Defining the Transcriptional Landscape during Cytomegalovirus Latency with Single-Cell RNA Sequencing

Defining the Transcriptional Landscape during Cytomegalovirus Latency with Single-Cell RNA Sequencing

Miri Shnayder[1,4], Aharon Nachshon[1,4], Benjamin Krishna[3], Emma Poole[3], Alina Boshkov[1], Amit Binyamin[1], Itay Maza[2], John Sinclair[3], Michal Schwartz[1*] and Noam Stern-Ginossar[1*]

[1] Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

[2]Department of Gastroenterology, Rambam Health Care Campus & Bruce Rappaport School of Medicine, Technion- Institute of Technology, Haifa, Israel.

[3]Department of Medicine, Addenbrooke's Hospital, University of Cambridge, Cambridge CB20QQ, UK

[4] These authors contributed equally to this work

* To whom correspondence should be addressed; michalsc@weizmann.ac.il or noam.stern-ginossar@weizmann.ac.il

Abstract

Primary infection with human cytomegalovirus (HCMV) results in a lifelong infection due to its ability to establish latent infection, with one characterized viral reservoir being hematopoietic cells. Although reactivation from latency causes serious disease in immunocompromised individuals, our molecular understanding of latency is limited. Here, we delineate viral gene expression during natural HCMV persistent infection by analyzing the massive RNA-seq atlas generated by the Genotype-Tissue Expression (GTEx) project. This systematic analysis reveals that HCMV persistence *in-vivo* is prevalent in diverse tissues. Notably, we find only viral transcripts that resemble gene expression during various stages of lytic infection with no evidence of any highly restricted latency-associated viral gene expression program. To further define the transcriptional landscape during HCMV latent infection, we also used single cell RNA-seq and a tractable experimental latency model. In contrast to some current views on latency, we also find no evidence for any highly restricted latency-associated viral gene expression program. Instead, we reveal that latency-associated gene expression largely mirrors a late lytic viral program albeit at much lower levels of expression. Overall, our work has the potential to revolutionize our understanding of HCMV persistence and suggests that latency is governed mainly by quantitative changes, with a limited number of qualitative changes, in viral gene expression.

Importance

The human cytomegalovirus is a prevalent pathogen, infecting most of the population worldwide and establishing lifelong latency in its hosts. Although reactivation from latency causes significant morbidity and mortality in immunocompromised hosts, our molecular understanding of the latent state remains limited. Here we examine the viral gene expression during natural and experimental latent HCMV infection on a transcriptome wide level. In contrast to the classical views on herpesvirus latency, we find no evidence for a restricted latency-associated viral gene expression program. Instead, we reveal that latency- gene expression largely resembles a late lytic viral profile albeit at exceedingly lower levels of expression. Taken together, our data transform the current view of HCMV persistence and suggest that latency is mainly governed by quantitative rather than qualitative changes in viral gene expression.

Introduction

Human cytomegalovirus (HCMV) is a ubiquitous pathogen that, like all herpes viruses, can establish latent infection that persists for the lifetime of the host. In healthy individuals infection rarely causes any significant clinical symptoms due to a robust immune response (1, 2). In contrast, primary infection or reactivation from latency can result in serious and often life-threatening disease in immunocompromised individuals (3–5). Latent infection is, therefore, a key part of viral persistence and latently infected cells are a clear threat when the immune system is suppressed. Despite this, our molecular understanding of HCMV latency state is still limited.

HCMV is tightly restricted to humans, however in its host it has extremely wide cell tropism (6), and many kinds of cells can be productively infected, including fibroblasts, epithelial cells and smooth muscle cells (7). In contrast, latent infection was so far characterized only in cells of the early myeloid lineage, including CD34+ hematopoietic progenitor cells (HPCs) and CD14+ monocytes (8). It was further established that terminal differentiation of HPCs and CD14+ monocytes to dendritic cells or macrophages triggers virus reactivation from latency (9–13). This differentiation-dependent reactivation of latent virus is thought to be mediated by changes in post-translational modification of histones around the viral major immediate-early promoter (MIEP)(11, 14–17). These modifications drive the viral major immediate-early (IE) gene expression, resulting in reactivation of the full viral lytic gene program cascade and the

production of infectious virions (11). Thus, the cellular environment is a key factor in determining the outcome of HCMV infection.

During productive lytic infection, HCMV expresses hundreds of different transcripts and viral gene expression is divided into three waves of expression IE, early, and late (6, 18, 19). The maintenance of viral genome in latently infected cells, is thought to be associated with expression of a much smaller number of viral genes relative to lytic infection (20–25) in the general absence of IE gene expression. Due to their therapeutic potential, significant attention has been drawn to a few latency- associated viral gene products, but the possibility that additional viral transcripts contribute to latency regulation remains unclear.

The earliest studies that looked for latency-associated gene expression identified a number of transcripts arising from the MIEP region of HCMV but no function was assigned to them (26–28). More systematic mapping of latency-associated transcripts was conducted with the emergence of microarray technology. Two studies detected a number of viral transcripts in experimentally latently infected myeloid progenitor cells (29, 30). The latent transcripts reported by these studies were not entirely overlapping, yet these findings were used as a guideline for targeted efforts to identify latent gene products. Interrogating the viral transcriptome in natural persistent infection is highly challenging since viral genomes are maintained in extremely few cells, at very low copy numbers, and viral genes are expected to be expressed in low levels. Nevertheless, subsequent work detected a number of these transcripts during natural latency (21, 22,

25), mainly using high sensitivity approaches such as nested PCR, building a short list of viral genes that is generally accepted to represent a distinct transcriptional profile during latent infection. These genes include UL138, UL81-82ast (LUNA), US28, as well as a splice variant of UL111A, which encodes a viral interleukin 10 (31–37).

More recently, RNA-seq was applied to map latency associated viral transcripts (38). This study revealed a wider viral gene expression profile that included two long non-coding RNAs (lncRNAs), RNA4.9 and RNA2.7 as well as the mRNAs encoding replication factors UL84 and UL44 (38). In a recent study a targeted enrichment platform was applied to study the transcriptome of HCMV latent infection in both experimental and natural samples revealing even a broader gene expression profile (39).

Such genome-wide analyses are highly informative as they measure the expression of all transcripts in an unbiased manner. However, a major limitation is that they portray a mean expression in cell population, without reflecting intra-population heterogeneity. In the case of latent HCMV infection models, this can be highly misleading since it is hard to exclude the possibility that a small, undesired population of cells, is undergoing lytic replication and thus can easily introduce “lytic noise”. This effect can be especially significant for viral genes that are highly expressed during lytic infection such as lncRNAs (19). Finally, the low frequency of natural latent cells is a major hurdle for global quantitative analysis of naturally latently infected cells.

To overcome the problem of scarcity of natural latent cells, we took advantage of the massive human RNA-seq atlas generated by the Genotype-Tissue Expression (GTEx)

consortium (40). Through analysis of 435 billion RNA reads, we did not find any evidence for a restricted latency associated viral gene program. Instead, in several tissues we captured low-level expression of viral transcripts that resembles gene expression at late stages of lytic infection. Next, to directly explore viral gene expression in a controlled latently infected cell population we turned to the established myeloid lineage experimental systems. By using single cell RNA-seq (scRNA-seq) we unbiasedly characterize the HCMV latency program of both experimentally latently infected CD14+ monocytes and CD34+ HPCs, overcoming the impediment of cell population variability. Surprisingly, in contrast to the existing view in the field, we find no strong evidence for a specific latency associated viral gene expression signature of specific viral genes. Instead, we reveal that in HCMV latency models, whilst there is little detectable IE expression, there is low-level expression of viral genes that largely resembles the late lytic stage viral gene expression profile. Our analyses thus redefine HCMV latent gene expression program and suggest mainly quantitative rather than qualitative changes that help determine latency. Our work illustrates how new genomic technologies can be leveraged to reevaluate complex host-pathogen interactions.

Results

No evidence for a restricted latency-associated viral gene expression program in natural HCMV infection

The proportion of infected mononuclear cells in seropositive individuals was estimated at 1:10,000-25,000 with a copy number of 2-13 genomes per infected cell (41). Given that transcription of viral genes is expected to be low in these cells, immense amount of sequencing data is required to capture viral transcripts. We thus took advantage of the Genotype-Tissue Expression (GTEx) database, a comprehensive atlas containing massive RNA-seq data across human tissues that were obtained postmortem, from otherwise healthy individuals (40). We analyzed HCMV reads in 9,416 RNA-seq samples from 549 individuals covering 31 tissues and containing more than 433 billion reads (Fig. S1A, B). In 40 samples we obtained only reads that aligned to a 229bp region in the IE promoter (Fig. S1C). Since the sequence in these reads matches the sequence of the HCMV promoter commonly used in vectors rather than the sequence observed in the majority of clinical samples (Fig. S1D), we concluded these reads may originate from a contamination and excluded them from further analysis.

Reassuringly, the number of samples that contained HCMV reads and the number of HCMV reads, were significantly higher in samples originating from seropositive individuals (Fig. 1A, Pval=0.0467 and Pval < 10^{-55} respectively, hypergeometric test). HCMV reads were found in 6 out of 2,210 seronegative samples, however all of them

contained only one viral read per sample. Therefore, this was used as a threshold and viral reads from samples containing less than two viral reads were filtered out in further analysis (data from all samples is summarized in Table S1A).

HCMV genomes have been detected in HPCs and in additional cells throughout the myeloid lineage (42, 43). Consequently, the blood and the hematopoietic system are a major focus in research of HCMV persistence. Analysis of the GTEx database provides an exceptional opportunity to unbiasedly assess HCMV prevalence in various tissues. Interestingly analysis of the abundance of HCMV reads in different tissues revealed that ovaries, blood, adipose tissue and lung had the highest percentage of samples containing viral reads (Fig. S1E) as well the highest normalized number of viral reads (Fig. 1B). Since the GTEx database did not contain RNA-seq data from bone marrow where HPCs reside, we performed RNA-seq on two HPC samples from HCMV positive individuals and surveyed additional 25 RNA-seq samples of HPCs from healthy individuals (Table S1D). Although we analyzed over 1.5 billion aligned RNA-seq reads we did not detect any viral reads in these samples (Fig. 1B).

Next, we analyzed the viral gene expression as reflected by the HCMV reads we identified in natural samples, including in this analysis only samples that contained more than 4 HCMV reads. Hierarchical clustering revealed that the samples could be subdivided into two groups based on the pattern of viral gene expression (Fig. 1C).

172 The first group (group I) was composed of samples that were dominated by transcripts
173 that are the most highly expressed during late stage of lytic infection, e.g. RNA2.7,
174 RNA4.9, RNA1.2 and UL22A (Fig. 1C and S1F). Indeed, when we compared the viral gene
175 expression of these samples to RNA-seq data we collected along lytic infection of
176 fibroblasts, we obtained high correlation to late stages of infection (R=0.97, Fig. 1D and
177 Fig. S2A). This correlation suggests these viral reads that were identified in natural
178 settings resemble late stage lytic gene expression program.
179 The second group (group II) is composed of samples that express *bona fide* immediate
180 early genes, e.g. UL123, US3 and UL36 as well as US33A which is the most highly
181 expressed transcript early in infection (18), and importantly had very limited expression
182 of transcripts that are abundant at late stage of lytic infection (Fig. 1C and FigS2B).
183 Therefore, we speculate these samples may reflect the onset of viral reactivation, a
184 state in which IE genes are transcribed but the full viral gene program is still suppressed.
185 Supporting this notion, viral gene expression of these samples correlated best with
186 lytically infected fibroblasts at 5 hours post infection (hpi) (R=0.55) (Fig. 1D and Fig.
187 S2B). This IE expression positive state may represent cells exiting from latency,
188 consistent with the view that reactivation goes through a stage of IE gene activation.
189 Since the tissues we analyzed were obtained postmortem, it is possible that
190 postmortem-related physiological events led to HCMV reactivation and IE gene
191 expression. To assess this hypothesis we inspected the time postmortem at which the
192 tissue was collected (data is provided by GTEx (40)). Samples in group II were not
193 enriched for long waiting time before tissue collection or any other clinical technical

details (Fig. S2C and tables S1B and S1C). In addition, there were no differences in the time interval of tissue collection between samples that contained HCMV reads and those that did not (Fig. S2D). These results suggest that the HCMV gene expression pattern we captured is likely independent of the trauma that occurred after death. Importantly, although we were able to identify HCMV transcripts, we were not able to identify tissue or blood samples that provide evidence for any highly restricted latency-associated viral gene expression program that differs from lytic viral gene expression. Since viral gene expression is expected to be very low in latent cells, a possible explanation for this is that a non-targeted sequencing approach may not detect these rare transcripts despite great sequencing depth.

Single cell transcriptomic analysis of latently infected CD14+ monocytes

Although in natural samples we detected only low level viral gene expression pattern that resembles the lytic gene expression program, the cellular heterogeneity in these samples does not allow us to distinguish whether we are analyzing latently infected cells, or rare cells in which productive infection is taking place. Consequently, we next moved to characterize the viral transcriptome in experimental models of HCMV latency. Since these models rely on primary hematopoietic cells that may vary in their differentiation state and may also contain heterogeneous populations, we took advantage of the emergence of single cell RNA-seq (scRNA-seq) technologies (44, 45). This high-resolution profiling of single cell transcriptomes allowed us to delineate the

nature of HCMV latency program in the best studied latent reservoir, hematopoietic cells.

Freshly isolated CD14+ human monocytes were infected with an HCMV TB40E strain containing an SV40 promoter driven GFP (TB40E-GFP) (46). This strain allows short-term detection of GFP-tagged latently infected cells, as in these cells GFP expression is efficiently detected at 2 dpi and then GFP signal gradually declines. Despite GFP levels in monocytes being much lower compared to those in lytic infection, the GFP expression allowed us to confirm that the majority of cells were indeed infected (Fig. S3A). To validate latent infection in our experimental settings, we analyzed by quantitative real-time PCR (qRT-PCR) the gene expression pattern of the well-studied latency associated gene, UL138 and of the immediate early gene, IE1 at 4 days post infection (dpi). Infected monocytes expressed relatively high level of UL138 while showing only trace level of IE1 transcript (Fig. 2A), thus manifesting the hallmark of latent infection (29, 31, 32, 37, 47, 48). Differentiation of these infected monocytes into dendritic cells resulted in detectable IE expression as well as production of infectious virions (Fig. 2B and Fig. 2C), thus demonstrating that our CD14+ cells are latently infected.

Next, HCMV infected CD14+ cells were single cell sorted without further selection at 3,4, 5, 6, 7 and 14 dpi and their transcript levels were measured using massively parallel 3' scRNA-seq (MARS-seq) (49). Analysis of the entire transcriptome was performed on 3,655 CD14+ infected cells in which we could detect 15,812 genes out of which 171

were HCMV transcription units (see material and methods and Fig. S3B for distribution of reads and genes over the cell population). Projection of the cells using t-distributed stochastic neighbor embedding (t-SNE) analysis revealed that most of the cells constitute a large heterogeneous but continuous population and only a small group forms a distinct population (Fig. 3A). When we calculated the percentage of reads that align to the HCMV genome in each of the cells, it became evident that the viral transcripts constitute >10% of the total reads in the small distinct population (Fig. 3A). Reassuringly, when performing the t-SNE analysis by using only cellular gene expression, we obtained the same structure, confirming we are looking at two different cell states (Fig. S4A). The small population likely represents a lytic infection state, and the rest of the monocytes, which are the vast majority, exhibit very low to undetectable, diverse viral gene expression levels, indicating that they likely represent latently infected cells. This distribution, showing a clear separation between two groups of cells exhibiting very different levels of viral gene expression, confirms the purity of the single-cell isolation and the dominance of latent cells in the population of CD14+ infected cells (Fig. 3A).

HCMV latency associated gene expression in CD14+ monocytes and CD34+ HPCs resembles late lytic gene expression program

To assess the heterogeneity in HCMV latently infected monocytes, we combined the data from all 3,655 cells and clustered them on the basis of their host and viral gene expression profiles into 6 clusters (Fig. 3B) (clustering method was previously described (50)). Notably, also in this approach, the cells exhibiting high viral expression levels,

representing lytic infection state, were clustered together and the most differential genes that were highly expressed in this cluster were almost exclusively viral genes (cluster 1, Fig. 3B, top panel). On the other hand, the rest of the cells exhibited very low levels of viral gene expression in varying degrees and the highly expressed differential genes in these five clusters were all cellular genes (Fig. 3B, lower panel and Table S2A).

These clusters were consistent with the t-SNE analysis, with cluster 1 overlapping with the distinct population probably representing lytic infection state (Fig. S4B). Indeed, by comparing the viral gene expression pattern of cells from this cluster to lytically infected monocyte-derived macrophages or fibroblasts we could confirm that they exhibit comparable programs (Fig. S4C). Unexpectedly, although the lytic and latent cells represent two very separable cell states (Fig. S4A), latent cells from all clusters, show viral gene expression profile that to large extent resembles the late lytic expression profile (cluster 1), with the dominant difference being the *level* of viral gene expression but not in the *identity* of the viral genes (Fig. 4A). The only viral genes that their deviation from this correlation was statistically significant, and were relatively higher in latent cells, were the exogenous GFP (False discovery Rate (FDR)=7.10^{-19}) which is driven by the strong SV40 promoter, the lncRNA, RNA2.7 (FDR<10^{-100}), which is the most abundant transcript, and a transcript encoding for UL30 (FDR =6.10^{-8}), a poorly characterized coding gene (19) (Table S2B).

We also examined whether viral gene expression program varies between the different populations of latently infected cells defined by the different clusters, by assessing the

correlation between lytic cells (cluster 1) and each of the five other clusters. We found that viral gene expression profiles of all clusters were correlated to some extent with the lytic cells (cluster 1) (Fig. S4D). The correlation coefficient declined with the reduction in number of viral reads, as expected, however throughout the different clusters only very few viral genes were significantly higher in latent cells composing these clusters (table S2C).

Interestingly, the continuous decline in viral gene expression appears tightly related to the time along infection and is also reflected in the separation into different clusters (Fig 3B and Fig S5). This gradual repression suggests progressive silencing of viral gene expression along latent infection as has been previously demonstrated (29, 30).

Importantly, by calculating the background noise in the single cell data (materials and methods), we confirmed that the results are not skewed by possible cross contamination in the single-cell data from the few lytic cells we have in our experiments (Fig. S6).

Overall, this analysis indicates that to large extent the viral gene expression program during experimental latency mirrors the viral gene expression program in late stage of lytic infection, albeit expressed at much lower levels.

It is noteworthy that these unexpected results do not contradict previous analyses of latent cells, as we observe latent infection to be associated with overall low levels of viral gene expression and with high levels of UL138 relative to IE1. Importantly, this high

UL138/IE1 ratio is also evident at late stages but not at early stages of lytic infection (Fig. 4B).

It was previously demonstrated that HCMV virions contain virus-encoded mRNAs (51, 52). To exclude the possibility that the transcripts we capture originate from input mRNAs that are carried in by virions, we infected CD14+ monocytes with untreated or UV-inactivated viruses and evaluated the levels of RNA2.7 and RNA4.9 at 5 dpi. The expression of both transcripts was over 30-fold lower in the cells infected with UV-inactivated virus compared to cells infected with untreated virus (Fig. 4C). In addition, viral transcripts levels at 5 hpi were much lower than at 5dpi (Fig. 4D), illustrating that the viral transcripts that we capture during latency result from de novo expression and are not the result of input mRNAs.

We next examined viral gene expression in experimentally infected CD34+ HPCs, which are another well-characterized site of latent HCMV infection (43, 53). CD34+ cells were infected with TB40E-GFP virus in the same manner as CD14+ monocytes, and used for generation of scRNA-libraries at 4 dpi. We initially used MARS-seq (49) to measure the transcriptome of infected HPCs, however in CD34+ viral gene expression was significantly lower and out of 424 cells we sequenced, viral transcripts could be detected in only 12 cells (Table S2E). We therefore moved to 10X genomics drop-seq platform that allows simultaneous analysis of thousands of cells. We analyzed the transcriptome of 7,634 experimentally infected HPCs in 366 of which we identified viral transcripts (see material and methods and Fig. S3C for distribution of reads and genes over the cell

321 population). Projection of cells using t-SNE analysis revealed heterogeneous populations
322 and cells that expressed viral transcripts were distributed throughout these populations
323 (Fig. 5A). Analysis of the 366 cells that expressed viral transcripts revealed low
324 expression levels and, as in CD14+ monocytes, the low viral gene expression we
325 measured in these cells correlated with the expression pattern of late stage of lytic
326 infection (comparing CD34 to cells to cluster 1, Fig. 5B). Also here, only for few
327 transcripts the deviation from this correlation was statistically significant, these included
328 RNA2.7 and UL30 (Table S2E).
329
330 A recent transcriptome mapping done on experimentally infected CD34+ revealed a
331 broader profile of gene expression than was previously appreciated (39). Importantly,
332 comparison of the viral expression profile using this independent dataset to expression
333 profile of late lytic fibroblasts from the same study also revealed significant correlation
334 (R=0.91 and R=0.89, Figure S7). Over all, our results and analysis show that during
335 experimental latent infection there is no well-defined latency associated viral gene
336 expression signature, but rather these cells are characterized by gradual repression of
337 viral gene expression with low level expression of a program largely resembling late lytic
338 infection stages.

Discussion

Despite the clinical importance of HCMV latency, the mechanisms involved in viral genome maintenance and reactivation are poorly understood. An important step in deciphering these mechanisms is to characterize viral transcripts that are expressed during latent infection in an unambiguous manner. To address this challenge we examined HCMV infection by comprehensive analysis of RNA-seq data from diverse human tissues and further used scRNA-seq to analyze gene expression of latently infected CD14+ monocytes and CD34+ HPCs. Surprisingly, our measurements demonstrate that in both natural HCMV infection and in experimental latency models there is no evidence of a unique latency-associated gene expression program but instead we describe viral gene expression pattern that is largely similar to late stage of lytic infection at exceedingly low levels. Although these results are surprising given the prevalent notion that HCMV latency involves a restricted gene expression program, evidence for broader viral gene expression was indicated in several previous genome-wide studies (29, 30, 38, 39).

Examination of HCMV infection by analyzing viral gene expression in diverse human tissues uncovered two patterns of gene expression; the first is composed of samples that contain viral transcripts that are abundant at late stage of lytic infection and the second is composed of samples with a restrictive gene expression pattern that includes mainly IE transcripts. The samples that contain late viral transcripts could reflect low-level expression that originates from few latent cells or the existence of scarce lytic cells

361 in these tissues. Since cells expressing viral transcript are very rare, it is currently
362 impossible to distinguish between these two scenarios.
363
364 The samples that contained mainly IE transcripts are interesting as they may reflect a
365 snapshot of viral gene expression during reactivation *in-vivo*, in natural human samples.
366 Although we did not observe any difference in the time interval from death until these
367 samples were collected, it remains possible that this restricted IE gene expression
368 occurred postmortem or due to the associated trauma (54). Regardless of the conditions
369 that initiated this restrictive IE gene expression, this state may imply that *in-vivo* exit
370 from latency goes through a phase in which IE genes are activated. The IE expression
371 pattern we find was seen mostly in blood samples but not solely. While speculative, the
372 restrictive IE gene expression in these cells may suggest that there is a threshold that
373 needs to be crossed (perhaps the accumulation of enough IE proteins) before
374 temporally controlled viral gene expression program can start. Indeed, this idea is
375 entirely consistent with differentiation of CD34+ cells *ex-vivo* to immature DCs resulting
376 in cells permissive for IE1 expression but not virus production (11) and with the
377 detection of IE1 expression without infectious virus production in immature DCs isolated
378 from healthy seropositive carriers (55). A similar model was proposed for HSV-1
379 reactivation from latency, where accumulation and localization of VP16 was suggested
380 to regulate the onset of full reactivation program (56).
381

382 Our analysis of natural samples also suggests that HCMV persistence is widespread
383 throughout the body, as we found viral gene expression in diverse human tissues.
384 Previous studies have shown the presence of viral genomes in tissues outside the blood
385 and hematopoietic system (57–60). Our data provide some evidence for viral gene
386 expression in various tissues. The tissue in which we found the highest levels of viral
387 transcripts was the lung which is consistent with recent results showing that HCMV DNA
388 could be identified in the lung (60), and in alveolar macrophages (9) and that HCMV
389 reactivation is often manifested clinically as pneumonitis (61, 62). The cellular
390 heterogeneity in tissue samples precludes any conclusion about the cellular sites of
391 HCMV infection in these natural samples.
392
393 Our inability to detect a restricted latency-associated gene expression program in this
394 systematic survey of natural samples motivated us to examine the viral gene expression
395 in the best studied latency experimental systems using single cell analysis. Notably, our
396 results challenge the view of latency as being a specific virally restricted program, and
397 highlight rather a quantitative aspect of viral gene expression that is likely governed by
398 the host cell. Unbiased transcriptome analyses of HPCs and monocytes latently infected
399 with HCMV either experimentally or naturally have been previously performed using
400 both microarrays and next generation sequencing (29, 35, 38). The list of expressed
401 genes emerging from these different studies included dozens of viral transcripts. The
402 recent study by Cheng et al (39) revealed an even broader profile of gene expression
403 during hematopoietic cell infection. By using recombinant viruses that establish a latent

404 or a replicative infection in HPCs, this study identified a class of low expressed genes
405 that are differentially expressed in latent vs. replicative states of infection and suggested
406 these genes may have a role in regulating latency. Our analysis of this dataset further
407 reveals a significant correlation between viral gene expression in latent HPCs and viral
408 gene expression in late lytic fibroblasts. This correlation provides an important
409 independent validation to our finding that viral gene expression during latency to large
410 extent resembles the program seen during late stage lytic infection.
411
412 The significant advantage of scRNA-seq, especially in the case of viral infection, is that
413 we can unbiasedly determine the existence of different cell populations and exclude the
414 possibility that the expression profile is skewed by a small group of cells. Importantly,
415 the clustering approach used in this study allows us to validate that the viral gene
416 expression profile is not related to viral expression levels. Although the correlation
417 coefficient is declining with the reduction in number of viral reads, the decline in viral
418 gene expression level is progressive and suggests continuous repression of viral gene
419 expression along latent infection. Thus we see expression profiles that correlate with
420 late stages of lytic infection even in the clusters that have almost undetectable levels of
421 viral gene expression.
422
423 At the present sampling depth and coverage efficiency, our analysis of CD14+ cells can
424 detect subpopulations of 0.3% (11–12 cells) or higher. Therefore, although we cannot
425 exclude the possibility that a very small population of cells are in a different state and

426 will harbor a different, more restricted, viral gene expression program, if such cells exist
427 they would be rare.
428
429 Our analyses reveal differences in cellular gene expression that are associated with
430 differences in the levels of viral gene expression. These differences could stem from
431 variation in cell maturation state that restricts viral gene expression or alternatively they
432 could reflect virally-induced changes in the host environment. Future work will help to
433 distinguish between these two options.
434
435 The results we obtained for both CD14+ and CD34+ progenitors were qualitatively
436 similar, however the relative levels of viral transcripts in CD34+ progenitors were
437 significantly lower, suggesting that these cells are by nature much more repressive.
438 These results are in line with previous studies showing that MIEP is more repressed in
439 CD34+ cells (63). Likewise, in natural latency we were unable to detect any viral
440 transcripts by examining more than 1.5 billion RNA-seq reads from CD34+ cells. In
441 contrast, by examining 3 billion RNA-seq reads from the blood we identified 378 viral
442 reads from 18 samples. These results suggest that viral gene expression is more
443 restricted in CD34+ progenitors both in natural and in experimental settings and further
444 support the notion that the host cell environment plays a major role in dictating the
445 latency state.
446

447 An essential step in understanding HCMV latency is deciphering the importance of viral
448 transcripts and proteins to latency maintenance and to the ability of the virus to
449 reactivate. Based on the view that only a limited number of genes are expressed during
450 HCMV latency, only several candidates for viral functions that may control HCMV
451 latency have been studied. These include UL138 (31, 32), astUL81-82/LUNA (34, 48),
452 UL111A/LAcmvIL-10 (33, 35) and US28 (36, 37). Despite the lack of a clear restricted
453 latency-associated expression program, our results do not undermine the importance of
454 these factors to HCMV latency, rather add many additional candidate genes. Two
455 appealing candidates are RNA2.7 and UL30. RNA2.7 is the most abundant transcript in
456 both lytic and latent cells, but in our measurements RNA2.7 relative expression in latent
457 cells was constantly higher than expected when comparing to the lytic profile. RNA2.7
458 was demonstrated to protect infected cells from mitochondria-induced cell death (64),
459 but its role in latency was never tested. UL30 transcript was suggested to encode for
460 *UL30A*, which is conserved among primate cytomegaloviruses, and expressed from a
461 nonconventional initiation codon (ACG) (18, 19) but its functional role was never
462 studied. Future work will have to delineate the importance of the different transcripts
463 we detected to regulating latency.
464
465 Overall, our experiments and analyses start to challenge the dogma that all
466 herpesviruses express a highly restricted latency-associated program and suggest that
467 HCMV latency is more associated with quantitative shifts rather than qualitative
468 changes in viral gene expression. Although the relevance of these viral transcripts to

latency should be further studied, our findings provide a potential new context for deciphering virus-host interactions underlying HCMV lifelong persistence.

Acknowledgments

We thank Yosef Shaul, Schraga Schwartz, Igor Ulitsky, Rotem Sorek, Ian Mohr and Stern-ginossar lab members for critical reading of the manuscript. We thank Eain A. Murphy for the TB40E-GFP virus strain. We thank Elad Chomsky, Yaara Arkin, Hadas Keren-Shaul and Efrat Hagai for technical assistance. This research was supported by the EU-FP7-PEOPLE career integration grant, the Israeli Science Foundation (1073/14), Infect-ERA (TANKACY) and the European Research Council starting grant (StG-2014-638142). NS-G is incumbent of the skirball career development chair in new scientist.

Figures legend

Figure 1: Viral gene expression during natural persistent infection

(A) Box plot showing number of HCMV reads per sample in HCMV seronegative and HCMV seropositive samples (B) Bar plot showing distribution of total sequenced reads in different tissues, color coding reflects the number of viral reads normalized to total number of sequenced reads in each tissue (Number of HCMV reads/10^9 total aligned reads). Viral reads from samples containing less than 2 viral reads were filtered out. Data for all samples was obtained from GTEx (40, 65) except CD34+ data that were collected from 25 different NCBI GEO datasets (Table S1D). (C) Hierarchical clustering of natural samples, with more than 4 HCMV reads, according to viral gene expression. The samples are portioned into 2 groups: group I and group II. Upper panel color coding indicates tissue origin of each sample. The heatmap in the lower panel shows expression level of representative differentially expressed genes in each sample. (D) Heatmap showing correlations between viral gene expression program from natural samples from both groups (I and II) and experimental lytically infected fibroblasts at different time points post infection.

Figure 2: Establishment of HCMV latency in CD14+ monocytes

(A) Monocytes and monocyte-derived macrophages were infected with HCMV strain TB40E-GFP at an MOI of 5. RNA was collected at 4 days post infection (dpi) from the

latent monocytes and 5 hours post infection (hpi) from lytic monocyte-derived macrophages and was analyzed by RT-qPCR for the transcript levels of UL138 and IE1. Expression was normalized to the human Anxa5 transcript. Means and error bars (showing standard deviations) represent three measurements. (B) Monocytes were latently infected with TB40E-GFP at an MOI of 5. Three dpi cells were either differentiated into dendritic cells (reactivated DCs) or left undifferentiated (latent monocytes) and 2 days post terminal differentiation reactivation was visualized by GFP and IE1/2 staining. Representative fields are presented. (C) Monocytes were latently infected with TB40E-GFP at an MOI of 5. At 3 dpi cells were either differentiated to dendritic cells (reactivated DCs) or left undifferentiated (latent monocytes). Two days post terminal differentiation cells were co-cultured with primary fibroblasts and GFP-positive plaques were counted. Number of positive plaques per 100,000 monocytes or monocyte-derived dendritic cells is presented. Cell number and viability were measured by Trypan blue staining prior to plating. Means and error bars (showing standard deviations) represent two experiments.

518 *Figure 3: scRNA-seq analysis of latently infected CD14+ monocytes*

519

520 Single cell RNA sequencing analysis of 3,655 cells from a cell population of latently
521 infected monocytes. CD14+ monocytes were infected with HCMV (TB40E-GFP) and
522 analyzed at 3, 4, 5, 6, 7 and 14 dpi. (A) t-SNE plot of all 3,655 single cells based on host
523 and viral gene expression. Color bar shows the percentage of viral reads from total
524 reads per cell. (B) Heatmap showing clustering analysis of 3,655 single cells rows show
525 expression of 176 most differential genes (32 out of 171 detected viral transcripts, 144
526 out of 15,812 detected cellular transcripts). Bar over the upper panel shows the number
527 of reads obtained for each cell (log scale). Upper panel shows the most abundant viral
528 genes, lower panel indicates most differential host genes and bars under the heatmap
529 indicate the percentage of viral reads from total reads and dpi for each cell. Cells are
530 partitioned into 6 distinct clusters (C1-6) based on gene expression profiles and ordered
531 by the relative abundance of viral reads, from high to low. Number of cells in each
532 cluster is shown in parentheses next to the cluster number.

533

534 *Figure 4: Transcriptional program in latently infected CD14+ monocytes*

535

536 (A) Scatterplot showing read number of viral genes in latent monocytes (defined as cells
537 in which the proportion of viral reads was below 0.5% of total reads) versus lytic cells
538 (cells from cluster 1). Horizontal and vertical error bars indicate 95% non-parametric
539 bootstrap confidence interval across cells. (B) Relative expression of IE1 and UL138

transcripts in RNA-seq data from lytic fibroblasts at 5 and 72 hpi. (C) Relative RNA expression level of viral RNA2.7 (left panel) and RNA4.9 (right panel) in monocytes infected with untreated or UV inactivated virus, measured by qRT-PCR at 5 dpi. A representative analysis of two independent experiments is shown. (D) RNA expression level of viral RNA2.7 (left panel) and RNA4.9 (right panel), relative to –RT samples, in infected monocytes, measured by qRT-PCR at 5 hours and 5 days post infection. Means and error bars (showing standard deviations) represent three measurements. A representative analysis of two independent experiments is shown.

Figure 5: scRNA-seq analysis of latently infected CD34+ progenitor cells

Single cell RNA sequencing analysis of 7,634 cells randomly sampled from a cell population of latently infected HPCs. CD34+ HPCs were infected with HCMV (TB40E-GFP) and analyzed at 4 dpi (A) t-SNE projection of all 7,634 single cells based on host and viral gene expression. Color bar shows the level of viral gene expression as percentage of total reads per cell. (B) Scatter plot showing read number of all viral genes in the latently infected CD34+ progenitors versus lytic cells. Horizontal and vertical error bars indicate 95% non-parametric bootstrap confidence interval across cells.

Supplementary figures and tables

Figure S1: Detection of HCMV reads in natural samples

(A) Distribution of the number of total aligned reads per sample in samples from the GTEx dataset. (B) Distribution of the number of HCMV aligned reads per sample in positive samples from the GTEx dataset. (C) RNA-seq reads from GTEx samples aligned to the MIEP region of HCMV genome colored by sample. (D) Alignment of RNA-seq reads from GTEx samples and sequences of 101 clinical isolates to the MIEP region (positions 175,493 and 175,494 in the viral genome). Base variation from the reference (Merlin strain, which is identical in these sites to the CMV promoter that is used in plasmids) is indicated by a color corresponding to the substituting base. Color legend is on the right. (E) Percentage of samples containing HCMV reads in different tissues. Viral reads from samples containing less than 2 viral reads were filtered out. (F) Genome browser view showing aligned reads from samples assigned to group I or group II in genome regions coding for abundant genes in these groups.

Figure S2: Clustering according to HCMV reads in natural samples

(A) Scatterplot showing read number of viral genes in group I samples from the GTEx database versus lytic fibroblasts 72 hours post infection. (B) Scatterplot showing read number of viral genes in group II samples from the GTEx database versus lytic fibroblasts

5 hours post infection. (C and D) Violin plots showing the time of sample harvesting (measured in minutes after death) versus (C) sample assignment to gene expression group (I or II) and (D) presence or absence of HCMV specific reads in the sample.

Figure S3: Validation of infection and scRNA library composition

(A) Flow cytometry analysis showing GFP expression level in population of CD14+ monocytes infected with TB40-GFP at 2 dpi. (B and C) Bar plots showing distribution of number of reads per cell (left) and number of genes per cell (right) in scRNA-seq data of (B) infected CD14+ monocytes and (C) CD34+ HPCs.

Figure S4: scRNA-seq analysis of latently infected CD14+ monocytes.

(A) t-SNE plot of all 3,655 single cells based on host gene expression. Color bar shows the percentage of viral reads from total reads per cell. (B) t-SNE plot of 3,655 single latently infected monocytes based on host and viral gene expression (as shown in Fig. 3A) depicting the separation to 6 clusters as shown in Fig 3B. (C) Scatterplot showing read number of all viral genes in cells from cluster 1 versus lytically infected monocyte derived macrophages at 4 dpi (left panel) or fibroblasts at 3 dpi (right panel). (D) Scatterplot showing read number of all viral genes in cells from clusters 2-6 (labeled on

y-axis) versus cells from cluster 1. Horizontal and vertical error bars indicate 95% nonparametric bootstrap confidence interval across cells.

Figure S5: scRNA-seq analysis of latently infected CD14+ cells clustered by days post infection (dpi)

Single cell RNA sequencing analysis of 3,655 cells from a cell population of latently infected monocytes. CD14+ monocytes were infected with HCMV (TB40E-GFP) and analyzed at 3, 4, 5, 6, 7 and 14 dpi. The heatmap shows clustering analysis of 3,655 single cells reflecting expression of all viral genes detected. Cells are partitioned into 6 clusters (C1-6) according to the day post infection (dpi). Number of cells in each cluster is shown in parentheses next to the cluster number. The bar above the heatmap shows the total reads number for each cell (log scale).

Figure S6: Assessment of lytic noise effect on gene expression in CD14+ scRNA-seq libraries

Scatter plot showing read number of all viral genes in (A) latent cells (defined as cells in which the proportion of viral reads was below 0.5% of total reads) and in (B) cells from clusters 2-6 (labeled on y-axis) versus cells from cluster 1. Analysis was done after

627 exclusion of cells in which viral read counts were lower than the noise cut-off level (See
628 materials and methods). Horizontal and vertical error bars indicate 95% non-parametric
629 bootstrap confidence interval across cells.
630
631
632 *Figure S7: Analysis of viral gene expression in latent and ly*tic *samples reported by Cheng*
633 e*t al., 2017* (39).
634
635 Scatterplots showing expression of each detected gene in latent (at 6 dpi) vs. lytic
636 samples at 48 hpi (left) and 72 hpi (right). X and Y values for each gene represent its
637 percentage out of all viral reads. Values for each gene were calculated as a mean of two
638 donors, error bars indicate SD.
639
640 **Supplementary tables:**
641 *Table S1: Analysis of natural inection*
642 *(A) Summary of HCMV reads in GTEx samples*
643 Columns indicate: Sample ID, subject ID, HCMV Sero Status, number of reads (in
644 millions), number of aligned reads (in millions), number of HCMV reads, number of
645 HCMV reads exluding the MIEP region transcript. Columns I to end indicate the number
646 of reds for each indicated gene.
647
648 *(B) Attributes of GTEx seropositive samples*

649 The detailed description of what each column represents can be found at:

650 ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000424/phs000424.v7.p2/pheno_variable

651 _summaries/phs000424.v7.pht002742.v7.p2.GTEx_Subject_Phenotypes.var_report.xml

652

653 *(C) Attributes of GTEx seropositive subjects*

654 The detailed description of what each column represents can be found at:

655 ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000424/phs000424.v5.p1/pheno_variable

656 _summaries/phs000424.v5.pht002743.v5.p1.GTEx_Sample_Attributes.var_report.xml

657

658 *(D) Analysis of publicly available CD34+ RNA-seq datasets*

659 Columns indicate: Data set ID, Sample file ID, Cell type, number of reads in indicated

660 sample, number of aligned reads.

661

662

663 *Table S2: scRNA-seq analysis*

664

665 *(A) Differential genes in latently infected monocytes*

666 Columns indicate: Gene annotation, the cluster with the highest expression of the

667 indicated gene, expression level of the indicated gene in the cluster where it is most

668 highly expressed (relative to the expression of all other genes in the same cluster),

669 number of reads for the indicated gene in clusters 1-6, total number of reads for the

670 indicated gene across all clusters.

(*B) Transcripts enriched in latent CD14+ monocytes*

Columns indicate: gene annotation, number of reads in lytic cells (cluster 1), number of viral reads in latent cells (cells in which less than 0.5% of the reads originated from the virus), mean and SD of the number of reads for an indicated gene in the latent cells according to bootstrap analysis (see material and methods), Z-score, false discovery rate (FDR).

(C) Transcripts enriched in latent cells from clusters 2-6 compared to lytic cells

Columns indicate: gene annotation, number of reads in cluster 1, number of reads in the indicated cluster, mean and SD of the number of reads for each specified gene in the indicated cluster according to bootstrap analysis (see material and methods), Z-score, false discovery rate (FDR).

(*D) Viral genes detected in latently infected CD34+ HPCs by MARS-seq analysis*

The number of reads identified for each of the detected viral genes, in each of the cells. Cell sum- indicates the total number of viral reads per cell. Gene Sum- indicates the total number of reads detected for each viral gene.

(E) Transcripts enriched in latent CD34+ HPCs

691 Columns indicate: gene annotation, number of reads in lytic cells (cluster 1 in CD14+
692 analysis), number of viral reads in infected HPCs, mean and SD of the number of reads
693 for an indicated gene in latent CD34+ HPCs according to bootstrap analysis (see material
694 and methods), Z-score, false discovery rate (FDR).

Materials and Methods

Cells and virus stocks

Primary CD14+ monocytes were isolated from fresh venous blood, obtained from healthy donors, using Lymphoprep (Stemcell Technologies) density gradient followed by magnetic activated cell sorting with CD14+ magnetic beads (Miltenyi Biotec). Cryopreserved Bone Marrow CD34+ Cells were obtained from Lonza. Alternatively, fresh CD34+ cells were purified from umbilical cord blood of healthy donors. Isolation was done using Lymphoprep (Stemcell Technologies) density gradient followed by magnetic activated cell sorting with CD34+ magnetic beads (Miltenyi Biotec). CD34+ and CD14+ cells were cultured in X-Vivo15 media (Lonza) supplemented with 2.25mM L-glutamine at 37^{o}C in 5% CO2 (66).

Human foreskin fibroblasts (HFF) (ATCC CRL-1634) and retinal pigmented epithelial cells (RPE-1) (ATCC CRL-4000) were maintained in DMEM with 10% fetal bovine serum (FBS), 2mM L-glutamine, and 100 units/ml penicillin and streptomycin (Beit-Haemek, Israel).

The bacterial artificial chromosome (BAC)-containing the clinical strain TB40E with an SV40-GFP tag (TB40E-GFP) was described previously (67, 68). This strain lacks the US2-US6 region, and therefore these genes were not included in our analysis. Virus was propagated by electroporation of infectious BAC DNA into HFF cells using the Amaxa P2 4D-Nucleofector kit (Lonza) according to the manufacturer's instructions. Viral stocks

were concentrated by ultracentrifugation at 70000xg, 4^{o}C for 40 minutes. Infectious virus yields were assayed on RPE-1 cells.

Infection and reactivation procedures

For experimental latent infection models, CD14+ monocytes and CD34+ HPCs were infected with HCMV strain TB40E-GFP at MOI of 5. Cells were incubated with the virus for 3 hours, washed and supplemented with fresh media. To assess infection efficiency, a sample of the infected cell population was FACS analyzed for GFP expression at 2 dpi. For single cell experiments cells were isolated without further selection; CD14+ cells were harvested at 3, 4, 5, 6, 7 and 14 dpi and CD34+ HPCs were harvested at 4 dpi. Lytic infection was carried out on primary fibroblasts and monocyte-derived macrophages obtained by growing CD14+ monocytes in 50ng/ml PMA containing media for 2 days. For reactivation assays, infected monocytes were differentiated into dendritic cells (DCs) at 3 dpi by incubation with granulocyte-macrophage CSF and interleukin-4 (Peprotech) at 1,000 U/ml for 5 days, followed by stimulation with 500 ng/ml of LPS (Sigma) for 48 hours (as previously described in (66)). Release of infectious virions was assayed by co-culturing of 100,000 differentiated and non-differentiated infected monocytes at the end of the differentiation procedure with HFF cells for 10 days and quantification of GFP positive plaques. Cell number and viability were measured by Trypan blue staining prior to plating.

For UV inactivation, the virus was irradiated in a Stratalinker 1800 (Stratagene) with 200 mjoules.

736 *Immunofluorescence*

737 Cells were fixed in 4% paraformaldehyde for 10 minutes, permeabilized with 0.1% Triton
738 X-100 in PBS for 10 minutes and blocked in 10% normal goat serum in PBS. Detection of
739 IE-1 was performed by immunostaining with anti-IE1 antibodies (1:100, Abcam
740 ab53495), followed by goat anti-mouse antibody (1:200, AlexaFluor647, Invitrogen
741 A21235) and Hoechst nuclear stain. Cells were visualized in a Zeiss Axioobserver
742 fluorescent microscope.

743 *qRT-PCR*

744 Total RNA was extracted using Tri-Reagent (Sigma) according to manufacturer's
745 protocol. cDNA was prepared using qScript cDNA Synthesis Kit (Quanta Biosciences)
746 according to manufacturer's protocol. Real time PCR was performed using the SYBR
747 Green PCR master-mix (ABI) on a real-time PCR system QuantStudio 12K Flex (ABI) with
748 the following primers (forward, reverse):
749
750 IE1 (GGTGCTGTGCTGCTATGTCTC, CATGCAGATCTCCTCAATGC)
751 UL138 (GTGTCTTCCCAGTGCAGCTA, GCACGCTGTTTCTCTGGTTA)
752 RNA 2.7 (TCCTACCTACCACGAATCGC, GTTGGGAATCGTCGACTTTG)
753 RNA 4.9 (GTAAGACGGGCAAATACGGT, AGAGAACGATGGAGGACGAC)
754 Anxa 5 (AGTCTGGTCCTGCTTCACCT, CAAGCCTTTCATAGCCTTCC)
755

Single cell sorting and MARS-seq RNA library construction

Single cell sorting and library preparation were conducted according to the massively parallel single-cell RNA-seq (MARS-seq) protocol, as previously described (49). In brief, cells from latently infected populations of CD14+ monocytes and CD34+ HPCs were FACS sorted into wells of 384 well capture plates containing 2 µl of lysis buffer and reverse transcription (RT) indexed poly(T) primers, thus generating libraries representing the 3' of mRNA transcripts. Four empty wells were kept in each 384-well plate as a no-cell control during data analysis. Immediately after sorting, each plate was spun down to ensure cell immersion into the lysis solution, snap frozen on dry ice and stored at −80 °C until processed. Barcoded single-cell capture plates were prepared with a Bravo automated liquid handling platform (Agilent). For generation of RNA library, mRNA from cells sorted into capture plates was converted into cDNA and pooled using an automated pipeline. The pooled sample was then linearly amplified by T7 in vitro transcription, and the resulting RNA was fragmented and converted into a sequencing-ready library by tagging the samples with pool barcodes and Illumina sequences during ligation, RT, and PCR. Each pool of cells was tested for library quality and concentration was assessed as described earlier (49).

RNA sequencing of lytic cells

For generation of a reference lytic RNA library used in the single cell experiments, monocyte-derived macrophages or primary fibroblasts were infected with TB40E-GFP

776 virus at MOI of 5 and used for library preparation at 4 dpi. The libraries were generated
777 from a samples of ~10,000 cells according to the MARS-seq protocol (49).

778 The lytic fibroblasts derived RNA-seq libraries used as reference in analysis of the
779 natural samples were previously described (18).

780 *Single cell library construction using 10x platform*

781 Cell suspensions at a density of 700 cells/μl in PBS + 0.04% BSA were prepared for single
782 cell sequencing using the Chromium Single Cell 3' Reagent Version 2 Kit and Chromium
783 Controller (10x Genomics, CA, USA) as previously described (69). Briefly, 9,000 cells per
784 reaction were loaded for Gel Bead-in-Emulsion (GEM) generation and barcoding. GEM-
785 RT, post GEM-RT cleanup and cDNA amplification were performed to isolate and amplify
786 cDNA for library construction. Libraries were constructed using the Chromium Single Cell
787 3' Reagent Kit (10x Genomics, CA, USA) according to manufacturer's protocol. Library
788 quality and concentration were assessed according to manufacturer's instructions.

789 *Sequencing*

790 RNA-Seq libraries (pooled at equimolar concentration) were sequenced using NextSeq
791 500 (Illumina), at median sequencing depth of ~45,000 reads per cell for MARS-Seq and
792 ~32,000 reads per cell for 10x. Read parameters were: Read1: 72 cycles and Read2: 15
793 cycles, for MARS-seq and Read1: 26 cycles, Index1: 8 cycles, and Read2: 58 cycles for
794 10x.

795 *MARS-seq CD14+ analysis*

796 The analysis of the MARS-seq data was done with the tools described in (49) and (50).
797 The reference was created from the hg19 and TB40E (NCBI EF999921.1) strain of HCMV.
798 The transcription units of the virus were based on NCBI annotations, with some changes
799 based on the alignment results. This includes merging together several transcripts
800 (taking into account that the library maps only the 3' of transcripts), and adding some
801 antisense transcripts. Reads assignment to wells was based on the batch barcode (4bps)
802 and the well barcode (7bp), and removing reads with low quality of the barcodes. The
803 read itself (37bp) was aligned to the reference using bowtie2 (70), and the counting of
804 the reads per gene is done based on unique molecular identifiers (UMIs) (8bp). For each
805 batch the leakage noise level was estimated by comparing the number of UMIs in the 2
806 empty wells, to the total number of UMIs in the batch. Batches with high noise level
807 (>8%) were discarded. Wells with number of reads < 1000 were discarded. The number
808 of wells that were used for further analysis is 3655. Genes with low total number of
809 reads (< 10), with low variability (variance / mean < 1.1), and also ribosomal protein and
810 histones were excluded. By using a multiplicative probabilistic model, and expectation-
811 maximization like optimization procedure, the 3,655 cells were clustered to 6 clusters.
812 The model includes a regularization parameter (=0.5) simulating additional uniform
813 reads to all genes. The clusters are ordered according to the viral content from high to
814 low.

When analyzing correlation in gene expression the error bars represent 95% confidence interval that were calculated by 10,000 bootstrap iteration of the cells in each one of the clusters. The t-SNE plot of the MARS-seq CD14+ cells was calculated with the R package (71), after down sampling each cell to 1000 UMIs.

To exclude background noise, in each one of the batches, all cells with number of viral reads below 3 times the estimated noise at this batch, were excluded.

To estimate the p-value of getting number of reads n, in cluster B, under the null hypothesis of same expression program as in cluster A, a semi parametric bootstrap method was used. First the probability of sampling UMIs for each viral gene was calculated according to the gene expression in cluster A. Then each bootstrap simulation consists of a parametric step and a-parametric step. The parametric step is, for each cell in cluster B, to sample number of UMIs according to the actual number of read in this cell, with distribution over the genes according to the probabilities calculated from cluster A. Then the a-parametric step is a usual bootstrap sampling of the cells in cluster B, and calculate the total number of reads in this cluster B. After doing this simulation 1,000 times, for each viral gene, the mean and the standard deviation of the number of reads in cluster B, under the null hypothesis was calculated. Based on this value, the Z-score of the actual value n was calculated, and a p-value was calculated assuming normal distribution of the number of reads under the null hypothesis. Lastly, these p-values were adjusted for multiple testing, and just the genes with false discovery rate (FDR) of < 0.01 are reported in tables S2B and S2C.

836 *GTEx and GEO analysis*

837 All RNA-Seq, paired end GTEx samples available on July 2016 were used for the analysis.
838 The reference genome that was used was based on hg19 and Merlin strain of HCMV
839 (NC_006273.2). Bowtie2 (70) was used for alignment with the default parameters,
840 besides the additional flag --local. Pairs with mapping quality less than 30 were
841 excluded. Pairs with only one read aligned to the Merlin sequence were excluded. For
842 each sample possible PCR duplications were removed. The counting of the alignments to
843 the genes was done with HTSeq-count (72). Annotation of gff files is based on NCBI data,
844 with some adjustment taking into account correcting for the non-stranded library. The
845 clustering for Fig. 1C and 1D were generated with GENE-E (73). The analysis of the
846 CD34+ GEO samples was carried out in the same way. The list of datasets that were
847 used is presented in Table S1D.

848 *10X CD34+ Data analysis*

849 We used CellRanger v2.0.0 (74) software with the default settings to process the FASTQ
850 files. The reference was created with the mkref CellRanger command, based on the
851 CellRanger human hg19 reference, and TB40E (NCBI EF999921.1) as was used in the
852 analysis of the MARS-seq data. The de-multiplexing of the Illumina files, and the analysis
853 done with the CellRanger commands mkfastq and count respectively, The raw reads
854 data was extracted with the CellRanger R Kit (74). The t-SNE plot is based on the
855 coordinates calculated by the count command.

Analysis of data from Cheng et al., 2017

The files containing the number of viral reads per samples were downloaded from GSE99823. Full details are given in (39). Briefly, lung fibroblasts (MRC-5), and CD34+ cells from few donors were infected with HCMV TB40E strain, and extracted RNA were sequenced (paired end). The computational pipeline includes trimming and QC with Trim Galore, alignment with Tophat2, and reads counting with HTSeq. In the presented correlation figure, just wild type samples, without any selection were used. For each sample, the number of reads was normalized to percentage of viral expression, and then for the two CD34+ samples, the mean and standard deviation of the percentage were calculated and are displayed in Fig. S7 vs. the percentage viral expression of the HFF sample.

Ethical statement

All fresh peripheral blood samples were obtained after approval of protocols by the Weizmann Institutional Review Board (IRB application 92-1) and umbilical cord blood of anonymous healthy donors was obtained in accordance with local Helsinki committee approval (#RMB-0452-15). Informed written consent was obtained from all volunteers and all experiments were carried out in accordance with the approved guidelines.

Data availability

All next generation sequencing data files were deposited on Gene Expression Omnibus under accession number GSE101341.

References

1. Pass RF, Stagno S, Britt WJ, Alford CA. 1983. Specific cell-mediated immunity and the natural history of congenital infection with cytomegalovirus. Journal of Infectious Diseases 148:953–961.

2. Zanghellini F, Boppana SB, Emery VC, Griffiths PD, Pass RF. 1999. Asymptomatic Primary Cytomegalovirus Infection: Virologic and Immunologic Features. The Journal of Infectious Diseases 180:702–707.

3. Griffiths PD. 2010. Cytomegalovirus in intensive care. Reviews in Medical Virology 20:1–3.

4. N S, Kirby KA, Rubenfeld GD, Leisenring WM, Bulger EM, Neff MJ, Gibran NS, Huang M-L, Hayes TKS, Corey L, Boeckh M. 2008. Cytomegalovirus Reactivation in Critically Ill Immunocompetent Patients. Jama 300:413.

5. Crough T, Khanna R. 2009. Immunobiology of Human Cytomegalovirus: from Bench to Bedside. Clinical Microbiology Reviews 22:76–98.

6. Mocarski E, Shenk T, Griffiths P, Pass R. Cytomegaloviruses, p 1960–2014. virology, 6th ed Lippincott Williams & Wilkins, 2013.

7. Sinzger C, Digel M, Jahn G. 2008. Cytomegalovirus cell tropism. Current Topics in Microbiology and Immunology 325:63–83.

8. Smith MS, Goldman DC, Bailey AS, Pfaffle DL, Kreklywich CN, Spencer DB, Othieno FA, Streblow DN, Garcia JV, Fleming WH, Nelson JA. 2010. Granulocyte-colony stimulating factor reactivates human cytomegalovirus in a latently infected

humanized mouse model. Cell Host and Microbe2010/09/14. 8:284–291.

9. Poole E, Juss JK, Krishna B, Herre J, Chilvers ER, Sinclair J. 2015. Alveolar macrophages isolated directly from human cytomegalovirus (HCMV)-seropositive individuals are sites of HCMV reactivation in vivo. Journal of Infectious Diseases 211:1936–1942.

10. Taylor-Wiedeman J, Sissons P, Sinclair J. 1994. Induction of Endogenous Human Cytomegalovirus Gene Expression after Differentiation of Monocytes from Healthy Carriers. Journal of Virology 68:1597–1604.

11. Reeves MB, MacAry PA, Lehner PJ, Sissons JGP, Sinclair JH. 2005. Latency, chromatin remodeling, and reactivation of human cytomegalovirus in the dendritic cells of healthy carriers. Proceedings of the National Academy of Sciences 102:4140–4145.

12. Huang MM, Kew VG, Jestice K, Wills MR, Reeves MB. 2012. Efficient Human Cytomegalovirus Reactivation Is Maturation Dependent in the Langerhans Dendritic Cell Lineage and Can Be Studied using a $CD14^{+}$ Experimental Latency Model. Journal of Virology2012/06/01. 86:8507–8515.

13. Soderberg-Naucler C, Streblow DN, Fish KN, Allan-Yorke J, Smith PP, Nelson JA. 2001. Reactivation of Latent Human Cytomegalovirus in CD14+ Monocytes Is Differentiation Dependent. Journal of Virology 75:7543–7554.

14. Reeves MB. 2005. An in vitro model for the regulation of human cytomegalovirus latency and reactivation in dendritic cells by chromatin remodelling. Journal of

General Virology.

15. Reeves MB. 2011. Chromatin-mediated regulation of cytomegalovirus gene expression. Virus Research 157:134–143.

16. Meier JL. 2001. Reactivation of the human cytomegalovirus major immediate-early regulatory region and viral replication in embryonal NTera2 cells: role of trichostatin A, retinoic acid, and deletion of the 21-base-pair repeats and modulator. Journal of virology 75:1581–93.

17. Gan X, Wang H, Yu Y, Yi W, Zhu S, Li E, Liang Y. 2017. Epigenetically repressing human cytomegalovirus lytic infection and reactivation from latency in THP-1 model by targeting H3K9 and H3K27 histone demethylases. Plos One.

18. Stern-Ginossar N, Weisburd B, Michalski A, Le VTK, Hein MY, Huang S-X, Ma M, Shen B, Qian S-B, Hengel H, Mann M, Ingolia NT, Weissman JS. 2012. Decoding Human Cytomegalovirus. Science 338:1088–1093.

19. Gatherer D, Seirafian S, Cunningham C, Holton M, Dargan DJ, Baluchova K, Hector RD, Galbraith J, Herzyk P, Wilkinson GWG, Davison AJ. 2011. High-resolution human cytomegalovirus transcriptome. Proceedings of the National Academy of Sciences of the United States of America2011/11/24. 108:19755–60.

20. Sinclair J, Sissons P. 2006. Latency and reactivation of human cytomegalovirus. Journal of General Virology 87:1763–1779.

21. Poole E, Sinclair J. 2015. Sleepless latency of human cytomegalovirus. Medical Microbiology and Immunology 204:421–429.

939 22. Sinclair JH, Reeves MB. 2013. Human cytomegalovirus manipulation of latently
940 infected cells. Viruses2013/11/29. 5:2803–2824.

941 23. Dupont L, Reeves MB. 2016. Cytomegalovirus latency and reactivation: recent
942 insights into an age old problem. Reviews in Medical Virology 26:75–89.

943 24. Goodrum F. 2016. Human Cytomegalovirus Latency: Approaching the Gordian
944 Knot. Annual Review of Virology 3:333–357.

945 25. Slobedman B, Cao JZ, Avdic S, Webster B, McAllery S, Cheung AK, Tan JC,
946 Abendroth A. 2010. Human cytomegalovirus latent infection and associated viral
947 gene expression. Future Microbiology 5:883–900.

948 26. Kondo K, Kaneshima H, Mocarski ES. 1994. Human cytomegalovirus latent
949 infection of granulocyte-macrophage progenitors. Proceedings of the National
950 Academy of Sciences of the United States of America 91:11879–83.

951 27. Kondo K, Mocarski ES. 1995. Cytomegalovirus latency and latency-specific
952 transcription in hematopoietic progenitors. Scandinavian journal of infectious
953 diseases Supplementum 99:63–7.

954 28. Kondo K, Xu J, Mocarski ES. 1996. Human cytomegalovirus latent gene expression
955 in granulocyte-macrophage progenitors in culture and in seropositive individuals.
956 Proceedings of the National Academy of Sciences of the United States of America
957 93:11137–42.

958 29. Goodrum FD, Jordan CT, High K, Shenk T. 2002. Human cytomegalovirus gene
959 expression during infection of primary hematopoietic progenitor cells: A model

960 for latency. Proceedings of the National Academy of Sciences2002/11/29.
961 99:16255–16260.

962 30. Cheung AKL, Abendroth A, Cunningham AL, Slobedman B. 2006. Viral gene
963 expression during the establishment of human cytomegalovirus latent infection in
964 myeloid progenitor cells. Blood 108:3691–3699.

965 31. Goodrum F, Reeves M, Sinclair J, High K, Shenk T. 2007. Human cytomegalovirus
966 sequences expressed in latently infected individuals promote a latent infection in
967 vitro. Blood2007/04/19. 110:937–945.

968 32. Petrucelli A, Rak M, Grainger L, Goodrum F. 2009. Characterization of a Novel
969 Golgi Apparatus-Localized Latency Determinant Encoded by Human
970 Cytomegalovirus. Journal of Virology 83:5615–5629.

971 33. Jenkins C, Abendroth A, Slobedman B. 2004. A Novel Viral Transcript with
972 Homology to Human Interleukin-10 Is Expressed during Latent Human
973 Cytomegalovirus Infection. Journal of Virology 78:1440–1447.

974 34. Bego M, Maciejewski J, Khaiboullina S, Pari G, St. Jeor S. 2005. Characterization of
975 an Antisense Transcript Spanning the UL81-82 Locus of Human Cytomegalovirus.
976 Journal of Virology 79:11022–11034.

977 35. Cheung AKL, Gottlieb DJ, Plachter B, Pepperl-Klindworth S, Avdic S, Cunningham
978 AL, Abendroth A, Slobedman B. 2009. The role of the human cytomegalovirus
979 UL111A gene in down-regulating CD4+T-cell recognition of latently infected cells:
980 Implications for virus elimination during latency. Blood 114:4128–4137.

981 36. Humby MS, O'Connor CM. 2016. Human Cytomegalovirus US28 Is Important for
982 Latent Infection of Hematopoietic Progenitor Cells. Journal of Virology 90:2959–
983 2970.

984 37. Krishna BA, Spiess K, Poole EL, Lau B, Voigt S, Kledal TN, Rosenkilde MM, Sinclair
985 JH. 2017. Targeting the latent cytomegalovirus reservoir with an antiviral fusion
986 toxin protein. Nature Communications 8:14321.

987 38. Rossetto CC, Tarrant-Elorza M, Pari GS. 2013. Cis and Trans Acting Factors
988 Involved in Human Cytomegalovirus Experimental and Natural Latent Infection of
989 CD14 (+) Monocytes and CD34 (+) Cells. PLoS Pathogens 9:e1003366.

990 39. Cheng S, Caviness K, Buehler J, Smithey M, Nikolich-Žugich J, Goodrum F. 2017.
991 Transcriptome-wide characterization of human cytomegalovirus in natural
992 infection and experimental latency. Proceedings of the National Academy of
993 Sciences.

994 40. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G,
995 Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S,
996 Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker
997 L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B,
998 Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive
999 C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R,
1000 Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M,
1001 Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D,

MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struewing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF. 2013. The Genotype-Tissue Expression (GTEx) project. Nature Genetics 45:580–585.

41. Slobedman B, Mocarski ES. 1999. Quantitative Analysis of Latent Human Cytomegalovirus. Journal of Virology 73:4806–4812.

42. Taylor-Wiedeman J, Sissons JGP, Borysiewicz LK, Sinclair JH. 1991. Monocytes are a major site of persistence of human cytomegalovirus in peripheral blood mononuclear cells. Journal of General Virology 72:2059–2064.

43. Mendelson M, Monard S, Sissons P, Sinclair J. 1996. Detection of endogenous human cytomegalovirus in CD34+ bone marrow progenitors. Journal of General Virology 77:3099–3102.

44. Wang Y, Navin NE. 2015. Advances and Applications of Single-Cell Sequencing Technologies. Molecular Cell 58:598–609.

45. Ciuffi A, Rato S, Telenti A. 2016. Single-cell genomics for virology. Viruses 8:123.

46. O'Connor CM, Vanicek J, Murphy EA. 2014. Host MicroRNA Regulation of Human Cytomegalovirus Immediate Early Protein Translation Promotes Viral Latency. Journal of Virology2014/03/07. 88:5524–5532.

47. Lau B, Poole E, Krishna B, Montanuy I, Wills MR, Murphy E, Sinclair J. 2016. The Expression of Human Cytomegalovirus MicroRNA MiR-UL148D during Latent Infection in Primary Myeloid Cells Inhibits Activin A-triggered Secretion of IL-6. Scientific Reports 6:31205.

48. Reeves MB, Sinclair JH. 2010. Analysis of latent viral gene expression in natural and experimental latency models of human cytomegalovirus and its correlation with histone modifications at a latent promoter. Journal of General Virology 91:599–604.

49. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. 2014. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science2014/02/18. 343:776–779.

50. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weiner A, David E, Cohen N, Lauridsen FKB, Haas S, Schlitzer A, Mildner A, Ginhoux F, Jung S, Trumpp A, Porse BT, Tanay A, Amit I. 2015. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163:1663–1677.

51. Bresnahan WA. 2000. A Subset of Viral Transcripts Packaged Within Human Cytomegalovirus Particles. Science 288:2373–2376.

52. Terhune SS, Schroer J, Shenk T. 2004. RNAs Are Packaged into Human Cytomegalovirus Virions in Proportion to Their Intracellular Concentration. Journal of Virology 78:10390–10398.

53. von Laer D, Meyer-koenig U, Serr A, Finke J, Kanz L, Fauser AA, Neumann-haefelin D, Brugger W, Hufert FT. 2014. Detection of cytomegalovirus DNA in CD34+ cells from blood and bone marrow 4086–4090.

54. Glaser R, Kiecolt-Glaser JK. 2005. Science and society: Stress-induced immune dysfunction: implications for health. Nature Reviews Immunology 5:243–251.

55. Reeves MB, Sinclair JH. 2013. Circulating Dendritic Cells Isolated from Healthy Seropositive Donors Are Sites of Human Cytomegalovirus Reactivation In Vivo. Journal of Virology 87:10660–10667.

56. Kim JY, Mandarino A, Chao M V, Mohr I, Wilson AC. 2012. Transient Reversal of Episome Silencing Precedes VP16- Dependent Transcription during Reactivation of Latent HSV-1 in Neurons. PLoS Pathog 8.

57. Chen T, Hudnall SD. 2006. Anatomical mapping of human herpesvirus reservoirs of infection. Modern Pathology 19:726–737.

58. Hendrix RM, Wagenaar M, Slobbe RL, Bruggeman CA. 1997. Widespread presence of cytomegalovirus DNA in tissues of healthy trauma victims. Journal of Clinical Pathology 50:59–63.

59. Harkins LE, Matlaf LA, Soroceanu L, Klemm K, Britt WJ, Wang W, Bland KI, Cobbs CS. 2010. Detection of human cytomegalovirus in normal and neoplastic breast epithelium. Herpesviridae 1:8.

60. Gordon CL, Miron M, Thome JJC, Matsuoka N, Weiner J, Rak MA, Igarashi S, Granot T, Lerner H, Goodrum F, Farber DL. 2017. Tissue reservoirs of antiviral T cell immunity in persistent human CMV infection. The Journal of Experimental Medicine 214:jem.20160758.

61. Ljungman P, Hakki M, Boeckh M. 2010. Cytomegalovirus in hematopoietic stem cell transplant recipients. Infectious Disease Clinics of North America 24:319–337.

62. Santos CAQ, Brennan DC, Yusen RD, Olsen MA. 2015. Incidence, Risk Factors and Outcomes of Delayed-onset Cytomegalovirus Disease in a Large Retrospective Cohort of Lung Transplant Recipients. Transplantation 99:1658–1666.

63. Saffert RT, Penkert RR, Kalejta RF. 2010. Cellular and Viral Control over the Initial Events of Human Cytomegalovirus Experimental Latency in CD34+ Cells. Journal of Virology 84:5594–5604.

64. Reeves MB, Davies AA, McSharry BP, Wilkinson GW, Sinclair JH. 2007. Complex I Binding by a Virally Encoded RNA Regulates Mitochondria-Induced Cell Death. Science 316:1345–1348.

65. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G,

1086 Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira
1087 PG, Ongen H, Rivas MA, Battle A, Mostafavi S, Monlong J, Sammeth M, Mele M,
1088 Reverter F, Goldmann JM, Koller D, Guigo R, McCarthy MI, Dermitzakis ET,
1089 Gamazon ER, Im HK, Konkashbaev A, Nicolae DL, Cox NJ, Flutre T, Wen X,
1090 Stephens M, Pritchard JK, Tu Z, Zhang B, Huang T, Long Q, Lin L, Yang J, Zhu J, Liu
1091 J, Brown A, Mestichelli B, Tidwell D, Lo E, Salvatore M, Shad S, Thomas JA,
1092 Lonsdale JT, Moser MT, Gillard BM, Karasik E, Ramsey K, Choi C, Foster BA, Syron
1093 J, Fleming J, Magazine H, Hasz R, Walters GD, Bridge JP, Miklos M, Sullivan S,
1094 Barker LK, Traino HM, Mosavel M, Siminoff LA, Valley DR, Rohrer DC, Jewell SD,
1095 Branton PA, Sobin LH, Barcus M, Qi L, McLean J, Hariharan P, Um KS, Wu S, Tabor
1096 D, Shive C, Smith AM, Buia SA, Undale AH, Robinson KL, Roche N, Valentino KM,
1097 Britton A, Burges R, Bradbury D, Hambright KW, Seleski J, Korzeniewski GE,
1098 Erickson K, Marcus Y, Tejada J, Taherian M, Lu C, Basile M, Mash DC, Volpi S,
1099 Struewing JP, Temple GF, Boyer J, Colantuoni D, Little R, Koester S, Carithers LJ,
1100 Moore HM, Guan P, Compton C, Sawyer SJ, Demchok JP, Vaught JB, Rabiner CA,
1101 Lockhart NC, Ardlie KG, Getz G, Wright FA, Kellis M, Volpi S, Dermitzakis ET. 2015.
1102 The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation
1103 in humans. Science 348:648–660.

1104 66. Cobbs CS, Matlaf L, Harkins LE. 2014. Human Cytomegaloviruses 1119:165–196.

1105 67. Sinzger C, Hahn G, Digel M, Katona R, Sampaio KL, Messerle M, Hengel H,
1106 Koszinowski U, Brune W, Adler B. 2008. Cloning and sequencing of a highly
1107 productive, endotheliotropic virus strain derived from human cytomegalovirus

TB40/E. Journal of General Virology 89:359–368.

68. O'Connor CM, Murphy EA. 2012. A Myeloid Progenitor Cell Line Capable of Supporting Human Cytomegalovirus Latency and Reactivation, Resulting in Infectious Progeny. Journal of Virology 86:9854–9865.

69. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Shuga J, Montesclaros L, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Ness KD, Beppu LW, Joachim Deeg H, McFarland C, Valente WJ, Ericson NG, Stevens EA, Radich JP, Hindson BJ, Bielas JH. 2016. Massively parallel digital transcriptional profiling of single cells 1 2. Phone 8:667–3170.

70. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods 9:357–9.

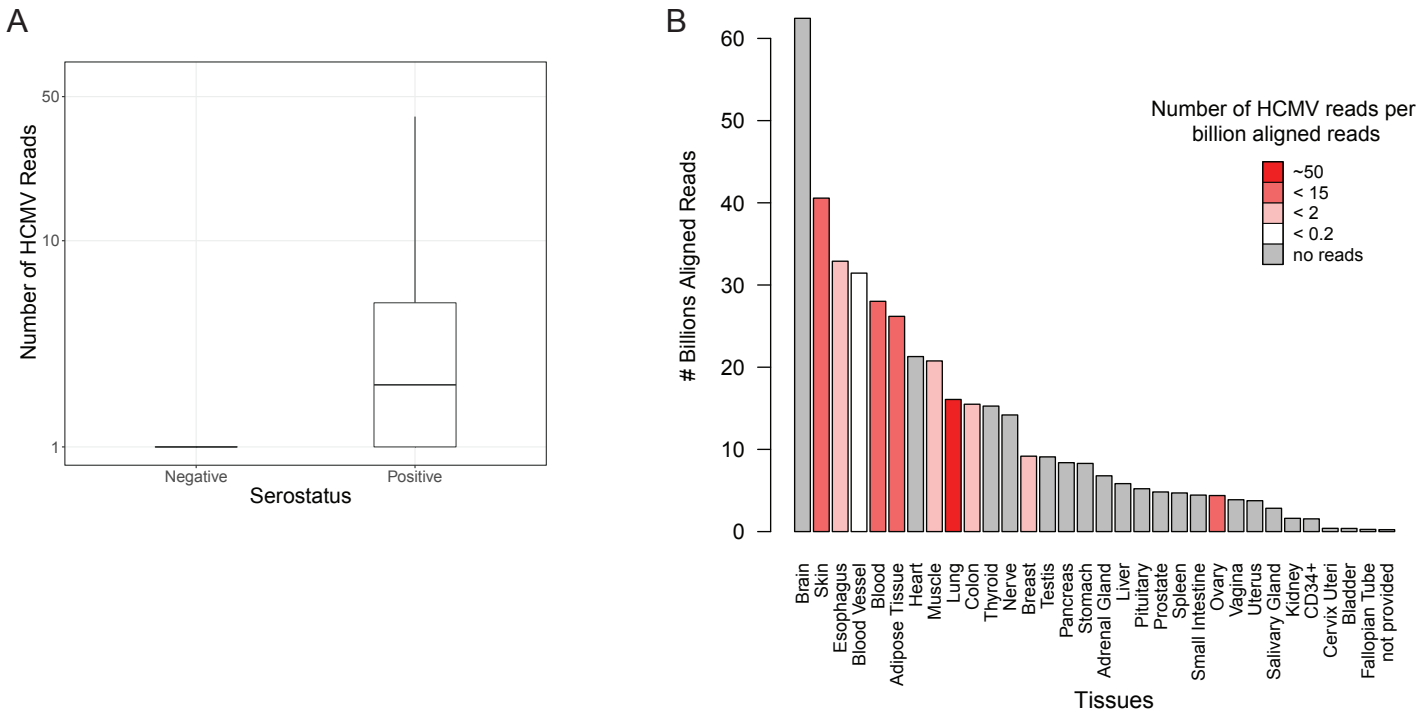
71. Rtsne.

72. Anders S, Pyl PT, Huber W. 2015. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169.

73. GENE-E.

74. 10x CellRanger 2.0.0.

Figure 1


A
Number of HCMV Reads
50
10
1
Negative
Positive
Serostatus
B
Billions Aligned Reads
60
50
40
30
20
10
0
Number of HCMV reads per billion aligned reads
~50
< 15
< 2
< 0.2
no reads
Brain
Skin
Esophagus
Blood Vessel
Blood
Adipose Tissue
Heart
Muscle
Lung
Colon
Thyroid
Nerve
Breast
Testis
Pancreas
Stomach
Adrenal Gland
Liver
Pituitary
Prostate
Spleen
Small Intestine
Ovary
Vagina
Uterus
Salivary Gland
Kidney
CD34+
Cervix Uteri
Bladder
Fallopian Tube
not provided
Tissues

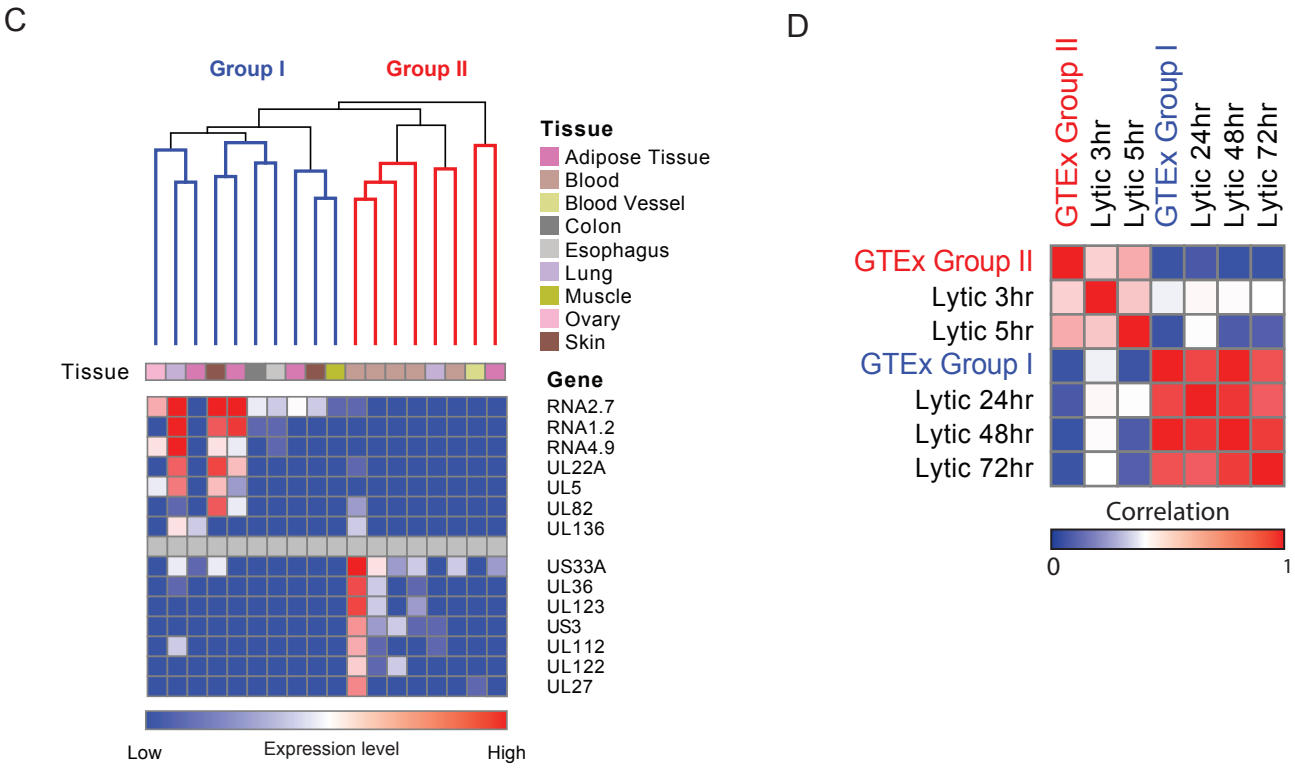

C
Group I
Group II
Tissue
Tissue
Adipose Tissue
Blood
Blood Vessel
Colon
Esophagus
Lung
Muscle
Ovary
Skin
Gene
RNA2.7
RNA1.2
RNA4.9
UL22A
UL5
UL82
UL136
US33A
UL36
UL123
US3
UL112
UL122
UL27
Low
Expression level
High
D
GTEx Group II
Lytic 3hr
Lytic 5hr
GTEx Group I
Lytic 24hr
Lytic 48hr
Lytic 72hr
Correlation
0
1

Figure 2

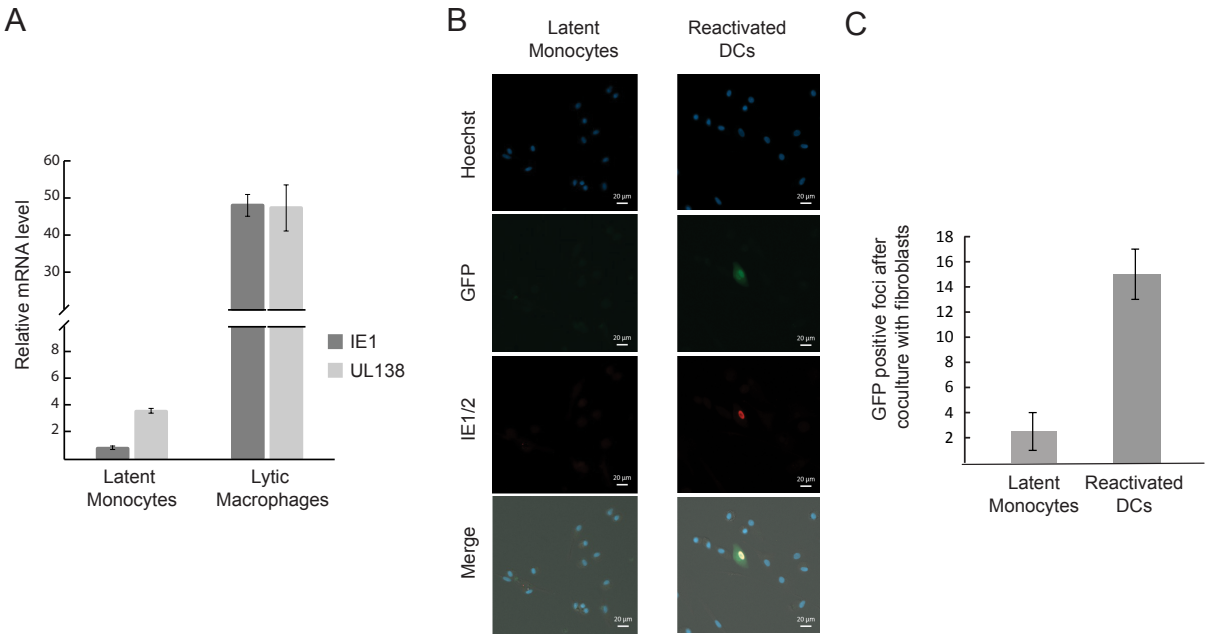

A
Relative mRNA level
60
50
40
30
8
6
4
2
IE1
UL138
Latent Monocytes
Lytic Macrophages
B
Latent Monocytes
Reactivated DCs
Hoechst
GFP
IE1/2
Merge
20 µm
C
GFP positive foci after coculture with fibroblasts
18
16
14
12
10
8
6
4
2
Latent Monocytes
Reactivated DCs

Figure 3

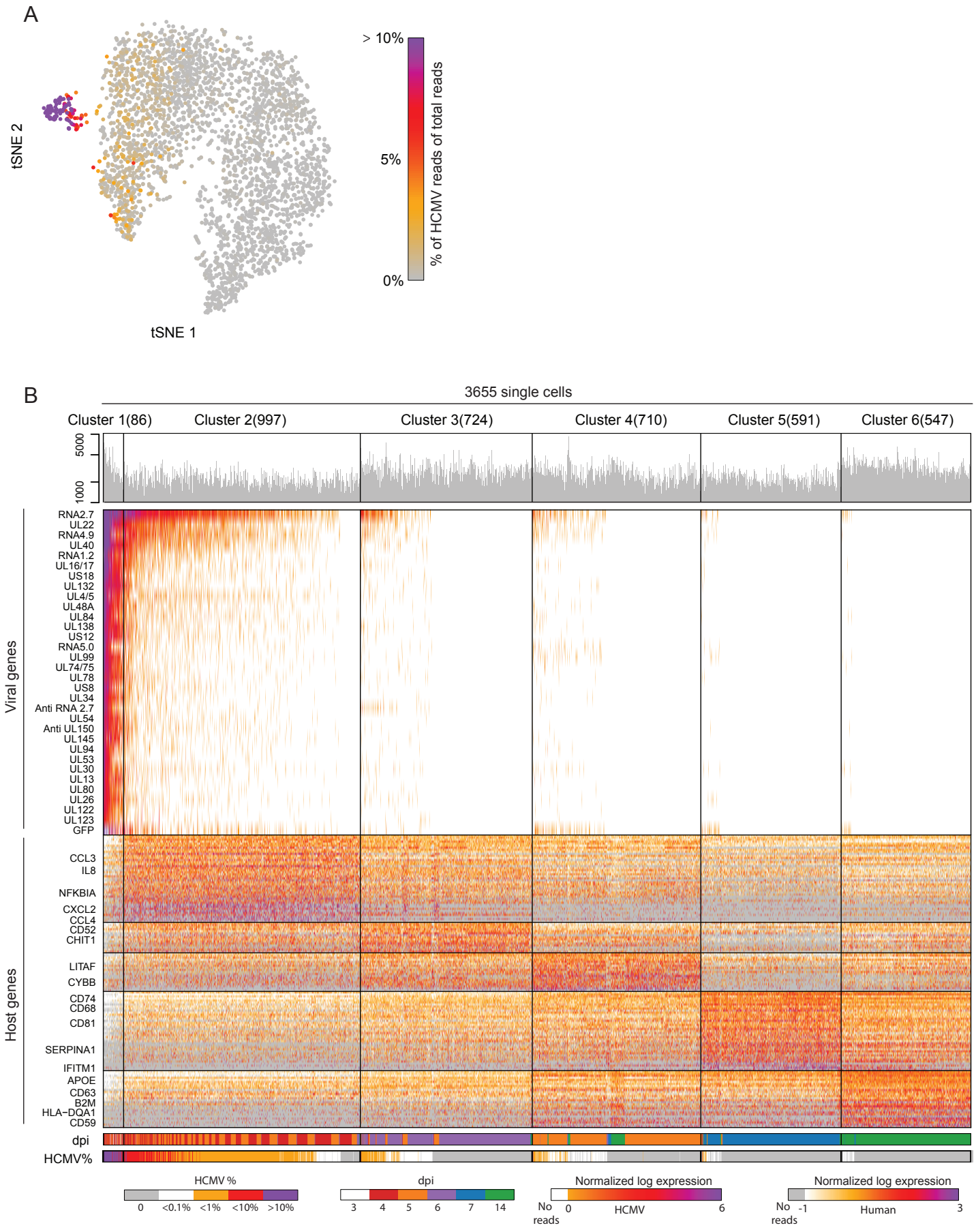

A
tSNE 2
tSNE 1
> 10%
5%
0%
% of HCMV reads of total reads
B
3655 single cells
Cluster 1(86)
Cluster 2(997)
Cluster 3(724)
Cluster 4(710)
Cluster 5(591)
Cluster 6(547)
5000
1000
Viral genes
RNA2.7
UL22
RNA4.9
UL40
RNA1.2
UL16/17
US18
UL132
UL4/5
UL48A
UL84
UL138
US12
RNA5.0
UL99
UL74/75
UL78
US8
UL34
Anti RNA 2.7
UL54
Anti UL150
UL145
UL94
UL53
UL30
UL13
UL80
UL26
UL122
UL123
GFP
Host genes
CCL3
IL8
NFKBIA
CXCL2
CCL4
CD52
CHIT1
LITAF
CYBB
CD74
CD68
CD81
SERPINA1
IFITM1
APOE
CD63
B2M
HLA−DQA1
CD59
dpi
HCMV%
HCMV %
0
<0.1%
<1%
<10%
>10%
dpi
3
4
5
6
7
14
Normalized log expression
No reads
0
HCMV
6
Normalized log expression
No reads
-1
Human
3

Figure 4

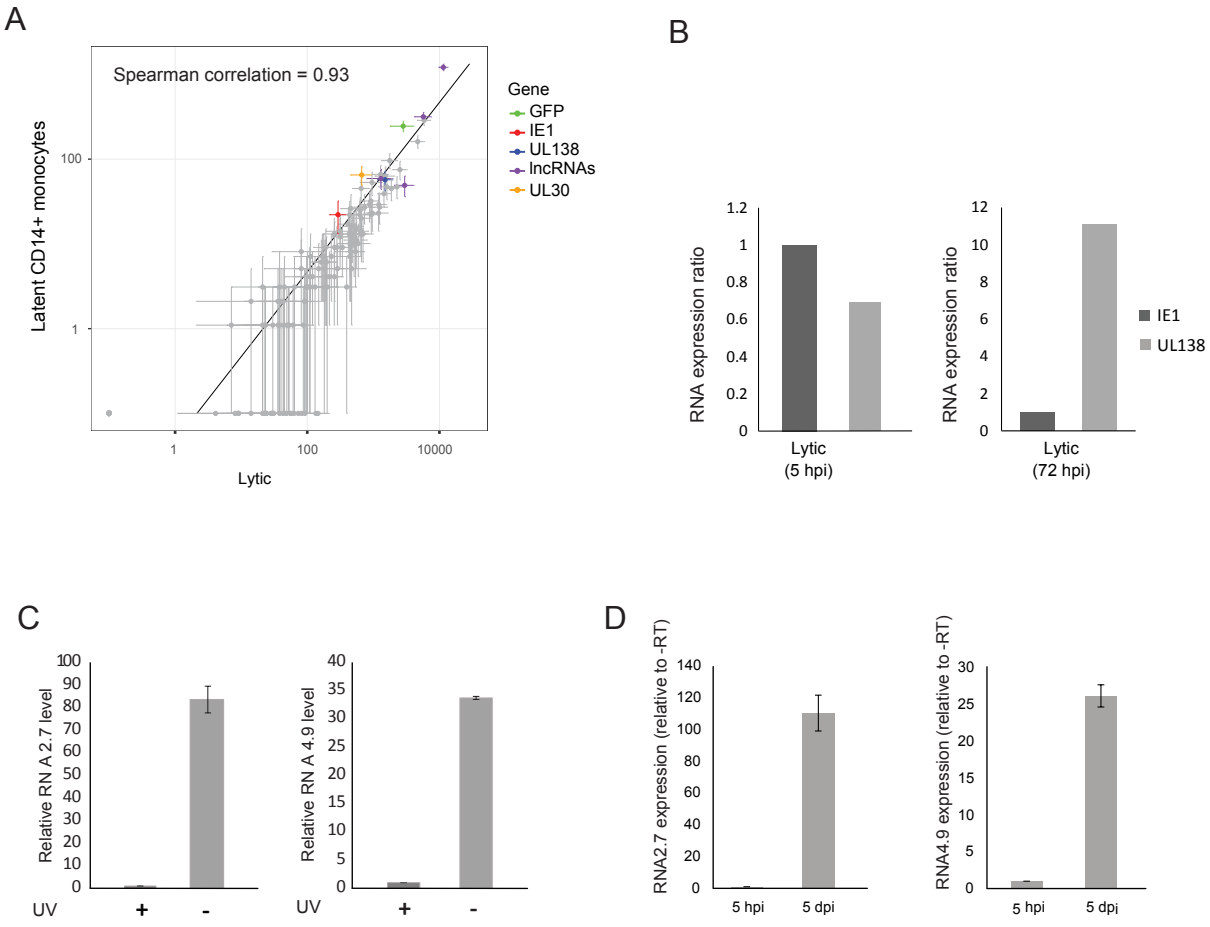

A
Spearman correlation = 0.93
Latent CD14+ monocytes
Lytic
Gene
GFP
IE1
UL138
lncRNAs
UL30
B
RNA expression ratio
Lytic
(5 hpi)
Lytic
(72 hpi)
IE1
UL138
C
Relative RN A 2.7 level
Relative RN A 4.9 level
UV
+
-
D
RNA2.7 expression (relative to -RT)
RNA4.9 expression (relative to -RT)
5 hpi
5 dpi

Figure 5

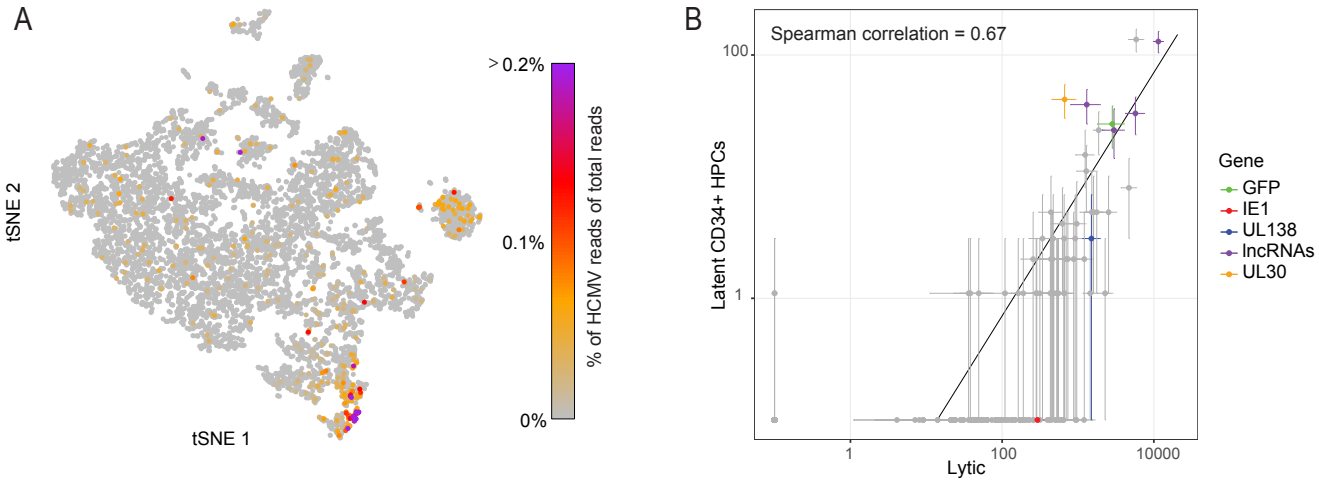
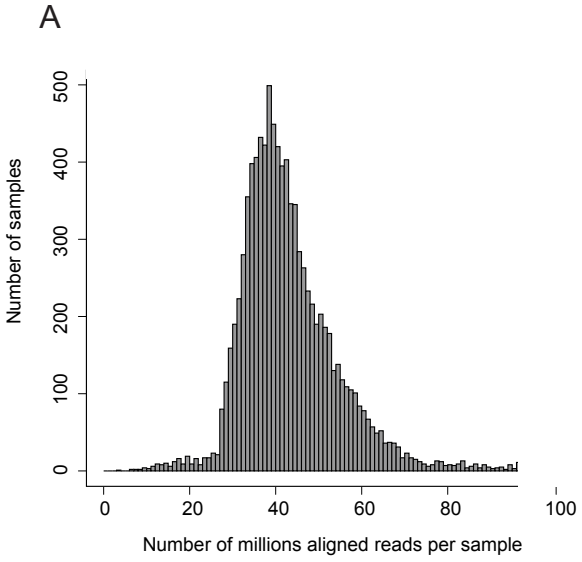
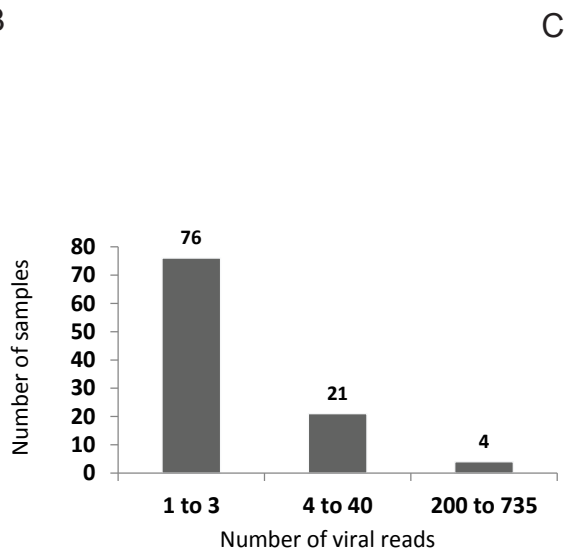
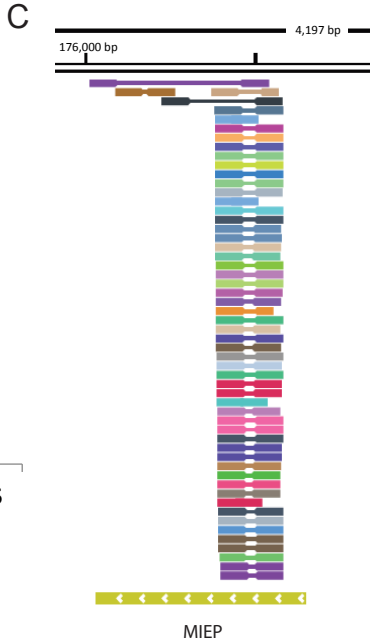
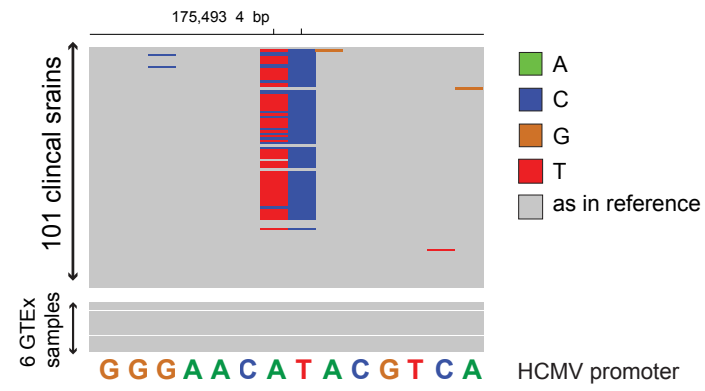

A
tSNE 2
tSNE 1
> 0.2%
0.1%
0%
% of HCMV reads of total reads
B
Spearman correlation = 0.67
Latent CD34+ HPCs
Lytic
100
1
1
100
10000
Gene
GFP
IE1
UL138
lncRNAs
UL30

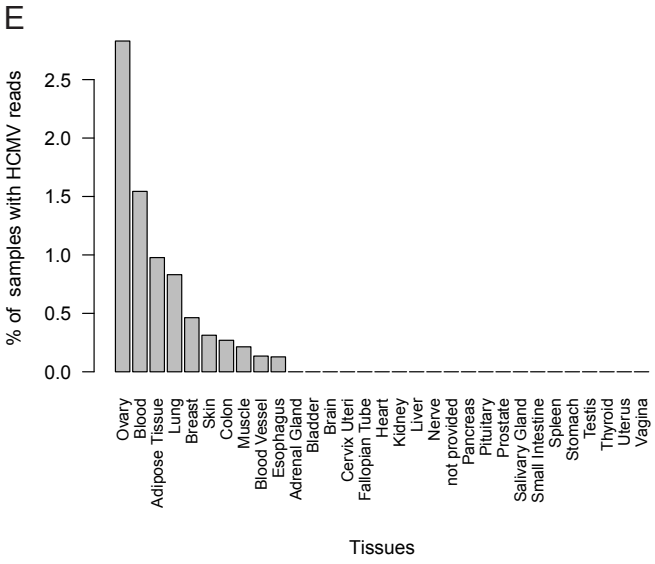
Figure S1


A
Number of samples
0
100
200
300
400
500
0
20
40
60
80
100
Number of millions aligned reads per sample


B
Number of samples
80
70
60
50
40
30
20
10
0
76
21
4
1 to 3
4 to 40
200 to 735
Number of viral reads


C
4,197 bp
176,000 bp
MIEP


D
175,493 4 bp
101 clincal srains
6 GTEx samples
G G G A A C A T A C G T C A
A
C
G
T
as in reference
HCMV promoter


E
% of samples with HCMV reads
0.0
0.5
1.0
1.5
2.0
2.5
Ovary
Blood
Adipose Tissue
Lung
Breast
Skin
Colon
Muscle
Blood Vessel
Esophagus
Adrenal Gland
Bladder
Brain
Cervix Uteri
Fallopian Tube
Heart
Kidney
Liver
Nerve
not provided
Pancreas
Pituitary
Prostate
Salivary Gland
Small Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Tissues

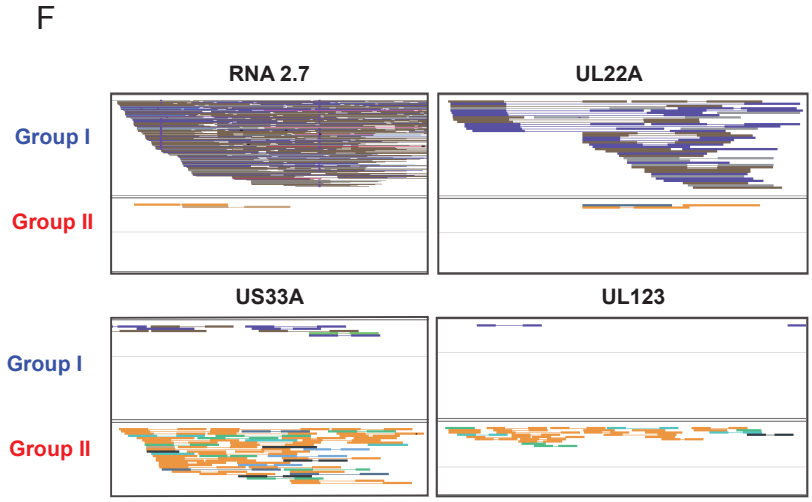

F
RNA 2.7
UL22A
US33A
UL123
Group I
Group II
Group I
Group II

Figure S2

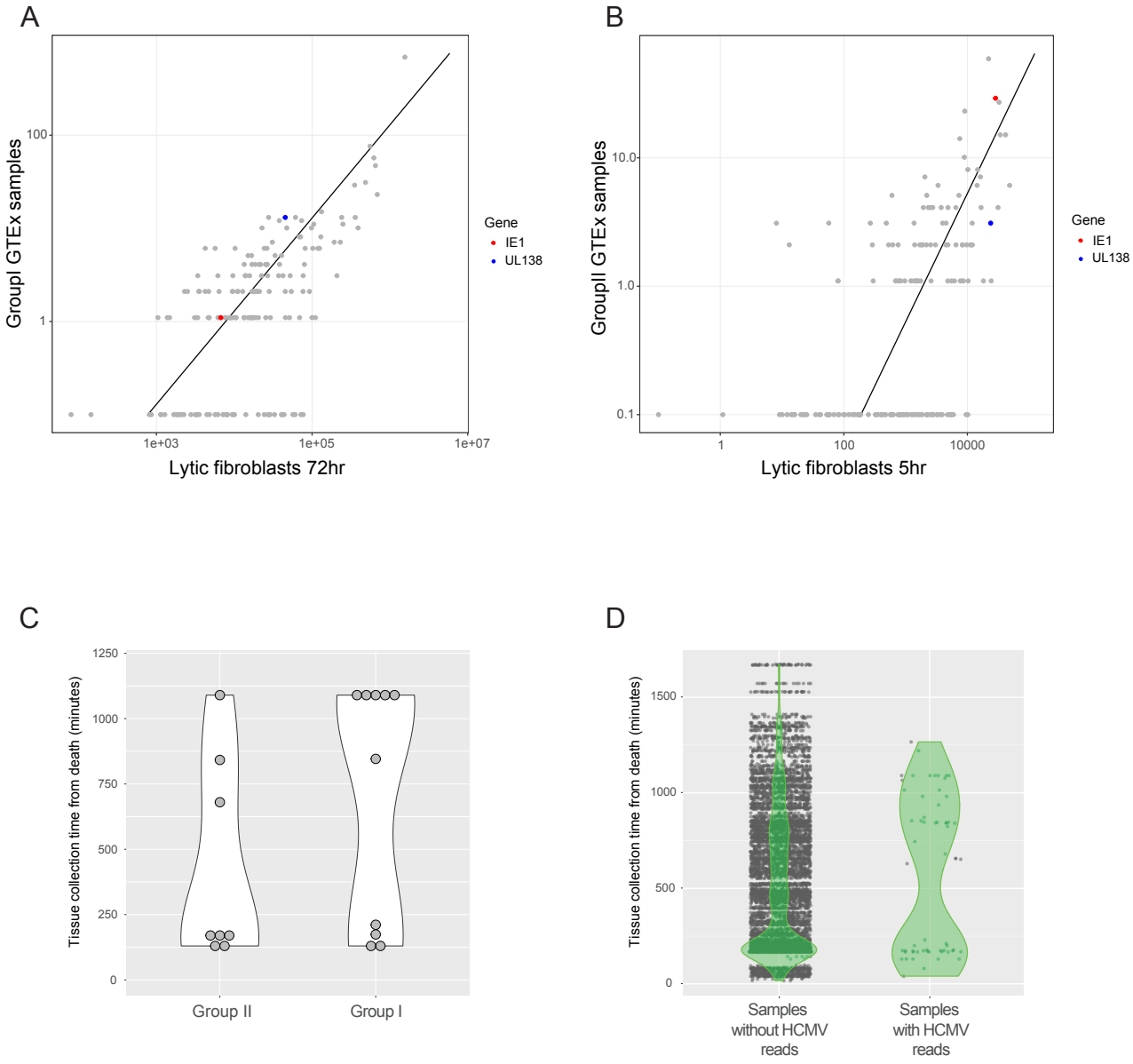
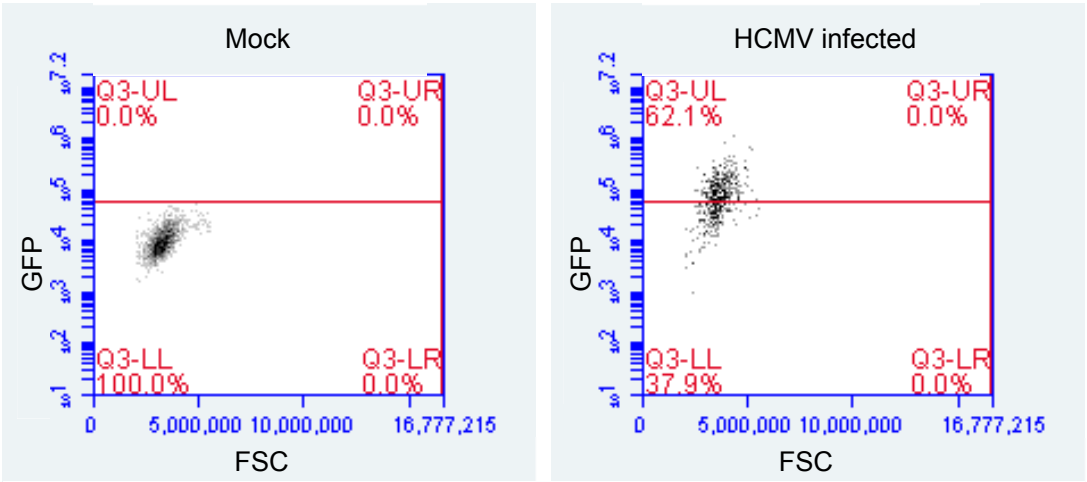
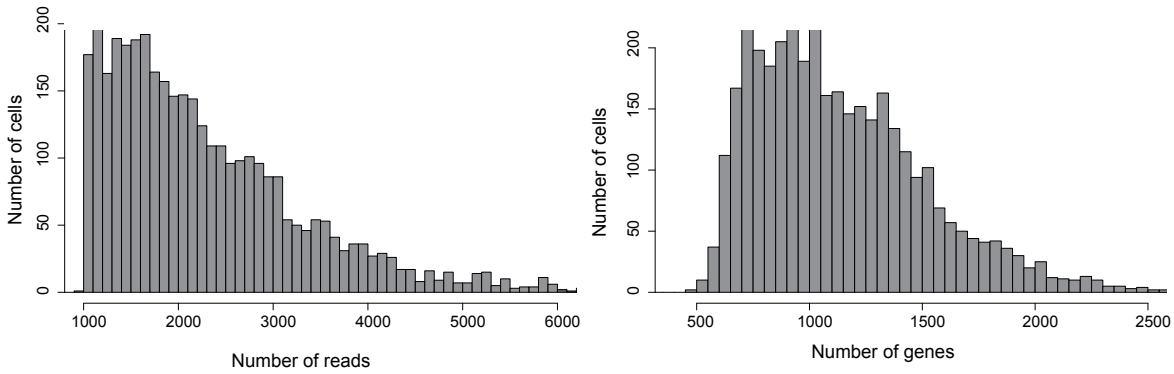

A
GroupI GTEx samples
100
1
1e+03
1e+05
1e+07
Lytic fibroblasts 72hr
Gene
IE1
UL138
B
GroupII GTEx samples
10.0
1.0
0.1
1
100
10000
Lytic fibroblasts 5hr
Gene
IE1
UL138
C
Tissue collection time from death (minutes)
1250
1000
750
500
250
0
Group II
Group I
D
Tissue collection time from death (minutes)
1500
1000
500
0
Samples without HCMV reads
Samples with HCMV reads

Figure S3

A



B



C

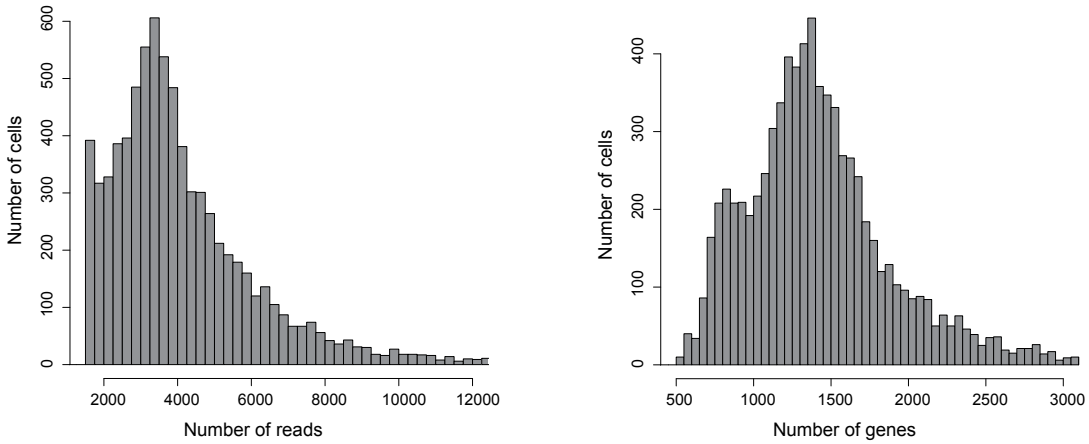


Figure S4

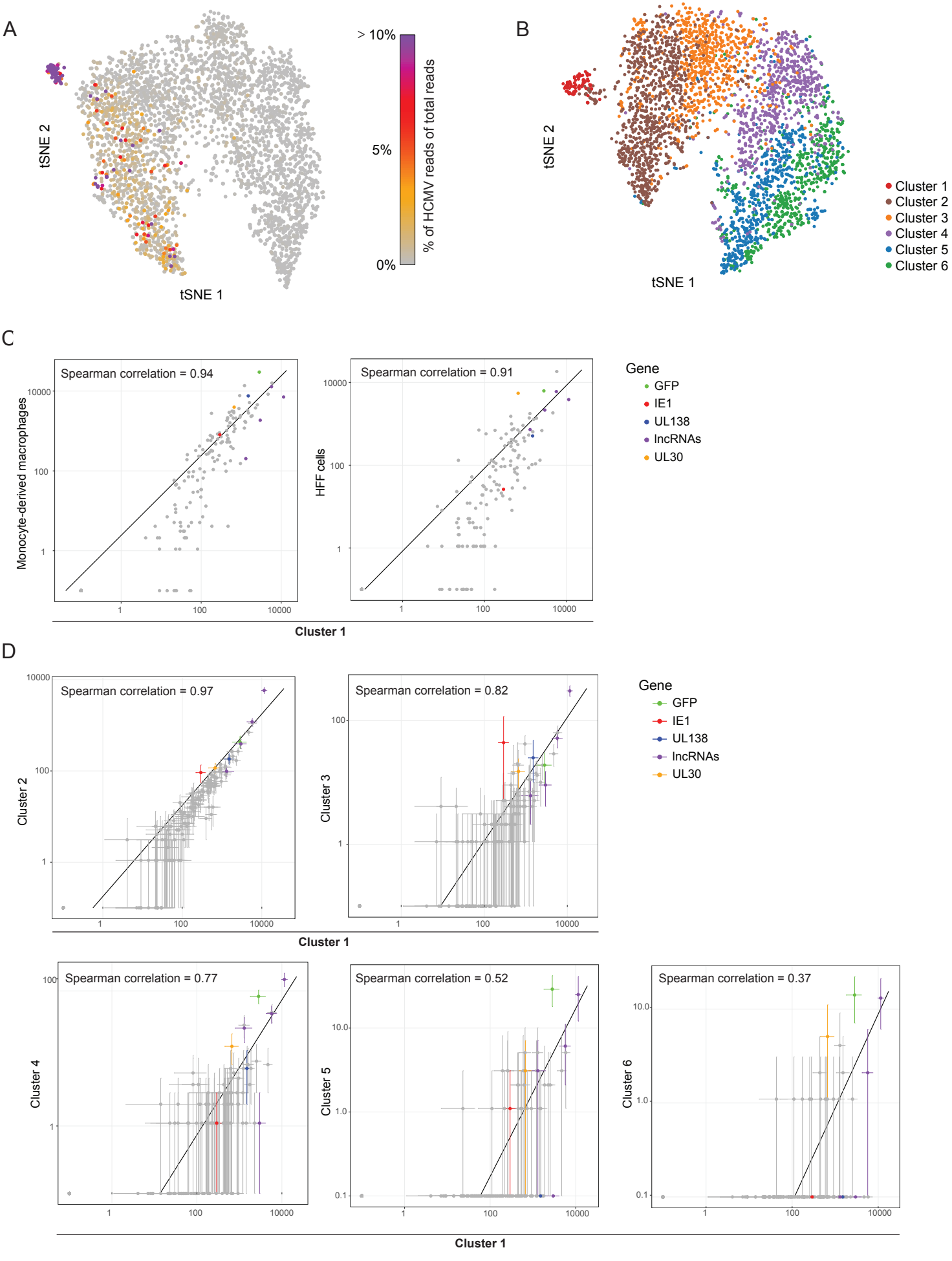

A
tSNE 2
tSNE 1
> 10%
5%
0%
% of HCMV reads of total reads
B
tSNE 2
tSNE 1
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
C
Spearman correlation = 0.94
Spearman correlation = 0.91
Monocyte-derived macrophages
HFF cells
Cluster 1
Gene
GFP
IE1
UL138
lncRNAs
UL30
D
Spearman correlation = 0.97
Spearman correlation = 0.82
Spearman correlation = 0.77
Spearman correlation = 0.52
Spearman correlation = 0.37
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 1

Figure S5

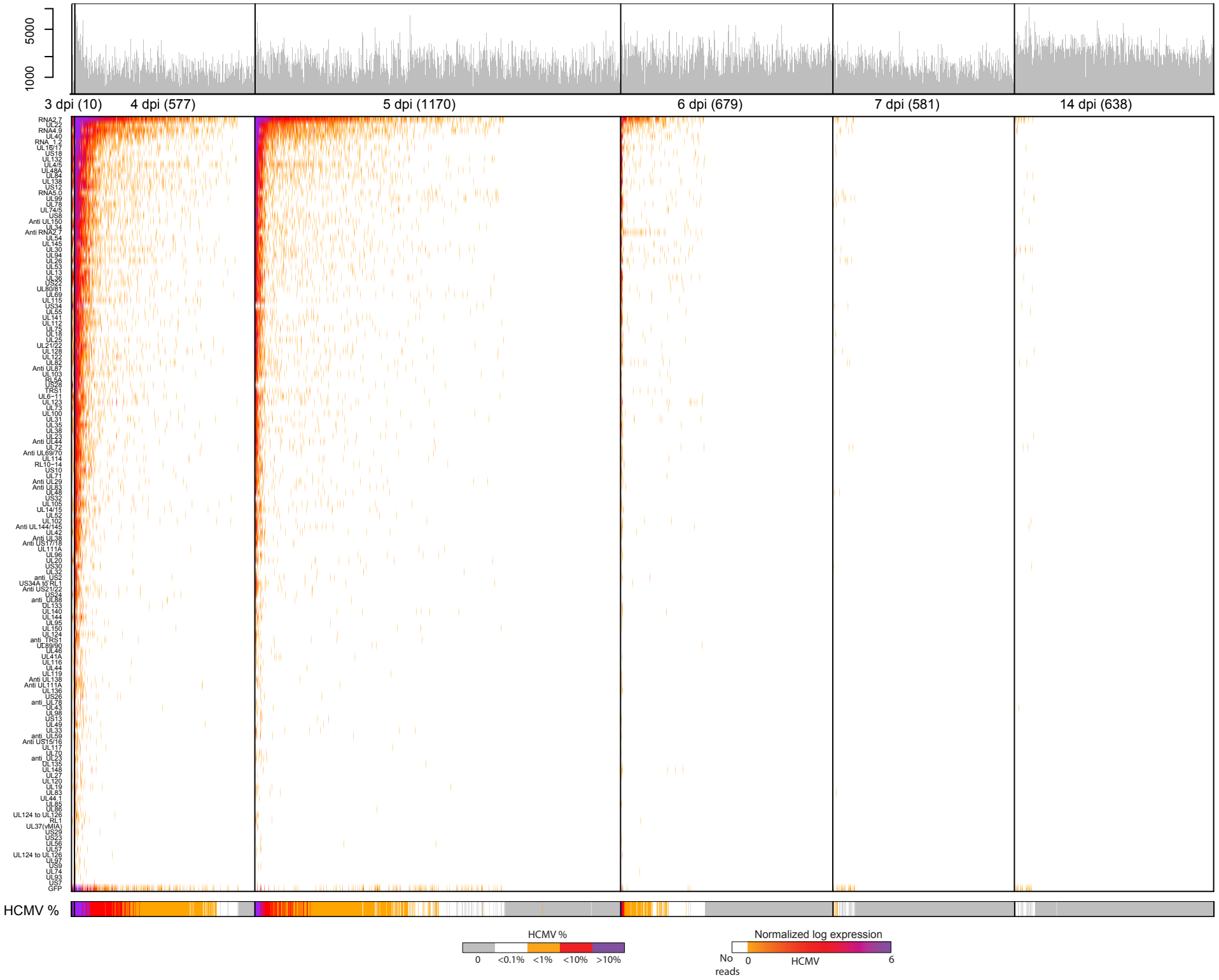

5000
1000
3 dpi (10)
4 dpi (577)
5 dpi (1170)
6 dpi (679)
7 dpi (581)
14 dpi (638)
RNA2.7
UL22
RNA4.9
UL40
RNA 1.2
UL18/17
US18
UL132
UL4/5
UL48A
UL84
UL138
US12
RNA5.0
UL99
UL78
UL74/5
US8
Anti UL150
UL34
Anti RNA2.7
UL54
UL145
UL30
UL94
UL26
UL53
UL13
UL36
US22
UL80/81
UL69
UL115
US34
UL55
UL141
UL112
UL75
UL18
UL25
UL21/22
UL128
UL122
UL82
Anti UL87
UL103
UL5A
US28
TRS1
UL6~11
UL123
UL73
UL100
UL31
UL35
UL38
UL23
Anti UL44
UL72
Anti UL69/70
UL114
RL10~14
US10
UL71
Anti UL29
Anti UL83
UL48
US32
UL105
UL14/15
UL52
UL102
Anti UL144/145
UL42
Anti UL38
Anti US17/18
UL111A
UL96
UL20
US30
UL32
anti_US2
US34A to RL1
Anti US21/22
US24
anti_UL88
DL133
UL140
UL144
UL95
UL150
UL124
anti_TRS1
UL89/90
UL46
UL41A
UL116
UL44
UL119
Anti UL138
Anti UL111A
UL136
US26
anti_UL78
UL43
UL98
US13
UL49
UL33
anti_UL59
Anti US15/16
UL117
UL70
anti_UL23
DL135
UL148
UL27
UL120
UL19
UL83
UL44.1
UL85
UL86
UL124 to UL126
RL1
UL37(vMIA)
US29
US23
UL56
UL57
UL124 to UL126
UL97
US9
UL74
UL93
US7
GFP
HCMV %
HCMV %
0
<0.1%
<1%
<10%
>10%
Normalized log expression
No reads
0
HCMV
6

Figure S6

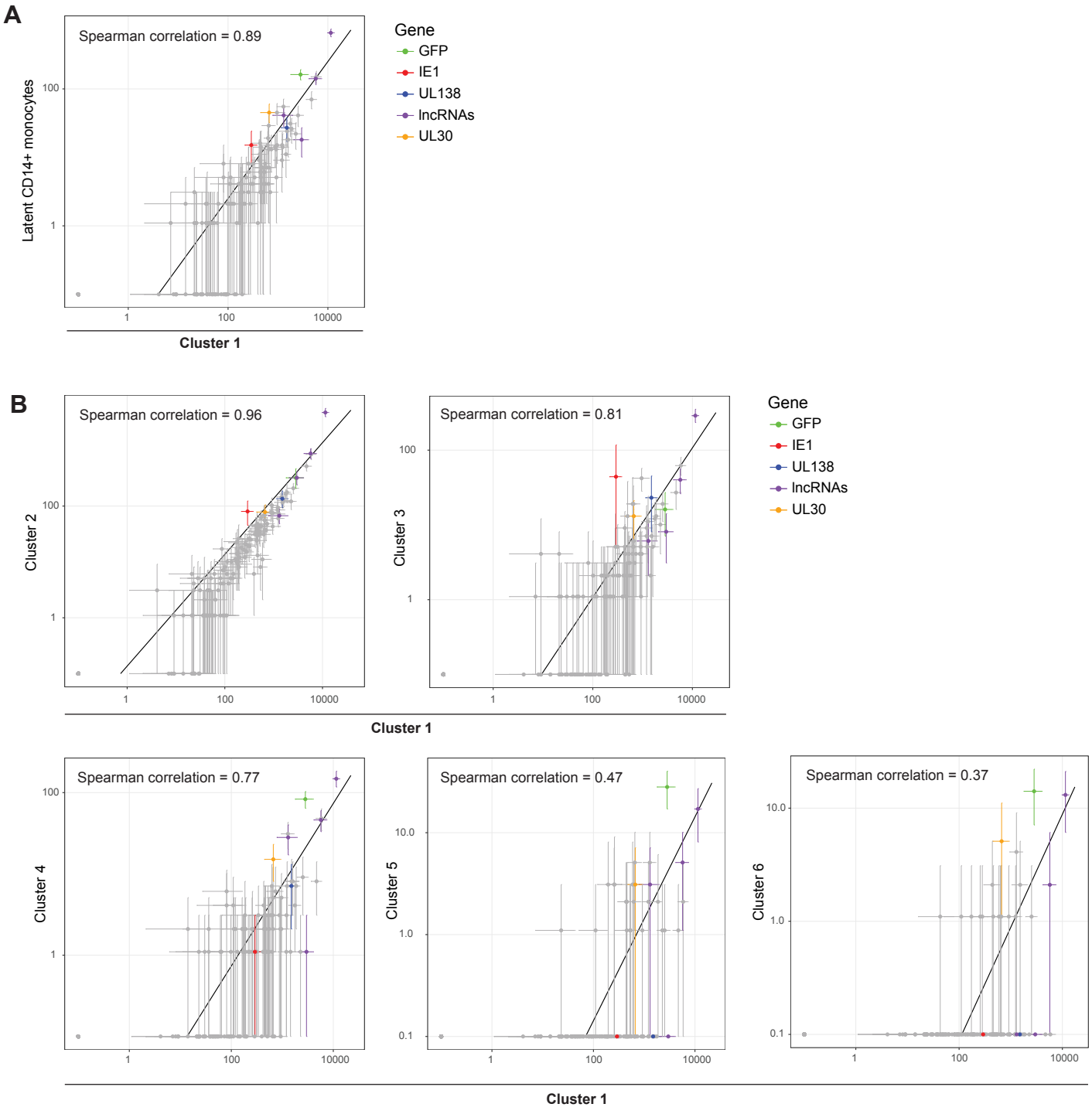
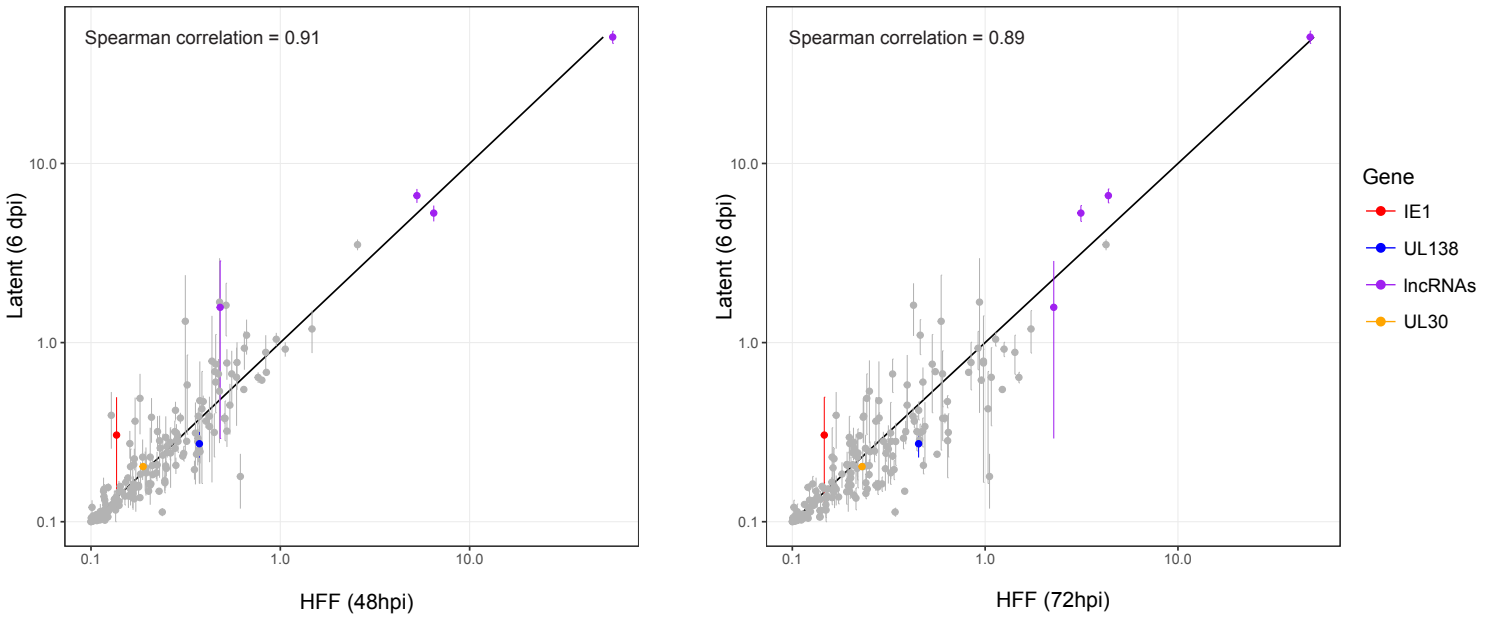

A
Spearman correlation = 0.89
Latent CD14+ monocytes
Cluster 1
Gene
GFP
IE1
UL138
lncRNAs
UL30
B
Spearman correlation = 0.96
Cluster 2
Spearman correlation = 0.81
Cluster 3
Cluster 1
Spearman correlation = 0.77
Cluster 4
Spearman correlation = 0.47
Cluster 5
Spearman correlation = 0.37
Cluster 6
Cluster 1

Figure S7


Spearman correlation = 0.91
Latent (6 dpi)
10.0
1.0
0.1
0.1
1.0
10.0
HFF (48hpi)
Spearman correlation = 0.89
Latent (6 dpi)
10.0
1.0
0.1
0.1
1.0
10.0
HFF (72hpi)
Gene
IE1
UL138
lncRNAs
UL30