Identifying differential read density in functional genomics data

Elena Weiß

München 2024

Identifying differential read density in functional genomics data

Elena Weiß

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Elena Weiß
aus Niedernhausen

München, den 13.08.2024

Erstgutachter: Prof. Dr. Caroline C. Friedel
Zweitgutachter: Prof. Dr. Nico Pfeifer
Drittgutachter: Prof. Dr. Florian Erhard
Tag der mündlichen Prüfung: 12.12.2024

Eidesstattliche Versicherung

Hiermit erkläre ich, Elena Weiß, an Eides statt, dass die vorliegende Dissertation ohne unerlaubte Hilfe gemäß Promotionsordnung vom 12.07.2011, § 8, Abs. 2 Pkt. 5, angefertigt worden ist.

München, den 13.08.2024

. . . . . .

Elena Weiß

Acknowledgement

"You must trust and believe in people or life becomes impossible." - Anton Chekhov

I would like to thank all the people who accompanied me during my time as a PhD student being thus very special to me and of whom I shall always keep very fond memories.

First, I would like to express my deepest gratitude to Prof. Dr. Caroline Friedel for the opportunity to work on interesting scientific questions in her laboratory and write my PhD thesis. She provided me with valuable guidance right on. Our goal-oriented discussions always helped me to stay focused. As my mentor, Prof. Friedel often counterbalanced me. When I pushed too hard, she brought me back to the ground and neutralized my panic. When I lacked motivation, she set interesting and realistic goals to achieve.

I thank Prof. Dr. Nico Pfeifer and Prof. Dr. Florian Erhard for serving as reviewers for my thesis and Prof. Dr. Sven Strickroth and Prof. Dr. Marie-Christine Jakobs for being chairman and deputy chairman of my thesis committee. I would like to thank our longstanding collaboration partners Prof. Dr. Lars Dölken and Dr. Thomas Hennig who provided me with necessary data and helped me to interpret the results with their biological expertise. Not to forget, Dr. Adam W. Whisnant for reviewing my scientific publications and giving critical feedback. I felt very lucky to be a member of the DEEP-DV research unit during the entire time of my research activities, gaining a comprehensive knowledge and lots of inspiration and motivation. Our monthly meetings and annual conferences enabled me to practice the dissemination of my scientific work.

I express special thanks to Prof. Dr. Zimmer for informing all the colleagues about the latest interesting publications thus extending our horizon beyond our own research scope. Our weekly meetings were always held in a fruitful and friendly atmosphere, thanks to all colleagues. My sincere appreciation, acknowledgement, and respect to my colleague and faithful friend Armin Hadziahmetovic, who has always supported me during the hardest time of my PhD project and celebrated all the bright moments with me. I cannot help but express my deepest respect and gratefulness to our former colleague Dr. Gergely Csaba who I was lucky to have a chance to work with and to learn from. I highly admire his (bio)informatics skills, his mindset toward work and his cheerful nature in everyday life.

Last but not least, I am deeply thankful to my family who always believed in me, supported, motivated and encouraged me. My family provided me with nuggets of wisdom, helping me through my thesis and to overcome myself both professionally and personally.

Zusammenfassung

Die Einführung von Hochdurchsatz-Sequenzierungsmethoden der nächsten Generation (NGS) ermöglicht nun genomweite Untersuchungen zur Genregulation in bisher nicht gekanntem Detailgrad und Maßstab. Dies hat jedoch auch die Komplexität der Analysen erhöht und zu neuen Herausforderungen für die Bioinformatik geführt. Darunter wurde die Identifizierung von Unterschieden in der Verteilung von Reads in einzelnen genomischen Regionen, zum Beispiel in Genen oder Promotoren, bisher nur wenig adressiert.

In dieser Arbeit haben wir neuartige Methoden entwickelt, um genomische Regionen mit signifikanten Veränderungen in der Verteilung von Reads zu identifizieren. Dies wurde motiviert durch Analysen von Veränderungen im promoter-proximalen Pausieren der Polymerase II (Pol II) und der Chromatin-Zugänglichkeit bei einer lytischen Infektion mit dem Herpesvirus Typ 1 (HSV-1). Bestehende Methoden zur Analyse von Veränderungen im Pausieren von Pol II konzentrieren sich ausschließlich auf Verhältnisse von Read-Zahlen innerhalb bestimmter Fenster und können nicht zwischen einfachen Zu- oder Abnahmen des Pausierens oder komplexeren Veränderungen in der Pol II-Verteilung unterscheiden. Um diese Einschränkung zu überwinden, haben wir zuerst einen Ansatz entwickelt, der auf dem Clustern von Pol II Profilen um Promotoren basiert. Dabei zeigte sich, dass Pol II bei den meisten Wirtsgenen bei einer HSV-1-Infektion häufiger an stromabwärts gelegenen Stellen pausiert. Da diese erste Methode immer noch nicht in der Lage war, Veränderungen in der Verteilung von Reads auf Einzelgen-Ebene zu erkennen, haben wir RegCFinder entwickelt, um Unterregionen von genomischen Fenstern zu bestimmen, die Unterschiede in der Verteilung von Reads zwischen zwei Bedingungen aufweisen. RegCFinder ermöglicht es, sich auf beliebige gewünschte Regionen für jede Art von Sequenzierungsdaten zu konzentrieren. Wir haben RegCFinder angewendet, um Veränderungen der Chromatin-Zugänglichkeit und der Positionierung der Nukleosomen in Promotorregionen während einer HSV-1-Infektion zu untersuchen. Dies enthüllte eine wesentliche Veränderung der Chromatin-Architektur bei einer HSV-1-Infektion für die Mehrheit der Wirtsgene. Hier führte die Relaxation von +1-Nukleosomen zu einer Verbreiterung der zugänglichen Chromatin-Regionen an der Transkriptions-Startstelle (TSS) in Abhängigkeit vom Transkriptionsniveau in nicht infizierten Zellen.

Zusammenfassend liefert diese Arbeit neue Beiträge zur Identifizierung von Änderungen in der Read-Dichte auf Gen-Ebene für jede Art von Sequenzierungsdaten. Die Anwendung dieser Methoden lieferte darüber hinaus neue Erkenntnisse über die Infektion mit einem weitverbreiteten menschlichen Krankheitserreger.

Summary

The introduction of high-throughput next-generation sequencing (NGS) methods now allows genome-wide investigations into gene regulation at unprecedented detail and scale. Yet, this also increased complexity of analyses and has introduced new challenges for bioinformatics. One major challenge that has hardly been addressed so far is the identification of differences in read distributions on individual genomic regions, e.g., genes or promoters.

In this thesis, we developed novel methods to identify genomic regions exhibiting significant changes in read distributions. This was motivated by analyses of changes in promoter-proximal pausing of polymerase II (Pol II) and chromatin accessibility in lytic Herpes virus type 1 (HSV-1) infection. Existing methods for analyzing changes in Pol II pausing focus only on ratios of read counts within specific windows and cannot distinguish between simple increases or decreases in pausing or more complex changes in Pol II distribution. To address this limitation, we first developed an approach based on clustering Pol II occupancy profiles around promoters. This revealed a prevalent delay of Pol II pausing to more downstream sites for most host genes in HSV-1 infection. Since this first method was still incapable of detecting changes in read distributions at single-gene level, we developed RegCFinder to determine subregions of genomic windows that exhibit differences in read distributions between two conditions. RegCFinder allows to focus on any regions of interest for any type of sequencing assay. We applied RegCFinder to investigate changes of chromatin accessibility and nucleosome positioning in promoter regions during HSV-1 infection. This uncovered a major change in chromatin architecture upon HSV-1 infection for the majority of host genes. Here, relaxation of +1 nucleosome positions led to a broadening of accessible chromatin regions at the transcription start site (TSS) into downstream regions depending on transcription levels in uninfected cells.

In summary, this thesis presents novel contributions toward identifying changes in read density at gene level for any type of sequencing data. Furthermore, application of these methods provided new insights into infection with a common human pathogen.

Contents

Chapter 1

Introduction

1.1 Functional genomics

Functional genomics refers to a multitude of approaches studying a biological organism and the interplay of its individual molecular units [1–4]. Functional genomics approaches aim at determining gene functions and their specific role in the context of certain conditions [2–5]. In particular, analysis of gene expression under different contexts allows to systematically evaluate the linkage between genotypes and phenotypes and to relate events in healthy or disease-affected cells to the genomic sequence [1, 2, 6–8].

The introduction of next-generation sequencing (NGS) has proven to be crucial for functional genomics studies [3, 4, 9–11]. NGS has elevated analyses from single genes to genome-wide levels and has increased resolution to nucleotide level [6, 8]. Prior to the development of NGS, Sanger sequencing was introduced in 1977, being the gold standard for determining nucleic acid sequences for many years [12, 13]. This method utilizes natural DNA synthesis to build a complementary strand by incorporating one of the four amino nucleobases [14]. In the early days of Sanger sequencing, radioactively labeled dideoxynucleotide triphosphate (ddNTP) was used as a non-reversible chain terminator [11]. It was later simplified by using fluorescence labeling together with a laser-based detection, which increased the overall efficiency of that method [11, 15–17]. Due to this progress, entire genomes could be sequenced, which led to a rapid evolvement of the functional genomics field [2]. The first organism for which a complete genome sequence was determined was *Haemophilus influenzae* in 1995 [18]. Soon, genome sequences of several other organisms, including *Drosophila Melanogaster*, *Homo Sapiens*, and *Mus Musculus*, were determined [19].

While it was previously believed that the genome size is proportional to the complexity of the organism. The completion of the human genome project in 2003, however, made it clear that this assumption is not valid anymore [20]. This is known as the C-value paradox [3, 21–25]. Currently, the 3.2 gigabase (Gb) long human reference genome is estimated to encode for ~ 20,000 protein-coding genes [26]. Remarkably, protein-coding genes constitute the smallest part of the genome with around 59 million base pairs (bp) (~

1 % of the human genome). The largest part of the genome consists of intergenic regions and non-coding genes with biological functions that are still poorly understood [27–30].

The development of NGS revolutionized and massively sped up genome sequencing [3, 4, 9–11]. NGS introduced massively parallel sequencing of short DNA fragments, thus increasing throughput substantially [31]. The original DNA fragments are first amplified and then millions of sequencing reactions are conducted simultaneously [32]. In this way, sequencing costs are significantly reduced, while the amount of sequencing data is increased at the same time [33, 34].

The capability to decode DNA sequences at large scale resulted in various approaches to elucidate gene functions on a genome-wide scale [35]. Indeed, functional genomics considers more than just the bare DNA sequence but employs various *omics* methods [8, 36, 37]. These refer to the analyses of different molecular entities, including the genome and the transcriptome [38]. Analyses are focusing on transcription, translation, protein-protein interactions, and epigenetic regulations [1]. Following the introduction of NGS, functional genomics studies have created a huge amount of primary data obtained with different sequencing assays, some of which are outlined in Section 1.3 [8, 39, 40]. Several large projects such as the ENCODE project extensively used omics approaches to elucidate functional elements in the human genome and, as a result, created a large encyclopedia of data [41, 42]. Further large-scale projects used NGS to quantify genetic variations, to characterize the genotype-phenotype linkage, and to characterize the epigenome among others [3, 4, 43–46]. The recent introduction of third-generation sequencing (TGS) has further revolutionized the field. TGS can sequence a single molecule without the need of fragmentation or polymerase chain reaction amplification in real-time [47–50].

1.2 Biological background

Transcription of DNA into RNA represents the first step of gene expression, followed by the translation of RNA into proteins (Figure 1.1a). This sequence of events is known as the central dogma of molecular biology and is assumed to be unidirectional [51]. The transcription cycle itself consists of three phases: initiation, elongation, and termination. The key transcriptional player for most genes is the RNA polymerase II (Pol II), an enzyme synthesizing the RNA. Transcription is initiated when Pol II recognizes the transcription start site (TSS), but is quickly paused after 20 to 60 nucleotides [52]. This promoter-proximal Pol II pausing is critical, because many initiated transcripts are aborted and thus do not reach the elongation phase. In fact, the failure rate is high as only around 10 % of transcripts are elongated and reach their transcription termination sites (TTS) [53–55]. Transcription termination is triggered by a polyadenylation (poly(A)) signal sequence that is highly conserved among eukaryotes. The central part of the poly(A) signal consists of an AAUAAA sequence (or variants thereof), which is typically enclosed by specific up- and downstream sequence elements [56–58]. Recognition of the transcription termination signal causes cleavage and polyadenylation of the nascent transcript. As a consequence, nascent RNA downstream of the cleavage site is degraded and Pol II disassociates from

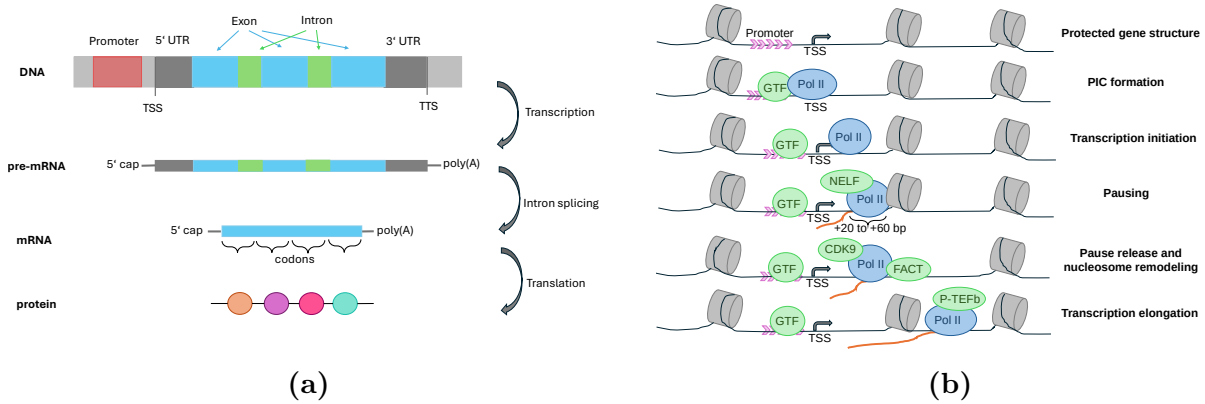


Figure 1.1 Gene structure and transcription. (a) The promoter region is located upstream of a gene and contains specific binding sites for Pol II and TFs. The TSS marks the beginning of transcription, which continues till the transcription termination site (TTS). Following transcription from 5' end to 3' end, a 5' cap and a poly(A) tail are added to the resulting pre-mRNA and introns are spliced out. The mature mRNA is then translated into a protein. (b) The DNA is winded around histones to build the nucleosome complexes (grey cylinders). First, general transcription factors (GTF) assemble at promoter regions and Pol II is recruited to form the PIC. Pol II then starts transcribing several nucleotides (nt) of DNA and quickly pauses after 20 to 60 nt, near the +1 nucleosome. Pol II pausing is maintained by the negative elongation factor NELF. The recruitment of the cyclin-dependent kinase (CDK) 9 subunit of the positive transcription elongation factor b (P-TEFb) complex then promotes pause release of Pol II. Meanwhile, the +1 nucleosome is disassembled and again reassembled after passing of Pol II, facilitated by the histone chaperone FACT. Then, Pol II continues with transcription elongation, mainly mediated by the P-TEFb.

the DNA strand [59, 60].

The transcription cycle is a highly complex process that requires many more transcription factors (TF) in addition to Pol II [61, 62]. General TFs accumulate during initiation and form the pre-initiation complex (PIC) [63–67]. The PIC recruits and guides Pol II to the TSS (Figure 1.1b) [68]. Specific binding sites in the promoter region define the position where TFs can bind and hence are important for either enhancing or suppressing gene expression [3, 69, 70]. Through their ability to bind DNA in a sequence-specific manner, TFs are major players in transcription regulation [71]. By (de)acetylation of histones, TFs are also capable to modify the chromatin structure [72] and their binding to DNA sequences can impact nucleosome positions [73, 74].

Notably, nucleosomes also impact the transcription cycle, because they represent transcription barriers for Pol II and must be disassembled and reassembled during transcription [75–77]. A nucleosome represents the basic unit of chromatin, a complex of DNA wrapped

around a histone (Figure 1.1b) [78–80]. Histones are octamer protein complexes consisting of the subunits H3, H4, H2A, and H2B, and are linked to the DNA via the H1 histone [80–82]. Promoter regions are generally depleted from nucleosomes and thus referred to as nucleosome-free regions (NFR). In contrast, gene bodies are covered with nucleosomes. In general, nucleosomes are very well-positioned around promoters and less so with increasing distance to the TSS [83–86]. The very first nucleosomes up- and downstream of the TSS are referred to as -1 and +1 nucleosomes, respectively. Depending on nucleosome compaction, the chromatin state is commonly classified as closed (heterochromatin) or open (euchromatin) and the chromatin structure also impacts the transcription regulation [87–91]. Closed chromatin prevents TF to bind in promoter regions and can lead to gene silencing [87]. Thus, nucleosomes contribute to gene regulation.

Within the scope of this thesis, changes in the transcription cycle and the chromatin architecture for the host during infection by herpes simplex virus type I (HSV-1) were investigated. HSV-1 is a large DNA virus that belongs to the family of *herpesviridae* with more than hundred herpesviruses [92–94]. In addition to HSV-1, there are only eight other herpesviruses infecting humans [95–97]. Due to its short replication cycle, HSV-1 is often taken as model virus to study the early phase of DNA virus infection. HSV-1 is known for causing cold sores and fever blisters [98–100], but it can also cause lethal diseases like encephalitis [101–104]. To date, HSV-1 infection is still not curable, but its symptoms can be treated [105]. Once HSV-1 infects a host, it persists for the lifetime of the host [106]. According to the World Health Organization (WHO), more than two thirds of the world population aged below 50 are infected by HSV-1, with similar rates estimated by the Robert Koch Institute in Germany [107, 108]. The virus is highly infectious and orally transmitted across epithelial cells [98, 109–111]. When infecting a cell, the viral envelope merges with the host membrane and releases its capsid into the cytoplasm [98, 110]. The virus can reach and infect host neuronal cells of the nervous system where it remains in an inactive (latent infection) state [112]. Reactivation from latency can be triggered by different conditions such as stress or a weakened immune system [113–120]. This remarkable characteristic of Herpesviridae to switch between active (lytic) and latent infection is the target of many investigations but yet poorly understood [114, 121].

The HSV-1 virus has a linear double-stranded DNA genome that is divided into two unique segments and four repeat regions [110, 122, 123]. It encodes for at least 121 open reading frames (ORF) with more than hundred gene products [110, 124]. The genes are classified into immediate early (IE), early (E), and late (L) groups according to the point of time of their expression [110, 125]. In general, IE genes (ICP0, ICP4, ICP22, ICP27, and ICP47) are responsible for recruiting host transcription factors and regulatory proteins to start viral gene expression [126–128]. In particular, the viral gene ICP27 is known for interfering with Pol II transcription termination downstream of 3' gene ends resulting in disruption of transcription termination (DoTT) [129]. Recently, it was uncovered that read-through transcription following DoTT is accompanied by increased chromatin accessibility downstream of 3' gene ends (downstream open chromatin region (dOCRs)), requiring the viral gene ICP22 [130]. In addition to the effects of the viral IE genes, the viral host shutoff (*vhs*) protein contributes to the host transcriptional shutoff by degrading host and viral

mRNAs [98, 131].

1.3 Functional genomics assays important for this thesis

RNA sequencing (RNA-seq) quantifies gene expression on the RNA level by measuring transcript abundancies [36, 132, 133]. To make use of the whole-genome sequencing techniques described above, RNA is first reverse transcribed into complementary DNA (cDNA), which is then sequenced [134, 135]. RNA-seq is usually applied under different conditions, so that differentially regulated genes can be discovered. More recent sequencing techniques such as global run-on sequencing (GRO-seq) or precision nuclear run-on sequencing (PRO-seq) focus on mapping actively engaged Pol II sites with base pair resolution, especially at gene 5' ends [136–138]. These methods can also be used to analyze Pol II pausing or to measure the transcription rate of Pol II [139].

Identification and quantification of protein-DNA interactions that contribute to gene regulation can be performed by chromatin immuno-precipitation coupled to DNA sequencing (ChIP-seq) [36, 140, 141]. For this purpose, DNA-binding proteins are first crosslinked to DNA and chromatin is then fragmented. An antibody specific to a selected DNA-binding protein is used to immunoprecipitate the DNA-protein complex [142]. After that, the complex is purified to obtain the DNA, which is then sequenced. Alignment of reads to the genome then reveal the protein binding sites [143]. ChIP-seq can also be used to analyze histone modifications, DNA methylations, or nucleosome localizations [85, 144, 145]. Epigenomic states can be inferred from ChIP-seq of histone modifications [146]. For genome-wide quantification of chromatin accessibility, ATAC-seq (assays for transposase-accessible chromatin coupled to high-throughput sequencing) can be used [3, 147–150]. This method utilizes the hyperactive Tn5 transposase to target accessible DNA by cutting and inserting specific adaptors for sequencing, known as tagmentation [87, 151, 152]. Recently, ChIP-mentation was introduced as an improvement to ChIP-seq which combines tagmentation with immunoprecipitation [153]. Tagmentation is carried out directly on bead-bound chromatin, so that only DNA purification is required beforehand. Consequently, it offers a rapid solution with minimal cell number requirements and high resolution to map transcription factor binding sites and histone modifications [153].

1.4 Challenges for bioinformatics analyses of functional genomics data

1.4.1 Standard approaches

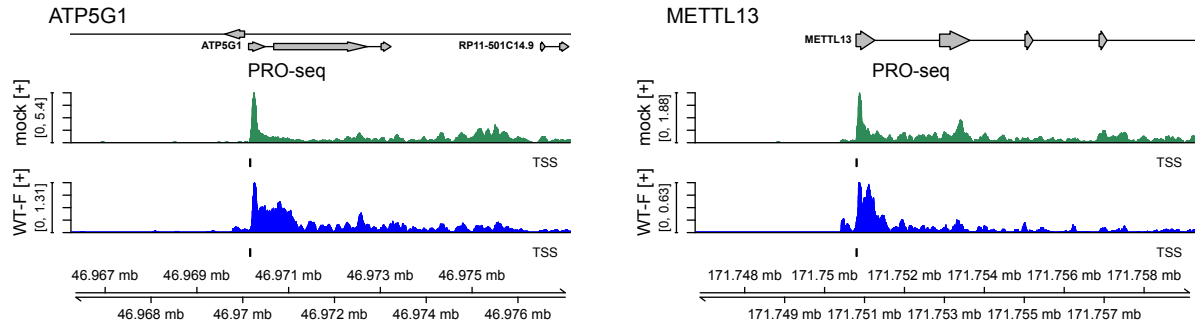
Regardless of the sequencing assay being utilized, resulting raw sequencing reads are first preprocessed, which usually includes quality control, read trimming, and finally read alignment [154–156]. Here, the introduction of NGS has led to new challenges for data handling

and bioinformatics analyses, in particular for mapping sequencing reads to the genome [157, 158]. Fragments obtained by NGS are shorter but more abundant and, therefore, accurate and fast methods had to be developed to map the fragments back to their original location on the DNA sequence [9, 11, 31, 32, 40, 159–162]. Common alignment tools include Bowtie [163] or BWA [164] for short-read genome alignment, and STAR [165], TopHat [166], or ContextMap [167] for spliced read alignment [161]. The mapping provides each read with its genomic position, so that read counts per gene or transcript can be determined. Usually, read counts are the basis for further downstream analyses investigating gene function, regulation, and biological mechanisms [37]. Downstream analysis commonly includes differential expression analysis (DEA) [168, 169]. Several tools for DEA are available, including DESeq2, edgeR, and limma [168–170]. These tools mainly differ in the way dispersion is estimated [168].

DEA methods for RNA-seq data estimate log2 fold-changes and the statistical significance based on the change in total read counts for a particular feature, e.g., genes or promoter regions. However, if only the distribution of reads for the corresponding genomic window is altered without any significant change in total number of reads (e.g., due to Pol II pausing changes), these changes are not detected. Differential methods for ChIP-seq data analyses suffer from the same problem. Identification of TF binding sites is usually carried out by peak calling methods [171], e.g., MACS2 [172], BCP [173], or GEM [174], whereas genomic annotation of peaks is typically achieved by HOMER [175], MEME [176], or ChIPseeker [177]. Some approaches first identify peaks by using peak calling algorithms and then assess differences in read counts within peaks between different conditions by using DEA methods [178], e.g., DiffBind [179], or PeakSeq [180]. Here, only total counts for peak regions are evaluated, while changes in read distributions within these peaks are neglected. A second class of ChIP-seq analyses approaches use a segmentation approach that relies on a sliding window, e.g., diffReps [181], THOR [178], or CisGenome [182]. Again, only the total number of read counts within the current sliding window are considered, so that changes in the distribution of reads of with the same or similar number of total reads remain undetected [178, 181, 183].

1.4.2 Pausing index

One approach that explicitly aims to quantify changes in distributions is the pausing index (PI), which was specifically developed to quantify changes in promoter-proximal Pol II pausing. PIs are calculated as the ratio between normalized read counts in the promoter window and those of the gene body window. For this purpose, normalized read counts can, e.g., be determined as number of reads per kilobase per million mapped reads (RPKM) [184]. The promoter window p typically includes the TSS, while the gene body window b includes the gene without the promoter window. Importantly, promoter and gene body windows lack an exact definition and choices of promoter and gene body windows strongly affect PI analysis. In general, PI for a gene g is calculated as follows:



(a) **(b)**

Figure 1.2 Changes in Pol II pausing for selected genes in HSV-1 infection. Read coverage for gene promoters in PRO-seq data on the sense strand for uninfected (green) and HSV-1 strain F (WT-F) infected (blue) cells for the genes (a) ATP5G1 and (b) METTL13. Read coverage is normalized to a total number of mapped reads and averaged between replicates. The used TSS is indicated by the vertical line below read coverage tracks. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns, and direction is indicated by the arrowhead. The genomic coordinates are shown at the bottom. Read coverage plots were created with the R Bioconductor Package *Gviz* [189].

$$PI(g) = \frac{RPKM_{[p_{start},p_{end}]}}{RPKM_{[b_{start},b_{end}]}}. \tag{1.1}$$

Unfortunately, PI analysis only reports changes in the distribution of reads between promoter and gene body windows, but fails to detect altered read count distributions within specified genomic windows. Moreover, a reduction of PI, normally interpreted as a loss of Pol II pausing, can also originate from delayed pausing beyond the promoter window.

We previously utilized PI analysis to analyze Pol II pausing changes upon inhibition of CDK11 [185]. CDK11 is a kinase that phosphorylates the C-terminal domain of Pol II [186]. PI analysis showed an increase of pausing upon CDK11 inhibition for more than 70 % of genes. PI analysis was also utilized to analyze pausing changes in HSV-1 infection, which is the subject of the first article included in this thesis [128, 187, 188]. Here, we applied PI analysis for a first assessment of Pol II pausing changes after 3 h of HSV-1 infection. Indeed, while this showed a reduction in PI for around 90 % of host genes, inspection of read coverages for the corresponding PRO-seq data showed that the reduction in PI did not simply represent reduced pausing (illustrated in Figure 1.2). Reduced PI for the gene ATP5G1 reflected a broadened Pol II occupancy downstream of the TSS in read coverage data (Figure 1.2a). Meanwhile, the apparently reduced PI for the gene METTL13 reflected a secondary pause site downstream of the TSS in read coverage data (Figure 1.2b). Thus, PI analysis was not able to distinguish between simple reductions in occupancy and more complex alterations.

1.4.3 Metagene analysis

A more widely used method to analyze changes in the distribution of reads is the metagene analysis, which displays the average read distribution over a set of genes for specified genomic regions (for an example see Figure 2.1b). This method is applicable to any number of conditions. Metagene analyses have the advantage that they can illustrate various changes of read distributions. This allows identifying general trends but has the drawback that changes affecting only a minority of genes can be missed. Unfortunately, the method is only applicable for sufficiently large gene sets due to the poor signal-to-noise ratio per gene. Hence, metagene analyses cannot be performed at the level of individual genes in contrast to PI analysis. Our group previously utilized metagene analysis to investigate effects of CDK12 inhibition and could show that long genes suffer from a Pol II processivity defect [190]. Furthermore, we have applied metagene analysis to inspect Pol II pausing changes upon HSV-1 infection first for all host genes, which confirmed a general decrease of Pol II at the TSS and a broadening of Pol II signal into downstream regions [188]. However, more complex deviations from that general change in Pol II occupancy affecting only a subset of genes were not identified. We could solve this issue by clustering genes according to their read distributions and subsequently applying metagene analyses on these grouped gene sets as outlined in Section 2.1.

Identification of differential genomic regions between conditions has also been pursued by methods developed for identifying differential chromatin modifications such as diffReps [181]. diffReps conducts a *de novo* search for differential chromatin domains (DCD) utilizing a sliding window approach [181]. Windows with significant differences in total number of reads are selected and overlapping windows are subsequently merged. Unfortunately, this approach does not identify changes in the distribution of reads within the window. Moreover, the *de novo* discovery cannot be targeted to genomic regions and hence, multiple testing correction has to be very stringent due to the high number of performed tests. Similarly, methods such as XCAVATOR [191] for investigating copy number variants (CNV) based on read depth (RD) aim at identifying differentially covered genomic regions. However, their underlying assumption for CNV events differs from changes commonly observed in functional genomics assays as CNV events are characterized by sudden shifts in RD between consecutive genomic windows [191]. Ultimately, none of the available methods allowed analyses of read distribution changes suited for our purpose. State-of-the-art methods either lacked gene level analysis or were only capable of identifying changes in total number of reads rather than changes in distributions of reads. For this reason, we developed a novel method called RegCFinder which is presented in the second article included in this thesis and applied in the third article [192].

1.5 Thesis outline

In this thesis, I developed and applied different methodologies to investigate changes in read distributions in functional genomics data which resulted in the three articles included

in this thesis. Here, the focus was on changes in Pol II pausing and chromatin accessibility at promoter regions upon HSV-1 infection. In the following, I briefly summarize the content of each contributing article. My contribution to the work is described in Section 1.6.

Contributing article for Section 2.1

Weiß E, Hennig T, Graßl P, Djakovic L, Whisnant A.Ws., Jürges C.S., Koller F, Kluge M, Erhard F, Dölken L, and Friedel C.C. HSV-1 Infection Induces a Downstream Shift of Promoter-Proximal Pausing for Host Genes. *Journal of Virology*, 97(5):e00381–23, April 2023

In the first article outlined in Section 2.1, we examined changes in promoter-proximal Pol II pausing upon HSV-1 infection compared to uninfected cells using previously published PRO-seq data. Usually, Pol II pauses shortly downstream of the TSS which is also reflected as a strong and well-defined peak of Pol II occupancy near the TSS in PRO-seq data. Previous studies on Pol II pausing in HSV-1 infection reported a reduction of Pol II pausing at promoters [127, 128]. Our initial PI analysis also suggested a loss of pausing. However, metagene analysis in combination with hierarchical clustering uncovered more complex changes in Pol II pausing behavior than anticipated so far. During infection, Pol II occupancy peaks broadened and/or shifted toward downstream regions. By refining the well-known metagene analysis by a preceding clustering of genes, we were able to identify a new aspect of transcriptional changes upon HSV-1 infection for host genes. We conducted further downstream analyses and could exclude several biological mechanisms as the cause for delayed Pol II pausing.

Contributing article for Section 2.2

Weiß E and Friedel C.C. RegCFinder: targeted discovery of genomic subregions with differential read density, *Bioinformatics Advances*, 3(1):vbad085, June 2023

In Section 2.2, I present RegCFinder, our novel bioinformatics method to identify genomic subregions with differences in read distributions. RegCFinder operates on a single-gene level and is targeted by specifying genomic windows of interest as input. The method is realized as a workflow for the workflow management system (WMS) Watchdog and can be easily used by bioinformation and wet lab scientists with limited programming expertise due to automatic deployment of software dependencies with Conda [193, 194]. We demonstrated the wide applicability of RegCFinder by applying it on two different real-world scenarios. With the help of RegCFinder we reanalyzed PRO-seq data of HSV-1 infection regarding Pol II pausing, which indeed confirmed the general downstream shift of Pol II pausing. However, gene-level results revealed very different extents in the delay of Pol II pausing. Furthermore, we could distinguish a subset of genes with read-in transcription as well as a set of genes with proper pausing loss. Likewise, re-analysis of the ChIP-seq data for CDK12 inhibition confirmed that long genes are preferentially affected by a loss of Pol II at 3' gene ends [190].

Contributing article for Section 2.3

Weiß E, Whisnant A.W., Hennig T, Djakovic L, Dölken L, and Friedel C.C. HSV-1 infection induces a downstream shift of the +1 nucleosome, *bioRxiv 2024.03.06.583707; doi: https://doi.org/10.1101/2024.03.06.583707*

In the article described in Section 2.3, we investigated changes in chromatin accessibility in promoter regions upon HSV-1 infection. By applying RegCFinder to a large ATAC-seq data set of wild type (WT) and null-mutant HSV-1 infections, we uncovered three major patterns of chromatin changes: increased chromatin accessibility either downstream or upstream of the TSS, and decreased chromatin accessibility around the TSS. Through further downstream analyses of gene expression data, we were able to show that increases upstream of the TSS effectively also represented downstream changes, only for genes located on the opposite strand in bidirectional promoters. Moreover, the extent of changes in chromatin accessibility is dependent on gene expression prior to infection. Since chromatin accessibility is tightly associated with nucleosome positioning, we further investigated ChIPmentation data of the noncanonical histone variant H2A.Z, which is enriched at +1 and -1 nucleosomes [195]. This uncovered that +1 nucleosomes are shifted downstream upon HSV-1 infection, thus extending the NFR at promoter regions.

1.6 Contribution

I was the leading bioinformatics author on all three contributing articles. Under the supervision of Prof. C. Friedel, I developed and applied methods for bioinformatics analyses of published experimental data and data provided by collaboration partners from the laboratory of Prof. L. Dölken (T. Hennig, L. Djakovic, A.W. Whisnant). For the first article [188], I substantially extended preliminary analyses by P. Graßl performed in her Bachelor's Thesis. Furthermore, I supervised F. Koller in her Bachelor's Thesis, which provided sequence motif analyses for that article. M. Kluge provided support for metagene analyses and C. S. Jürges and Prof. Erhard provided transcript start site annotations. For the second article [192], I developed and implemented the RegCFinder method based on suggestions and under the supervision of Prof. C. Friedel and performed the evaluation of RegCFinder on published sequencing data of PRO-seq and ChIP-seq assays. For the third article [196], I applied the newly developed RegCFinder on published ATAC-seq data and H2A.Z ChIPmentation data provided by the collaboration partners mentioned above. For all three articles, I wrote the first draft of the articles. I was supported in the revision by C. Friedel, L. Dölken, and T. Henning for the first article, by C. Friedel for the second article, and by C. Friedel, A. W. Whisnant, T. Hennig, and L. Dölken for the third article.

Chapter 2

Summary of contributing articles

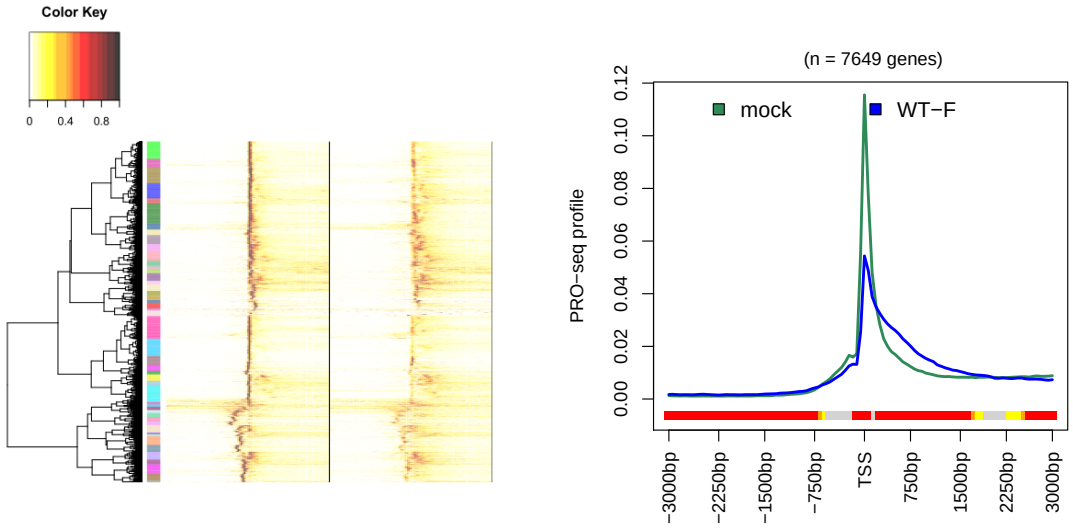
2.1 Changes in promoter-proximal Pol II pausing during HSV-1 infection

2.1.1 Biological motivation

Previous studies of HSV-1 infection showed a wide-spread shutoff of host transcription mediated by viral IE genes and enhanced by the viral *vhs* protein [129, 131, 197]. Host transcription shutoff is particularly affected by a global reduction of Pol II occupancy from host genes [126–128, 198]. The article presented in this section [188] was based on a re-analysis of published PRO-seq data for mock and WT-F HSV-1 infection at 3 h post infection (p.i.) from a study by Birkenheuer *et al.* [128]. This re-analysis was initially motivated by manual inspection of mapped reads for individual genes in a genome browser (IGV) which showed different types of changes in read distributions at gene starts (Figure 1.2).

2.1.2 TSS identification during HSV-1 infection

For an accurate genome-wide analysis of changes at gene starts, the used TSS first had to be precisely and correctly identified for each gene. Because genes can have more than one annotated transcript start, we determined the dominant TSS per gene based on PRO-seq and PROcap-seq data of flavopiridol-treated uninfected human foreskin fibroblasts (HFF) [199]. Flavopiridol inhibits CDK9 which is necessary for Pol II pause release. Thus, Pol II is arrested in a paused state at the TSS and TSS identification is facilitated [195, 200–203]. Candidate TSS were first identified using the iTiSS program developed by Jürges et al. [204]. Then, the candidate TSS were reduced to a consensus set that only included consistent start sites confirmed by both data sets. In addition, TSS were restrained to be located near an annotated gene start ($\leq$ 500 bp) to obtain high confidence sites. Finally, 7,650 unique TSS were selected using only the highest expressed TSS for each gene. The selected TSS matched very well the Pol II promoter peaks in the PRO-seq data by Birkenheuer *et*



(a) **(b)**

Figure 2.1 Transcription start sites in PRO-seq data. (a) Heatmap showing Pol II occupancy profiles in PRO-seq data on sense strand for mock and WT-F infection (separated by vertical black line) after hierarchical clustering analysis. For this purpose, PRO-seq profiles for mock and WT-F infection were first concatenated and then clustered according to Euclidean distances and Ward's clustering criterion. The cutoff on the hierarchical clustering dendrogram was chosen to obtain 50 clusters (marked by colored rectangles between the dendrogram and heatmap). Pol II profiles for all analyzed genes are shown for a 6 kb promoter window centered around the TSS identified in PROcap-seq and PRO-seq data. (b) Metagene plot of PRO-seq profiles in sense direction in the 6 kb promoter windows for mock (dark green) and WT-F HSV-1 infection (dark blue). The colored bar at the bottom indicates the significance of paired Wilcox tests for each position comparing normalized coverage between the two conditions. *P*-values were adjusted for multiple testing with the Bonferroni method; color code: red = adj. *P*-value $\leq 10^{-15}$, orange = adj. *P*-value $\leq 10^{-10}$, yellow = adj. *P*-value $\leq 10^{-3}$.

al. (Figure 2.1a), better than to gene starts annotated by Ensembl.

2.1.3 Identifying groups of genes with altered PRO-seq profiles

As a first analysis, we performed standard PI analysis to assess the differences in Pol II pausing between mock and WT-F infection. This analysis identified a reduced PI for WT-F in comparison to mock for the majority of host genes (~ 90 %). However, as outlined in the introduction, PI analysis cannot distinguish between a proper loss of pausing and other pausing changes. In the next step, a metagene analysis suggested more complex alterations in the Pol II occupancy (Figure 2.1b). Instead of a mere reduction of Pol II occupancy at the TSS, Pol II occupancy was significantly broadened toward downstream

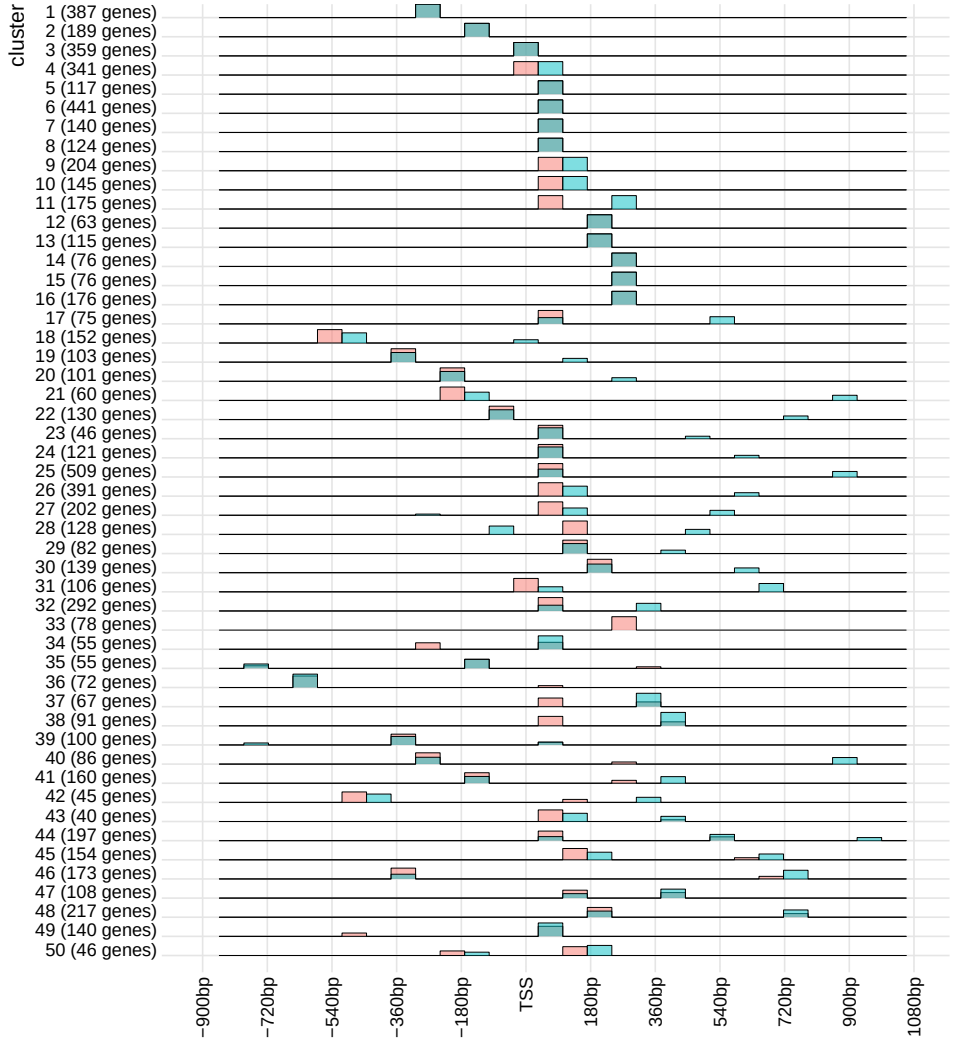
sites. Unfortunately, metagene analysis on all genes did not allow distinguishing subgroups of genes with distinct changes. Thus, we combined metagene analysis with a clustering approach to identify genes with similar changes in read distributions. We applied hierarchical clustering on concatenated Pol II promoter occupancy profiles for mock and WT-F infection of individual genes and selected a cutoff on the clustering dendrogram that resulted in 50 clusters identified by manual inspection of the clustering dendrogram (Figure 2.1a). This analysis identified distinct changes of Pol II pausing, including secondary minor pause sites originating downstream of TSS in HSV-1 infection (as shown for the selected gene in Figure 1.2b). The other identified patterns of changes included reduced pausing at the TSS but broadened Pol II occupancy in downstream direction, two equally high pausing peaks, and a second pause site downstream higher than at the TSS.

To characterize the 50 gene clusters, we identified the main peaks of Pol II occupancy within the promoter window. For this purpose, the major peak was first identified as the global maximum. Subsequently, the next highest local peaks up- and downstream of the major peak were identified using the find_peaks function in the R *ggpmisc* package. Furthermore, secondary peaks were required to be sufficiently far from the borders of the 6 kb promoter window (i.e., between -1,800 bp and +4,800 bp relative to the TSS). In addition, the difference between the height of the secondary peak and the minimum value between the major and secondary peak should reach at least 10 % of the major peak height, while the height of the secondary peak should reach 20 % of the major peak. If secondary peaks reached 95 % of the major peak height, both peaks were considered as equally high. 28 clusters showed new secondary pause sites downstream of the TSS with a mean distance of 480 bp to the major peak (Figure 2.2). For eight of these clusters, the secondary downstream peak was at least as high as the major peak. In total, pausing at the original TSS was reduced for 48 clusters and broadened toward downstream sites for 45 clusters (Figure 2.2a).

To analyze a possible correlation between gene function and changes in Pol II pausing, over-representation analysis for Gene Ontology (GO) terms was performed for each cluster [205, 206]. However, no significant enrichment for biological functions or molecular processes could be identified. Enrichment analysis for transcription factor binding motifs from TRANSFAC showed that cluster 32 was strongly enriched for G-rich or C-rich motifs. As GC content and GC skew have been shown to be correlated to pausing [187], we investigated the GC content for each cluster. While this confirmed GC-richness of pause sites, the GC content did not appear to be correlated to distinct pausing changes.

2.1.4 Integration with further functional genomics data

To investigate the time when the Pol II pausing changes manifested during infection, we studied two more recently published PRO-seq studies by Birkenheuer *et al.* [198, 207] which included three time points (1.5 h, 3 h, 6 h) of HSV-1 WT-F infection [198]. Metagene analysis for each time point for the same 50 clusters revealed that Pol II occupancy was shifted downstream at 3 h and 6 h p.i. While changes between 3 h and 6 h p.i. were negligible, significant differences were observed between 1.5 and 3 h p.i. with decreased



(a)

Peak pattern	mock		WT-F 3 h p.i.	
	no. clusters	no. genes	no. clusters	no. genes
One TSS peak	33	5728	21	3347
Additional minor peak downstream of major TSS peak	13	1461	17	2986
Two approximately equally high peaks	2	232	4	296
Additional downstream peak higher than TSS peak	0	0	7	834
none of the above	2	228	1	46

(b)

Figure 2.2 Pol II peak identification across clusters. (a) Positions and relative heights of peaks identified in PRO-seq profiles in sense direction in mock (light red) and WT-F (turquoise) HSV-1 infection for all 50 clusters. Darker turquoise indicates a common peak between mock and WT-F infection. The relative peak height was computed by dividing the peak height by the sum of all peak heights for the same condition. (b) Statistics on the number of clusters and number of genes with different types of identified peaks across the clusters. Identified peaks are classified according to their number, location, and relative height.

occupancy at the major peak and increased broadening downstream of the TSS peak at 3 h p.i. compared to 1.5 h p.i. In addition, the study by Birkenheuer *et al.* [198, 207] included infection with the ΔICP22 null mutant and its repair virus for the same three time points. ICP22 was previously identified to inhibit host transcription elongation by interacting with CDK9 [186, 208]. Therefore, we investigated whether the presence of ICP22 was necessary for changes in Pol II pausing in HSV-1 infection. However, both ΔICP22 and repair virus infection showed similar alterations in Pol II pausing. Generally, no significant differences were detected between ΔICP22 and repair virus infection at 3 h and 6 h p.i. with the only exception being two clusters where the TSS peak was slightly reduced in ΔICP22 infection. Thus, ICP22 was not required to induce delayed Pol II pausing downstream of the TSS.

Furthermore, the additionally emerging Pol II pausing sites during HSV-1 infection could result from alternative transcription initiation. Therefore, we investigated the presence of alternative TSSs for all clusters either detected in flavopiridol-treated PROcap-seq or PRO-seq data or being present in the human genome annotation from Ensembl. For most clusters, less than 15 % of genes showed an alternative TSS at positions with additional Pol II peaks. This was independent of whether experimentally or annotated TSS were considered. In addition, we re-analyzed recently published data of cRNA-seq and directional RNA-seq (dRNA-seq) data, which are enriched for 5' transcript ends, of mock and HSV-1 WT strain 17 (WT-17) infection [124]. Here, cRNA-seq was performed for 1, 2, 4, 6, and 8 h post HSV-1 infection and dRNA-seq was performed for 8 h post HSV-1 infection. Metagene analyses for all clusters showed peaks co-located with the TSS peaks in PRO-seq data present in mock infection for both data sets. However, no peaks were detected that coincided with newly emerging PRO-seq peaks downstream of the TSS upon HSV-1 infection.

Pol II pausing is mediated by its key regulator, the negative elongation factor NELF [209–211], and rapid depletion of NELF induces a shift of pausing sites toward downstream regions [210, 211]. Since NELF is depleted from some host genes in HSV-1 infection, NELF depletion could cause the delay in Pol II pausing [126]. We thus reanalyzed data for 0, 1, 2, and 4 h auxin-induced NELF degradation [210] and again performed metagene analyses of our 50 clusters. This showed that changes in Pol II pausing could not be explained by the loss of NELF.

2.2 Identifying regions with differential read density using RegCFinder

2.2.1 Motivation and overview

The second article included in this thesis presents RegCFinder, which we developed to address the problems experienced with PI and metagene analysis [192]. The main objective of RegCFinder is to identify regions with differential read density at single-gene level. It can process any type of functional genomics sequencing assay as input, e.g., ATAC-seq or ChIP-seq. By specifying certain genomic regions as input, RegCFinder can be targeted

to a wide range of applications. This not only includes promoter regions but also gene ends or viral genomes. RegCFinder is implemented as a workflow for the WMS Watchdog [193, 194] and the required software is deployed using Conda [212]. If a computing cluster is available, Watchdog can be used to schedule the input regions to be analyzed in parallel. RegCFinder makes use of the linear solution to the well-known "all maximum scoring subsequences" (AMSS) problem and is thus highly efficient [213]. It allows the comparison of two conditions, each having two or more replicates, and uses DEXSeq for the statistical analyses of the identified differential regions to provide them with fold-changes and significance [214]. Analyses with more than two conditions require the comparison of each condition with a common reference.

2.2.2 Method

RegCFinder takes aligned sequencing data in BAM format and a set of windows of interest W (genomic regions in adapted BED format) as input. It is designed for two conditions (i.e., c_1: control, c_2: test) with two or more replicates each (samples s_{11}, ..., s_{1k} and s_{21}, ..., s_{2k} with $k \geq 2$ as the number of replicates). For each sample s_{ck} and each input window $w \in W$, read counts per sequence position $r_{cs}^{w}(i)$ are determined and then normalized to obtain the read density $d_{cs}^{w}(i)$:

$$d_{cs}^{w}(i) = \frac{r_{cs}^{w}(i)}{\sum_{i \in w} r_{cs}^{w}(i)} \text{for } s \in [1:k], i \in w. \tag{2.1}$$

The read densities $d_{cs}^{w}(i)$ for each replicate are then averaged to read densities $d_{c}^{w}(i)$ for each condition $c \in \{1, 2\}$:

$$d_{c}^{w}(i) = \frac{1}{k} \sum_{s=1}^{k} d_{cs}^{w}(i) \text{ for } i \in w. \tag{2.2}$$

Based on these read densities, RegCFinder then aims to identify subregions of the input windows where the density is higher in one condition than in the other. In a perfect case, where read distributions look like those exemplified in Figure 2.3a, differential subregions could simply be determined by searching for the intersection points of the two curves. Unfortunately, read distributions are noisier and look more similar to the example in Figure 2.3b. Therefore, we formulated the problem as an instance of the all maximum scoring subsequences (AMSS) problem [213]. This problem searches for all nonoverlapping maximal scoring subsequences (MSS) on a sequence $X = (x_1, \ldots, x_n)$ of real numbers. A subsequence m of X is an MSS if two requirements are fulfilled:

1. All subsequences of m must have a smaller score than m

2. No supersequence of m fulfills the first condition.

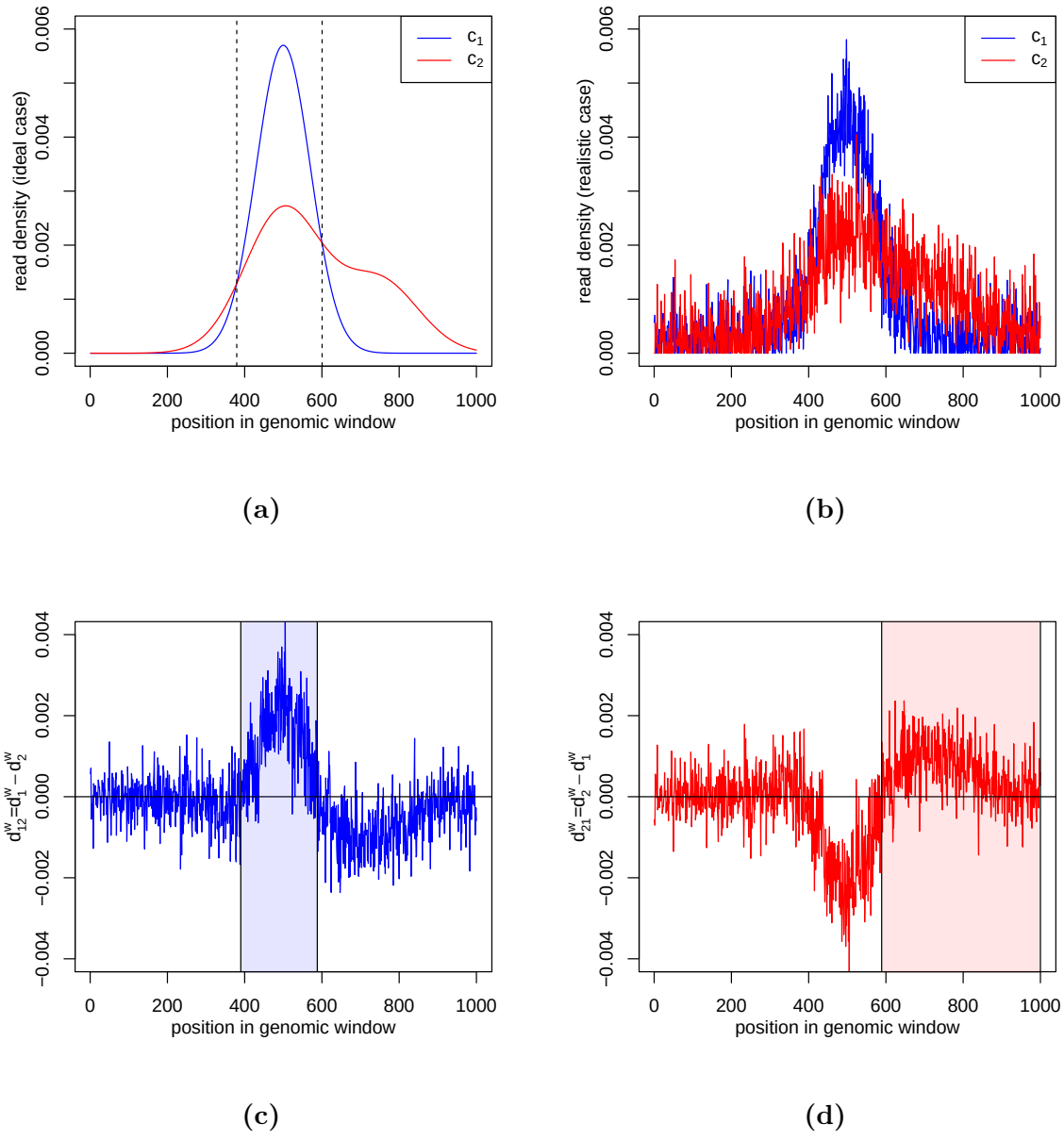


Figure 2.3 Identification of subregions with differences in read distributions. (a) Simplified illustration of ideal read densities for two conditions. (b) Illustration of realistic read distributions with noise. (c, d) Differential regions are identified by calculating $d_{12}^w = d_1^w - d_2^w$ (c) and $d_{21}^w = d_2^w - d_1^w$ (d) and then identifying regions with predominantly positive values in these sequences (shaded regions).

For this purpose, the score $S_{i,j}$ for a subsequence is defined as the sum of each element within this corresponding subsequence $(x_i, \ldots, x_j)$ of X:

$$S_{i,j} = \sum_{l=i}^{j} x_l \tag{2.3}$$

Notably, every positive element is contained in one MSS. A linear time algorithm for this problem was developed by Ruzzo and Tompa in 1999 [213]. To apply the MSS problem, the read densities $d_c^w(i)$ for the two conditions are subtracted from each other to calculate the two sequences $d_{12}^w := d_1^w - d_2^w$ and $d_{21}^w := d_2^w - d_1^w$. This is illustrated in the Figures 2.3c and 2.3d. In addition, a pseudo-count adjusted by a parameter ρ (default $\rho = 1$) is subtracted from each element in the sequences. We introduced this pseudocount to prevent RegCFinder from identifying long, meaningless MSS with many zero elements surrounded by positive elements due to sparse read counts (e.g., $Y = (1, 0, 0, 0, 0, 0, 0, 0, 0, 1)$). ρ can be used to tune the length of resulting MSS, with high values of ρ leading to shorter MSS. The final two sequences for computation of MSS are then defined as follows:

$$X_{st}^w(i) = 100 \times d_{st}^w(i) - \frac{\rho}{|w|} \text{ for } 1 \leq i \leq |w|,\ s, t \in \{1, 2\}, s \neq t, \tag{2.4}$$

Read densities are multiplied by 100 to obtain larger numbers. As a consequence, long MSS are penalized by a linear function $p(\lambda) = \frac{\rho}{|w|}(\lambda)$, where λ is the length of the MSS. Application of the algorithm by Ruzzo and Tompa [213] then generates two sets of MSS: M_{12} for X_{12}^w, and M_{21} for X_{21}^w. In the next step, short MSS consisting only of one or a few positive elements are filtered. For this purpose, a randomization approach is implemented in RegCFinder for which each of the input sequences X is repeatedly (by default 1000 times) permuted randomly (Figure 2.4a). MSS are then determined for each of the randomly permuted sequences (Figure 2.4c). The set of MSS from the original sequences (Figure 2.4b) are then filtered, keeping only MSS with scores better than the maximum for any MSS on the randomized sequences (Figure 2.4d). The final set of MSS for an input window w is obtained by merging the filtered MSS M_{12} and M_{21} into a final set M. In case of overlaps, the MSS with the higher score is retained.

Finally, the significance of each MSS $\in M$ is assessed by DEXSeq to obtain fold-changes and P-values. Originally, DEXSeq was developed to identify differential exon usage [214]. To meet the requirements of DEXSeq, RegCFinder creates a new extended annotation file. Here, genes are the initial input windows $w \in W$ and exons are defined as identified MSS from M as well as filler regions between these subregions that are defined as introns. In this way, significance of each MSS is estimated with regards to the background window w. In the end, RegCFinder provides a table with all identified MSS $\in M$ (also denoted as regions of change, short RegC) per input window $w \in W$, and the corresponding information about significance, fold-change, genomic coordinates, score, and the condition in which read density is higher.

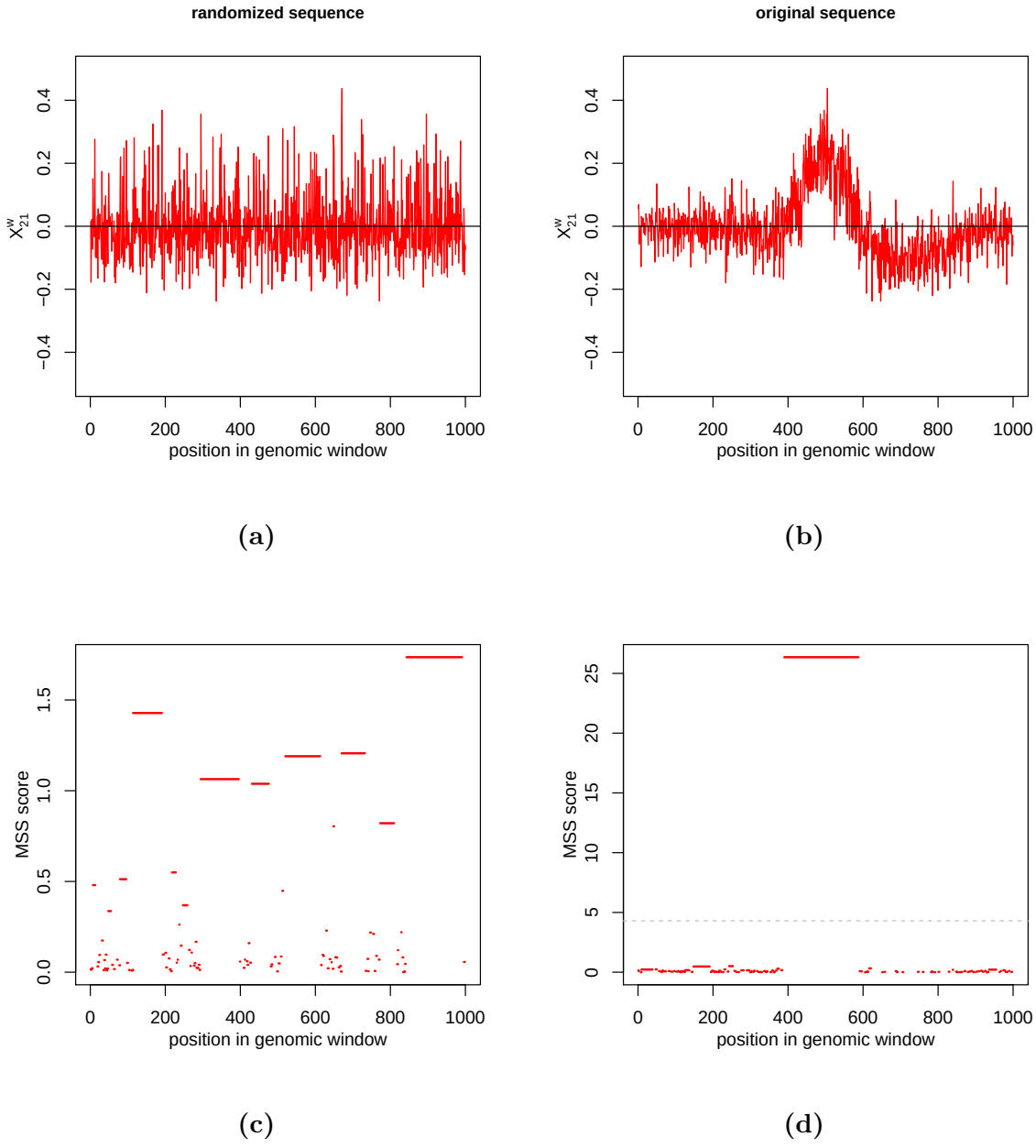


Figure 2.4 Filtering MSS by randomization. (a) Randomized read density difference sequence based on the original sequence X_{21}^w for that window. (b) Original read density difference sequence for that window. (c) All MSS identified on the randomized sequence distribution. The x-axis shows the position and the y-axis the score. (d) All MSS identified on the original sequence. The grey dotted line indicates the threshold for retaining an MSS determined by the MSS with maximum score on the randomized sequences.

2.2.3 Results

Real-life application

We reanalyzed promoter-proximal Pol II pausing upon HSV-1 infection on a gene level by applying RegCFinder on the 6 kb promoter windows used for the analysis described in the first article (see Section 2.1). For the set of 7,650 genes, RegCFinder identified 7,621 significant differential RegC for 6,958 genes (multiple testing adjusted P-value ≤ 0.01, 96 % of genes). We also evaluated RegCFinder with ten randomizations instead of the default 1,000. This led to more identified RegC but at the same time to a smaller fraction of significant RegC. Thus, increased significance can be achieved by reducing sensitivity and vice versa.

The location of RegC across promoter regions for genes which have at least one significant RegC are illustrated in the heatmap in Figure 2.5. Based on RegC locations in the promoter region, genes were clustered into eleven groups. Compared to our previous metagene analyses discussed in Section 2.1, this permitted detection of more diverse changes of Pol II occupancy. Not only did we confirm the delay of Pol II pausing into downstream regions (clusters 2, 5-11), but we also observed increased Pol II occupancy upstream of some genes (clusters 3, 4). Subsequent analyses of these clusters revealed that increased Pol II occupancy upstream of genes is due to read-through transcription originating from upstream genes. Moreover, RegCFinder results allowed to distinguish between clusters with increased read coverage only shortly downstream of the TSS (9, 10, part of 11) and clusters with increased read coverage extended till 3 kb downstream of TSS (clusters 5-8, part of 11). For the latter case, inspection of individual genes identified examples for both a very long delay in pausing and increased elongation due to loss of pausing.

To further demonstrate the wide applicability of RegCFinder, we also applied it to a ChIP-seq study for Pol II and Ser2 phosphorylation (P-Ser2) upon DMSO treatment and 4.5 h inhibition of CDK12 [190]. This study previously showed that CDK12 inhibition induces a Pol II processivity defect, which is accompanied by a shift of read coverage from gene 3' ends into gene bodies [190]. Our analysis was performed for whole genes with additional 3 kb upstream of the TSS and downstream of the TTS, respectively. In general, RegCFinder confirmed the loss of Pol II from gene 3' ends and the shift of P-Ser2 peaks from gene 3' ends into gene bodies. Moreover, clustering of genes based on their concatenated RegC profiles for Pol II and P-Ser2 into 15 groups revealed that long genes were more strongly affected by premature transcription termination.

Comparison to other tools

We evaluated our method in comparison with other existing approaches aiming to detect differentially covered regions, namely XCAVATOR (developed for CNV detection, see Section 1.4.3) and diffReps (developed for DCD localization, see Section 1.4.3) [181, 191]. Analyses of these methods were carried out with the same data as for RegCFinder (see above), i.e., Pol II pausing upon HSV-1 infection on PRO-seq data and analysis of Pol II processivity upon CDK12 inhibition on ChIP-seq data. XCAVATOR did not identify any significant differential regions for any of the two experiments. Most likely, this can

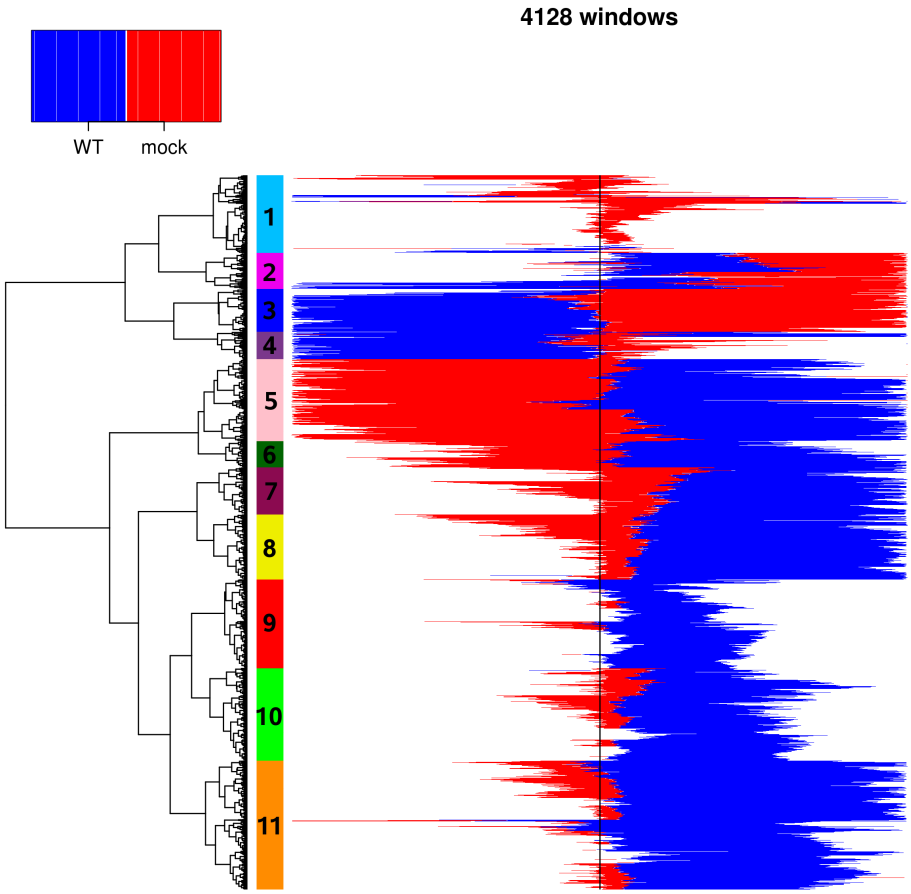


Figure 2.5 Regions with differences in read densities in PRO-seq data between mock and WT-F HSV-1 infection identified by RegCFinder. Heatmap showing the location of identified RegC in 6 kb promoter regions. RegC are colored in red and blue if they have higher read density in mock or WT-F, respectively. White regions represent RegC that were not significant or regions without differences between conditions. The vertical black line indicates the TSS. Genes were clustered hierarchically into eleven groups according to their RegC pattern within promoter regions based on Euclidean distances and Ward's clustering criterion. Clusters are indicated by the colored and numbered rectangles between the dendrogram and the heatmap.

be explained by the lack of sudden shifts in read coverage expected for CNVs. Regions identified by diffReps were filtered for those overlapping the analyzed promoter regions as its sliding window approach did not allow for a targeted analysis. Most of the identified regions for both studies represented only absolute changes in read coverage and changes were only determined in an unstranded manner. Hence, both tools were unsuitable for identifying regions with differential read distributions.

2.3 HSV-1 infection induces a downstream shift of +1 nucleosomes

2.3.1 Biological motivation

Previously, it was shown that HSV-1 infection impacts chromatin architecture downstream of gene 3' ends [130, 215]. Since Pol II pausing is linked to nucleosome positioning [216, 217], we investigated changes in chromatin accessibility in promoter regions for the third study included in this thesis [196]. Using RegCFinder, we analyzed recently published ATAC-seq data [130] for mock and HSV-1 WT-17 infection as well as null mutant infections for ICP0, ICP22, ICP27, and *vhs*. Analysis of these null mutant infections allowed studying the individual role of each gene with regard to potential chromatin changes in infection. Furthermore, we also analyzed an ATAC-seq time-course experiment of WT infection for 1, 2, 4, 6, and 8 h p.i. that enabled us to investigate when in the course of HSV-1 infection changes in chromatin accessibility are established [215]. In addition, we analyzed new ChIPmentation data performed at the lab of our collaboration partner Lars Dölken for mock and WT HSV-1 infection with an antibody recognizing the histone variant H2A.Z. H2A.Z is especially enriched in the +1 and -1 nucleosomes that surround the TSS and define the borders of the NFR at the TSS [195].

2.3.2 Widespread extension of the NFR in HSV-1 infection

We first performed pairwise comparisons of HSV-1 WT and null mutant infections against mock on the ATAC-seq data to identify chromatin changes in promoter regions during HSV-1 infection. Here, the same 6 kb promoter windows were used as input for RegCFinder as for the analysis described in Section 2.2.3. We restricted subsequent analyses to genes that exhibited at least one significant RegC for any of the performed comparisons, yielding 4,981 genes in total. Analysis of RegC locations across HSV-1 infection variants in a heatmap revealed very consistent changes between WT and null mutant infections with increased chromatin accessibility generally observed downstream of the TSS for the majority of host genes (Figure 2.6). We then utilized a clustering approach for the identification of distinct patterns of changes in chromatin accessibility. Here, clustering was performed based on the concatenated vectors of RegC locations within promoter regions for all pairwise analyses. The obtained 14 clusters are visualized in Figure 2.6. We then applied metagene analyses for each cluster to characterize the chromatin accessibility changes. Metagene plots were augmented to show the fraction of promoter windows in the cluster with a RegC at each position (see Figure 2.7). This identified three major patterns of changes in chromatin accessibility at promoters: I increased chromatin accessibility downstream of the TSS (example in Figure 2.7a), II increased chromatin accessibility upstream of the TSS (example in Figure 2.7b), and III increased chromatin accessibility, both, up- and downstream of the TSS. Notably, almost all clusters displayed increased chromatin accessibility downstream of the TSS (pattern I) and only three clusters (919 genes) showed the reverse picture (pattern II).

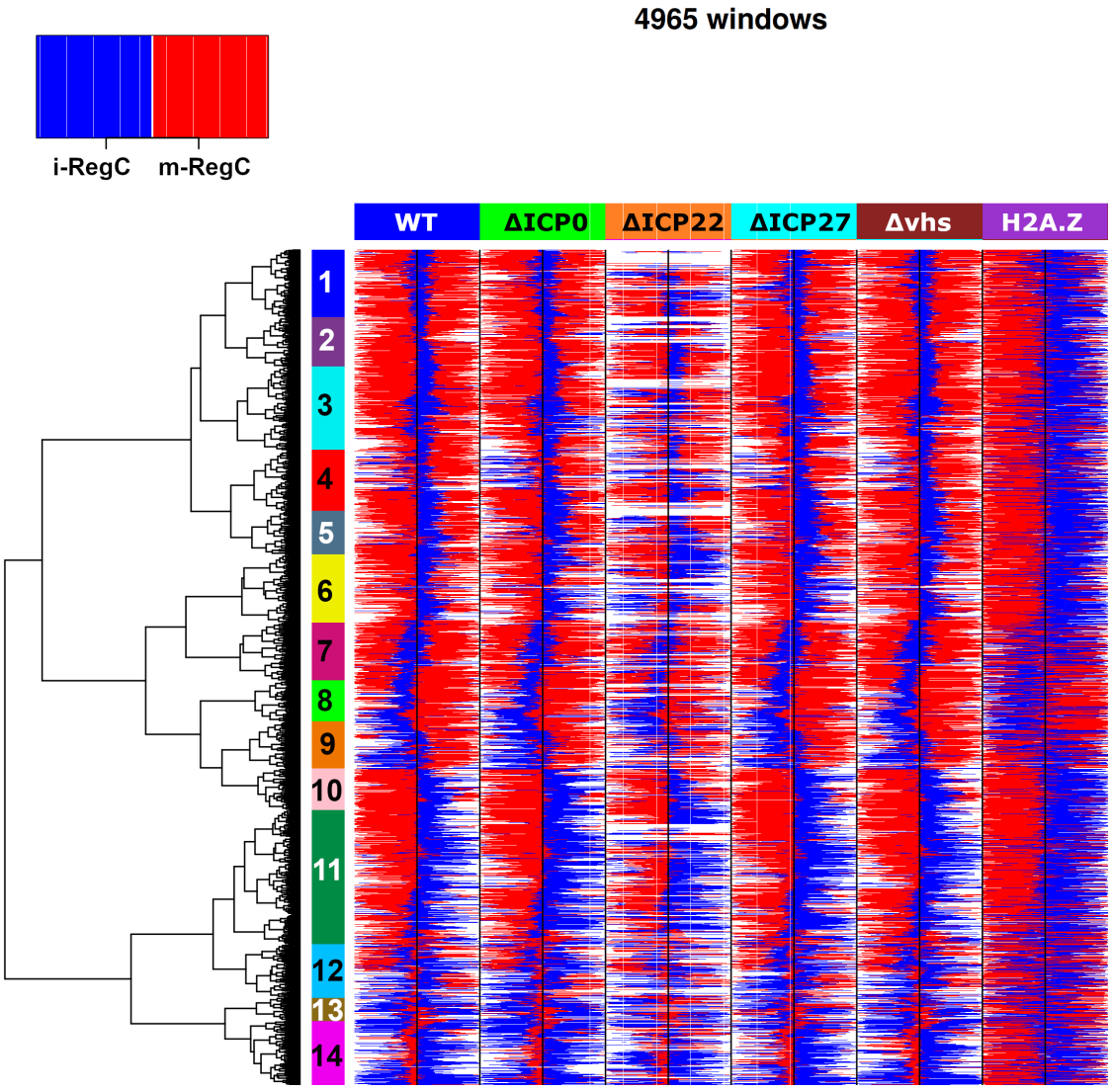
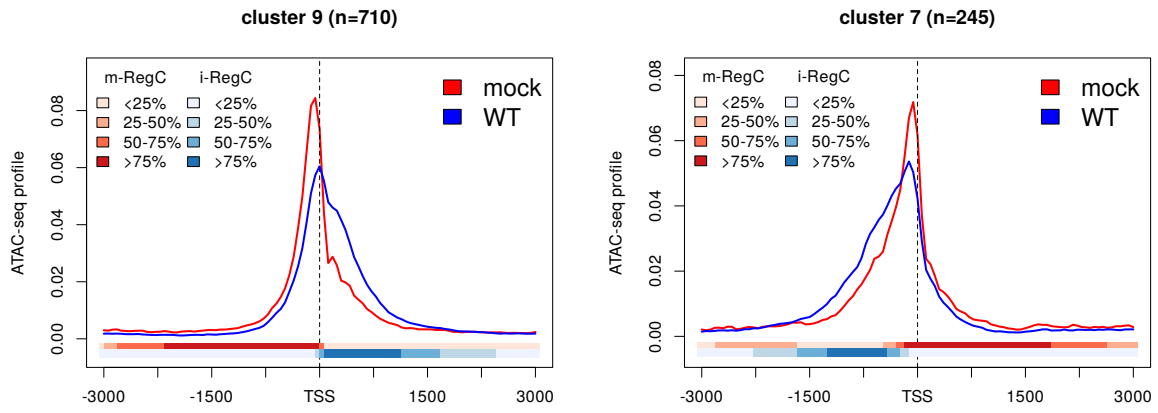


Figure 2.6 Regions with chromatin accessibility and H2A.Z distribution changes in promoters during HSV-1 infection. The heatmap shows locations of RegC identified by RegCFinder in ATAC-seq and H2A.Z ChIPmentation profiles in WT (ATAC-seq and H2A.Z ChIPmentation) and null mutant HSV-1 infections (ATAC-seq only). RegC are colored in red and blue for regions with higher read density in mock or HSV-1 infection, respectively. White color represents regions without significant differences. 6 kb promoter regions centered around TSS (vertical black lines) are concatenated for all six comparisons performed (separated by vertical black line). Genes were clustered hierarchically into 14 groups according to their RegC pattern in ATAC-seq data based on Euclidean distances and Ward's clustering criterion. Clusters are indicated by the colored and numbered rectangles between the dendrogram and the heatmap.



(a) **(b)**

Figure 2.7 Augmented metagene analysis of selected clusters from Figure 2.6. Metagene plot showing the average ATAC-seq profile for mock (red) and WT (blue) HSV-1 infection in a 6 kb promoter window. The colored bands below the metagene curves indicate the percentage of genes having a RegC at that position in mock (m-RegC) or WT (i-RegC), respectively. (a) shows the average profile for cluster 9 as an example for pattern I with broadened chromatin accessibility downstream of the TSS. (b) shows the average profile for cluster 7 as an example for pattern II with broadened chromatin accessibility upstream of the TSS.

To determine the onset time of these changes, we applied RegCFinder to pairwise comparisons of each point of the infection ATAC-seq time-course and mock infection. This revealed that changes in chromatin accessibility were beginning to manifest already at 4 h p.i., with changes becoming more pronounced until 8 h p.i. Although changes at 8 h p.i. were less pronounced than those from the first ATAC-seq experiment which was performed at 8 h p.i., they confirmed the changes in chromatin accessibility as a result of an independent experiment. We also analyzed changes in chromatin accessibility for HSV-1 infection with peracetic acid (PAA) treatment, which inhibits viral replication and reduces relocation of Pol II to viral genomes [198, 218–220]. This showed that PAA substantially, but not completely, reduced the broadening in chromatin accessibility downstream of the TSS.

2.3.3 Downstream shift of +1 nucleosomes

To test whether the increase in chromatin accessibility up- or downstream of the TSS upon HSV-1 infection reflected changes in nucleosome positions, we investigated ChIPmentation data for the histone variant H2A.Z. Nucleosomes, in particular those around the TSS, are very well-positioned and function as natural barriers to transcription [221, 222]. We applied RegCFinder to H2A.Z ChIPmentation data for mock and WT infection and compared this

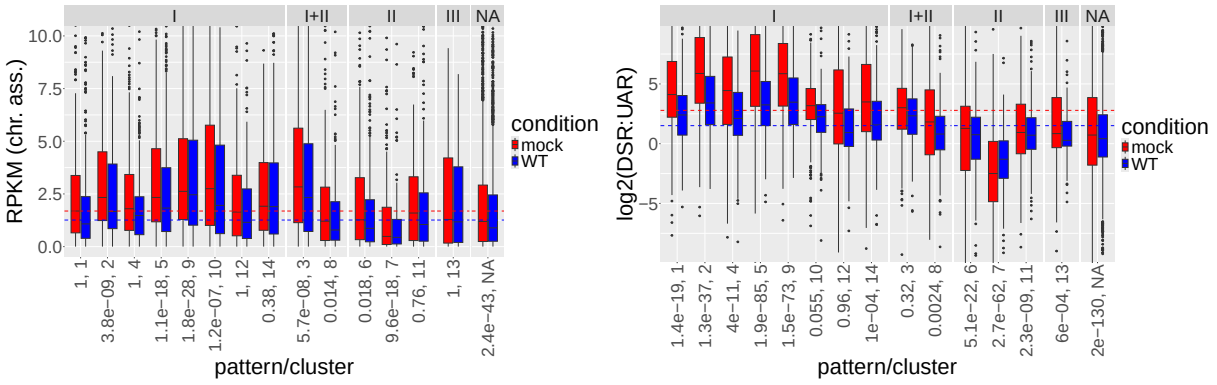
to the ATAC-seq results (see Figure 2.6). RegC identified on H2A.Z largely reflected the major patterns detected in ATAC-seq data, indicating that H2A.Z occupancy of the +1 nucleosome was shifted downstream of the TSS for the majority of genes in analogy to pattern I in ATAC-seq data. For pattern II cluster 8 and part of clusters 6 and 7, H2A.Z occupancy of the -1 nucleosome was shifted upstream of the TSS matching changes in ATAC-seq data. This was confirmed by metagene analyses for H2A.Z of these clusters. To summarize the identified changes in promoter regions, HSV-1 infection leads to extensions of NFR, mostly toward downstream regions, as +1 nucleosomes are shifted downstream of the TSS. Exceptions with extended NFR upstream of the TSS showed upstream shifted -1 nucleosomes, respectively.

2.3.4 The link between transcription and chromatin changes

In yeast, it was shown that loss of Pol II leads to a relaxation of +1 nucleosome positions shifting them toward downstream sites [145]. We thus hypothesized that the global loss of host transcription in HSV-1 infection leads to the observed changes in +1 and -1 nucleosome positions. Consistent with previously reported loss of transcription in HSV-1 infection [127, 128], gene expression analysis (in terms of RPKM) in chromatin-associated RNA-seq data for mock and 8 h post WT HSV-1 infection showed reduced transcriptional activity upon HSV-1 infection across all clusters (Figure 2.8a) [131]. However, RPKM levels for pattern II clusters were notably lower than those for other genes, both before and after infection. Further analyses confirmed that genes which are more strongly expressed prior to infection exhibit larger downstream broadening of the NFR. In contrast, lowly expressed genes showed no or only small changes in nucleosome positions.

Since pattern II essentially resembled a mirrored image of pattern I, we hypothesized that these might represent bidirectional promoters. Thus, for each cluster we analyzed the promoters for presence of annotated antisense gene starts within 1 kb upstream of the respective TSS. This showed that pattern II clusters were highly enriched for bidirectional promoters. To clarify which direction of transcription was dominant for each promoter, we quantified gene expression as RPKM for the window downstream of the TSS in sense direction (DSR) relative to the window upstream of the TSS in antisense direction (UAR). The distributions of log2 ratios of DSR to UAR for all clusters are shown in Figure 2.8b. This showed that pattern II clusters generally exhibited low log2 DSR:UAR ratios. Low positive or even negative log2 ratios of DSR to UAR indicated strong transcription in antisense direction which, in many cases, was the dominant direction of transcription. Thus, pattern II effectively represented the same type of changes as pattern I, with downstream shifts of +1 nucleosomes on the opposite strand.

To demonstrate that inhibition of transcription alone can induce downstream shifts of +1 nucleosomes, we analyzed published H2A.Z ChIP-seq data with and without α-amanitin treatment [223]. α-amanitin inhibits RNA synthesis by degrading Rbp1, the largest Pol II subunit. This prevents Pol II translocation and reduces Pol II occupancy at the TSS [224–229]. In contrast, H2A.Z has been shown to accumulate at the TSS upon α-amanitin treatment [223]. RegCFinder and metagene analyses confirmed very similar



(a) **(b)**

Figure 2.8 Gene expression and antisense transcription for genes with different changes in chromatin accessibility. (a) Boxplots of gene expression (RPKM) in chromatin-associated RNA in mock (red) and WT (blue) infection for all clusters, grouped according to the three major patterns. Median gene expression values for all genes for mock and WT are indicated by horizontal red and blue dashed lines, respectively. Value below cluster numbers on the x-axis indicate adjusted *P*-values for Wilcoxon rank sum test comparing gene expression levels of the corresponding cluster against values of all other analyzed genes. *P*-values are adjusted for multiple testing with the Bonferroni method. The NA group refers to genes for which no significant RegC were identified in their promoters by RegCFinder. (b) Boxplots showing log2 ratios of DSR to UAR in mock (red) and WT (blue) infection for all clusters, grouped according to the three major patterns. DSR was calculated on the region from the TSS to 3 kb downstream of the TSS and UAR in the region from 3 kb upstream of the TSS to the TSS. Median log2 ratios for mock and WT infection across all genes are indicated by dashed red and blue horizontal lines, respectively.

changes upon α-amanitin treatment as during HSV-1 infection. In general, clusters with nucleosome shifts up- or downstream of the TSS showed even more pronounced shifts upon α-amanitin treatment than upon WT HSV-1 infection. In conclusion, our analyses showed that downstream shifts of +1 nucleosomes are likely a consequence of the loss of Pol II from the host genome in HSV-1 infection.

Chapter 3

Discussion and outlook

3.1 Analyzing changes in read distributions

Differential analyses are an important aspect of functional genomics and many methods have previously been developed for this purpose. For instance, this includes DESeq2 [168] and edgeR [169] for differential gene expression analysis, DEXSeq [214] for differential exon usage or diffReps [181] for detecting differential chromatin modifications. However, state-of-the-art methods commonly do not consider changes in the distribution of reads. Identification of such differences in read distributions was the key objective of this thesis. By focusing on detecting these changes at the level of individual genes or smaller groups of genes, we aimed to address the problems of standard metagene analyses. Metagene plots visualize the mean read coverage over a set of genomic windows, e.g., genes or promoter regions. While they easily can combine multiple conditions and different experiments and allow easy identification and interpretation of changes in read distributions, they are not designed for analyses at the level of single genes. As metagene plots represent only mean read coverage changes, changes for single genes can deviate significantly from these mean changes. Furthermore, evaluation of changes is commonly based on visual inspection without a precise quantification.

In contrast to metagene analyses, PI analysis quantifies read distribution changes between promoter and gene body at gene level, but the results depend strongly on the size of promoter and gene body windows. Since this method compares absolute read count ratios between two windows, it can identify reductions or increases of read counts in either one of the windows, but cannot distinguish more complex changes in read distributions. As a consequence, we were not able to properly distinguish delayed Pol II pausing with this method.

As a first approach to address this issue, we developed a method based on clustering read distribution profiles (contributing article for Section 2.1 [188]). While this provided novel results regarding Pol II pausing changes in HSV-1 infection, it still did not allow analysis at single-gene level. We thus developed RegCFinder, a novel method to analyze changes in read distributions at the level of individual genes and identify genomic regions with

differential read density (contributing article for Section 2.2 [192]). This overcame many of the above-mentioned limitations of currently available methods. Moreover, the extent and significance of changes in the identified regions are quantified. Since it is implemented as a workflow for the WMS Watchdog, RegCFinder can make use of Watchdog features, including automatic deployment of software with conda, parallel distribution on a computer cluster, progress monitoring, and error detection.

We evaluated RegCFinder against the alternative approaches either originally developed for detection of copy number variation (CNV) (XCAVATOR [191]) or identification of differential chromatin domains (DCD) (diffReps [181]). However, their original purposes did not suit our objectives to identify changes in read distributions. Hence, these tools did not yield meaningful results for our target applications.

An important feature of RegCFinder is its applicability to any functional genomics sequencing assay. By specifying genomic input regions, the search can be targeted without any limit for size or number. Furthermore, RegCFinder makes use of the linear run time of the Ruzzo and Tompa algorithm [213] for solving the "all maximum scoring subsequences" (AMSS) problem and is thus highly efficient. Finally, its independence of the type of sequencing assays and its standardized results allows for an easy data integration throughout and after the analysis. It should be noted that RegCFinder is designed to compare only two conditions against each other. If more conditions need to be analyzed, a common reference is required, however, this is generally the case for differential methods.

3.2 Applications of RegCFinder

Novel insights into HSV-1 infection

Throughout the contributing articles within this thesis, we investigated changes in Pol II occupancy and nucleosome positioning upon HSV-1 infection and uncovered more complex changes than previously anticipated. While the downstream delay of Pol II pausing was originally discovered with metagene analysis combined with clustering, we showed that the analyses at single-gene level facilitated by RegCFinder allowed identification of further types of changes. Since transcription is tightly associated with the chromatin structure and nucleosome positioning, RegCFinder was applied to ATAC-seq and H2A.Z ChIPmentation data of HSV-1 infection (contributing article for Section 2.3 [196]). This revealed that the NFR at promoters was commonly extended in downstream direction, which was associated with downstream shifts of +1 nucleosomes. Furthermore, we showed that genes with strong expression prior to infection were affected more strongly than less expressed genes. Previously, it was shown that Pol II pausing is enhanced by strong positioning of the +1 nucleosome, while less well positioned +1 nucleosomes enhance Pol II pause release [230]. Moreover, depletion of NELF leads to reduced Pol II pausing at promoter-proximal sites and increased pausing at further downstream sites near +1 nucleosomes [210]. Hence, the observed downstream shifts of +1 nucleosomes could provide a possible explanation for the delay in Pol II pausing upon HSV-1 infection. In support of this hypothesis, changes in +1 nucleosome positioning occurred already at 4 h p.i. but not yet at 2 h p.i., similarly to

Pol II pausing where changes were visible at 3 h p.i. but not yet at 1.5 h p.i. Importantly, these changes were independent of ICP22, which mediated chromatin changes downstream of genes.

Beyond promoters and HSV-1 infection
Since RegCFinder was developed to allow inspection of any genomic region, it is applicable beyond the promoter regions that were the focus of this thesis. For instance, at the moment we are currently applying RegCFinder to investigate premature transcription termination upon HSV-1 infection again using the PRO-seq data from the study by Birkenheuer *et al.* [128]. However, for this study, we are focusing on input windows surrounding poly(A) sites. In addition, RegCFinder was already successfully utilized to investigate epigenetic changes induced by Kaposi's sarcoma-associated herpesvirus (KSHV). Here, RegCFinder was able to detect changes in chromatin structure on both the host and the KSHV genome.

3.3 Conclusion

In summary, the RegCFinder method developed as part of this thesis represents a new bioinformatics approach to analyze functional genomics data. With the help of RegCFinder, we not only extended the current understanding of HSV-1 infection but made general observations on the linkage between transcription and chromatin architecture. In particular, we also showed that loss of Pol II from human genomes generally appears to lead to a downstream shift of +1 nucleosomes and an extension of the NFR. This had previously only been reported for yeast [231]. Since RegCFinder can be targeted to any type of genomic feature, including viral genomes and 3' gene ends, it will be highly useful for a wide range of applications beyond those presented in this thesis.

Glossary

+1 nucleosome First nucleosome downstream of the TSS. 4

ATAC-seq Deep sequencing technology to assess chromatin accessibility. 5

C-value paradox Organismal complexity does not correlate with genome size. 1

Central dogma of molecular biology Unidirectional synthesis from DNA to mRNA to proteins. 2

ChIP-seq Deep sequencing technology for identification of protein-DNA interactions. 5

ChIPmentation Combination of ChIP-seq and tagmentation by Tn5 transposase. 5

cRNA-seq RNA-seq based on circulization of RNA fragments for mapping of 5' transcript ends. 15

DEXSeq Method for inferring differential exon usage in RNA-seq data. 16

dRNA-seq Directional RNA-seq based on selective cloning and sequencing of 5' ends of cap-protected RNA molecules. 15

ENCODE Encyclopedia of DNA elements of all functional elements in the human genome. 2

GRO-seq Global run-on sequencing for mapping active RNA Polymerase II. 5

Metagene analysis Method for aggregating and normalizing a group of features regarding their read coverage within a given window and visualizing their common shape. 8

Mock infection Cells infected in the same medium but without the virus expression, functions as healthy control. 11

Pol II pausing Promoter-proximal RNA polymerase II pausing. 2

PRO-seq Precision run-on sequencing for mapping active RNA pOlymerase II with single-base resolution. 5

PROcap-seq Variation of PRO-seq for mapping transcription initiation sites with base-pair resolution. 11

Read-through Transcription beyond TTS due to unrecognized termination signal. 4

RegC Regions of change, maximal subregions identified by the AMSS algorithm. 18

RegCFinder *de novo* discovery of genomic regions with differential read density. 8

RNA-seq Deep sequencing technology for transcriptome profiling. 5

Acronyms

AMSS All maximum scoring subsequences. 16

BAM Binary alignment map format. 16

BED Browser extensible data format. 16

bp Base pair. 1

CDK Cyclin-dependent kinsase. 3

cDNA Complementary deoxyribonucleic acid. 5

CNV Copy number variant. 8

DCD Differential chromatin domains. 8

ddNTP Dideoxynucleotide triphosphate. 1

DEA Differential expression analysis. 6

DNA Deoxyribonucleic acid. 1

dOCR Downstream open chromatin region. 4

DoTT Disruption of transcription termination. 4

DSR Downstream sense region. 25

Gb Gigabase. 1

GO Gene ontology. 13

HSV-1 Herpes simplex virus type I. 4

IE Immediate early genes of HSV-1 virus. 4

IGV Integrative genomics viewer. 11

mRNA Messenger RNA. 3

MSS Maximum scoring subsequences. 16

NELF Negative elongation factor. 3

NFR Nucleosome-free region. 4

NGS Next generation sequencing. 1

ORF Open reading frame. 4

P-TEFb Positive transcription elongation factor b. 3

p.i. post infection. 11

PAA Peracetic acid. 24

PI Pausing index. 6

PIC Pre-initiation complex. 3

Pol II RNA polymerase II. 2

poly(A) Polyadenylation site. 2

RD Read depth. 8

RNA Ribonucleic acid. 2

RPKM Reads per kilobase million. 6

TF Transcription factor. 3

TGS Third generation sequencing. 2

TSS Transcription start site. 2

TTS Transcription termination site. 2

UAR Upstream antisense region. 25

vhs Viral host shutoff protein. 5

WHO World health organization. 4

WMS Workflow management system. 9

WT Wild-type. 10

WT-17 HSV-1 wild-type strain 17 infection. 15

WT-F HSV-1 wild-type strain F infection. 7

References

[1] S Kaushik, S Kaushik, and D Sharma. Functional genomics. In S Ranganathan, M Gribskov, K Nakai, and C Schönbach, editors, *Encyclopedia of bioinformatics and computational biology*, pages 118–133. Academic Press, Oxford, 2019.

[2] T Lappalainen. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Research*, 25(10):1427–1431, 2015.

[3] DJ Morris-Rosendahl. A glossary of relevant genetic terms. *Dialogues in clinical neuroscience*, 12(1):116–120, 2010.

[4] PS Cooper, D Lipshultz, WT Matten, SD McGinnis, S Pechous, ML Romiti, T Tao, M Valjavec-Gratian, and EW Sayers. Education resources of the national center for biotechnology information. *Briefings in Bioinformatics*, 11(6):563–569, 2010.

[5] DL Mattson. Functional genomics. In W Walz, editor, *Integrative Physiology in the Proteomics and Post-Genomics Age*, pages 7–25. Humana Press, 2005.

[6] L Steinmetz and R Davis. Maximizing the potential of functional genomics. *Nature Reviews Genetics*, 5(3):190–201, 2004.

[7] J Pevsner. *Bioinformatics and functional genomics*. John Wiley and Sons, 2015.

[8] T Werner. Next generation sequencing in functional genomics. *Briefings in Bioinformatics*, 1(5):499–511, 2010.

[9] S Bao, R Jiang, W Kwan, B Wang, X Ma, and YQ Song. Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, 56(6):406–414, 2011.

[10] S Goodwin, J McPherson, and W McCombie. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

[11] O Morozova and MA Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–64, 2008.

[12] F Sanger, S Nicklen, and AR Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[13] BM Crossley, J Bai, A Glaser, R Maes, E Porter, ML Killian, T Clement, and K Toohey-Kurth. Guidelines for Sanger sequencing and molecular assay monitoring. *Journal of Veterinary Diagnostic Investigation*, 32(6):767–775, 2020.

[14] F Sanger and AR Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3): 441–448, 1975.

[15] LM Smith, JZ Sanders, RJ Kaiser, P Hughes, C Dodd, CR Connell, and LE Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071): 674–679, 1986.

[16] JM Prober, GL Trainor, RJ Dam, FW Hobbs, CW Robertson, RJ Zagursky, and K Baumeister. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science*, 238(4825):336–341, 1987.

[17] H Swerdlow, JZ Zhang, DY Chen, HR Harke, R Grey, SL Wu, NJ Dovichi, and C Fuller. Three DNA sequencing methods using capillary gel electrophoresis and laser-induced fluorescence. *Analytical Chemistry*, 63(24):2835–41, 1991.

[18] RD Fleischmann et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995.

[19] S Broder and JC Venter. Whole genomes: The foundation of new biology and medicine. *Current Opinion in Biotechnology*, 11(6):581–585, 2000.

[20] J Vosseberg, JJE van Hooff, M Marcet-Houben, A van Vlimmeren, LM van Wijk, T Gabaldón, and B Snel. Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nature Ecology and Evolution*, 5(1):92–100, 2021.

[21] L Pray. Eukaryotic genome complexity. *Nature Education*, 1(1):96, 2008.

[22] GM Cooper. *The complexity of eukaryotic genomes*. Sunderland, 2000.

[23] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[24] MW Hahn and GA Wray. The G-value paradox. *Evolution and Development*, 4(2): 73–75, 2002.

[25] I Choi, EC Kwon, and NS Kim. The C- and G-value paradox with polyploidy, repeatomes, introns, phenomes and cell economy. *Genes and Genomics*, 42(7):699–714, 2020.

[26] P Amaral, S Carbonell-Sala, FM De La Vega, T Faial, A Frankish, T Gingeras, R Guigo, JL Harrow, AG Hatzigeorgiou, R Johnson, TD Murphy, M Pertea, KD Pruitt, S Pujar, H Takahashi, I Ulitsky, A Varabyou, CA Wells, M Yandell,

P Carninci, and SL Salzberg. The status of the human gene catalogue. *Nature*, 7981 (622):41–47, 2023.

[27] A Piovesan, F Antonaros, L Vitale, P Strippoli, MC Pelleri, and M Caracausi. Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, 12 (1):315, 2019.

[28] XW Xu, XH Zhou, RR Wang, et al. Functional analysis of long intergenic non-coding RNAs in phosphate-starved rice using competing endogenous RNA network. *Scientific Reports*, 6:20715, 2016.

[29] JC Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[30] S Lu, J Zhang, X Lian, L Sun, K Meng, Y Chen, Z Sun, X Yin, Y Li, J Zhao, T Wang, G Zhang, and QY He. A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Research*, 47(15):8111–8125, 2019.

[31] T Hu, N Chitnis, D Monos, and A Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, 2021.

[32] JS Reis-Filho. Next-generation sequencing. *Breast Cancer Research*, 11(Suppl 3): S12, 2009.

[33] KV Voelkerding, SA Dames, and JD Durtschi. Next-generation sequencing: From basic research to diagnostics. *Clinical Chemistry*, 55:641–658, 2009.

[34] EM Bunnik and KG Leroch. An introduction to functional genomics and systems biology. *Advances in Wound Care*, 2(9):490–498, 2013.

[35] P Hieter and M Boguski. Functional genomics: It's all how you read it. *Science*, 278 (5338):601–602, 1997.

[36] L Luo, M Gribskov, and S Wang. Bibliometric review of ATAC-Seq and its application in gene expression. *Briefings in Bioinformatics*, 23(3):bbac061, 2022.

[37] R Pereira, J Oliveira, and M Sousa. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *Journal of Clinical Medicine*, 9 (1):132, 2020.

[38] SG Oliver. Functional genomics: Lessons from yeast. *Philosophical Transactions of the Royal Society London B Biological Sciences*, 357(1417):17–23, 2002.

[39] KO Mutz, A Heilkenbrinker, M Lönne, JG Walter, and F Stahl. Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology*, 24(1): 22–30, 2013.

[40] DS Horner, G Pavesi, T Castrignanò, P D'Onorio De Meo, S Liuni, M Sammeth, E Picardi, and G Pesole. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, 11(2):181–197, 2010.

[41] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 498(7414):958–975, 2012.

[42] J Ou, H Liu, J Yu, MA Kelliher, LH Castilla, ND Lawson, and LJ Zhu. ATACseqQC: A Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics*, 19(1):169, 2018.

[43] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[44] PM Visscher, NR Wray, Q Zhang, P Sklar, MI McCarthy, MA Brown, and J Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[45] Roadmap Epigenomics Consortium, A Kundaje, W Meuleman, J Ernst, M Bilenky, A Yen, A Heravi-Moussavi, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

[46] CE Romanoski, CK Glass, HG Stunnenberg, L Wilson, and G Almouzni. Epigenomics: Roadmap for regulation. *Nature*, 518(7539):314–6, 2015.

[47] T Xiao and W Zhou. The third generation sequencing: The advanced approach to genetic diseases. *Translational Pediatrics*, 9(2):163–173, 2020.

[48] K Athanasopoulou, MA Boti, PG Adamopoulos, PC Skourou, and A Scorilas. Third-generation sequencing: The spearhead towards the radical transformation of modern genomics. *Life*, 12(1):30, 2021.

[49] EL van Dijk, Y Jaszczyszyn, D Naquin, and C Thermes. The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681, 2018.

[50] B Searle, M Müller, T Carell, and A Kellett. Third-generation sequencing of epigenetic DNA. *Angewandte Chemie*, 135(14):e202215704, 2023.

[51] F Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[52] EA Bowman and WG Kelly. RNA polymerase II transcription elongation and Pol II CTD Ser2 phosphorylation: A tail of two kinases. *Nucleus*, 5(3):224–36, 2014.

[53] Liu X, WL Kraus, and X Bai. Ready, pause, go: Regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends in Biochemical Sciences*, 40(9):516–25, 2015.

[54] H Kwak, NJ Fuda, LJ Core, and JT Lis. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, 339(6122):950–3, 2013.

[55] X Darzacq, Y Shav-Tal, V de Turris, Y Brody, SM Shenoy, RD Phair, and RH Singer. In vivo dynamics of RNA polymerase II transcription. *Nature Structural and Molecular Biology*, 14(9):796–806, 2007.

[56] Q Yang and S Doublié. Structural biology of poly(A) site definition. *Wiley Interdisciplinary Reviews: RNA*, 2(5):732–47, 2011.

[57] M Legendre and D Gautheret. Sequence determinants in human polyadenylation site selection. *BMC Genomics*, 4(1):7, 2003.

[58] NJ Proudfoot. Ending the message: poly(A) signals then and now. *Genes and Development*, 25(17):1770–82, 2011.

[59] JN Kuehner, EL Pearson, and C Moore. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Review Molecular Cell Biology*, 12 (5):283–94, 2011.

[60] O Porrua and D Libri. Transcription termination and the control of the transcriptome: Why, where and how to stop. *Nature Review Molecular Cell Biology*, 16(3): 190–202, 2015.

[61] S Hahn. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structure and Molecular Biology*, 11(5):394–403, 2004.

[62] J Vaquerizas, S Kummerfeld, S Teichmann, and NM Luscombe. A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, 2009.

[63] TW Sikorski and S Buratowski. The basal initiation machinery: Beyond the general transcription factors. *Current Opinion in Cell Biology*, 21(3):344–351, 2009.

[64] DB Nikolov and SK Burley. RNA polymerase II transcription initiation: A structural view. *Proceedings of the National Academy of Sciences*, 94(1):15–22, 1997.

[65] X Chen and Y Xu. Structural insights into assembly of transcription preinitiation complex. *Current Opinion in Structural Biology*, 75:102404, 2022.

[66] L Farnung and SM Vos. Assembly of RNA polymerase II transcription initiation complexes. *Current Opinion in Structural Biology*, 73:102335, 2022.

[67] J Soutourina. Transcription regulation by the mediator complex. *Nature Reviews Molecular Cell Biology*, 19:262–274, 2018.

[68] S Malik and RG Roeder. Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators. *Nature Reviews Genetics*, 24(11):767–782, 2023.

[69] M Hampsey. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews*, 62(2):465–503, 1998.

[70] M Levine and R Tjian. Transcription regulation and animal diversity. *Nature*, 424 (6945):147–151, 2003.

[71] SA Lambert, A Jolma, LF Campitelli, PK Das, Y Yin, M Albu, X Chen, J Taipale, TR Hughes, and MT Weirauch. The human transcription factors. *Cell*, 172(4): 650–665, 2018.

[72] DS Latchman. *Eukaryotic transcription factors*. Academic Press, 2010.

[73] M Li, A Hada, P Sen, L Olufemi, MA Hall, BY Smith, S Forth, JN McKnight, A Patel, GD Bowman, B Bartholomew, and MD Wang. Dynamic regulation of transcription factors by nucleosome remodeling. *Elife*, 4:e06249, 2015.

[74] F Zhu, L Farnung, E Kaasinen, B Sahu, Y Yin, B Wei, SO Dodonova, KR Nitta, E Morgunova, M Taipale, P Cramer, and J Taipale. The interaction landscape between transcription factors and the nucleosome. *Nature*, 562(7725):76–81, 2018.

[75] TE O'Neill, M Roberge, and EM Bradbury. Nucleosome arrays inhibit both initiation and elongation of transcripts by bacteriophage T7 RNA polymerase. *Journal of Molecular Biology*, 223(1):67–78, 1992.

[76] CH Chang and DS Luse. The H3/H4 tetramer blocks transcript elongation by RNA polymerase II in vitro. *Journal of Biological Chemistry*, 272(37):23427–23434, 1997.

[77] JA Knezetic and DS Luse. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell*, 45(1):95–104, 1986.

[78] CL Peterson and MA Laniel. Histones and histone modifications. *Current Biology*, 14(14):R546–R551, 2004.

[79] RK McGinty and Song Tan. Nucleosome structure and function. *Chemical Reviews*, 115(6):2255–2273, 2015.

[80] GA Armeev, AK Gribkova, I Pospelova, GA Komarova, and AK Shaytan. Linking chromatin composition and structural dynamics at the nucleosome level. *Current Opinion in Structural Biology*, 56:46–55, 2019.

[81] RH Morse and RT Simpson. DNA in the nucleosome. *Cell*, 54(3):285–287, 1988.

[82] TJ Richmond and C Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–150, 2003.

[83] L Bai, G Charvin, ED Siggia, and FR Cross. Nucleosome-depleted regions in cell-cycle-regulated promoters ensure reliable gene expression in every cell cycle. *Developmental Cell*, 18(4):544–555, 2010.

[84] L Bai and AV Morozov. Gene regulation by nucleosome positioning. *Trends in Genetics*, 26(11):476–483, 2010.

[85] C Jiang and BF Pugh. Nucleosome positioning and gene regulation: Advances through genomics. *Nature Reviews Genetics*, 10(3):161–172, 2009.

[86] AK Singh and F Mueller-Planitz. Nucleosome positioning and spacing: From mechanism to function. *Journal of Molecular Biology*, 433(6):166847, 2021.

[87] Y Sun, N Miao, and T Sun. Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas*, 156:29, 2019.

[88] AR Mansisidor and VI Risca. Chromatin accessibility: Methods, mechanisms, and biological insights. *Nucleus*, 13(1):236–276s, 2022.

[89] JJ Hayes, DJ Clark, and AP Wolffe. Histone contributions to the structure of DNA in the nucleosome. *Proceedings of the National Academy of Sciences*, 88(15):6829–6833, 1991.

[90] A Wolffe. *Chromatin: Structure and function*. Academic press, 1998.

[91] K Maeshima, S Iida, MA Shimazoe, S Tamura, and S Ide. Is euchromatin really open in the cell? *Trends in Cell Biology*, 34(1):7–17, 2023.

[92] B Roizman and J Baines. The diversity and unity of herpesviridae. *Comparative Immunology, Microbiology and Infectious Diseases*, 14(2):63–79, 1991.

[93] RJ Whitley. Herpes simplex virus. In *Infections of the central nervous system*, pages 123–44, Philadelphia, 2004. Lippincott Williams and Wilkins.

[94] M Fatahzadeh and RA Schwartz. Human herpes simplex virus infections: Epidemiology, pathogenesis, symptomatology, diagnosis, and management. *Journal of the American Academy of Dermatology*, 57(5):737–763, 2007.

[95] RB Roizman and RJ Whitley. Herpes simplex viruses. In DM Knipe and PM Howley, editors, *Fields Virology*, pages 2501–2601, Philadelphia, 2007. Lippincott Williams and Wilkins.

[96] PE Pellet and RB Roizman. Herpesviridae. In DM Knipe and PM Howley, editors, *Fields Virology*, pages 1802–2128, Philadelphia, 2013. Wolters Kluwer Health/Lippincott Williams and Wilkins.

[97] D Ablashi, H Agut, R Alvarez-Lafuente, DA Clark, S Dewhurst, D DiLuca, L Flamand, N Frenkel, R Gallo, UA Gompels, P Höllsberg, S Jacobson, M Luppi, P Lusso, M Malnati, P Medveczky, Y Mori, PE Pellett, JC Pritchett, K Yamanishi, and T Yoshikawa. Classification of HHV-6A and HHV-6B as distinct viruses. *Archives of Virology*, 159(5):863–70, 2014.

[98] S Zhu and A Viejo-Borbolla. Pathogenesis and virulence of herpes simplex virus. *Virulence*, 12(1):2670–2702, 2021.

[99] B Grinde. Herpesviruses: Latency and reactivation – viral strategies and host response. *Journal of Oral Microbiology*, 5(1):22766, 2013.

[100] GR Bedadala, JR Palem, L Graham, JM Hill, HE McFerrin, and SC Hsia. Lytic HSV-1 infection induces the multifunctional transcription factor early growth response-1 (EGR-1) in rabbit corneal cells. *Virology Journal*, 8:262, 2011.

[101] S Petti and G Lodi. The controversial natural history of oral herpes simplex virus type 1 infection. *Oral Diseases*, 25(8):1850–1865, 2019.

[102] MJ Bradshaw and A Venkatesan. Herpes simplex virus-1 encephalitis in adults: Pathophysiology, diagnosis, and management. *Neurotherapeutics*, 13(3):493–508, 2016.

[103] M Astuto, CI Palermo, CM Costanzo, GC Ettorre, S Palmucci, C Franchina, R Russo, P Valastro, V Timpanaro, and G Scalia. Fatal pulmonary disease and encephalic complication in a man with HSV-1 Infection: A case report. *Journal of Clinical Virology*, 59(1):59–62, 2014.

[104] ME Marcocci, G Napoletani, V Protto, O Kolesova, R Piacentini, Puma DD Li, P Lomonte, C Grassi, AT Palamara, and G De Chiara. Herpes simplex virus-1 in the brain: The dark side of a sneaky infection. *Trends in Microbiology*, 28(10): 808–820, 2020.

[105] A Birkmann and H Zimmermann. HSV antivirals – current and future treatment options. *Current Opinion in Virology*, 18:9–13, 2016.

[106] J Esmann. The many challenges of facial herpes simplex virus infection. *Journal of Antimicrobial Chemotherapy*, 41(1):17–27, 2001.

[107] KJ Looker, AS Magaret, MT May, KM Turner, P Vickerman, SL Gottlieb, and LM Newman. Global and regional estimates of prevalent and incident herpes simplex virus type 1 infections in 2012. *PLoS One*, 10(10):e0140765, 2015.

[108] 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*, 17(1):471, 2017.

[109] SB Woo and SJ Challacombe. Management of recurrent oral herpes simplex infections. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 103:S12–e1, 2007.

[110] Y Shen and J Nemunaitis. Herpes simplex virus 1 (HSV-1) for cancer treatment. *Cancer Gene Therapy*, 13(11):975–992, 2006.

[111] MT Shieh and PG Spear. Herpesvirus-induced cell fusion that is dependent on cell surface heparan sulfate or soluble heparin. *Journal of Virology*, 68(2):1224–1228, 1994.

[112] G Smith. Herpesvirus transport to the nervous system and back again. *Annual Review of Microbiology*, 66:153–176, 2012.

[113] GC Perng and C Jones. Towards an understanding of the herpes simplex virus type 1 latency-reactivation cycle. *Interdisciplinary Perspectives on Infectious Diseases*, 2010:262415, 2010.

[114] JB Suzich and AR Cliffe. Strength in diversity: Understanding the pathways to herpes simplex virus reactivation. *Virology*, 522:81–91, 2018.

[115] R Glaser and J Kiecolt-Glaser. Stress-induced immune dysfunction: Implications for health. *Nature Reviews Immunology*, 5(3):243–251, 2005.

[116] DA Padgett, JF Sheridan, J Dorne, GG Berntson, J Candelora, and R Glaser. Social stress and the reactivation of latent herpes simplex virus type 1. *Proceedings of the National Academy of Sciences*, 95(12):7231–5, 1998.

[117] P Bruynseels, PG Jorens, HE Demey, H Goossens, SR Pattyn, MM Elseviers, J Weyler, LL Bossaert, Y Mentens, and M Ieven. Herpes simplex virus in the respiratory tract of critical care patients: A prospective study. *The Lancet*, 362(9395): 1536–41, 2003.

[118] NM Sawtell and RL Thompson. Rapid in vivo reactivation of herpes simplex virus in latently infected murine ganglionic neurons after transient hyperthermia. *Journal of Virology*, 66(4):2150–2156, 1992.

[119] F Cohen, ME Kemeny, KA Kearney, LS Zegans, JM Neuhaus, and MA Conant. Persistent stress as a predictor of genital herpes recurrence. *Archives of Internal Medicine*, 159(20):2430–2436, 1999.

[120] Y Chida and X Mao. Does psychosocial stress predict symptomatic herpes simplex virus recurrence? A meta-analytic investigation on prospective studies. *Brain, Behavior, and Immunity*, 23(7):917–925, 2009.

[121] IR Lehman and PE Boehmer. Replication of herpes simplex virus DNA. *Journal of Biological Chemistry*, 274(40):28059–28062, 1999.

[122] J Akhtar and D Shukla. Viral entry mechanisms, cellular and viral mediators of herpes simplex virus entry. *EBS Journal*, 276(24):7228–36, 2009.

[123] C Hulo, E de Castro, P Masson, L Bougueleret, A Bairoch, I Xenarios, and P Le Mercier. ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Research*, 39(Database issue):D576–82, 2011.

[124] AW Whisnant, CS Jürges, T Hennig, E Wyler, B Prusty, AJ Rutkowski, A L'hernault, L Djakovic, M Göbel, K Döring, J Menegatti, R Antrobus, NJ Matheson, FWH Künzig, G Mastrobuoni, C Bielow, S Kempa, C Liang, T Dandekar, R Zimmer, M Landthaler, F Grässer, PJ Lehner, CC Friedel, F Erhard, and L Dölken. Integrative functional genomics decodes herpes simplex virus 1. *Nature Communications*, 11(1):2038, 2020.

[125] C Prod'hon, I Machuca, H Berthomme, A Epstein, and B Jacquemont. Characterization of regulatory functions of the HSV-1 immediate-early protein ICP22. *Virology*, 226(2):393–402, 1996.

[126] T Rivas, JA Goodrich, and JF Kugel. The herpes simplex virus 1 protein ICP4 acts as both an activator and repressor of host genome transcription during infection. *Molecular and Cellular Biology*, 41(10):e0017121, 2021.

[127] RG Abrisch, TM Eidem, P Yakovchuk, JF Kugel, and JA Goodrich. Infection by herpes simplex virus 1 causes near-complete loss of RNA polymerase II occupancy on the host cell genome. *Journal of Virology*, 90(5):2503–2513, 2016.

[128] CH Birkenheuer, CG Danko, and JD Baines. Herpes simplex virus 1 dramatically alters loading and positioning of RNA polymerase II on host genes early in infection. *Journal of Virology*, 92(8):e02184–17, 2018.

[129] A Rutkowski, F Erhard, A L'Hernault, T Bonfert, M Schilhabel, C Crump, P Rosenstiel, S Efstathiou, R Zimmer, CC Friedel, and Lars Dölken. Widespread disruption of host transcription termination in HSV-1 infection. *Nature Communications*, 6: 7126, 2015.

[130] L Djakovic, T Hennig, K Reinisch, A Milić, AW Whisnant, K Wolf, E Weiß, T Haas, A Grothey, CS Jürges, M Kluge, E Wolf, F Erhard, CC Friedel, and L Dölken. The HSV-1 ICP22 protein selectively impairs histone repositioning upon Pol II transcription downstream of genes. *Nature Communications*, 14(1):4591, 2023.

[131] CC Friedel, AW Whisnant, L Djakovic, AJ Rutkowski, MS Friedl, M Kluge, and L Dölken. Dissecting herpes simplex virus 1-induced host shutoff at the RNA level. *Journal of Virology*, 95(3):10–1128, 2021.

[132] S Marguerat and J Bahler. RNA-seq: From technology to biology. *Cellular and Molecular Life Sciences*, 67(4):569–79, 2010.

[133] Z Dong and Y Chen. Transcriptomics: Advances and approaches. *Science China Life Sciences*, 56(10):960–967, 2013.

[134] U Nagalakshmi, Z Wang, K Waern, C Shou, D Raha, M Gerstein, and M Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–9, 2008.

[135] JZ Levin, M Yassour, X Adiconis, C Nusbaum, DA Thompson, N Friedman, A Gnirke, and A Regev. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, 7:709–15, 2010.

[136] LJ Core, JJ Waterfall, and JT Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–8, 2008.

[137] J Pel, WWY Choi, A Leung, G Shibahara, L Gelinas, M Despotovic, WL Ung, and A Marziali. Duplex proximity sequencing (Pro-Seq): A method to improve DNA sequencing accuracy without the cost of molecular barcoding redundancy. *PLoS One*, 13(10):e0204265, 2018.

[138] H Leng, CV Kasey, CS David, and G Yan. Alternative applications for distinct RNA sequencing strategies. *Briefings in Bioinformatics*, 16:629–639, 2015.

[139] P Maiuri, A Knezevich, A De Marco, D Mazza, A Kula, JG McNally, and A Marcello. Fast transcription rates of RNA polymerase II in human cells. *EMBO Reports*, 12 (12):1280–5, 2011.

[140] D Schmidt, MD Wilson, C Spyrou, GD Brown, J Hadfield, and DT Odom. ChIP-seq: Using highthroughput sequencing to discover protein-DNA interactions. *Methods*, 48 (3):240–8, 2009.

[141] R Mundade, HG Ozer, H Wei, L Prabhu, and T Lu. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, 13(18):2847–52, 2014.

[142] T Furey. ChIP–seq and beyond: New and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012.

[143] PJ Park. ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–80, 2009.

[144] H O'Geen, L Echipare, and PJ Farnham. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods in Molecular Biology*, 791: 265–86, 2011.

[145] A Weiner, TH Hsieh, A Appleboim, HV Chen, A Rahat, I Amit, OJ Rando, and N Friedman. High-resolution chromatin dynamics during a yeast stress response. *Molecular Cell*, 58(2):371–86, 2015.

[146] R Nakato and T Sakata. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods*, 187:44–53, 2021.

[147] JD Buenrostro, B Wu, HY Chang, and WJ Greenleaf. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109:21.29.1–21.29.9, 2015.

[148] J Wu, B Huang, H Chen, Q Yin, Y Liu, Y Xiang, B Zhang, B Liu, Q Wang, W Xia, W Li, Y Li, J Ma, X Peng, H Zheng, J Ming, W Zhang, J Zhang, G Tian, F Xu, Z Chang, J Na, X Yang, and W Xie. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*, 534(7609):652–657031, 2016.

[149] K Suryamohan and MS Halfon. Identifying transcriptional cisregulatory modules in animal genomes. *Wiley Interdisciplinary Reviews Developmental Biology*, 4(2):59–84, 2015.

[150] JD Buenrostro, PG Giresi, LC Zaba, HY Chang, and WJ Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–8, 2013.

[151] W Li, WU Tim, Y Cheng, Y Huang, L Mao, M Sun, C Qiu, L Zhou, and L Gao. Epigenetic application of ATAC-seq based on Tn5 transposase purification technology. *Genetics Research*, 2022:8429207031, 2022.

[152] WS Reznikoff. Transposon Tn5. *Annual Review of Genetics*, 42:269–86, 2008.

[153] C Schmidl, AF Rendeiro, NC Sheffield, and C Bock. ChIPmentation: Fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature Methods*, 12(10): 963–965, 2015.

[154] B He, R Zhu, H Yang, Q Lu, W Wang, L Song, X Sun, G Zhang, S Li, J Yang, G Tian, P Bing, and J Lang. Assessing the impact of data preprocessing on analyzing next generation sequencing data. *Frontiers in Bioengineering and Biotechnology*, 30(8): 817, 2020.

[155] R Bao, L Huang, J Andrade, W Tan, WA Kibbe, H Jiang, and G Feng. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Informatics*, 13(Suppl 2):67–82, 2014.

[156] X Liu, Z Yan, C Wu, Y Yang, X Li, and G Zhang. FastProNGS: Fast preprocessing of next-generation sequencing reads. *BMC Bioinformatics*, 20(1):345, 2019.

[157] Z Wang, M Gerstein, and M Snyder. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[158] S Pepke, B Wold, and A Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(11):S22–S32, 2009.

[159] NA Fonseca, J Rung, A Brazma, and JC Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, 2012.

[160] ML Metzker. Sequencing technologies – the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[161] K Reinert, B Langmead, D Weese, and DJ Evers. Alignment of next-generation sequencing reads. *Annual Review of Genomics and Human Genetics*, 16:133–151, 2015.

[162] A Magi, M Benelli, A Gozzini, F Girolami, F Torricelli, and ML Brandi. Bioinformatics for next generation sequencing data. *Genes*, 1(2):294–307, 2010.

[163] B Langmead, C Trapnell, M Pop, and SL Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3): R25, 2009.

[164] H Li and R Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[165] A Dobin, CA Davis, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and TR Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[166] C Trapnell, L Pachter, and SL Salzberg. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–11, 2009.

[167] T Bonfert, G Csaba, R Zimmer, and CC Friedel. A context-based approach to identify the most likely mapping for RNA-seq experiments. *BMC Bioinformatics*, 13 (Suppl 6):S9, 2012.

[168] MI Love, W Huber, and S Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[169] MD Robinson, DJ McCarthy, and GK Smyth. EdgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010.

[170] ME Ritchie, B Phipson, D Wu, Y Hu, CW Law, W Shi, and GK Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[171] S Ma and Y Zhang. Profiling chromatin regulatory landscape: Insights into the development of ChIP-seq and ATAC-seq. *Molecular Biomedicine*, 1(9), 2020.

[172] Y Zhang, T Liu, CA Meyer, J Eeckhoute, DS Johnson, BE Bernstein, C Nusbaum, RM Myers, M Brown, W Li, and XS Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008.

[173] H Xing, Y Mo, W Liao, and MQ Zhang. Genome-wide localization of protein-dna binding and histone modification by a bayesian change-point method with ChIP-seq data. *PLoS Computational Biology*, 8(7):e1002613, 2012.

[174] Y Guo, S Mahony, and DK Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Computational Biology*, 8(8):e1002638, 2012.

[175] S Heinz, C Benner, N Spann, E Bertolino, YC Lin, P Laslo, JX Cheng, C Murre, H Singh, and CK Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4):576–589, 2010.

[176] TL Bailey, J Johnson, JC Grant, and WS Noble. The MEME suite. *Nucleic Acids Research*, 43(W1):W39–W49, 2015.

[177] G Yu, L Wang, and Q He. ChIPseeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383, 2015.

[178] M Allhoff, K Seré, FJ Pires, M Zenke, and GI Costa. Differential peak calling of ChIP-seq signals with replicates with THOR. *Nucleic Acids Research*, 44(20):e153, 2016.

[179] R Stark and G Brown. DiffBind: Differential binding analysis of ChIP-Seq peak data. *Bioconductor*, 2011.

[180] J Rozowsky, G Euskirchen, RK Auerbach, ZD Zhang, T Gibson, R Bjornson, N Carriero, M Snyder, and MB Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology*, 27(1):66–75, 2009.

[181] L Shen, NY Shao, X Liu, I Maze, J Feng, and EJ Nestler. DiffReps: Detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS One*, 8(6):e65598, 2013.

[182] H Ji, H Jiang, W Ma, DS Johnson, RM Myers, and WH Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnology*, 26(11):1293–300, 2008.

[183] K Liang and S Keles. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*, 28(1):121–2, 2012.

[184] A Mortazavi, BA Williams, K McCue, L Schaeffer, and B Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–8, 2008.

[185] M Hluchý, P Gajdušková, I Ruiz de Los Mozos, M Rájecký, M Kluge, BT Berger, Z Slabá, D Potěšil, E Weiß, J Ule, Z Zdráhal, S Knapp, K Paruch, CC Friedel, and D Blazek. CDK11 regulates pre-mRNA splicing by phosphorylation of SF3B1. *Nature*, 609(7928):829–834, 2022.

[186] NF Isa, O Bensaude, NC Aziz, and S Murphy. HSV-1 ICP22 is a selective viral repressor of cellular RNA polymerase II-mediated transcription elongation. *Vaccines (Basel)*, 9(10):1054, 2021.

[187] JA Watts, J Burdick, J Daigneault, Z Zhu, C Grunseich, A Bruzel, and VG Cheung. Cis elements that mediate RNA polymerase II pausing regulate human gene expression. *The American Journal of Human Genetics*, 105(4):677–688, 2019.

[188] E Weiß, T Hennig, P Graßl, L Djakovic, AW Whisnant, CS Jürges, F Koller, M Kluge, F Erhard, L Dölken, and CC Friedel. HSV-1 infection induces a downstream shift of promoter-proximal pausing for host genes. *Journal of Virology*, 97(5): e0038123, 2023.

[189] F Hahne and R Ivanek. Statistical genomics: Methods and protocols. In E Mathé and S Davis, editors, *Statistical Genomics*, page 335–351, New York, 2016. Springer New York.

[190] MAP Chirackal, K Pilarova, M Kluge, K Bartholomeeusen, M Rajecky, J Oppelt, P Khirsariya, K Paruch, L Krejci, CC Friedel, and D Blazek. CDK12 controls G1/S progression by regulating RNAPII processivity at core DNA replication genes. *EMBO Reports*, 20(9):e47592, 2019.

[191] A Magi, T Pippucci, and C Sidore. XCAVATOR: Accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC Genomics*, 18(1):747, 2017.

[192] E Weiß and CC Friedel. RegCFinder: Targeted discovery of genomic subregions with differential read density. *Bioinformatics Advances*, 3(1):vbad085, 2023.

[193] M Kluge and CC Friedel. Watchdog - a workflow management system for the distributed analysis of large-scale experimental data. *BMC Bioinformatics*, 19(1):97, 2018.

[194] M Kluge, MS Friedl, AL Menzel, and CC Friedel. Watchdog 2.0: New developments for reusability, reproducibility, and workflow execution. *Gigascience*, 9(6):giaa068, 2020.

[195] RM Raisner, PD Hartley, MD Meneghini, MZ Bao, CL Liu, SL Schreiber, OJ Rando, and HD Madhani. Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. *Cell*, 123(2):233–248, 2005.

[196] E Weiß, AW Whisnant, T Hennig, L Djakovic, L Dölken, and CC Friedel. HSV-1 infection induces a downstream shift of the +1 nucleosome, 2024. URL `https://doi.org/10.1101/2024.03.06.583707`.

[197] AD Kwong and NIZA Frenkel. The herpes simplex virus virion host shutoff function. *Journal of Virology*, 63(11):4834–4839, 1989.

[198] CH Birkenheuer and JD Baines. RNA Polymerase II promoter-proximal pausing and release to elongation are key steps regulating herpes simplex virus 1 transcription. *Journal of Virology*, 94(5):e02035–19, 2020.

[199] M Parida, KA Nilson, M Li, CB Ball, HA Fuchs, CK Lawson, DS Luse, JL Meier, and DH Price. Nucleotide resolution comparison of transcription of human cytomegalovirus and host genomes reveals universal use of RNA polymerase II elongation control driven by dissimilar core promoter elements. *mBio*, 10(1):e02047–18, 2019.

[200] NF Marshall and DH Price. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *Journal of Biological Chemistry*, 270 (21):12335–8, 1995.

[201] SM Vos, L Farnung, M Boehning, C Wigge, A Linden, H Urlaub, and P Cramer. Structure of activated transcription complex Pol II-DSIF-PAF-SPT6. *Nature*, 560 (7720):607–612, 2018.

[202] Q Zhou, T Li, and DH Price. RNA polymerase II elongation control. *Annual Review of Biochemistry*, 81:119–43, 2012.

[203] C Laitem, J Zaborowska, NF Isa, J Kufs, M Dienstbier, and S Murphy. CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. *Nature Structural and Molecular Biology*, 22(5):396–403, 2015.

[204] CS Jürges, L Dölken, and F Erhard. Integrative transcription start site identification with iTiSS. *Bioinformatics*, 37(18):3056–3057, 2021.

[205] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–9, 2000.

[206] The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

[207] CH Birkenheuer, L Dunn, R Dufour, and JD Baines. ICP22 of herpes simplex virus 1 decreases RNA polymerase processivity. *Journal of Virology*, 96(5):e0219121, 2022.

[208] J Zaborowska, S Baumli, C Laitem, D O'Reilly, PH Thomas, P O'Hare, and S Murphy. Herpes Simplex Virus 1 (HSV-1) ICP22 protein directly interacts with cyclin-dependent kinase (CDK)9 to inhibit RNA polymerase II transcription elongation. *PLoS One*, 9(9):e107654, 2014.

[209] DA Gilchrist, S Nechaev, C Lee, SK Ghosh, JB Collins, L Li, DS Gilmour, and K Adelman. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes and Development*, 22(14): 1921–33, 2008.

[210] Y Aoi, ER Smith, AP Shah, EJ Rendleman, SA Marshall, AR Woodfin, FX Chen, R Shiekhattar, and A Shilatifard. NELF regulates a promoter-proximal step distinct from RNA Pol II pause-release. *Molecular Cell*, 78(2):261–274.e5, 2020.

[211] J Li, Y Liu, HS Rhee, SKB Ghosh, L Bai, BF Pugh, and DS Gilmour. Kinetic competition between elongation rate and binding of NELF controls promoter-proximal pausing. *Molecular Cell*, 50(5):711–722, 2013.

[212] Continuum Analytics. Conda, 2017. URL `https://conda.io`.

[213] WL Ruzzo and M Tompa. A linear time algorithm for finding all maximal scoring subsequences. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, pages 234–41, 1999.

[214] S Anders, A Reyes, and W Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–17, 2012.

[215] T Hennig, M Michalski, AJ Rutkowski, L Djakovic, AW Whisnant, MS Friedl, BA Jha, MAP Baptista, A L'Hernault, F Erhard, L Dölken, and CC Friedel. HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. *PLoS Pathogens*, 14(3):e1006954, 2018.

[216] TN Mavrich, C Jiang, IP Ioshikhes, X Li, BJ Venters, SJ Zanton, LP Tomsho, J Qi, RL Glaser, SC Schuster, DS Gilmour, I Albert, and BF Pugh. Nucleosome organization in the drosophila genome. *Nature*, 453(7193):358–62, 2008.

[217] DA Gilchrist, G Dos Santos, DC Fargo, B Xie, Y Gao, L Li, and K Adelman. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, 143(4):540–51, 2010.

[218] J Hay, SM Brown, AT Jamieson, FJ Rixon, H Moss, DA Dargan, and JH Subak-Sharpe. The effect of phosphonoacetic acid on herpes viruses. *Journal of Antimicrobial Chemotherapy*, A:63–70, 1977.

[219] Y Becker, Y Asher, Y Cohen, E Weinberg-Zahlering, and J Shlomai. Phosphonoacetic acid-resistant mutants of herpes simplex virus: Effect of phosphonoacetic acid on virus replication and in vitro deoxyribonucleic acid synthesis in isolated nuclei. *Antimicrobial Agents and Chemotherapy*, 11(5):919–22, 1977.

[220] RW Honess and DH Watson. Herpes simplex virus resistance and sensitivity to phosphonoacetic acid. *Journal of Virology*, 21(2):584–600, 1977.

[221] CM Weber, S Ramachandran, and S Henikoff. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Molecular Cell*, 53(5):819–830, 2014.

[222] VA Bondarenko, LM Steele, A Ujvári, DA Gaykalova, OI Kulaeva, YS Polikanov, DS Luse, and VM Studitsky. Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II. *Molecular Cell*, 24(3):469–79, 2006.

[223] A Lashgari, JF Millau, PÉ Jacques, and L Gaudreau. Global inhibition of transcription causes an increase in histone H2A.Z incorporation within gene bodies. *Nucleic Acids Research*, 45(22):12715–12722, 2017.

[224] O Bensaude. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription*, 2(3):103–108, 2011.

[225] PA Lowe. Levels of DNA-dependent RNA polymerases in herpes simplex virus-infected BHK21 C13 cells. *Virology*, 86(2):577–80, 1978.

[226] CM Preston and AA Newton. The effects of herpes simplex virus type 1 on cellular DNA-dependent RNA polymerase activities. *Journal of General Virology*, 33(3): 471–82, 1976.

[227] ON Kostopoulou, V Wilhelmi, S Raiss, S Ananthaseshan, MS Lindström, J Bartek, and C Söderberg-Naucler. Human cytomegalovirus and herpes simplex type I virus can engage RNA polymerase I for transcription of immediate early genes. *Oncotarget*, 8(57):96536–96552, 2017.

[228] KH Seifart and CE Sekeris. Alpha-amanitin, a specific inhibitor of transcription by mammalian RNA-polymerase. *Zeitschrift für Naturforschung B*, 24(12):1538–44, 1969.

[229] TJ Lindell, F Weinberg, PW Morris, RG Roeder, and WJ Rutter. Specific inhibition of nuclear RNA polymerase II by alpha-amanitin. *Science*, 170(3956):447–9, 1970.

[230] S Jimeno-González, M Ceballos-Chávez, and JC Reyes. A positioned +1 nucleosome enhances promoter-proximal pausing. *Nucleic Acids Research*, 43(6):3068–78, 2015.

[231] A Weiner, A Hughes, M Yassour, OJ Rando, and N Friedman. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Research*, 20(1):90–100, 2010.

Appendices

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

CELLULAR RESPONSE TO INFECTION

HSV-1 Infection Induces a Downstream Shift of Promoter-Proximal Pausing for Host Genes

Elena Weiß,[a] Thomas Hennig,[b] Pilar Graßl,[a] Lara Djakovic,[b] Adam W. Whisnant,[b] Christopher S. Jürges,[b] Franziska Koller,[a] Michael Kluge,[a] Florian Erhard,[b] Lars Dölken,[b,c] Caroline C. Friedel[a]

[a]Institute of Informatics, Ludwig-Maximilians-Universität München, Munich, Germany
[b]Institute for Virology and Immunobiology, Julius-Maximilians-University Würzburg, Würzburg, Germany
[c]Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Center for Infection Research (HZI), Würzburg, Germany

ABSTRACT Herpes simplex virus 1 (HSV-1) infection exerts a profound shutoff of host gene expression at multiple levels. Recently, HSV-1 infection was reported to also impact promoter-proximal RNA polymerase II (Pol II) pausing, a key step in the eukaryotic transcription cycle, with decreased and increased Pol II pausing observed for activated and repressed genes, respectively. Here, we demonstrate that HSV-1 infection induces more complex alterations in promoter-proximal pausing than previously suspected for the vast majority of cellular genes. While pausing is generally retained, it is shifted to more downstream and less well-positioned sites for most host genes. The downstream shift of Pol II pausing was established between 1.5 and 3 h of infection, remained stable until at least 6 hours postinfection, and was observed in the absence of ICP22. The shift in Pol II pausing does not result from alternative *de novo* transcription initiation at downstream sites or read-in transcription originating from disruption of transcription termination of upstream genes. The use of downstream secondary pause sites associated with +1 nucleosomes was previously observed upon negative elongation factor (NELF) depletion. However, downstream shifts of Pol II pausing in HSV-1 infection were much more pronounced than observed upon NELF depletion. Thus, our study reveals a novel aspect in which HSV-1 infection fundamentally reshapes host transcriptional processes, providing new insights into the regulation of promoter-proximal Pol II pausing in eukaryotic cells.

IMPORTANCE This study provides a genome-wide analysis of changes in promoter-proximal polymerase II (Pol II) pausing on host genes induced by HSV-1 infection. It shows that standard measures of pausing, i.e., pausing indices, do not properly capture the complex and unsuspected alterations in Pol II pausing occurring in HSV-1 infection. Instead of a reduction of pausing with increased elongation, as suggested by pausing index analysis, HSV-1 infection leads to a shift of pausing to downstream and less well-positioned sites than in uninfected cells for the majority of host genes. Thus, HSV-1 infection fundamentally reshapes a key regulatory step at the beginning of the host transcriptional cycle on a genome-wide scale.

KEYWORDS HSV-1 infection, RNA polymerase II pausing

Lytic herpes simplex virus 1 (HSV-1) infection exerts a profound shutoff of host gene expression. Two major contributors to this shutoff are the degradation of host and viral mRNAs by the virus host shutoff protein (*vhs*) (1, 2) and a general inhibition of the host transcriptional activity by HSV-1 (3–6). Efficient recruitment of RNA polymerase II (Pol II) and elongation factors from the host chromatin to replicating viral genomes leads to a substantial loss of Pol II occupancy from the host genome as early as 2 to 3 h postinfection (h p.i.) (3–6). By 8 h p.i., host transcriptional activity is estimated to be only 10 to 20% of uninfected cells (7). We previously showed that HSV-1 infection disrupts transcription termination for the majority but not all cellular genes, leading to

Editor Anna Ruth Cliffe, University of Virginia

Address correspondence to Lars Dölken, lars.doelken@vim.uni-wuerzburg.de, or Caroline C. Friedel, caroline.friedel@bio.ifi.lmu.de.

The authors declare no conflict of interest.

Received 10 March 2023
Accepted 3 April 2023

read-through transcription for tens of thousands of nucleotides beyond the poly(A) site (8). More recently, Rivas et al. (9) and Birkenheuer et al. (10) demonstrated that HSV-1 also impacts promoter-proximal pausing of Pol II on host genes. Following transcription initiation, Pol II pauses 20 to 60 nucleotides (nt) downstream of the transcription start site (TSS) (11, 12) as a consequence of one or more structural rearrangements within the transcription elongation complex (13). Pausing makes Pol II vulnerable to nucleosome-induced arrest, backtracking of the elongation complex along the DNA, and promoter-proximal premature termination (13). The elongation factor TFIIS can rescue Pol II from pausing and restart transcription by mediating cleavage of backtracked RNA (14). In contrast, 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) stabilize paused Pol II (15, 16). Phosphorylation of DSIF, NELF, and the Pol II C-terminal domain by the CDK9 subunit of the positive transcription elongation factor b (P-TEFb) is required for the release of paused Pol II into gene bodies and the switch to productive elongation (17–19). As a consequence, inhibition of P-TEFb by CDK9 inhibitors increases promoter-proximal pausing (20). While the facilitates chromatin transcription (FACT) histone chaperone complex has previously been reported to cooperate with P-TEFb to overcome NELF/DSIF-mediated inhibition of Pol II elongation (21), recent studies in *Drosophila* instead suggest a role of FACT in the maintenance of Pol II pausing, with FACT knockdown decreasing Pol II pausing (22).

The HSV-1 immediate early protein ICP22 inhibits Pol II transcription elongation by direct interaction with CDK9 (23), and ectopic expression of a short segment of ICP22 mimics the effects of P-TEFb inhibition on Pol II transcription (24). Moreover, ICP22 directly interacts with both FACT subunits (25) and ICP22 is required for the redistribution of FACT as well as the DSIF-subunit SPT5 and the elongation factor SPT6 to viral genomes (24). Consistent with ICP22 mimicking the effects of P-TEFb inhibition, Birkenheuer et al. (10) recently found that Pol II pausing was reduced for a subset of host genes in an ICP22-null mutant (ΔICP22) of HSV-1 compared to a repair virus derived from the null mutant with a genetically restored ICP22. For this purpose, they employed precise nuclear run-on followed by deep sequencing (PRO-seq), which sequences RNA that is actively transcribed by Pol II and depicts strand-specific Pol II transcriptional activity. Transcription initiation from most human gene promoters is bidirectional with productive transcription elongation occurring only in the sense direction (26–29). PRO-seq thus provides nucleotide-level resolution of Pol II activity and allows separating sense and antisense Pol II initiation and pausing. Birkenheuer et al. (5) previously also reported that Pol II levels at the promoter-proximal pause site were altered in a gene-specific manner in HSV-1 strain F (WT-F) infection compared to mock. However, they did not explicitly investigate Pol II pausing for host genes in WT-F infection but only for HSV-1 genes in a later study (30). When comparing promoter-proximal Pol II pausing between mock and HSV-1 strain KOS infection using Pol II chromatin immunoprecipitation sequencing (ChIP-seq), Rivas et al. (9) recently found that HSV-1 infection frequently reduced promoter-proximal Pol II pausing, at least for activated genes. This was largely dependent on ICP4. ICP4 is one of five immediate early proteins (including also ICP0, ICP22, ICP27, and ICP47) expressed shortly after infection and is necessary for the transcription of early and late viral genes (31). HSV-1-activated genes exhibited a greater increase in Pol II occupancy on gene bodies than on promoters, consistent with increased transcriptional elongation. For repressed genes, promoter-proximal Pol II pausing was increased independently of ICP4 with Pol II occupancy decreasing more strongly on gene bodies than in the promoter region.

Here, we report on a genome-wide investigation of the impact of HSV-1 infection on promoter-proximal pausing of all expressed host genes. This is based on a reanalysis of PRO-seq data for mock and 3-h p.i. WT-F infection from the study by Birkenheuer et al. (5). Our reanalysis revealed that HSV-1 infection only seemingly leads to a reduction of Pol II pausing for the majority of genes when using standard Pol II pausing index analyses. More detailed analyses, however, demonstrated that Pol II pausing is retained in HSV-1 infection for most host genes but shifted to sites further downstream

of the promoter. In contrast to well-defined Pol II pausing peaks at the TSS observed in mock infection, HSV-1 infection resulted in more varied and less well-positioned patterns of Pol II pausing. This included both broadening of Pol II pausing peaks into downstream regions for some genes as well as newly originating or increasing Pol II peaks at downstream sites for other genes. In summary, our study demonstrates that HSV-1 impacts promoter-proximal Pol II pausing in a more complex and unexpected manner than previously thought.

RESULTS

Widespread changes in promoter-proximal Pol II pausing during HSV-1 infection. The standard measure for quantifying promoter-proximal pausing is the so-called pausing index (PI) of a gene, which is calculated as the ratio of normalized read counts in a window around the TSS (=promoter window) divided by normalized read counts in a window on the gene body excluding the promoter. We thus started by performing a genome-wide PI analysis using the published PRO-seq data of mock and 3-h p.i. WT-F infection from the study by Birkenheuer et al. (5). Notably, PIs were also used by Rivas et al. (9) to quantify the effects of lytic HSV-1 infection on Pol II pausing from Pol II ChIP-seq and by Birkenheuer et al. (10) to determine differences in Pol II pausing between ΔICP22 and repair virus infection. Since annotated gene 5′ ends do not necessarily reflect the used TSS in a cell type and multiple alternative TSSs are often annotated, we first identified the dominantly used TSS for each gene from published PROcap-seq and PRO-seq data of flavopiridol-treated uninfected human foreskin fibroblasts (HFF) (32) (see Materials and Methods). PROcap-seq is a variation of PRO-seq that specifically maps Pol II initiation sites. Flavopiridol inhibits CDK9 and thus arrests Pol II in a paused state at the TSS (33) and allows also identifying the TSS for genes that are not or weakly paused in untreated cells. Consistent peaks in PROcap-seq and PRO-seq of flavopiridol-treated HFF provided an initial set of 136,090 putative TSS positions, which were further filtered to identify high-confidence sites by requiring a maximum distance of 500 bp to the nearest annotated gene. This identified 42,193 potential TSS positions for 7,650 genes (median number of TSS per gene = 4 with a median distance of 42 bp). For each gene, the TSS with the highest expression was selected for further analysis. Although the PRO-seq data by Birkenheuer et al. (5) was obtained in HEp-2 cells, for most genes the identified TSS in HFF matched very well to PRO-seq peaks in mock-infected HEp-2 cells (Fig. S1a in the supplemental material), better than gene 5′ ends annotated in Ensembl (Fig. S1b).

For PI calculation, normalized read counts were determined in a strand-specific manner as reads per kilobase million (RPKM) in the window from the TSS to TSS + 250 bp for the promoter region and from TSS + 250 bp to TSS + 2,250 bp (or the gene 3′ end if closer) for the gene body. It should be noted that there is no consensus on how to best define promoter and gene body windows for PI calculation and a wide range of alternative ranges have previously been used (see, e.g., references 34–36). Genes with zero reads in the promoter or gene body window in mock or WT-F 3 h p.i. were excluded, resulting in PIs for 7,056 genes (Data set S1 in the supplemental material). This analysis showed that for the vast majority of genes PIs were reduced upon 3-h p.i. HSV-1 infection compared to mock (Fig. 1a). Even with very lenient criteria for an increase in PI, i.e., a fold change >1 in HSV-1 infection compared to mock, only 763 genes (10.8%) showed an increase in PI upon HSV-1 infection (red in Fig. 1a). In contrast, 2,082 genes (29.5%) exhibited a slightly reduced PI (fold change <1 but ≥ 0.5, blue in Fig. 1a) and 4,211 genes (59.7%) showed a strongly reduced PI (fold change <0.5, i.e., more than 2-fold reduced, green in Fig. 1a). Thus, HSV-1 infection induces widespread changes in promoter-proximal Pol II pausing of host genes, resulting in PI reductions for almost all genes and strong PI reductions for the majority of genes. This is consistent with findings by Birkenheuer et al. (5) that Pol II occupancy at promoter-proximal regions was reduced for the majority of genes in WT-F infection compared to mock. While they also found a reduction of Pol II on gene bodies for most of these genes, they did not investigate the relative change between promoter-proximal regions and gene bodies, i.e., the change in PI.

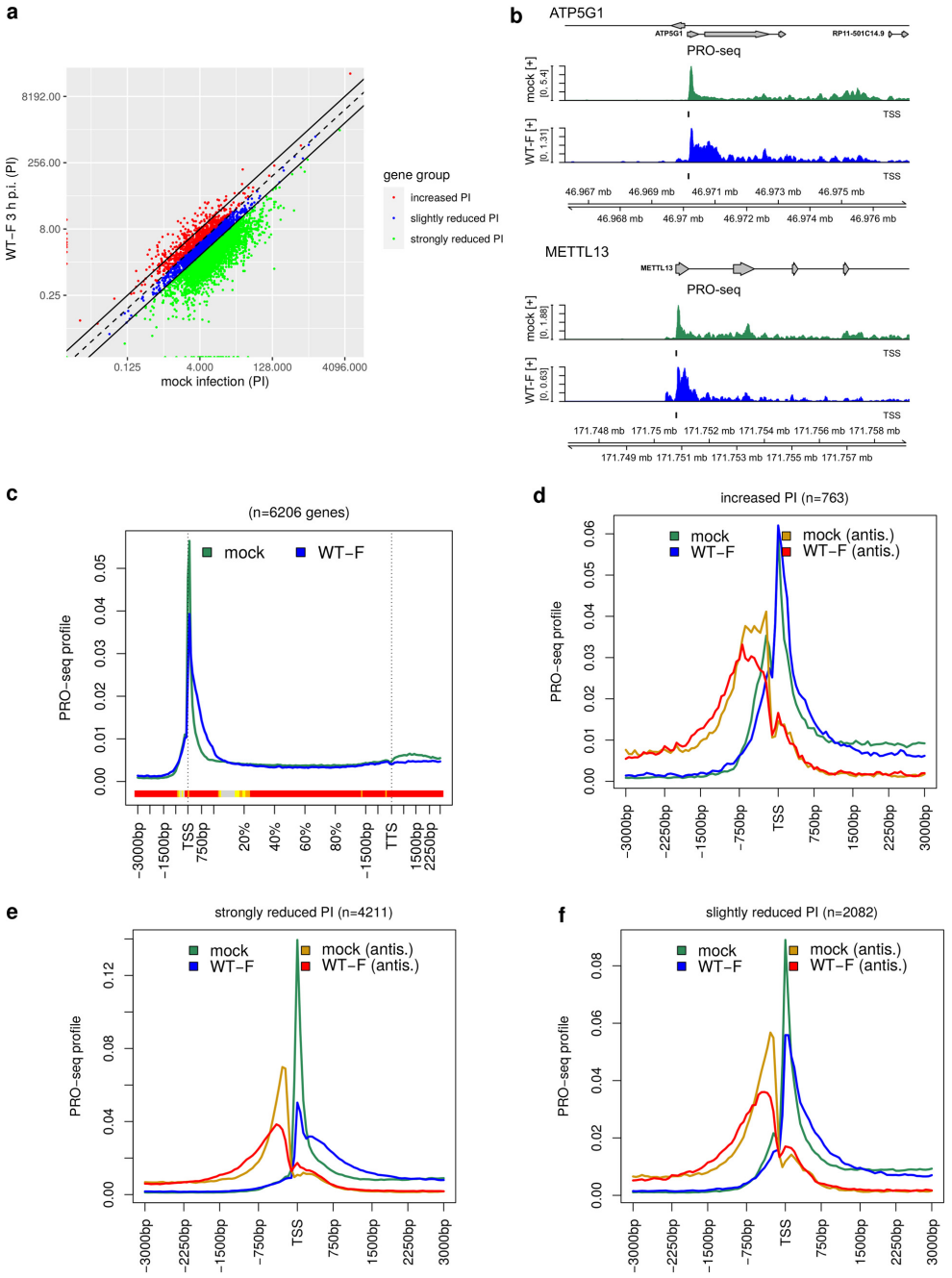


FIG 1 HSV-1 infection impacts promoter-proximal pausing of most host genes. (a) Scatterplots comparing pausing indices (PI) between mock and WT-F infection at 3 h p.i. The dashed line indicates equal PI values and solid lines a 2-fold change in PIs. Genes were divided into three

(Continued on next page)

HSV-1 infection shifts Pol II pausing to downstream sites for most host genes. A disadvantage of PI analyses is that PIs are not only impacted by increases or decreases in Pol II pausing with decreased or increased elongation across the gene body but also by any alteration in Pol II occupancy affecting the number of reads in either promoter or gene body windows. We thus next investigated read coverage in a genome viewer for several example genes with reduced PI. This indicated that changes in promoter-proximal Pol II pausing during HSV-1 infection were highly complex as exemplified by the ATP5G1 and METTL13 genes in Fig. 1b. Instead of narrow promoter-proximal PRO-seq peaks observed in mock infection, promoter peaks in HSV-1 infection often extended into the gene body by a few hundred nucleotides (e.g., ATP5G1) and/or additional downstream peaks were observed as, e.g., for METTL13. However, at the end of these extended/additional peaks, read levels dropped again to similarly low levels relative to the promoter peak as in uninfected cells. Read levels were not increased across the whole gene body relative to the promoter peak as would be expected with increased levels of elongation and productive transcription. The extended promoter peaks and additional downstream peaks in HSV-1 infection reduced PI values, as they extended >250 bp from the TSS into the gene body. Consequently, gene body RPKM, i.e., the denominator in PI calculation, was increased relative to the promoter RPKM, i.e., the numerator. It should be noted that overall Pol II occupancy was reduced both on the promoter and gene body for the majority of genes during HSV-1 infection as already reported by Birkenheuer et al. (5). To allow comparing the distribution of Pol II occupancy, not absolute levels, and visualize downstream shifts in pausing sites, different scales are used for mock and HSV-1 infection in read coverage plots like Fig. 1b. These reflect the highest values observed in mock and HSV-1 infection, respectively, for the selected genomic region.

To investigate whether such complex pausing changes were a global trend, we performed metagene analyses in ±3 kb windows around promoters for all 7,650 analyzed genes (Fig. S2a, excluding 1 gene without reads in some samples, significance analysis for differences in antisense and sense transcription between mock and WT-F infection in Fig. S2b and c, respectively) as well as separately for the three gene groups defined above based on PI changes (Fig. 1d to f, significance analyses shown in Fig. S2d to i). For metagene analyses, the 6 kb promoter windows for each gene were divided into 101 bins. PRO-seq read counts were determined for each bin in a strand-specific manner, normalized to sequencing depth, and averaged across replicates. Subsequently, bin values for each gene were normalized to sum up to 1 to obtain the Pol II occupancy profile around the promoter for each gene before averaging across all genes. This normalization allows comparing Pol II occupancy around the promoter between genes with different expression levels and makes the analysis independent of global

FIG 1 Legend (Continued)
groups according to changes in their PI: (i) increased PI in HSV-1 infection (fold change >1, 763 genes, red); (ii) slightly reduced PI in HSV-1 infection (fold change <1 but ≥0.5, 2,082 genes, blue); and (iii) strongly reduced PI in HSV-1 infection (fold change <0.5, 4,211 genes, green). (b) Read coverage around the TSS in PRO-seq data (sense strand only) for mock (dark green) and WT-F infection at 3 h p.i. (dark blue) for example genes with a reduction in PI upon HSV-1 infection. Read coverage was normalized to total number of mapped reads and averaged between replicates. The identified TSS used in the analysis is indicated by a short vertical line below each read coverage track. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns, and direction is indicated by arrowheads. Genomic coordinates are shown at the bottom. Figures are not centered around the TSS, but a larger region downstream of the TSS was included than upstream of the TSS. (c) Metagene plot showing the distribution of PRO-seq profiles from 3 kb upstream of the TSS to 3 kb downstream of the TTS in sense direction for mock (dark green) and 3-h p.i. WT-F infection (dark blue) for all analyzed genes with a gene length >3 kb. Here, regions from −3 kb to +1.5 kb of the TSS and from −1.5 kb to +3 kb of the TTS were divided into 90-bp bins, respectively, and the remainder of the gene body (+1.5 kb of TSS to −1.5 kb of TTS) into 100 bins of variable length to compare genes with different lengths. Shorter genes were excluded as regions around the TSS and TTS would overlap otherwise, resulting in 6,206 genes. The colored band below the metagene curves indicates the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between mock and 3-h p.i. WT-F infection. *P* values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. *P* value ≤ 10^{-15}, orange = adj. *P* value ≤ 10^{-10}, yellow = adj. *P* value ≤ 10^{-3}. (d to f) Metagene plots showing the distribution of PRO-seq profiles in sense (dark green and blue) and antisense (gold and red) direction from −3 kb to + 3 kb around the TSS for the three gene groups defined in panel a with increased PI (d), strongly reduced PI (e), and slightly reduced PI (f). Mock infection is shown in dark green (sense) and gold (antisense) and WT-F infection at 3 h p.i. in dark blue (sense) and red (antisense). For this purpose, the TSS ± 3 kb promoter window for each gene was divided into 101 bins, and PRO-seq read counts for each sample were determined for each bin, normalized to sequencing depth, and averaged across replicates. Subsequently, bin values for each gene were normalized to sum up to 1 to obtain the Pol II occupancy profile in the promoter window. PRO-seq profiles were determined separately for sense and antisense strands. Results of significance analyses are shown in Fig. S2d to i.

changes in Pol II occupancy between mock and HSV-1 infection. As a consequence, sharp, singular peaks at promoters are characterized by higher peak maxima (e.g., dark green curves in Fig. 1c to f), while broader peaks or multiple peaks have lower peak maxima (e.g., dark blue curves in Fig. 1c, e, and f). Normalization was performed independently for sense and antisense PRO-seq profiles; thus, the height of peaks does not reflect relative levels of sense versus antisense transcription but the distribution of sense and antisense transcription, respectively, around the TSS. To assess the significance of differences between two conditions, Wilcoxon signed-rank tests were performed for each bin in metagene plots comparing normalized coverage values for each gene between the two conditions across all genes. Multiple testing corrected *P* values are color coded at the bottom of metagene plots (red = adjusted [adj.] *P* value $\leq 10^{-15}$; orange = adj. *P* value $\leq 10^{-10}$; yellow = adj. *P* value $\leq 10^{-3}$). If >2 curves are included in metagene plots, significance results for pairwise comparisons are shown in the supplemental material.

Analysis of all genes already showed that PRO-seq profiles for both sense and antisense direction were significantly altered between mock and HSV-1 infection (Fig. S2a to c). In HSV-1 infection, lower Pol II occupancy was observed directly at the TSS and increased occupancy down- and upstream of the TSS for sense and antisense transcription, respectively. This was limited to within 2,250 bp of the TSS in both cases. Metagene analysis on complete genes from the promoter to downstream of the transcription termination site (TTS) confirmed that this relative increase in occupancy downstream of the TSS did not extend across gene bodies (Fig. 1c). Significance analysis showed highly significant differences in the distribution of Pol II occupancy between mock and WT infection for almost the complete gene body, with the notable exception of the region at the end of the promoter window, where increased relative Pol II occupancy in HSV-1 infection downstream of the TSS changed to decreased relative Pol II occupancy on the gene body. The reduction in Pol II occupancy downstream of the TTS during HSV-1 infection reflects the loss of Pol II pausing at the TTS associated with disruption of transcription termination previously reported in HSV-1 infection (8). Interestingly, genes with increased PI upon HSV-1 infection (red in Fig. 1a) only showed a small change in Pol II occupancy in sense direction at the promoter in HSV-1 infection (Fig. 1d; Fig. S2e). In contrast, genes with strong PI reduction upon HSV-1 infection showed a strong reduction of the major peak height at the TSS, a pronounced broadening of the peak into the gene body, and a second minor peak downstream the TSS (Fig. 1e; Fig. S2g). A similar but less pronounced effect was observed for genes with a weak reduction in PI, with a general broadening of the TSS peak but no minor peak (Fig. 1f; Fig. S2i). These changes in the distribution of Pol II occupancy explain the reduction in PIs as read counts in the gene body window are increased relative to read counts in the promoter window.

To identify groups of genes with distinct patterns of changes of Pol II occupancy around the TSS between HSV-1 and mock infection, we performed hierarchical clustering of genes based on their PRO-seq profiles in sense direction for both mock and HSV-1 infection (Fig. 2a). Since we wanted to ensure that genes with distinct patterns of changes were placed in different clusters, a stringent cutoff was applied on the clustering dendrogram to obtain 50 gene clusters at the cost of obtaining multiple clusters with similar patterns. While most clusters exhibited only a narrow peak at the TSS in mock infection (e.g., Fig. 2b; Fig. S3a, b, f to h), some already exhibited a second minor peak shortly after the TSS already before infection (e.g., Fig. 2c and d; Fig. S3c to e). In addition, a few clusters representing a total of 2,018 genes showed peaks in mock infection that were shifted relative to the TSS we had identified (e.g., Fig. 2d; Fig. S3e). In all cases, the peak was shifted at most 750 bp from the identified TSS and was commonly within 100 to 200 bp of the identified TSS. Since the position of peaks within clusters was highly similar due to the stringent clustering cutoff, analysis of individual clusters thus avoids confounding effects resulting from the misidentification of TSS positions. Typical Pol II occupancy changes upon HSV-1 infection included a reduction of peak height at the TSS with a broadening of the peak into the gene body (e.g., Fig. 2b) and changes in minor and major peak heights in case multiple peaks were

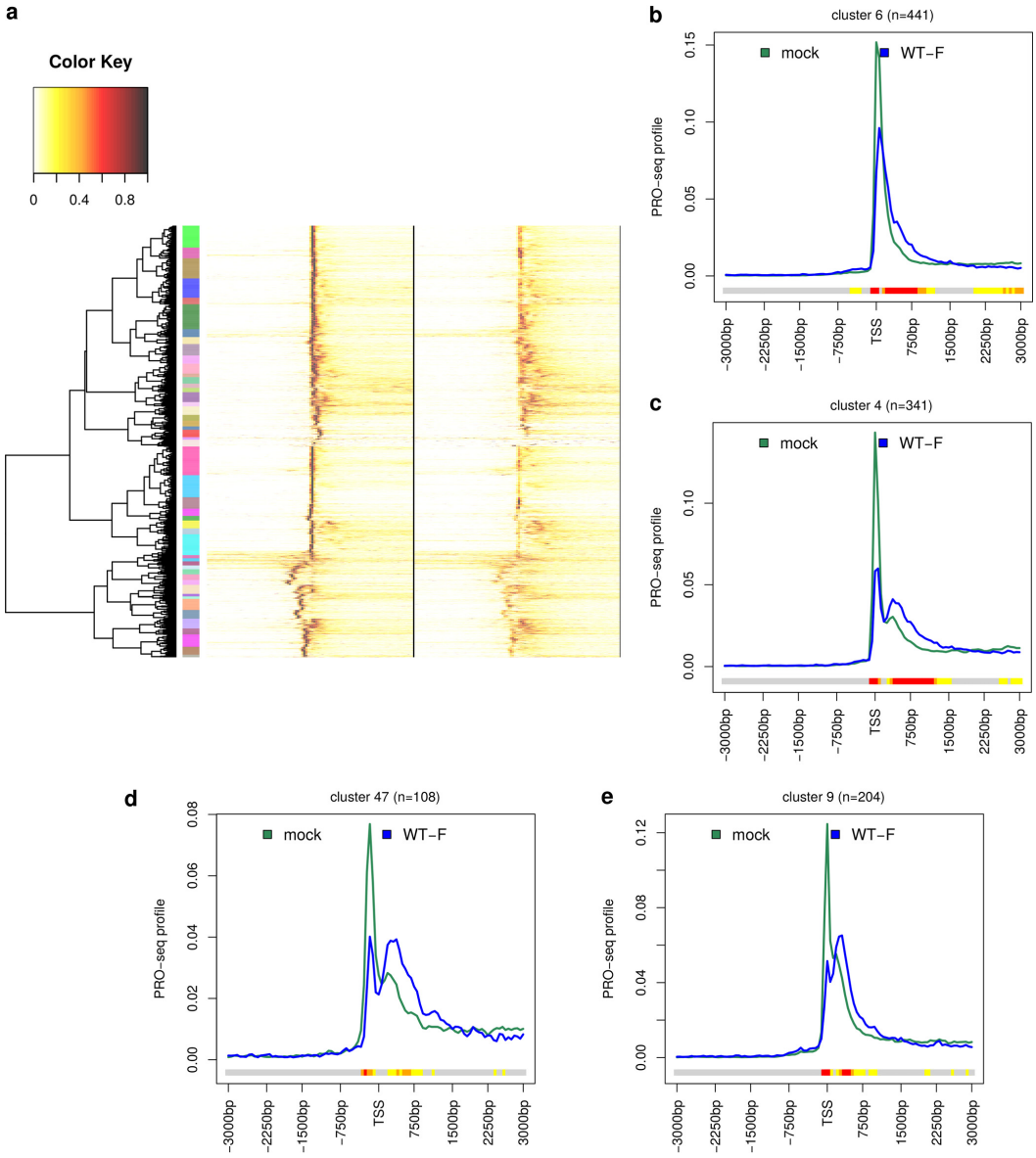


FIG 2 Distinct patterns of changes in promoter-proximal pausing upon HSV-1 infection. (a) Heatmap showing the result of the hierarchical clustering analysis of PRO-seq profiles in mock and WT-F infection. For clustering, PRO-seq profiles in sense direction for mock and WT-F infection were first concatenated and then divided by the maximum value in the concatenated profiles. This resulted in a value of 1 for the position of the highest peak in either mock or HSV-1 infection. Hierarchical clustering was performed according to Euclidean distances and Ward's clustering criterion, and the cutoff on the hierarchical clustering dendrogram was selected to obtain 50 clusters (marked by colored rectangles between the dendrogram and heatmap). Clusters are numbered from top to bottom. (b to e) Metagene plots of PRO-seq profiles on the sense strand in mock (dark green) and WT-F infection at 3 h p.i. (dark blue) for example clusters 6, 4, 47, and 9 in panel a. See Materials and Methods and Fig. 1 legend for an explanation of metagene plots. The colored bands below the metagene curves in each panel indicate the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between mock and 3-h p.i. WT-F infection. *P* values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. *P* value $\leq 10^{-15}$, orange = adj. *P* value $\leq 10^{-10}$, yellow = adj. *P* value $\leq 10^{-3}$.

already present before infection (e.g., Fig. 2c and d) as well as new peaks originating downstream of the TSS in HSV-1 infection (e.g., Fig. 2e). Figure S4 provides an overview on positions, number, and relative heights of Pol II occupancy peaks in mock and HSV-1 infection for each cluster. In total, 30 clusters shared the same major TSS peak between mock and HSV-1 infection, which included also clusters with only a reduction in peak height but no broadening of the TSS peak (cluster 20; Fig. S3f) and clusters without loss of Pol II pausing (clusters 21 and 27; Fig. S3g and h). Twenty-eight clusters showed a second peak in HSV-1 infection downstream of the TSS peak with a median distance of 480 bp to the major peak (e.g., Fig. 2c and d). For 9 of these 28 clusters, the major TSS peak differed between mock and HSV-1 infection (e.g., Fig. 2e). Almost all clusters exhibited a reduced Pol II peak height at the TSS (except clusters 27 and 21, 262 genes; Fig. S3g and h), and most clusters with a reduced Pol II peak height showed an extension of the peak into the gene body or an increased downstream peak (except clusters 7, 20, 23, 24, and 36, 480 genes). Read coverage plots for example genes from different clusters are shown in Fig. S5, and a UCSC genome browser session showing PRO-seq read coverage for all human genes separately for replicates is available at https://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Caroline+Friedel&hgS_otherUserSessionName=PROseq_HSV1.

Recently, the Baines lab also published PRO-seq data of 1.5-, 3-, and 6-h p.i. WT-F, ΔICP22, and its repair virus infection as well as 3-h p.i. WT-F infection with cycloheximide (CHX) treatment (10, 30, 37). This allowed investigation of both the progression of the changes in Pol II pausing during infection as well as the impact of ICP22. Metagene analyses confirmed the downstream shift of Pol II pausing at 3 and 6 h p.i. for both WT-F and the repair virus (Fig. 3a to d; Fig. S6 for significance analysis of pairwise comparisons for WT-F), with a reduction of the major peak height compared to 1.5 h p.i. and a broadening of the TSS peak or increasing or newly originating downstream peaks. Both 3 and 6 h p.i. differed significantly from 1.5 h p.i., but only a few differences were observed between 3 and 6 h p.i. (Fig. S6). Alterations in pausing were slightly less pronounced at 6 h p.i. than at 3 h p.i. for WT-F infection (Fig. 3a and b; Fig. S6), whereas they were slightly more pronounced at 6 h p.i than at 3 h p.i. for the repair virus infection (Fig. 3c and d). Taken together, these results indicate that HSV-1 infection impacts Pol II pausing already very early in infection (after 1.5 h p.i. but before 3 h p.i.) and that the downstream shift in Pol II pausing remains stable until at least 6 h p.i. ΔICP22 infection generally showed the same trend as the repair virus (Fig. 3e and f) with no significant differences downstream of the TSS between ΔICP22 and repair virus infection at 3 and 6 h p.i. (Fig. S7). For several clusters, a small reduction of the TSS peak was observed in ΔICP22 compared to repair virus infection, which was statistically significant for some of these clusters at either 3 h p.i. (clusters 1, 6, and 13) or 6 h p.i. (clusters 6 and 25). This is consistent with the observation by Birkenheuer et al. (10) that pausing indices were increased in the repair virus compared to ΔICP22 infection for 472 and 721 genes at 3 and 6 h p.i., respectively. In summary, these results show that ICP22 is not required for the downstream shift in Pol II pausing but leads to more retention of Pol II directly at the TSS for some genes. Unexpectedly, at 1.5 h p.i. the opposite effect was observed with a significantly increased Pol II pausing peak at the TSS in ΔICP22 infection and reduced Pol II levels downstream of the TSS. A possible explanation for this observation is that ΔICP22 infection progresses more slowly than repair virus infection such that small effects are already detectable in repair virus infection by 1.5 h but not in ΔICP22 infection. By 3 h p.i., the downstream shift in Pol II pausing is then well established in both viruses. Interestingly, while inhibition of protein translation by CHX during the first 3 h of WT-F infection significantly attenuated changes in pausing both at the TSS and downstream of the TSS, some Pol II pausing changes were still observed in the absence of viral protein translation compared to mock infection (Fig. S8). While most clusters showed increased Pol II levels shortly downstream of the TSS in WT-F infection with CHX treatment compared to mock, indicative of a small downstream shift in Pol II pausing, these differences were not statistically significant.

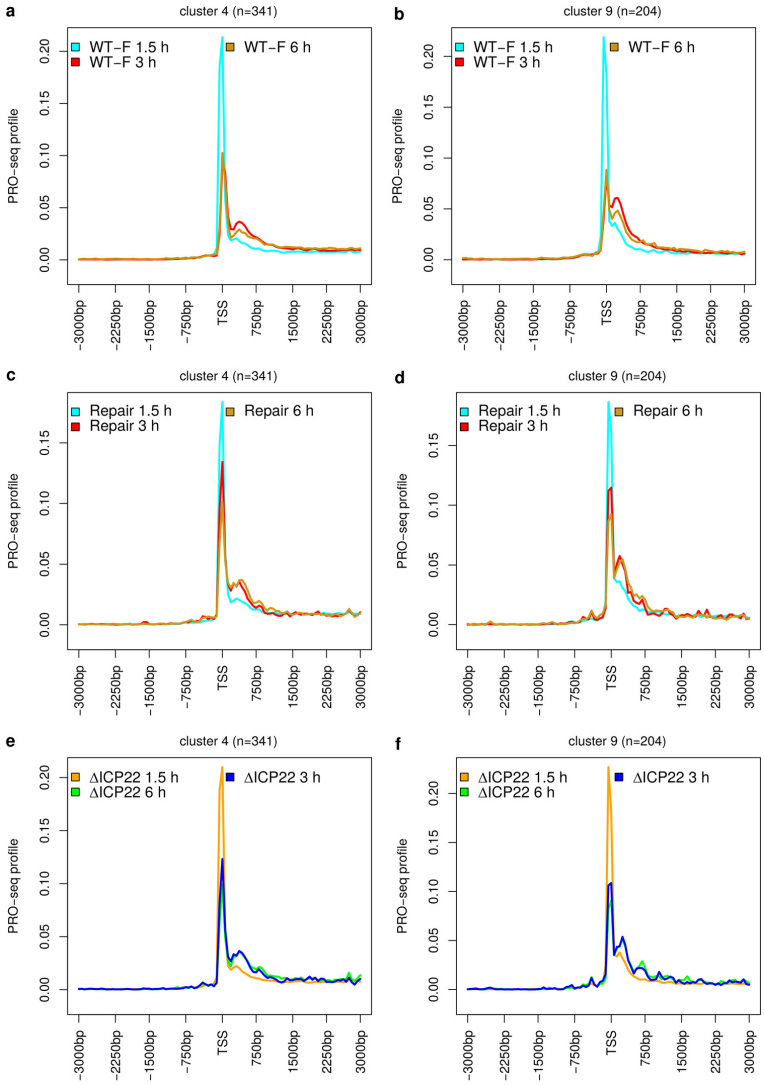


FIG 3 Downstream shift of Pol II pausing throughout the first 6 h of HSV-1 infection and impact of ICP22. Metagene plots of PRO-seq profiles on the sense strand for 1.5-, 3-, and 6-h p.i. WT-F infection (a and b), repair virus infection (c and d), and ΔICP22 infection (e and f) for example clusters in Fig. 2a. Metagene plots with significance analyses of pairwise comparisons between time points for WT-F infection for these and other example clusters can be found in Fig. S6. Metagene plots with significance analyses of pairwise comparisons between repair virus and ΔICP22 infection for the same time points for these and other example clusters can be found in Fig. S7.

However, for some clusters, Pol II occupancy was significantly reduced at or shortly upstream of the TSS and, for many clusters, it was significantly reduced further downstream of the TSS (>1.5 kb). Although the interpretation of these changes is not straightforward, they indicate a role of viral tegument proteins, e.g., VP16, which also interacts with P-TEFb (24), or a virus entry-induced stress or immune response in manipulating Pol II pausing.

To investigate whether the different patterns in Pol II pausing changes were correlated with gene function or transcription factor binding, we performed over- and underrepresentation analysis for Gene Ontology (GO) terms and transcription factor binding motifs from TRANSFAC for each cluster (Data set S2; adj. *P* value cutoff $<$ 0.001, a stringent *P* value cutoff was chosen to adjust for performing this analysis separately for 50 clusters). This revealed an enrichment of subunits of the spliceosomal snRNP complex, specifically U5 and U6 snRNAs, in cluster 27 (adj. *P* value $< 1.31 \times 10^{-7}$), one of the clusters without change in pausing. Since U6 snRNAs are transcribed by RNA polymerase III (38), not Pol II, we investigated the PRO-seq signal of the U6 snRNAs in a genome browser. While some signal was found, it commonly either started already upstream of the U6 snRNA locus or resulted from many reads mapping at the same position. Since U6 snRNA loci are repeated several times in the human genome, we performed a BLAT search for the 70-bp sequence covered by these reads. We found many occurrences of this sequence with few mismatches within Pol II transcribed regions, either in introns of protein-coding genes or in regions downstream of their 3′ end that are still reached by Pol II before RNA cleavage at the upstream poly(A) site. This suggests that these reads were mismapped to other U6 snRNA loci due to sequencing errors making them more similar to these other loci than their actual genomic origins. It is thus not surprising that these loci are enriched in cluster 27, which exhibits no changes between mock and WT-F infection. Nevertheless, they still represent only a very small fraction of this cluster (~7.5%). Cluster 16 was enriched for genes encoding proteins of the large ribosomal subunit (adj. *P* value $<$ 0.00057), but again these represented only a very small fraction (5.9%) of this cluster. No other over- or underrepresentation was observed. Thus, clusters identified based on changes in pausing did not represent functionally related gene groups. Interestingly, however, cluster 32 was strongly enriched for a number of G- or C-rich transcription factor binding motifs, with 90% of genes having a match for a long G-rich motif (GGGMGG GGSSGGGGGGGGGGGG, adj. *P* value $<$ 0.00025). In contrast, several A- and T-rich motifs (e.g., NNNNRNTAATTARY, adj. *P* value $< 6.94 \times 10^{-9}$) were underrepresented. The opposite effect was observed for cluster 6, with G-/C-rich motifs being under- and A-/T-rich motifs being overrepresented. A few G-/C-rich motifs were also underrepresented in cluster 10. Recently, Watts et al. showed that GC content is high around pause sites and that GC skew [= (G − C)/(G + C)] peaks in the 100 nt upstream of the pause site (39). Analysis of GC content and GC skew around the TSS for individual clusters indeed showed a high GC content for cluster 32 at and downstream of the TSS, although no peak in GC skew (Fig. S9a). In contrast, GC content was less increased around the TSS for cluster 6, while GC skew peaked at the TSS and was increased downstream of the TSS (Fig. S9b). As clusters 32 and 6 differed considerably regarding the change in Pol II pausing upon HSV-1 infection, this raises the possibility that sequence composition around the TSS could play a role in determining the changes in Pol II pausing upon HSV-1 infection. However, analysis of the other clusters did not reveal any consistent trend in GC content or GC skew that explained differences in Pol II pausing between clusters. The one consistent trend we observed was that clusters for which the major PRO-seq peak was significantly upstream of our identified TSS commonly exhibited a plateau of high GC content starting at the PRO-seq peak and extending to the TSS (e.g., Fig. S9c and d). It should be noted that the HSV-1 genome is highly GC rich (~68%, Fig. S9e) and almost 50% of viral genome positions have a GC content at least as high as observed at host pause sites (70%; reference 39). Consistently, many occurrences of the G/C-rich over- or underrepresented transcription factor binding motifs for clusters 6 and 32 can be found in the HSV-1 genome (Fig. S9e), with these occurrences being even more G/C-rich than the viral genome overall.

Considering the bidirectionality of transcription initiation at human promoters, we also performed metagene analyses of PRO-seq profiles in antisense direction. Antisense transcription initiation at bidirectional promoters commonly only results in short, unspliced, nonpolyadenylated, and unstable upstream antisense RNAs (uaRNAs) (40) that have highly heterogeneous 3′ ends (41). The metagene analyses of antisense PRO-seq profiles for all

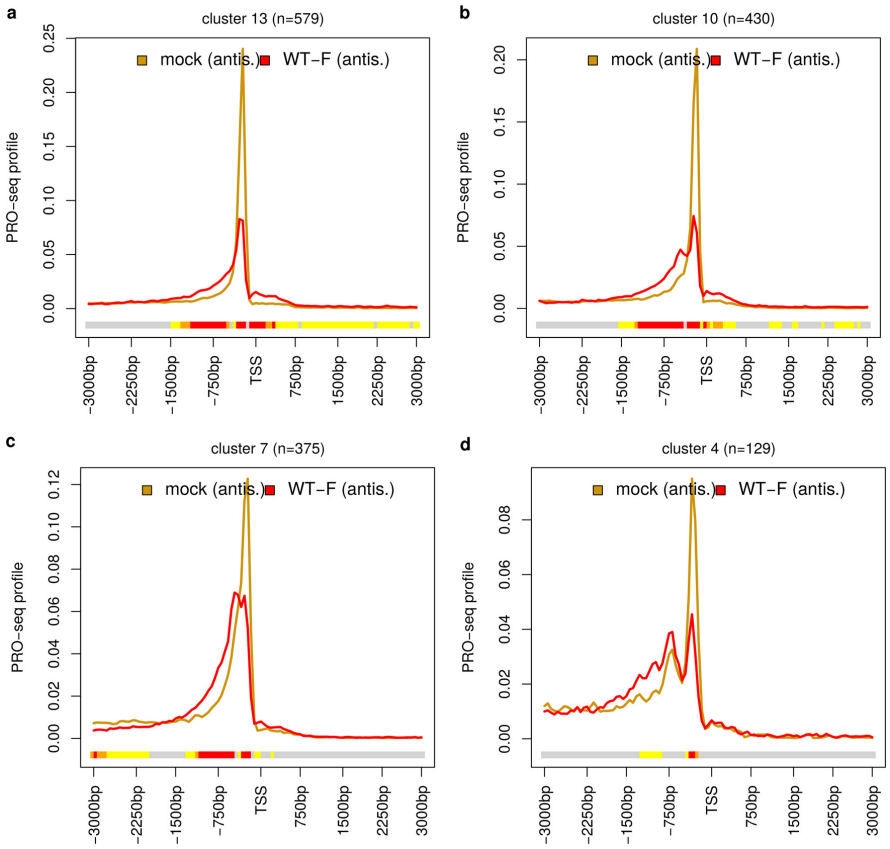


FIG 4 Changes in antisense promoter-proximal pausing in HSV-1 infection. Metagene plots of PRO-seq profiles on antisense direction in mock (gold) and WT-F infection at 3 h p.i. (red) for example clusters resulting from the hierarchical clustering of genes according to antisense PRO-seq profiles in mock and WT-F infection. Here, clustering was performed as described in Fig. 2a legend but applied to concatenated antisense PRO-seq profiles in mock and WT-F infection. Thus, clusters shown here differ from the clusters shown in all other figures. (a) The most common pattern observed for almost all clusters with a broadening of the antisense PRO-seq peak at the TSS. (b to d) The only three clusters that exhibit different patterns with additional peaks originating or increasing in antisense direction during infection. See Materials and Methods and Fig. 1 and 2 legends for the explanation of metagene plots.

genes as well as genes grouped by PI changes also showed a significant reduction in the antisense TSS peak height and a broadening of the peak in the antisense direction (Fig. 1d to f; Fig. S2). However, clustering of antisense PRO-seq profiles in mock and HSV-1 infection with the same approach as for sense profiles to obtain 50 clusters did not identify different patterns between clusters. Most of the 50 antisense clusters exhibited only the same pattern as the metagene analysis of all genes (e.g., Fig. 4a). Only two clusters (430 and 375 genes) showed a small secondary antisense peak originating in HSV-1 infection in addition to the broadening of the antisense signal upstream of the TSS (Fig. 4b and c). One other cluster (129 genes) showed a secondary peak that was already present in mock infection but increased relative to the TSS peak in WT infection (Fig. 4d); however, the increase at this secondary peak was not statistically significant.

Delayed pausing is not an artifact of *de novo* transcription initiation or read-through transcription. Since increasing or newly originating secondary PRO-seq peaks could also represent alternative transcription initiation, we next investigated the presence of alternative TSSs for all clusters in either the PROcap-seq and PRO-seq data

of flavopiridol-treated HFF or the human genome annotation. For most clusters, <15% of genes showed evidence for an alternative TSS at the additional peak positions in either flavopiridol-treated cells (e.g., Fig. S10a and c) or the genome annotation (e.g., Fig. S10b and d). This is in clear contrast to clusters in which the TSS identified from flavopiridol-treated HFF did not represent the dominant TSS in HEp-2 cells. Here, almost 50% of genes had an additional peak in flavopiridol-treated cells or an annotated transcript start at the position of the dominant TSS in HEp-2 cells (e.g., Fig. S10e and f). We furthermore investigated induction of alternative *de novo* transcription initiation downstream of the TSS during HSV-1 infection using cRNA-seq and directional RNA-seq (dRNA-seq) data of transcript 5′ ends for mock and HSV-1 strain 17 (WT-17) infection of HFF from our recent reannotation of the HSV-1 genome ($n = 2$ replicates) (42). cRNA-seq is based on circularization of RNA fragments. dRNA-seq is based on selective cloning and sequencing of the 5′ ends of cap-protected RNA molecules resistant to the 5′–3′-exonuclease XRN1. Both methods strongly enrich reads from 5′-RNA ends. cRNA-seq was performed for mock and 1-, 2-, 4-, 6-, and 8-h p.i. HSV-1 infection. dRNA-seq was performed for mock and 8-h p.i. HSV-1 infection with and without XRN1 treatment. Metagene analyses of cRNA- and dRNA-seq data showed clear peaks coinciding with the major PRO-seq peaks in mock infection and smaller peaks at minor PRO-seq peaks already present in mock infection (Fig. 5a to d; Fig. S11 and S12). In contrast, no (increased) peaks were observed at the positions of downstream PRO-seq peaks that increased or newly originated during HSV-1 infection. This was the case both early (2 and 4 h p.i. in cRNA-seq; Fig. 5a and b; Fig. S11) and later in infection (6 and 8 h p.i. in cRNA-seq, Fig. 5a and b and Fig. S11; 8 h p.i. in dRNA-seq, Fig. 5c and d; Fig. S12) around the 3 and 6 p.i. time points when the downstream shift of Pol II pausing was observed in the PRO-seq data.

Since cRNA- and dRNA-seq were obtained for WT-17 infection, while PRO-seq was performed for WT-F infection, we compared expression changes for host genes between WT-17 and WT-F infection to assess the similarity of virus-induced expression changes. For this purpose, we analyzed total RNA-seq data for WT-17 infection at 8 and 12 h p.i. and WT-F infection at 8 h and 12 h p.i. (Fig. S13). Total RNA-seq data for (i) mock and 8-h p.i. WT-17 infection were taken from our previous study (8), for (ii) mock and 12-h p.i. WT-17 infection from the study by Pheasant et al. (43), and for (iii) mock and 8- and 12-h p.i. WT-F infection from our recent study (44). Since RNA-seq data for WT-F at 8 and 12 h p.i. were obtained in the same experiment and thus expected to be more similar due to less technical noise, we included RNA-seq for WT-17 infection from two different sources to assess the extent of differences that can be ascribed to experimental noise rather than differences between virus strains. This analysis showed that gene expression fold changes compared to mock were highly correlated between WT-17 and WT-F infection both for 8 and 12 h p.i. (Fig. S13a and b), with most genes showing less than a 2-fold difference (indicated by gray lines). Comparison of fold changes for 8 and 12 h p.i. for WT-17 infection between separate experiments (Fig. S13c) showed that observed differences between WT-17 and WT-F were within the range of differences observed in separate experiments. In contrast, fold changes were highly similar between 8 and 12 h p.i. from the same experiment (Fig. S13d). Moreover, when comparing differentially expressed genes ($P < 0.01$) across time points and strains (Fig. S13e), we observed a high consistency across all four conditions, with differences reflecting more the source of the data than the HSV-1 strain. For instance, a group of genes (marked by a red rectangle in Fig. S13e) was partly downregulated in the 12-h p.i. WT-17 data from Pheasant et al. (43) but generally upregulated in the WT-17 8-h p.i and WT-F data from our lab. Since total RNA-seq reflects the cumulative effect of viral infection on host expression up to this time point, the similarity observed between WT-17 and WT-F infection late in infection confirms a strong concordance in virus-induced host expression changes up to this time point between the two strains. We conclude that changes in Pol II occupancy during HSV-1 infection are not due to alternative initiation at novel TSSs leading to capped transcripts. However, we cannot fully exclude that some may reflect

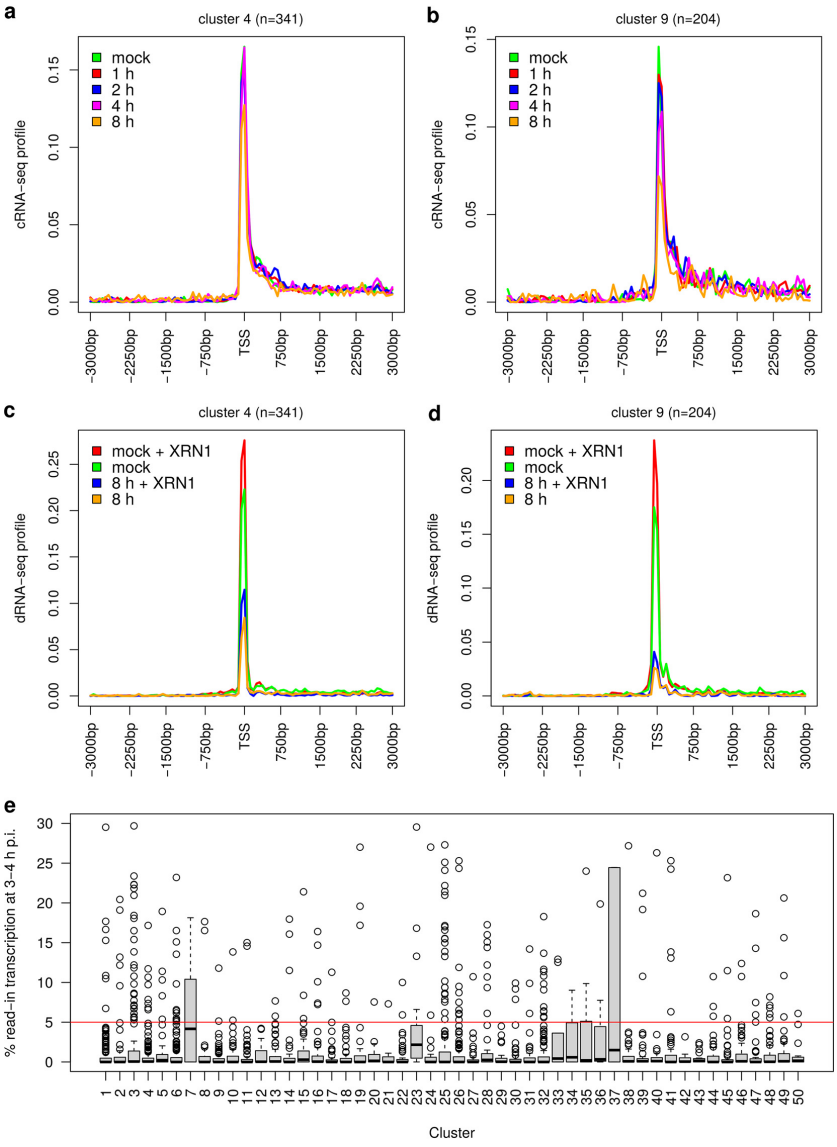


FIG 5 Delayed pausing is not an artifact of alternative *de novo* initiation or read-in transcription. (a to d) Metagene plots of cRNA-seq profiles on the sense strand in mock and WT-17 infection at 1, 2, 4, and 8 h p.i. (a and b) and dRNA-seq profiles on the sense strand in mock and 8-h p.i. WT-17 infection (c and d) with and without XRN1 treatment for example clusters 4 and 9, which show broadening of peaks or additional peaks originating or increasing in height in PRO-seq data during WT-F infection. For metagene plots of PRO-seq profiles for these clusters, see Fig. 2c and e. (e) Boxplots showing the distribution of read-in transcription at 3 to 4 h p.i. for genes in the 50 clusters identified from sense PRO-seq profiles. Boxes represent the range between the first and third quartiles for each cluster. Black horizontal lines in boxes show the median. The ends of the whiskers (vertical lines) extend the box by 1.5 times the interquartile range. Data points outside this range (outliers) are shown as small circles. The red horizontal line indicates the cutoff we previously used to determine that no read-in transcription is observed (≤5% read-in transcription). Metagene plots of PRO-seq profiles for clusters 7, 23, and 33 to 37 with some read-in transcription observed at 3 to 4 h p.i. are shown in Fig. S14.

abortive *de novo* initiation at novel initiation sites downstream of the TSS. This analysis also excludes Pol II creeping, which is observed upon H_2O_2 treatment (45), as the latter would lead to signals from capped transcripts increasing downstream of the TSS in the pausing region.

We previously showed that late in infection "read-in" transcription originating from disrupted transcription termination for an upstream gene commonly extends into downstream genes, which can be mistaken for induction of downstream genes (8). Although read-in transcription only affected very few genes within the first 4 h p.i., increased Pol II occupancy downstream of the TSS could potentially originate from read-in transcription. To quantify read-in transcription, we used our previously published 4-thiouridine sequencing (4sU-seq) time course for every hour of the first 8 h of lytic HSV-1 strain 17 (WT-17) infection of HFF (8). 4sU-seq sequences newly transcribed RNA obtained by labeling with 4sU in specific time intervals of infection (here: 1-h intervals for the first 8 h of lytic expression). Read-in transcription was quantified as previously described (see references 7 and 46 and Materials and Methods for details). In brief, we first calculated the percentage of upstream transcription (=transcription in a 5-kb window upstream of the gene 5′ end/gene expression) for mock infection and each 1-h window of HSV-1 infection. Subsequently, the percentage of read-in transcription was calculated by subtracting values in mock infection from values in each 1-h window of HSV-1 infection (multiplied by 100, negative values set to zero). This analysis included only genes with ≥5 kb to the next up- or downstream gene. By 3 to 4 h p.i., read-in transcription was essentially absent (i.e., much less than 5%) for almost all genes in nearly all clusters (Fig. 5e). Only a few clusters (clusters 7, 23, 33, to 37) exhibited a small extent of read-in transcription already this early in infection; however, these clusters did not exhibit substantial downstream shifts in Pol II occupancy or additional secondary peaks (Fig. S14). The largest of these clusters, cluster 7, indeed showed significantly increased Pol II occupancy in the sense direction upstream of the TSS in HSV-1 infection (Fig. S14a), consistent with read-in transcription. We conclude that read-in transcription extending (partially) into downstream genes does not explain extended TSS peaks or novel or increasing downstream peaks in Pol II occupancy observed in HSV-1 infection.

Delayed pausing in HSV-1 infection occurs downstream of secondary pause sites used upon NELF depletion. Recently, Aoi et al. (47) showed that rapid depletion of NELF, the key mediator of Pol II pausing, does not completely abolish pausing. Instead, Pol II is paused at a secondary more downstream pause site around the +1 nucleosome. Since Rivas et al. (9) showed an ICP4-dependent decrease of NELF in the promoter-proximal region of some HSV-1-activated genes, we reanalyzed PRO-seq data from the study of Aoi et al. (47) for 0-, 1-, 2-, and 4-h auxin-induced degradation of NELF for our 50 clusters to investigate whether changes in pausing upon NELF depletion showed similarities to changes of HSV-1 infection. For this purpose, we also used the TSS positions identified from the PROcap-seq and PRO-seq data of flavopiridol-treated HFF for the NELF degradation data. Although Aoi et al. performed PRO-seq in DLD-1 (colorectal adenocarcinoma) cells, our identified TSS matched well to PRO-seq peak positions for 0 h NELF degradation in these cells (Fig. S15a). This was also confirmed in metagene analyses for our 50 clusters (Fig. 6b and d; Fig. S15c to k).

In the metagene analyses, we indeed observed an increased second PRO-seq peak upon NELF degradation or a broadening of the first PRO-seq peak downstream of the TSS for a few of our clusters (e.g., Fig. 6; Fig. S15b to e). For most clusters, however, we only observed a reduction in the major peak height and a minor broadening of the peak into downstream regions (e.g., Fig. S15f to k). In either case, the changes in the distribution of Pol II occupancy in HSV-1 infection were much more pronounced than after NELF depletion, with more extensive broadening of peaks and new secondary peaks arising further downstream of the major peak. In summary, delayed pause sites in HSV-1 infection are further downstream than "normal" secondary pause sites at +1 nucleosome positions used upon NELF depletion.

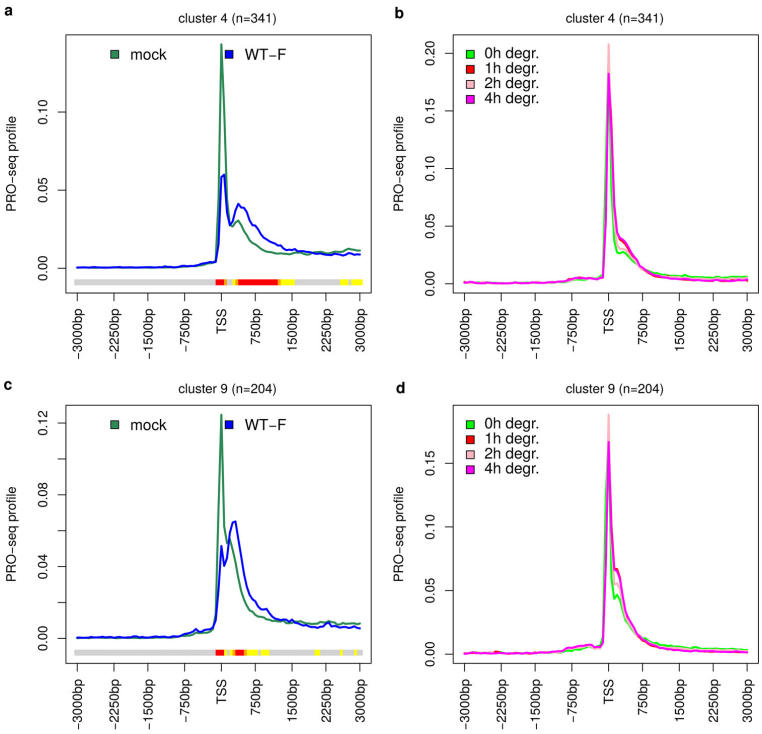


FIG 6 HSV-1 infection leads to stronger downstream shifts in pause sites than NELF depletion. Metagene plots around the TSS of PRO-seq profiles for mock and 3-h p.i. WT-F infection from the study of Birkenheuer et al. (5) (a and c) and 0-, 1-, 2-, and 4-h auxin-inducible degradation of NELF from the study by Aoi et al. (47) (b and d) for example clusters showing a broadening of the TSS peak (a and b) or a small downstream peak (c and d) upon NELF degradation. See Materials and Methods and Fig. 1 and 2 legends for the explanation of metagene plots.

DISCUSSION

Promoter-proximal Pol II pausing is a key regulatory step between transcription initiation and productive elongation. HSV-1 infection has previously been reported to dramatically impact Pol II positioning on host genes, including promoter-proximal regions (5). Promoter-proximal Pol II pausing on HSV-1 genes also plays a key role in HSV-1 transcription (30). Rivas et al. (9) recently reported that HSV-1 infection leads to a reduction in pausing indices for activated host genes and an increase in pausing indices at repressed host genes. While our reanalysis of PRO-seq data of mock and 3-h p.i. WT-F infection also showed a reduction of pausing indices for most expressed host genes during HSV-1 infection, it also illustrated that pausing indices are an inadequate measure of promoter-proximal Pol II pausing. Pausing indices are altered by any change in the distribution of Pol II between the promoter and gene body. Thus, more in-depth analyses are necessary to characterize changes in promoter-proximal Pol II pausing, not only during HSV-1 infection. Our metagene analyses revealed that HSV-1 infection does not lead to a simple reduction of promoter-proximal Pol II pausing with a relative increase of elongating Pol II on the whole gene body. Instead, we observed that Pol II pausing is retained for the vast majority of genes but is shifted to downstream pause sites. This is reflected in broadened Pol II promoter-proximal peaks that extend further into the gene body and newly originating or increasing downstream peaks. A fine-grained clustering analysis identified a wide range of different patterns

for different genes in HSV-1 infection, which contrasts with the sharp promoter-proximal peaks commonly observed in uninfected cells. This indicates that the positioning of shifted pause sites in HSV-1 infection is less well defined than that of "normal" pause sites in uninfected cells. Pronounced downstream shifts of Pol II pausing were only observed after 1.5 h p.i. but remained stable until at least 6 h p.i. Analysis of transcript start site profiling for early (2 and 4 h p.i.) and later (6 and 8 h p.i.) infection time points and of newly transcribed RNA in HSV-1 infection excluded that this was due to *de novo* initiation at downstream sites or read-in transcription originating from disrupted transcription termination for upstream genes.

Interestingly, analysis of promoter-proximal pausing in the antisense direction also showed a broadening of antisense TSS peaks upon HSV-1 infection, with antisense transcription extending further upstream of the TSS than in uninfected cells. However, secondary antisense peaks were only observed for a small fraction of genes (12.2%). It has been proposed that Pol II is particularly prone to pausing and termination during early elongation, specifically on AT-rich sequences often found upstream of promoters (48). Previously, we reported both widespread disruption of transcription termination in HSV-1 infection (8) and activation of antisense transcription at promoters and within gene bodies (49). It is thus tempting to speculate that activation of antisense transcription in HSV-1 infection could be linked to alterations in antisense Pol II pausing, potentially in combination with disruption of transcription termination in the antisense direction.

Nucleosomes represent a natural barrier to transcription and are disassembled before and reassembled after transcribing Pol II (50). Nucleosomes directly downstream of the TSS are generally well positioned at specific locations, in particular the +1 nucleosome, but less so further up- or downstream (51–53). In the presence of NELF, Pol II pausing occurs between the promoter and the +1 nucleosome, and strong positioning of the +1 nucleosome increases pausing (54). While NELF has previously been considered to be required for establishing Pol II pausing, rapid depletion of NELF using auxin-inducible degron does not abolish pausing (47). Instead, pausing appears to be a two-step process with Pol II transitioning from the first to a secondary pause site associated with +1 nucleosomes upon NELF depletion. As Rivas et al. (9) reported decreased levels of NELF at promoter regions of four activated genes tested by ChIP, depletion of NELF from host promoters may play a role in the downstream shift of promoter-proximal Pol II pausing in HSV-1 infection. Notably, a few clusters with additional downstream peaks observed in HSV-1 infection already showed small and much less pronounced peaks at these positions in mock infection. This supports the hypothesis that loss of pausing at major pause sites upon HSV-1 infection leads to pausing of Pol II at secondary downstream pause sites. However, a comparison of the effects of NELF degradation and HSV-1 infection for the 50 identified clusters showed that HSV-1 infection led to much more pronounced alterations in Pol II pausing and more extensive downstream shifts of pause sites than degradation of NELF. We conclude that NELF depletion at promoters alone is unlikely to explain the delay in Pol II pausing observed in HSV-1 infection.

The role of ICP22 in shaping Pol II pausing during WT HSV-1 infection remains an intriguing open question given its previously reported inhibition of the CDK9 subunit of P-TEFb (23). Moreover, lack of ICP22 during HSV-1 infection reduces promoter-proximal Pol II pausing of immediate early genes ICP4, ICP0, and ICP27 and some host genes compared to infection with a repair virus carrying ICP22 (10). We therefore investigated whether infection with an ICP22-null mutant (10) exhibited differences in Pol II pausing compared to the repair virus. While we observed small differences directly at the TSS for some genes, no significant differences between ΔICP22 and the repair virus were detected downstream of the TSS at 3 and 6 h p.i. These results confirm an effect of ICP22 on pausing directly at the TSS for some host genes but also show that ICP22 is not required for the widespread downstream shift of Pol II pausing. Interestingly, global inhibition of protein translation and thus *de novo* lytic viral gene expression by CHX during the first 3 h of HSV-1 infection increased pausing peaks directly at the TSS and largely abolished the downstream shift in Pol II pausing but still resulted in

differences in Pol II occupancy in the promoter region compared to mock infection. The HSV-1 tegument protein VP16 delivered with incoming virions contains a region with structural similarity to ICP22 that interacts with P-TEFb (24). Furthermore, the coexpression of VP16 and ICP22 restores phosphorylation of CDK9 pThr186, a mark of CDK9 activity, which is reduced when expressing ICP22 on its own (24). Based on these and previous findings, Isa et al. (24) proposed a model in which ICP22 activity leads to promoter-proximal stalling and premature termination of elongating Pol II on cellular genes, which is attenuated by VP16 to redirect cellular resources for transcription of viral immediate early genes. This raises the possibility that VP16 plays a role in HSV-1-mediated changes in Pol II pausing. Considering the many ways HSV-1 manipulates the host transcription factory, including other factors involved in Pol II pausing like FACT, SPT5, and SPT6 (required for RNA Pol II progression through +1 and subsequent nucleosomes; reference 55), it is likely that no single viral protein is solely responsible. As such, the downstream shift in Pol II pausing may simply be a by-product of other processes ongoing in HSV-1 infection, e.g., the general loss of Pol II and elongation factors from the host genome and their recruitment to viral genomes. Moreover, considering the high GC content of the HSV-1 genome (68%) similar to the GC content at host pause sites (70%; reference 39) and evidence that high GC content stabilizes DNA-RNA hybrids downstream of the pause site and in this way contributes to pausing (56), HSV-1 may need to manipulate host pausing factors to alleviate Pol II pausing at viral promoters and allow active elongation for viral genes. In this case, the downstream shift of Pol II pausing on host genes may simply be a bystander effect.

The functional impact of the observed changes in promoter-proximal Pol II pausing during HSV-1 infection also remains unclear. Rivas et al. (9) concluded that ICP4 activates host genes by promoting the release of paused Pol II into elongation. However, our analysis showed that Pol II is not fully released from pausing but pausing is shifted to downstream sites for most genes. Nevertheless, the global changes in pausing might still serve to promote increased elongation for a few genes as some genes strongly upregulated in total RNA indeed exhibited increased elongation rather than delayed pausing. This is exemplified by the JUNB gene in Fig. S16a. JUNB encodes the JunB subunit of the heterodimeric AP-1 transcription complex, which is composed of members of the JUN, FOS, ATF, and MAF protein families (57). JUNB is an immediate early gene induced rapidly and transiently by various stimuli and depletion of a NELF subunit increased JUNB expression both before and after induction by interleukin-6 stimulation (58), indicating that NELF-mediated pausing is involved in attenuating JUNB expression. HSV-1 infection has been shown to activate AP-1 binding activity via JNK/SAPK and p38 MAPK pathways, with JunB and JunD being the major AP-1 components by 11 h p.i. (59). While the role of AP-1 in HSV-1 infection has not been completely resolved, AP-1 has recently been reported to induce a gene encoding miR-24, a microRNA that dampens the host antiviral response to HSV-1 (60). On the other hand, the downstream shift in Pol II pausing might also have a negative effect on transcription for affected genes. Premature termination at promoter-proximal pausing sites is both an essential aspect of gene regulation and a response to the accumulation of Pol II stalling and arrest (61). It would thus be tempting to speculate that alterations of Pol II pausing lead to increased premature termination and in this way contribute to the loss of host transcriptional activity. Since transcripts terminated prematurely close to the TSS are generally unprocessed and nonpolyadenylated and thus rapidly degraded (62), they are unlikely to have any functional impact themselves. What lends some weight to this hypothesis is that a number of genes with shifted Pol II pausing play a role in antiviral responses. For instance, METTL3 (Fig. S16b) stabilizes IRF3 mRNA via N^6-methyladenosine modification, and type I interferon (IFN) induction, e.g., in response to HSV-1 infection, is impaired in METTL3 knockout cells (63). In contrast, overexpression of METTL3 enhanced type I IFN induction by HSV-1 (63). Similarly, the DExD-box RNA helicase DDX50 (Fig. S16c) activates the IRF3 signaling pathway following infection with RNA and DNA viruses, including an ICP0-null mutant of HSV-1 (64). Several other DExD/H-box helicases, many of which have

been identified as regulators of antiviral innate immunity (65), show shifts in Pol II pausing, e.g., DHX36 (Fig. S16d, involved in DNA virus sensing; reference 66) or DDX3X (Fig. S5a, contributes to IRF3 activation; reference 67). Notably, however, HSV-1 depends on optimal DDX3X protein levels for viral gene expression, replication, propagation, and infectivity and incorporates DDX3X proteins into mature particles (68–70). This highlights that it is difficult to fully appreciate the functional impact of downstream shifts in Pol II pausing on virus infection.

Finally, it should be noted that our study has also important implications for the analysis of functional genomics studies on HSV-1 and potentially other viral infections. As already observed in previous studies reporting on disruption of transcription termination or activation of antisense transcription observed upon HSV-1 infection (8, 49), standard sequencing data analysis methods are not designed and are thus insufficient to uncover previously unsuspected alterations in transcription. Thus, more in-depth analyses and customized methods are required. In summary, our study highlights a novel aspect in which HSV-1 infection fundamentally alters the host transcriptional cycle, which has implications for our understanding not only of HSV-1 infection but also of the maintenance of Pol II pausing in eukaryotic cells.

MATERIALS AND METHODS

Previously published sequencing data analyzed in this study. PROcap-seq and PRO-seq data of flavopiridol-treated uninfected HFF cells were taken from the study by Parida et al. (32) (GEO accession: GSE113394, samples GSM3104917 and GSM3104913). PRO-seq data for mock and WT-F infection at 3 h p.i. of HEp-2 cells were taken from the study by Birkenheuer et al. (5) (n = 3 replicates, GEO accession: GSE106126, samples GSM2830123 to GSM2830127). PRO-seq data of HEp-2 cells for 1.5, 3, and 6 h of WT-F, ΔICP22, and ICP22 repair virus infection and 3-h p.i. WT-F infection + CHX treatment were taken from studies by Birkenheuer et al. (30) and Dunn et al. (37) (n = 2 to 6, GEO accessions: GSE130342, samples GSM3736426 to GSM3736437; GSE169574, samples GSM5210187 to GSM5210194; GSE202363, samples GSM6112020 to GSM6112028). PRO-seq data for 0, 1, 2, and 4 h of auxin-induced degradation of NELF were taken from the study by Aoi et al. (47) (n = 1 apart from 0 h with n = 2, GEO accession: GSE144786, samples GSM4296314 to GSM4296316, GSM4296318, and GSM4296319). dRNA-seq data for mock and 8-h p.i. HSV-1 infection with and without XRN1 treatment and cRNA-seq data for mock and 1-, 2-, 4-, 6-, and 8-h p.i. HSV-1 infection of HFFF was taken from our previous study (42) (n = 2, GSE128324, samples GSM3671394 to GSM3671411). 4sU-seq data for mock and hourly intervals for the first 8 h of WT-17 infection of HFFF were taken from our previous study (8) (n = 2, GEO accession: GSE59717, samples GSM1444171 to GSM1444179, GSM1444185 to GSM1444193). Total RNA-seq for mock and WT-F and WT-17 infection at 8 and 12 h p.i. were taken from our previous studies (8, 44) (n = 2, GEO accession: GSE59717, samples GSM1444166, GSM1444170, GSM1444180, GSM1444193; GSE185239, samples GSM5608630 to GSM5608643 without PAA) and the study by Pheasant et al. (43) (n = 5, SRA accession: SRP168592, samples SRR8187008 to SRR8187014, SRR8186995 to SRR8186999).

Read alignment. The read alignment pipeline was implemented and run in the workflow management system Watchdog (71, 72). Published sequencing data were first downloaded from SRA using the sratoolkit version 2.10.8. Sequencing reads were aligned against the human genome (GRCh37/hg19) and human rRNA sequences using ContextMap2 version 2.7.9 (73) (using BWA as short read aligner (74) and allowing a maximum indel size of 3 and at most 5 mismatches). PRO-seq reads commonly contain parts of sequence adapters that cannot be aligned to the genome. While these can be removed before alignment using, e.g., cutadapt (75) as outlined in the protocol by Mahat et al. (76), we did not include it in our workflow as ContextMap2 automatically trims parts of reads that cannot be aligned to the genome. As a consequence, adapter sequences were automatically removed during alignment. For sequencing data of HSV-1 infection, alignment also included the HSV-1 genome (human herpesvirus 1 strain 17, GenBank accession code: JN555585). For the two repeat regions in the HSV-1 genome, only one copy was retained each, excluding nucleotides 1 to 9,213 and 145,590 to 152,222 from the alignment. SAM output files of ContextMap2 were converted to BAM files using samtools (77). Read coverage in bedGraph format was calculated from BAM files using BEDTools (78).

Data plotting and statistical analysis. All figures were created in R, and all statistical analyses were performed in R (79). Read coverage plots were created using the R Bioconductor package Gviz (80).

Transcription start site identification. We used the iTiSS program to identify candidate TSS in PROcap-seq and PRO-seq of flavopiridol-treated HFF (42, 81). For this purpose, iTiSS was run separately for each sample in the SPARSE_PEAK mode with standard parameters. Afterward, the iTiSS TSRMerger program was used to select only peaks that were identified in both samples within ±5 bp. Consistent peaks were only further considered if they were within 500 bp of the nearest annotated gene, and for each gene the TSS with the highest read count (weighted by the number of possible alignments for the read) was selected for further analyses.

Calculation of pausing indices. PRO-seq read counts in promoter windows (TSS to TSS + 250 bp) and gene bodies (TSS + 250 bp to TSS + 2,250 bp or gene 3′ end if closer) were determined using featureCounts (82) and gene annotations from Ensembl (version 87 for GRCh37) (83) in a strand-specific manner and normalized by the total number of reads and window lengths to obtain RPKM values. RPKM

values were averaged between replicates and genes with zero reads in either promoter or gene body window were excluded from the analysis. PI for a gene was then calculated as the ratio of promoter RPKM to gene body RPKM.

Metagene and clustering analysis. Metagene analyses were performed as previously described (84) using the R program developed for this previous publication (available with the Watchdog *binGenome* module in the Watchdog module repository (https://github.com/watchdog-wms/watchdog-wms-modules/)). For promoter region analyses, the regions −3 kb to +3 kb of the TSS were divided into 101-bp bins for each gene. For each bin, the average coverage per genome position was calculated in a strand-specific manner for PRO-seq data and bin read coverages were then normalized by dividing by the total sum of all bins. Metagene curves for each replicate were created by averaging results for corresponding bins across all genes and metagene plots and then showing the average metagene curves across replicates. Genes without any reads in any of the analyzed samples were excluded from the analysis. For metagene analyses on the whole gene, the regions from −3 kb to +1.5 kb of the TSS and from −1.5 kb to +3 kb of the TTS were divided into 90-bp bins, and the remainder of the gene body (+1.5 kb of TSS to −1.5 kb of TTS) into 100 bins of variable length to compare genes with different lengths. Genes with a gene length <3 kb were excluded as regions around the TSS and TTS would overlap otherwise. To determine the statistical significance of differences between average metagene curves for two conditions, paired Wilcoxon signed rank tests were performed for each bin comparing normalized coverage values for each gene for this bin between the two conditions. *P* values were adjusted for multiple testing with the Bonferroni method across all bins within each subfigure and are color-coded in the bottom track of subfigures: red = adj. *P* value $\leq 10^{-15}$, orange = adj. *P* value $\leq 10^{-10}$, yellow = adj. *P* value $\leq 10^{-3}$.

For hierarchical clustering analysis, PRO-seq profiles for each gene and condition were calculated for sense or antisense strand as for metagene analyses (without averaging across genes). PRO-seq profiles in promoter windows for mock and WT-F infection at 3 h p.i. were then concatenated and divided by the maximum value in the concatenated vector. Hierarchical clustering was performed using the hclust function in R according to Euclidean distances and Ward's clustering criterion. Peaks in metagene plots for each cluster were then determined in the following way: first, all local and global maxima and minima of metagene curves for each condition were identified for each cluster using the find_peaks function in the R ggpmisc package. The major peak was the global maximum. Subsequently, the next highest local maxima up- or downstream of the major peak were determined and retained as secondary peaks if (i) they were sufficiently removed from the borders of the 6 kb promoter window (i.e., within bins 30 to 80 of the 101 bins), (ii) the difference between the height of the secondary peak and the minimum value between the major and secondary peak was at least 10% of the major peak height, and (iii) the height of the secondary peak was at least 20% of the major peak height.

Over- and underrepresentation analysis. Over- and underrepresentation analysis of Gene Ontology (GO) terms and transcription factor binding motifs from TRANSFAC was performed for each cluster using the g:Profiler webserver (85) and the R package gprofiler2 (86), which provides an R interface to the webserver. *P* values were corrected for multiple testing using the Benjamini-Hochberg false discovery rate (87) and significant terms or motifs were identified at an adjusted *P* value cutoff of 0.001.

Calculation of GC content and GC skew. Genome sequences in the ±3kb around the TSS for each gene were extracted from the hg19 genome with twoBitToFa (http://genome.ucsc.edu/goldenPath/help/twoBit.html) and mean GC content and GC skew (G − C)/(G + C) was calculated in 100-bp sliding windows with steps of 1 bp as described by Watts et al. (39).

Differential gene expression analysis and quantification of read-in transcription. Number of fragments (=read pairs) per gene or in the 5 kb upstream of a gene were determined from mapped paired-end 4sU-seq reads in a strand-specific manner using featureCounts (82) and gene annotations from Ensembl (version 87 for GRCh37). For genes, all fragments overlapping exonic regions on the corresponding strand by ≥25bp were counted for the corresponding gene. For the 5-kb upstream regions, all fragments overlapping the 5 kb upstream of the gene 5′ end were counted. Fold changes in gene expression and statistical significance of changes were determined using DESeq2 (88), and *P* values were adjusted for multiple testing using the method by Benjamini and Hochberg (87). Gene expression and upstream transcriptional activity were quantified in terms of fragments per kilobase of exons per million mapped reads (FPKM). Only reads mapped to the human genome were counted for the total number of mapped reads for FPKM calculation. The percentage of read-in transcription was calculated as previously described (7, 46) for 7,271 genes that had no up- or downstream gene within 5 kb and were well expressed (average FPKM over replicates ≥1) in at least one time point of our 4sU-seq time course. For this purpose, the percentage of transcription upstream of a gene was first calculated separately for each replicate as percentage of upstream transcription = 100 × (FPKM in 5 kb upstream of gene)/(gene FPKM) and averaged between replicates. Second, the percentage of read-in at each 4sU-seq time point of infection was calculated as the percentage of upstream transcription in infected cells minus the percentage of downstream transcription in uninfected cells. Negative values were set to 0.

Code availability. Workflows for PI calculation, metagene analyses, clustering, and figure creation were implemented and run in Watchdog (71, 72) and are available at https://doi.org/10.5281/zenodo.7322848. Corresponding Watchdog modules are available in the Watchdog module repository (https://github.com/watchdog-wms/watchdog-wms-modules/).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, XLSX file, 0.3 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.3 MB.
SUPPLEMENTAL FILE 3, PDF file, 7 MB.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in the framework of the Research Unit FOR5200 DEEP-DV (443644894) project FR2938/11-1 and by grants FR2938/9-1 to C.C.F. and LD1275/6-1 to L.D.

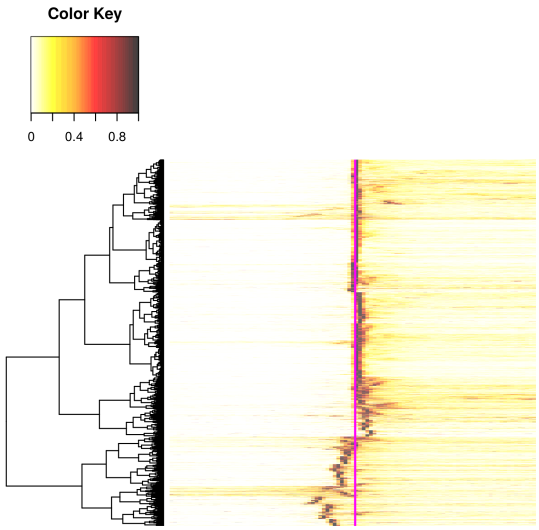
We declare no competing interests.

REFERENCES

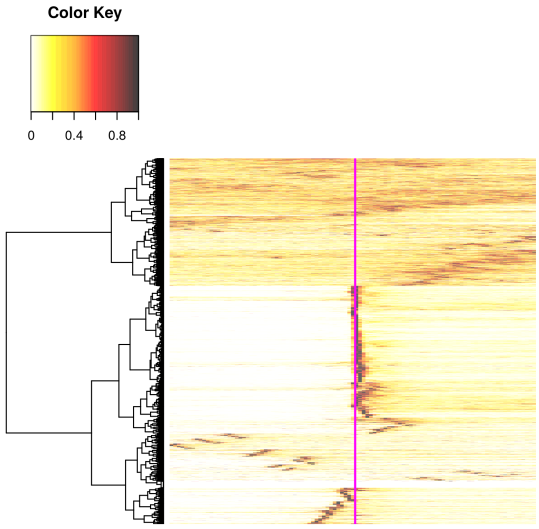
1. Kwong AD, Frenkel N. 1987. Herpes simplex virus-infected cells contain a function(s) that destabilizes both host and viral mRNAs. Proc Natl Acad Sci U S A 84:1926–1930. https://doi.org/10.1073/pnas.84.7.1926.
2. Oroskar AA, Read GS. 1989. Control of mRNA stability by the virion host shutoff function of herpes simplex virus. J Virol 63:1897–1906. https://doi.org/10.1128/JVI.63.5.1897-1906.1989.
3. Spencer CA, Dahmus ME, Rice SA. 1997. Repression of host RNA polymerase II transcription by herpes simplex virus type 1. J Virol 71:2031–2040. https://doi.org/10.1128/JVI.71.3.2031-2040.1997.
4. Abrisch RG, Eidem TM, Yakovchuk P, Kugel JF, Goodrich JA. 2015. Infection by herpes simplex virus 1 causes near-complete loss of RNA polymerase II occupancy on the host cell genome. J Virol 90:2503–2513. https://doi.org/10.1128/JVI.02665-15.
5. Birkenheuer CH, Danko CG, Baines JD. 2018. Herpes simplex virus 1 dramatically alters loading and positioning of RNA polymerase II on host genes early in infection. J Virol 92:e02184-17. https://doi.org/10.1128/JVI.02184-17.
6. Dremel SE, DeLuca NA. 2019. Herpes simplex viral nucleoprotein creates a competitive transcriptional environment facilitating robust viral transcription and host shut off. Elife 8:e51109. https://doi.org/10.7554/eLife.51109.
7. Friedel CC, Whisnant AW, Djakovic L, Rutkowski AJ, Friedl M-S, Kluge M, Williamson JC, Sai S, Vidal RO, Sauer S, Hennig T, Grothey A, Milić A, Prusty BK, Lehner PJ, Matheson NJ, Erhard F, Dölken L. 2021. Dissecting herpes simplex virus 1-induced host shutoff at the RNA level. J Virol 95:e01399-20. https://doi.org/10.1128/JVI.01399-20.
8. Rutkowski AJ, Erhard F, L'Hernault A, Bonfert T, Schilhabel M, Crump C, Rosenstiel P, Efstathiou S, Zimmer R, Friedel CC, Dölken L. 2015. Widespread disruption of host transcription termination in HSV-1 infection. Nat Commun 6:7126. https://doi.org/10.1038/ncomms8126.
9. Rivas T, Goodrich JA, Kugel JF. 2021. The herpes simplex virus 1 protein ICP4 acts as both an activator and repressor of host genome transcription during infection. Mol Cell Biol 41:e0017121. https://doi.org/10.1128/MCB.00171-21.
10. Birkenheuer CH, Dunn L, Dufour R, Baines JD. 2022. ICP22 of herpes simplex virus 1 decreases RNA polymerase processivity. J Virol 96:e02191-21. https://doi.org/10.1128/jvi.02191-21.
11. Gilmour DS, Lis JT. 1986. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. Mol Cell Biol 6:3984–3989. https://doi.org/10.1128/mcb.6.11.3984-3989.1986.
12. Rougvie AE, Lis JT. 1988. The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell 54:795–804. https://doi.org/10.1016/S0092-8674(88)91087-2.
13. Landick R. 2006. The regulatory roles and mechanism of transcriptional pausing. Biochem Soc Trans 34:1062–1066. https://doi.org/10.1042/BST0341062.
14. Sheridan RM, Fong N, D'Alessandro A, Bentley DL. 2019. Widespread backtracking by RNA Pol II is a major effector of gene activation, 5′ pause release, termination, and transcription elongation rate. Mol Cell 73:107–118.e4. https://doi.org/10.1016/j.molcel.2018.10.031.
15. Adelman K, Lis JT. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet 13:720–731. https://doi.org/10.1038/nrg3293.
16. Yamaguchi Y, Shibata H, Handa H. 2013. Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond. Biochim Biophys Acta 1829:98–104. https://doi.org/10.1016/j.bbagrm.2012.11.007.
17. Marshall NF, Price DH. 1995. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. J Biol Chem 270:12335–12338. https://doi.org/10.1074/jbc.270.21.12335.
18. Vos SM, Farnung L, Boehning M, Wigge C, Linden A, Urlaub H, Cramer P. 2018. Structure of activated transcription complex Pol II-DSIF-PAF-SPT6. Nature 560:607–612. https://doi.org/10.1038/s41586-018-0440-4.
19. Zhou Q, Li T, Price DH. 2012. RNA polymerase II elongation control. Annu Rev Biochem 81:119–143. https://doi.org/10.1146/annurev-biochem-052610-095910.
20. Laitem C, Zaborowska J, Isa NF, Kufs J, Dienstbier M, Murphy S. 2015. CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II–transcribed genes. Nat Struct Mol Biol 22:396–403. https://doi.org/10.1038/nsmb.3000.
21. Wada T, Orphanides G, Hasegawa J, Kim DK, Shima D, Yamaguchi Y, Fukuda A, Hisatake K, Oh S, Reinberg D, Handa H. 2000. FACT relieves DSIF/NELF-mediated inhibition of transcriptional elongation and reveals functional differences between P-TEFb and TFIIH. Mol Cell 5:1067–1072. https://doi.org/10.1016/s1097-2765(00)80272-5.
22. Tettey TT, Gao X, Shao W, Li H, Story BA, Chitsazan AD, Glaser RL, Goode ZH, Seidel CW, Conaway RC, Zeitlinger J, Blanchette M, Conaway JW. 2019. A role for FACT in RNA polymerase II promoter-proximal pausing. Cell Rep 27:3770–3779.e7. https://doi.org/10.1016/j.celrep.2019.05.099.
23. Zaborowska J, Baumli S, Laitem C, O'Reilly D, Thomas PH, O'Hare P, Murphy S. 2014. Herpes simplex virus 1 (HSV-1) ICP22 protein directly interacts with cyclin-dependent kinase (CDK)9 to inhibit RNA polymerase II transcription elongation. PLoS One 9:e107654. https://doi.org/10.1371/journal.pone.0107654.
24. Isa NF, Bensaude O, Aziz NC, Murphy S. 2021. HSV-1 ICP22 is a selective viral repressor of cellular RNA polymerase II-mediated transcription elongation. Vaccines 9:1054. https://doi.org/10.3390/vaccines9101054.
25. Fox HL, Dembowski JA, DeLuca NA. 2017. A herpesviral immediate early protein promotes transcription elongation of viral transcripts. mBio 8:e00745-17. https://doi.org/10.1128/mBio.00745-17.
26. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. 2008. The antisense transcriptomes of human cells. Science 322:1855–1857. https://doi.org/10.1126/science.1163853.
27. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. Science 322:1851–1854. https://doi.org/10.1126/science.1164096.
28. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. Science 322:1849–1851. https://doi.org/10.1126/science.1162253.
29. Seila AC, Core LJ, Lis JT, Sharp PA. 2009. Divergent transcription: a new feature of active promoters. Cell Cycle 8:2557–2564. https://doi.org/10.4161/cc.8.16.9305.
30. Birkenheuer CH, Baines JD. 2020. RNA Polymerase II promoter-proximal pausing and release to elongation are key steps regulating herpes simplex virus 1 transcription. J Virol 94:e02035-19. https://doi.org/10.1128/JVI.02035-19.
31. Watson RJ, Clements JB. 1980. A herpes simplex virus type 1 function continuously required for early and late virus RNA synthesis. Nature 285:329–330. https://doi.org/10.1038/285329a0.
32. Parida M, Nilson KA, Li M, Ball CB, Fuchs HA, Lawson CK, Luse DS, Meier JL, Price DH. 2019. Nucleotide resolution comparison of transcription of human cytomegalovirus and host genomes reveals universal use of RNA polymerase II elongation control driven by dissimilar core promoter elements. mBio 10:e02047-18. https://doi.org/10.1128/mBio.02047-18.
33. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, Rando OJ, Madhani HD. 2005. Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. Cell 123:233–248. https://doi.org/10.1016/j.cell.2005.10.002.

34. Williams LH, Fromm G, Gokey NG, Henriques T, Muse GW, Burkholder A, Fargo DC, Hu G, Adelman K. 2015. Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. Mol Cell 58:311–322. https://doi.org/10.1016/j.molcel.2015.02.003.
35. Jonkers I, Kwak H, Lis JT. 2014. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife 3:e02407. https://doi.org/10.7554/eLife.02407.
36. Day DS, Zhang B, Stevens SM, Ferrari F, Larschan EN, Park PJ, Pu WT. 2016. Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. Genome Biol 17:120. https://doi.org/10.1186/s13059-016-0984-2.
37. Dunn LEM, Birkenheuer CH, Dufour R, Baines JD. 2022. Immediate early proteins of herpes simplex virus transiently repress viral transcription before subsequent activation. J Virol 96:e01416-22. https://doi.org/10.1128/jvi.01416-22.
38. Egloff S, O'Reilly D, Murphy S. 2008. Expression of human snRNA genes from beginning to end. Biochem Soc Trans 36:590–594. https://doi.org/10.1042/BST0360590.
39. Watts JA, Burdick J, Daigneault J, Zhu Z, Grunseich C, Bruzel A, Cheung VG. 2019. cis Elements that mediate RNA polymerase II pausing regulate human gene expression. Am J Hum Genet 105:677–688. https://doi.org/10.1016/j.ajhg.2019.08.003.
40. Lykke-Andersen S, Žumer K, Molska EŠ, Rouvière JO, Wu G, Demel C, Schwalb B, Schmid M, Cramer P, Jensen TH. 2021. Integrator is a genome-wide attenuator of non-productive transcription. Mol Cell 81:514–529.e6. https://doi.org/10.1016/j.molcel.2020.12.014.
41. Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. 2015. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. Mol Cell 58:1101–1112. https://doi.org/10.1016/j.molcel.2015.04.006.
42. Whisnant AW, Jürges CS, Hennig T, Wyler E, Prusty B, Rutkowski AJ, L'hernault A, Djakovic L, Göbel M, Döring K, Menegatti J, Antrobus R, Matheson NJ, Künzig FWH, Mastrobuoni G, Bielow C, Kempa S, Liang C, Dandekar T, Zimmer R, Landthaler M, Grässer F, Lehner PJ, Friedel CC, Erhard F, Dölken L. 2020. Integrative functional genomics decodes herpes simplex virus 1. Nat Commun 11:2038. https://doi.org/10.1038/s41467-020-15992-5.
43. Pheasant K, Möller-Levet CS, Jones J, Depledge D, Breuer J, Elliott G. 2018. Nuclear-cytoplasmic compartmentalization of the herpes simplex virus 1 infected cell transcriptome is co-ordinated by the viral endoribonuclease vhs and cofactors to facilitate the translation of late proteins. PLoS Pathog 14:e1007331. https://doi.org/10.1371/journal.ppat.1007331.
44. Djakovic L, Hennig T, Reinisch K, Milic A, Whisnant A, Wolf K. 2021. The HSV-1 ICP22 protein selectively impairs histone repositioning upon Pol II transcription downstream of genes. Res Square https://doi.org/10.21203/rs.3.rs-998249/v1.
45. Nilson KA, Lawson CK, Mullen NJ, Ball CB, Spector BM, Meier JL, Price DH. 2017. Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. Nucleic Acids Res 45:11088–11105. https://doi.org/10.1093/nar/gkx724.
46. Hennig T, Michalski M, Rutkowski AJ, Djakovic L, Whisnant AW, Friedl M-S, Jha BA, Baptista MAP, L'Hernault A, Erhard F, Dölken L, Friedel CC. 2018. HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstreamof genes. PLoS Pathog 14:e1006954. https://doi.org/10.1371/journal.ppat.1006954.
47. Aoi Y, Smith ER, Shah AP, Rendleman EJ, Marshall SA, Woodfin AR, Chen FX, Shiekhattar R, Shilatifard A. 2020. NELF regulates a promoter-proximal step distinct from RNA Pol II pause-release. Mol Cell 78:261–274.e5. https://doi.org/10.1016/j.molcel.2020.02.014.
48. Vlaming H, Mimoso CA, Field AR, Martin BJE, Adelman K. 2022. Screening thousands of transcribed coding and non-coding regions reveals sequence determinants of RNA polymerase II elongation potential. Nat Struct Mol Biol 29:613–620. https://doi.org/10.1038/s41594-022-00785-9.
49. Wyler E, Menegatti J, Franke V, Kocks C, Boltengagen A, Hennig T, Theil K, Rutkowski A, Ferrai C, Baer L, Kermas L, Friedel C, Rajewsky N, Akalin A, Dölken L, Grässer F, Landthaler M. 2017. Widespread activation of antisense transcription of the host genome during herpes simplex virus 1 infection. Genome Biol 18:209. https://doi.org/10.1186/s13059-017-1329-5.
50. Petesch SJ, Lis JT. 2012. Overcoming the nucleosome barrier during transcript elongation. Trends Genet 28:285–294. https://doi.org/10.1016/j.tig.2012.02.005.
51. Bai L, Morozov AV. 2010. Gene regulation by nucleosome positioning. Trends Genet 26:476–483. https://doi.org/10.1016/j.tig.2010.08.003.
52. Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell 132:887–898. https://doi.org/10.1016/j.cell.2008.02.022.
53. Struhl K, Segal E. 2013. Determinants of nucleosome positioning. Nat Struct Mol Biol 20:267–273. https://doi.org/10.1038/nsmb.2506.
54. Jimeno-González S, Ceballos-Chávez M, Reyes JC. 2015. A positioned +1 nucleosome enhances promoter-proximal pausing. Nucleic Acids Res 43:3068–3078. https://doi.org/10.1093/nar/gkv149.
55. Žumer K, Maier KC, Farnung L, Jaeger MG, Rus P, Winter G, Cramer P. 2021. Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo. Mol Cell 81:3096–3109.e8. https://doi.org/10.1016/j.molcel.2021.05.028.
56. Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, Cramer P. 2017. CDK9-dependent RNA polymerase II pausing controls transcription initiation. Elife 6:e29736. https://doi.org/10.7554/eLife.29736.
57. Eferl R, Wagner EF. 2003. AP-1: a double-edged sword in tumorigenesis. Nat Rev Cancer 3:859–868. https://doi.org/10.1038/nrc1209.
58. Aida M, Chen Y, Nakajima K, Yamaguchi Y, Wada T, Handa H. 2006. Transcriptional pausing caused by NELF plays a dual role in regulating immediate-early expression of the junB gene. Mol Cell Biol 26:6094–6104. https://doi.org/10.1128/MCB.02366-05.
59. Zachos G, Clements B, Conner J. 1999. Herpes simplex virus type 1 infection stimulates p38/c-Jun N-terminal mitogen-activated protein kinase pathways and activates transcription factor AP-1. J Biol Chem 274:5097–5103. https://doi.org/10.1074/jbc.274.8.5097.
60. Sharma N, Wang C, Kessler P, Sen GC. 2021. Herpes simplex virus 1 evades cellular antiviral response by inducing microRNA-24, which attenuates STING synthesis. PLoS Pathog 17:e1009950. https://doi.org/10.1371/journal.ppat.1009950.
61. Noe Gonzalez M, Blears D, Svejstrup JQ. 2021. Causes and consequences of RNA polymerase II stalling during transcript elongation. Nat Rev Mol Cell Biol 22:3–21. https://doi.org/10.1038/s41580-020-00308-8.
62. Kamieniarz-Gdula K, Proudfoot NJ. 2019. Transcriptional control by premature termination: a forgotten mechanism. Trends Genet 35:553–564. https://doi.org/10.1016/j.tig.2019.05.005.
63. Chen J, Wei X, Wang X, Liu T, Zhao Y, Chen L, Luo Y, Du H, Li Y, Liu T, Cao L, Zhou Z, Zhang Z, Liang L, Li L, Yan X, Zhang X, Deng X, Yang G, Yin P, Hao J, Yin Z, You F. 2022. TBK1-METTL3 axis facilitates antiviral immunity. Cell Rep 38:110373. https://doi.org/10.1016/j.celrep.2022.110373.
64. Pallett MA, Lu Y, Smith GL. 2022. DDX50 is a viral restriction factor that enhances IRF3 activation. Viruses 14:316. https://doi.org/10.3390/v14020316.
65. Su C, Tang Y-d, Zheng C. 2021. DExD/H-box helicases: multifunctional regulators in antiviral innate immunity. Cell Mol Life Sci 79:2. https://doi.org/10.1007/s00018-021-04072-6.
66. Kim T, Pazhoor S, Bao M, Zhang Z, Hanabuchi S, Facchinetti V, Bover L, Plumas J, Chaperot L, Qin J, Liu Y-J. 2010. Aspartate-glutamate-alanine-histidine box motif (DEAH)/RNA helicase A helicases sense microbial DNA in human plasmacytoid dendritic cells. Proc Natl Acad Sci U S A 107:15181–15186. https://doi.org/10.1073/pnas.1006539107.
67. Gu L, Fullam A, Brennan R, Schröder M. 2013. Human DEAD box helicase 3 couples IκB kinase ε to interferon regulatory factor 3 activation. Mol Cell Biol 33:2004–2015. https://doi.org/10.1128/MCB.01603-12.
68. Khadivjam B, Stegen C, Hogue-Racine M-A, El Bilali N, Döhner K, Sodeik B, Lippé R. 2017. The ATP-dependent RNA helicase DDX3X modulates herpes simplex virus 1 gene expression. J Virol 91:e02411-16. https://doi.org/10.1128/JVI.02411-16.
69. Loret S, Guay G, Lippé R. 2008. comprehensive characterization of extracellular herpes simplex virus type 1 virions. J Virol 82:8605–8618. https://doi.org/10.1128/JVI.00904-08.
70. Stegen C, Yakova Y, Henaff D, Nadjar J, Duron J, Lippé R. 2013. Analysis of virion-incorporated host proteins required for herpes simplex virus type 1 infection through a RNA interference screen. PLoS One 8:e53276. https://doi.org/10.1371/journal.pone.0053276.
71. Kluge M, Friedel CC. 2018. Watchdog–a workflow management system for the distributed analysis of large-scale experimental data. BMC Bioinformatics 19:97. https://doi.org/10.1186/s12859-018-2107-4.
72. Kluge M, Friedl M-S, Menzel AL, Friedel CC. 2020. Watchdog 2.0: new developments for reusability, reproducibility, and workflow execution. GigaScience 9:giaa068. https://doi.org/10.1093/gigascience/giaa068.
73. Bonfert T, Kirner E, Csaba G, Zimmer R, Friedel CC. 2015. ContextMap 2: fast and accurate context-based RNA-seq mapping. BMC Bioinformatics 16:122. https://doi.org/10.1186/s12859-015-0557-5.
74. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

75. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. J Comput Biol 24:1138–1143, 2017. https://doi.org/10.1089/cmb.2017.0096.
76. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ, Lis JT. 2016. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc 11:1455–1476. https://doi.org/10.1038/nprot.2016.086.
77. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10:giab008. https://doi.org/10.1093/gigascience/giab008.
78. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
79. R Core Team. 2022. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
80. Hahne F, Ivanek R. 2016. Visualizing genomic data using Gviz and bioconductor, p 335–351. *In* Mathé E, Davis S (eds), Statistical genomics: methods and protocols. Springer, New York.
81. Jürges CS, Dölken L, Erhard F. 2021. Integrative transcription start site identification with iTiSS. Bioinformatics 37:3056–3057. https://doi.org/10.1093/bioinformatics/btab170.
82. Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi.org/10.1093/bioinformatics/btt656.
83. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Martinez JG, Guijarro-Clarke C, Gymer A, Hardy M, Hollis Z, Hourlier T, Hunt T, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Marugán JC, Mohanan S, Mushtaq A, Naven M, Ogeh DN, Parker A, Parton A, Perry M, Piližota I, Prosovetskaia I, et al. 2022. Ensembl 2022. Nucleic Acids Res 50:D988–D995. https://doi.org/10.1093/nar/gkab1049.
84. Chirackal Manavalan AP, Pilarova K, Kluge M, Bartholomeeusen K, Rajecky M, Oppelt J, Khirsariya P, Paruch K, Krejci L, Friedel CC, Blazek D. 2019. CDK12 controls G1/S progression by regulating RNAPII processivity at core DNA replication genes. EMBO Rep 20:e47592. https://doi.org/10.15252/embr.201847592.
85. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 47:W191–W198. https://doi.org/10.1093/nar/gkz369.
86. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. 2020. gprofiler2–an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler [version 2; peer review: 2 approved]. F1000Res 9:709. https://doi.org/10.12688/f1000research.24956.1.
87. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
88. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8.

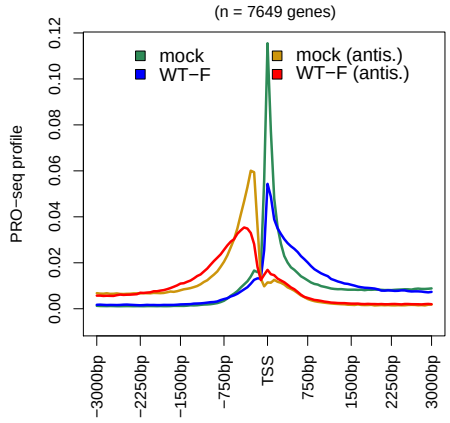


(a)

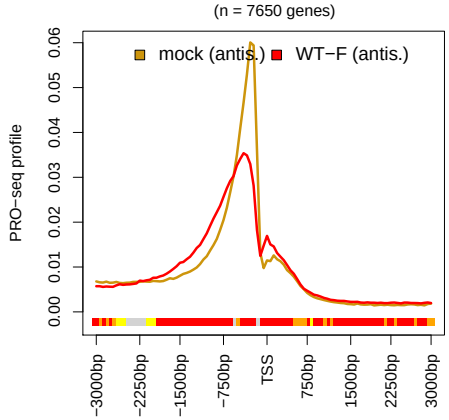


(b)

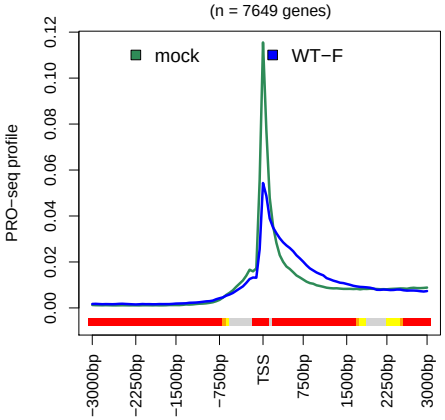
Fig. S1 Heatmaps of PRO-seq profiles on the sense strand in mock infection in a window of ±3 kb around **(a)** the TSS positions identified from PROcap-seq and PRO-seq data of flavopiridol-treated HFF or **(b)** annotated gene 5' ends. For this purpose, PRO-seq profiles were divided by the maximum value in the ±3 kb promoter window, resulting in a value of 1 for the position of the highest peak in PRO-seq profiles. Hierarchical clustering of normalized PRO-seq profiles for all genes was performed using the *hclust* function in R according to Euclidean distances and Ward's clustering criterion. The central position in the promoter window (= the identified TSS) is marked by a vertical magenta line.



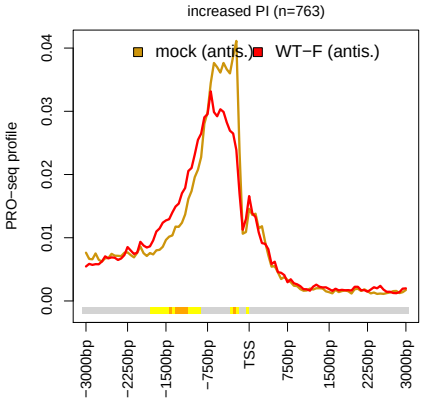
(a)



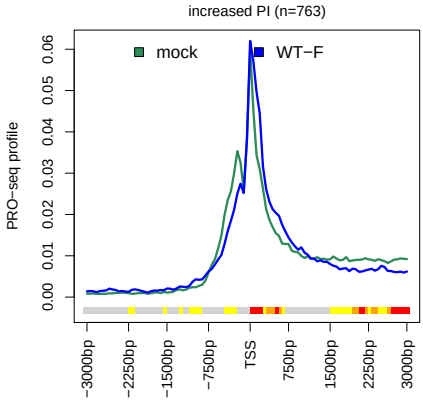
(b)



(c)



(d)



(e)

(Continued on next page)

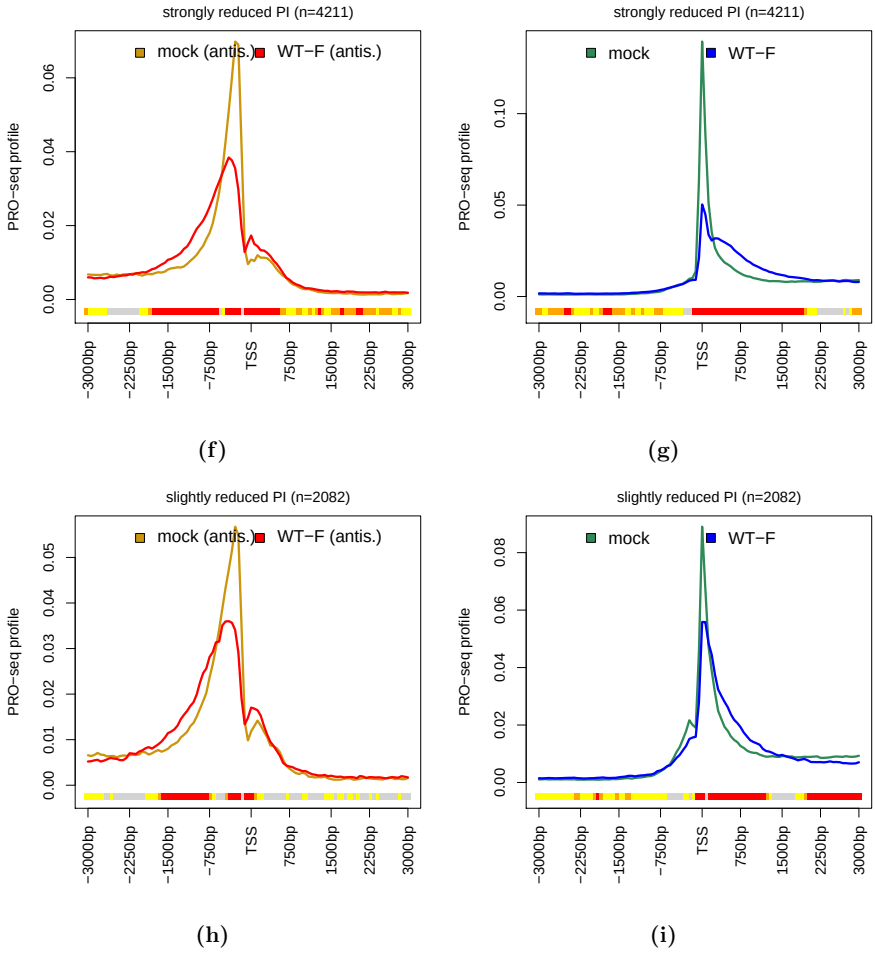
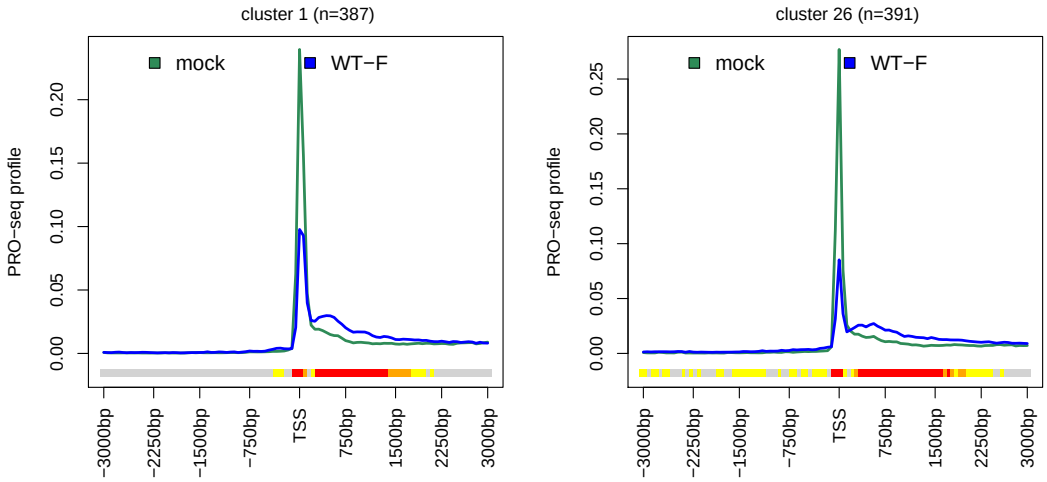
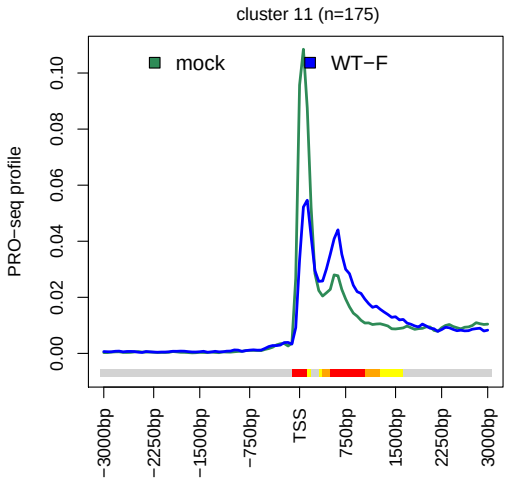


Fig. S2 (a-c) Metagene plot showing the distribution of PRO-seq profiles in sense (dark green and blue) and antisense (gold and red) direction from -3 kb to +3 kb around the TSS for all analyzed genes for mock infection (dark green and gold) and WT-F 3 h p.i. infection (dark blue and red). One gene without reads on the sense strand in some of the analyzed samples was excluded. **(b)** and **(c)** show metagene curves from **(a)** separately for antisense **(b)** and sense **(c)** direction. The color track at the bottom indicates the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between mock and WT-F 3 h p.i. infection. P-values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. p-value $\leq 10^{-15}$, orange = adj. p-value $\leq 10^{-10}$, yellow = adj. p-value $\leq 10^{-3}$. **(d-i)** Metagene plots showing the distribution of PRO-seq profiles separately for antisense **(d,f,h)** and sense **(e,g,i)** direction for genes with increased PI **(d,e)**, strongly reduced PI **(f,g)** and slightly reduced PI **(h,i)**. The color track at the bottom indicates the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between mock and WT-F 3 h p.i. infection.

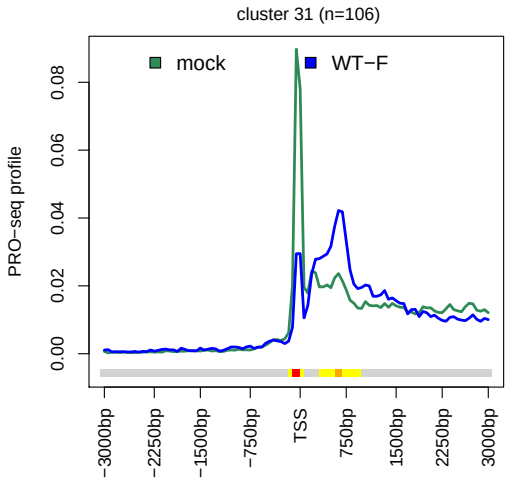


(a)

(b)



(c)



(d)

(Continued on next page)

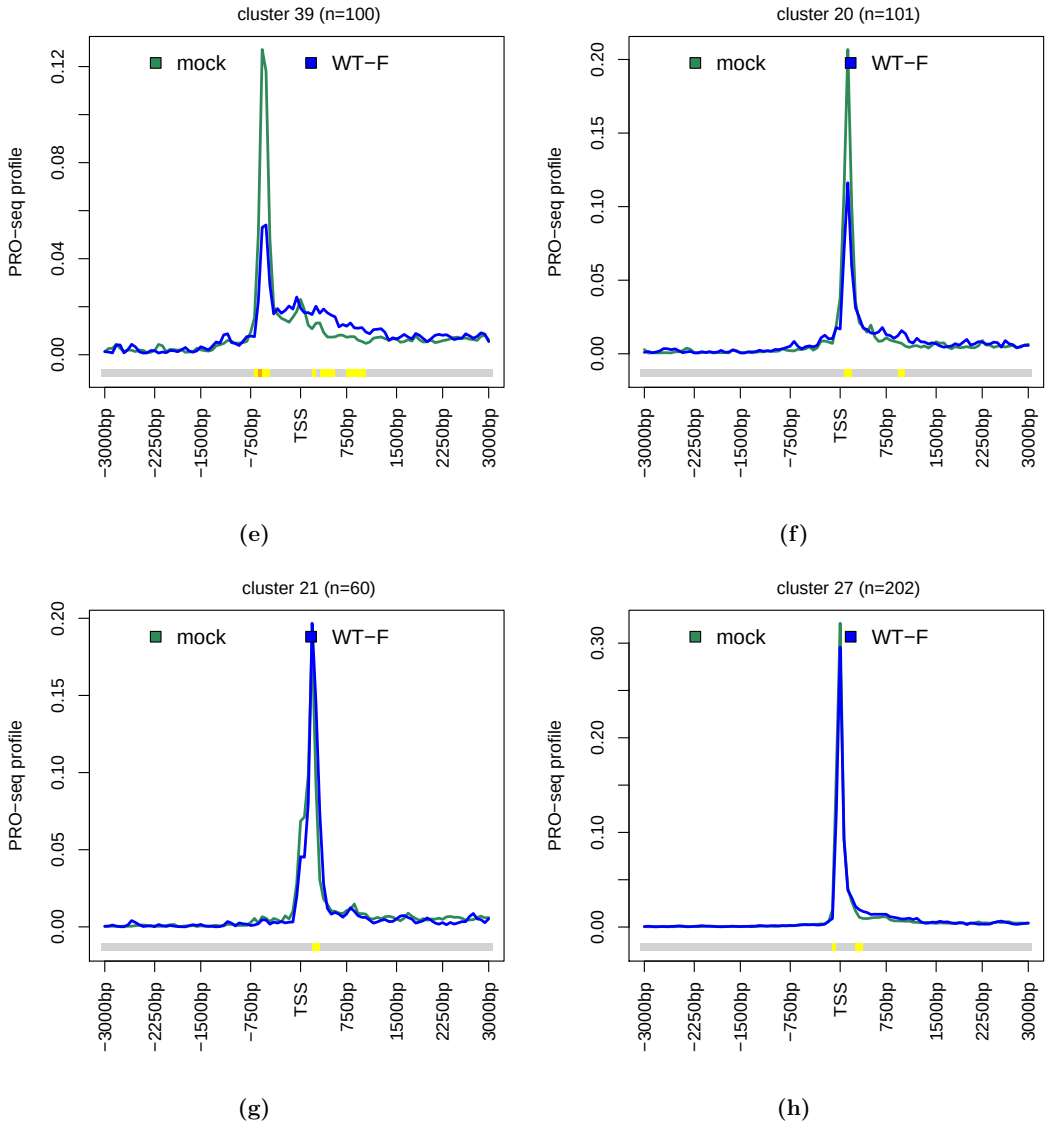
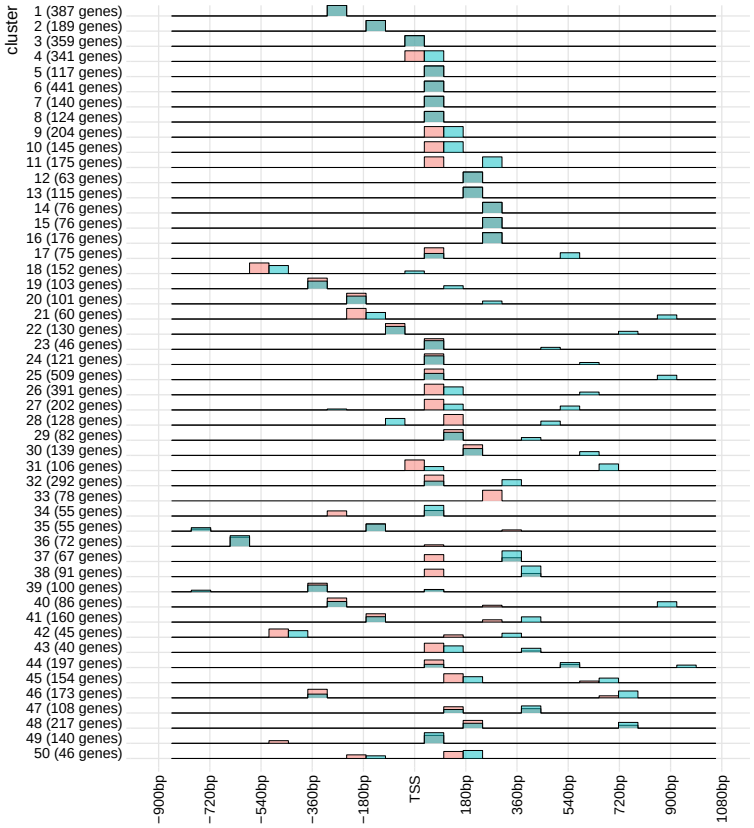


Fig. S3 Metagene plots showing the PRO-seq profile in sense direction from -3 kb to +3 kb around the TSS for mock infection (dark green) and WT-F 3 h p.i. infection (dark blue) separately for example clusters. Cluster numbers and number of genes in each cluster are indicated on top of subfigures. The color track at the bottom of each subfigure indicates the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between mock and WT-F 3 h p.i. infection. P-values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. p-value $\leq 10^{-15}$, orange = adj. p-value $\leq 10^{-10}$, yellow = adj. p-value $\leq 10^{-3}$.

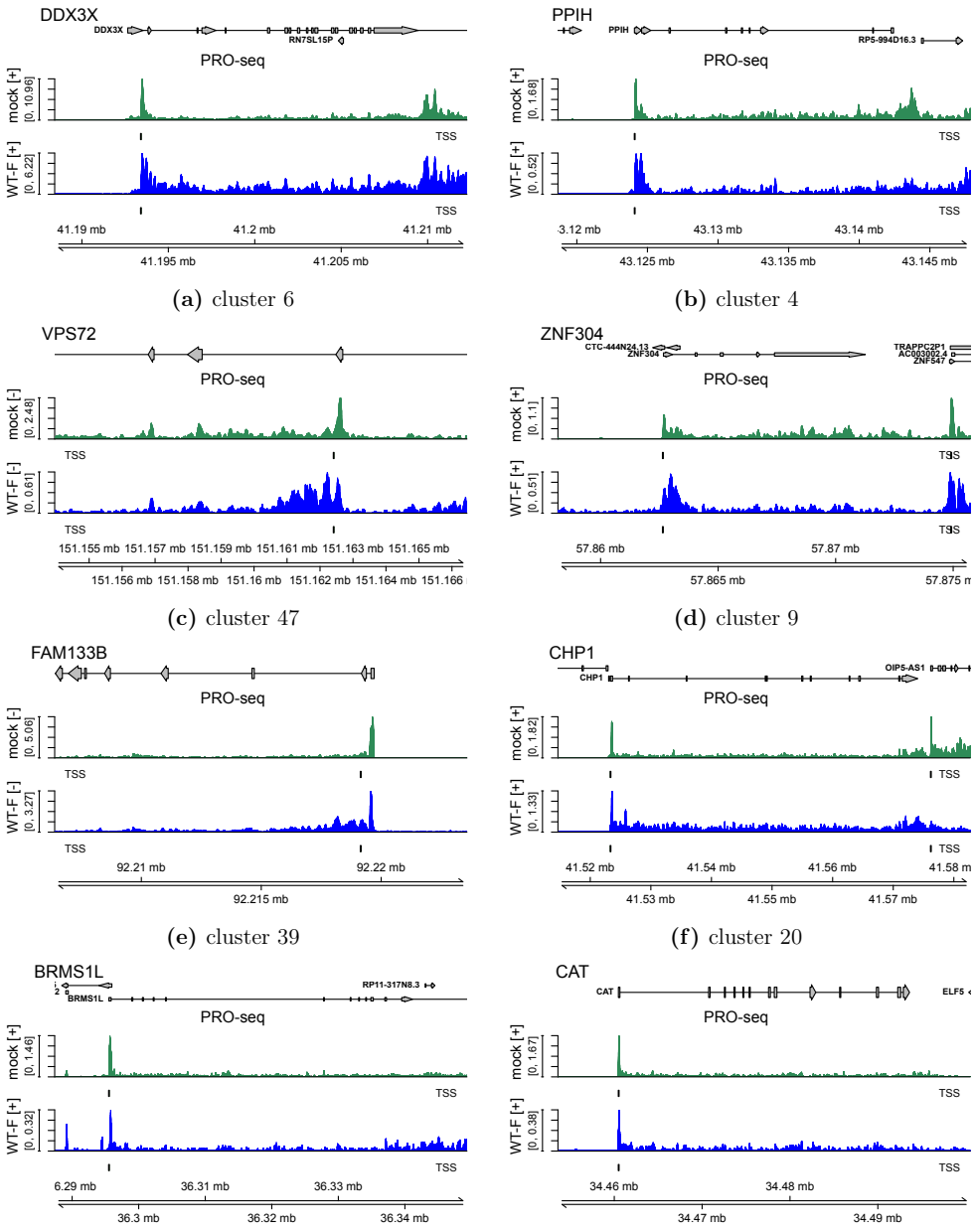


(a)

Peak pattern	mock		WT-F 3 h p.i.	
	no. clusters	no. genes	no. clusters	no. genes
One TSS peak	33	5728	21	3347
Additional minor peak downstream of major TSS peak	13	1461	17	2986
Two approximately equally high peaks	2	232	4	296
Additional downstream peak higher than TSS peak	0	0	7	834
none of the above	2	228	1	46

(b)

Fig. S4 (a) Positions, number, and relative heights of peaks identified in PRO-seq profiles in sense direction for the 50 clusters. Mock infection is shown in light red and WT-F 3 h p.i. infection in turquoise. Darker turquoise indicates that a peak is present at the same position in mock and WT-F 3 h p.i. infection. The relative peak height is calculated as the peak height divided by the sum of all peak heights for the same condition. Thus, a single peak has a value of 1, two equally high peaks both have a value of 0.5, and so on. **(b)** Statistics on the number of clusters and number of genes with different types of peak patterns defined by the number and relative height of peaks for mock and WT-F 3 h p.i. infection shown in **(a)**.



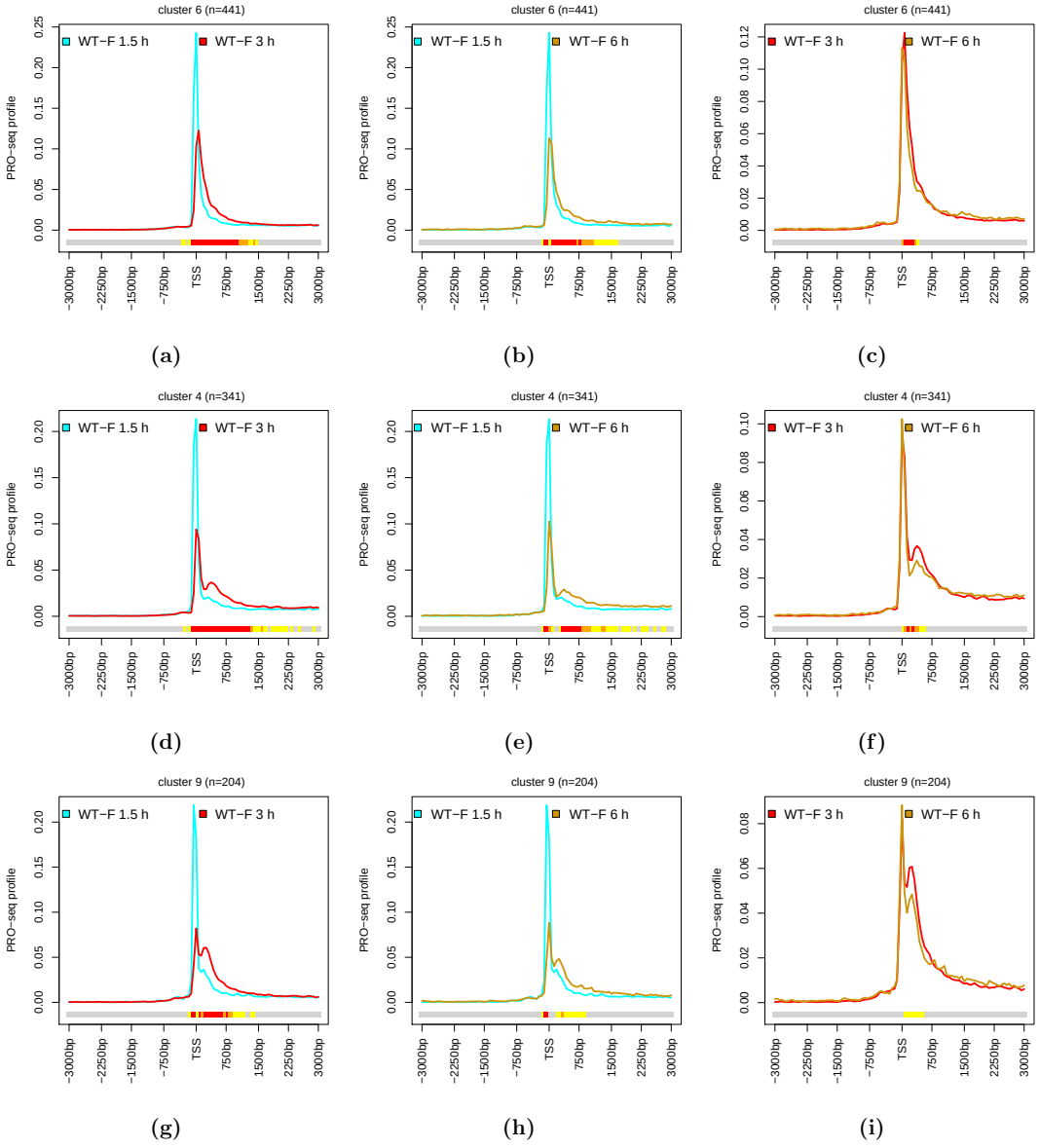
(a) cluster 6 **(b)** cluster 4

(c) cluster 47 **(d)** cluster 9

(e) cluster 39 **(f)** cluster 20

(g) cluster 21 **(h)** cluster 27

Fig. S5 Read coverage around the TSS in PRO-Seq data (sense strand only) for mock (green) and WT-F infection (blue) at 3 h p.i. for example genes (gene name of the selected gene on the top left) in different clusters (cluster number shown below subfigures). Read coverage was normalized to total number of mapped reads and averaged between replicates. The identified TSS used in the analysis is indicated by a short vertical line below each read coverage track. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns and direction is indicated by arrowheads. Genomic coordinates are shown on the bottom. Please note that figures are not centered around the TSS, but a larger region downstream of the TSS was included than upstream of the TSS.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(Continued on next page)

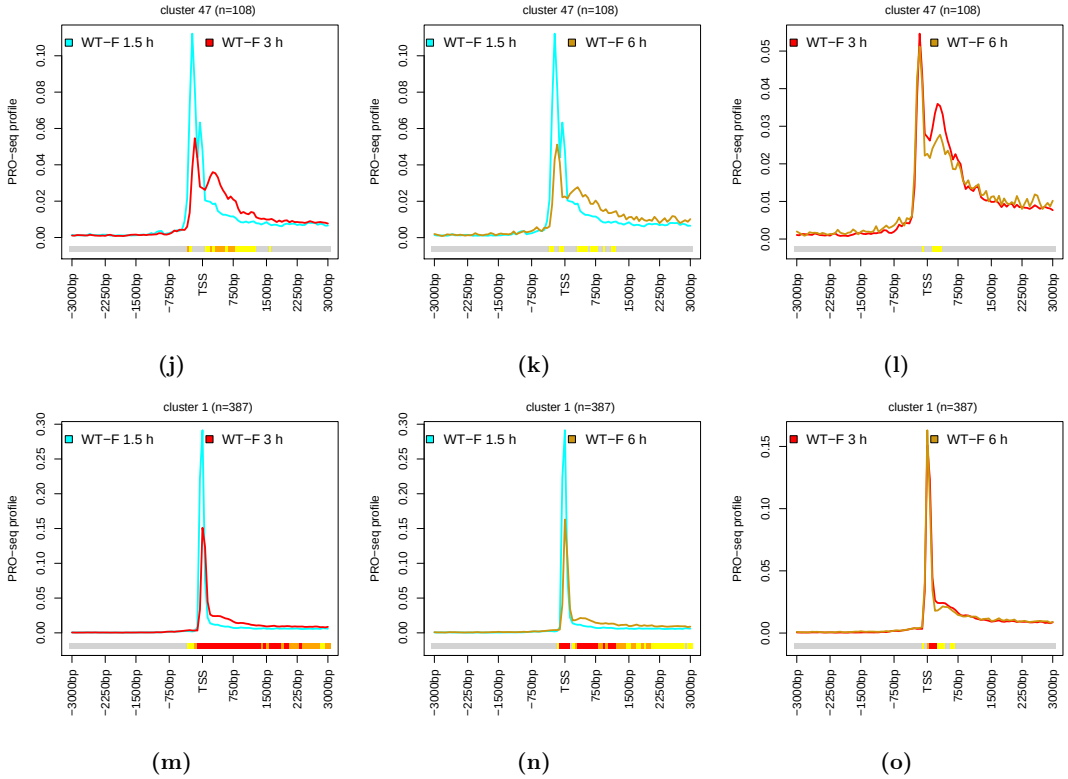
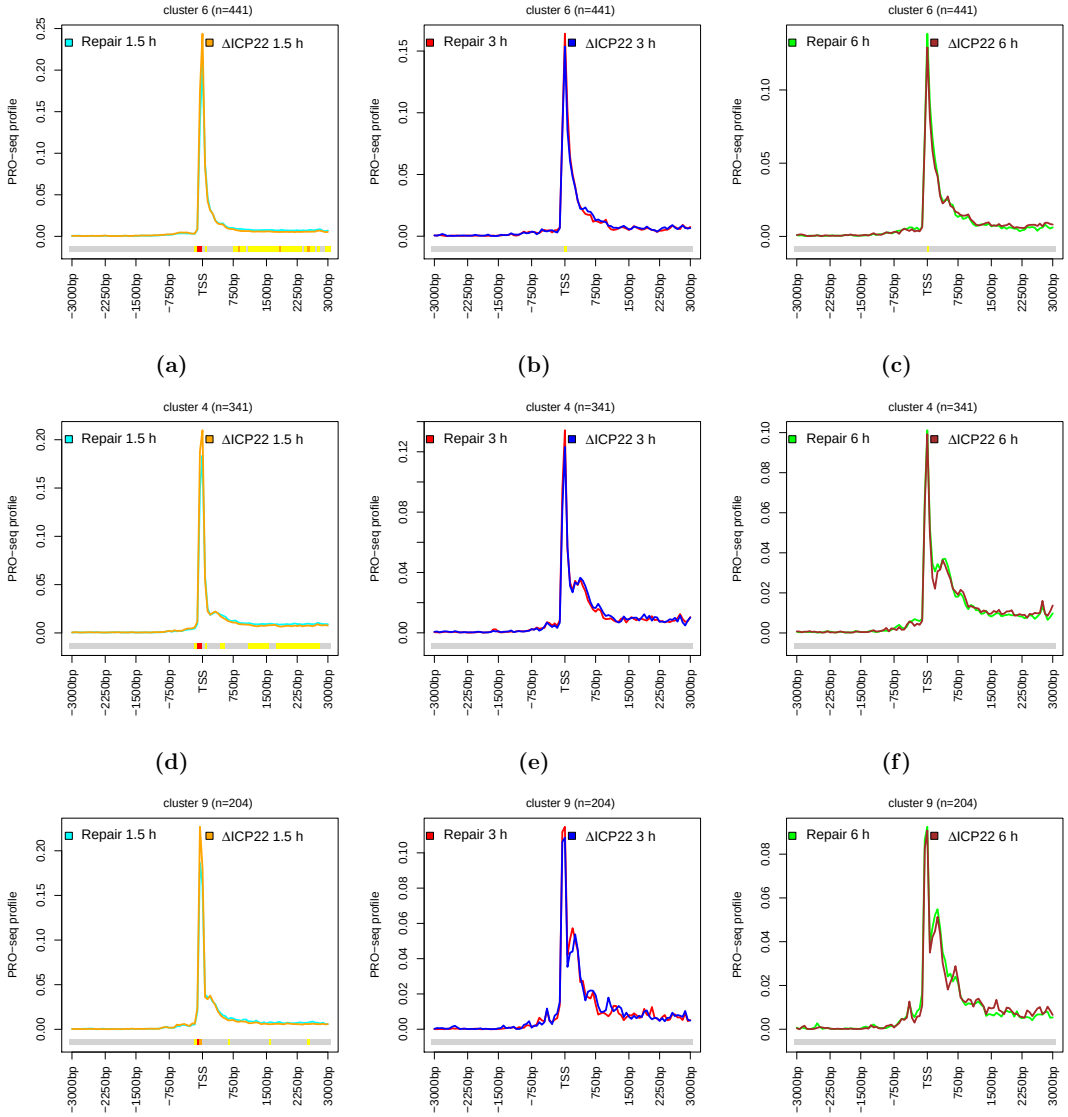


Fig. S6 Metagene plots showing the PRO-seq profile in sense direction from -3 kb to +3 kb around the TSS for the pairwise comparisons of WT-F 1.5, 3 and 6 h p.i. infection for example clusters. Cluster numbers and number of genes in each cluster are indicated on top of subfigures. The color track at the bottom of each subfigure indicates the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between the two time-points of WT-F infection. P-values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. p-value $\leq 10^{-15}$, orange = adj. p-value $\leq 10^{-10}$, yellow = adj. p-value $\leq 10^{-3}$.

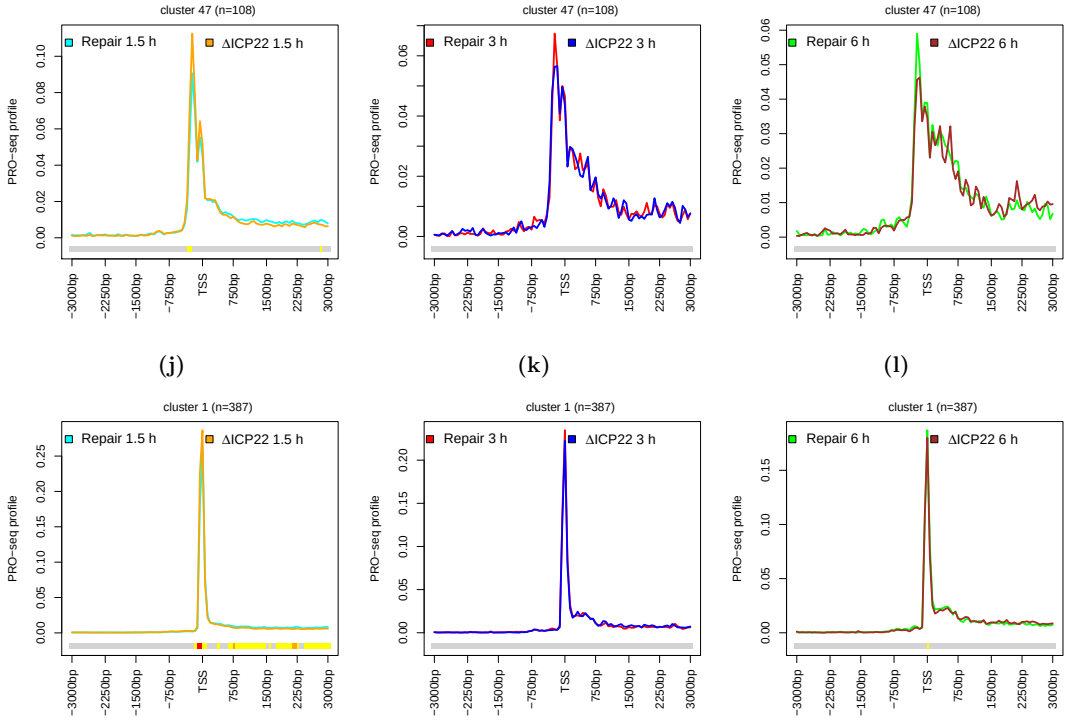


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

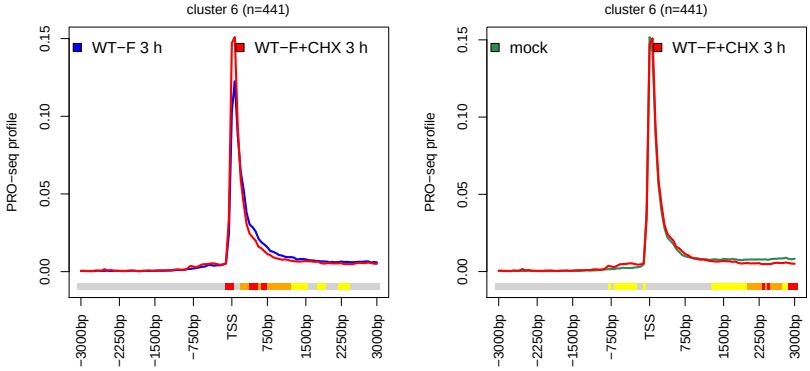
(Continued on next page)



(j) **(k)** **(l)**

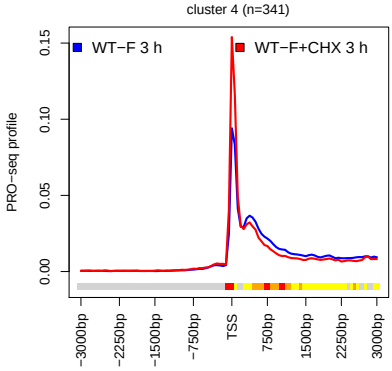
(m) **(n)** **(o)**

Fig. S7 Metagene plots showing the PRO-seq profile in sense direction from -3 kb to +3 kb around the TSS for the pairwise comparisons of ΔICP22 and repair virus infection at 1.5 (left column), 3 (middle column) and 6 h (right column) for example clusters. Cluster numbers and number of genes in each cluster are indicated on top of subfigures. The color track at the bottom of each subfigure indicates the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between the two time-points of WT-F infection. P-values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. p-value $\leq 10^{-15}$, orange = adj. p-value $\leq 10^{-10}$, yellow = adj. p-value $\leq 10^{-3}$.

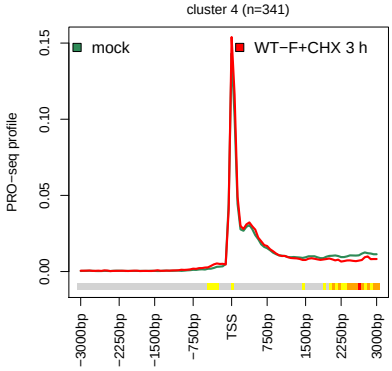


(a)

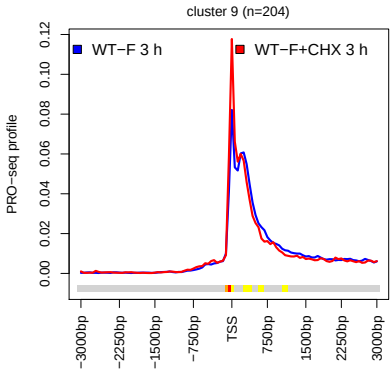
(b)



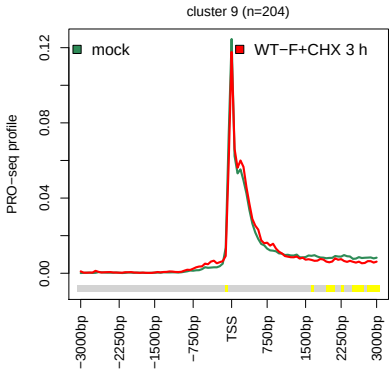
(c)



(d)



(e)



(f)

(Continued on next page)

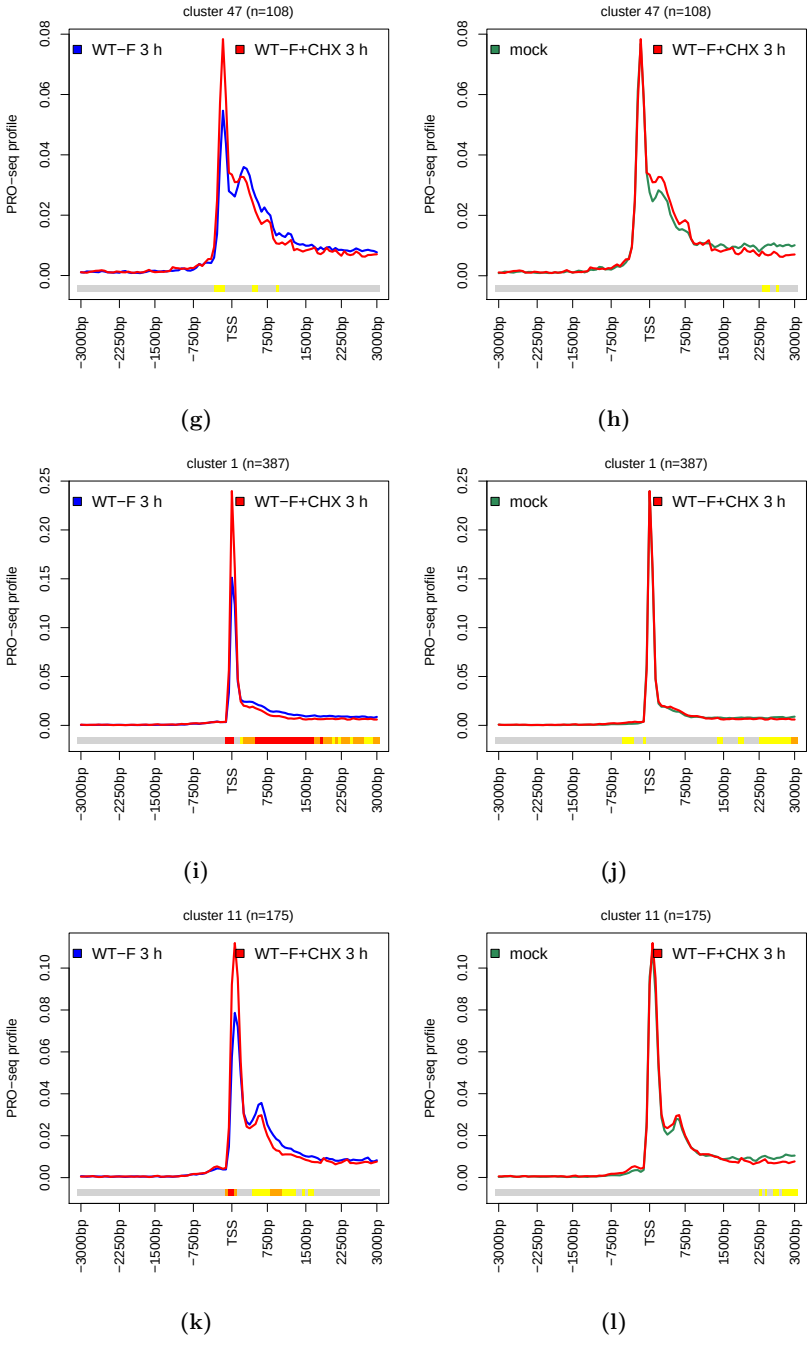
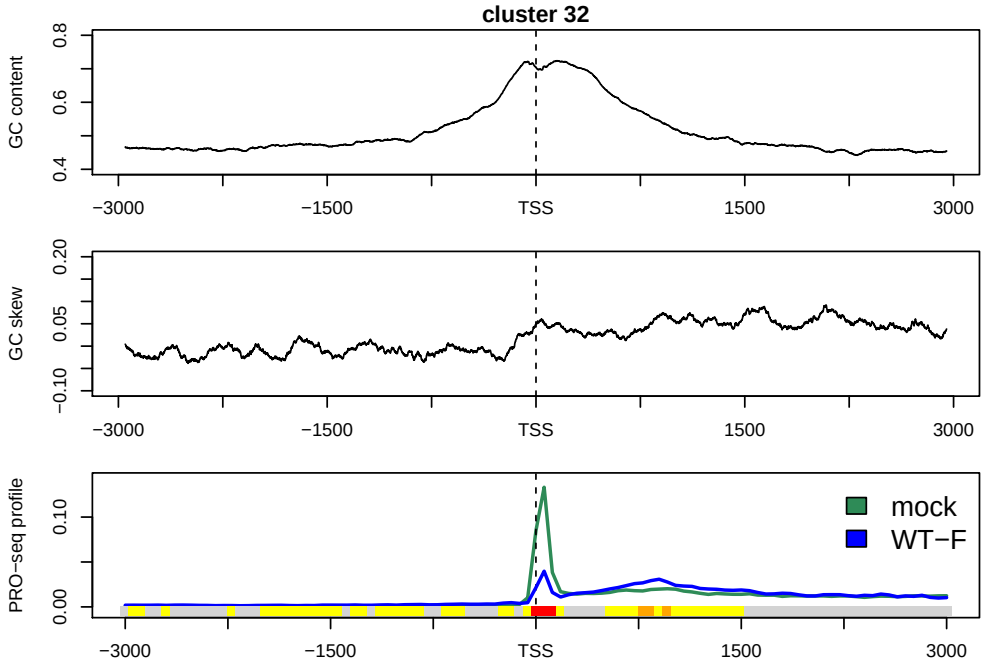
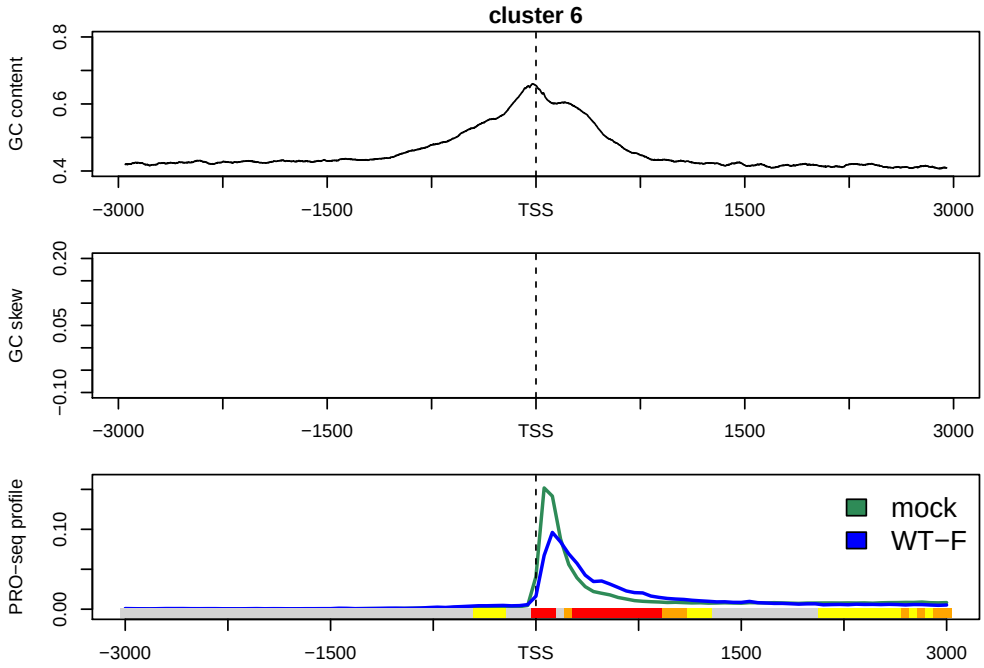


Fig. S8 Metagene plots showing the PRO-seq profile in sense direction from -3 kb to +3 kb around the TSS for WT-F 3 h p.i. ± CHX (left column) and for mock infection and WT-F 3 h p.i.+CHX (right column) for example clusters. Cluster numbers and number of genes in each cluster are indicated in subfigures. The color track at the bottom of subfigures indicates the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between the two conditions. P-values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. p-value $\leq 10^{-15}$, orange = adj. p-value $\leq 10^{-10}$, yellow = adj. p-value $\leq 10^{-3}$.

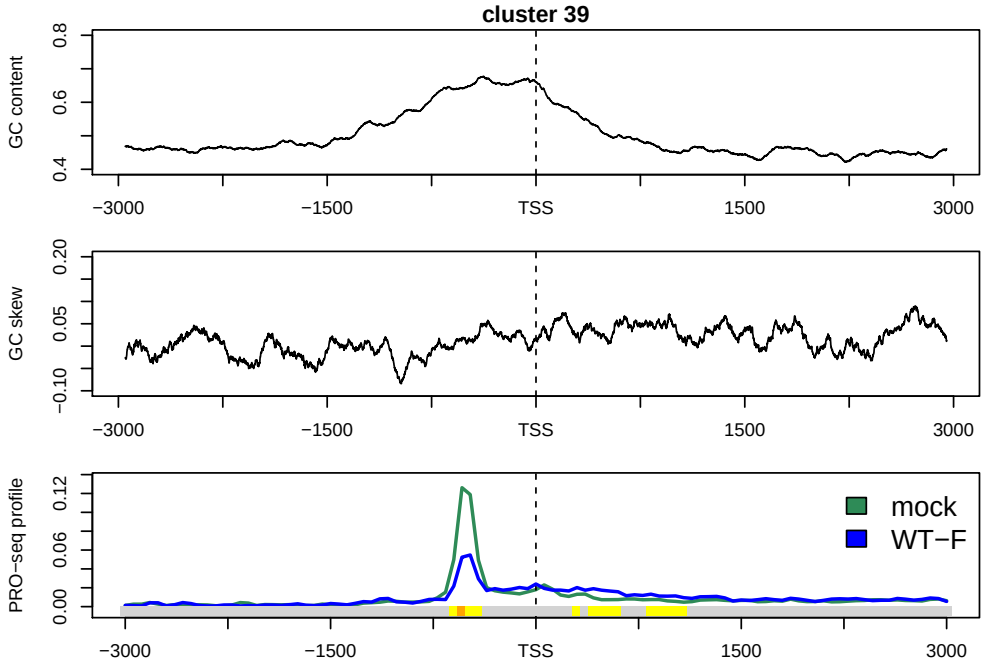


(a)

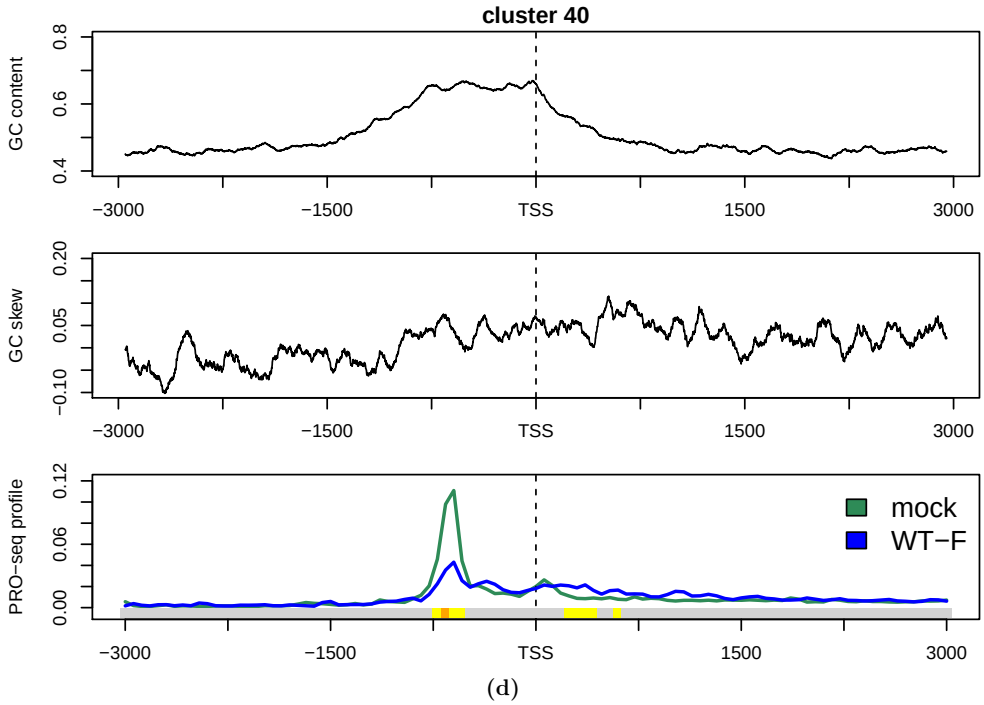


(b)

(Continued on next page)

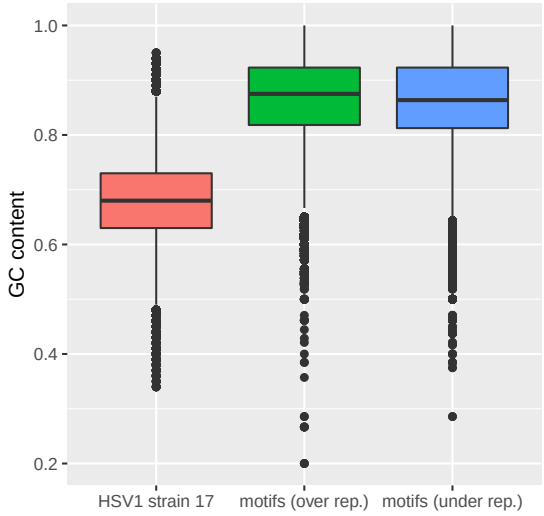


(c)



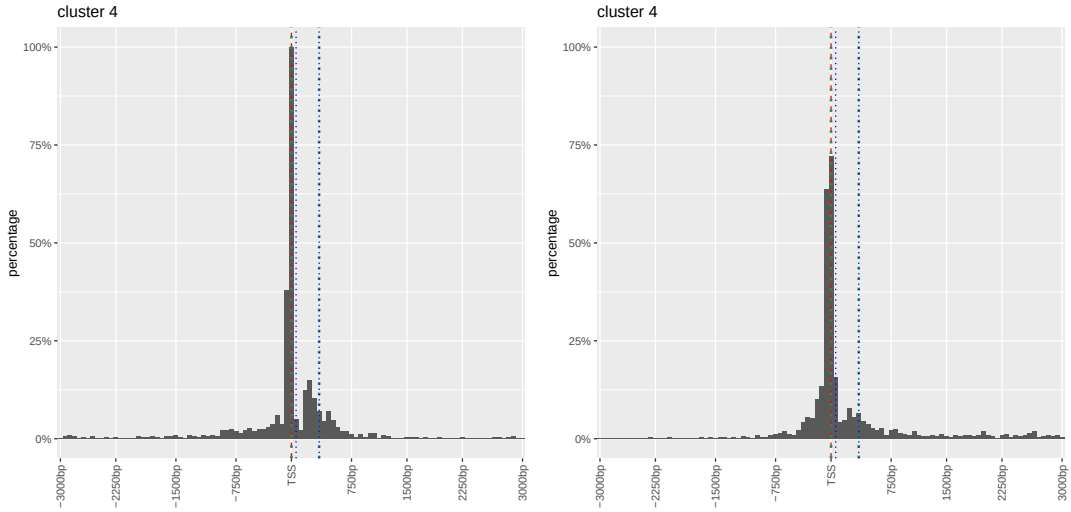
(d)

(Continued on next page)



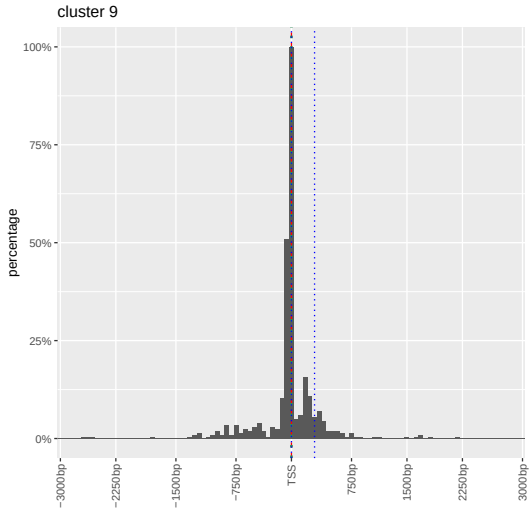
(e)

Fig. S9 (a-d) GC content and GC skew in promoter regions for example clusters (cluster numbers shown on top of subfigures). For each gene GC content and GC skew was determined in 100 bp sliding windows from -3 kb of the TSS to +3 kb of the TSS. Values for each sliding window were then averaged across genes in this cluster. The bottom panel of each subfigure shows the PRO-seq profiles in mock and WT-F 3 h p.i. infection for comparison. The color tracks at the bottom of PRO-seq panels indicate the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between the two time-points of WT-F infection. P-values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. p-value $\leq 10^{-15}$, orange = adj. p-value $\leq 10^{-10}$, yellow = adj. p-value $\leq 10^{-3}$. **(e)** Boxplots showing the distribution of the GC content determined in sliding windows of length 100 bp on the HSV-1 genome (GenBank accession JN555585.1, red) and the GC content of motif occurrences in the HSV-1 genome for transcription factor motifs found to be either over-represented (green) or under-represented (blue) in Clusters 6 and 32. Motif occurrences were determined using the fimo function of the MEME suite.

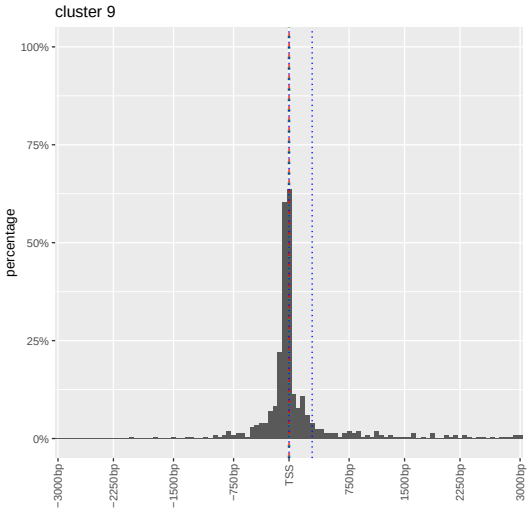


(a)

(b)

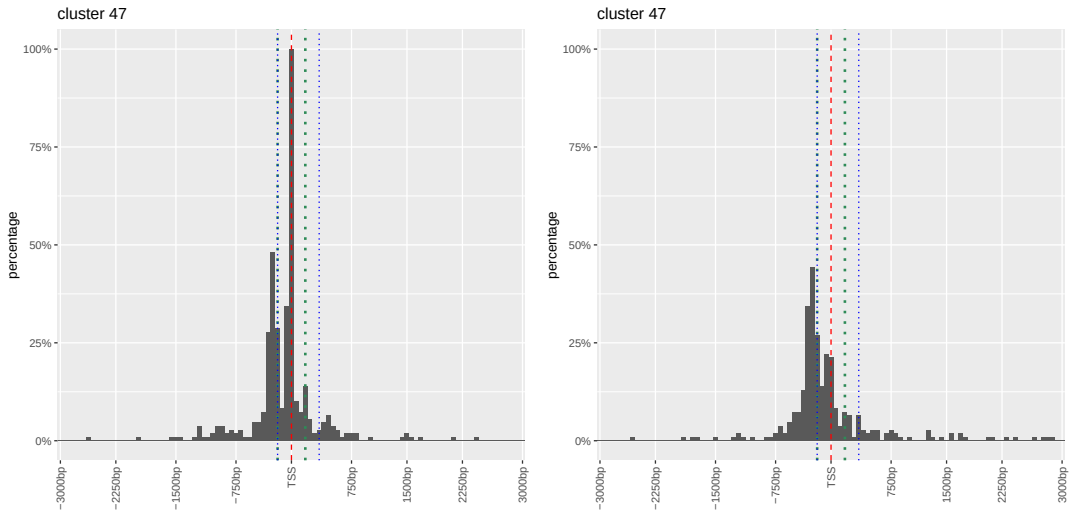


(c)



(d)

(Continued on next page)



(e) **(f)**

Fig. S10 (a,c,e) Percentage of genes exhibiting a peak in the PROcap-seq and PRO-seq data of flavopiridol-treated HFF at particular positions for example clusters (indicated on the top left of subfigures). This includes all identified peaks for a gene not just the major peak used for identifying the TSS. For this purpose, the region ± 3 kb around the identified peak was divided into bins of 60 bp and for each bin the percentage of genes with a peak falling into this bin were calculated. The red dashed vertical line marks the identified TSS. Green dotted vertical lines indicate peak positions in mock infection and blue dotted vertical lines peak positions in WT-F infection at 3 h p.i. **(b,d,f)** Percentage of genes exhibiting an annotated TSS in each 60 bp bin around the identified TSS for example clusters (indicated on the top left of subfigures). TSS and peak positions in WT-F infection at 3 h p.i. are indicated as in **(a,c,e)**

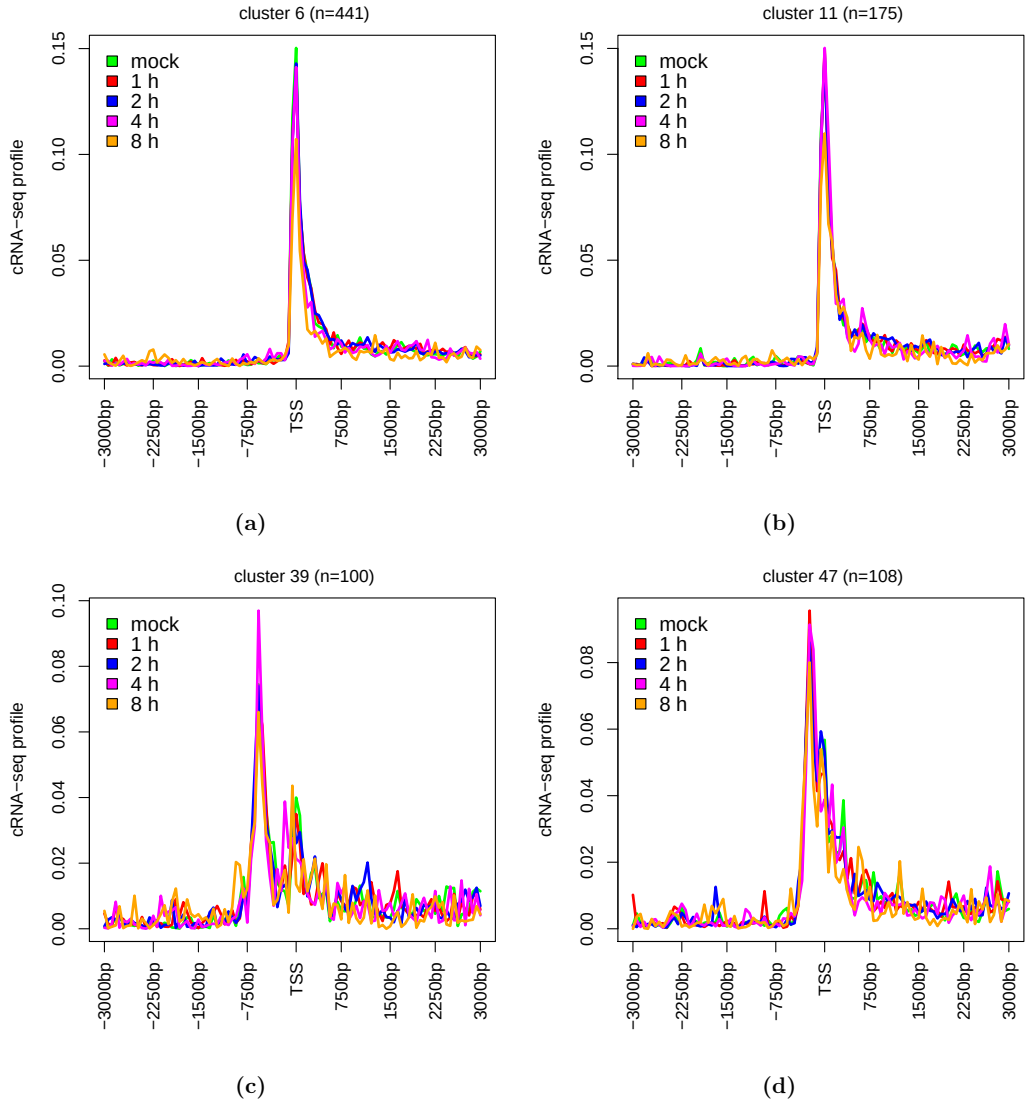


Fig. S11 Metagene plots of cRNA-seq profiles on the sense strand in mock and WT-17 infection at 1, 2, 4, and 8 h p.i. for example Clusters 6, 11, 39, and 47, which show broadening of peaks or additional peaks originating or increasing in height in PRO-seq data during WT-F infection. For metagene plots of PRO-seq profiles for these clusters see Fig. 2 and Fig. S3.

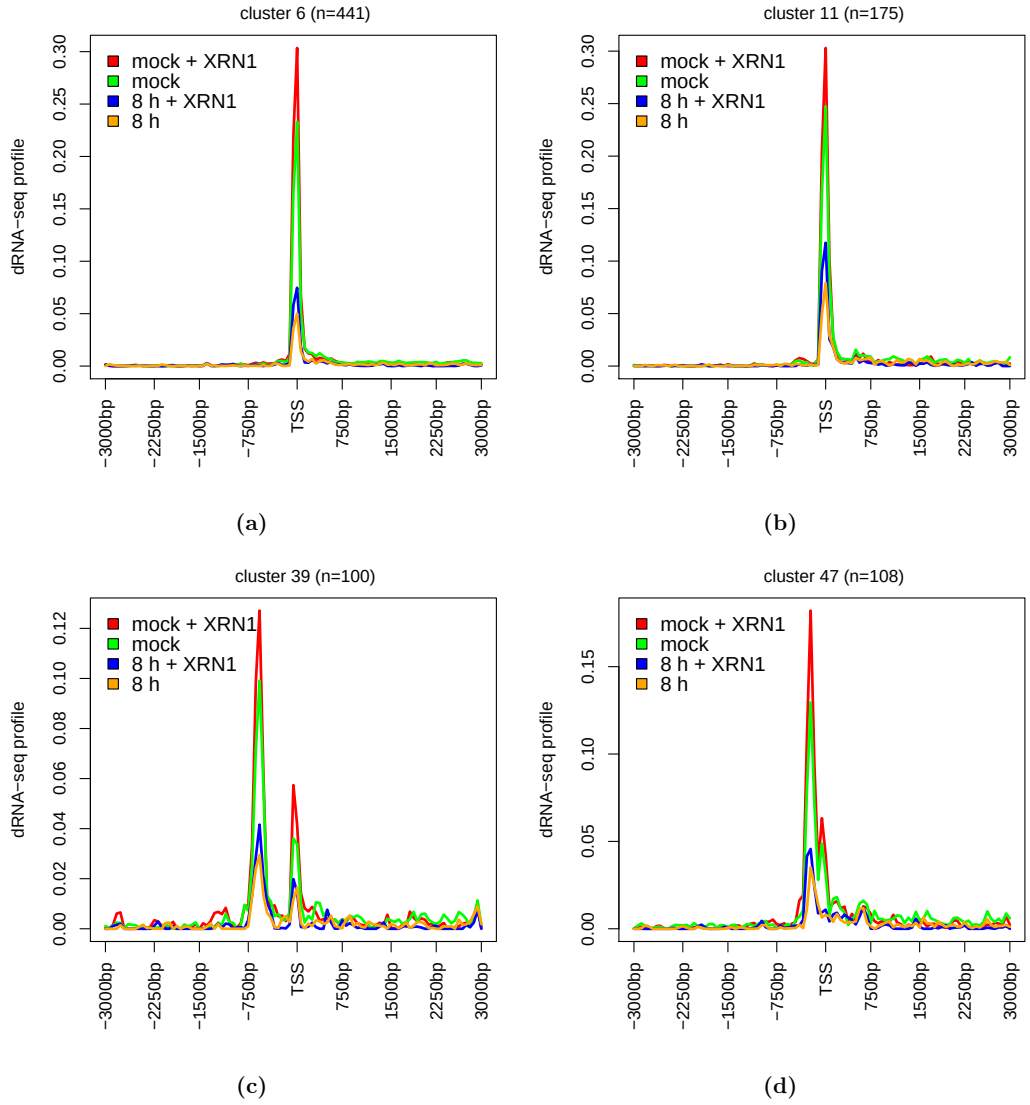


Fig. S12 Metagene plots of dRNA-seq profiles on the sense strand in mock and WT-17 8 h p.i. infection with and without XRN1 treatment for example Clusters 6, 11, 39, and 47, which show broadening of peaks or additional peaks originating or increasing in height in PRO-seq data during WT-F infection. For metagene plots of PRO-seq profiles for these clusters see Fig. 2 and Fig. S3.

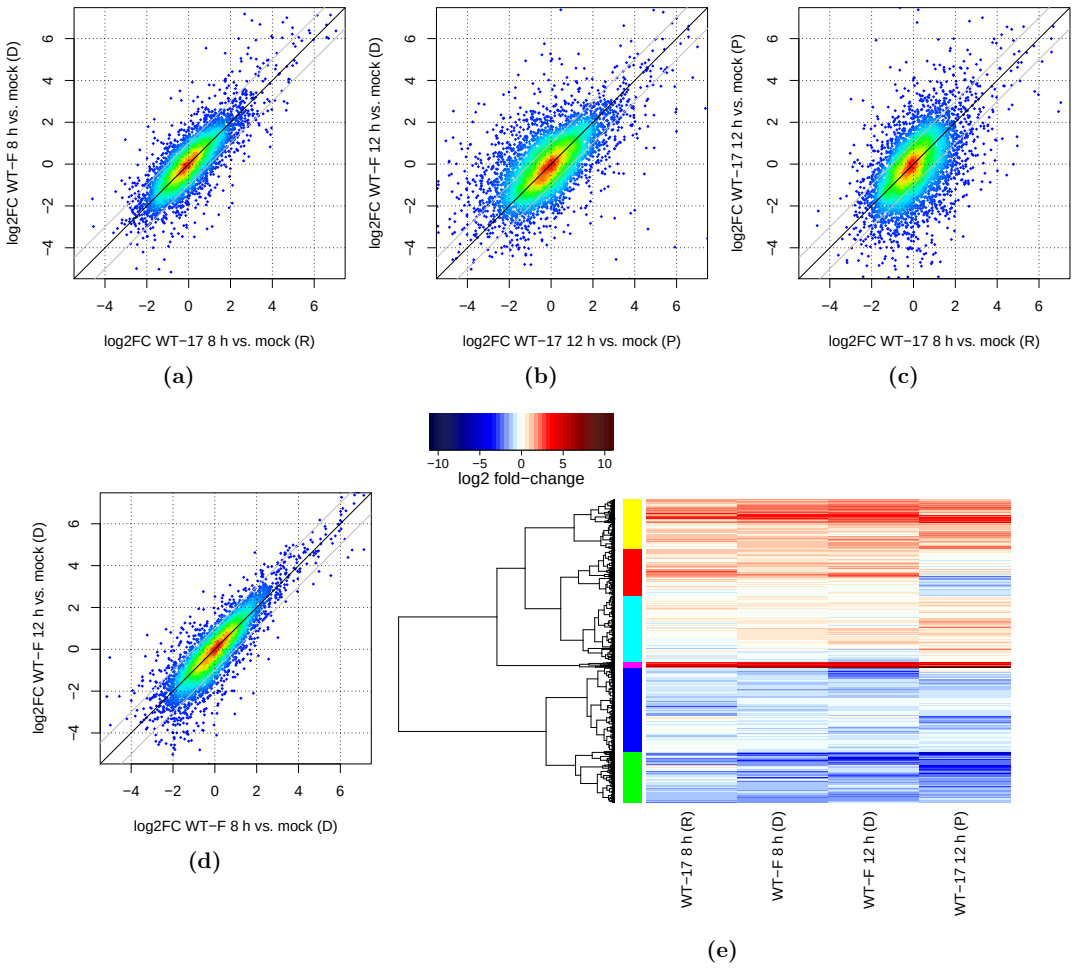
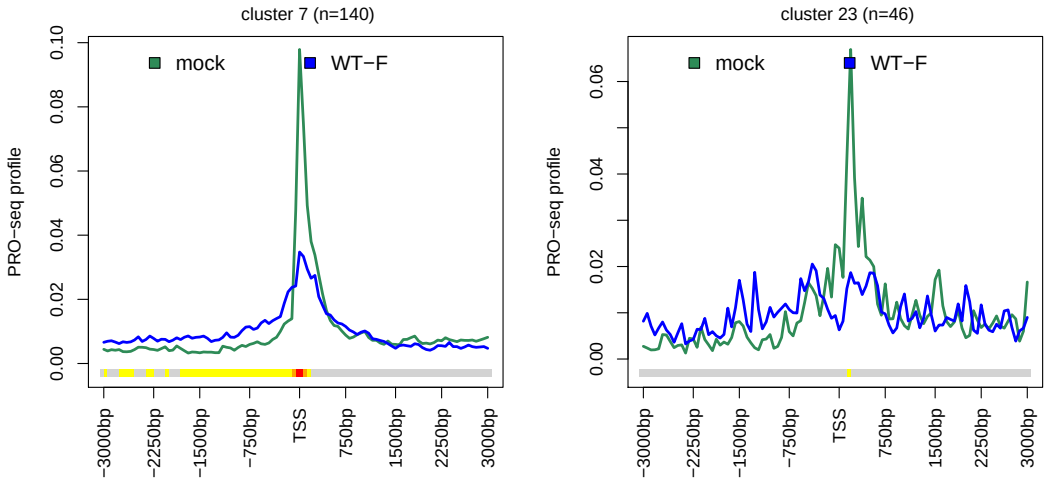
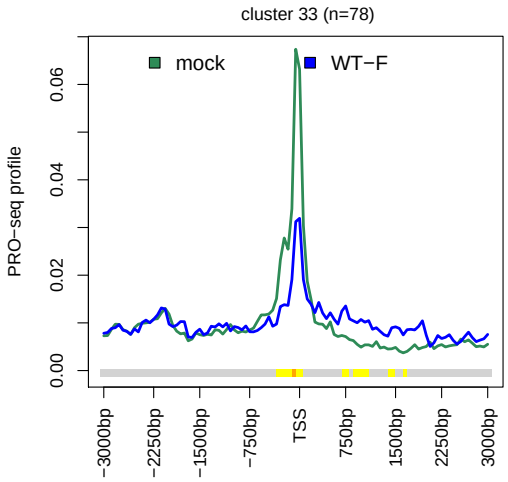


Fig. S13 (a-d) Scatter plots comparing log2 fold-changes (log2FC) in gene expression for analyzed genes between mock and WT-F or WT-17 infection at 8 or 12 h p.i. from the studies of Rutkowski *et al.* (WT-17 8 h p.i., R), Djakovic *et al.* (WT-F 8 and 12 h p.i., D) and Pheasant *et al.* (WT-17 12 h p.i., P). Colors indicate density of points from low (blue) to high (red). Black lines indicate the diagonal and gray lines a fold-change of 2. (e) Heatmap showing log2 fold-changes for WT-F or WT-17 infection at 8 or 12 h p.i. compared to mock for all genes differentially expressed (multiple testing adjusted p-value < 0.01) in at least one virus strain or time-point of infection. Hierarchical clustering was performed in R using Euclidean distances and Ward's clustering criterion. Six broad clusters were identified and are marked by colored rectangles on the right.

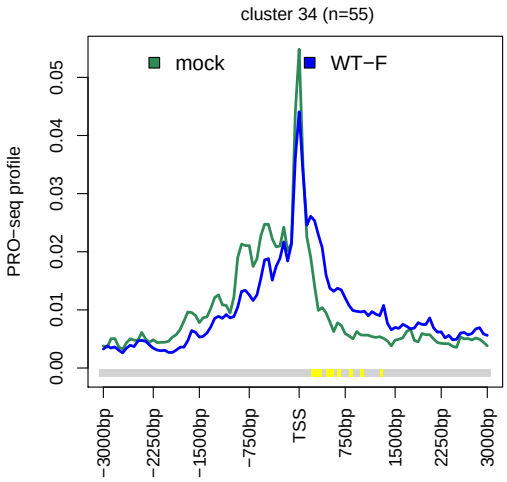


(a)

(b)



(c)



(d)

(Continued on next page)

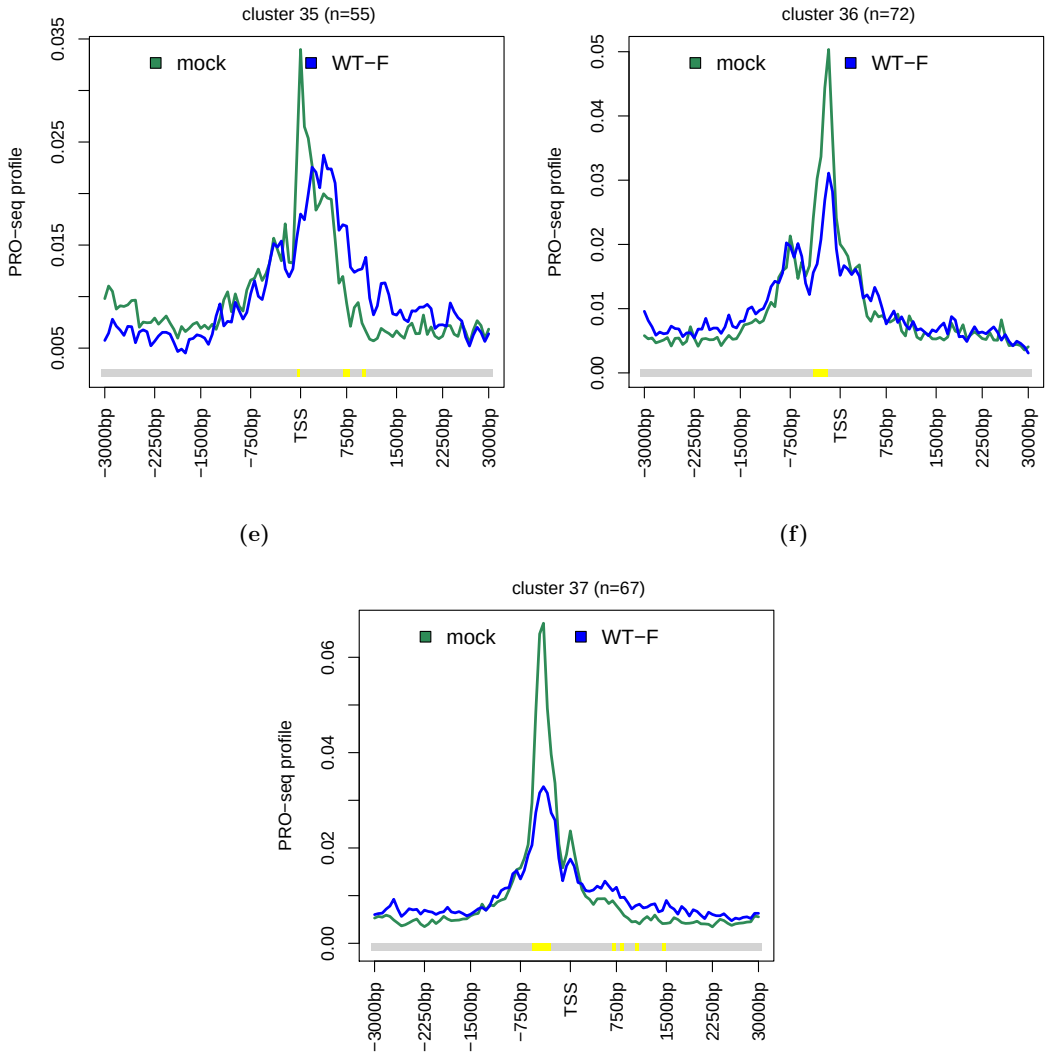
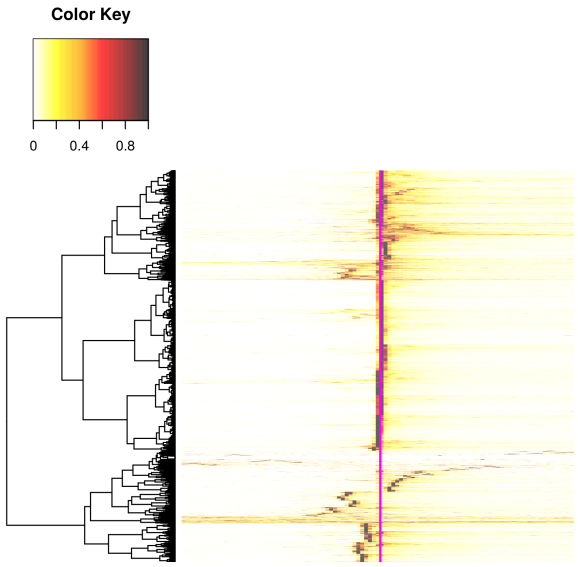
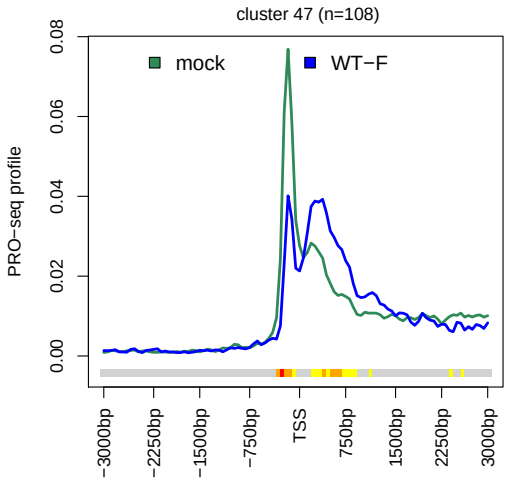


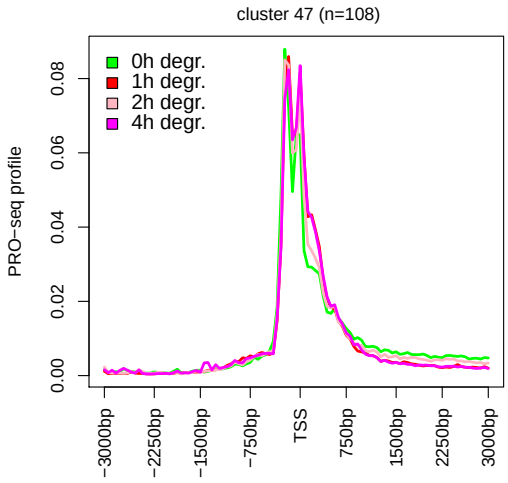
Fig. S14 Metagene plots showing the PRO-seq profile in sense direction from -3 kb to +3 kb around the TSS for mock infection (dark green) and WT-F 3 h p.i. infection (dark blue) separately for Clusters 7, 23, and 33 to 37, which exhibit a small extent of read-in transcription in 3-4 h p.i. 4sU-seq (see Fig. 5). Cluster numbers and number of genes in each cluster are indicated on top of subfigures. The color track at the bottom of each subfigure indicates the significance of paired Wilcoxon tests comparing the normalized PRO-seq coverages of genes for each bin between mock and WT-F 3 h p.i. infection. P-values are adjusted for multiple testing with the Bonferroni method within each subfigure; color code: red = adj. p-value $\leq 10^{-15}$, orange = adj. p-value $\leq 10^{-10}$, yellow = adj. p-value $\leq 10^{-3}$.



(a)

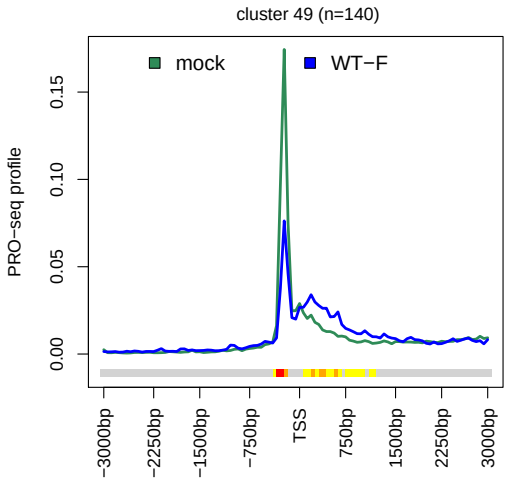


(b)

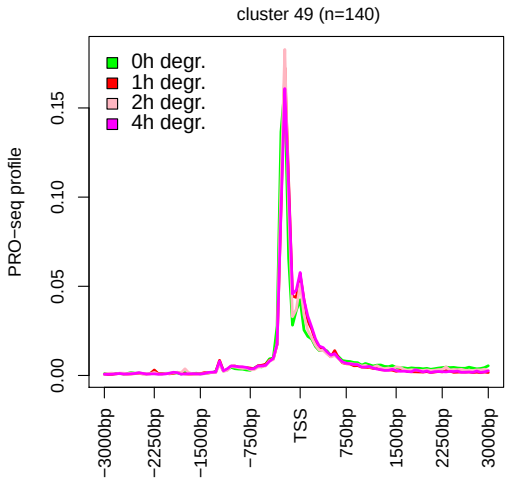


(c)

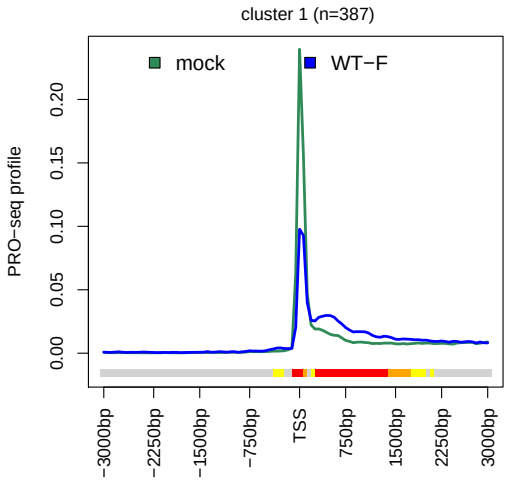
(Continued on next page)



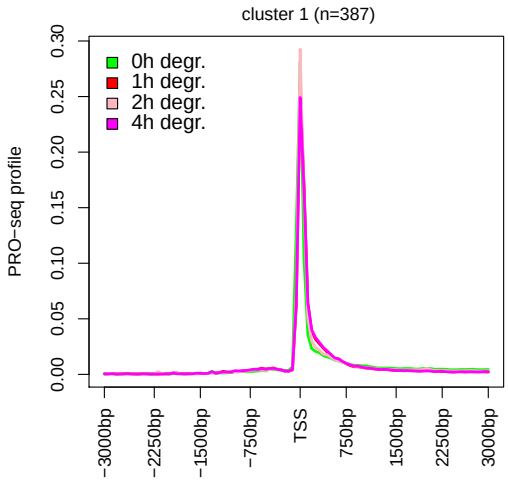
(d)



(e)



(f)



(g)

(Continued on next page)

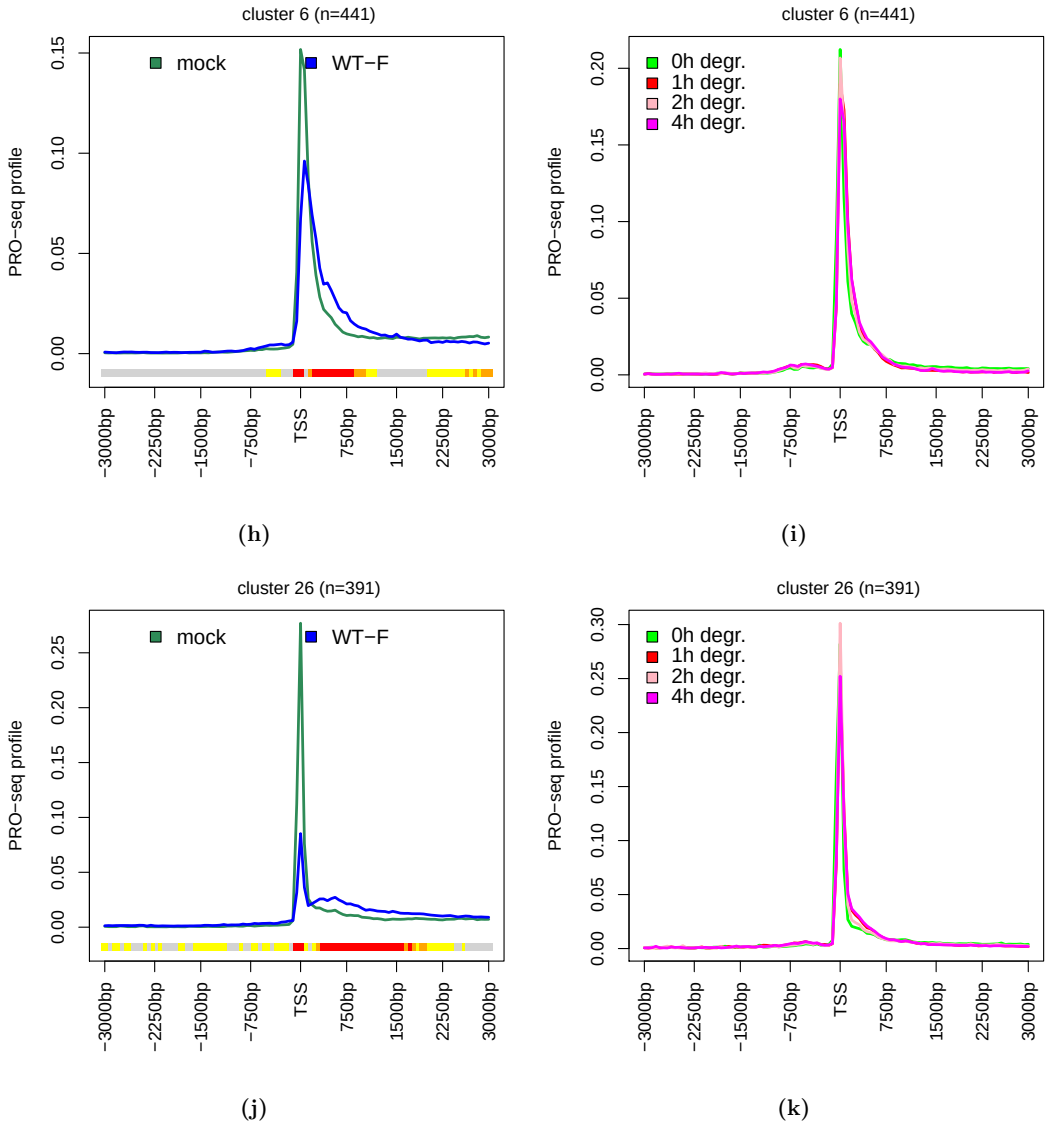


Fig. S15 (a) Heatmaps of PRO-seq profiles for 0 h auxin-inducible degradation of NELF from the study by Aoi *et al.* in a window of ±3 kb around the TSS positions identified from PROcap-seq and PRO-seq data of flavopiridol-treated HFF. For this purpose, PRO-seq profiles were divided by the maximum value in the ±3 kb promoter window, resulting in a value of 1 for the position of the highest peak in PRO-seq profiles. Hierarchical clustering of normalized PRO-seq profiles for all genes was performed using the *hclust* function in R according to Euclidean distances and Ward's clustering criterion. The central position in the promoter window (= the TSS identified in flavopiridol-treated HFF) is marked by a vertical magenta line. **(b-k)** Metagene plots around the TSS of PRO-Seq profiles for mock and WT-F 3 h p.i. infection from the study of Birkenheuer *et al.* (left column) and 0, 1, 2, and 4 h auxin-inducible degradation of NELF from the study by Aoi *et al.* (right column) for example clusters showing **(b-e)** an increased downstream peak or **(f-k)** only a reduced and slightly broadened TSS peak upon NELF degradation.

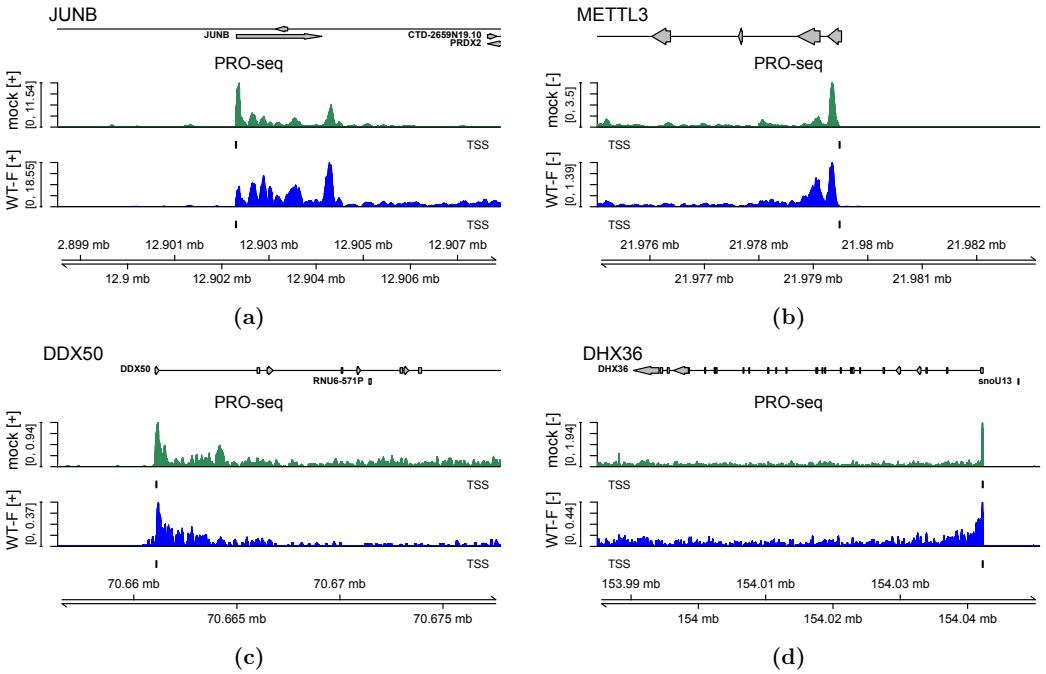


Fig. S16 Read coverage around the TSS in PRO-Seq data (sense strand only) for mock (green) and WT-F infection (blue) at 3 h p.i. for example host genes (gene name of the selected gene on the top left) mentioned in the discussion. Read coverage was normalized to total number of mapped reads and averaged between replicates. The identified TSS used in the analysis is indicated by a short vertical line below each read coverage track. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns and direction is indicated by arrowheads. Genomic coordinates are shown on the bottom. Please note that figures are not centered around the TSS, but a larger region downstream of the TSS was included than upstream of the TSS.

Bioinformatics Advances, 2023, **00**, vbad085
https://doi.org/10.1093/bioadv/vbad085
Advance access publication 4 July 2023
Original Article

OXFORD

Gene Regulation

RegCFinder: targeted discovery of genomic subregions with differential read density

Elena Weiß [1] **and Caroline C. Friedel** [1,*]

[1]Institute of Informatics, Ludwig-Maximilians-Universität München, Amalienstr. 17, Munich 80333, Germany

*To whom correspondence should be addressed.

Associate Editor: Thomas Lengauer

Abstract

Motivation: To date, no methods are available for the targeted identification of genomic subregions with differences in sequencing read distributions between two conditions. Existing approaches either only determine absolute read number changes, require predefined subdivisions of input windows or average across multiple genes.

Results: Here, we present RegCFinder, which automatically identifies subregions of input windows with differences in read density between two conditions. For this purpose, the problem is defined as an instance of the all maximum scoring subsequences problem, which can be solved in linear time. Subsequently, statistical significance and differential usage of identified subregions are determined with DEXSeq. RegCFinder allows flexible definition of input windows to target the analysis to any regions of interests, e.g. promoters, gene bodies, peak regions and more. Furthermore, any type of sequencing assay can be used as input; thus, RegCFinder lends itself to a wide range of applications. We illustrate the usefulness of RegCFinder on two applications, where we can both confirm previous results and identify interesting gene subgroups with distinctive changes in read distributions.

Availability and implementation: RegCFinder is implemented as a workflow for the workflow management system Watchdog and available at: https://github.com/watchdog-wms/watchdog-wms-workflows/

Contact: caroline.friedel@bio.ifi.lmu.de

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Functional genomics assays using high-throughput sequencing provide unparalleled opportunities for investigating cellular processes at unprecedented detail. To name just two examples, ChIP-seq and similar assays allow genome-wide mapping of DNA–protein interactions, e.g. for transcription factors and histones (Johnson *et al.*, 2007). Precision nuclear run-on analysis (PRO-seq) sequences nascent RNA 3' ends and thus allows studying active RNA Polymerase II (Pol II) transcription in a strand-specific manner (Mahat *et al.*, 2016).

To identify differences between conditions probed with functional genomics assays, numerous computational and statistical methods have been developed. One commonly used approach employs differential analysis methods on read count data, e.g. DESeq2 (Love *et al.*, 2014), to determine log2 fold-changes in read counts and statistical significance for selected genomic windows. These windows can be either user-defined, e.g. windows around the transcription start site (TSS) for investigating promoter-proximal Pol II pausing, or identified using peak calling, e.g. for differential transcription factor binding analysis. This approach is implemented in the Bioconductor package DiffBind (Ross-Innes *et al.*, 2012), which applies DESeq2 or edgeR after identifying a consensus peak set for all samples. This method considers only the total number of reads in each genomic window, but not how the reads are distributed. Thus, a change in the distribution of reads within a window, e.g. due to changes in Pol II pausing or occupancy of DNA binding proteins, without (significant) changes in the total number of reads would not be identified as differential.

To identify read distribution changes between conditions for particular types of genomic windows (e.g. promoters, gene bodies), metagene plots are commonly used. These show the average read distribution profile for sets of genomic regions. While they allow identifying general trends, individual genomic windows can deviate substantially from the general trend and changes affecting only a minority of genomic windows are often missed. Furthermore, observed changes can only be correlated to other properties of individual genomic windows, such as, e.g. gene length or sequence composition, by subdividing windows into subgroups based on these other properties and then performing metagene analyses separately for subgroups. We previously used this approach to show that CDK12 inhibition triggers a Pol II processivity defect preferentially for long, poly(A)-signal-rich genes (Chirackal Manavalan *et al.*, 2019). More recently, we used clustering of read distribution profiles for promoter windows in PRO-seq data for Herpes Simplex virus 1 (HSV-1) infection to identify subsets of genes with different Pol II pausing changes during HSV-1 infection (Weiß *et al.*, 2023). Metagene analyses of clusters showed that HSV-1 infection induced a downstream shift of Pol II pausing for the majority of host genes. While these metagene analyses provided novel insights in the general impact of CDK12 inhibition and

Received: June 14, 2023. **Editorial Decision:** June 26, 2023. **Accepted:** July 3, 2023

HSV-1 infection, respectively, they did not admit more detailed analyses at single-gene level.

For analysis of Pol II pausing at single-gene level, pausing indices (PIs) have been previously established as a standard metric. The PI of a gene is calculated as the ratio of normalized read counts in a window around the TSS (= promoter window) divided by normalized read counts on the gene body excluding the promoter. PIs are easy to calculate and are altered by changes in the distribution of Pol II occupancy around the TSS and on the gene body. However, while a reduction of PIs is commonly considered evidence of a loss or reduction of Pol II pausing, it can also originate from shifts or an extension of the pausing region downstream of the TSS, as observed, e.g. for HSV-1 infection. Thus, PI changes can be easily misinterpreted and results depend on how wide promoter windows are defined.

De novo identification of regions with differential use between conditions can be performed with diffReps (Shen *et al.*, 2013), which was developed for identifying differential chromatin modification sites from ChIP-seq data. It first employs a sliding window approach along the complete genome to identify windows with a significant difference in read counts between two conditions and then merges overlapping windows. While diffReps does not require predefined regions of interest as input, it also does not allow targeting the differential analysis to specific genomic windows of interests. Such genomic windows could not only be peak regions but also promoters, gene bodies, enhancer regions or other types of genomic windows depending on the biological question. While these windows are covered by the sliding windows, the position of the sliding windows relative to the start and end of the regions of interest varies. This complicates the comparison between different regions of interest. Furthermore, all sliding windows are included in the differential analysis, increasing the number of statistical tests performed (one per sliding window). Accordingly, more stringent multiple testing correction is required, which reduces the sensitivity of the approach.

Detection of differentially covered genomic windows is also commonly performed when identifying copy number variants (CNVs) based on read depth (RD). However, these approaches generally assume sudden shifts in RD between consecutive genomic windows with different copy numbers (Magi *et al.*, 2017), which is often not observed for differential regions in functional genomics assays.

Here, we present RegCFinder, a novel method for identifying subregions (= **Reg**ions of **C**hange) of input windows with differences in read distributions between two conditions. RegCFinder can be applied to any type of functional genomics assay and any user-defined genomic windows, such as peak regions, promoters, genes, enhancers, etc. RegCFinder proceeds in two steps: First, regions of change are identified for each input window by reducing this problem to the problem of finding all non-overlapping maximal scoring subsequences in a sequence of real numbers. Second, fold-changes in the relative use of these regions and statistical significance of fold-changes are calculated using DEXSeq (Anders *et al.*, 2012). DEXseq was originally developed to determine differential exon usage from exon counts in multiple RNA-seq replicates for different conditions. This identifies exons of a gene whose relative use compared with the other exons of the same gene increases or decreases. By redefining "genes" as input windows and "exons" as identified regions of change with "filler" regions in-between, we can assess whether there is a statistically significant difference in the use of these regions across replicates of two conditions. We evaluate the usefulness of RegCFinder on the tasks of identifying changes in Pol II pausing upon HSV-1 infection and Pol II processivity upon CDK12 inhibition at single-gene level and compare it against diffReps and XCAVATOR, an RD-based approach for CNV detection. This confirmed both general results from our previous studies, but identified subsets of genes with different characteristics not evident from metagene analyses.

2 Methods

2.1 General idea

The input to RegCFinder is a set of user-defined genomic windows W (in BED format) and aligned read data (in BAM format) for two conditions c_1 and c_2 with two or more replicates each (samples $s_{11},\dots,s_{1k}$ and $s_{21},\dots,s_{2k}$ with $k \geq 2$ the number of replicates). RegCFinder then first calculates density functions of the read distribution for each condition and each window in the following way: First, the number of reads mapping to each position $i \in w$ for each window $w \in W$ are determined for each sample. This results in values $r^w_{11}(i),\dots,r^w_{1k}(i), r^w_{21}(i),\dots,r^w_{2k}(i)\ \forall i \in w\ \forall w \in W$. Second, read densities $d^w_{11},\dots,d^w_{1k}, d^w_{21},\dots,d^w_{2k}$ are calculated for all $w \in W$ and for each condition $j \in \{1,2\}$ as

$$d^w_{js}(i) = \frac{r^w_{js}(i)}{\sum_{i \in w} r^w_{js}(i)} \text{ for } s \in [1:k], i \in w. \tag{1}$$

Finally, the average density across replicates for each condition is calculated $\forall w \in W$ as

$$d^w_j(i) = \frac{1}{k}\sum_{s=1}^{k} d^w_{js}(i) \text{ for } i \in w. \tag{2}$$

In the ideal case, densities would be continuous functions as exemplified in Figure 1a but due to noise in the sequencing data, densities are more likely to look as shown in Figure 1b.

The key idea behind RegCFinder is to identify subregions of input windows in which the density is higher in one condition than in the other. In the example in Figure 1a, this would be straightforward as we could simply determine the borders of regions as those positions $i \in w$ for window w, where the two density functions cross, i.e. where $d^w_1(i) = d^w_2(i)$ (or alternatively where $d^w_1(i) - d^w_2(i) = 0$, shown as dashed lines in Fig. 1a). However, in the realistic scenario exemplified in Figure 1b, density functions can cross several times splitting the window into many small regions. To prevent this, we want to tolerate some crossovers between densities within a region, provided that the density for one condition tends to be higher than the density for the other condition for most of the region. In terms of the differences between conditions, i.e. $d^w_{12} := d^w_1 - d^w_2$ (example in Fig. 1c) or $d^w_{21} := d^w_2 - d^w_1$ (example in Fig. 1d), this means that we aim to identify regions in which either d^w_{12} or d^w_{21} contains mostly positive values with few negative values in-between (colored rectangles in Fig. 1c and d).

This reduces our problem to a well-known computational problem, i.e. the problem of finding all maximum scoring subsequences (AMSS) in a sequence of real numbers, which can be solved in linear time (Ruzzo and Tompa, 1999). In the following sections, we describe the AMSS problem and how we use the solution to this problem in RegCFinder.

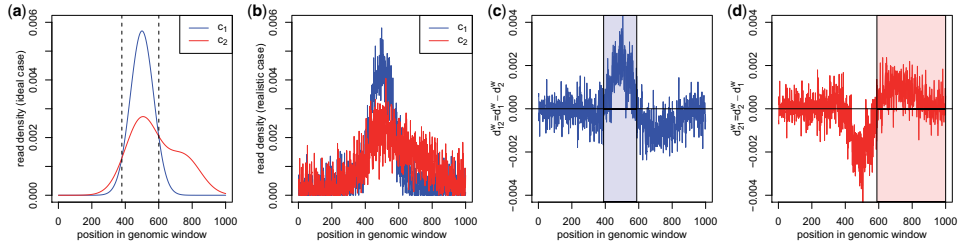


Figure 1. Example illustrating the RegCFinder approach. (**a**) RegCFinder aims to identify regions with differences in read distributions between two conditions. In the ideal case shown here, this can easily be done by identifying the intersection points of the corresponding density functions. (**b**) Sequencing noise, however, leads to noisy read densities (d_1^w and d_2^w for condition 1 and 2, respectively) with multiple intersection points. Differences in density functions are calculated from the densities in (b) resulting in sequences $d_{12}^w := d_1^w - d_2^w$ (**c**) and $d_{21}^w := d_2^w - d_1^w$ (**d**). Maximal scoring subsequences (MSS) are calculated for these sequences and filtered based on randomization (see also Fig. 2). The shaded rectangles mark the final MSS

2.2 All maximum scoring subsequences

The input to the AMSS problem is a sequence $X = (x_1, \ldots, x_n)$ of real numbers. The score $S_{i,j}$ of a subsequence $(x_i, \ldots, x_j)$ of X is defined as

$$S_{i,j} = \sum_{l=i}^{j} x_l. \tag{3}$$

The score of an empty subsequence is 0.

A maximum scoring subsequence (MSS) of X is then defined as a subsequence $m = (x_i, \ldots, x_j)$ of X with the following properties:

1) All proper subsequences $m' = (x_k, \ldots, x_l)$ of m have a lower score than m, i.e. $S_{k,l} < S_{i,j}$.
2) No proper supersequence of m fulfills property 1.

As the empty sequence is also a subsequence of any subsequence of X, $S_{i,j} > 0$ for any MSS of X. Furthermore, the MSS of X are disjoint and every $x_i \in X$ with $x_i > 0$, i.e. every positive element of X, is contained in an MSS of X. The AMSS problem is simply the problem of finding **all** MSS of X. A linear-time algorithm for solving this problem was developed by Ruzzo and Tompa (1999).

2.3 Calculation of MSS in RegCFinder

RegCFinder implements the algorithm by Ruzzo and Tompa to separately calculate all MSS for the sequences $d_{12}^w = (d_{12}^w(1), \ldots, d_{12}^w(|w|))$ and $d_{21}^w = (d_{21}^w(1), \ldots, d_{22}^w(|w|))$ for each window w with two small modifications. First, we multiply each element in the sequence by 100 to obtain larger values and, second, we subtract a small value from each element in the sequence. This serves to prevent long stretches of zeroes in a sequence, which would lead to artificially long MSS if the sequence contains positive elements at the end of these long stretches of zeroes. For instance, for $X = (1, 0, 0, 0, 0, 0, 0, 0, 0, 1)$, the only MSS is the complete sequence.

The final input sequences for window w, for which MSS are calculated in RegCFinder, are thus defined as

$$X_{st}^w(i) = 100 \times d_{st}^w(i) - \frac{\rho}{|w|} \text{ for } 1 \le i \le |w|, s, t \in \{1,2\}, s \ne t. \tag{4}$$

Here, ρ is a pseudocount, which is set to 1 by default. Since $\sum_{i=1}^{|w|} d_s^w(i) = 1$ for $s \in \{1,2\}$, we have for $s, t \in \{1,2\}, s \ne t$ that

$$\begin{aligned} \sum_{i=1}^{|w|} X_{st}^w(i) &= 100 \times \sum_{i=1}^{|w|} d_{st}^w(i) - \sum_{i=1}^{|w|} \frac{\rho}{|w|} \\ &= 100 \times \left(\sum_{i=1}^{|w|} d_s^w(i) - \sum_{i=1}^{|w|} d_t^w(i) \right) - \rho = -\rho. \end{aligned} \tag{5}$$

As a consequence, long MSS are penalized with a linear penality function $p(\lambda) = \frac{\rho}{|w|} \times \lambda$, where λ is the length of the MSS. Thus, higher values of ρ lead to shorter MSS.

2.4 Filtering and merging of MSS

In the following, let M_{12} be the set of MSS determined for X_{12}^w and M_{21} be the set of MSS for X_{21}^w. As noted above, any positive element of the input sequence X is contained in an MSS. Thus, even without long stretches of mostly positive values, MSS will be identified. In the worst case, every positive element will be an MSS of length 1, e.g. as for the sequence $X = (\underline{1}, -2, \underline{1}, -2, \underline{1}, -2, \underline{1})$ (MSS underlined). Furthermore, even random sequences can contain long MSS if they contain a sufficient number of large positive elements larger than absolute values of most negative values (see Fig. 2a and b).

To identify and remove such MSS that are no better than random, we repeatedly randomly permute each of the input sequences X by sampling $|X|$ times from X without replacement (default = 1000 randomizations). We then identify MSS for the randomized sequences and remove all MSS for the original sequence with a score less than or equal to the maximum MSS score identified for any of the randomized sequences. In case of the example shown in Figure 2c (calculated from the example in Fig. 1d), all identified MSS are removed with a score below the gray dotted line in Figure 2d. Thus, for the example in Figure 1 only one region would be identified in which the density for condition 2 is higher than the density of condition 1 (red transparent rectangle in Fig. 1d). In this way, RegCFinder also filters regions with only small differences in the distributions (e.g. the region left of the left dashed vertical line in Fig. 1a).

After filtering MSS, M_{12} and M_{21} are merged to obtain the final regions of change. In case of overlapping MSS $m_{12} \in M_{12}$ and $m_{21} \in M_{21}$, only the MSS with the highest score is retained. Regions with higher density in condition 1 (c_1) than in condition 2 (c_2) are denoted as $c_1 > c_2$ regions and regions

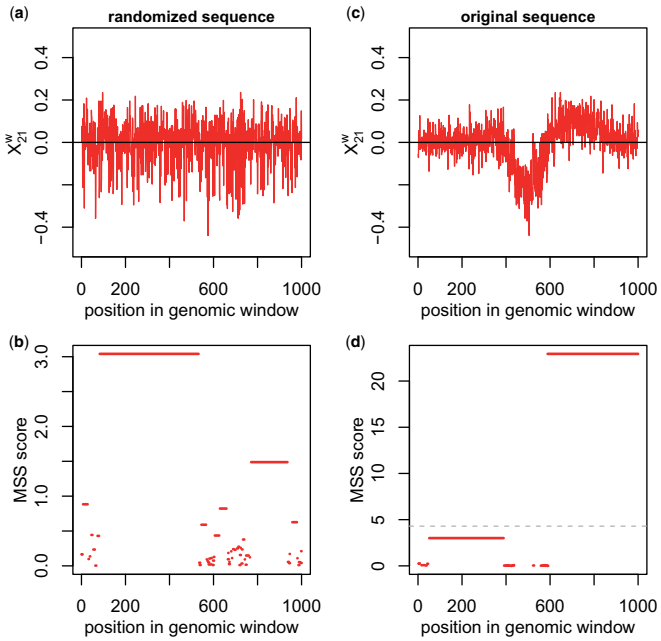


Figure 2. (**a**) Randomized and (**c**) original input sequence obtained from the example in Figure 1b and d. (**b**, **d**) MSS identified on the randomized (b) and original (d) sequence. Each MSS is shown as a horizontal line covering the positions in the genomic window shown on the *x*-axis. The *y*-axis position of the MSS indicates its score. The dashed line in (d) shows the maximum MSS score across 10 randomizations. All MSS with a score less or equal to this score are discarded, resulting in the identification of the one region of change shown in Figure 1d

with higher density in condition 2 than in condition 1 are denoted as $c_2 > c_1$ regions.

2.5 Significance assessment using DEXSeq

To assess statistical significance of differential and log2 fold-changes in relative use of the identified regions given the number of reads in each replicate, RegCFinder uses DEXSeq (Anders *et al.*, 2012). For this purpose, a new annotation file is generated for DEXSeq with "genes" defined as the input windows. The "exons" consist of the identified regions of change and "filler" regions representing the genomic regions within the window between and around regions of change. For the example in Figure 1, five exons would be defined: two regions of change (blue and red rectangle in Figure 1c and d, respectively) and three filler regions (i) on the left of the first region, (ii) on the right of the second region and (iii) between the two regions. Read counts for exons are determined using featureCounts (Liao *et al.*, 2014) on input BAM files. DEXSeq results are included in the final output file (in TSV format), which contains coordinates of the identified regions, their MSS score and their log2 fold-change and multiple testing adjusted *P*-value from DEXSeq.

3 Results

3.1 Input data

We evaluated RegCFinder on two datasets, which we previously investigated using metagene analyses to identify differences in Pol II pausing (Weiß *et al.*, 2023) and Pol II processivity (Chirackal Manavalan *et al.*, 2019), respectively. The first dataset was obtained with PRO-seq for mock infection and 3 h post wild-type (WT) HSV-1 infection (three replicates each) by Birkenheuer *et al.* (2018). Here, mock infection means that cells were exposed to the same medium as the HSV-1-infected cells that lacked virus. We previously used clustering of read distributions around the TSS and metagene analyses of clusters to show that WT HSV-1 infection leads to a downstream shift in Pol II pausing for most host genes (Weiß *et al.*, 2023). Unfortunately, the metagene analysis did not allow further investigating this effect at single-gene level. PI analysis on this dataset showed widespread reductions in PIs for most genes, which can be misinterpreted as a loss in Pol II pausing and increased elongation.

The second dataset consisted of ChIP-seq data for Pol II and Ser2 phosphorylations (P-Ser2) of the Pol II carboxy-terminal domain (CTD) in a cell line expressing an analog-sensitive version of the CDK12 kinase (Chirackal Manavalan *et al.*, 2019). This analog-sensitive CDK12 is inhibited by the ATP analog 3-MB-PP1. Three replicates were obtained each with either DMSO (Ctl) or 3-MB-PP1 (Inhi) treatment for 4.5 h. We previously reported that CDK12 inhibition induces a Pol II processivity defect characterized by a loss of ChIP-seq read coverage toward 3' ends of predominantly long, poly(A)-signal-rich genes and a shift of the terminal P-Ser2 peak into the gene body (Chirackal Manavalan *et al.*, 2019) (see Supplementary Fig. S7a for an example). Matching RNA-seq

of nuclear RNA showed that this was associated with premature transcription termination at poly(A) signals within gene bodies.

Data preprocessing is described in the Supplementary Material.

3.2 Changes in Pol II pausing during HSV-1 infection

We first applied RegCFinder to the PRO-seq samples of mock and WT HSV-1 infection for promoter windows defined as ±3 kb around the TSS of 7650 genes (with the default of 1000 randomizations). These TSS were previously identified from PROcap-seq and PRO-seq data of flavopiridol-treated uninfected human foreskin fibroblasts (Weiß *et al.*, 2023) (see Supplementary Material). PROcap-seq is a variation of PRO-seq that specifically maps Pol II initiation sites. Flavopiridol inhibits CDK9, which is required for the switch to active elongation, and arrests Pol II in a paused state at the promoter. This allows also identifying the TSS for genes that are not or weakly paused in untreated cells.

RegCFinder identified a total of 7621 regions of change for the input windows. For 7201 regions, a *P*-value was calculated by DEXSeq. The remaining 420 regions were excluded by DEXSeq in the independent filtering step, which excludes regions with low read counts. For 6958 regions (96.7% of regions with *P*-values), a significant change was observed (multiple testing adjusted *P*-value $\leq$0.01). For comparison, a test with only 10 randomizations instead of the default 1000 randomizations identified 10 528 regions of change for which DEXSeq calculated *P*-values, but only 88% of these were statistically significant (adj. $P \leq 0.01$). Thus, by increasing the number of randomizations, significance of results can be improved at the cost of reduced sensitivity.

Of the 7201 regions of change with *P*-values, 3414 had a higher density in mock than WT (mock>WT regions) and 3787 had a higher density in WT (WT>mock regions). Fold-changes determined by DEXSeq were consistent with the RegCFinder predictions as mock>WT regions had positive log2 fold-changes in the comparison of mock versus WT and WT>mock regions had negative log2 fold-changes (Supplementary Fig. S1). In contrast, the median log2 fold-change of the 10 072 filler regions was close to zero and only 1697 filler regions (17%) showed a statistically significant change. Notably, with 10 randomizations, only 8% of the filler regions were statistically significant. Since more randomizations lead to more stringent filtering, some correct regions of change identified with 10 randomizations are thus filtered with 1000 randomizations and instead included as filler regions.

Figure 3a visualizes the location of the identified regions within the input promoter windows for the 4128 windows containing at least one region with a *P*-value from DEXSeq. Here, each row represents one window and each column a position in the window (from −3 kb upstream of the TSS to +3 kb downstream of the TSS, red = mock>WT regions, blue = WT>mock regions, white = filler regions or region without DEXSeq *P*-value). Windows were clustered according to Euclidean distances and Ward's clustering criterion. A cut-off on the clustering dendogram was chosen manually to obtain the 11 clusters marked in Figure 3a. Supplementary Figure S2 shows log2 fold-changes in mock versus WT determined with DEXseq for regions with adj. $P \leq 0.01$ for windows ordered as in Figure 3a. The latter confirms the high consistency between log2 fold-changes determined by DEXseq and the type of region determined by RegCFinder.

Figure 3a reveals both interesting subgroups as well as general trends. With some exceptions (clusters 1–4), WT>mock regions extended downstream of the TSS, while mock>WT regions were located around or upstream of the TSS. However, there were strong differences with regard to how far downstream of the TSS the WT>mock regions extended. For clusters 9, 10 and parts of cluster 11, the WT>mock regions ended well before the 3' end of the promoter windows. This is consistent with a downstream shift of Pol II pausing and confirmed by inspection of read distributions for

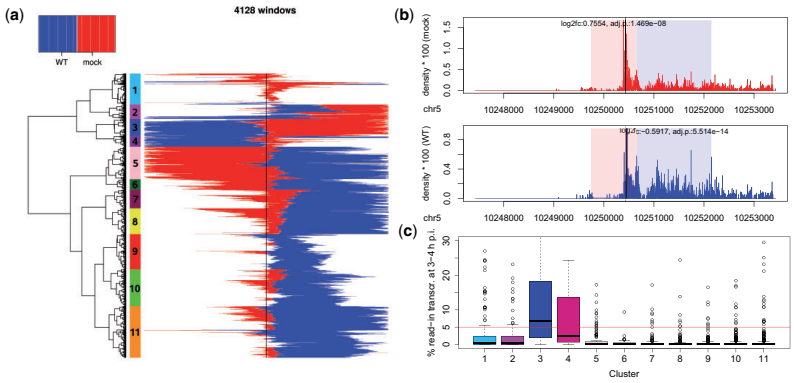


Figure 3. (**a**) Heatmap showing the location and type of identified regions of change (red = mock>WT, blue = WT>mock) for 4128 windows with at least one region with *P*-values determined by DEXSeq. For details, see the main text. The TSS is indicated by a black vertical line. (**b**) Read density in mock (red, top panel) and WT (blue, bottom panel) infection for an example from cluster 10. Windows are shown in 5'–3' direction, i.e. regions up- and downstream of the TSS are to the left and right, respectively, of the TSS (black vertical line). Red shaded rectangle = mock>WT region, blue shaded rectangle = WT>mock regions. Log2 fold-changes and adj. *P*-values in mock versus WT infection are shown in the top panel for mock>WT regions and in the bottom panel for WT>mock regions. (**c**) Boxplots showing the distribution of % read-in transcription (for definition, see the text) at 3–4 h p.i. for the 11 clusters marked in (a). The red horizontal line indicates the cutoff (5%) we previously used for identifying genes with read-in transcription

individual genes (Fig. 3b) and metagene analyses (Supplementary Fig. S3a). For other clusters (5–8, partly 11), WT>mock regions extended to or close to the 3' end of promoter windows suggesting either a further downstream shift of pause sites or increased elongation due to a loss of pausing. While the metagene analyses suggest the former (Supplementary Fig. S3b and c), inspection of individual genes identifies both examples for increased elongation (Supplementary Fig. S4a) and increased (relative) use of downstream pause sites (Supplementary Fig. S4b). This reflects the limits of metagene analyses due to averaging across genes with potentially diverse patterns.

Interestingly, clusters 3 and 4 had long WT>mock regions upstream of the TSS. We previously showed using 4sU-seq that HSV-1 infection disrupts transcription termination, leading to extensive read-through transcription beyond poly(A) sites that can extend for tens-of-thousands of nucleotides into intergenic regions and into downstream genes (Rutkowski *et al.*, 2015). 4sU-seq is based on labeling newly transcribed RNA with 4-thiouridine (4sU) in specific time intervals (here: 1 h intervals during infection) followed by sequencing of labeled RNA. Read-through transcription extending into a downstream gene is denoted as "read-in transcription" and calculated as expression in the 5-kb upstream of the gene start divided by gene expression (Hennig *et al.*, 2018). [Formal definition: % read-in transcription = 100 × fragments per million mapped reads (FPKM) in the 5 kb upstream of the gene start divided by the gene FPKM. Values for uninfected cells are subtracted from values for infected cells and negative values are set to 0]. Analysis of read-in transcription previously determined using 4sU-seq for mock and 3–4 h p.i. HSV-1 infection (Rutkowski *et al.*, 2015) showed significant read-in transcription for clusters 3 and 4 (Fig. 3c, metagene plots in Supplementary Fig. S3d and e). Thus, WT>mock regions identified by RegCFinder upstream of the TSS reflect read-in transcription for these genes.

3.3 Impact of CDK12 inhibition on Pol II processivity

As a second analysis, we applied RegCFinder to ChIP-seq data for Pol II and P-Ser2 with DMSO (Ctl) or 3-MB-PP1 (Inhi) treatment for 4.5 h. Here, we used windows covering complete genes from −3 kb upstream of the TSS to +3 kb downstream of the transcription termination site (TTS). We included only genes with a distance ≥5 kb to the next up- and downstream gene (=8086 gene windows).

More regions of change were identified for P-Ser2 (10 218 with *P*-values calculated by DEXseq) than for Pol II (7065) and a larger fraction of P-Ser2 regions were significant (90%) than of Pol II regions (83%). This can be explained by the fact that Pol II ChIP-seq reads are often concentrated in the promoter region due to Pol II pausing, resulting in lower coverage on gene bodies (Yu *et al.*, 2015). In contrast, P-Ser2 reads more evenly cover the gene body with a less prominent peak shortly downstream of the TTS (see e.g. Supplementary Fig. S7a). Again, log2 fold-changes determined by DEXSeq were consistent with the direction of change identified by RegCFinder (Supplementary Fig. S5).

For 3405 and 4639 genes at least one statistically significant region of change was identified for Pol II and P-Ser2, respectively. Here, 74% and 86%, respectively, of these genes contained both a significant Inhi>Ctl and Ctl>Inhi region. The heatmap in Figure 4 visualizes the identified regions of change in Pol II and P-Ser2 for the 3534 genes for which at least one region of change was identified in both Pol II and P-Ser2 similar to Figure 3. Here, the positions of the identified regions relative to the 5' (left) and 3' end of the window (right) are shown. Regions were clustered as described above and 15 clusters were obtained from the clustering dendogram (marked in Fig. 4). Results were generally highly consistent between Pol II and P-Ser2 ChIP-seq, with regions of changes of the same type identified at similar positions.

Clusters 1–5 exhibited the pattern we expected from our previous metagene analysis, which showed a loss of Pol II from gene 3' ends and a shift of 3' end P-Ser2 peaks into the gene body upon CDK12 inhibition (Chirackal Manavalan *et al.*, 2019). Accordingly, Inhi>Ctl regions (red), i.e. regions with a relative increase upon inhibitor treatment, were located closer to the TSS and Ctl>Inhi regions (blue) were located closer to the TTS. An example gene for cluster 1 is shown in Supplementary Figure S7a. In this case, the Inhi>Ctl region found for P-Ser2 approximately matched the terminal P-Ser2 peak shifted upstream upon CDK12 inhibition.

Previously, we found that longer genes were more strongly affected by CDK12 inhibition and loss of Pol II at gene 3' ends resulted in a reduction in nuclear RNA levels for corresponding genes (Chirackal Manavalan *et al.*, 2019). As illustrated in the example gene in Supplementary Figure S7a, this is not due to down-regulation of the complete gene, but rather due to premature transcription termination leading to a loss of reads in the 3' end region of the gene. Consistent with this, clusters 4 and 5, for which Inhi>Ctl regions ended relatively close to the gene start, indicating a strong shift of Pol II from the gene 3' end toward the gene 5' end, contained longest genes (Supplementary Fig. S6). Moreover, clusters 1–5 all showed a strong reduction in nuclear RNA levels upon CDK12 inhibition and stronger reduction was observed for clusters with Inhi>Ctl regions ending closer to the TSS (Fig. 5).

For cluster 6, Inhi>Ctl regions in Pol II and P-Ser2 cover a large fraction of the gene and are followed by only short or no Ctl>Inhi regions at the 3' end. Thus, the Pol II processivity defect is only noticeable close to the gene 3' end. Consistently, these genes were short and little reduction in nuclear RNA was observed. In contrast, cluster 8 showed similar patterns as observed for the example in Supplementary Figure S7a, i.e. an Inhi>Ctl region between two Ctl>Inhi regions for P-Ser2. This central Inhi>Ctl region reflects the shift of the terminal P-Ser2 peak into the gene body. Cluster 13 represents the one example with strong divergence between the Pol II and P-Ser2 results. Manual inspection of example genes showed generally low read coverage on gene bodies, in particular for Pol II, explaining the low consistency between Pol II and P-Ser2 results.

The remaining clusters (clusters 7, 9–11, 14 and 15) showed Inhi>Ctl regions downstream of relatively long Ctl>Inhi regions in either P-Ser2 alone or in both Pol II and P-Ser2. The Inhi>Ctl regions were not followed by additional downstream Ctl>Inhi regions or only very short ones. Notably, genes in these clusters did not show reduced expression upon CDK12 inhibition, with some even showing an increase in expression. An example gene from cluster 15 is shown in Supplementary Figure S7b. Here, no premature transcription termination is observed and this gene is even weakly up-regulated in nuclear RNA (log2 fold-change 0.36,

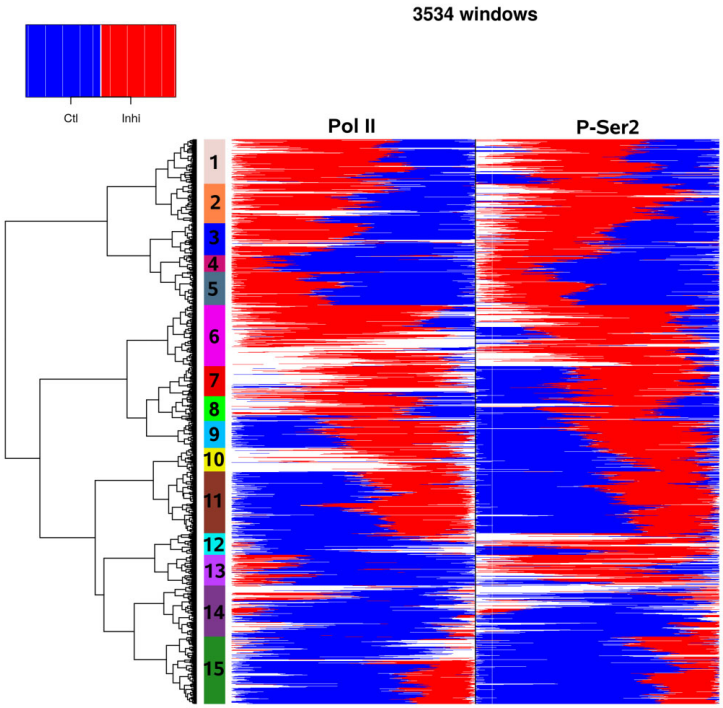


Figure 4. Heatmap showing the location of identified regions of change in Pol II and P-Ser2 ChIP-seq data for control (Ctl) and CDK12 inhibitor (Inhi) treatment. Location is shown relative to the window start and end. Inhi>Ctl regions are indicated in red and Ctl>Inhi regions are indicated in blue. Filler regions and regions without a *P*-value calculated by DEXSeq are shown in white. For more details, see the main text

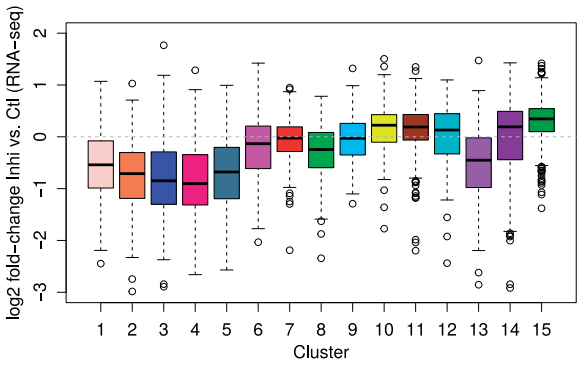


Figure 5. Boxplots showing the distribution of log2 fold-changes in nuclear RNA between CDK12 inhibitor and control treatment for the clusters from Figure 4

adj. $P = 6.69 \times 10^{-6}$). The Inhi>Ctl region identified by RegCFinder is located upstream of the terminal P-Ser2 peak. While the summit position of the P-Ser2 peak is unchanged, read distributions upstream of this peak clearly differ between control and CDK12 inhibition.

In summary, RegCFinder identified interesting subsets of genes that are not fully explained by our existing model of the effects of CK12 inhibition on Pol II processivity developed based on metagene analyses. One such subset is cluster 2, which shows a strong shift of the terminal P-Ser2 peak into

the gene body that cannot be explained by gene length alone. Moreover, 46% of genes shown in Figure 4 (i.e. clusters 7, 9–11, 14 and 15) showed changes in Pol II occupancy upon CDK12 inhibition that do not appear to be directly linked to premature transcription termination and warrant further research.

3.4 Comparison to competing approaches

We compared RegCFinder against two alternative approaches that focus on identifying differentially covered regions in the genome: XCAVATOR, a RD-based approach for CNV detection (Magi *et al.*, 2017), and diffReps, developed for detecting differential chromatin modification sites (Shen *et al.*, 2013). Full details of the comparison can be found in the Supplementary Material. In brief, XCAVATOR did not identify any differential regions on the PRO-seq and ChIP-seq data, likely due to the absence of sudden shifts in read coverage. diffReps mostly identified regions with absolute changes in RD. Consequently, results on the PRO-seq data differed strongly. On the ChIP-seq data, diffReps identified a subset of regions identified by RegCFinder on input genes.

4 Discussion

In this article, we present RegCFinder, a new approach for determining differences between two conditions in sequencing data. In contrast to previous approaches, it focuses on identifying differences in the distribution of reads at single-gene, or rather single-window, level. Given a set of input windows defined by the user, RegCFinder identifies subregions of these input windows, the so-called regions of change, in which one condition has a higher read density than the other condition. For this purpose, the problem is defined as an instance of the AMSS problem, which can be solved efficiently in linear time (Ruzzo and Tompa, 1999).

Since this problem definition considers only the distribution of reads within the input windows but not the absolute read numbers, statistical significance and log2 fold-changes in the relative use of each identified region of change compared with the rest of this input window are calculated with DEXSeq from read counts. Since we also include "filler" regions that are not part of any identified regions of change in the input for DEXSeq, this step is not limited to identifying regions of change as statistically significant. Indeed, in the two applications shown in this article, 8–17% of filler regions showed a statistically significant change. Nevertheless, this is much lower than the 84–96% of the regions of change identified by RegCFinder that were statistically significant. Furthermore, we observed a trade-off between sensitivity and specificity of RegCFinder that can be tuned by adjusting the number of randomizations used for filtering the identified MSS. Fewer randomizations result in less stringent filtering while more randomizations lead to filtering of some of the truly differential regions, which will then be included as filler regions. Notably, for significance analysis, DEXSeq can be replaced with any other method with a similar purpose.

Since RegCFinder is both agnostic to how input windows are defined by the user and what type of sequencing data is provided as input, it lends itself to a wide range of applications. Here, we illustrated the usefulness of RegCFinder on two applications: (i) Pol II pausing changes upon WT HSV-1 infection analyzed using PRO-seq data and (ii) changes in Pol II processivity upon CDK12 inhibition analyzed using ChIP-seq of Pol II and P-Ser2. In both cases, we previously used metagene analyses on subgroups of genes either defined (i) by clustering of PRO-seq read distributions for mock and WT infection (Weiß *et al.*, 2023) or (ii) based on gene length or differential gene expression (Chirackal Manavalan *et al.*, 2019). While the metagene analyses already yielded interesting novel insights, we were frustrated by their limitations regarding the analysis of individual genes, which motivated the development of RegCFinder. Other approaches, in particular PIs, were also unsatisfactory as they could not distinguish between "normal" Pol II pausing changes with increased elongation and the downstream shifts of pause sites in WT HSV-1 infection.

The analysis of PRO-seq data of mock and WT HSV-1 infection confirmed the downstream shift of Pol II pausing to less well-defined downstream pause sites for a large fraction of genes (clusters 9, 10 and partly 11 in Fig. 3). For other genes (clusters 5–8 and partly 11), increased read density in WT infection downstream of the TSS extended until or close to the end of the promoter window. An investigation of example genes suggested that at least some of these genes may not actually exhibit delayed Pol II pausing but rather increased elongation on the whole gene body. Thus, RegCFinder now allows more detailed analyses of these genes and their characteristics compared with other genes for which pausing is retained at downstream sites. Finally, RegCFinder also identified genes with read-in transcription originating from disrupted transcription termination of an upstream gene. Thus, no prior filtering of these genes was necessary.

Similarly, the analysis of Pol II and P-Ser2 ChIP-seq upon CDK12 inhibition confirmed our previous observations for a large fraction of genes. However, we also identified a large number of genes with different patterns of changes in the Pol II and P-Ser2 distribution that open up new avenues of investigation into the role of CDK12 not evident from the metagene analyses.

RegCFinder also provides new possibilities for integrating different data types. First, location of identified regions of change for two or more types of data (e.g. different ChIP-seq antibodies, different sequencing assays) or different experiments can be easily compared with the heatmap approach shown in Figure 4. Second, the new DEXSeq annotation created from identified regions of change for one data type or experiment can be directly used to calculate differential use on a different data type or in a different experiment. For instance, one could analyze if P-Ser2 regions showed the same differential use in Pol II ChIP-seq data or ChIP-seq data for other CTD phosphorylations, elongation factors, histone modifications or in nuclear RNA-seq data.

The limitations of RegCFinder should also be noted. First, it is designed for pairwise comparisons of conditions. Thus, it cannot be used for segmentation of input windows given only one condition and comparison of three or more conditions requires comparison to a common reference (similar to other differential approaches like differential gene expression or exon usage). Second, depending on the noise level in the data, the precise border positions of regions of change may be difficult to determine using the RegCFinder approach and thus will likely vary between biological replicates or different RegCFinder parameter settings. In cases in which borders can be more accurately determined by other means, e.g. reads crossing splice junctions in case of RNA-seq, it may thus be better to use these other means or complement the

RegCFinder results with such other information. Finally, RegCFinder is targeted toward applications in which the read distribution provides information, such as functional genomics approaches based on short read sequencing, like ChIP-seq, PRO-seq, ATAC-seq and more. While in principle RegCFinder could also be applied to read densities obtained from long-read sequencing, the applications usually addressed by long-read sequencing likely are not suited for the RegCFinder approach.

In summary, RegCFinder implements a novel approach for identifying genomic regions with differences in read density between two conditions. Due to its flexibility regarding the definition of input windows and the type of input sequencing data, we believe it will be of broad use for a wide range of biological questions.

Author Contributions

Elena Weiß (Formal analysis, Investigation, Methodology, Software [lead], Visualization, Writing—original draft, Writing—review & editing [equal]) and Caroline C. Friedel (Formal analysis [equal], Funding acquisition [lead], Investigation [equal], Methodology [equal], Project administration [lead], Supervision [lead], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal])

Software and data availability

RegCFinder was implemented as a workflow for the workflow management system Watchdog (Kluge and Friedel, 2018; Kluge *et al.*, 2020) and uses Conda for automatic software deployment (https://docs.conda.io). The RegCFinder workflow is available in the Watchdog workflow repository at https://github.com/watchdog-wms/watchdog-wms-workflows together with a detailed README file on installing and running the workflow (including installation instructions for Watchdog) and in- and output file formats. All modules used in this workflow, including pre-existing modules (featureCounts, mergeFeatureCounts and DEXSeq) and modules newly developed for RegCFinder (amss, quant CurveScore, both implemented in R, preDexseq, implemented as a bash script) are available in the Watchdog module repository at https://github.com/watchdog-wms/watchdog-wms-modules. Download links and more detailed documentation for Watchdog can be found at https://www.bio.ifi.lmu.de/watchdog. The data underlying this article are available in Gene Expression Omnibus at https://www.ncbi.nlm.nih.gov/geo/ and can be accessed with accession GSE106126 (PRO-seq data for mock and HSV-1 infection) and accession GSE120072 (ChIP-seq data of CDK12 inhibition).

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the framework of the Research Unit FOR5200 DEEP-DV (443644894) project FR 2938/11-1.

Conflict of interest

None declared.

References

Anders,S. *et al.* (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.

Birkenheuer,C.H. *et al.* (2018) Herpes simplex virus 1 dramatically alters loading and positioning of RNA polymerase II on host genes early in infection. *J. Virol.*, **92**, e02184-17.

Chirackal Manavalan,A.P. *et al.* (2019) Cdk12 controls G1/S progression by regulating RNAPII processivity at core DNA replication genes. *EMBO Rep.*, **20**, e47592.

Hennig,T. *et al.* (2018) HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. *PLoS Pathog.*, **14**, e1006954.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein–DNA interactions. *Science (New York, N.Y.)*, **316**, 1497–1502.

Kluge,M. and Friedel,C.C. (2018) Watchdog—a workflow management system for the distributed analysis of large-scale experimental data. *BMC Bioinformatics*, **19**, 97.

Kluge,M. *et al.* (2020) Watchdog 2.0: new developments for reusability, reproducibility, and workflow execution. *GigaScience*, **9**, giaa068.

Liao,Y. *et al.* (2014) Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Love,M. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Magi,A. *et al.* (2017) XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC Genomics*, **18**, 747.

Mahat,D.B. *et al.* (2016) Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.*, **11**, 1455–1476.

Ross-Innes,C.S. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.

Rutkowski,A.J. *et al.* (2015) Widespread disruption of host transcription termination in HSV-1 infection. *Nat. Commun.*, **6**, 7126.

Ruzzo,W.L. and Tompa,M. (1999) A linear time algorithm for finding all maximal scoring subsequences. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, Heidelberg, pp. 234–241.

Shen,L. *et al.* (2013) diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS ONE*, **8**, e65598.

Weiß,E. *et al.* (2023) HSV-1 infection induces a downstream shift of promoter-proximal pausing for host genes. *J. Virol.*, **97**, e0038123.

Yu,M. *et al.* (2015) RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II. *Science (New York, N.Y.)*, **350**, 1383–1386.

RegCFinder: targeted discovery of genomic subregions with differential read density

Supplementary Material

Elena Weiß and Caroline C. Friedel

1 Data preprocessing

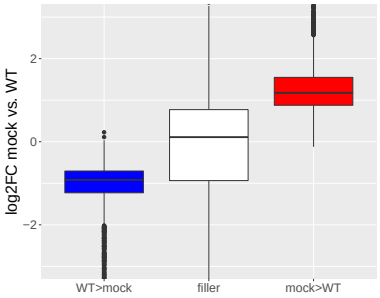
PRO-seq data were aligned against the human genome (GRCh37/hg19), human rRNA sequences, and the HSV-1 genome (GenBank accession code: JN555585) using ContextMap2 version 2.7.9 [Bonfert et al., 2015] (using BWA [Li and Durbin, 2009] as short read aligner and allowing a maximum indel size of 3 and at most 5 mismatches). For the two repeat regions in the HSV-1 genome, only one copy was retained each, excluding nucleotides 1–9,213 and 145,590–152,222 from the alignment. ChIP-seq reads were aligned to the human genome (hGRCh38/g38) using BWA [Li and Durbin, 2009]. Reads with an alignment score < 20 were discarded. SAM output files of aligners were converted to BAM files using samtools [Danecek et al., 2021].

Log2 fold-changes in nuclear RNA for CDK12 inhibition vs. control were taken from our previous publication [Chirackal Manavalan et al., 2019].

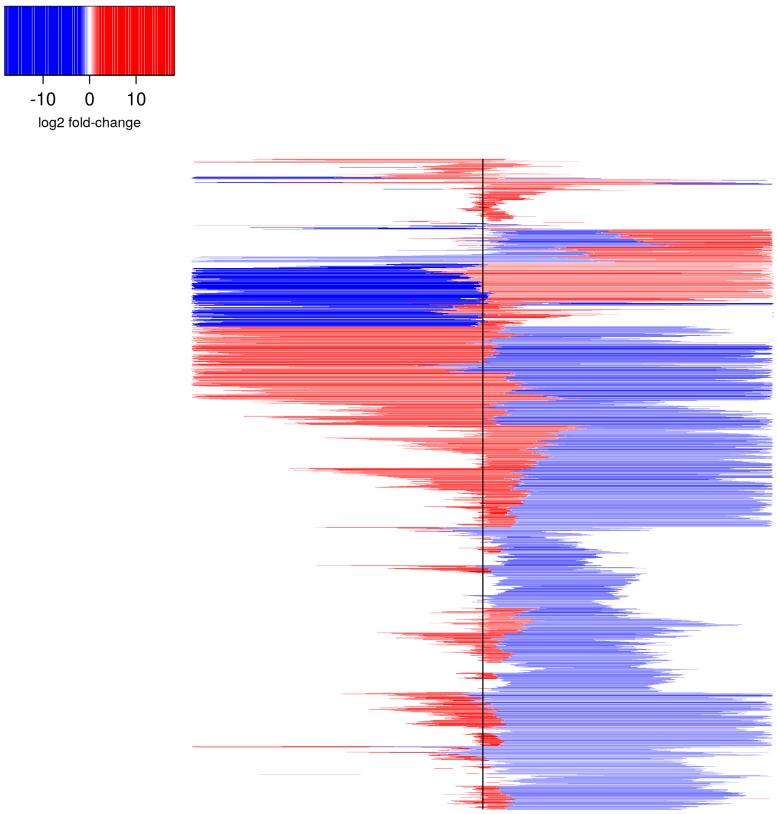
2 Identification of TSS

To identify TSS from the PROcap-seq and PRO-seq of flavopiridol-treated cells, the iTiSS program [Jürges et al., 2021] was run separately for each sample in the SPARSE_PEAK mode with standard parameters. Afterward, the iTiSS TSRMerger program was used to select only peaks that were identified in both samples within ±5 bp resulting in 136,090 putative TSS positions. These were further filtered by requiring a maximum distance of 500 bp to the nearest annotated gene. This resulted in 42,193 potential TSS positions for 7,650 genes. The TSS with the highest expression was selected for each gene.

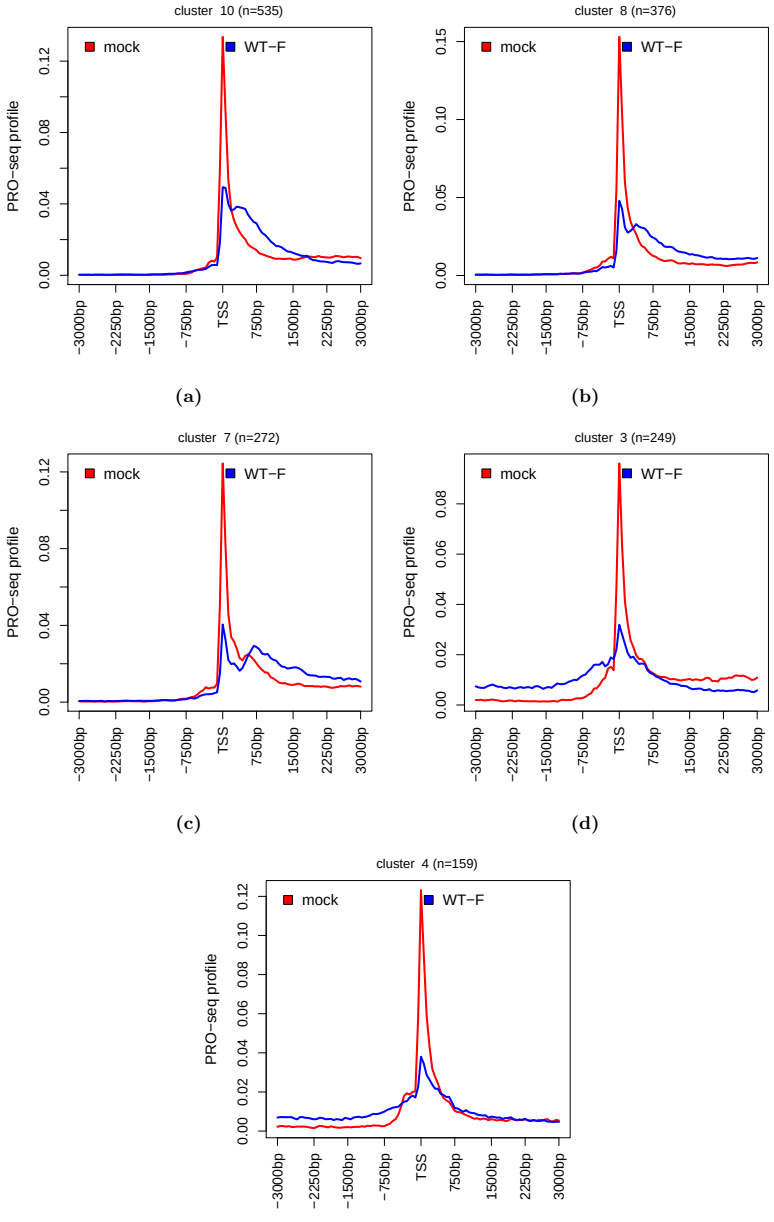
3 Supplementary Figures



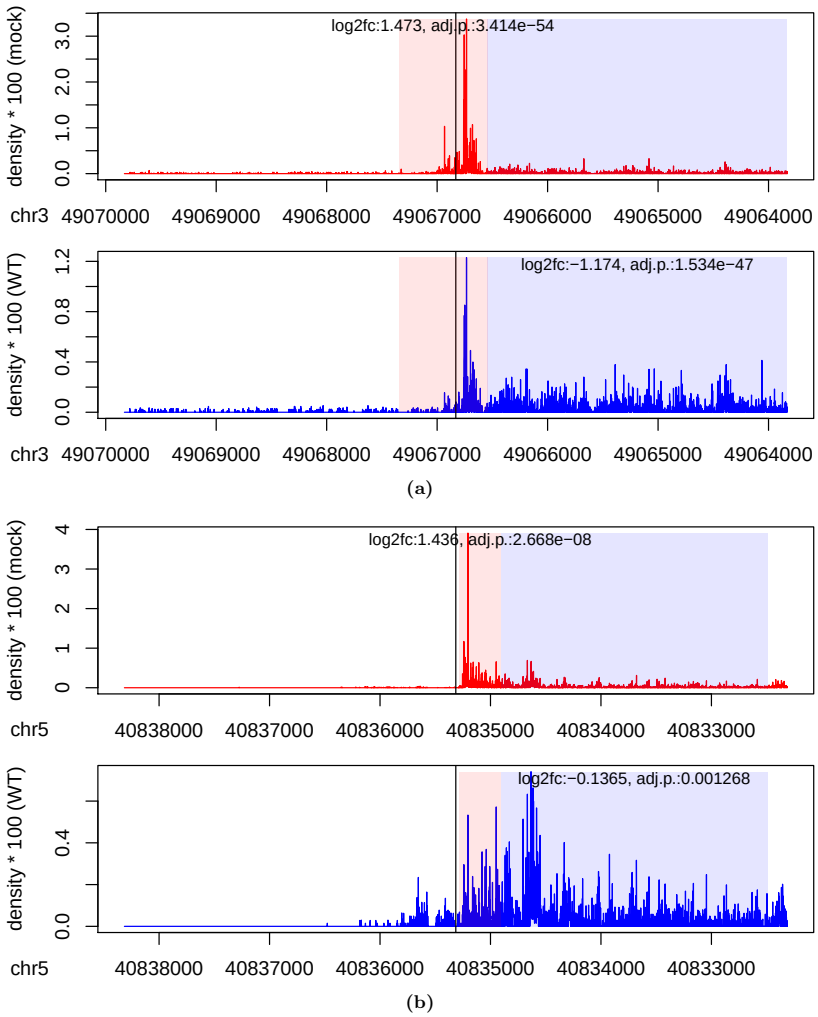
Supplementary Fig. S1 Boxplot showing log2 fold-changes for mock vs. WT infection determined with DEXseq for WT>mock, mock>WT regions and filler regions. Positive log2 fold-changes indicate increased use in mock and negative log2 fold-changes increased use in WT.



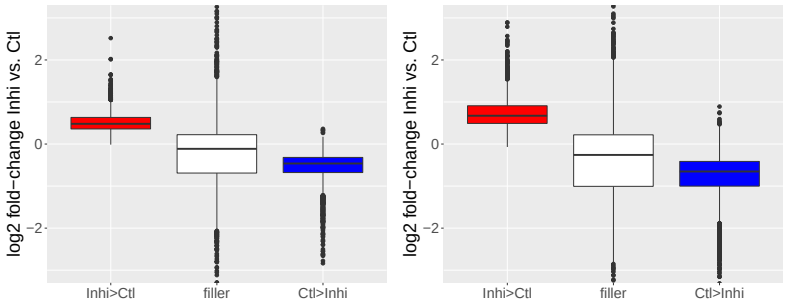
Supplementary Fig. S2 Heatmap illustrating log2 fold-changes in mock vs. WT determined with DEXseq for regions of change with adj. $p \leq 0.01$. The order of windows is the same as in Fig. 3a. Here, all regions with log2 fold-change ≤ -2 or log2 fold-change ≥ 2 are shown with the darkest blue or red, respectively. Filler regions and regions with no p-value or adj. $p > 0.01$ are shown in white.



Supplementary Fig. S3 Metagene curves in the ± 3 kb around the TSS for mock (red) and WT (blue) infection for selected clusters from Fig. 3 in the main manuscript (cluster numbers and number of genes in each cluster on top of subfigures). For this purpose, the regions -3 kb to +3 kb of the TSS were divided into 101 bp bins for each gene. For each bin, the average coverage per genome position was calculated in a strand-specific manner for PRO-seq data and bin read coverages were then normalized by dividing by the total sum of all bins. Metagene curves for each replicate were created by averaging results for corresponding bins across all genes and then showing the average metagene curves across replicates.

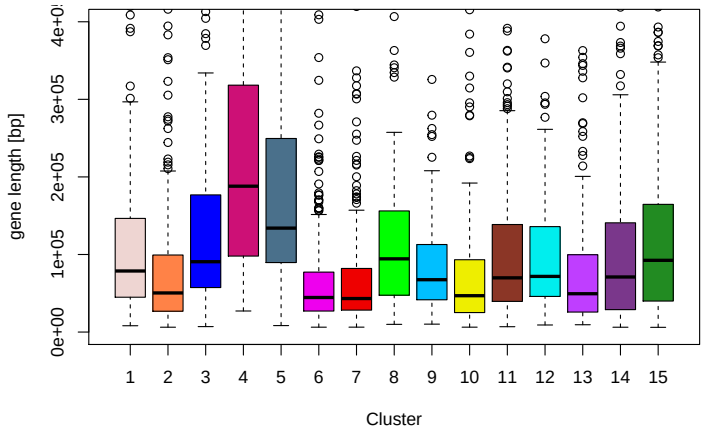


Supplementary Fig. S4 Read density in mock (red, top panel) and WT (blue, bottom panel) infection for example windows from cluster 8. The black vertical line marks the TSS and windows are shown in 5' to 3' direction, i.e. regions up- and downstream of the TSS are to the left and right, respectively, of the vertical line. mock>WT regions are marked by a red shaded rectangle and WT>mock regions by a blue shaded rectangle. Log2 fold-changes and adj. p-values in mock vs. WT infection are shown in the top panel for mock>WT regions and in the bottom panel for WT>mock regions.



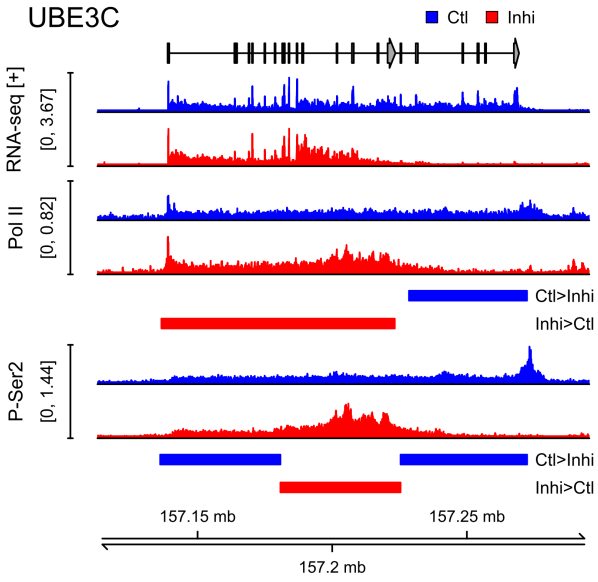
(a) **(b)**

Supplementary Fig. S5 Boxplots showing log2 fold-changes for CDK12 inhibitor treatment (Inhi) vs. control (Ctl) determined with DEXseq for Inhi>Ctl, Ctl>Inhi and filler regions determined from Pol II (a) and P-Ser2 ChIP-seq (b). Positive log2 fold-changes indicate higher use in Inhi and negative log2 fold-changes higher use in Ctl.

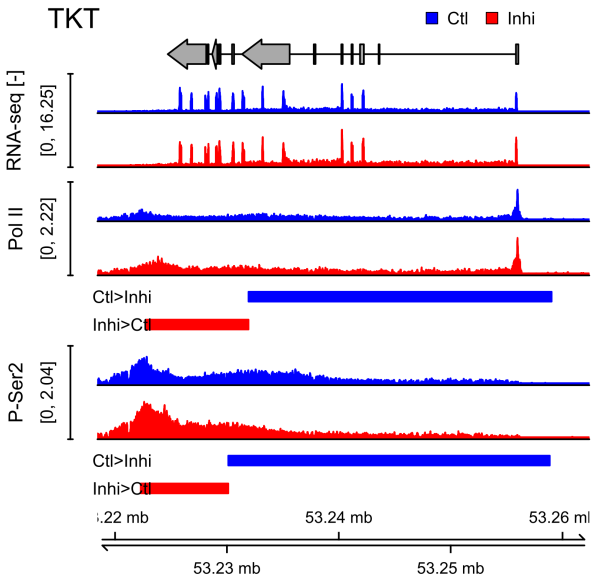


Supplementary Fig. S6 Boxplots showing the distribution of gene lengths for the clusters shown in Fig. 4.

.



(a)



(b)

Supplementary Fig. S7 Read coverage plots showing nuclear RNA-seq data on the respective strand and Pol II and P-Ser2 ChIP-seq data for example genes ((a) UBE3C, from cluster 1 in Fig. 4, (b) TKT, from cluster 15 in Fig. 4) for control (Ctl, blue) and CDK12 inhibitor treatment (Inhi, red). Read counts were normalized to the total number of mapped reads per sample and averaged between replicates. Blue and red boxes below Pol II and P-Ser2 tracks indicate identified regions of change in Pol II and P-Ser2 ChIP-seq data, respectively. Exon (boxes) and intron (lines) structure of corresponding genes is shown on top of subfigures, with gene strand indicated by arrowheads.

4 Comparison to competing approaches

4.1 Comparison to XCAVATOR

Since tools for discovery of copy number variants (CNVs) based on identifying changes in read depth (RD) appear to be potentially useful for identifying regions with differential read distributions, we first evaluated XCAVATOR [Magi et al., 2017] for our applications. XCAVATOR is an RD-based approach that performs differential CNV detection between two conditions (instead of one condition against a reference genome) and is able to also detect multiple copy amplifications. Thus, it appeared most appropriate for identifying genomic regions with differential read distributions among the CNV tools we researched.

Unfortunately, application of XCAVATOR to both PRO-seq data for mock and WT HSV-1 infection from the study by Birkenheuer et al. [2018] and ChIP-seq data for Pol II and P-Ser2 with DMSO (Ctl) or 3-MB-PP1 (Inhi) treatment for 4.5 h from the study by Chirackal Manavalan et al. [2019] did not identify any differential regions. To test that we ran XCAVATOR correctly, we applied it also to low-coverage whole genome shotgun read data from the 1000 genomes project (for individuals NA12878 and NA12815) [1000 Genomes Project Consortium, 2015]. This test did identify ∼17,000 differential regions, which excludes usage errors on our side. The likely explanation why XCAVATOR does not identify any differential regions on the PRO-seq and ChIP-seq data is that changes observed on these data do not follow the assumptions underlying RD-based tools for detecting CNVs. In brief, RD-based tools generally assume that the number of reads mapping to any region of the reference genome (=read count) follows a Poisson distribution and is proportional to the copy number of this region. Thus, the copy number for a genomic region can be estimated from consecutive windows, which should show similar fold-changes to the reference or between conditions (taking into account noise) if they have the same copy number. Furthermore, sudden shifts between consecutive windows with different copy numbers are expected. In contrast, changes in read counts for consecutive subwindows of the differential regions that RegCFinder aims to identify may in- or decrease (see e.g. Fig. 1c,d) and no sudden shifts are observed at the end of the differential regions. It is thus not surprising that XCAVATOR (and similar tools) cannot recover these types of differential regions, for which they were not developed.

4.2 Comparison to diffReps

We also compared RegCFinder against diffReps [Shen et al., 2013]. diffReps also pursues a sliding window-based approach on the whole genome to identify differential regions similar to RD approaches for CNV detection. However, it first determines the significance of a change within each sliding window separately using a negative binomial distribution. Subsequently, diffReps merges overlapping significant windows and recalculates significance of merged windows. This does not require (approximately) the same fold-changes for all subwindows of a differential region and no sudden shifts at the end of regions.

We applied diffReps to both the PRO-seq data and ChIP-seq data using default parameters (in particular, window size = 1000, step size = 100). For this purpose, sequence alignments in BAM format were first converted to the BED format required by diffReps using the bamtobed utility of BEDTools [Quinlan and Hall, 2010]. Since diffReps uses a sliding window approach across the complete genome, no target windows can be defined. To compare diffReps results against RegCFinder, we then analyzed the overlap of significant regions identified by diffReps to input windows used for RegCFinder. Furthermore, since diffReps was developed for ChIP-seq data, it does not consider read strand. Thus, the diffReps analysis for the PRO-seq data was performed in an unstranded manner.

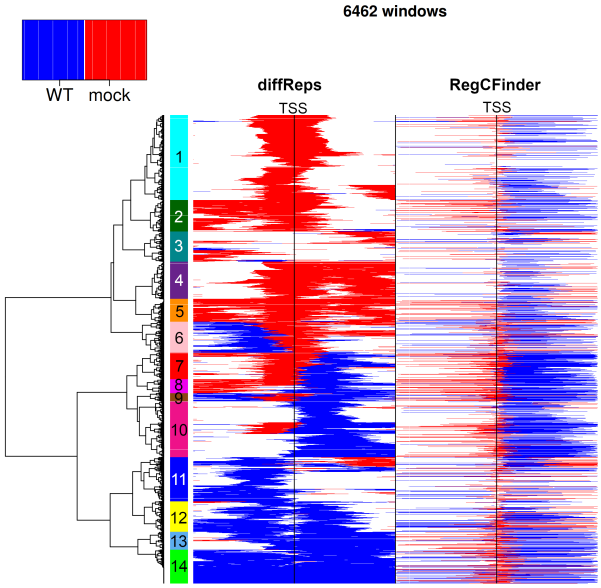
4.3 Results on PRO-seq data for HSV-1 infection

Application of diffReps to PRO-seq data for mock and WT HSV-1 infection identified a total of 71038 differential regions, with 30022 of these determined as up-regulated by diffReps in WT infection and 41016 as down-regulated. Almost all of these (70299 = 99%) were significant at an adjusted p-value cutoff of 0.01 (p-value calculated by diffReps). To directly compare results against RegCFinder, we evaluated the 9040 differential regions identified by diffReps that overlapped the 7650 promoter windows used as input for RegCFinder. Here, 6462 of the promoter windows overlapped with at least one differential region identified by diffReps. Location of these differential windows is visualized in the heatmap in Supplementary Fig. S8 on the left side in the same way as for the RegCFinder results in Fig. 3a . The right side of Supplementary Fig. S8 shows regions identified by RegCFinder for these windows. Here, “WT” (blue) indicates regions found to be up-regulated in HSV-1 infection by diffReps or WT>mock

regions from RegCFinder and "mock" (red) down-regulated regions identified by diffReps or mock>WT regions from RegCFinder. To identify distinct patterns in the regions identified by diffReps, we clustered the heatmap for diffReps results according to Euclidean distances and Ward's clustering criterion. The RegCFinder regions in Supplementary Fig. S8 are ordered according the clustering on the diffReps regions. This comparison showed that, with the exception of (parts of clusters) 7-10, the patterns of changes identified by diffReps differed strongly from the changes identified by RegCFinder. If not noted explicitly otherwise, differential regions mentioned below are differential regions identified by diffReps.

Here, clusters 1-5 represented promoter windows containing almost only differential regions down-regulated in WT compared to mock. For most of these windows (clusters 1, 2, 4 and 5), these regions covered the TSS and upstream (in case of clusters 2 and 5) or downstream regions (in case of clusters 4 and 5). Birkenheuer et al. previously showed using these PRO-seq data that HSV-1 infection leads to a loss of Pol II both at gene promoters and gene bodies for the majority of human genes [Birkenheuer et al., 2018]. This suggests that diffReps identified reduced presence of Pol II at promoters in clusters 1, 2, 4 and 5 as well as gene bodies in clusters 4 and 5, which was confirmed by manual inspection in a genome viewer for example genes from these clusters (Supplementary Fig. S9a,b). Here, diffReps did not identify the relative increases downstream of the TSS in HSV-1 infection found by RegCFinder for many of these genes.

For diffReps clusters 2 and 5, for which down-regulated regions were also found upstream of TSS, inspection indicated that this represented a reduction in antisense transcription from these promoters (Supplementary Fig. S9c) or other close-by promoters on the opposite strand (Supplementary Fig. S9d). Such pairs of close-by promoters on opposite strands with transcription diverging from these promoters are denoted as divergent promoters. Notably, for most human promoters, transcription initiation is



Supplementary Fig. S8 Heatmap showing the location and type of differential regions identified by diffReps (left side) or RegCFinder (right side) for 6462 windows with at least one differential region determined by diffReps (adjusted p-value ≤ 0.01). The TSS is indicated by black vertical lines in both cases. Windows were clustered according to the location of diffReps differential regions and RegCFinder results are shown according to this order (Color scheme: red = regions down-regulated according to diffReps or mock>WT regions according to RegCFinder, blue=regions up-regulated according to diffReps or WT>mock regions according to RegCFinder).

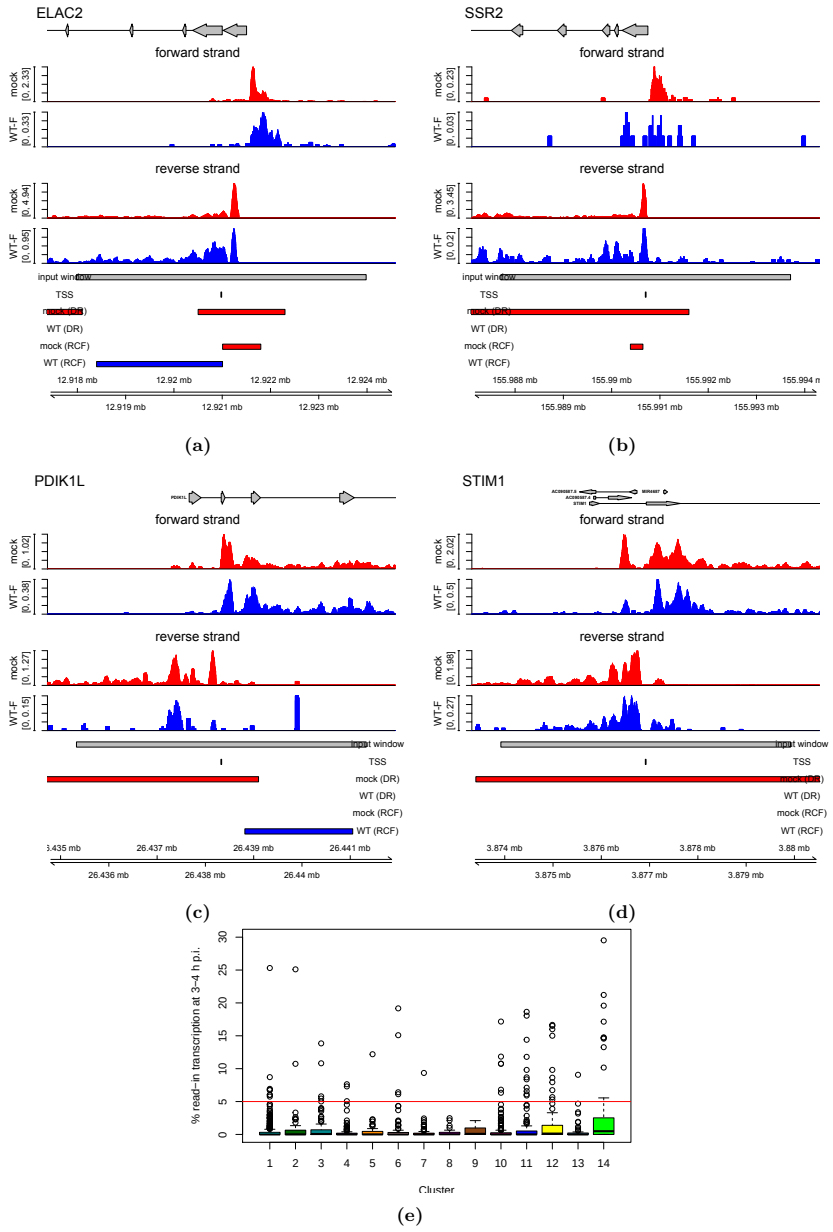
bidirectional but elongation occurs only in the sense direction while antisense transcription is quickly terminated [He et al., 2008, Preker et al., 2008, Seila et al., 2008, 2009]. Reduced levels of Pol II at human promoters during HSV-1 infection thus likely also lead to reduced antisense transcription. While PRO-seq data is strand-specific, diffReps does not consider strand as noted above and thus cannot distinguish between changes in sense and antisense transcription. In summary, diffReps clusters 1-5 predominantly reflect the loss of Pol II on host genes and thus changes in absolute levels of Pol II, but not differences in the distribution of Pol II.

Clusters 6, 9, 11-14 represent windows for which regions up-regulated in HSV-1 infection are found upstream of the TSS by diffReps. Analysis of read-in transcription (Supplementary Fig. S9e) and manual inspection of example genes indicated that up-regulated regions upstream of the TSS partly resulted from either (i) read-in transcription from upstream genes (in particular for some windows in clusters 12 and 14, example in Supplementary Fig. S10a), (ii) divergent promoters (Supplementary Fig. S10b) or (iii) promoters for which (non-productive) antisense transcription continued further into upstream regions before it was terminated (Supplementary Fig. S10c). Consistent with (i), RegCFinder also identified WT>mock regions upstream of the TSS for some windows in diffReps clusters 6, 11, 12 and 14. However, enrichment of read-in transcription for any diffReps clusters was by far not as high as for RegCFinder clusters 3 and 4 (see Fig. 3c), for which a substantial fraction of genes showed read-in >5%, i.e. greater than the cutoff we previously used to distinguish genes with read-in transcription. Thus, RegCFinder clusters 3 and 4 more specifically identify genes with read-in transcription than any of the clusters from the diffReps analysis.

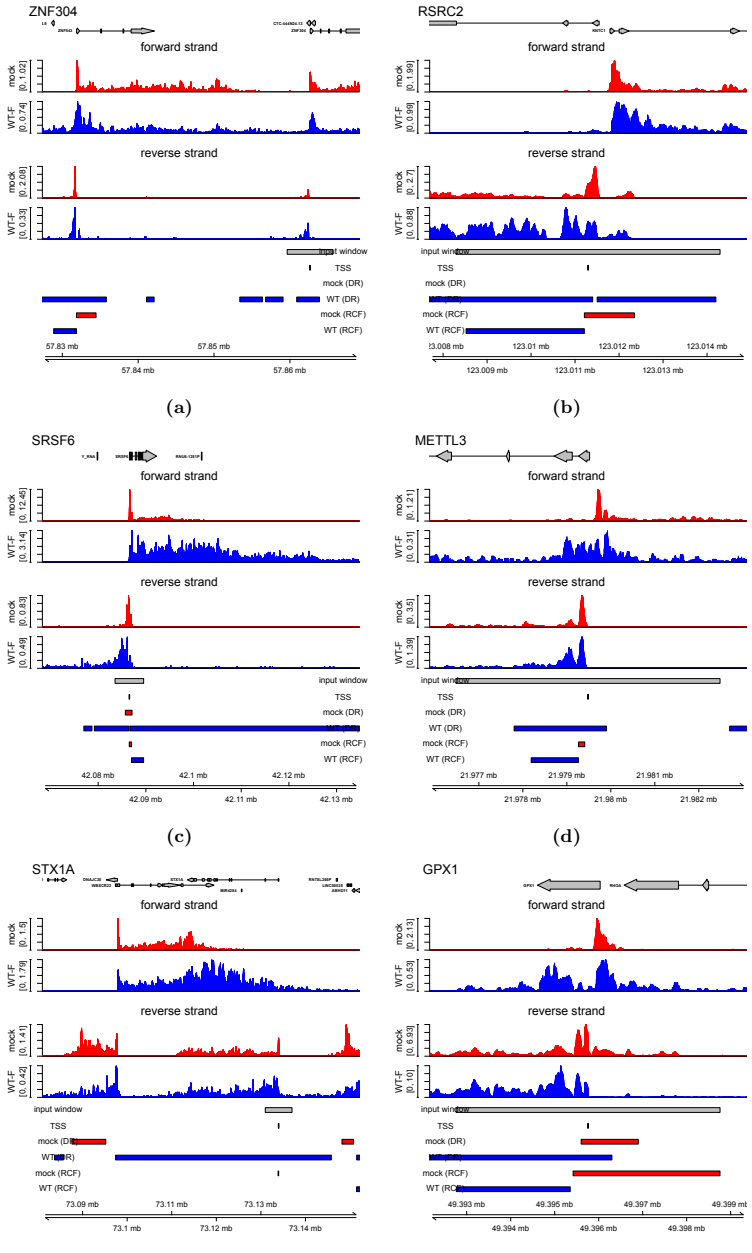
Observations (ii) and (iii) are consistent both with the broadening of antisense Pol II pausing peaks we previously reported [Weiß et al., 2023] and previously reported activation of antisense transcription in HSV-1 infection [Wyler et al., 2017]. Both phenomena lead to increased transcription further upstream of the TSS. However, up-regulated upstream regions were not consistently identified by diffReps for all genes with extended antisense transcription. They were also not identified by RegCFinder, as it performs the analysis in a strand-specific manner. To identify changes in antisense Pol II distribution, input windows for the opposite strand would have to be used.

Up-regulated regions downstream of the TSS observed in clusters 7-10 and 12-14 represented increased Pol II levels on the gene body. Generally, this was either due to (i) increased elongation along the whole gene and potentially downstream of the gene due to read-through transcription beyond poly(A) sites (example in Supplementary Fig. S10c), (ii) a downstream shift in Pol II pausing (Supplementary Fig. S10d) or in a few cases (iii) read-through transcription from a downstream gene on the opposite strand (Supplementary Fig. S10e). However, (i) and (ii) were also observed for windows for which diffReps only identified down-regulated regions (clusters 1, 2, 4 and 5, e.g. Supplementary Fig. S9a). In contrast, RegCFinder often identified WT>mock regions for genes in these clusters.

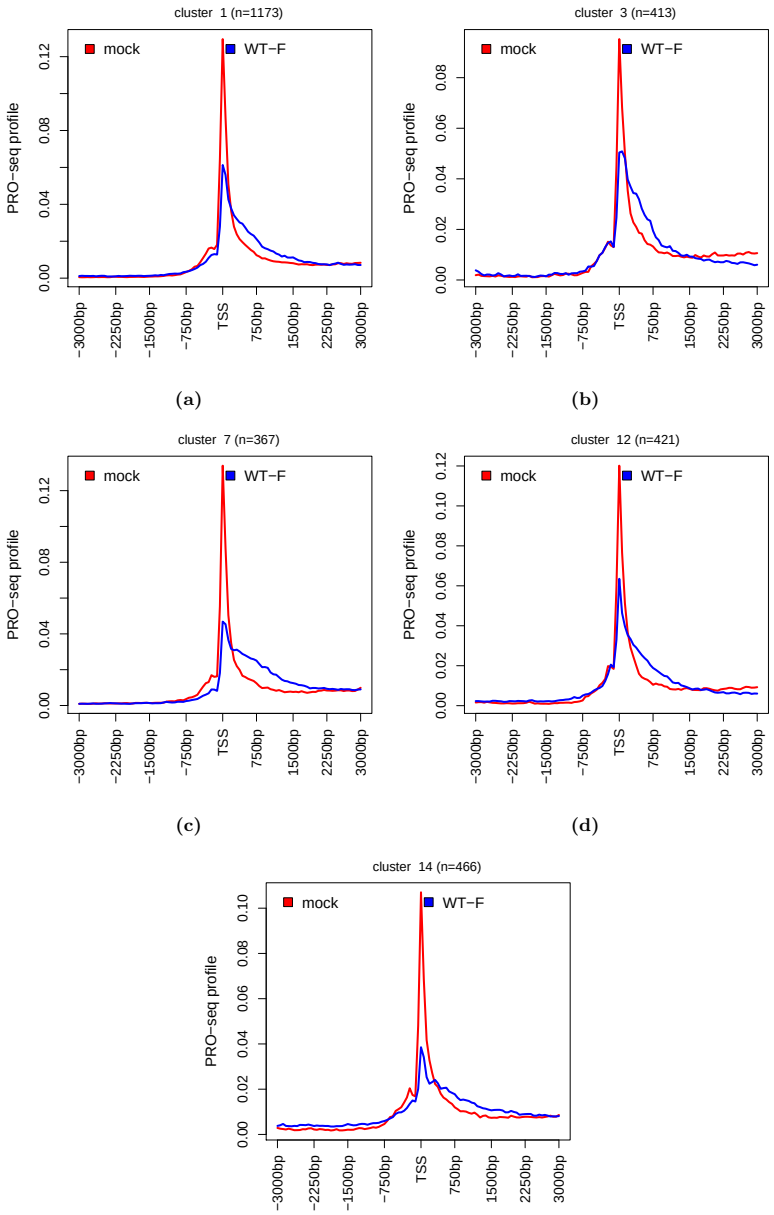
Accordingly, metagene plots for all diffReps clusters showed a downstream broadening of the PRO-seq signal and differed only in the extent of broadening (Supplementary Fig. S11). Furthermore, consistent with presence of read-in transcription for some genes in clusters 12 and 14, some small increases upstream of the TSS were observed in metagene plots for these clusters (Supplementary Fig. S11d,e). In contrast, metagene plots for RegCFinder clusters showed more diverse patterns, including not only a downstream broadening of PRO-seq peaks but increased upstream levels for RegCFinder clusters 3 and 4 (Supplementary Fig. S3d,e), reflecting the read-in transcription observed for many genes of these clusters, and increased downstream peaks for clusters 7, 8 and 10 (Supplementary Fig. S3a-c) Finally, up-regulated regions identified by diffReps often included the major Pol II peak at the TSS, which was reduced relative to downstream regions (Supplementary Fig. S10b-d,f). In summary, these observations suggest that diffReps predominantly does not detect the change in read distributions but rather absolute changes in read depth.



Supplementary Fig. S9 (a-d) Read coverage plots showing PRO-seq coverage in mock (red) and WT HSV-1 infection (blue) separately on the positive and negative strand for example genes. Input windows (gray), differential regions identified by diffReps (DR, red=down-regulated in HSV-1 infection, blue=up-regulated in HSV-1 infection) and regions of change identified by RegCFinder (RCF, red=mock>WT, blue=WT>mock) are shown below read coverage tracks. Exon (boxes) and intron (lines) structure of genes in this genomic region is shown on top of subfigures, with gene strand indicated by arrowheads. The central gene for which the promoter window was defined is indicated in the top left of subfigures. (e) Boxplots showing the distribution of the % read-in transcription for the 14 clusters obtained for differential regions identified by diffReps in Supplementary Fig. S8. For details on calculation of read-in transcription, see legend to Fig. 3c.



Supplementary Fig. S10 Read coverage plots as in Supplementary Fig. S9 for example genes: (a) example gene (ZNF304, on the right side) with read-in transcription from an upstream gene (ZNF543, on the left side); (b) two genes (RSRC2, KNTC1) with divergent promoters on opposite strands; (c) a gene (SRSF6) with increased elongation along the whole gene and read-through transcription; (d) a gene (METTL3) with a downstream shift in Pol II pausing; (e) a gene (STX1A, reverse strand) for which read-through transcription is observed for a downstream gene on the opposite strand (WBSCR22, forward strand); (f) a gene (GPX1) with a downstream shift in Pol II pausing, for which the up-regulated region identified by diffReps includes the TSS.



Supplementary Fig. S11 Metagene curves in the ± 3 kb around the TSS for mock (red) and WT HSV-1 (blue) infection for selected clusters from Supplementary Fig. S8 (cluster numbers and number of genes in each cluster on top of subfigures). For a description on how metagene curves were calculated see caption to Supplementary Fig. S3.

4.4 Results on ChIP-seq data for CDK12 inhibition

Application of diffReps to Pol II and P-Ser2 ChIP-seq data identified 23215 differential regions on the Pol II ChIP-seq data (23004 with adj. p-value < 0.01). 11590 of these were up-regulated and 11625 down-regulated. 10778 ($\sim$46%) of the differential regions overlapped 3630 of the gene windows we had defined as input for RegCFinder (Supplementary Fig. S12a). For the P-Ser2 ChIP-seq data 42064 differential regions were identified (41909 with adj. p-value < 0.01), with 15762 region up-regulated and 26302 down-regulated. 20313 ($\sim$48%) of these differential windows overlapped 4776 of the gene windows (Supplementary Fig. S12b).

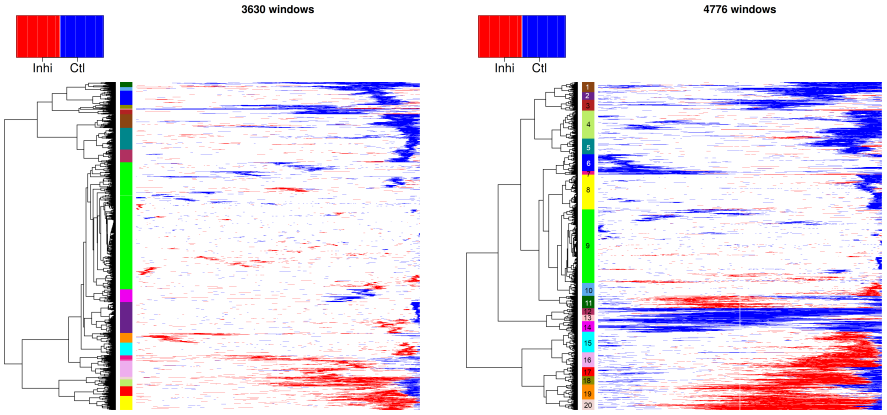
Clustering analysis of the location heatmaps indicated that in both cases a large fraction of gene windows (39% and 22% for the Pol II and P-Ser2 ChIP-seq data, respectively) belonged to one large cluster with no specific pattern regarding the location of differential regions. This was even the case when selecting a relatively stringent clustering cutoff for the P-Ser2 data, which resulted in 20 clusters. As the differential regions identified by diffReps on the Pol II ChIP-seq data were fewer and covered smaller fractions of input gene windows, we focused on the diffReps results on the P-Ser2 data for the comparison against RegCFinder. Here, more distinctive patterns were observed and many clusters showed down-regulation at or close to the end of genes upon inhibitor treatment (clusters 1, 4, 5, 8, 11, 12, 14-18 and 20) as expected from our previous study [Chirackal Manavalan et al., 2019]. Some, but not all of these clusters also exhibited regions up-regulated upon inhibitor treatment upstream of these down-regulated regions (clusters 15-20).

Supplementary Fig. S12c shows a direct comparison of differential regions identified by diffReps (left) and RegCFinder (right) on the P-Ser2 data. diffReps identified on average twice as many regions per window (4.25 on average) than RegCFinder (2.16 on average). Often this was due to diffReps identifying multiple small regions where RegCFinder identified one large region covering all of these (see e.g. Supplementary Fig. S13a-e). Thus, some sort of clustering of close-by differential regions from diffReps would have to be applied to group these together into "meta-regions" for downstream analyses.

Apart from the large, non-specific cluster 9, RegCFinder generally identified the same differential regions as diffReps – but commonly as one continuous region rather than the multiple smaller regions identified by diffReps – as well as additional up- and/or downstream regions (e.g. Supplementary Fig. S13b,d,f). Manual inspection of example genes confirmed that in many cases a clear relative increase in reads was observed in these additional regions identified by RegCFinder (e.g. Inhi>Ctl regions in Supplementary Fig. S13b,f). Other additional differential regions reflect a shift in the relative distribution of reads from these regions to other regions in the window (e.g. Ctl>Inhi regions in Supplementary Fig. S13d,e). That diffReps does not identify these regions as differential likely reflects the different objective of diffReps, which focuses on identifying regions with differences in ChIP-seq enrichment rather than changes in the distribution of reads within particular windows.

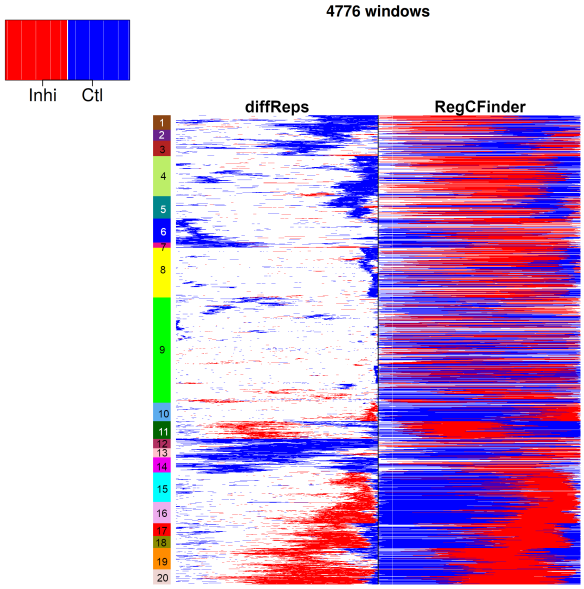
For the large non-specific cluster 9, diffReps identified only short (relative to gene length) differential regions and no consistent patterns with no similarity to RegCFinder results. diffReps cluster 9 contained very long genes (Supplementary Fig. S14a, e.g. Supplementary Fig. S13g). Consistently, genes from RegCFinder clusters 4 and 5, which represented the clusters with the longest genes in the RegCFinder clustering analysis, were strongly enriched in diffReps cluster 9. Here, diffReps cluster 9 contained around 48% of RegCFinder clusters 4 and 5 genes, but diffReps results did not reflect the distinctive pattern identified by RegCFinder with a relative increase in P-Ser2 close to the gene start upon CDK12 inhibition and a decrease of P-Ser2 further downstream. Thus, the loss of P-Ser2 signal towards gene ends in particular for long genes is not as clearly reflected in diffReps results as in RegCFinder results. In part, this is likely due to the fact that these genes are also among the most lowly expressed genes (according to nuclear RNA-seq data, Supplementary Fig. S14b). Thus, diffReps likely does not identify these differential regions due to low read counts (see e.g. Supplementary Fig. S13h).

In summary, our comparison shows that diffReps identifies only a subset of differential regions with changes in the P-Ser2 distribution identified by RegCFinder on the input gene windows, while RegCFinder recovers most differential regions identified by diffReps.



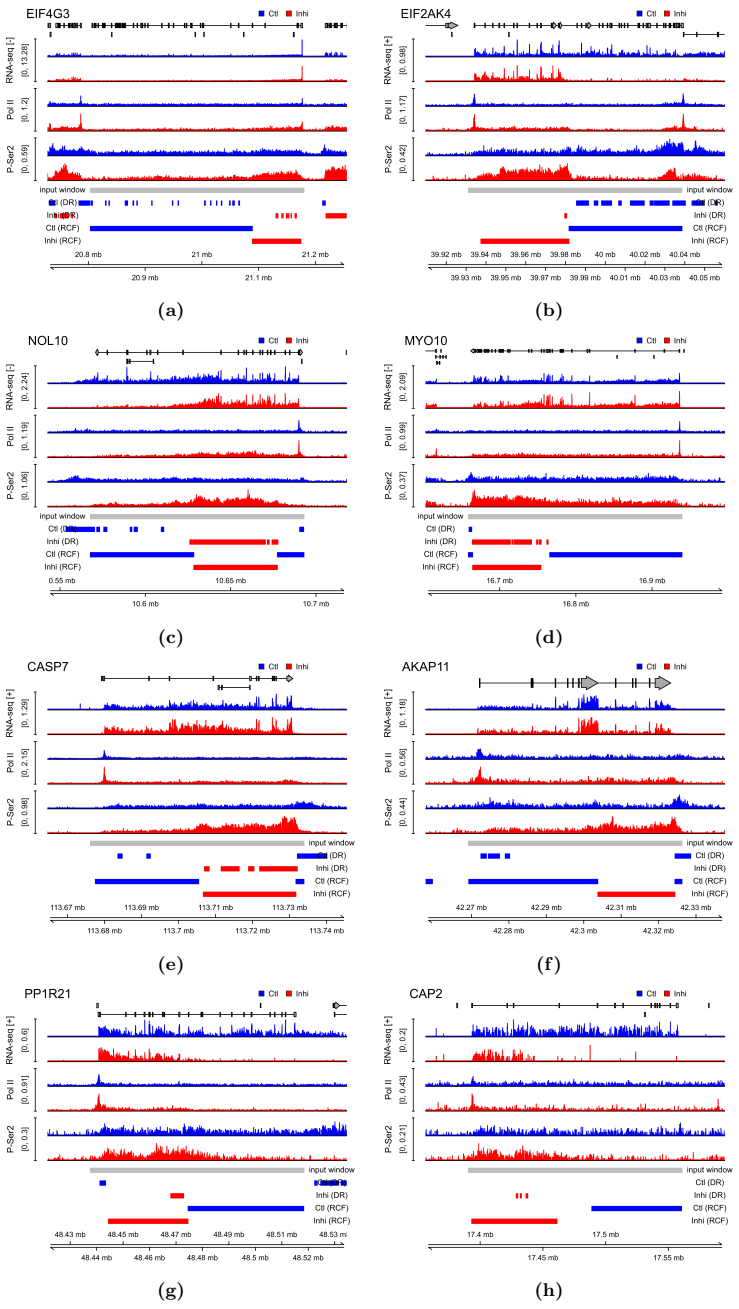
(a)

(b)

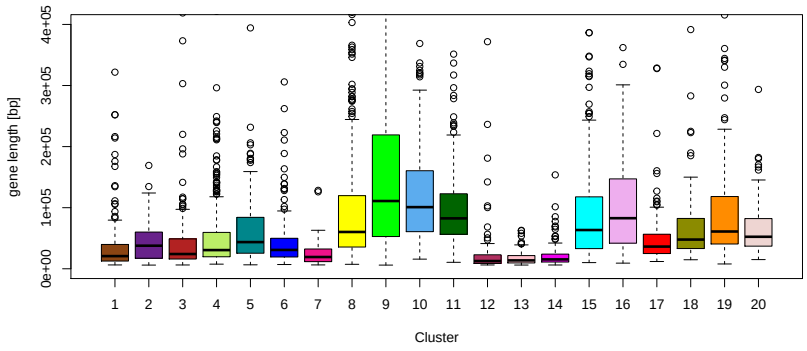


(c)

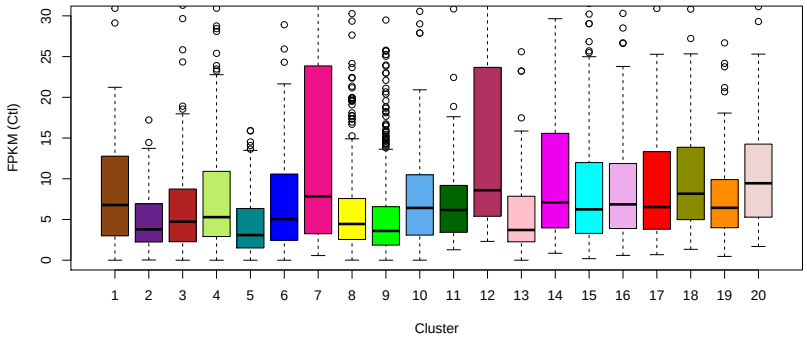
Supplementary Fig. S12 (a,b) Heatmap showing the location and type of differential regions identified by diffReps on the Pol II (a) and P-Ser2 (b) ChIP-seq data for windows with at least one differential region determined by diffReps (adjusted p-value ≤ 0.01). (c) Heatmap showing the location and type of differential regions identified by diffReps (left side) or RegCFinder (right side) for the P-Ser2 ChIP-seq data for windows with at least one differential region determined by diffReps (adjusted p-value ≤ 0.01). Windows were ordered according to the clustering in (b). Color scheme: blue = regions down-regulated according to diffReps upon inhibitor treatment or Ctl>Inhi regions according to RegCFinder, red=regions up-regulated according to diffReps or Inhi>Ctl regions according to RegCFinder. Cluster numbers are indicated on the left.



Supplementary Fig. S13 Read coverage plots showing nuclear RNA-seq data on the respective strand and Pol II and P-Ser2 ChIP-seq data for example genes for control (Ctl, blue) and CDK12 inhibitor treatment (Inhi, red). Read counts were normalized to the total number of mapped reads per sample and averaged between replicates. Input windows (gray), differential regions identified by diffReps (DR, blue=down-regulated upon inhibitor treatment, red=up-regulated upon inhibitor treatment) and regions of change identified by RegCFinder (RCF, blue=Ctl>Inhi, red=Inhi>Ctl) are shown below read coverage tracks. Exon (boxes) and intron (lines) structure of corresponding genes is shown on top of subfigures, with gene strand indicated by arrowheads. Gene symbols of the central gene for which the window was defined are shown on the top left.



(a)



(b)

Supplementary Fig. S14 Boxplots showing distribution of (a) gene length and (b) gene expression (calculated as fragments per million mapped reads =: FPKM) in nuclear RNA-seq data for control samples for the 20 clusters identified from the location heatmap for differential regions identified by diffReps on the P-Ser2 data from Supplementary Fig. S12b.

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

C. H. Birkenheuer, C. G. Danko, and J. D. Baines. Herpes simplex virus 1 dramatically alters loading and positioning of RNA Polymerase II on host genes early in infection. *Journal of virology*, 92, Apr. 2018.

T. Bonfert, E. Kirner, G. Csaba, R. Zimmer, and C. C. Friedel. ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC bioinformatics*, 16:122, Apr. 2015.

A. P. Chirackal Manavalan, K. Pilarova, M. Kluge, K. Bartholomeeusen, M. Rajecky, J. Oppelt, P. Khirsariya, K. Paruch, L. Krejci, C. C. Friedel, and D. Blazek. Cdk12 controls G1/S progression by regulating RNAPII processivity at core DNA replication genes. *EMBO reports*, 20:e47592, Sept. 2019.

P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of samtools and bcftools. *GigaScience*, 10: giab008, Feb. 2021.

Y. He, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, and K. W. Kinzler. The antisense transcriptomes of human cells. *Science (New York, N.Y.)*, 322:1855–1857, 2008.

C. S. Jürges, L. Dölken, and F. Erhard. Integrative transcription start site identification with iTiSS. *Bioinformatics*, 37:3056–3057, 2021.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25:1754–1760, July 2009.

A. Magi, T. Pippucci, and C. Sidore. XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC genomics*, 18: 747, 2017.

P. Preker, J. Nielsen, S. Kammler, S. Lykke-Andersen, M. S. Christensen, C. K. Mapendano, M. H. Schierup, and T. H. Jensen. RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, 322:1851–1854, 2008.

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

A. C. Seila, J. M. Calabrese, S. S. Levine, G. W. Yeo, P. B. Rahl, R. A. Flynn, R. A. Young, and P. A. Sharp. Divergent transcription from active promoters. *Science*, 322:1849–1851, 2008.

A. C. Seila, L. J. Core, J. T. Lis, and P. A. Sharp. Divergent transcription: a new feature of active promoters. *Cell cycle*, 8:2557–2564, 2009. ISSN 1551-4005. doi: 10.4161/cc.8.16.9305.

L. Shen, N.-Y. Shao, X. Liu, I. Maze, J. Feng, and E. J. Nestler. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PloS one*, 8:e65598, 2013.

E. Weiß, T. Hennig, P. Graßl, L. Djakovic, A. W. Whisnant, C. S. Jürges, F. Koller, M. Kluge, F. Erhard, L. Dölken, and C. C. Friedel. HSV-1 infection induces a downstream shift of promoter-proximal pausing for host genes. *Journal of virology*, page e0038123, 2023.

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*, 18:209, 2017.

HSV-1 infection induces a downstream shift of the +1 nucleosome

Elena Weiß[1], Adam W. Whisnant[2,3], Thomas Hennig[2,3], Lara Djakovic[2], Lars Dölken[2,3,4#], Caroline C. Friedel[1#]

[1] Institute of Informatics, Ludwig-Maximilians-Universität München, Amalienstr. 17, 80333 Munich, Germany

[2] Institute for Virology and Immunobiology, Julius-Maximilians-University Würzburg, Versbacher Straße 7, 97078 Würzburg, Germany

[3] Institute for Virology, Medizinische Hochschule Hannover, Carl-Neuberg-Str. 1, 30625 Hannover, Germany

[4] Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Center for Infection Research (HZI), 97080 Würzburg, Germany

[#] Address correspondence: lars.doelken@uni-wuerzburg.de; Caroline.Friedel@bio.ifi.lmu.de

Running Title: HSV-1-induced downstream shift of the +1 nucleosome

Abstract

Herpes simplex virus 1 (HSV-1) infection induces a loss of host transcriptional activity and widespread disruption of host transcription termination, which leads to an induction of open chromatin downstream of genes. In this study, we show that lytic HSV-1 infection also leads to an extension of chromatin accessibility at promoters into downstream regions. This is most prominent for highly expressed genes and independent of the immediate-early proteins ICP0, ICP22, and ICP27 and the virion host shutoff protein *vhs*. ChIPmentation of the noncanonical histone variant H2A.Z, which is strongly enriched at +1 and -1 nucleosomes, indicated that these chromatin accessibility changes are linked to a downstream shift of +1 nucleosomes. In yeast, downstream shifts of +1 nucleosomes are induced by RNA Polymerase II (Pol II) degradation. Accordingly, irreversible depletion of Pol II from genes in human cells using α-amanitin altered +1 nucleosome positioning similar to lytic HSV-1 infection. Consequently, treatment with phosphonoacetic acid (PAA) and knockout of ICP4, which both prevent viral DNA replication and alleviate the loss of Pol II from host genes, largely abolished the downstream extension of accessible chromatin in HSV-1 infection. In the absence of viral DNA replication, doxycycline-induced expression of ICP27, which redirects Pol II from gene bodies into intergenic regions by disrupting transcription termination, induced an attenuated effect that was further enhanced by co-expression of ICP22. In summary, our study provides strong evidence that HSV-1-induced depletion of Pol II from the host genome leads to a downstream shift of +1 nucleosomes at host promoters.

Importance

Lytic herpes simplex virus 1 (HSV-1) infection leads to a profound host transcription shutoff. Loss of RNA Polymerase II (Pol II) in yeast has previously been shown to relax +1 nucleosome positioning to more thermodynamically favorable sites downstream of transcription start sites. Here, we show that a similar phenomenon is likely at play in lytic HSV-1 infection. Sequencing of accessible chromatin revealed a widening of nucleosome-free regions at host promoters into downstream regions. By mapping genome-wide positions of the noncanonical histone variant H2A.Z enriched at +1 and -1 nucleosomes, we demonstrate a downstream shift of +1 nucleosomes for most cellular genes in lytic HSV-1 infection. As chemical depletion of Pol II from genes also leads to a downstream shift of +1 nucleosomes in human cells, changes in chromatin architecture at promoters in HSV-1 infection are likely a consequence of HSV-1-induced loss of Pol II activity from the host genome.

Introduction

Herpes simplex virus 1 (HSV-1) is one of nine human herpesviruses, and more than half of the global population are latently infected with HSV-1 (1-4). HSV-1 does not only cause the common cold sores but is also responsible for life-threatening diseases, particularly in the immunocompromised (5). The HSV-1 genome consists of 152 kilobases (kb) of double-stranded DNA and encodes for at least 121 large open reading frames (ORFs) and >100 small ORFs (6). Upon lytic infection, four of the five viral immediate early (IE) genes, i.e., ICP0, ICP4, ICP22 and ICP27, hijack the host gene expression machinery to ensure viral replication during early stages of infection (7-11). This results in the recruitment of RNA polymerase II (Pol II) from the host chromatin to viral genomes and a general shutoff of host transcription (7, 9, 11-13). Pol II recruitment to viral genomes is predominantly facilitated by the viral IE protein ICP4 (13), which effectively sequesters Pol II into the viral replication compartments (14). The global host transcription shutoff is further exacerbated by ICP22-mediated inhibition of host transcription elongation (15) and the viral *vhs* (virion host shutoff) protein. *vhs* is a nuclease delivered by the tegument of the incoming viral particles that cleaves both viral and host mRNAs (16-18). We recently demonstrated that *vhs* continuously degrades about 30 % of host mRNAs per hour during the first 8 h of high multiplicity lytic infection of primary human fibroblasts (19).

HSV-1 further disturbs host transcription by inducing a widespread disruption of host transcription termination (DoTT) resulting in read-through transcription for tens of thousands of nucleotides (nt) beyond cellular polyadenylation (pA) sites (20). Read-through transcription commonly extends into downstream genes, leading to a seeming transcriptional induction of these genes. DoTT negatively affects host gene expression in two ways: First, as read-through transcripts are not exported (21), DoTT prevents the translation of newly transcribed mRNAs

and thus dampens the host transcriptional response to infection. Second, it redirects a large fraction of ongoing Pol II transcription from host gene bodies into downstream intergenic regions and, thus, further exacerbates the depletion of Pol II from host genes. We recently showed that ICP27 is sufficient for inducing DoTT (22). Interestingly, transcription downstream of genes (DoG) is also induced in cellular stress responses (21, 23-25). Accordingly, infection with an ICP27-null mutant still induces read-through transcription, albeit at much lower levels, presumably due to a virus-induced cellular stress response (26). In contrast to stress-induced DoGs, HSV-1-induced DoTT is associated with a massive increase of chromatin accessibility downstream of genes (denoted as downstream open chromatin regions (dOCRs)) detectable by ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) (21). dOCR induction requires strong absolute levels of transcription downstream of genes, thus it is predominantly observed for highly transcribed genes with strong read-through. Treatment with phosphonoacetic acid (PAA) during HSV-1 infection, which inhibits viral DNA replication (27, 28), results in increased dOCR induction (29). As viral DNA redirects Pol II from the host chromatin, PAA treatment mitigates the loss of Pol II from the host genome and thus increases absolute levels of transcription on host genes and – importantly for dOCR induction – read-through regions. We recently showed that ICP22 is both required and sufficient for inducing dOCRs in the presence of ICP27-induced read-through (29).

As we previously only investigated chromatin accessibility *downstream* of genes, we now further explore our previously published ATAC-seq data to examine changes in chromatin accessibility around host gene promoters. This revealed a broadening of open chromatin regions at promoters into regions downstream of the transcription start sites (TSS) for most host genes. While it was consistently observed in the absence of ICP0, ICP22, ICP27, and *vhs*, PAA treatment and ICP4 knockout, both of which inhibit viral DNA replication and alleviate the HSV-1-induced loss of host transcription, substantially reduced the extension of open

chromatin at the TSS. Weiner *et al.* previously showed that Pol II depletion in yeast induces downstream shifts of +1 nucleosomes, in particular for highly expressed genes (30). Transcribing Pol II disturbs histone-DNA interactions and, at high transcription rates, leads to eviction of nucleosomes from the chromatin (31). Weiner *et al.* proposed that Pol II depletion relaxes chromatin, allowing nucleosomes to shift to more thermodynamically favorable sites. The noncanonical histone H2A variant H2A.Z is strongly enriched at gene promoters at +1 and -1 nucleosome positions (32, 33), likely as H2A.Z deposition increases nucleosome mobility and makes DNA more accessible to the transcriptional machinery (34). We thus performed ChIPmentation for H2A.Z during lytic HSV-1 infection to map +1 nucleosome positions. This showed that changes in chromatin accessibility in HSV-1 infection reflected a downstream shift of +1 nucleosomes. We observed similar changes in H2A.Z occupancy upon treatment with α-amanitin, a deadly toxin found in Amanita mushrooms, which inhibits transcription by preventing translocation of Pol II and triggering degradation of Rpb1, the largest subunit of Pol II (35). Our findings thus provide strong evidence that depletion of Pol II from host genes during HSV-1 infection leads to the downstream shift of +1 nucleosomes to more thermodynamically favorable sites.

Results

Widespread extension of chromatin-free regions downstream of transcription start sites

To investigate changes in chromatin accessibility within promoter regions during HSV-1 infection, we re-analyzed our recently published ATAC-seq data (29) for mock and wild-type (WT) HSV-1 infection of human fetal foreskin fibroblasts (HFF) as well as infection with null mutant viruses for the HSV-1 IE proteins ICP0, ICP22, and ICP27 as well as the tegument-delivered late protein *vhs*. Samples for mock, WT, Δ*vhs*, and ΔICP27 infection were collected at 8 h p.i., while samples for ΔICP0 and ΔICP22 infection were collected at 12 h p.i. To assess chromatin accessibility changes in promoter regions, we identified genomic regions with differential ATAC-seq read density using our recently published RegCFinder method (36). RegCFinder searches for genomic subregions for which the read distribution differs between two conditions (e.g., mock and WT infection). It is targeted towards regions of interest by specifying genomic windows as input (**Sup. Fig 1a** in supplemental material). For each input window, subregions that show differences in the distribution of reads within this window are identified. Statistical significance and log2 fold-changes for identified subregions are determined by DEXSeq (37). We defined the windows of interest as the ± 3kb promoter region around the TSS of 7,649 human genes. These TSSs were previously defined in our analysis of promoter-proximal Pol II pausing in HSV-1 WT infection (38) based on a re-analysis of published precision nuclear run-on sequencing (PRO-seq) and PROcap-seq (a variation of PRO-seq) data of flavopiridol-treated uninfected HFF. Inhibition of the CDK9 subunit of the positive transcription elongation factor b (P-TEFb) by flavopiridol arrests Pol II in a paused state at the TSS, allowing the mapping of Pol II initiation sites with both PRO- and PROcap-seq. Filtering for high-confidence TSSs confirmed by both data sets and within a maximum distance of 500 bp to the nearest annotated gene on the same strand identified the major TSS for 7,649 genes. Using these 6 kb promoter windows as input, we applied RegCFinder to our

ATAC-seq data for WT and all null-mutant infections in comparison to mock to identify promoter subregions that show differences in the distribution of open chromatin between infection and mock.

It should be noted that dOCRs can extend into downstream genes following very strong read-through transcription. We nevertheless did not filter promoter windows beforehand to exclude genes for which dOCRs extended from an upstream gene into their promoter window for two reasons: (i) Since dOCRs represent changes in chromatin accessibility, RegCFinder should, by design, detect them; (ii) The farther the distance from the upstream gene, the less pronounced are changes in chromatin accessibility in dOCR regions, making it difficult to distinguish them from background. Any approach we tested to filter promoter windows overlapping with dOCRs from upstream genes before the RegCFinder analysis either excluded too many or too few windows. Thus, we decided to first run RegCFinder and then investigate which detected changes were due to the induction of dOCRs from upstream genes.

RegCFinder identified 23,000-24,000 differential subregions (in the following also denoted as RegC, short for regions of change) for each virus infection compared to mock (WT: 23,357 RegC, ΔICP0: 23,532, ΔICP22: 24,640, ΔICP27: 23,523, Δ*vhs*: 23,716, **Sup. Fig 1b**). Between 29 and 37 % of these showed a statistically significant difference in read density within the corresponding 6 kb promoter windows upon infection (multiple testing adjusted p-value (adj. p.) ≤ 0.01, WT: 8,552 RegC, ΔICP0: 6,740, ΔICP22: 5,710, ΔICP27: 7,188, Δ*vhs*: 6,860). These represented between 3,129 (ΔICP22 infection) and 4,391 (WT) genes (**Sup. Fig 1c**). The lower fraction of statistically significant differential regions in ΔICP22 infection is likely due to the relatively low number of ATAC-seq reads mapping to the host genome compared to the other infections (~3.8-fold fewer, see **Sup. Table 1** in supplemental material). The main reason for

this was the high fraction of reads mapping to the HSV-1 genome in ΔICP22 infection (70-94%) compared to WT infection (~52%). We confirmed the higher proportion of viral reads in ΔICP22 infection compared to WT in an independent experiment for both 8 and 12 h infection with the ΔICP22 mutant and its parental WT strain F (WT-F, **Sup. Table 1**). Previously, McSwiggen *et al.* showed that the viral genome remains largely nucleosome-free and thus highly accessible (14). In contrast, the human genome is predominantly inaccessible except at promoters, gene bodies of highly expressed genes and enhancers (39). Accordingly, ATAC-seq coverage on the HSV-1 genome is essentially uniform (**Sup. Fig 2a,b**). The exception are the inverted repeat regions as we masked the terminal repeat copies from read alignment. Thus, the internal repeats exhibit approximately twice the coverage of the unique HSV-1 genome regions. The flat ATAC-seq coverage is observed for all null mutants, indicating that viral chromatin accessibility is not dependent on individual viral proteins.

Fig 1a visualizes the positions of identified RegC for the 4,981 promoter windows containing at least one statistically significant (adj. p. $\leq$ 0.01) differential region for at least one virus infection compared to mock. Log2 fold-changes for statistically significant regions are illustrated in **Sup. Fig 1d**. Here, blue indicates that the relative read density in that subregion is increased during virus infection compared to mock (infection>mock, denoted as i-RegC in the following), and red indicates that relative read density in that subregion is higher in mock compared to infection (mock>infection, denoted as m-RegC). Strikingly, our results showed a highly homogeneous picture for WT and null-mutant infections compared to mock, with changes in the same direction observed at approximately the same locations. To identify distinct patterns of changes in chromatin accessibility, we performed hierarchical clustering on the RegC location heatmap in **Fig 1a.** We selected a cutoff on the dendrogram to obtain 14 clusters identified by visual inspection. To visualize the average read density for each cluster, we performed metagene analyses of promoter windows combined with statistics on RegC locations

for each cluster (**Fig 1b,d,f,g, Sup. Fig 3**). Each 6 kb input window was divided into 101 bins of ~59 bp for metagene analyses. For each bin, ATAC-seq read counts were determined, normalized to sequencing depth, normalized to sum up to 1 for each input window, averaged across all windows in each cluster, and then averaged across replicates. In addition, we show which fraction of windows have an m- (red) or i-RegC (blue) for each bin. Based on the location heatmap and the metagene analyses, we identified three major patterns of changes in chromatin accessibility around promoters during WT infection: (I) ATAC-seq peaks at the TSS that shift and/or broaden into regions downstream of the TSS during HSV-1 infection (clusters 1, 2, 4, 5, 9, 10, 12, 14, a total of 3,472 genes, **Fig 1b, Sup. Fig 3a-g,** examples in **Fig 1c, Sup. Fig 4a-h**), (II) ATAC-seq peaks at the TSS that shift and/or broaden into regions upstream of the TSS (clusters 6, 7, 11, a total of 919 genes, **Fig 1d, Sup. Fig 3h,i,** examples in **Fig 1e, Sup. Fig 4i-k**), and (III) an increase in chromatin accessibility both up- and downstream of the TSS peak in infection compared to mock (cluster 13, 126 genes, **Fig 1f,** example in **Sup. Fig 4l**). Two clusters (3 and 8, a total of 464 genes) exhibited a combination of patterns I and II with an extension of the TSS peak in both up- and downstream direction (**Fig 1g, Sup. Fig 3j, Sup. Fig 4m,n**).

Most chromatin accessibility changes at promoters are independent of dOCR induction

First, we evaluated which observed changes were due to dOCRs extending into promoter regions. For this purpose, we determined the fraction of promoter windows in each cluster that overlapped with dOCRs of the 1,296 genes for which we previously showed consistent dOCR induction across different HSV-1 strains (29) (**Fig 2a**). dOCR regions in mock, WT and all null mutant infections (average across two replicates) were identified as previously described (21, 29) (see also Materials and Methods). It is important to note that some relatively short dOCRs are already observed in mock infection, but these substantially extend during HSV-1 infection in read-through regions. In this analysis, we also included ATAC-seq data obtained in the same

experiment for 8 h p.i. WT infection combined with PAA treatment (WT+PAA) to identify dOCR regions that are not as clearly detectable without PAA treatment but may still bias promoter analyses. WT (± PAA) and null mutant infections are ordered according to the overall extent of dOCR induction in **Fig 2a**. While no differences to mock infection were observed for ΔICP22 infection as expected, some clusters showed enrichment for dOCR overlaps upon WT(± PAA), ΔICP0, ΔICP27, and Δ*vhs* infection. This increased with the overall extent of dOCR induction in these viruses. Notably, since read-through in ΔICP27 infection is strongly reduced, dOCR induction is also reduced – but not abolished – as dOCRs require strong read-through transcription.

However, no cluster had more than 20% of promoter windows overlapping with dOCRs even with PAA treatment, and most had <5% overlap, indicating that dOCRs represent only a very small fraction of observed changes in chromatin accessibility around promoters. Most importantly, the extension of accessible chromatin regions at promoters was also observed upon infection with an ICP22-null mutant. Moreover, in contrast to dOCRs, which increase upon PAA treatment, the extension of open chromatin at promoters was strongly reduced by PAA treatment, indicating that it is not linked to dOCRs (**Fig 2b,c**, **Sup. Fig 5-18**). Even for clusters 12-14, which exhibited some enrichment for dOCRs, most differential regions were also detected in ΔICP22 infection (**Sup. Fig 16-18**) and thus not linked to dOCR induction. This was confirmed in the independent ATAC-seq experiment for 8 and 12 h infection with WT-F and its ΔICP22 mutant, which also confirmed the chromatin changes at promoters for a different HSV-1 strain (**Sup. Fig 19**). Notably, there was little change between 8 and 12 h infection in both WT-F and ΔICP22 infection. The absence of ICP0, ICP27, or *vhs* also had little impact on changes in chromatin accessibility around promoters (**Sup. Fig 5-18**) despite the longer duration of infection for the ICP0-null mutant and the different parental virus strains for the

null mutants. These observations suggest that there is an upper limit to the extent of chromatin accessibility changes at promoters in HSV-1 infection. We conclude that neither dOCR induction nor the activity of ICP0, ICP22, ICP27, or *vhs* alone explains most observed changes in chromatin accessibility at promoters in HSV-1 infection.

Extension of accessible chromatin regions around promoters is linked to transcription

To correlate observed changes to potential transcription factor binding, we next performed motif discovery for novel and known motifs in m- and i-RegC grouped either by cluster or pattern compared to the background of all promoter windows using HOMER (40) but found no enriched motifs. Functional enrichment analysis for Gene Ontology (GO) terms for each cluster also yielded no significant results, except for cluster 13 (pattern III). Cluster 13 was enriched for "mRNA splicing via the spliceosome" and related terms, however, this was mainly due to several snRNA genes in this cluster. Accordingly, cluster 13 was significantly enriched for snRNA genes (**Fig 2d**, adj. p. = 3×10^{-16}, see Materials and Methods). In contrast, pattern I clusters tended to contain high fractions of protein-coding genes but no or very few snRNAs. Manual investigation of these few snRNAs either showed no shift or an overlap with other protein-coding genes. Thus, most snRNAs either showed no significant changes or a relative increase in chromatin accessibility on both sides of the TSS (= pattern III in cluster 13) but no down- or upstream shifts in chromatin accessibility. Notably, snRNAs are transcribed by RNA Polymerase III (Pol III), not Pol II. However, snRNA loci and other genes not transcribed by Pol II (rRNAs, tRNAs) are repeated several times in the human genome. The quality of read mappings to these loci is thus insufficient to draw a definitive conclusion. Consequently, few snRNA (55), almost no rRNA (12) and no tRNA loci were included in our analysis. In summary, this analysis provides clear evidence for an extension of accessible chromatin at promoters of many genes transcribed by Pol II, particularly protein-coding genes.

Strikingly, cluster 7 (= the most pronounced case of pattern II) was significantly enriched for antisense transcripts (adj. p. = 5×10^{-15}). The other pattern II clusters 6 and 11, as well as cluster 8 (combined pattern I + II), also contained a relatively high fraction of antisense transcripts. Consistently, we found that cluster 7 (adj. p. = 3.8×10^{-14}) and to a lesser degree clusters 6 and 11 (adj. p. = 0.006 and 0.041, respectively) were enriched for bidirectional promoters, i.e., promoters containing an annotated gene start within 1 kb upstream of the TSS of the target gene (= the gene around whose TSS the promoter window was originally defined, **Fig 2e,** see **Fig 1e** for an example). We thus hypothesized that pattern II essentially represented the mirror image of pattern I (mirrored at a vertical axis through the TSS) with the "downstream" broadening/shift of chromatin accessibility occurring in antisense direction for bidirectional promoters. To confirm this, we calculated the sense-to-antisense transcription ratio in promoter windows for all genes. For this purpose, we analyzed RNA-seq data of chromatin-associated RNA, which depicts nascent transcription, for mock and 8 h p.i. WT HSV-1 infection from our recent study (19). We also included in this analysis the 2,668 genes (denoted as NA group) without significant chromatin accessibility changes that were excluded from **Fig 1a**. Interestingly, metagene analyses for these genes also revealed a slight broadening of the TSS peak into downstream regions. However, this was much less pronounced than for pattern I clusters and, therefore, not detected by RegCFinder (**Sup. Fig 20a**). To calculate the sense-to-antisense transcription ratio, promoter windows were divided into the regions down- and upstream of the TSS, and expression in chromatin-associated RNA was determined in sense direction for the downstream region (=DSR) and in antisense direction for the upstream region (=UAR). We then compared log2 ratios of DSR to UAR expression between clusters (**Fig 3a)**. Positive log2(DSR:UAR) ratios indicate that sense transcription downstream of the TSS is stronger than antisense transcription upstream of the TSS, and negative values indicate the opposite. Strikingly, all pattern II clusters had significantly lower DSR to UAR ratios than the remaining genes, while pattern I tended to have considerably higher ratios. While for pattern II

clusters 6 and 11 median values were still positive, values for cluster 7 with the most pronounced pattern II were commonly negative. Thus, transcription upstream of the TSS in antisense direction was the dominant mode of transcription for cluster 7 promoters, which means pattern II essentially just represents pattern I for the genes on the antisense strand.

Next, we investigated differences in gene expression changes in WT compared to mock infection between patterns (**Sup. Fig 20b**). This showed a significant difference (adj. $p < 0.001$) for cluster 1, for which gene expression was more strongly down-regulated than for all other genes. No other consistent trends were observed between the different patterns. In contrast, gene expression (quantified as FPKM = fragments per kilobase million mapped reads, **Fig 3b**) differed considerably between patterns. Most pattern I clusters as well as cluster 3 (combined pattern I+II) exhibited median expression levels above average in mock (**Fig 3b**), which was statistically significant for 5 clusters. In contrast, clusters 6, 7 (pattern II) and 8 (combined pattern I+II) exhibited relatively low expression values, with cluster 7 genes having significantly lower expression. The latter is consistent with these genes being less expressed than their antisense counterpart in these bidirectional promoters. Genes without significant chromatin accessibility changes (NA group) also showed significantly lower expression, albeit not as low as cluster 7. These differences between clusters/patterns were generally maintained in HSV-1 infection at lower overall expression levels, consistent with the absence of differences in gene expression fold-changes between most clusters. When stratifying analyzed genes into five equal-sized groups according to FPKM in chromatin-associated RNA in uninfected cells, we observed that the fraction of genes with at least one significant RegC increased with gene expression (**Fig. 3c**). Consistently, pattern II was significantly enriched among the lowliest expressed genes and depleted among genes with medium to high expression (**Fig. 3d**). In contrast, cluster 5, 9 and 10 with pattern I were enriched among the most highly expressed genes. Metagene analyses on the five gene expression groups confirm this trend, with the

325 downstream extension of chromatin accessibility increasing with gene expression (**Fig. 3e-g**).
326 Interestingly, pattern I cluster 1, which showed a significantly higher reduction in gene
327 expression during HSV-1 infection than the other analyzed genes, was most frequent among
328 genes with low to high expression but weakly significantly depleted among the most highly
329 expressed genes. In summary, these results indicate that HSV-1 infection extends the accessible
330 chromatin around promoters in the dominant direction of transcription for most host genes.
331 Here, highly expressed genes and moderately expressed genes with stronger transcription
332 reduction in HSV-1 infection are most strongly affected. Notably, although highly expressed
333 genes do not generally exhibit a stronger reduction of transcription *relative* to their original
334 expression levels in mock, the *absolute* drop in Pol II occupancy between mock and HSV-1
335 infection is more pronounced than for more lowly expressed genes. Thus, our results show
336 parallels to observations for yeast that Pol II depletion induces downstream shifts of +1
337 nucleosomes, which extends nucleosome-free regions and accordingly accessible chromatin at
338 promoters, most prominently for highly expressed genes (30). We thus hypothesized that the
339 loss of Pol II during HSV-1 infection causes the extension of accessible chromatin at promoters
340 during HSV-1 infection.
341
342 **Changes in chromatin accessibility manifest between 4 and 6 h of HSV-1 infection**
343 To investigate how early in infection changes in chromatin accessibility around promoters can
344 be detected, we ran RegCFinder on an ATAC-seq time-course for 1, 2, 4, 6, and 8 h p.i. WT
345 infection from our previous publication (21) (**Fig 4a,b**). While barely any significant
346 differential regions were identified at 1 and 2 h p.i., 1,963 significant regions in 1,249 promoter
347 windows were identified by 4 h p.i. and 7,558 significant regions in 3,989 promoter windows
348 by 8 h p.i. (**Fig 4b**). Log2 fold-changes for significant differential regions generally reflected
349 the patterns observed in the analysis for WT and null mutant infections (**Fig 4a**). The same
350 applied when we separately determined fold-changes and significance on the time-course data

for the differential regions determined for WT vs. mock from our initial analysis (**Sup. Fig 21**) and when we performed metagene analyses for the clusters from **Fig 1a (Fig 4d-f**, **Sup. Fig 22**, example genes in **Sup. Fig 23)**. Although the changes in chromatin accessibility at 8 h p.i. of the time-course were less pronounced than for the original WT vs. mock comparison at 8 h p.i., this nevertheless confirms the changes in chromatin accessibility around promoters in HSV-1 infection in an independent experiment. The lower fraction of viral reads at 8 h p.i. in the time-course experiment (~25%, **Sup. Table 1**) than both for the first WT experiment (~52%) and WT-F at 8 h (~61%) suggests a slower progression of infection in the time-course experiment, which likely explains why the effect was less pronounced. Again ATAC-seq read coverage on the HSV-1 genome remained essentially uniform throughout the time-course (**Sup. Fig 2c,d**).

The time-course analysis also reveals that these changes begin to manifest by 4 h p.i., with 1,249 genes already showing a significant change. Moreover, genes that showed an early effect (by 4 h p.i. or earlier) were significantly more highly expressed than genes that showed an effect by 6 h p.i. or later (**Fig 4c**). The lowest expression was observed for genes with a significant effect in the first WT experiment but not in the time-course. This again confirms the link between expression levels of a gene and the change in chromatin accessibility. Notably, we previously showed that transcriptional activity on host genes drops substantially in the first 4 h of infection to only 40% of transcription in uninfected cells, which was further halved until 8 h p.i. (19). Thus, the onset of the extension of accessible chromatin regions at promoters into downstream (for pattern I) and upstream (for pattern II) regions follows the onset of the drop in host transcription during infection. This provides further evidence for the hypothesis that loss of Pol II from host genes drives changes in chromatin accessibility at promoters. Since the slower progression of virus infection in the time-course experiment likely leads to a less pronounced loss of Pol II from host genes, this would explain the differences in effect between the time-course and the null mutant experiment.

377

378 **ICP4 knockout reduces, and combined expression of ICP22 and ICP27 induces the**
379 **extension of accessible chromatin**

380 We already showed above that PAA treatment substantially reduced the broadening and/or
381 downstream shift of accessible chromatin regions at promoters (**Fig 2b,c, Sup. Fig 5-18**). This
382 was also observed for 8 and 12 h p.i. PAA treatment of WT-F in our independent ATAC-seq
383 experiment with little differences between the two time-points (**Sup. Fig 24**). Control
384 experiments with mock ± PAA at 8 and 12 h showed no changes in chromatin accessibility at
385 promoters (**Sup. Fig 25**). We previously found that PAA alleviates the depletion of Pol II from
386 host genomes by inhibiting viral DNA replication (29). This provides further evidence that Pol
387 II depletion leads to the observed shifts in chromatin accessibility at promoters. The original
388 ATAC-seq experiment with null mutant infections also included infection with an ICP4 null
389 mutant at 8 h p.i., which had not been previously published. Since ICP4 facilitates recruitment
390 of Pol II to viral replication compartments (13, 14) and Pol II depletion from host promoters is
391 not observed in ΔICP4 infection (13, 41), we now also investigated changes in chromatin
392 accessibility in ΔICP4 infection. Strikingly, while we still observed the extension of accessible
393 chromatin in down- (**Fig 5a**, **Sup. Figs 26a-g,k,l**) or upstream direction (**Fig 5b**, **Sup Fig.**
394 **26h,i**), it was similarly reduced in ΔICP4 infection compared to WT infection as upon PAA
395 treatment. The fraction of viral reads in ATAC-seq experiments was comparably low at 3-4%
396 in both ΔICP4 and WT+PAA infection (**Sup. Table 1**). This is consistent with previous reports
397 that there is no viral DNA replication in the absence of ICP4 (42, 43). Moreover, high chromatin
398 accessibility in HSV-1 infection has been shown to be independent of ICP4 (13) and ATAC-
399 seq read coverage remained essentially uniform (**Sup. Fig 2a,b**), thus low ATAC-seq read
400 numbers in ΔICP4 infection are not due to reduced chromatin accessibility of viral genomes.

401

We next tested for enrichment in promoter regions of ICP4 binding sites on the human genome previously identified with ChIP-seq by Dremel *et al.* (41). This showed a significant enrichment (adj. p. = 1.1×10^{-5}) of ICP4 binding sites for genes with significant RegC (compared to the NA group of genes without any significant changes) and among these, a weakly significant enrichment in clusters 9 (pattern I, adj. p = 0.0046) and 11 (pattern II, adj. p.=0.0018). Notably, the only cluster with less frequent ICP4 binding than the NA group was cluster 13 (pattern III), which did not show shifts in chromatin accessibility. Thus, promoters of genes with significant shifts in chromatin accessibility are more frequently bound by ICP4 than genes without shifts, providing evidence that Pol II depletion from host genomes by ICP4 contributes to this phenomenon. However, since PAA does not inhibit ICP4 synthesis (27), the strongly reduced chromatin changes at promoters in WT + PAA infection indicate that in absence of viral DNA replication ICP4 activity is not sufficient to induce the full extent of chromatin accessibility changes.

Nevertheless, the question remains why reduced changes in chromatin accessibility are observed even in absence of ICP4. A possible explanation is that activities of other immediate-early proteins, which are still expressed and active in the absence of ICP4 (42-44), lead to an effective depletion of Pol II from host gene bodies. Two candidates for this are ICP27, which redirects Pol II transcription from gene bodies into intergenic regions by disrupting transcription termination (20), and ICP22, which inhibits host transcription elongation (15). In particular, HSV-1-induced disruption of transcription termination has a massive effect on remaining host transcriptional activity, with ~50% of newly transcribed RNA reads originating from intergenic regions (20). To investigate whether ICP27 and/or ICP22 expression alone can induce such changes, we re-analyzed our recently published Omni-ATAC-seq (a recent improvement of ATAC-seq (45)) data for telomerase-immortalized human foreskin fibroblasts (T-HFs) that express either ICP22 (T-HF-ICP22 cells) or ICP27 (T-HF-ICP27 cells) in isolation or

combination (T-HF-ICP22/ICP27 cells) upon doxycycline (dox) exposure. While we did not identify any significant RegC upon ICP22 expression alone, ICP27 expression alone led to significant changes for >2,000 genes (**Fig 5c**, **Sup. Fig 27a,b**). Combined ICP22 and ICP27 expression led to more pronounced results, with the same patterns observed as for WT infection but in an attenuated manner (**Fig 5c-e**, **Fig 27**). Thus, the combined activity of ICP22 and ICP27 is sufficient to induce a moderate extension of chromatin accessibility at promoters. RNA-seq analysis performed in parallel to ATAC-seq in the same experiment showed that ICP27 was not as strongly expressed in the T-HF-ICP22/ICP27 cells upon dox exposure as in the T-HF-ICP27 cells (~3-fold less). In contrast, ICP22 was much more strongly expressed upon dox exposure in T-HF-ICP22/ICP27 cells than in the T-HF-ICP22 cells (>5-fold more). This indicates that direct effects mediated by ICP22 are important for further extending chromatin accessibility at promoters rather than any indirect effects via enhancement of ICP27 expression and thus read-through.

Nevertheless, our results showed that the pronounced changes observed in WT infection require viral DNA replication but neither ICP22 nor ICP27 expression. Notably, although ICP27 is required for optimal viral DNA replication, knockout of ICP27 does not completely abolish viral DNA replication (46). Consistent with this, ~23% of ATAC-seq reads have a viral origin in ΔICP27 infection in contrast to only 3-4% in ΔICP4 and WT + PAA infection (**Sup. Table 1, Sup. Fig 2a,b**). This is comparable to the 8 h p.i. time-point in the time-course experiment. Furthermore, depletion of Pol II from host promoters is still observed in ΔICP22 and ΔICP27 infection (41). We conclude that the common feature between the different experimental conditions that exhibit promoter chromatin changes is the depletion of RNA Pol II from host promoter regions either by a generalized loss of Pol II or by Pol II translocation into downstream genomic regions upon disruption of transcription termination.

HSV-1 infection induces a downstream shift of +1 nucleosomes

The HSV-1-induced changes in chromatin accessibility at promoters indicate a broadening of the nucleosome-free region around promoters in either sense (pattern I) or antisense (pattern II) direction. The extension of nucleosome-free regions requires shifts in +1 or -1 nucleosome positions, and depletion of Pol II has been shown to induce a downstream shift of +1 nucleosomes in yeast (30). We, thus, performed ChIPmentation in HFF for mock and WT infection at 8 h p.i. with an antibody recognizing the C-terminal part of the non-canonical histone H2A.Z (n=3 replicates). H2A.Z is highly enriched at gene promoters at -1 and +1 nucleosome positions (32, 33) and is encoded by two genes, whose protein products H2A.Z.1 and H2A.Z.2 differ by only three amino acids. While their genomic occupancy patterns are similar, there are quantitative differences, with H2A.Z.1 being more abundant at active promoters than H2A.Z.2 (47). Since the C-terminal regions of H2A.Z.1 and H2A.Z.2 differ only by the last amino acid, the antibody used for ChIPmentation recognizes both isoforms. A metagene analysis of all analyzed promoter windows showed the expected distribution of H2A.Z occupancy in mock infection with two peaks on both sides of the TSS, corresponding to -1 and +1 nucleosome positions (**Sup. Fig 28a**).

Application of RegCFinder to H2A.Z ChIPmentation data generally identified a relative increase in H2A.Z downstream of the TSS during infection and a relative decrease upstream of the TSS for pattern I and combined pattern I+II clusters (**Fig 6a**, log2 fold-changes for differential regions shown in **Sup. Fig 28b**, examples in **Fig 6b**, **Sup. Fig 29a-g,k,l**). Metagene analyses showed that this reflected a downstream shift and broadening of +1 nucleosome peaks as well as a relative increase of +1 nucleosome peaks compared to -1 nucleosome peaks (**Fig 6d**, **Sup. Fig 28c-i,m,n**). In contrast, cluster 7 (most pronounced pattern II cluster) showed the opposite trend with relative increases in H2A.Z upstream of the TSS and decreases downstream of the TSS (example in **Fig 6c**). Consistent with pattern II representing pattern I in antisense

direction, the metagene analysis showed an upstream shift and broadening of the -1 nucleosome peak and a relative increase of the -1 nucleosome peak compared to the +1 nucleosome peak (**Fig 6e**). However, this was only observed for a few genes in cluster 6 and cluster 11 (**Fig 6a**, **Sup. Fig 28j,k**, examples in **Sup. Fig 29h,i**), which fits with the observation that pattern II in the ATAC-seq data was also much less pronounced for these clusters. In contrast to pattern I and II, pattern III in the ATAC-seq data (cluster 13) was not associated with distribution changes in H2A.Z (**Sup. Fig 25l**, example in **Sup. Fig 29j**), thus it is not shaped by shifts in +1 and/or -1 nucleosome positioning. This is consistent with pattern III being at least partly associated with dOCR induction for upstream genes, which we previously linked to impaired nucleosome repositioning following Pol II transcription downstream of genes (29).

In summary, our data confirm that the broadening of chromatin accessibility around promoters results from down- and upstream shifts of +1 or -1 nucleosomes, respectively. Here, the shift direction depends on whether sense or antisense transcription represents the dominant direction of transcription at this promoter. Notably, a recent study in yeast proposed that -1 nucleosomes should be considered +1 nucleosomes for antisense transcription (48). This suggests that the loss of Pol II from host promoters during HSV-1 infection induces a shift of +1 and -1 nucleosome positions similar to what was previously observed for yeast. To confirm that inhibition of transcription alone can lead to shifts of +1 and -1 nucleosomes in human cells, we re-analyzed published H2A.Z ChIP-seq data of HCT116 cells with and without α-amanitin treatment from the recent study by Lashgari *et al*. (49). They showed that α-amanitin treatment leads to reduced Pol II levels at gene TSSs and increased incorporation of H2A.Z at TSSs of transcribed genes. Here, Pol II ChIP and RT-qPCR of control genes showed that α-amanitin treatment reduced Pol II signal at the TSS to 10-20% of untreated cells and transcriptional activity to less than 40%. This is not much lower than what has previously been reported for HSV-1 infection: Abrisch *et al.* found that Pol II occupancy on gene bodies upon 4 h HSV-1

infection was reduced on average to around 20% (9), and we previously estimated that transcriptional activity was reduced to 20-40% between 4 and 8 h p.i. (19). Notably, in contrast to DRB and flavopiridol, which inhibit transcription via arresting Pol II in a paused state at the promoter, α-amanitin induces a loss of Pol II from the host genome, thus making it a better model of HSV-1-induced Pol II depletion.

Lashgari *et al*. previously only analyzed changes in absolute H2A.Z levels at promoters and did not focus on changes in +1 and -1 nucleosome positioning. Metagene and RegCFinder analyses of the H2A.Z ChIP-seq data indeed showed a similar trend upon α-amanitin treatment as during HSV-1 infection (**Fig 6f,g, Sup. Fig 30,** examples are shown in **Sup. Fig 31**). Pattern I and pattern I+II clusters showed a massive broadening of the +1 nucleosome peak into downstream regions, even more pronounced than what is observed upon HSV-1 infection. In addition, they also exhibited a (less pronounced) broadening of -1 nucleosome peaks. Similarly, pattern II clusters showed a strong broadening of both -1 and +1 nucleosome peaks into upstream regions. In the case of cluster 7 (strongest pattern II), the broadening was much more pronounced for the -1 nucleosome (i.e., the +1 nucleosome in antisense direction) than for the +1 nucleosome (i.e., the -1 nucleosome in antisense direction). This analysis confirms that the depletion of Pol II from promoters leads to a downstream shift of +1 nucleosomes in human cells. In summary, our results demonstrate that the shift of +1 and -1 nucleosomes in the direction of transcription in HSV-1 infection is a consequence of Poll II depletion from host chromatin.

Discussion

Previously, we showed that HSV-1 infection disrupts chromatin architecture downstream of genes with strong read-through transcription. In this study, we reveal that chromatin architecture is also substantially altered at host gene promoters during HSV-1 infection. Here, 57% of genes showed a statistically significant change in the distribution of chromatin

accessibility at promoters in WT HSV-1 infection, and metagene analyses indicated similar but less pronounced changes even for genes without statistically significant changes. In essence, we identified three types of changes: Most genes showed a shift and/or broadening of accessible chromatin into regions downstream of the TSS (pattern I). In contrast, ~900 genes showed the opposite trend with a shift and/or broadening of accessible chromatin into regions upstream of the TSS (pattern II). A further ~460 genes exhibited a combination of patterns I and II, with the broadening of accessible chromatin in the downstream direction being more pronounced than in the upstream direction. Only a small set of genes (126 genes) showed relative increases in chromatin accessibility up- and downstream of the TSS, which could partly be attributed to overlaps with dOCRs from upstream genes and partly to an enrichment for short non-coding RNAs not transcribed by Pol II, for which promoter windows were longer than the actual gene. We thus did not further investigate this pattern.

As patterns I and II were still observed in the absence of ICP22, which is necessary for dOCR induction, these are not artifacts of dOCRs extending into downstream genes. On the contrary, PAA treatment, which increases dOCR induction, substantially reduced – though not completely abolished – patterns I and II. Furthermore, knockout of neither ICP0, ICP22, ICP27, nor *vhs* substantially affected patterns I and II. In contrast, ICP4 knockout, similar to PAA treatment, also alleviated the down- and upstream broadening of chromatin accessibility. Both PAA treatment and ICP4 knockout largely abolish viral DNA replication and consequently alleviate the depletion of Pol II from host genes, leading us to hypothesize that this HSV-1-induced loss of Pol II causes the widespread extension of open chromatin at host gene promoters. This hypothesis was further supported by the observation that combined dox-induced expression of ICP22 and ICP27 (and to a lesser degree dox-induced ICP27 expression alone) led to some broadening of chromatin accessibility at promoters despite knockout of neither of these proteins affecting the HSV-1-induced changes. Knockout of ICP22 or even

ICP27 does not sufficiently abolish viral DNA replication and thus has only a minor effect on the depletion of Pol II from host genomes and, consequently, on chromatin accessibility changes. In contrast, in the absence of viral DNA replication, ICP22 and ICP27 (and potentially other viral factors) sufficiently reduce Pol II levels on gene bodies to induce an alleviated form of those chromatin accessibility changes. Notably, we determined that the more pronounced changes upon co-expression of ICP27 with ICP22 compared to expression of ICP27 alone are likely due to the direct effects of ICP22 on transcription. ICP22 interacts with P-TEFb, several transcriptional kinases, as well as elongation factors such as the FACT complex to inhibit transcription elongation of cellular genes (15). Moreover, both the relatively late onset of chromatin accessibility changes between 4 and 8 h p.i. in our time-course ATAC-seq analysis and the generally reduced levels of changes in the time-course, which had a lower fraction of viral reads at 8 h p.i., confirm that significant depletion of Pol II is necessary to observe substantial effects on chromatin accessibility. On the other hand, the little differences observed between 8 and 12 h of infection and the null mutants of ICP0, ICP22, ICP27 and *vhs* also suggest that there is an upper limit to the extent of chromatin accessibility changes at promoters in HSV-1 infection.

Analysis of transcriptional activity using RNA-seq of chromatin-associated RNA revealed that pattern I and II are indeed linked to transcription, with pattern I associated with more highly expressed genes and pattern II associated with bidirectional promoters with strong antisense transcription on the opposite strand. In the case of cluster 7, for which pattern II was most pronounced, antisense transcription was much higher than sense transcription. This was not due to the widespread induction of antisense transcription in HSV-1 infection, which we previously reported on (50), but these genes already exhibited strong transcription in antisense direction in uninfected cells originating from bidirectional promoters. Consistently, cluster 7 was enriched for genes annotated as antisense, and more than a third of these promoters contained an

583 annotated gene starting on the opposite strand within 1 kb upstream of the TSS of the target
584 gene. Thus, pattern II is essentially just pattern I for the genes on the opposite strand in
585 bidirectional promoters.
586
587 Since the broadening of chromatin accessibility in promoter regions suggested an extension of
588 nucleosome-free regions at promoters, we mapped +1 and -1 nucleosome positions by
589 ChIPmentation of the H2A.Z histone variant enriched at these nucleosomes. Analysis of H2A.Z
590 occupancy indeed showed a downstream shift of +1 nucleosomes for pattern I genes in HSV-1
591 infection and an upstream shift of -1 nucleosomes for genes with the most pronounced pattern
592 II (cluster 7). This further confirmed that pattern II just represents pattern I for genes on the
593 antisense strand. It is also consistent with recent results from yeast by Bagchi *et al.* that -1
594 H2A.Z-containing nucleosomes should be considered as +1 nucleosomes for antisense
595 transcription (48). Downstream shifts of the +1 nucleosome to more thermodynamically
596 favorable sites have previously been reported upon Pol II degradation in yeast by Weiner *et al.*
597 (30), in particular for highly expressed genes. Re-analysis of published H2A.Z ChIP-seq data
598 for α-amanitin-induced depletion of Pol II from gene promoters indicates that depletion of Pol
599 II alone is sufficient to induce downstream shifts of +1 nucleosomes in human cells. Moreover,
600 for genes with dominant antisense transcription (cluster 7), it resulted in upstream shifts of -1
601 nucleosomes. Although 24 h α-amanitin treatment – as used for ChIP-seq – potentially leads to
602 other effects beyond losing genome-bound Pol II that may explain chromatin changes, rapid
603 degradation of Pol II with an inducible degron system has also been shown to increase
604 chromatin dynamics similar to α-amanitin (51). The broadening of H2A.Z peaks in both HSV-
605 1 infection and upon α-amanitin treatment, respectively, indicate a less precise positioning of
606 corresponding nucleosomes. This was more pronounced upon α-amanitin treatment and also
607 observed for -1 nucleosomes, suggesting that loss of Pol II upon α-amanitin treatment is more
608 pronounced than in HSV-1 infection.

609

610 One open question remaining concerns the biological significance of the observed changes in
611 nucleosome positioning. Our analysis of gene expression changes did not suggest an effect on
612 differential gene expression, rather the opposite with strong down-regulation leading to the
613 same effect for less strongly expressed genes (cluster 1) as for highly expressed genes that are
614 not more down-regulated than other genes. However, the downstream shifts in +1 nucleosomes
615 may provide an explanation for the downstream shift of Pol II pausing we recently reported for
616 HSV-1 infection (38). The +1 nucleosome was shown to play a role in promoter-proximal
617 Pol II pausing between the promoter and the +1 nucleosome (52). In particular, the +1
618 nucleosome represents a 2nd barrier to Pol II pause release independent of the main pausing
619 factor negative elongation factor (NELF). Upon NELF depletion, Pol II stops at a 2nd pausing
620 region around the +1 nucleosomal dyad-associated region (53). We previously observed that
621 Pol II pausing in HSV-1 infection is shifted to more downstream and less well-positioned sites
622 for the majority of genes (38), consistent with +1 nucleosome positioning also appearing less
623 well-positioned upon HSV-1 infection in our H2A.Z ChIPmentation data. Alternatively, the
624 shifts in +1 nucleosomes might also be linked to the mobilization of H1, H2 (including H2A.Z),
625 and H4 histones during HSV-1 infection (54-56), which has been proposed to serve as a source
626 of histones for viral chromatin assembly. Nevertheless, even if the changes in nucleosome
627 positioning upon Pol II depletion are of no further functional consequence for HSV-1 infection,
628 they are highly relevant for any functional genomics studies on chromatin architecture in
629 HSV-1 infection or other conditions that deplete Pol II from the genome. If not properly taken
630 into account, the broadening of accessible chromatin at promoters may be mistaken, e.g., for
631 differential transcription factor binding, which could lead to wrong conclusions. Together with
632 read-through transcription and dOCRs extending into downstream genes, this represents one
633 more example of how HSV-1 infection confounds standard functional genomics analyses. In

any case, our study highlights that HSV-1 infection impacts chromatin architecture at promoters independently of the widespread changes downstream of genes mediated by ICP22.

Materials and Methods

Previously published sequencing data analyzed in this study

ATAC-seq for mock and WT HSV-1 infection (strain 17, 8 h p.i.), WT infection with 8 h PAA treatment, infection with ICP0-null mutant (ΔICP0, strain 17, (57), 12 h p.i.), ICP22-null mutant (ΔICP22, R325, strain F, (58), 12 h p.i.), ICP27-null mutant (ΔICP27, strain KOS, (59), 8 h p.i.) and *vhs*-null mutant (Δ*vhs*, strain 17, (60)) in of HFF were taken from our recent publication (29) (n=2 apart from ΔICP22 infection with n=4, GEO accession: GSE185234). This experiment also included infection with an ICP4-null mutant (ΔICP4, n12, strain KOS, (42), 8 h p.i.) which had not previously been published. ATAC-seq data for mock and WT HSV-1 infection of HFF at 1, 2, 4, 6 and 8 h p.i. WT infection (n=2 replicates, GEO accession: GSE100611) and chromatin-associated RNA-seq data for mock and 8 h p.i. WT infection (GEO accession: GSE100576) were taken from our previous publication (21). H2A.Z ChIP-seq data of untreated and α-amanitin-treated HCT116 cells were taken from the study by Lashgari *et al.* (49) (n=2, GEO accession: GSE101427).

ChIPmentation, library preparation and sequencing

HFF were purchased from ECACC and cultured in Dulbecco's Modified Eagle Medium (DMEM, ThermoFisher #41966052) supplemented with 10% (v/v) Fetal Bovine Serum (FBS, Biochrom #S0115), 1× MEM Non-Essential Amino Acids (ThermoFisher #11140050) and 1% penicillin/streptomycin. Two days prior to infection, two million HFF cells were seeded in 15 cm dishes. On the day of infection, cells had expanded to ~80% confluency. Cells were infected with the respective viruses as described in the results section (n=3). At 8 p.i., cells were fixed for 10 minutes at room temperature by adding 1% formaldehyde (final) directly to the medium.

Cells were scraped in 1mL of ice-cold 1x PBS containing protease inhibitor cocktail (1x) (Roche #11836153001) with an additional 1mM phenylmethylsulfonyl fluoride (PMSF). Cells were pelleted at 1500 rpm for 20 min at 4 °C. Supernatant was aspirated and cell pellets were frozen in liquid N_2.

Cell pellets were resuspended in 1.5 mL 0.25% [w/v] SDS sonication buffer (10 mM Tris pH=8.0, 0.25% [w/v] SDS, 2 mM EDTA) with 1x protease inhibitors and 1 mM additional PMSF and incubated on ice for 10 minutes. Cells were sonicated in fifteen 1 minute intervals, 25% amplitude, with Branson Ultrasonics SonifierTM S-450 until most fragments were in the range of 200-700 bp as determined by agarose gel electrophoresis. Two million cells used for the preparation of the ChIPmentation libraries were diluted 1:1.5 with equilibration buffer (10 mM Tris, 233 mM NaCl, 1.66% [v/v] Triton X-100, 0.166% [w/v] sodium deoxycholate, 1 mM EDTA, protease inhibitors) and spun at 14,000x g for 10 minutes at 4 °C to pellet insoluble material. Supernatant was transferred to a new 1.5 mL screw-cap tube and topped up with RIPA-LS (10 mM Tris-HCl pH 8.0, 140 mM NaCl, 1 mM EDTA pH 8.0, 0.1% [w/v] SDS, 0.1% [w/v] sodium deoxycholate, 1% [v/v] Triton X-100, protease inhibitors) to 200 µL. Input and gel samples were preserved. Lysates were incubated with 1µg/IP of anti-H2A.Z antibody (Diagenode, #C15410201) on a rotator overnight at 4 °C.

Dependent on the added amount of antibody, the amount of Protein A magnetic beads (ThermoFisher Scientific #10001D) was adjusted (e.g., for 1-2 µg of antibody/IP = 15 µL of beads) and blocked overnight with 0.1% [w/v] bovine serum albumin in RIPA buffer. On the following day, beads were added to the IP samples for 2 h on a rotator at 4 °C to capture the antibody-bound fragments. The immunoprecipitated chromatin was subsequently washed twice with 150 µL each of ice-cold buffers RIPA-LS, RIPA-HS (10 mM Tris-HCl pH 8.0, 50 0mM NaCl, 1 mM EDTA pH 8.0, 0.1% [w/v] SDS, 0.1% [v/v] sodium deoxycholate, 1% [v/v] Triton

X-100), RIPA-LiCl (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA pH 8.0, 0.5% [w/v] sodium deoxycholate, 0.5% [v/v] Nonidet P-40) and 10 mM Tris pH 8.0 containing protease inhibitors. Beads were washed once more with ice-cold 10 mM Tris pH 8.0 lacking inhibitors and transferred into new tubes.

Beads were resuspended in 25 µL of the tagmentation reaction mix (Nextera DNA Sample Prep Kit, Illumina) containing 5 µL of 5x Tagmentation buffer, 1 µL of Tagment DNA enzyme, topped up with H_2O to the final volume and incubated at 37 °C for 10 minutes in a thermocycler. Beads were mixed after 5 minutes by gentle pipetting. To inactivate the Tn5 enzyme, 150 µL of ice-cold RIPA-LS was added to the tagmentation reaction. Beads were washed twice with 150 µL of RIPA-LS and 1x Tris-EDTA and subjected to de-crosslinking by adding 100 µL ChIPmentation elution buffer (160 mM NaCl, 40 µg/mL Rnase A (Sigma-Aldrich #R4642), 1x Tris-EDTA (Sigma #T9285) and incubating for 1h at 37 °C followed by overnight shaking at 65 °C. The next day, 4 mM EDTA and 200 µg/mL Proteinase K (Roche, #03115828001) were added, and samples incubated for another 2h at 45 °C with 1000 rpm shaking. Supernatant was transferred into a new tube and another 100 µL of ChIPmentation elution buffer was added for another hour at 45 °C with 1000 rpm shaking. DNA was isolated with MinElute PCR Purification Kit (Qiagen #28004) and eluted in 21 µL of H_2O.

DNA for the final library was prepared with 25 µL NEBNext Ultra II Q5 Master Mix, 3.75 µL IDT custom primer i5_n_x (10 µM); 3.75 µL IDT custom primer i7_n_x (10 µM); 3.75 µL H_2O and 13.75 µL ChIPmentation DNA. The Cq value obtained from the library quantification, rounded up to the nearest integer plus one additional cycle, was used to amplify the rest of the ChIPmentation DNA. Library qualities were verified by High Sensitivity DNA Analysis on the Bioanalyzer 2100 (Agilent) before performing sequencing on NextSeq 500 (paired-end 35bp

711 reads) at the Core Unit Systemmedizin, Würzburg, Germany. All samples were sequenced at
712 equimolar ratios.
713
714 *Read alignment*
715 The read alignment pipeline was implemented and run in the workflow management system
716 Watchdog (61, 62) as already previously described (38). Public sequencing data were
717 downloaded from SRA using the sratoolkit version 2.10.8. Sequencing reads were aligned
718 against the human genome (GRCh37/hg19), the HSV-1 genome (Human herpesvirus 1 strain
719 17, GenBank accession code: JN555585) and human rRNA sequences using ContextMap2
720 version 2.7.9 (63) (using BWA as short read aligner (64) and allowing a maximum indel size
721 of 3 and at most 5 mismatches). For the two repeat regions in the HSV-1 genome, only one
722 copy was retained each, excluding nucleotides 1–9,213 and 145,590–152,222 from the
723 alignment. SAM output files of ContextMap2 were converted to BAM files using samtools
724 (65). Read coverage in bedGraph format was calculated from BAM files using BEDTools (66).
725 Subregions of promoter windows (TSS ± 3 kb) with differential read coverage in ATAC-seq
726 and H2A.Z ChIPmentation/-seq data were determined with RegCFinder (36). For this purpose,
727 RegCFinder was applied to ATAC-seq data for all pairwise comparisons of mock to WT and
728 null mutant infections as well as mock to each timepoint of infection for the time-course data
729 and to H2A.Z ChIPmentation and ChIP-seq data for the comparison of mock and WT infection
730 as well as untreated and α-amanitin-treated HCT116 cells.
731

Quality control

Statistics on numbers of mapped reads and reads mapped to human and HSV-1 genomes were determined with samtools (65). Promoter/Transcript body (PT) scores were determined with ATACseqQC (66). For peak calling, BAM files with mapped reads were converted to BED format using BEDvTools (67) and peaks were determined from these BED files using F-Seq with default parameters (68). The fraction of reads in peaks (FRiP) was calculated with featureCounts (69) using identified peaks as annotation. Annotation of peaks relative to genes was performed using ChIPseeker (70).

Data plotting and statistical analysis

All figures were created in R and all statistical analyses were performed in R (67). Read coverage plots were created using the R Bioconductor package Gviz (68).

Metagene and clustering analysis

Metagene analyses were performed as previously described (69) using the R program developed for this previous publication (available with the Watchdog *binGenome* module in the Watchdog module repository (https://github.com/watchdog-wms/watchdog-wms-modules/)). For promoter region analyses, the regions −3 kb to +3 kb of the TSS were divided into 101 equal-sized bins for each gene. For each bin, the average coverage per genome position was calculated and bin read coverages were then normalized by dividing by the total sum of all bins. Metagene curves for each replicate were created by averaging results for corresponding bins across all genes in a cluster/group and then averaged across replicates. For hierarchical clustering analysis, RegCFinder profiles of differential regions were calculated for each promoter window and comparison by setting each position in an m-RegC to 1, in an i-RegC to -1 and all other positions to 0. RegCFinder profiles of each promoter window for each comparison were concatenated into one row in the matrix. Hierarchical clustering of the resulting matrix was then

758 performed using the hclust function in R according to Euclidean distances and Ward's
759 clustering criterion.
760
761 *Analysis of downstream open chromatin regions (dOCRs)*
762 Open chromatin regions (OCRs) were determined from ATAC-seq data by first converting
763 BAM files with mapped reads to BED format using BEDTools (70) and then determining
764 enriched regions from these BED files using F-Seq with default parameters (71). dOCRs for
765 individual genes were calculated from OCRs as previously described (21, 29). In brief, dOCRs
766 are determined for genes by first assigning all OCRs overlapping with the 10 kb downstream
767 of a gene to this gene. Second, OCRs starting at most 5 kb downstream of the so far most
768 downstream OCR of a gene are also assigned to this gene. In both steps, individual OCRs can
769 be assigned to multiple genes. The second step is iterated until no more OCRs can be assigned.
770 The dOCR of a gene is then defined as the region from the gene 3'end to the end of the most
771 downstream OCR assigned to this gene.
772
773 *Gene expression analysis*
774 Number of fragments (=read pairs) per gene were determined from mapped paired-end RNA-
775 seq reads in a strand-specific manner using featureCounts (72) and gene annotations from
776 Ensembl (version 87 for GRCh37). For genes, all fragments overlapping exonic regions on the
777 corresponding strand by ≥ 25bp were counted for the corresponding gene. Fold-changes in gene
778 expression and statistical significance of changes were determined using DESeq2 (73) and p-
779 values were adjusted for multiple testing using the method by Benjamini and Hochberg (74).
780 Gene expression was quantified in terms of fragments per kilobase of exons per million mapped
781 reads (FPKM). Only reads mapped to the human genome were counted for the total number of
782 mapped reads for FPKM calculation.

783

784 *Motif discovery and enrichment analysis*

785 Motif discovery was performed for each cluster separately for the i- and m-RegC regions using
786 findMotifsGenome.pl script of the Homer suite (40), with our 7,649 input promoter windows
787 as background. For this purpose, we used the hg19 annotation provided by Homer and
788 automated trimming of input windows was disabled. Significant motifs were identified at a
789 q-value (=False Discovery Rate (FDR) calculated with the Benjamini-Hochberg method (74))
790 cutoff of 0.01. Over-representation of Gene Ontology (GO) terms was performed separately for
791 each cluster using the g:Profiler webserver (75) and the R package gprofiler2 (76), which
792 provides an R interface to the webserver. As background gene list, the genes corresponding to
793 our 7,649 input promoter windows were provided. P-values were corrected for multiple testing
794 using the Benjamini-Hochberg method (74) and significantly over-represented GO terms were
795 identified at an multiple testing adjusted p-value cutoff of 0.001. Enrichment of gene types
796 (obtained from the Ensembl annotation (version 87 for GRCh37)) within clusters was
797 determined using one-sided Fisher's exact tests (with alternative = greater). Enrichment (odds-
798 ratio >1) or depletion (odds-ratio < 1) of ICP4 binding sites from the study of Dremel *et al.* (41)
799 within clusters as well as enrichment and depletion of clusters within gene groups with different
800 expression levels were determined with two-sided Fisher's exact tests. P-values were always
801 corrected for multiple testing using the Benjamini-Hochberg method.

802

803 ## Acknowledgements

804 This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research
805 Foundation, www.dfg.de) in the framework of the Research Unit FOR5200 DEEP-DV
806 (443644894) project FR 2938/11-1 to C.C.F. and by the European Research Council (ERC,
807 https://erc.europa.eu) project ERC-2021-CoG 101041177 – DecipherHSV to L.D.

References

1. Roizman B KD, RJ W. Herpes Simplex Viruses. In: Knipe DM HP, editor. Fields Virology. 5th. Philadelphia: Lippincott Williams & Wilkins; 2007. p. 2501-601.

2. Pellet PE RB. Herpesviridae. In: Knipe DM HP, editor. Fields Virology. 6th. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2013. p. 1802-2128.

3. Ablashi D, Agut H, Alvarez-Lafuente R, Clark DA, Dewhurst S, DiLuca D, et al. Classification of HHV-6A and HHV-6B as distinct viruses. Arch Virol. 2014;159(5):863-70.

4. Korr G, Thamm M, Czogiel I, Poethko-Mueller C, Bremer V, Jansen K. 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 Infect Dis. 2017;17(1):471.

5. Whitley RJ, Roizman B. Herpes simplex virus infections. Lancet. 2001;357(9267):1513-8.

6. Whisnant AW, Jurges CS, Hennig T, Wyler E, Prusty B, Rutkowski AJ, et al. Integrative functional genomics decodes herpes simplex virus 1. Nat Commun. 2020;11(1):2038.

7. Spencer CA, Dahmus ME, Rice SA. Repression of host RNA polymerase II transcription by herpes simplex virus type 1. J Virol. 1997;71(3):2031-40.

8. Rice SA, Davido DJ. HSV-1 ICP22: hijacking host nuclear functions to enhance viral infection. Future Microbiol. 2013;8(3):311-21.

9. Abrisch RG, Eidem TM, Yakovchuk P, Kugel JF, Goodrich JA. Infection by Herpes Simplex Virus 1 Causes Near-Complete Loss of RNA Polymerase II Occupancy on the Host Cell Genome. J Virol. 2015;90(5):2503-13.

10. Dembowski JA, DeLuca NA. Selective recruitment of nuclear factors to productively replicating herpes simplex virus genomes. PLoS Pathog. 2015;11(5):e1004939.

832 11. Birkenheuer CH, Danko CG, Baines JD. Herpes Simplex Virus 1 Dramatically Alters
833 Loading and Positioning of RNA Polymerase II on Host Genes Early in Infection. J Virol.
834 2018;92(8).
835 12. Rivas HG, Schmaling SK, Gaglia MM. Shutoff of Host Gene Expression in Influenza
836 A Virus and Herpesviruses: Similar Mechanisms and Common Themes. Viruses.
837 2016;8(4):102.
838 13. Dremel SE, DeLuca NA. Herpes simplex viral nucleoprotein creates a competitive
839 transcriptional environment facilitating robust viral transcription and host shut off. Elife.
840 2019;8.
841 14. McSwiggen DT, Hansen AS, Teves SS, Marie-Nelly H, Hao Y, Heckert AB, et al.
842 Evidence for DNA-mediated nuclear compartmentalization distinct from phase separation.
843 eLife. 2019;8:e47098.
844 15. Isa NF, Bensaude O, Aziz NC, Murphy S. HSV-1 ICP22 Is a Selective Viral Repressor
845 of Cellular RNA Polymerase II-Mediated Transcription Elongation. Vaccines.
846 2021;9(10):1054.
847 16. Kwong AD, Frenkel N. Herpes simplex virus-infected cells contain a function(s) that
848 destabilizes both host and viral mRNAs. Proc Natl Acad Sci U S A. 1987;84(7):1926-30.
849 17. Oroskar AA, Read GS. Control of mRNA stability by the virion host shutoff function
850 of herpes simplex virus. J Virol. 1989;63(5):1897-906.
851 18. Feng P, Everly DN, Jr., Read GS. mRNA decay during herpesvirus infections:
852 interaction between a putative viral nuclease and a cellular translation factor. J Virol.
853 2001;75(21):10272-80.
854 19. Friedel CC, Whisnant AW, Djakovic L, Rutkowski AJ, Friedl MS, Kluge M, et al.
855 Dissecting Herpes Simplex Virus 1-Induced Host Shutoff at the RNA Level. J Virol.
856 2021;95(3).

857 20. Rutkowski AJ, Erhard F, L'Hernault A, Bonfert T, Schilhabel M, Crump C, et al.
858 Widespread disruption of host transcription termination in HSV-1 infection. Nat Commun.
859 2015;6:7126.
860 21. Hennig T, Michalski M, Rutkowski AJ, Djakovic L, Whisnant AW, Friedl MS, et al.
861 HSV-1-induced disruption of transcription termination resembles a cellular stress response but
862 selectively increases chromatin accessibility downstream of genes. PLoS Pathog.
863 2018;14(3):e1006954.
864 22. Wang X, Hennig T, Whisnant AW, Erhard F, Prusty BK, Friedel CC, et al. Herpes
865 simplex virus blocks host transcription termination via the bimodal activities of ICP27. Nat
866 Commun. 2020;11(1):293.
867 23. Vilborg A, Passarelli MC, Yario TA, Tycowski KT, Steitz JA. Widespread Inducible
868 Transcription Downstream of Human Genes. Mol Cell. 2015;59(3):449-61.
869 24. Rosa-Mercado NA, Zimmer JT, Apostolidi M, Rinehart J, Simon MD, Steitz JA.
870 Hyperosmotic stress alters the RNA polymerase II interactome and induces readthrough
871 transcription despite widespread transcriptional repression. Mol Cell. 2021;81(3):502-13 e4.
872 25. Hadar S, Meller A, Saida N, Shalgi R. Stress-induced transcriptional readthrough into
873 neighboring genes is linked to intron retention. iScience. 2022;25(12):105543.
874 26. Wu N, Watkins SC, Schaffer PA, DeLuca NA. Prolonged gene expression and cell
875 survival after infection by a herpes simplex virus mutant defective in the immediate-early genes
876 encoding ICP4, ICP27, and ICP22. Journal of Virology. 1996;70(9):6358-69.
877 27. Honess RW, Watson DH. Herpes simplex virus resistance and sensitivity to
878 phosphonoacetic acid. Journal of Virology. 1977;21(2):584-600.
879 28. Becker Y, Asher Y, Cohen Y, Weinberg-Zahlering E, Shlomai J. Phosphonoacetic
880 Acid-Resistant Mutants of Herpes Simplex Virus: Effect of Phosphonoacetic Acid on Virus
881 Replication and In Vitro Deoxyribonucleic Acid Synthesis in Isolated Nuclei. Antimicrobial
882 Agents and Chemotherapy. 1977;11(5):919-22.

883 29. Djakovic L, Hennig T, Reinisch K, Milic A, Whisnant AW, Wolf K, et al. The HSV-1
884 ICP22 protein selectively impairs histone repositioning upon Pol II transcription downstream
885 of genes. Nat Commun. 2023;14(1):4591.
886 30. Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N. High-resolution nucleosome
887 mapping reveals transcription-dependent promoter packaging. Genome Res. 2010;20(1):90-
888 100.
889 31. Schwabish MA SK. Evidence for eviction and rapid deposition of histones upon
890 transcriptional elongation by RNA polymerase II. Mol Cell Biol. 2004;24(23):10111-7.
891 32. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution
892 profiling of histone methylations in the human genome. Cell. 2007;129(4):823-37.
893 33. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, et al. Histone
894 variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell.
895 2005;123(2):233-48.
896 34. Li S, Wei T, Panchenko AR. Histone variant H2A.Z modulates nucleosome dynamics
897 to promote DNA accessibility. Nature Communications. 2023;14(1):769.
898 35. Bensaude O. Inhibiting eukaryotic transcription. Which compound to choose? How to
899 evaluate its activity? Transcription. 2011;2(3):103-8.
900 36. Weiss E, Friedel CC. RegCFinder: targeted discovery of genomic subregions with
901 differential read density. Bioinform Adv. 2023;3(1):vbad085.
902 37. Anders S RA, Huber W. Detecting differential usage of exons from RNA-seq data.
903 Genome Research. 2012;22(10):2008-17.
904 38. Weiss E, Hennig T, Grassl P, Djakovic L, Whisnant AW, Jurges CS, et al. HSV-1
905 Infection Induces a Downstream Shift of Promoter-Proximal Pausing for Host Genes. J Virol.
906 2023;97(5):e0038123.
907 39. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The
908 accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75-82.

909 40. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations
910 of lineage-determining transcription factors prime cis-regulatory elements required for
911 macrophage and B cell identities. Mol Cell. 2010;38(4):576-89.
912 41. Dremel SE, Sivrich FL, Tucker JM, Glaunsinger BA, DeLuca NA. Manipulation of
913 RNA polymerase III by Herpes Simplex Virus-1. Nature Communications. 2022;13(1):623.
914 42. DeLuca NA, Schaffer PA. Physical and functional domains of the herpes simplex virus
915 transcriptional regulatory protein ICP4. Journal of Virology. 1988;62(3):732-43.
916 43. DeLuca NA, McCarthy AM, Schaffer PA. Isolation and characterization of deletion
917 mutants of herpes simplex virus type 1 in the gene encoding immediate-early regulatory protein
918 ICP4. Journal of Virology. 1985;56(2):558-70.
919 44. Dixon RA, Schaffer PA. Fine-structure mapping and functional analysis of temperature-
920 sensitive mutants in the gene encoding the herpes simplex virus type 1 immediate early protein
921 VP175. Journal of Virology. 1980;36(1):189-203.
922 45. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna
923 S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of
924 frozen tissues. Nature Methods. 2017;14(10):959-62.
925 46. Uprichard SL, Knipe DM. Herpes simplex ICP27 mutant viruses exhibit reduced
926 expression of specific DNA replication genes. Journal of Virology. 1996;70(3):1969-80.
927 47. Greenberg RS, Long HK, Swigut T, Wysocka J. Single Amino Acid Change Underlies
928 Distinct Roles of H2A.Z Subtypes in Human Syndrome. Cell. 2019;178(6):1421-36 e24.
929 48. Bagchi DN, Battenhouse AM, Park D, Iyer VR. The histone variant H2A.Z in yeast is
930 almost exclusively incorporated into the +1 nucleosome in the direction of transcription.
931 Nucleic Acids Research. 2020;48(1):157-70.
932 49. Lashgari A, Millau J-F, Jacques P-É, Gaudreau L. Global inhibition of transcription
933 causes an increase in histone H2A.Z incorporation within gene bodies. Nucleic Acids Research.
934 2017;45(22):12715-22.

935 50. Wyler E, Menegatti J, Franke V, Kocks C, Boltengagen A, Hennig T, et al. Widespread
936 activation of antisense transcription of the host genome during herpes simplex virus 1 infection.
937 Genome Biology. 2017;18(1):209.
938 51. Nagashima R, Hibino K, Ashwin SS, Babokhov M, Fujishiro S, Imai R, et al. Single
939 nucleosome imaging reveals loose genome chromatin networks via active RNA polymerase II.
940 Journal of Cell Biology. 2019;218(5):1511-30.
941 52. Jimeno-González S, Ceballos-Chávez M, Reyes JC. A positioned +1 nucleosome
942 enhances promoter-proximal pausing. Nucleic Acids Research. 2015;43(6):3068-78.
943 53. Aoi Y, Smith ER, Shah AP, Rendleman EJ, Marshall SA, Woodfin AR, et al. NELF
944 Regulates a Promoter-Proximal Step Distinct from RNA Pol II Pause-Release. Molecular Cell.
945 2020;78(2):261-74.e5.
946 54. Conn KL HM, Schang LM. Linker histones are mobilized during infection with herpes
947 simplex virus type 1. J Virol. 2008;82:8629–46.
948 55. Conn KL HM, Schang LM. Core histones H2B and H4 are mobilized during infection
949 with herpes simplex virus 1. J Virol. 2011;85:13234–52.
950 56. Conn KL, Schang LM. Chromatin Dynamics during Lytic Infection with Herpes
951 Simplex Virus 1. Viruses. 2013;5(7):1758-86.
952 57. Stow ND, Stow EC. Isolation and Characterization of a Herpes Simplex Virus Type 1
953 Mutant Containing a Deletion within the Gene Encoding the Immediate Early Polypeptide
954 Vmw110. Journal of General Virology. 1986;67(12):2571-85.
955 58. Post LE, Roizman B. A generalized technique for deletion of specific genes in large
956 genomes: a gene 22 of herpes simplex virus 1 is not essential for growth. Cell. 1981;25(1):227-
957 32.
958 59. Smith IL, Hardwicke MA, Sandri-Goldin RM. Evidence that the herpes simplex virus
959 immediate early protein ICP27 acts post-transcriptionally during infection to regulate gene
960 expression. Virology. 1992;186(1):74-86.

961 60. Fenwick ML, Everett RD. Inactivation of the Shutoff Gene (UL41) of Herpes Simplex
962 Virus Types 1 and 2. Journal of General Virology. 1990;71(12):2961-7.
963 61. Kluge M, Friedel CC. Watchdog - a workflow management system for the distributed
964 analysis of large-scale experimental data. BMC Bioinformatics. 2018;19(1):97.
965 62. Kluge M, Friedl MS, Menzel AL, Friedel CC. Watchdog 2.0: New developments for
966 reusability, reproducibility, and workflow execution. Gigascience. 2020;9(6).
967 63. Bonfert T, Kirner E, Csaba G, Zimmer R, Friedel CC. ContextMap 2: fast and accurate
968 context-based RNA-seq mapping. BMC Bioinformatics. 2015;16:122.
969 64. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
970 transform. Bioinformatics. 2009;25(14):1754-60.
971 65. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years
972 of SAMtools and BCFtools. Gigascience. 2021;10(2).
973 66. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
974 features. Bioinformatics. 2010;26(6):841-2.
975 67. R Core Team. R: A Language and Environment for Statistical Computing. Vienna,
976 Austria2022.
977 68. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. Methods
978 Mol Biol. 2016;1418:335-51.
979 69. Chirackal Manavalan AP, Pilarova K, Kluge M, Bartholomeeusen K, Rajecky M,
980 Oppelt J, et al. CDK12 controls G1/S progression by regulating RNAPII processivity at core
981 DNA replication genes. EMBO Rep. 2019;20(9):e47592.
982 70. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
983 features. Bioinformatics. 2010;26(6):841-2.
984 71. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for
985 high-throughput sequence tags. Bioinformatics. 2008;24(21):2537-8.

986 72. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for
987 assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.
988 73. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
989 RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
990 74. Benjamini Y HY. Controlling the False Discovery Rate: A Practical and Powerful
991 Approach to Multiple Testing. Journal of the Royal Statistical Society Series B
992 (Methodological). 1995;57(1):289-300.
993 75. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web
994 server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic
995 Acids Res. 2019;47(W1):W191-W8.
996 76. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2 -- an R package for
997 gene list functional enrichment analysis and namespace conversion toolset g:Profiler.
998 F1000Res. 2020;9.

999

1000 ## Figures

1001 **Fig 1: HSV-1 infection impacts chromatin accessibility around promoters for the majority**
1002 **of host genes.**

1003 **(a)** Heatmap illustrating the location of differential regions (m- and i-RegC) identified by
1004 RegCFinder on the ATAC-seq data for WT, ΔICP0, ΔICP22, ΔICP27, and Δ*vhs* infection
1005 compared to mock infection. Results are shown for 4,981 (out of 7,649) promoter windows
1006 containing at least one statistically significant (adj. p. ≤ 0.01) differential region for at least one
1007 virus infection compared to mock. Red and blue colors represent m-RegC (mock>infection)
1008 and i-RegC (infection>mock) locations, respectively, within promoter windows. Here, each
1009 heatmap row shows results for the same input window in different comparisons of virus

infection to mock. Colored rectangles on top of the heatmap indicate which virus infection was compared to mock. Black vertical lines in the center of each part of the heatmap indicate the position of the TSS. Hierarchical clustering was performed according to Euclidean distances and Ward's clustering criterion, and the cutoff on the hierarchical clustering dendrogram was selected to obtain 14 clusters (marked by colored rectangles between the dendrogram and heatmap and numbered from top to bottom as indicated). Log2 fold-changes for differential regions are shown in **Sup. Fig 1d. (b)** Metagene plot showing the average ATAC-seq profile around the TSS ± 3 kb in mock (red) and WT (blue) infection for cluster 5, an example for pattern I. For a description of metagene plots, see Materials and Methods. The colored bands below the metagene curves in each panel indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position. **(c)** Read coverage in a ±3.6 kb window around the TSS in ATAC-seq data of mock (red) and WT (blue) infection, for example, gene FBXO28 with pattern I. Read coverage was normalized to the total number of mapped reads for each sample and averaged between replicates. A short vertical line below the read coverage track for mock infection indicates the TSS. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns, and the direction of transcription is indicated by arrowheads. Below the read coverage track for WT infection, m-RegC (red bars) and i-RegC (blue bars) are shown for the comparison of WT vs. mock infection. **(d)** Metagene plot as in **(b)** for cluster 7, the most pronounced case of pattern II. **(e)** Read coverage plot as in **(c)** for example gene ARHGAP1 with pattern II. Here, the promoter window also contains the TSS of the ZNF408 gene on the opposite strand, and the accessible chromatin region is extended upstream of the ARHGAP1 TSS, i.e., downstream of the ZNF408 TSS. **(f,g)** Metagene plots as in **(b)** for **(f)** cluster 13, which exhibits pattern III, and **(g)** cluster 8, one of two clusters exhibiting the combined I+II pattern.

Fig 2: Changes in chromatin accessibility at promoters are mostly independent of dOCR induction.

(a) Bar plots showing the percentage of promoter regions for each cluster overlapping with dOCRs downstream of the 1,296 genes for which we previously showed consistent dOCR induction across different HSV-1 strains (29). dOCR regions in mock, WT (± PAA treatment), and null mutant infection were calculated as previously described (29) (see Materials and Methods for further details), and the overlaps of promoter windows to dOCR regions originating from an upstream gene were determined for each cluster. WT (± PAA treatment) and null mutant infection are ordered according to the overall extent of dOCR induction. Only a few clusters (12-14) showed some enrichment for dOCRs from upstream genes, which was abolished in ΔICP22 infection. **(b-c)** Metagene curves showing average ATAC-seq profiles in promoter windows for mock (red), WT (blue), WT+PAA (green), and ΔICP22 (orange) infection for clusters 5 **(b)** and 7 **(c)**. (**d**) Barplot showing the percentage of genes in each cluster annotated as either protein-coding, long intervening/intergenic noncoding RNAs (lincRNA), snRNA, antisense, or others in Ensembl. The significance of enrichment of each gene type in each cluster was determined using a one-sided Fisher's exact test (with alternative = greater), and p-values were corrected for multiple testing using the Benjamini-Hochberg method. Adj. p. < 0.05 are indicated in the corresponding field of the barplot. (**e**) The percentage of bidirectional promoters in each cluster was calculated as the percentage of promoter windows containing a protein-coding, lincRNA, or antisense gene (according to Ensembl annotation) on the opposite strand to the target gene starting within 1 kb upstream of the TSS of the target gene. Enrichment and significance analysis and multiple testing correction were performed as for (**d**) and adj. p. < 0.05 are indicated on top of bars.

Fig 3: Changes in chromatin accessibility at promoters are linked to transcription.

(a,b) Boxplots showing the distribution of **(a)** log2(DSR:UAR) ratios and **(b)** gene expression (FPKM) values in chromatin-associated RNA in mock (red) and 8 h p.i. WT (blue) HSV-1 infection for all clusters (grouped by pattern, cluster numbers are shown below each boxplot) and remaining genes without significant chromatin accessibility changes (NA group). DSR was calculated as the expression in chromatin-associated RNA for the region downstream of the TSS in the sense direction of the target gene and UAR as the expression in the upstream region in the antisense direction. P-values for Wilcoxon rank sum tests comparing values in mock infection for each cluster against all other analyzed genes are indicated below cluster numbers and were corrected for multiple testing using the Bonferroni method. **(c)** Barplot showing the percentage of genes with very low, low, medium, high and very high gene expression with at least one significant differential region (RegC). Here, gene expression cutoffs for the five groups were determined such that each group contains the same number of genes and are indicated below the labels on the x-axis. Multiple testing corrected p-values for two-sided Fisher's exact tests comparing the fraction of genes with RegC between subsequent gene expression groups are indicated on top of bars. **(d)** Percentage of genes within each cluster among those genes with at least one significant RegC for the five gene expression groups from **(c)**. Significance of enrichment or depletion for each cluster in each gene group was determined with two-sided Fisher's exact tests and multiple testing correction was performed with the method by Benjamini and Hochberg. Adj. p. < 0.05 are shown and are underlined in case the cluster is enriched and not underlined if it is depleted. Clustered are ordered and colored according to pattern (red-yellow: pattern I, green: pattern I + II, blue: pattern II, magenta: pattern III). **(e-g)** Metagene curves showing average ATAC-seq profiles in promoter windows for mock (red) and WT infection (blue) for genes with very low, medium, and very high expression from **(b)**.

Fig. 4: Changes in chromatin accessibility begin to manifest between 4 and 8 h p.i. depending on gene expression

(a) Heatmap showing the log2 fold-changes determined with DEXSeq on the ATAC-seq time-course data (1, 2, 4, 6 and 8 h p.i. WT infection compared to mock infection) for statistically significant differential regions (m- and i-RegC) identified by RegCFinder. For comparison, log2 fold-changes for the WT vs. mock comparison from **Fig 1a** are also shown. Each row represents the results for one of the input windows included in **Fig 1a**. Colored rectangles on top indicate the time-point of infection or whether the original WT vs. mock comparison is shown. Statistically significant differential regions are colored according to the log2 fold-change determined by DEXSeq. Here, the color scale is continuous between -1 and 1, and all log2 fold-changes >1 are colored the same red, and all log2 fold-changes <1 the same blue. Promoter windows are ordered as in **Fig 1a**, and clusters are annotated by colored and numbered rectangles on the left. **(b)** Number of statistically significant differential regions (m- and i-RegC) and number of genes with at least one statistically significant differential region identified for each time-point of infection in the ATAC-seq time-course data. **(c)** Boxplot showing the distribution of gene expression (FPKM) values in chromatin-associated RNA for mock and 8 h p.i. WT infection for genes with significant changes in chromatin accessibility in promoter windows (i) at 4 h p.i. or earlier, (ii) at 6 h p.i. but not yet at 4 h p.i., (iii) at 8 h p.i. but not yet at 6 h p.i. and (iv) in the analysis shown in **Fig 1a**, but not yet at 8 h p.i. in the time-course ATAC-seq experiment. P-values for Wilcoxon rank sum tests comparing FPKM values in mock infection between subsequent groups are also indicated. **(d-f)** Metagene curves of ATAC-seq profiles in mock infection (red) and all time-points of infection (blue shades) from the time-course experiment, for example, clusters 5 (pattern I, **d**), 7 (pattern II, **e**) and 13 (pattern III, **f**). The colored bands below the metagene curves indicate the percentage of genes having an i- or m-RegC (blue or red, respectively) at that position in the comparison of 8 h p.i. to mock

1109 infection from the time-course experiment. Metagene plots for all clusters are shown in **Sup.**
1110 **Fig 22**.

1111 **Fig 5: Chromatin accessibility changes upon ICP4 knockout and dox-induced combined**
1112 **ICP22 and ICP27 expression**

1113 **(a-b)** Metagene curves showing average ATAC-seq profiles in promoter windows for mock
1114 (red), WT (blue), WT+PAA (green) and ΔICP4 (violet) infection for clusters 5 **(a)** and 7 **(b)**.
1115 Metagene plots for all other clusters are shown in **Sup. Fig 26**. **(c)** Heatmap showing the
1116 location of differential regions (m- and i-RegC) identified by RegCFinder for T-HF-ICP27 cells
1117 and T-HF-ICP22/ICP27 cells upon dox exposure. Each row represents the results for one of the
1118 input windows included in **Fig 1a**. Promoter windows are ordered as in **Fig 1a** and clusters are
1119 annotated as colored and numbered rectangles on the left. Log2 fold-changes for differential
1120 regions are shown in **Sup. Fig 27b**. **(d-e)** Metagene curves showing average ATAC-seq profiles
1121 in promoter windows for T-HF-ICP22/ICP27 cells with (blue) and without (red) dox exposure
1122 for clusters 5 **(d)** and 7 **(e)**. The colored bands below the metagene curves indicate the
1123 percentage of genes having an m-RegC (red, decreased upon dox exposure) or i-RegC (blue,
1124 increased upon dox exposure) or at that position. Metagene plots for all clusters are shown in
1125 **Sup. Fig 27**.

1126

1127 **Fig 6: HSV-1 infection and depletion of Pol II from human genes lead to a downstream**
1128 **shift of H2A.Z-containing +1 nucleosomes.**

1129 **(a)** Heatmap showing the location for differential regions (m- and i-RegC) identified in the WT
1130 vs. mock comparison on the ATAC-seq (left half, identical to left-most part of **Fig 1a**) and
1131 H2A.Z ChIPmentation data (right half) for the promoter windows included in **Fig 1a.**
1132 Differential regions with mock>infection (m-RegC) are marked in red, and differential regions
1133 with infection>mock (i-RegC) in blue. The order of promoter windows is the same as in **Fig**

1a, and clusters are indicated by colored and numbered rectangles on the left. Log2 fold-changes for differential regions are shown in **Sup. Fig 28b**. **(b,c)** Read coverage in a ±3.6 kb window around the TSS in H2A.Z ChIPmentation data for mock (red) and WT (blue) infection, for example, genes with **(b)** pattern I (FBXO28) and **(c)** pattern II (ARHGAP1). For descriptions of read coverage plots, see caption to **Fig 1**. **(d-g)** Metagene plots of H2A.Z profiles for mock and WT infection **(d,e)** and untreated and α-amanitin-treated HCT116 cells **(f,g)** for clusters 5 (pattern I, **d,f**) and 7 (pattern II, **e,g**). Metagene plots for other clusters can be found in **Sup. Fig 28** and **30**, respectively. The colored bands below the metagene curves in each panel indicate the percentage of genes having m- or i-RegC at that position. For **(f,g)**, m-RegC (green) are differential regions with relative read coverage higher in untreated cells and i-RegC (violet) differential regions with relative read coverage higher upon α-amanitin-treatment.

Fig 1

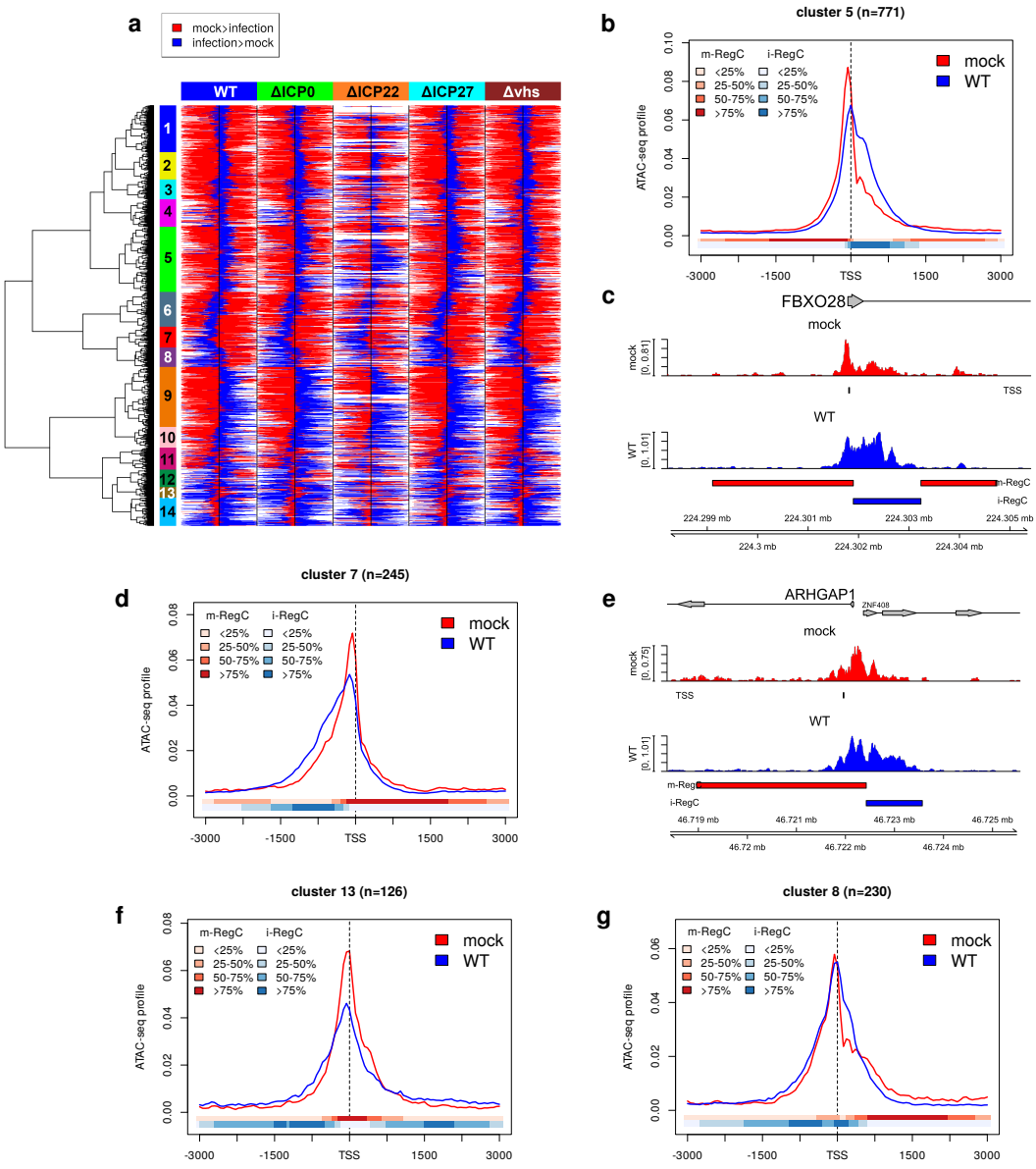

a
mock>infection
infection>mock
WT
ΔICP0
ΔICP22
ΔICP27
Δvhs
1
2
3
4
5
6
7
8
9
10
11
12
13
14
b
cluster 5 (n=771)
m-RegC
i-RegC
<25%
25-50%
50-75%
>75%
mock
WT
ATAC-seq profile
0.00
0.02
0.04
0.06
0.08
0.10
-3000
-1500
TSS
1500
3000
c
FBXO28
mock
[0, 0.81]
TSS
WT
[0, 1.01]
m-RegC
i-RegC
224.299 mb
224.301 mb
224.303 mb
224.305 mb
224.3 mb
224.302 mb
224.304 mb
d
cluster 7 (n=245)
e
ARHGAP1
ZNF408
[0, 0.75]
46.719 mb
46.721 mb
46.723 mb
46.725 mb
46.72 mb
46.722 mb
46.724 mb
f
cluster 13 (n=126)
g
cluster 8 (n=230)

Fig 2

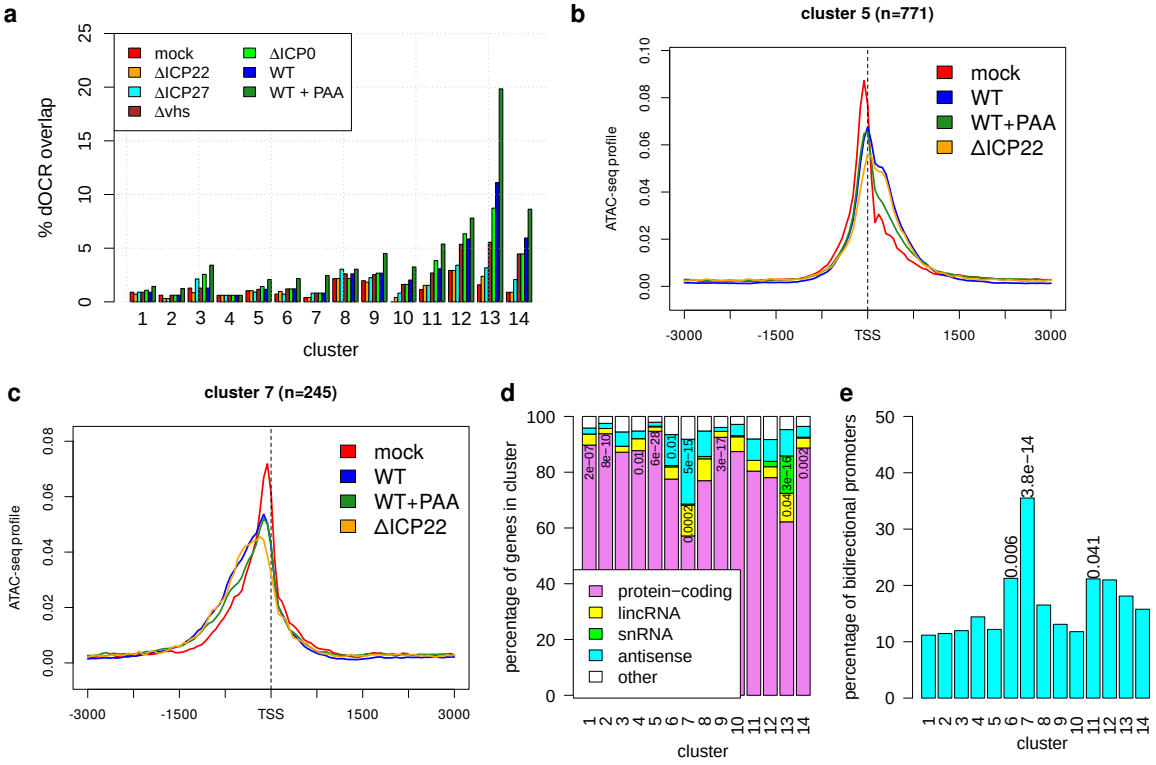

a
% dOCR overlap
mock
ΔICP22
ΔICP27
Δvhs
ΔICP0
WT
WT + PAA
cluster
b
cluster 5 (n=771)
ATAC-seq profile
mock
WT
WT+PAA
ΔICP22
-3000
-1500
TSS
1500
3000
c
cluster 7 (n=245)
d
percentage of genes in cluster
protein−coding
lincRNA
snRNA
antisense
other
e
percentage of bidirectional promoters
0.006
3.8e−14
0.041

Fig 3

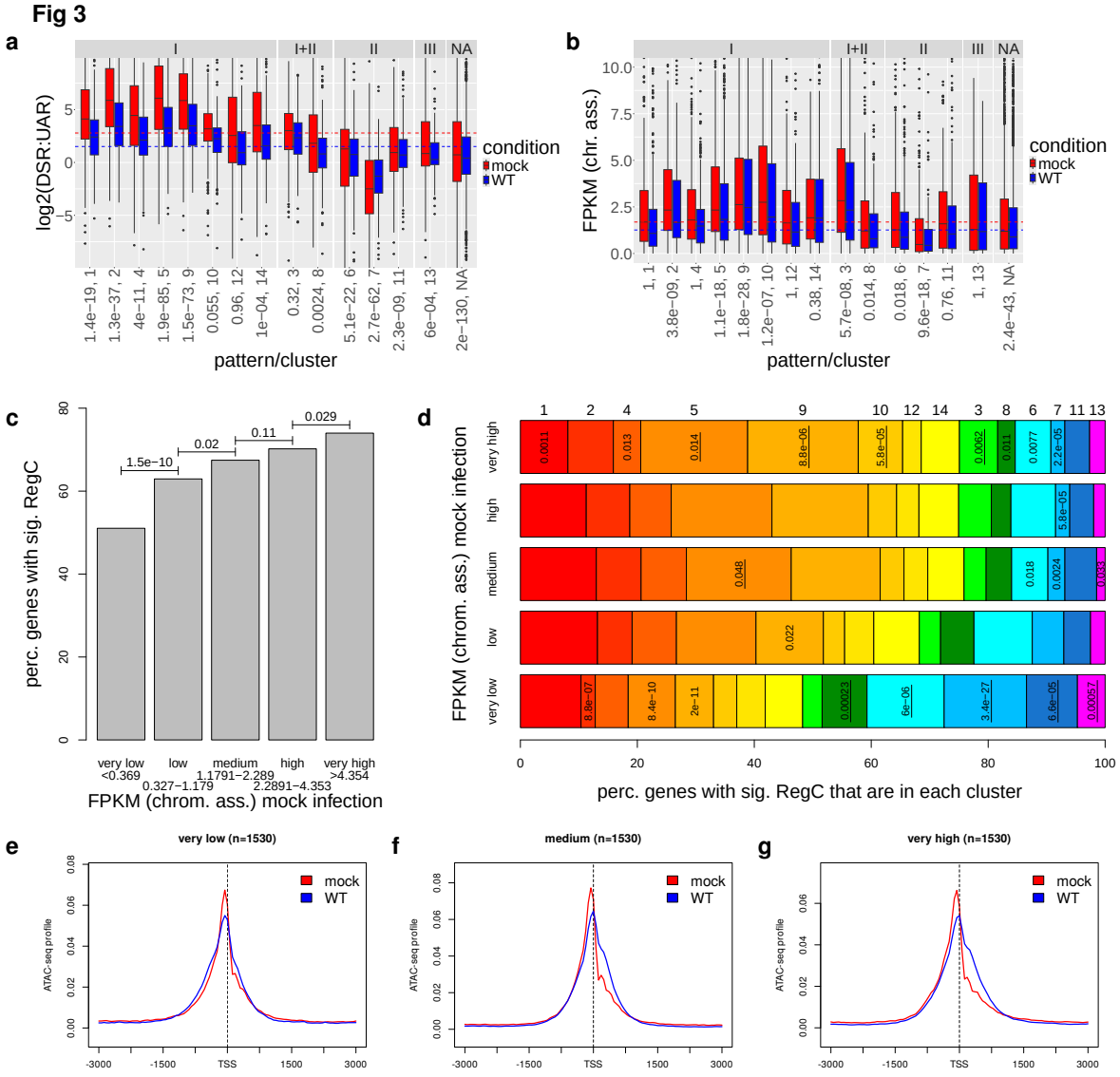

a
log2(DSR:UAR)
I
I+II
II
III
NA
condition
mock
WT
pattern/cluster
b
FPKM (chr. ass.)
pattern/cluster
c
perc. genes with sig. RegC
1.5e−10
0.02
0.11
0.029
very low
low
medium
high
very high
FPKM (chrom. ass.) mock infection
d
FPKM (chrom. ass.) mock infection
very high
high
medium
low
very low
perc. genes with sig. RegC that are in each cluster
e
very low (n=1530)
f
medium (n=1530)
g
very high (n=1530)
ATAC-seq profile
TSS

Fig 4

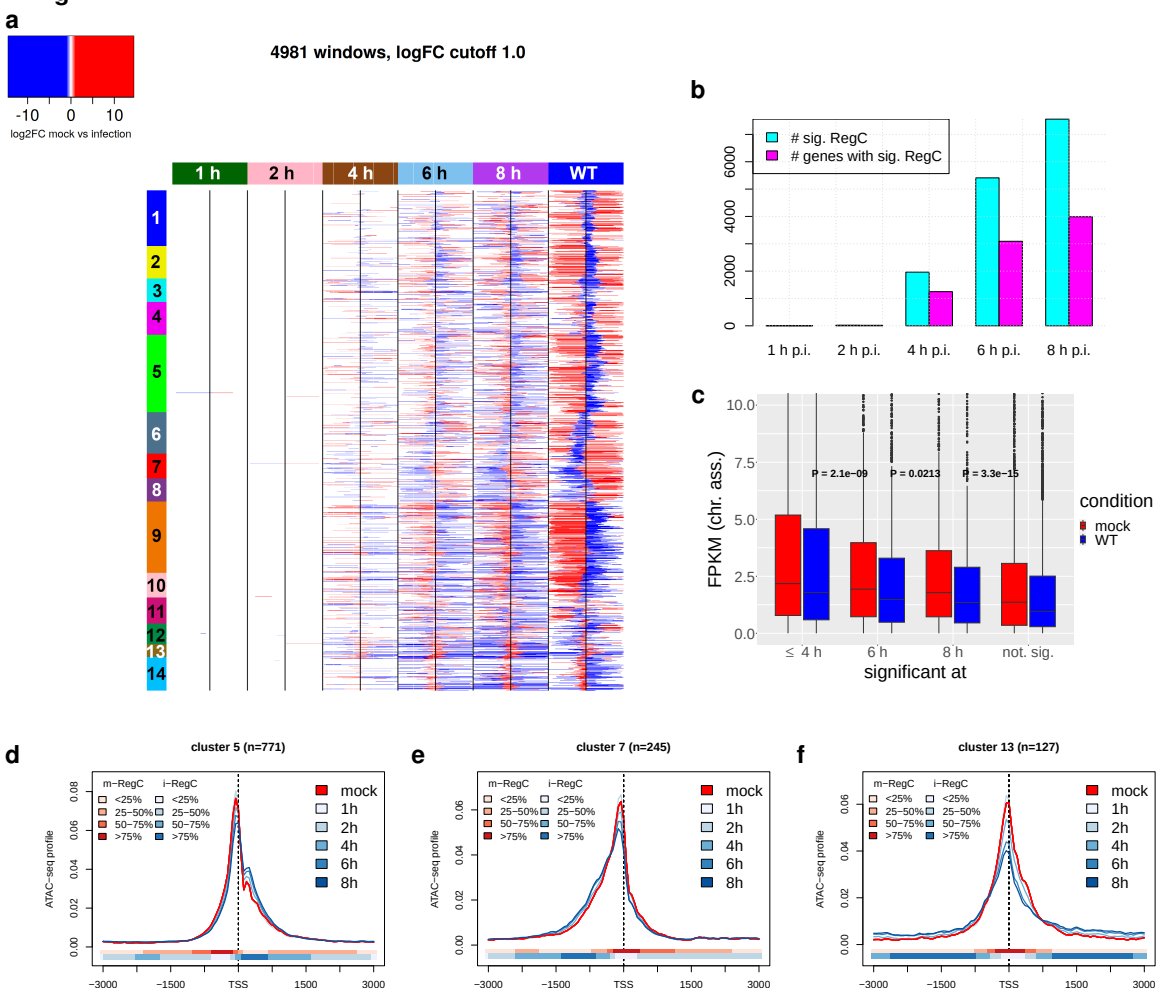

a
4981 windows, logFC cutoff 1.0
-10 0 10
log2FC mock vs infection
1 h
2 h
4 h
6 h
8 h
WT
1
2
3
4
5
6
7
8
9
10
11
12
13
14
b
sig. RegC
genes with sig. RegC
0
2000
4000
6000
1 h p.i.
2 h p.i.
4 h p.i.
6 h p.i.
8 h p.i.
c
FPKM (chr. ass.)
10.0
7.5
5.0
2.5
0.0
P = 2.1e−09
P = 0.0213
P = 3.3e−15
≤ 4 h
6 h
8 h
not. sig.
significant at
condition
mock
WT
d
cluster 5 (n=771)
e
cluster 7 (n=245)
f
cluster 13 (n=127)
m−RegC
i−RegC
<25%
25−50%
50−75%
>75%
mock
1h
2h
4h
6h
8h
ATAC−seq profile
−3000
−1500
TSS
1500
3000

Fig 5

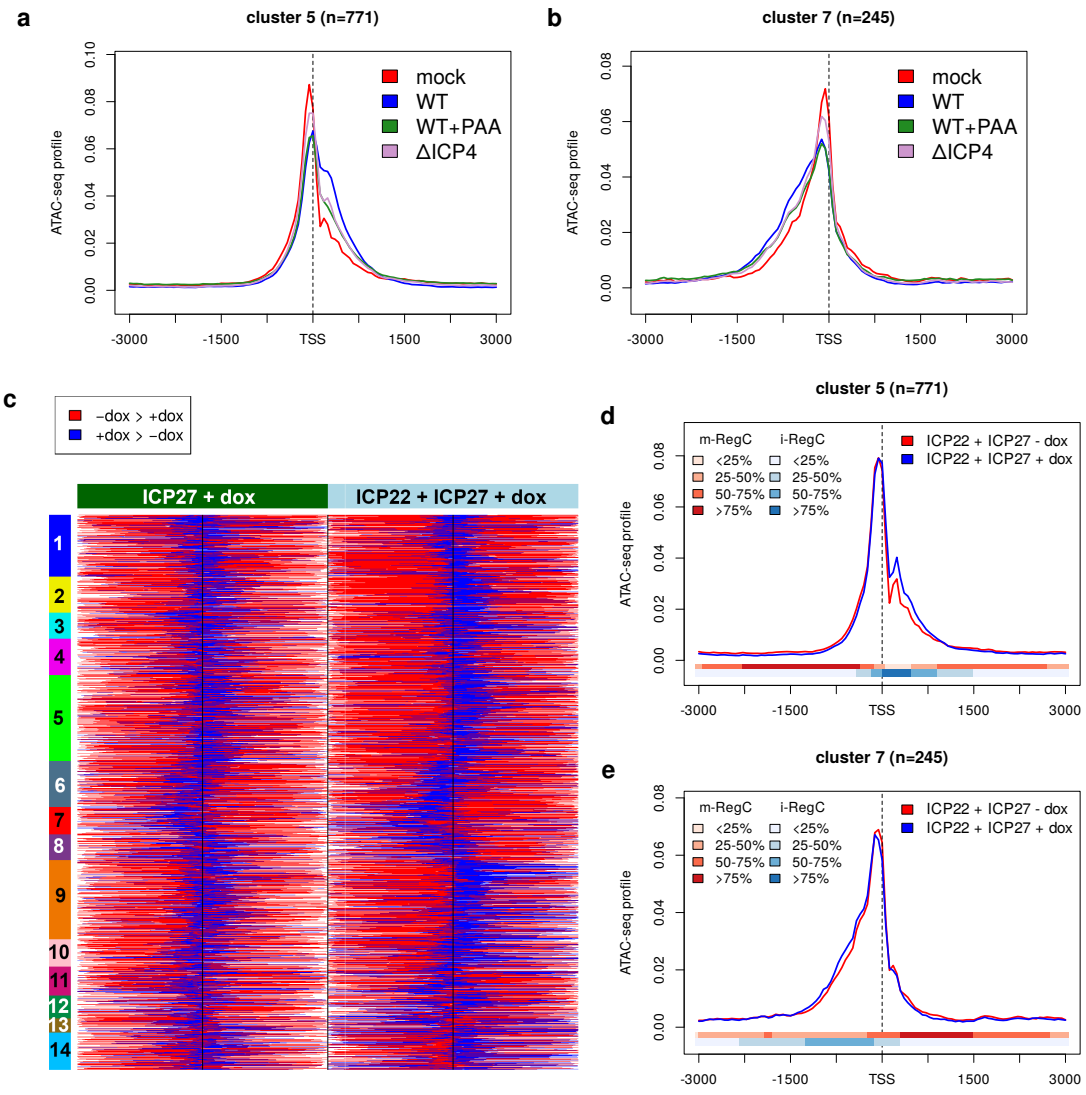
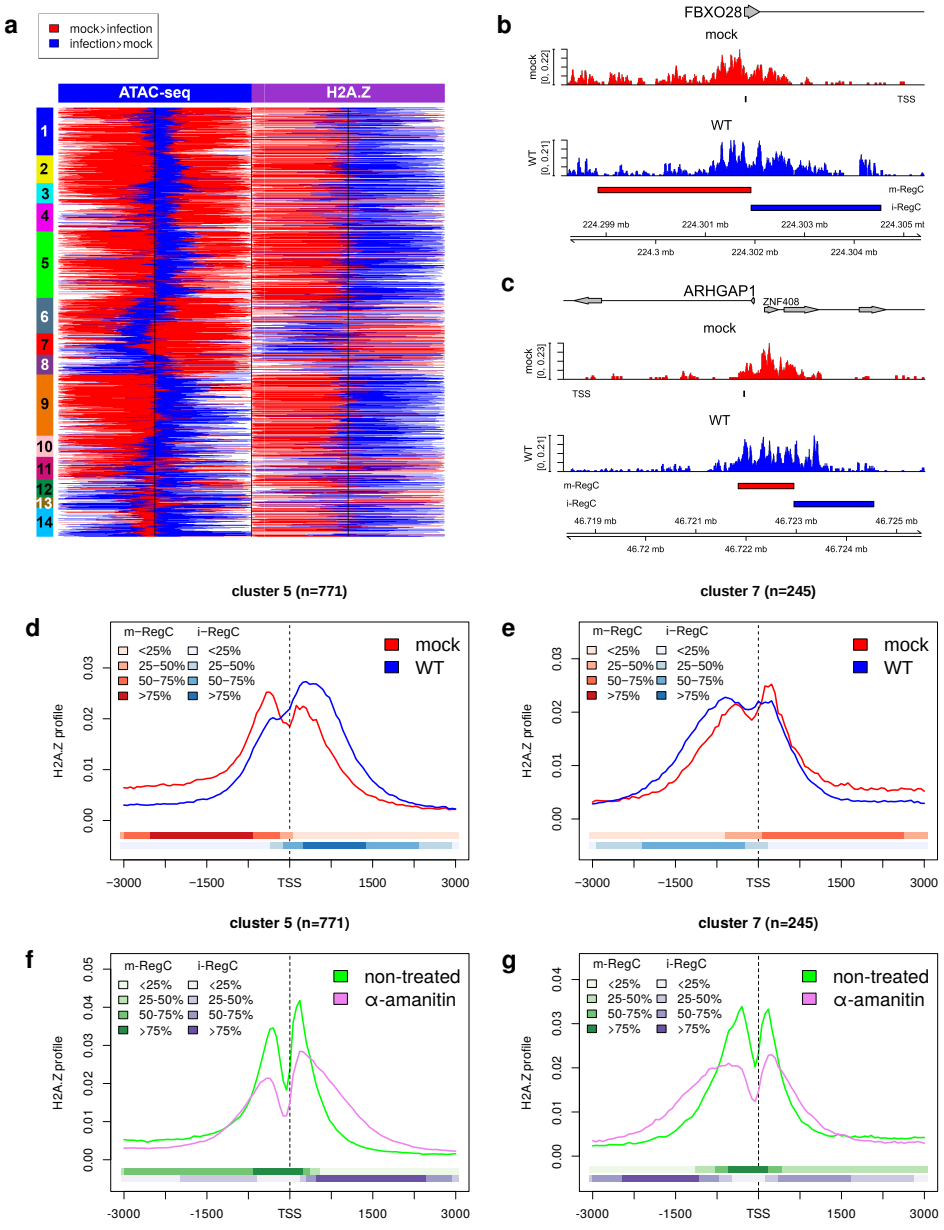
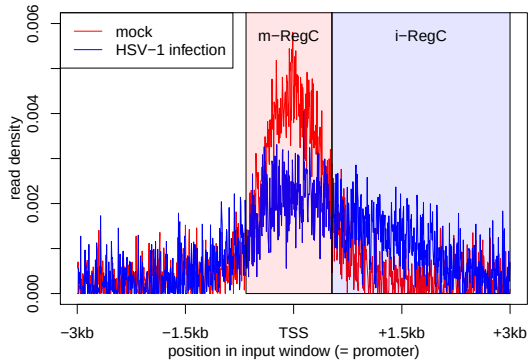

a
cluster 5 (n=771)
ATAC-seq profile
mock
WT
WT+PAA
ΔICP4
-3000
-1500
TSS
1500
3000
b
cluster 7 (n=245)
c
−dox > +dox
+dox > −dox
ICP27 + dox
ICP22 + ICP27 + dox
d
cluster 5 (n=771)
m-RegC
i-RegC
<25%
25-50%
50-75%
>75%
ICP22 + ICP27 - dox
ICP22 + ICP27 + dox
e
cluster 7 (n=245)

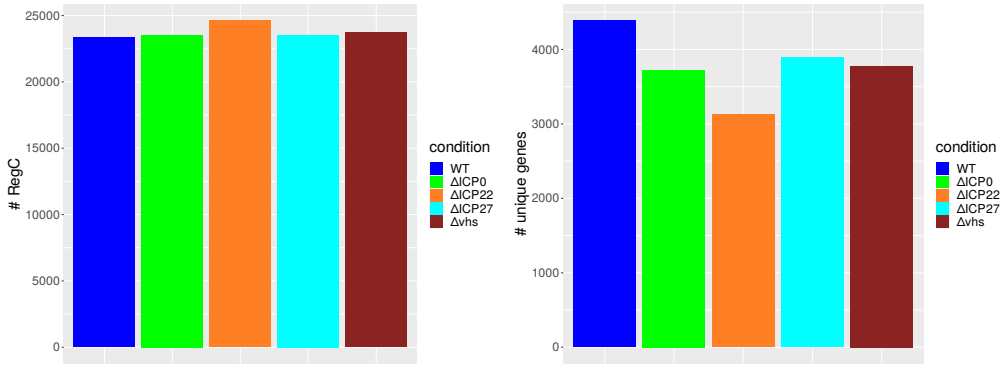
Fig 6


a
mock>infection
infection>mock
ATAC-seq
H2A.Z
b
FBXO28
mock
WT
m-RegC
i-RegC
c
ARHGAP1
ZNF408
mock
WT
m-RegC
i-RegC
cluster 5 (n=771)
cluster 7 (n=245)
d
e
m−RegC i−RegC
<25% <25%
25−50% 25−50%
50−75% 50−75%
>75% >75%
mock
WT
H2A.Z profile
−3000 −1500 TSS 1500 3000
f
g
m-RegC i-RegC
<25% <25%
25-50% 25-50%
50-75% 50-75%
>75% >75%
non-treated
α-amanitin
-3000 -1500 TSS 1500 3000

Supplementary Figures

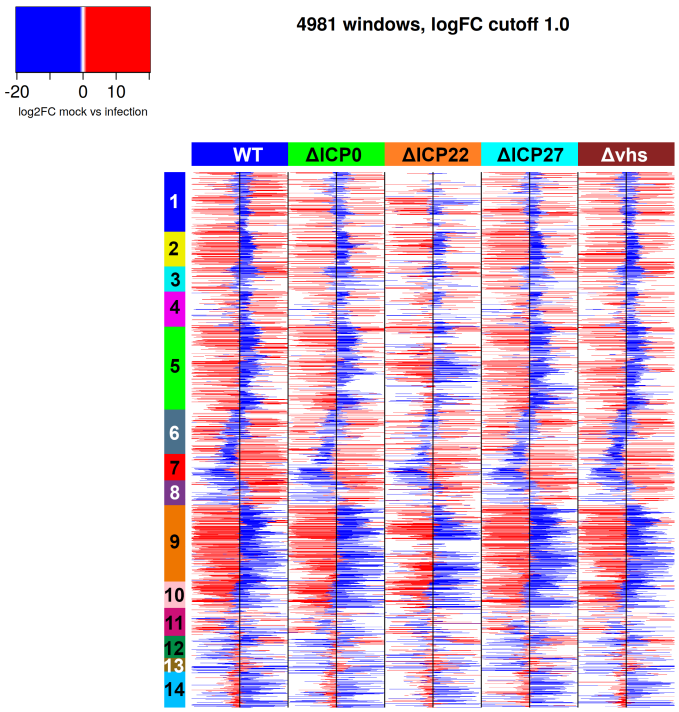


(a)



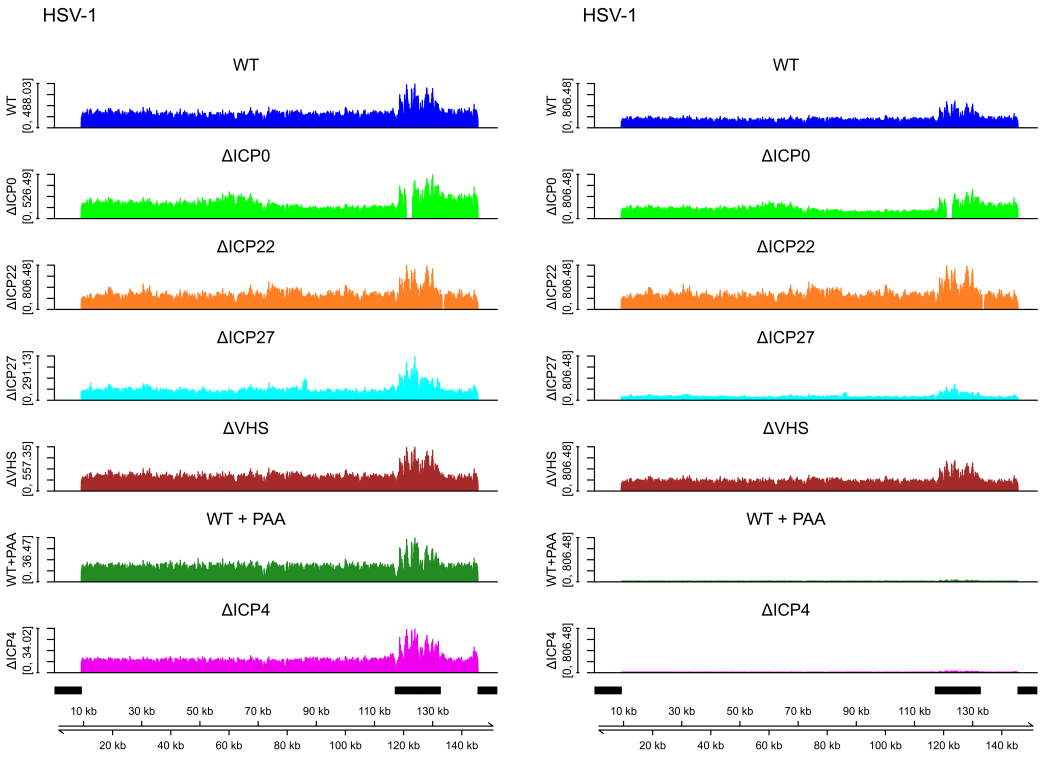
(b) **(c)**

(Continued on next page)



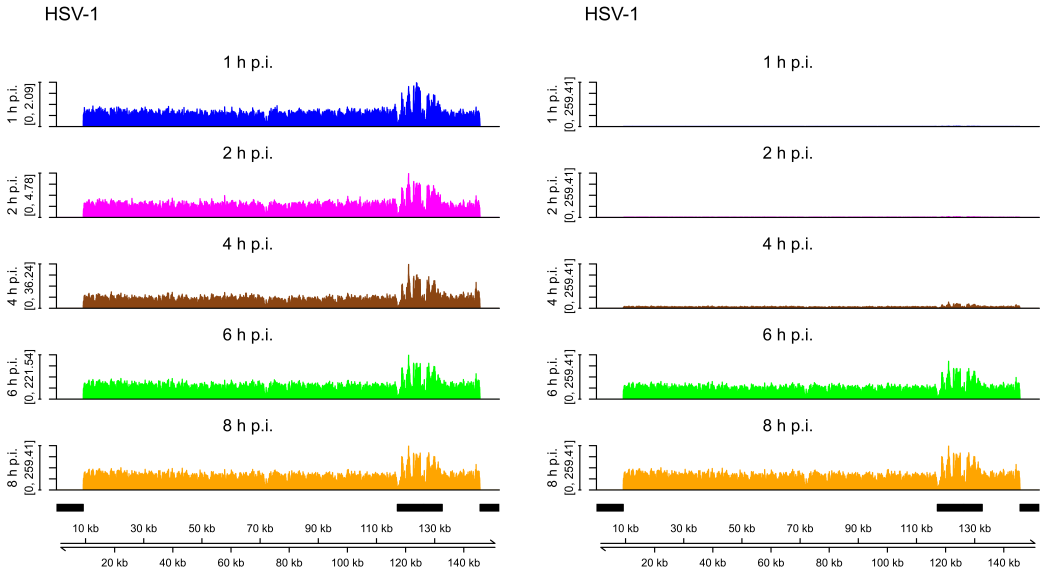
(d)

Sup. Fig 1 (a) Overview on the RegCFinder approach (Weiss and Friedel, 2023). The key objective of RegCFinder is to identify subregions of an input window that show a relative increase in read density within the input window in one condition compared to a second condition. In our application, this means identifying subregions of 6 kb promoter windows that exhibit a relative increase in read density in mock compared to HSV-1 infection (red shaded region, denoted as m-RegC) or vice versa (blue shaded region, denoted as i-RegC). In this example, an m-RegC (mock>infection) is identified around the TSS and an i-RegC (infection>mock) downstream of the TSS. This would reflect a relative decrease of read density at the TSS and a relative increase downstream of the TSS during infection. Statistical significance of changes in read density between identified subregions of input windows (including identified m- and i-RegC as well as filler regions between them) is determined with DEXSeq (Anders *et al.*, 2012). **(b)** Barplot showing the number of identified differential regions (= m- and i-RegC) for WT and null mutant infections in comparison to mock infection. **(c)** Barplot showing the number of unique genes with at least one statistically significant (multiple testing adjusted p-value (adj. p.) ≤ 0.01) differential region for WT and null mutant infection in comparison to mock infection. **(d)** Heatmap visualizing log2 fold-changes for the identified differential regions shown in **Fig. 1a**. For this purpose, results for the same input window for the different comparisons of WT or null mutant virus infections to mock are concatenated in one row of the heatmap matrix. Colored rectangles on top of columns indicate which virus infection was compared to mock. Black vertical lines in the center of each comparison indicate the position of the TSS. Regions corresponding to statistically significant differential regions are colored according to the log2 fold-change in mock vs. infection determined by DEXSeq. Here, the color scale is continuous between -1 and 1 and log2 fold-changes > 1 are colored the same red and log2 fold-changes < 1 the same blue. Promoter windows are ordered as in **Fig. 1a** and clusters from **Fig. 1a** are annotated as colored and numbered rectangles on the left.

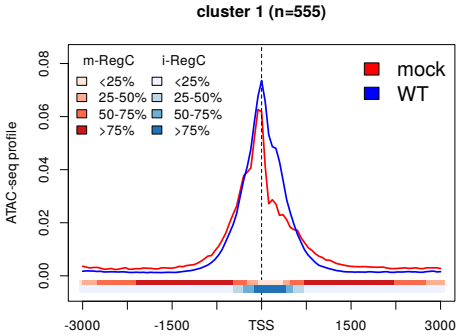


(a) **(b)**

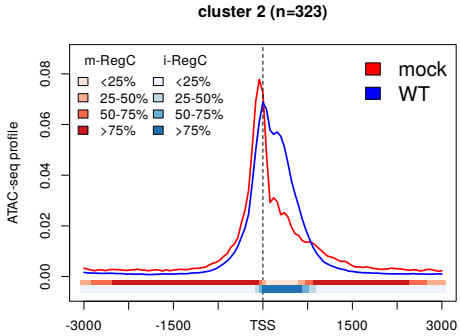
(Continued on next page)



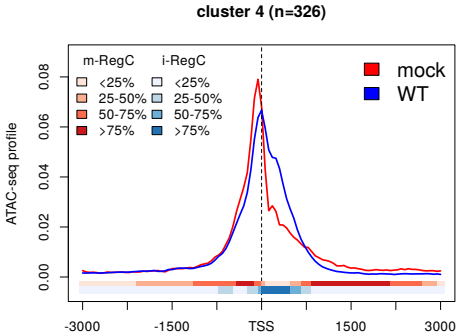
Sup. Fig 2 Read coverage on the HSV-1 genome for **(a,b)** WT (blue), ΔICP0 (green), ΔICP22 (orange), ΔICP27 (cyan), Δ*vhs* (brown), WT+PAA infection (dark green) and ΔICP4 infection (magenta) from the first ATAC-seq experiment and **(c,d)** 1 h (blue), 2 h (magenta), 4 h (brown), 6 h (green), and 8 h p.i. (orange) HSV-1 infection from the ATAC-seq time-course experiment. Read coverage was normalized to total number of mapped reads for each sample and averaged between replicates. For **(a)** and **(c)**, the y-axis range was determined independently for each condition and for **(b)** and **(d)** the same y-range was chosen for all conditions based on the maximum observed coverage. The black rectangles above the genome coordinates depict the position of the inverted repeat regions in the HSV-1 genome. The terminal repeat copies at the start and end of the genome were masked from read alignment, resulting in all reads from the inverted repeats mapping to the internal repeat copies.



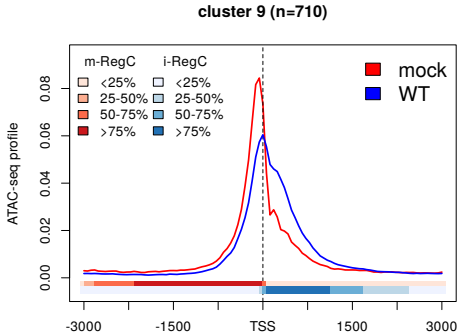
(a) pattern I



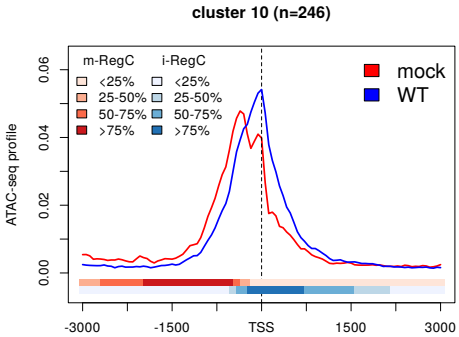
(b) pattern I



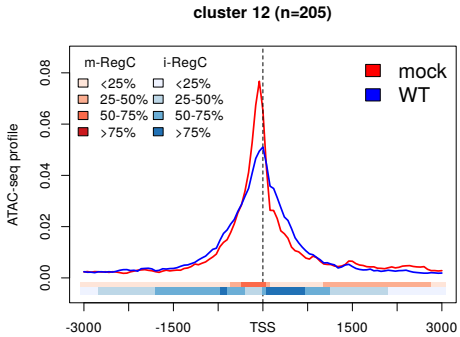
(c) pattern I



(d) pattern I

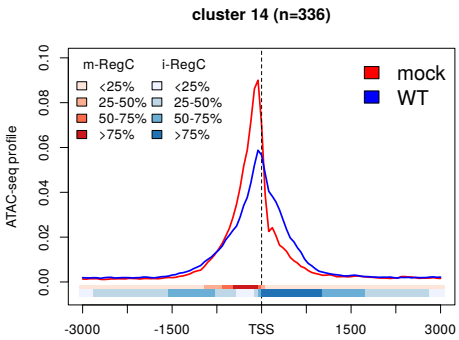


(e) pattern I

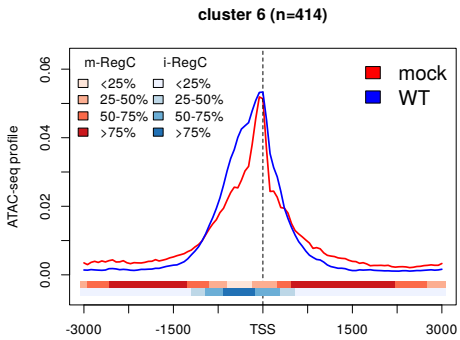


(f) pattern I

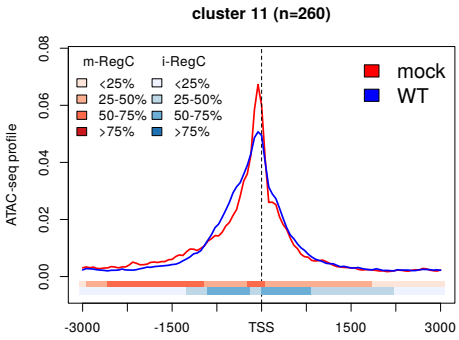
(Continued on next page)



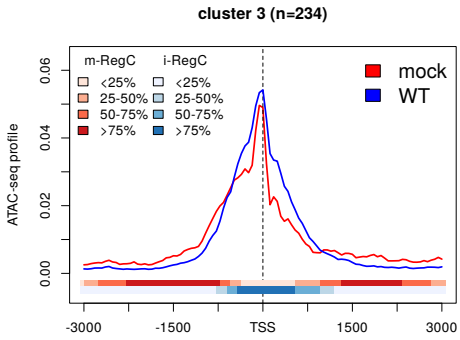
(g) pattern I



(h) pattern II

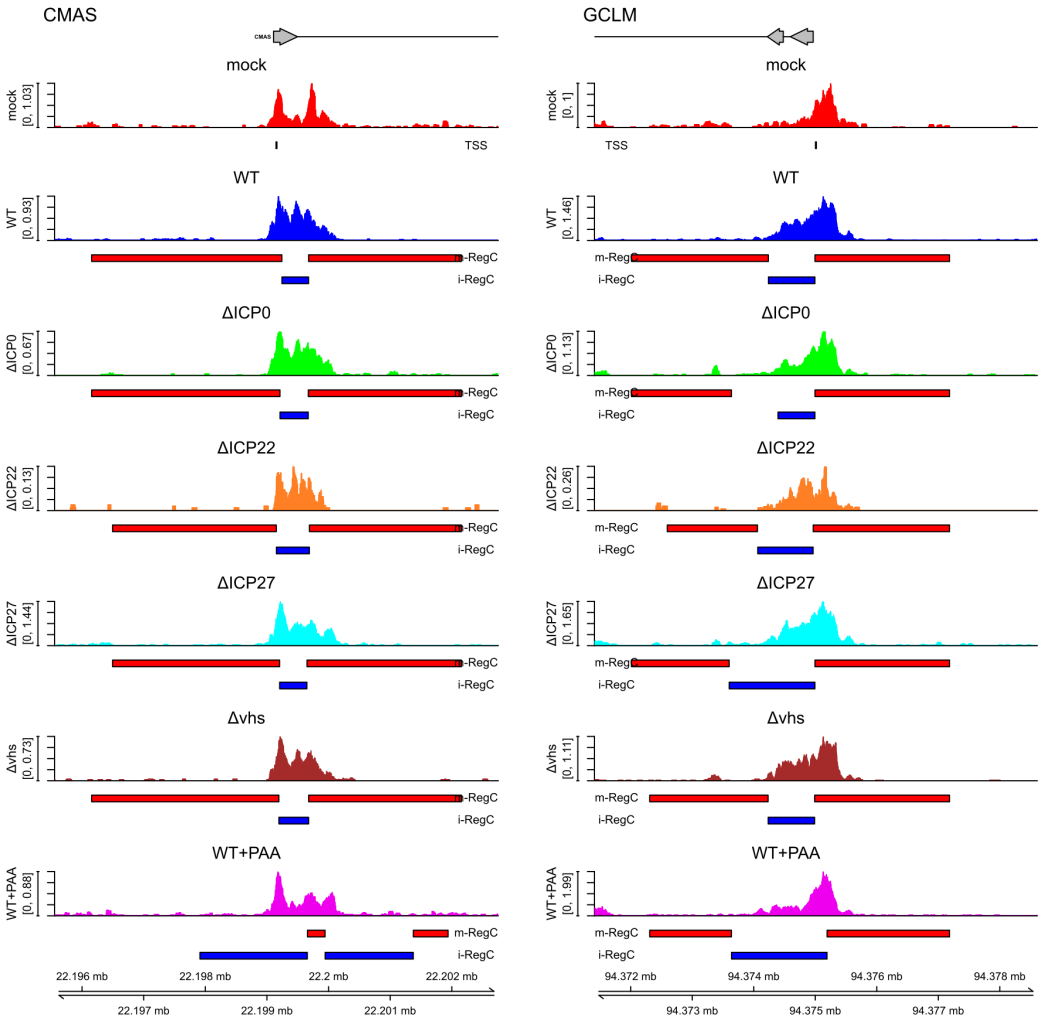


(i) pattern II



(j) pattern I + II

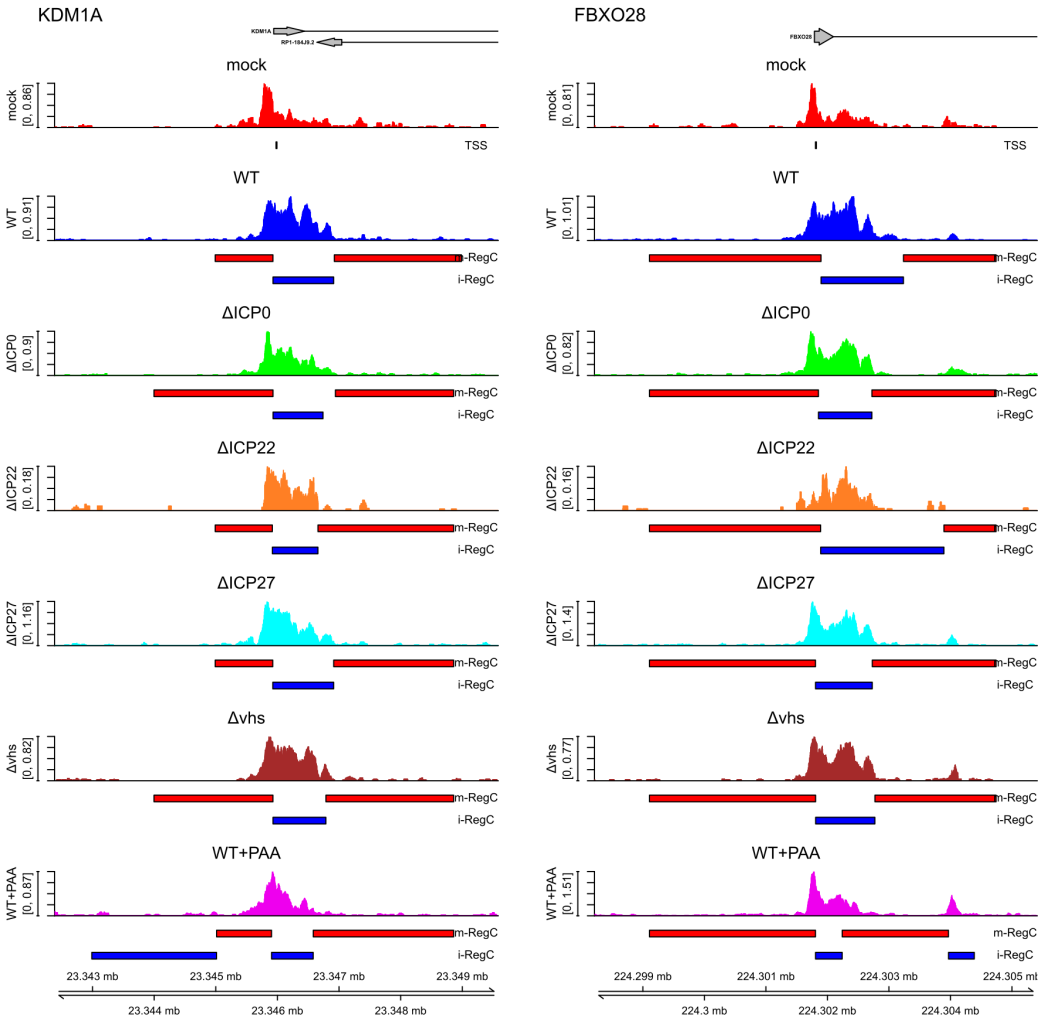
Sup. Fig 3 Metagene plots showing ATAC-seq profiles around the TSS ± 3 kb in mock (red) and WT (blue) infection for all clusters apart from clusters 5, 7, 8, and 13, which are shown in **Fig. 1**. See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the comparison of WT infection to mock. Subfigures **(a-g)** show clusters with pattern I with a shift and/or broadening of the TSS peak into downstream regions, while **(h-i)** show clusters that exhibit the mirror pattern II with a shift and/or broadening of the TSS peak into upstream regions. Subfigure **(j)** shows a cluster with combined patterns I and II.



(a) pattern I, cluster 1

(b) pattern I, cluster 2

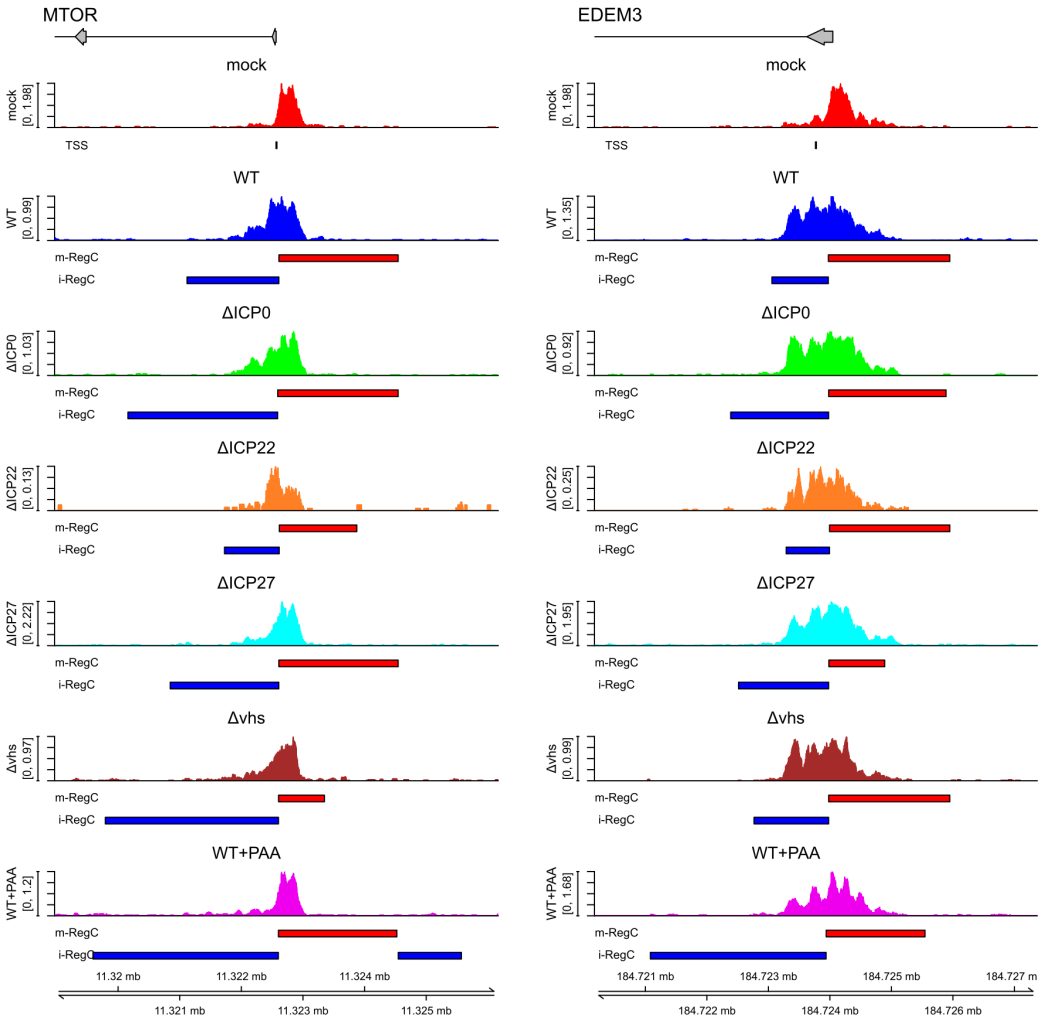
(Continued on next page)



(c) pattern I, cluster 4

(d) pattern I, cluster 5

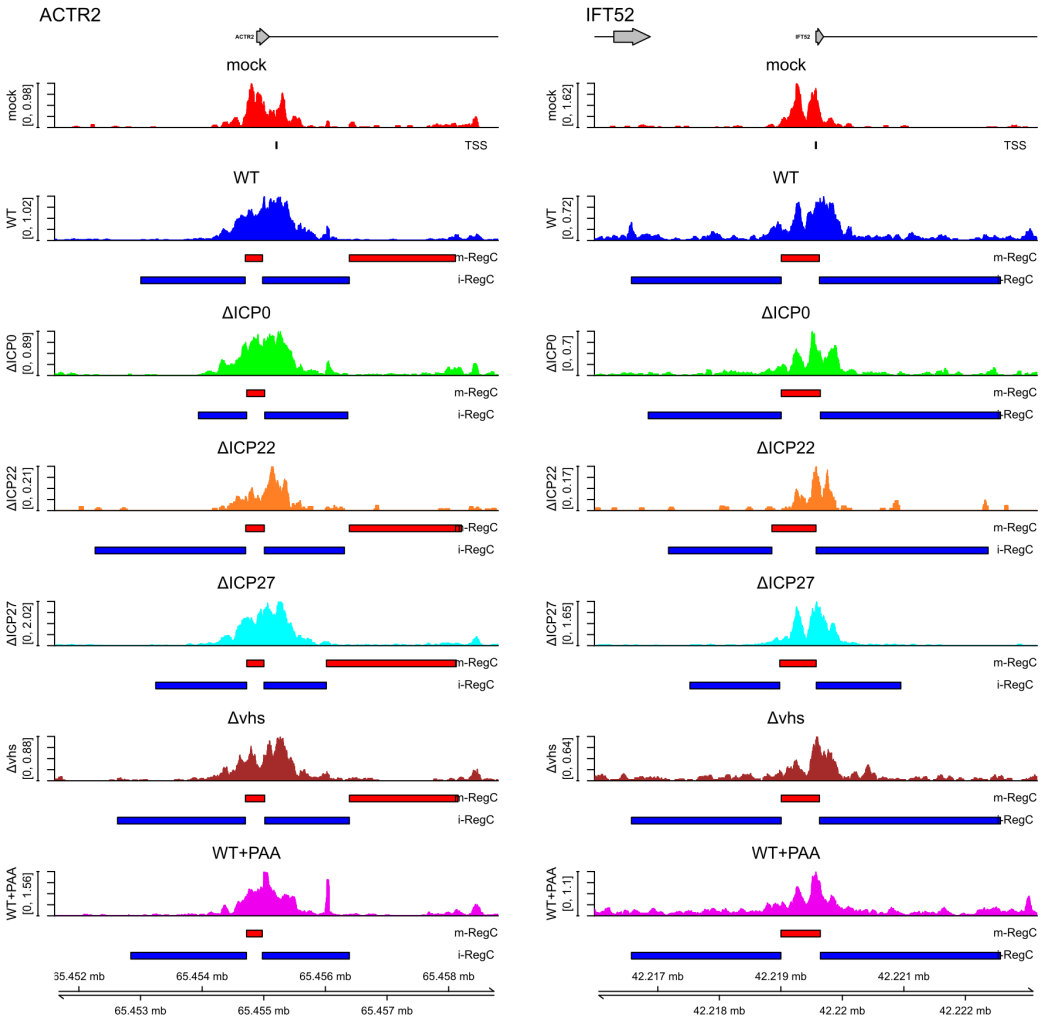
(Continued on next page)



(e) pattern I, cluster 9

(f) pattern I, cluster 10

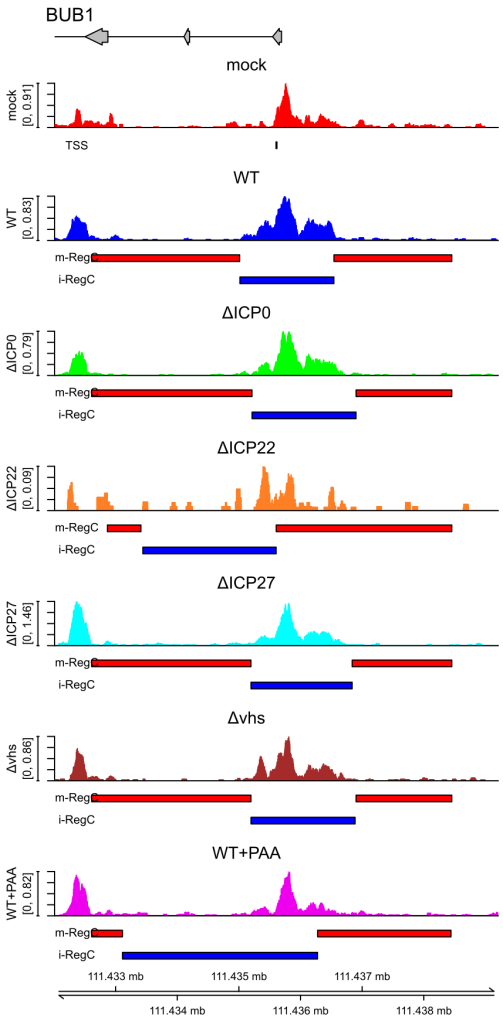
(Continued on next page)



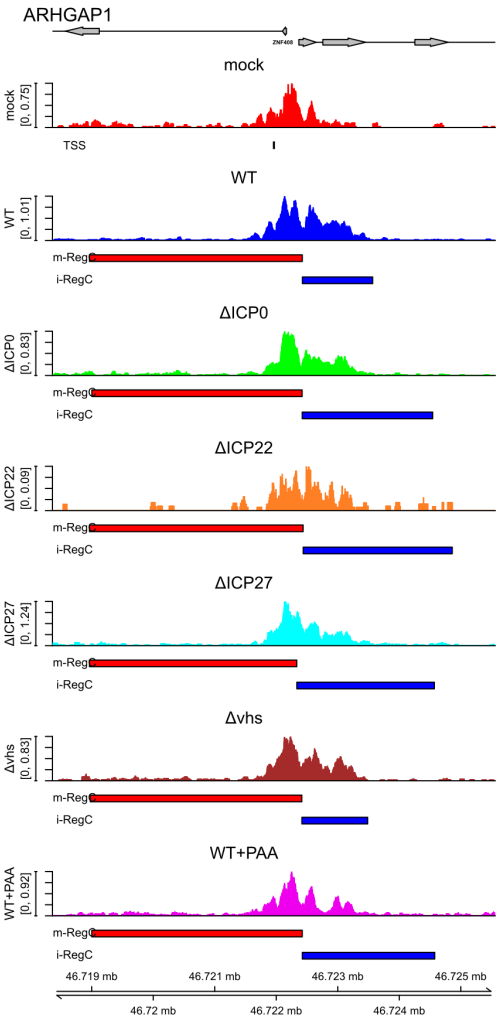
(g) pattern I, cluster 12

(h) pattern I, cluster 14

(Continued on next page)

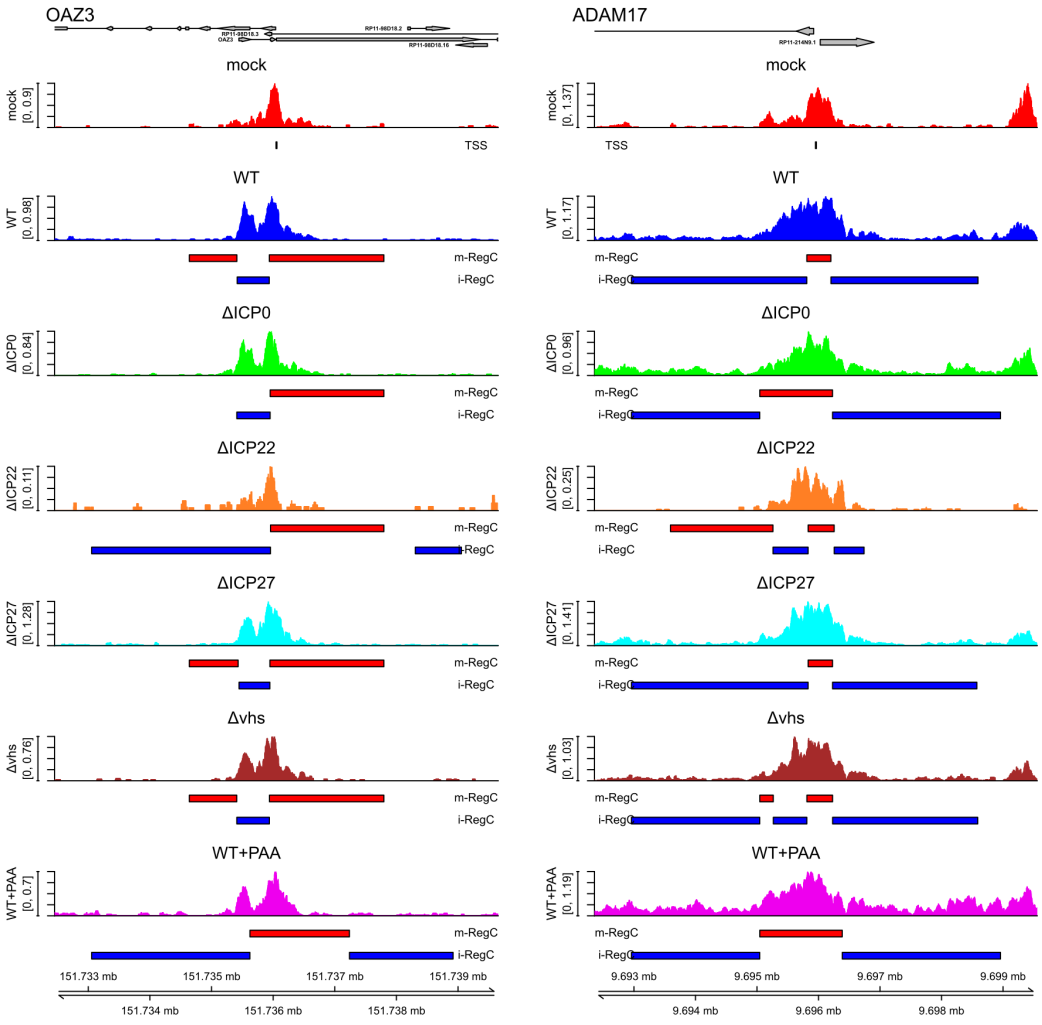


(i) pattern II, cluster 6



(j) pattern II, cluster 7

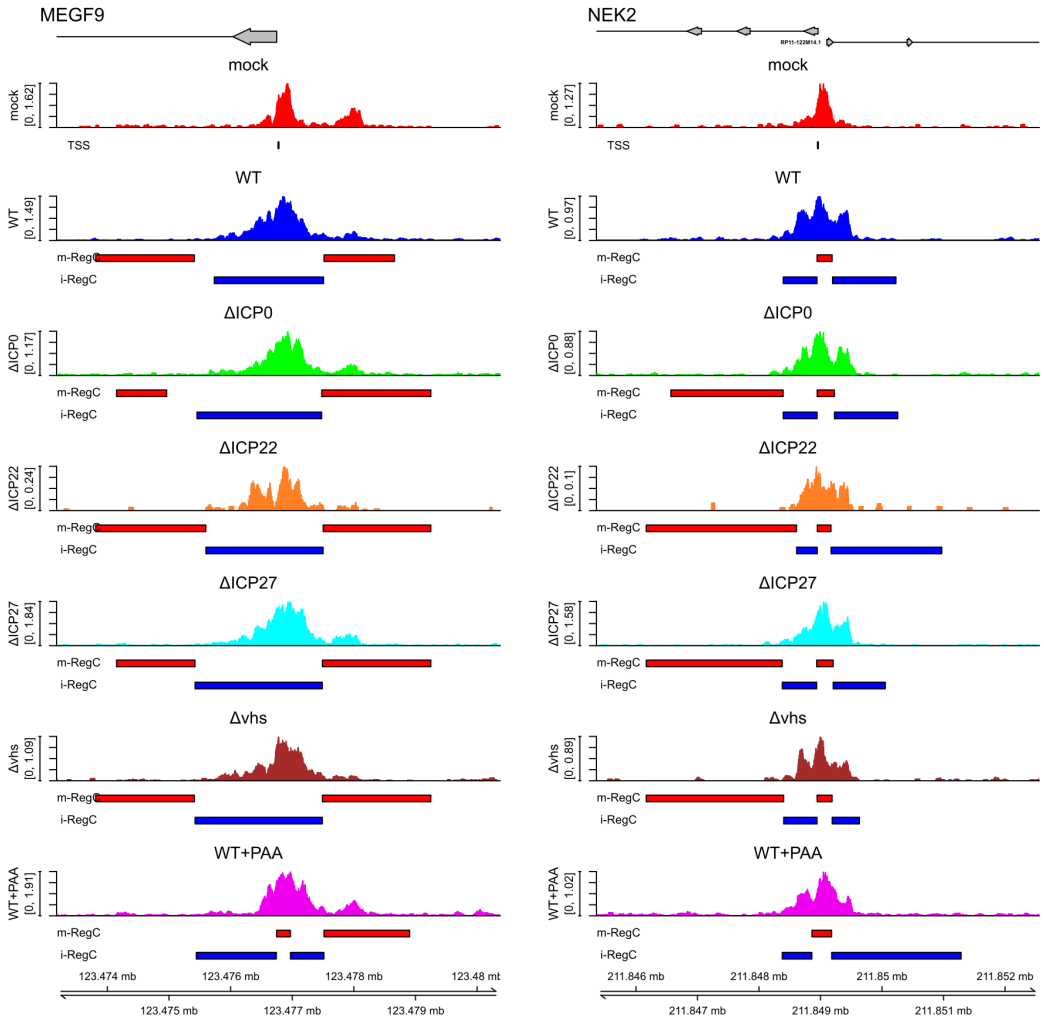
(Continued on next page)



(k) pattern II, cluster 11

(l) pattern III, cluster 13

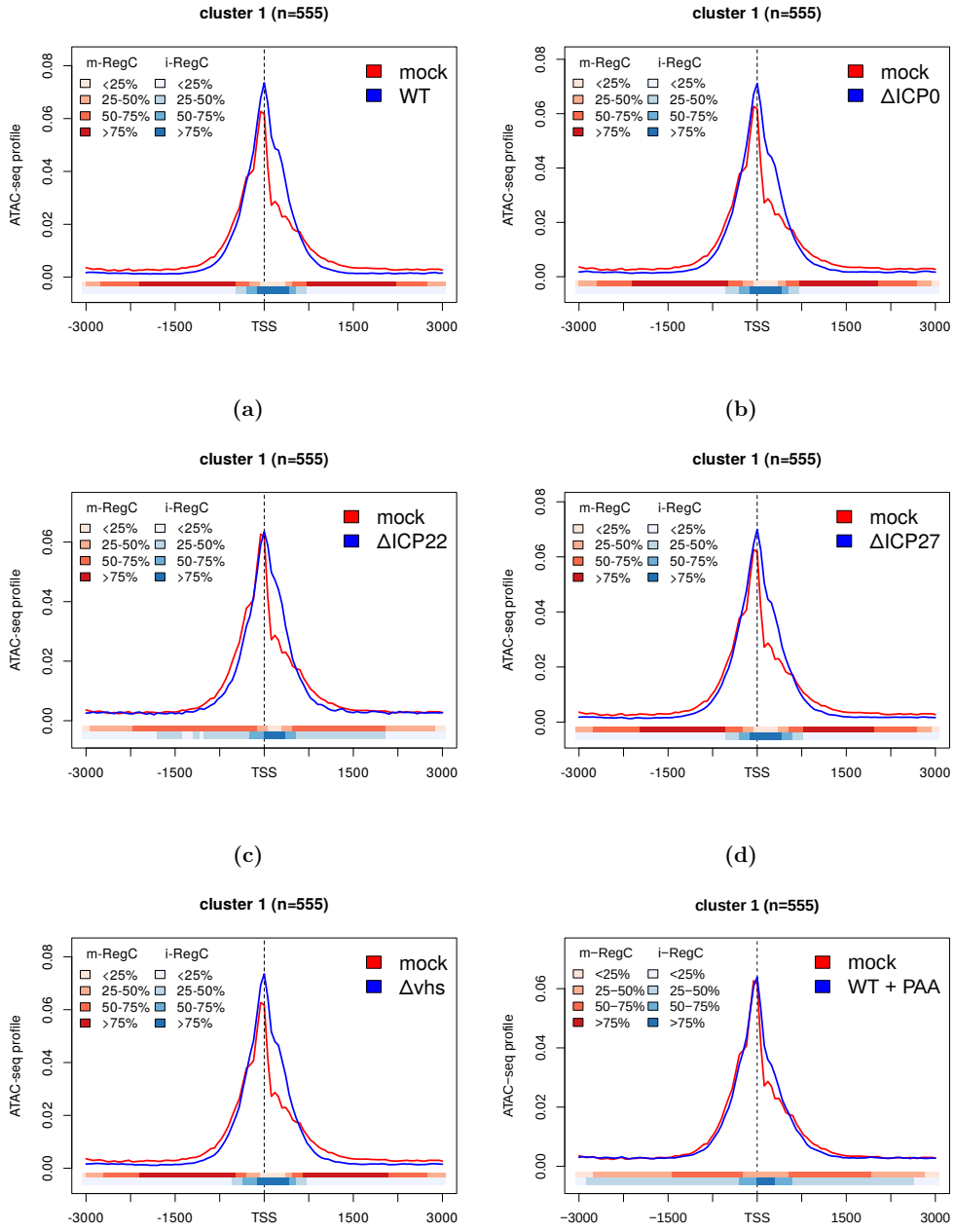
(Continued on next page)



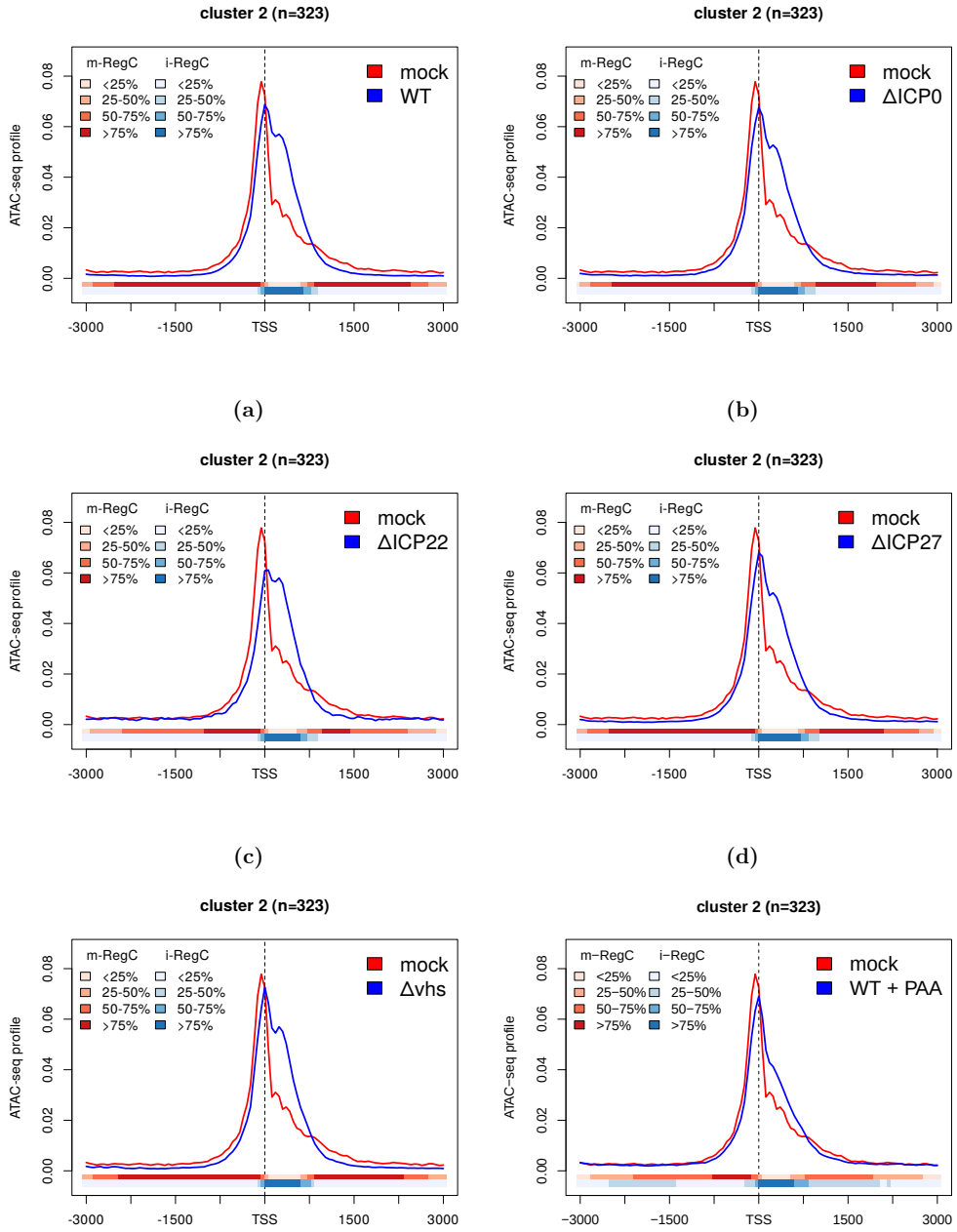
(m) pattern I + II, cluster 3

(n) pattern I + II, cluster 8

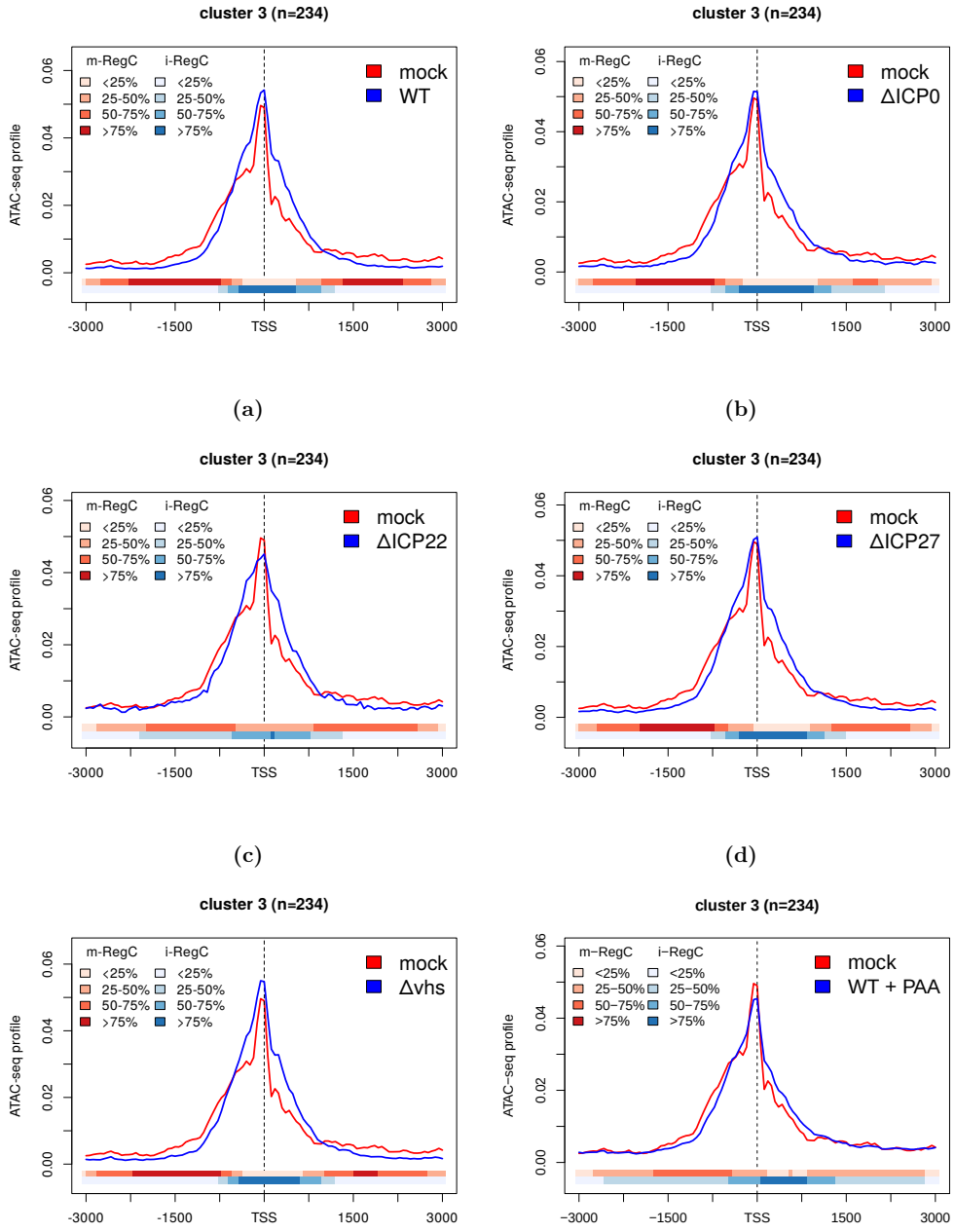
Sup. Fig 4 Read coverage in a ± 3.6 kb window around the TSS in ATAC-seq data of mock (red), WT (blue), ΔICP0 (green), ΔICP22 (orange), ΔICP27 (cyan), Δ*vhs* (brown) and WT+PAA infection (magenta) for example genes for all patterns and clusters. Pattern and cluster are indicated below each subfigure. Read coverage was normalized to total number of mapped reads for each sample and averaged between replicates. The TSS used in the analysis is indicated by a short vertical line below the read coverage track for mock infection. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns, and direction of transcription is indicated by arrowheads. The name of the gene whose promoter window was analyzed is indicated in larger font on the top left and – if not clear from the context – beside the gene annotation. Names for other genes overlapping the input window are indicated next to these genes if necessary. Below each read coverage track m- (red bars) and i-RegCs (blue bars) are indicated for the comparison of the corresponding virus infection to mock.



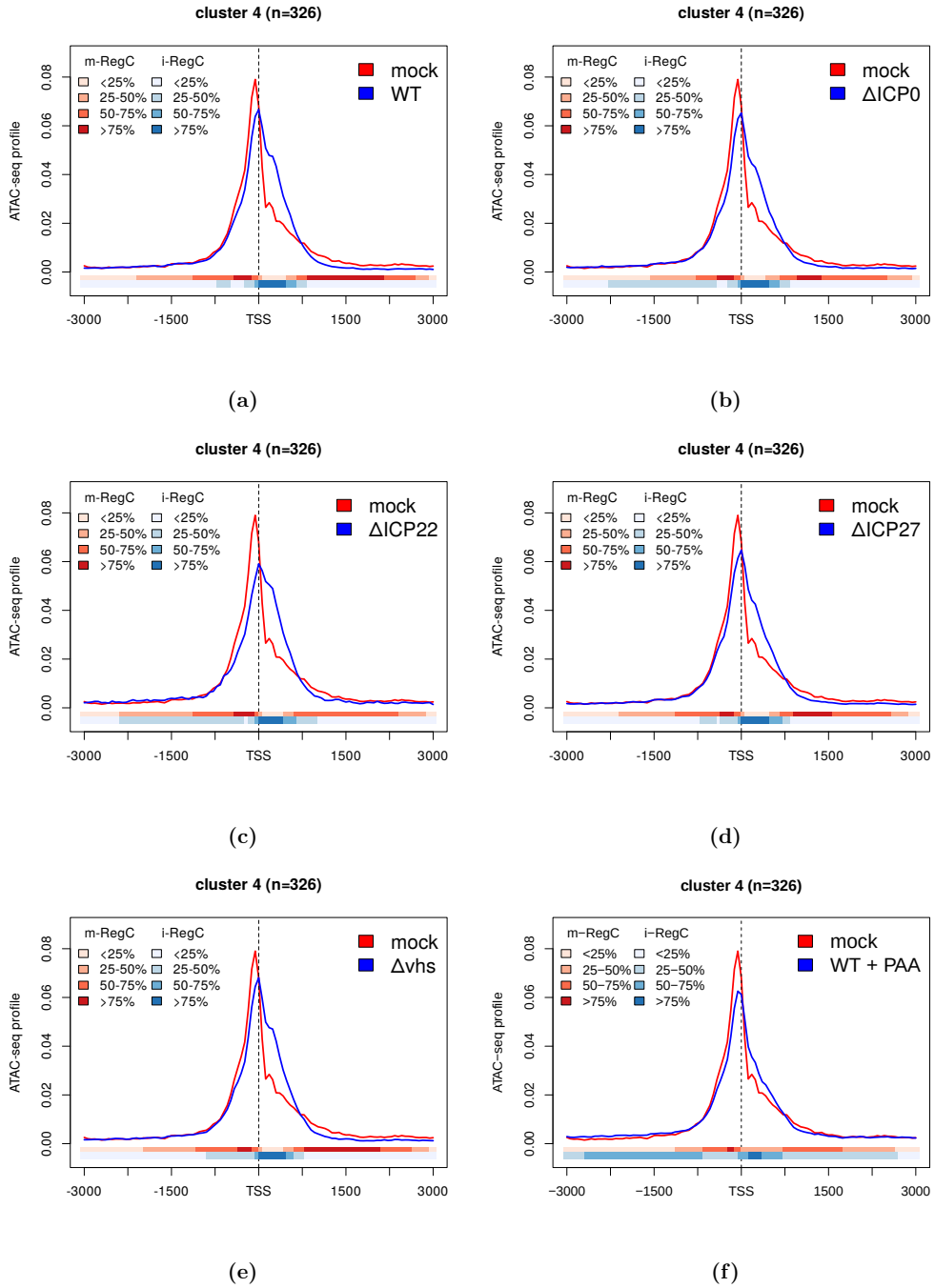
Sup. Fig 5 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 1 (pattern I). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



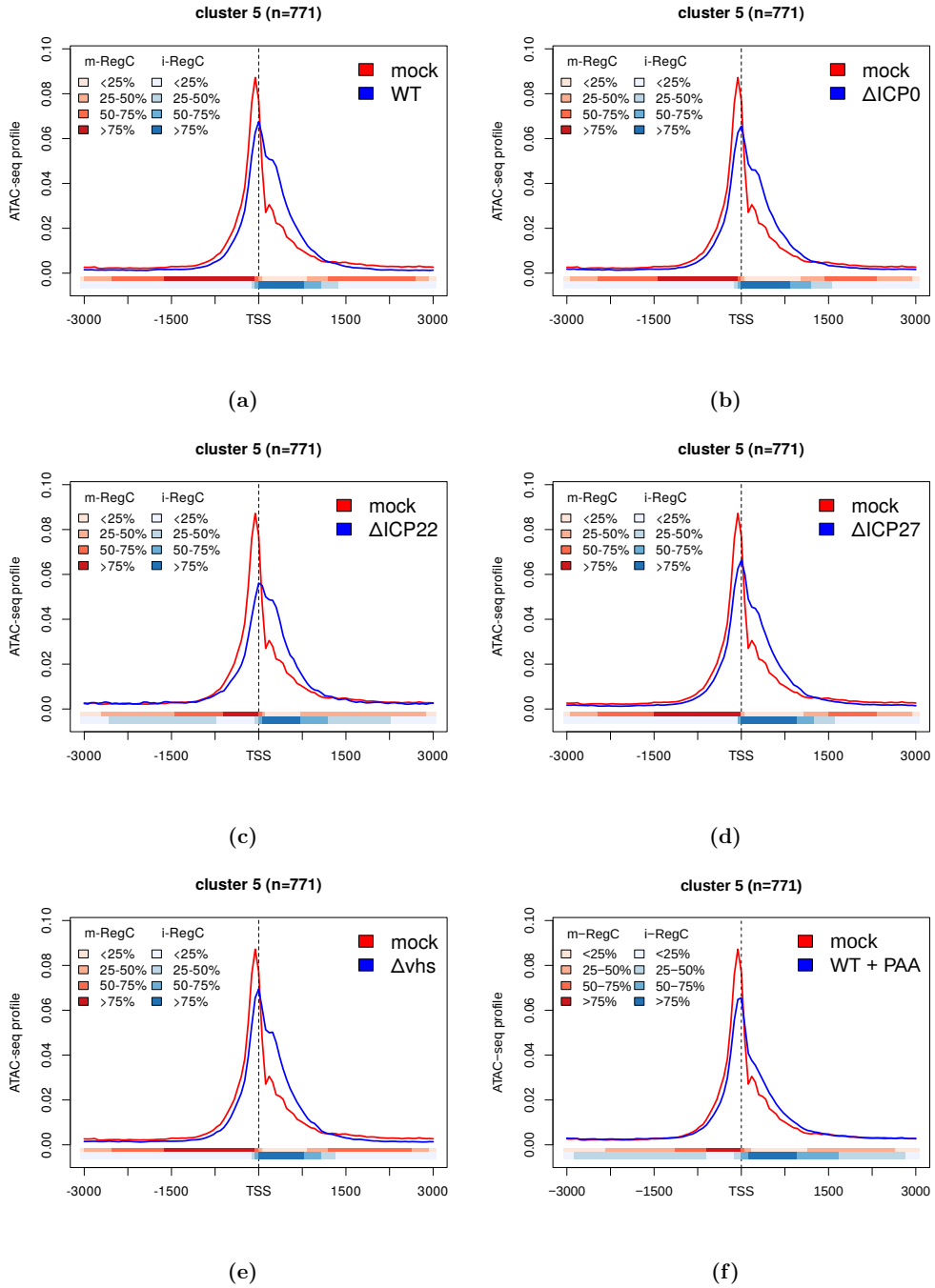
Sup. Fig 6 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 2 (pattern I). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



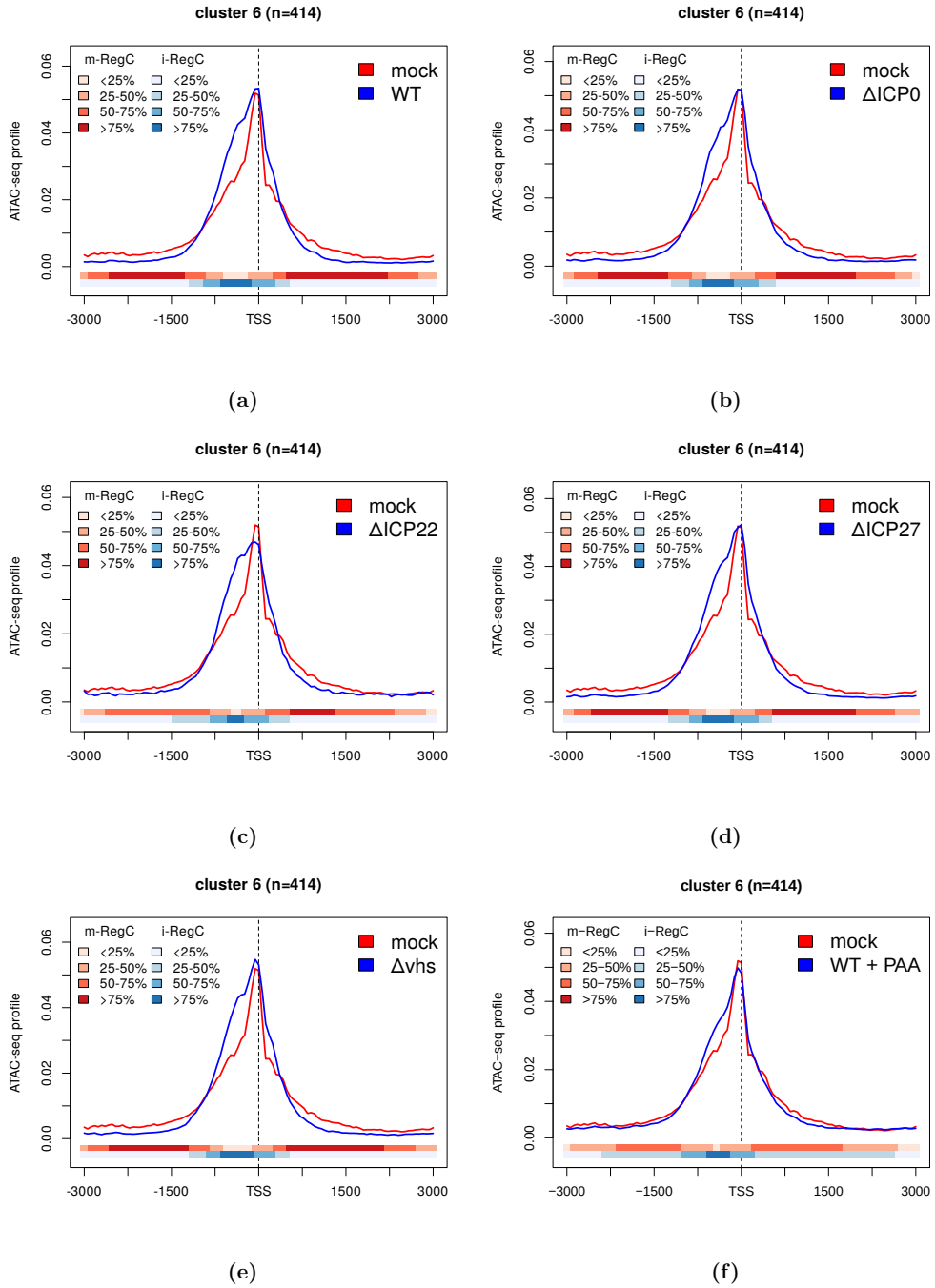
Sup. Fig 7 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 3 (combined pattern I + II). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



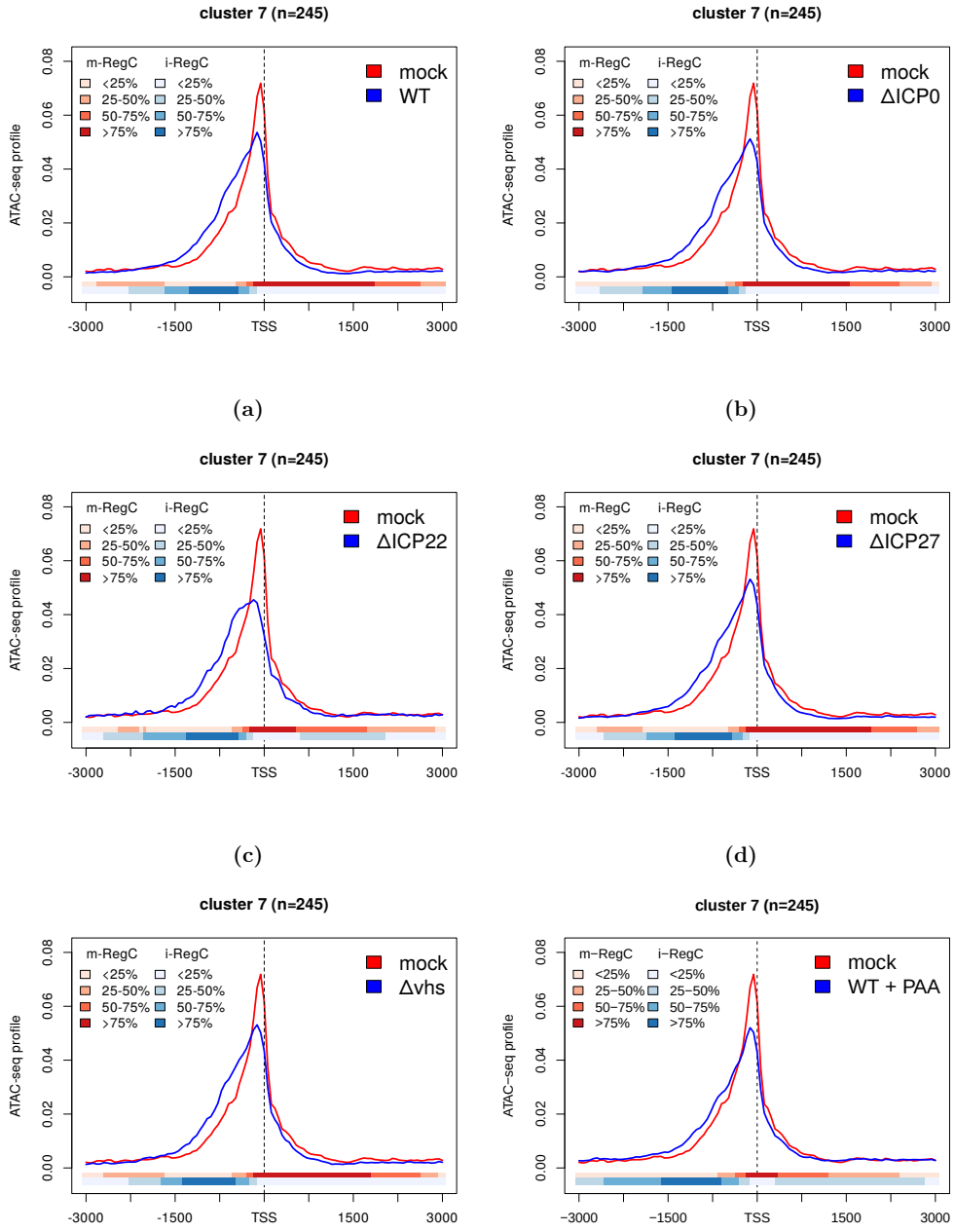
Sup. Fig 8 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 4 (pattern I). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



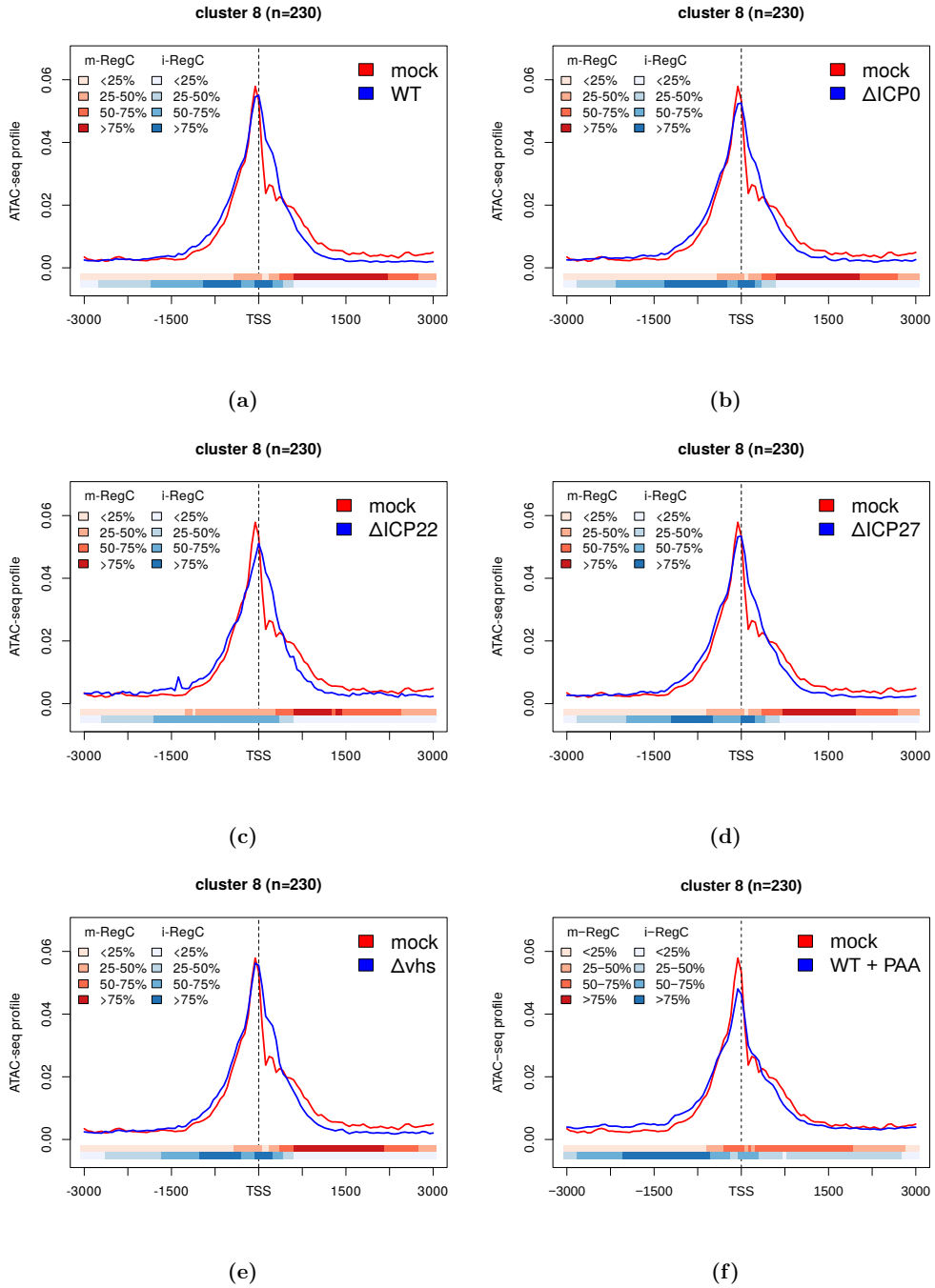
Sup. Fig 9 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 5 (pattern I). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



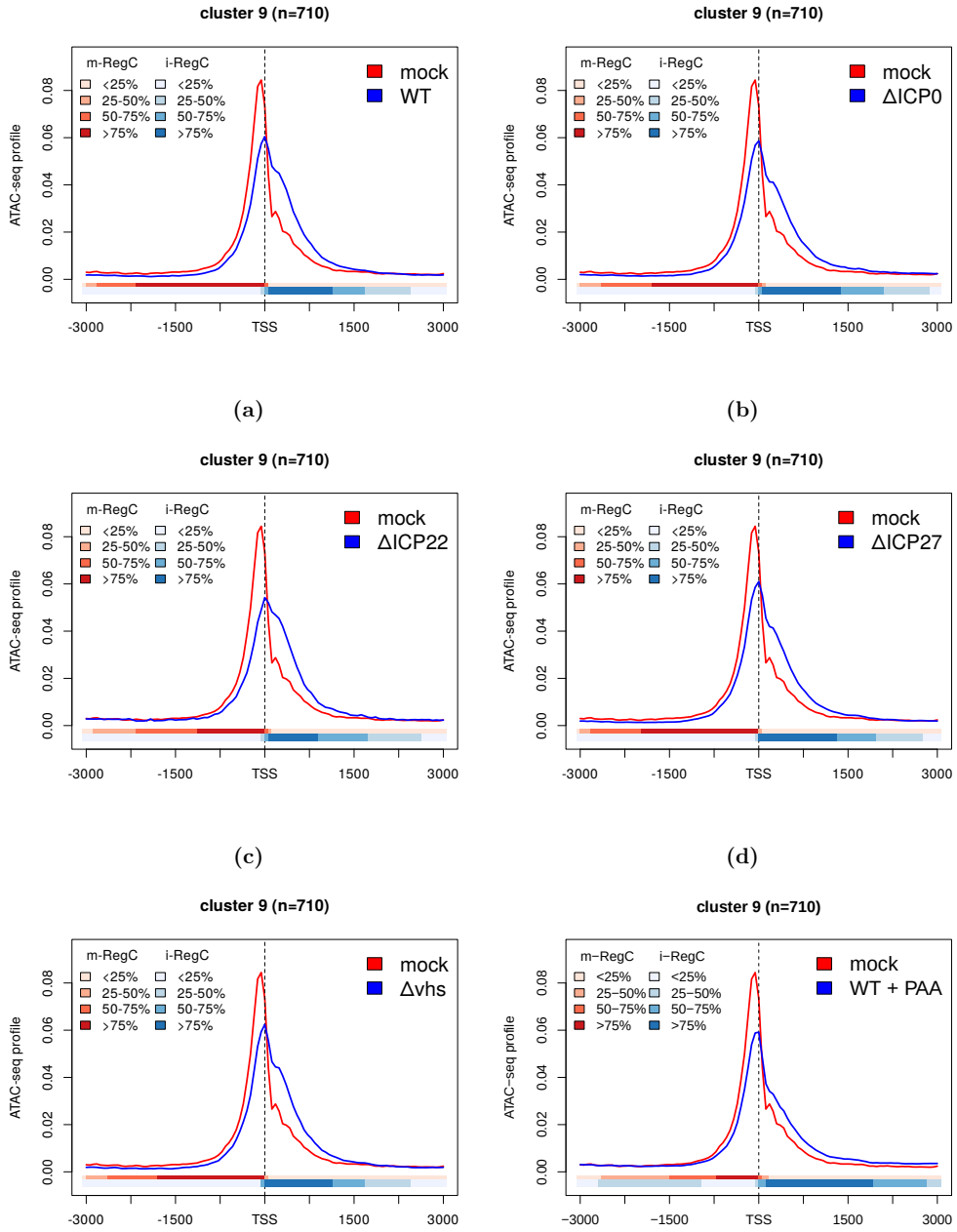
Sup. Fig 10 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 6 (pattern II). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



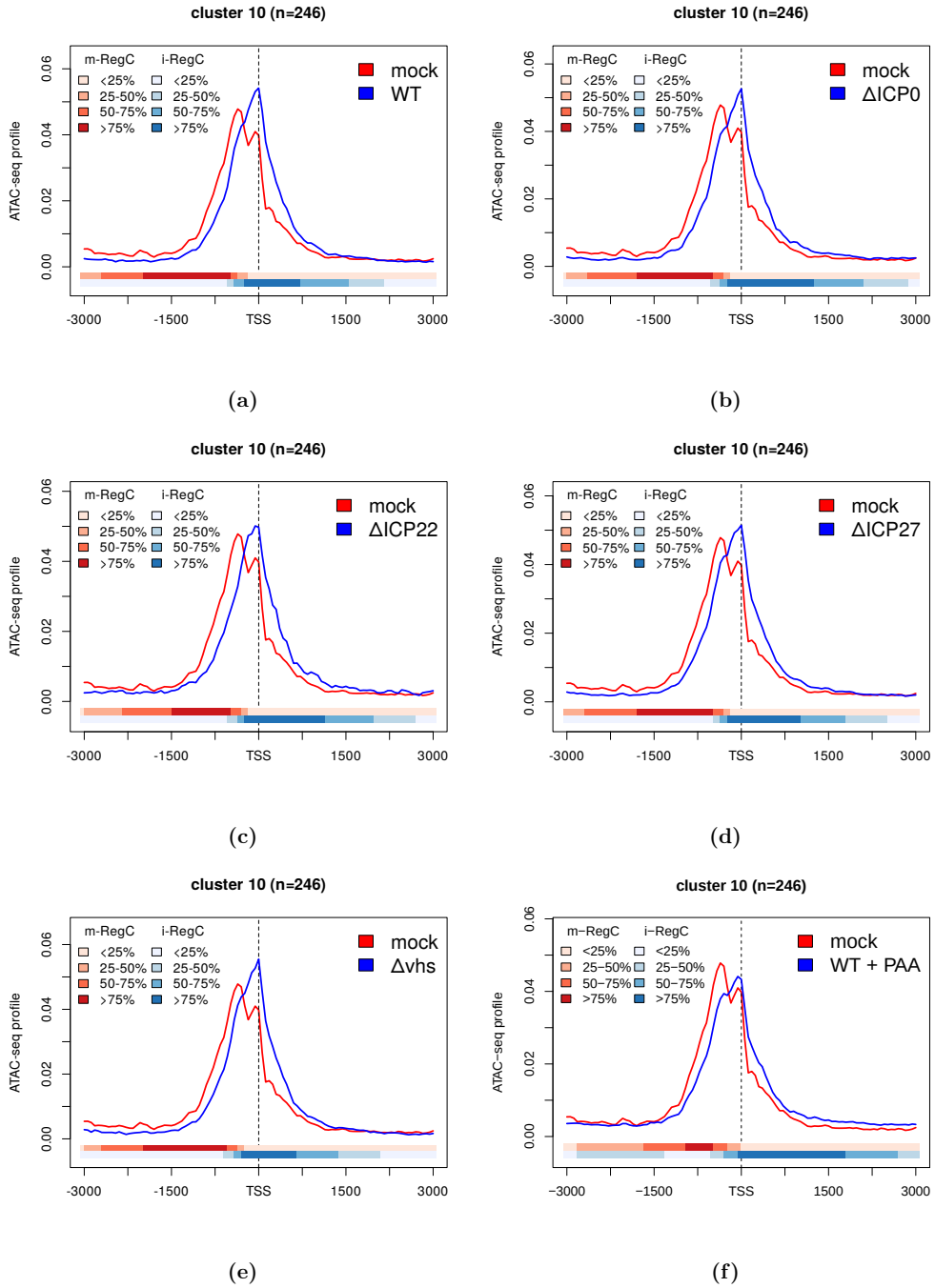
Sup. Fig 11 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 7 (pattern II). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



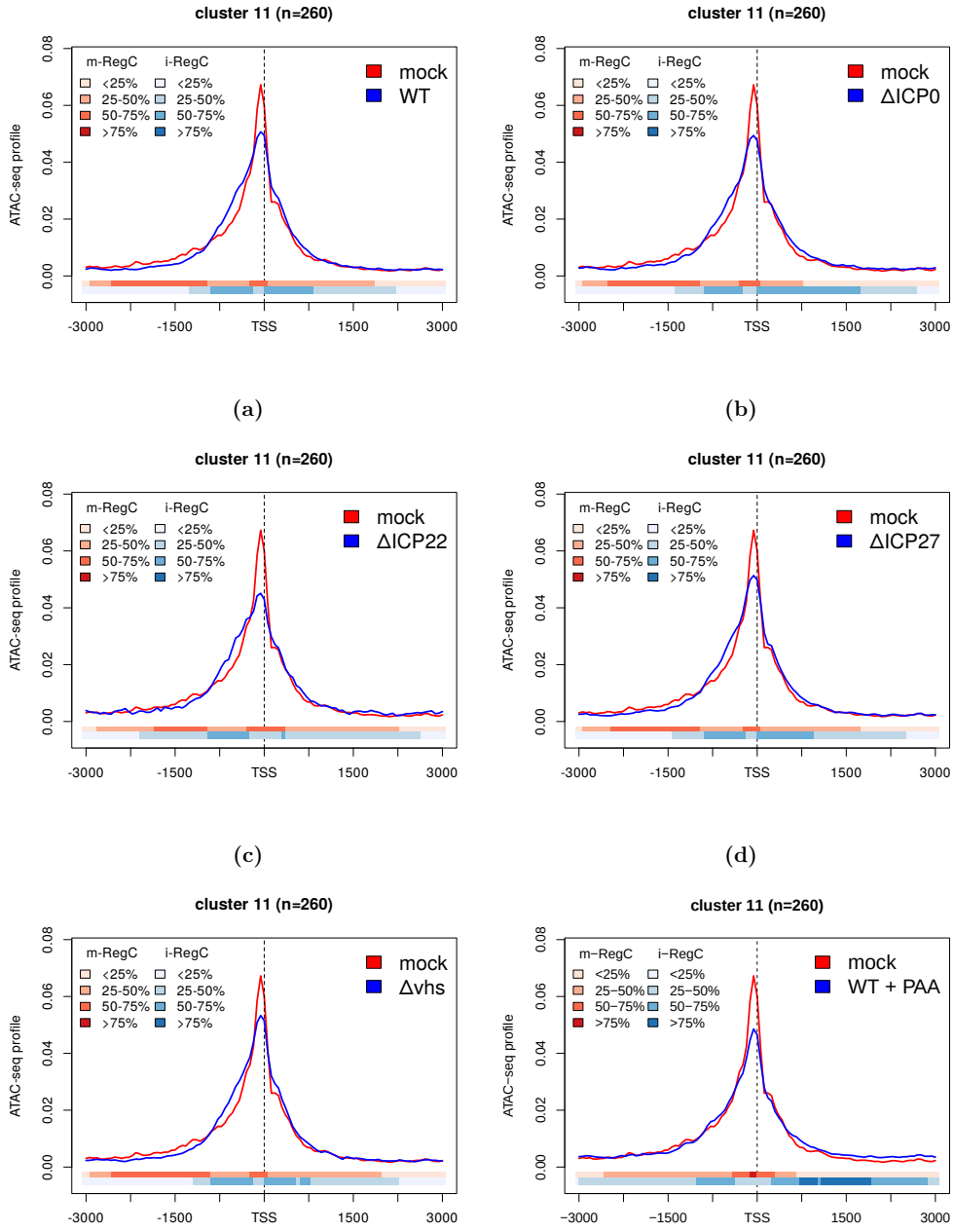
Sup. Fig 12 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 8 (combined pattern I + II). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



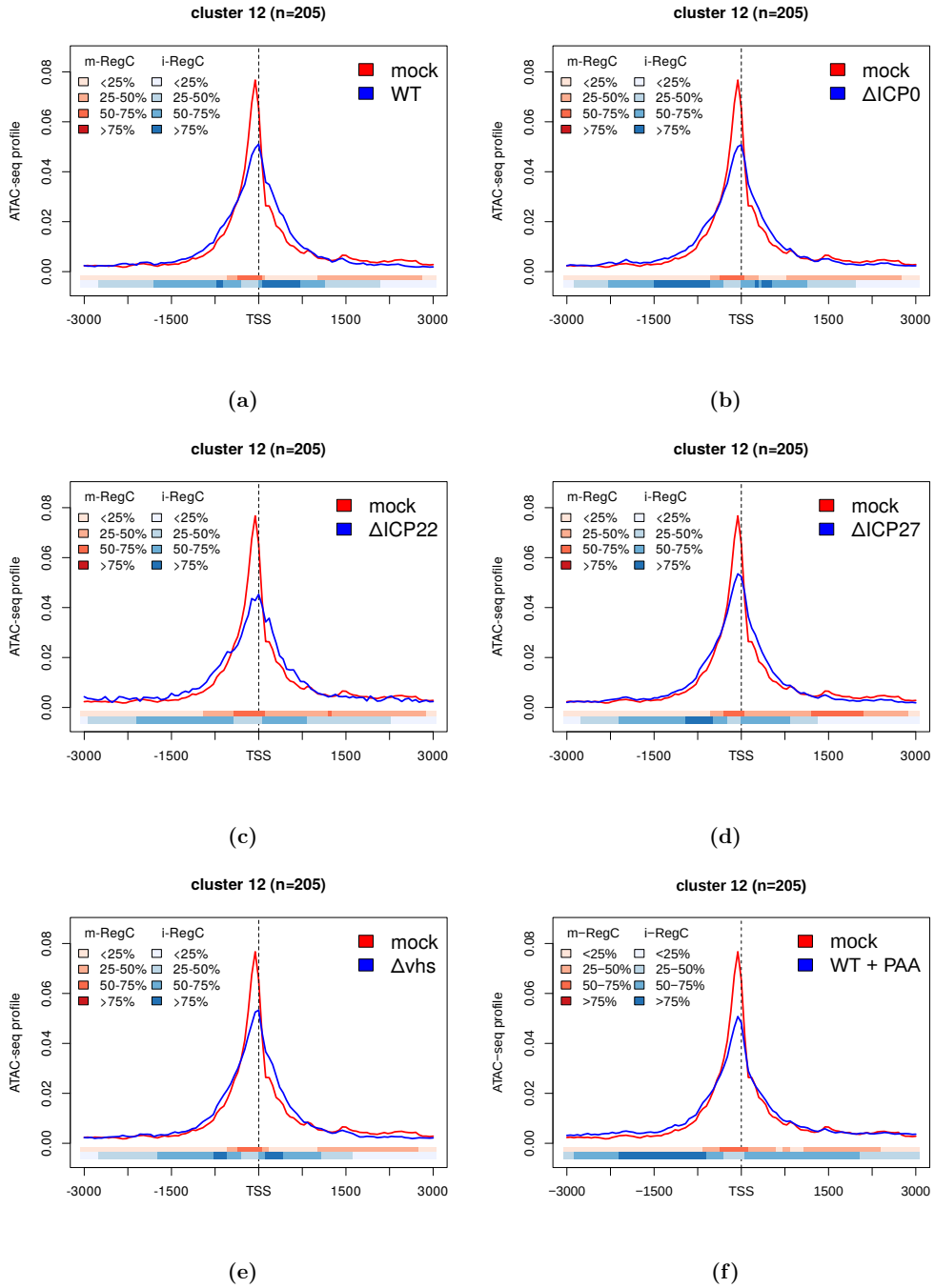
Sup. Fig 13 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 9 (pattern I). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



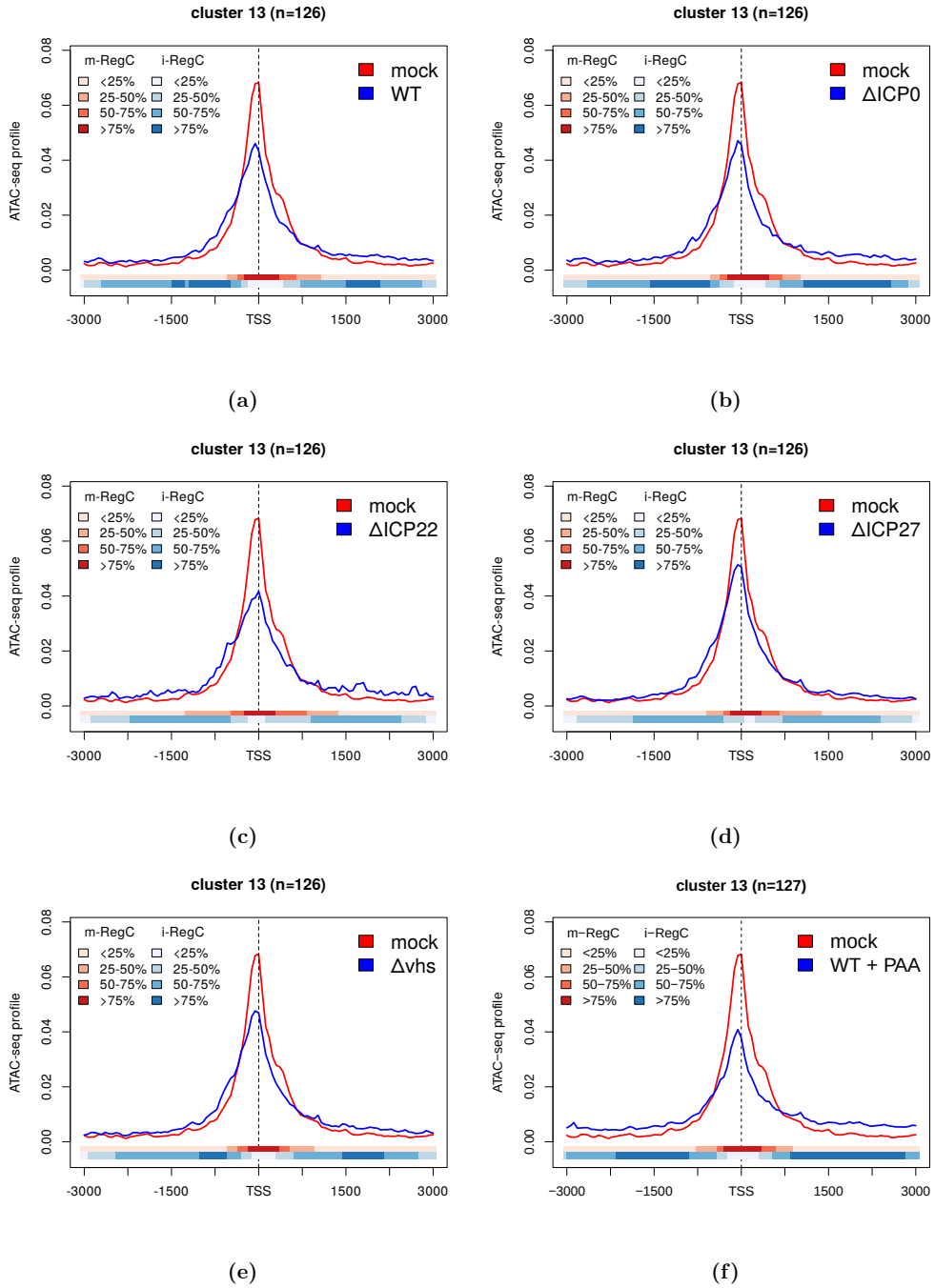
Sup. Fig 14 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 10 (pattern I). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



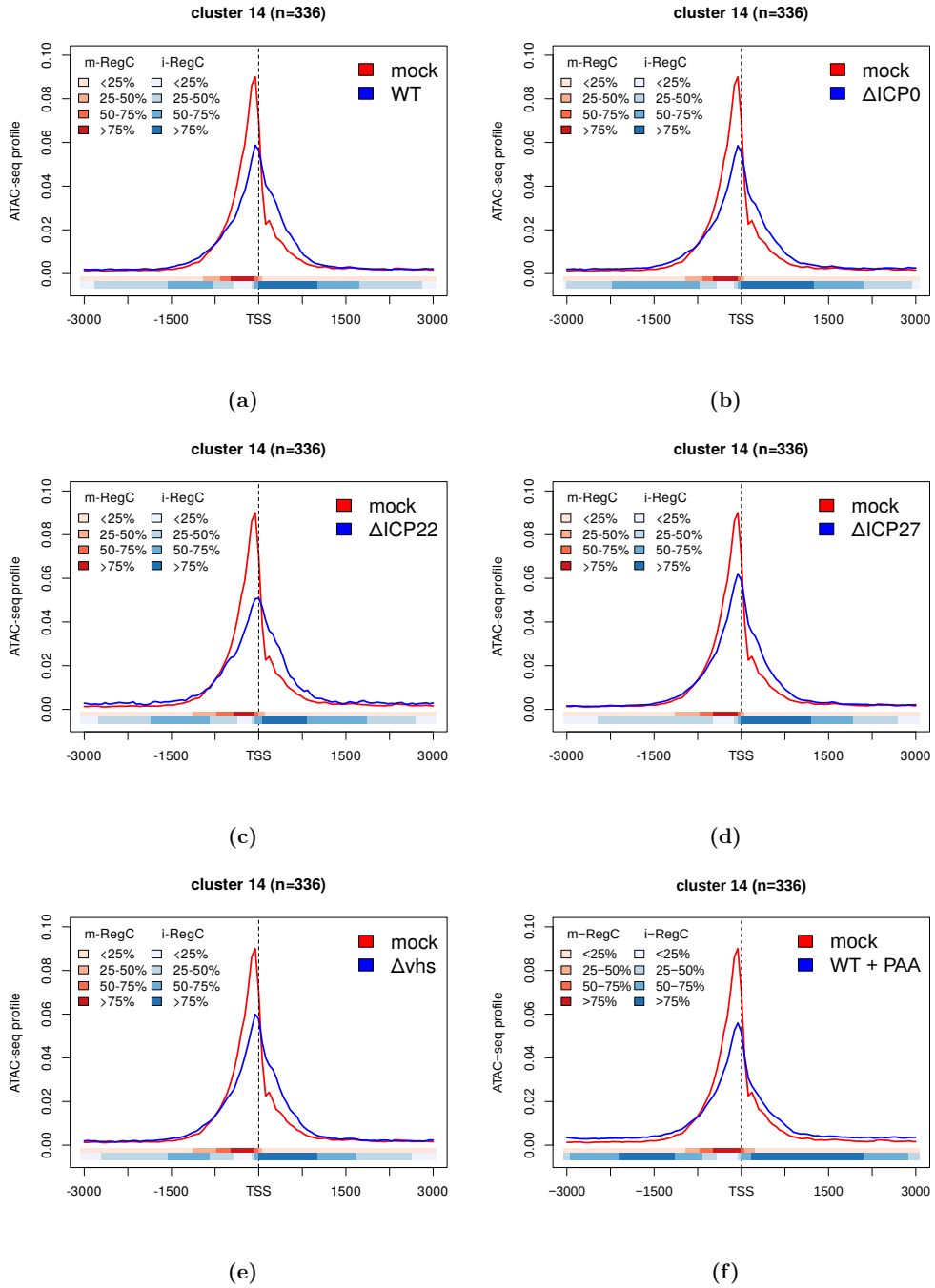
Sup. Fig 15 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 11 (pattern II). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



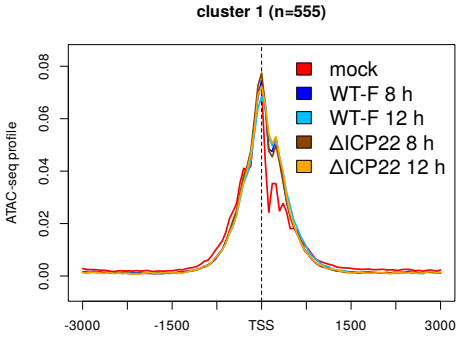
Sup. Fig 16 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 12 (pattern I). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



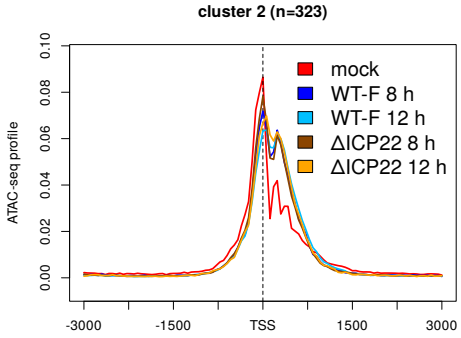
Sup. Fig 17 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 13 (pattern III). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



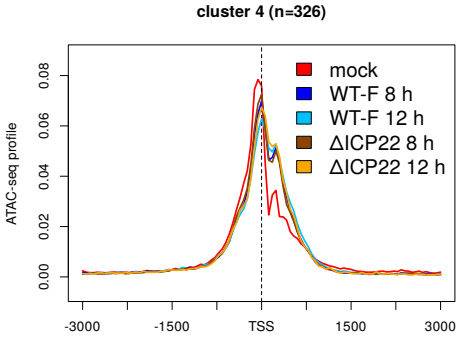
Sup. Fig 18 Metagene plots of ATAC-seq profiles in mock as well as **(a)** WT infection, **(b-e)** null mutant infections and **(f)** WT+PAA infection for cluster 14 (pattern I). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the corresponding comparison of WT or null mutant virus infection to mock.



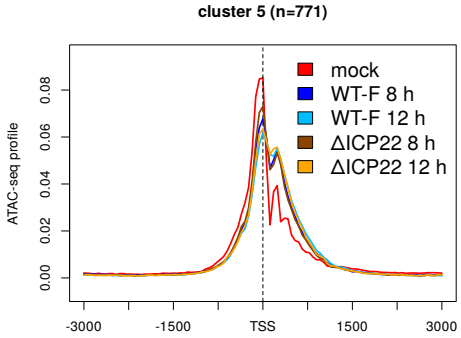
(a) pattern I



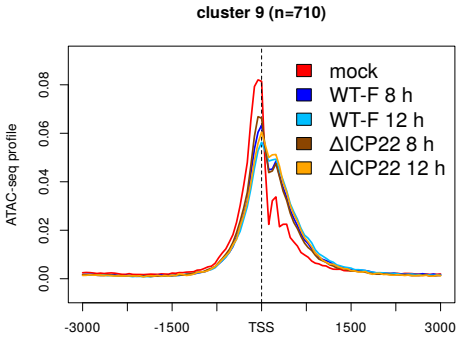
(b) pattern I



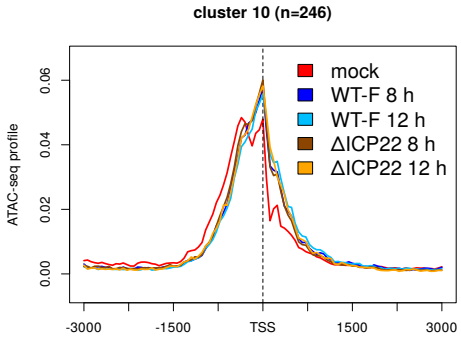
(c) pattern I



(d) pattern I

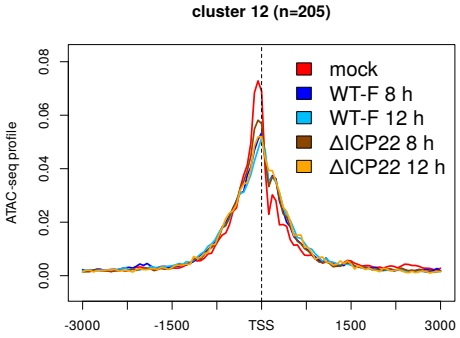


(e) pattern I

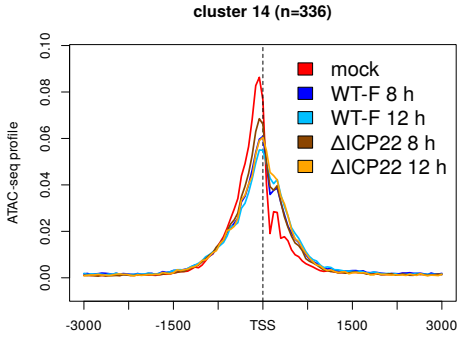


(f) pattern I

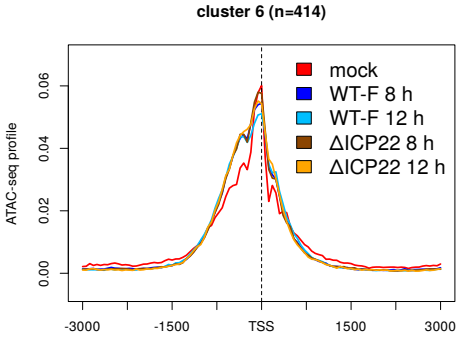
(Continued on next page)



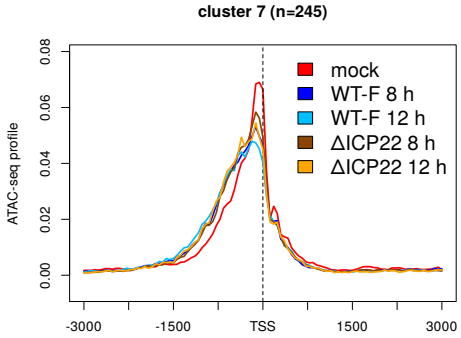
(g) pattern I



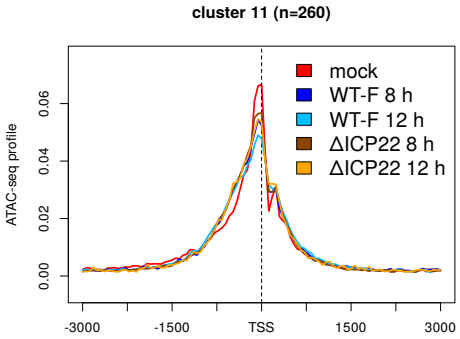
(h) pattern I



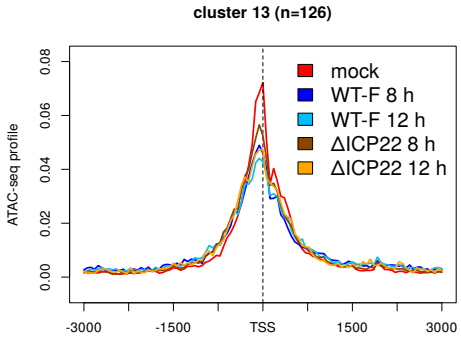
(i) pattern II



(j) pattern II

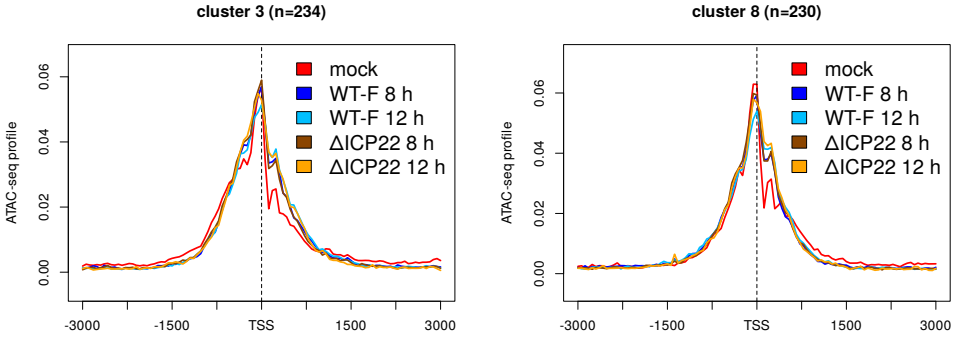


(k) pattern II



(l) pattern III

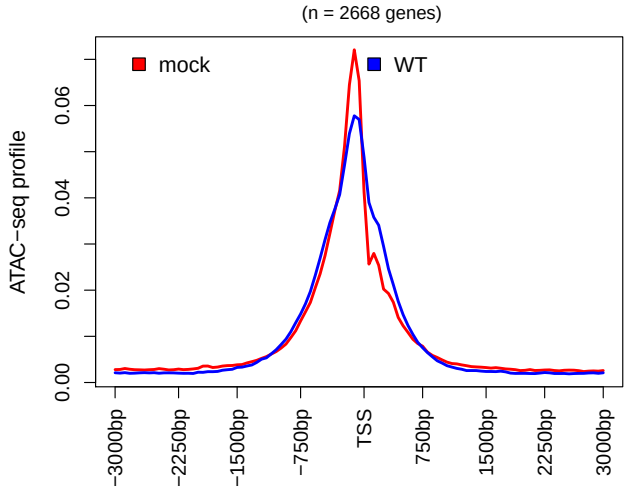
(Continued on next page)



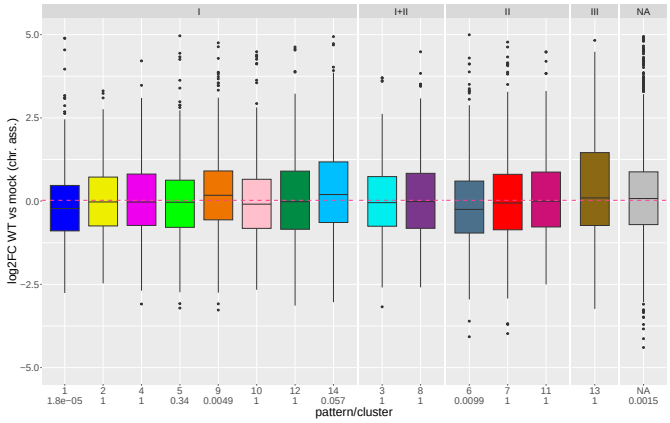
(m) pattern I + II

(n) pattern I + II

Sup. Fig 19 Metagene plots of ATAC-seq profiles in mock, WT strain F (WT-F) and ΔICP22 infection at 8 and 12 h p.i. for all clusters (indicated on top of subfigures) ordered according to patterns (indicated below subfigures). See Materials and Methods for a detailed description of metagene plots.

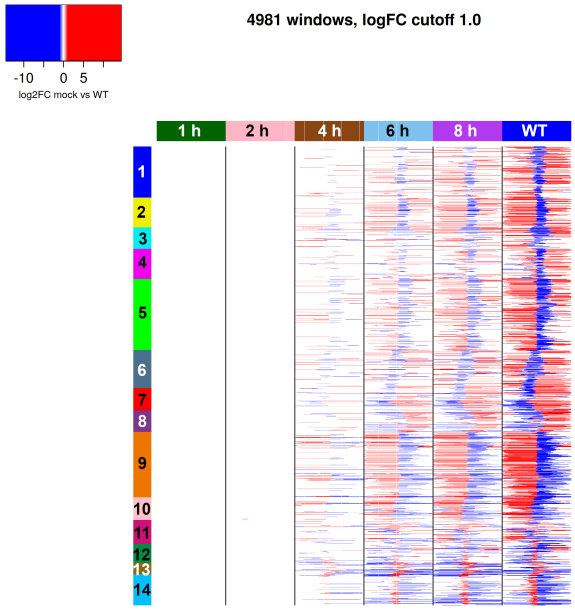


(a)

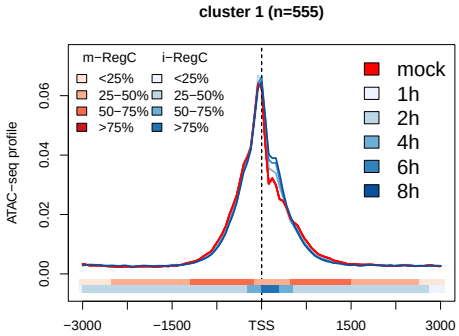


(b)

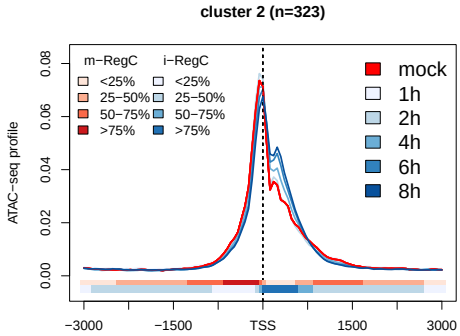
Sup. Fig 20 (a) Metagene plot of the ATAC-seq profiles for mock (red) and WT (blue) infection for all promoter windows for which no significant RegC was identified in the ATAC-seq data for any of the WT or null mutant infections compared to mock (denoted as NA group in **(b)**). These genes were not included in **Fig. 1a** and most analyses in this article. See Materials and Methods for an explanation of metagene plots. **(b)** Boxplot showing the distribution of log2 fold-changes in chromatin-associated RNA for 8 h p.i. WT infection compared to mock for all clusters (grouped by pattern) as well as genes without significant differential regions (NA group). P-values for Wilcoxon rank sum tests comparing log2 fold-changes for each group against all other analyzed genes were corrected for multiple testing using the Bonferroni method and are shown below each gene group.



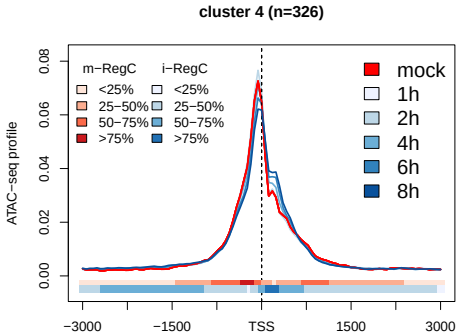
Sup. Fig 21 Heatmap showing the log2 fold-changes determined with DEXSeq on the ATAC-seq time-course data (1, 2, 4, 6 and 8 h p.i. WT infection compared to mock) for the differential regions (m- and i-RegC) determined in the WT vs. mock comparison shown in **Fig. 1a**. Each row shows log2 fold-changes for the same m- and i-RegCs at the different time-points of the time-course compared to mock. For comparison, log2 fold-changes for the WT vs. mock comparison from **Fig. 1a**/**Sup. Fig. 1d** are also shown. Colored rectangles on top indicate the time-point or whether the WT vs. mock comparison is shown. Statistically significant differential regions are colored according to the log2 fold-change determined by DEXSeq. Here, the color scale is continuous between -1 and 1 and all log2 fold-changes > 1 are colored the same red and all log2 fold-changes < 1 the same blue. Promoter windows are ordered as in **Fig. 1a** and clusters from **Fig. 1a** are annotated as colored and numbered rectangles on the left.



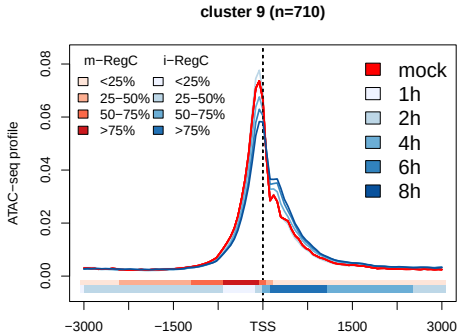
(a) pattern I



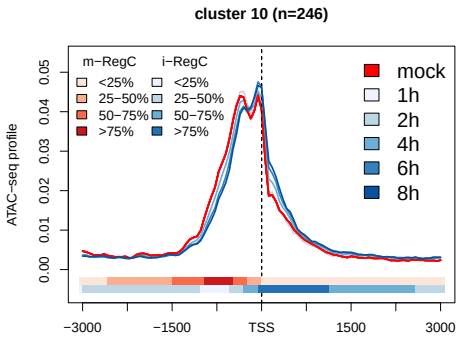
(b) pattern I



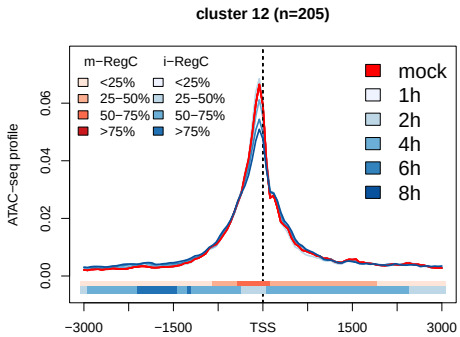
(c) pattern I



(d) pattern I

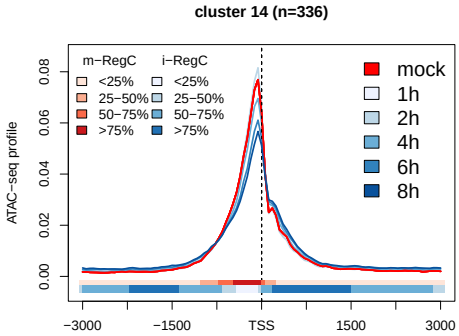


(e) pattern I

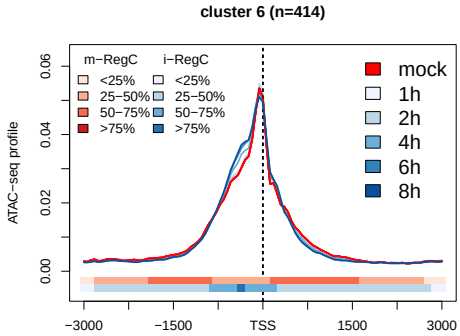


(f) pattern I

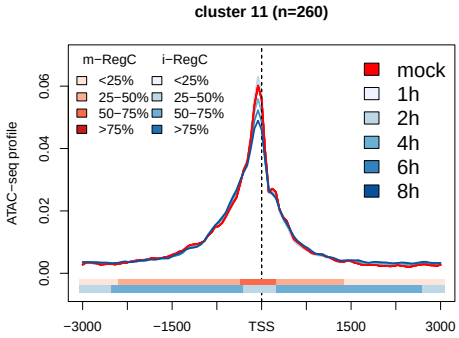
(Continued on next page)



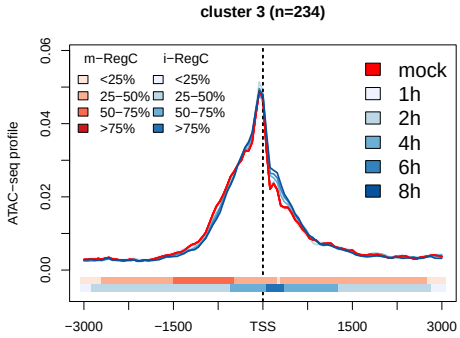
(g) pattern I



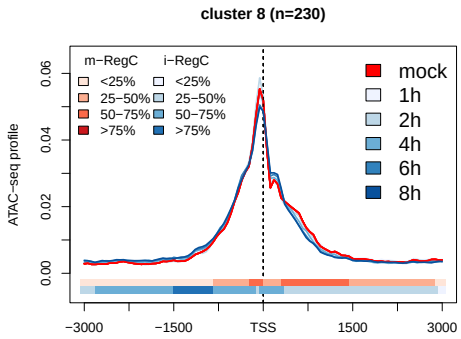
(h) pattern II



(i) pattern II

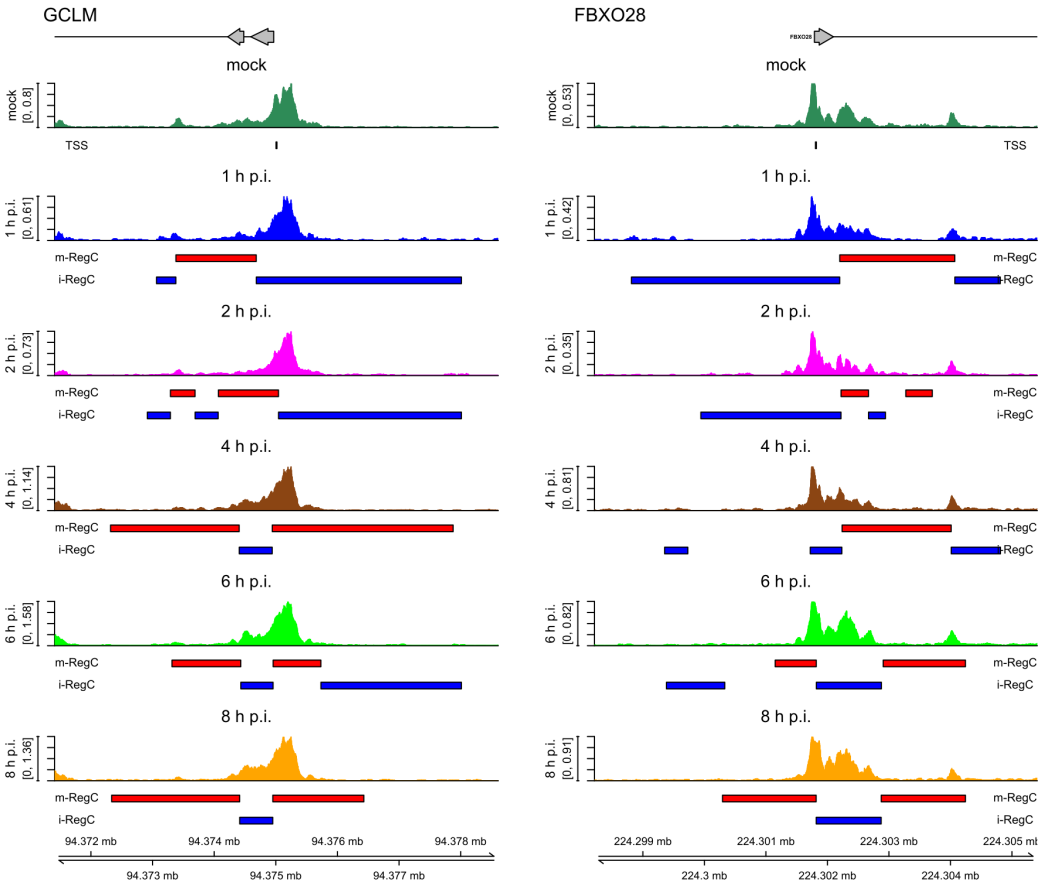


(j) pattern I + II



(k) pattern I + II

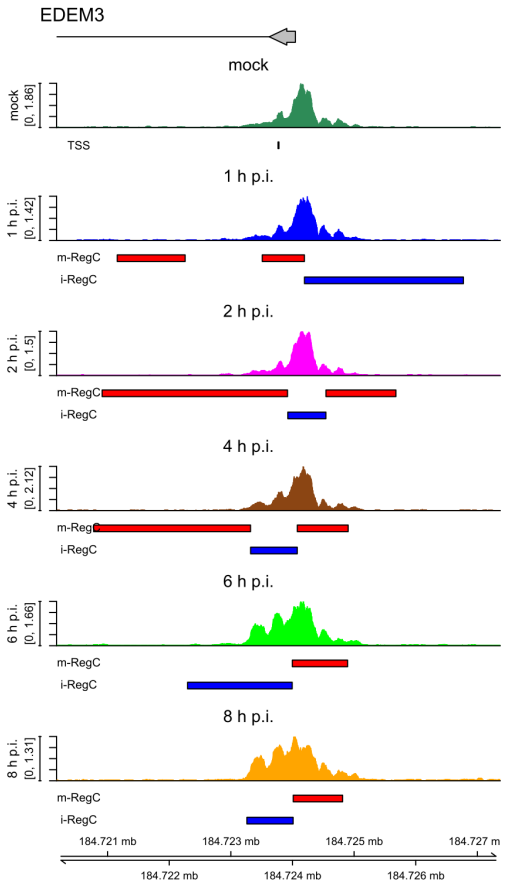
Sup. Fig 22 Metagene plots of ATAC-seq profiles in mock infection (red) and all time-points of infection (blue shades) from the ATAC-seq time-course experiment for all clusters (indicated on top of subfigures) ordered according to patterns (indicated below subfigures). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC (red or blue, respectively) at that position for the comparison of 8 h p.i. to mock.



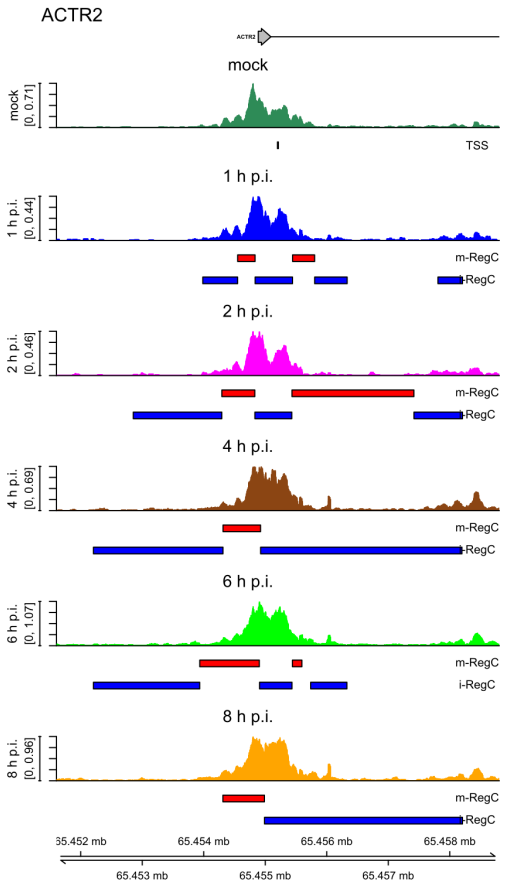
(a) pattern I, cluster 2

(b) pattern I, cluster 5

(Continued on next page)

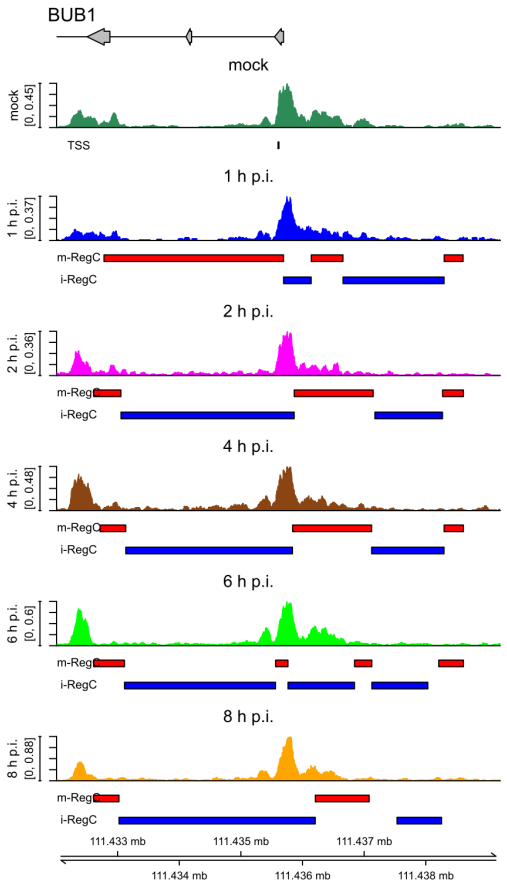


(c) pattern I, cluster 10

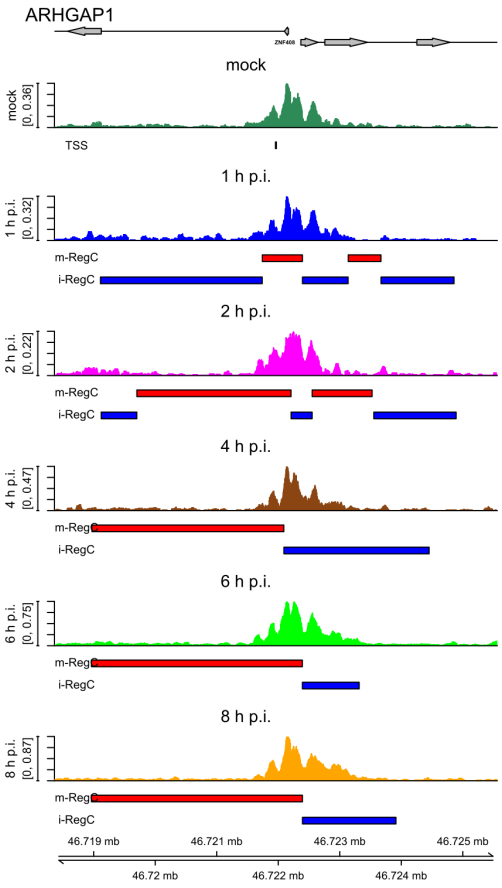


(d) pattern I, cluster 12

(Continued on next page)

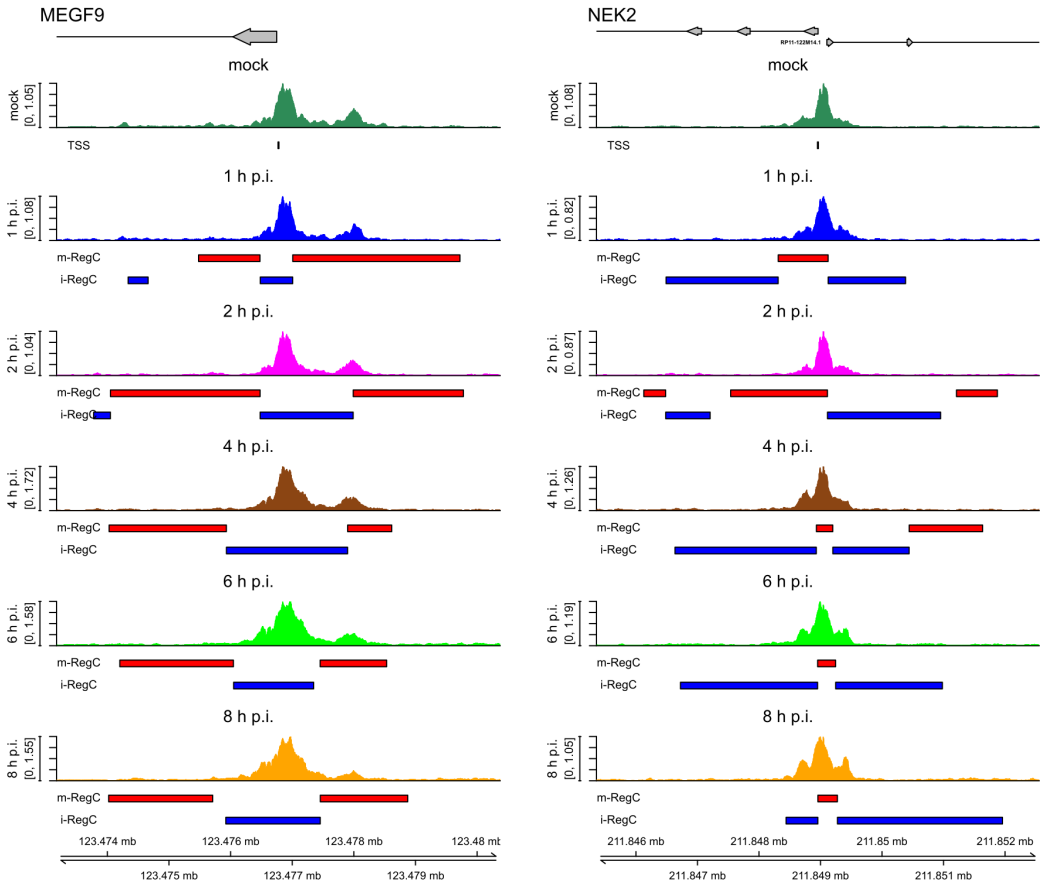


(e) pattern II, cluster 6



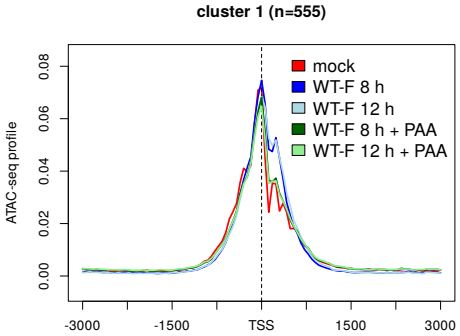
(f) pattern II, cluster 7

(Continued on next page)

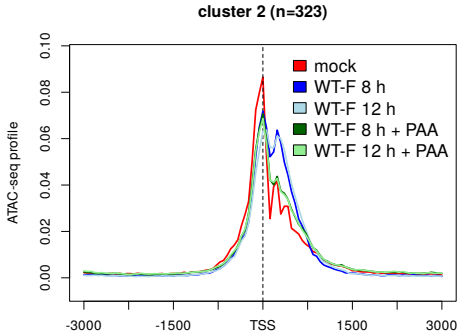


(g) pattern I + II, cluster 3 **(h)** pattern I + II, cluster 8

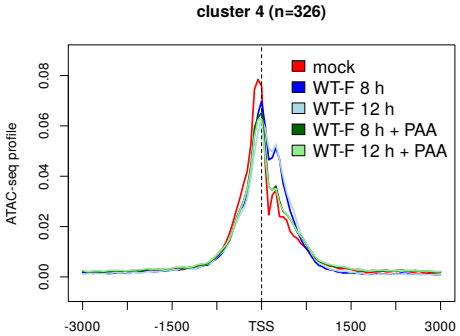
Sup. Fig 23 Read coverage in a ± 3.6 kb window around the TSS in ATAC-seq data for mock (dark green), 1 h (blue), 2 h (magenta), 4 h (brown), 6 h (green), and 8 h p.i. (orange) for example genes with pattern I **(a-d)**, pattern II **(e,f)** and combined patterns I and II **(g,h)**. Read coverage was normalized to total number of mapped reads for each sample and averaged between replicates. The TSS used in the analysis is indicated by a short vertical line below the read coverage track for mock infection. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns, and direction of transcription is indicated by arrowheads. The name of the gene whose promoter window was analyzed is indicated in larger font on the top left and sometimes beside the gene annotation. Names for other genes overlapping the input window are also indicated. Below each read coverage track m- (red bars) and i-RegCs (blue bars) are indicated for the comparison of the corresponding time-point of infection to mock. Corresponding read coverage plots in ATAC-seq data for mock, WT, ΔICP0, ΔICP22, ΔICP27, Δ*vhs* and WT+PAA infection are shown in **Sup. Fig. 4**.



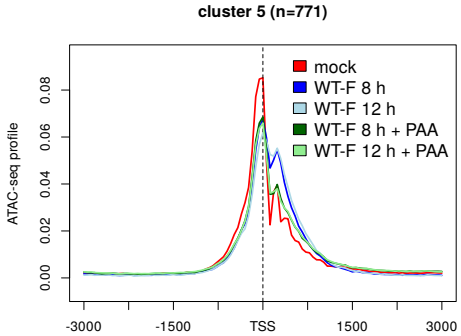
(a) pattern I



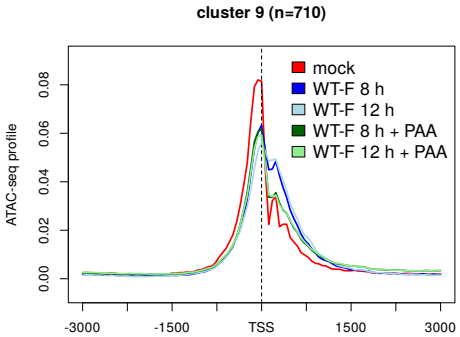
(b) pattern I



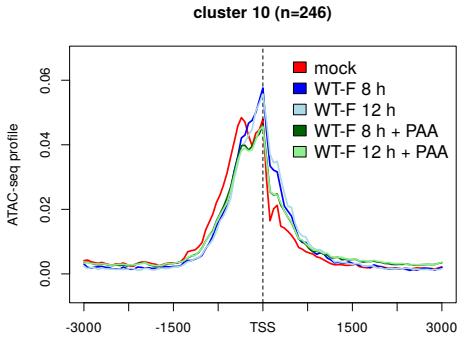
(c) pattern I



(d) pattern I

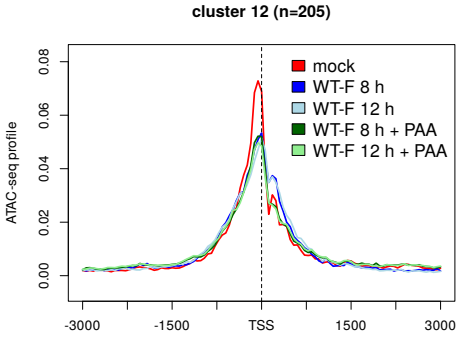


(e) pattern I

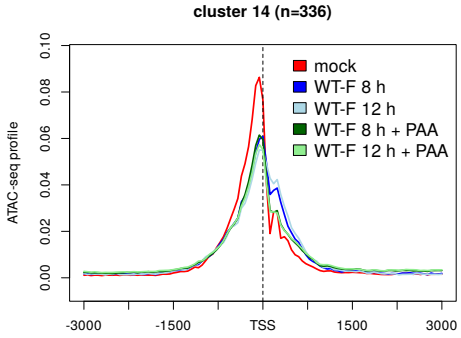


(f) pattern I

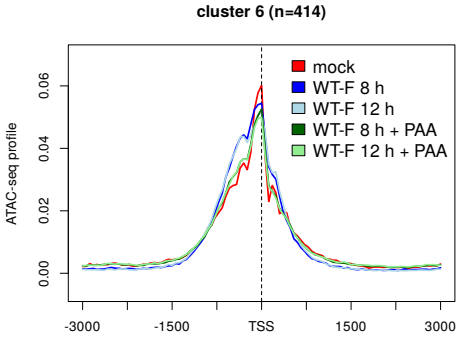
(Continued on next page)



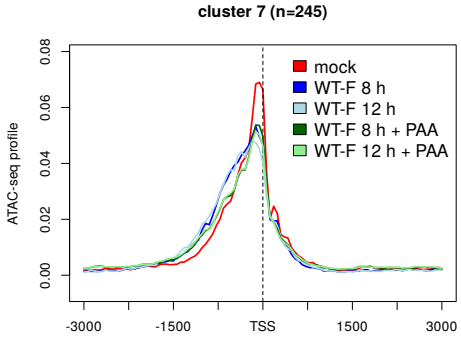
(g) pattern I



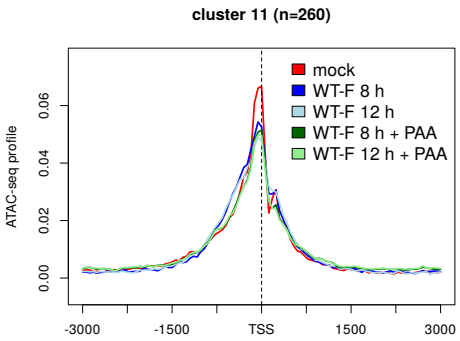
(h) pattern I



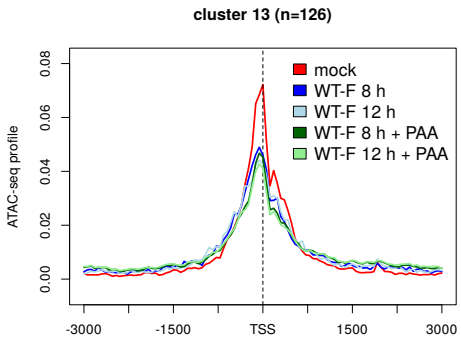
(i) pattern II



(j) pattern II

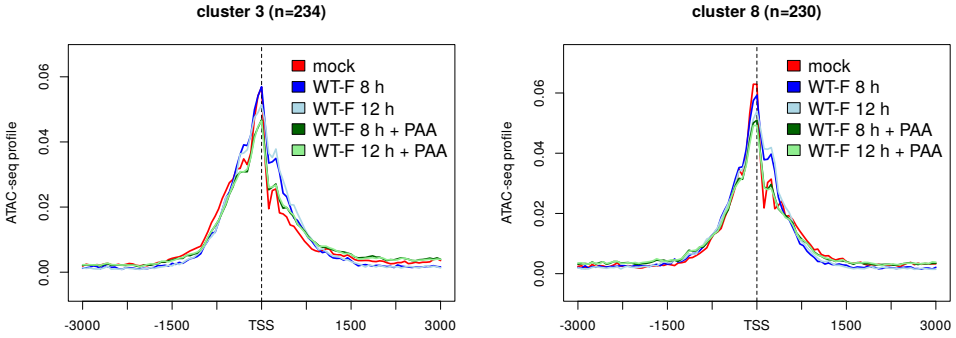


(k) pattern II



(l) pattern III

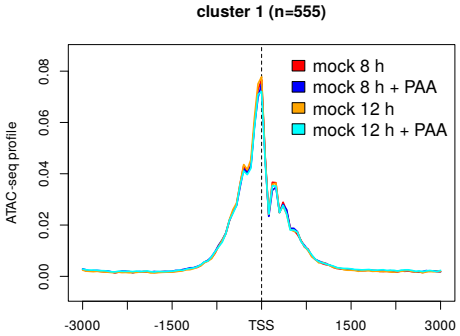
(Continued on next page)



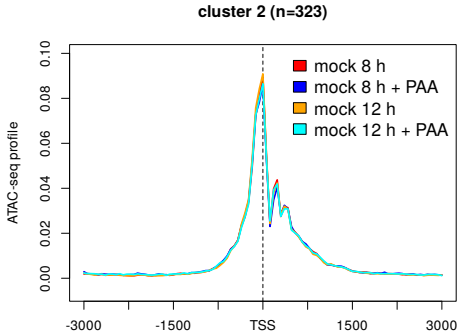
(m) pattern I + II

(n) pattern I + II

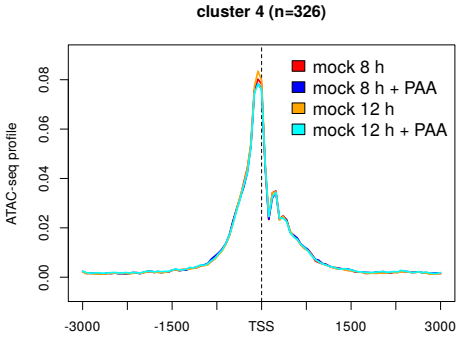
Sup. Fig 24 Metagene plots showing ATAC-seq profiles in mock and WT-F infection at 8 and 12 h p.i. ± PAA for all clusters (indicated on top of subfigures) ordered according to patterns (indicated below subfigures). See Materials and Methods for a detailed description of metagene plots.



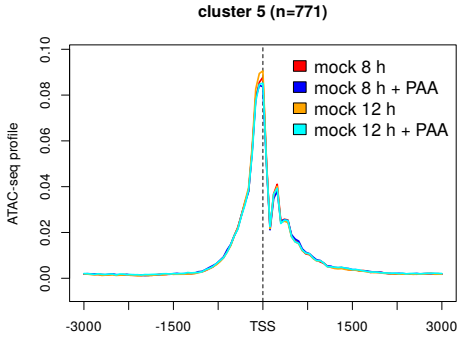
(a) pattern I



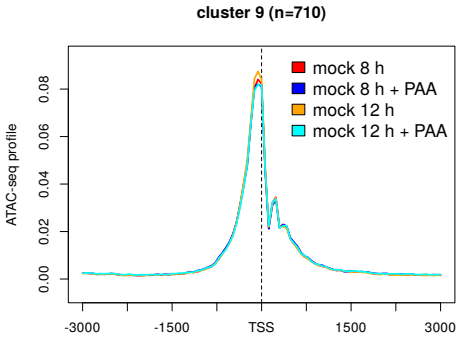
(b) pattern I



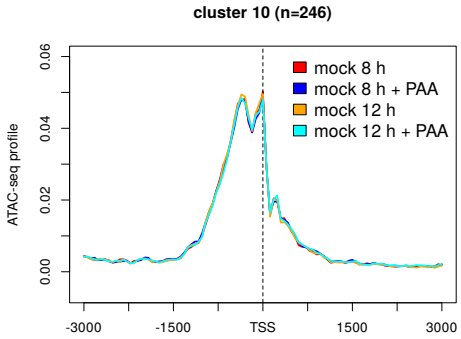
(c) pattern I



(d) pattern I

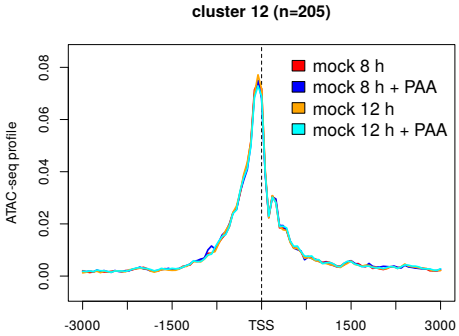


(e) pattern I

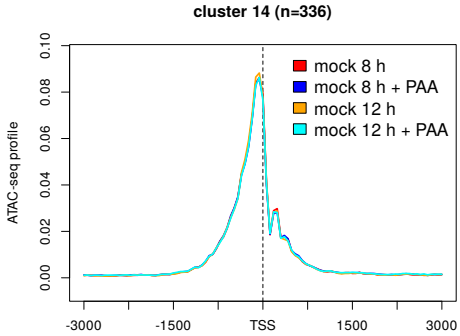


(f) pattern I

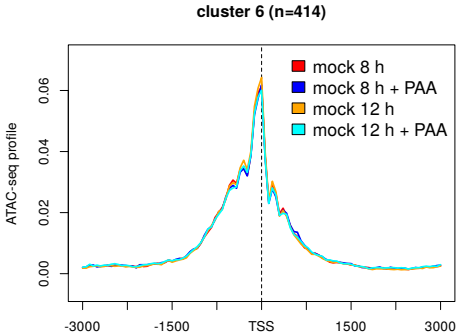
(Continued on next page)



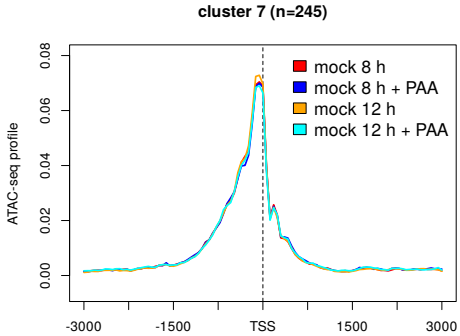
(g) pattern I



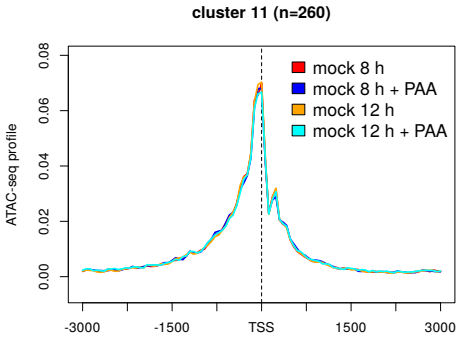
(h) pattern I



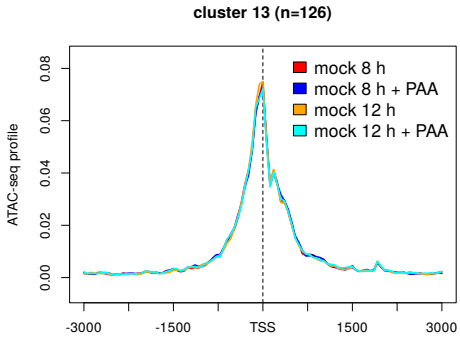
(i) pattern II



(j) pattern II

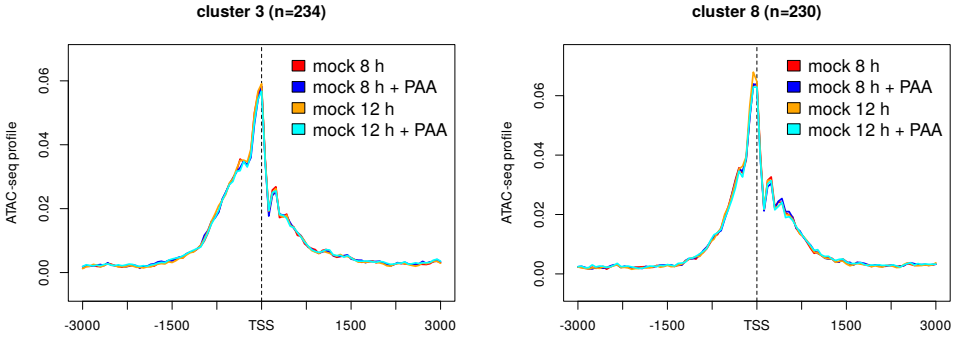


(k) pattern II



(l) pattern III

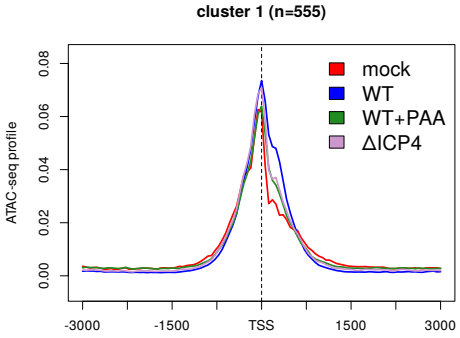
(Continued on next page)



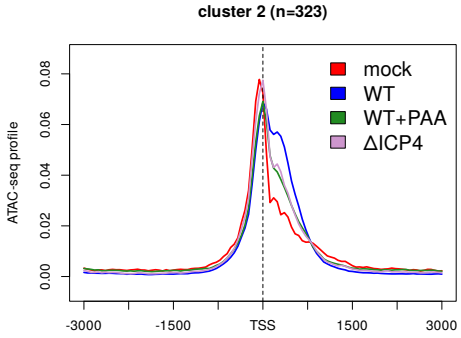
(m) pattern I + II

(n) pattern I + II

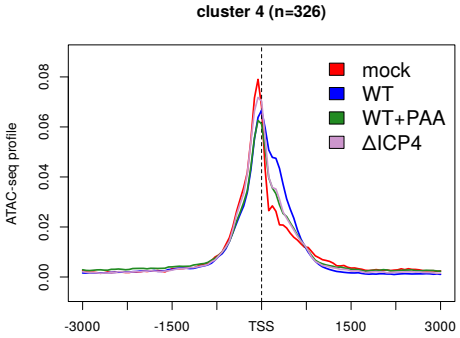
Sup. Fig 25 Metagene plots showing ATAC-seq profiles in mock infection at 8 and 12 h p.i. ± PAA for all clusters (indicated on top of subfigures) ordered according to patterns (indicated below subfigures). See Materials and Methods for a detailed description of metagene plots.



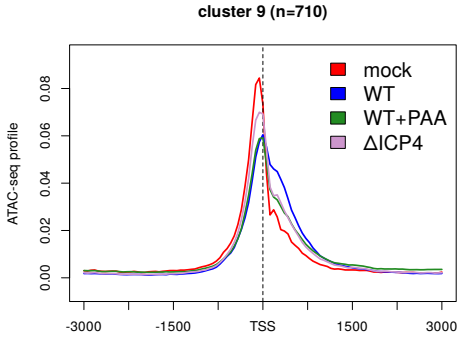
(a) pattern I



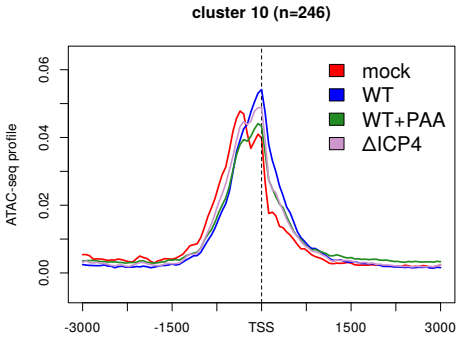
(b) pattern I



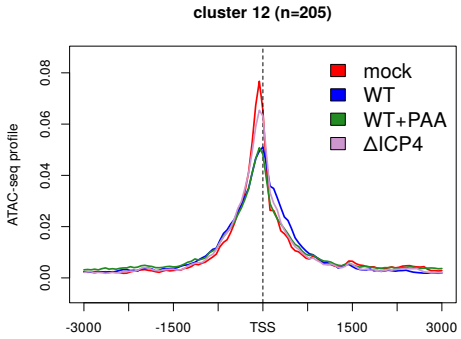
(c) pattern I



(d) pattern I

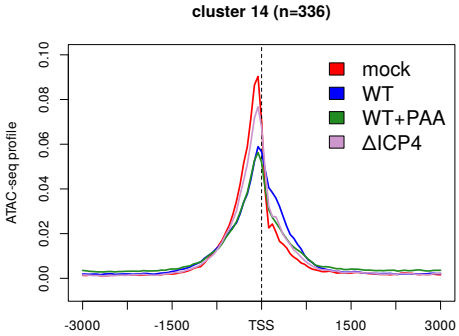


(e) pattern I

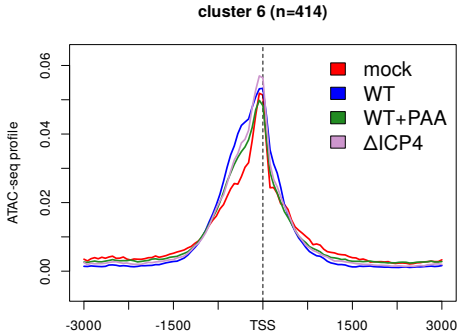


(f) pattern I

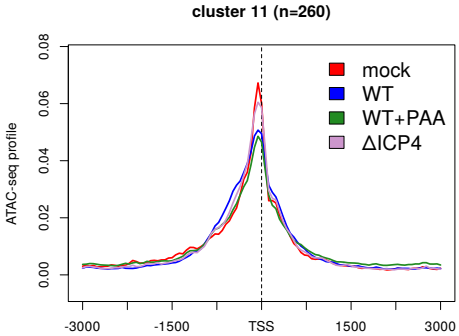
(Continued on next page)



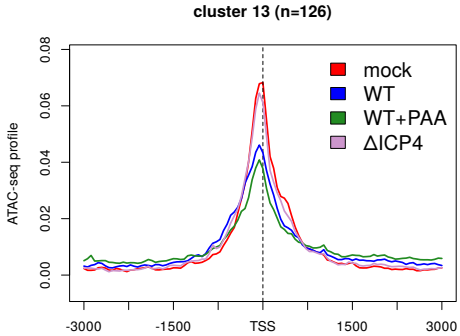
(g) pattern I



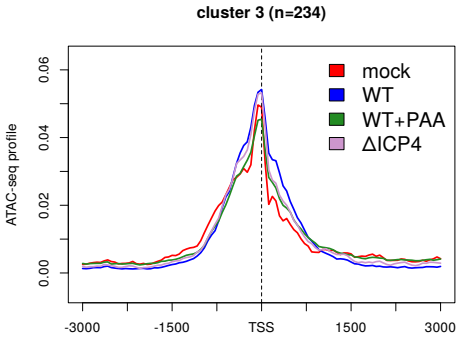
(h) pattern II



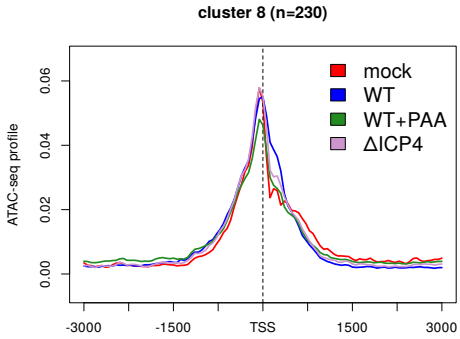
(i) pattern II



(j) pattern III

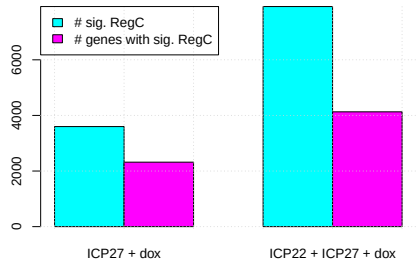


(k) pattern I + II

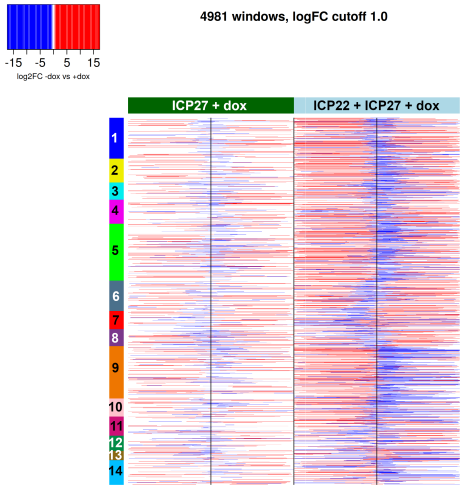


(l) pattern I + II

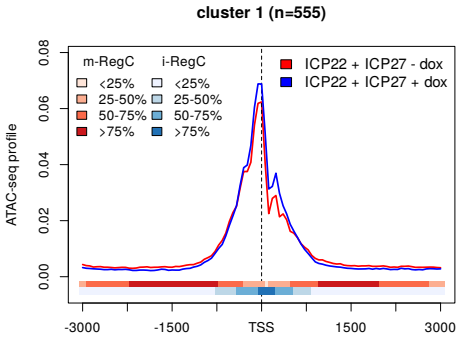
Sup. Fig 26 Metagene plots of ATAC-seq profiles for mock (red), WT (blue), WT+PAA (green) und ΔICP4 infection (violet) for all clusters (indicated on top of subfigures) ordered according to patterns (indicated below subfigures). See Materials and Methods for a detailed description of metagene plots.



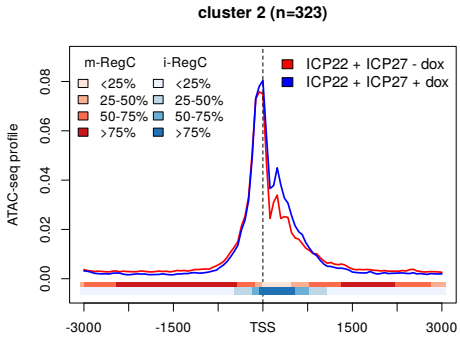
(a)



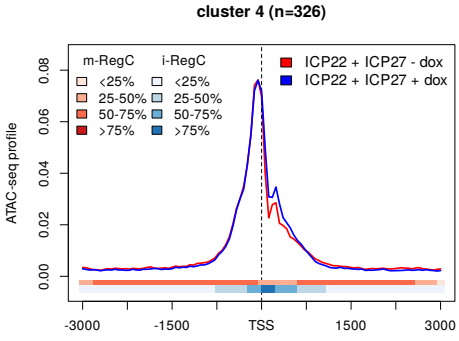
(b)



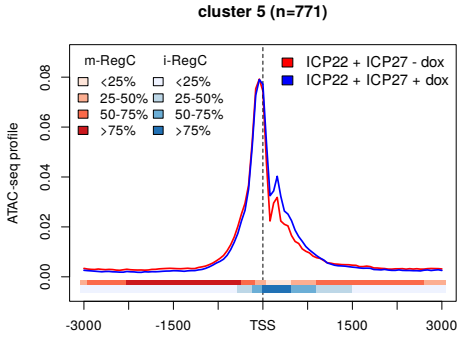
(c) pattern I



(d) pattern I

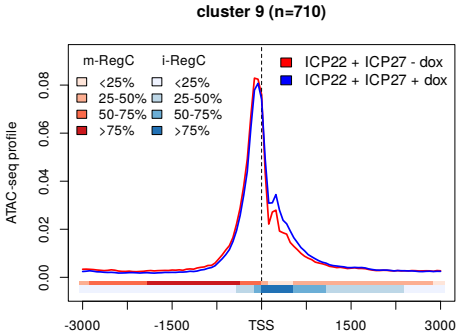


(e) pattern I

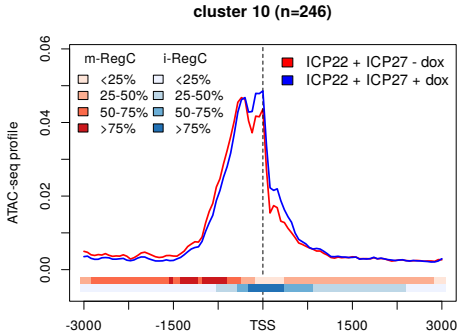


(f) pattern I

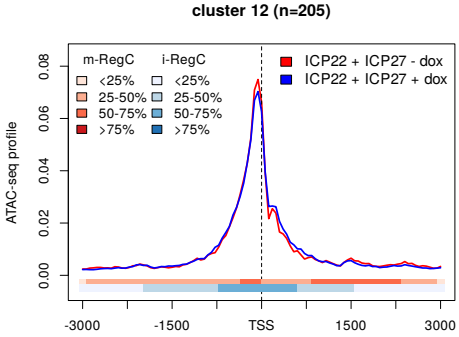
(Continued on next page)



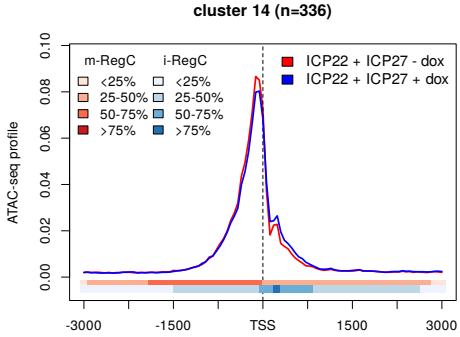
(g) pattern I



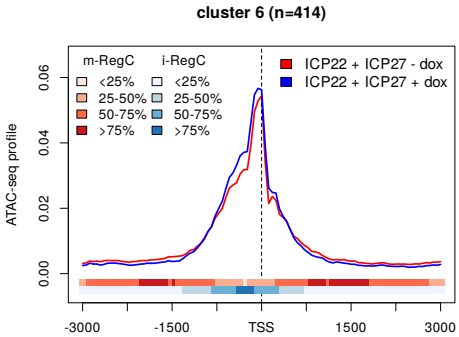
(h) pattern I



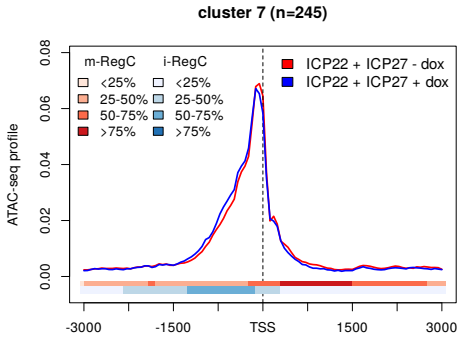
(i) pattern I



(j) pattern I

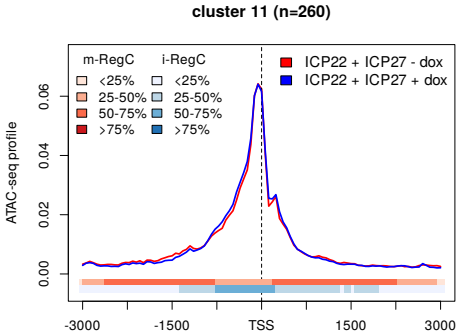


(k) pattern II

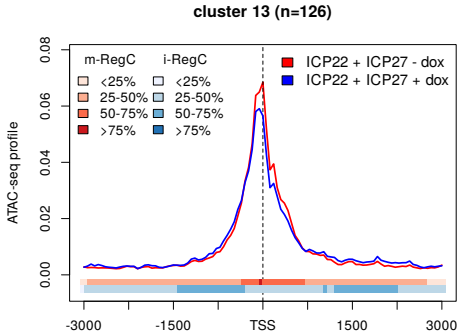


(l) pattern II

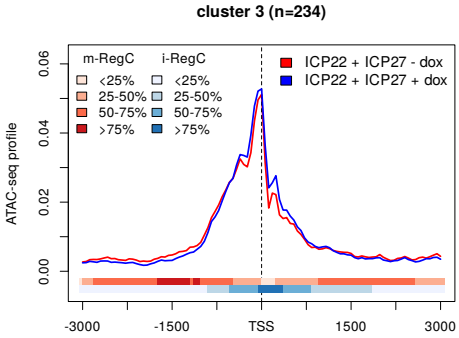
(Continued on next page)



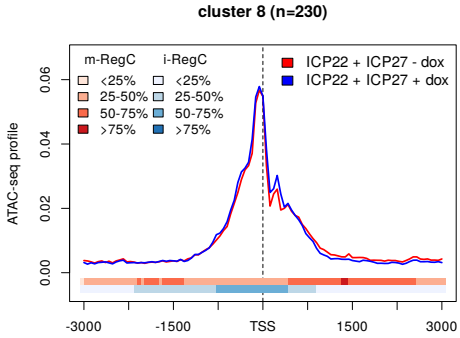
(m) pattern II



(n) pattern III

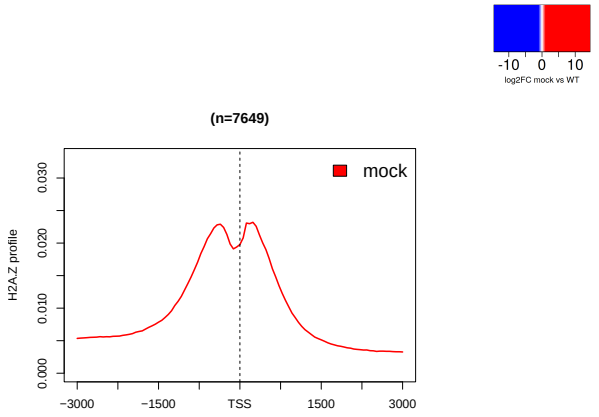


(o) pattern I + II

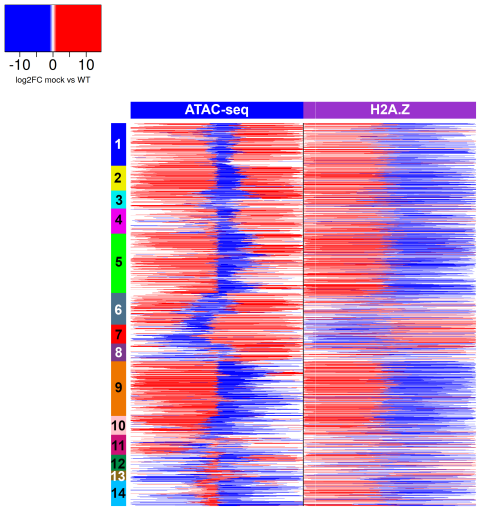


(p) pattern I + II

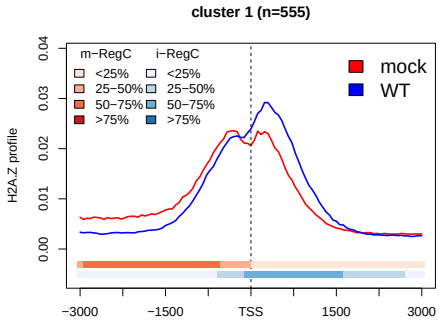
Sup. Fig 27 (a) Number of significant RegC and genes with significant RegC identified by RegCFinder for T-HF-ICP27 cells (left) and T-HF-ICP22/ICP27 cells (right) upon dox exposure. **(b)** Heatmap visualizing log2 fold-changes for the differential regions identified for T-HF-ICP27 cells (left half) and T-HF-ICP22/ICP27 cells (right half) upon dox exposure for the promoter windows included in **Fig. 1a**. One row of this heatmap represents results for a particular input window. Black vertical lines in the center of each half of the heatmap indicate the position of the TSS. Regions corresponding to statistically significant differential regions (adj. p. < 0.01) are colored according to the log2 fold-change determined by DEXSeq. Here, the color scale is continuous between -1 and 1 and log2 fold-changes > 1 are colored the same red and log2 fold-changes < 1 the same blue. Promoter windows are ordered as in **Fig. 1a** and clusters from **Fig. 1a** are shown as colored and numbered rectangles on the left. **(c-p)** Metagene plots of ATAC-seq profiles for T-HF-ICP22/ICP27 cells with (blue) and without (red) dox exposure for all clusters (indicated on top of subfigures) ordered according to patterns (indicated below subfigures). See Materials and Methods for a detailed description of metagene plots. The colored bands below the metagene curves indicate the percentage of genes having an m-RegC (red, decreased upon dox exposure) or i-RegC (blue, increased upon dox exposured) or at that position.



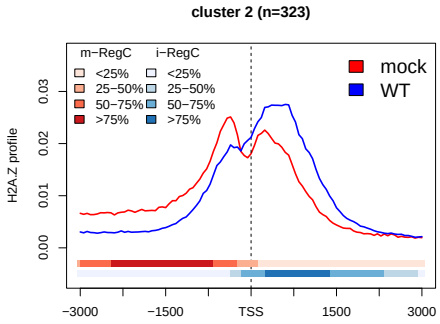
(a)



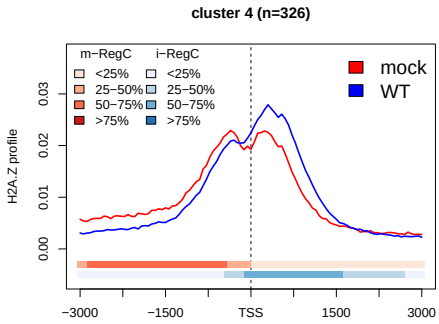
(b)



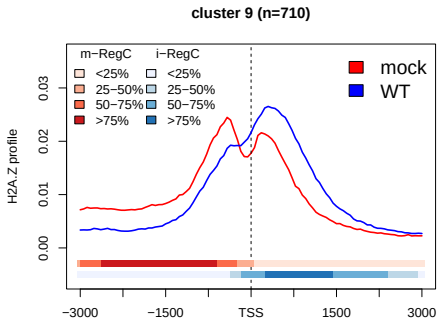
(c) pattern I



(d) pattern I

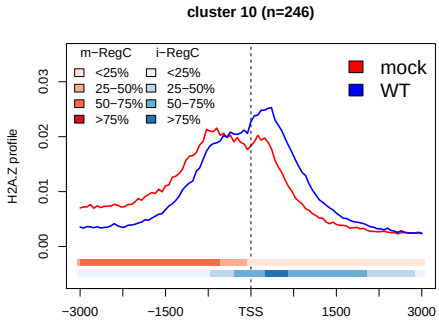


(e) pattern I

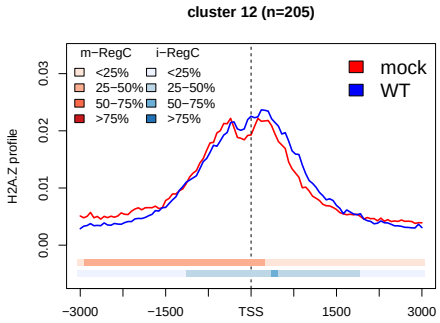


(f) pattern I

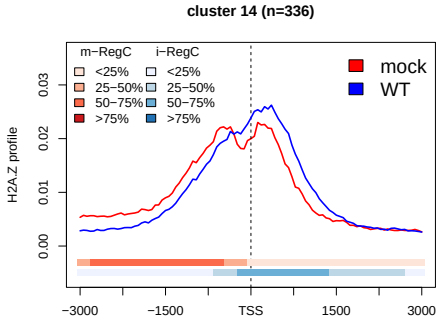
(Continued on next page)



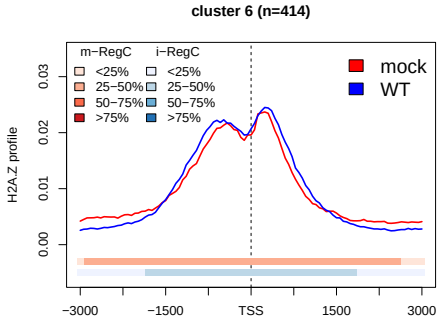
(g) pattern I



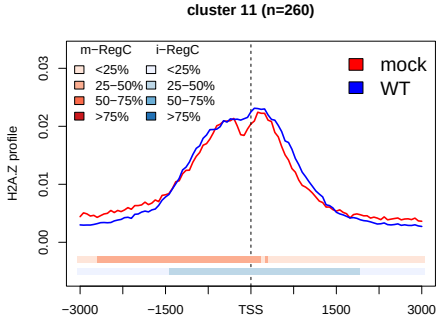
(h) pattern I



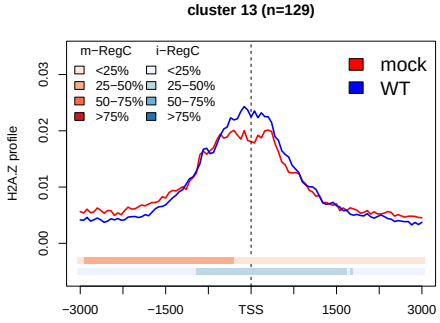
(i) pattern I



(j) pattern II

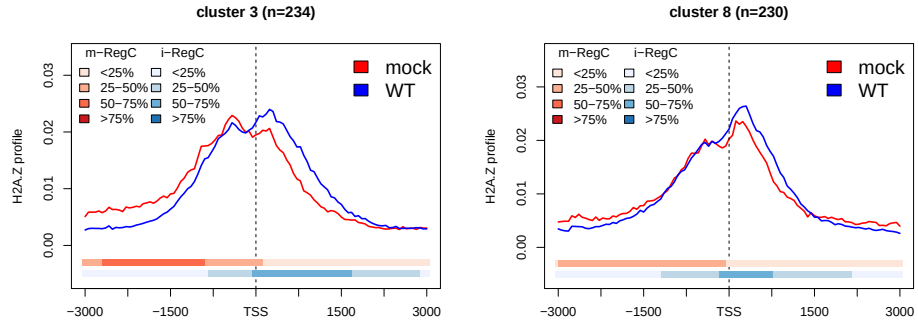


(k) pattern II



(l) pattern III

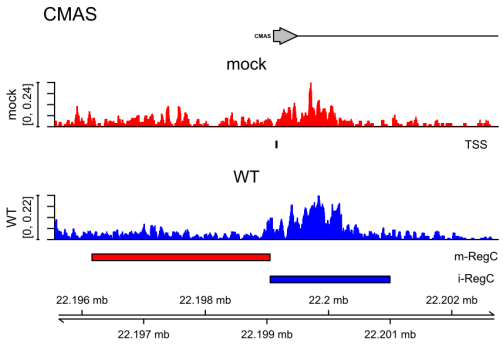
(Continued on next page)



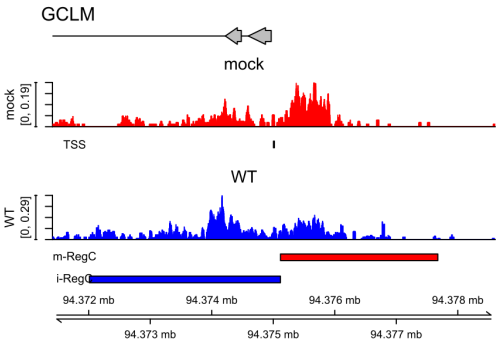
(m) pattern I + II

(n) pattern I + II

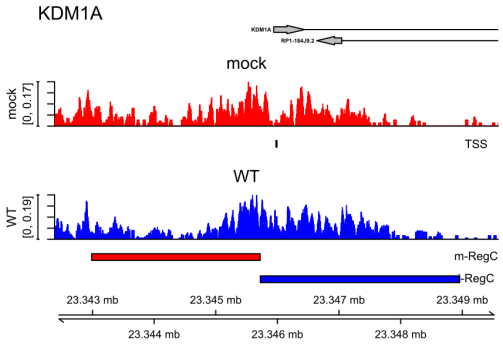
Sup. Fig 28 (a) Metagene plot of H2A.Z profiles for mock infection for all analyzed promoter windows. See Materials and Methods for an explanation of metagene plots. **(b)** Heatmap showing log2 fold-changes for differential regions (m- and i-RegC) identified in the WT vs. mock comparison on the ATAC-seq (left half) and H2A.Z ChIPmentation data (right half) for the promoter windows included in **Fig. 1a**. Statistically significant (adj. p. < 0.01) differential regions are colored according to the log2 fold-change determined by DEXSeq. Here, the color scale is continuous between -1 and 1 and log2 fold-changes > 1 are colored the same red and log2 fold-changes < 1 the same blue. Promoter windows are ordered as in **Fig. 1a** and clusters from **Fig. 1a** are annotated as colored and numbered rectangles on the left. **(c-n)** Metagene plots of H2A.Z profiles in mock (red) and WT (blue) infection for clusters with pattern I (**c-i**, cluster 5 is shown in **Fig. 4d**), pattern II (**j,k**, cluster 7 is shown in **Fig. 4e**), pattern III **(l)** and combined patterns I and II **(m,n)**. See Materials and Methods for an explanation of metagene plots. The colored bands below the metagene curves in each panel indicate the percentage of genes having an m- or i-RegC at that position in the comparison of WT vs. mock in the H2A.Z ChIPmentation data.



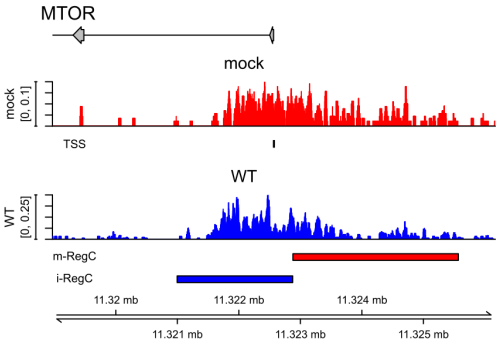
(a) pattern I, cluster 1



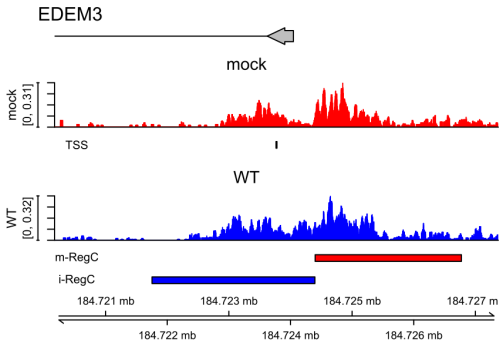
(b) pattern I, cluster 2



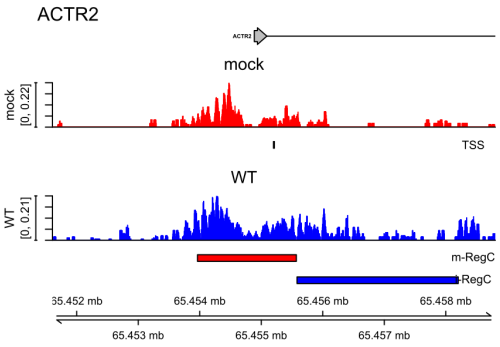
(c) pattern I, cluster 4



(d) pattern I, cluster 9

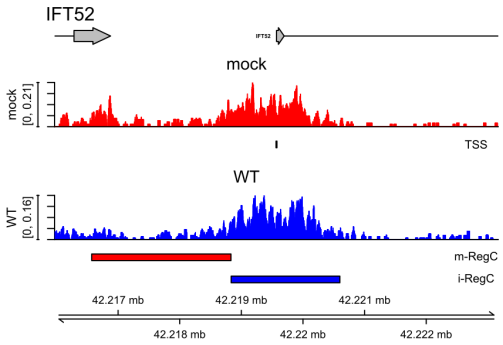


(e) pattern I, cluster 10

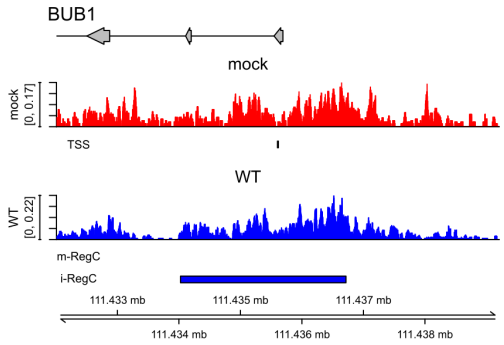


(f) pattern I, cluster 12

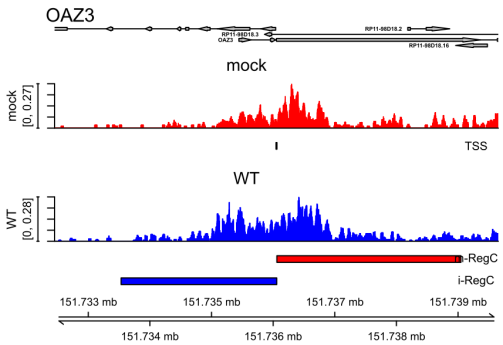
(Continued on next page)



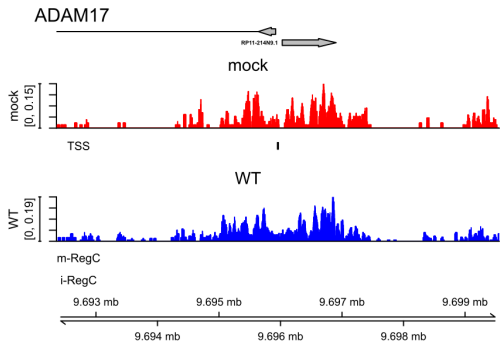
(g) pattern I, cluster 14



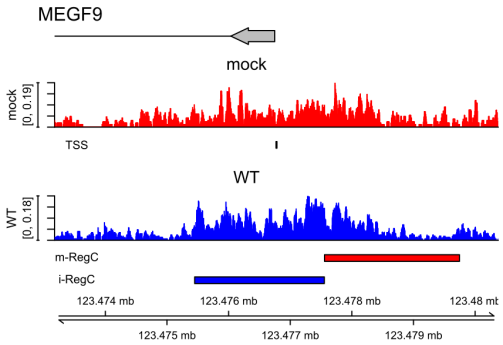
(h) pattern II, cluster 6



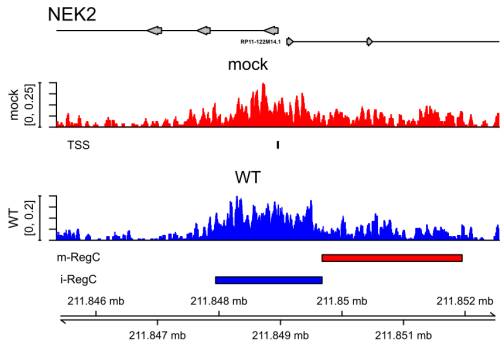
(i) pattern II, cluster 11



(j) pattern III, cluster 13



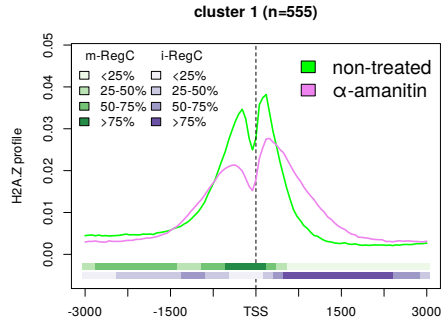
(k) pattern I + II, cluster 3



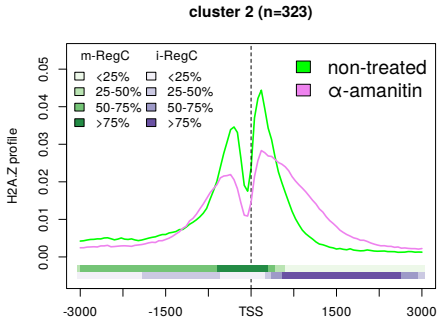
(l) pattern I + II, cluster 8

(Continued on next page)

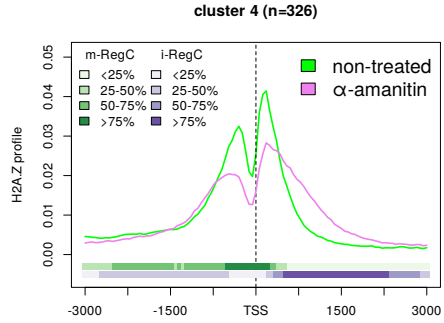
Sup. Fig 29 Read coverage in a ± 3.6 kb window around the TSS in H2A.Z ChIPmentation data for mock (red) and WT (blue) infection for example genes with pattern I **(a-g)**, pattern II **(h,i)**, pattern III **(j)** and combined patterns I and II **(k,l)**. Read coverage was normalized to total number of mapped reads for each sample and averaged between replicates. The TSS used in the analysis is indicated by a short vertical line below the read coverage track for mock infection. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns, and direction of transcription is indicated by arrowheads. The name of the gene whose promoter window was analyzed is indicated in larger font on the top left and – if not clear from the context – beside the gene annotation. Names for other genes overlapping the input window are also indicated if necessary. Below each read coverage track m- (red bars) and i-RegCs (blue bars) are indicated for the comparison of WT vs. mock om the H2A.Z ChIPmentation data. Corresponding read coverage plots in ATAC-seq data for mock, WT, ΔICP0, ΔICP22, ΔICP27, Δ*vhs* and WT+PAA infection are shown in **Sup. Fig. 4**.



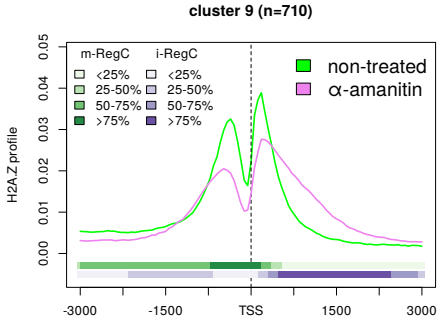
(a) pattern I



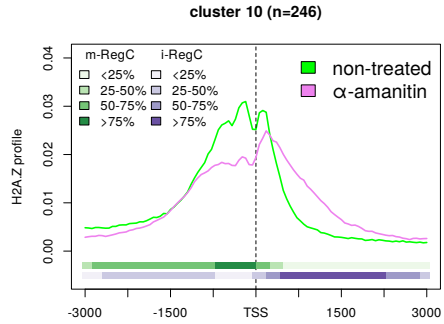
(b) pattern I



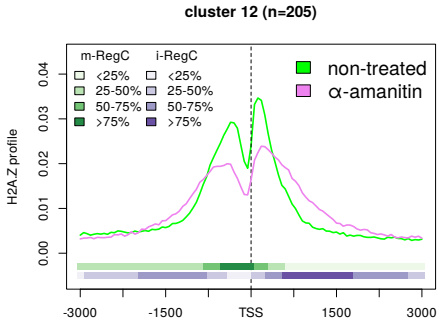
(c) pattern I



(d) pattern I

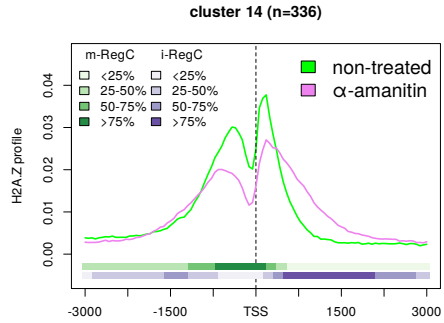


(e) pattern I

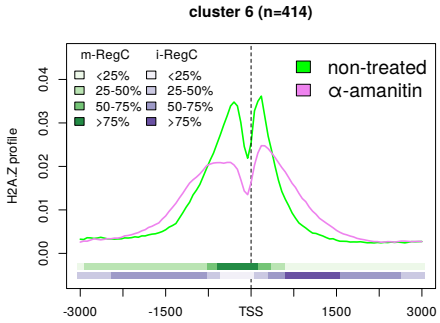


(f) pattern I

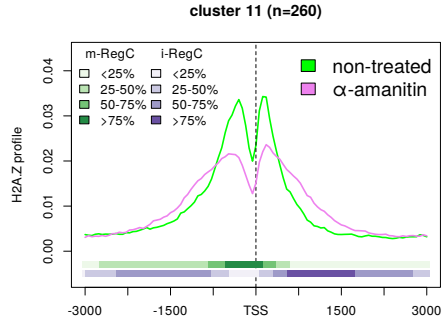
(Continued on next page)



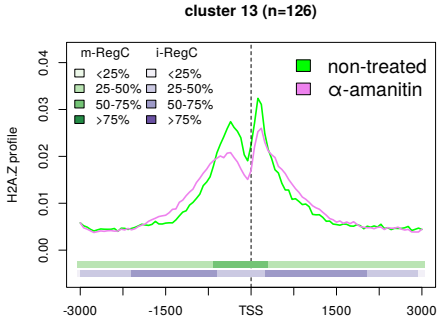
(g) pattern I



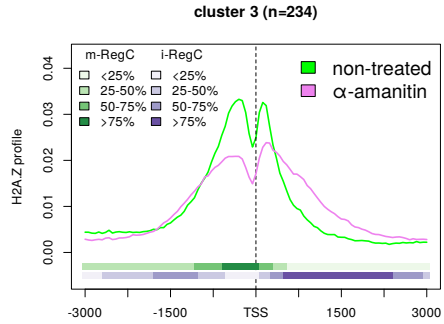
(h) pattern II



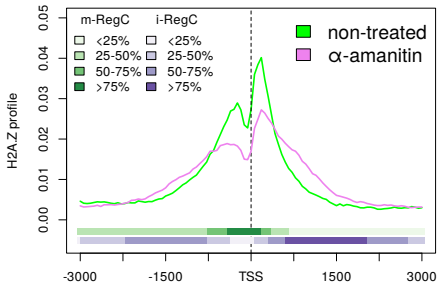
(i) pattern II



(j) pattern III

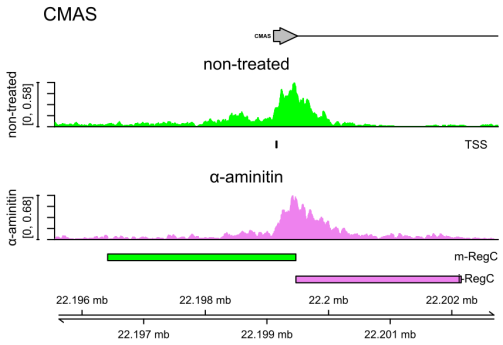


(k) pattern I + II

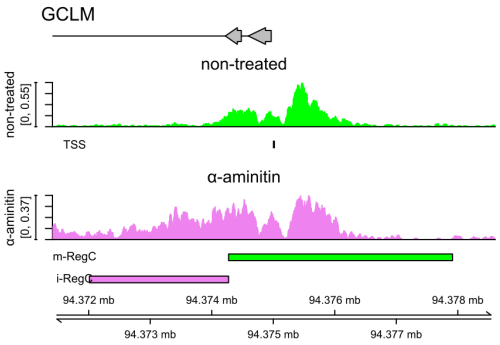


(l) pattern I + II

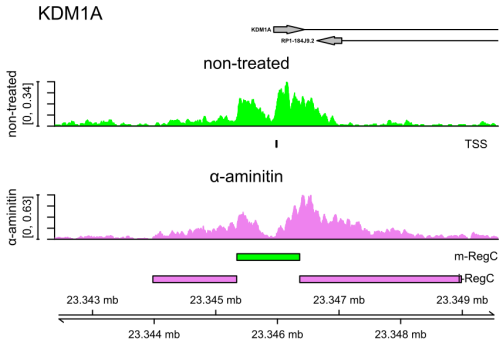
Sup. Fig 30 Metagene plots of H2A.Z profiles for untreated (green) and α-amanitin-treated (violet) HCT116 cells for clusters with pattern I (**c-g**, cluster 5 is shown in **Fig. 4f**), pattern II (**h,i**, cluster 7 is shown in **Fig. 4g**), pattern III **(j)** and combined patterns I and II **(k,l)**. See Materials and Methods for an explanation of metagene plots. The colored bands below the metagene curves in each subfigure indicate the percentage of genes having an m- or i-RegC at that position in the comparison of α-amanitin-treatment vs. no treatment in the H2A.Z ChIP-seq data. Here, m-RegC are differential regions with relative read density higher in untreated cells and i-RegC differential regions with relative read density higher upon α-amanitin-treatment.



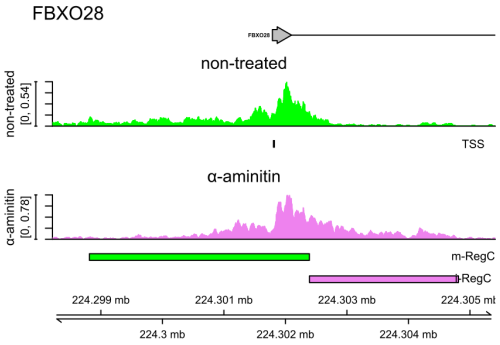
(a) pattern I, cluster 1



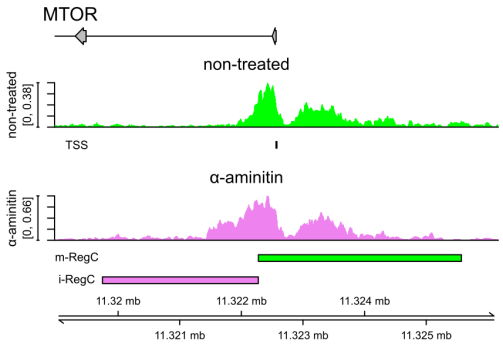
(b) pattern I, cluster 2



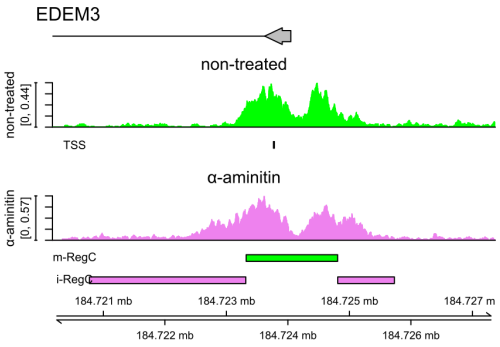
(c) pattern I, cluster 4



(d) pattern I, cluster 5

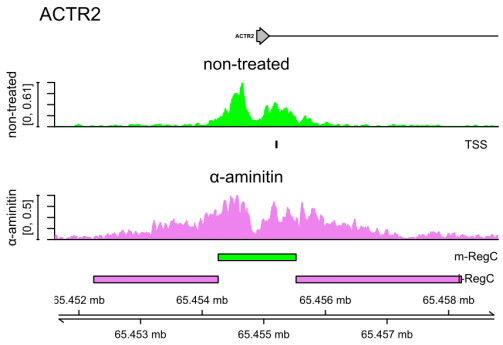


(e) pattern I, cluster 9

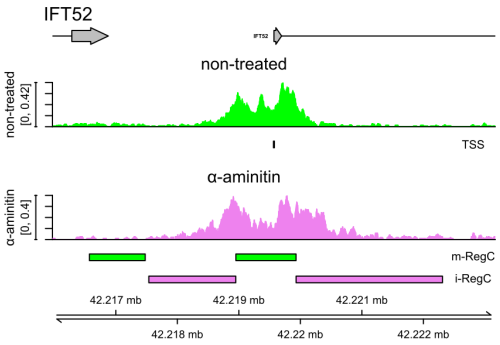


(f) pattern I, cluster 10

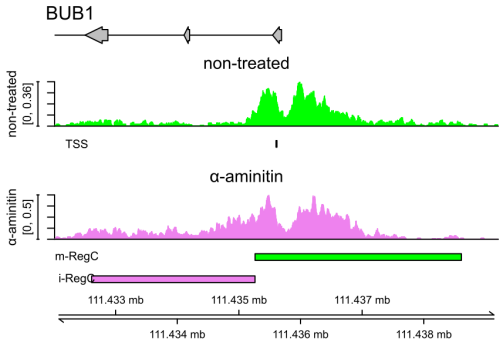
(Continued on next page)



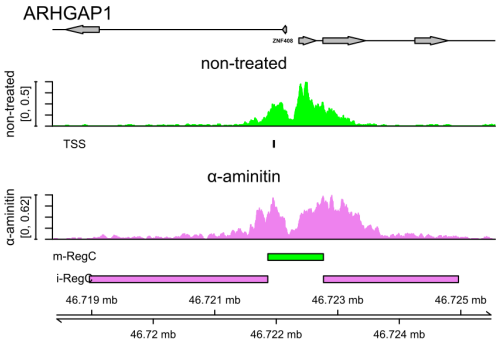
(g) pattern I, cluster 12



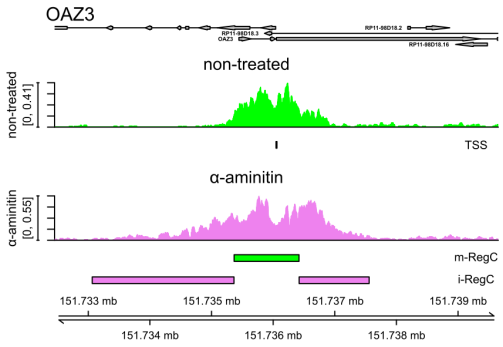
(h) pattern I, cluster 14



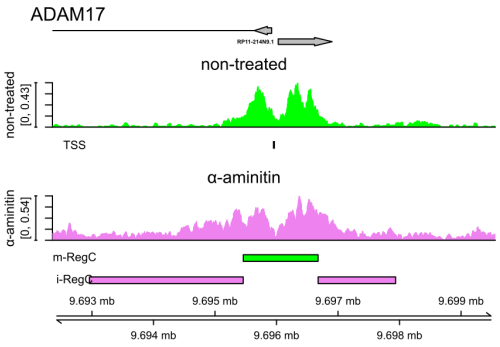
(i) pattern II, cluster 6



(j) pattern II, cluster 7

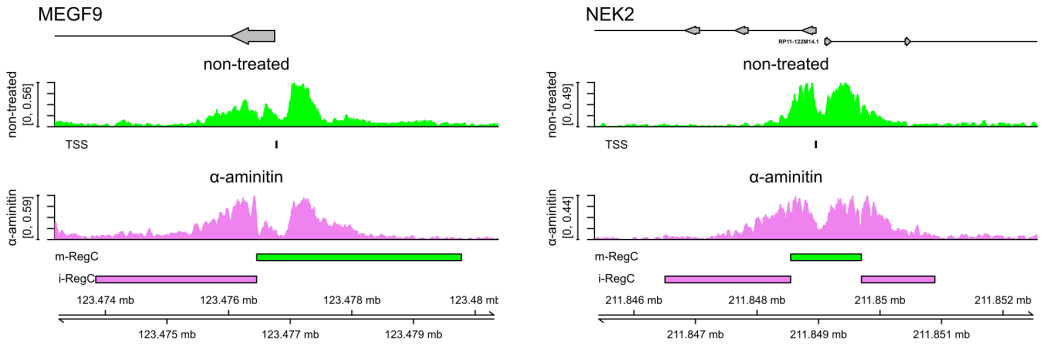


(k) pattern II, cluster 11



(l) pattern III, cluster 13

(Continued on next page)



(m) pattern I + II, cluster 3

(n) pattern I + II, cluster 8

Sup. Fig 31 Read coverage in a ± 3.6 kb window around the TSS in the H2A.Z ChIP-seq data for untreated (green) and α-amanitin-treated (violet) HCT116 cells for example genes for all patterns and clusters. Pattern and cluster are indicated below each subfigure. Read coverage was normalized to total number of mapped reads for each sample and averaged between replicates. The TSS used in the analysis is indicated by a short vertical line below the read coverage track for untreated cells. Gene annotation is indicated at the top. Boxes represent exons, lines represent introns, and direction of transcription is indicated by arrowheads. The name of the gene whose promoter window was analyzed is indicated in larger font on the top left and – if not clear from the context – beside the gene annotation. Names for other genes overlapping the input window are also indicated when necessary. Below each read coverage track m-RegCs (i.e. differential regions with relative read coverage higher in untreated cells) and i-RegC (i.e. differential regions with relative read coverage higher upon α-amanitin-treatment) are indicated.