WELCOME TO THE **ANTIGEN-SPECIFIC REVOLUTION.**

ANTYGEN™

BE SPECIFIC™

Understanding Your HuScan™ and VirScan™ Results

Methods summary based on the PhIP-Seq pipeline built in the Laboratory of H. Benjamin Larman at Johns Hopkins University.

By Tyler Hulett & Daniel Monaco

PhIP-Seq: HuScan™
Human proteome phage-display

PhIP-Seq: VirScan™
Human virome phage-display

Grand Seromics™
Multi-dimensional antibody discovery

Updated July 21st, 2020

CDI LABS
NEXT**GEN** PROTEOMICS™

Overview

T7 bacteriophage immunoprecipitation sequencing (PhIP-Seq) is a method of multiplexed analysis that combines high-throughput DNA sequencing with next-generation proteomics to determine an individual's unique fingerprint of antigen-specific antibodies. PhIP-Seq was developed at Harvard University by CDI Scientific Advisory Board member Ben Larman, Steve Elledge, and colleagues in 2011. This work created HuScan™, a synthetic representation of the complete human proteome as a T7 bacteriophage display library. They built upon this work in 2015 by creating VirScan™, a synthetic representation of the complete human virome: the universe of known infectious viruses. Recently, data analysis has been improved by the AntiViral Antibody Response Deconvolution Algorithm (AVARDA), which deconvolutes antibody cross-reactivity between closely related viruses.

Selected HuScan™ Publications:

Mohan D et al. (2018) PhIP-Seq-characterization of serum antibodies-using oligonucleotide-encoded-peptidomes. Nat Protoc 3:1958-1978.

Larman et al.. (2013) PhIP-Seq-characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. J Autoimmun 43:1-9.

Larman et al. (2011) Autoantigen discovery with a synthetic human peptidome. Nat Biotechnol 29:535–541.

Selected VirScan™ Publications:

Mina, M. J. et al. (2019) Measles virus infection diminishes preexisting-antibodies that offer protection from-other pathogens. Science 366, 599–606.

Xu GJ et al. (2015) Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome. Science 348:(6239):aaa0698.

Isnard P et al. (2019) Temporal virus serological profiling of kidney graft recipients using VirScan. Proc Natl Acad Sci USA 116:10899-10904.

Monaco, D. et al (2018). Deconvoluting Virome-Wide Antiviral Antibody Profiling Data. BioRxiv. https://doi.org/10.1101/333625

July 21st, 2020

Welcome to the antigen-specific revolution!

If you're reading this – you know how much remains to be discovered about the adaptive immune system.

We may have powerful new tools like checkpoint blockade and CAR-T – but they're crude compared to the natural symphony of antigen-specific immunity. Unfortunately, immunology remains art as much as a science.

Why is this? I think our problem is that unlike nucleotide sequencing – high-throughput monitoring tools for antigen-specific immunity remain underdeveloped. And immunology without antigen-specificity data is like trying to understand planetary motion without a concept of gravity.

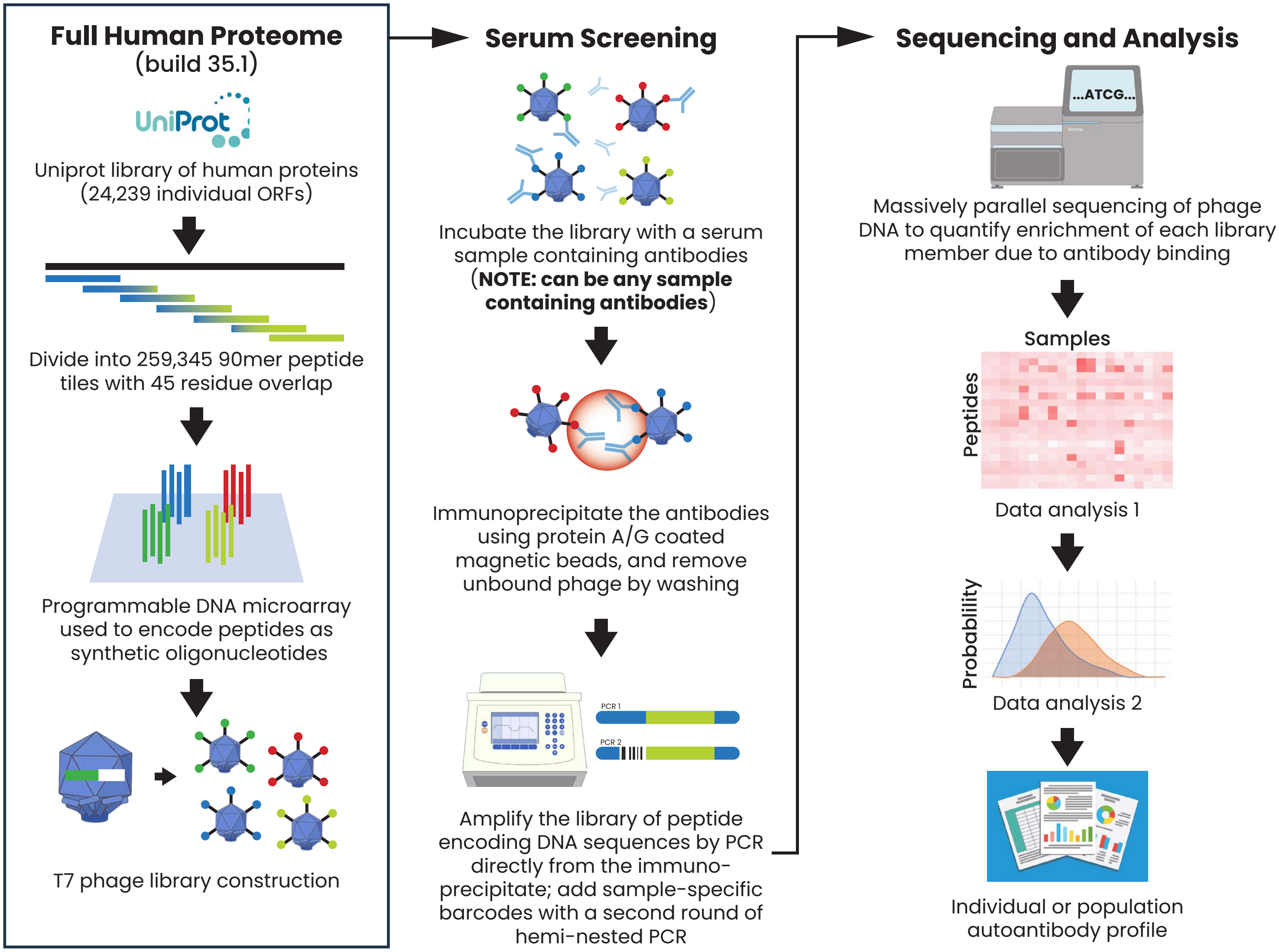
I believe that given the right antigen-specific tools, immunologists can make profound new discoveries. I hope that PhIP-Seq becomes one of these tools, and after a few years of hard work by Scott Paschke, Ignacio Pino, Dan Eichinger, Shaohui Hu, Jose Irizarry, Pedro Ramos, and several others here at CDI Antygen – we are excited to be providing you access to HuScan & VirScan.

These technologies are the result of a decade's work by Ben Larman, Steve Elledge, Gabe Roman-Melendez, Daniel Monaco, and many others at Harvard University & Johns Hopkins University. I followed this work for years before I joined CDI, and am honored to be helping get PhIP-Seq out into the hands of more scientists.

If you have any questions about your data or understanding the outputs and processes described in this document – please don't hesitate to contact me personally,

Tyler

Tyler Hulett, PhD
Director, Biomarker Development
tyler@cdi.bio



Full Human Proteome (build 35.1)

Uniprot library of human proteins (24,239 individual ORFs)

Divide into 259,345 90mer peptide tiles with 45 residue overlap

Programmable DNA microarray used to encode peptides as synthetic oligonucleotides

T7 phage library construction

Serum Screening

Incubate the library with a serum sample containing antibodies (**NOTE: can be any sample containing antibodies**)

Immunoprecipitate the antibodies using protein A/G coated magnetic beads, and remove unbound phage by washing

Amplify the library of peptide encoding DNA sequences by PCR directly from the immunoprecipitate; add sample-specific barcodes with a second round of hemi-nested PCR

Sequencing and Analysis

Massively parallel sequencing of phage DNA to quantify enrichment of each library member due to antibody binding

Data analysis 1

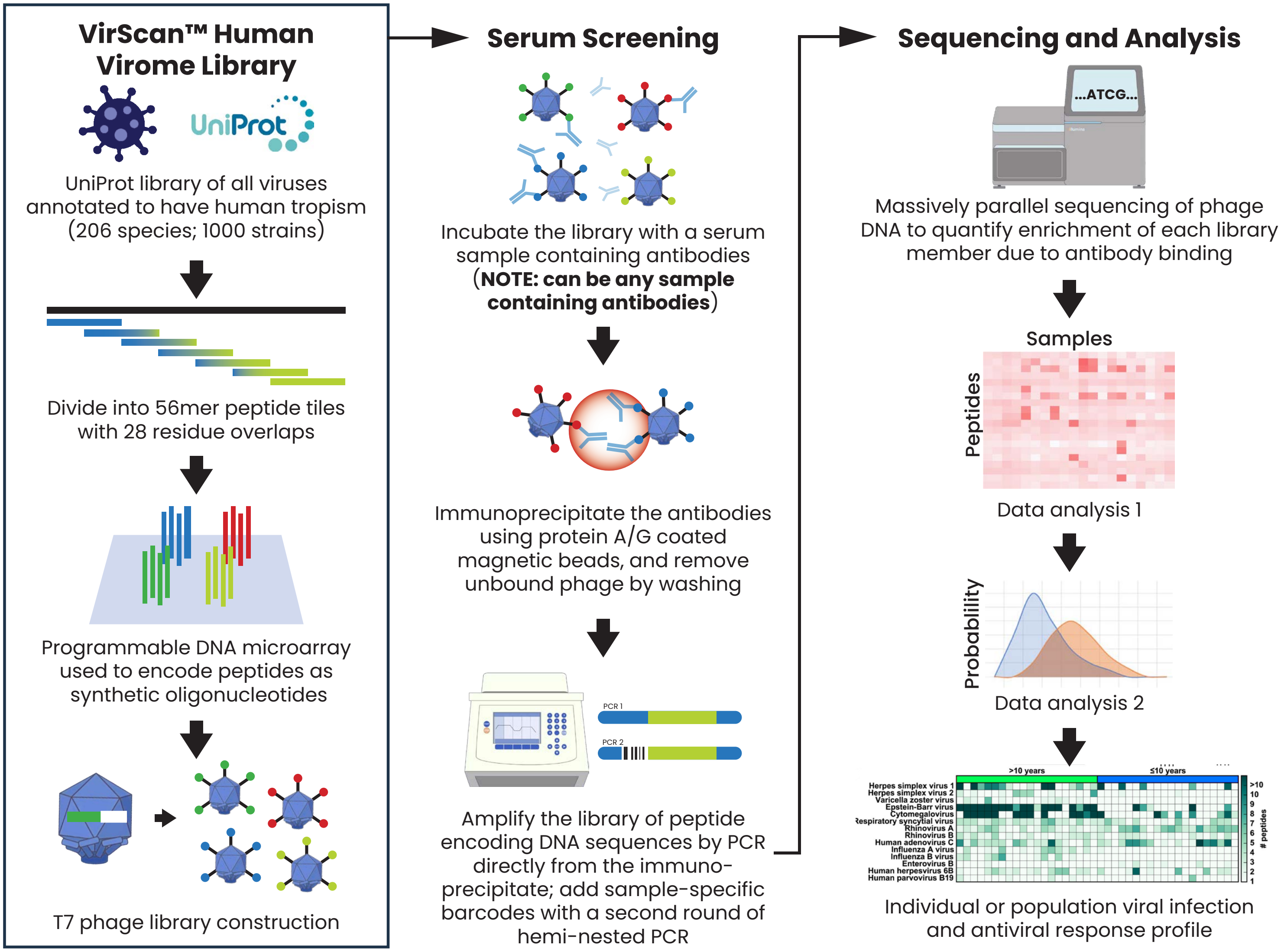
Data analysis 2

Individual or population autoantibody profile

Overview of HuScan™ methods.

The original HuScan PhIP-Seq library was created by downloading viral protein sequences from UniProt, using software to tile overlapping sequences, tuning tiled sequences for translation by the phages and their bacterial hosts, and then ordering the sequences as synthetic oligonucleotides from a vendor as described (Larman, et al, 2011). This oligonucleotide library was PCR-amplified with adaptors for cloning and inserted into a T7 phage display vector that was expanded in E. Coli. The expanded T7 phage library quality is confirmed by sequencing to have >90% of the library within one log of overall average clonal frequency. An aliquot from this library is then reacted with diluted patient serum or other antibody-containing fluid. Bound antibodies are immunoprecipitated with protein A/G beads, the precipitate amplified by PCR, and the sequences quantified by a next-generation sequencing and analysis pipeline that compares patient-sample IP read counts to negative controls with no antibody input (mock-IPs) in the context of overall clonal frequency of individual peptides in the parent library. Output data are then created at both the peptide and whole-protein level. A more detailed description of this process is available (Mohan, et al, 2018).

A typical HuScan™ service involves case and control serum or plasma samples. These are heat deactivated and undergo a protein A/G pulldown assay, PCR amplification, and next-generation sequencing. Raw sequence data are run through a normalization and quantitation pipeline as previously described (Mohan et al., 2018). Raw pipeline outputs are then provided to customers as normalized read counts data tables for both whole proteins and individual 90mer peptides.



VirScan™ Human Virome Library

UniProt library of all viruses annotated to have human tropism (206 species; 1000 strains)

Divide into 56mer peptide tiles with 28 residue overlaps

Programmable DNA microarray used to encode peptides as synthetic oligonucleotides

T7 phage library construction

Serum Screening

Incubate the library with a serum sample containing antibodies (**NOTE: can be any sample containing antibodies**)

Immunoprecipitate the antibodies using protein A/G coated magnetic beads, and remove unbound phage by washing

Amplify the library of peptide encoding DNA sequences by PCR directly from the immuno-precipitate; add sample-specific barcodes with a second round of hemi-nested PCR

Sequencing and Analysis

Massively parallel sequencing of phage DNA to quantify enrichment of each library member due to antibody binding

Data analysis 1

Data analysis 2

Individual or population viral infection and antiviral response profile

Overview of VirScan™ methods.

The original VirScan PhIP-Seq library was created by downloading viral protein sequences from UniProt, using software to tile overlapping sequences, tuning tiled sequences for translation by the phages and their bacterial hosts, and then ordering the sequences as synthetic oligonucleotides from a vendor as described (Xu, et al, 2020). This oligonucleotide library was PCR-amplified with adaptors for cloning and inserted into a T7 phage display vector that was expanded in E. Coli. The expanded T7 phage library quality is confirmed by sequencing to have >90% of the library within one log of overall average clonal frequency. An aliquot from this library is then reacted with diluted patient serum or other antibody-containing fluid. Bound antibodies are immunoprecipitated with protein A/G beads, the precipitate amplified by PCR, and the sequences quantified by a next-generation sequencing and analysis pipeline that compares patient-sample IP read counts to negative controls with no antibody input (mock-IPs) in the context of overall clonal frequency of individual peptides in the parent library. Output data are then created at both the peptide and whole-protein level. A more detailed description of this process is available (Mohan, et al, 2018).

A typical VirScan™ service involves case and control serum or plasma samples. These are heat deactivated and undergo a protein A/G pulldown assay, PCR amplification, and next-generation sequencing. Raw sequence data are run through a normalization and quantitation pipeline as previously described (Mohan et al., 2018). Raw pipeline outputs are then provided to customers as normalized read counts data tables for both whole proteins and individual 56mer peptides.

Analysis Pipeline Summary

The core readout of a PhIP-Seq assay begins with sequence counts of individual peptide-bearing bacteriophages precipitated by antibodies in your sample; this data is acquired by a DNA sequencer from PCR-amplified sample precipitates. Significant enrichment is determined in relationship to mock-immunoprecipitation internal controls – the same assay performed without antibodies.

*****Note: Unless specifically described otherwise, all output file structures are the same for HuScan™, VirScan™, or both together.**

Each file will contain sample columns and rows with unique protein or peptide IDs. All 'protein level' results represent a sum of peptide values (prosum) or maximal peptide value (promax) for all peptides across that protein. In annotated files, these will contain parent protein gene symbols, full descriptive protein names, accession numbers for UniProt, and accession numbers for NCBI RefSeq. Peptide-level annotations will include specific peptide sequences and the sequential position of that peptide within its parent protein.

We generally recommend you begin your analysis with the file:

***_Hits_foldchange_promax_annotated.tsv**

This file contains protein-level metadata annotations alongside the score for the 'top' peptide result from that protein. It has been filtered to only show results if they have been deemed statistically significant for that sample versus the 'beads only' mock-IP controls. From there, you're likely to wish to proceed to:

_Hits_foldchange_annotated.tsv

This provides a peptide-level version of the same results.

Many other interesting output files are provided with varying levels of processing. All are described in the following document.

Single-library HuScan or VirScan Studies: You will receive a single folder containing individual *.tsv & *.csv data tables separated by tabs or columns. Excepting AVARDA readouts for VirScan which is in a separate nested folder with unique file structure, all files will combine all samples across your experiment alongside protein A/G 'beads only' internal controls. Each sample or control will receive its own column. AVARDA is included in the VirScan folder. Individual data outputs can then easily be opened in Excel for data tidying and use in statistical programs.

Combined-library HuScan + VirScan Studies: You will receive a parent folder including files combining results from both HuScan & VirScan assays as individual *.tsv & *.csv data tables separated by tabs or columns. Excepting AVARDA readouts for VirScan which is in a separate nested folder with unique file structure, all files will combine all samples across your experiment alongside protein A/G 'beads only' internal controls. Each sample or control will receive its own column. Individual data outputs can then easily be opened in Excel for data tidying and use in statistical programs. Individualized HuScan & VirScan results will be in separate nested folders; AVARDA is included the VirScan folder. Individual data outputs can then easily be opened in Excel for data tidying and use in statistical programs.

Full PhIP-Seq Pipeline Details

1. Prior to data processing, a **HuScan™** and / or **VirScan™** T7 bacteriophage display immunoprecipitation wet-lab assay is performed with provided antibody samples alongside 'beads only' mock-IP (protein A/G magnetic beads without antibody present).

2. RAW PhIP-Seq FASTQ data files are acquired via next generation sequencing of PCR-amplified precipitates to determine which phages are captured in this experiment. Demultiplexing (demux) of sample barcodes then creates individual FASTQ files for each sample.

3. Individualized sample FASTQ results are processed with alignment software to create ***counts files***, a family of output files counting the number times each peptide encoding sequence is found in each sample FASTQ file. This alignment is done using a simplified exact matching approach to report how many times a peptide's first 50 encoding nucleotides was sequenced.

*****Note: technical replicates should have higher concordance within these raw FASTQ *counts files* than any other output file.**

This step results in tab separated (.tsv) output data tables. All files have a single column for each study sample alongside 'beads only' mock-IP internal controls:

***_Counts.tsv –** File listing the number of times a peptide is found in the sample FASTQ as identified by its unique 50 nucleotide leader sequence.

***_Counts_annotated.tsv –** File listing the number of times a peptide is found in the sample FASTQ as identified by its unique 50 nucleotide leader sequence with metadata annotations.

***_Counts_prosum.tsv –** File listing sum of all peptide counts for each whole protein as identified by accession number.

***_Counts_prosum_annotated.tsv –** File listing sum of all peptide counts for each whole protein as identified by accession number with metadata annotations.

4. Starting from ***counts files***, further processing is used to generate ***EdgeR enrichment files*** and ***fold change files*** from the counts file. This process uses EdgeR, a software package developed for RNAseq analysis which uses a negative binomial model. EdgeR uses average and variance data from the protein A/G 'beads only' mock-IPs to determine a p-value for each peptide within a sample (ie: how significantly different a sample peptide count signal is versus decoy immunoprecipitation internal controls).

This step results in tab separated (.tsv) output data tables. All files have a single column for each study sample alongside 'beads only' mock-IP internal controls:

***_EdgeR.tsv –** File listing –log10 of p-value calculated by edgeR differential algorithm comparing each sample's peptide counts to mock IP control counts and identified by its unique 50 nucleotide leader sequence.

***_EdgeR_annotated.tsv –** File listing –log10 of p-value calculated by edgeR differential algorithm comparing each sample's peptide counts to mock IP control counts and identified by its unique 50 nucleotide leader sequence with metadata annotations.

***_EdgeR_promax.tsv –** File listing maximal value of –log10 of edgeR p-values found across peptides from an individual protein as identified by accession number.

***_EdgeR_promax_annotated.tsv –** File listing maximal value of –log10 of edgeR p-values found across peptides from an individual protein as identified by accession number with metadata annotations.

***_FoldChange.tsv –** File listing fold change of peptide counts compared against median of counts for mock IPs identified by its unique 50 nucleotide leader sequence.

***_FoldChange_annotated.tsv –** File listing fold change of peptide counts compared against median of counts for mock IPs identified by its unique 50 nucleotide leader sequence with metadata annotations.

5. Using the previously generated ***counts files***, ***EdgeR enrichment files***, and ***fold change files,*** generate the ***master hits files*** and ***derivative hits files***. The pipeline uses three criteria to define peptide 'hits:'

1) A raw ***counts file*** value >= 15 for that peptide within that individual sample.

2) An ***EdgeR enrichment file*** –log10(p-value) >= 3 for that peptide within that individual sample.

3) A ***fold change file*** value >=5 for that peptide within that individual sample.

Peptides which pass the 'hit' requirements are output in the ***master hits files***. **This step results in tab separated (.tsv) output data tables.** All files have a single column for each study sample alongside 'beads only' mock-IP internal controls:

***_Hits.tsv –** File listing which peptides are significantly enriched compared to noise, reported as binary output. Value of '1' indicates 'is a hit;' all other peptides are classified as 0 for 'not a hit.'

***_Hits_annotated.tsv –** File listing which peptides are significantly enriched compared to noise, reported as binary output. Value of '1' indicates 'is a hit;' all other peptides are classified as 0 for 'not a hit.' Includes metadata annotations.

***_Polyclonal.tsv –** File listing the total number of entirely unique (non-overlapping) significantly enriched 'is a hit' peptides for each protein. Indicates polyclonality and epitope spreading of antibody reactivity to different sections of the same protein; created from *_Hits.tsv file.

***_Polyclonal_annotated.tsv –** File listing the total number of entirely unique (non-overlapping) significantly enriched 'is a hit' peptides for each protein. Indicates polyclonality and epitope spreading of antibody reactivity to different sections of the same protein; created from *_Hits.tsv file. Includes metadata annotations.

6. The above ***master hits file*** *_Hits.tsv is then used to filter all previous output files to create ***derivative hits files***. These ***derivative hits files*** preserve their original column values for all 'hits' from previous unfiltered outputs, but replace all 'non-hits' with a placeholder value that indicates no significant change from background noise. This results in the ***hits counts files***, ***hits EdgeR enrichment files***, and finally ***hits fold change files***.

*****Note: Best results are often obtained by starting downstream analysis from the *hits fold change files* created during this step of processing.**

This step results in tab separated (.tsv) output data tables. All files have a single column for each study sample alongside 'beads only' mock-IP internal controls:

List of *hits counts files*:

***Hits_counts.tsv** – File listing the number of times a peptide is found in the sample FASTQ as identified by its unique 50 nucleotide leader sequence. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_Counts.tsv

* **Hits_counts_annotated.tsv** – File listing the number of times a peptide is found in the sample FASTQ as identified by its unique 50 nucleotide leader sequence with metadata annotations. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_Counts_annotated.tsv

*** Hits_counts_prosum.tsv –** File listing sum of all peptide counts for each whole protein as identified by accession number. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_Counts_prosum.tsv

*** Hits_counts_prosum_annotated.tsv –** File listing sum of all peptide counts for each whole protein as identified by accession number with metadata annotations. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_Counts_prosum_annotated.tsv

<u>List of *hits EdgeR enrichment files*:</u>

*** Hits_enrichment.tsv –** File listing –log10 of p-value calculated by edgeR differential algorithm comparing each sample's peptide counts to mock IP controls and identified by its unique 50 nucleotide leader sequence. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_EdgeR.tsv

***Hits_enrichment_annotated.tsv –** File listing –log10 of p-value calculated by edgeR differential algorithm comparing each sample's peptide counts to mock IP controls and identified by its unique 50 nucleotide leader sequence with metadata annotations. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_EdgeR_annotated.tsv

***Hits_enrichment_promax.tsv –** File listing maximal value of –log10 of edgeR p-values found across peptides from an individual protein as identified by accession number. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and EdgeR_promax.tsv

***Hits_enrichment_promax_annotated.tsv –** File listing maximal value of –log10 of edgeR p-values found across peptides from an individual protein as identified by accession number with metadata annotations. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_EdgeR_promax_annotated.tsv

<u>List of *hits fold change files*:</u>

***_Hits_foldchange.tsv –** File listing fold change of peptide counts compared against median of counts for mock IPs identified by its unique 50 nucleotide leader sequence. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_FoldChange.tsv

***_Hits_foldchange_annotated.tsv –** File listing fold change of peptide counts compared against median of counts for mock IPs identified by its unique 50 nucleotide leader sequence with metadata annotations. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_FoldChange_annotated.tsv

***_Hits_foldchange_promax.tsv –** File listing maximal fold change value of all peptides across an individual protein identified by accession number. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_FoldChange.tsv

***_Hits_foldchange_promax_annotated.tsv –** File listing maximal fold change value of all peptides across an individual protein with metadata annotations. Only reported if that result is significantly enriched compared to noise. Derived from *_Hits.tsv and *_FoldChange_annotated.tsv

6. In the case of a **HuScan™** assay, the above concludes the list of file processing & delivered files. For assays involving **VirScan™,** an output called **AVARDA** is next created from the ***_Hits_foldchange.tsv** file. AVARDA stands for AntiViral Antibody Response Deconvolution Algorithm and is a multi-module software package for analyzing VirScan™ datasets. The AVARDA output provides a probabilistic assessment of infection at species-level resolution by considering alignment of all library peptides to each other and to all other human viruses. This addresses the issue of antibody cross-reactivity between closely related viruses with conserved sequence domains. The current primary output is a table listing the most significant viral infections found within each sample. A full overview of the AVARDA program can be found here.

This step results in comma separated (.csv) output data tables nested within the AVARDA folder of your VirScan™ output. File structures vary as described:

***AVARDA_breadth.csv –** Matrix listing the number of non-overlapping enriched peptides that were assigned to a given virus within this sample by the AVARDA algorithm.

***AVARDA_compiled_full_output.csv –** Compiled tabular version of all sample outputs for the AVARDA algorithm, reported only cases where a virus was detected in sample.

***AVARDA_evidence_pep.csv –** Matrix listing identity of peptides associated with a given virus.

***AVARDA_post_p_value.csv –** Matrix listing p-value of evidence for a given viral infection.

***AVARDA_post_p_value_BH.csv –** Matrix listing p-value of evidence for given viral infection BH test corrected

***AVARDA_unfiltered_evidence_number.csv –** Matrix listing total number of enriched peptides that were assigned to a given virus.

***SAMPLE_NAME.csv –** Table listing all AVARDA metrics for a given individual sample.

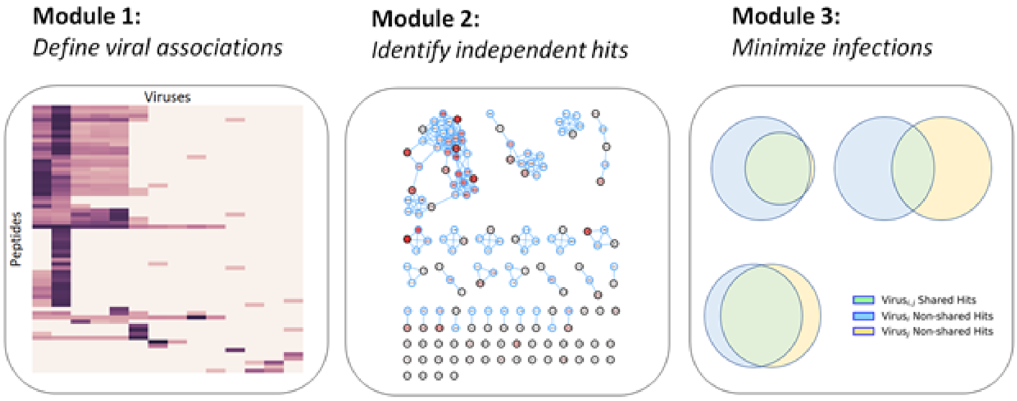
7. In the case of a **VirScan™** assay, this concludes sample processing and outputs. If you ordered combined **HuScan™** and **VirScan™** assay, additional processing then combines outputs from the two data libraries. Your data will be delivered with this 'combined' output in the main root folder. All file names and processing methods are identical to the above VirScan & HuScan outputs except stitched together to create a combined result. Individual VirScan & HuScan outputs are still provided individually within nested folders.

Overview of AVARDA.

1) tBlastn align all peptide hits against all human viruses.

2) Use network graph to identify greatest set of non-overlapping peptide enrichments.

3) Remove overlapping peptides with binomial statistics as null models.


Module 1:
Define viral associations
Viruses
Peptides
Module 2:
Identify independent hits
Module 3:
Minimize infections
Virus$_{i,j}$ Shared Hits
Virus$_i$ Non-shared Hits
Virus$_j$ Non-shared Hits

CDI Laboratories, Inc.

Puerto Rico:

12 W Mendez Vigo Avenue
Mayaguez, PR 00680
USA
939-280-5293
844-539-6296 (Toll Free)

Baltimore:

855 N. Wolfe St., Suite 603
Baltimore, MD 21205
USA
443-388-8029

Technical and Sales Support:

844-539-6296 (Toll Free US/Canada)
sales@cdi.bio (Sales inquiries)
support@cdi.bio (Product, Services and Technical Support inquiries)

Product Co-Development and Licensing

spaschke@cdi.bio

ANTYGEN™
BE SPECIFIC™

CDI LABS
NEXTGEN PROTEOMICS™