**[SieveSifter or SieveViz or ?]: A web-based tool for visualizing the sieve analyses of HIV-1 vaccine efficacy trials**

Andrew Fiore-Gartland1,2, Nick Kullman3, Allan de Camp1, Graham Clenaghan3, Wayne Yang3, Craig Magaret1, Paul Edlefsen1, Peter Gilbert1

1Vaccine and Infection Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA

2eScience Institute, University of Washington, Seattle, WA

3University of Washington, Seattle, WA

Corresponding author: Andrew Fiore-Gartland, agartlan@fredhutch.org

**Abstract**

At the conclusion of a HIV-1 vaccine efficacy trial a characterization of the “breakthrough” viruses that infect trial participants can help to elucidate the mechanisms of imperfect or heterogeneous protection. The approach, known as “sieve analysis”, can identify the functional specificities of vaccine-induced immune responses by comparing the genetic sequences of viruses isolated from vaccine versus placebo recipients. We have created an interactive web-based visualization and database tool for exploring the results of sieve analyses performed on two major efficacy trials: (1) the HVTN 502/Step trial, and (2) the MHRP/Thai Health Ministry RV144 trial. The tool acts simultaneously as a portal for organizing and sharing the viral sequence data generated for these trials and as a platform for discovery of pertinent sieve effects as ongoing follow-up studies of the vaccines continue to test and validate sieve analysis findings.

http://sieve.fredhutch.org/viz

**1. Introduction and motivation**

A major goal of HIV-1 vaccine efficacy trials is to identify immune correlates of risk and protection [Gilbert, Plotkin, Haynes, Tomaras, correlates papers etc.]. The identification of vaccine-induced immune responses that correlate with risk of infection among vaccine recipients generates hypotheses about the mechanisms of protection and can aid in the rational design and iterative development of novel vaccine candidates. A complementary approach, compares the genetic sequences of the “breakthrough” infections in the vaccine versus placebo groups, to identify evidence of vaccine-specific immune pressure [Gilbert, Edlefsen, Rolland sieve analysis reviews?]. The underlying hypothesis is that vaccine-induced immune responses will more effectively target viruses that are similar to the vaccine immunogens. The approach has been given the name sieve analysis since the vaccine-induced immune responses act like a sieve, blocking some viruses while letting others break through. Statistically significant differences between the randomized vaccine and placebo treatment groups generate hypotheses about the vaccine-induced immune specificities that mediated partial or heterogenous protection in the trial  [Gilbert, Self, Ashby].

A fundamental challenge in sieve analysis is the identification of potentially small footprints of vaccine-induced pressure (1 AA to ~50 AA) within an HIV-1 genome of >3000 amino acids. Combined with the fact that HIV-1 vaccine efficacy trials have historically accumulated fewer than 150 cases overall [Vax004, Vax003, Step, 503, RV144, 505] there is low statistical power to identify amino acid sites under vaccine pressure. For example, significant signatures of immune pressure were detected in the RV144 vaccine trial, the first HIV-1 vaccine to show partial efficacy [RV144 primary]. The primary sieve analysis focused on known contact sites of Env V1/V2-specific IgG antibodies [Edlefsen, Rolland], which had been elicited by the vaccine and were known to correlate inversely with risk among vaccinees. At one site in Env V1/V2 (HXB2 K169), sequences from vaccine recipients were more likely to mismatch the vaccine immunogen compared to placebo. Follow-up experiments have since shown that mutations at this site abrogate binding of vaccine-elicited V2-specific IgG, thus helping to validate the sieve analysis and further support a hypothesis that V2-specific antibodies mediated the observed efficacy [Zolla-Pazner, Tomaras]. However, in the more comprehensive whole-protein Env sieve analysis that followed this site and the other significant site of a sieve effect in V1/V2 failed to pass multiplicity adjustment (q < 0.2; [Edlefsen comprehensive RV144]). This demonstrated the importance of leveraging experimental data in sieve analysis and suggested that evidence of additional sieve effects in the existing sequence data may be later detected in the context of additional immunological data. Indeed, follow-up work has since revealed that V3-specific IgG may also have played a role in the protection [Zolla-Pazner and Edlefsen; Gilbert and Corey RV144 correlates review]. Though in many scientific fields it is possible to reproduce experimental results to validate hypotheses, there are significant barriers [ethical, budgets, sociopolitical, etc.] that make it challenging if not impossible to repeat human HIV-1 vaccine efficacy trials. This makes it even more critical to share data widely and ensure that it is fully mined for potential insight.

To facilitate future discovery and a deeper understanding of vaccine responses through sieve analysis, we have created a publicly accessible and interactive web-based tool for exploring the sieve analyses of HIV-1 vaccine efficacy trials. The visualization allows immunologists, virologists and biostatisticians alike to intuitively navigate site-wise comparisons of sequences isolated from vaccine and placebo recipients. At publication the tool includes data and results from the HVTN 502/Step trial [Rolland/Decamp and Buchbinder] as well as the MHRP and Thai Ministry of Health RV144 trial [Edlefsen comprehensive ms]. As results become available the sieve analyses of the HVTN 503/Phambili and HVTN 505 efficacy trials will also be included in the tool.

**2. Implementation**

Fundamentally a sieve analysis is a statistical test comparing the viral sequences isolated from participants in the randomized vaccine and placebo treatment groups. The statistic for comparison is typically a sequence-based genetic distance representing the (dis)similarity of the vaccine immunogen and the breakthrough virus. For this visualization we focus on a site-wise “mismatch” or Hamming distance that indicates whether or not a breakthrough sequence matches the vaccine immunogen sequence at each amino acid site. Though an infected participant harbors many different HIV-1 quasispecies and there may be a sequence for several of these variants, the primary sieve analyses have focused on single representative sequences from each infection (see primary manuscripts for details). The treatment groups are compared at each site using a pooled variance t-statistic and a p-value is generated by permutation test (10,000 permutations). The tool also provides false-discovery rate (FDR) adjusted q-values across each protein.

The visualization has four primary frames (**Fig. 1**): (A) a sequence navigator showing the site-wise sieve analysis results from a single trial, for a single HIV-1 protein immunogen and providing the ability to select sets of sites, (B) a summary table of the selected sites, (C) a summary chart showing vaccine versus placebo distances over the selected sites and (D) a series of stacked bar charts showing the distribution of amino acids at each selected site.

In the selection interface (**Fig. 1A**), each bar represents an amino acid and the sequence can be zoomed and panned to move fluidly through the data. The height of each bar is used to encode a statistic about the site, including the t-statistic, p-value, t-statistic, q-value or entropy. Color is used to show amino acid groupings ([Taylor chemistry? or what?]). A selection is made by holding shift, clicking and dragging across the bars. This allows the user to specify groups of sites that may represent a cellular or humoral epitope and to visualize a combined vaccine distance. Alternatively, a selection can be made by entering the HXB2 position of the sites of interest. Selections can easily be shared with collaborators using a custom URL query that can be generated after a selection is made.

Summaries of the selected sites are provided in a table and distance chart (**Fig. 1B,C**). The table shows the t-statistic, p-value and q-value of the sieve analysis at each site. It also shows entropy of each site, calculated over the vaccine and placebo breakthrough sequences. The stacked bar charts (**Fig. 1D**) shows the distribution of amino acids at each selected site with bar illustrating the fraction of breakthrough infections with each amino acid. The color encoding of the bars can be changed and the resulting chart can be exported in the Scalable Vector Graphics (SVG) format for incorporation in publication quality figures.

The visualization is constructed using HTML5 and Javascript using the D3 visualization library and is compatible with most modern web browsers. The code has been made available (https://github.com/nkullman/SIEVE) and is released under the MIT license. [The results are all contained in a publicly available MySQL/PostgreSQL database which also has a browsable frontend for downloading sequence datasets for sieve analysis].

[Optional TODO: tool allows for uploading, analyzing and visualizing nove sieve datasets]

**3. References**

[Generate in Word later]

**[Figure 1 will be a screenshot with each of the four frames highlighted and labeled with the corresponding letter]**

**Figure 1. SieveSifter layout**. A sieve analysis of a single immunogen/protein for an HIV-1 vaccine efficacy trial is visualized in four primary frames: (A) a sequence navigator showing the site-wise sieve analysis results (B) a summary table of the selected sites, (C) a summary chart of vaccine versus placebo group distances over the selected sites and (D) stacked bar charts showing the distribution of amino acids at each selected site.

Examples of other Application Notes from the journal include [CiVi](http://bioinformatics.oxfordjournals.org/content/31/17/2867.full) and [Tax4Fun](http://bioinformatics.oxfordjournals.org/content/31/17/2882.full).

The Bioinformatics Journal defines Application Notes as follows:

**Application Notes** (up to 2 pages; this is approx. 1300 words or 1000 words plus one figure) Applications Notes are short descriptions of novel software or new algorithm implementations, databases and network services (web servers, and interfaces). Software or data must be freely available to non-commercial users. Availability and Implementation must be clearly stated in the article. Authors must also ensure that the software is available for a full TWO YEARS following publication. Web services must not require mandatory registration by the user. Additional Supplementary data can be published online-only by the journal. This supplementary material should be referred to in the abstract of the Application Note. If describing software, the software should run under nearly all conditions on a wide range of machines. Web servers should not be browser specific. Application Notes must not describe trivial utilities, nor involve significant investment of time for the user to install. The name of the application should be included in the title.

**Sections from final SIEVE paper**

3.1 Selection

The selection interface (figure 8) aims to allow quick navigation through the sequence while highlighting sites of potential interest. Each bar represents an amino acid, and the sequence can be zoomed and panned to move through the data fluidly. The height of the bars representing amino acids is used to encode statistics about the site: p-value and entropy are currently supported and other user-specified statistics could be included in future extensions of SIEVE.

A selection is made by holding shift, clicking and dragging across the bars. A selected bar is distinguished from non-selected bars through color (opacity) and position encoding - selected bars are raised above the heights of non-selected bars. Trials during development of SIEVE showed that this choice of encoding allowed for detection of selected sites even when the zoom extent is at minimum and the entire overview sequence is visible. Because of this, the user can comfortably select multiple groups of sites across the genome. Since proteins fold, non-contiguous groups in the genome may be spatially close, so such selections are necessary for studying recognition of HIV by antibodies.

3.2 Site-wise Comparisons

Once a selection is made, stacked bar charts are generated for each site in the selection (see figure 9). The stacked bar charts show the distribution of amino acids at the site for both the vaccine and placebo groups. The presence of an amino acid at a site different from the vaccine’s amino acid at that site is known as a mismatch. Mismatches have proven to be a useful gauge of a site’s importance in sieve analysis [2]. The stacked bar chart figures were designed as a direct replacement for those appearing in figure 7, but allowing easier comparison between the two groups.

The color scheme used throughout is a known encoding for amino acids which gives chemically similar acids similar colors. Other options are provided, including colors used in WebLogo and a 20-category color scheme included in D3.js which maximizes the perceptual difference between acids.

If a researcher finds an interesting chart, this part of the tool can be exported as an SVG, preserving the formatting options. The exported file can be easily annotated using an SVG editor and converted to other formats for use in publications.

3.3 Group Data

Group statistics about a selection are shown in two ways: a chart representing the distribution of mismatch counts, and a table showing summary statistics such as joint entropy, as well as site-specific statistics for all sites selected. These are seen in figure 10.

The data set in the chart is the number of mismatched amino acids in the current selection for each patient. There are two options to view this data, either as a pyramid chart (as shown in figure 10) showing a histogram of mismatch counts, or as a box plot.

The table in figure 10 can either show statistics about entropy, which measures how much variety there is in the number of amino acids seen at a particular site or selection, or the raw mismatch counts.

3.4 Configuration Options and Alternate Selection

A small number of configuration options are available to the user to increase the flexibility of the tool while preventing it from being overwhelming. These options are displayed in figure 11. Of particular importance is the ability to change the measure encoded by the height of the bars in the selection chart. These measures are used to indicate sites where interesting behavior is observed, so being able to view

a quick overview of what sites are important according to different metrics is critical to the task of exploratory analysis.

The two text fields in figure 11 are alternative selection mecha- nisms. In the first text field, a user can select sites by entering them as a comma separated list of positions or ranges of positions (based on the indexing used in the HXB2 reference HIV strain). The second text field selects all sites with a p-value less than or equal to the threshold entered. So, if a user enters 0.05 into the second field, the selection will be changed to the set of sites with p<0.05.