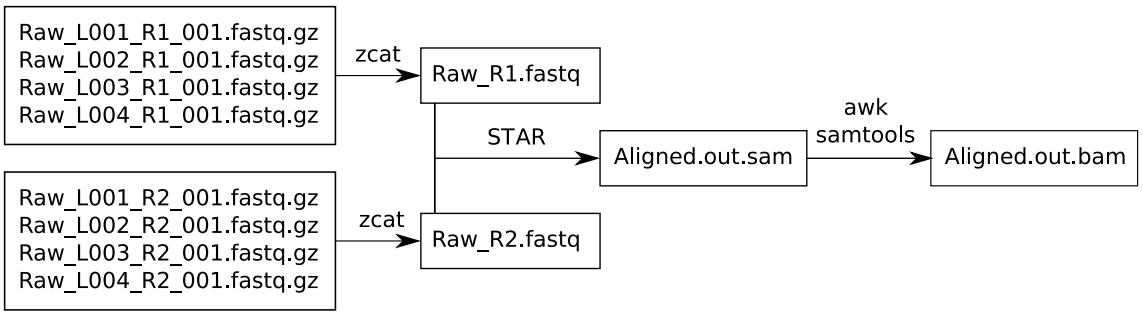


Abbildung 2.1.: Aufarbeitung der Rohdaten

Die Rohdaten wurden mit Hilfe von 4sU-seq (s. Kap 1.2.1) und ATAC-seq (s. Kap. 1.2.2), welche beide auf NGS aufbauen, gewonnen. Um die Rohdaten für andere Pro-

gramme wie etwa Peak-Caller lesbar zu machen, müssen diese zuerst konvertiert und gemappt werden:
Die Rohdaten liegen pro Versuchsbedingung als 8 Rohdatensätze im FASTQ Format vor. Dabei steht L001-L004 für die verschiedenen Bahnen einer „flow cell“. Der Inhalt dieser Dateien wurde mit dem „Bourne-again shell“ (Bash) Befehl zcat aneinandergereiht. Da es sich um Daten aus „Paired-End Sequencing“ (PES) handelt, wurde dieser Schritt zweimal durchgeführt (für R1 und R2). Bei PES werden DNA-Fragmente von beiden Enden aus sequenziert. Dies muss beim Mappen auf das Genom berücksichtigt werden.
Das Mapping der Reads wurde mit „Spliced Transcripts Alignment to a Reference“ (STAR) [27] auf das Referenzgenom „Genome Reference Consortium Human Build 37“ (GRCh38) durchgeführt und erzeugt eine „Sequence Alignment Map“ (SAM). Reads können RNA repräsentieren und RNA kann gespliced werden, was STAR berücksichtigen kann. Repräsentieren Reads DNA, wie es bei ATAC-seq der Fall ist, so ist das Berücksichtigen von Splicing-Events nicht sinnvoll. Deshalb wird STAR mit den Parametern „–scoreGap -1000“ und „–scoreGapNoncan -1000“ aufgerufen, wodurch das Berücksichtigen von Splicing unterdrückt wird. Danach wird die SAM mit SAMtools in das „Binary Sequence Alignment Map“ (BAM) Format konvertiert, sortiert und indexiert (siehe Abb. 2.1).

2.1.3. Aufstellung der Rohdaten

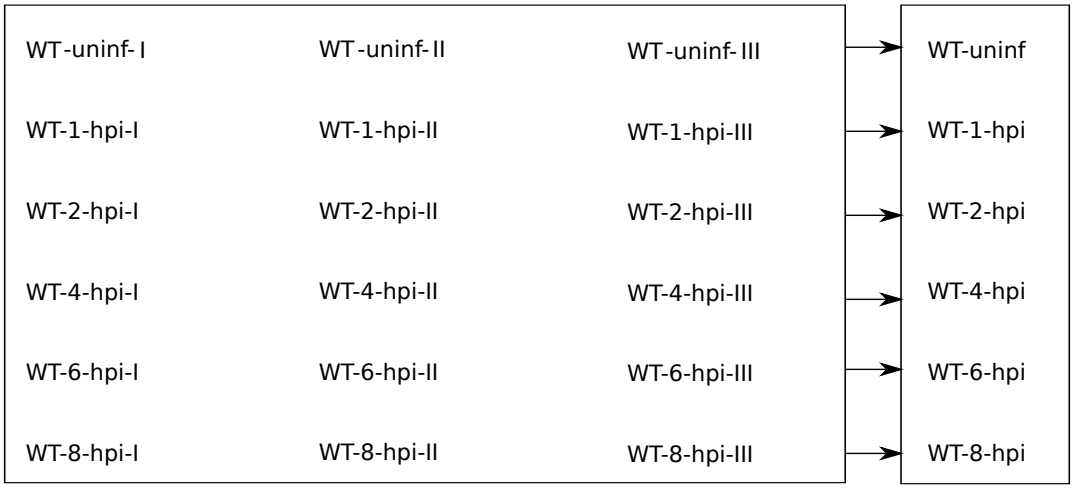


Abbildung 2.2.: Aufstellung der Rohdaten.
Die 24 Versuchsbedingungen zu HSV-1 setzen sich aus 3 Replikaten mit jeweils 6 verschiedenen Zeitpunkten p.i. zusammen, die zusätzlich über jeden Zeitpunkt gepoolt werden.

Das oben beschriebene Verfahren wurde für die 18 Versuchsbedingungen der HSV-1-Infektion durchgeführt (siehe Abb. 2.2. Dabei steht I-III für die jeweiligen Wiederholungen des Experiments (Replikate), uninfziert (uninf.) für die Kontrollbedingung bzw. uninfzierte Zellen, WT für den Wild-Typ und n h p.i. für n Stunden nach HSV-1 Infektion). Weiterhin wurden die Reads der Replikate I-III zu den jeweiligen Zeitpunkten gepoolt. So enthält z. B. der Datensatz WT-8-hpi die Reads aus WT-8-hpi-I, WT-8-hpi-II und WT-8-hpi-III.
Die Aufbereitung und Aufstellung der Daten für Salz- und Hitzestress entsprechen im Wesentlichen dem Vorgehen nach HSV-1 Infektion mit folgenden Unterschieden: Es gibt vier anstatt von drei Replikaten für die Kontrollbedingung uninf. und 2 Replikate für die Postinterventionsbedingungen. Diese Replikate wiederum erstecken sich nicht bis 8 h post infektionem, sondern bis 2 h post interventionem (p.i.).
Wenn nicht anders gekennzeichnet, wurden die Abbildungen im Folgenden auf Grundlage der gepoolten Reads für die einzelnen Replikate erstellt. Das heißt, dass z. B. für die Daten aus uninfizierten Zellreihen die Reads der Replikate I-III addiert wurden. Anschließend wurde mit den gepoolten Daten weitergearbeitet.

2.1.4. Subset von 3684 Genen aus GRCh38

Ein Teil der Auswertung wird nur anhand eines Subsets von 3684 Genen durchgeführt. Dies hat den Hintergrund, dass die Wechselwirkungen von Genen in Genclustern eine Auswertung erschweren. Daher werden in diesem Subset keine Gene berücksichtigt, die zu keinem Zeitpunkt exprimiert wurden. Dies betrifft rund die Hälfte der Gene. Weiterhin wurden Gene ausgeschlossen, die sich in weniger als 5 kbp Upstream oder Downstream zu einem anderen Gen auf dem selben Strang befinden. Weiterhin wurden Gene mit mehr als 10 % Read-in ausgeschlossen. Dieses Vorgehen ist analog zu dem von Hennig et al. (2018) [22]. Das Subset ist als GTF-Datei unter der DOI 10.5281/zenodo.4710625 veröffentlicht.

2.2. Auswahl des Peak-Caller

Die etablierten Peak-Caller sind nicht für ATAC-seq Daten optimiert, sondern für „DNase I hypersensitive sites sequencing“ (DNAse-seq) und/oder ChIP-seq. Trotzdem unterscheiden sich ihre Ergebnisse bereits bei DNAse-seq erheblich [28]. Daher wurde in einem ersten Schritt eruiert, welcher Peak-Caller die durch Rutkowski et al. (2015) beschriebenen Veränderungen nach HSV-1 Infektion am besten wiedergibt bzw. welche Parameter eine Differenzierung zwischen uninfizierten und infizierten

Replikaten erlauben [12]. Es werden die folgenden vier Annahmen getroffen:

Erstens korreliert die Transkription von Genen mit offenem Chromatin im Downstreambereich dieser Gene. Durch DoTT werden 8 h p.i. bereichsweise 100 kbp Downstream transkribiert [12], die in 0 h p.i. nicht transkribiert werden. Daher wird hier erwartet, dass auch die Länge der Peaks 8 h p.i. im Vergleich zu 0 h p.i. zunimmt.
Zweitens erstreckt sich 8 h p.i. ein langer Peak über eine Sequenz, in der bei uninfizierten Zellen (aufgrund von Rauschen) mehrere kurze Peaks annotiert wurden, so nimmt die Summe der Peaks ab. Zudem werden aufgrund des „host shut-off“ Gene herunterreguliert. Diese reduzierte Transkription lässt auch eine Reduzierung des offenen Chromatins und damit der absoluten Anzahl der Peaks erwarten.
Drittens haben Rutkowski et al. (2015) gezeigt, dass die Größe von dOCR p.i. zunimmt [12]. Betrachtet man die Berechnung von dOCR (s. Kap. 1.4.4) ist offensichtlich, dass lange und zahlreiche Peaks die Größe von dOCR begünstigen.
Viertens sollten die Zuordnungen der Peaks plausibel sein. Das heißt, die Peaks sollten bei stichprobenartiger Betrachtung im Genom-Viewer mit Bereichen zusammenfallen, die eine erhöhte Anzahl an Reads aufweisen.

Zusammenfassend wird ein Peak-Caller gesucht, der 8 h p.i. eine reduzierte Anzahl von Peaks findet, die verlängert und plausibel sind und zur Berechnung von größeren dOCR führen.

2.2.1. Peak-Caller Kandidaten

ENCODE ATAC-seq Pipeline

„Assay for Transposase Accessible Chromatin with high-throughput sequencing“ (ATAC-seq) Pipeline [29, 30] wurde bzw. wird im Rahmen des Forschungsprojekts „Encyclopedia of DNA Elements“ (ENCODE) an der Stanford University entwickelt. ATAC-seq Pipeline ist von den hier genannten Peak-Callern der jüngst entwickelteste und lag zum Zeitpunkt der Datenverarbeitung noch als Prototyp vor. Seit dem 2. November 2020 liegt die Version 1.9.1 vor, wobei die finale Revision und volle Implementierung in „ENCODE Uniform Processing Pipelines series“ zu diesem Zeitpunkt noch ausstand. ATAC-seq Pipeline verbindet bereits bestehende Programme. So kann es mit u. a. Bowtie 2 Reads „alignen“ und benutzt dann „Model-based Analysis for Chromatin Immunoprecipitation DNA-Sequencing 2“ (MACS2), um Peaks zu annotieren. Dabei werden die Parameter weniger restriktiv eingestellt, wodurch

viele Peaks gefunden werden. Hinterher wird „Irreproducible Discovery Rate“ (IDR) genutzt, um einen sinnvollen Grenzwert für die Signifikanz der Peaks zu finden und diese dementsprechend zu filtern.

ATAC-seq Pipeline wurde mit folgenden „flags“ aufgerufen:

-chrz	Referenzdatei für die Chromosomengrößen
-enable_idr	IDR wird verwendet
-species hg38	Genome Reference Consortium Human Build 38 wird als Referenzgenom verwendet

ATAC-seq Pipeline berücksichtigt Paired-End-Daten standardmäßig.

F-Seq

„Feature Density Estimator for High-Throughput Sequence Tags“ (F-Seq) [31, 32] wurde im Terry Furey Lab der „University of North Carolina at Chapel Hill“ entwickelt.
Boyle et al. (2008) beschreiben die Funktion von F-Seq folgendermaßen [31]: F-Seq nutzt eine Kerndichteschätzung, um Regionen mit einer hohen Read-Dichte zu identifizieren. Dabei wird daran, wie gut sich eine Normalverteilung an die Reads anmodellieren lässt, entschieden, an welchen Stellen Peaks annotiert werden. Im Detail wird die Wahrscheinlichkeitsdichtefunktion mit der Kerndichteschätzung

$$\hat{p}(x) = \frac{1}{nb} \sum_{i=1}^{n} K(\frac{x - x_i}{b}) \tag{2.1}$$

berechnet, wobei n Reads auf einem Chromosom der Länge L verteilt sind [31]. Der Bandbreitenparameter b bestimmt die Glätte der Dichteschätzung und *K()* ist die „Gaussian kernel function“ mit dem Median 0 und der Varianz 1. Der Benutzer kann b frei wählen, wobei 600 der Default-Wert ist und größere Werte eine höhere Glätte bedeuten. Es würde die Rechenleistung aktuell handelsüblicher Computer übersteigen, die Dichtefunktion gleichzeitig über jedem Punkt eines Chromosoms zu berechnen. Daher wird eine Standard Fenstergröße w als Funktion von b und dem „Gaussian kernel“ kalkuliert, sodass gilt [31]:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{w}{b})^2} > min(floatingpoint) \tag{2.2}$$

Außerdem wird ein Grenzwert für die Signifikanz der Peaks berechnet. [31] Die-

ser Grenzwert wird in der Tabelle der Peaks als „Score“ bezeichnet und man kann nach ihm filtern, um die Anzahl der falsch-positiv annotierten Peaks (hoher Score) oder die Anzahl der falsch-negativ annotierten Peaks (niedriger Score) zu reduzieren.

F-Seq wurde in der Entwicklung mit einem Datensatz mit weniger als 9 Millionen Reads getestet [32]. Bei größeren Datenmengen können über 12 GB „Random-Access Memory“ (RAM) zur Schätzung der „estimated fragment size“ beansprucht werden. Dies entfällt bei der Bearbeitung von Paired-End-Daten, da hier die Größe der sequenzierten Fragmente bereits definiert ist.
F-Seq benötigt Dateien im „Browser Extensible Data“ (BED) Format als Input. Diese benötigen aufgrund fehlender Komprimierung erheblich mehr Speicherplatz als Dateien im BAM Format. Damit ist F-Seq für große Datenmengen wenig geeignet. Zudem benötigt man zusätzliche Software wie den bffBuilder und den GEM Mapper. Die minimale Peaklänge wird in der weiteren Auswertung auf 10 Basenpaar(e) (bp) begrenzt und entspricht damit der von Hotspot.

F-Seq wurde mit folgenden „flags“ aufgerufen:

-b Dateipfad	Pfad zur Hintergrunddatei für „mapability“
-f 0	Für Paired-End-Daten
-of bed	Ausgabe von Daten im BED-Format

Hotspot

Hotspot [33, 34] wurde im Rahmen des Forschungsprojekts ENCODE an der „University of Washington“ bis 2014 entwickelt.
Hotspot entscheidet daran, wie gut sich eine Binominalverteilung an die Reads anmodellieren lässt, an welchen Stellen Peaks annotiert werden. Weiterhin werden genomweite Unterschiede in der Readdichte anhand eines „local background model“ berücksichtigt. Hotspot errechnet einen „False Discovery Rate“ (FDR) Wert, nach dem die Peaks gefiltert werden, um weniger falsch-positive Peaks zu erhalten. Hotspot ist mit bis zu drei Tagen Laufzeit der langsamste getestete Peak-Caller. Die Optionen werden nicht beim Aufrufen des Befehls im Terminal eingegeben, sondern müssen davor zuerst in einer Konfigurationsdatei gespeichert werden.

Hotspot wurde standardmäßig aufgerufen.

MACS2

„Model-based Analysis for Chromatin Immunoprecipitation DNA-Sequencing 2“ (MACS2) [35, 36] wurde in Harvard im Xiaole Shirley Liu's Lab entwickelt und ist der etablierteste Peak-Caller. MACS2 identifiziert Peaks anhand einer Poisson-Verteilung, wobei sich überschneidende Peaks zusammengefügt werden. MACS2 kann mit einer BAM als einzigen Input aufgerufen werden, ohne dass andere Dateien als Input angegeben werden müssen. Damit ist MACS2 der am einfachsten zu bedienende Peak-Caller. Mit der FDR Funktion kann eine „controll.bam“ angegeben werden. Dies wäre z. B. der DNA-Input ohne „Chromatin Immunoprecipitation“, wenn man „Chromatin Immunoprecipitation DNA-Sequencing“ untersucht. Die Reads der „controll.bam“ werden in die Score-Berechnung mit einbezogen, um die Anzahl an falsch-positiv annotierten Peaks zu reduzieren. Vom Gebrauch dieser Funktion wurde in dieser Arbeit jedoch abgesehen, um die Kontroll- und Interventionsbedingungen isoliert voneinander betrachten zu können.

MACS2 wurde mit folgenden „flags“ aufgerufen:

–broad	„broad peak calling“
–broad-cutoff 0.1	„q-value“ cutoff Wert für „broad peaks“
-f BAMPE	Paired-End wird berücksichtigt
-g hs	Es wird das vorkompilierte humane Genom genutzt

SPP

„ChIP-seq processing pipeline“ (SPP) [37] wurde in Harvard im Park Lab entwickelt. Im Gegensatz zu den anderen Peak-Callern ist bei SPP die Verwendung einer „controll.bam“ (s. MACS2) obligat. Vom Gebrauch von SPP wurde in dieser Arbeit abgesehen, um die Kontroll- und Interventionsbedingungen isoliert voneinander betrachten zu können.

ZINBA

„Zero-Inflated Negative Binomial Algorithm“ (ZINBA) [38] ist ein Gemeinschaftsprojekt des „Department of Biostatistics“ und des Lieb Lab an der „University of North Carolina at Chapel Hill“. ZINBA wurde zuerst in Betracht gezogen, dann aber ausgeschlossen, da es im Vergleich zu F-Seq, Hotspot und MACS2 370-fach langsamer läuft, den 4,5-fachen Arbeitsspeicher verbraucht, eine geringere „True Positive Rate“ (TPR) aufweist und auf GitHub nicht vertreten ist [28].

Tabelle 2.1.: Übersicht über die Peak-Caller

Eigenschaften und Optionen	ATAC-seq Pipeline	F-Seq	Hotspot	MACS2
Version	0.99999e 2016-08-26 06:34	1.85	v4.1	2.1.1.20160309
Latest commit	23.12.2017	31.03.2016	12.04.2014	05.05.2017
Mappability	Nein	Ja	Ja	Nein
Tag length	Ja	Ja	Ja	Ja
Omit regions	Nein	Nein	Ja	Nein
BAM Input	Ja	Nein	Ja	Ja
BED Input	Ja	Ja	Ja	Ja
FASTQ Input	Ja	Nein	Nein	Nein
broad peaks	Nein	Nein	Nein	Ja
Thresholding	IDR/p-val	SD(σ)	FDR	p-val/FDR
Optimiert für	ATAC-seq DNAse-seq	ChIP-seq DNAse-seq	ChIP-seq DNAse-seq	ChIP-seq
Zitierungen	7	279	583	4184

Übersicht über die Peak-Caller:
Version: Welche Version für diese Arbeit genutzt wurde.
latest commit [39–42]: Zu diesem Zeitpunkt wurde der Peak-Caller zuletzt auf GitHub in Form eines sog. commit, also einer aufgezeichneten Änderung des repository, aktualisiert (stand 29.12.2017).
Mappability: Wird eine Mappability-Datei benötigt, die angibt, welche Sequenzen wie eindeutig zugeordnet werden können?
Tag length: Sind Angaben zur Länge der tags möglich?
Omit regions: Kann man Stellen im Genom angeben, die z. B. wegen repetitiven Sequenzen vom Peak-Calling ausgeschlossen werden?
Input: Kann der Peak-Caller BAM, BED oder FASTQ Format als Input lesen?
Broad Peaks: Hat das Programm eine Voreinstellung, um breite Peaks zu berücksichtigen?
Thresholding: Kann ein Wert für IDR, FDR, „probability value“ (p-val) oder die Standardabweichung (SD) angegeben werden, um die Qualität der Peaks festzulegen?
Optimiert für: Für welche Sequenzierungsmethode wurde der Peak-Caller ursprünglich optimiert?
Zitierungen: Wie oft wurde die zugehörige Facharbeit bis zum 29.12.2017 zitiert?

2.3. „Pipeline for ATAC-seq and 4sU-seq plotting“ (PASsUS)

Um die ATAC-seq- und 4sU-seq-Daten zu den verschiedenen Zeitpunkten p.i. miteinander zu vergleichen, wurde im Rahmen dieser Dissertation die „Pipeline for ATAC-seq and 4sU-seq plotting“ (PASsUS) in der Programmiersprache R [43] geschrieben. Diese kann den $\log_2$-fold-change der ATAC-seq bzw. der 4sU-seq der behandelten Zellen bezüglich der Unbehandelten in verschiedenen Downstreambereichen errechnen und erstellt Grafiken zur Veranschaulichung. Alternativ können anstatt der $\log_2$-fold-changes auch die absoluten „Reads Per Kilobase Million“ (RPKM) oder die Differenzen der RPKM zwischen den Interventions- und Kontrolldaten berechnet werden.

2.3.1. Bedienung und Funktion von PASsUS

Um PASsUS über R aufrufen zu können, müssen folgende flags gesetzt sein:

„starts“ und „ends“

Mit **starts** (vector) werden die Startpunkte der Downstreambereiche angegeben, die untersucht werden sollen und mit **ends** (integer) die Länge ebendieser.

„inputGTF“

Bei **inputGTF** (string) muss der Dateipfad zu einer „General Transfer Format“ (GTF) Datei angegeben werden. Um zu berücksichtigen, dass die ATAC-seq- und 4sU-seq-Daten auf unterschiedliche „Genome Reference Consortium“ (GRC) Versionen gemapped werden können, wird in inputGTF_ATAC und inputGTF_4sU unterschieden.
In einem ersten Schritt werden die Downstreambereiche anhand der in der GTF-Datei annotierten Genendpunkte berechnet und in eine neue GTF-Datei gespeichert. Dabei werden alle Einträge entfernt, die nicht dem Feature-Type „Gene“ entsprechen. Unter Berücksichtigung des Strangs wird aus dem Endpunkt eines Genes und der Summe aus **starts** und 1 der Anfang des jeweiligen Downstreambereiches berechnet.
Aus dem Endpunkt eines Genes und der Summe aus **starts**, 1 und **ends** wird das Ende des jeweiligen Downstreambereiches berechnet. Das Addieren von 1 soll bewirken, dass der Downstreambereich nicht mit dem Gen überlappt. Dadurch kann

es bei Genen auf dem Minusstrang vorkommen, dass der Downstreambereich mit einer negativen Zahl annotiert ist, was einer späteren Weiterverarbeitung mit featureCounts im Wege steht. Deshalb werden diese Einträge entfernt.
Ist bei **subtract_regions** (string) der Dateipfad zu einer BED-Datei angegeben, so werden die Intervalle aus der Datei mit Hilfe von „bedtools subtract“ [44] unter Verwendung der Standardeinstellungen entfernt. Dies wird hier genutzt, um die Peaks der uninfizierten Replikate in p.i.-Replikaten auszuschließen, was wiederum z. B. „Transcription Factor Binding Sites“ (TFBS) aus der Auswertung entfernt. Da das Augenmerk auf größeren Bereichen mit offenem Chromatin liegt, könnten kleine Bereiche mit vielen Reads das Bild verzerren.
Analog dazu ruft **intersect_regions** (string) „bedtools intersect“ [44] unter Verwendung der Standardeinstellungen auf. Dies erlaubt z. B. nur Reads zu berücksichtigen, die in p.i. Peaks enthalten sind.

„pre_normalize“

Jedes Replikat unterliegt Verzerrungen aufgrund unterschiedlicher Bedingungen während der Durchführung der Experimente. Deshalb müssen die Replikate transformiert werden, um sie direkt vergleichbar zu machen. Dies geschieht durch Skalierung anhand von Merkmalen, die zwischen den Replikaten die gleiche Ausprägung aufweisen sollten.
Da unterschiedliche Downstreambereiche eines Replikats auf dieselbe Art und Weise normalisiert werden können und dies damit nur einmalig notwendig ist, kann man mit **pre_normalize** (logical) wählen, ob die für die Normalisation notwendige Datei erstellt werden soll.
Die 4sU-seq Daten werden nach rRNA normalisiert, deren Menge entweder aus Report Dateien importiert wird oder selbst berechnet werden kann. Die ATAC-seq Daten werden nach mtDNA normalisiert. Hierfür wird die Anzahl der mtDNA-Reads mit Hilfe von SAMtools [45] ausgelesen (samtools view -c -F260 file MT). Bei diesem Schritt braucht das Programm die Angabe einer gtf.genes.tab Datei.
Die Tabelle enthält in der ersten Spalte die ermittelten Reads und in der zweiten Spalte die daraus berechneten Normalisierungsfaktoren, wobei zuerst die Faktoren für ATAC-seq nach Zeit sortiert aufgelistet werden und darunter die Faktoren für 4sU-seq. Die Zuordnung der einzelnen Zeilen zu den Versuchsbedingungen erfolgt also anhand der Anzahl von als Input gegebenen BAM-Dateien.

„count_reads“

Anschließend wird im **count_reads** (logical) Schritt featureCounts [46] genutzt, um zu zählen, wie viele ATAC-seq Reads bzw. 4sU-seq Reads im entsprechenden Downstreambereich vorhanden sind (featureCounts -t gene -s 2 -p -a files). Dabei werden Reads, welche nicht eindeutig einem Bereich zugeordnet werden können, nicht gezählt.

„no_overlap“

Da sich der Read-through teilweise über 100 kbp erstreckt, reicht er auch über andere Gene hinaus. Ist **no_overlap** (logical) aktiviert, so werden alle Gene ausgeschlossen, bei denen sich zwischen Beginn des Gens und Ende des zu betrachtenden Downstreambereiches noch ein anderes Gen oder ein anderer Downstreambereich befindet. Da dies in Abhängigkeit zum Betrachtungsbereich steht, führen große Werte für **starts** oder **ends** zum Ausschluss von vielen Genen. Betrachtet man z. B. den Downstreambereich 200-205 kbp, sind bis auf 29 Gene alle ausgeschlossen.

„exp_RPKM_cut-offs“

Die Ergebnisse können für jedes Gen in Abhängigkeit von dem RPKM Wert 8 h p.i. stratifiziert werden. Mit **exp_RPKM_cut-offs** (vector) werden die cut-off-Werte zwischen den einzelnen Strata angegeben.

„draw_plots“

Nachdem alle Vorbereitungen getroffen sind, werden die Daten nun im Schritt **draw_plots** (logical) geplottet.
Zunächst werden die Gene auf alle Gene in einer xlsx-Datei, die mit **input_xlsx** (string) angegeben wird, reduziert. Dies hat den Hintergrund, dass Gene, welche zu keinem Zeitpunkt exprimiert sind, in der Auswertung nicht hilfreich sind und daher ausgeschlossen werden können. Außerdem werden aus dieser Datei weitere Informationen zum Filtern der Gene gewonnen. So kann z. B. nach Read-through und RPKM-Werten gefiltert werden.
Dann wird durch Division der vorher erstellten Faktoren, welche in „pre_normalize“ erstellt wurden, normalisiert. Der $\log_2$-fold-change wird unter Nutzung von LFC [26] berechnet. LFC errechnet eine sinnvolle Anzahl an Pseudocounts. Dies hat den Vorteil, dass weniger Verzerrung durch eine willkürliche Wahl der Anzahl an Pseudocounts entsteht. Der $\log_2$-fold-change wird für jeweils aus den Reads eines be-

stimmten Zeitraums p.i. (Dividend) im Verhältnis zu den Reads 0 h p.i. (Divisor) berechnet.
Die Datentabellen werden dann für das Plotten mit ggplot [47] umgeformt und zur Dokumentation in einem Ordner für temporäre Dateien abgelegt.

„comparison“

Mit **comparison** (character) kann gesteuert werden, ob der Vergleich zwischen den ATAC-seq- und den 4sU-seq-Daten standardmäßig anhand der $\log_2$-fold-changes oder alternativ mit den absoluten RPKM-Werten oder der Differenz der RPKM-Werte zwischen Interventions- und Kontrollbedingung durchgeführt wird.

2.3.2. Quelltext und Beispielaufruf

Der Quelltext sowie ein Beispielaufruf von PASsUS sind auf https://zenodo.org/ unter der DOI 10.5281/zenodo.4710625 veröffentlicht.

3. Ergebnisse

3.1. Vergleich der Peak-Caller

Wie in Kap. 1.5 beschrieben, wurden die Peak-Caller auf die ATAC-seq-Daten angewendet, um zu eruieren, welcher Peak-Caller die Veränderungen der Chromatinorganisation 8 h p.i. am treffendsten beschreibt. Idealerweise sollte bei jeder OCR ein Peak annotiert werden. Dadurch sollten sich Peaks und OCRs im Verlauf einer Infektion in gleicher Weise verändern.

3.1.1. Globale Eigenschaften der Peaks

Zuerst wurden die globalen Eigenschaften der Peaks anhand ihrer statistischen Maßzahlen verglichen.

F-Seq und Hotspot zeigten eine Mindestlänge der Peaks von 10 bp. MACS2 schnitt bei der Verlängerung der OCRs 8 h p.i. am besten ab, da sich der Mittelwert der Peak-Längen fast verdoppelt hat.
Bei F-Seq hingegen verdoppelte sich die Summe der in Peaks annotierten Basenpaare. Die Anzahl der annotierten Peaks nahm bei F-Seq zu und bei MACS2 ab (s. Tab. 3.1).
Laut F-Seq lagen unter Kontrollbedingungen 5,64 % und 8 h p.i. 11 % des menschlichen Genoms (rund 3 200 Mbp) in Form von OCRs vor.
Für die Betrachtung der empirischen Verteilungsfunktionen der einzelnen Peak-Caller ist zu beachten, dass die einzelnen Graphen eine unterschiedliche Anzahl an Peaks repräsentieren (s. Abb.3.1). Hier schnitten ATAC-seq Pipeline und MACS2 am besten ab. 8 h p.i. liegt eine Rechtsverschiebung der Kurven vor. Damit wurde die erwartete Verlängerung der Peaks dargestellt. F-Seq schnitt schlechter ab, da eine Zunahme der OCR-Längen nur sehr gering und nur für Peaks mittlerer Länge zu sehen war. Hotspot schnitt schlecht ab: Es zeigte sich eine Abnahme der OCR-Längen.

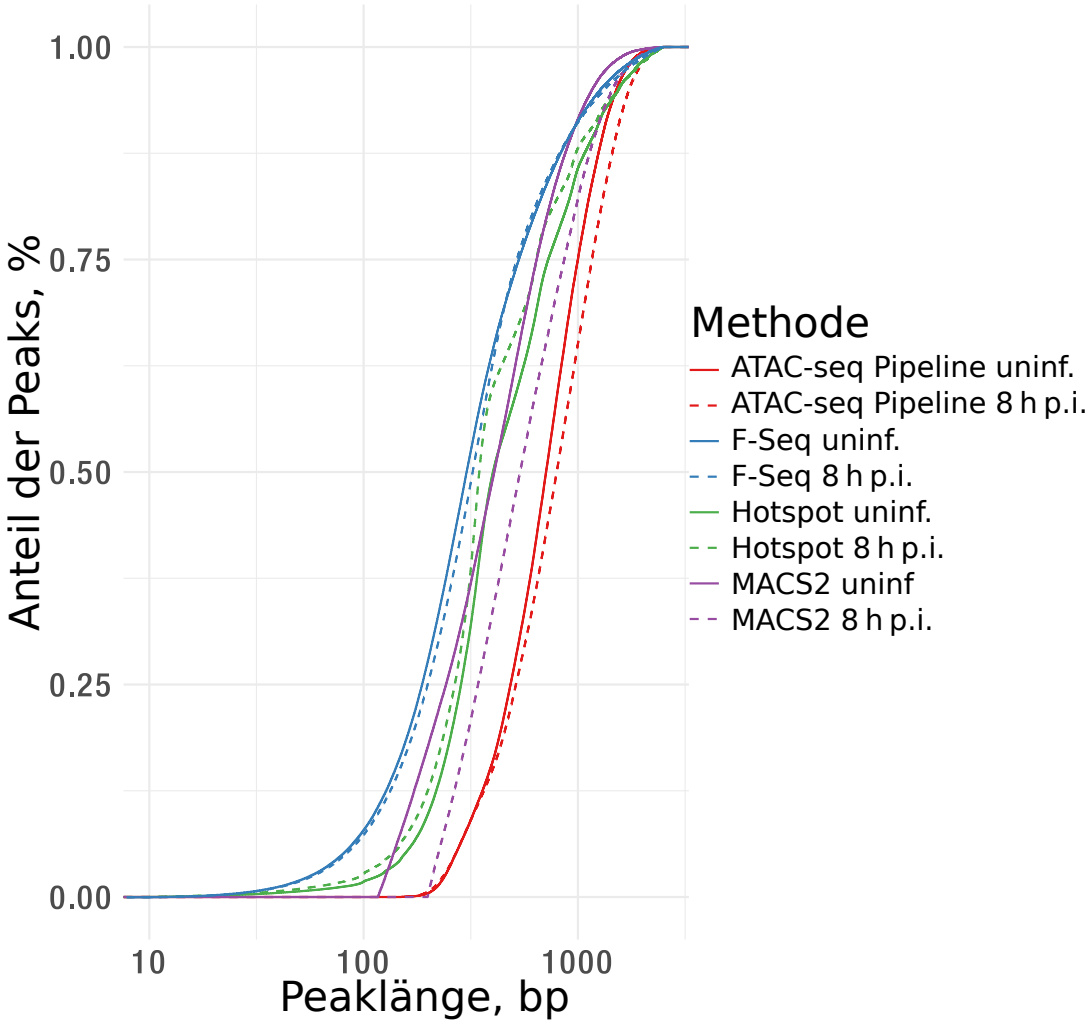


Abbildung 3.1.: Dargestellt sind Peak-Längen bei uninfizierten Zellen (durchgezogene Linien) und 8 h p.i. (gestrichelte Linien), welche durch die verschiedenen Peak-Caller berechnet wurden. Auf der Abszisse wurde die Peaklänge in Basenpaaren aufgetragen und auf der Ordinate, welcher Anteil der Peaks kürzer ist als die entsprechende Länge. Bei ATAC-seq Pipeline und MACS2 8 h p.i. fand eine Rechtsverschiebung der Kurve statt, was einer Verlängerung der Peaks nach Infektion entspricht. Bei der Betrachtung mittellanger Peaks von F-Seq konnte eine leichte Rechtsverschiebung festgestellt werden. Die Kurve der für 8 h p.i. mit Hotspot berechneten Peaks war linksverschoben. F-Seq annotierte in Relation zu den anderen Peak-Callern mehr kurze Peaks.

Ob die von Peak-Caller-A annotierten Peaks an derselben Position liegen, wie die von Peak-Caller-B annotierten Peaks, ist ein Qualitätsindiz: Wenn Peak-Caller OCRs perfekt detektieren würden, würden sich die Peaks von allen Peak-Callern komplett überschneiden. Inwiefern dies zutrifft, wurde im nächsten Schritt untersucht:

Tabelle 3.1.: Verteilung der Basenpaare in Peaks und Anzahl der Peaks.

Uninf.:	ATAC-seq Pipeline	F-Seq	Hotspot	MACS2
Min. bp	150	10	10	117
1st Qu.	485	187	286	243
Median	710	303	408	409
Mean	778	496	619	1096
3rd Qu.	1000	544	754	647
Max. bp	6889	105.619	25.076	5.079.063
Sum bp	65×10^6	203×10^6	141×10^6	115×10^6
Peaks	180×10^3	410×10^3	227×10^3	105×10^3
8 h p.i.:	ATAC-seq Pipeline	F-Seq	Hotspot	MACS2
Min. bp	150	10	10	199
1st Qu.	514	199	269	345
Median	798	326	350	545
Mean	881	577	592	1973
3rd Qu.	1167	544	659	885
Max. bp	5660	243.183	25.093	5.464.850
Sum bp	50×10^6	396×10^6	157×10^6	96×10^6
Peaks	141×10^3	686×10^3	266×10^3	48×10^3

Verteilung der Basenpaare in Peaks und Anzahl der Peaks.
Min. bp / Max. bp: Anzahl der Basenpaare im kürzesten / längsten Peak.
1st Qu. / Median / 3rd Qu.: Anzahl der Basenpaare im Peak, der länger ist als 25 % / 50 % / 75 %) der restlichen Peaks.
Mean: Mittelwert der Länge aller Peaks in Basenpaaren.
Sum bp.: Summe aller Basenpaare, die allen Peaks zugeordnet wurden.
Peaks: Anzahl der durch einen Peak-Caller als Peak markierten Bereiche.
Die Menge der den Peaks zugeordneten Basenpaare nach Infektion war bei MACS2 und ATAC-seq Pipeline reduziert, während sie bei F-Seq und Hotspot zunahm. Der Median und der Mittelwert hatte sich für keinen der Peak-Caller wesentlich verändert, wobei er sich bei allen Peak-Callern bis auf MACS2 zumindest geringfügig vergrößerte. MACS2 ist zudem der einzige Peak-Caller, bei dem sich die Mindestlänge der Peaks zwischen den Versuchsbedingungen änderte.

In absoluten Zahlen hatten die von F-Seq annotierten Peaks die größte positionelle Schnittmenge mit von anderen Peak-Callern annotierten Peaks. Gleichzeitig hatten sie auch die größte Differenzmenge. (s. Abb. 3.2 und 3.3) Dies lag unter anderem daran, dass die Summe der Peak-Längen in bp für F-Seq das Zwei- bis Dreifache betrug als für die anderen Peak-Caller (s. Tabelle 3.1). Bei F-Seq wurden 32 % (bzw. 52 % 8 h p.i.) der Basenpaare nicht von anderen Peak-Callern markiert. Dies bedeutet entweder, dass F-Seq als einziger Peak-Caller diese OCRs erkannte oder dass

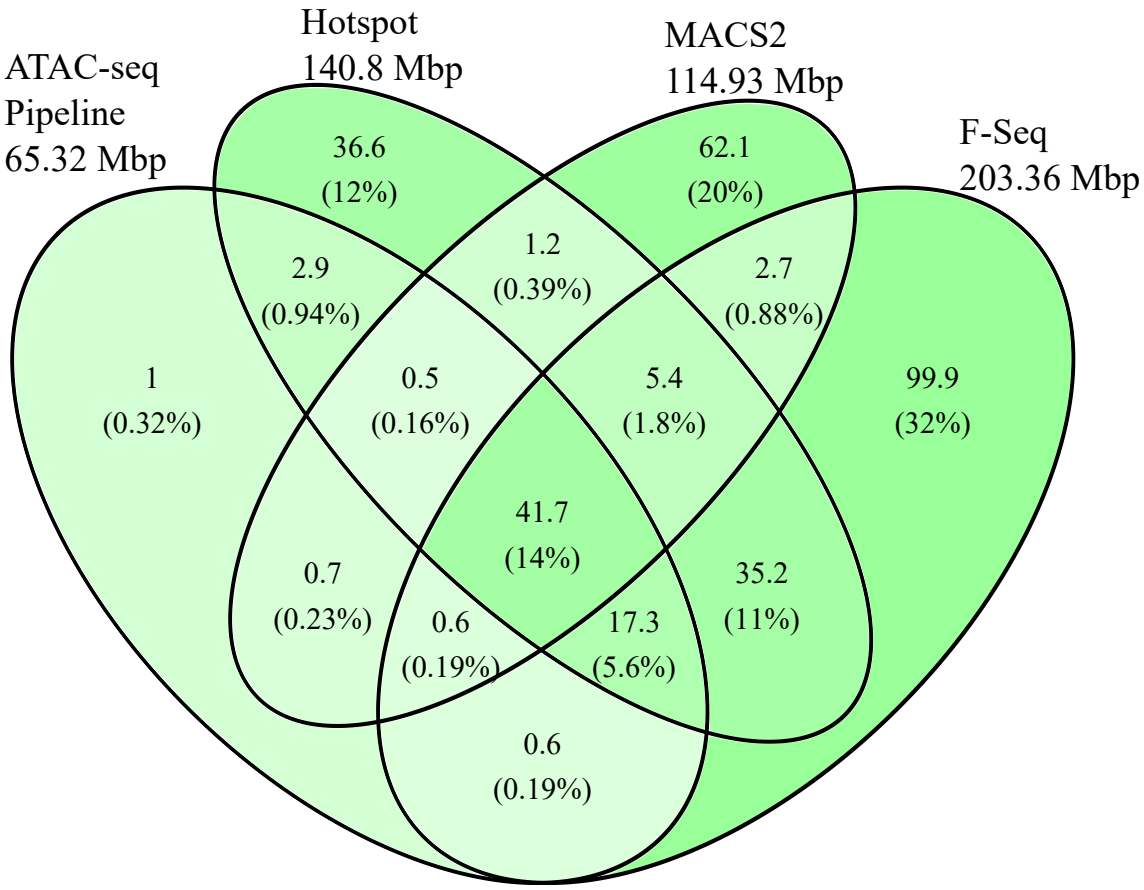


Abbildung 3.2.: Positionelle Übereinstimmung der den Peaks zugeordneten Mbp für uninfizierte Zellen und prozentualer Anteil.
Über dem Venn-Diagramm ist zu sehen, wie viele Basenpaare die einzelnen Peak-Caller den jeweiligen Peaks zugeordnet haben. Dies entspricht der Summe der Peak-Längen in der Einheit Basenpaare. Die Zellen des Venn-Diagramms geben die Schnittmengen und Differenzmengen zwischen den Peaks der verschiedenen Peak-Caller in Mbp und deren prozentualen Anteil an.

hier keine OCRs vorlagen und F-Seq falsch-positive Peaks annotiert hat. Stichproben mit dem Genom-Viewer deuteten auf Letzteres hin (vgl. 3.4). Trotzdem schien F-Seq den Read-through für 8 h p.i. durch lange Peaks am besten erfasst zu haben. Ähnliches galt für Hotspot, wobei hier die Übereinstimmung mit den anderen Peak-Callern noch geringer war. Die von ATAC-seq Pipeline als Teil eines Peaks annotierten Basenpaare waren nur zu 0.32 % bzw. 0.012 % nicht von anderen Peak-Callern als Teil eines Peaks annotiert. Daher ist davon auszugehen, dass ATAC-seq Pipeline wenig falsch-positive Peaks errechnete.

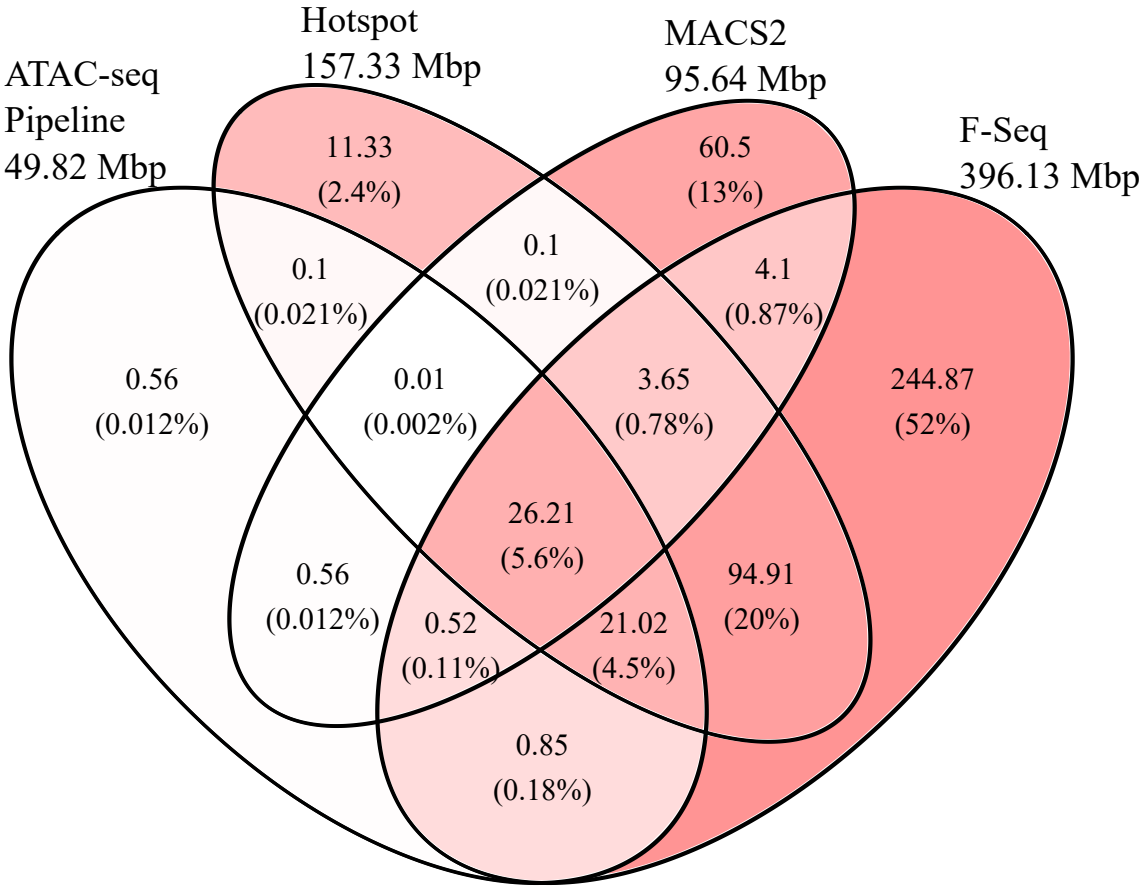


Abbildung 3.3.: Positionelle Übereinstimmung der den Peaks zugeordneten Mbp 8 h p.i. und prozentualer Anteil.
Über dem Venn-Diagramm ist zu sehen, wie viele Basenpaare die einzelnen Peak-Caller den jeweiligen Peaks zugeordnet haben. Dies entspricht der Summe der Peak-Längen in der Einheit Basenpaare. Die Zellen des Venn-Diagramms geben die Schnittmengen und Differenzmengen zwischen den Peaks der verschiedenen Peak-Caller in Mbp und deren prozentualen Anteil an.

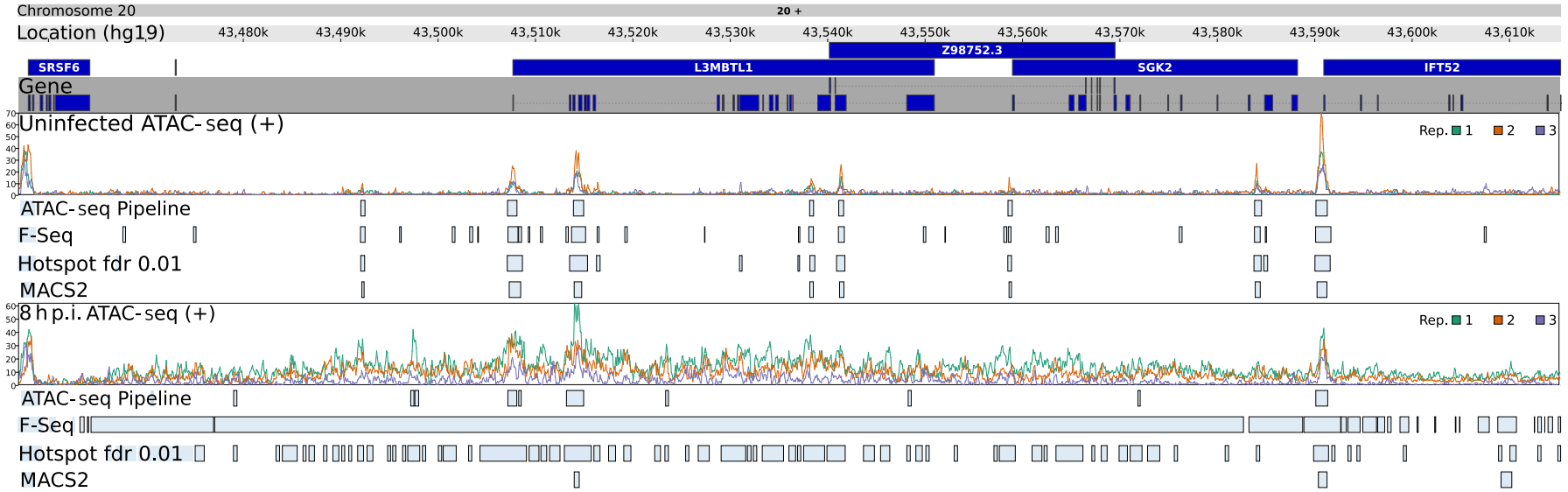


Abbildung 3.4.: Vergleich der Peak-Caller im Genom Viewer. Zu sehen ist von oben nach unten der Bereich des 20. Chromosoms von rund 43.458k bis 43.615k nach GRCh38 bzw. hg19. Die Genregionen mit den entsprechenden Genen sind in den dunkelblauen Rechtecken dargestellt. Es folgt zuerst für uninfizierte Zellen die Transkriptionsaktivität aufgeteilt nach den 3 Replikaten und die von den verschiedenen Peak-Callern zugeordneten Peaks (gepoolte Daten, hellblaue Rechtecke). Unten sind die Spuren für 8 h p.i. zu sehen. ATAC-seq-Pipeline und MACS2 scheinen unter den Standardeinstellungen eher Peaks zu markieren, die sich deutlich aus der Umgebung abheben.

Die Schnittmenge aller Peak-Caller sank vom Zustand uninfizierter Zellen zum Zustand 8 h p.i. von 14 % auf 5,6 %. Die Anzahl der Basenpaare, welche nur von ATAC-seq, Hotspot oder MACS2 in Peaks annotiert wurden, nahm ab. Dahingegen nahm die Anzahl der Basenpaare, welche sowohl von F-Seq als auch einem weiteren Peak-Caller in Peaks annotiert wurden, zu. Lässt man F-Seq außen vor, so sank die Anzahl der von allen anderen 3 Peak-Callern in Übereinstimmung annotierten Peaks weiterhin erheblich im Verlauf der Infektion. Damit lag eine im Verlauf der Infektion steigende Uneinigkeit der Peak-Caller vor, welche für F-Seq besonders stark ausgeprägt war. Ob diese Sonderstellung von F-Seq darauf zurückzuführen ist, dass die von F-Seq annotierten Peaks entweder einen besonders hohen oder einen besonders niedrigen prädiktiven Wert für das Vorhandensein von OCRs aufwiesen, lässt sich hieraus nicht ableiten.

SRSF6 ist ein Paradebeispiel für ein Gen, mit ausgeprägtem Read-through (s. Abb. 3.4). SRSF6 zeigt beispielhaft, dass die von den Peak-Callern annotierten Peaks für die uninfizierten Zellen größere Übereinstimmungen aufweisen, als 8 h p.i.. F-Seq hat lange Bereiche von offenem Chromatin 8 h p.i. durch lange Peaks hervorragend erfasst. Hotspot erfasste dies noch durch mehrere kürzere Peaks mittelmäßig. Im Gegensatz dazu annotierten ATAC-seq Pipeline und MACS2 nur kurze Ansammlungen von Reads als Peak, welche sich von der unmittelbaren Umgebung abhoben.

3.1.2. Vergleich zwischen den „downstream Open Chromatin Region(s)“ (dOCR)

Nachdem nun die Basiseigenschaften der Peaks miteinander verglichen wurden, wurde im nächsten Schritt geprüft, ob sich die aus den Peaks abgeleitete Hilfsgröße dOCR eignet, eine Trennschärfe zwischen uninfizierten Zellen und 8 h p.i. Replikaten zu erzielen. Diese Trennschärfe könnte genutzt werden, um zu untersuchen, ob Experimente mit Salz- und Hitzestress ähnliche Auswirkungen auf die Chromatinorganisation haben, wie die lytische HSV-1-Infektion.
Die Berechnung von dOCR ermöglichte für Hotspot eine geringe und für F-Seq eine hohe Trennschärfe zwischen uninfizierten Zellen und 8 h p.i.. Bei ATAC-seq und MACS2 konnte keine Trennschärfe erreicht werden (s. Abb. 3.5 und Abb. 3.6)
Aufgrund der iterativen Berechnung von dOCR (s. Kap. 1.4.4) begünstigte eine größere Anzahl von Peaks größere Werte für dOCR. Dies bedingte eine höhere Trennschärfe, wenn alle annotierten Peaks berücksichtigt wurden und nicht nur jene Peaks mit einem hohen Score. Daher wurden im Weiteren alle Peaks berücksichtigt, auch wenn dadurch eine höhere Anzahl falsch-positiver Peaks einbezogen wurden.

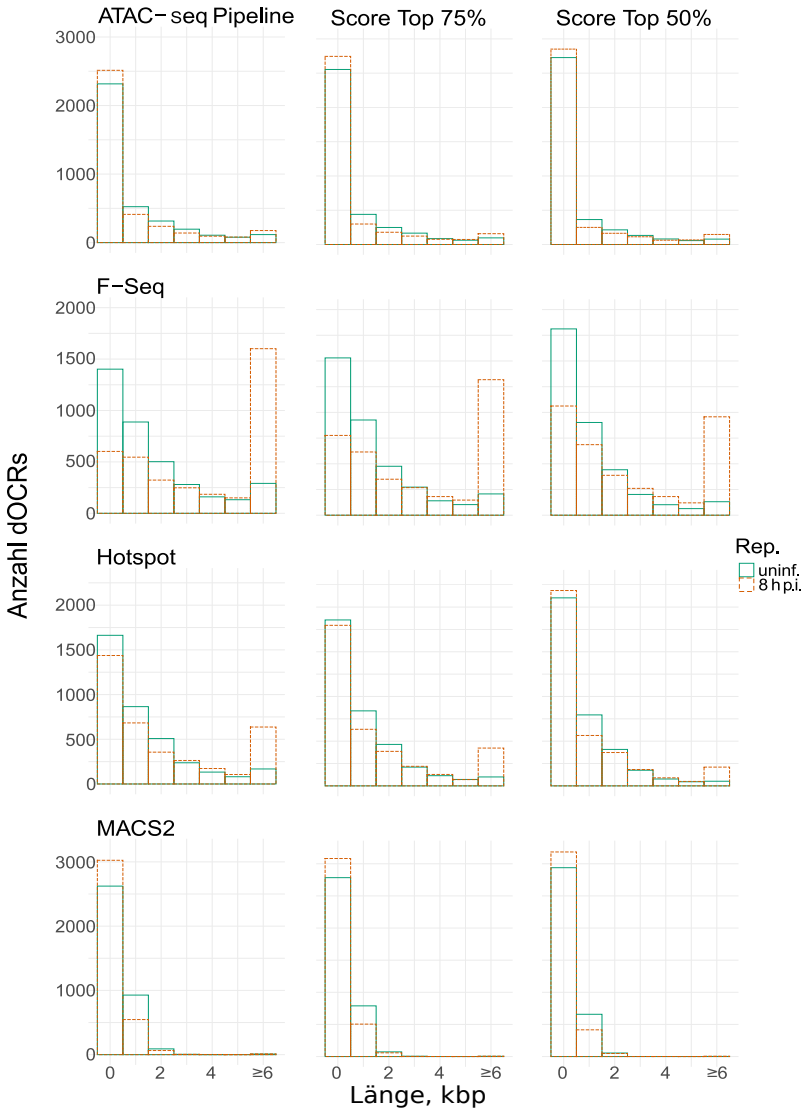


Abbildung 3.5.: Längenverteilung der dOCR-Längen für die verschiedenen Peak-Caller und Scores, 3684 Gene berücksichtigt.
1. Zeile ATAC-seq Pipeline, 2. Zeile F-Seq, 3. Zeile Hotspot, 4. Zeile MACS2. Die linke Spalte bildet dOCR-Längen auf Basis von allen Peaks ab. Die Mittlere auf der Basis von Peaks, bei denen der Score größer war als die dritte Quartile. Die Rechte auf der Basis von Peaks, bei denen der Score größer war als der Median. Auf den Abszissen wurden die entsprechenden Längen der dOCR angegeben, wobei im letzten Balken alle dOCR-Längen zusammengetragen wurden, die länger als 6 kbp waren. Auf den Ordinaten wurde die Anzahl der dOCR-Längen aufgetragen, welche eine bestimmte Länge überschreiten.

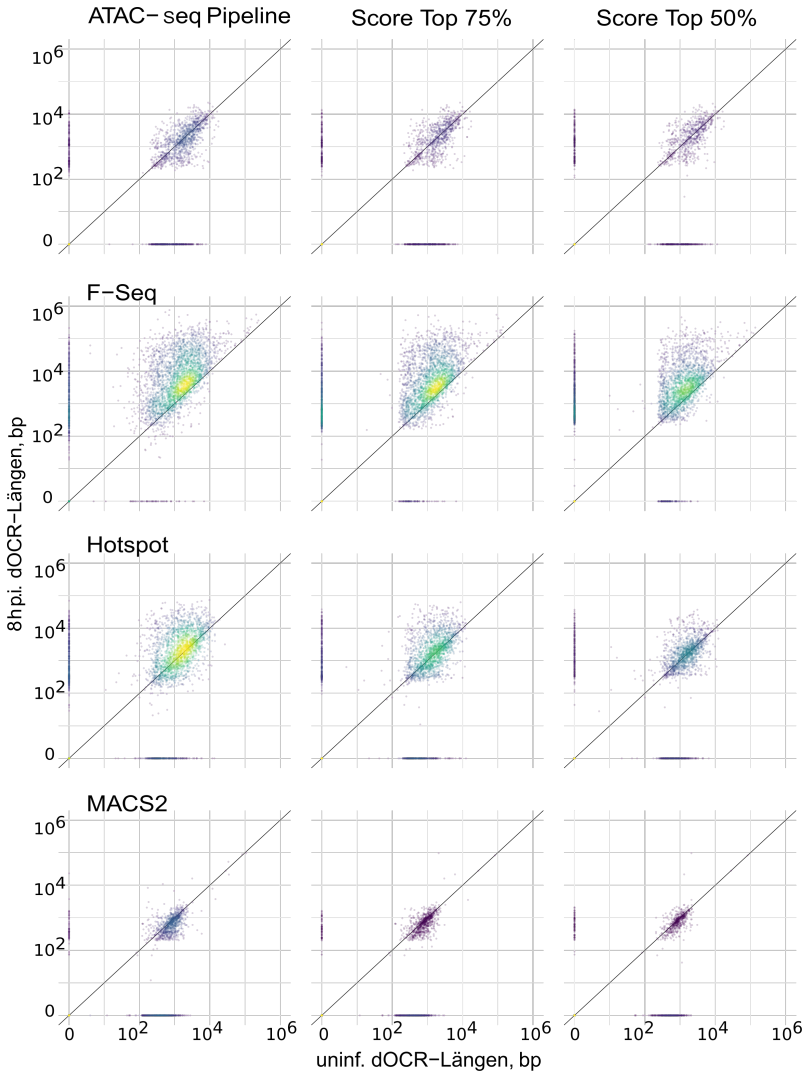


Abbildung 3.6.: Längenverteilung der dOCR-Längen für die verschiedenen Peak-Caller und Scores, 3684 Gene berücksichtigt
1. Zeile ATAC-seq Pipeline, 2. Zeile F-Seq, 3. Zeile Hotspot, 4. Zeile MACS2. Die linke Spalte bildet dOCR-Längen auf der Basis von allen Peaks ab. Die Mittlere auf der Basis von Peaks, bei denen der Score größer war als die dritte Quartile. Die Rechte auf der Basis von Peaks, bei denen der Score größer war als der Median.
Auf den Abszissen wurden die entsprechende dOCR-Längen für uninfizierte Zellen angegeben. Auf den Ordinaten wurden dOCR-Längen für 8 h p.i. aufgetragen.
Die Geraden entsprechen den Hauptdiagonalen. Die Helligkeit der Farbe entspricht der Dichte an Datenpunkten, d.h. gelb symbolisiert hier im Gegensatz zu blau eine hohe Punktdichte.

Da ATAC-seq Pipeline beim annotieren von Peaks auf MACS2 zurückgreift, wiesen beide Peak-Caller auch eine ähnliche Verteilung der dOCR-Längen auf (s. Abb. 3.7). Im Gegensatz dazu streuten die dOCR-Längen bei Hotspot und besonders bei F-Seq stärker nach oben. Dieser Effekt verstärkte sich 8 h p.i., was die höhere Trennschärfe zur Folge hatte.
Bei MACS2 ergab sich auch ein Unterschied zwischen uninfizierten Zellen und 8 h p.i.. Die von MACS2 berechnete minimale Peak-Länge betrug für die Kontrollbedingung 117 bp und 8 h p.i. 199 bp (s. Abb. 3.7). Die auf Basis von MACS2 berechneten dOCR-Längen waren nicht ausgeprägt rechtsschief. Das heißt, die Streuung der dOCR-Längen war hier nicht zu Gunsten einer erhöhten Anzahl von besonders langen dOCR-Längen hin schief. Bei Hotspot und besonders bei F-Seq war die Streuung der dOCR-Längen für 8 h p.i. zu Gunsten einer erhöhten Anzahl von besonders langen dOCR-Längen hin schief. Außerdem war eine diskrete Linie der Datenpunkte aufgrund der minimalen Peak-Länge, wie es für MACS2 der Fall war, nicht abgrenzbar; die minimale Peak-Länge war mit 10 bp geringer und bei vielen kurzen Peaks sinkt die Wahrscheinlichkeit, dass ein dOCR nur einen Peak enthält.
Zudem wurde der Korrelationskoeffizient nach Spearman berechnet. Dieser war für ATAC-seq Pipeline und MACS2 mit 0,78 für uninfizierte Zellen bzw. 0.85 8 h p.i. erwartungsgemäß am höchsten. Es zeigte sich eine Ähnlichkeit zwischen F-Seq und Hotspot, die mit 0.67 für uninfizierte Zellen und 0.68 8 h p.i. als einziges Paar keinen Abfall des Korrelationskoeffizienten 8 h p.i. zeigten. Dies verdeutlicht erneut, dass ATAC-seq Pipeline und MACS2 im Besonderen, aber auch die anderen Peak-Caller, nicht für die Analyse von langen OCRs optimiert sind.

3.1.3. Tiefer gehende Analyse mit F-Seq

F-Seq wies eine gute Trennschärfe zwischen uninfizierten Zellen und 8 h p.i. auf (s. Abb. 3.5 und 3.6). Dies führte zu der Entscheidung, für die weitere Auswertung F-Seq zu verwenden. Daher wird F-Seq hier genauer beschrieben.
F-Seq annotierte unter Berücksichtigung aller Peaks eine hohe Rate an falsch-positiven Peaks. Dies heißt, dass F-Seq an Positionen Peaks annotiert, an denen bei genauerem Hinsehen keine OCRs vorliegen (s. Abb. 3.2 und 3.3). Um dies genauer zu untersuchen, wurden Daten zu verschiedenen Schwellenwerten für den Score angefertigt (s. Abb. 3.8).

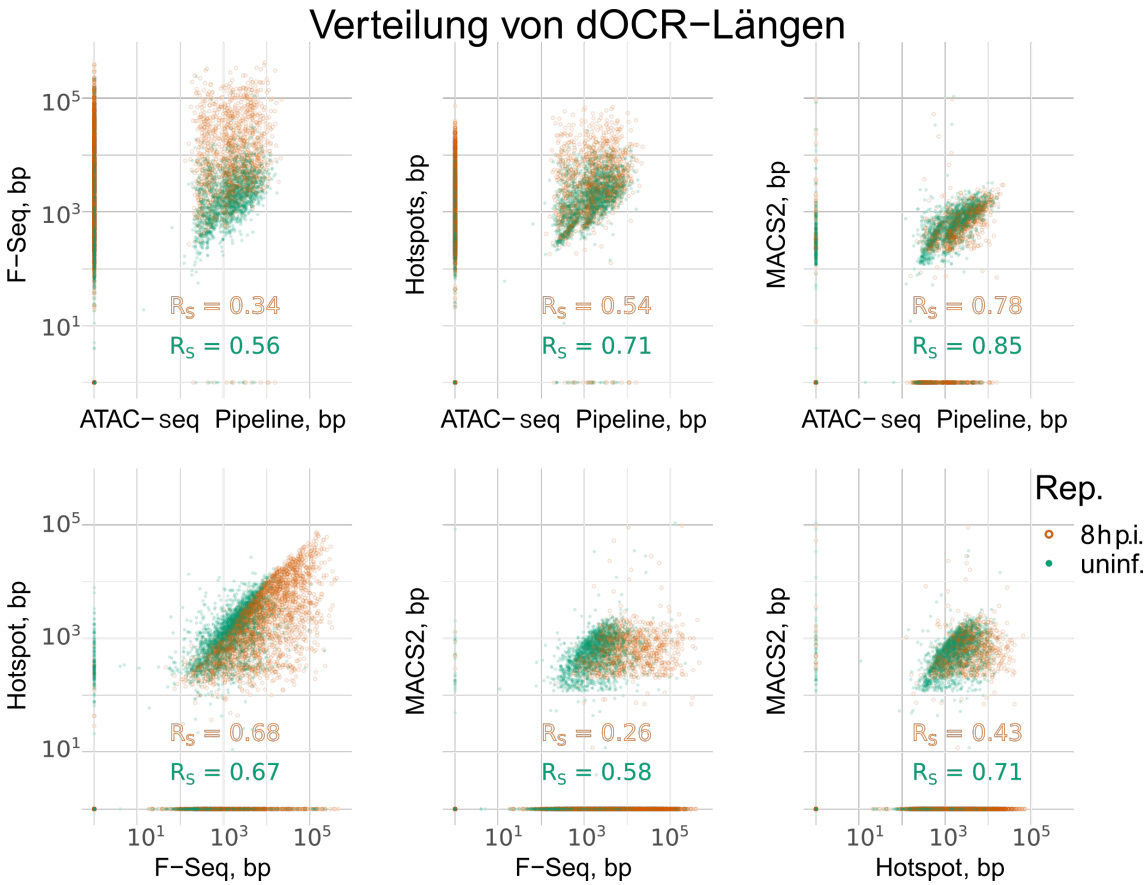


Abbildung 3.7.: Verteilung von dOCR für verschiedene Peak-Caller im Vergleich.
1. Zeile: F-seq, Hotspot und MACS2 gegen ATAC-seq Pipeline aufgetragen.
2. Zeile: Hotspot und MACS2 gegen F-Seq aufgetragen, MACS2 gegen Hotspot aufgetragen.
Jeder Graph enthält jeweils 3684 Beobachtungen für uninfizierte Zellen und 8 h p.i..
Auf den Ordinaten und Abszissen wurden die dOCR-Längen in bp für die jeweiligen Peak-Caller aufgetragen. Zudem wird die Spearman Korrelation für uninfizierte Zellen und 8 h p.i. getrennt angegeben ($p < 0.001$).

Mit zunehmend großen Schwellenwerten nahmen die Peak-Längen zu, das Verhältnis von Peak-Längen der uninfizierten Zellen zu Peak-Längen 8 h p.i. nahm jedoch ab. Dies heißt, dass größere OCRs mit einer höheren Treffsicherheit laut Score erkannt wurden. Jedoch ermöglicht dies keine bessere Differenzierung zwischen uninfizierten und infizierten Replikaten.

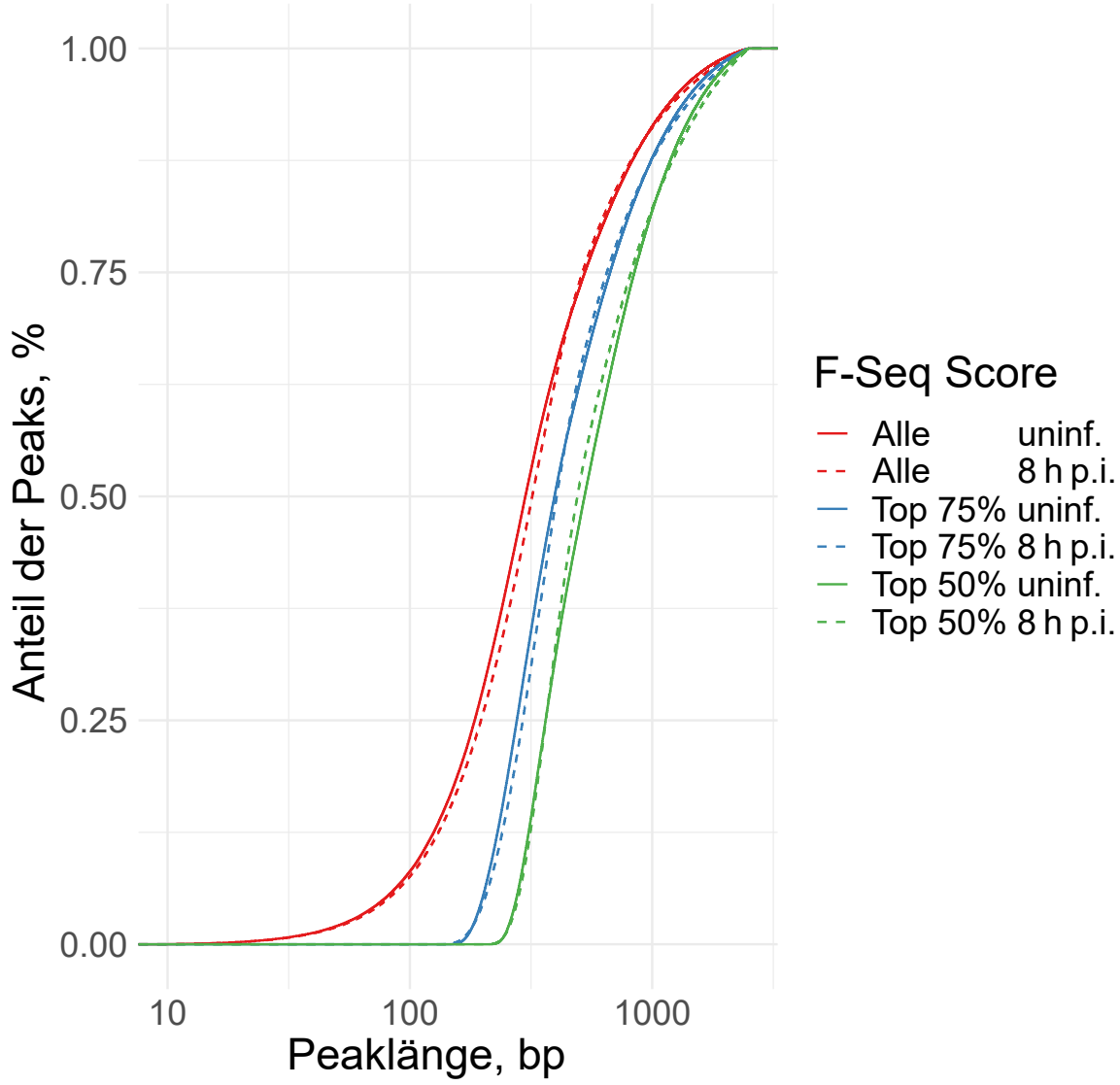


Abbildung 3.8.: Peak-Längen für uninfizierte Zellen (durchgezogene Linien) und 8 hp.i. (gestrichelte Linien) für F-Seq mit verschiedenen Schwellenwerten für den Score.
Auf der Abszisse wurde die Peaklänge in Basenpaaren aufgetragen und auf der Ordinate, welcher Anteil der Peaks kürzer ist als die entsprechende Länge. Die Schwellenwerte für den Score wurden so gewählt, dass die rote Kurve alle Peaks und die blaue bzw. grüne Kurve die 75 % bzw. 50 % der Peaks mit dem höchsten Score zeigen. Je größer der Grenzwert für den Score war, desto größer wird der Anteil der Kurve, welcher 8 hp.i. im Vergleich zu uninfizierten Zellen linksverschoben ist.

F-Seq annotierte 8 h p.i. im Vergleich zum uninfizierten Replikat mehr Peaks (686 000 zu 410 000, 167,3 %). Dies galt unabhängig davon, ob man alle Peaks betrachtete oder nur Peaks in Genen oder im Downstreambereich (s. Abb. 3.9.1–4). Die Verteilung der Peaks alleine erlaubte keine Aussagen über die OCR oder DoTT. Daher wurden auf der Basis der Peaks verschiedene Hilfsgrößen eingeführt (s. Kap. 1.4.4 und 1.4.5).
Der Vorteil der aggregierten Peak-Längen im 5 kbp Downstreambereich ist, dass das Ausmaß des offenen Chromatins in diesem kurzen Bereich nur gering überschätzt werden kann (s. Abb. 3.10.1). Dahingegen kommt es zu einer Unterschätzung des Ausmaßes von offenem Chromatin. Im 5 kbp Downstreambereich wurden für 694 von 3684 Genen (18,8 %) keine Peaks annotiert.
Für aggregierte Peak-Längen im 50 kbp Downstreambereich wird das Ausmaß des offenen Chromatins hingegen überschätzt. Im 50 kbp Downstreambereich wurden für nur 49 (uninfizierte Zellen) bzw. 24 (8 h p.i.) Gene keine Peaks annotiert.
Die Hilfsgröße dOCR reduziert die Nachteile eines festen Betrachtungsrahmens (s. Abb. 3.10.3). Die dOCR bieten eine gute Trennschärfe zwischen uninfizierten Zellen und 8 h p.i. und betrugen für 647 (uninfizierte Zellen) bzw. 245 (8 h p.i.) Gene 0 bp. Damit war das Ausmaß der Überschätzung geringer als für das 50 kbp Fenster.

3.1.4. Abschließende Beurteilung

F-Seq eignet sich als einziger Peak-Caller hervorragend zur Berechnung von dOCR. Die von F-Seq annotierten Peaks waren sowohl für die ATAC-seq-Daten uninfizierter Zellen als auch infizierter Zellen überwiegend plausibel. Dahingegen spricht die unplausibel hohe Anzahl an in Peaks annotierten Basenpaaren (11 % des Genoms) in Kombination mit der geringen Schnittmenge mit anderen Peak-Callern (s. Abb. 3.3) für eine hohe Anzahl falsch-positiv annotierter Peaks.
ATAC-seq Pipeline scheint breite Peaks zwar weniger gut beschrieben zu haben, wird aber aufgrund seiner guten Überschneidung mit anderen Peak-Callern empfohlen, wenn eine niedrige Rate an falsch-positiven Peaks notwendig ist. Aufgrund der vielen Basenpaare, die entweder nur durch Hotspot oder MACS2 Peaks zugeordnet wurden, scheinen diese Peak-Caller weniger geeignet. Zudem nahm bei MACS2 und ATAC-seq Pipeline die Summe der Basenpaare 8 h p.i. ab und bei Hotspot fehlte eine Verlängerung des durchschnittlichen Peaks, was für ein über weite Bereiche offenes Chromatin untypisch scheint.

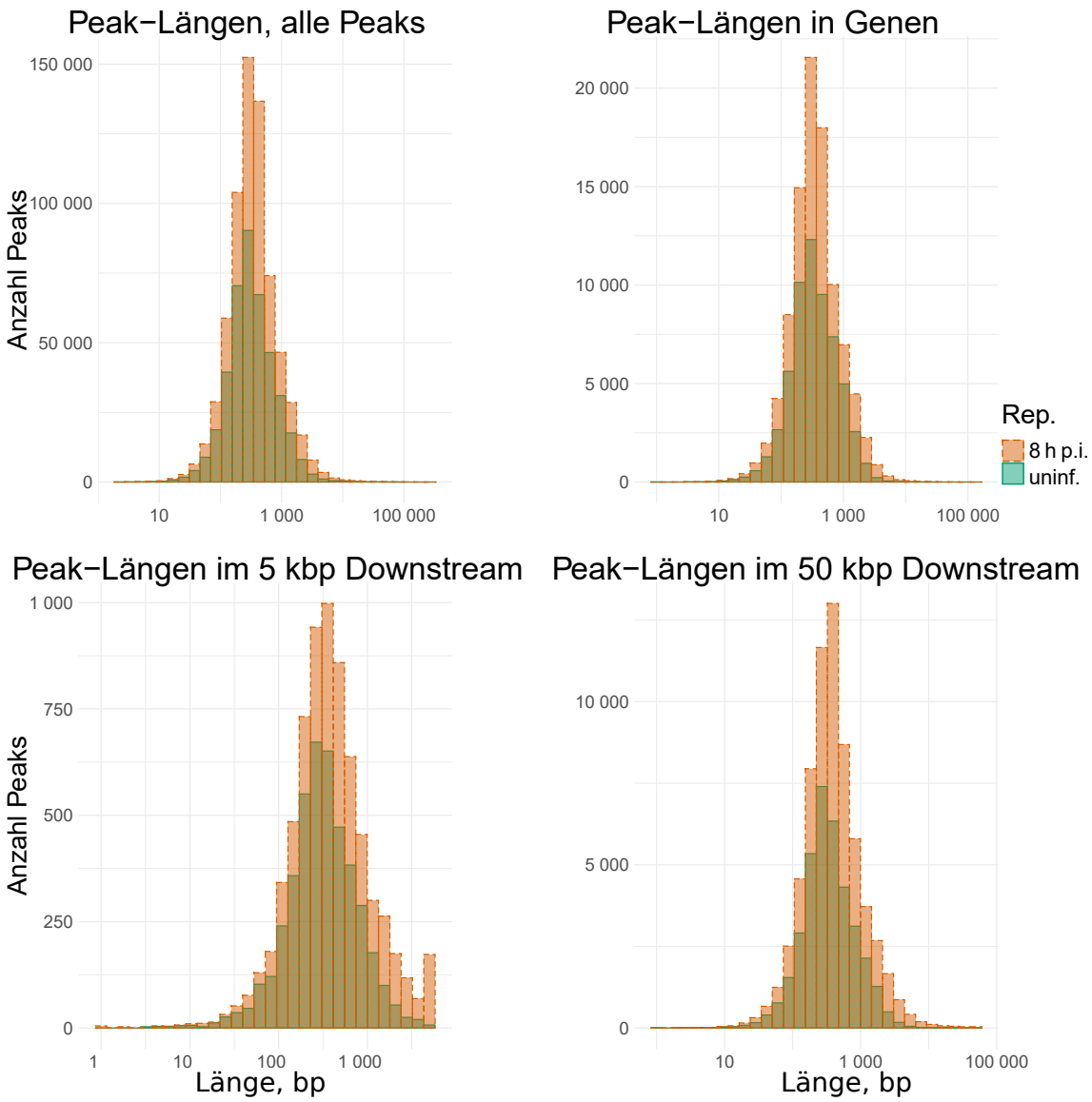


Abbildung 3.9.: Längenverteilung der Peaks für F-Seq an verschiedenen Positionen.
Auf den Abszissen wurden die Peak-Längen in Basenpaaren aufgetragen. Auf den Ordinaten wurden die Anzahlen von Peaks aufgetragen. Von links nach rechts und oben nach unten:
1: Peak-Längen aller Peaks für F-Seq einzeln, ungeachtet der Position.
2: Peak-Längen von Peaks in Genen. Es wurden alle Gene berücksichtigt.
3 und 4: Peak-Längen einzeln im 5 kbp bzw. 50 kbp Downstreambereich aller Gene.

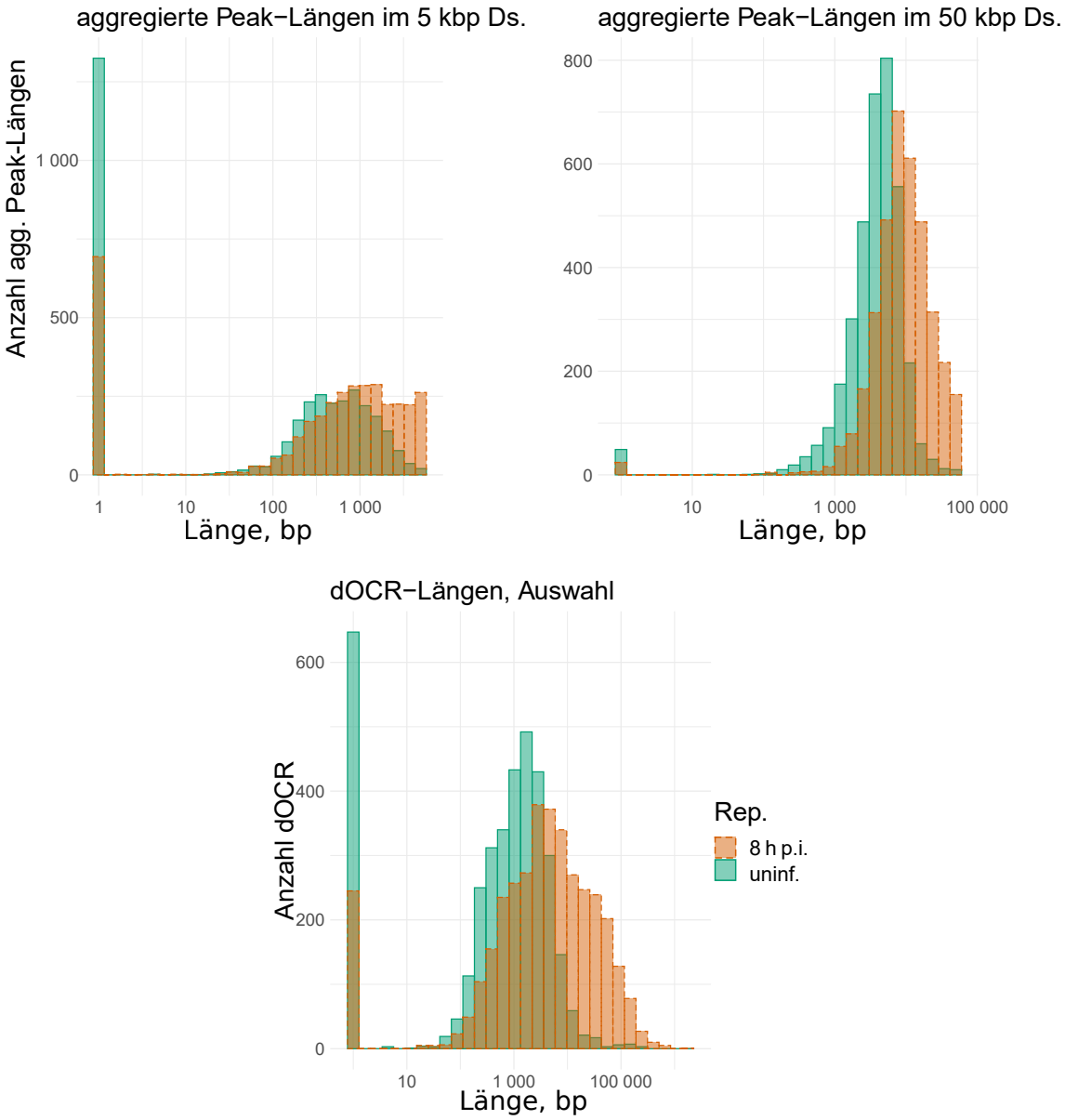


Abbildung 3.10.: Längenverteilung der „aggregierte Peaks“ (agg. Peaks) bzw. dOCR für F-Seq.
Auf den Abszissen wurden die Längen in bp der agg. Peaks bzw. dOCR aufgetragen. Auf den Ordinaten wurden die Anzahlen von agg. Peaks bzw. dOCR aufgetragen. Von links nach rechts und oben nach unten:
1 und 2: Aggregierte Peak-Längen für den 5 kbp bzw. 50 kbp Downstreambereich (Ds) 3684 ausgewählter Gene.
Enthält ein Downstreambereich keinen Peak, so entspricht die aggregierte Peak-Länge für das entsprechende Gen null.
3: Anzahl der dOCR-Längen für 3684 vorausgewählte Gene.

3.2. Zusammenhänge zwischen DoTT, OCRs, dOCR und agg. Peaks bei lytischer HSV-1 Infektion

Im letzten Kapitel wurde auf die Eigenschaften und Plausibilität von dOCR und agg. Peaks eingegangen. Im nächsten Schritt wurden die Zusammenhänge mit DoTT und Read-through untersucht.

3.2.1. Read-through, agg. Peaks und dOCR

Um zu überprüfen, ob Read-through mit agg. Peaks bzw. dOCR zusammenhängt, wurden agg. Peaks und dOCR in Abb. 3.11 gegeneinander aufgetragen. Es fällt auf, dass der prozentuale Read-through weder mit den agg. Peaks noch mit dOCR korrelierte. Dies galt sowohl für uninfizierte Zellen als auch für 8 h p.i..
Anscheinend bewirkt die Infektion sowohl einen vermehrten Read-through als auch vermehrte dOCR, ohne dass diese beiden Effekte mit der hier gezeigten Methodik in Korrelation gebracht werden konnten. Die Betrachnung von dOCR war auch hier der Betrachtung von agg. Peaks überlegen, denn sie ermöglichte sowohl die Erkennung von Genen ohne OCRs im Downstreambereich als auch die Erkennung von Genen mit vielen OCRs im Downstreambereich als solche.

3.2.2. Read-through, ATAC-seq-RPKM und 4sU-seq-RPKM

In Kap. 3.2.1 konnte anhand der aus Peaks abgeleiteten Größen keine Korrelation zwischen Read-through und dOCR gezeigt werden. Da hierbei nur die Information über die Position und Länge der Peaks genutzt wurde, nicht aber über die Signalstärke der Peaks in Form von Reads, wurden im nächsten Schritt die RPKM-Werte im Downstreambereich betrachtet, um einen eventuellen Zusammenhang zwischen DoTT und OCRs im Downstream zu eruieren.

Da Differenzen und Quotienten die Daten auf unterschiedliche Weise gewichten, wurden sowohl die absoluten RPKM-Werte, als auch deren Differenzen und $\log_2$-fold-changes untersucht. Dabei stellte sich heraus, dass sich diese drei Betrachtungsweisen im Streudiagramm nur gering unterscheiden. Allenfalls die Streuung der $\log_2$-fold-changes für ATAC-seq war zu Gunsten langer dOCR hin schief (s. Abb. 3.12). Weiterhin unterscheidete sich die Korrelation zwischen ATAC-seq-RPKM und 4sU-seq-RPKM nicht in Abhängigkeit vom prozentualen Read-through. Hierbei korrelierten ATAC-seq-RPKM und 4sU-seq-RPKM unter Betrachtung der RPKM-Quotienten

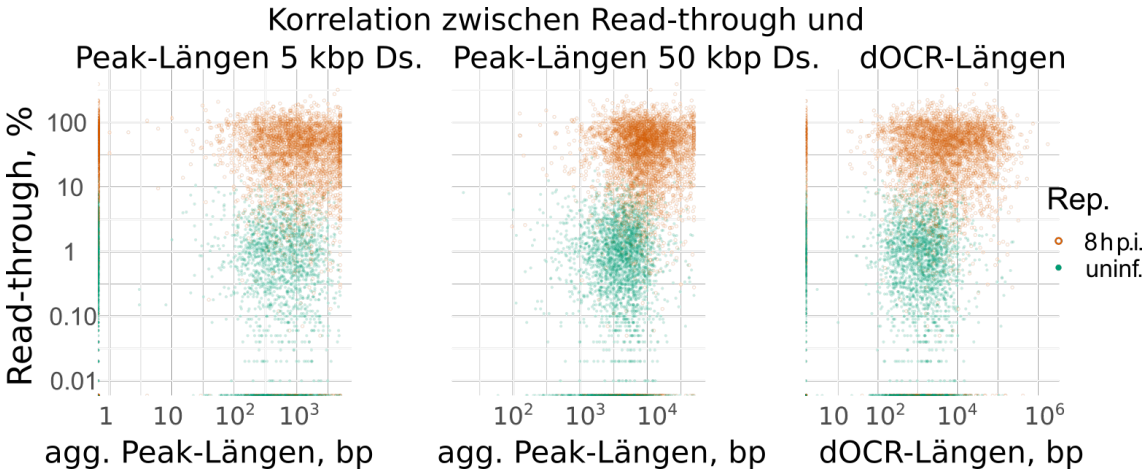


Abbildung 3.11.: Korrelation zwischen Read-through und aggregierten Peak-Längen bzw. dOCR-Längen.:
Jeder Datenpunkt symbolisiert ein Gen. Auf der Abszisse wurden die dOCR-Längen bzw. die aggregierten Peak-Längen in bp aufgetragen. Auf der Ordinate wurde der Read-through Wert der entsprechenden Gene aufgetragen.
Es wurden alle von F-Seq annotierten Peaks größer 10 bp verwendet, ungeachtet des Scores oder der Peak-Länge.
Die erste und zweite Abbildung von links zeigen die aggregierten Peak-Längen für den 0-5 bzw. 0-50 kbp Downstreambereich (Ds). Dies erkennt man auch an den Datenpunkten für 8 h p.i., welche 0-5 bzw. 0-50 kbp nicht überschreiten und daher für diese Werte eine diskrete Linie parallel zur Ordinate bilden, welche in der rechten Abbildung fehlt. Die rechte Abbildung zeigt die dOCR, die wiederum kein direktes oberes Limit haben. Für dOCR und den 0-5 kbp Downstreambereich finden sich zahlreiche Datenpunkte direkt auf der Ordinate, die jene Gene darstellen, welche im 0-5 kbp Downstreambereich keine Peaks aufwiesen. Dies trat im 0-50 kbp Downstreambereich nicht auf.

am geringsten.

Bemerkenswerterweise korrelierten die ATAC-seq- und 4sU-seq-RPKM unstratifiziert geringer als nach Read-through stratifiziert, wobei sich zwischen den Read-through-Kategorien (0–20 %, 20–40 %, 40–60 %, 60–80 %, >80 % Read-through) kein Trend abzeichnete. Darüber hinaus korrelierten die ATAC-seq-Daten nicht mit dem prozentualen Read-through. Betrachtete man nur die Gene mit Read-through über 80 %, so betrug die Spearman Korrelation zwischen den absoluten ATAC-seq-RPKM und dem prozentualem Read-through $R_s \approx 0.02$. Betrachtete man nur die Gene mit Read-through unter 20 %, so betrug die Spearman Korrelation zwischen den absoluten ATAC-seq-RPKM und dem prozentualem Read-through $R_s \approx 0.08$. Für die hier gezeigten Korrelationen galt $p<0.001$.

Der Read-through ist daher kein geeignetes Maß, um DoTT mit OCRs im Downstreambereich in Verbindung zu bringen. Stattdessen eignen sich die RPKM-Werte im Downstreambereich. Nicht der prozentuale Anteil der mRNAs, die von DoTT betroffen sind, sondern die absolute Transkriptionsaktivität bestimmt also das Ausmaß von dOCR-Induktion in den Downstreambereichen von Genen.

3.2.3. Read-through, ATAC-seq-RPKM und 4sU-seq-RPKM für Gene mit dOCR größer als 110 kbp.

Nachdem in Kap. 3.2.1 keine Korrelation zwischen Read-through und dOCR gefunden wurde und in Kap. 3.2.2 festgestellt wurde, dass sich absolute RPKM-Werte am ehesten dazu eignen, eine Korrelation zwischen OCRs und Read-through herzustellen, erfolgte als Nächstes eine Subgruppenanalyse mit dem Ziel der Maximierung dieser Korrelation. Im Anschluss kann diese Subgruppe als Stichprobe dienen, um manuell im Genom-Viewer nach Gemeinsamkeiten und Besonderheiten dieser Gene zu suchen. Der Korrelationskoeffizienten zwischen ATAC-seq-RPKM und Read-through war für Gene ab einem dOCR von mehr als 110 kbp am höchsten (Spearman Korrelation $R_s \approx 0.21$). Diese Korrelation bestand im 5 kbp Downstreambereich für Gene mit einem dOCR von größer als 15 kbp (373 Gene) und erreichte sein Maximum zwischen einem dOCR von 100 kbp und 120 kbp Länge (133 Gene). Für diese Teilmenge an Genen zeigte sich eine erhöhte Korrelation zwischen ATAC-seq-RPKM bzw. 4sU-seq-RPKM (Spearman Korrelation für Differenzen $R_s \approx 0.59$) (s. Abb. 3.13).
Aufgrund der geringen Anzahl von Genen in den einzelnen Read-through-Kategorien fielen die 95 %-KI groß aus (n = 31 (30, 25, 22, 25) Gene für einen Read-through von 0-20 (20-40, 40-60, 60-80, >80) Prozentpunkten bei einem dOCR (8 h p.i.) > 110 kbp (s. Kap. 1.4.4)).

Optimierte man nicht anhand der dOCR, sondern anhand des Downstreambereiches, in welchem RPKM gezählt wurden, stellte sich bei 0-75 kbp eine Korrelation von $R_s \approx 0.04$ zwischen ATAC-seq-RPKM und Read-through ein (3684 Gene). Für ATAC-seq-RPKM und 4sU-seq-RPKM betrug dann die Spearman Korrelation $R_s \approx 0.39$. Optimierte man sowohl anhand der dOCR als auch anhand des betrachteten Downstreambereiches, stellte sich bei dOCR größer als 110 kbp und 0-75 kbp betrachtetem Downstreambereich eine Korrelation von $R_s \approx 0.34$ zwischen ATAC-seq-RPKM und Read-through ein (133 Gene). In diesem Fall betrug die Spearman Korrelation zwischen ATAC-seq-RPKM und 4sU-seq-RPKM $R_s \approx 0.59$.

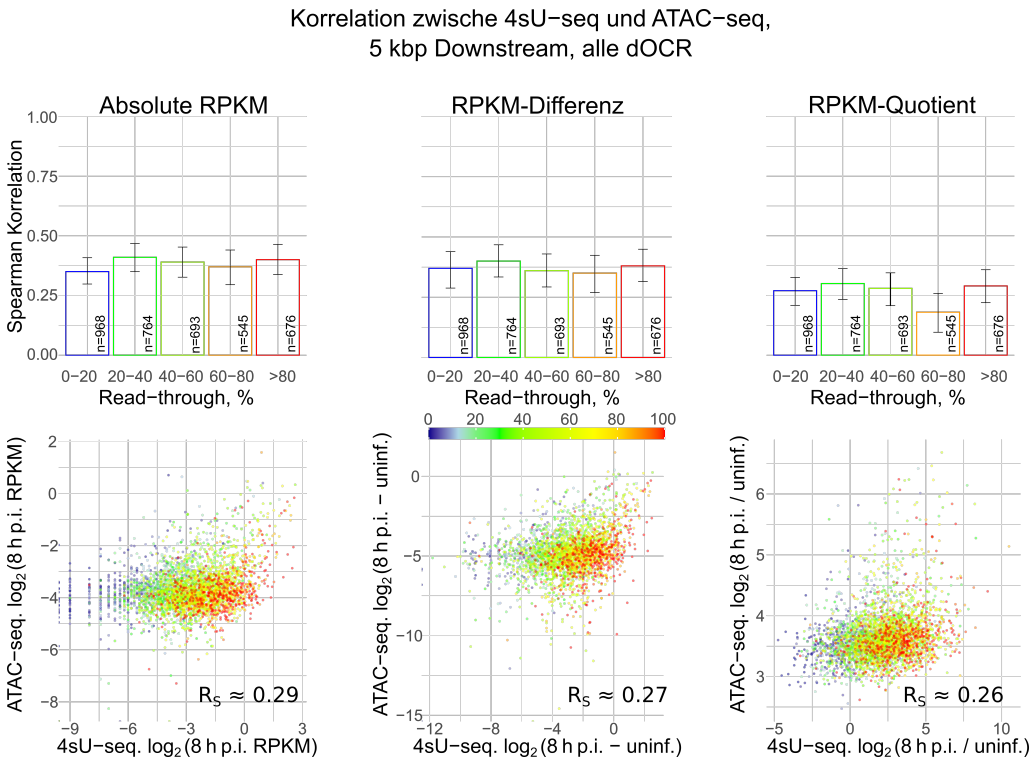


Abbildung 3.12.: Korrelation zwischen ATAC-seq-RPKM und 4sU-seq-RPKM, nach Read-through stratifiziert, im 0-5 kbp Downstreambereich, für Gene mit dOCR (8 h p.i.) ≥ 0 kbp .
Balkendiagramme: Korrelation zwischen 4sU-seq und ATAC-seq, stratifiziert nach dem prozentualen Read-through in 20 %-Schritten für n = 3684 Gene. Zudem ist das 95 %-KI gezeigt. Von links nach rechts handelt es sich um die absoluten RPKM-Werte, die Differenz der absoluten RPKM-Werte zwischen 8 h p.i. und uninfizierten Zellen und den $\log_2$-fold-change zwischen 8 h p.i. und uninfizierten Zellen.
Streudiagramme: Jeder Datenpunkt symbolisiert eines der 3684 Gene. Auf der Ordinate wurden die ATAC-seq-Werte der entsprechenden Gene aufgetragen, auf der Abszisse die 4sU-seq-Werte.
Von links nach rechts handelt es sich um die absoluten RPKM-Werte (Spearman Korrelation $R_s \approx 0.29$), die Differenz der absoluten RPKM-Werte ($R_s \approx 0.27$) zwischen 8 h p.i. und uninfizierten Zellen und den $\log_2$-fold-change ($R_s \approx 0.26$) zwischen 8 h p.i. und uninfizierten Zellen.
Der prozentuale Read-through korrelierte nicht mit den ATAC-seq-Werten, auch dann nicht, wenn man nach Read-through-Werten stratifizierte.
Die Einfärbung kodiert den Read-through. Gene mit einem Read-through von über 100 % werden mit 100 %-Read-through dargestellt.

Betrachtete man Stichproben im Genom-Viewer, zeigte sich, dass es sich trotz der

in Kap. 2.1.4 beschriebenen Vorselektion bei dieser Subgruppe von Genen um Gene in Genclustern handelt (s. Abb. 3.29). Gencluster sind ein Confounder für erhöhte dOCR, Read-through, ATAC-seq-RPKM und 4sU-seq-RPKM in Downstreambereichen.

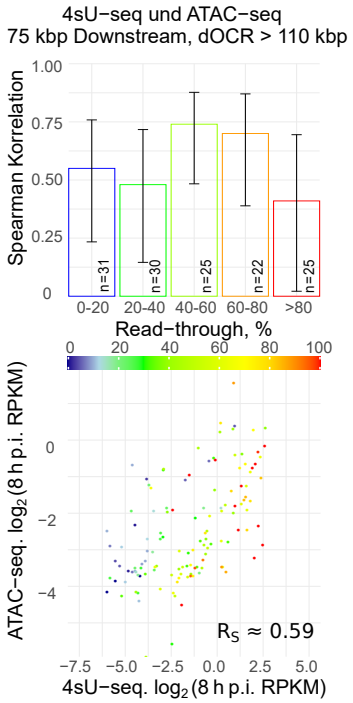


Abbildung 3.13.: Korrelation zwischen ATAC-seq-RPKM und 4sU-seq-RPKM, nach Read-through stratifiziert, im 0-75 kbp Downstreambereich, für Gene mit dOCR (8 h p.i.) > 110 kbp.
Balkendiagramme: Korrelation zwischen 4sU-seq und ATAC-seq, stratifiziert nach dem prozentualen Read-through in 20 %-Schritten für n = 133 Gene. Zudem ist das 95 %-KI gezeigt. Es handelt sich um die absoluten RPKM-Werte.
Streudiagramme: Jeder Datenpunkt symbolisiert eines von 133 Genen. Auf der Abszisse wurden die absoluten RPKM für 4sU-seq aufgetragen. Auf der Ordinate wurden die absoluten RPKM für ATAC-seq aufgetragen (Spearman Korrelation $R_s \approx 0.59$).
Die Spearman Korrelation zwischen ATAC-seq und Read-through betrug $R_s \approx 0.21$. Die Mindestlänge von 110 kbp für dOCR resultierte aus einer Optimierung auf diesen Korrelationskoeffizienten. Der Read-through wurde farblich kodiert. Gene mit einem Read-trough von über 100 % werden als 100 %-Read-through dargestellt.

3.3. Analyse der DoTT und OCRs bei lytischer HSV-1 Infektion

Nachdem sich in Kap. 3.2.2 die ATAC-seq- und 4sU-seq-RPKM-Werte als vorteilhaft erwiesen haben, um Zusammenhänge im Downstreambereich zu erkennen, wurden darauf aufbauend verschiedene Downstreambereiche untersucht. Dadurch können Erkenntnisse gewonnen werden, wie weit die Auswirkungen von DoTT in den Downstreambereich reichen. Es wurde das in Kap. 2.1.4 beschriebene Subset von 3684 Genen betrachtet.

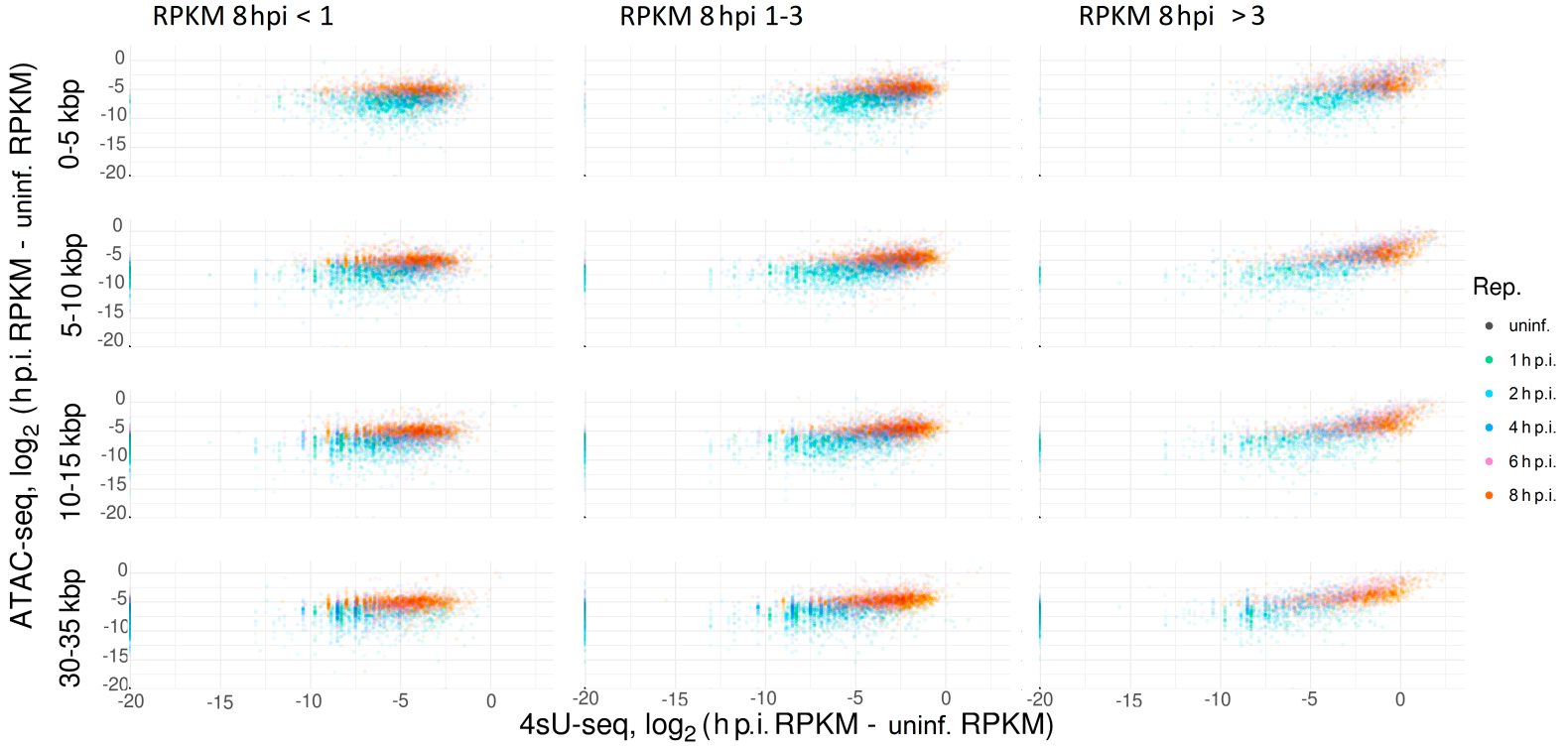
3.3.1. Analyse mit Augenmerk auf die Transkriptionsrate 8 h p.i.

Die Daten wurden mit Hilfe von PASsUS wie in Kap. 2.3 beschrieben geplottet. Die Betrachtung anhand von absoluten RPKM (s. Abb. 3.14), von RPKM-Differenzen (s. Abb. 3.16) und von RPKM-Quotienten (s. Abb. 3.18) führten qualitativ in mehreren Punkten zu den gleichen Ergebnissen:
Erhöhte RPKM im Downstreambereich, welche auf DoTT und OCR schließen lassen, waren 8 h p.i. besonders stark bei Genen mit einer Transkriptionsrate von über 3 RPKM ausgeprägt (s. Abb. 3.14 – 3.19). Dies zeigte sich im Downstreambereich in unmittelbarer Nähe zu Genen (0-5 kbp) und verlierte sich in weiter Ferne (200-205 kbp), wobei bei Letzterem die Interpretierbarkeit eingeschränkt ist, da sich hierbei der Downstreambereich über andere Gene erstreckte. Durch diese Überlappung kam es auch zu Genen, bei denen der errechnete $\log_2$-fold-change für DoTT und OCR 8 h p.i. geringer ist als 0 h p.i. (s. auch Kap. 4.2.2).
Von der Verwendung des no-overlap-Modus von PASsUS (Details siehe Ende Kap. 2.3) wurde abgesehen, da bereits durch die Vorauswahl nur Gene berücksichtigt wurden, in denen in der 5 kbp Umgebung auf dem selben Strang keine anderen Gene liegen. Bemerkenswert ist jedoch, dass für den 200-205 kpb Downstreambereich nur noch 29 Gene einen Downstreambereich aufwiesen, der kein anderes Gen schnitt. Daher ist dieser Bereich kritisch zu betrachten und hängt vermutlich eher mit zufälligen und generellen Veränderungen der Chromatinorganisation als mit dem Ausgangsgen zusammen.

Abbildung 3.14.: DoTT und OCR in absoluten RPKM in Abhängigkeit von der Transkriptionsrate 8 h p.i. und den Downstreambereichen (Teil 1). Die Graphen zeigen je Zeile einen anderen Downstreambereich und je Spalte eine andere basale Transkriptionsrate gemessen in RPKM 8 h p.i.. Jeder Punkt steht hierbei für die entsprechenden Werte eines Gens zu einem bestimmten Zeitpunkt p.i. (grün 1 h p.i., rot 8 h p.i.). Damit ist jedes Gen für jeden Zeitpunkt jeweils einmal vertreten. Bei den einzelnen Graphen wurde auf der Abszisse die RPKM-Werte der 4sU-seq-Reads zu verschiedenen Zeitpunkten nach der Infektion als Maß für die DoTT aufgetragen. Analog dazu wurden auf der Ordinate die ATAC-seq-Daten als Maß für die Offenheit des Chromatins abgebildet.

Abbildung 3.15.: DoTT und OCR in absoluten RPKM in Abhängigkeit von der Transkriptionsrate 8 h p.i. und den Downstreambereichen (Teil 2). Der Vergleich der Spalten zeigt, dass DoTT und OCR mit stark transkribierten Genen korrelieren. Dies zeigt sich vorwiegend im Downstreambereich in unmittelbarer Nähe zum Gen (0-5 kbp) und weniger in weiter Ferne (200-205 kbp), wobei bei Letzterem die Interpretierbarkeit eingeschränkt ist, da sich hierbei der Downstreambereich über andere Gene erstreckt.

Abbildung 3.16.: DoTT und OCR in RPKM-Differenzen in Abhängigkeit von der Transkriptionsrate 8 h p.i. und den Downstreambereichen (Teil 1). Die Graphen zeigen je Zeile einen anderen Downstreambereich und je Spalte eine andere basale Transkriptionsrate gemessen in RPKM 8 h p.i.. Jeder Punkt steht hierbei für die entsprechenden Werte eines Gens zu einem bestimmten Zeitpunkt p.i. (grün 1 h p.i., rot 8 h p.i.). Damit ist jedes Gen für jeden Zeitpunkt jeweils einmal vertreten. Bei den einzelnen Graphen wurde auf der Abszisse die Differenz aus der Summe der 4sU-seq-Reads für uninfizierte Zellen subtrahiert von der Summe der 4sU-seq-Reads zu den verschiedenen Zeitpunkten post infectionem als $\log_2$-fold-change aufgetragen. Vereinfacht gesagt handelt es sich hierbei um ein Maß für die DoTT. Analog dazu wurden auf der Ordinate die ATAC-seq-Daten als Maß für die Offenheit des Chromatins aufgetragen.



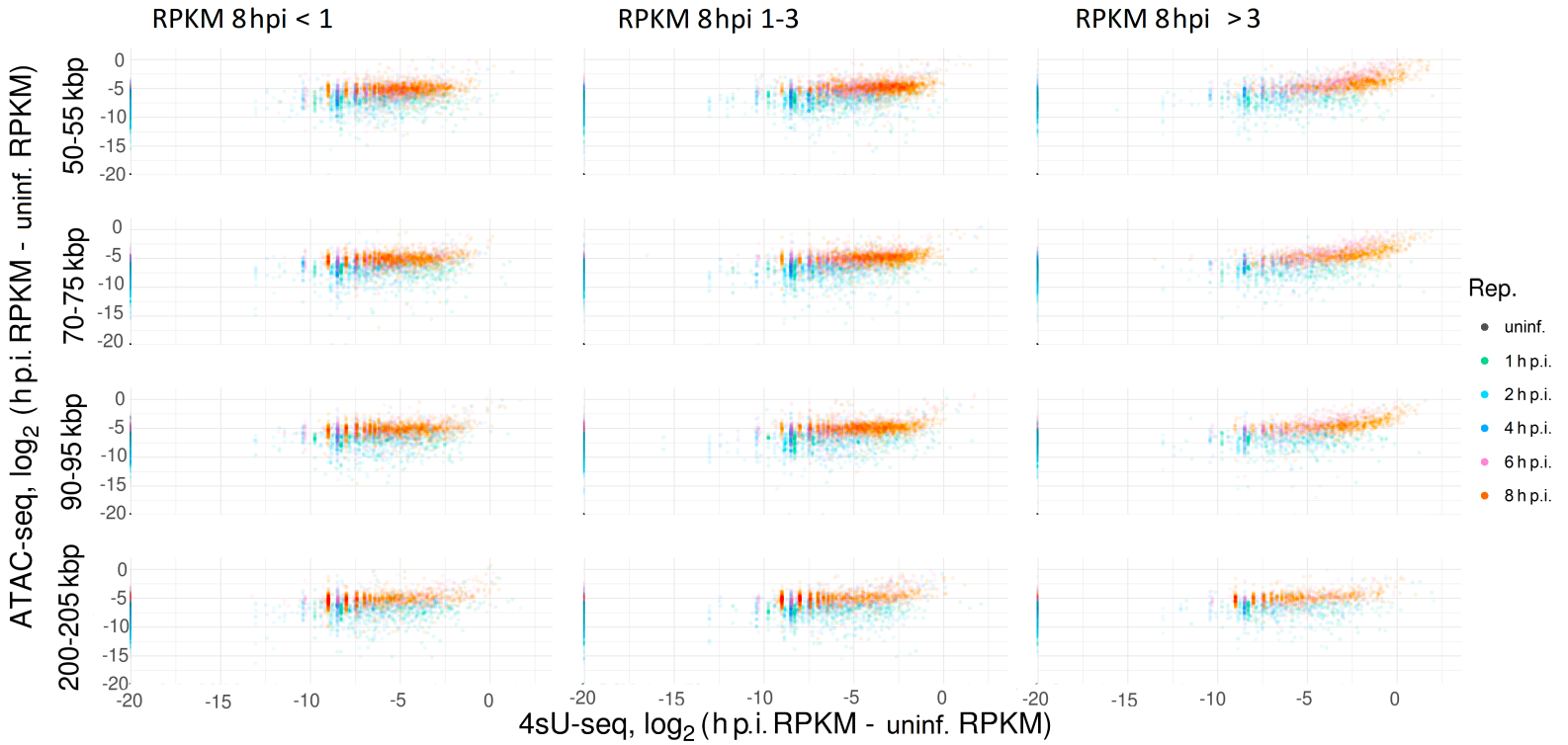
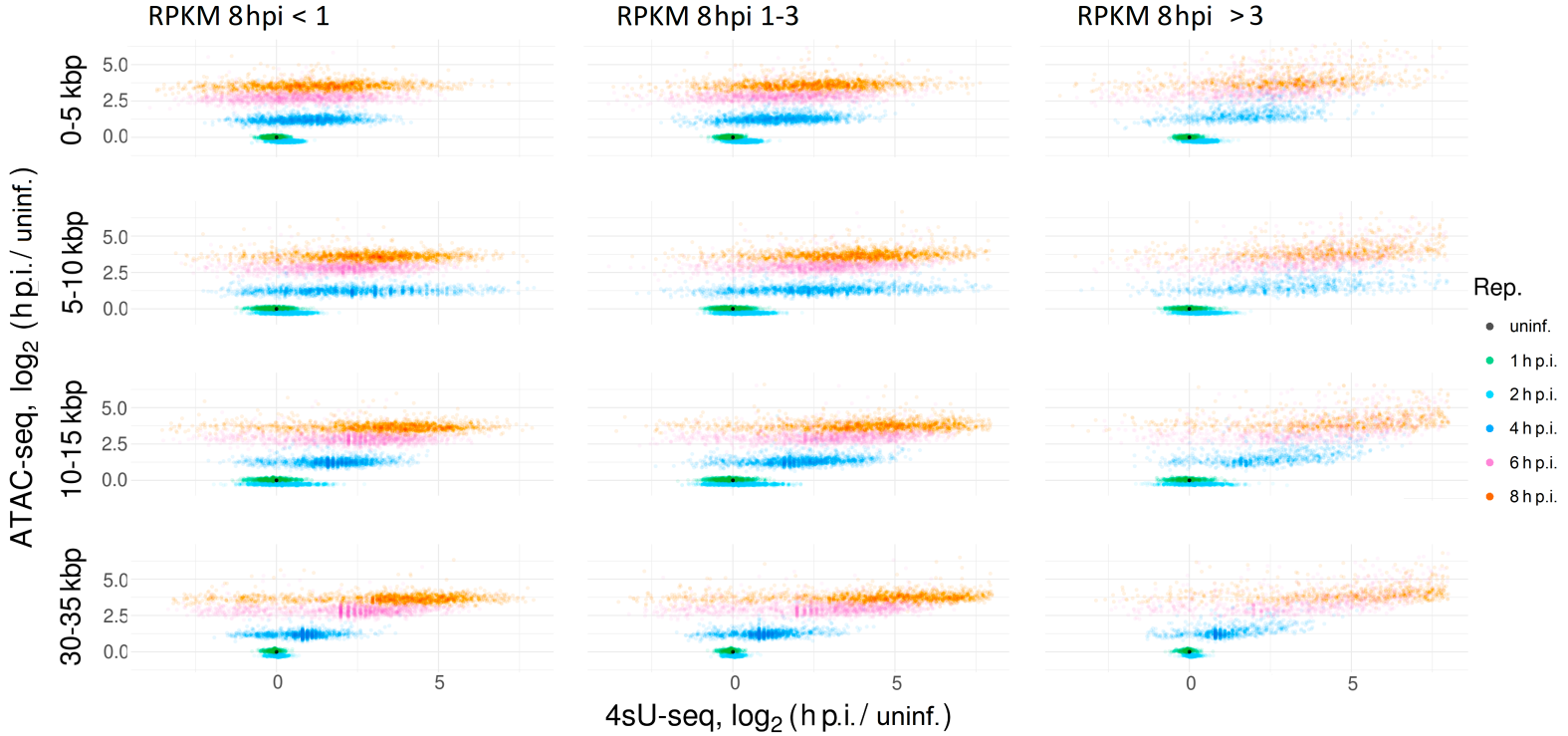


Abbildung 3.17.: DoTT und OCR in RPKM-Differenzen in Abhängigkeit von der Transkriptionsrate 8 h p.i. und den Downstreambereichen (Teil 2). Der Vergleich der Spalten zeigt, dass DoTT und OCR mit stark transkribierten Genen korrelieren. Dies zeigt sich vorwiegend im Downstreambereich in unmittelbarer Nähe zum Gen (0-5 kbp) und weniger in weiter Ferne (200-205 kbp), wobei bei Letzterem die Interpretierbarkeit eingeschränkt ist, da sich hierbei der Downstreambereich über andere Gene erstreckt.

Abbildung 3.18.: DoTT und OCR in RPKM-Quotienten in Abhängigkeit von der Transkriptionsrate 8 h p.i. und den Downstreambereichen (Teil 1). Die Graphen zeigen je Zeile einen anderen Downstreambereich und je Spalte eine andere basale Transkriptionsrate gemessen in RPKM 8 h p.i.. Jeder Punkt steht hierbei für die entsprechenden Werte eines Gens zu einem bestimmten Zeitpunkt p.i. (grün 1 h p.i., rot 8 h p.i.). Damit ist jedes Gen für jeden Zeitpunkt jeweils einmal vertreten. Bei den einzelnen Graphen wurde auf der Abszisse die Quotienten aus der Summe der 4sU-seq-Reads zu verschiedenen Zeitpunkten nach der Infektion dividiert durch die Summe der 4sU-seq-Reads zum Zeitpunkt 0 h p.i. als $\log_2$-fold-change aufgetragen. Vereinfacht gesagt handelt es sich hierbei um ein Maß für die DoTT. Analog dazu wurden auf der Ordinate die ATAC-seq Daten als Maß für die Offenheit des Chromatins angegeben.



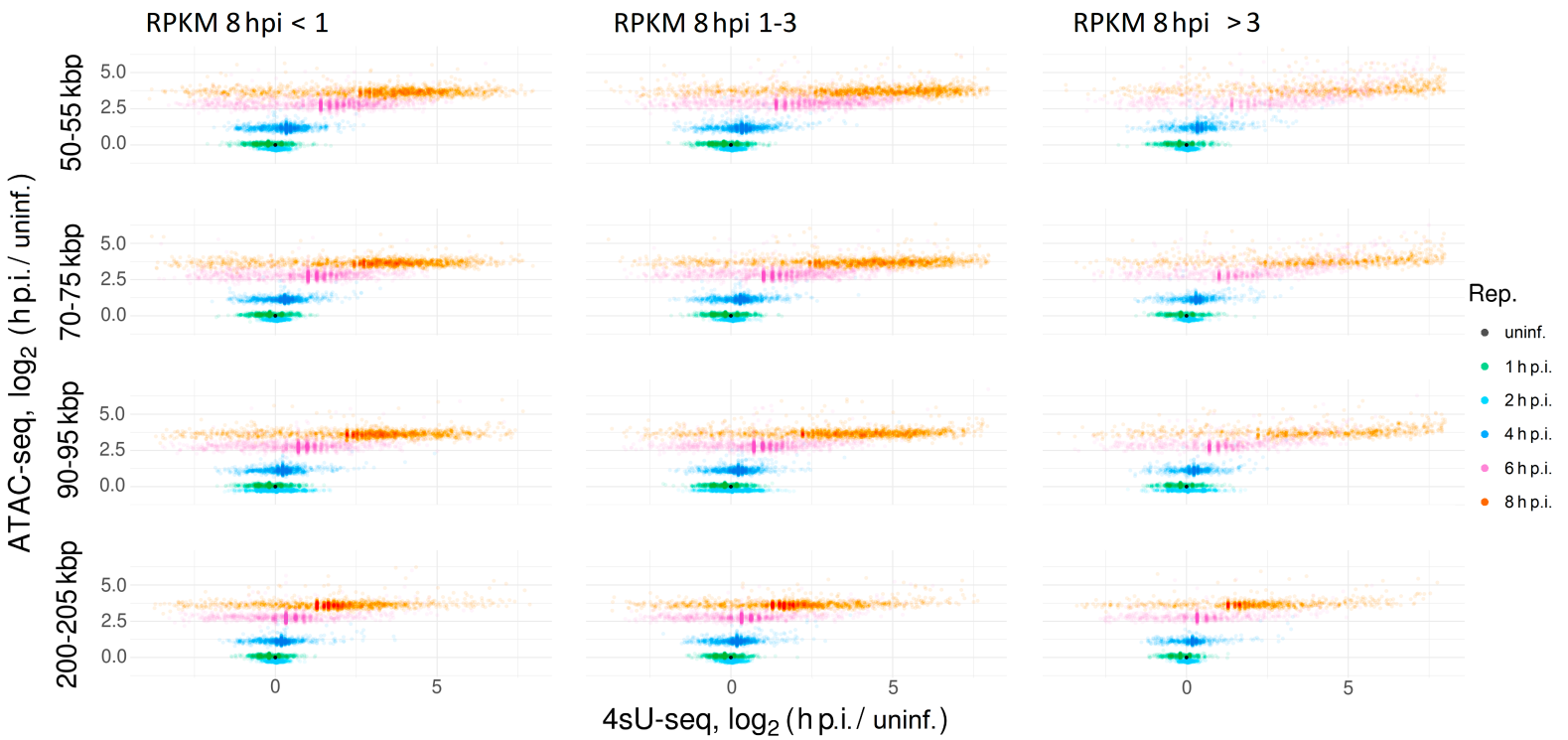


Abbildung 3.19.: DoTT und OCR in RPKM-Quotienten in Abhängigkeit von der Transkriptionsrate 8 h p.i. und den Downstreambereichen (Teil 2). Der Vergleich der Spalten zeigt, dass DoTT und OCR mit stark transkribierten Genen korrelieren. Dies zeigt sich vorwiegend im Downstreambereich in unmittelbarer Nähe zum Gen (0-5 kbp) und weniger in weiter Ferne (200-205 kbp), wobei bei Letzterem die Interpretierbarkeit eingeschränkt ist, da sich hierbei der Downstreambereich über andere Gene erstreckt.

Die Betrachtung anhand von absoluten RPKM wies eine höhere Varianz auf als die Betrachtung anhand von RPKM-Differenzen. Dies liegt daran, dass bei der Bildung von Differenzen bei ähnlichem Betrag von Minuend und Subtrahend geringe Werte resultieren. Damit tendierten die Differenzen zur Mitte hin und streuten weniger, als die absoluten RPKM. Die Betrachtung anhand von RPKM-Quotienten wies eine höhere Trennschärfe im Bezug auf die Zeit nach Infektion auf, als die anderen Betrachtungsweisen. Dies liegt an der Berechnung von Pseudocounts durch LFC. Außerdem lagen bei den RPKM-Quotienten die Maxima von ATAC-seq-RPKM und 4sU-seq-RPKM nicht im unmittelbaren Downstreambereich, sondern nahmen mit Abstand zu Genen zu. Die Ergebnisse der RPKM-Differenzen und RPKM-Quotienten waren sich auch anhand vom Spearman-Korrelationskoeffizient sehr ähnlich. Anhand der ATAC-seq Daten z. B. betrug er im 0-5 kbp Downstream $R_s \approx 0{,}867$ (95 % KI 0,859-0,875), im 5-10 kbp Downstreambereich $R_s \approx 0{,}849$ (95 % KI 0,839-0,857) und im 10-15 kbp Downstreambereich $R_s \approx 0{,}844$ (95 % KI 0,834-0,857).

3.3.2. Analyse mit Augenmerk auf die verschiedenen Downstreambereiche

Durch Mitteln der RPKM-Werte aller Gene eines bestimmten Downstreambereiches zu einem bestimmten Zeitpunkt nach Infektion lassen sich Unterschiede zwischen den verschiedenen Downstreambereichen weiter veranschaulichen, besser abgrenzen und kompakter darstellen (s. Abb. 3.14 – 3.19). Während bei den absoluten RPKM-Werten und Differenzen der Effekt für DoTT und OCR im unmittelbaren Downstreambereich am stärksten zu sein schien, nahm dieser bei den Quotienten bis zu dem 30 kbp–35 kbp Downstreambereich zu.
Bei den absoluten RPKM-Werten und Differenzen dieser stimmte außerdem überein, dass 2 h p.i. keine nennenswerte DoTT oder OCR vorlagen. Das Chromatin erreichte 6 h p.i. seine maximale Offenheit, während die DoTT noch bis 8 h p.i. zunahm. Bei den Quotienten konnten 2 h p.i. leicht gesteigerte 4sU-seq-Werte festgestellt werden. Die unterschiedlichen Sättigungseffekte von DoTT und OCR waren weniger stark ausgeprägt.
Für alle Betrachtungsweisen stimmte überein, dass OCR auch noch im 200 kbp Downstreambereich auftritt, während DoTT hier keine bzw. kaum Effekte zeigte.

Diese Darstellung ermöglicht es, den Verlauf von DoTT und OCRs über die Zeit und die Entfernung in den Downstreambereich abzubilden und vergleichbar zu ma-

chen. Daher wurde die Darstellung im Weiteren zur Untersuchung des Hitze- und Salzstresses verwendet.

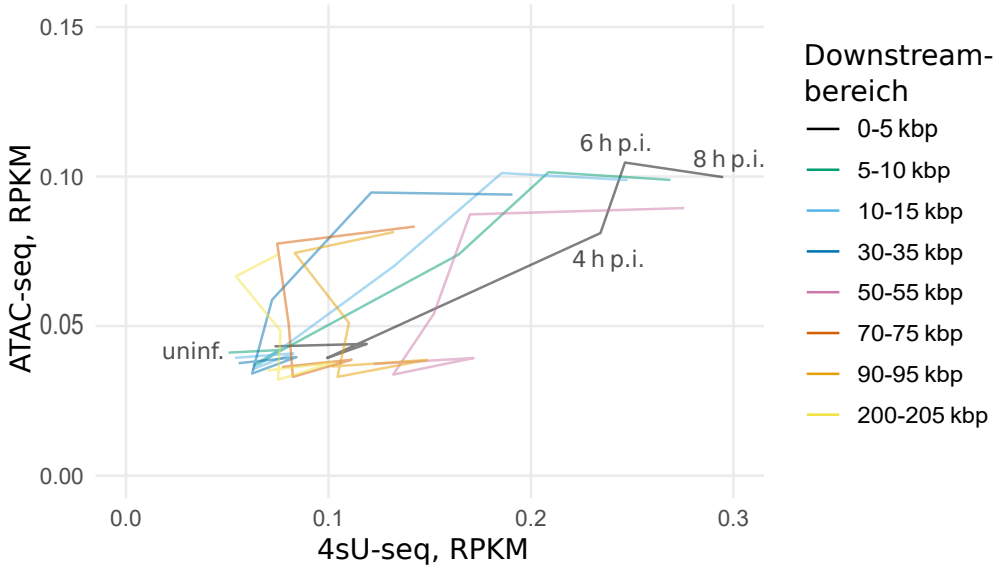


Abbildung 3.20.: 4sU-seq und ATAC-seq in absoluten RPKM in Abhängigkeit vom Zeitpunkt p.i. und den Downstreambereichen über alle Gene gemittelt.
Auf der Abszisse wurde der Mittelwert aller 4sU-RPKM zu verschiedenen Zeitpunkten nach der Infektion aufgetragen. Vereinfacht gesagt handelt es sich hierbei um ein über alle betrachteten Gene gemitteltes Maß für die DoTT.
Analog dazu wurden auf der Ordinate die ATAC-seq-Werte als Maß für die Offenheit des Chromatins aufgetragen. Der Graph enthält 8 verschiedenfarbige Kurven für die jeweiligen Downstreambereiche. Jede Kurve besteht aus 6 Knotenpunkten, die den Mittelwerten der RPKM-Werte von links nach rechts $0, 1, 2, 4, 6$ und 8 h p.i. entsprechen.
Die Offenheit des Chromatins nahm im Downstreambereich 8 h p.i. im Vergleich zu 6 h p.i. nicht weiter zu. DoTT und OCRs scheinen im 0-5 kbp Downstreambereich am stärksten ausgeprägt gewesen zu sein und nahmen mit zunehmendem Abstand ab, wobei OCR weniger abnahm.

Aus dem vorherigen Kapitel und diesem Kapitel geht hervor, dass die Betrachtungsweisen der absoluten RPKM, RPKM-Differenzen und RPKM-Quotienten sich in dem Subset aus 3684 Genen qualitativ sehr ähnlich sind. Daher wurden für die Analyse der DoTT und OCRs bei Hitze- und Salzstress nur die Differenzen der RPKM dargestellt.

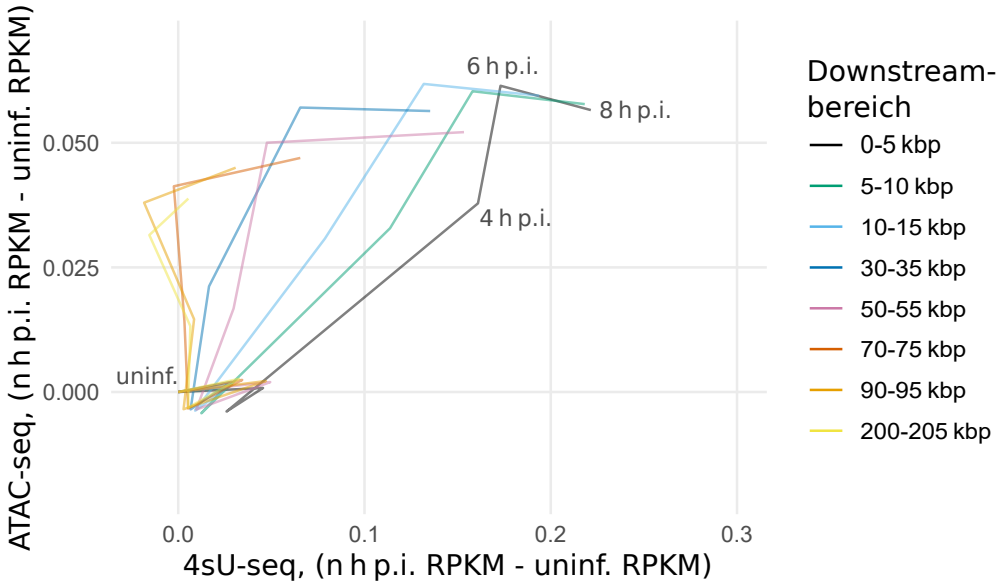


Abbildung 3.21.: 4sU-seq und ATAC-seq RPKM-Differenzen in Abhängigkeit vom Zeitpunkt p.i. und den Downstreambereichen über alle Gene gemittelt.
Auf der Abszisse wurde der Mittelwert aller RPKM-Differenzen der 4sU-Reads zum Zeitpunkt 0 h p.i. subtrahiert von den 4sU-Reads zu verschiedenen Zeitpunkten p.i. aufgetragen. Vereinfacht gesagt handelt es sich hierbei um ein über alle betrachteten Gene gemitteltes Maß für die DoTT.
Analog dazu wurden auf der Ordinate die ATAC-seq-Werte als Maß für die Offenheit des Chromatins aufgetragen.
Der Graph enthält 8 verschiedenfarbige Kurven für die jeweiligen Downstreambereiche. Jede Kurve besteht aus 6 Knotenpunkten, die den Mittelwerten der RPKM-Werte von links nach rechts 0, 1, 2, 4, 6 und 8 h p.i. entsprechen. Da 2 h p.i. keine nennenswerte DoTT oder OCRs vorliegen, lassen sich die Knotenpunkte erst 4 h p.i. unterscheiden.
Die Offenheit des Chromatins nahm im Downstreambereich 8 h p.i. im Vergleich zu 6 h p.i. nicht weiter zu. DoTT und OCRs schienen im 0-5 kbp Downstreambereich am stärksten ausgeprägt gewesen zu sein und nahmen mit zunehmendem Abstand ab, wobei OCRs weniger abnahmen.

3.4. Analyse der DoTT und OCRs bei Hitze- und Salzstress

Im Folgenden wurden die Erkenntnisse aus den vorhergegangenen Kapiteln auf Hitze- und Salzstress übertragen.

Da in Bezug auf von Peak-Callern abgeleiteten Daten die größte Trennschärfe zwi-

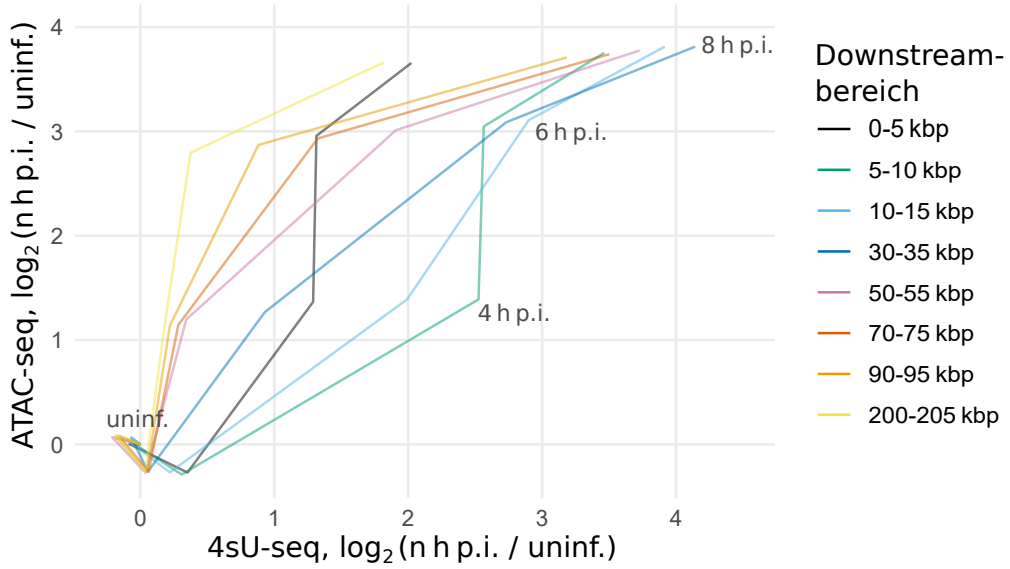


Abbildung 3.22.: 4sU-seq und ATAC-seq in RPKM-Quotienten in Abhängigkeit vom Zeitpunkt p.i. und den Downstreambereichen über alle Gene gemittelt.
Auf der Abszisse wurde der Mittelwert aller $\log_2$-fold-changes aus dem Quotienten der Summe der 4sU-seq-Reads zu verschiedenen Zeitpunkten nach der Infektion geteilt durch die Summe der 4sU-seq-Reads zum Zeitpunkt 0 h p.i. aufgetragen. Vereinfacht gesagt handelt es sich hierbei um ein über alle betrachteten Gene gemitteltes Maß für die DoTT.
Analog dazu wurden auf der Ordinate die ATAC-seq-Werte als Maß für die Offenheit des Chromatins aufgetragen.
Der Graph enthält 8 verschiedenfarbige Kurven für die jeweiligen Downstreambereiche. Jede Kurve besteht aus 6 Knotenpunkten, die den Mittelwerten der $\log_2$-fold-changes für von links nach rechts $0, 1, 2, 4, 6$ und 8 h p.i. entsprechen.
Da 2 h p.i. keine nennenswerte DoTT oder OCRs vorlagen, lassen sich die Eckpunkte erst 4 h p.i. unterscheiden. Die Offenheit des Chromatins schien 8 h p.i. im Vergleich zu 6 h p.i. bereits wieder abzunehmen.
DoTT und OCRs schienen im 30-35 kbp Downstreambereich am stärksten ausgeprägt zu sein und nahmen im 50-55 kbp Downstreambereich bereits ab, wobei ihre Auswirkungen auch noch im 200-205 kbp Downstreambereich zu erkennen waren.

schen uninfizierten Zellen und 8 h p.i. Replikaten durch dOCR-Berechnung mittels F-Seq erreicht wurde, wurde dieses Verfahren auf Hitze und Salzstress übertragen (s. Abb. 3.23). Es zeigte sich wie auch schon von Hennig et al. (2018) beschrieben keine Erhöhung der Offenheit des Chromatins für sowohl Hitzestress als auch für Salzstress. [22]

Ein ähnliches Bild zeigte sich bei Betrachtung der aus RPKM-Differenzen gewonnenen Daten (s. Abb. 3.24): Beim Hitzestress waren die 4sU-seq-RPKM als Maß für DoG 2 h p.i. analog zu HSV-1-Infektion 8 h p.i. ausgeprägt, während die ATAC-seq-RPKM als Maß für OCR nicht erhöht waren. Dies zeigt, dass die Erhöhung der Chromatinoffenheit keine Vorraussetzung für DoG ist, wie es auch schon Hennig et al. (2018) anhand von dOCR beschrieben hat [22]. Der Betrag der 4sU-seq-RPKM bei Salzstress als Maß für DoG war 2 h p.i. rund halb so groß wie bei HSV-1-Infektion 8 h p.i.; der Betrag der ATAC-seq-RPKM bei Salzstress als Maß für OCR war 1 h p.i. weniger als ein Fünftel so groß wie bei HSV-1-Infektion 8 h p.i. und fiel 2 h post interventionem wieder ab. Ein Abfallen der ATAC-seq-RPKM mit zunehmender Entfernung vom 3'-Ende fehlte (Vergleich Abb. 3.23 und 3.21).

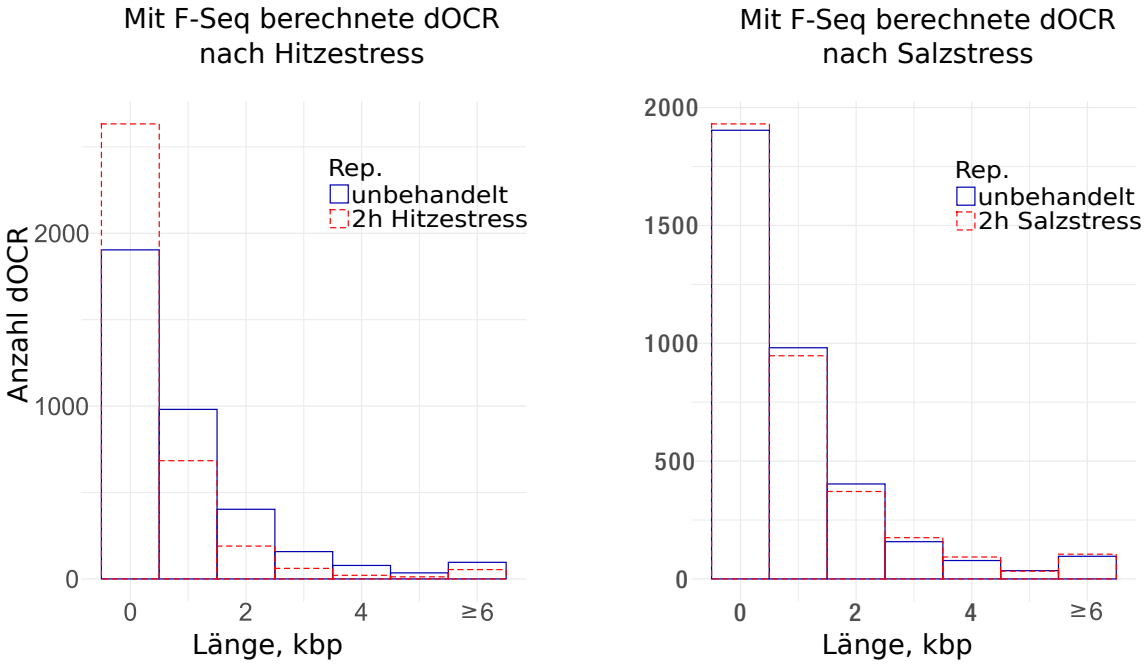


Abbildung 3.23.: Längenverteilung der dOCR-Längen für Hitze- und Salzstress, 3684 Gene berücksichtigt.
Auf den Abszissen wurden die entsprechenden Längen der dOCR angegeben, wobei im letzten Balken alle dOCR-Längen zusammengetragen wurden, die länger als 6 kbp sind. Auf den Ordinaten wurde die Anzahl der dOCR-Längen aufgetragen, welche eine bestimmte Länge überschritten.

Mit PASsUS wurde ein Programm geschaffen, dass diese Art der Auswertung weitgehend automatisiert. Durch die Betrachtung von RPKM sind sensitive Untersuchungen möglich, welche die Ergebnisse von dOCR validieren.

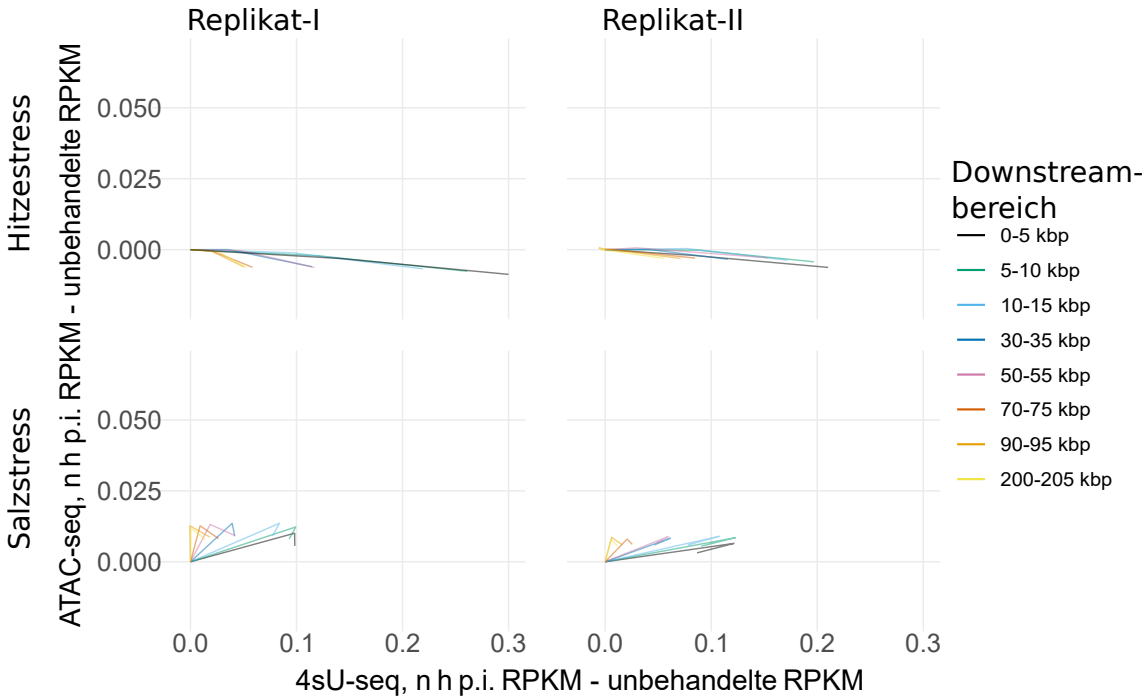


Abbildung 3.24.: DoG und OCR in RPKM-Differenzen in Abhängigkeit vom Zeitpunkt p.i. und den Downstreambereichen über alle Gene gemittelt für Hitzestress (oben) und Salzstress (unten), Replikat I und II.

Auf der Abszisse wurde der Mittelwert aller RPKM-Differenzen der 4sU-seq-Reads zum Zeitpunkt 0 h p.i. subtrahiert von den 4sU-seq-Reads zu verschiedenen Zeitpunkten p.i. aufgetragen. Vereinfacht gesagt handelt es sich hierbei um ein über alle betrachteten Gene gemitteltes Maß für die DoG.
Analog dazu wurden auf der Ordinate die ATAC-seq-Werte als Maß für die Offenheit des Chromatins aufgetragen.
Der Graph enthält 8 verschiedenfarbige Kurven für die jeweiligen Downstreambereiche. Jede Kurve besteht aus 3 Knotenpunkten, die den Mittelwerten der RPKM-Werte von links nach rechts 0, 1 und 2 h p.i. entsprechen.
Beachtet man die Skala für ATAC-seq, fällt auf, dass die Offenheit des Chromatins nur unwesentlich beeinflusst war, während für 4sU-seq-Daten ein ähnlicher Trend wie bei HSV-I bestand.

3.5. Selektierte Beispiele im Genom Viewer

Die Abbildungen zu selektierten Beispielen im Genom Viewer befinden sich im Anschluss an dieses Kapitel (s. Abb. 3.25 – 3.29). Der Einfachheit halber wurde jedes Beispiel weitestgehend identisch aufgebaut.
Zuerst wird der Plusstrang dargestellt.
Die Abbildungen enthalten von oben nach unten zunächst Informationen über den genau gezeigten Bereich, bestehend aus dem Chromosom, der Region und der dort beinhalteten Gen-Regionen und Gene.
Danach folgen die ATAC-seq-Reads mit den jeweilig annotierten Peaks für die drei uninfizierten Replikate der HSV-I-Infektion, so wie die gepoolten ATAC-seq-Reads für Zellstress zum ungestressten Zeitpunkt (falls dargestellt). Es folgen die 4sU-seq-Reads der zwei uninfizierten Replikate und die gepoolten 4sU-Reads für Zellstress zum ungestressten Zeitpunkt.
Dies wiederholt sich für die Daten post interventionem, wobei hier Hitze- und Salzstress getrennt voneinander dargestellt werden (falls dargestellt).
Zuletzt wiederholt sich all dies für den Minusstrang. Bei Betrachtungen von Genen auf dem Plusstrang werden die Reads auf dem Minusstrang nicht dargestellt und vice versa.
Das Gen, auf welches in der jeweiligen Abbildung das Augenmerk liegt, ist in den Abbildungen links oben zu finden, wenn es auf dem Plusstrang liegt und rechts unten, wenn es auf dem Minusstrang liegt.

3.5.1. SRSF3 und SRSF6

„Serine And Arginine Rich Splicing Factor 3 bzw. 6“ (SRSF3 bzw. SRSF6) sind Paradebeispiele für Read-through, da sie diesen im großen Maße aufweisen und nicht in Genclustern liegen (s. Abb. 3.25 und 3.26).

Für sowohl SRSF3 als auch SRSF6 lies sich zeigen: Bei den ATAC-seq-Reads der uninfizierten Zellen waren Peaks mit hohem Score in Genen annotiert, die teilweise mit den transkribierten Bereichen zusammenfielen. Weiterhin waren außerhalb der Genregionen Peaks annotiert, von denen die Größten entweder mit Genen auf dem Minusstrang oder mit Promotoren und anderen Kontrollsequenzen zusammenfielen. Manche der annotierten Peaks fielen jedoch in das Rauschen.
8 h p.i. lies sich in den Downstreambereichen ein großer Anstieg der 4sU-Reads für HSV-I-Infektion und ein geringerer Anstieg für Hitze- und Salzstress zeigen. Der

Anstieg entsprach dem Read-through. Dieser korrelierte nur für HSV-I-Infektion mit einem Anstieg der ATAC-seq-Reads. Bemerkenswerterweise fiel besonders bei SRSF3 (Abb. 3.25), aber auch bei SRSF6 (Abb. 3.25), ein Unterschied zwischen dem Verhalten von ATAC-seq-Reads und 4sU-seq-Reads im Downstream auf. Die 4sU-seq-Reads hatten ihr Maximum wie erwartet im Gen und nahmen von dort aus mehr oder weniger gleichmäßig über den Verlauf des Read-throughs ab.
Die ATAC-seq-Reads hingegen waren am 3'-Ende des Gens gering und nahmen bis in den 30 kbp Downstreambereich zu, bevor sie ihr Maximum erreichten. Bei SRSF3 (Abb. 3.26) ließ sich zudem erkennen, dass die Abnahme der 4sU-seq-Reads der Abnahme der ATAC-seq-Reads vorausging. Am Ende des gezeigten Bereiches waren die 4sU-seq-Reads bereits fast auf das Level vor Infektion gesunken, während die ATAC-seq-Reads noch erhöht waren. Dies entspricht den Zusammenhängen, die mit PASsUS gezeigt wurden (s. Abb. 3.21).

3.5.2. GAPDH und ACTB

Glycerinaldehyd-3-phosphat-Dehydrogenase (GAPDH) und Beta-actin (ACTB) sind Beispielgene ohne Read-through, bzw. Gene mit einem Read-through von unter 10 % (s. Abb. 3.27 und 3.28).

Für beide lässt sich erneut sagen: Bei den ATAC-seq-Reads der uninfizierten Zellen waren Peaks in Genen zu erkennen, die teilweise mit den transkribierten Bereichen zusammenfielen. Weiterhin waren außerhalb der Genregionen Peaks annotiert, von denen die Größten mit Promotoren und anderen Kontrollsequenzen zusammenfielen. Manche der annotierten Peaks fielen auch hier in das Rauschen.
Hier wird besonders deutlich, wie wichtig es ist, beide DNA-Stränge zu betrachten. Bei GAPDH schienen schon bei uninfizierten Zellen OCRs im Downstream vorzuliegen. Diese fielen jedoch mit Genen auf dem Minusstrang zusammen.
8 h p.i. ließ sich im Downstream kein Anstieg der 4sU-seq-Reads für HSV-I-Infektion, Hitze- und Salzstress zeigen, was dem fehlenden Read-through entspricht. Ein Anstieg der ATAC-seq-Reads wie bei SRSF3 und SRSF6 blieb demnach auch weitestgehend aus, wobei sich jedoch für GAPDH zumindest die Konfiguration der OCRs geändert hat und für ACTB ein Bild vorlag, als gäbe es in den 10-15 kbp Downstreambereich einen Read-through.
Auch hier reichten Erhöhungen der ATAC-seq-Reads wieder weiter nach distal als Erhöhungen der 4sU-seq-Reads.

3.5.3. SLC30A5

„Solute Carrier Family 30 Member 5 “ (SLC30A5) ist interessant, da es 8 h p.i. mit 1,7 Mbp den größten dOCR von allen Genen aufwies (s. Abb. 3.29). Dies lag unter anderem daran, dass der entsprechende Downstreambereich viele Gene enthält, was jedoch nicht zum Ausschluss von SLC30A5 geführt hat, da dies nicht für den Sensestrang 5 kbp nach dem 3’ Ende zutrifft.
Bereits bei uninfizierten Zellen hatte der Downstreambereich von SLC30A5 massive OCRs, die teilweise auf die umliegenden Gene zurückzuführen waren. Zudem wurden im Bereich 69 700 k–69 800 k von F-Seq zahlreiche Peaks annotiert, die unplausibel waren. Weiterhin problematisch war die Ungewissheit darüber, ob die dem SLC30A5 folgenden Gene selbst neuen Read-through induzierten.

Damit ist SLC30A5 ein Beispiel für die Probleme, die eine Analyse über eine Vielzahl von Genen hinweg mit sich bringt. Problematisch ist die große Heterogenität der Gene. Gene unterscheiden sich nicht nur in Bezug auf Länge, transkriptionaler Aktivität und deren verschiedenen Beeinflussungen. Besonders ihre Position in Bezug auf Gencluster und Strang ist für diese Art der Auswertung von Bedeutung. Daher stellen viele Gene einen Spezialfall dar.

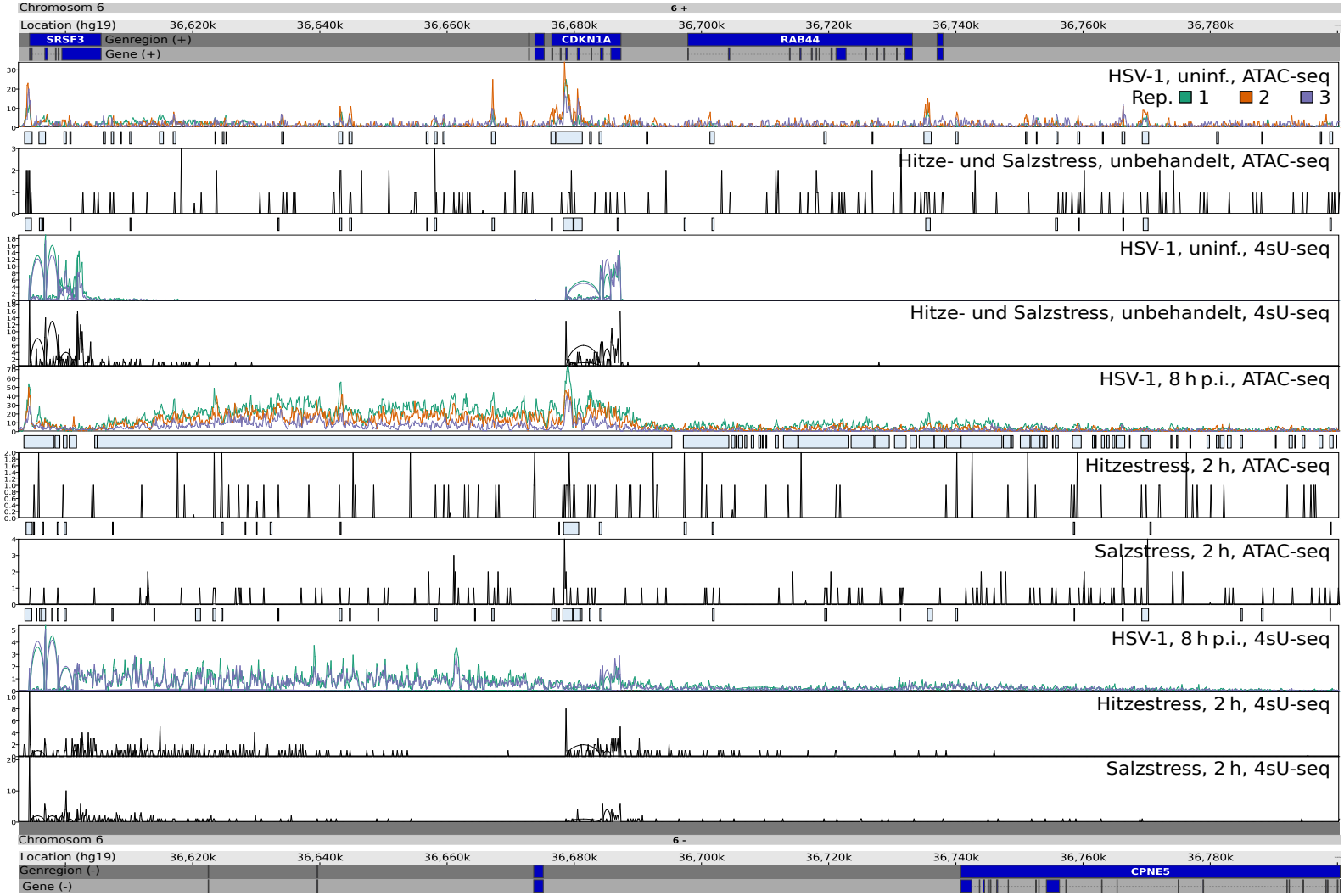


Abbildung 3.25.: ATAC-seq- und 4sU-seq-Daten am Beispiel von SRSF3 (Erklärung s. Kap. 3.5).

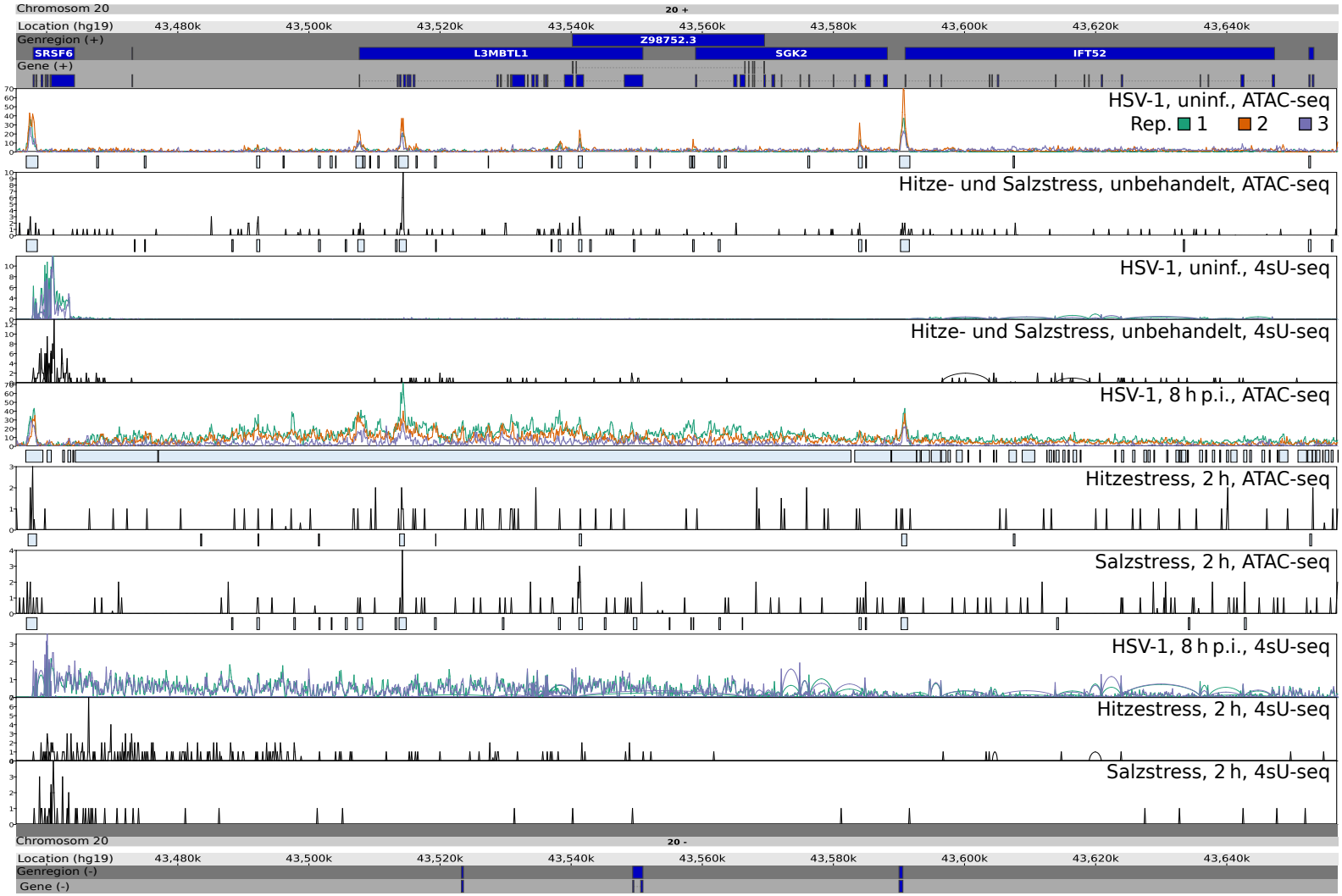


Abbildung 3.26.: ATAC-seq- und 4sU-seq-Daten am Beispiel von SRSF6 (Erklärung s. Kap. 3.5).

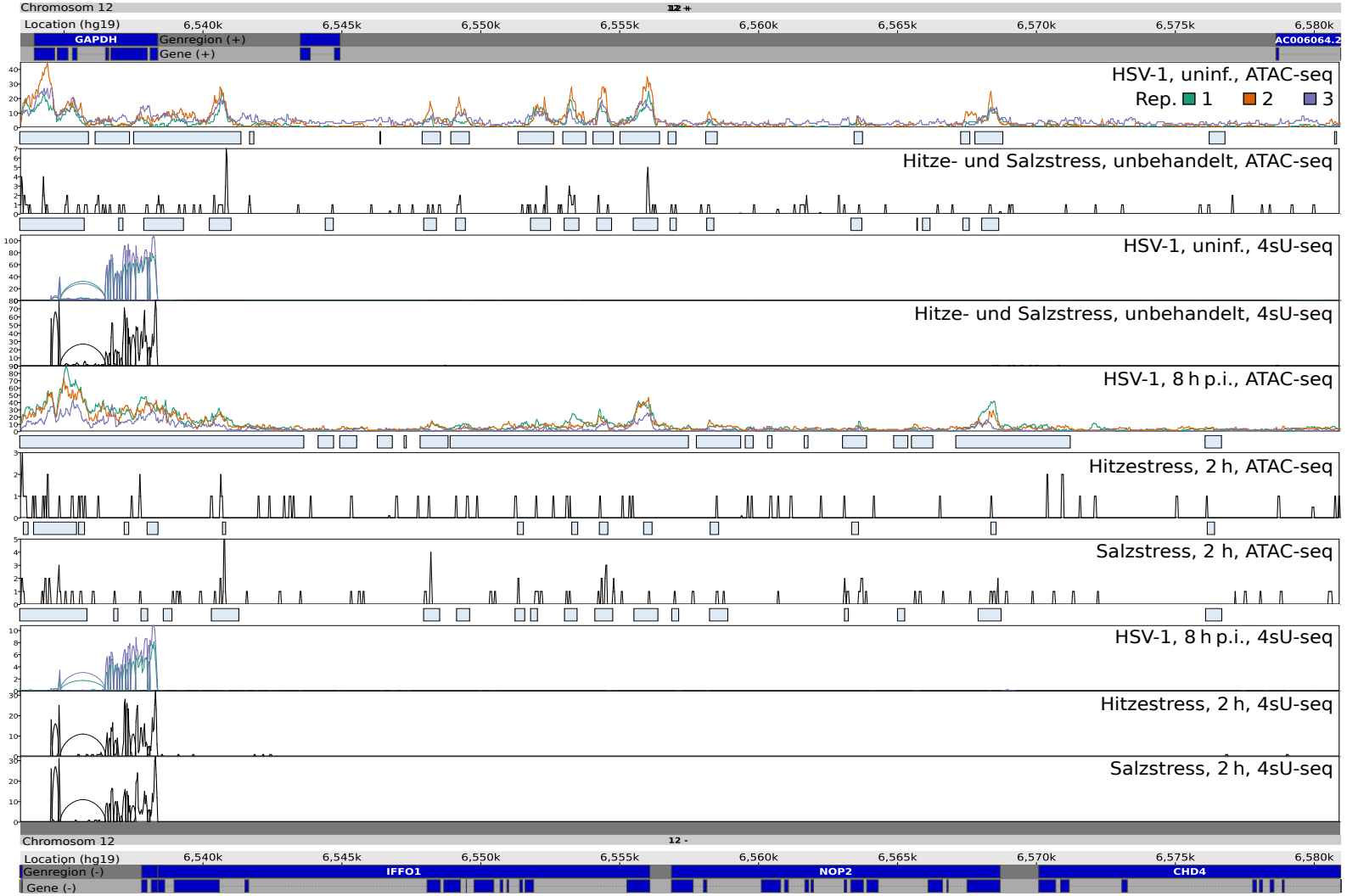


Abbildung 3.27.: ATAC-seq- und 4sU-seq-Daten am Beispiel von GAPDH (Erklärung s. Kap. 3.5).

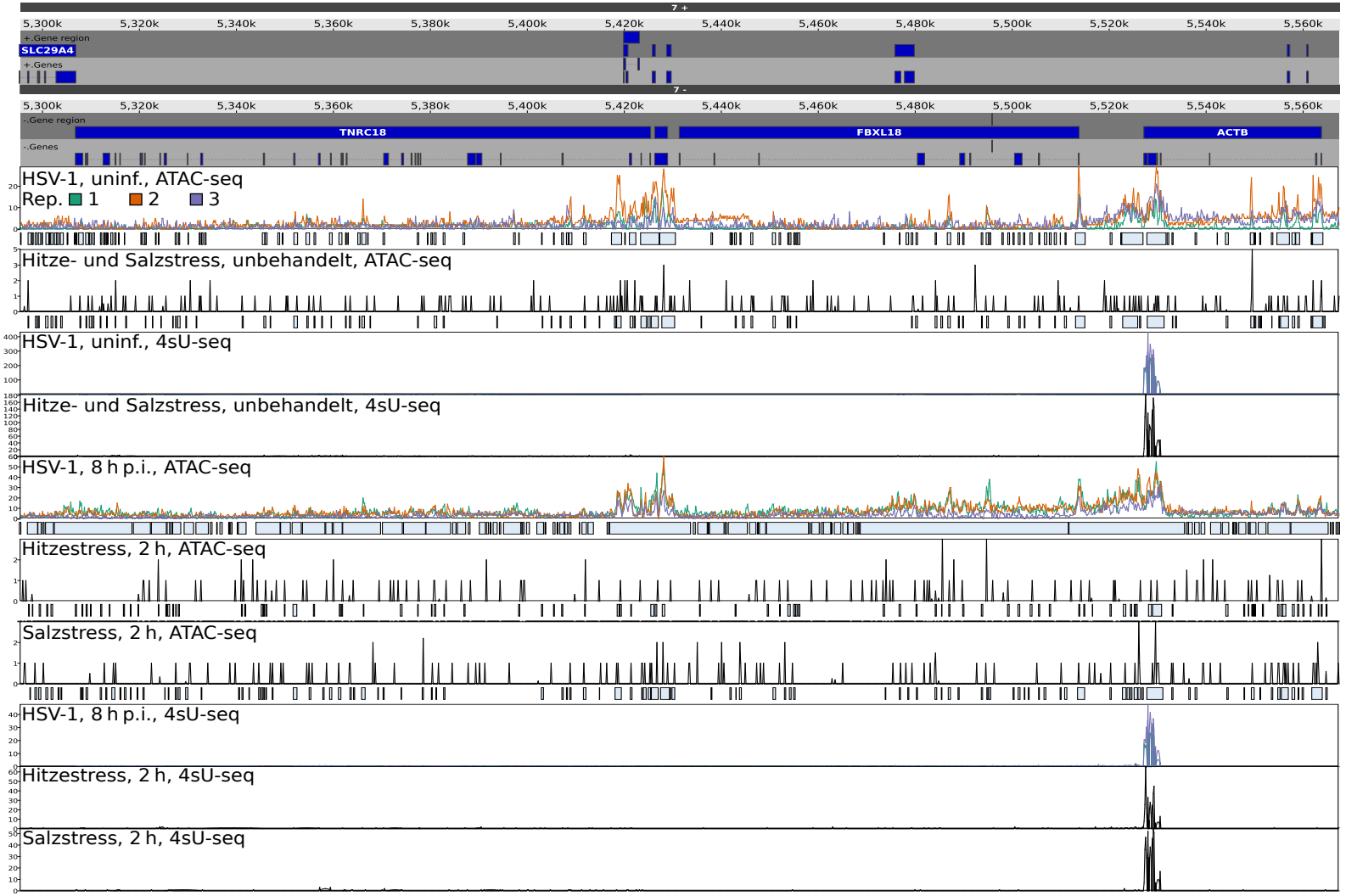


Abbildung 3.28.: ATAC-seq- und 4sU-seq-Daten am Beispiel von ACTB (Erklärung s. Kap. 3.5).

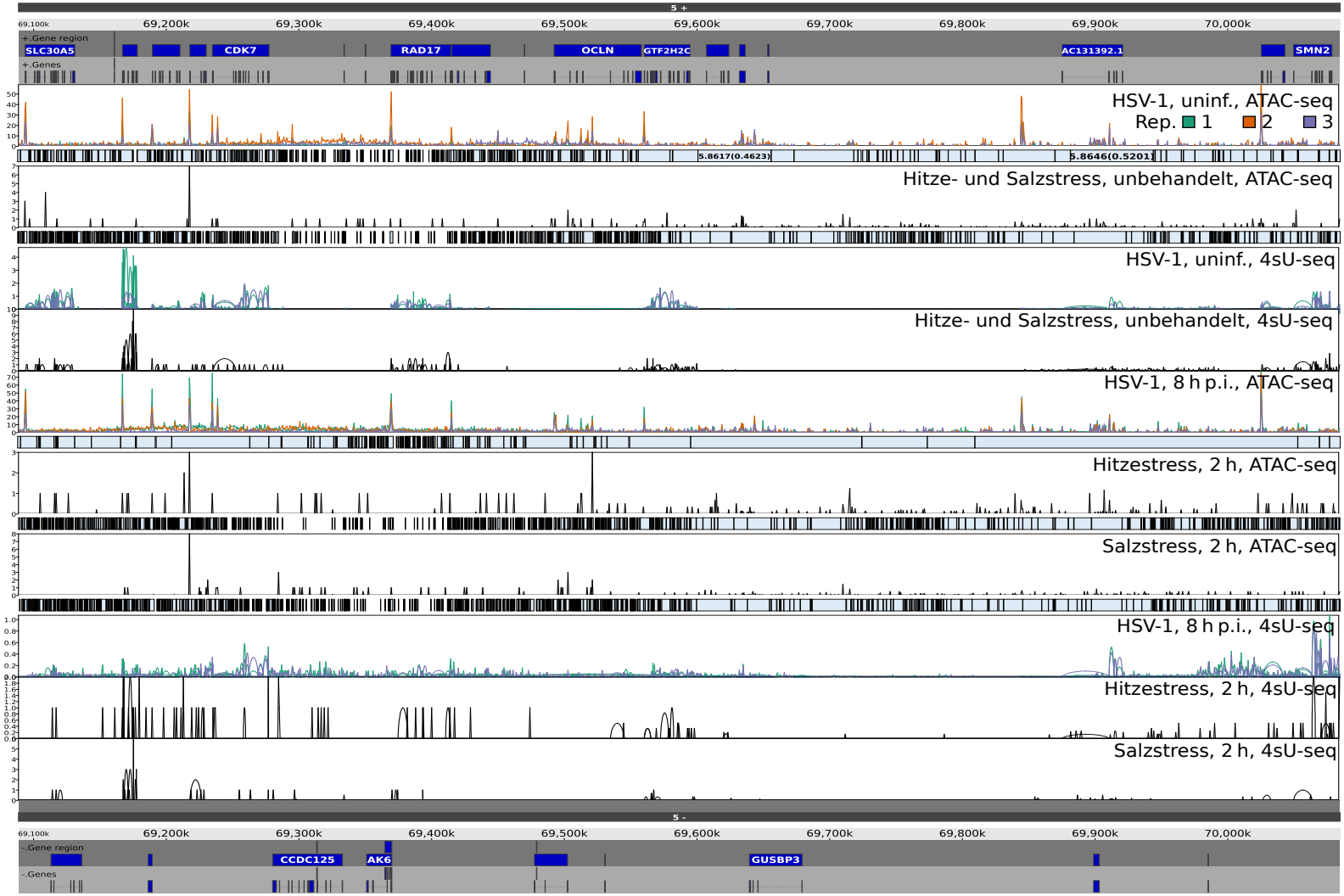


Abbildung 3.29.: ATAC-seq- und 4sU-seq-Daten am Beispiel von SLC30A5 (Erklärung s. Kap. 3.5).

4. Diskussion

4.1. Peak-Caller

Da die etablierten Peak-Caller nicht für ATAC-seq-Daten sondern für DNAse-seq und/oder ChIP-seq optimiert sind und sich ihre Ergebnisse bereits bei DNAse-seq erheblich unterscheiden [28], wurde in einem ersten Schritt eruiert, welcher Peak-Caller die durch Rutkowski et al. (2015) beschriebenen Veränderungen nach HSV-1 Infektion am besten wiedergibt bzw. welche Parameter eine Differenzierung zwischen uninfizierten und infizierten Replikaten erlauben (s. Kap. 2.2) [12].
Obwohl die Daten darauf hinwiesen, dass ATAC-seq Pipeline die geringste Rate an falsch-positiven Peaks annotiert, wurde für die Analyse der Chromatinorganisation F-Seq verwendet, da sich mit F-Seq über dOCR die größte Trennschärfe zwischen uninfizierten Zellen und 8 h p.i. erzielen ließ (s. Kap 1.4.2). Es stellte sich heraus, dass keiner der verwendeten Peak-Caller für sich genommen für die Anforderungen in einem befriedigenden Maß geeignet war. F-Seq annotierte viele falsch-positive Peaks und ATAC-seq Pipeline viele falsch-negative Peaks. Dies war nicht zuletzt dem Umstand geschuldet, dass F-Seq nicht für die Auswertung von ATAC-seq Daten entwickelt wurde und es sich bei ATAC-seq Pipeline noch um einen Prototypen handelte. In den Daten spiegelte sich dies darin wider, dass die Peak-Caller 8 h p.i. rund ein Drittel weniger Überschneidung zeigten als für uninfizierte Zellen (siehe Abb. 3.2 und 3.3). Eine stichprobenartige Untersuchung der von F-Seq annotierten Peaks zeigte stellenweise erhebliche Mängel. Von einem Bias durch unzureichende Optimierung der Peak-Caller ist somit auszugehen.
F-Seq wurde bei PASsUS verwendet, um Bereiche auszuschließen, in denen vor der Infektion ein starkes Signal vorhanden war. Eine Berücksichtigung der DNA-Bindungsstellen würde zur Unterschätzung des $\log_2$-fold-changes führen (s. Kap. 4.2). Da F-Seq dazu tendierte, falsch-positive Peaks zu annotieren, besonders wenn kein Score-Schwellenwert vorgenommen wurde, könnte ein übermäßiger Ausschluss von Bereichen einen Bias eingebracht haben.

Es wurde postuliert, dass DoTT aufgrund konfluierender Peaks zu einer Verlängerung der Peak-Längen sowie zu einer Abnahme der Peak-Anzahl führt (s. Kap. 2.2). Beide Annahmen trafen auf ATAC-seq Pipeline und MACS2 zu (s. Tab. 3.1). Jedoch ließ sich hierdurch keine Trennschärfe zwischen uninfizierten Zellen und 8 h p.i. erzielen. ATAC-seq Pipeline und MACS2 erfassten mit DoTT einhergehende Erweiterung der OCRs unzureichend (s. Abb. 3.4).
Einerseits könnte dies an der Konfiguration der Peak-Caller gelegen haben, welche das Auffinden von kurzen Peaks begünstigen haben könnte. Eine lange OCR kann als Rauschen interpretiert werden, obwohl eigentlich ein Peak annotiert werden sollte (s Kap. 2.2.1). Andererseits könnten auch die Annahmen falsch sein, dass es unter Infektion zu einer Verlängerung der Peak-Längen sowie zu einer Abnahme der Peak-Anzahl kommt. Unter der Annahme, dass der Downstreambereich 200-205 kbp so weit vom Gen entfernt liegt, dass er eher eine zufällige Stichprobe aus dem Genom als einen Downstreambereich repräsentiert, sprechen die Daten dafür, dass das Chromatin 8 h p.i. im Durchschnitt offener vorliegt als bei uninfizierten Zellen. Würde man Peaks auf Daten berechnen, die nur Rauschen darstellen, so würde man bei passender Grobheit des Rauschens und Konfiguration des Peak-Callers viele kleine Peaks finden. Auf eine ähnliche Art und Weise könnte das offenere Chromatin 8 h p.i. kurze Peaks begünstigt haben.
Daher ist es notwendig, Peaks zu poolen. Der Gewinn an Trennschärfe ist anhand der dOCR in Abbildung 3.10 zu erkennen.

Weiterhin bieten die Peak-Caller die Möglichkeit, durch Parameter ihre Passgenauigkeit zu erhöhen. Ein Peak-Caller, der DoTT und OCRs unter Standardeinstellungen schlecht beschrieben hat, könnte dazu theoretisch gut in der Lage sein, wenn er anders konfiguriert werden würde. Von einer Feinadjustierung durch Parameter wurde abgesehen, da sich durch die Kombination dieser eine unzählige Anzahl an Möglichkeiten ergibt und eine objektive Bewertung den Rahmen dieser Arbeit gesprengt hätte.

4.2. „Pipeline for ATAC-seq and 4sU-seq plotting“ (PASsUS)

PASsUS setzt die RPKM von ATAC-seq und 4sU-seq ins Verhältnis zueinander. Obwohl damit keine Peaks verglichen werden, werden dennoch von F-Seq annotierte Peaks verwendet. Wie in Kap. 2.3 erläutert, werden mit **subtract_regions** Bereiche aus der Auswertung ausgeschlossen, in denen bereits bei uninfizierten Zellen

Peaks annotiert wurden. Diese enthalten z. B. „Transcription Factor Binding Sites“ (TFBS). Dies hat den Hintergrund, dass eine TFBS mit 1000 Reads über 10 bp sonst mit einer OCR mit 1000 Reads über 1000 bp gleichgewichtet werden würde. Der Peak-Caller kann nicht unterscheiden, ob Chromatin aufgrund von TFBS oder anderen Gründen offen ist.
Liegt z. B. Gen-A im Downstreambereich von Gen-B und ist bei uninfizierten Zellen stark transkribiert und damit offen, gilt folgendes: Wenn Gen-A 8 h p.i. herunterreguliert wurde und das Chromatin weniger offen vorliegt, wäre auch für Gen-B die Offenheit des Chromatins im Downstreambereich herabgesetzt. Es wäre jedoch falsch, dieses für Gen-B zu berücksichtigen, da die Veränderung auf Gen-A zurückzuführen ist und nicht auf eine Veränderung des DoTT bzw. die DoG Transkription von Gen-B.
Die gleichzeitige Verwendung von **subtract_regions** und **intersect_regions** für uninfizierte Zellen ist nicht sinnvoll. Subtrahiert man von einer Menge A zuerst eine Menge B, so ist die Schnittmenge aus der entstehenden Menge und der Ausgangsmengen leer. Damit erübrigt sich auch, in einem späteren Schritt Differenzen oder Quotienten auf Basis dieser Menge zu bilden. Für die absoluten Werte p.i. hat sich gezeigt, dass die Korrelation zwischen 4sU- und ATAC-seq-Daten geringer ausfällt, wenn man **intersect_regions** nutzt, um nur die Reads in Peaks zu zählen.

4.2.1. Normalisierung

Wie in Kapitel 2.3.1 beschrieben, ist eine Normalisierung der Reads sowohl für ATAC-seq als auch für 4sU-seq notwendig. Dies hat den Hintergrund, dass Datensets aus Sequenzierungen unterschiedlich groß ausfallen. Dadurch sind Reads zwischen verschiedenen Datensets nicht direkt vergleichbar. Anhand von Merkmalen, die zwischen verschiedenen Sequenzierungen gleich groß ausfallen sollten, lassen sich die Reads jedoch skalieren, sodass sich eine Vergleichbarkeit erreichen lässt. Für 4sU-seq wurde die rRNA verwendet. Es wird davon ausgegangen, dass sich die tatsächliche rRNA nicht zwischen den Replikaten geändert hat. Ändert sich die gemessene rRNA zwischen den Replikaten um einen Faktor, so kann eine Vergleichbarkeit der Replikate durch Division bzw. Multiplikation mit diesem Faktor hergestellt werden. Bei ATAC-seq lässt sich das gleiche Prinzip für die mtDNA anwenden.

4.2.2. Negative log_2-fold-changes

Manche Gene wiesen 8 h p.i. einen negativen log_2-fold-change auf (s. Kap. 3.3). Dies würde bei oberflächlicher Betrachtung darauf hinweisen, dass in Bezug auf DoTT bei diesen Genen physiologisch mehr DoTT vorgelegen hätte als 8 h p.i.. Zieht man aber in Betracht, dass dieser Effekt hauptsächlich in Genen festzustellen war, bei denen distal weitere Gene liegen, so lässt sich schlussfolgern, dass dies der Berechnung bei Überlappung geschuldet war.

4.3. Downstream OCRs (dOCR)

Die dOCR, wie sie von Hennig et al. (2018) beschrieben wurde, ist nicht zu verwechseln mit einer Region im engeren Sinne, wie zum Beispiel einer OCR [22]. Vielmehr ist dOCR eine Hilfsgröße, die durch zusammenlegen von Peaks entsteht. So könnte ein dOCR von 1000 bp sowohl einen Peak der Länge 1000 darstellen als auch 100 Peaks der Länge 10, die über 500 kbp verteilt sind.
Dies verdeutlicht, dass die Größe von dOCR nicht mit der tatsächlichen Länge von offenem Chromatin im Downstreambereich verwechselt werden darf.
Die Funktion zur Berechnung von dOCR addiert alle Peak-Längen im 10 kbp Downstreambereich und alle weiteren Peak-Längen von Peaks, die zum letzten addierten Peak einen Abstand von höchsten 5 kbp haben, sofern es in den ersten 10 kbp einen Peak gab. Dabei kann nicht unterschieden werden, auf welchen Strang eine OCR zurückzuführen ist. Dies stellt ein Problem dar, da offenes Chromatin immer beide Stränge betrifft. Ein Gen kann also von einem bis zu 10 kbp langen Downstreambereich mit geschlossenem Chromatin gefolgt sein und trotzdem eine große dOCR haben, wenn vor dem 10 000 bp noch ein Peak beginnt, auf den weitere in einem Abstand geringer als 5 kbp folgen. Es sollte klar sein, dass ein solches dOCR nicht auf DoTT zurückzuführen ist.
Zudem kommt es vor, dass sich mehrere Gene das Ende eines langen dOCR teilen. Zwischen SLC30A5 und CDK7 liegen auf demselben Strang noch fünf weitere Gene und es war alleine anhand der dOCR nicht möglich zuzuordnen, welches dieser Gene kausal mit dem dOCR zusammenhängt (s. Abb. 3.29). Der Ausschluss von Genen, die einen hohen Read-in oder Gene im 5 kbp Umfeld hatten, war jedoch ein Schritt in diese Richtung.
In diesem Sinne überschätzt dOCR das Ausmaß von OCRs im Downstream. Auf der anderen Seite berücksichtigt dOCR die Intervalle zwischen Peaks nicht, wodurch der Wert von dOCR geringer ausfallen kann als der tatsächlich von DoTT betroffene

Bereich.
Die 10 kbp als Startfenster und die 5 kbp für die Iteration könnten reduziert werden, um dOCR weniger zu überschätzen.
Diese Probleme müssen auch bei der Interpretation der Subgruppenanalyse besonders großer dOCR bedacht werden (s. Abb. 3.13). Hier konnte im 5 kbp Downstreambereich für 133 Gene mit einem dOCR größer als 110 kbp eine Spearman Korrelation von $R_s \approx 0.21$ zwischen ATAC-seq-Reads und prozentualem Read-through berechnet werden. Für eine Mindestlänge für dOCR von 15 kbp konnte jedoch keine Korrelation gezeigt werden. Dabei zeigten gerade große dOCR-Längen eine Überlappung und damit Verzerrung mit bzw. durch andere Gene, wie man an ACTB (s. Abb. 3.28) und SLC30A5 (s. Abb. 3.29) sehen kann.
Dies deutet darauf hin, dass Gencluster ein Confounder für dOCR und hohen Read-through sind. Da zum Zeigen der Korrelation zwischen absoluten ATAC-seq-Reads und prozentualem Read-through notwendig war, von 3684 Genen nur 133 zu berücksichtigen, sollte dieses Ergebnis nicht überbewertet werden. Dies spiegelt sich auch in den großen Konfidenzintervallen wider. Wie bereits erwähnt waren n = 31 (30, 25, 22, 25) Gene für einen Read-through von 0-20 (20-40, 40-60, 60-80, >80) Prozentpunkten bei einem dOCR (8 h p.i.) > 110 kbps bemerkenswert: Wenn es keine Gencluster gäbe, würde man erwarten, dass nur Gene mit hohem Read-through zu großen dOCR führen. Dies spiegelte sich aber keinesfalls in der hier gezeigten Verteilung wieder. Dies stützt die These, dass Gencluster ein Confounder für große dOCR und hohen Read-through sind.

Weiterhin fällt auf, dass große dOCR darauf angewiesen sind, dass der Peak-Caller viele Peaks annotiert hat. Dies bestätigt sich darin, dass ATAC-seq Pipeline und MACS2 wenige Peaks und kleine dOCR im Vergleich zu F-Seq und Hotspot begünstigten.
F-Seq waren 8 h p.i. nicht nur die größten dOCR zugeschrieben, sondern es annotierte auch die meisten Peaks. Theoretisch könnte dies auf ein erhöhtes Rauschen 8 h p.i. zurückzuführen sein (Vgl. Abb. 3.4). Denn wenn durch Rauschen alle 5 kbp ein falsch-positiver Peak annotiert werden würde, so würde dies zu großen dOCR führen. Dies würde jedoch bei starker Ausprägung zum Fehlen von kurzen dOCR führen, was nicht der Fall war. Weiterhin würden die dOCR dann nicht mehr mit anderen Methoden korrelieren.

Besonders für die Berechnung von ATAC-seq $\log_2$-fold-changes ist es von Bedeu-

tung, Bereiche mit offenem Chromatin in uninfizierten Zellen und damit TFBS auszuschließen (s. Kap. 4.2). Für dOCR wurde eine vergleichbare Korrektur nicht vorgenommen. Möglicherweise besteht hier noch Raum für Optimierungen.

Trotz aller Kritik zeigte sich die Nützlichkeit der dOCR-Berechnung in der daraus resultierenden Trennschärfe zwischen uninfizierten Zellen und jenen in der späten Infektion. Dabei war die dOCR-Berechnung mit einem erheblich geringerem Rechenaufwand verbunden als die RPKM-Analyse mittels PASsUS.

4.4. DoTT bzw. DoG und OCR

Der Großteil an DoTT und OCRs spielten sich im Bereich der ersten 75 kbp des Downstreambereiches ab (s. Abb. 3.22). Auffällig ist, dass beim HSV-1 Wildtyp unter Betrachtung der RPKM-Differenzen die Offenheit des Chromatins 6 h p.i. ein Maximum erreichte, während DoTT weiterhin zunahm (s. Abb 3.21). Auch sonst spricht der nicht lineare Kurvenverlauf für eine hohe Komplexität der Zusammenhänge zwischen den beiden Phänomenen. Im Zuge der lytischen HSV-1 Infektion kommt es zu einem zunehmenden Verlust von Pol-II und zellulärer Transkription [12, 48]. Dies erklärt aller Voraussicht nach die Abnahme von dOCR von 6 auf 8 h p.i., während DoTT bei der verbleibenden Transkription zunimmt. Eine weitere Erklärung könnte sein, dass beide Phänomene durch unterschiedliche virale Proteine verursacht werden, die sich in ihrer Expressionskinetik unterschiedlich verhalten. Es konnte festgestellt werden, dass das Maximum der ATAC-seq-Reads und 4sU-seq-Reads gemessen an den RPKM-Quotienten nicht im unmittelbaren Downstreambereich lag, sondern dass diese mit Abstand zum 3'-Ende bis zu 30 kbp zunahmen (s. Abb. 3.22). Dieses Ergebnis bestätigt sich im Genomviewer zumindest für die ATAC-seq-Daten (s. Abb. 3.25). An dieser Stelle sei darauf hingewiesen, dass die Quotientenbildung Änderungen in Bereichen mit weniger Reads überpräsentiert, während Differenzenbildung diese unterrepräsentiert.
Aus rein biochemischen Überlegungen über die Pol-II ergibt ein Zunehmen der Reads nach distal wenig Sinn, da die Pol-II sich aus einem Gen heraus in den Downstreambereich bewegt und daher einen distalen Bereich nicht stärker, also häufiger, transkribieren kann, als einen näher gelegenen.
Die Ursache hierfür bei ATAC-seq-Reads könnte rein spekulativ darin liegen, dass die Histonrepositionierung im Gen geringfügiger gestört ist als im intergenomischen Bereich. Unter diesem Gesichtspunkt müsste der Effekt in Genclustern weniger prominent sein (s. Abb. 3.29). Zudem geht die physiologische Termination damit einher,

dass die Offenheit des Chromatins wieder reduziert wird. Das Bild lässt sich also am besten dadurch erklären, dass eine Teilmenge der Pol-II auch 8 h p.i. noch physiologisch terminiert und damit das Chromatin im unmittelbaren Downstreambereich relativ geschlossen hält.

Sowohl „Facilitates Chromatin Transcription“ (FACT), welches an der Histonrepositionierung beteiligt ist, als auch die Histonklassen, sind evolutionär alte und hoch konservierte Proteine. Da HSV-1 eine Vielzahl von Wirbeltieren als auch Wirbellosen infizieren kann, sind evolutionär konservierte Gene gute Kandidaten, um eine Wechselwirkung zwischen Host und Virus zu suchen. All dies spricht dafür, dass FACT und/oder Histone über eine gestörte Histonrepositionierung mit OCR kausal verbunden sein könnten und damit ein gutes Ziel für weitere Forschung sind.
OCRs reichen weiter in die Downstreambereiche hinein als DoTT (s. Abb. 3.20 – 3.22). Dies betrifft sogar den Downstreambereich 200 kbp bis 205 kbp. Dieser ist so weit vom Ursprungsgen entfernt, dass die hier gemessenen Read-Dichten eher zufälligen Stichproben entsprechen, als dass sie wirklich einen Zusammenhang zum Ursprungsgen hätten. Nur 29 Gene haben bis 205 kbp Downstream keine Überlappung mit einem anderem Gen. Dies könnte darauf hindeuten, dass die Offenheit des Chromatins über das komplette Genom gemittelt nach Infektion zugenommen hat. Mit zunehmender Nähe eines Downstreambereiches zu dem entsprechenden Gen ist jedoch davon auszugehen, dass die Änderungen im Downstreambereich von dem Gen ausgehen. Der Trend, dass offenes Chromatin im Downstreambereich über DoTT hinausgeht, bestätigte sich auch im Genomviewer (s . Abb. 3.4) und steht im scheinbaren Widerspruch zu der Tatsache, dass DoTT die Voraussetzung für ausgeprägte OCRs im Downstream ist.

Das schnelle Erreichen eines Plateaus der ATAC-seq-RPKM suggeriert, dass ein relativ geringes Maß an Transkription bei gestörter Wiedereinfügung der Histone bereits zu einem hohen Maß an offenem Chromatin führt (s. Abb. 3.21 und 3.22). Eine hohe Steigerung der Transkription führt dann nur noch zu einer geringen Steigerung der ATAC-seq-Reads. Dies könnte darauf hindeuten, dass ATAC-seq eine geringere dynamische Breite als 4sU-seq aufweist (d.h. es gab zumindest in den hier vorliegenden Datensets weniger Abstufungen zwischen der geringsten und stärksten Ausprägung). Dem steht entgegen, dass erstens die Aktivität der Tn5 Transposase in Abhängigkeit der räumlichen Zugänglichkeit des Chromatins Abstufungen hat und zweitens die ATAC-seq-Daten aus einem Medium von ungefähr 50.000 Zellen

gewonnen wurden und damit ein Mittel dieser darstellen.

Das Fehlen von vermehrten OCRs im Downstream bei Hitzestress (s. Kap. 3.4) unterstreicht die funktionelle Unabhängigkeit von DoTT gegenüber OCR. Daher dürfte es hilfreich sein, bei der Suche nach den molekularen Mechanismen der dOCR 8 h p.i. neben OCR bei DoTT auch auf Gründe für OCR im Allgemeinen zu schauen. Möglicherweise lässt sich durch eine gezielte Manipulation von FACT und/oder Histonen die Histonrepositionierung stören und damit während Zellstress dOCR induzieren, womit sich molekulare Mechanismen entschlüsseln ließen.
Eine geringe Erhöhung der Offenheit des Chromatins bei Salzstress wurde noch nicht berichtet und unterscheidet sich von den aus dOCR abgeleiteten Daten (s. Abb. 3.24). Jedoch wird diese Beobachtung nicht durch Beispiele im Genomviewer gestützt. Außerdem fehlt ein Abfallen der ATAC-seq-RPKM mit zunehmender Entfernung vom 3'-Ende. Dies spricht gegen eine Erhöhung der Offenheit des Chromatins im Zusammenhang mit Veränderungen der Transkription. Sollte es sich hierbei um einen Fehler in z. B. der Messung oder der Normalisierung handeln, so dürfte es sich um einen systematischen Fehler handeln, da die einzelnen Replikate in ihm übereinstimmen.

4.5. Betrachtungsweise von Quotienten und Differenzen

PASsUS ermöglicht es, Abbildungen sowohl unter Zuhilfenahme von Quotienten aus RPKM als auch unter Zuhilfenahme von Differenzen aus RPKM anzufertigen (s. Kap. 3.3). Dies geschah aus folgenden Überlegung heraus:
Erstens könnte ein RPKM-Quotient zu einer Überschätzung führen, wenn der Bereich, der im Divisor betrachtet wird, sehr wenige RPKM enthält. Bei einer geringen RPKM-Differenz zwischen dem im Divisor und Dividenden betrachteten Bereich kommt es zu einem großen Quotienten. Zweitens könnten RPKM-Quotienten zu einer Unterschätzung führen, wenn sowohl der im Divisor als auch der im Dividenden betrachtete Bereich sehr viel RPKM enthält. Die absolute RPKM-Differenz könnte dadurch relativ groß ausfallen, während der Quotient gering ausfällt.

Aus Kap. 3.3.2 geht jedoch hervor, dass diese Extremfälle nicht zu relevanten Verzerrungen der beiden Betrachtungsweisen geführt haben. Dies ist erstens darauf zurückzuführen, dass mit LFC die Auswahl der Pseudocounts gegen diese Art der

Verzerrung optimiert wurde. Die Daten bestätigen, dass sich bei keinem der 3684 Genen die RPKM-Differenz und der RPKM-Quotient stark unterscheidet. Zweitens wurden eine selektierte Subgruppe von exprimierten Genen betrachtet. Dies bedeutet zwar nicht zwangsläufig, dass auch der Downstreambereich der 3684 Gene exprimiert wurde, trotzdem vermeidet es den Effekt, den nicht exprimierte Gene gehabt hätten. Drittens ist davon auszugehen, dass sich die Überschätzungen und Unterschätzungen zumindest teilweise ausbalanciert haben.

4.6. Bedeutung der Analysen für Zellbiologie und Virologie

HSV-1 ist in den vergangenen Jahren zunehmend ein Model-System zur Untersuchung der Terminierung der Transkription geworden. In der lytischen Infektion unterliegt die Genexpression sowohl viralen als auch antiviralen Mechanismen. DoG Transkription ist im Gegensatz zur Erhöhung von offenem Chromatin im Downstreambereich ein Phänomen, dass sich sowohl in lytischer Infektion als auch in Zellstress vorfinden lässt. Da DoTT zum „host shut-off“ beiträgt, lässt sich annehmen, dass HSV-1 mit DoTT einen zellulären Mechanismus instrumentalisiert. Im Gegensatz dazu ist die Erhöhung der Chromatinoffenheit im Downstream in Verbindung mit DoG kein Phänomen, welches sich bei Hitze- und Salzstress findet.
Durch Untersuchung der RPKM-Werte wurde gezeigt, dass die Offenheit des Chromatins in weiter distal liegende Downstreambereiche reicht als DoTT (s. Abb. 3.21). Dies konnte anhand von dOCR alleine nicht gezeigt werden, da zur Untersuchung von dOCR ATAC-seq-Daten betrachtet wurden. Die Analysen mittels RPKM zeigten, dass das Ausmaß von DoTT in RPKM gemessen das Ausmaß von DoG um ein Vielfaches überstieg. Sowohl die Analysen mittels RPKM als auch mittels dOCR sind geeignet, um eine Trennschärfe zwischen unbehandelten Replikaten und Replikaten 8 h p.i. zu erzielen. Damit bieten die Analysen mittels RPKM eine zusätzliche Validierung des Vorgehen mittels dOCR.

Bei lytischer HSV-1 Infektion wird von DoTT und bei Zellstress von DoG Transkription gesprochen. Die von RPKM abgeleiteten Daten legen aber Nahe, dass sich diese beiden Phänomene nicht nur in ihrer Ursache unterscheiden, sondern auch in ihrer Ausprägung. Dies bleibt durch weitere Untersuchungen zu bestätigen oder zu widerlegen.

5. Zusammenfassung

Aktuell ist kein Peak-Caller zur Erkennung sehr langer OCRs hinreichend optimiert (s. Kap. 3.1.1). Trotzdem ließen die bereits existierenden Programme, in erster Linie F-Seq, eine sinnvolle Auswertung von ATAC-seq-Daten zu (s. Kap. 3.1.3). Das Errechnen der Hilfsgröße „downstream Open Chromatin Region(s)“ (dOCR) half dabei eine Trennschärfe zwischen uninfizierten Zellen und 8 h p.i. zu erreichen (s. Kap. 3.6). Dies ließ sich durch die Analyse der RPKM mittels PASsUS validieren.
Die ATAC-seq- und 4sU-seq-Reads korrelierten 8 h p.i. leicht bis mittelmäßig miteinander (Spearman Korrelation $R_s \approx 0.29$ für den 5 kbp Downstreambereich bzw. 0.59 für den 15 kbp Downstreambereich für Gene mit dOCR größer als 50 kbp, $p < 0.001$, s. Abb. 3.12), wobei die ATAC-seq-Reads nur mit dem Read-through korrelierten, wenn man Gene mit großem dOCR betrachtete (Spearman Korrelation $R_s \approx 0.21$ für den 5 kbp Downstreambereich, $p < 0.001$). Dabei ist zu beachten, dass dOCR die tatsächliche Offenheit des Chromatins im Downstreambereich überschätzte. Für knapp ein Fünftel der 3684 untersuchten Gene war 8 h p.i. dOCR größer als Null, obwohl sich laut F-Seq im 5 kbp Downstreambereich keine OCR fand (s. Kap. 3.10). Die dOCR-Länge und der prozentuale Read-through korrelierten nicht bis kaum miteinander (Spearman Korrelation $R_s < 0.01$ für den 5 kbp Downstreambereich bzw. 0.21 für den 15 kbp Downstreambereich für Gene mit dOCR größer als 50 kbp, $p < 0.001$, s. Abb. 3.13.)
Während die Offenheit des Chromatins 6 h p.i. ihr Maximum erreichte und sich danach zu reduzieren begann, nahm das Maß an „Disruption of Transcription Termination“ (DoTT) weiterhin zu (s. Kap. 3.3.2). Andererseits stiegen bei DoTT die ATAC-seq-Reads erst distaler als die 4sU-seq-Reads an und fielen ebenfalls erst distaler ab (s. Kap. 4.4).
Für Hitze und Salzstress bestätigte sich, dass DoTT ähnlich wie bei einer HSV-1-Infektion auftrat, jedoch nicht von dOCR begleitet wurde (s. Kap. 3.4). An dieser Stelle bietet sich die Erforschung der molekularen Mechanismen an.

6. Literaturverzeichnis

[1] R. D. Kornberg, "Chromatin structure: A repeating unit of histones and dna," *Science*, vol. 184, no. 4139, pp. 868–871, 1974.

[2] O. Bell, V. K. Tiwari, N. H. Thoma, and D. Schubeler, "Determinants and dynamics of genome accessibility," *Nat Rev Genet*, vol. 12, no. 8, pp. 554–64, 2011.

[3] G. Löffler and P. C. Heinrich, *Biochemie und Pathobiochemie*. Springer-Lehrbuch, Berlin ; Heidelberg ; Berlin ; Heidelberg: Springer ; Springer, 2014.

[4] W. Luo, A. W. Johnson, and D. L. Bentley, "The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model," *Genes Dev*, vol. 20, pp. 954–965, Apr 2006.

[5] N. J. Proudfoot, "Transcriptional termination in mammals: Stopping the rna polymerase ii juggernaut," *Science*, vol. 352, no. 6291, 2016.

[6] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao, "Combinatorial patterns of histone acetylations and methylations in the human genome," *Nat Genet*, vol. 40, pp. 897–903, Jul 2008.

[7] L. Marcinowski, M. Lidschreiber, L. Windhager, M. Rieder, J. B. Bosse, B. Rädle, T. Bonfert, I. György, M. de Graaf, O. Prazeres da Costa, P. Rosenstiel, C. C. Friedel, R. Zimmer, Z. Ruzsics, and L. Dölken, "Real-time transcriptional profiling of cellular and viral gene expression during lytic cytomegalovirus infection," *PLoS Pathog*, vol. 8, p. e1002908, Sep 2012.

[8] L. Dolken, Z. Ruzsics, B. Radle, C. C. Friedel, R. Zimmer, J. Mages, R. Hoffmann, P. Dickinson, T. Forster, P. Ghazal, and U. H. Koszinowski, "High-resolution gene expression profiling for simultaneous kinetic parameter analysis of rna synthesis and decay," *Rna*, vol. 14, no. 9, pp. 1959–72, 2008.

[9] L. Windhager, T. Bonfert, K. Burger, Z. Ruzsics, S. Krebs, S. Kaufmann, G. Malterer, A. L'Hernault, M. Schilhabel, S. Schreiber, P. Rosenstiel, R. Zimmer, D. Eick, C. C. Friedel, and L. Dölken, "Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution," *Genome Res*, vol. 22, pp. 2031–2042, Oct 2012.

[10] B. Rädle, A. J. Rutkowski, Z. Ruzsics, C. C. Friedel, U. H. Koszinowski, and L. Dölken, "Metabolic labeling of newly transcribed RNA for high resolution gene expression profiling of RNA synthesis, processing and decay in cell culture," *J Vis Exp*, Aug 2013.

[11] J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf, "Atac-seq: A method for assaying chromatin accessibility genome-wide," *Current protocols in molecular biology*, vol. 109, pp. 21.29.1–21.29.9, 2015.

[12] A. J. Rutkowski, F. Erhard, A. L'Hernault, T. Bonfert, M. Schilhabel, C. Crump, P. Rosenstiel, S. Efstathiou, R. Zimmer, C. C. Friedel, and L. Dolken, "Widespread disruption of host transcription termination in hsv-1 infection," *Nature Communications*, vol. 6, p. 7126, 2015.

[13] K. W. Savin, B. G. Cocks, F. Wong, T. Sawbridge, N. Cogan, D. Savage, and S. J. V. J. Warner, "A neurotropic herpesvirus infecting the gastropod, abalone, shares ancestry with oyster herpesvirus and a herpesvirus associated with the amphioxus genome," vol. 7, no. 1, p. 308, 2010.

[14] C. Mahiet, A. Ergani, N. Huot, N. Alende, A. Azough, F. Salvaire, A. Bensimon, E. Conseiller, S. Wain-Hobson, M. Labetoulle, and S. Barradeau, "Structural variability of the herpes simplex virus 1 genome in vitro and in vivo," vol. 86, no. 16, pp. 8592–8601, 2012.

[15] A. W. Whisnant, C. S. Jürges, T. Hennig, E. Wyler, B. Prusty, A. J. Rutkowski, A. L'hernault, L. Djakovic, M. Göbel, K. Döring, J. Menegatti, R. Antrobus, N. J. Matheson, F. W. H. Künzig, G. Mastrobuoni, C. Bielow, S. Kempa, C. Liang, T. Dandekar, R. Zimmer, M. Landthaler, F. Grässer, P. J. Lehner, C. C. Friedel, F. Erhard, and L. Dölken, "Integrative functional genomics decodes herpes simplex virus 1," *Nat Commun*, vol. 11, p. 2038, 04 2020.

[16] S. H. James, J. S. Sheffield, and D. W. Kimberlin, "Mother-to-child transmission of herpes simplex virus," *J Pediatric Infect Dis Soc*, vol. 3 Suppl 1, pp. S19–23, 2014.

[17] G. Korr, M. Thamm, I. Czogiel, C. Poethko-Mueller, V. Bremer, and K. Jansen, "Decreasing seroprevalence of herpes simplex virus type 1 and type 2 in germany leaves many people susceptible to genital infection: time to raise awareness and enhance control," *BMC Infectious Diseases*, vol. 17, p. 471, 2017.

[18] N. Suttorp, M. Mielke, W. Kiehl, and B. Stück, *Infektionskrankheiten*. Stuttgart: Thieme, 2004.

[19] S. L. Gottlieb, B. K. Giersing, J. Hickling, R. Jones, C. Deal, and D. C. Kaslow, "Meeting report: Initial world health organization consultation on herpes simplex virus (hsv) vaccine preferred product characteristics, march 2017," *Vaccine*, 2017.

[20] S. A. Harris and E. A. Harris, "Herpes simplex virus type 1 and other pathogens are key causative factors in sporadic alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 48, no. 2, pp. 319–353, 2015.

[21] M. P. Nicoll, J. T. Proença, and S. Efstathiou, "The molecular basis of herpes simplex virus latency," *FEMS microbiology reviews*, vol. 36, no. 3, pp. 684–705, 2012.

[22] T. Hennig, M. Michalski, A. J. Rutkowski, L. Djakovic, A. W. Whisnant, M.-S. Friedl, B. A. Jha, M. A. P. Baptista, A. L'Hernault, F. Erhard, L. Dölken, and C. C. Friedel, "Hsv-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes," *PLoS pathogens*, vol. 14, no. 3, pp. e1006954–e1006954, 2018.

[23] A. Vilborg, N. Sabath, Y. Wiesel, J. Nathans, F. Levy-Adam, T. A. Yario, J. A. Steitz, and R. Shalgi, "Comparative analysis reveals genomic features of stress-induced transcriptional readthrough," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 40, pp. E8362–E8371, 2017.

[24] E. Wyler, J. Menegatti, V. Franke, C. Kocks, A. Boltengagen, T. Hennig, K. Theil, A. Rutkowski, C. Ferrai, L. Baer, L. Kermas, C. Friedel, N. Rajewsky, A. Akalin, L. Dölken, F. Grässer, and M. Landthaler, "Widespread activation of antisense transcription of the host genome during herpes simplex virus 1 infection," *Genome biology*, vol. 18, no. 1, pp. 209–209, 2017.

[25] F. Erhard, "Genomic data integration platform (gedi)." `https://github.com/erhard-lab/gedi`, 2017.

[26] F. Erhard, "Estimating pseudocounts and fold changes for digital expression measurements," *Bioinformatics*, 2018.

[27] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[28] H. Koohy, T. A. Down, M. Spivakov, and T. Hubbard, "A comparison of peak callers used for dnase-seq data," *PLOS One*, vol. 9, no. 5, p. e96303, 2014.

[29] C. D. Scharer, E. L. Blalock, B. G. Barwick, R. R. Haines, C. Wei, I. Sanz, and J. M. Boss, "Atac-seq on biobanked specimens defines a unique chromatin accessibility structure in naive sle b cells," *Scientific Reports*, vol. 6, p. 27030, 2016.

[30] ENCODE: "ATAC-seq Data Standards and Prototype Processing Pipeline", unter: https://www.encodeproject.org/atac-seq/ (abgerufen am 29.12.2017).

[31] A. P. Boyle, J. Guinney, G. E. Crawford, and T. S. Furey, "F-seq: a feature density estimator for high-throughput sequence tags," *Bioinformatics*, vol. 24, no. 21, pp. 2537–2538, 2008.

[32] Terry Furey Lab: "F-Seq: A Feature Density Estimator for High-Throughput Sequence Tags", unter: http://fureylab.web.unc.edu/software/fseq/ (abgerufen am 29.12.2017).

[33] S. John, P. J. Sabo, R. E. Thurman, M.-H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, and J. A. Stamatoyannopoulos, "Chromatin accessibility predetermines glucocorticoid receptor binding patterns," *Nature Genetics*, vol. 43, p. 264, 2011.

[34] University of Washington Human and Mouse ENCODE Center: "Hotspot and the SPOT data quality metric", unter: http://fureylab.web.unc.edu/software/fseq/ (abgerufen am 29.12.2017).

[35] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of chip-seq (macs)," *Genome Biology*, vol. 9, no. 9, p. R137, 2008.

[36] Juni 2011. Xiaole Shirley Liu's Lab: "Model-based Analysis for ChIP-Seq ", unter: http://fureylab.web.unc.edu/software/fseq/ (abgerufen am 29.12.2017).

[37] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park, "Design and analysis of chip-seq experiments for dna-binding proteins," *Nature biotechnology*, vol. 26, no. 12, pp. 1351–1359, 2008.

[38] N. U. Rashid, P. G. Giresi, J. G. Ibrahim, W. Sun, and J. D. Lieb, "Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions," *Genome Biology*, vol. 12, no. 7, p. R67, 2011.

[39] J. Lee, G. Christoforo, G. Christoforo, C. Foo, C. Probert, A. Kundaje, N. Boley, kohpangwei, M. Dacre, and D. Kim, "kundajelab/atac_dnase_pipelines: 0.3.3," Dec. 2016. unter: https://doi.org/10.5281/zenodo.211733 (abgerufen am 29.12.2017).

[40] aboyle. "A Feature Density Estimator for High-Throughput Sequence Tags", unter: https://github.com/aboyle/F-seq (abgerufen am 29.12.2017).

[41] rthurman. "Hotspot is a program for identifying genomic regions of local enrichment of short-read sequence tags.", unter: https://github.com/rthurman/hotspot (abgerufen am 29.12.2017).

[42] taoliu. "MACS – Model-based Analysis of ChIP-Seq", unter:https://github.com/taoliu/MACS (abgerufen am 29.12.2017).

[43] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[44] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–2, 2010.

[45] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, 2009.

[46] Y. Liao, G. K. Smyth, and W. Shi, "featurecounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–30, 2014.

[47] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[48] C. H. Birkenheuer, C. G. Danko, J. D. Baines, and R. M. Sandri-Goldin, "Herpes simplex virus 1 dramatically alters loading and positioning of rna polymerase ii on host genes early in infection," *Journal of Virology*, vol. 92, no. 8, pp. e02184–17, 2018.

7. Abbildungsverzeichnis

8. Tabellenverzeichnis

Eidesstattliche Erklärung

Hiermit versichere ich, die vorliegende Abschlussarbeit selbstständig und nur unter Verwendung der von mir angegebenen Quellen und Hilfsmittel verfasst zu haben. Sowohl inhaltlich als auch wörtlich entnommene Inhalte wurden als solche kenntlich gemacht. Die Arbeit hat in dieser oder vergleichbarer Form noch keinem anderem Prüfungsgremium vorgelegen.

Datum: _____ Unterschrift: _____

Danksagungen

Hiermit danke ich Prof. Dr. Lars Dölken für die Chance, diese Arbeit am Institut für Virologie und Immunobiologie der Universität Würzburg erstellen zu können. Weiterhin gilt Herrn Jun.-Prof. Dr. Florian Erhard ganz herzlicher Dank für das Entwerfen des Themas und die stets freundliche Unterstützung bei der Durchführung der Arbeit. Beide sind sehr vorbildliche Betreuer, welche stets bei Fragen zu Verfügung standen und Ihre Verantwortung gegenüber mir als Doktorand über das übliche Maß hinaus erfüllt haben. Zuletzt möchte ich mich noch bei dem Kollegium des Instituts für Virologie und Immunobiologie bedanken. Wohin ich auch gekommen bin, traf ich überall hilfsbereite Personen, welche immer freundlich und respektvoll miteinander umgegangen sind.