Document S1

Identification of viral transcriptional regulators (vTRs)

We created a catalog of vTRs using a combined manual curation and computational approach (**Figure S1**). First, a set of candidate vTRs was obtained based on extensive literature searches. Next, a second set of candidate vTRs was obtained by scanning for putative DNA-binding domains using HMMER in the set of 4,026,372 protein sequences encoded by human viruses contained in the UniProt database, using a previously described approach (Lambert et al., 2019). Evidence that a given virus infects humans is provided in **Table S2**.

The two sets of candidate vTRs were next combined into a single list. This list was then used to identify additional candidate vTRs based on orthology. Specifically, we used the amino acid sequences of each putative vTR in the list as a BLASTP query against the 4,026,372 viral protein sequences obtained from UniProt. To establish BLASTP cutoffs to use in these searches, we generated randomly shuffled viral protein sequences with varying levels of sequence identity to the original query protein (90% identical sequences, 80%, 70%, etc.), and queried these random proteins against the full set of viral proteins using BLASTP. Based on this randomized BLASTP analysis, we chose cutoffs of (local) alignment > 40%, coverage >50%, and BLASTP E-value <1E-05, since no random protein exceeded these values in our BLASTP runs.

We next identified groups of orthologous putative vTRs across the strains of a given viral species by clustering the protein amino acid sequences using CD-HIT (Fu et al., 2012). Sequences were grouped into a single cluster using previously defined lenient cutoffs of >60% protein identity and a CD-HIT tolerance level of 5. The resulting clusters of putative vTRs were manually verified and curated by four co-authors (X.L., T.H., M.T.W., and J.I.F.B.) based on available literature evidence.

Finally, each of the resulting set of putative vTRs was again used as a BLASTP query, using the same BLASTP parameter settings; however, in this step, the proteins were only compared to other proteins within the same viral species. This step was performed to identify any additional vTRs that might have been missed due to thresholding issues in the initial BLASTP search. The resulting set of putative vTRs was again manually verified and curated by four co-authors (X.L., T.H., M.T.W., and J.I.F.B.) to produce the final list. Each vTR has been annotated with its viral family, species, vTR name, vTR aliases, protein IDs, binding ligand (DNA, RNA, or none), known functions, role(s) in addition to transcriptional regulation, role 'summary' ('primary' or 'secondary'), and available experimental evidence (ChIP, ChIP-seq, crystal structure, reporter assay, EMSA, 'other'). All of this information is provided in **Table S2** and on the project web site (https://vtr.cchmc.org).

Protein-protein interaction networks

Protein-protein interactions (PPIs) between vTRs and human proteins were downloaded from VirusMentha on 11/06/2019 (Calderone et al., 2015), VirusHostNet on 04/23/2020

(Guirimand et al., 2015), and a recent study on SARS-CoV-2 (https://www.biorxiv.org/content/10.1101/2020.03.22.002386v3). VirusHostNet data sources were filtered to only include a subset of experimental types: acetylation reaction; adenylate cyclase complementation; ADP ribosylation reaction; affinity chromatography technology; affinity technology; anti bait coimmunoprecipitation; anti-tag coimmunoprecipitation; beta-lactamase complementation; bimolecular fluorescence complementation; bioluminescence resonance energy transfer; biophysical; circular dichroism; classical fluorescence spectroscopy coimmunoprecipitation; colocalization by fluorescent probes cloning; competition binding cross-linking study; cytoplasmic complementation assay; dephosphorylation reaction; electrophoretic mobility shift assay; electrophoretic mobility supershift assay; enzyme linked immunosorbent assay; experimental interaction detectionfar western blotting; filamentous phage display; fluorescence polarization spectroscopy; fluorescent resonance energy transfer; gal4 vp16 complementation; GST pull-down; His pull-down; identification by mass spectrometry; in vitro; interactome parallel affinity capture; isothermal titration calorimetry; luminescence based mammalian interactome mapping; maltose binding protein tag; mammalian protein protein interaction trap; mass spectrometry studies of complexes; molecular interaction; myc-tag coimmunoprecipitation; nuclear magnetic resonance; one-hybrid; peptide massfingerprinting; phage display; protein array; protein complementation assay; protein kinase assay; protein three hybrid; pull-down; sumoylation reaction; surface plasmon resonance; tandem affinity purification; tap tag coimmunoprecipitation; two-hybrid; two-hybrid array; two-hybrid pooling approach; validated two-hybrid; x-ray crystallography; yeast display.

Gene ontology enrichment analyses for human proteins that interact with vTRs from each viral family were performed in www.pantherdb.org (Mi et al., 2013) using GO-Slim Biological Process and PANTHER pathways. Only viral families with at least five vTRs were analyzed. Statistical significance was determined by Fisher's exact test using an FDR < 0.05. For each cluster, the term with the highest fold enrichment that applies to at least five interacting human proteins and has no more than 2,000 genes belonging to the term was selected. In addition, the GO-Slim Biological Process was determined for the sets of human proteins that interact with at least three primary or three secondary vTRs. Only terms with at least five interacting human proteins and no more than 2,000 genes belonging to the term were included.

Protein-protein interaction networks between human proteins and DNA virus or RNA virus vTRs were then constructed. Human protein hubs that interact with at least five and three DNA virus and RNA virus vTRs were included in the DNA virus and RNA virus protein-protein interaction networks, respectively.

ChIP-seq data collection, processing, and analysis

vTR ChIP-seq datasets were identified in the Gene Expression Omnibus (GEO) and downloaded from the NCBI Sequence Read Archive (SRA). All datasets were uniformly processed and analyzed using an in-house automated pipeline. Briefly, the SRA files were converted to fastq files and QC was performed using FastQC (v0.11.2) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In cases where FastQC detects adapter sequences, fastq files were passed to Trim Galore (v0.4.2) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), a wrapper script that

runs cutadapt (v1.9.1) (DOI:10.14806/ej.17.1.200) to remove the detected adapter sequence from the reads. When available, reads were combined between replicates prior to genome alignment and peak calling. The quality-controlled reads were aligned to the reference human genome (hg19/GRCh37) using bowtie2 (v2.3.4.1) (https://github.com/BenLangmead/bowtie2/releases). Aligned reads (in .bam format) were then sorted using samtools (v1.8.0) (http://samtools.sourceforge.net/) and duplicate reads were removed using picard (v1.89) (https://broadinstitute.github.io/picard/). Finally, peaks were called using MACS2 (v2.1.0) (https://github.com/taoliu/MACS) (callpeak -g hs -q 0.01 -t bamfile). When available, corresponding control/input datsets, were used to call peaks with MACS2 (callpeak -g hs -q 0.01 -t bamfile -c CONTROL.bam). ENCODE blacklist regions were removed from the called peaks using the hg19-blacklist.bed.gz file available at https://github.com/Boyle-Lab/Blacklist/tree/master/lists/Blacklist_v1/.

Each ChIP-seq dataset was manually evaluated for quality by M.T.W. based on several criteria: the number of peaks in the human genome, enrichment for expected DNA-binding motifs in these peaks (where applicable), peak overlap with related and/or relevant protein ChIP-seq datasets (as measured by RELI (Harley et al., 2018), which is described below), peak overlap with related and/or relevant ATAC-seq or histone mark ChIP-seq datasets (as measured by RELI), signal-to-noise ratio (as measured by FRiP scores and manual inspection of UCSC Genome Browser wiggle tracks), and proximity of peaks to relevant human genes (as measured by pathway enrichment scores produced by GREAT http://great.stanford.edu/). All criteria were considered in the final decision to include or exclude a given dataset. We emphasize that exclusion of a given dataset does

not mean that it lacks utility or biological signal - it merely indicates that it did not meet our rigorous QC standards. We also note that some of the datasets might involve vTRs that do not bind to the human genome under the conditions analyzed.

To calculate the overlap between the ChIP-seq peaks of two datasets, we used RELI (Harley et al., 2018). RELI systematically intersects the peak regions for a given dataset with all other ChIP-seq dataset peak regions, and the number of intersecting peaks is counted. The overlap between a given pair of ChIP-seq datasets was obtained by dividing the number of overlapping peaks by the total number of peaks present in the smaller dataset. vTR ChIP-seq peaks from all datasets were examined for enriched human transcription factor binding site motif instances using the HOMER suite of tools (Heinz et al., 2010), modified to include the set of human motifs contained in the CisBP database, build 2.0 (Lambert et al., 2019). In Figure 4C, motifs shown are among the top five motif families (based on HOMER p-value) in any individual ChIP-seq dataset. Normalized –log p-values are shown, such that a value of 100 means that the motif has the best –log p-value for the given vTR, and 50% indicates half of the best –log p-value.