Supplementary Text S1: Procedure for the integrated computational pipeline to investigate associations between pathogen-host protein-protein interactions and functional changes underlying gene expression data in host cells

====Software requirement=======

The pipeline has been tested on a computer server with the following software installed on Ubuntu 14.04.2. Statistical test is carried out using RStudio on Windows 8.

(1) TopHat v2.0.12

(2) HTseq v0.6.1

(3) RSeQC v2.5

(4) DESeq v2.1.8.1

(5) RStudio v0.99.448 with R version 3.2.1

(6) samtools v0.1.19-44428cd

==========

Source dataset

Four FASTQ-formatted sequence files were downloaded under the following EMBL accession numbers (Sample names represented in parentheses): SRR049678 (A2h), SRR049683 (B2h), SRR049679 (A4h), SRR049684 (B4h).

Read type: 50 bp single-end and colorspace reads

Strand-specific cDNA library

Sequencer: AB SOLiD System

Data collection and formatting

Download data from EMBL and then use *csfqGzip2csfastaQual.pl* to prepare input file for read mapping using Tophat

```
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR049/SRR049678/SRR049678.fastq.gz -O A2h.fastq.gz
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR049/SRR049679/SRR049679.fastq.gz -O A4h.fastq.gz
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR049/SRR049683/SRR049683.fastq.gz -O B2h.fastq.gz
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR049/SRR049684/SRR049684.fastq.gz -O B4h.fastq.gz
```

Next, the following command lines will be executed for each sample with the same command options.

```
$ perl ~/install/perl/csfqGzip2csfastaQual.pl A2h.fastq.gz &
```

Two output files A2h.csfasta and A2h.qual are produced and optionally compressed to save disk space

```
$ gzip A2h.csfasta
$ gzip A2h.qual
```

Read mapping to a combined genome of mixed-species

Create bowtie1 index for the combination of both reference genomes from human (GRCh38 from Ensembl release 77, 3,096,649,726) and virus (NCBI RefSeq accession number NC_006998), respectively.

```
$ wget ftp://ftp.ensembl.org/pub/release-77/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
$ perl ~/install/perl/formatChrName.pl Homo_sapiens.GRCh38.dna.primary_assembly.fa hg38.fa &
$ rm Homo_sapiens.GRCh38.dna.primary_assembly.fa
```

The FASTA file of the complete reference genome of VACV is downloaded from http://www.ncbi.nlm.nih.gov/nuccore/66275797 and designated as *vacv.fa*. Open the file in a text editor NotePad++, the FASTA header of the VACV genome is simplified with "NC_006998", which is consistent with the following analysis of read counting.

Build the index on a combined genome of mix species

```
$ cat hg38.fa vacv.fa > mixGenome.fa
$ nohup nice -n 19 bowtie-build -C mixGenome.fa mixGenome &
```

Next, read mapping to the above indexed genome using Tophat2

```
$ nohup nice -n 19 tophat2 -p 16 --bowtie1 --color --quals --max-multihits 1 -o alignment-out-A2h ../combinedSpeciesIndex/mixGenome A2h.csfasta.gz A2h.qual.gz > mixSpecies-A2h.nohup &
$ nohup nice -n 19 tophat2 -p 16 --bowtie1 --color --quals --max-multihits 1 -o alignment-out-A4h ../combinedSpeciesIndex/mixGenome A4h.csfasta.gz A4h.qual.gz > mixSpecies-A4h.nohup &
```

```
$ nohup nice -n 19 tophat2 -p 16 --bowtie1 --color --quals --max-multihits 1 -o alignment-out-B2h ../combinedSpeciesIndex/mixGenome B2h.csfasta.gz B2h.qual.gz > mixSpecies-B2h.nohup &
$ nohup nice -n 19 tophat2 -p 16 --bowtie1 --color --quals --max-multihits 1 -o alignment-out-B4h ../combinedSpeciesIndex/mixGenome B4h.csfasta.gz B4h.qual.gz > mixSpecies-B4h.nohup &
```

Checking strandness of reads per sample

Download the human hg38 UCSC gene model and add its link in the directory containing the output read alignment .bam file

```
$ curl -L http://sourceforge.net/projects/rseqc/files/BED/Human_Homo_sapiens/hg38_UCSC_knownGene.bed.gz/download > hg38_UCSC_knownGene.bed.gz
$ ln -s ../geneModel/human/hg38_UCSC_knownGene.bed .
$ mv accepted_hits.bam A2h.bam
$ samtools index A2h.bam
$ infer_experiment.py -r hg38_UCSC_knownGene.bed -s 3000000 -i A2h.bam > A2h.strandSpecific.txt
```

In this case, read strandness of two replicates B2h and B4h sampled from B condition are detected to be "+-,-+" (read mapped to '+' strand indicates parental gene on '-' strand; read mapped to '-' strand indicates parental gene on '+' strand). For both samples, a wrapper script is created to invert read strand orientation in the SAM file.

```
$ samtools view -h -o B2h.sam B2h.bam &
$ perl ~/install/perl/invertStrandInSam.pl B2h.sam | samtools view -bS - > B2h-invert-strand.bam &
```

Read counting of human genes

The human gene model is from the R package TxDb.Hsapiens.UCSC.hg38.knownGene. Four alignment files (A2h.bam, A4h.bam, B2h-invert-strand.bam, B4h-invert-strand.bam) are used as input files for the R script *human_readCounting.R*.

The output tabular delimited text file *human-readcount.txt* containing read count per gene across samples is used as input file for differential gene expression analysis and gene-set enrichment test implemented by the R scripts human-DE-analysis.R and human-GSEA-analysis.R

Analysis of differential expression on human genes

Differential expression analysis using the R wrapper script human-DE-analysis.R

Two tab-delimited output files will be produced

1) human-expressedGene.txt: count data of expressed genes across samples (Expressed genes are extracted based on read count more than 5 in at least two samples)

2) human-DEA-output.txt: summary of differential expression analysis on all expressed genes tested. The file contains ten fields.

Gene-set analysis

Gene-set enrichment analysis based on the human gene expression data produced above. The source gene-sets/pathways (Human_GOBP_AllPathways_no_GO_iea_March_24_2015_entrezgene.gmt) are downloaded from the Bader lab (http://baderlab.org/GeneSets).

To eliminate the side effect of large or small gene-sets, the gene-set composed of genes more than 500 or less than 5 is removed. The gene-set names are formatted for the following application. The output file *humanRefGSready.gmt* is used as input file for GAGE.

```
$ perl /path-to-the-script/filter-GO-set.pl
Human_GOBP_AllPathways_no_GO_iea_March_24_2015_entrezgene.gmt
$ perl /path-to-the-script/formatGSname.pl filtered-pathway-all.gmt
humanRefGSready.gmt
```

Next, an R script *human-GSEA-analysis.R* is created to apply GAGE for the gene-set enrichment test. Two output files geneset.up.list and geneset.down.list are generated. The wrapper script extractEnrichedGS.pl is used to extract the significantly enriched gene-sets with *q*-value set to 0.1. The cutoff can be tuned if more replicates per group/condition are available.

```
$ perl /path-to-the-script/extractEnrichedGS.pl humanRefGSready.gmt geneset.up.list
geneset.down.list enrichedResults.txt
```

Counting reads per gene in the viral genome using HT-seq

The alignment file .bam generated above contains reads aligned to both human and viral genomes. The reads mapped to the viral genome are then extracted. In addition, as the viral gene is not alternatively spliced, the spliced alignments are removed if the N symbol is present in the CIGAR string of the SAM file.

```
$ samtools view -h A2h.bam | awk '($6 !~ /N/ && $3 ~ /NC_006998/) || $1 ~ /^@/' |
samtools view -bS - > A2h-vacv.bam
$ samtools index A2h-vacv.bam
```

We can make a double check on the strandness of reads mapped to the viral genome. Retrieve the gene mode of the viral reference genome from NCBI in the GFF format. In addition, gene information including gene names, product, and locus tag, is extracted and output to a tabular delimited file vacvGeneID.txt. Please note, three atypical gene names annotated as "temporal expression: early" are manually replaced with their locus-tag.

```
$ wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/Vaccinia_virus_uid15241/NC_006998.gff -O
vacv.gff
$ perl ~/install/perl/parseGeneInfoFromGFF.pl vacv.gff > vacvGeneID.txt
```

Convert GFF format to BED format required by RSeQC application. The name of the viral genome in the output BED file should be consistent with the header of the corresponding genome in the FASTA file.

```
$ perl ~/install/perl/gff2bed.pl vacv.gff NC_006998 > vacv.bed
$ infer_experiment.py -r vacv.bed -i A2h-vacv.bam > A2h.strandSpecific.txt
```

The resulting output file A2h.strandSpecific.txt shows the read strandness. If "++,--" is inferred, the HTseq option -s yes will be activated, which enables the reads mapped to the same strand as the feature for single-end reads. Conversely, if "+-,-+" is inferred, -s reverse will be activated.

As protein-coding regions may be overlapped in the viral genome, we set the overlap resolution mode to intersection-nonempty.

```
$ perl ~/install/perl/gff2gff.pl vacv.gff NC_006998 > vacv.chrName.gff
$ htseq-count -f bam -i Name -t CDS -m intersection-nonempty -s yes -q A2h-vacv.bam
vacv.chrName.gff > A2h-output-read-count.txt
$ htseq-count -f bam -i Name -t CDS -m intersection-nonempty -s yes -q A4h-vacv.bam
vacv.chrName.gff > A4h-output-read-count.txt
$ htseq-count -f bam -i Name -t CDS -m intersection-nonempty -s reverse -q B2h-vacv.bam
vacv.chrName.gff > B2h-output-read-count.txt
$ htseq-count -f bam -i Name -t CDS -m intersection-nonempty -s reverse -q B4h-vacv.bam
vacv.chrName.gff > B4h-output-read-count.txt
```

Merge read count per sample into a single count table. We first copy each output file into a newly created directory *mergeCount* and then execute the following command line at the current working directory. The output file virus-read-count.txt is readily used as input for differential expression analysis

```
$ perl ~/install/perl/combineCountTable.pl mergeCount > virus-read-count.txt
```

Differential expression analysis on viral genes using DEseq2

We run the R wrapper script *virus-DE-analysis.R*. The other input file for the script is vacvGeneID.txt, which provides the mappings of RefSeq identifiers to viral gene names. The output file is virus-DEA-output.txt. DEGs are extracted with the following cutoffs: |log2 fold change| > 1 and false discovery rate (q-value) < 0.1. Then the amino acid sequences of DEGs are extracted and output to a FASTA file.

```
$ perl /path-to-the-script/extDEGlist.pl virus-DEA-output.txt > viralDEG.list
$ wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/Vaccinia_virus_uid15241/NC_006998.faa
$ perl /path-to-the-script/extDEGseq.pl NC_006998.faa viralDEG.list > viralDEG.faa
```

Prediction of host and pathogen protein interactions

Submit the above protein FASTA file to the HPIDB database (http://agbase.msstate.edu/hpi/main.html) to search for homologous host-pathogen interactions using BLAST against a database of viral proteins. Blast query parameters include: E-value < 1e-10, at least 60% sequence identity and 70% query coverage, and the top five hits.

A zipped file containing output result files is downloaded from the website. The homologous and unique HPIs between vaccinia virus and human are parsed from the tab delimited file *_pr_res.tsv

```
$ perl /path-to-the-script/parseHPIresults.pl p43ota1439316612_pr_res.tsv > parsedHPIs.txt
```

The mapping of UniProt accessions of the hit human proteins to Entrez gene ID is detected by using db2db (http://biodbnet.abcc.ncifcrf.gov/db/db2db.php). The results are downloaded as a text file and renamed to bioDBnet_db2db_output.txt. Entrez gene IDs are added by a wrapper script.

```
$ perl /path-to-the-script/Uniprot2EntrezID.pl bioDBnet_db2db_output.txt parsedHPIs.txt >
HPIs.addEntrezID.txt
```

Adding symbols for host and viral genes using the R script *addGeneSymbol.R* with two input files *vacvGeneID.txt* and *HPIs.addEntrezID.txt*. The output file *HPIs.addGeneSymbol.txt*.

Among the predicted human proteins interacted with viral DEGs, the expressed human genes are extracted as the signature gene-set for the subsequent association analysis.

```
$ perl /path-to-the-script/intersaction.pl human-expressedGene.txt
HPIs.addGeneSymbol.txt > HPIs.expressedGene.txt
```

Visualization of enrichment map

Three input files are required for plotting an enrichment map, two of which are produced in the above step: the reference gene-sets of human genes in the .gmt file *humanRefGSready.gmt* and the significantly enriched gene-sets in the file *enrichedResults.txt*. Gene expression data is formatted into a .gct file using the perl script below. The output file *.gct contains count data of expressed genes across samples, which can be used as input file for heatmap visualization in the Cytoscape plugin Enrichment map.

```
$ perl /path-to-the-script/txt2gct.pl human-expressedGene.txt > expressedHumanGene.gct
```

For the set of host genes showing evidence for interaction with viral DEGs between 2h and 4h post-infection, a post-query analysis is performed to detect associations between host genes and enriched pathways/gene-sets based on the number of overlapped genes between gene-sets. In this case, the hypergeometric test is carried out with the significance level set to 0.001. Moreover, intersection of genes between the reference GMT and user-provided expression set is chosen.

```
$ perl /path-to-the-script/make_query_set_for_EM.pl HPIs.expressedGene.txt > hostgene-signatureGS.gmt
```

Creation of a comprehensive association table, including the information below

1. viral gene names (RefSeq and Gene name) with log2 fold change, p-value, q-value
2. PPI link to the human partner proteins
3. the membership of the related human genes to gene sets used. One human genes may be assigned to multiple gene sets

Using the R script *mergeTabViaRefSeqID.R* to merge two table files produced above, *virus-DEA-output.txt* and *HPIs.expressedGene.txt*, based on the shared column of RefSeq IDs of viral genes. The output file is *viralDEA.mergeHPI.txt*.

Next, map gene-set id in the reference gene-set file humanRefGSready.gmt to Entrez ID of human genes in the file *viralDEA.mergeHPI.txt*. A single gene ID may be assigned to multiple gene-sets

```
$ perl /path-to-the-script/EntrezGeneid_link_GS.pl humanRefGSready.gmt viralDEA.mergeHPI.txt > viralDE-PPI-host-GS.tab
```

Lastly, merge two output files *viralDE-PPI-host-GS.tab* and *enrichedResults.txt* based on the intersection of the column of gene-set IDs

```
$ perl /path-to-the-script/intersactGS.pl enrichedResults.txt viralDE-PPI-host-GS.tab viralDE-PPI-host-enrichedGS.tab
```

For further biological interpretation, we group the viral genes and host genes belonging to the same gene-set based on the file *viralDE-PPI-host-enrichedGS.tab* produced above

```
$ perl /path-to-the-script/clusterGene2gs.pl viralDE-PPI-host-enrichedGS.tab > geneset-centric.txt
```

The sub-list of the interested gene-set cluster can be extracted.

The text file cluster-33GS-immune-response.list contains 33 functionally associated gene-sets on immune response.

```
$ perl /path-to-the-script/extractGSsubList.pl cluster-33GS-immune-response.list geneset-centric.txt > cluster-33GS-immune-response.tab
```