# The determinants of African Swine Fever Virus Virulence – the Georgia 2007/1 strain and the host macrophage response

Gwenny Cackett[a§], Raquel Portugal[b§] , Dorota Matelska[a], Linda Dixon[b*] and Finn Werner*[a]

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

[b]Pirbright Institute, Ash Road, Pirbright, Surrey, GU24 0NF, United Kingdom

[§] have contributed equally to this work

Short Title:

The ASFV Georgia 2007/1 Strain Transcriptome

#Address correspondence to linda.dixon@pirbright.ac.uk and f.werner@ucl.ac.uk.

## Abstract [217 words]

African swine fever virus (ASFV) has a major global economic impact. With a case fatality in domestic pigs approaching 100%, it currently presents the largest threat to animal farming. Although genomic differences between attenuated and highly virulent ASFV strains have been identified, the molecular determinants for virulence at the level of gene expression have remained opaque. Here we characterise the transcriptome of ASFV genotype II Georgia 2007/1 (GRG) during infection of the physiologically relevant host cells, porcine macrophages. In this study we applied Cap Analysis Gene Expression sequencing (CAGE-seq) to map the 5' ends of viral mRNAs at 5 and 16 hpi. A bioinformatics analysis of the sequence context surrounding the transcription start sites (TSSs) enabled us to characterise the global early and late promoter landscape of GRG. We compared transcriptome maps of the GRG isolate and the lab-attenuated BA71V strain that highlighted GRG virulent-specific transcripts belonging to multigene families including two newly characterised MGF 100 genes I7L and I8L. Structural homology modelling suggest that I7L and I8L encode unorthodox SH2 domain proteins with the potential to interfere with the host's immune response. In parallel, we monitored transcriptome changes in the infected host cells, which showed a pro-inflammatory immune response with the upregulation of NF-kB activated genes, innate immunity, as well as lysosome components including S100 proteins.



## Author Summary [196 words]

African swine fever virus (ASFV) causes a haemorrhagic fever in domestic pigs and wild boar with mortality rates approaching 100%, for which there are no approved vaccines or antivirals. The highly-virulent ASFV Georgia 2007/1 strain was the first isolated when ASFV spread from Africa to the Caucasus region in 2007. From here it has spread through Eastern Europe, and more recently across Asia. We have used an RNA-based next generation sequencing technique called CAGE-seq to map the starts of viral genes across the ASFV Georgia 2007/1 strain DNA genome. This has allowed us to investigate how it controls its viral gene expression during different stages of infection in macrophage cells. We have characterised which genes are expressed at different levels during early or late stages of infection, and compared them to the non-virulent ASFV-BA71V strain to identify key genes that enhance virulence. We have discovered new genes, and predicted the likely roles of uncharacterised genes during ASFV infection. In parallel we have investigated how the host cells respond to ASFV infection, which has revealed how the virus early on

suppresses components of the host immune response to ultimately win the arms race against its porcine host.

## Introduction [1,192 words]

ASFV originated in Sub-Saharan Africa where it remains endemic. However, following the introduction in 2007 of a genotype II isolate to Georgia (1) and subsequent spread in Russia and Europe. The virus was then introduced to China in 2018 (2), from here it spread rapidly across Asia, strongly emphasizing this disease as a severe threat to global food security. ASFV is the only characterised member of the Asfarviridae family (3) in the recently classified Nucleocytoviricota (ICTV Master Species List 2019.v1) phylum (4,5). ASFV has a linear double-stranded DNA (dsDNA) genome of ~170–193 kbp encoding ~150–~200 open reading frames (ORFs). Little is currently known about either the transcripts expressed from the ASFV genome or the mechanisms of ASFV transcription. Much of what is known about transcription is extrapolated from vaccinia virus (VACV), a distantly-related Nucleocytoviricota member, from the Poxviridae family (6). ASFV encodes a eukaryotic-like 8-subunit RNA polymerase (RNAP), an mRNA capping enzyme and poly-A polymerase, all of which are carried within mature virus particles. These virions are transcription competent upon solubilisation in vitro (7,8) and support mRNA modification by including a 5'-methylated cap and a 3' poly-adenylated (polyA) tail of ~33 nucleotide-length (8,9).

Viral genes are typically classified according to their temporal expression patterns. ASFV genes have historically been categorised as 'immediate early' when expressed immediately following infection, as 'early genes' following viral protein synthesis, as 'intermediate genes' after viral DNA replication, or as 'late genes' thereafter. The temporal regulation of transcription is likely enabled by different sets of general transcription initiation factors that recognise distinct early (EPM) or late (LPM) promoter elements, as we previously investigated in the ASFV-BA71V strain (10), and address further in this study. EPM recognition is likely enabled by the ASFV homologue of heterodimeric VACV early transcription factor (VETF), consisting of D1133L (D6) and G1340L (A7) gene products (11,12). Both are late genes, i. e. synthesised late during infection and packaged into virus particles. The ASFV LPM is less well defined than the EPM, but a possible initiation factor involved in its recognition is the viral homolog of eukaryotic-like ASFV-encoded TATA-binding protein (TBP), expressed during early infection. By analogy with the VACV system, additional factors including homologs of A1, A2 and G8 may also contribute to late transcription (6)

We have recently carried out a detailed and comprehensive ASFV whole genome expression analysis using complimentary next-generation sequencing (NGS) results and computational approaches to characterise the ASFV transcriptome following BA71V infection of Vero cells at 5 hpi and 16 hpi post-infection (hpi) (10). Most of our knowledge about the molecular biology of ASFV, including gene expression, has been derived from attenuated virus strains, such as BA71V infecting Vero tissue culture cells (9,10). These model systems provide convenient models to study the replication cycle but have deletions of many genes that are not essential for replication, but have important roles in virulence within its natural porcine hosts. (13–15). To date 24 ASFV genotypes have been identified in Africa (15–22), while all strains spreading across Asia and Europe belong to the Type II genotype. Most of these are highly virulent in domestic pigs and wild boar, including the ASFV Georgia 2007/1 (GRG) (23), and the Chinese ASFV Heilongjiang, 2018 (Pig/HLJ/18) (24) isolates. Though, a number of less virulent isolates have been identified in wild boar in the Baltic States and domestic pigs in China (25–28). It is crucial to understand the similarities and commonalities between ASFV strains, and to characterise the host response to these in order to understand the molecular determinants for ASFV pathogenicity. Information about the gene content and genome organisation can be gained from comparing virus genome sequences. However, only functional genomics such as transcriptome analyses can provide information about the differences in gene expression programmes and the host responses to infection.

On the genome level, most differences between virulent (e. g. GRG) and attenuated (e.g. lab-attenuated BA71V) ASFV strains reside towards the genome termini. Figure 1a shows a whole genome comparison of GRG (left) and BA71V (right) strains with the sequence conservation colour coded in different shades of blue. The regions towards the end of the chromosome are more dynamic compared to the central region of the chromosome that is highly conserved, as genes at the termini are prone to deletion, duplication, insertion and fusion (16,29). Most of the GRG-specific genes are expressed early during infection (early genes are colour coded blue in the outer arch of Figure 1a) and belong to Multi-Gene Families (MGFs, purple in the inner arch). The functions of many MGF members remain poorly understood, though variation among MGFs is linked to virulence (30). MGF 110 is not thought necessary for virulence in pigs due to few members being present in virulent isolate genomes (16), but are highly expressed both on the mRNA (10) and protein level (31), suggesting MGF 110 holds importance during infection. Overall, the functions of MGF 360 and 505 members are better characterised, playing a role in evading the host type I interferon (IFN) response (14,32–36). In summary, comparing the expression of ASFV genes, especially

MGFs between the virulent GRG- and the lab adapted BA71V strains, is fundamental in identification of virulence factors and better MGF characterisation.

Macrophages are the primary target cells for ASFV, they are important immune effector cells that display remarkable plasticity allowing efficient response to environmental signals (37). They can activate specific functional programs, which can be divided into two main groups: classically activated macrophages (M1) present during acute infections and alternatively activated macrophages (M2) (38). Infection of macrophages with virulent ASFV has been shown to inhibit the expression of IFN, cytokines, chemokines, adhesion molecules and other immunomodulatory genes, thereby interfering with M1 macrophage function (30). Our understanding of why virulent ASFV isolates like GRG can infect macrophages and win the battle with the host immune system, while attenuated strains like BA71V struggle to do so, currently relies on genomic comparison. Little is known about how host macrophages respond to infection apart from a microarray study of primary swine macrophage cells infected with GRG (39), and an RNA-seq study of whole blood isolated from pigs infected either with a low pathogenic ASFV isolate, OURT 88/3, or the highly pathogenic GRG (40).

Here we applied CAGE-seq to characterise the transcriptome of the highly virulent GRG isolate (23), in primary porcine macrophages, the biologically relevant target cells for ASFV infection. We have investigated the differential gene expression patterns of viral mRNAs at 5- and 16 hpi and mapped their promoter motifs. Importantly, we have compared the expression levels and temporal regulation of genes conserved in both strains. With a few exceptions, both mRNA expression levels and temporal regulation of the conserved genes are surprisingly similar between BA71V and GRG. This confirms that it is not deregulation of their conserved genes, but the virulent isolate-specific genes, which are the key determinants for ASFV virulence. Most of these genes are MGF members, likely involved in suppression of the host immune-response. The transcriptome analysis of the porcine macrophages upon GRG infection reflects a pro-inflammatory immune response with the upregulation of in particular NF-kB activated genes, but also innate immunity related and lysosome components.

# Results [4,801 words]

## Genome-wide Transcription Start Site-Mapping

We infected primary porcine alveolar macrophages with ASFV GRG at a high multiplicity of infection (MOI 5.0), isolated total RNA at 5 hpi and 16 hpi and sequenced using CAGE-seq (Supplementary Table 1a). The

resulting mRNA 5' ends were mapped to the GRG genome (Figure 1b) resulting in the annotation of 229 and 786 TSSs at 5 and 16 hpi, respectively (Figure 1c and d, from Supplementary Table 1b and c, respectively). The majority of TSSs were identified within 500 bp upstream of the start codon of a given ORF, a probable location for a *bona fide* gene TSS. The strongest and closest TSSs upstream of ORFs were annotated as 'primary' TSS (pTSS, listed in Supplementary Table 1d) and in this manner we could account TSS for 177 out of 189 GRG ORFs annotated in the FR682468.1 genome. TSSs signals below the threshold for detection included MGF_110-11L, C62L, and E66L, the remainder being short ORFs designated as 'ASFV_G_ACD', predicted solely from the FR682468 genome sequence (23). The E66L ORF was originally predicted from only the BA71V genome sequence, but likewise undetectable with CAGE-seq (10), making its expression unlikely. Our TSS mapping identified novel ORFs (nORFs) downstream of the TSS, which were included in the curated GRG genome map (Supplementary Table 1d includes pTSSs of annotated ORFs and nORFs in gene feature file or 'GFF' format). In addition to ORF-associated TSSs, some were located within ORFs (intra-ORF or ioTSS), or in between them (inter-ORF TSS), and all detected TSSs are listed in Supplementary Table 1b-c.

## Expression of GRG genes during Early and Late Infection

Having annotated TSSs across the GRG genome, we quantified the viral mRNAs originating from pTSSs from CAGE-seq data, normalising against the total number of reads mapping to the ASFV genome (i. e. RPM or reads per million mapped reads per sample). We compared gene expression between early and late infection, and simplistically defined genes as 'early' or 'late' if they are significantly down- or upregulated (respectively), using DESeq2 (41). In summary, 165 of the 177 detectable genes were differentially expressed (adjusted p-value or padj < 0.05, Supplementary Table 1e). Those showing no significant change were D345L, DP79L, I8L, MGF_100-1R, A859L, QP383R, B475L, E301R, DP63R, C147L, and I177L. 87 of those 165 differentially expressed genes were significantly downregulated, thus representing the 'early genes', while 78 of the 165 genes were upregulated or 'late genes'. The majority of MGFs were early genes, apart from MGF 505-2R, MGF360-2L and MGF 100-1L (Figure 2a). Figure 2b shows the expression patterns of GRG-exclusively expressed genes, which we defined as only having a detectable CAGE-seq TSS in GRG, and not in BA71V (regardless of presence in the BA71V genome). These unsurprisingly, consist of many MGFs (18), all of which were early genes (Figure 2b), barring MGF 100-1L.

We extracted the top twenty most highly expressed genes of GRG (as RPM) during 5 hpi (Figure 2c) and 16 hpi (Figure 2d) post-infection. Ten genes are shared between both top 20 lists: MGF 110-3L, A151R,

MGF 110-7L, MGF 110-5L-6L, I73R, 285L, CP312R, ASFV_G_ACD_00600, MGF 110-4L, and CP204L. It is important to note that the relative expression values (RPM) for genes at 5 hpi are significantly higher than those at 16 hpi. This is consistent with our observations in the BA71V strain (10) and due to the increase in global transcript levels during late infection discussed below. Supplementary Table 1f includes all the GRG annotated ORFs, their TSS locations during early and late infection, their relative distances if these TSS locations differ, and their respective Untranslated Region (UTR) lengths.

## GRG and BA71V Share Strong Similarity between Conserved Gene Expression

Next we carried out a direct comparison of mRNA levels from the 132 conserved genes between the virulent GRG and attenuated BA71V (10) strain making use of our previously published CAGE-seq data. The relative transcript levels (RPM) of the genes conserved between the two strains showed a significant correlation at 5 hpi (Figure 3a) and 16 hpi (Figure 3b), supported by the heatmap in Supplementary Figure 1, the RPM for each gene, across both time-points and replicates, showing a strong congruence between the two strains. Of the 132 conserved genes, 125 showed significant differential expression in both strains. 119 of these 125 showed the same down- or up-regulated patterns of significant differential expression from 5 hpi to 16 hpi (Figure 3c, early genes in blue, late genes in red). The exceptions are D205R, CP80R, C315R, NP419L, F165R, and DP148R (MGF 360-18R), encoding RNA polymerase subunits RPB5 and RPB10 (14), Transcription Factor IIB (TFIIB) (14), DNA ligase (42), a putative signal peptide-containing protein, and a virulence factor (43), respectively. The ASFV-TFIIB homolog (C315R) is classified as an early gene in GRG but not in BA71V, in line with the predominantly early-expressed TBP (B263R), its predicted interaction partner. It is worth noting however, that D205R, CP80R, and C315R are close to the threshold of significance, with transcripts being detected at both 5 hpi and 16 hpi (Supplementary Table 1e).

## Increased and pervasive transcription during late infection

During late infection of BA71V (10), we noted an increase in genome-wide mRNA abundance, as well as an increasing number of TSSs and transcription termination sites, reminiscent of pervasive transcription observed during late Vaccinia virus (44). To quantify and compare the global mRNA increase both in BA71V and GRG, we calculated the ratio of read coverage at 16 hpi versus 5 hpi (log2 transformed ratio of RPM) per nucleotide across the viral genome (Figure 4a, increase shown above- and decrease below the x-axis). This dramatic increase is due to the overall increase of virus mRNAs present, which is visible in both strains (Figure 4b), with a ~2 fold increase in GRG from 5 hpi to 16 hpi, versus ~8 fold in BA71V (Figure 4c).

188 This observation can at least in part be attributed to the larger number of viral genomes during late
189 infection, with increased levels of viral RNAP and associated factors available for transcription, following
190 viral protein synthesis. Viral DNA-binding proteins, such as histone-like A104R (45), may remain associated
191 with the genome originating from the virus particle in early infection. This could suppress spurious
192 transcription initiation, compared to freshly replicated nascent genomes that are highly abundant in late
193 infection. In order to test whether the increased mRNA levels correlated with the increased number of
194 viral genomes in the cell, we determined the viral genome copy number by using quantitative PCR (qPCR
195 against the p72 capsid gene sequence) using purified total DNA from infected cells isolated at 0 hpi, 5 hpi
196 and 16 hpi, and normalized values to the total amount of input DNA. Using this approach, we observed
197 genome copy levels that were consistent from 0 hpi to 5 hpi, followed by a substantial increase at 16 hpi,
198 which was more pronounced in BA71V infection (Figure 4d). This corresponded to a 15-fold increase in
199 GRG genome copy numbers from late, compared to early times post-infection of porcine macrophages,
200 and a 30-fold increase in BA71V during infection of Vero cells (Figure 4e). In summary, the ASFV
201 transcriptome changes both qualitatively and quantitatively as infection progresses, and the increase of
202 virus mRNAs during late infection is accompanied by the dramatic increase in viral genome copies.
203 Interestingly, the increase in viral transcripts and genome copies was less dramatic in the virulent GRG
204 strain.

205 ## Correcting the bias of temporal expression pattern

206 The standard methods of defining differential gene expression are well established in transcriptomics
207 using programs like DESeq2 (41). This is a very convenient and powerful tool which captures the nuances
208 of differential expression in complex organisms. However, virus transcription is often characterised by
209 more extreme changes, typically ranging from zero to millions of reads. Furthermore, in both BA71V and
210 GRG strains the genome-wide mRNA levels and total ASFV reads increase over the infection time course
211 (Figure 4 and Supplementary Table 1a). As a consequence, such normalisation against the total mapped
212 transcripts per sample (RPM) generates overestimated relative expression values at 5 hpi, and
213 understates those at 16 hpi (10). In order to validate the early-late expression patterns derived from CAGE-
214 seq, we carried out RT-PCR for selected viral genes, as this signal is proportionate to the number of specific
215 mRNAs regardless of the level of other transcripts – with the minor caveat that it can pick up readthrough
216 transcripts from upstream genes. We tested differentially expressed conserved genes including GRG early-
217 (MGF505-7R, MGF505-9R, NP419L), and D345L which showed stable relative expression values (RPM
218 values in Figure 1e). All selected genes showed a consistently stronger RT-PCR signal during late infection

in both BA71V and GRG (Figure 5a-d). The exception is NP419L whose levels were largely unchanged, and this is an example of how a gene whose transcript levels remain constant would be considered downregulated, when almost all other mRNA levels increase (Figure 5b).

The standard normalisation of NGS reads against total mapped reads (RPM) is regularly used as it enables a statistical comparison between samples and conditions, subject to experimental variations (46). Keeping this in mind, we used an additional method of analysing the 'raw' read counts to represent global ASFV transcript levels that are not skewed by the normalisation against total mapped reads. Figure 5 shows a side-by-side comparison of RT-PCR results, and the CAGE-seq data normalised (RPM) or expressed as raw counts, beneath each RT-PCR gel. Unlike CAGE-seq, RT-PCR will detect transcripts originating from read-through of transcripts initiated from upstream TSS including intra-ORF TSS (ioTSSs). To detect such 'contamination' we used multiple primer combinations in upstream and downstream segments of the gene (Figure 5c, cyan and yellow arrows) to capture and account for possible variations. Overall, our comparative analyses shows that the normalised data (RPM) of early genes such as MGF505-7R and 9R indeed skews and overemphasises their early expression, while the raw counts are in better agreement with the mRNA levels detected by RT-PCR. In contrast, late genes such as NP419L and D345L would be categorised as late using all three quantification methods, in agreement with GRG CAGE-seq but not BA71V from Figure 3c. We validated the expression pattern of the early GRG-specific gene MGF360-12L (Figure 5e). While the RPM values indicated a very strong decrease in mRNA levels from early to late time points, the decrease in raw counts was less pronounced and more congruent with the RT-PCR analysis, showing a specific signal with nearly equal intensity during early and late infection. Lastly, we used qRT-PCR to quantify C315R transcript levels, as this was close to the early vs late threshold, (a log2fold change of 0 in Figure 3c), which showed again that qRT-PCR better agreed with the raw counts.

## An improved temporal classification of ASFV genes

Based on the considerations above, we prepared a revised classification of temporal gene expression of the genes conserved between the two strains based on raw counts. The heatmap in Figure 6a shows the mRNA levels at early and late infection stages of BA71V and GRG strains (all in duplicates) with the genes clustered into five subcategories (1 to 5, Figure 6a) according to their early and late expression pattern, which are shown in Figure 6b. Genes that are expressed at high or intermediate levels during early infection but that also show high or intermediate mRNA levels during late infection are classified as 'early' genes belonging to cluster-1 (8 genes, levels: high to high, H-H), cluster-4 (33 genes, mid to mid, M-M)

and cluster-5 (16 genes, low-mid to low-mid, LM-LM). Genes with low or undetectable mRNA levels during early infection, which increase to intermediate or high levels during late infection are classified as 'late' genes and belong to cluster-2 (15 genes, low to high, L-H) and cluster-3 (60 genes, low to mid, L-M), respectively. Overall, the clustered heatmap based on raw counts shows a similar but more emphasised pattern compared to the normalised (RPM) data (compare Figure 6 and Supplementary Figure 1). Calculating the percentage of reads per gene, which can be detected at 16 hpi compared to 5 hpi, reveals only a small number of genes have most ( ≥70%) of their reads originating during early infection: 30 genes in the GRG strain and 5 genes in the BA71V strain. For over half of the BA71V-GRG conserved genes, 90-100 % of reads can be detected during late infection (Figure 6c). For all GRG genes, this generates a significant difference between the raw counts per gene between time-points (Figure 6d).

Below we discuss specific examples of genes subcategorised in specific clusters. I73R is among the top twenty most-expressed genes during both early and late infection according to the normalised RPM values (Figure 2c and d) resides in cluster-1 (H-H) (Figure 6a). While I73R is expressed during early infection, the mRNA levels remain high with >1/3 of all reads detected during late infection in both strains when calculated as raw counts (34 % in GRG and 45 % in BA71V). This new analysis firmly locates I73R into cluster-1 (H-H) and is classified confidently as early gene. Notably, our new approach results in biologically meaningful subcategories of genes that are likely to be coregulated, e. g. the eight key genes that encode the ASFV transcription system including RNAP subunits RPB1 (NP1450L), RPB2 (EP1242L), RPB3 (H359L), RPB5 (D205R), RPB7 (D339L) and RPB10 (CP80R), the transcription initiation factor TBP (B263R) and the capping enzyme (NP868R) belong to cluster-4 (M-M), and transcription factors TFIIS (I243L) and TFIIB (C315R) belong to cluster-5 (LM-LM). The overall mRNA levels of cluster-4 and -5 genes are different, but remain largely unchanged during early and late infection, consistent with the transcription machinery being required throughout infection. In contrast, the mRNAs encoding the transcription initiation factors D6 (D1133L) and A7 (G1340L) are only present at low levels during early- but increase during late infection and thus belong to cluster-3 (L-M), classifying them as late genes. This is meaningful since the heterodimeric D6-A7 factor is packaged into viral particles (7), presumably during the late stage of the infection cycle. The mRNAs of the major capsid protein p72 (B646L) and the histone-like-protein A104R (45,47) follow a similar late pattern but are present at even higher levels during late infection and therefore belong to cluster-2 (L-H).

## Architecture of ASFV promoter motifs

In order to characterise early promoter motifs (EPM) in the GRG strain, we extracted sequences 35 bp upstream of all early gene TSSs and carried out multiple sequence alignments. As expected, this region shows a conserved sequence signature in good agreement with our bioinformatics analyses of EPMs in the BA71V strain, including the correct distance between the EPM and the TSS (9-10 nt from the EPM 3' end) and the 'TA' motif characteristic of the early gene Initiator (Inr) element (Figure 7a) (10). A motif search using MEME (48) identified a core (c)EPM motif with the sequence 5'-AAAATTGAAT-3' (Figure 7b), within the longer EPM. The cEPM is highly conserved and is present in almost all promoters controlling genes belonging to cluster-1, -4 and -5 (Supplementary Table 3). A MEME analysis of sequences 35 bp upstream of late genes (Figure 7c), provided a 17-bp AT-rich core late promoter motif (cLPM, Figure 7d), however, this could only be detected in 46 of the late promoters.

In an attempt to improve the promoter motif analyses and deconvolute putative sequence elements further, we probed the promoter sequence context of the five clusters (clusters 1-5 in Figure 7e-i, respectively) of temporally expressed genes with MEME (Supplementary Table 3). The early gene promoters of clusters-1 (H-H), -4 (M-M) and -5 (LM-LM) are each associated with different expression levels, and all of them contain the cEPM located 15-16 nt upstream of the TSS with two exceptions that are characterized by relatively low mRNA levels (Figure 7k). Interestingly, cluster-2 (L-H) promoters are characterized by a conserved motif with significant similarity to eukaryotic TATA-box promoter element that binds the TBP-containing TFIID transcription initiation factor (Figure 7f highlighted with red bracket, detected via Tomtom (49) analysis of the MEME motif output). Cluster-3 (L-M) promoters contain a long motif akin to the cLPM, derived from searching all late gene promoter sequences, and which is similar to the LPM identified in BA71V (Figure 7d and g, green bracket). All motifs described in the cluster analysis above could be detected with statistically significance (p-value < 0.05) via MEME, in every gene in each respective cluster with only two exceptions: MGF 110−3L from cluster-1, and MGF 360-19R from cluster-4, for the latter see details below.

## Updating Genome Annotations using Transcriptomics Data

TSS-annotation provides a useful tool for re-annotating predicted ORFs in genomes like ASFV (10) where many of the gene products have not been fully characterized and usually rely on prediction from genome sequence alone. We have provided the updated ORF map of the GRG genome in GFF format (Supplementary Table 1f). This analysis identified an MGF 360-19R ortholog (Figure 8), demonstrating how

transcriptomics enhances automated annotation of ASFV genomes by predicting ORFs from TSSs. The MGF 360-19R was included in subsequent DESeq2 analysis showing it was not highly nor significantly differentially expressed (Supplementary Table 1e). Another important feature is the identification of intra-ORF TSSs (ioTSSs) within MGF 360-19R that potentially direct the synthesis of N-terminally truncated protein variants expressed either during early or late infection. The presence of EPM and LPM promoter motifs lends further credence to the ioTSSs (Figure 8). Similar truncation variants were previously reported for I243L and I226R (50) and in BA71V (10). In addition, we detected multiple TSSs within MGF 360-19R encoding very short putative novel ORFs (nORF) 5, 7 or 12 aa residues long; since these ioTSSs were present in both early and late infection they are not all likely to be due to pervasive transcription during late infection.

We investigated the occurrence of ioTSS genome wide and uncovered many TSSs with ORFs downstream that were not annotated in the GRG genome (Supplementary Table 2a). These ORFs could be divided into sub-categories: in-frame truncation variants (Supplementary Table 2b, akin to MGF 360-19R in Figure 8), nORFs (Supplementary Table 2c), and simply mis-annotated ORFs. All updated annotations are found in Supplementary Table 1f. Putative truncation variants generated from ioTSSs were predominantly identified during late infection, suggesting these could be a by-product of pervasive transcription. Therefore, those detected early or throughout infection are perhaps more interesting, they span a variety of protein functional groups, and many gene-products are entirely uncharacterised (Figure 9a). The truncation variants additionally showed a size variation of 5'-UTRs between the ioTSSs and downstream start codon (Figure 9b). An example of a mis-annotation would be CP204L (Phosphoprotein p30, Figure 9c) gene that is predicted to be 201 residues long. The TSS determined by CAGE-seq and validated by Rapid Amplification of cDNA Ends (5'-RACE) is located downstream of the annotated start codon; based on our results we reannotated the start codon of CP204L which results in a shorter ORF of 193 amino acids (Figure 9c).

Our GRG TSS map led to the discovery of many short nORFs, which are often overlooked in automated ORF annotations due to a minimum size, e. g. 60 residues in the original BA71V annotation (14). Some short ORFs have been predicted for the GRG genome including those labeled 'ASFV_G_ACD' in the Georgia 2007/1 genome annotation (18). However, their expression was not initially supported by experimental evidence, though we have now demonstrated their expression via CAGE-seq (Figure 2b, Supplementary Table 1e). We have now identified TSSs for most of these short ORFs, indicating at minimum they are

transcribed. As described above, we noted that TSSs were found throughout the genome in intergenic regions in addition to those identified upstream of the 190 annotated GRG ORFs (including MGF 360-19R, Supplementary Table 2c). Our systematic, genome-wide approach identified 175 novel putative short ORFs. BLASTP (51) alignments showed that 13 were homologous to ORFs predicted in other strains, including DP146L and pNG4 from BA71V . We validated the TSSs for these candidates using 5'-RACE, which demonstrates the presence of these mRNAs and their associated TSSs at both time-points (Figure 9f and g, respectively), compared to our CAGE-seq data (Figure 9f and g, respectively).

## Identification of Functional Domains in Uncharacterised Genes

Our understanding of the ASFV genome is hampered by the large number of genes with unknown functions. We attempted to remedy this limitation by systematically identifying conserved domains of 47 MGF members and 46 uncharacterised ASFV-GRG genes. These candidates included ten genes that are among the highest-expressed genes and ten genes whose protein products are present in viral particles (7). The MGF 100 genes form the smallest multigene family include three short (100–150 aa) paralogs located at both genome ends (right, R and left, L): 1R, 2L (DP141L in BA71V), and 3L (DP146L in BA71V) (29). We predicted the two highly similar GRG ORFs called I7L and I8L (51% sequence identity) to be additional members of the MGF 100 family (Figure 10a). The conservation of I7L and I8L MGF 100 members in more virulent and specifically porcine-infecting ASFV strains, combined with their early expression in GRG from our data, and their deletion reducing virulence in swine (52), suggests I7L and I8L play an important role during early infection of porcine macrophages hosts . Through extensive HHpred searches (53), we found that they all include a Src Homology 2 (SH2) domain, with detectable similarity to the N-terminal domain of *C. elegans* Cell death abnormality protein 2, ced-2 (54) (11% seq. id. between I7L and ced-2, E = 0.002 (Figure 10b). SH2 domains, including ced-2, are important protein-protein interaction domains that interact with phospho-tyrosine (pTyr) residues as a part of larger signaling cascades (55). We subsequently generated computational homology models of I7L and I8L, and interestingly both have lost the invariant arginine ßB5 from the canonical pTyr binding pocket, suggesting either a different function or a different mode for recognition of phosphorylated amino acid residues (Figure 10c). In that context it is noteworthy that ASFV genomes encode a putative kinase R298L, which is homologous to vaccinia virus B1R gene (56). Both I7L and I8L show similar overall expression levels to the MGF 100 members. 1L and 1R are already annotated in the GRG genome (Supplementary Figure 1e), I7L and I8L are both early genes, while MGF 100-1L and 1R are late and not significantly changing, respectively.

We characterised another gene product, C717R, which is expressed at relatively low levels (< 200 RPM), but upregulated from early to late infection (Supplementary Figure 1e). C717R includes a central domain that is similar to serine/threonine kinases (18% amino acid seq. id. to Vps34, pdb id: 5DFZ over 221 aa, E = 8.4e-15 (Supplementary Figure 3). Besides Asfarviridae, it is conserved also in various Kaumoebavirus and Faustovirus strains. Computational homology modelling shows that, similar to R298L, the canonical ATP-binding loop ('GxGxxG') has been lost, while the catalytic loop ('HRD' motif) is conserved in a slightly modified form (as 'HAD', Supplementary Figure 3). Also, activation segment residues (typically 'DFG') are not conserved in C717R homologs ('DRN' in ASFV).

Finally, we investigated the K421R gene product. ASFV transcripts are thought to be polyadenylated by a polyA polymerase (C475L), which is highly conserved with other poxvirus and mimivirus viral polyA polymerases or vPAPs (57). Surprisingly, we discovered that Asfarviridae and related genomes (Abalone asfarvirus, Pacmanvirus, and Faustovirus) encode another vPAP member, represented by K421R in ASFV GRG (12% seq. id. to vPAP, pdb id: 4P37, E = 2.5e-60, (Supplementary Figure 4a). Computational homology modelling of K421R paralogues suggest that they contain the N-terminal dimerisation helix that is characteristic for homodimeric self-processive vPAPs (58) but have lost two out of three acidic active site residues (Supplementary Figure 4b). Consequently, they are predicted to be unable to coordinate metal ions necessary for polyA addition and carry out an unusual unknown function. K421R is a late gene, as the transcript is not detected at 5 hpi and increases to ~550 RPM at 16 hpi, while C475L (polyA polymerase) is also expressed late, but at ~3 times the level of K421R (Supplementary Table 1e). Even more intriguing is the fact that both vPAPs are packaged in viral particles (7), suggesting both may be required for early gene expression.

## The response of the porcine macrophage transcriptome to ASFV infection

In order to evaluate the impact of ASFV on the gene expression of the host cell, we analysed transcriptomic changes of infected porcine macrophages using the CAGE-seq data from 0 (uninfected cells), at 5 and 16 hpi. We annotated 9,384 macrophage-expressed protein-coding genes with CAGE-defined TSSs (Supplementary Table 4). Although primary macrophages are known to vary largely in their transcription profile, the CAGE-seq reads were highly similar between samples (Spearman's correlation coefficients ≥ 0.77).

As TSSs are not well annotated for the swine genome, we annotated them *de novo* using our CAGE-seq data with the RECLU pipeline. 37,159 peaks could be identified, out of which around half (18,575) matched

399 unique CAGE-derived peaks annotated in Robert et al. (59) i.e. they were located closer than 100 nt to the
400 previously described peaks. Mapping CAGE-seq peaks to annotated swine protein-coding genes led to
401 identification of TSSs for 9,384 macrophage-expressed protein-coding genes (Supplementary Table 4). The
402 remaining 11,904 swine protein-coding genes did not have assigned TSSs, and therefore their expression
403 levels were not assessed. The majority of genes were assigned with multiple TSSs, and these TSS-assigned
404 genes, were many critical functional macrophage markers, including genes encoding 56 cytokines and
405 chemokines (including CXCL2, PPBP, CXCL8 and CXCL5 as the most highly expressed), ten S100 calcium
406 binding proteins (S100A12, S100A8, and S100A9 in the top expressed genes), as well as interferon and
407 TNF receptors (IFNGR1, IFNGR2, IFNAR1, IFNAR2, IFNLR1, TNFRSF10B, TNFRSF1B, TNFRSF1A, etc.), and
408 typical M1/M2 marker genes such as TNF, ARG1, CCL24, and NOS2 (Supplementary Table 5). The mRNA
409 levels of genes of interest were verified using RT-PCR (Figure 11f).

410 The 9,384 genes with annotated promoters were subjected to differential expression analysis using
411 DESeq2 to compare the three time points (0, 5 and 16) in a pairwise manner. Expression of only 25 host
412 genes was significantly deregulated between 0 and 5 hpi, compared to 652 genes between 5 hpi and 16
413 hpi, and 1325 genes between 0 and 16 hpi (at FDR of 0.05). This implies that the major host transcriptomic
414 response to ASFV does not occur during the early phase of infection. Based on the pairwise comparisons,
415 we could distinguish major response profiles of the host genes. Late response genes, whose expression
416 was significantly deregulated both between 0 and 16 hpi and 5 and 16 hpi, and early response genes,
417 whose expression was significantly deregulated both between 0 and 5 hpi, but not later (Figure 11a). The
418 latter category included only 20 genes, whereas more than 500 genes showed the late differentially
419 regulated response: 344 genes were up-regulated, and 180 genes were down-regulated. Comparison of
420 differences between expression levels at the three time points indicate that macrophage differentially
421 expressed transcription programs start mostly after 5 hpi (Figure 11b and c) i. e. if a gene's expression
422 changes, it has usually happens between 5 hpi and 16 hpi. The upregulated late response genes with
423 highest expression levels included several S100 calcium binding proteins. In contrast, initially high
424 expression of important cytokines (including CCL24, CXCL2, CXCL5 and CXCL8) significantly decreased from
425 5 hpi to 16 hpi (Figure 11d).

426 To investigate the transcriptional response pathways and shed light on possible transcription factors
427 involved in the macrophage response to ASFV infection, we searched for DNA motifs enriched in
428 promoters of the four categories of deregulated genes in Figure 11a. Both late response promoter sets
429 were significantly enriched with motifs, some of which contained sub-motifs known to be recognised by

human transcription factors (Supplementary Figure 2). The highest-scored motif found in promoters of upregulated genes contained a sub-motif recognised by a family of human interferon regulatory factors (IRF9, IRF8 and IRF8, Supplementary Figure 2a) that play essential roles in the anti-viral response. Interestingly, both upregulated and downregulated promoters (Supplementary Figure 2b and c, respectively) were enriched with extended RELA/p65 motifs. p65 is a Rel-like domain-containing subunit of the NF-kappa-B complex, regulated by I-kappa-B, whose analog is encoded by ASFV. This pathway being a known target for ASFV in controlling host transcription (60–63).

To understand functional changes in the macrophage transcriptome, we also performed gene set enrichment analysis using annotations of human homologs. The top enriched functional annotations in the upregulated late response genes include glycoproteins and disulfide bonds, transmembrane proteins, innate immunity, as well as positive regulation of inflammatory response (Figure 11e). In contrast, sterol metabolism, rRNA processing, cytokines, TNF signalling pathway, inflammatory response as well as innate immunity were the top enriched functional clusters among the downregulated late response genes. Interestingly, the genes associated with innate immunity appear overrepresented in both up- and downregulated gene subsets, yet cytokines are 8-fold enriched only in the downregulated genes.

## Discussion [2,366 words]

In order to shed light on the gene expression determinants for ASF virulence, we focussed our analyses on the similarities and differences in gene expression between a highly virulent Georgia 2007/1 isolate and a nonvirulent, lab-adapted strain (BA71V). Previous annotation identified 125 ASFV ORFs that are conserved between all ASFV strain genomes irrespective of their virulence (15). These represent a ‘core’ set of genes required for the virus to produce infectious progeny and include gene products like those involved in virus genome replication, virion assembly, RNA transcription and modification, these are located in the central region of the genome (Figure 1a). Besides these essential genes, about one third are non-essential genes for replication, but have roles in evading host defence pathways. Some genes are conserved between isolates, but not necessarily essential core genes, for example apoptosis inhibitors: Bcl-2 family member A179L and IAP family member A224L. Other non-essential genes, especially MGF members, vary in number between isolates. Our transcriptomics analysis captured 119 genes both shared between the BA71V and GRG genomes, and also match expression patterns during early and late infection, according to CAGE-seq (Figure 3, Figure 4a-c). Outliers include DP148R, which is obvious, given its promoter region is deleted in BA71V, and its coding region is interrupted by a frame shift mutation,

460 therefore functional protein expression unlikely. DP148R is a non-essential, early-expressed virulence
461 factor in the Benin 97/1 strain (43) – consistent with our GRG data. Many additional GRG genes, lost from
462 BA71V are MGFs, which are mostly upregulated during early infection and located at the ends of the linear
463 genome (Figure 1a). MGFs have evolved on the virus genome by gene duplication, and do not share
464 significant similarity to other proteins, though some conserved domains, including ankyrin repeats are
465 present in some MGF 360 and 505 family members (16,18).

466 We have predicted the function of two novel MGF 100 members to be SH2 domain-containing protein
467 that have lost the canonical pTyr-binding pocket. Although SH2 domains are primarily specific to
468 eukaryotes, rare cases of horizontally transferred SH2 domains, found in viruses, are implicated in
469 hijacking host cell pTyr signalling (64). A large family of 'super-binding' SH2 domains were discovered in
470 Legionella. Its members, including single SH2 domain-proteins are likely effector proteins during infection
471 (65). Interestingly, loss of MGF 100 members was observed during the process of adapting a virulent
472 Georgia strain to grow in cultured cell lines (66). We also identified a further MGF 100 member in the GRG
473 genome as one of our nORFs, a partial 100-residue copy of DP146L (MGF 100-3L) (Supplementary Table
474 2c). Unlike its annotated MGF 100-1L and MGF 100-1R cousins it was downregulated from 5 hpi to 16 hpi
475 (Supplementary Table 1e). Together with I7L and I8L GRG encodes a total of 5 MGF 100 genes.

476 The Georgia 2007/1 genome was recently re-sequenced which identified a small number of genome
477 changes affecting mapped ORFs and identified new ORFs (17). Adjacent to the covalently cross-linked
478 genome termini, the BA71V genome contains terminal inverted repeats of >2 kbp, in which two short
479 ORFs were identified (DP93R, DP86L). These were not included in previous GRG sequence annotations,
480 however our nORFs included a 55-residue homolog of DP96R, which was a late, but not highly expressed
481 gene. These are yet further examples of how transcriptomics aid in improving ASFV genome annotation.
482 Functional data is available for only a few of proteins coded by ORFs not conserved between BA71V and
483 GRG. This includes the p22 protein (KP177R), which is expressed on the cell membrane during early
484 infection, and also incorporated into the virus particle inner envelope. The function of the KP177R-like
485 GRG gene I10L has not been studied, but may provide an antigenically divergent variant of P22, enabling
486 evasion of the host immune response (18). We found KP177R was highly expressed at 16 hpi, while I10L
487 was also expressed late, but at much lower levels. Their function is unknown, though the presence of an
488 atypical SH2 domain indicates possible roles in signalling pathways (7,18,67).

MGF 110 members are among the highest expressed genes during early infection both in GRG (this study), and in BA71V (10), suggesting high importance during infection, at least in porcine macrophages and Vero cells, respectively. However, MGF 110 remains poorly characterised, and 13 orthologues were identified thus far, with numbers present varying between isolates (29). MGF 110 proteins possess cysteine-rich motifs, optimal for an oxidizing environment as found in the ER lumen or outside the cell, and MGF 110-4L (XP124L) contains a KDEL signal for retaining the protein in the endoplasmic reticulum (ER) (68). Since highly virulent isolates have few copies of these genes (for example, only 5 in the Benin 97/1 genome), it was assumed they are not importance for virulence in pigs (16), but their high expression warrant further investigation.

There is, good evidence that MGF 360 and 505 carry out important roles in evading the host type I interferon (IFN) response - the main host antiviral defence pathway (32). Evidence for the role of MGF 360 and 505 genes in virulence from deletions in tissue-culture adapted and field attenuated isolates as well as targeted gene deletions This correlated with induction of the type I interferon response, which itself is inhibited in macrophages infected with virulent ASFV isolates (33–35). Deletions of these MGF 360, and 505 genes also correlated with an increased sensitivity of ASFV replication, to pre-treatment of the macrophage cells with type I IFN (36). Thus, the MGF 360 and 505 genes have roles in inhibiting type I IFN induction and increasing sensitivity to type I IFN. However, it remains unknown if these MGF 360 and MGF 505 genes act synergistically or if some have a more important role than others type I IFN suppression. Our DESeq2 analysis did show that members of both these families showed very similar patterns of early expression (Figure 2 and Figure 3), conserved cEPM-containing promoters, and almost exclusive presence in clusters-1 (H-H), -4 (M-M), and -5 (LM-LM) (Figure 6 and Figure 7), consistent with ASFV prioritising inhibition of the host immune response during early infection.

An interesting pattern which emerged during our CAGE-seq analysis was the clear prevalence of ioTSSs within the ORFs, especially in MGFs (Figure 8 and Figure 9). However, it is not clear whether subsequent in-frame truncation variants generate stable proteins, nor what their function could be. Perhaps even more interesting was the discovery of 176 nORFs (including MGF 360-19R), with clear TSSs according to CAGE-seq, highlighting the power of transcriptomics to better annotate sequenced genomes. We were able to detect previously unannotated genes from other strains, and partial duplications of genes already encoded in GRG (Supplementary Table 2).

518 The increase in transcription across the ASFV genome during late infection (10), appears ubiquitous. At
519 least 50 genes have previously been investigated in single gene expression studies using Northern blot or
520 primer extension (for review see references (10,69). Transcripts from over two thirds of these genes were
521 detected during late infection, and a quarter had transcripts detected during both early and late infection.
522 Therefore, clear evidence using several techniques now support this increase in ASFV transcripts at late
523 times post-infection. It is not entirely clear whether it is due to pervasive transcription, high mRNA stability
524 or a combination of factors. However, there is a correlated increase in viral genome copies, potentially
525 available as templates for pervasive transcription. The increase in genome copies is more pronounced in
526 BA71V compared to GRG, which likewise is reflected in the increase in transcripts during late infection
527 (Figure 4).

528 Our transcriptomic analysis of the porcine macrophage host revealed 522 genes whose expression
529 patterns significantly changed between 5 and 16 hrs post-infection (Figure 11a) and only 20 genes were
530 found to change from 0 and 5 hpi. In aggregate, this reflects a relatively slow host response to ASFV
531 infection following expression of early ASFV genes. We observed mild downregulation of some genes e.g.
532 ACTB coding for ß-actin, eIF4A, and eIF4E (Supplementary Table 5), resembling patterns previously shown
533 by RT-qPCR (70). The macrophage transcriptome mainly shuts down immunomodulation between 5 hpi
534 to 16 hpi post-infection; cytokines appeared highly expressed at 5 hpi, but downregulated from 5 hpi to
535 16 hpi. Of the 54 cytokine genes we detected, expression of thirteen was decreased: four interleukin genes
536 (IL1A, IL1B, IL19, IL27), four pro-inflammatory chemokines (CCL24, CXCL2, CXCL5, CXCL8), and tumor
537 necrosis factor (TNF) genes. Since inflammatory responses serve as the first line of host defense against
538 viral infections, viruses have developed ways to neutralise host pro-inflammatory pathways. ASFV
539 encodes a structural analog of IκB, A238L, which was proposed to act as a molecular off-switch for NFκB-
540 targeted pro-inflammatory cytokines (62). In our study, A238L is one of the most expressed ASFV genes
541 at 5 hpi, but significantly downregulated afterwards (Figure 2c). Accordingly, swine homologs of human
542 NFκB target genes were significantly over-represented (3.8 fold) among downregulated macrophage
543 genes (Fisher's exact p-value < 1e-5, based on human NFκB target genes from https://www.bu.edu/nf-
544 kb/gene-resources/target-genes/). Downregulated genes include interleukins 1A, 1B, and 8, and 27 (IL1A,
545 IL1B, CXCL8, IL27), TNF, as well as a target for common nonsteroidal anti-inflammatory drugs,
546 prostaglandin-endoperoxide synthase 2 (PTGS2 or COX-2). Interestingly, promoters of both up- and
547 downregulated genes contained a motif with the sequence preferentially recognised by the human p65-
548 NFκB complex (71) (Supplementary Figure 2). Expression of TNF, a well-known marker gene for acute

549 immune reaction and M1 polarisation, was recorded at a high level at 0 and 5 hpi, but significantly
550 dropped at 16 hpi. It has been already shown that ASFV inhibits transcription of TNF and other
551 proinflammatory cytokines. (62). On the other hand, the downregulation of TNF stands in contrast to
552 previous results from ASFV-E75 strain-infected macrophages in vitro, where TNF expression increased
553 significantly after 6 hpi (72). Therefore, the different time courses of TNF expression induced by the
554 moderately virulent E75 and more virulent Georgia strain may reflect different macrophage activation
555 programs (38).

556 Four S100 family members are among the host genes that are upregulated after 5 hpi (Figure 11b)
557 including S100A8, S100A11, S100A12, and S100A13. S100A8 and S100A12 are among the most highly
558 expressed genes on average throughout infection. S100 proteins are calcium-binding cytosolic proteins
559 that are released and serve as a danger signal, and stimulate inflammation (73). Once released from the
560 cell, S100A12 and S100A8 function as endogenous agonists to bind TLR4 and induce apoptosis and
561 autophagy in various cell types (73). S100A8 and S100A9 were also found in the RNA-seq whole blood
562 study as the top upregulated upon infection of the pigs with Georgia 2007/1, but not of a low pathogenic
563 ASFV isolate OURT 88/3 (40).

564 Two previous studies described global swine transcriptome changes upon ASFV infection using short read
565 sequencing (Illumina): the RNA-seq described above (40), and a microarray study of primary swine
566 macrophage cell cultures infected with the GRG strain, at six post-infection time points (39). Although
567 these varied in designs and selected methods, results of these works both give some indication into the
568 main host immune responses and ways how ASFV could evade them. The latter microarray study indicated
569 similar suppression of inflammatory response after 16 hpi as we observed in this study, with expression
570 of many cytokines down-regulated relatively to non-infected macrophages (39). Neither study obtained
571 the read-depth and nucleotide resolution (obtainable with CAGE-seq), to investigate differential
572 expression of both the virus and host, the former being especially difficult in a compact genome where
573 transcription read-through can undermine results from classical RNA-sequencing techniques (10,74). A
574 recent investigation into ASFV RNA transcripts using long-read based Oxford Nanopore Technologies
575 (ONT) – provides fascinating insight into their length and read-through heterogeneity, but unfortunately
576 lacked the coverage for in-depth host transcript analysis alongside it (75).

577 Here we have demonstrated that CAGE-seq is an exceptionally powerful tool for quantifying relative
578 expression of viral genes across the ASFV genome, as well as making direct comparison between strains

for expression of shared genes, and further highlighting the importance of highly-expressed but still functionally uncharacterised viral genes. CAGE-seq conveniently circumvents the issue in compact viral genomes like those of ASFV and VACV, of transcripts reading through into downstream genes which cannot be distinguished from classical short-read RNA-seq (10,40,76). Furthermore, it enables us to effectively annotate genome-wide the 5' ends of capped viral transcripts, and thus TSSs of viral genes, and subsequently their temporal promoters. We have now expanded on promoter motifs we previously described (Figure 7), to identify 5 clusters of genes (Figure 6), with distinct patterns of expression. Three of these clusters (-1: high to high levels, -4: mid to mid, and -5 low-mid to low-mid) have slightly differing promoters, with a highly conserved core EPM. This is akin to the early gene promoter of VACV (74) for VETF recognition and early gene transcription initiation (77,78). We have found late genes can be categorised into two types that either increase from low to extremely high expression levels (e. g. p72-encoding B646L) in cluster-2, or from low to medium expression levels in cluster 3 (e. g VETF-encoding genes). The promoters of these genes show resemblance to the eukaryotic TATA-box (79) or the BA71V LPM (10), respectively. Our analysis additionally shows the potential for a variety of non-pTSSs: alternative ones used for different times in infection, ioTSSs which could generate in-frame truncation variants of ORFs, sense or antisense transcripts relative to annotated ORFs, and finally TSSs generating nORFs, which predominantly have no known homologs.

In summary, it is becoming increasingly clear that the transcriptomic landscape of ASFV during infection is far more complex than originally anticipated. Much of this raises further questions about the basal mechanisms underlying ASFV transcription and how it is regulated over the infection time course. Which subsets of initiation factors enable the RNAPs to recognise early and late promoters? Does ASFV include intermediate genes, and what factors enables their expression? What is the molecular basis of the pervasive transcription during late infection? The field of ASFV transcription has been understudied and underappreciated, and considering the severe threat that ASF poses for the global food system and -food security, we now need to step up and focus our attention and resources to study the fundamental biology of ASFV to develop effective antiviral drugs and vaccines.

# Methodology [2057 words]

## GRG-Infection of Macrophages and RNA-extraction

Primary porcine alveolar macrophage cells collected from two animals following approval by the local Animal Welfare and Ethical Review Board at The Pirbright Institute. Cells were seeded in 6-well plates ($2x10^6$ cells/well) with RPMI medium (with GlutaMAX), supplemented with 10% Pig serum and 100 IU/ml penicillin, 100 µg/ml streptomycin. They were infected as 2 replicate wells for 5 hpi or 16 hpi with a multiplicity of infection (MOI) of 5 of the ASFV Georgia 2007/1 strain, while uninfected cells were seeded in parallel as a control. Total RNA was extracted according to manufacturer's instructions for extraction with Trizol Lysis Reagent (Thermo Fisher Scientific and the subsequent RNAs were resuspended in 50µl RNase-free water and DNase-treated (Turbo DNAfree kit, Invitrogen). RNA quality was assessed via Bioanalyzer (Agilent 2100). 5 µg of each sample was ethanol precipitated before sending to CAGE-seq (Kabushiki Kaisha DNAFORM, Japan). Samples were named as follows: uninfected cells (C1-0 hpi and C2-0h), at 5 hpi post-infection (samples G1-5h and G2-5h), and at 16 hpi post-infection (G3-16h and G4-16h).

## CAGE-sequencing and Mapping to GRG and *Sus scrofa* Genomes

Library preparation and CAGE-sequencing of RNA samples was carried out by CAGE-seq (Kabushiki Kaisha DNAFORM, Japan). Library preparation produced single-end 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, washed 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. FastQC (80) analysis was carried out on all FASTQ files at Kabushiki Kaisha DNAFORM and CAGE-seq reads showed consistent read quality across their read-length, therefore, were mapped in their entirety to the GRG genome (FR682468.1) in our work using Bowtie2 (81), and *Sus scrofa* (GCF_000003025.6) genome with HISAT2 (81,82) by Kabushiki Kaisha DNAFORM.

## Transcription Start Site-mapping Across Viral GRG Genome

CAGE-seq mapped sample BAM files were converted to BigWig (BW) format with BEDtools (83) genomecov, to produce per-strand BW files of 5' read ends. Stranded BW files were input for TSS-prediction in RStudio (84) with Bioconductor (85) package CAGEfightR (86). Genomic feature locations were imported as a TxDb object from FR682468.1 genome gene feature file (GFF3). CAGEfightR was used to quantify the CAGE reads mapping at base pair resolution to the GRG genome - at CAGE TSSs, separately for the 5 hpi and 16 hpi replicates. TSS values were normalized by tags-per-million for each sample, pooled, and only TSSs supported by presence in both replicates were kept. TSSs were assigned to clusters, if within 25 bp of one another, filtering out pooled, RPM-normalized TSS counts below 25 bp for 5 hpi samples, or 50 bp for 16 hpi, and assigned a 'thick' value as the highest TSS peak within that cluster. A higher cut-off for 16 hpi was used to minimise the extra noise of pervasive transcription observed during late infection (10). TSS clusters were assigned to annotated FR682468.1 ORFs using BEDtools intersect, if its highest point ('thick' region) was located within 500 bp upstream of an ORF, 'CDS' if within the ORF, 'NA' if no annotated ORF was within these regions. Multiple TSSs located within 500 bp of ORFs were split into subsets: 'Primary' cluster subset contained either the highest scoring CAGEfightR cluster or the highest scoring manually-annotated peak (when manual ORF corrections necessary), and the highest peak coordinate was defined as the primary TSS (pTSS) for an ORF. Further clusters associated with these ORFs were classified as 'non-primary', with their highest peak as a non-primary TSS (npTSS). If the strongest TSS location was intra-ORF, without any TSSs located upstream of the ORF, then the ORF was manually re-defined as starting from the next ATG downstream.

## DESeq2 Differential Expression Analysis of GRG Genes

For analysing differential expression with the CAGE-seq dataset, a GFF was created with BEDtools extending from the pTSS coordinate, 25 bp upstream and 75 bp downstream, however, in cases of alternating pTSSs this region was defined as 25 bp upstream of the most upstream pTSS and 75 bp downstream of the most downstream pTSS. HTSeq-count (87) 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. (41)) and those regions showing changes with an adjusted p-value (padj) of <0.05 were considered significant. A caveat of this 'early' or 'late' definition is that it is a binary definition of whether a gene is up- or downregulated between conditions (time-points),

662 relative to the background read depth of reads which map to the genome in question. Further analysis of
663 ASFV genes used their characterised or predicted functions, from the VOCS tool database
664 (https://4virology.net/) (88,89) entries for the GRG genome.

665 ## Quantification of viral genome copies at different time points of infection

666 Porcine lung macrophages were seeded and infected as described above. *Vero* cells were similarly
667 cultured in 6-well plates in DMEM medium supplemented with 10% Fetal calf serum, 100 IU/ml penicillin
668 and 100 µg/ml streptomycin, when semi-confluent they were infected with MOI 5 of Ba71V. Immediately
669 after infection (after 1h adsorption period, considered '0 hpi), or at 5 hpi, and 16 hpi, the supernatant was
670 removed and nucleic acids were extracted using the Qiamp viral RNA kit (Qiagen) and quantified using a
671 NanoDrop spectrophotometer (ThermoFisher Scientific). For quantification of viral genome copy
672 equivalents, 50 ng of each nucleic acid sample was used in qPCR with primers and probe targeting the
673 viral capsid gene B646L. As previously described (90), standard curve quantification qPCR was carried out
674 on a Mx3005P system (Agilent Technologies) using the primers CTGCTCATGGTATCAATCTTATCGA and
675 GATACCACAAGATC(AG)GCCGT and probe 5'-(6-carboxyfluorescein [FAM])-
676 CCACGGGAGGAATACCAACCCAGTG-3'-(6-carboxytetramethylrhodamine [TAMRA]).

677 ## Analysis of mRNA levels by RT-PCR and quantitative real time PCR (qPCR)

678 RNA from GRG or Ba71V infected macrophages, or *Vero* cells respectively, or from uninfected cell controls,
679 was collected at the different time points post-infection with Trizol, as described above. RNA was reverse
680 transcribed (800 ng RNA per sample) using SuperScript III First-Strand Synthesis System for RT-PCR and
681 random hexamers (Invitrogen). For PCR, cDNAs were diluted 1:20 with nuclease free water and 1 µl each
682 sample was amplified in a total volume of 20 µl using Platinum™ Green Hot Start PCR Master Mix
683 (Invitrogen) and 200 nM of each primer. Annealing temperatures were tested for each primer pair in
684 gradient PCR to determine the one optimal for amplification. Supplementary Table 7a shows the primers
685 used for each gene target, the amplicon size, PCR reaction conditions, and NCBI accession numbers for
686 sequences used primer design. PCRs were then performed with limited cycles of amplification to have a
687 semi-quantitative comparison of transcript abundance between infection timepoints (by not reaching the
688 maximum product amplification plateau). Amplification products were viewed using 1.5% agarose gel
689 electrophoresis.

690 C315R transcript levels were assessed by qPCR, using housekeeping gene glyceraldehyde-3-phosphate
691 dehydrogenase (GAPDH) expression was used for normalisation. Primer details and the qPCR

amplification program are shown in Supplementary Table 7b (GAPDH primers used for *Vero* cells were previously published by Melchjorsen et al., 2009 (91)). Primers were used at 250 nM concentration with Brilliant III Ultra-Fast SYBR® Green QPCR Master Mix (Agilent 600882), 1 µl cDNA in 20 µl (1:20) total reaction volumes, and qPCRs carried out in Mx3005P system (Agilent Technologies). Similar amplification efficiencies (97-102%) for all primers had been observed upon amplification of serially diluted cDNA samples, and the relative expression at each timepoint of infection was calculated using the formula $2^{\Delta Ct}$ ($2^{Ct\_GAPDH-Ct\_C315R}$).

## ASFV Promoter Motif Analysis

DESeq2 results were used to categorise ASFV genes into two simple sub-classes: early; 87 genes downregulated from early to late infection and late; the 78 upregulated from early to late infection. These characterised gene pTSSs were then pooled with the nORF pTSSs, and sequences upstream and downstream of the pTSS were extracted from the GRG genome in FASTA format using BEDtools. Sequences 35 bp upstream of and including the pTSSs were analysed using MEME software (http://meme-suite.org) (92), searching for 5 motifs with a maximum width of 20 nt and 27 nt, respectively (other settings at default). The input for MEME motif searches included sequences upstream of 134 early pTSSs (87 genes and 47 nORFs) for early promoter searching, while 234 late pTSSs (78 genes and 156 nORFs) were used to search for late promoters. For analysis of conserved motifs upstream of the five clusters described in Figure 6a-b, sequences were extracted in the same manner as above, but grouped according to their cluster. MEME motif searches were carried out for sequences in each cluster, searching for 3 motifs, 5-36 bp in length, with zero or one occurrence per sequence ('zoops' mode).

## Identification of TSSs by rapid amplification of cDNA ends - 5'RACE

For 5'RACE of GRG genes DP146L, pNG4 and CP204L we designed the gene specific primers (GSP) shown in Supplementary Table 7c, and used the kit: "5´ RACE System for Rapid Amplification of cDNA Ends" (Invitrogen), according to manufacturer instructions. Briefly, 150 ng RNA from either 5 hpi or 16 hpi macrophages (one of the replicate RNA samples used for CAGE-seq) was used for cDNA synthesis with GSP1 primers, followed by degradation of the mRNA template with RNase Mix, and column purification of the cDNA. A homopolymeric tail was added to the cDNA 3'ends with Terminal deoxynucleotidyl transferase, which allowed PCR amplification with an "Abridged Anchor Primer" (AAP) from the 5'RACE kit and a nested GSP2 primer. A second PCR was performed over an aliquot of the previous, with 5'RACE "Abridged Universal amplification Primer" (AUAP), and an additional nested primer GSP3, except for pNG4

where GSP2 was re-used due to the small predicted size of the amplicon. Platinum™ Green Hot Start PCR Master Mix (Invitrogen) was used for PCR and products were run in 2% agarose gel electrophoresis (see Supplementary Table 7c for expected sizes). Efficient recovery of cDNA from the purification column requires a product of at least 200 bases and therefore, due to the small predicted size of pNG4 transcripts its GSP1 primer was extended at the 5' end with an irrelevant non-annealing sequence of extra 50 nt in order to create a longer recoverable product.

## CAGE-seq Analysis for the *Sus scrofa* Genome

Analyses of TSS-mapping, gene expression and motif searching with CAGE-seq reads mapped to the *Sus scrofa* 11.1 genome were carried out by DNAFORM (Yokohama, Kanagawa, Japan). The 5' ends of CAGE-seq reads were utilised as input for the Reclu pipeline (93) with a cutoff of 0.1 RPM, and irreproducible discovery rate of 0.1. 37,159 total CAGE-seq peaks could be identified, of which around half (16,720) match unique CAGE peaks previously identified by Roberts et al. (59) (i.e. within 100 nt of any of them). TSSs for 9,384 protein-coding genes (out of 21,288) were annotated de novo from the CAGE-defined TSSs (Supplementary Table 4).

Protein-coding genes with annotated TSSs (9,384 out of 21,288) were then subjected to differential expression analysis. CAGE-seq reads were summed up over all TSSs assigned to a gene and compared between two time points using edgeR (94) at maximum false discovery rate of 0.05. The full list of host genes with annotated promoters together with their estimated expression levels is provided in Supplementary Table 5. Gene set enrichment analysis was performed with the DAVID 6.8 Bioinformatics Resources (95), using best BLASTP (96) human hits (from the UniProt (97) reference human proteome). The 9,331 genes with human homologs were used as a background, and functional annotations of the four major expression response groups (late/early up-/down-regulated genes) were clustered in DAVID 6.8 using medium classification stringency. MEME motif searches were conducted for promoters of four differentially regulated subsets of host genes, as defined in Figure 11a. Promoters sequences were extended 1000 bp upstream and 200 bp downstream of TSSs, searched with MEME (max. 10 motifs, max. 100 bp long, on a given strand only, zero or one site per sequence, $E < 0.01$), and then compared against known vertebrate DNA motifs with Tomtom (p-value < 0.01).

## Data Availability

Raw sequencing data are available on the Sequence Read Archive (SRA) database under BioProject: PRJNA739166. This also includes CAGE-seq data aligned to the ASFV-GRG (FR682468.1 *Sus scrofa* (GCF_000003025.6) genomes (see methods above) in BAM format. Available for review via the link below:

https://dataview.ncbi.nlm.nih.gov/object/PRJNA739166?reviewer=390lg85cohvh81llto5gr1d22n

## Acknowledgements

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). 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). Research in the ASFV group at The Pirbright Institute is funded by the Biotechnology and Biological Sciences Research Council (BBS/E/I/0007030 and BBS/E/I/0007031). The authors are grateful to all members of the RNAP lab and Tine Arnvig for critical reading of the manuscript.

## References

1. Gogin A, Gerasimov V, Malogolovkin A, Kolbasov D. African swine fever in the North Caucasus region and the Russian Federation in years 2007-2012. Virus Res. 2013 Apr 1;173(1):198–203.

2. Zhou X, Li N, Luo Y, Liu Y, Miao F, Chen T, et al. Emergence of African Swine Fever in China, 2018. Transbound Emerg Dis. 2018 Dec 1;65(6):1482–4.

3. Alonso C, Borca M, Dixon L, Revilla Y, Rodriguez F, Escribano JM, et al. ICTV Virus Taxonomy Profile: Asfarviridae. J Gen Virol. 2018 May 1;99(5):613–4.

4. Koonin E V., Yutin N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology. 2010;53(5):284–92.

5. Yutin N, Koonin E V. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. Virol J. 2012 Aug 14;9(1):161.

6. Broyles SS. Vaccinia virus transcription. J Gen Virol. 2003 Sep 1;84(9):2293–303.

7. Alejo A, Matamoros T, Guerra M, Andrés G. A proteomic atlas of the African swine fever virus particle. J Virol. 2018 Sep 5;JVI.01293-18.

776 8. Salas ML, Kuznar J, Viñuela E. Polyadenylation, methylation, and capping of the RNA synthesized
777 in vitro by African swine fever virus. Virology. 1981 Sep 1;113(2):484–91.

778 9. Rodríguez JM, Salas ML. African swine fever virus transcription. Vol. 173, Virus Research. Elsevier
779 B.V.; 2013. p. 15–28.

780 10. Cackett G, Matelska D, Sýkora M, Portugal R, Malecki M, Bähler J, et al. The African Swine Fever
781 Virus Transcriptome. J Virol. 2020 Feb 19;94(9).

782 11. Iyer LM, Balaji S, Koonin E V., Aravind L. Evolutionary genomics of nucleo-cytoplasmic large DNA
783 viruses. Virus Res. 2006 Apr;117(1):156–84.

784 12. Yutin N, Wolf YI, Raoult D, Koonin E V. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters
785 of orthologous genes and reconstruction of viral genome evolution. Virol J. 2009 Dec 17;6(3):223.

786 13. Rodríguez JM, Moreno LT, Alejo A, Lacasta A, Rodríguez F, Salas ML. Genome Sequence of African
787 Swine Fever Virus BA71, the Virulent Parental Strain of the Nonpathogenic and Tissue-Culture
788 Adapted BA71V. Munderloh UG, editor. PLoS One. 2015 Nov 30;10(11):e0142889.

789 14. Yáñez RJ, Rodríguez JM, Nogal ML, Yuste L, Enríquez C, Rodriguez JF, et al. Analysis of the complete
790 nucleotide sequence of African swine fever virus. Virology. 1995 Apr 1;208(1):249–78.

791 15. Dixon LK, Chapman DAG, Netherton CL, Upton C. African swine fever virus replication and
792 genomics. Virus Res. 2013;173(1):3–14.

793 16. Chapman DAG, Tcherepanov V, Upton C, Dixon LK. Comparison of the genome sequences of non-
794 pathogenic and pathogenic African swine fever virus isolates. J Gen Virol. 2008 Feb 1;89(2):397–
795 408.

796 17. Forth JH, Forth LF, King J, Groza O, Hübner A, Olesen AS, et al. A deep-sequencing workflow for the
797 fast and efficient generation of high-quality African swine fever virus whole-genome sequences.
798 Viruses. 2019;11(9).

799 18. Chapman DAG, Darby AC, da Silva M, Upton C, Radford AD, Dixon LK. Genomic analysis of highly
800 virulent Georgia 2007/1 isolate of African swine fever virus. Emerg Infect Dis. 2011 Apr;17(4):599–
801 605.

802 19. Farlow J, Donduashvili M, Kokhreidze M, Kotorashvili A, Vepkhvadze NG, Kotaria N, et al. Intra-

epidemic genome variation in highly pathogenic African swine fever virus (ASFV) from the country of Georgia. Virol J. 2018 Dec 14;15(1):190.

20. Mazur-Panasiuk N, Woźniakowski G, Niemczuk K. The first complete genomic sequences of African swine fever virus isolated in Poland. Sci Rep. 2019 Dec 1;9(1):3–5.

21. Granberg F, Torresi C, Oggiano A, Malmberg M, Iscaro C, De Mia GM, et al. Complete genome sequence of an African swine fever virus isolate from Sardinia, Italy. Genome Announc. 2016;4(6):1220–36.

22. Wang Z, Jia L, Li J, Liu H, Liu D. Pan-Genomic Analysis of African Swine Fever Virus. Virologica Sinica. Science Press; 2019. p. 1–4.

23. Rowlands RJ, Michaud V, Heath L, Hutchings G, Oura C, Vosloo W, et al. African swine fever virus isolate, Georgia, 2007. Emerg Infect Dis. 2008 Dec;14(12):1870–4.

24. Zhao D, Liu R, Zhang X, Li F, Wang J, Zhang J, et al. Replication and virulence in pigs of the first African swine fever virus isolated in China. Emerg Microbes Infect. 2019 Jan 1;8(1):438–47.

25. Zani L, Forth JH, Forth L, Nurmoja I, Leidenberger S, Henke J, et al. Deletion at the 5'-end of Estonian ASFV strains associated with an attenuated phenotype. Sci Reports 2018 81. 2018 Apr 25;8(1):1–11.

26. Gallardo C, Nurmoja I, Soler A, Delicado V, Simón A, Martin E, et al. Evolution in Europe of African swine fever genotype II viruses from highly to moderately virulent. Vet Microbiol. 2018 Jun 1;219:70–9.

27. Pershin A, Shevchenko I, Igolkin A, Zhukov I, Mazloum A, Aronova E, et al. A Long-Term Study of the Biological Properties of ASF Virus Isolates Originating from Various Regions of the Russian Federation in 2013–2018. Vet Sci 2019, Vol 6, Page 99. 2019 Dec 6;6(4):99.

28. E S, Z Z, Z W, X H, X Z, L W, et al. Emergence and prevalence of naturally occurring lower virulent African swine fever viruses in domestic pigs in China in 2020. Sci China Life Sci. 2021 May 1;64(5):752–65.

29. Imbery J, Upton C. Organization of the multigene families of African Swine Fever Virus. Fine Focus. 2017;3(2):155–70.

30. Netherton CL, Connell S, Benfield CTO, Dixon LK. The Genetics of Life and Death: Virus-Host Interactions Underpinning Resistance to African Swine Fever, a Viral Hemorrhagic Disease. Front Genet. 2019 May 3;10(MAY):402.

31. Keßler C, Forth JH, Keil GM, Mettenleiter TC, Blome S, Karger A. The intracellular proteome of African swine fever virus. Sci Rep. 2018 Oct 2;8(1):14714.

32. Randall RE, Goodbourn S. Interferons and viruses: An interplay between induction, signalling, antiviral responses and virus countermeasures. Vol. 89, Journal of General Virology. Microbiology Society; 2008. p. 1–47.

33. Afonso RCL, Piccone ME, Zaffuto KM, Neilan J, Kutish GF, Lu Z, et al. African swine fever virus multigene family 360 and 530 genes affect host interferon response. J Virol. 2004;78:1858–64.

34. Neilan JG, Zsak L, Lu Z, Kutish GF, Afonso CL, Rock DL. Novel Swine Virulence Determinant in the Left Variable Region of the African Swine Fever Virus Genome. J Virol. 2002 Apr 1;76(7):3095–104.

35. Reis AL, Abrams CC, Goatley LC, Netherton C, Chapman DG, Sanchez-Cordon P, et al. Deletion of African swine fever virus interferon inhibitors from the genome of a virulent isolate reduces virulence in domestic pigs and induces a protective response. Vaccine. 2016 Sep 7;34(39):4698–705.

36. Golding JP, Goatley L, Goodbourn S, Dixon LK, Taylor G, Netherton CL. Sensitivity of African swine fever virus to type I interferon is linked to genes within multigene families 360 and 505. Virology. 2016 Jun 1;493:154–61.

37. Mosser DM, Edwards JP. Exploring the full spectrum of macrophage activation. Vol. 8, Nature Reviews Immunology. NIH Public Access; 2008. p. 958–69.

38. Roy S, Schmeier S, Arner E, Alam T, Parihar SP, Ozturk M, et al. Redefining the transcriptional regulatory dynamics of classically and alternatively activated macrophages by deepCAGE transcriptomics. Nucleic Acids Res. 2015;43(14):6969–82.

39. Zhu JJ, Ramanathan P, Bishop EA, O'Donnell V, Gladue DP, Borca M V. Mechanisms of African swine fever virus pathogenesis and immune evasion inferred from gene expression changes in infected swine macrophages. PLoS One. 2019;14(11).

857 40. Jaing C, Rowland RRR, Allen JE, Certoma A, Thissen JB, Bingham J, et al. Gene expression analysis
858 of whole blood RNA from pigs infected with low and high pathogenic African swine fever viruses.
859 Sci Rep. 2017 Dec 31;7(1):10115.

860 41. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq
861 data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.

862 42. Hammond JM, Kerr SM, Smith GL, Dixon LK. An African swine fever virus gene with homology to
863 DNA ligases. Nucleic Acids Res. 1992 Jun 11;20(11):2667–71.

864 43. Reis AL, Goatley LC, Jabbar T, Sanchez-Cordon PJ, Netherton CL, Chapman DAG, et al. Deletion of
865 the African Swine Fever Virus Gene DP148R Does Not Reduce Virus Replication in Culture but
866 Reduces Virus Virulence in Pigs and Induces High Levels of Protection against Challenge. J Virol.
867 2017 Dec 15;91(24).

868 44. Yang Z, Martens CA, Bruno DP, Porcella SF, Moss B. Pervasive initiation and 3'-end formation of
869 poxvirus postreplicative RNAs. J Biol Chem. 2012 Sep 7;287(37):31050–60.

870 45. Frouco G, Freitas FB, Coelho J, Leitão A, Martins C, Ferreira F. DNA-Binding Properties of African
871 Swine Fever Virus pA104R, a Histone-Like Protein Involved in Viral Replication and Transcription. J
872 Virol. 2017;91(12).

873 46. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of
874 best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17:13.

875 47. García-Escudero R, Viñuela E. Structure of African Swine Fever Virus Late Promoters: Requirement
876 of a TATA Sequence at the Initiation Region. J Virol. 2000 Sep 1;74(17):8176–82.

877 48. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif
878 discovery and searching. Nucleic Acids Res. 2009 Jul 1;37(Web Server):W202–8.

879 49. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W. Quantifying similarity between motifs.
880 Genome Biol. 2007 Feb 26;8(2):R24.

881 50. Rodríguez JM, Salas ML, Viñuela E. Intermediate class of mRNAs in African swine fever virus. J Virol.
882 1996 Dec;70(12):8584–9.

883 51. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server.

884 Nucleic Acids Res. 2004;32(WEB SERVER ISS.):W526–31.

885 52. J Z, Y Z, T C, J Y, H Y, L W, et al. Deletion of the L7L-L11L Genes Attenuates ASFV and Induces
886 Protection against Homologous Challenge. Viruses. 2021 Feb 1;13(2).

887 53. Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein Sequence Analysis Using
888 the MPI Bioinformatics Toolkit. Curr Protoc Bioinforma. 2020 Dec 1;72(1).

889 54. Kang Y, Xu J, Liu Y, Sun J, Sun D, Hu Y, et al. Crystal structure of the cell corpse engulfment protein
890 CED-2 in Caenorhabditis elegans. Biochem Biophys Res Commun. 2011 Jul 1;410(2):189–94.

891 55. Kaneko T, Huang H, Zhao B, Li L, Liu H, Voss CK, et al. Loops govern SH2 domain specificity by
892 controlling access to binding pockets. Sci Signal. 2010 May 4;3(120).

893 56. Baylis SA, Banham AH, Vydelingum S, Dixon LK, Smith GL. African swine fever virus encodes a serine
894 protein kinase, which is packaged into virions. J Virol. 1993;67:4549–56.

895 57. Aravind L, Koonin E V. DNA polymerase beta-like nucleotidyltransferase superfamily: identification
896 of three new families, classification and evolutionary history. Nucleic Acids Res. 1999 Apr
897 1;27(7):1609–18.

898 58. Priet S, Lartigue A, Debart F, Claverie JM, Abergel C. MRNA maturation in giant viruses: Variation
899 on a theme. Nucleic Acids Res. 2015;43(7):3776–88.

900 59. Robert C, Kapetanovic R, Beraldi D, Watson M, Archibald AL, Hume DA. Identification and
901 annotation of conserved promoters and macrophage-expressed genes in the pig genome. BMC
902 Genomics. 2015 Nov 18;16(1).

903 60. Ganchi PA, Sun SC, Greene WC, Ballard DW. A novel NF-kappa B complex containing p65
904 homodimers: implications for transcriptional control at the level of subunit dimerization. Mol Cell
905 Biol. 1993 Dec;13(12):7826–35.

906 61. Dixon LK, Abrams CC, Bowick G, Goatley LC, Kay-Jackson PC, Chapman D, et al. African swine fever
907 virus proteins involved in evading host defence systems. In: Veterinary Immunology and
908 Immunopathology. 2004. p. 117–34.

909 62. Powell PP, Dixon LK, Parkhouse RM. An IkappaB homolog encoded by African swine fever virus
910 provides a novel mechanism for downregulation of proinflammatory cytokine responses in host

911 macrophages. J Virol. 1996;70(12):8527–33.

912 63. Granja AG, Sánchez EG, Sabina P, Fresno M, Revilla Y. African swine fever virus blocks the host cell
913 antiviral inflammatory response through a direct inhibition of PKC-theta-mediated p300
914 transactivation. J Virol. 2009 Jan 15;83(2):969–80.

915 64. Takeya T, Hanafusa H. DNA sequence of the viral and cellular src gene of chickens. II. Comparison
916 of the src genes of two strains of avian sarcoma virus and of the cellular homolog. J Virol.
917 1982;44(1):12–8.

918 65. Kaneko T, Stogios PJ, Ruan X, Voss C, Evdokimova E, Skarina T, et al. Identification and
919 characterization of a large family of superbinding bacterial SH2 domains. Nat Commun. 2018 Dec
920 1;9(1).

921 66. Krug PW, Holinka LG, O'Donnell V, Reese B, Sanford B, Fernandez-Sainz I, et al. The progressive
922 adaptation of a georgian isolate of African swine fever virus to vero cells leads to a gradual
923 attenuation of virulence in swine corresponding to major modifications of the viral genome. J Virol.
924 2015 Feb 15;89(4):2324–32.

925 67. Camacho A, ViÑuela E. Protein p22 of African swine fever virus: An early structural protein that is
926 incorporated into the membrane of infected cells. Virology. 1991 Mar 1;181(1):251–7.

927 68. Netherton C, Rouiller I, Wileman T. The subcellular distribution of multigene family 110 proteins
928 of African swine fever virus is determined by differences in C-terminal KDEL endoplasmic reticulum
929 retention motifs. J Virol. 2004 Apr 1;78(7):3710–21.

930 69. Cackett G, Sýkora M, Werner F. Transcriptome view of a killer: African swine fever virus. Vol. 48,
931 Biochemical Society Transactions. Portland Press Ltd; 2020. p. 1569–81.

932 70. Quintas A, Pérez-Núñez D, Sánchez EG, Nogal ML, Hentze MW, Castelló A, et al. Characterization
933 of the African Swine Fever Virus Decapping Enzyme during Infection. Jung JU, editor. J Virol. 2017
934 Dec 15;91(24):e00990-17.

935 71. Kunsch C, Ruben SM, Rosen CA. Selection of optimal kappa B/Rel DNA-binding motifs: interaction
936 of both subunits of NF-kappa B with DNA is required for transcriptional activation. Mol Cell Biol.
937 1992 Oct;12(10):4412–21.

938 72. Gómez del Moral M, Ortuño E, Fernández-Zapatero P, Alonso F, Alonso C, Ezquerra A, et al. African
939 Swine Fever Virus Infection Induces Tumor Necrosis Factor Alpha Production: Implications in
940 Pathogenesis. J Virol. 1999 Mar 1;73(3):2173–80.

941 73. Xia C, Braunstein Z, Toomey AC, Zhong J, Rao X. S100 proteins as an important regulator of
942 macrophage inflammation. Vol. 8, Frontiers in Immunology. Frontiers Media S.A.; 2018.

943 74. Yang Z, Bruno DP, Martens CA, Porcella SF, Moss B. Simultaneous high-resolution analysis of
944 vaccinia virus and host cell transcriptomes by deep RNA sequencing. Proc Natl Acad Sci U S A. 2010
945 Jun 22;107(25):11513–8.

946 75. Olasz F, Tombácz D, Torma G, Csabai Z, Moldován N, Dörmő Á, et al. Short and Long-Read
947 Sequencing Survey of the Dynamic Transcriptomes of African Swine Fever Virus and the Host Cells.
948 Front Genet. 2020 Jul 28;11:2020.02.27.967695.

949 76. Yang Z, Bruno DP, Martens CA, Porcella SF, Moss B. Genome-Wide Analysis of the 5' and 3' Ends of
950 Vaccinia Virus Early mRNAs Delineates Regulatory Sequences of Annotated and Anomalous
951 Transcripts. J Virol. 2011 Jun;85(12):5897–909.

952 77. Gershon PD, Moss B. Early transcription factor subunits are encoded by vaccinia virus late genes.
953 Proc Natl Acad Sci U S A. 1990 Jun 1;87(11):4401–5.

954 78. Li J, Broyles SS. Recruitment of vaccinia virus RNA polymerase to an early gene promoter by the
955 viral early transcription factor. J Biol Chem. 1993;268(4):2773–80.

956 79. Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK. TATA element recognition by the TATA
957 box-binding protein has been conserved throughout evolution. Genes Dev. 1999 Dec
958 15;13(24):3217–30.

959 80. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. Babraham, England:
960 Babraham Bioinformatics;

961 81. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr
962 4;9(4):357–9.

963 82. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat
964 Methods. 2015 Apr 9;12(4):357–60.

965 83. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
966 Bioinformatics. 2010 Mar 15;26(6):841–2.

967 84. RStudio Team. RStudio: Integrated Development for R. Boston, MA: RStudioe, Inc; 2016.

968 85. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-
969 throughput genomic analysis with Bioconductor. Nat Methods. 2015 Feb 1;12(2):115–21.

970 86. Thodberg M, Thieffry A, Vitting-Seerup K, Andersson R, Sandelin A. CAGEfightR: Cap Analysis of
971 Gene Expression (CAGE) in R/Bioconductor. bioRxiv. 2018 Apr 28;310623.

972 87. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing
973 data. Bioinformatics. 2015 Jan 15;31(2):166–9.

974 88. Upton C, Slack S, Hunter AL, Ehlers A, Roper RL, Rock DL. Poxvirus orthologous clusters: toward
975 defining the minimum essential poxvirus genome. J Virol. 2003 Jul 1;77(13):7590–600.

976 89. Tu SL, Upton C. Bioinformatics for Analysis of Poxvirus Genomes. In: Methods in Molecular Biology.
977 Humana Press Inc.; 2019. p. 29–62.

978 90. DP K, SM R, GH H, SS G, PJ W, LK D, et al. Development of a TaqMan PCR assay with internal
979 amplification control for the detection of African swine fever virus. J Virol Methods. 2003
980 Jan;107(1):53–61.

981 91. J M, H K, R C, J R, S M, SR P, et al. Differential regulation of the OASL and OAS1 genes in response
982 to viral infections. J Interferon Cytokine Res. 2009 Apr 1;29(4):199–207.

983 92. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in
984 biopolymers. Proceedings Int Conf Intell Syst Mol Biol. 1994;2:28–36.

985 93. Ohmiya H, Vitezic M, Frith MC, Itoh M, Carninci P, Forrest ARR, et al. RECLU: A pipeline to discover
986 reproducible transcriptional start sites and their alternative regulation using capped analysis of
987 gene expression (CAGE). BMC Genomics. 2014 Apr 25;15(1):269.

988 94. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression
989 analysis of digital gene expression data. Bioinformatics. 2009 Nov 11;26(1):139–40.

990 95. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using

DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.

96. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

97. Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

98. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W187-91.

99. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004 May 12;14(6):1188–90.

100. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017–8.

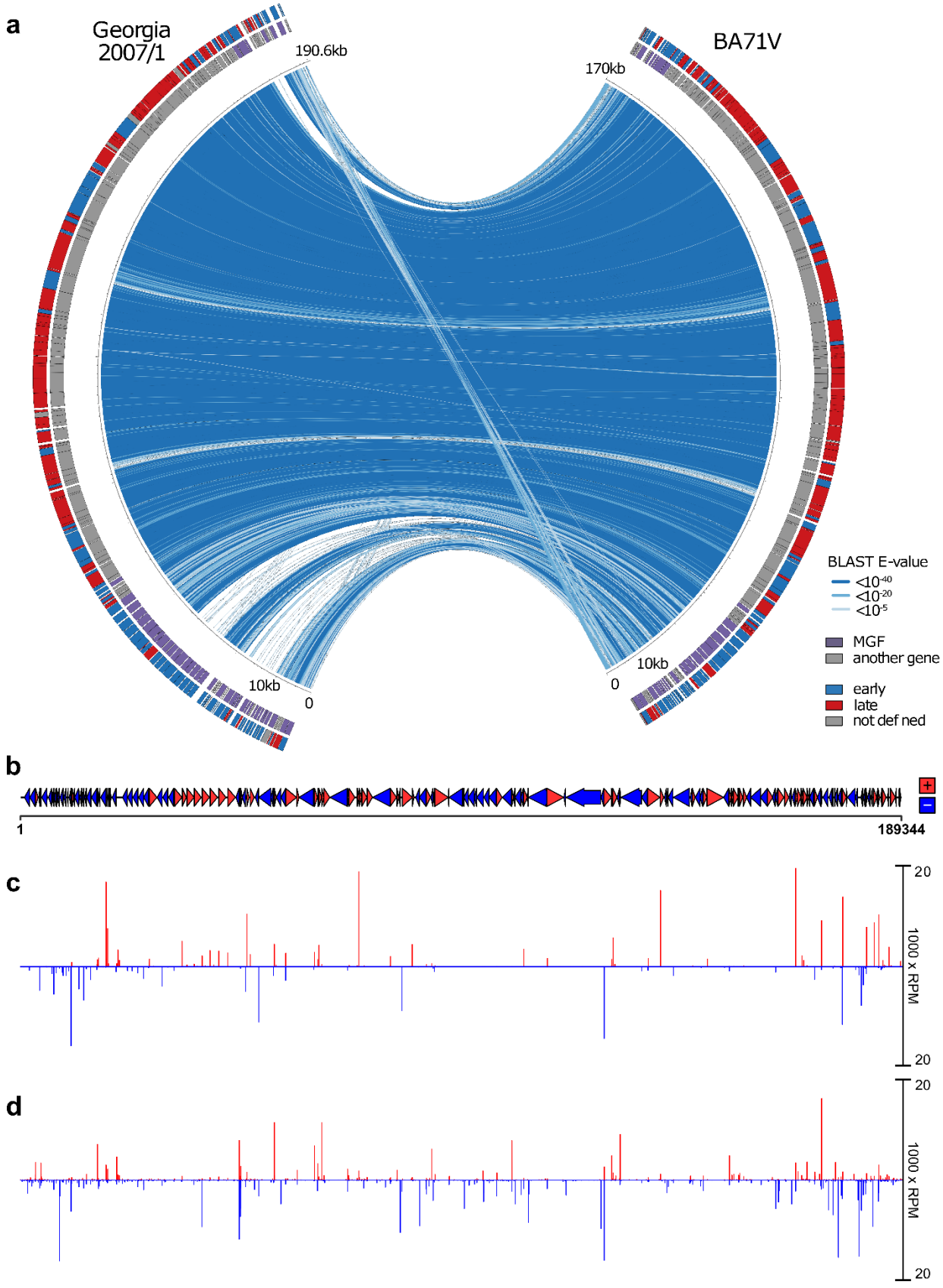
101. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012 May 1;40(10):4288–97.

102. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995 Jan;57(1):289–300.

103. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020 Jan 1;48(D1):D682–8.

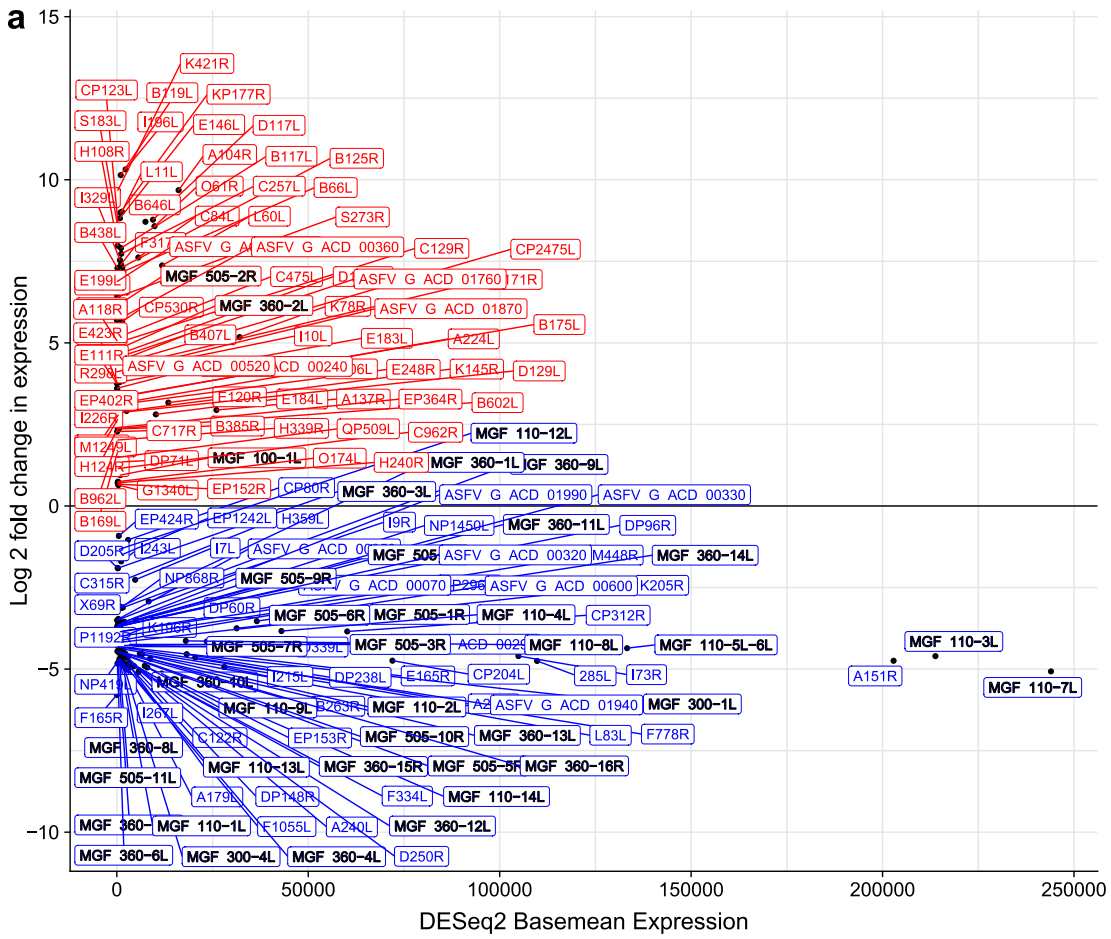
104. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, et al. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 2003 Jan 1;31(1):28–33.

Figures




1012

Figure 1. Functional genome annotation of ASFV GRG. (a) Comparison between the genomes of BA71V and GRG, generated with Circos (http://circos.ca/). Blue lines represent sequence conservation (Blast E-values per 100 nt). The Inner ring represents genes defined as MGF members (purple), and all others (grey). The outer ring shows annotated genes which we have defined as early or late according to downregulation or upregulation between 5 hpi and 16 hpi from DESeq2 analysis. (b) 189 GRG annotated ORFs are represented as arrows and coloured according to strand. CAGE-seq peaks across the GRG genome at 5 hpi (c) and 16 hpi (d), normalized coverage reads per million mapped reads (RPM) of 5' ends of CAGE-seq reads. The coverage was capped at 20000 RPM for visualisation, though multiple peaks exceeded this. DeepTools (98), was used to convert bam files to bigwig format and imported into Rstudio for visual representation via packages ggplot, ggbio, rtracklayer, and gggenes was used to generate the ORF map in (b).






1024

Figure 2. Summary of GRG gene expression (a) Expression profiles for 164 genes for which we annotated pTSSs from CAGE-seq and which showed significant differential expression. Log2 fold change and basemean expression values were from DESeq2 analysis of raw counts (see methods). Genes are coloured according to their log2 fold change in expression as red (positive: upregulated from 5 hpi to 16 hpi) or blue (negative: downregulated). MGFs are emphasised with a black outline to highlight their overrepresentation in the group of downregulated genes. (b) Expression profiles for 43 genes (excluding nORFs) only detected as being expressed in GRG and not BA71V, format as in (a). (c) Expression (RPM) of 20 highest-expressed genes at 5 hpi, error bars represent standard deviation between replicates. (d) Expression (RPM) of 20 highest-expressed genes at 16 hpi pi, error bars are the standard deviation between replicates.

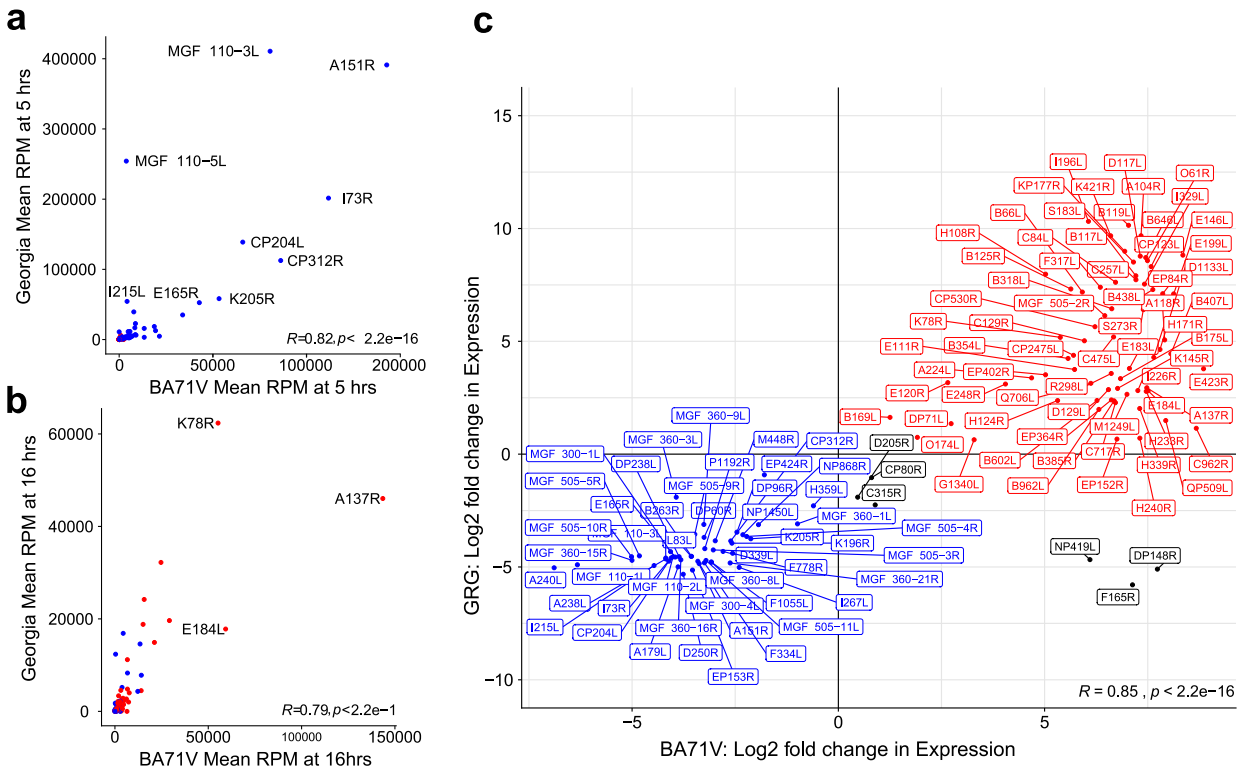



Figure 3. Comparison of gene expression profiles for genes shared between GRG and BA71V. Scatter plots of mean RPM across replicates for shared genes at 5 hpi (a) and 16 hpi (b), coloured according to whether genes show significant downregulation (blue), or upregulation (red) according to DESeq2 analysis in GRG. In both (b) and (c) genes with RPM values above 40000 RPM in either strain are labelled. (c) Comparison of log2 fold change in expression values of genes in GRG and BA71V, in blue are downregulated (early)

genes in both strains, red are upregulated (late) genes in both strains, while the genes which disagree in their differential expression patterns between strains are in black. R represents the Pearson Correlation coefficient for each individual plot in (a), (b), and (c).

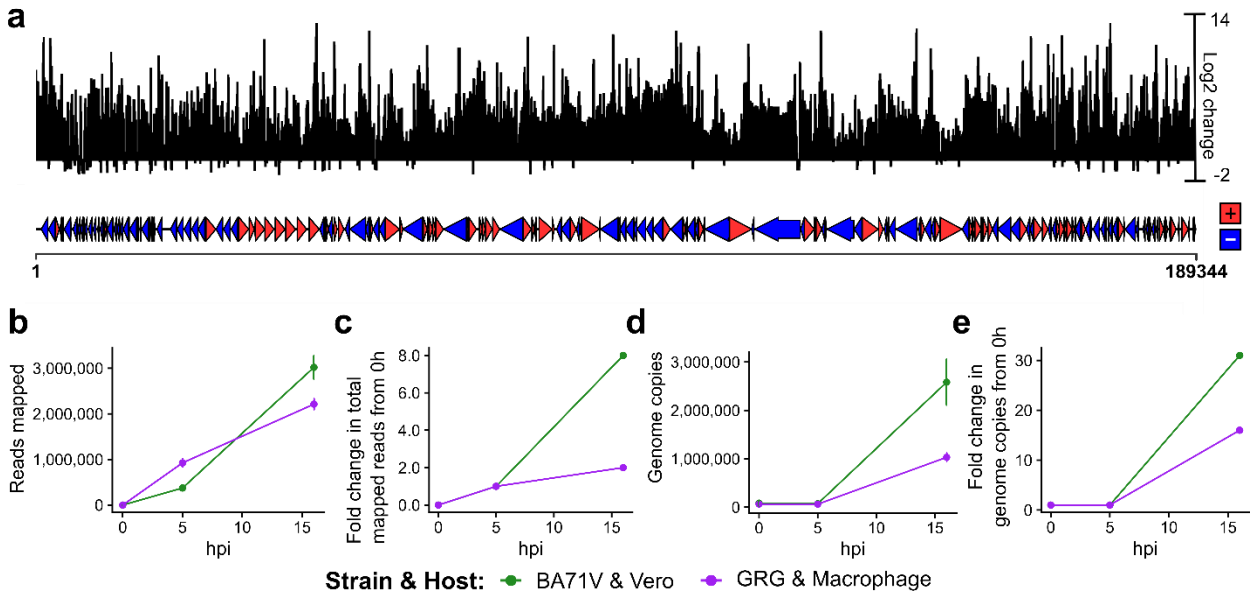
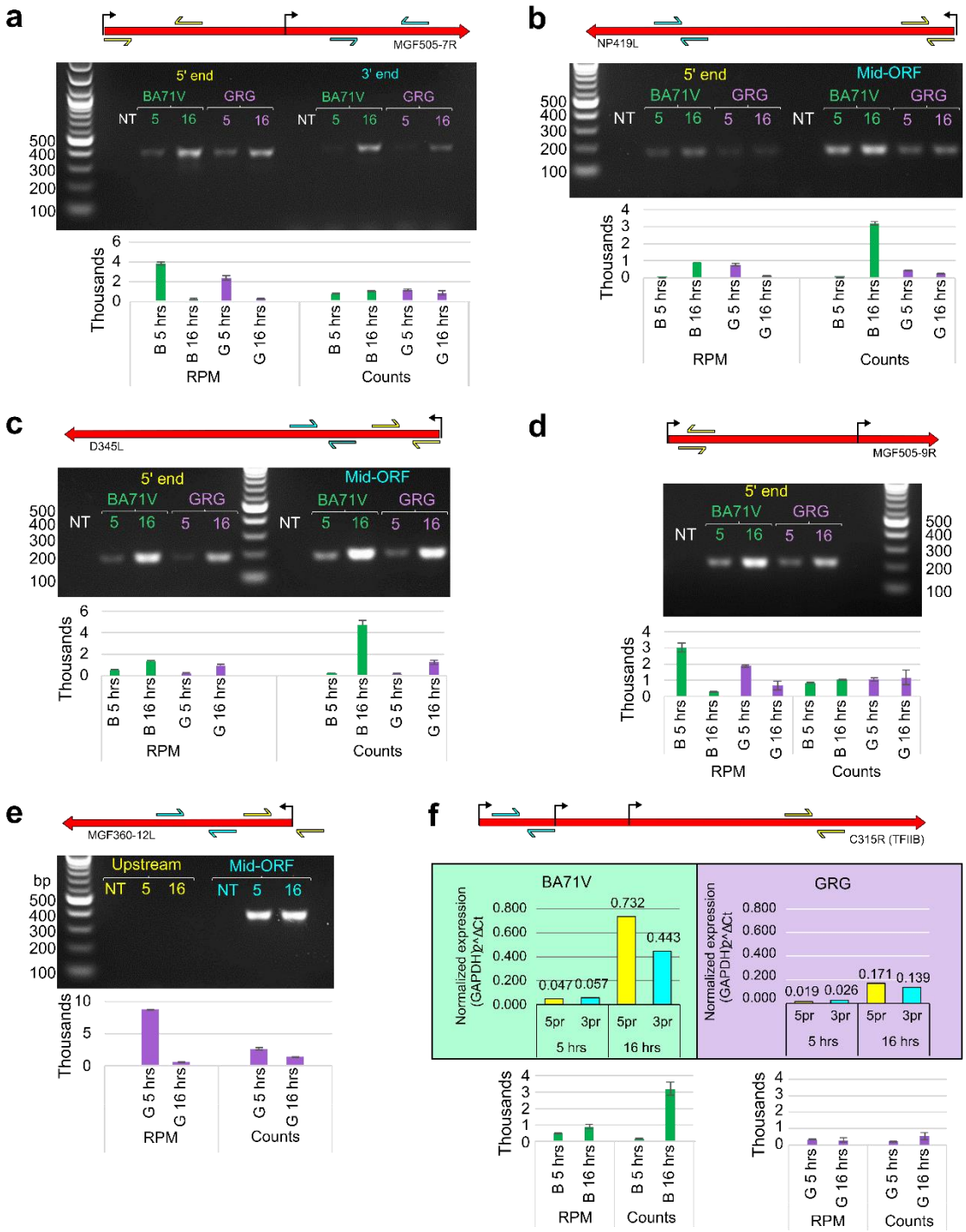



Figure 4. Increase in virus genome copy number mRNA levels during late infection. (a) The 'log2 change' represents log2 of the ratio of CAGE-seq reads (normalised per million mapped reads) at 16 hpi vs. 5 hpi per nucleotide across the genome. Alignment comparisons and calculations were done with deepTools (98). (b) Replicate means of CAGE-seq reads mapped to either the BA71V (green) or GRG (purple) genomes throughout infection. (c) Fold change in CAGE-seq reads during infection, calculated via mean value across 2 replicates, but with the assumption number of reads at 0 hpi is 0, therefore dividing by values from 5 hpi. (d) Change in genome copies from DNA qPCR of B646L gene, dividing by value at 0 hpi to represent '1 genome copy per infected cell'. (e) Fold change in genome copies present at 0 hpi , 5 hpi and 16 hpi from qPCR in (d). (d) calculated as for (c), but with actual vales for 0 hpi.




1054

1055 Figure 5. RT-PCR results of genes for comparison to CAGE-seq data from (a) MGF505-7R, (b) NP419L, (c)
1056 D345L, (d) MGF360-12L, (e) MGF505-9R, and (f) qRT-PCR results of C315R (ASFV-TFIIB). (NT = no template
1057 control). For each panel at the top is a diagrammatic representation of each gene's TSSs (bent arrow,
1058 including both pTSS and ioTSSs), annotated ORF (red arrow), and arrow pairs in cyan or yellow represent

the primers used for PCR (see methods for primer sequences). Beneath each PCR results are bar charts representing the CAGE-seq results as either normalised (mean RPM) or raw (mean read counts) data, error bars show the range of values from each replicate.

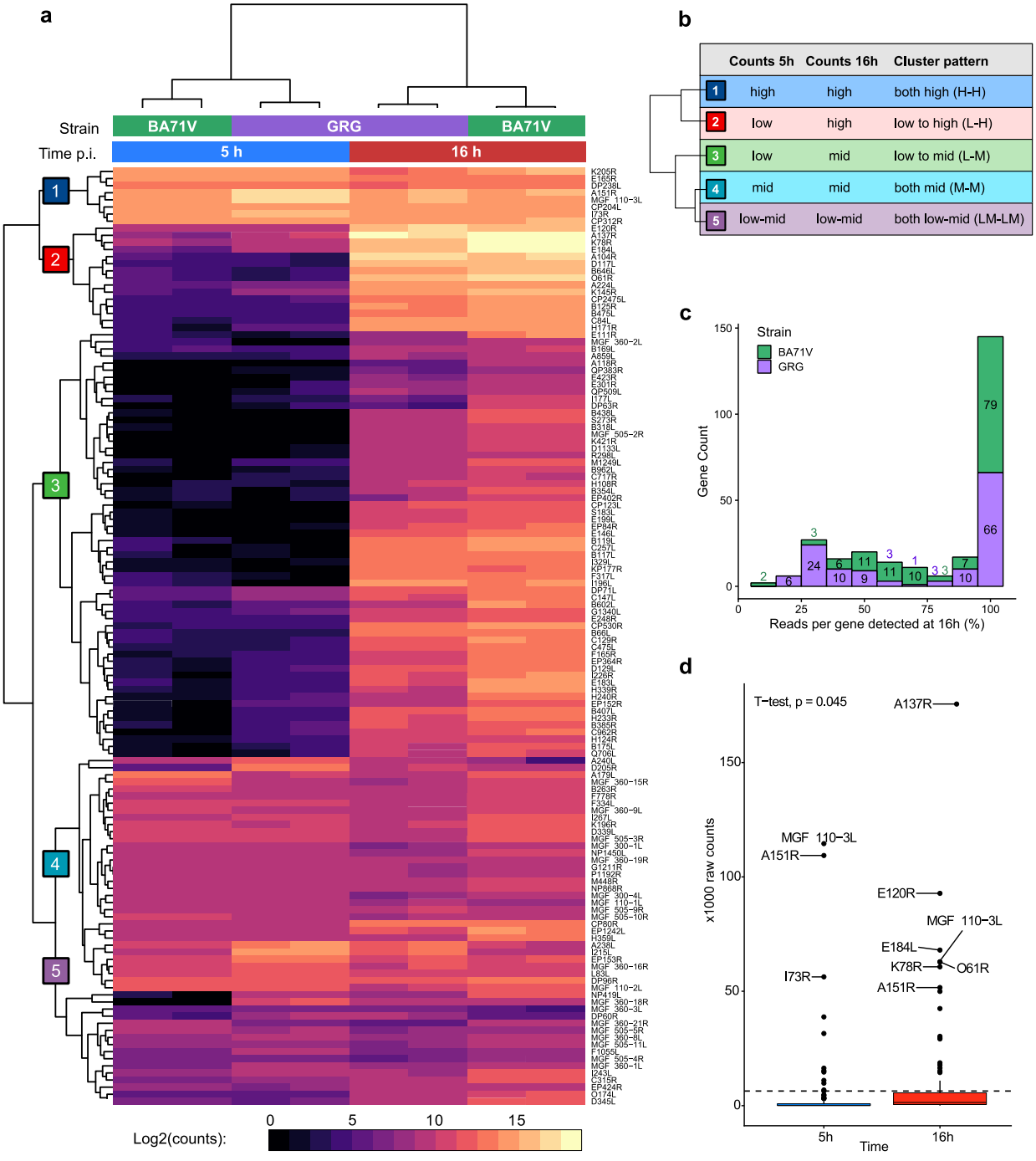


Figure 6. Comparison of the raw read counts for genes shared between BA71V and GRG. (a) clustered heatmap representation of raw counts for genes shared between BA71V and GRG, generated with pheatmap. (b) broad patterns represented by genes in the 5 clusters indicated in (a). (c) histogram showing the percentage of the total raw reads per gene which are detected at 16 hpi vs. 5 hpi post-infection, and comparing the distribution of percentages between GRG and BA71V. (d) Mean read counts from GRG at 5 hpi vs 16 hpi replicates, showing a significant increase (T-test, p-value: 0.045) from 5 hpi to 16 hpi.

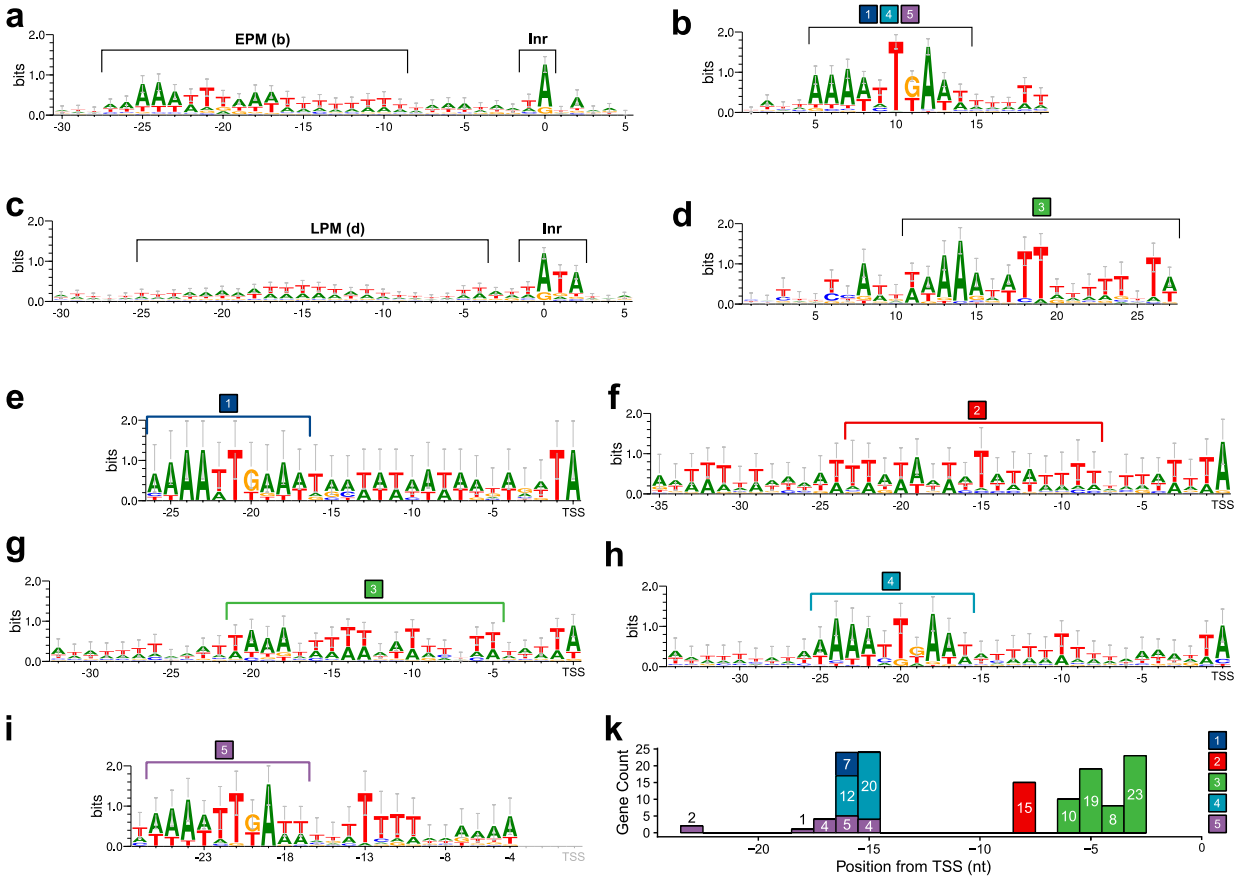



Figure 7. Promoter motifs and initiators detected in early and late ASFV GRG TSSs including alternative TSSs and those for nORFs. (a) Consensus of 30 bp upstream and 5 bp downstream of all 134 early TSSs including nORFs, with the conserved EPM (10) and Inr annotated. (b) 30 bp upstream and 5 bp downstream of all 234 late gene and nORFs TSSs, with the LPM and Inr annotated (c) The conserved EPM detected via MEME motif search of 35 bp upstream for 133 for 134 early TSSs (E-value: 3.1e-069). The conserved LPM detected via MEME motif search of 35 bp upstream for 46 for 234 late gene TSSs (E-value: 2.6e-003). The locations of the EPM shown in (b) and LPM shown in (d) are annotated with brackets in (a)

and (b), respectively. Motifs detected via MEME search of 35 bp upstream of genes in clusters from Figure 6: cluster 1 (7 genes, E-value: 9.1e-012), 2 (15 genes, E-value: 2.6e-048), 3 (60 genes, E-value: 1.0e-167), 4 (32 genes, E-value: 4.7e-105), 5 (16 genes, E-value: 5.7e-036), are shown in e-i, respectively. For ease of comparison, (e), (g), (i) and (f), (h) are aligned at TSS position. All motifs were generated using Weblogo 3 (99). (k) shows the distribution of MEME motif-end distances, from last nt (in coloured bracket), to their respective downstream TSSs.

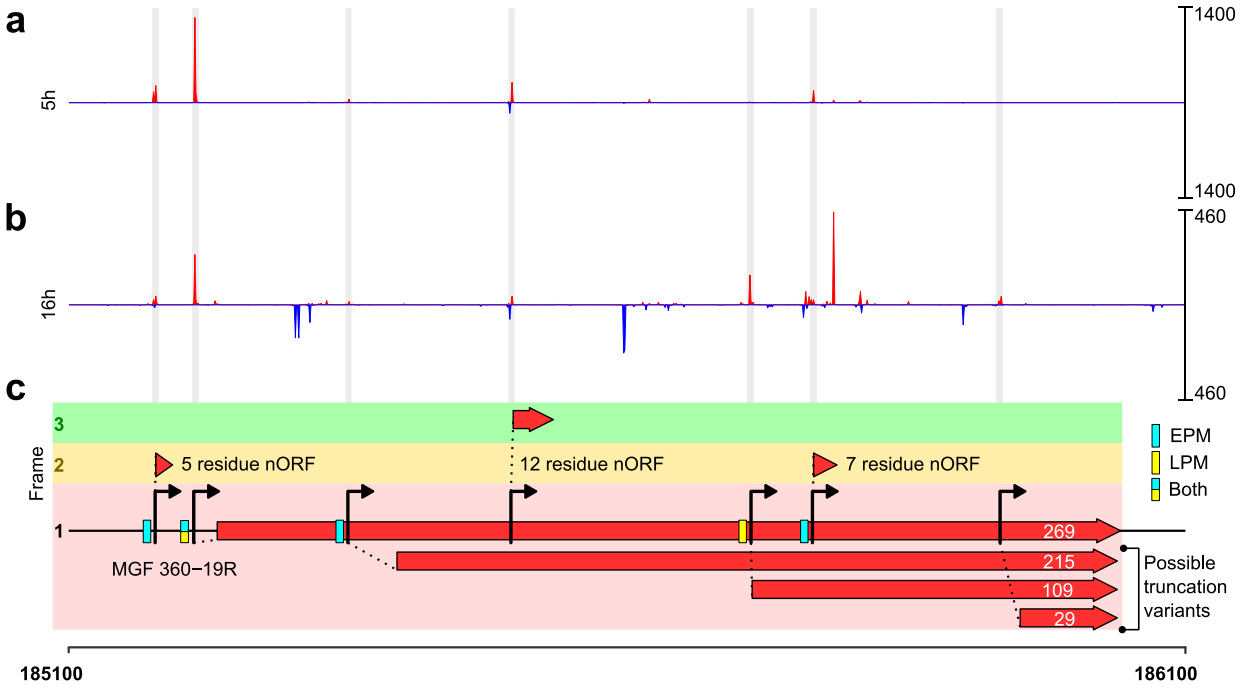
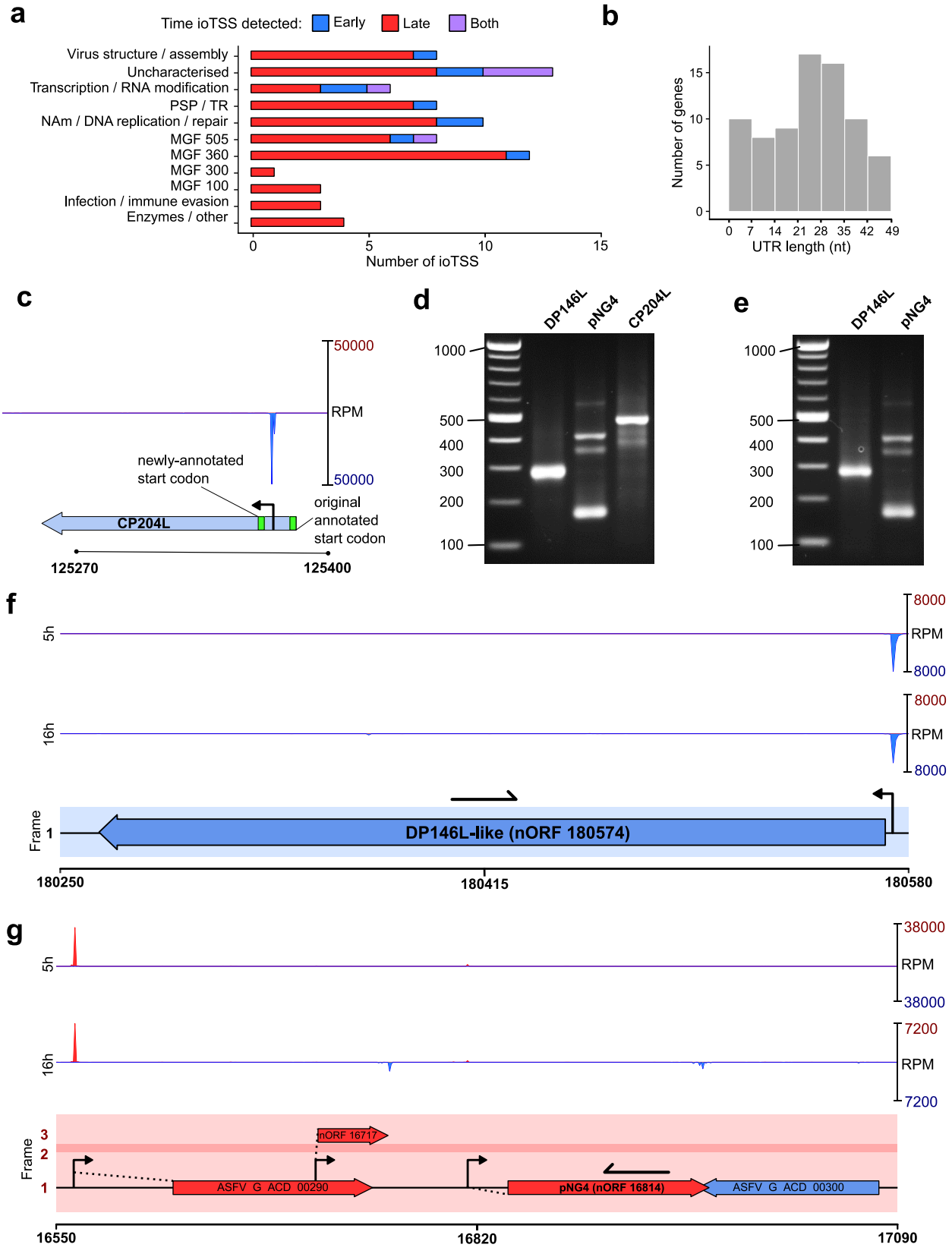



Figure 8. The TSSs of MGF 360-19R. Panels (a) 5 hpi and (b) 16 hpi show CAGE-seq 5' end data from these time-points, in red are reads from the plus strand and blue from the minus strand, the RPM scales are on the right. (c) TSSs are annotated with arrows if they can generate a minimum of 5 residue-ORF downstream, and grey bars indicate where they are located on the CAGE-seq coverage in (a) and (b). ORFs identified downstream of TSSs are shown as red arrows (visualized with R package gggenes), including three short nORFs out of frame with MGF 360-19R. Also shown are three in-frame truncation variants, from TSSs detected inside the full-length MGF 360-19R 269-residue ORF, downstream of its pTSS at 185213. Blue or yellow boxes upstream of TSSs indicate whether the EPM or LPM (respectively) could be detected within 35 nt upstream of the TSS using FIMO searching (100).




1094

Figure 9. Summary of intra-ORF TSSs (ioTSSs) and nORFs detected in the GRG genome, further information in Supplementary Table 2. (a) Summarises the gene types in which ioTSSs were detected, showing an overrepresentation of MGFs, especially from families 360 and 505, furthermore, the majority of ioTSSs are detected at 16 hpi. (b) For ioTSSs in-frame with the original, summarised are the subsequent UTR lengths i.e. distance from TSS to next in frame ATG start codon, which could generate a truncation variant. (c) Example of a miss-annotation for CP204L, whereby the pTSS is downstream the predicted start codon. (d) and (e) show the results of 5'RACE for three genes (DP146L, pNG4, and CP204L, see methods for primers), at 5 hpi and 16 hpi, respectively. Examples of genome regions around DP146L (f) and pNG4 (g), wherein ioTSSs were detected with capacity for altering ORF length in subsequent transcripts, and therefore protein output. Primers used for 5'RACE for DP146L and pNG4 are represented as black arrows in (f) and (g), respectively.

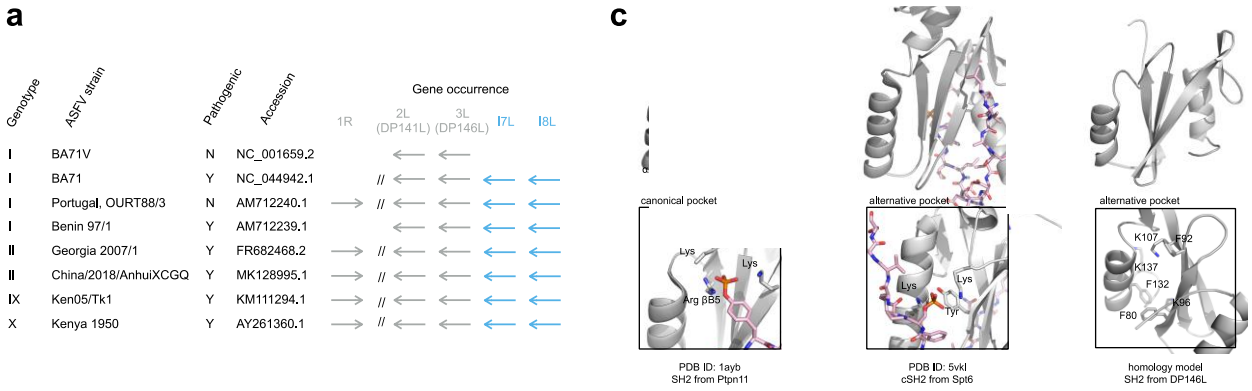



Figure 10. Function prediction of MGF 100 genes. (a) Occurrence of MGF 100 genes in genomes of selected ASFV strains. Genome of the China/2018 strain missed annotation of MGF 100-3L gene, which is located at positions 180315–180617 on the minus strand. (b) Structure-guided multiple sequence alignment of selected MGF 100 members and SH2 domains. Secondary structure for MGF 100 members was predicted with PSIPRED (springs, A-helices; arrows, B-strands). (c) Structures of SH2 domains (from right to left): canonical recognising pTyr (PDB ID 1ayb, mouse Ptpn11), atypical recognising pSer (PDB ID 5vkl, yeast

Spt6), and homology model of DP146L from ASFV strain BA71V (based on PDB IDs 4xey and 4fl3). Binding pockets are shown in the bottom panels.

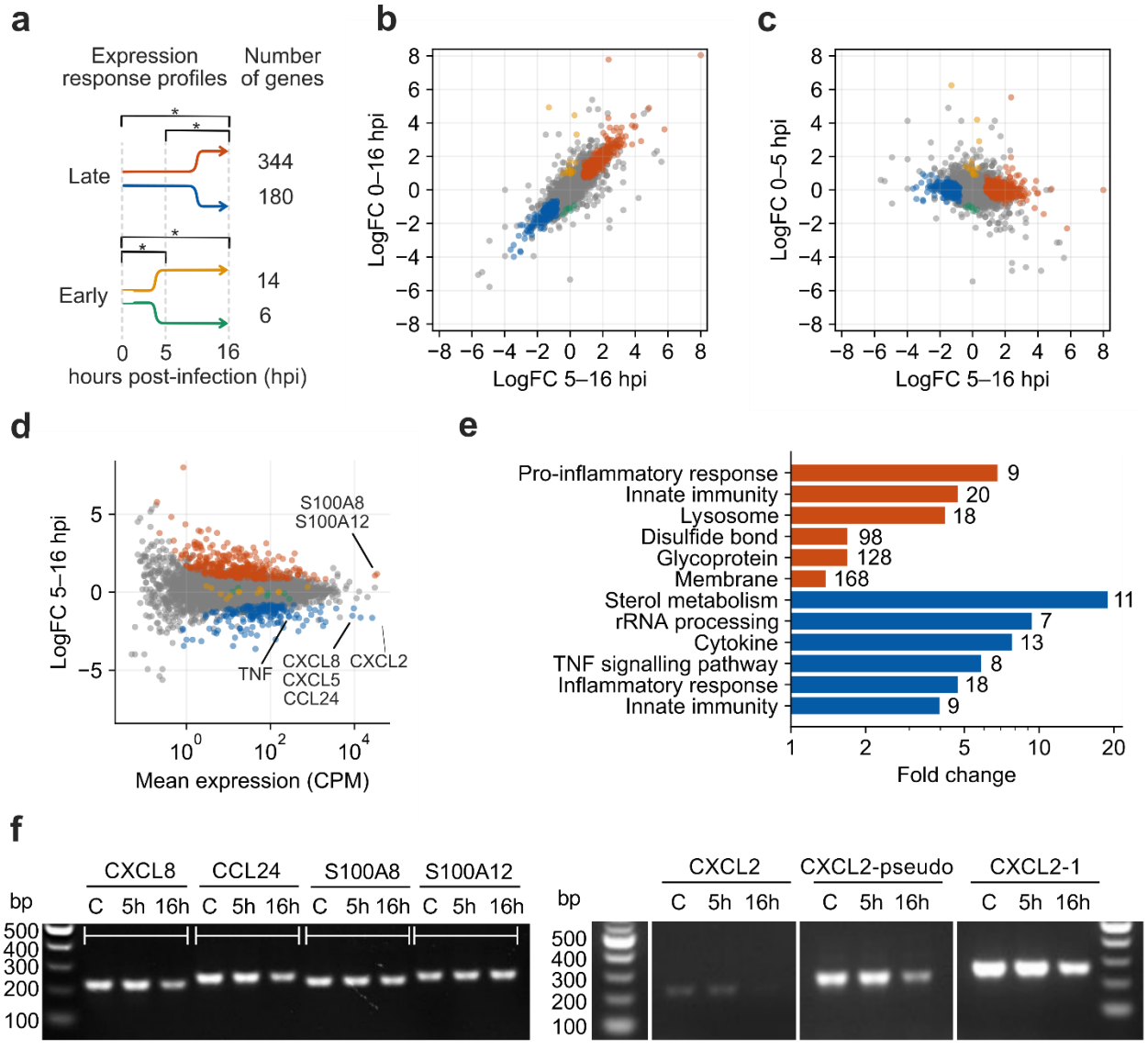
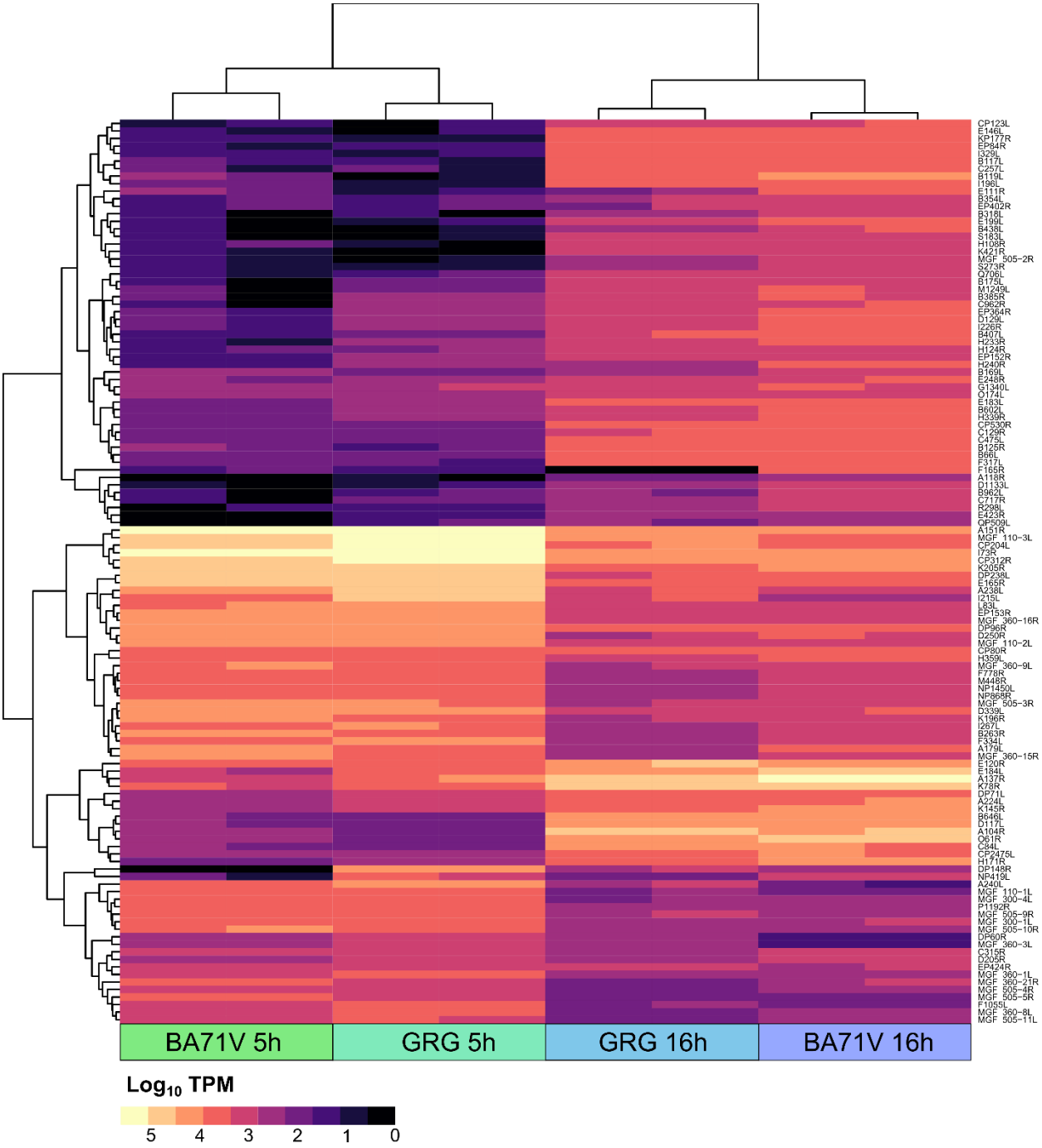



Figure 11. Changes in the swine macrophage transcriptome upon ASFV GRG infection. (a) Major expression response profiles of the pig macrophage transcriptome. Late response genes are significantly deregulated (false discovery rate < 0.05) in one direction both between 0 and 16 hpi as well as between 5 and 16 hpi, but not between 0 and 5 hpi. Early response genes are significantly deregulated in one direction both between 0 and 5 hpi as well as 0 and 16 hpi, but not between 5 and 16 hpi. (b) Relationship of log fold changes (logFC) of TSS-derived gene expression levels of the total 9,384 swine genes expressed in macrophages between 5–16 hpi and 0–16 hpi. Colors correspond to the response groups from the panel

a. (c) Relationship of log fold changes of TSS-derived gene expression levels of the total 9,384 swine genes expressed in macrophages between 5–16 hpi and 0–5 hpi. (d) MA plot of the TSS-derived gene expression levels between 5 and 16 hpi based on differential expression analysis with edgeR (94,101). (e) Representative overrepresented functional annotations of the upregulated (red) and downregulated (blue) macrophage genes following late transcription response (Benjamini-corrected p-value lower than 0.05). Numbers on the right to the bars indicate total number of genes from a given group annotated with a given annotation. (f) RT-PCR of four genes of interest indicated in (d). 'C' is the uninfected macrophage control, NTC is the Non Template Control for each PCR, excluding template DNA. See methods for primers used.

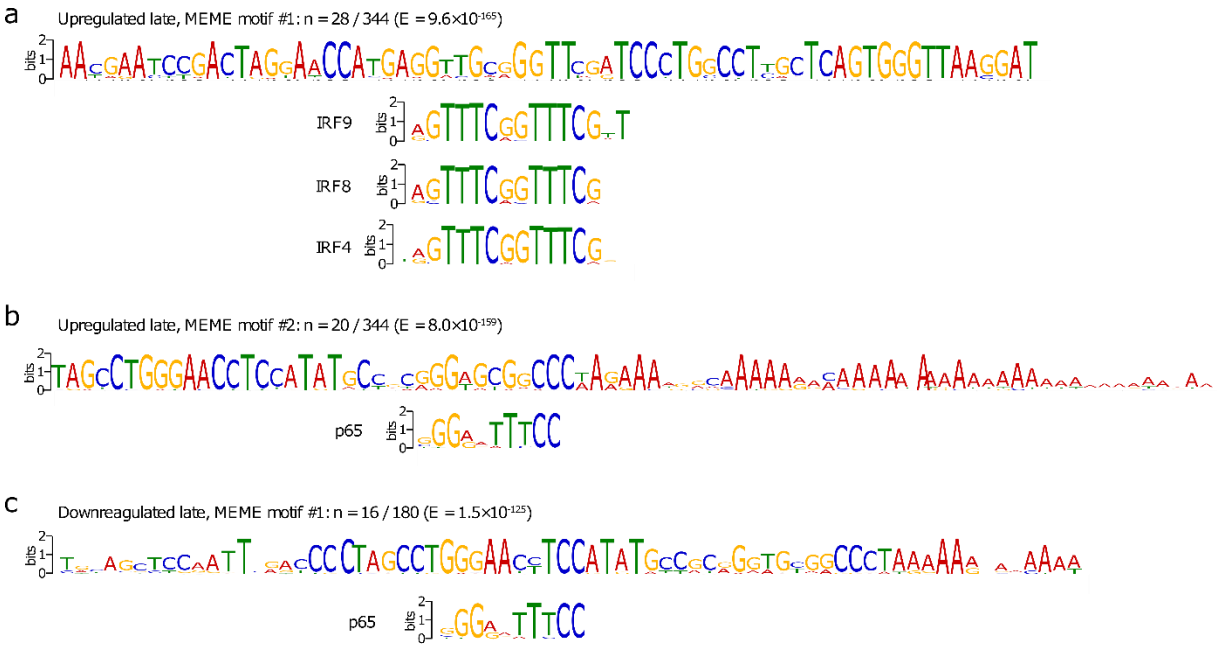
1132 # Supplementary Figures



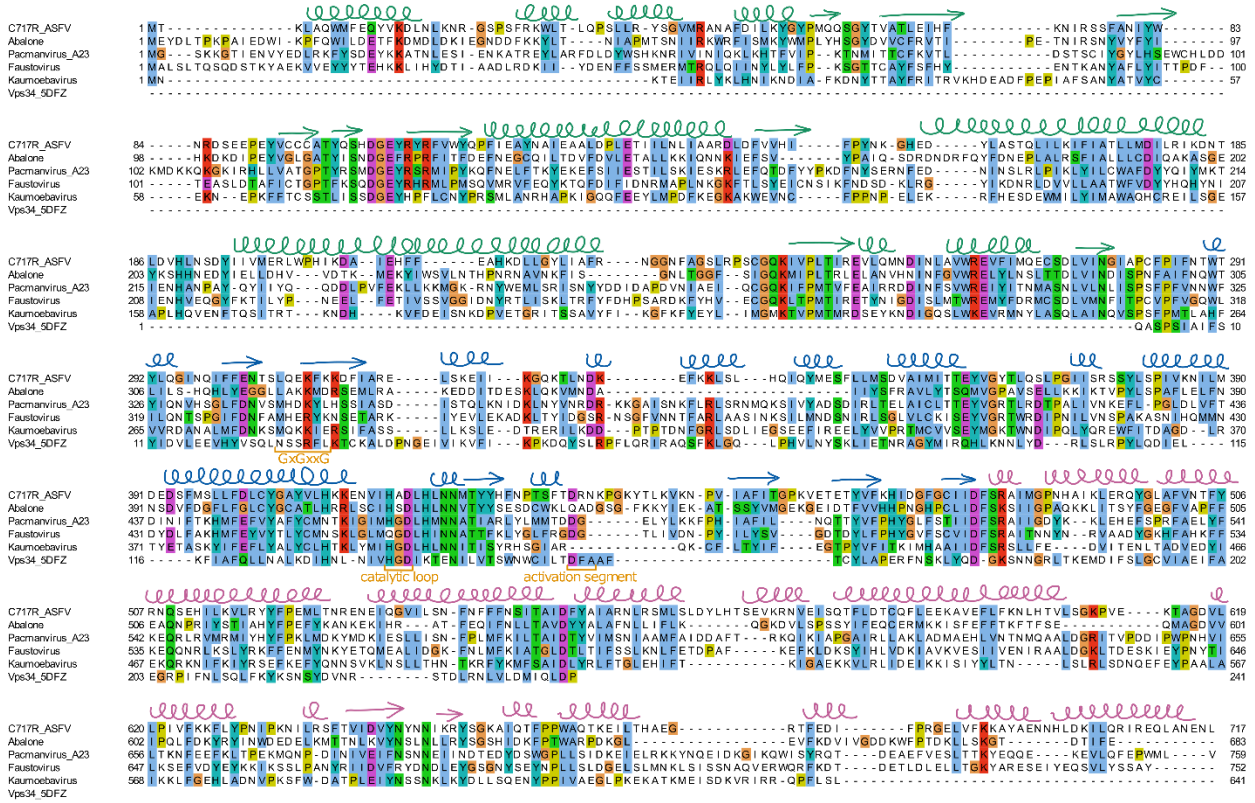

1133

1134 Supplementary Figure 1. Clustered heatmap of ASFV gene expression for the genes shared between
1135 BA71V and GRG that showed significant differential expression. R package 'pheatmap'-generated
1136 clustered heatmap of per-gene RPM values for genes shared between GRG and BA71V, across time-points

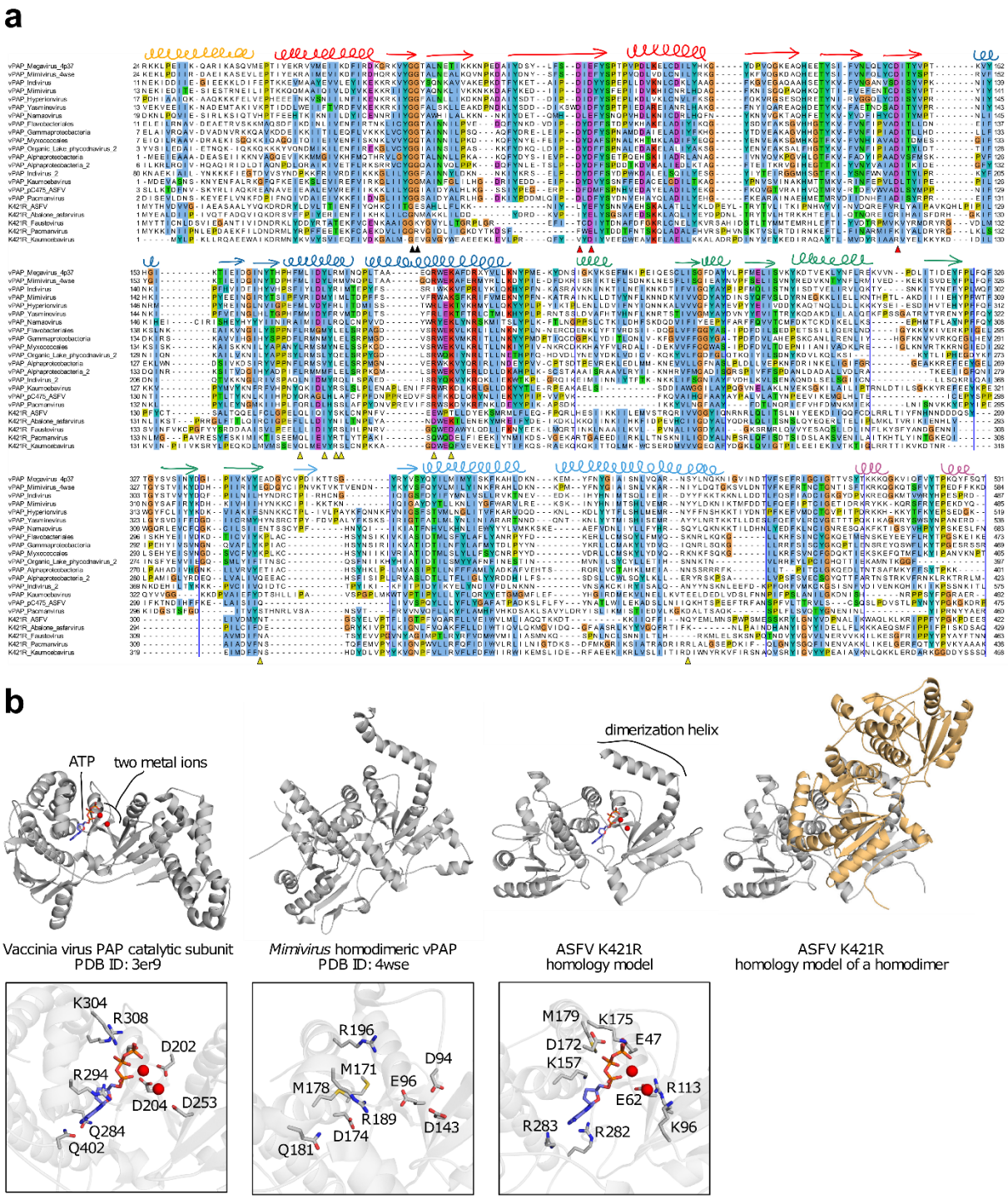
(5 hpi and 16 hpi), strains, with biological replicates as separate columns. Gene names for each row are listed on the right and both rows and columns were clustered according to Euclidean distance.




Supplementary Figure 2. Comparison of top-scored MEME motifs enriched in promoters of deregulated host genes to sequences recognised by human transcription factors: a) Motif found in 28 upregulated gene promoters similar to sequences recognised by human interferon response factors (JASPAR accessions MA0653.1: IRF9, MA0652.1: IRF8, MA1419.1: IRF4). b) Motif found in 20 upregulated gene promoters similar to sequences recognised by human p65/RELA protein (JASPAR accessions MA0107.1). c) Motif found in 16 downregulated gene promoters similar to sequences recognised by human p65/RELA protein (JASPAR accessions MA0107.1).




Supplementary Figure 3. Structure-guided multiple sequence alignment of selected C717R homologs and the kinase domain of Vps34 (PDB ID 5dfz). Secondary structure for C717R from African Swine Fever Virus, GRG (C717R-ASFV) was predicted with PSIPRED (springs, A-helices; arrows, B-strands). Colors of the secondary structure elements correspond to predicted domains: unknown N-terminal domain (green), kinase domain (blue), and alpha-helical C-terminal domain (pink). The C717R homologs were aligned with MAFFT and Vps34 was aligned based on HHpred mapping.




Supplementary Figure 4. Function prediction of K421R. A. Structure-guided multiple sequence alignment of selected K421R homologs and viral poly(A) polymerases (vPAP). Secondary structure was predicted with PSIPRED (springs, A-helices; arrows, B-strands). Colors of the secondary structure elements

correspond to predicted domains: dimerization domain (yellow), catalytic domain (red and blue), duplicated domain (green and blue), and C-terminal domain (purple). The ASFV-K421R sequence corresponds to K421R from GRG strain. Full set of sequences was aligned using MAFFT, following by manual corrections using known and predicted secondary structures. Red, black and, yellow triangles depict acidic residues from the described vPAP active site, the GG motif conserved in NTases, and residues usually coordinating ATP, respectively. Blue vertical lines denote insertions of non-conserved residues. B. Homology model of ASFV K421R compared to Vaccinia and *Mimivirus* vPAPs, shown in a cartoon representation (top panel) together with zoom-in views on the potential active sites (bottom panel). Five alternative models were generated with Modeller using templates of homodimeric *Megavirus chiliensis* and *Mimivirus* vPAPs (PDB IDs 4p37 and 4wse) with symmetric restraints, and the best one selected using the DOPE (Discrete Optimized Protein Energy) method.

## Supplementary Tables

Supplementary Table 1. (a) Summary of CAGE-seq reads mapping to the ASFV and *Sus scrofa* genomes using Bowtie2 and HISAT2, respectively (see methods). Bed file output of clusters detected from separate CAGEfightR analysis from CAGE-seq data at 5 hpi (b) and 16 hpi (c) post-infection. (d) Table summarizing pTSS locations in GFF format for either annotated genes or nORFs. TSSs for MGF 100-3L, MGF 110-11L, MGF 300-2R, E66L, C62L, and KP93L were not detected. (e) Results from DESEq2 analysis of CAGE-seq peaks at TSSs - both those for primary gene TSSs and those for novel ORFs. DESeq2 output is explained in Love et al. (41). The DESeq2 default adjusted p-value is a Wald test p-value with Benjamini–Hochberg correction (102). Per sample coverage is reported in transcripts per million mapped reads (TPM) in the final four columns. (f) Updated genome coordinates of annotated ORFs in the GRG genome, and for 5 hpi and 16 hpi each gene's stage-specific pTSS locations, with their nt distances relative to one another. The Untranslated Region or UTR length for each gene is shown (nt distance from each pTSS to the ORF start codon), and lastly whether the gene is differently expressed: early, late or not-classified (NC).

Supplementary Table 2. Summary of TSSs detected which potentially generate novel ORFs, unannotated in the GRG genome by identifying CAGEfightR-annotated TSSs with an unannotated ORF encoded downstream. (a) All putative unannotated ORF-generating TSSs. (b) The ioTSSs - overlapping in-frame with annotated ORFs, with potential for generating truncation variants. (c) TSSs with potential for encoding

ORFs downstream (down to 5 residues) which are not among the 189 ORFs annotated in the FR682468.1 genome.

Supplementary Table 3. Results from MEME motif (48) analysis of DNA sequences 35 bp upstream of conserved genes between ASFV and BA71V split into 5 clusters described in Figure 6.

Supplementary Table 4. List of CAGE-seq peaks (i.e., CAGE-derived transcription start sites, TSSs) found in *Sus scrofa* ASFV-infected macrophages. TSSs were mapped to the nearest *S. scrofa* 11.1 Ensembl (103) protein-coding genes using the RECLU pipeline (93). The read coverage was normalised to counts (reads) per million (CPM).

Supplementary Table 5. TSS-based transcript levels of *S. scrofa* macrophage-expressed genes. The read coverage was summed over all TSSs assigned to a given gene and normalised to counts (reads) per million (CPM). Significantly down- or upregulated genes were found with edgeR (94,101) at false discovery rate of 0.05.

Supplementary Table 6. Functional categories enriched in swine macrophage-expressed genes down- or upregulated at 16 hpi p.i. found with DAVID 6.8 Bioinformatics Resources (95), using best BLASTP (96,104) human hits. Supplementary Table 7. (a) ASFV and host genes whose expression level was assessed by RT-PCR. Primer pair sequences, PCR conditions and Accession Numbers of the sequences used for primer design are shown. (b) Primer details and qPCR conditions used for quantification of C315R gene expression. (c) Genes analyzed with 5'RACE and details of the primers used. The predicted locations of the genes on the GRG complete genome (Accession FR682468) as well as the expected length of upstream untranslated regions (5'UTR) according to the detected TSSs in CAGE-seq are shown. The final expected amplicon sizes in the last column are the result of PCR with GSP3 nested primer (GSP2 in pNG4) plus 36 bp added to the final sequence as a result of the 5'RACE polynucleotide tail addition and amplification with "Abridged Anchor Primer" (AAP). GSP1 primer from pNG4 was extended with 50 non-annealing nucleotides (small caps italic) to increase the cDNA length.