JVI Journals.ASM.org

Deep Sequencing of HIV-Infected Cells: Insights into Nascent Transcription and Host-Directed Therapy

Xinxia Peng,[a,b] Pavel Sova,[a,b] Richard R. Green,[a,b] Matthew J. Thomas,[a,b] Marcus J. Korth,[a,b] Sean Proll,[a,b] Jiabao Xu,[c] Yanbing Cheng,[c] Kang Yi,[c] Li Chen,[c] Zhiyu Peng,[c,d] Jun Wang,[c] Robert E. Palermo,[a,b] Michael G. Katze[a,b]

Department of Microbiology, University of Washington, Seattle, Washington, USA[a]; Washington National Primate Research Center, Seattle, Washington, USA[b]; BGI-Shenzhen, Shenzhen, China[c]; BGI-Guangzhou, Guangzhou, China[d]

ABSTRACT

Polyadenylated mature mRNAs are the focus of standard transcriptome analyses. However, the profiling of nascent transcripts, which often include nonpolyadenylated RNAs, can unveil novel insights into transcriptional regulation. Here, we separately sequenced total RNAs (Total RNAseq) and mRNAs (mRNAseq) from the same HIV-1-infected human CD4$^+$ T cells. We found that many nonpolyadenylated RNAs were differentially expressed upon HIV-1 infection, and we identified 8 times more differentially expressed genes at 12 h postinfection by Total RNAseq than by mRNAseq. These expression changes were also evident by concurrent changes in introns and were recapitulated by later mRNA changes, revealing an unexpectedly significant delay between transcriptional initiation and mature mRNA production early after HIV-1 infection. We computationally derived and validated the underlying regulatory programs, and we predicted drugs capable of reversing these HIV-1-induced expression changes followed by experimental confirmation. Our results show that combined total and mRNA transcriptome analysis is essential for fully capturing the early host response to virus infection and provide a framework for identifying candidate drugs for host-directed therapy against HIV/AIDS.

IMPORTANCE

In this study, we used mass sequencing to identify genes differentially expressed in CD4$^+$ T cells during HIV-1 infection. To our surprise, we found many differentially expressed genes early after infection by analyzing both newly transcribed unprocessed pre-mRNAs and fully processed mRNAs, but not by analyzing mRNAs alone, indicating a significant delay between transcription initiation and mRNA production early after HIV-1 infection. These results also show that important findings could be missed by the standard practice of analyzing mRNAs alone. We then derived the regulatory mechanisms driving the observed expression changes using integrative computational analyses. Further, we predicted drugs that could reverse the observed expression changes induced by HIV-1 infection and showed that one of the predicted drugs indeed potently inhibited HIV-1 infection. This shows that it is possible to identify candidate drugs for host-directed therapy against HIV/AIDS using our genomics-based approach.

Recently, we reported the first transcriptome deep sequencing (RNAseq) analysis of a CD4$^+$ T cell line infected with HIV-1 (1). We observed both the dramatic expansion of viral mRNA expression and the widespread differential expression of host genes. Particularly, we observed a striking discordance between a small set (~1% of detected genes) of differentially expressed (DE) host genes and the large amount of viral RNAs (~20% of total mappable reads) present at 12 h postinfection (hpi), the earliest time point studied. Given the large quantity of viral RNA detected, this apparent silencing of the host transcriptional response at 12 hpi is intriguing, considering that the expression of multiple host transcription factors is already significantly altered at 12 hpi (1).

In that study, we adopted mRNAseq, the typical application of RNAseq focusing on polyadenylated [poly(A)+] mature mRNAs through the use of poly(T) priming (1), which is also standard for microarray analysis. By its design, mRNAseq leaves out the nonpolyadenylated [poly(A)−] fraction of the mammalian transcriptome, which includes many noncoding RNAs (ncRNAs) and transcripts known to encode proteins such as histones (2, 3). A more recent report from the ENCODE project shows that poly(A)− transcripts can be found in most protein-coding genes (4). Also, studies sequencing total RNAs or subcellular fractions of total RNAs have shown the capture of nascent pre-mRNAs in the poly(A)− fraction of the mammalian transcriptome (5, 6). Further, recent studies show that a clearly detectable lag exists between pre-mRNA and mRNA production in response to stimuli such as macrophage activation by lipopolysaccharide (LPS) (6) or tumor necrosis factor (TNF) (7) and epithelial cell stimulation by epidermal growth factor (8). Also, splicing is indicated as a contributing factor to the delay between pre-mRNA and mRNA production (6, 7), and HIV-1 infection can modulate host RNA splicing machineries (9). Together, we reasoned that early host transcriptional changes may be more evident by measuring nascent pre-mRNAs than mature mRNAs.

Total RNAseq, an alternative application of RNAseq, directly

Received 17 March 2014 Accepted 17 May 2014
Published ahead of print 21 May 2014
Editor: G. Silvestri
Address correspondence to Michael G. Katze, honey@uw.edu.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.00768-14.

doi:10.1128/JVI.00768-14

sequences total RNA (depleted of rRNA), therefore covering both the poly(A)+ and poly(A)− fractions of the transcriptome. Compared to mRNAseq, Total RNAseq generates many more short sequencing reads originating from regions outside known exons, i.e., introns or intergenic regions (3, 5, 10). Normally, the large amount of reads mapped to introns significantly complicates important RNAseq applications such as mature mRNA assembly and isoform quantification. However, these intronic reads offer a distinct benefit of detecting nascent transcription, i.e., incompletely spliced and nonpolyadenylated immature transcripts (5, 6). Despite the complementarities between mRNAseq and Total RNAseq, existing studies still use either of two RNAseq approaches, due to factors such as high cost and analytical needs. Since the real impact of choosing a single RNAseq approach for specific biological questions is unknown, it is valuable to contrast mRNAseq and Total RNAseq to see if broader early host transcription changes could be detected at the pre-mRNA than at the mRNA level.

To further investigate the early host response to HIV-1 infection, including poly(A)− ncRNAs and nascent pre-mRNAs, in this study we performed Total RNAseq analysis of the same HIV-1-infected $CD4^+$ T cell samples as those that we previously used for mRNAseq analysis (1). Here we show that many poly(A)− transcripts were differentially expressed during HIV-1 infection, which were accurately detected by Total RNAseq but not mRNAseq. Surprisingly, we identified much broader (about 8-fold-greater) host transcriptional changes at 12 hpi by Total RNAseq than mRNAseq. We found that the initiation of the early response to HIV-1 infection was largely independent of viral replication, as treatment of cells with nonreplicating HIV-1 virions induced transcriptional changes of similar trends. We also identified over 1,000 long ncRNAs differentially expressed during HIV-1 infection. The systematic characterization of early transcriptional changes, by integrating large-scale transcription factor chromatin immunoprecipitation (CHiP)-seq data and a human tissue mRNAseq expression compendium, enabled predictions of underlying regulators such as transcription factors and long ncRNAs. We then identified drugs capable of reversing the early transcriptional changes induced by HIV-1 infection and utilized the reversed drug expression profiles to refine regulator predictions. With these predicted regulators, we computationally derived regulatory programs for the induction of early transcriptional changes from a compendium of published expression data and validated their predictability. Finally, we showed experimentally that lycorine, one of the drugs predicted by our computational analyses, potently inhibited HIV-1 infection of $CD4^+$ T cells.

MATERIALS AND METHODS

Cell culture, viral infection, and drug treatment. Infection of the human $CD4^+$ T cell line SUP-T1 with HIV-1 strain LAI, cell and virus propagation, and sample collection for RNAseq were described in reference 1. Briefly, SUP-T1 cells were obtained from the American Type Culture Collection (CRL-1942) and propagated in RPMI 1640 medium (Gibco) supplemented with 10% fetal bovine serum (HyClone), penicillin (100 U/ml), streptomycin (100 g/ml), and GlutaMAX-I. HIV-1 strain LAI (catalog no. 2522) was obtained from the NIH AIDS Research and Reference Reagent Program (Germantown, MD) and propagated in SUP-T1 cells. U373-MAGI-CXCR4CEM cells were obtained from M. Emerman through the AIDS Research and Reference Reagent Program to test virus stock titers (11). Typical titers reached 10^7 infectious units per ml. Infections were carried out at a multiplicity of infection (MOI) of 5 and performed in triplicate. Mock-infected samples received SUP-T1 cell conditioned medium and were also performed in triplicate. Inactivation of HIV-1 by UV irradiation was described in reference 12. Briefly, UV-inactivated HIV-1 was generated by irradiating HIV preparations for 5 min, a dose we found sufficient to abrogate viral replication in U373-MAGI-CXCR4 cells and in SUP-T1 cells as detected by viral mRNA load TaqMan quantitative reverse transcription-PCR (qPCR) (13), with GAPDH (glyceraldehyde-3-phosphate dehydrogenase) transcript serving as an internal control (forward primer, 5′GGCCTCCAAGGAGTAAGACC3′; reverse primer, 5′AGGGGTCTACATGGCAACTG3′). For qPCR assays of long noncoding RNAs (lncRNAs), an independent time course of infections of SUP-T1 with HIV-1 LAI was carried out as described in reference 14. The infectious doses in both infections were optimized to achieve 100% infected cells at 24 hpi with ~50% cell viability as measured by trypan blue exclusion assay. Infected cells were visualized by immunofluorescence assay with rabbit HIV-1SF2 p24 antiserum kindly provided by BioMolecular Technologies through the AIDS Research and Reference Reagent Program.

To test the effect of the drugs predicted to reverse HIV-1-induced expression changes, we first infected 1×10^6 SUP-T1 cells with HIV-1 LAI at an MOI of 0.1 for 1 h. After washing away the virus inoculum, cells were suspended in medium containing dimethyl sulfoxide (DMSO) or lycorine and cultured for 24 h. Cells were pelleted, and total RNA was isolated and reversely transcribed into cDNA (using the QuantiTect kit; Qiagen), and intracellular viral RNA was quantified as described previously (13), with GAPDH transcript serving as an internal control (forward primer, 5′GGCCTCCAAGGAGTAAGACC3′; reverse primer, 5′AGGGGTCTACATGGCAACTG3′). Cells were also deposited on a microscopic slide, and immunofluorescent Gag staining was carried out as described previously (1).

For influenza virus infections of a polarized human bronchial epithelial cell line (Calu-3), three avian-origin influenza A viruses (IAVs) isolated from fatal human cases, i.e., strains A/Anhui/01/2013 (H7N9) (here Anhui01), A/Netherland/219/2003 (H7N7) (here NL219), and A/Vietnam/1203/2004 (H5N1) (here VN1203), and a seasonal human virus, A/Panama/2007/1999 (H3N2) (here Pan99), were grown in the allantoic cavities of 10-day-old embryonated hen's eggs for 24 to 28 h at 37°C for avian-origin viruses or for 48 h at 34°C for the H3N2 virus. Allantoic fluids from multiple eggs were pooled, clarified by centrifugation, aliquoted, and stored at −70°C. The propagation, polarization, and infection of Calu-3 cells were carried out as described in references 15 and 16. Calu-3 cell sample collection and RNA isolation were described in reference 16. All research with avian viruses was conducted under biosafety level 3 containment, including enhancements required by the U.S. Department of Agriculture and the Select Agent Program (http://www.cdc.gov/od/ohs/biosfty/bmbl5/bmbl5toc.htm).

Transcriptome sequencing analysis. For Total RNAseq analysis, approximately 20 μg of total RNA from each sample was submitted to the Beijing Genomics Institute (BGI) for sequencing. In brief, rRNAs were depleted using the RiboMinus human/mouse transcriptome isolation kit (Invitrogen, CA). rRNA-depleted RNAs were fragmented, and cDNA synthesis was primed using random hexamers. Short fragments were purified for an average insert size of 200 nucleotides (nt) and then ligated with proprietary adapters. The (2 × 90)-nt paired-end sequencing was done using an Illumina HiSeq 2000. The mRNAseq data were acquired using the exact same samples as reported in reference 1.

Read mapping and differential expression analysis. The read mapping was carried out essentially as described in reference 1. Briefly, we mapped short reads to human ribosomal sequences to remove potential rRNA sequences using the short-read aligner software Bowtie (17). We then mapped the remaining unmapped reads to the HIV genome (GenBank accession no. K02013) using the gapped aligner software TopHat (18), which predicts HIV splicing junctions and maps intron-spanning reads to known splicing junctions. To quantify transcript expression, we mapped all reads that remained unmapped to the human reference genome (hg19, build GRCh37, downloaded from the UCSC genome

browser [http://genome.ucsc.edu]) using TopHat. RefSeq transcript annotations were supplied to facilitate the mapping of reads spanning known splicing junctions. Gene level quantification was obtained using HT-seq (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html). The differential expression analysis was performed using edgeR (19) separately for Total RNAseq and mRNAseq data sets. For each data set, for the downstream analyses, we kept only those genes with at least 10 raw read counts in at least three biological samples and which were considered detected in this study. Clustering and other statistical analyses were performed using R (http://www.r-project.org/). For visualization, BAM files were generated using TopHat and SAMtools (20) and displayed using the UCSC Genome Browser.

Compilation of custom human genome annotation. We created a custom human genome annotation using a hybrid approach. First, we combined a published catalog of human long noncoding RNAs (21) with a reference human annotation recently released by the ENCODE project (Gencode v17). Independent of these known annotations, we reconstructed transcripts in each sample, using Cufflinks (22) based on reads mapped to the human genome, separately for mRNAseq and Total-RNAseq data sets. Only uniquely mapped reads were used, as this was meant to complement existing annotation. Transcripts assembled from individual samples were merged together using Cuffmerge. All newly assembled transcripts were checked against the known annotations as described above using the tool Cuffcompare from Cufflinks, and only those assembled transcripts with the class code "u" (unknown, intergenic transcript) were added into the combined known annotations as predicted novel transcripts.

Identification of genes preferably detected by Total RNAseq relative to mRNAseq. To investigate if we captured nonpolyadenylated transcripts through Total RNAseq analysis, we compared for the same gene in the same sample the read counts from Total RNAseq analysis to that from mRNAseq analysis. We reasoned that if nonpolyadenylated transcripts were transcribed from a gene, Total RNAseq analysis would consistently collect more short reads than mRNAseq analysis. To facilitate the comparison, first, for each sample by each RNAseq analysis, the raw gene read counts were scaled by the total gene read counts. Next, for each gene, we counted the number of samples in which the scaled read count from Total RNAseq analysis was 1.5-fold greater or more than that from the corresponding mRNAseq analysis. A gene was considered enriched in Total RNAseq data if it had more (1.5-fold or higher) reads in more than one-half of the total samples.

qPCR. RNA was reverse transcribed using the QuantiTect reverse transcription kit (Qiagen, Valencia, CA), and the resulting cDNA was diluted 50 times. Primers for gene targets of interest were designed using the Primer 3 program (http://frodo.wi.mit.edu/primer3). PCR was run on an ABI Prism 7900HT sequence detection system in triplicate per each sample and target. The relative change in transcript abundance was calculated using the $\Delta\Delta C_T$ method (where C_T is the threshold cycle) with GAPDH as an internal gene reference. We inspected the expression change of GAPDH in our RNAseq data, and the expression change was negligible. We also checked against a second calibrator (IRF-3) that did not change expression upon HIV-1 infection based on both the RNAseq data collected here and an unpublished microarray data set, and the results were identical to those derived from the GAPDH-controlled experiment. Intracellular viral RNA load was quantified as previously described (13). The expression changes of a set of 46 host genes were quantified using qPCR. These genes were previously selected spanning a range of values for the validation of the results from mRNAseq analysis, but only the results from 34 of 46 genes were reported (1).

Transcription factor binding site enrichment analysis with ChIP-seq data. ChIP-seq identified transcription factor (TF) binding sites from the ENCODE project (23) were downloaded from the UCSC genome browser (http://genome.ucsc.edu). A TF was considered found in the promoter region of a gene if one of its binding sites was located within the window from 1,000 bp upstream to 100 bp downstream of its annotated transcriptional start site. Considering that the ENCODE ChIP-seq data were a pool of data from different cell types, we limited our analysis to those genes that had at least one Pol II binding site located in their promoter regions, indicating that they were actually transcribed in the assayed cells. This requirement also ensured that the annotations of transcriptional start sites were less likely incorrect. A hypergeometric test was used to test if the binding sites of a TF were enriched in the promoters of a list of differentially expressed genes. For the list of genes differentially expressed at 12 hpi, we limited our analysis to protein-coding genes, as they were mostly (over 85%) protein-coding genes and the differences in the frequencies of gene biotypes between this set and the rest could bias the analysis. For lncRNAs, the enrichment analysis was performed for all differentially expressed lncRNAs (12 and 24 hpi results combined), as the list for 12 hpi was too small. Also, a similar analysis was done for the list of lncRNAs that were not differentially expressed during HIV-1 infection, in order to filter out TFs for which the enrichment might be due to differences in gene biotypes. Three lncRNA categories were excluded from the analysis, i.e., 3prime_overlapping_ncrna, antisense, sense_intronic, and sense_overlapping, because their genomic overlapping with other annotated genes introduced ambiguities in assigning binding sites.

Functional enrichment analysis. Functional analysis was performed using Ingenuity Pathways Analysis (IPA; Ingenuity Systems, Inc.). The Ensembl gene identifier was used to map each gene to its corresponding molecule in the Ingenuity Pathways Knowledge Base, a curated repository of biological interactions and functional annotations. The *P* value associated with a function or a pathway was calculated using the right-tailed Fisher exact test. For all analyses, IPA-generated *P* values were adjusted using the Benjamini-Hochberg multiple testing correction. Enriched functions from IPA analysis were custom processed to summarize redundant entries, i.e., different functions containing the same or similar sets of genes, or a subset of genes in other functions. Specifically, a similarity measurement between two functions was defined as the ratio between the number of overlapping genes and the total number of genes in one of two functions with a smaller number of genes. The corresponding distance was defined as 1 minus the calculated similarity measure. Hierarchical clustering analysis was performed on the all-against-all distance matrix using the single-linkage clustering method to group IPA functional categories into function groups. The most significant *P* value among all functions within each function group was defined as the *P* value for the function group. These function groups were further filtered if the genes in a function group were mostly covered by more-significant (smaller *P* value) function groups. A similar summarization and filtering process was done for enriched canonical pathways.

Association of long ncRNA with functions enriched in genes differentially expressed at 12 hpi. We downloaded the alignment files of the mRNAseq read alignment of 24 human tissues and cell types from the Human lincRNA Catalog website (http://www.broadinstitute.org/genome_bio/human_lincrnas/?q=home) and then quantified gene expression in each tissue in the same way as we did for the RNAseq data collected here. We added mRNAseq read counts collected from our control samples (the average of 3 mock-infected replicates from12 hpi as a single data column) representing T cells to the human tissue collection. Only those genes with at least 10 raw read counts in at least 3 tissues were kept for the downstream analysis. The normalized read count, in counts per million (cpm), for each tissue was obtained using edgeR (19). After normalization, we further limited the following correlation analysis to those genes with nonzero entries across at least 60% of tissues. The normalized read cpms were $\log_2$ transformed before the calculation of correlations, which was done using the bicor function provided by the WGCNA package (24). For each lncRNA differentially expressed at 12 hpi, we ranked all other genes by their correlation coefficients with the lncRNA. The ranked list was analyzed using the GseaPreranked tool provided by gene set enrichment analysis (GSEA) (25) to obtain the functions that were highly correlated with the lncRNA of interest. For GSEA analysis, the human gene symbol was used to map each gene to the corresponding genes in the reference func-

tion annotation database, here the function groups enriched in 12 hpi DE genes, which were derived from the IPA analysis as described above and the gene modules as described below.

cmap database. To search for drugs that reversed the early expression changes induced by HIV-1 infection, we used the publicly available Connectivity Map (cmap) database (build 02) (26). cmap is a collection of genome-wide transcriptional data from cultured human cells treated with 1,309 different compounds. In cmap, the basic unit of data is defined as an instance, which is a pair of a single treatment of a compound and the corresponding control and the list of genes ordered by their extent of differential expression between the treatment and control. The same compound can be tested multiple times (under the same or different conditions), each of which is considered an instance. Given a query signature, i.e., a list of up- and downregulated genes, a value between +1 and −1 (called the connectivity score) is calculated for each of the instances in cmap. A high positive connectivity score indicates that the corresponding compound induced the expression of the query signature. A high negative connectivity score indicates that the corresponding compound reversed the expression of the query signature. Instances are rank ordered by descending connectivity score. The top-ranked instance (i.e., a score of +1) is said to be the most positively connected with the query signature. The bottom-ranked instance (i.e., a score of −1) is the most negatively connected with the query signature. For each compound, cmap calculates a measure of the enrichment of its instances in either end of the order list of all instances in cmap and a permutation *P* value for that enrichment score. By default, cmap returns a list of the top 20 compounds best connected (positively and negatively) with the query signature, ordered in ascending order of *P* values and then ascending order of (absolute) enrichment. We used the 500 most upregulated and 500 most downregulated genes at 12 hpi for a query signature. Genes were mapped to Affymetrix HG-U133A probe sets using the Affymetrix NetAffy batch query tool to query the cmap database.

Gene module construction, regulatory model learning, and validation. We manually queried NCBI GEO for human microarray data sets related to HIV infection and/or $CD4^+$ T cells. To minimize potential technical complications, we kept only data sets using the same microarray platform, Affymetrix Human Genome U133 Plus 2.0 array (accession number GPL570 in GEO). For each data set, we downloaded the raw cel files and reprocessed the data in the same way using the "rma" function in R/affy package (27). Probe set to gene mapping was obtained using the Affymetrix NetAffy batch query tool, and probe sets mapped to the same gene were averaged. Within each data set, we identified one control condition (untreated, uninfected, time point zero, etc.) and converted the expression measures to $\log_2$ ratios between each sample and the mean of the control samples to focus on the effects of perturbations and to minimize potential technical differences across different data sets. This also facilitated the downstream interpretation. We also averaged technical replicates and replicates of cell line samples to maximally retain biological variations.

To find gene modules, we combined the obtained $\log_2$ ratios from each data set and performed weighted coexpression network analysis using the R package WGCNA (24). We used the "signed" network with the value of 6 for power and biweight midcorrelation "bicor" for correlation calculation. For each gene model, we then learned predictive models from the compiled expression data. To derive predictive models, we used a machine learning procedure called "elastic net," which is implemented by the R package "glmnet" (28). For each gene module, the predictor variables were a set of candidate predicators/regulators, and the response variable was the median expression changes of genes for each condition. In the case of 10 identified TFs as candidate regulators, we fitted a linear regression model (family "gaussian") with the lasso penalty (alpha = 1), as there were no strong correlations among predictor variables. The fitting procedure generates a set of selected regulators and the corresponding regression coefficients for the linear regression model. In the case of the expanded set of 115 TFs and lncRNAs as candidate regulators, we fitted a linear regression model with the elastic net penalty. We compiled the expanded list of candidate regulators based on the Gene Ontology (TF, GO:0003700) and Gencode (lncRNA) annotation, and we excluded the ones that were not differentially expressed at 12 hpi during HIV infection. For each gene module, we searched a series of values between 0 and 1 for alpha and chose the one that gave the minimum mean cross-validation error. Here we used 10-fold cross-validation, i.e., the samples were randomly partitioned into 10 sets, of which 9 sets were used to learn a predictive model, which was subsequently used to blindly predict the outcome in the 10th set. This process was repeated iteratively 10 times, and the model with the minimum mean cross-validation error was identified and subjected to the predictability assessment. To ensure that the learning and evaluation were robust, we first randomly divided the 173 conditions/samples into 10 subsets and learned 10 predictive models by leaving out one of the subsets each time. Then we used the learned model to predict the overall expression changes of the gene module during HIV-1 infection.

Nucleotide sequence accession numbers. The RNAseq data from this publication have been submitted to NCBI's GEO database (http://www.ncbi.nlm.nih.gov/geo) and assigned the identifier GSE53993.

RESULTS

Total RNAseq uncovers nonpolyadenylated transcripts differentially expressed in HIV-1-infected $CD4^+$ T cells. Previously, we infected SUP-T1 cells, a human $CD4^+$ T lymphoblast cell line, with HIV-1 LAI under conditions that gave a synchronous infection in ~100% of the cells, and we collected samples at 12 and 24 h postinfection (hpi) (1). To extend our prior mRNAseq analysis of poly(A)+ transcripts into less-studied poly(A)− transcripts, we sequenced total RNAs (Total RNAseq) from the same set of samples plus additional samples from cells treated with UV-inactivated virions (see Table S1 in the supplemental material). Accordingly, we expanded human gene annotation by combining the recent annotation provided by ENCODE (29), a published catalog of human large intergenic noncoding RNAs (lincRNAs, one category of long ncRNAs) (21), and unannotated intergenic transcripts reconstructed from this RNAseq data using Cufflinks (22). As shown in Table 1, over 40% of genes in the expanded annotation encode ncRNAs and intergenic transcripts, allowing us to better cover less-studied ncRNAs along with well-annotated protein-coding genes.

We then quantified host gene expression changes during HIV-1 infection, separately using mRNAseq and Total RNAseq (see Materials and Methods). In total, 11,094 genes were differentially expressed (false-discovery rate [FDR] adjusted *P* value, <0.05) at one or more time points following HIV infection, by either or both of the RNAseq methods (Table 1). Due to the much deeper (2.2-fold-higher, on average) sequencing coverage of host nonribosomal transcripts by mRNAseq, the expression of 2,159 DE genes was detected only by mRNAseq. Yet, there were still 165 DE genes detected only by Total RNAseq, indicating that these genes produced exclusively poly(A)− transcripts. Not surprisingly, these genes were enriched with ncRNAs and intergenic transcripts while the majority of DE genes detected only by mRNAseq were protein-coding genes (Fig. 1A).

Since many human genes, including protein-coding ones, are found in both poly(A)+ and poly(A)− RNA fractions (4), we investigated for which genes the measuring of both RNA fractions (by Total RNAseq) would improve the detection of differentially expressed genes. To this end, we first identified genes for which Total RNAseq tended to generate consistently higher expression abundances than the parallel mRNAseq analysis across the same samples,

TABLE 1 Summary of the expanded human gene annotation and the numbers of differentially expressed genes identified

Gene biotype[a]	Annotation	No. of differentially expressed genes identified in cells infected with:					
		Intact HIV			UV-inactivated HIV		
		Total	12 hpi	24 hpi	Total	12 hpi	24 hpi
Coding	20,317	8,570	1,660	8,189	1,118	826	425
Intergenic	359	248	64	234	51	44	17
lncRNA	16,927	1,098	98	1,074	45	29	18
sncRNA	8,527	29	4	26	6	0	6
Pseudogene	14,138	1,144	107	1,115	96	27	74
Other	841	5	0	5	1	1	0
Total	61,109	11,094	1,933	10,643	1,317	927	540

[a] Gene biotype was based on the GENCODE classification. For lncRNA, processed_transcript, lincRNA,3prime_overlapping_ncrna, antisense, sense_intronic, sense_overlapping. For sncRNA, snRNA, snoRNA, misc_RNA, miRNA. For pseudogene, polymorphic_pseudogene, pseudogene, IG_V_pseudogene, TR_V_pseudogene. For "other," IG_V_gene, TR_C_gene, TR_J_gene, TR_V_gene, rRNA.

an indication of the existence of poly(A)− transcripts. We found a set of 3,264 genes (of 17,316 genes detected overall) that had expression abundances 1.5-fold greater or higher in Total RNAseq than the corresponding mRNAseq across more than 50% (6/11) of available samples, hence considered enriched by Total RNAseq. Compared to the rest of detected genes, it contained relatively fewer (62% versus 75%) protein-coding genes, and more (23% versus 12%) ncRNAs or intergenic genes, including known poly(A)− transcribed genes such as TERC and RMRP genes (see Fig. S1 in the supplemental material). Next, we examined the differential expression *P* values of DE genes in

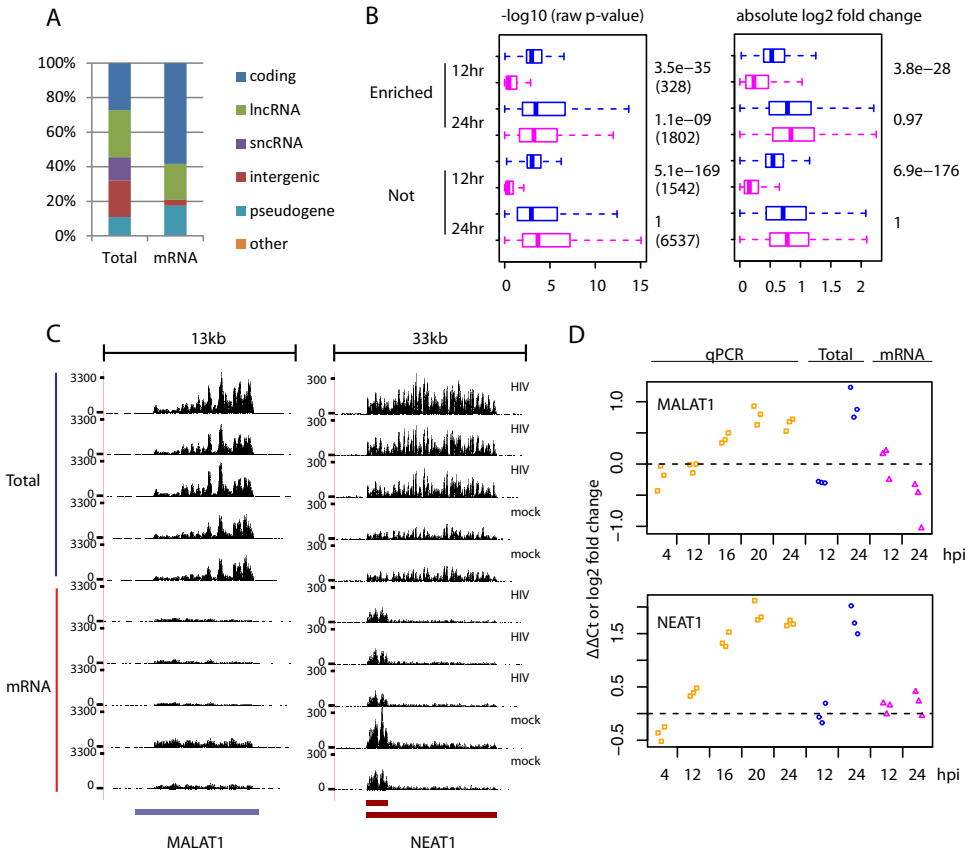

FIG 1 Comparison of Total RNAseq versus mRNAseq detection of poly(A)− transcripts. Only genes from the list of 11,094 DE genes are included here. (A) Percentages of gene biotypes whose expressions were detected only by either Total RNAseq (Total) or mRNAseq (mRNA). The classification of gene biotypes is shown in Table 1. (B) Metrics (left, $-\log_{10}$ raw *P* values; right, absolute $\log_2$-fold changes) of differential expression obtained from Total RNAseq (blue) versus mRNAseq (pink) data separately for genes enriched (Enriched) versus those not enriched (Not) by Total RNAseq. The numbers on the right side show the *P* values from Wilcoxon tests comparing metrics from Total RNAseq versus mRNAseq analysis. For the boxplot, the whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range from the box. (C) Raw read coverage (*y* axis) across the MALAT1 and NEAT1 genes (*x* axis) in samples collected at 24 hpi. (D) Temporal expression changes of MALAT1 and NEAT1 measured by qPCR and expression changes obtained by the two RNAseq methods. For qPCR, a mixture of oligo(dT) and random primers was used for cDNA synthesis.

this Total RNAseq enriched set. Interestingly, these DE genes still tended to have better *P* values in Total RNAseq than mRNAseq data, even though their expression was detected by both RNAseq methods (Fig. 1B). For some genes, such as NEAT1 and MALAT1 genes, that are known to be transcribed as poly(A)− transcripts, we observed clear discrepancies in expression changes between Total RNAseq and mRNAseq measurements, and independent analyses by qPCR agreed with Total RNAseq (Fig. 1C and D). In summary, these results convincingly demonstrated for the first time that many poly(A)− host transcripts were differentially expressed during HIV-1 infection in $CD4^+$ T cells. The first application of combined Total RNAseq and mRNAseq enabled the discovery of those genes that were differentially expressed during HIV-1 infection but not adequately addressed by mRNAseq due to poly(A)− transcripts.

Total RNAseq reveals broad early host transcriptional changes not detected by mRNAseq. Similar to our previous report (1), even with this expanded annotation we still identified only a small set (220) of DE genes at 12 hpi by mRNAseq. Unexpectedly, we found many more (1,801) DE genes at 12 hpi by Total RNAseq. Several comparisons showed that this difference was not a technical artifact. First, at 12 and 24 hpi, Total RNAseq measurements agreed well (Pearson correlation coefficient, *r*, 0.90 to 0.98) with separate qPCR measurements over a large set of genes (46 genes) that were selected independent of Total RNAseq data (see Fig. S2A in the supplemental material). Second, at 24 hpi, the overall expression changes of DE genes agreed between mRNAseq and Total RNAseq (see Fig. S2B in the supplemental material). These findings indicate that the 12 hpi expression changes detected by Total RNAseq are accurate.

We categorized the 1,933 DE genes identified at 12 hpi sequentially into four exclusive subgroups: subgroup a, genes with an adjusted *P* value of <0.05 by mRNAseq; subgroup b, genes with a raw *P* value of <0.2 by mRNAseq; subgroup c, genes enriched in Total RNAseq as described above; and subgroup d, the rest (Fig. 2A). Subgroups a and b were intended to cover DE genes found by mRNAseq (a) or likely found by mRNAseq (b), and subgroup c was for genes inherently not well detected by mRNAseq, e.g., poly(A)− transcripts. Interestingly, for both subgroups a and b, the expression changes at 12 hpi measured by mRNAseq versus by Total RNAseq were highly consistent (Fig. 2B). Again, this agreement verifies the significant expression changes at 12 hpi detected by Total RNAseq.

To investigate the reasons for the apparent discrepancies between mRNAseq and Total RNAseq for subgroup d (Fig. 2B), we counted reads mapped to introns for each gene and performed differential expression analysis in the same way as we did for reads mapped to exons. We reasoned that the additional immature transcripts captured by Total RNAseq could drive the observed differences (5, 6). Intriguingly, for subgroup d, the expression changes measured by reads mapped to introns were positively correlated ($r = 0.6$) with the expression changes measured by reads mapped to exons (Fig. 2C). In addition, 12 hpi expression changes were positively correlated ($r = 0.7$) with expression changes at 24 hpi. This indicates that for subgroup d, transcriptional changes that occurred at 12 hpi did not affect the amount of mature mRNAs until a later time point such as 24 hpi (Fig. 2D). A similar trend was also evident for subgroup b, suggesting that mature mRNAs in this subgroup were affected by transcriptional changes observed at 12 hpi by Total RNAseq, but not as significantly as for genes in subgroup a (Fig. 2C).

To investigate if these subgroups represent distinct biological functions, we identified biological functions significantly enriched in all 12 hpi DE genes. For each enriched function, we then tabulated the number of genes from each subgroup. In total, we found 14 major biological functions enriched in 12 hpi DE genes (see Table S2 in the supplemental material). For most of these 14 enriched functions, genes spread across multiple subgroups, but with different degrees of relative contributions from each subgroup. For example, subgroup a had relatively more genes related to T-cell differentiation, while genes associated with mitochondrial dysfunction or CTLA4 signaling tended to be from subgroup d (see Fig. S4 in the supplemental material). This suggests that genes across different subgroups were part of the same biological processes but regulated distinctly as illustrated by different patterns of expression changes. In summary, by contrasting Total RNAseq and mRNAseq measurements, we show that thousands of host genes in $CD4^+$ T cells had undergone significant transcriptional changes as early as 12 hpi, but for most of these genes, changes in the abundance of mature mRNAs were not detected until 24 hpi.

The initiation of many of the HIV-mediated early host transcriptional changes, including the suppression of genes associated with T cell functionality, is not dependent upon viral replication. To investigate the mechanisms driving the expression changes observed at 12 hpi, we evaluated the expression changes in SUP-T1 cells treated with UV-inactivated, nonreplicating HIV virions. We performed this experiment because the interaction of cells with nonreplicating HIV virions (or HIV envelope protein gp120 alone) can trigger many intracellular molecular events (30, 31). Strikingly, nonreplicating HIV virions induced expression changes similar to those of intact HIVs at 12 hpi, though the magnitude of expression changes was smaller (Fig. 3A). Functional enrichment analysis also showed that similar functional categories (such as T cell functionality) were enriched in the two lists of DE genes (see Table S3 in the supplemental material). These results indicate that the initiation of early host responses to HIV-1 infection in $CD4^+$ T cells was largely independent of viral replication.

To investigate if any master regulators were driving the early host transcriptional changes, we searched for transcription factors (TFs) with binding sites enriched in the promoters of 12 hpi DE genes. Using the large-scale TF CHiP-seq data from ENCODE (23), we found that the binding sites of 58 TFs were enriched (hypergeometric test *P* value, <0.01) in the promoters of the 12 hpi DE genes. Clustering analysis indicated that the binding sites of some of these TFs tended to co-occur in the promoters of different subsets of 12 hpi DE genes (see Fig. S5 in the supplemental material), suggesting that some TFs functioned together. Alternatively, the enrichment of binding sites of nonfunctional TFs could be merely due to the co-occurrence of their binding sites in the same promoters along with that of some other functional TFs.

Next, we explored if any of these enriched TFs were associated with the transcriptional changes induced by both the intact HIV-1 and UV-inactivated viruses. We examined 10 enriched TFs from subgroups a and b of 12 hpi DE genes as described above. Since the levels of mature mRNAs encoding these transcription factors also changed at 12 hpi, we reasoned that their regulatory activities were also more likely to be modulated. Together, these 10 TFs had at least one binding site in 75% (1,446/1,933) of 12 hpi DE genes, indicating their broad regulatory impacts on the early transcriptional changes. Among them, MYC was the most downregulated TF at 12 hpi in samples from HIV-1-infected cells and in samples

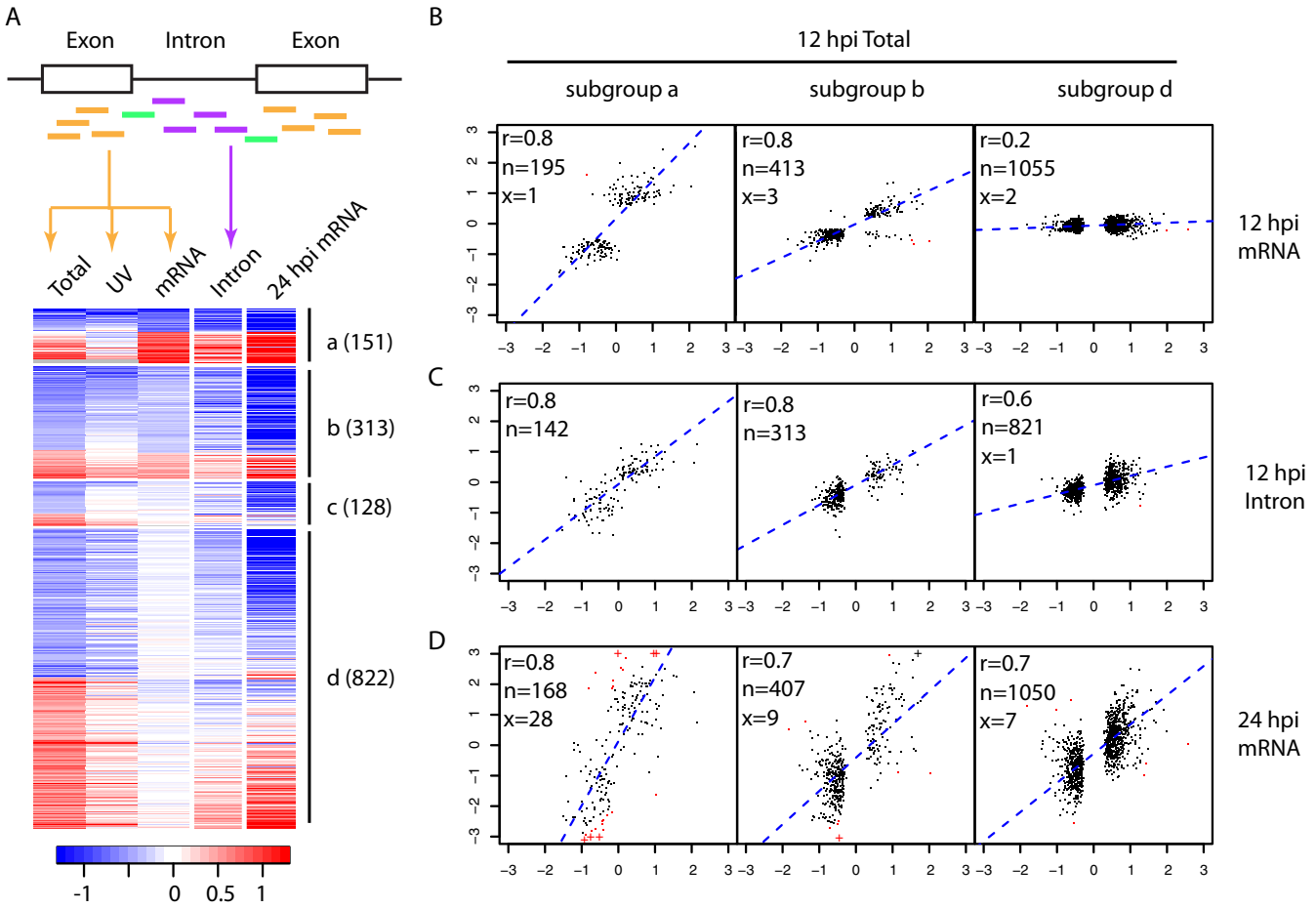

FIG 2 Comparison of expression changes at 12 hpi measured by Total RNAseq versus mRNAseq. (A) Genes (rows of the heatmap) differentially expressed at 12 hpi were divided into four subgroups (a to d) as described in the text. The numbers of genes in each subgroup are in parentheses. Colors represent $\log_2$ infection/mock ratios, blue for downregulation and red for upregulation. The columns are infection/mock ratios for Total RNAseq of HIV infection (Total), Total RNAseq of cells treated with UV-inactivated virions (UV), mRNAseq of HIV infection (mRNA), intronic read counts from Total RNAseq of HIV infection (Intron) at 12 hpi, and mRNAseq of HIV infection at 24 hpi (24 hpi mRNA). To assist the visualization of intronic results, the heatmap shows only 1,414 genes with introns detected (at least 10 reads in at least 3 samples) at 12 hpi by Total RNAseq (Fig. S3 in the supplemental material shows a heatmap with all 1,933 DE genes). The scatterplots of panels B to D are based on the full list of DE genes at 12 hpi. (B to D) Scatterplots between $\log_2$ infection/mock ratios by 12 hpi Total RNAseq (*x* axis) versus that by 12 hpi mRNAseq (B, *y* axis), intronic read counts from 12 hpi Total RNAseq (C, *y* axis), and 24 hpi mRNAseq (D, *y* axis), separately for subgroups a, b, and d. In the top left corner of each scatterplot, the Pearson correlation coefficient (r) and the numbers of genes included in (n, black) or excluded from (x, red) the calculation of correlation are given. The exclusion criterion was a $\log_2$ infection/mock ratio difference larger than 4. A $\log_2$ infection/mock ratio larger than 3 was truncated to 3 and is indicated with the symbol "+."

from cells treated with UV-inactivated virions and had binding sites in the promoters of one-half (1,000/1,933) of the 12 hpi DE genes (Fig. 3B). MYC binding sites were also enriched in the DE genes derived from cells treated with UV-inactivated virions at 12 hpi, consistent with HIV-1 gp120 downregulation of MYC proteins in human mesangial cells (32). NFKB1 and FOS are two other enriched TFs with documented interactions with gp120 (33), and the expression changes of FOS in cells treated with UV-inactivated virions were in the same direction as in HIV-1-infected cells. The genes encoding two additional TFs, YY1 and ELK4, were among the genes differentially expressed in primary peripheral blood mononuclear cells treated with gp120 (31). However, EGR1 was the most upregulated TF in HIV-1-infected cells but was downregulated in cells treated with UV-inactivated virions, even though EGR1 binding sites were enriched in both cases, indicating that additional regulation likely occurred during intact HIV-1 virus infection. To the best of our knowledge, this is the first report that contrasts the host response to intact HIV-1 viruses versus nonreplicating virions by whole-transcriptome analysis. The results show that the initiation of many of the early host responses was regulated in a replication-independent manner, involving master transcription factors likely triggered by early interactions between HIV-1 and host proteins.

Long noncoding RNAs are associated with HIV-mediated early host transcriptional changes. Long ncRNAs (lncRNAs) have emerged as a new class of important regulators in various diseases, including HIV-1 infection (34). Compared to our prior analysis of mRNAseq data (1), here we identified many more (1,098) DE lncRNAs (Table 1), due to the expanded lncRNA annotation and the added Total RNAseq data. To better evaluate the kinetics of lncRNA expression changes, we used qPCR to profile 11 annotated lncRNAs along with 8 newly identified intergenic

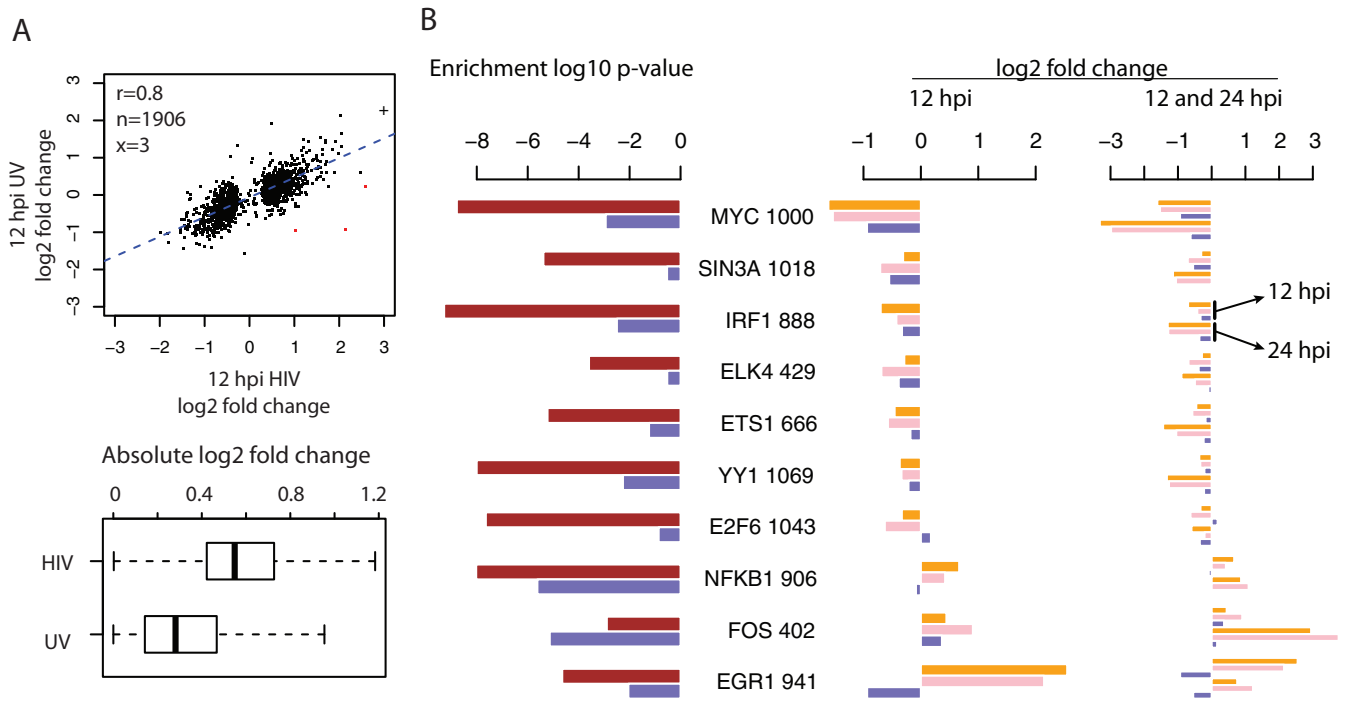

FIG 3 Comparisons of expression changes induced by intact HIV-1 infection versus treatment with UV-inactivated virions at 12 hpi and enriched TFs. (A) Top, scatterplot of infection/mock ratios in cells infected by intact HIV-1 (*x* axis) versus treated with UV-inactivated virions (*y* axis) for 12 hpi DE genes. In the top left corner, the Pearson correlation coefficient (r) and the numbers of genes included in (n, black) or excluded from (x, red) the correlation analysis are given. The exclusion criterion was a $\log_2$ infection/mock ratio difference larger than 4. A $\log_2$ infection/mock ratio larger than 3 was truncated to 3 and is indicated with the symbol "+." Bottom, boxplots of the absolute values of $\log_2$ ratios shown above. For the boxplot, the whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range from the box. (B) Enrichment of TFs in the promoters of DE genes detected in cells infected by intact HIV-1 viruses versus treated by UV-inactivated virions at 12 hpi. Left, enrichment *P* values of 10 TFs using DE genes separately from HIV-1 infection (brown) and the treatment with UV-inactivated virions (blue) at 12 hpi. Next to each TF name is the number of 12 hpi DE genes with its binding sites. Middle, the $\log_2$ infection/mock ratios of enriched TFs at 12 hpi, in the order of mRNAseq, Total RNAseq, and UV-inactivated virions. Right, the $\log_2$ infection/mock ratios of enriched TFs at both 12 and 24 hpi, showing that in general the increased changes at 24 hpi during HIV-1 infection were absent in the treatment of UV-inactivated virions.

loci and 2 annotated pseudogenes across a finer time course of HIV-1 infection (Fig. 4A). As expected, all examined lncRNAs were significantly differentially expressed at 24 hpi, confirming the RNAseq results. Overall, the expression of these lncRNAs tended to be monotonically up- or downregulated throughout the course of HIV-1 infection. Interestingly, even with this small set of lncRNAs, we observed four distinct expression change patterns, separated by the earliest time point showing significant change (*t* test *P* value, <0.05) and the direction of expression change (Fig. 4A), indicating that the expression of these lncRNAs was tightly regulated during HIV-1 infection.

To investigate if the differential expression of these lncRNAs was exclusive to HIV-1 infection, we profiled a subset of these same lncRNAs in human airway epithelial cells infected with one of four strains of influenza virus, including highly pathogenic H5N1 and recent H7N9 viruses. Surprisingly, almost all of these lncRNAs were significantly differentially expressed during at least one of the influenza infections, and the directions of expression changes induced by influenza viruses were mostly similar to that in HIV-1 infection (Fig. 4B). These results indicate that many of the lncRNAs differentially expressed during HIV-1 infection are likely related to certain host responses triggered during different virus infections.

Next, we investigated how lncRNAs were involved in the early host response, given that there were 98 lncRNAs significantly differentially expressed at 12 hpi. For this, we performed lncRNA function prediction by a "guilt-by-association" analysis similar to that described in reference 35. We first obtained a collection of human tissue mRNAseq data (21), which allowed us to recalculate the expression of protein-coding and -noncoding genes consistently, in the same samples, and with the same annotation used here. Next, for each 12 hpi DE lncRNA, we calculated the correlations of the lncRNA tissue expression levels to protein-coding genes. We then applied gene set enrichment analysis (GSEA) to identify functional categories enriched in highly correlated protein-coding genes to infer the biological functions associated with the lncRNA. As shown in Fig. S6A in the supplemental material, individual lncRNAs were strongly associated with the functions enriched in 12 hpi DE genes such as T cell functionalities and mitochondrial dysfunction. These results indicate that these less-studied lncRNAs are actually part of the early host response to HIV-1 infection.

To explore how lncRNA differential expression was regulated, we performed a similar TF binding site enrichment analysis on all DE lncRNAs (see Materials and Methods). We found that 12 TF binding sites were enriched ($P < 0.05$) in the promoters of DE lncRNAs but not in non-DE lncRNAs (see Fig. S6B in the supplemental material). Among the TFs identified, BCL11A is involved in negative regulation of gene expression and T cell differentiation (Gene Ontology annotation). Interestingly, three of the TFs, JUN,

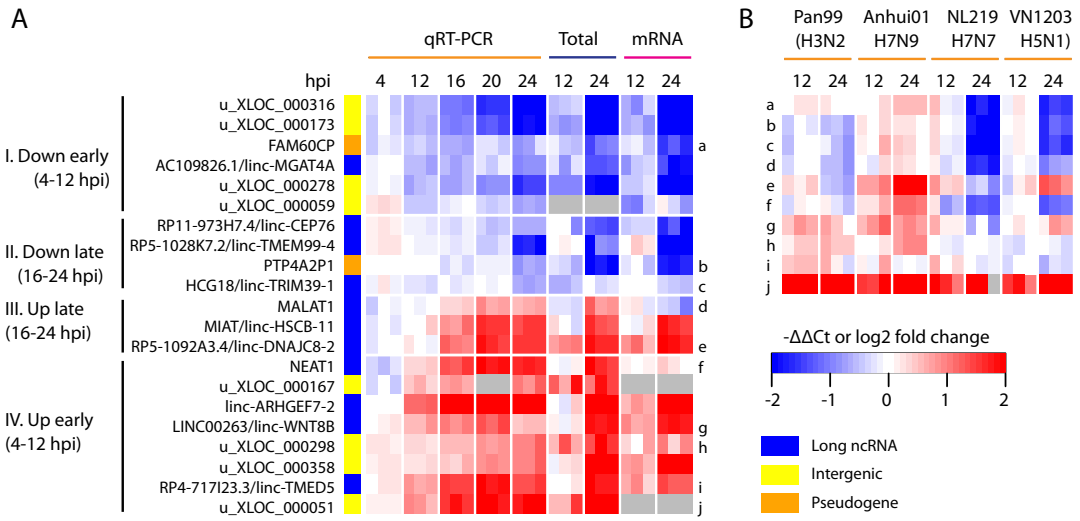

FIG 4 Long ncRNAs differentially expressed in HIV-1-infected CD4$^+$ T cells. (A) Temporal expression profiles of lncRNA, pseudogene, and unannotated intergenic transcripts during HIV-1 infection. Red on the heatmap indicates upregulation during HIV-1 infection, and blue indicates downregulation. Genes were clustered based on the earliest time point at which significant differential expression was detected by qPCR (*t* test *P* value, <0.05). (B) qPCR measurements of temporal expression changes of a subset of ncRNAs shown in panel A (letter or number labels on the right of the heatmap) in human airway epithelial cell line Calu-3 cells infected with one of four strains of influenza A virus, each from a different serotype as shown in parentheses.

FOS, and FOSL1, are related to the AP-1 complex, and the genes encoding these TFs were upregulated at 24 hpi. AP-1 is composed of a mixture of homo- and heterodimers formed between Jun and Fos proteins, and the upregulation of AP-1 during HIV infections and its interaction with various HIV proteins have been extensively studied (33). The enrichment of AP-1 binding sites in lncRNAs has also been observed recently (36, 37), and our results show that AP-1 regulates lncRNAs in a specific context, i.e., HIV-1-infected CD4$^+$ T cells. Overall, this analysis indicates that many of the DE lncRNAs were likely direct targets of transcription factors, providing an alternative mechanism to connect these less-studied lncRNAs with other functions.

Use of Connectivity Map data to refine predicted transcription factor regulation of the early host response. Through the computational analyses above, we predicted specific TFs and lncRNAs as regulators of early host transcriptional changes. To assess and refine our predictions, we searched a large collection of drug transcriptional signatures in cell lines using Connectivity Map (cmap) (26), to see if any drug could reverse the expression changes induced by HIV-1 infection at 12 hpi. We reasoned that if our predicted regulatory relationships were robust, we would likely observe reversed expression changes in targets when the direction of expression change in the corresponding regulator was flipped. With a query signature, here the 500 most upregulated and 500 most downregulated genes at 12 hpi, cmap returned drugs that induced similar and opposite transcriptional changes relative to the query gene signature. We then looked for specific drugs with high negative connectivity scores, an indication that the drug reversed the expression changes induced by HIV-1 infection. Lycorine, the highest-ranked drug with a negative connectivity score, had an average connectivity score of −0.699 over 5 different tests, ranging from −0.629 to −0.860 (see Table S4 in the supplemental material). Within the list of the top 20 ranked drugs, 5 other drugs also had negative connectivity scores: carbimazole (−0.537), gabazine (SR-95531) (−0.416), methiazole (Prestwick-1080) (−0.263), diazoxide (−0.334), and theobromine (−0.385).

Next, we compared the actual gene expression changes induced by lycorine against that by HIV-1 infection at 12 hpi. Of the 1,279 genes differentially expressed at 12 hpi during HIV-1 infection that had common identifiers mapped across the RNAseq and microarray platforms, most (68%, 870) had reversed expression changes (Fig. 5A). This list of genes with reversed expression changes included 32 TFs (of the 58 identified above) with binding sites enriched in the promoters of 12 hpi DE genes. For each of these 32 TFs, we then evaluated if its predicted targets were enriched in the genes with reversed expression changes, indicating that the predicted TF-target relationships were recapitulated here. In total, 14 (44%) of 32 TFs had their predicted targets enriched (hypergeometric test *P* value, <0.05) in the genes with reversed expression changes (see Table S5 in the supplemental material). Interestingly, those TFs which themselves were also differentially expressed at 12 hpi during HIV-1 infection tended to be more likely to have targets enriched in genes with reversed expression changes (Fig. 5B). Specially, all 4 (100%) TFs that had mRNA level changes at 12 hpi (subgroups a and b shown in Fig. 2) had targets enriched in the genes with reversed expression changes. Therefore, the use of both mRNA differential expression and binding site enrichment at 12 hpi significantly improved the precision of our TF predictions. We then pruned the initial list of 58 enriched TFs to those 10 TFs that already covered most (75%) of the 12 hpi DE genes as noted above.

Derivation of regulatory programs to predict the early host transcriptional changes upon HIV-1 infection. Next, we investigated how these computationally identified TFs could regulate the observed early host response. To do so, we first compiled a compendium of human expression data that are related to HIV infection and/or CD4$^+$ T cells (354 microarrays of a single platform from GEO; see Table S6 in the supplemental material). Using this compendium of expression data (without the RNAseq data collected in this study), we then constructed a coexpression network (24) for the genes that were identified above as differentially expressed at 12 hpi during HIV infection. From the coexpression net-

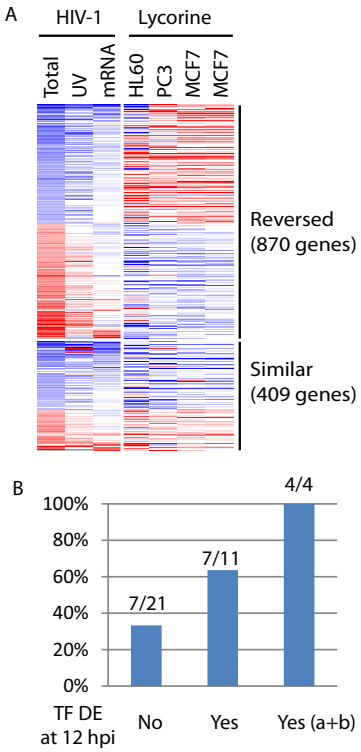

FIG 5 Assessment of TF predictions for HIV-1 infection. (A) Side by side comparison of expression changes induced by HIV-1 infection (12 hpi) or by lycorine treatment. Colors represent $\log_2$ infection/mock or treatment/control ratios: blue for downregulation and red for upregulation. Columns represent individual conditions: Total RNAseq of HIV-1 infection (Total) or UV-inactivated virion treatment (UV), mRNAseq of HIV-1 infection (mRNA), and lycorine treatment of three different cell lines (one repeat for MCF7). Lycorine-induced expression profiles were from Connectivity Map. (B) Percentages of 32 identified TFs that themselves had reversed expression changes after lycorine treatment also had targets enriched in genes with reversed expression changes after lycorine treatment. These 32 TFs were grouped into three bins based on their own DE results at 12 hpi: not DE (No), DE (Yes), and DE and subgroups a and b shown in Fig. 2 [Yes (a+b)], which is a subset of DE (Yes).

work, we identified 9 gene modules, i.e., clusters of highly interconnected genes. In total, 1,614 of 1,933 12 hpi DE genes were covered by the microarray platform, 1,292 of which were covered by 9 gene modules together. As shown in Fig. 6A, genes within each module had highly similar expression profiles across different conditions, and each module represented a distinct expression profile.

Both the patterns and the specific conditions of expression changes of gene modules offered additional insights into the regulation of genes perturbed during HIV-1 infection (Fig. 6A). For example, across both $CD4^+$ T cell activation time courses, the genes in the module D1 were upregulated, but the genes in the module D4 were downregulated, suggesting that genes in both modules are related to $CD4^+$ T cell activation but regulated differently. Further, both modules tended to be upregulated in $CD4^+$ T cells and lymph nodes isolated from HIV-infected patients, though more obviously for the module D4, indicating that the expression of these genes is also perturbed *in vivo*. Genes in the module U1 also tended to be upregulated in $CD4^+$ T cells and lymph nodes isolated from HIV-infected patients. However, their expression changes in two types of activated $CD4^+$ T cells were opposite: downregulated during regulatory T cell activation but upregulated during T effector cell activation, suggesting that the upregulation of these genes during HIV-1 infection is likely driven by regulatory mechanisms similar to those that occur during T effector cell activation. Interestingly, the genes of this module also had a tendency to be downregulated in the $CD4^+$ T cells from HIV-resistant patients, hinting that reversing their upregulation upon HIV infection might confer resistance to infection.

Next, we explored if the identified TFs could predict the expression changes of gene modules during HIV infection, a strong indication of their regulatory roles. To do so, for each gene model we attempted to learn a predictive model from this compendium of expression data. Then, we used the learned model to predict the overall expression changes of the gene module during HIV-1 infection, with identified TFs as predictors. To derive the predictive model, we used a machine learning procedure called elastic net (28), which automatically selects relevant features from high-dimensional data and generates a predictive model with the lowest error through cross-validation. Here, we used 10-fold cross-validation, i.e., the samples were randomly partitioned into 10 sets, of which 9 sets were used to learn a predictive model, which was subsequently used to blindly predict the outcome in the 10th set. This process was repeated iteratively 10 times, and the model with the minimum mean cross-validation error was identified and subjected to the predictability assessment.

For each gene module, we built predictive models with those 10 identified TFs as candidate predictors and the median expression change of genes of the module as the response variable (Fig. 6B, step 3; see also Fig. S7 in the supplemental material). Very promisingly, we found that for all 6 downregulated gene modules, the learned models correctly predicted their downregulation (Fig. 6B, step 4; see also Fig. S8 and S9 in the supplemental material), and 3 of them had predicted values very close to the median expression change, the metric used to learn the predictive model. For comparison, we expanded the set of candidate predictors to all 115 annotated TFs and lncRNAs that were identified as differentially expressed 12 hpi during HIV infection and covered by the microarray platform. Interestingly, the models learned with the expanded set of candidate predictors generated similar predicted values for these 6 gene modules (Fig. 6B, step 4; see also Fig. S8 and S9 in the supplemental material), indicating that the small set of identified TFs was sufficient to achieve optimal predictability. However, for the module U1, the models derived from the expanded set of candidate regulators did correctly predict its upregulation at 12 hpi (see Fig. S8 in the supplemental material), indicating that more regulators remain to be identified in addition to those 10 TFs that were selected for this gene module. Across the predictive models of different gene modules, the common TFs tended to have varied regression coefficients (see Fig. S7 in the supplemental material), representing different configurations of multiple transcription factors for synergistic control of each gene module. For example, YY1 had the most negative regression coefficients for the module U1, indicating that it negatively regulates the module's expressions. But YY1 had the most positive regression coefficients for modules D4, D5, and D6, an indication of strong positive regulatory effects. ELK4 and ETS1 were not selected for module D4, suggesting no or very little regulatory effect, but they had relatively large regression coefficients for

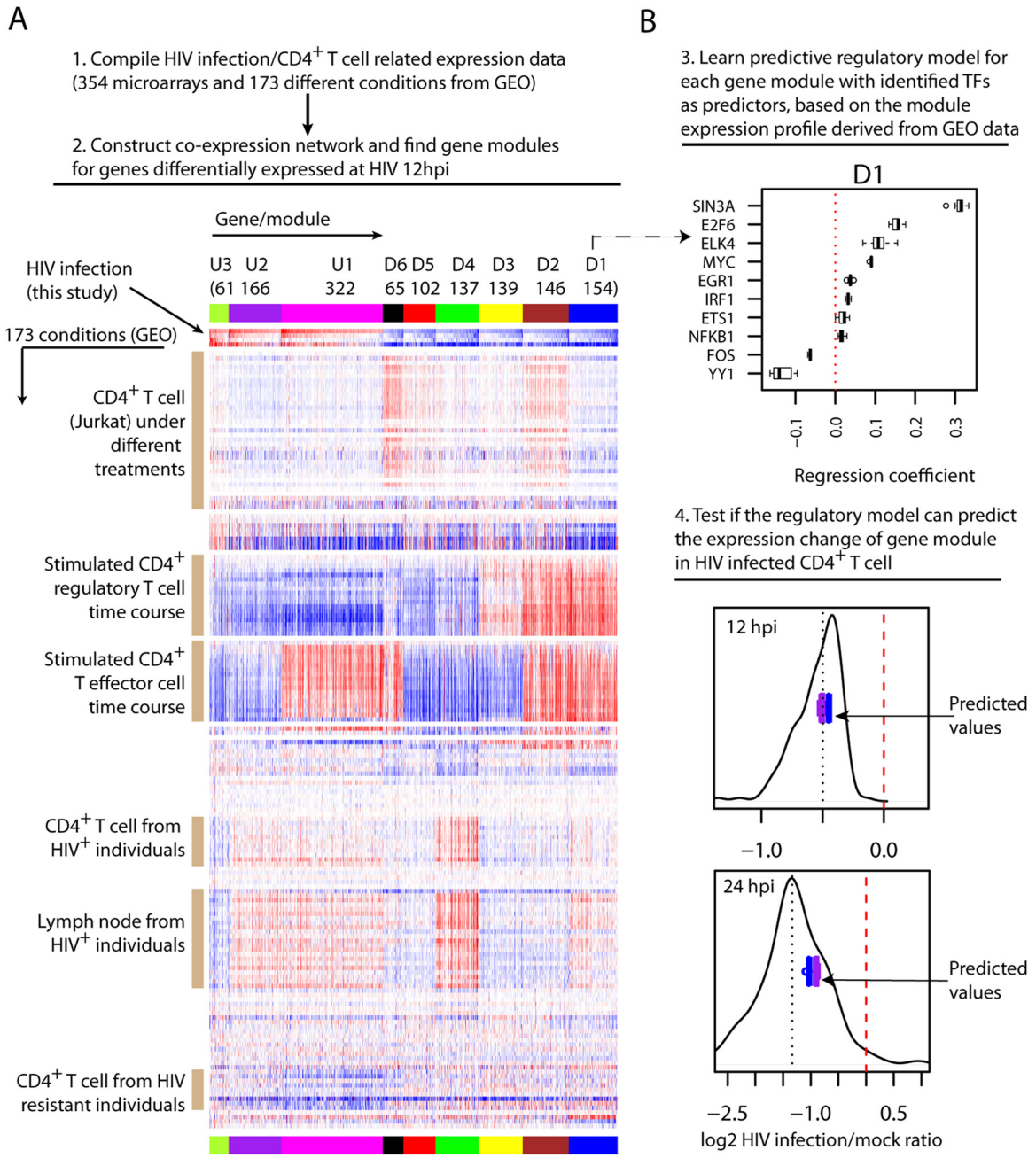

FIG 6 Gene module construction, regulatory model learning, and validation. (A) Identification of gene modules through coexpression network analysis. Top, computational approach for gene module identification. Bottom, expression profile overview of gene modules. The rows of the heatmap are experimental conditions with labels for selected conditions on the left. The HIV-1-induced expression changes measured in the present study are shown in the following order (top to bottom): 12 hpi Total RNAseq, 12 hpi UV-inactivated virions, 12 hpi mRNAseq, and 24 hpi mRNAseq. The columns are genes, and their assignment to each gene module is highlighted by the top colored band (repeated at bottom). The size and name of each gene module are shown above the colored band. (B) Learning and validation of regulatory models for gene modules. Top, an example showing the regression coefficients of the predictive models learned for gene model D1 with 10 identified TFs as candidate regulators (the dotted line indicates a regression coefficient value of zero). To assess if the learning and evaluation were robust, we randomly divided the 173 conditions/samples into 10 subsets and learned 10 predictive models, but leaving out one of the subsets each time (as described in the text). The boxplot for each TF shows the distribution of its regression coefficients across the 10 predictive models. For the boxplot, the whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range from the box. Bottom, comparison of the predicted values and the measured expression changes for gene module D1 during HIV-1 infection, separately at 12 hpi (Total RNAseq) and 24 hpi (mRNAseq). Each of the 10 learned models generated one predicted expression change for the gene module. The boxplot in purple shows the spread of 10 predicted values from the models using 10 identified TFs as candidate regulators, and the boxplot in blue shows the models with the expanded set of 115 candidate regulators (102 TFs and 13 lncRNAs) as predictors. The black line shows the distribution of measured expression changes of genes in module D1, with the black dotted line indicating the median expression change and the red dashed line indicating no change.

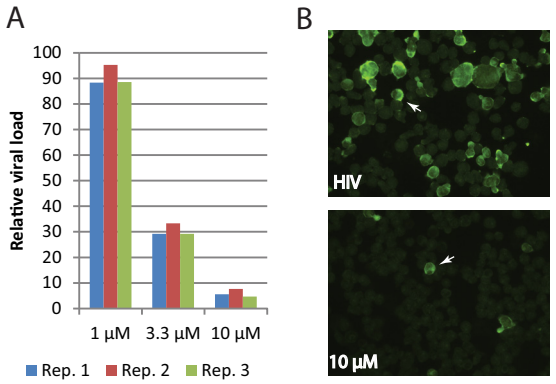

FIG 7 Lycorine inhibits HIV-1 viral replication. (A) Relative viral loads in HIV-1-infected cells treated with lycorine. SUP-T1 cells were infected with HIV-1 LAI at a multiplicity of infection (MOI) of 0.1 for 1 h. After washing away the virus inoculum, cells were suspended in medium containing DMSO or lycorine and cultured for 24 h. Viral load was quantified by qPCR with GAPDH as internal control. The relative viral load in cells treated with lycorine was calculated as the percentage of the viral load in similarly infected cells without drug (DMSO only). (B) Gag staining of cell smears from untreated (top, DMSO) and treated (bottom, 10 μM lycorine) samples collected at 24 hpi showing HIV-1 infection. To guide visualization, one HIV-1-infected cell is highlighted by a white arrow in each image.

modules D2, D3, and D6, an indication of large effects there. In summary, these results provide additional evidence that the computationally identified TFs are likely regulators, and they provide novel insights into the modular regulation of the early transcriptional changes induced by HIV infection.

Lycorine, a drug identified using a host transcriptional signature, potently inhibits HIV-1 replication in CD4$^+$ T cells. As shown above, the enrichment of relevant TFs was based on the prediction that lycorine could reverse the majority of the host transcriptional changes in CD4$^+$ T cells induced by HIV-1 infection. The gene module analysis indicated that reversed host expression changes might be associated with HIV resistance (module U1 in Fig. 6). We therefore reasoned that lycorine should inhibit HIV-1 infection, and we tested this prediction experimentally. As shown in Fig. 7A, a diminishing amount of HIV-1 viral RNAs was detected in SUP-T1 cells treated with increasing concentrations of lycorine, from about 10% loss at 1 μM lycorine to over a 90% viral load loss in cells treated with 10 μM lycorine. Striking differences in viral loads were also visible by comparing the amount of Gag staining in cell smears (Fig. 7B). Strong to moderate staining of Gag antigen was present in approximately one-quarter of untreated cells, while less than one percent of cells treated with 10 μM lycorine showed Gag staining. Further analyses showed that the decrease of HIV-1 viral loads was not the result of direct killing of host cells by lycorine, while lycorine had inhibitory effect on cell proliferation (see Fig. S10 in the supplemental material). A manual search in PubChem (SID 56463667) showed that lycorine can target many host genes, such as SMAD3 gene, which is annotated as a negative regulator of cell proliferation. Lycorine also inhibits NFKB activation in T cells (PubChem CID 11972533; AIDs 489035, 489041, and 489033) and NFKB was upregulated here and identified as a regulator (Fig. 3). More studies are needed to elucidate the detailed mechanism driving lycorine's inhibitory effect on HIV-1 infection, but these results convincingly demonstrate the relevance of using the combined total and mRNA transcriptome analysis to fully capture the early host response and the feasibility of developing host-based therapies for HIV-1 infection through systems analysis of the host response.

DISCUSSION

Due to technical constraints, the relevance of unconventional transcripts, such as long ncRNAs and poly(A)− transcripts, to HIV infection has not been systematically investigated. Here, we used RNAseq to generate an unbiased profiling of host transcriptome changes in response to HIV infection. The uniqueness of our approach is that we combined mRNAseq and Total RNAseq to quantify both poly(A)+ and poly(A)− transcripts, in contrast to standard transcriptome analysis focusing on poly(A)+ mature mRNAs using microarray or mRNAseq. Through a side-by-side comparison, we showed that HIV-1 infection induced the differential expression of many poly(A)− transcripts, which were accurately detected by Total RNAseq but not by the more frequently used mRNAseq. Since most annotated human genes transcribe both poly(A)+ and poly(A)− transcripts (4), targeting only the poly(A)+ fraction significantly compromises transcriptome analysis due to incomplete coverage. As evident here, one big challenge for Total RNAseq is the need to deplete abundant rRNAs. Fortunately, protocols for more-efficient rRNA depletion have been developed, which will considerably improve Total RNAseq coverage of nonribosomal transcripts.

Combined Total RNAseq and mRNAseq analysis revealed the striking finding that Total RNAseq detected a widespread early host response to HIV-1 infection at 12 hpi that was not detected by mRNAseq. Our analysis indicated that the discrepancy was driven mainly by nascent transcripts captured by Total RNAseq and the detectable delay between pre-mRNA transcription initiation and the production of mRNAs at the 12 h time point. Several recent studies have shown that a clearly detectable lag exists between pre-mRNA and mRNA production in response to stimuli such as LPS or TNF activation of macrophages (6, 7) or epidermal growth factor stimulation of epithelial cells (8). The observation that splicing is a significant factor contributing to the delay between pre-mRNA and mRNA production (6, 7) is particularly notable since HIV-1 infection can modulate host RNA splicing machineries (1, 9). These findings suggest that the analyses of nascent transcripts will likely offer a more accurate picture of the kinetics of transcriptional induction than mRNA analyses in the case of HIV-1 infection.

The discovery of large numbers of differentially expressed lncRNAs during HIV-1 infection presents new challenges and opportunities in AIDS research. Zhang et al. reported that the knockdown of the lncRNA NEAT1, also upregulated here (Fig. 1D), enhanced HIV-1 production by increasing the export of Rev-dependent instability element-containing HIV-1 mRNAs from the nucleus to the cytoplasm (34). This example clearly demonstrates the relevance of lncRNAs to HIV-1 infection, but more-systematic approaches are needed to efficiently prioritize lncRNAs for mechanistic studies. Here, we leveraged a human tissue expression compendium to infer lncRNA function through correlated expression of functionally annotated genes. The effectiveness of this guilt-by-association strategy has been illustrated previously, using expression data collected from custom-designed ncRNA microarrays (35) or simply reannotating existing microarrays (38). With RNAseq or comparable technologies, fu-

ture studies will have more complete coverage of both ncRNAs and coding genes. Previously, we found mouse lncRNAs differentially expressed during severe acute respiratory syndrome coronavirus and influenza virus infection (10), and here lncRNAs differentially expressed during HIV-1 infection were also differentially expressed during influenza infection. Combining gene expression data from different virus infection studies will therefore likely improve the statistical power for ncRNA functional inference and facilitate the prioritization of HIV-1 infection-related lncRNAs.

The more complete characterization of early transcriptional changes induced by HIV-1 infection offers at least two unique benefits. First, it enables investigation of underlying molecular regulatory mechanisms. Our strategy relied on the integration of multiple large-scale orthogonal data sets, including CHiP-seq TF binding sites, a tissue expression compendium, an HIV-protein interaction database, and a compendium of related expression data compiled from GEO. Combining the expression profiles induced by nonreplicating virions and TF predictions, our analyses show that early transcriptional changes are largely initiated in a manner independent of viral replication, and we identified specific master regulators triggered by HIV-host interactions. Next, we combined network analysis and machine learning techniques to show that identified regulators were indeed predictive of the early transcriptional changes. Not only do these results provide additional evidence for the relevance of predicted regulators, but also they offer novel insights into transcriptional regulation. In particular, we showed how those host genes were regulated in a modular manner under the synergistic control of multiple regulators, and not by a particular regulator. Also we inferred that less-studied long ncRNAs are potentially additional regulators, which has not been done routinely. The integration of the emerging ncRNAs is made feasible because the generation of genome-scale data by RNAseq or CHiP-seq is independent of gene annotation. These emerging ncRNAs may represent a new class of biomarkers for HIV disease progression or drug targets.

Another benefit to our approach is that it can be used to discover candidate drugs that could be repurposed for treating HIV-1 infection. In this study, we used Connectivity Map (26), a large collection of drug transcription profiles. With the expression signature of HIV-1 infection identified at 12 hpi, we found several drugs capable of reversing the gene expression changes induced by HIV-1 to various degrees. In particular, lycorine, an alkaloid compound found in plants, was predicted to reverse about 70% of the gene expression changes induced by HIV-1 infection at 12 hpi. Remarkably, we showed experimentally that lycorine potently inhibited HIV-1 infection of $CD4^+$ T cells. This result convincingly demonstrates that the principle strategy that we developed will be effective in discovering more host-based antiviral therapies, since the databases of drug profiles are continuously growing. Lycorine has previously been reported to have antiviral activity against flaviviruses (39) and even against HIV-1 in human T cells (40), but by unknown or virus-based mechanisms. Our results suggest that lycorine's regulatory impact on the host response is a likely mechanism of action, which is a novel finding. Since targeting the host response is very different from targeting viral proteins, lycorine and its derivatives (and other drugs identified here) could be further investigated for treating HIV-1 infection. Also, the results showing that reversing the expression changes of a subset of host genes as defined here can confer resistance to HIV infection indicate that we have identified candidate host restriction factors for HIV infection, which are worthy of more investigation. Here, we used the reversed expression profiles to assess and refine the predicted regulators. Therefore, these drugs can also be used experimentally as perturbations to better understand HIV biology.

In conclusion, by contrasting transcriptome sequencing by total RNA and mRNA, we show that there are widespread nascent host transcriptional changes early after HIV-1 infection of $CD4^+$ T cells but the production of mature mRNAs is largely delayed. Using integrative computational analysis, we uncovered possible regulatory mechanisms driving these transcriptional changes. This more complete characterization of the early host response also enabled the discovery of promising drugs for treating HIV-1 infection. This study provides a framework for better understanding HIV biology through iterations of systems level analyses, which will guide the development of targeted intervention of HIV-1 infection in the future.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health Office of the Director (P51 OD010425 and R24 OD011172) and by the DAIDS Reagent Resource Support Program for AIDS Vaccine Development, Quality Biological, Gaithersburg, Maryland, Division of AIDS contract N01-A30018.

We thank Lynn Law for editorial assistance and Terrence Tumpey for sharing influenza virus-infected samples.

REFERENCES

1. **Chang ST, Sova P, Peng X, Weiss J, Law GL, Palermo RE, Katze MG.** 2011. Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and regulation of noncoding RNA expression in a CD4+ T cell line. mBio **2**(5):pii:e00134-11. http://dx.doi.org/10.1128/mBio.00134-11.
2. **Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL.** 2011. Genomewide characterization of non-polyadenylated RNAs. Genome Biol. **12**:R16. http://dx.doi.org/10.1186/gb-2011-12-2-r16.
3. **Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, Reaman G, Milos P, Arceci RJ, Thompson JF, Triche TJ.** 2010. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. BMC Biol. **8**:149. http://dx.doi.org/10.1186/1741-7007-8-149.
4. **Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras TR.** 2012. Landscape of transcription in human cells. Nature **489**:101–108. http://dx.doi.org/10.1038/nature11233.
5. **Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L.** 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. **18**:1435–1440. http://dx.doi.org/10.1038/nsmb.2143.
6. **Bhatt DM, Pandya-Jones A, Tong AJ, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST.** 2012. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell **150**:279–290. http://dx.doi.org/10.1016/j.cell.2012.05.043.
7. **Hao S, Baltimore D.** 2013. RNA splicing regulates the temporal order of

TNF-induced gene expression. Proc. Natl. Acad. Sci. U. S. A. **110:**11934–11939. http://dx.doi.org/10.1073/pnas.1309990110.

8. **Zeisel A, Kostler WJ, Molotski N, Tsai JM, Krauthgamer R, Jacob-Hirsch J, Rechavi G, Soen Y, Jung S, Yarden Y, Domany E.** 2011. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. Mol. Syst. Biol. **7:**529. http://dx.doi.org/10.1038/msb.2011.62.
9. **Dowling D, Nasr-Esfahani S, Tan CH, O'Brien K, Howard JL, Jans DA, Purcell DF, Stoltzfus CM, Sonza S.** 2008. HIV-1 infection induces changes in expression of cellular splicing factors that regulate alternative viral splicing and virus production in macrophages. Retrovirology **5:**18. http://dx.doi.org/10.1186/1742-4690-5-18.
10. **Peng X, Gralinski L, Armour CD, Ferris MT, Thomas MJ, Proll S, Bradel-Tretheway BG, Korth MJ, Castle JC, Biery MC, Bouzek HK, Haynor DR, Frieman MB, Heise M, Raymond CK, Baric RS, Katze MG.** 2010. Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. mBio **1**(5):pii: e00206-10. http://dx.doi.org/10.1128/mBio.00206-10.
11. **Vodicka MA, Goh WC, Wu LI, Rogel ME, Bartz SR, Schweickart VL, Raport CJ, Emerman M.** 1997. Indicator cell lines for detection of primary strains of human and simian immunodeficiency viruses. Virology **233:**193–198. http://dx.doi.org/10.1006/viro.1997.8606.
12. **Chang ST, Thomas MJ, Sova P, Green RR, Palermo RE, Katze MG.** 2013. Next-generation sequencing of small RNAs from HIV-infected cells identifies phased microRNA expression patterns and candidate novel microRNAs differentially expressed upon infection. mBio **4:**e00549–12. http://dx.doi.org/10.1128/mBio.00549-12.
13. **Li CC, Seidel KD, Coombs RW, Frenkel LM.** 2005. Detection and quantification of human immunodeficiency virus type 1 p24 antigen in dried whole blood and plasma on filter paper stored under various conditions. J. Clin. Microbiol. **43:**3901–3905. http://dx.doi.org/10.1128/JCM.43.8.3901-3905.2005.
14. **Navare AT, Sova P, Purdy DE, Weiss JM, Wolf-Yadlin A, Korth MJ, Chang ST, Proll SC, Jahan TA, Krasnoselsky AL, Palermo RE, Katze MG.** 2012. Quantitative proteomic analysis of HIV-1 infected CD4+ T cells reveals an early host response in important biological pathways: protein synthesis, cell proliferation, and T-cell activation. Virology **429:**37–46. http://dx.doi.org/10.1016/j.virol.2012.03.026.
15. **Zeng H, Goldsmith C, Thawatsupha P, Chittaganpitch M, Waicharoen S, Zaki S, Tumpey TM, Katz JM.** 2007. Highly pathogenic avian influenza H5N1 viruses elicit an attenuated type I interferon response in polarized human bronchial epithelial cells. J. Virol. **81:**12439–12449. http://dx.doi.org/10.1128/JVI.01134-07.
16. **Li C, Bankhead A, III, Eisfeld AJ, Hatta Y, Jeng S, Chang JH, Aicher LD, Proll S, Ellis AL, Law GL, Waters KM, Neumann G, Katze MG, McWeeney S, Kawaoka Y.** 2011. Host regulatory network response to infection with highly pathogenic H5N1 avian influenza virus. J. Virol. **85:**10955–10967. http://dx.doi.org/10.1128/JVI.05792-11.
17. **Langmead B, Trapnell C, Pop M, Salzberg SL.** 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10:**R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25.
18. **Trapnell C, Pachter L, Salzberg SL.** 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25:**1105–1111. http://dx.doi.org/10.1093/bioinformatics/btp120.
19. **McCarthy DJ, Chen Y, Smyth GK.** 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. **40:**4288–4297. http://dx.doi.org/10.1093/nar/gks042.
20. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.
21. **Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL.** 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. **25:**1915–1927. http://dx.doi.org/10.1101/gad.17446611.
22. **Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L.** 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28:**511–515. http://dx.doi.org/10.1038/nbt.1621.
23. **Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M.** 2012. Architecture of the human regulatory network derived from ENCODE data. Nature **489:**91–100. http://dx.doi.org/10.1038/nature11245.
24. **Langfelder P, Horvath S.** 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9:**559. http://dx.doi.org/10.1186/1471-2105-9-559.
25. **Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.** 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. **102:**15545–15550. http://dx.doi.org/10.1073/pnas.0506580102.
26. **Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR.** 2006. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science **313:**1929–1935. http://dx.doi.org/10.1126/science.1132939.
27. **Gautier L, Cope L, Bolstad BM, Irizarry RA.** 2004. affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics **20:**307–315. http://dx.doi.org/10.1093/bioinformatics/btg405.
28. **Friedman J, Hastie T, Tibshirani R.** 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat Softw. **33:**1–22. http://www.jstatsoft.org/v33/i01/.
29. **Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ.** 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. **22:**1760–1774. http://dx.doi.org/10.1101/gr.135350.111.
30. **Holm GH, Gabuzda D.** 2005. Distinct mechanisms of CD4+ and CD8+ T-cell activation and bystander apoptosis induced by human immunodeficiency virus type 1 virions. J. Virol. **79:**6299–6311. http://dx.doi.org/10.1128/JVI.79.10.6299-6311.2005.
31. **Cicala C, Arthos J, Martinelli E, Censoplano N, Cruz CC, Chung E, Selig SM, Van Ryk D, Yang J, Jagannatha S, Chun TW, Ren P, Lempicki RA, Fauci AS.** 2006. R5 and X4 HIV envelopes induce distinct gene expression profiles in primary peripheral blood mononuclear cells. Proc. Natl. Acad. Sci. U. S. A. **103:**3746–3751. http://dx.doi.org/10.1073/pnas.0511237103.
32. **Kapasi AA, Fan S, Singhal PC.** 2001. Role of 14-3-3epsilon, c-Myc/Max, and Akt phosphorylation in HIV-1 gp 120-induced mesangial cell proliferation. Am. J. Physiol. Renal Physiol. **280:**F333–F342.
33. **Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG.** 2009. Human immunodeficiency virus type 1, human protein interaction database at NCBI. Nucleic Acids Res. **37:**D417–D422. http://dx.doi.org/10.1093/nar/gkn708.
34. **Zhang Q, Chen CY, Yedavalli VS, Jeang KT.** 2013. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. mBio **4:**e00596-12. http://dx.doi.org/10.1128/mBio.00596-12.
35. **Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES.** 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature **458:**223–227. http://dx.doi.org/10.1038/nature07672.
36. **Garmire LX, Garmire DG, Huang W, Yao J, Glass CK, Subramaniam S.** 2011. A global clustering algorithm to identify long intergenic non-coding RNA—with applications in mouse macrophages. PLoS One **6:**e24051. http://dx.doi.org/10.1371/journal.pone.0024051.
37. **Sui W, Lin H, Peng W, Huang Y, Chen J, Zhang Y, Dai Y.** 2013. Molecular dysfunctions in acute rejection after renal transplantation revealed by integrated analysis of transcription factor, microRNA and long

noncoding RNA. Genomics **102:**310–322. http://dx.doi.org/10.1016/j.ygeno.2013.05.002.

38. **Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, Zhao G, Luo H, Bu D, Zhao H, Skogerbo G, Wu Z, Zhao Y.** 2011. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res. **39:**3864–3878. http://dx.doi.org/10.1093/nar/gkq1348.
39. **Zou G, Puig-Basagoiti F, Zhang B, Qing M, Chen L, Pankiewicz KW, Felczak K, Yuan Z, Shi PY.** 2009. A single-amino acid substitution in West Nile virus 2K peptide between NS4A and NS4B confers resistance to lycorine, a flavivirus inhibitor. Virology **384:**242–252. http://dx.doi.org/10.1016/j.virol.2008.11.003.
40. **Szlavik L, Gyuris A, Minarovits J, Forgo P, Molnar J, Hohmann J.** 2004. Alkaloids from Leucojum vernum and antiretroviral activity of Amaryllidaceae alkaloids. Planta Med. **70:**871–873. http://dx.doi.org/10.1055/s-2004-827239.