The African Swine Fever Virus Transcriptome

Gwenny Cackett[1], Dorota Matelska[1], Michal Sýkora[1,3], Raquel Portugal[2], Michal Malecki[4], Jürg Bähler[1,5], Linda Dixon[2] and Finn Werner[*1]

[1] Institute for Structural and Molecular Biology, Darwin Building, University College London, Gower Street, London WC1E 6BT, United Kingdom

[2] Pirbright Institute, Ash Road, Pirbright, Surrey, GU24 0NF, UK

[3] Institute of Molecular Genetics, Czech Academy of Sciences, Prague, CZ

[4] Current affiliation: Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland

[5] Institute of Healthy Ageing, Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, United Kingdom

*correspondence should be addressed to linda.dixon@pirbright.ac.uk and f.werner@ucl.ac.uk

List of abbreviations:

Abbreviation	Definition
ASFV	African swine fever virus
NCLDV	Nucleocytoplasmic Large DNA Virus
ORF	Open reading frame
MGF	Multigene family
VACV	Vaccinia virus
RNAP	RNA polymerase
Pol II	RNA polymerase II
TBP	TATA-binding protein
TFIIB	transcription initiation factor II B

ETF	early transcription factor
TSS	transcription start site
p/np TSS / TTS	Primary/non-primary TSS/TTS
TTS	transcription termination site
CAGE-seq	cap analysis gene expression sequencing
UTR	untranslated region
NGS	Next generation sequencing
pNG	Putative novel gene
TU	Transcription unit
Inr	Initiator
EPM	Early promoter motif
UCE	Upstream Control Element
LPM	Late promoter motif
sncRNA	Small non-coding RNA
BRE	B-recognition element
ChIP	chromatin immunoprecipitation
CTSS	CAGE-seq TSS
SRA	Sequence Read Archive

15

Abstract

African Swine Fever Virus (ASFV) causes haemorrhagic fever in domestic pigs, presenting the biggest global threat to animal farming in recorded history. Despite its importance, little is known about the mechanisms and regulation of ASFV transcription. Using RNA sequencing methods, we have determined total RNA abundance, transcription start sites and transcription termination sites at single nucleotide-resolution. This allowed us to characterise DNA consensus motifs of early and late ASFV core promoters, as well as a poly-thymidylate sequence determinant for transcription termination. Our results demonstrate that ASFV utilises alternative transcription start sites between early and late stages of infection, and that ASFV-RNAP undergoes promoter-proximal transcript slippage at 5' ends of transcription units, adding quasi templated AU- and AUAU-5' extensions to mRNAs. Here we present the first much-needed genome-wide transcriptome study that provides unique insight into ASFV transcription and serves as a resource to aid future functional analyses of ASFV genes which are essential to combat this devastating disease.

Importance

African swine fever virus (ASFV) causes incurable and often lethal haemorrhagic fever in domestic pigs. In 2019, ASF presents an acute and global animal health emergency that has the potential to devastate entire national economies as effective vaccines or antiviral drugs are not currently available (Food and Agriculture Organization of the UN). With major outbreaks ongoing in Eastern Europe and Asia urgent action is needed to advance our knowledge about the fundamental biology of ASFV, including the mechanisms and temporal control of gene expression. A thorough understanding of RNAP and transcription factor function, and the sequence context of their promoter motifs, as well as accurate knowledge of which genes are expressed when and the amino acid sequence of the encoded proteins, is direly needed for the development of antiviral drugs and vaccines.

Introduction

ASFV is the sole characterised member of *Asfarviridae* (1), a family resembling others in the Nucleocytoplasmic Large DNA Viruses (NCLDV) and *Megavirales* order (2, 3). *Asfarviridae* also include the uncharacterised Abalone asfarvirus (NCBI:txid2654827), while the Faustoviruses show similarity to ASFV but have larger genomes and infect amoeba (*Vermamoeba vermiformis*) (4). ASFV originated in East Sub-Saharan Africa where it remains endemic, it crossed continents to Georgia in 2007 (5) and its subsequent spread in Europe and to Asia 2018 (6) has resulted in the current emergency situation. ASFV has a linear double-stranded DNA (dsDNA) genome of ~170–194 kbp encoding ~150–170 open reading frames (ORFs). Genomic variation between strains predominantly originates from loss or gain of genes at the genome termini among members of multigene families (MGFs) (7). Despite its global economic importance, little is known about ASFV transcription, but it is believed to be related to the vaccinia virus (VACV) system (8–10), a distantly-related NCLDV and *Poxviridae* family member (11).

We have focused our analysis on the BA71V strain (170,101 bp genome, with 153 annotated ORFs (12, 13), because this is the most well-studied ASFV strain regarding viral molecular biology including gene expression and mRNA modification (10, 14). Based on a paradigm of the vaccinia virus, several stages of ASFV gene expression have been hypothesised in the literature including immediate early-, early-, intermediate- and late genes (10, 15–17). However, the experimental evidence for four discrete gene expression stages in ASFV leaves room for improvement, though the presence of two alternative subsets of transcription initiation factors strongly supports the notion of at least two discrete stages, early and late, likely at pre- and post-replicative stages of the virus life cycle. Previous individual gene expression studies have made use of chemical inhibitors to inhibit replication or protein synthesis (10, 15, 16). While valid tools when used with care (18), the application of these chemicals is not unproblematic due to the possibility of indirect pleiotropic effects. E.g. the nucleotide analogue cytosine arabinoside (AraC) can be incorporated into DNA and while at low concentrations mostly inhibiting replication, it can interfere with the action of many DNA-binding

65 enzymes including RNA polymerases, transcription factors as well as topoisomerase (19). In light of
66 this, in this study we chose to characterise transcription unadulterated by chemical inhibitors.
67 ASFV inhabits the eukaryotic cytoplasm and appears to be self-sufficient in terms of transcription and
68 modification of viral mRNA. It encodes an RNA polymerase (RNAP), a poly-A polymerase and an
69 mRNA capping enzyme, importantly, extracts obtained from mature virus particles are fully
70 transcription competent (10, 20, 21). The basal ASFV transcription machinery resembles the
71 eukaryotic RNAPII system encompassing an (8-subunit) ASFV-RNAP and distant relatives of the TATA-
72 binding protein (TBP), the transcription initiation factor II B (TFIIB) and the elongation factor TFIIS (8,
73 9, 13). ASFV also encodes a histone-like DNA binding protein pA104R and ASFV topoisomerase II
74 (pP1192R) which collaborate to generate DNA-binding and supercoiling activity (22). Of particular
75 interest is the possibility that the ASFV-RNAP gains promoter-specificity in terms of temporal (early
76 or late) gene expression dependent on the association with either TBP/TFIIB-like or virus-specific
77 factors including those encoded by ASFV BA71V genes D1133L and G1340L, which are homologous to
78 the D6 and A7 (respectively) early transcription factor (ETF) heterodimer (23, 24) from VACV.
79 Promoter consensus motifs for early and late ASFV genes have not been characterised on a genome-
80 wide scale, or in great detail, with the exception of an AT-rich sequence motif upstream of the p72
81 gene transcription start site (TSS) and some other late genes, as well as a consistently AT-rich region
82 overlapping the TSS (25). Importantly, information about the temporal ASFV gene expression, TSS
83 and transcription termination sites (TTS) is not available (10, 11).
84 We have applied a combination of NGS techniques including RNA-seq, RNA 5'-end (cap analysis gene
85 expression sequencing or 'CAGE-seq') and RNA 3'-end (3' RNA-seq) determination. We report (i) the
86 ASFV transcriptome map showing differences in gene expression between early and late infection, (ii)
87 a genome-wide TSS map that has allowed us to define early and late ASFV promoter consensus
88 motifs as well as 5'-mRNA leaders, and (iii) a genome-wide TTS map that provides novel insights into
89 the mechanism of transcription termination in ASFV. Figure 1 is a genome-wide map visualising our
90 results from TSS-mapping and differential gene expression in ASFV.

Results

Overview of the ASFV transcriptome

A transcriptome is defined by the overall expression levels of transcripts, and their 5' and 3' termini. We carried out RNA-seq, CAGE-seq and 3' RNA-seq in order to characterise these parameters during early and late ASFV infection, when combined they inform about the ASFV transcriptome and DNA sequence signatures associated with transcription initiation and termination. The processed data are compiled in an assembly hub and can be publicly accessed in the UCSC Genome Browser by the following address: https://bit.ly/2TazQxK.

Vero cells were infected with BA71V, and viral RNA was extracted at 5h and 16h post-infection. These time points were chosen based on a previous report of a small subset of genes that were experimentally characterised using nuclease S1 mapping and primer-extension analysis (10, 26). Bowtie 2 (27) mapping of the RNA-seq, CAGE-seq and 3' RNA-seq reads (summarised in Supplementary Table 1) showed a strong correlation between replicates (Pearson correlation coefficient $r \geq 0.9$), with one exception of RNA-seq from 16h (r of 0.74 and 0.84 for two strands, data not shown). Figure 2a provides a whole-genome view of mapped reads from all three Next Generation Sequencing (NGS) approaches, while a selection of individual examples of TSSs and TTSs at single-nucleotide resolution is shown in Figure 2 b-e. The sequencing depth of the RNA-seq approach was more than sufficient to determine significant changes in ASFV transcription (i.e. reads) at early and late infection due the small genome size (170 kb). The majority of CAGE-seq reads (i.e. TSSs) were located upstream and proximal to ORF start codons. A subset of late infection TSSs mapped to more distant locations between ORFs or within ORFs, these are caused by pervasive transcription, mRNA de-capping and -degradation followed by re-capping, or BA71V genome mis-annotations (28–31). The increased background of TSSs were more noticeable during late infection (Figure 2a, 'CAGE-seq 16h') and likely due to pervasive transcription, a phenomenon that has been observed in humans (32) and in VACV (28). The cause of this low-level and genome-spanning

transcription is unclear but has been attributed to an open chromatin structure in cellular organisms (33). In viral genomes it may reflect differences between nascent, newly replicated genomic DNA during late infection and genomic DNA still associated with histone-like proteins (such as A104R) just released from the virus particle during early infection.

Mapping of ASFV Primary Transcription Start Sites

Following mapping of CAGE-seq reads to the ASFV-BA71V genome, we located regions with an enrichment of reads corresponding to the 5' ends of transcripts and thereby the TSS. We detected a 779 clusters of CAGE-seq signals, and CAGE-seq clusters upstream annotated ORFs were manually investigated to confirm that they represent 'primary' TSSs (pTSSs) based on peak height, proximity to the ORF initiation codon, and coverage from our complementing RNA-seq data. We identified pTSSs fulfilling these criteria upstream of 151 BA71V ORFs, thus only two genes, E66L and C62L, were not found associated with a pTSS. Overall, our data showed good agreement with previously individually mapped TSSs of 44 ORFs (Supplementary Table 2). Not all of the ~780 clusters were located within 500 bp upstream of ASFV ORFs, but within, or in the antisense orientation relative to ORF coding sequences (Figure 3a). We reannotated eleven ORFs based on gene-internal TSS and RNA-seq reads (Table 1, and I177L example in Figure 3b); we provide a novel gene feature file based on our revised annotations, (Supplementary GFF).

Several genes have a *bona fide* pTSS upstream of the annotated start codon and an alternative TSS residing within the including J64R (Figure 2d) and B169L (Figure 3b). The alternative downstream TSS of J64R is weaker and specific to 16h p.i., compared to the upstream pTSS. Our genome-wide CAGE results are confirmed by previous analysis of individual genes such as I243L (26), which was shown to have distinct TSSs for different stages of infection (Figure 4a). I243L encodes a homologue of the Pol II transcript cleavage factor TFIIS, that is highly conserved between archaea, eukaryotes and among NCLDV members albeit with limited domain conservation (34). TFIIS has dual functions, it stabilises transcription initiation complexes, and reactivates stalled elongation complexes by transcript cleavage (35, 36). The late TSS is located downstream of the I243L start codon, and the utilisation of

the next Methionine codon would result in a TFIIS variant lacking 52 N-terminal amino acid residues (Figure 4b). While the early and long transcripts encode the fully functional three-domain TFIIS factor, the late and short transcripts encode a truncation variant lacking the N-terminal domain that is responsible for initiation functions of TFIIS. In essence, the TFIIS variants expressed during early and late infection would have a different functionality. We identified seven further genes with alternative pTSSs during early and late infection (Table 2). In most cases, the re-annotated (single pTSS downstream of start codon) or alternative pTSSs (multiple pTSSs, some downstream of start codon) did not substantially alter the ORF protein products, except for re-annotated I177L and alternative pTSSs of B169L, two putative transmembrane proteins (Figure 3b)(13, 20).

Novel Genes Supported by Sequencing Data

28 TSSs in our CAGE-seq data set were not associated with annotated ORFs (Supplementary Table 3) and seven of these pTSSs were associated with transcripts that encode short ORFs, which we call putative novel genes (pNGs). These encode polypeptides of 25–56 AA length that were missed in the initial BA71V ORF prediction as only ORFs ≥ 60 AA were annotated (13). Five pNG ORFs showed limited similarity to short ORF-encoding genes from other ASFV strains, while pNG5 showed no clear similarity (Table 3). Interestingly, pNG6 was homologous to KP93L which is already encoded by BA71V, but barely expressed according to our data. In contrast, pNG6 was highly expressed at 5h (Supplementary Table 4). Figure 3c illustrates the features of pNG1 and pNG3, with distinct TSS and TTS, and robust RNA-seq read coverage across the entire gene. All pNGs had the same orientation as neighbouring downstream genes (Figure 1), and five of the seven pNGs transcripts terminated promptly, i.e. were associated with a drop of reads following a 5–8 nucleotide thymidylate sequence (Figure 3c and (10, 16)). All these observations support the notion that these transcription units are new *bona fide* genes.

Highly expressed ASFV genes during Early and Late Infection

In order to gain insights into expression of individual genes, we quantified mRNA levels obtained by CAGE-seq and compared the most abundant mRNAs at early and late time points (Figure 5a). Supplementary Table 4 summarises expression of all detected ASFV-BA71V genes including the newly annotated pNGs. For this purpose, we temporarily re-defined ASFV gene transcription units (TUs) as regions spanning from pTSS to stop codon (as proxy for TTS, see below), and quantified TU expression based on RNA-seq data (Figure 5b, Supplementary Table 5), which closely reflected the CAGE-seq analysis. The highly expressed genes matched those identified in the viral proteome of infected tissue cultures determined by mass spectrometry (highlighted in Figure 5a-b) (37). Six genes in the top-20 highly expressed genes were common during early and late infection (CP312R, A151R, K205R, Y118L, pNG1, I73R). While their expression decreases from early to late infection (see below), these genes are clearly expressed throughout, suggestive of a multistage expression pattern. Considering their high levels of expression, they are likely important throughout infection which makes them interesting candidates as potential drug- or vaccine target. However, four (out of six) have an unknown function (Figure 5a) and await functional investigation.

Differential Expression of early and late ASFV Genes

We characterised differential expression of ASFV genes between early and late infection by comparing separate DESeq2 analyses of CAGE-seq and RNA-seq datasets (Figure 5c and d, respectively). Based on RNA-seq data, 103 ASFV TUs showed significant differential expression (adjusted *p*-value < 0.05), with 47 genes down- and 56 genes up-regulated during the progression from early to late infection. Henceforth, we focused on the CAGE-seq dataset because the reads are associated with the nascent transcription start sites and thus cannot arise from transcription readthrough from upstream genes (unlike mRNA quantification using RNA-seq) which would complicate the analyses. RNA-seq also had the disadvantage of a lower sequencing depth and thus lower apparent sensitivity compared to CAGE-seq. Indeed, the CAGE-seq identified 149 genes as significantly differentially expressed with 65 downregulated genes and 84 upregulated genes (Figure

191 5c). Naturally this is not a binary classification i.e. genes that are upregulated during late infection do
192 not have zero reads during early infection and *vice versa*. Interestingly, the relative expression levels
193 of early genes at 5 h p.i. appeared significantly higher than late genes at 16 h p.i. (Figure 6a). This is
194 due to normalisation of the reads and the increase of steady state levels of all transcripts during late
195 infection, which can be seen from the sequence alignment rates (Supplementary Table 1). While the
196 number of reads mapping to early genes during early infection is lower than the reads mapping to
197 late genes during late infection, the total number of reads mapping to *all* ASFV genes is higher during
198 late infection. The per-gene FPM values and differential expression analyses are normalised for ASFV-
199 mapped sequencing depth, which therefore reduces this background and emphasises highly
200 expressed genes during early infection. Overall, we did observe a greater and cleaner contrast in
201 expression of the genes during early compared to late infection. The expression of the least
202 expressed genes at 5h p.i. was more consistent and closer to zero than those at 16h p.i. (Figure 6b).
203 The most highly expressed genes at both time points were more similar, though relative expression
204 of the most expressed genes at 5h p.i. was higher than at 16h p.i (Figure 6c). In summary, it appears
205 ASFV maintains a tighter control of gene expression during early infection compared to late, in as
206 much as early genes are highly expressed and late genes show low or no expression, while during
207 late infection the total mRNA levels increase, which results in a greater change of absolute late
208 mRNA levels but lower relative levels of late mRNAs.
209 In order to stringently analyse differential expression in ASFV we identified the genes which showed
210 the same pattern of differential expression according to separate DESeq2 analyses of the CAGE-seq
211 and RNA-seq datasets. This minimises any potential biases from each of these complementing
212 techniques. 101 genes showed significant differential expression according to both independent
213 techniques, and the changes in expression were significantly correlated between these genes
214 (Spearman's rank correlation coefficient $\rho = 0.73$, Figure 6d). Only a small number of genes, ten out
215 of 101, showed a discrepancy between the two methods (DP63R, I329L, NP419L, B66L, A224L, E248R,
216 O174L, D345L, C315R and NP1450L), leaving 91 genes confidently classified as early (36) and late (55)

genes. Supplementary Table 6 provides details of these 91 genes, their functions, and whether previously detected in viral particles (20). The 91 genes with correlated differential expression were assigned with functional categories based on their annotation in the VOCS database (38) complemented with ASFVdb (39) (Figure 6e). Around one fifth of early and late genes were classified as 'uncharacterised' without any functional predictions. The transition between 5 h and 16 h post infection is characterised by a significant up-regulation of genes important for viral morphology and structure, but also the overall diversity of differentially expressed genes changed. A significant difference was seen in the multigene family members; they constitute nearly a half of the early genes, but only one (MGF 505-2R) among late genes. ORFs annotated as having a 'transmembrane region' (TR) or a 'putative signal peptide' (PSP) were also overrepresented in late infection (Fisher Test: $p < 0.05$); they remain poorly characterised beyond a domain prediction and 9 proteins (out of 12) of these ORFs could be detected in BA71V virions by mass spectrometry (20).

Architecture of ASFV Gene Promoters and Consensus Elements

The genome-wide TSS map combined with information about their differential temporal utilisation allowed us to analyse the sequence context of TSSs and thereby characterise the consensus motifs and promoter architecture of our clearly defined 36 early and 55 late genes. Eukaryotic RNA pol II core promoters are characterised by a plethora of motifs, including TATA boxes and BRE elements, and the Initiator (Inr). The former two interact with initiation factors TBP and TFIIB, while the latter interacts with RNA pol II (40). Alignment of regions immediately surrounding pTSSs in the BA71V genome revealed several interesting ASFV promoter signatures: the Inr element overlapping the TSS is a feature that distinguishes between early and late gene promoters (Figure 7a and b, respectively). The early gene Inr is a TA(+1)NA tetranucleotide motif (where N has no nucleotide preference, Figure 7c), while the late gene Inr shows a strong preference for the sequence TA(+1)TA (Figure 7d), that is not to be confused with the TBP-binding TATA box. Our late Inr consensus motif is in good agreement with those of 20 previously characterised late gene TSSs (10, 25). To search for additional promoter elements that likely interact with transcription initiation factors, we extended our search to include

243 sequences up to 40 bp upstream of the TSS. Analysis with MEME and FIMO software (41, 42)
244 identified and located a significant 19-nt motif (Figure 7e) located ~10 bp upstream of pTSSs for 36
245 (out of 36) early gene promoter sequences (Figure 7f), which we have called the Early Promoter
246 Motif (EPM). Our EPM is related to the VACV early gene promoter motif ('Upstream Control Element'
247 or UCE) (43, 44) as well as the yeast Virus-Like Element (VLE) promoters (45). However, the EPM is
248 not limited to the 36 early genes, since a FIMO software (42) motif search identified the EPM within
249 60 bp upstream of a much larger subset of 81 TSS/TUs including pNGs and alternative pTSSs, four of
250 which were the early alternative pTSS for I243L, B169L, J154L and CP80R Importantly. The limited
251 distance distribution between the EPM and TSS is indicative of constraints defined by distinct
252 protein-DNA interactions, e.g. by transcription initiation factors binding upstream of the TSS and
253 ASFV-RNAP engaging with promoter DNA and TSS (Figure 7f). Figure 7g illustrates expression profiles
254 of all genes with an EPM upstream according to FIMO, the majority showing a negative log2 fold
255 change between 5h and 16h. Since MGF members were overrepresented as early genes (Figure 6e),
256 we searched directly for the EPM among the FIMO hits. 23 of the 29 MGF members with mapped
257 pTSSs were associated with the EPM element including a consistent early expression and spacing
258 relative to their TSS (Figure 7h-i), which suggests that MGF genes are under the control of their own
259 promoters.
260 Using the same approach, we searched for promoter sequence motifs associated with late genes.
261 MEME identified a conserved motif upstream of only 17 (out 55) late genes, which we called Late
262 Promoter Motif (LPM, Figure 8a). The spacing (4–12 bp) between the LPM and TSS shows a much
263 greater diversity compared to the EPM (Figure 8b), though genes with the LPM were consistently
264 upregulated (Figure 8c). A Tomtom (46) search identified the LPM motif as a match for 28 distinct
265 motifs including the canonical TATA-box (*p*-value: 2.85e-03, E-value: 5.16e+00, Figure 8d). However,
266 this was not a strong hit and these motifs only bear a limited resemblance to each other except for
267 their AT-rich bias.

ASFV mRNAs have 5' leader regions

Early and late genes in ASFV vary with regard to the length of 5' UTRs i.e. the distance between the 5' mRNA end and the translation start codon. The 5' UTRs of late genes are significantly shorter and have a higher AT-content compared to early genes (p-value < 0.05, Figure 8e-f). Surprisingly, a subset of late gene CAGE-seq reads extended upstream of the assigned TSSs and were not complementary to the DNA template strand sequence. In order to rule out any mapping artefacts, we trimmed the CAGE-seq reads by removing the upstream 25 nt and aligned them to the genome at the 5' boundary of the reads. This did not significantly impair the mapping statistics but highlighted that nearly half of the annotated TSSs (74/158) among both early and late genes are associated with mRNAs that have short 5' extensions (or leaders), including seven genes with multiple TSSs (Supplementary Table 7). Most 5' leaders consist of two- or four nucleotides (Figure 9a) and the presence of the 5' leaders was not correlated with early or late expression (Figure 9b). The most common sequence motif in sequencing reads is AT (33% and 71% of early and late genes, respectively) and ATAT (7% in late genes, Figure 9c). In order to investigate any potential sequence-dependency of the mRNAs associated with AU- and AUAU-5' leaders, we scrutinised the template DNA sequence downstream of the TSS and found that all TUs, contained the motif ATA at positions +1 to +3 (Figure 9d). This suggests that the formation of AU-leaders is generated by RNA polymerase slippage on the first two nucleotides of the initial A(+1)TANNN template sequence, generating AUA(+1)UANNN or AUAUA(+1)UANNN mRNAs. A different but related slippage has been observed in the VACV transcription system, where all post-replicative mRNAs contain short polyA leaders which are associated with consensus Inr TAAAT motif (28).

Transcription termination of ASFV-RNAP

Previous mapping of mRNA 3' ends has revealed a conserved sequence motif consisting of ≥7 thymidylate residues in the template, which is consistent with 3' end formation *via* transcription termination like the RNA polymerase III paradigm (16, 47). To investigate the genome-wide sequence context of ASFV transcription termination, we used 3' RNA-seq sequencing to obtain the sequences

immediately preceding ASFV mRNA poly(A) tails, generating a complete map of mRNA 3' end peaks (Figure 2a). Using a similar approach as pTSS mapping, CAGEfightR detected a total of 657 termination site clusters, 212 TTSs within 1000 bp downstream of 1–3 ORFs. Because multiple ORFs had more than one cluster within that region (Supplementary Table 8), we defined 114 primary TTSs (pTTS) as the TTS with the highest CAGEfightR-score in closest proximity to a stop codon; we classified the 98 remaining peaks as non-primary TTSs (npTTS). We identified a highly conserved poly-T signal within 10 bp upstream of 126 TTSs (83 pTTSs, 43 npTTSs) that was characterised by ≥4 consecutive T residues (Figure 10a), with the ultimate residue located on or 2 bp upstream of the ultimate T residue in the motif (Figure 10b). The remaining 86 TTSs were not associated with any recognisable sequence motif besides a single T residue 1 bp upstream of the TTS. Our results are in good agreement with a previous S1 nuclease mapping of 6 coding mRNAs, but less so with 17 proposed TTSs which were predicted based on transcript length estimates relative to upstream transcription start sites (Supplementary Table 2). This may be because only ≥7 consecutive Ts in the template were included to serve as terminators. Our results demonstrate that the total number of consecutive Ts of the poly T motif can vary, with poly T tracts of CAGE-early genes being longer than those of late genes (Figure 10c). Finally, we observed differences between CAGE-early and CAGE-late gene termination, in as much as poly T terminators were overrepresented in CAGE-early and underrepresented in CAGE-late genes (Figure 10d). The 3' UTRs (i.e. nt length from translation stop codon to pTTS) of CAGE-late genes were significantly longer compared to CAGE-early genes (Figure 10e), in good agreement with previous studies on a small number of mRNAs which showed ASFV transcripts tended to be longer and more variable in length during late infection (Supplementary Table 2). ORFs are spaced closely in the ASFV genome, and scrutiny of RNA-seq reads reveal a limited extent of transcription readthrough from upstream ORFs into downstream ORFs likely due to leaky termination (Cackett and Werner, unpublished observations). However, any additional downstream ORFs generated aberrantly by transcription readthrough would not be able to be translated since

there is no evidence of ASFV utilising internal ribosome entry sites (IRES) that would be required to enable cap-independent translation (7).

Discussion

Here we report the first comprehensive ASFV transcriptome study at single-nucleotide resolution. The mapping of 158 TSS and 114 TTS for 159 ASFV genes allowed us to reannotate the BA71V genome. Our results provide detailed information about differential gene expression during early and late infection, the sequence motifs for early and late gene promoters (EPM and LPM, and Inr elements) and terminators (poly-T motif), and evidence quasi-templated 'AU' RNA-5' tailing by the ASFV-RNAP.

We have discovered seven novel putative genes, some of which are highly conserved with the aggressively virulent strains (Georgia 2007/1 and Belgium 2018/1) that have caused the current outbreak in Europe (Table 3). This suggests that BA71V has more genes in common with its virulent cousins than initially thought.

Our results demonstrate that the majority of ASFV genes show some degree of differential expression from early to late infection (Figure 1). Interestingly, our CAGE-seq results demonstrate that early genes are expressed at relatively higher levels during early infection, than late genes during late infection (Figure 6a-c). Future experiments including spike-in controls are needed to confidently quantify the absolute mRNA levels during early- and late infection (48). The RNA sequencing methods used here quantify the steady-state RNA levels and not RNA synthesis rates, and without information about ASFV mRNA stability it is not possible to distinguish between early mRNAs retained until late infection and early genes being transcribed at later stages. Nascent ASFV mRNA synthesis rates and half-lives could be determined using techniques including TT-seq (49) or by using transcription inhibitors including actinomycin D (50). Frustratingly, many of the highly expressed genes are uncharacterised (Figure 5a). These gene products are important candidates for further functional characterisation and may emerge as promising targets for vaccine development.

We have shown that MGFs show a distinct downregulation from early to late infection, while genes annotated as transmembrane region or putative signal peptides (though poorly characterised beyond this), along with structural or viral morphology genes, are overrepresented in late infection (Figure 6e). Our CAGE-analysis also identified TSS signals unlikely to serve as primary TSS for annotated genes (Figure 3a and Supplementary Table 9); these could provide a rich hunting ground for small non-coding (snc)RNAs. One TSS cluster associated with an sncRNA gene (at 71,302 on the BA71V genome) was previously reported by Dunn et al. (51) as ASFVsRNA2, that is encoded in the antisense orientation relative to the ASFV RNA polymerase subunit RPB6-encoding gene. Further investigation of antisense sncRNAs in the BA71V transcriptome may uncover further examples of riboregulation, i.e. a more complex method of modulating its own or host gene expression beyond the protein level.

While eukaryotic Pol II and archaeal RNAP critically rely on initiation factors TBP and TFIIB for transcription initiation on all mRNA genes, bacterial RNAP obtains specificity for subsets of gene promoters by associating with distinct sigma factors (52). ASFV-RNAP is related to archaeal and eukaryotic RNA polymerases, detailed phylogenetic analyses reveal that the RPB1 subunit is most closely related to the RNA polymerase I homologue (3, 45, 53). However, transcription initiation of early and late genes appears to be directed by two distinct sets of general initiation factors and their cognate DNA recognition motifs, as our TSS mapping demonstrates. The first feature of all ASFV promoters is the Inr element, a tetranucleotide motif overlapping the TSS with an A-residue serving as initiating nucleotides similar to most RNAP systems. The similarity of early and late gene Inr sequences, is likely because the Inr makes sequence-specific contacts with amino acid sidechains of the two largest RNAP subunits (RPB1 and 2). The EPM and LPM are located upstream of the TSS, both are AT-rich, though distinct in sequence (Figure 7e and 8a). The distance distribution of EPM is narrow (located 9–10 bp upstream of the TSS) while the distance between the LPM and TSS shows greater variation and is located closer (4–6 bp) to the TSS. The high sequence and distance conservation of the EPM, especially exemplified for early expressed MGFs (Figure 7h-i), emphasises the EPM's role in tight control of transcription during early infection. Considering the close

relationship between ASFV and VACV, we posit that the EPM is recognised by heterodimeric ASFV-BA71V D1133L/G1340L initiation factor (VACV D6/A7) (11) consistent with the late expression of these genes (Figure 6d, also ref (54)). Presence of D1133L/G1340L gene-products along with RNAP in viral particles (20) provides a system that is primed to initiate ASFV transcription of early genes.

ASFV-TBP (B263R) is an early gene and ASFV-TFIIB (C315R) is expressed throughout infection. We propose the LPM is utilised by ASFV-TBP and -TFIIB homologues, neither of which were detected in virions (20). A functional comparison of the LPM to the classical Pol II core promoter elements BRE/TATA-box is compelling. However, the tight spacing between the LPM and TSS is incompatible with the overall topology of a classical eukaryotic and archaeal TATA-TBP-TFIIB-RNA pol II preinitiation complex (PIC), where the BRE/TATA promoter elements are located ~ 24 bp upstream of the TSS (55). Considering low sequence conservation between cellular and ASFV-TBP (8) and unusual spacing of LPM and Inr, the structure of ASFV LPM-TBP-TFIIB-RNAP PIC is likely very different from canonical RNA pol II PICs. Additionally, factors including ASFV B175L and B385R may contribute to the PIC, as was proposed for VACV-A1 and A2 (56, 57). At this stage, we cannot rule out a limited overlap between early and late genes without additional information including insights into pre- and post-replicative gene expression pattern, mRNA stability of early and late genes, and knowledge about all regulatory factors that enable the temporal regulation of ASFV transcription. To unequivocally attribute factors to their cognate binding motifs genome-wide, a chromatin immunoprecipitation (ChIP) approach is required; the results may be full of surprises and have the potential to shed light on multistage gene expression pattern including the possibility of a more complex promoter architecture where some genes are under the control of more than one promoter.

An in-depth characterisation of the global gene regulation in ASFV with a higher temporal resolution is essential to assess how closely ASFV follows the cascade-like patterns of VACV (11). While two genes have been proposed to be intermediate genes in ASFV, both of them are also expressed during

intermediate and late (I226R), and during early, intermediate and late stages (I243L). Thus, there is no hard evidence of genes that are specifically expressed during the intermediate stage (26). A combination of a reversible replication inhibitor and a conditionally regulated late transcription factor has been successfully used to characterise intermediate gene expression in VACV (58). Such an approach might also be useful to identify intermediate ASFV genes - and help us refine the LPM that in our current analysis could reflect a combination of late- and intermediate gene promoters'.

We found several examples of alternative, gene-internal, TSS utilisation with the potential to increase the complexity of the viral proteome; protein variants which may provide the means to generate distinct functionalities, which has also been described in VACV by Yang et al. (28). Our TSS mapping uncovered a form of transcript slippage by the ASFV-RNAP occurring on promoters that start with an A(+1)TA motif, where mRNAs are extended by one or two copies of the dinucleotide AU. This is reminiscent of VACV, where late gene transcripts containing a poly-A 5' UTR (28) are associated with improved translation efficiency and reduced reliance on cap-dependent translation initiation (59, 60); likewise, distinct functional attributes of polyA leaders in translation have been documented in eukaryotes (61). Whether the 5' AU- and AUAU-tailing is a peculiarity of the ASFV-RNAP initiation, or whether these mRNA 5' leaders have any functional implications, remains to be investigated. The structural determinants underlying RNAP slippage are interactions between the template DNA sequence and the RNAP and/or transcription initiation factors; the differential use of distinct initiation factors for the transcription of early and late ASFV genes may account for difference in leader sequences.

The mechanisms underlying transcription termination of multisubunit RNAP are diverse (62, 63). Our analyses of genome-wide ASFV RNA-3' ends allowed the mapping of the ASFV 'terminome'. Over half of mRNA 3' ends are characterised by a stretch of seven U residues, with the TTS mostly coinciding with the last T residue in the template DNA motif - in good agreement with ASFV terminators that have been individually mapped (15, 16). In contrast, VACV appears to utilise a motif ~ 40 nt upstream of the mRNA 3' ends (64, 65). In essence, the ASFV-RNAP is akin to archaeal RNAPs and RNA pol III,

where a poly-U stretch is the sole *cis*-acting motif without any RNA secondary structures characteristic of bacterial intrinsic terminators (63). The pTTSs without any association with poly-U motifs are still likely to represent *bona fide* termination sites, since RNA-seq reads were decreasing towards these termination sites, despite no clear conserved sequence motif. However, ASFV does encode several (VACV-related) RNA helicases that have been speculated to facilitate transcription termination and/or mRNA release (10, 66). Future functional studies will address the molecular mechanisms of termination including the role of putative termination factors.

Understanding the molecular mechanisms of the ASFV transcription system is not only of academic interest. Unless effective vaccines in conjunction with antiviral treatments against ASFV are developed, a large proportion of the global pig population is projected to die in the context of this terrible disease (OIE, https://www.oie.int). The rational design of drugs that target the gene expression machinery is crucially reliant on our knowledge about the ASFV-RNAP, the basal factors that govern its function, and the DNA sequences they interact with, while vaccine development benefits from the intricate knowledge about gene expression patterns. Our results directly contribute to these burning issues for animal husbandry.

436 # Methods

437 ## RNA Sample Extraction from Vero Cells infected with BA71V

438 *Vero* cells (Sigma-Aldrich, cat #84113001) were grown in 6-well plates, plates and were infected in 2
439 replicate wells for 5h or 16h with a multiplicity of infection of 5 of the ASFV BA71V strain, collected in
440 Trizol Lysis Reagent (Thermo Fisher Scientific) separately, after growth medium was removed.
441 Infected cells were collected at 5h post-infection (samples for RNA-seq: S3-5h and S4-5h, CAGE-seq:
442 S1-5h and S2-5h and 3' RNA-seq: E-5h_1 and E-5h_1), and at 16h post-infection (RNA-seq: S5-16h
443 and S6-16h, CAGE-seq: S3-16h and S4-16h, and 3' RNA-seq: L-16h_1, L-16h_1). RNA was extracted
444 according to manufacturer's instructions for Trizol extraction and the subsequent RNA-pellets were
445 resuspended in 50µl RNase-free water and DNase-treated (Turbo DNAfree kit, Invitrogen). RNA
446 quality was assessed *via* Bioanalyzer (Agilent 2100), before ethanol precipitation. For CAGE-seq and
447 3' RNA-seq, samples were sent to *CAGE-seq* (Kabushiki Kaisha DNAFORM, Japan) and Cambridge
448 Genomic Services (Department of Pathology, University of Cambridge, Cambridge, UK), respectively.

449 ## RNA-seq, CAGE-seq and 3' RNA-seq Library Preparations and Sequencing

450 For RNA-seq, samples were resuspended in 100µl RNase-free water, and polyA-enriched using the
451 BIOO SCIENTIFIC NEXTflex™ Poly(A) Beads kit according to manufacturer's instructions and quality
452 was assessed *via* Bioanalyzer. NEXTflex™ Rapid Directional qRNA-Seq™ Kit was utilised to produce
453 paired-end indexed cDNA libraries from the polyA-enriched RNA samples, according to the
454 manufacturer's instructions. Per-sample cDNA library concentrations were calculated *via* Bioanalyzer,
455 and Qubit Fluorometric Quantitation (Thermo Fisher Scientific). Sample S3-5h, S4-5h, S5-16h and S6-
456 16h cDNA libraries were twice separately sequenced on Illumina MiSeq generating 75 bp reads
457 (Supplementary Table 1) and 12 FASTQ files.
458 Library preparation and CAGE-sequencing of RNA samples S1-5h, S2-5h, S3-16h and S4-16h was
459 carried out by *CAGE-seq* (Kabushiki Kaisha DNAFORM, Japan). Library preparation produce single-end
460 indexed cDNA libraries for sequencing: in brief, this included reverse transcription with random

primers, oxidation and biotinylation of 5' mRNA cap, followed by RNase ONE treatment removing RNA not protected in a cDNA-RNA hybrid. Two rounds of cap-trapping using Streptavidin beads, washing away uncapped RNA-cDNA hybrids. Next, RNase ONE and RNase H treatment degraded any remaining RNA, and cDNA strands were subsequently released from the Streptavidin beads and quality-assessed *via* Bioanalyzer. Single strand index linker and 3' linker was ligated to released cDNA strands, and primer containing Illumina Sequencer Priming site was used for second strand synthesis. Samples were sequenced using the Illumina NextSeq 500 platform producing 76 bp reads (Supplementary Table 1).

3' RNA-seq was carried out with samples E-5h_1, E-5h_2, L-16h_1 and L-16h_2 using the Lexogen QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina according to manufacturer's instructions. Library preparation and sequencing were carried out Cambridge Genomic Services (Department of Pathology, University of Cambridge, Cambridge, UK) on a single NextSeq flowcell producing 150 bp (Supplementary Table 1).

Sequencing Quality Checks and Mapping to ASFV and Vero Genomes

FastQC (67) analysis was carried out on all FASTQ files: for RNA-seq FASTQ files were uploaded to the web-platform Galaxy (www.usegalaxy.org/) (68, 69) and all reads were trimmed by the first 10 and last 1 nt using FASTQ Trimmer (70). After read-trimming, FASTQ files originating from the same RNA samples were concatenated. RNA-seq reads were mapped to the ASFV-BA71V (NC_001659.2) and *Vero* (GCF_000409795.2) genomes using Bowtie 2 directly after trimming (27), with alignments output in SAM file format. FASTQ analysed CAGE-seq reads showed consistent read quality across the 76 bp reads, except for the nucleotide 1. This was an indicator of the 5' mRNA methylguanosine due to the reverse transcriptase used in library preparation (71), therefore, the reads were mapped in their entirety to the ASFV-BA71V (U18466.2) and *Vero* (GCF_000409795.2) genomes.

FASTQC analysed 3' RNA-seq reads showed relatively varying and poorer quality after nucleotide 65. Cutadapt (72) was utilised to extract only fastq reads with 18 consecutive A's at the 3' end followed by the sample i7 Illumina adapter, selecting only for reads containing the 3' mRNA end and the polyA

tail. The 18A-adapter sequences were then trimmed and FASTQC-analysed reads were mapped *via* Bowtie2 to ASFV-BA71V (U18466.2) and *Vero* (GCF_000409795.2) genomes.

CAGE Analysis, TSS-Mapping

CAGE-seq mapped sample BAM files were converted to BigWig (BW) format with BEDtools (73) genomecov, to produce per-strand BW files of 5' read ends. Stranded BW files were input for TSS-prediction in RStudio (74) with Bioconductor (75) package CAGEfightR (76). Genomic feature locations were imported as a TxDb object from U18466.2 genome gene feature file (GFF3), modified to include C44L (12). CAGEfightR was used to quantify the CAGE tag transcripts mapping at base pair resolution to the ASFV-BA71V genome - at CAGE TSSs (CTSSs). CTSS values were normalized by tags-per-million for each sample, pooled and only CTSSs supported by presence in ≥2 samples were kept. CTSSs were assigned to clusters, merging CTSSs within 50 bp of one another, filtering out pooled, TPM-normalized CTSS counts below 25, and assigned a 'thick' value as the highest CTSS peak within that cluster. CTSS clusters were assigned to annotated U18466.2 ORFs (if clusters were between 300 bp upstream and 200 bp downstream of an ORF). Clusters were classified 'tssUpstream' if located within 300 bp upstream of an ORF, 'proximal' if located within 500 bp of an ORF, 'CDS' if within the ORF, 'NA' if no annotated ORF was within these regions (excepting pNG), and antisense if within these regions but antisense relative to the ORF.

Cluster classification was not successful in all cases, therefore, manual adjustment was necessary. Integrative Genomics Viewer (IGV) (77) was used to visualise BW files relative to the BA71V ORFs, and incorrectly classified clusters were corrected. Clusters with the 'tssUpstream' classification were split into subsets for each ORF. 'Primary' cluster subset contained either the highest scoring CAGEfightR cluster or the highest scoring manually-annotated peak, and the highest peak coordinate was defined as the primary TSS (pTSS) for an ORF. Further clusters associated these ORFs were classified as 'non-primary', highest peak as a non-primary TSS (npTSS).

If the strongest CTSS location was intra-ORF and corroborated with RNA-seq coverage, then the ORF was re-defined as starting from the next ATG downstream. For the 28 intergenic CTSSs, IGV was used

513 to visualise if CAGE BW peaks were followed by RNA-seq coverage downstream, and whether the
514 transcribed region encode a putative ORF using NCBI Open Reading Frame Finder (78).

515 ## TTS-Mapping

516 TTSs were mapped in a similar manner to TSSs and CAGEfightR was utilised as above to locate
517 clusters of 3' RNA-seq peaks, though differed in some respects: input BigWig files contained the 3'
518 read-end coverage extracted from BAM files using BEDtools genomecov. Clusters were detected for
519 the 3' RNA-seq peaks in the same manner as before, except merging clusters < 25 nt apart, which
520 detected a total of 567 clusters. BEDtools was used to check whether the highest point of each
521 cluster (TTS) was within 500 bp or 1000 bp downstream of annotated ORFs and pNGs. TTSs were
522 then filtered out if 10 nt downstream of the 3' end had ≥ 50% As, to exclude clusters potentialy
523 originated from miss-priming. TTS clusters for pNG3 and pNG4 were initially filtered out but included
524 in final 212 TTSs due to their strong RNA-seq agreement. In cases of multiple TTS clusters per gene
525 we defined the highest CAGEfightR-scored one within 1000 bp downstream of ORFs as primary
526 (pTTS) unless no clear RNA-seq coverage was shown, or manually annotated from the literature for
527 O61R (15).

528 ## DESeq2 Differential Expression Analysis of ASFV Genes

529 A new GFF was produced for investigating differential expression of ASFV genes across the genome
530 with changes from the original U18466.2.gff: for all 151 ASFV ORFs which had identified pTSSs, we
531 defined their transcription unit as beginning from the pTSS coordinate to ORF end. Since no pTSS was
532 identified for ORFs E66L and C62L these entries were left as ORFs within the GFF, while the 7 putative
533 pNGs were defined as their pTSS down to the genome coordinate at which the RNA-seq coverage
534 ends. In 8 cases where genes had alternative pTSSs for the different time-points the TUs were
535 defined as the most upstream pTSS down to the ORF end. For analysing differential expression with
536 the CAGE-seq dataset a GFF was created with BEDtools extending from the pTSS coordinate, 25 bp
537 upstream and 75 bp downstream, however, in cases of alternating pTSSs this TU was defined as 25

bp upstream of the most upstream pTSS and 75 bp downstream of the most downstream pTSS. HTSeq-count (79) was used to count reads mapping to genomic regions described above for both the RNA- and CAGE-seq sample datasets. The raw read counts were then used to analyse differential expression across these regions between the time-points using DESeq2 (default normalisation described by Love et al. (80)) and those regions showing changes with an adjusted *p*-value (padj) of <0.05 were considered significant. Further analysis of ASFV genes used their characterised or predicted functions as found in the VOCS tool database (https://4virology.net/) (38, 81) or ASFVdb (39) entries for the ASFV-BA71V genome.

Early and Late Promoter Analysis

DESeq2 results were used to categorise ASFV genes into two simple sub-classes: early; genes downregulated from early to late infection and late; those upregulated from early to late infection. For those with newly annotated pTSSs (151 including 7 pNGs but excluding 15 alternative pTSSs), sequences 30 bp upstream and 5 bp downstream were extracted from the ASFV-BA71V genome in FASTA format using BEDtools. The 36 Early, 55 Late and all 166 pTSSs (including alternative ones) at once were analysed using MEME software (http://meme-suite.org) (82), searching for 5 motifs with a width of 10–25 nt, other settings at default. Significant motifs (E-value < 0.05) detected *via* MEME were submitted to a following FIMO (42) search (*p*-value cut-off < 0.0001) of 60 nt upstream of the total 166 pTSS sequences (including pNGs and alternative pTSSs), and Tomtom software (46) search (UP00029_1, Database: uniprobe_mouse) to find similar known motifs.

Data Availability

Sequencing data from RNA-seq, CAGE-seq and 3' RNA-seq are available on Sequence Read Archive (SRA), BioProject: PRJNA590857 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA590857?reviewer=e597nf6o3r2hk5r45a5hqr9d). The processed data for two replicates are visualized in an UCSC Genome Browser [pmid: 24227676] and can be accessed at https://bit.ly/2TazQxK. The tracks include corrected gene annotations (primary TSSs, primary TTSs, and ORF coordinates), raw coverage of 5' ends (CAGE-seq) and 3' ends

(3'-RNA-seq), and RPKM values for the RNA-seq data. Coverage for the forward and reverse strands are shown in blue and red, respectively.

Results from differential gene expression analysis with DESeq2 of CAGE-seq and RNA-seq are found in Supplementary Tables 4 and 5, respectively. The 91 genes showing the same pattern of differential expression according to both of these NGS techniques are found in Supplementary Table 6. Details of non-templated extensions detected from CAGE-seq are in Supplementary Table 7. CAGEfightR-detected cluster peaks from 3' RNA-seq after removal of those arriving from polyA miss-priming are described in Supplementary Table 8. All 779 CAGEfightR-detected cluster peaks from CAGE-seq are listed in Supplementary Table 9.

Acknowledgements and Funding

Research in the RNAP laboratory at UCL is funded by a Wellcome Investigator Award in Science *'Mechanisms and Regulation of RNAP transcription'* to FW (WT 207446/Z/17/Z) and to JB [WT 095598/Z/11/Z]. GC is funded by the Wellcome Trust ISMB 4-year PhD programme *'Macromolecular machines: interdisciplinary training grounds for structural, computational and chemical biology'* (WT 108877/B/15/Z). The authors are grateful to all members of the RNAP lab and Tine Arnvig for critical reading of the manuscript.

Competing Interests

The authors declare that no competing interests exist.

584 ## References

585 1. Alonso C, Borca M, Dixon L, Revilla Y, Rodriguez F, Escribano JM, Consortium IR. 2018. ICTV
586 Virus Taxonomy Profile: Asfarviridae. J Gen Virol 99:613–614.
587 2. Koonin E V, Yutin N. 2010. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA
588 viruses. Intervirology 53:284–92.
589 3. Yutin N, Koonin E V. 2012. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA
590 viruses of eukaryotes. Virol J 9:161.
591 4. Reteno DG, Benamar S, Khalil JB, Andreani J, Armstrong N, Klose T, Rossmann M, Colson P,
592 Raoult D, La Scola B. 2015. Faustovirus, an asfarvirus-related new lineage of giant viruses
593 infecting amoebae. J Virol 89:6585–94.
594 5. Gogin A, Gerasimov V, Malogolovkin A, Kolbasov D. 2013. African swine fever in the North
595 Caucasus region and the Russian Federation in years 2007-2012. Virus Res 173:198–203.
596 6. Zhou X, Li N, Luo Y, Liu Y, Miao F, Chen T, Zhang S, Cao P, Li X, Tian K, Qiu H-J, Hu R. 2018.
597 Emergence of African Swine Fever in China, 2018. Transbound Emerg Dis 65:1482–1484.
598 7. Dixon LK, Chapman DAG, Netherton CL, Upton C. 2013. African swine fever virus replication
599 and genomics. Virus Res 173:3–14.
600 8. Kinyanyi D, Obiero G, Obiero GFO, Amwayi P, Mwaniki S, Wamalwa M. 2018. In silico
601 structural and functional prediction of African swine fever virus protein-B263R reveals
602 features of a TATA-binding protein. PeerJ 6:e4396.
603 9. Rodriguez JM, Salas ML, Viñuela E. 1992. Genes homologous to ubiquitin-conjugating proteins
604 and eukaryotic transcription factor SII in African swine fever virus. Virology 186:40–52.
605 10. Rodríguez JM, Salas ML. 2013. African swine fever virus transcription. Virus Res 173:15–28.
606 11. Broyles SS. 2003. Vaccinia virus transcription. J Gen Virol 84:2293–2303.
607 12. Kollnberger SD, Gutierrez-Castañeda B, Foster-Cuevas M, Corteyn A, Parkhouse RME. 2002.
608 Identification of the principal serological immunodeterminants of African swine fever virus by
609 screening a virus cDNA library with antibody. J Gen Virol 83:1331–1342.

13. Yáñez RJ, Rodríguez JM, Nogal ML, Yuste L, Enríquez C, Rodriguez JF, Viñuela E. 1995. Analysis of the complete nucleotide sequence of African swine fever virus. Virology 208:249–78.

14. Rodríguez JM, Moreno LT, Alejo A, Lacasta A, Rodríguez F, Salas ML. 2015. Genome Sequence of African Swine Fever Virus BA71, the Virulent Parental Strain of the Nonpathogenic and Tissue-Culture Adapted BA71V. PLoS One 10:e0142889.

15. Almazán F, Rodríguez JM, Angulo A, Viñuela E, Rodriguez JF. 1993. Transcriptional mapping of a late gene coding for the p12 attachment protein of African swine fever virus. J Virol 67:553–6.

16. Almazán F, Rodríguez JM, Andrés G, Pérez R, Viñuela E, Rodriguez JF. 1992. Transcriptional analysis of multigene family 110 of African swine fever virus. 66.

17. Breese SS, DeBoer CJ. 1966. Electron microscope observations of African swine fever virus in tissue culture cells. Virology 28:420–428.

18. Oda KI, Joklik WK. 1967. Hybridization and sedimentation studies on “early” and “late” vaccinia messenger RNA. J Mol Biol 27:395–419.

19. Zhang X, Kiechle FL. 2004. Cytosine arabinoside substitution decreases transcription factor-DNA binding element complex formation. Arch Pathol Lab Med 128:1364–1371.

20. Alejo A, Matamoros T, Guerra M, Andrés G. 2018. A proteomic atlas of the African swine fever virus particle. J Virol JVI.01293-18.

21. Salas ML, Kuznar J, Viñuela E. 1981. Polyadenylation, methylation, and capping of the RNA synthesized in vitro by African swine fever virus. Virology 113:484–491.

22. Frouco G, Freitas FB, Coelho J, Leitão A, Martins C, Ferreira F. 2017. DNA-Binding Properties of African Swine Fever Virus pA104R, a Histone-Like Protein Involved in Viral Replication and Transcription. J Virol 91.

23. Iyer LM, Balaji S, Koonin E V., Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 117:156–84.

24. Yutin N, Wolf YI, Raoult D, Koonin E V. 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses:

clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223.

25. García-Escudero R, Viñuela E. 2000. Structure of African swine fever virus late promoters: requirement of a TATA sequence at the initiation region. J Virol 74:8176–82.

26. Rodríguez JM, Salas ML, Viñuela E. 1996. Intermediate class of mRNAs in African swine fever virus. J Virol 70:8584–9.

27. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

28. Yang Z, Martens CA, Bruno DP, Porcella SF, Moss B. 2012. Pervasive initiation and 3'-end formation of poxvirus postreplicative RNAs. J Biol Chem 287:31050–60.

29. Yang Z, Bruno DP, Martens CA, Porcella SF, Moss B. 2011. Genome-wide analysis of the 5' and 3' ends of vaccinia virus early mRNAs delineates regulatory sequences of annotated and anomalous transcripts. J Virol 85:5897–909.

30. Schoenberg DR, Maquat LE. 2009. Re-capping the message. Trends Biochem Sci 34:435–42.

31. Mukherjee C, Patil DP, Kennedy BA, Bakthavachalu B, Bundschuh R, Schoenberg DR. 2012. Identification of cytoplasmic capping targets reveals a role for cap homeostasis in translation and mRNA stability. Cell Rep 2:674–84.

32. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris K V., Morillon A, Rozowsky JS, Gerstein MB, Wahlestedt C, Hayashizaki Y, Carninci P, Gingeras TR, Mattick JS. 2011. The reality of pervasive transcription. PLoS Biol 9:e1000625.

33. Castelnuovo M, Stutz F. 2015. Role of chromatin, environmental changes and single cell heterogeneity in non-coding transcription and gene regulation. Curr Opin Cell Biol. Elsevier Ltd.

34. Mirzakhanyan Y, Gershon PD. 2017. Multisubunit DNA-Dependent RNA Polymerases from Vaccinia Virus and Other Nucleocytoplasmic Large-DNA Viruses: Impressions from the Age of Structure. Microbiol Mol Biol Rev 81:e00010-17.

35. Kim B, Nesvizhskii AI, Rani PG, Hahn S, Aebersold R, Ranish JA. 2007. The transcription

662 elongation factor TFIIS is a component of RNA polymerase II preinitiation complexes. Proc
663 Natl Acad Sci 104:16068–16073.
664 36. Awrey DE, Shimasaki N, Koth C, Weilbaecher R, Olmsted V, Kazanis S, Shan X, Arellano J,
665 Arrowsmith CH, Kane CM, Edwards AM. 1998. Yeast transcript elongation factor (TFIIS),
666 structure and function. II: RNA polymerase binding, transcript cleavage, and read-through. J
667 Biol Chem 273:22595–605.
668 37. Keßler C, Forth JH, Keil GM, Mettenleiter TC, Blome S, Karger A. 2018. The intracellular
669 proteome of African swine fever virus. Sci Rep 8:14714.
670 38. Upton C, Slack S, Hunter AL, Ehlers A, Roper RL, Rock DL. 2003. Poxvirus orthologous clusters:
671 toward defining the minimum essential poxvirus genome. J Virol 77:7590–600.
672 39. Zhu Z, Meng G. 2019. ASFVdb: An integrative resource for genomics and proteomics analyses
673 of African swine fever. bioRxiv 670109.
674 40. Butler JEF, Kadonaga JT. 2002. The RNA polymerase II core promoter: a key component in the
675 regulation of gene expression. Genes Dev 16:2583–92.
676 41. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009.
677 MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208.
678 42. Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif.
679 Bioinformatics 27:1017–8.
680 43. Davison AJ, Moss B. 1989. Structure of vaccinia virus early promoters. J Mol Biol 210:749–769.
681 44. Yang Z, Bruno DP, Martens CA, Porcella SF, Moss B. 2010. Simultaneous high-resolution
682 analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing. Proc Natl
683 Acad Sci U S A 107:11513–8.
684 45. Sýkora M, Pospíšek M, Novák J, Mrvová S, Krásný L, Vopálenský V. 2018. Transcription
685 apparatus of the yeast virus-like elements: Architecture, function, and evolutionary origin.
686 PLOS Pathog 14:e1007377.
687 46. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W. 2007. Quantifying similarity between

motifs. Genome Biol 8:R24.

47. Arimbasseri AG, Rijal K, Maraia RJ. 2013. Comparative overview of RNA polymerase II and III transcription cycles, with focus on RNA polymerase III termination and reinitiation. Transcription. Taylor and Francis Inc.

48. Chen K, Hu Z, Xia Z, Zhao D, Li W, Tyler JK. 2016. The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses.

49. Schwalb B, Michel M, Zacher B, Hauf KF, Demel C, Tresch A, Gagneur J, Cramer P. 2016. TT-seq maps the human transient transcriptome. Science (80-) 352:1225–1228.

50. Kuznar J, Salas ML, Viñuela E. 1980. DNA-dependent RNA polymerase in African swine fever virus. Virology 101:169–75.

51. Dunn LEM, Ivens A, Netherton CL, Chapman DAG, Beard PM. 2019. Identification of a functional small non-coding RNA encoded by African swine fever virus. bioRxiv 865147.

52. Kazmierczak MJ, Wiedmann M, Boor KJ. 2005. Alternative Sigma Factors and Their Roles in Bacterial Virulence. Microbiol Mol Biol Rev 69:527–543.

53. Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. 2019. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. Proc Natl Acad Sci U S A 116:19585–19592.

54. Yáñez RJ, Rodríguez JM, Boursnell M, Rodriguez J, Viñuela E. 1993. Two putative African swine fever virus helicases similar to yeast 'DEAH' pre-mRNA processing proteins and vaccinia virus ATPases D11L and D6R. Gene 134:161–174.

55. Hahn S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. Nat Struct Mol Biol 11:394–403.

56. Knutson BA, Liu X, Oh J, Broyles SS. 2006. Vaccinia virus intermediate and late promoter elements are targeted by the TATA-binding protein. J Virol 80:6784–93.

57. Broyles SS, Knutson BA. 2010. Poxvirus transcription. Future Virol 5:639–650.

58. Yang Z, Reynolds SE, Martens CA, Bruno DP, Porcella SF, Moss B. 2011. Expression Profiling of

714 the Intermediate and Late Stages of Poxvirus Replication † Downloaded from. J Virol 85:9899–
715 9908.
716 59. Dhungel P, Cao S, Yang Z. 2017. The 5'-poly(A) leader of poxvirus mRNA confers a translational
717 advantage that can be achieved in cells with impaired cap-dependent translation. PLoS Pathog
718 13:e1006602.
719 60. Mulder J, Robertson ME, Seamons RA, Belsham GJ. 1998. Vaccinia virus protein synthesis has
720 a low requirement for the intact translation initiation factor eIF4F, the cap-binding complex,
721 within infected cells. J Virol 72:8813–9.
722 61. Shirokikh NE, Spirin AS. 2008. Poly(A) leader of eukaryotic mRNA bypasses the dependence of
723 translation on initiation factors. Proc Natl Acad Sci U S A 105:10738.
724 62. Kuehner JN, Pearson EL, Moore C. 2011. Unravelling the means to an end: RNA polymerase II
725 transcription termination. Nat Rev Mol Cell Biol.
726 63. Santangelo TJ, Cubonová L, Skinner KM, Reeve JN. 2009. Archaeal intrinsic transcription
727 termination in vivo. J Bacteriol 191:7102–7108.
728 64. Howard ST, Ray CA, Patel DD, Antczak JB, Pickup DJ. 1999. A 43-nucleotide RNA cis-acting
729 element governs the site-specific formation of the 3' end of a poxvirus late mRNA. Virology
730 255:190–204.
731 65. Shuman S, Moss B. 1988. Factor-dependent transcription termination by vaccinia virus RNA
732 polymerase. Evidence that the cis-acting termination signal is in nascent RNA. J Biol Chem
733 263:6220–5.
734 66. Freitas FB, Frouco G, Martins C, Ferreira F. 2019. The QP509L and Q706L superfamily II RNA
735 helicases of African swine fever virus are required for viral replication, having non-redundant
736 activities. Emerg Microbes Infect 8:291–302.
737 67. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. 0.11.7.
738 Babraham Bioinformatics, Babraham, England.
739 68. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D,

740 Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche E,
741 Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J. 2016. The Galaxy platform for
742 accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids
743 Res 44:W3–W10.
744 69. Gruening BA. 2104. Galaxy wrapper.
745 70. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy Team. 2010.
746 Manipulation of FASTQ data with Galaxy. Bioinformatics 26:1783–1785.
747 71. Potter J, Zheng W, Lee J. 2003. Thermal stability and cDNA synthesis capability of Superscript
748 III reverse transcriptase. Focus (Madison) 25:19–24.
749 72. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing
750 reads. EMBnet.journal 17:10.
751 73. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
752 features. Bioinformatics 26:841–842.
753 74. RStudio Team. 2016. RStudio: Integrated Development for R. 3.4.3. RStudioe, Inc, Boston, MA.
754 75. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto
755 L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J,
756 Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L,
757 Morgan M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. Nat
758 Methods 12:115–121.
759 76. Thodberg M, Thieffry A, Vitting-Seerup K, Andersson R, Sandelin A. 2018. CAGEfightR: Cap
760 Analysis of Gene Expression (CAGE) in R/Bioconductor. bioRxiv 310623.
761 77. Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-
762 performance genomics data visualization and exploration. Brief Bioinform 14:178–192.
763 78. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml
764 LM, Sequeira E, Tatusova TA, Wagner L. 2003. Database resources of the National Center for
765 Biotechnology. Nucleic Acids Res 31:28–33.

766 79. Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput
767 sequencing data. Bioinformatics 31:166–169.
768 80. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for
769 RNA-seq data with DESeq2. Genome Biol 15:550.
770 81. Tu SL, Upton C. 2019. Bioinformatics for Analysis of Poxvirus Genomes, p. 29–62. *In* Methods
771 in Molecular Biology. Humana Press Inc.
772 82. Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover
773 motifs in biopolymers. Proceedings Int Conf Intell Syst Mol Biol 2:28–36.
774 83. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2-A
775 multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.
776 84. Kettenberger H, Armache K-J, Cramer P. 2003. Architecture of the RNA Polymerase II-TFIIS
777 Complex and Implications for mRNA Cleavage. Cell 114:347–357.
778 85. Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: A Sequence Logo Generator.
779 Genome Res 14:1188–1190.
780 86. Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus
781 sequences. Nucleic Acids Res 18:6097–6100.
782 87. Andrés G, García-Escudero R, Viñuela E, Salas ML, Rodríguez JM. 2001. African swine fever
783 virus structural protein pE120R is essential for virus transport from assembly sites to plasma
784 membrane but not for infectivity. J Virol 75:6758–68.
785 88. NCBI Resource Coordinators NR. 2016. Database resources of the National Center for
786 Biotechnology Information. Nucleic Acids Res 44:D7-19.

787
788
789
790

791 **Figure 1. Annotated genome of ASFV-BA71V indicating transcription start sites (TSS) and early and late**
792 **genes.** The map includes 153 previously annotated as well as novel genes identified in this study and their
793 differential expression pattern from early to late infection from DESeq2 (80) analysis. Early genes (upregulated,
794 highlighted in dark blue) and late genes (upregulated, dark red) were differentially expressed according to both
795 RNA-seq and CAGE-seq approaches. The pale blue and pale red marking indicates a negative (early,
796 downregulated) or positive (late, upregulated) log2 fold change in expression according to both CAGE-seq and
797 RNA-seq, but is only statistically significant (adjusted *p*-value < 0.05) from CAGE-seq, due to its higher
798 sequencing depth and unlike RNA-seq, is not affected by transcription readthrough. Colour coding in white
799 suggests ambivalency of early and late expression patterns, i.e. not statistically significant according to either of
800 the methods, or only according to RNA-seq. These also include ten genes with reversed differential expression
801 between CAGE-seq and RNA-seq results. The map was visualised with the R package gggenes.
802

803 **Figure 2. The ASFV transcriptome including transcription start sites and termination sites.** (a) Whole genome
804 view of normalized coverage counts per million (CPM) of RNA-seq, 5' CAGE-seq and 3' RNA-seq reads. The
805 coverage was capped at 16000 CPM. 153 BA71V annotated ORFs are represented as arrows and coloured
806 according to strand. Peak cluster shape example from F1055L 5' CAGE-seq ends (b) and 3' RNA-seq ends (c)
807 showing a wide multi-peaked distribution, and J64R 5' CAGE-seq (d) and 3' RNA-seq (e) showing a narrow peak
808 distribution.
809

810 **Figure 3. Transcriptome mapping aids the reannotation of the ASFV BA71V genome**. (a-left) Summary bar
811 graph of CAGEfightR TSS clusters and their locations relative to the 153 annotated BA71V ORFs. (a-right) Types
812 of CAGEfightR clusters detected and the distribution of their respective CAGEfightR scores. (b) Two examples of
813 ORFs requiring re-annotation following pTSS identification downstream of annotated start codon, encoding
814 shorter ORFs from the pTSS (I177L, above) or during one expression stage (B169L, below). (c) Examples of two
815 putative novel genes (pNG3, left and pNG1 right) annotated with the normalized RNA-seq and CAGE-seq read
816 coverage (CPM) and their genome neighbourhood.

817 **Figure 4. Analysis of alternative pTSS usage in I243L.** (a) Close up of TSSs (CAGE-seq alignments) on the minus
818 strand at the start of the I243L ORF. Symbols indicate the TSS sites for early (▼), intermediate (●)and late (▽)
819 gene expression according to Rodríguez *et al*. (26), while E, I and L indicate their respective pTSS positions
820 concluded from our data. The first 21 AA residues of the annotated I243L ORF are shown, in yellow is the re-
821 annotated ORF which could be encoded in transcripts initiating from both our annotated Early pTSS. (b)
822 ClustalW multiple sequence alignment coloured by percentage identity between sequences, illustrated with
823 Jalview (83), of TFIIS homologues from ASFV (I243L, UniProt: P27948), *A. thaliana* (Q9ZVH8), *D. melanogaster*
824 (P20232), human (P23193), mouse (P10711) and *S. cerevisiae* (P07273). *S. cerevisiae* TFIIS domain locations
825 according to Kettenberger *et al*. (84) are shown below the alignment and acidic (DE) catalytic residues are
826 indicated with★. ASFV-TFIIS start codons encoded from alternative transcription start sites are labelled as in
827 (b).

828

829 **Figure 5. Gene expression of ASFV genes during early and late infection.** (a) FPKM values for 20 most highly
830 expressed ASFV TUs according to CAGE-seq at 5h (left) and 16h (right) post-infection. Genes highlighted in
831 maroon indicate those encoding proteins which were also found in the 20 most-abundantly expressed AFSV
832 proteins during infection of either WSL-HP, HEK293 or *Vero* cells according to proteome analysis done by
833 Keßler *et al*. (37). Gene functions are shown after their name with TR and PSP referring to predicted
834 transmembrane region and putative signal peptide, respectively. (b) 20 most-expressed genes during early
835 (green) and late (blue) infection according to RNA-seq data over gene TU, defined from TSS to ORF stop codon.
836 (c) MAplot from DESeq2 analysis of CAGE-seq representing the DESeq2 base mean of transcript levels versus
837 their log2 fold change, with significantly differentially expressed genes in purple (adjusted p-value < 0.05). (d)
838 MAplot representing expression of ASFV TUs including pNGs from DESeq2 analysis of RNA-seq data.

839

840 **Figure 6. Relative expression during infection stages and defining early and late genes.** (a) Boxplot mean FPM
841 values for the early and late genes at early and late infection, respectively. Outliers are labelled with their gene
842 names and Wilcoxon rank sum test showed the mean FPM of early genes during early infection was
843 significantly greater than that of late genes during late infection (p-value: 1.865e-06). (b) Distribution of gene
844 expressed for the least and most expressed genes during early and late infection. Genes in the 15th percentile

845 for their mean FPM values from each time-point, being below an early FPM threshold of 7.56 (blue) and late
846 FPM of 199.64 (red). (c) Genes in the 85th percentile for their mean FPM values from each time-point, being
847 above an early FPM threshold of 8148.91 (blue) and late FPM of 4706.27 (red). In dark blue and dark red are
848 medians for the plotted expression values for early and late infection respectively. (d) Scatter plot comparing
849 log2fold changes of the 101 significantly differentially expressed genes in common between RNA-seq and
850 CAGE-seq. Labels were coloured according to their significant upregulation or downregulation from RNA-seq.
851 (e) Pie chart of gene functional categories downregulated from 5 h to 16 h (36 early genes) and upregulated
852 from 5 h to 16 h (55 late genes). Fisher test carried out on gene counts for functional groups between early and
853 late infection, for this all MGF members were pooled into the 'MGFs' functional group.
854
855 **Figure 7. Initiator and promoter sequence signatures of ASFV genes.** WebLogo 3 (85, 86) of aligned early (a)
856 and late (b) sequences surrounding the Inr (+1) from −35 to +10, with gradients representing the basepair
857 conservation of the EPM (blue-white), Inr (purple-white) and LPM (peach-white). WebLogo 3 consensus motif
858 with error-bars, of the 36 early (c) and 55 late (d) gene sequences surrounding their respective pTSSs (5 nt up-
859 and downstream), i.e. initiator (Inr) motif. (e) EPM located upstream of all 36 of our classified early genes
860 according to MEME motif search (E-value: 8.2e-021), FIMO with a threshold of *p*-value < 1.0 E-4 then identified
861 at least one iteration of this motif upstream of 81 ASFV genes. (f) Distances of the EPM motif 3' end (nt 19)
862 relative to the 78 pTSSs (alternative pTSSs excluded). (4). (g) Expression profiles from DESeq2 analysis (log2fold
863 change vs. base mean expression) of genes with only an EPM from the FIMO search of 60 bp upstream of
864 pTSSs. Genes for which FIMO detected both EPM and LPM upstream of pTSSs were excluded. Genes in blue
865 showed a negative log2 fold change (early genes) and in red a positive log2fold change (regardless of
866 significance). (h) Expression profiles as in c. for the 26 MGFs where an EPM was detected upstream. (i)
867 Distances of the EPM motif 3' end (nt 19) relative to the MGF pTSSs.
868
869 **Figure 8. Promoter motif upstream of ASFV late genes.** (a) The LPM detected upstream of 17 of our classified
870 late genes from MEME motif search (E-value: 1.6e-003). (b) Distances from a FIMO search (threshold *p*-value <
871 1.0E-4) identified the LPM upstream of 53 ASFV genes (excluding those with alternative pTSSs), motif distances
872 from pTSSs are represented as a bar chart. (c) Expression profiles as in c. of genes with only an LPM from the
873 FIMO search of 60 bp upstream of pTSSs. (d) The eukaryotic TATA-box motif which was one of 28 hits in a

Tomtom search of the LPM. (e) 5' UTR lengths in nt of the 91 early (mean: 39, median: 14) or late (mean: 25, median: 9) classified ASFV genes, starting from the most upstream pTSS (in the case of alternating pTSSs) until the first ATG start codon nt, represented as a notched boxplot. 9 genes with 5' UTR's above 80 nt were excluded from the boxplot: QP509L (92 nt long), pNG2 (105 nt), I267L (110 nt), B318L (118 nt), C44L (131 nt), DP141L (165 nt), pNG1 (223 nt), EP402R (242 nt) and A118R (332 nt). (f) Percentage AT content of early (mean: 69.0, median: 70.9) and late (mean: 81.7, median: 83.3) 5' UTRs, omitting those of 0 length.

Figure 9. Investigating ASFV-RNAP slippage. (a) Frequency of different lengths of template-free extensions in early and late stage samples. (b) Relationship between the length of templated 5' UTRs and fraction of template-free extensions. Gene 5' UTRs split into 36 early (blue), 55 late (orange) and not-classified ('NC', green). (c) Frequency of most common template-free extensions in the early and late stage samples. (d) Sequence logo of region surrounding TSS of 'AU' and 'AUAU'-extended transcripts.

Figure 10. ASFV transcription termination. (a) WebLogo 3 motif of 10 nt upstream and 10 downstream of all pTTS and npTTSs with a polyT upstream with ≥4 consecutive Ts based on126 TTSs. (b) Distance from 3' terminal T in polyT motif to the TTS (median). (c) The distribution of polyT lengths among 126 polyT TTSs (median: 7), split into expression stages according to CAGE-seq differential expression analysis (NC: not-classified), showing late gene polyTs are shorter in length (Wilcoxon rank sum test, *p*-value: 0.0216). (d) Distribution of gene expression types among the 83 polyT pTTSs and 31 non-polyT pTTSs. Dotted lines labels indicate Fisher test *p*-values of gene types between the two pTTS-types, classified from CAGE-seq. (e) 55 Early and 53 late gene 3' UTR lengths from stop codon to pTTS (Wilcoxon rank sum test, *p*-value: 0.003).

895 Tables

ORF	Strand	pTSS Coordinate	Corrected Start Codon Coordinate	ORF Length	Comment
K93L	-	2131	2122	83	Alternative ATG codon 30 nt downstream. Another strong TSS was detected at 2037- whose transcripts would encode a 36 AA protein.
F165R	+	42354	42359	136	Alternative ATG codon 63 nt downstream.
C84L	-	64618	64492 | 64616	38 | 76	38 AA ORF was in-frame with original C84L start codon. 76 AA ORF encoded from first ATG after pTSS.
G1211R	+	96370	96377	1207	Alternative ATG codon 12 nt downstream.
CP204L	-	108573	108567	196	Alternative ATG codon 24 nt downstream.
CP312R	+	110491	110501	307	Alternative ATG codon 15 nt downstream.
I177L	-	L: 157857	157849	66	Strongest pTSS only detected in late time-point.
DP93R	+	167971	167980	83	Alternative ATG codon 30 nt downstream.
EP402R	+	56862	57104 | 56991	115 | 148	Encodes 115 AA in-frame with original EP402R start codon. 148 AA alternative ORF encoded from first ATG after pTSS.
B169L	**-**	**E: 80983 | L: 81025**	**81018 | 80745**	**169 | 78**	**Late pTSS can produce full-length B169L and early pTSS: 78 AA.**
I243L	**-**	**E: 155122 I: 155124**	**E/I: 155119 L: 154969**	**243 | 191**	**Late pTSS produces shorter transcript with closest downstream ATG encoding a**

		L: 155115			**shorter 191 AA protein.**

Table 1. Summary of ASFV genes where pTSS locations guided the re-annotation of ORFs. For B169L and I243L, the letters E, I and L refer to alternative pTSSs from early, intermediate and late infection, respectively reported by Rodríguez *et al.* (26).

Gene	**Early pTSS**	**Late pTSS**	**Function**
X69R	11315	11280	Uncharacterised
J154R	14174	14150	MGF 300-2R
EP1242L	53125	53135	ASFV-RPB2
C315R	70137	70131	ASFV-TFIIB
CP80R	110208	110191	ASFV-RPB10
D345L	129357	129257	Lambda-like exonuclease (7)
E120R	150949	150911	Structural protein (87)

Table 2. Alternative pTSS usage during early and late ASFV infection. List of ASFV genes with alternative pTSSs used in early and late infection.

Putative Gene	Strand	Transcription Start*	Transcription End**	Putative protein length (AA)	Similarity according to NCBI Blast	Gene-End oligoT (nt)
pNG1	+	13053	13435	25	13 residues had 92% identity to ASFV-G-ACD-00350 (AZP54308.1), E-value: 0.11	6
pNG2	-	30091	29827	50	50 residues had 100% identity with ASFV26544 00600 (AKM05534.1)	8
pNG3	+	12664	12896	44	38 residues had 59% identity to ASFV-G-ACD-00290 (AZP54130.1), E-value: 0.13	6
pNG4	+	10583	10835	44	42 residues had 65% identity with ASFV-G-ACD-00290 (AZP54130.1), E-value: 1e-09.	6
pNG5	+	29817	30080	31	No significant similarity.	None
pNG6	+	167005	167336	56	56 residues aligned with 40% identity to pKP93L (AIY22188.1), E-value: 6e-07	5
pNG7	+	10484	10616	32	32 residues aligned with a 31 AA hypothetical protein with from ASFV Belgium 2018/1 (BioProject: PRJEB31287) 87% identity (VFV47940.1), E-value: 8e-10.	3

Table 3. Details of seven novel ASFV candidate genes. NCBI ORFfinder and BLAST were used to predict the putative encoded ORFs and subsequently analysed for putative homologous sequences (78, 88). *: defined as pTSS from CAGE-seq. **: defined from 3' RNA-seq, underlined transcription ends defined from only RNA-seq. pNG5 is in the antisense orientation relative to pNG2, and the RNA-3' end of pNG6 is fuzzy according to RNA-seq and may overlap with DP42R. pNG7 is overlapping pNG4 on the same strand.