VOLARE:
Visual analysis of disease-associated microbiome-immune system interplay

Janet Siebert[1,2*], Charles Preston Neff[1], Jennifer Schneider[1], Brent Palmer[1], Catherine Lozupone[1], Carsten Görg[1]

[1] Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA, [2]CytoAnalytics, Denver, CO 80113

*To whom correspondence should be addressed.

Abstract:

Background: Associations between the microbiome and the immune system are well documented. However, elucidating specific mechanisms is an active research area. High-dimensional "omic" assays such as 16S ribosomal RNA (rRNA) sequencing and CyTOF immunophenotyping support hypotheses generation, identifying possible interactions that can be further explored experimentally. Linear regression between microbial and host immune features is useful for quantifying relationships between microbes and immune readouts. But, vetting dozens of significant associations can be cumbersome, especially when a project involves experts from different disciplines. In order to facilitate communication and sense-making across disciplines, we performed a design study for visual analysis of these relationships with a goal of helping researchers prioritize results for experimental follow-up.

Results: Using data from paired 16S ribosomal RNA (rRNA) sequencing and CyTOF immunophenotyping on gut biopsy samples from people with and without HIV, we fit a regression model to each microbe:immune cell pair, accounting for differences by disease status. We used permutation testing to control the false discovery rate, resulting in a "top table" of microbe:immune cell pairs. After identifying essential tasks in the further analysis of this top table, we designed VOLARE (Visualization Of LinEar Regression Elements), a web application that integrates a searchable top table, a network summarizing this table, sparkline-inspired graphs of fitted regression models, and detailed regression plots showing sample-level detail. We applied this application to two case studies--microbiome:immune cell data from gut biopsies and microbiome:cytokine data from fecal samples.

Conclusions: Systematically integrating microbiome-immune system data through linear regressions and presenting the top table results in an interactive environment supports the Shneiderman mantra ("Overview first, filter, details-on-demand"). Our approach allows domain experts to control the analysis of their results, empowering them to screen dozens of candidate relationships with ease. Our contributions include characterizing the exploration of microbiome-immune system data in a team science context, and the support of an associated workflow by integrating existing visualization approaches.

Availability: R scripts, the web application, and sample data are available at https://sourceforge.net/projects/cytomelodics

Contact: jsiebert@acm.org

1 Background

Research teams have learned that aberrations in the gut microbiome are associated with diseases, such as inflammatory bowel disease (1), type 1 diabetes (2), asthma (3), multiple sclerosis (4), rheumatoid arthritis (5), and HIV (6,7). All of these diseases are characterized by aberrant immune responses. HIV-1 infection is associated with alterations in both the gut microbiome and immune cell repertoire. However, it is unknown if these alterations drive or impact one another. Preliminary research suggests mechanistic relationships between immune cell subsets, gut microbes, and disease (8). For example, *Bacteroides fragilis* produces zwitterionic polysaccharides (ZPSs), including polysaccharide A (PSA), which induces the IL-10 production by regulatory T cells (9). IL-10 is also induced through the interaction of *Faecalibacterium prausnitzii* with dendritic cells (10). Segmented filamentous bacteria are associated with the accumulation of IL-17 and IL-22 producing CD4+ T cells (11), which in turn are associated with murine autoimmune arthritis (12). Based on these examples, we might reasonably speculate that there are other undiscovered microbe-immune cell-cytokine relationships, and that they might differ in health and disease.

Exploring the relationships between microbes and immune cell subsets demands expertise from both immunologists and microbiologists, and tools that enable these experts to navigate and explore these data in a team science context. Using paired 16S rRNA microbiome and CyTOF (Time of Flight Mass Cytometry) immune cell repertoire from mucosal biopsy of HIV-infected individuals and controls, we performed linear regressions coupled with permutation testing to identify pairs of microbe genera and immune cell subsets differentially related by disease state. This approach yielded a "top table" of statistically significant associations, similar to those generated in the analysis of gene expression data.

However, each row in our top table represents a sophisticated relationship across a microbe:immune subset pair, not well captured by the test statistic alone. Visual analysis supports vetting of the results for scientific relevance. Vetting includes comparing the relationship in health (HIV-) and disease (HIV+), identifying whether the relationship is driven by one or more outliers, and assessing whether the detailed regression plot is convincing. Furthermore, a particular analyte can appear in multiple relationships. For example, the microbe *Dialister* might be related to 6 immune cell subsets that are each in turn also associated with other microbes. Thus, the identification and inspection of subnetworks of interest is an important task in the visual analysis.

To elucidate the requirements of domain experts in exploring this data, we conducted a visualization design study using Munzner's nested model methodology (13). With input from our collaborating microbiologist and immunologists, our design study led to a web application, VOLARE, (Visualization Of LineAr Regression Elements). This application integrates a searchable top table, sparkline-inspired plots representing the linear models, detailed regression plots, and a network illustrating hubs within the top table. After using VOLARE to analyze microbe:immune cell data from mucosal biopsies, we then analyzed microbe:protein data from fecal samples, thus illustrating the generalizability of the application.

One of our contributions is the characterization of this team science context, in which the omes themselves interrogate distinct but interacting biomes (in this case, the microbiome and the immunome). Consequently, the relationships that span the omes are fundamentally cross-disciplinary and challenging for a single researcher to interpret. This challenge is not unique to microbiome/host relationships. Studies that interrogate relationships between a microbiome and metabolome are facilitated by interactions between individuals who have expertise in these respective domains. Thus, a successful approach must facilitate cross-disciplinary analysis and communication. A second contribution is the visual encoding of the top table and associated regression elements, leveraging existing visualization techniques. We extend the top table, a fundamental tool of single-omic analysis, to two omes. We supplement it with a sparkline of the regression model, from which the researcher can drill down to a detailed regression plot illustrating both regression fit and sample-level detail. The table itself is summarized by an interactive network. An additional strength of our approach is that it is broadly applicable to studies that include data from two or more high-throughput assays.

Related work

VOLARE complements existing applications that are designed to visualize the results of individual microbiome and immune cell assays, regressions, and biological networks, but is geared towards the exploration of integrative analysis of multiple assays. We put VOLARE into context with other applications in this section.

Visualizing results from single assays

Microbiome: Data from high-throughput sequencing of fixed and variable regions of 16S rRNA can be used to identify the microbes present in a biological sample. The relative abundance of the various mi-

crobes characterizes the microbial community of the sample. Qiime is one application that provides tools for both processing the sequence data and generating graphs that capture features of sets of data (14). Web applications like Calypso (15) and Seed (16) generate graphs from processed data, which generally consists of a taxa file (relative abundance values by sample) and a metadata file characterizing each sample. Common visualizations include normalized stacked bar charts, illustrating the taxa that comprise a set of samples; and 2- or 3-D principal coordinates plots (17), illustrating the similarity among various groups of samples. Calypso also provides some machine learning support, such as hierarchical clustering and regression analysis.

Immune cell subsets: Historically, immunologists have identified immune cell subsets using combinations of markers that bind to proteins on or inside the cells. Using a series of 2-D scatterplots and GUI tools, the scientists define regions (known as gates) that identify cells of interest. The regions often indicate presence or absence of one or more markers, and are thus labeled positive or negative for those markers, e.g. CD4+ICOS+. Traditional flow cytometry uses fluorescent tags on the markers, identifying 4 to 11 markers per sample. Mass cytometry (CyTOF) uses metal isotopes and time of flight spectrometry, identifying 30 to 50 markers per sample. Clearly, more cell subsets can be identified with mass cytometry, either through traditional manual gating or automated clustering approaches (18,19). Chester and Maecker provide an accessible overview of some automated approaches (20). Both traditional and automated approaches provide a combination of feature extraction and visualization. However, the visualization tends to focus on one sample or two samples at a time. Citrus supports the comparison of two pools of data, where samples representing the same experimental group (e.g. treated, untreated) are combined (18).

Our work is focused on identifying patterns across omes. As such, it differs from platform specific tools, such as Qiime for analyzing 16S sequencing, or approaches such as tSNE, SPADE, and Citrus for visualizing patterns in CyTOF data. In addition, these platform-specific tools may perform the feature extraction step of identifying and quantifying analytes, be they sequences that have been assigned to a microbe taxonomy or clusters based on immune markers. Our work assumes such identification and quantification has been performed by a platform-appropriate pipeline. Thus, we are able to maintain a separation of concerns between feature extraction and visualization. This allows us to focus on rich visual analysis tools that we can apply to other omes.

Visualizing regression models

One body of related work is that of visualizing regression models. Breheny and Burchett provide a brief summary of over 40 years of such work in the introduction of their generalized approach for regression visualization, the R package visreg (21). They distinguish between plotting models to illustrate the fitted model and plotting models to diagnose assumptions. Like them, we are focused on plotting models to illustrate fit. In general, visualizing model fit focuses on illustrating the results of a single regression model at a time. As such, there is limited emphasis on interactive visualization. In contrast, we consider dozens of fitted regression models concurrently. While each model considers a different pair of analytes, the models are of a common form. VOLARE allows the analyst to quickly inspect several detailed regression plots in a single view.

Visualizing biological networks: hairballs and heatmaps

Another body of related work is biological network visualization (22), represented by such works as RenoDOI (23) and Panther (24). These approaches tend to emphasize genomic relationships (e.g. genes and gene products, genes and transcription factors) and are commonly interactive. One of the challenges they tackle is filtering a very large number of relationships to a smaller, more manageable set that can be explored by a domain expert. In some settings, the relationships are assumed to be reliable, supported by multiple lines of evidence, such as co-occurrence in a publication or pathway or a straightforward experimental construct such as cell line:drug interaction. Both biological networks and microbe:immune system interactions have been represented as heatmaps, with color gradients illustrating p-values or correlation coefficients (23,25,26).

In contrast, rather than consider hundreds or thousands of relationships, we are focused on dozens. Navigating the "hairball" is less of a concern in this top table domain. Furthermore, in comparing human microbiome repertoire to immune cell repertoire, discovered relationships may well be considered speculative. The detailed regression plots allow the domain experts to assess the credibility of the relationships. The underlying detail, readily visualized, is essential to full appreciation of the top table results. Ease of exploration of this detail also sets VOLARE apart from heatmaps. While a heatmap provides an overall gestalt view, the underlying detail is rarely surfaced.

2 Methods

Data preparation

To compare microbiome to proteins in fecal samples, we combined data for 35 study participants into a single file consisting of 71 microbes and 17 proteins. We derived a linear regression model of the form Mb ~ Cohort + Protein + Cohort x Protein and compared those results to those from a reduced model, Mb ~ Cohort. We considered all possible 1,207 microbe:protein pairs, and surfaced the 58 pairs with $p < 0.05$ for exploration in VOLARE. Given these pairs, we organized both the regression results and underlying detail into a single JSON file. Regression results consisted of the names of each analyte in the pair, the observed F statistic, and the endpoints of the two fitted regression lines. Underlying detail consisted of the cohort membership of each study participant and the observed data for each of the analytes that appeared in the 58 pairs. Additional details are included in the R scripts.

To compare microbiome to immune cell repertoire in gut biopsy samples, we combined data for 18 study participants into a single file consisting of 54 microbiome genera with non-zero relative abundance values for at least 9 of 18 samples, and 103 immune cell subsets. To account for the complex correlation structure within the data, we derived a null distribution for the F statistic, F*, by permuting the class labels 5,000 times for 5,562 analyte pairs. Using a false discovery rate of 2%, we identified 126 results with $Fobs > F^*_{95\text{th percentile}}$ (27,28).

3 Results

The initial goal of the domain experts was to identify which microbes are associated with which immune cell subsets, accounting for differences in disease status (HIV positive or negative). To address the question "is the relationship between any particular microbiome genera (Mb) and any particular immune cell subset (Ic) different based on disease status?" we used a partial F test comparing the linear regression model, Mb ~ HIV + Ic + HIV x Ic, to a reduced model, Mb ~ HIV. We chose a linear regression framework because it is well established and extremely flexible. The results are readily quantifiable, with established procedures for assessing statistical significance (28). The regression framework also supports a variety of covariates (e.g. age, sex) and study populations, unlike differential correlation (29) or mutual information (30). As with microarray or RNA-Seq analysis, such as that performed with limma (31), this yielded a "top table" of candidate results. While the top table is an important step in the omics analysis workflow, it is only a starting point for more detailed exploration.

For example, one important question researchers ask is which of the dozens of results in the table are the most promising for follow up. Factors that influence this prioritization include credibility of the regression result, appreciation of the biological role of at least one of the analytes in the pair, and the ability to interrogate the relationship in an *in vitro* experiment. For example, if the microbe is readily available, the researcher can co-culture it with immune cells and assess immunological responses such as cytokine production, cell proliferation, and cell differentiation. To better appreciate the challenges faced by domain experts in analyzing this multiomic top table, we collaborated formed a multi-disciplinary team on the University of Colorado Medical School campus. Our team includes a faculty member, a postdoctoral researcher, and a research assistant from the Allergy and Clinical Immunlogy/Infectious Disease Flow Cytometry Facility, our immunology domain experts; one faculty member from the division of Biomedical Informatics and Personalized Medicine, our expert in microbiology; and a student and faculty member from the Computational Bioscience Program with expertise in statistics and data visualization. Hereafter, we refer to the microbiologists and immunologists as domain experts, and the computational bioscience program members as the investigators.

Task analysis

We developed VOLARE through an iterative design process, with the investigators working closely with the domain experts to gain an understanding of the domain vocabularies, identify analysis tasks and understand the overall analysis workflow, and gather feedback on designs. Initially, we conducted two structured interviews. One interview included both an immunologist and a microbiologist, and specifically targeted the cross-domain analysis aspects and team-science context. Another interview included a one-on-one session with an immunologist, allowing us to better distill domain-specific aspects of the analysis. We then created a first system prototype and iteratively refined it through domain-expert input from multiple feedback sessions. Finally, we distributed the application to our collaborators and observed as they analyzed their data using VOLARE. Throughout this design process, we attended weekly lab meetings that allowed us to better appreciate vocabulary, existing data analysis approaches, and various research concerns.

We identified the following four tasks:

T1: Explore relationships between an analyte of interest and associated analytes in the other ome. For example, the microbiologist might want to know which immune cell populations are associated with *Bacteroides*, while the immunologist might want to know which microbes are associated with cell populations that are positive for ICOS, a T cell co-stimulator related to CD28 (32) and associated with response to anti-CTLA4 immunotherapy (33). While "association" is a somewhat generic term, associations of interest include positive and negative correlations, with support for differing correlations by study cohorts, such as disease status or lifestyle. Analysis of 16S rRNA sequencing of human gut microbial communities results in the detection of relative abundance of thousands of microbial genera. CyTOF characterizes more than 100 immune cell subsets. Each approach measures analytes that are relatively well understood and analytes that are not. Well-understood analytes from one domain may shed insight on the other domain.

T2: Discover patterns of association based on disease status. While some microbes and immune cell subsets may be positively (or negatively) correlated regardless of disease status, there may be other associations that are different by disease status. For example, there may be an association in the healthy gut but not in the diseased gut.

T3: Compare detailed regression plots for several analyte pairs. Regression plots show which analyte pairs have convincing relationships, with credible magnitudes of readouts and well-fitting regression lines. They also indicate presence of outliers. The detailed plot illustrates goodness of fit, presence of outliers, and magnitude of readouts. The domain expert can use these to assess whether or not the relationships would be convincing to other experts in his field.

T4: Identify highly connected "hub" analytes. A particular microbe may be connected to a number of immune cell subsets. Such a microbe might be particularly influential in immune system responses. Conversely, a particular immune cell subset might be connected to a number of microbes.

Visual design

The linear regression framework associates a microbe with an immune readout, while accounting for disease status, in a single model. The corresponding detailed regression plot facilitates cross-disciplinary communication, potentially connecting a known analyte in one domain to an unknown analyte in another. The top table is one of the key elements generated by a "differential expression" analysis in an -omics data set. This table contains a listing of the best results and associated statistics and metrics. However, it is static, obscures underlying detail, and fails to highlight relationships among the analytes. When comparing omes, these relationships can shed insight on biological processes.

Since the top table is such a fundamental element of omics analysis, we built VOLARE around the table. To support **T1**, exploring which analytes in one ome are related to those in another ome, we added an interactive filter function to the top table. When the domain expert enters a microbe or immune marker, the table automatically displays only those relationships that match the phrase. While we could have represented the top table as a matrix or heatmap, the textual and numeric details of the table are essential to communicate the results of the statistical analysis. Furthermore, we can easily add columns to the table, placing additional derived data in context.

To support **T2**, discovering patterns of association based on disease status (Figure 1), we embedded a sparkline-inspired graph of the fitted regression lines in the top table (34). We call this graphic the "mPlot" or microPlot. Making this derived data attribute easily accessible enables the domain expert to scan the table to quickly assess what analytes are involved in what sorts of relationships. As such, it also functions as a small multiple display. The addition of the mPlot enhances an otherwise text-heavy display with a valuable visual element. The mPlot illustrates the regression model using line tilt and color. While the same data could be represented by numeric values for slope, such an encoding would be less conducive to visual analysis. Furthermore, the magnitude of the analyte readouts (and thus the slopes) can vary widely across the data set. The mPlot normalizes the magnitudes by plotting the relationship in a consistently sized glyph, regardless of the magnitude. Figure 2 provides three different mPlot examples, with different interpetations. Figure 2A illustrates a relationship in which the microbe and immune readout are associated in one cohort but not the other, possibly because the microbe is not present in one of the cohorts. Figure 2B illustrates positive association in one cohort and negative association in the other, which might suggest differing biological mechanisms in health and disease. Figure 2C illustrates a much smaller dynamic range of both analytes in one cohort than the

other. While this could be driven by a single outlier, it also could indicate truly different ranges in both analytes across the two cohorts. Thus, even though the mPlot provides a valuable glimpse of the relationship between the analytes, underlying detail is required to fully vet the relationship.

To provide this detail and to better support **T3**, comparing several different analyte pairs, we provide a detailed regression plot in response to clicking the mPlot. Multiple plots can be displayed in the same view. This detailed plot illustrates each data point, colored to indicate disease status, and the corresponding regression fit. Intriguingly, the encoding of a detailed regression plot necessary to convey statistical detail aligns well with best practices of visual encoding. Each point is grounded in a common two-dimensional space, groups are indicated by color, and the group-specific fitted regression lines leverage tilt (35). To support **T4**, identifying highly connected hub analytes, we present a network that summarizes the relationships in the top table. Each node is an analyte (purple = immune cell subset, green = microbe), with each edge indicating a row in the top table. This is an efficient use of screen real estate conveying both high-degree analytes and their relationships. Alternatively, we could have summarized the network with a histogram illustrating analyte degree by assay, but this would not have included the relationships between analytes. Interestingly, while a top table from a gene expression study can be summarized by a pathway analysis, there is no pre-existing pathway data for microbe:immune cell relationships. Thus, the network visually summarizes the top table in this context. Taken together, these encodings support the Shneiderman mantra of overview first, zoom and filter, then details on demand (36). The network and top table provide the overview. The mPlots provide a pre-zoomed representation. The top table itself can be filtered, and the detailed plots are available on demand.

System architecture

Our task analysis revealed that the processes of (1) organizing multiomic data for analysis, (2) performing thousands of pairwise linear regressions and generating an associated top table, and (3) analyzing the results in the top table are fundamentally distinct, separable, and performed by people with different areas of expertise. These distinctions are reflected in our system architecture as shown in Figure 3. Data generated by assay-specific analytical approaches are combined into a single file, with each row representing a sample and each column representing an analyte. Data for different assays can be contributed by different team members, each with assay-specific expertise. Regressions and permutation testing, which require statistical and computational expertise, are performed in R. Supporting details for the candidate results are organized into a single JSON file using the jsonlite library (37). Biological domain experts analyze top table results. VOLARE is implemented as a stand-alone JavaScript application, leveraging the D3 library (38). Because it reflects the fundamental separation of concerns, this architecture can be applied to a wide variety of omes. It also results in a visual analysis environment that is quite responsive to user input, since the computationally intensive analysis has been previously performed and summarized.

Generic workflow

Figure 4 provides a schematic of a representative analysis workflow. The researcher identifies an analyte of interest based on prior knowledge, network community, or mPlot trends. He then filters the table to display the summary results for the subset of results that include this analyte. He then inspects the detailed plots, which may in turn lead to the identification of a new analyte of interest. A specific example of this workflow is discussed in section titled Microbiome:cytokine relationships**.**

Validation

We applied VOLARE to two data sets, each from a different study. We originally developed VOLARE to support the analysis of microbiome:immune cell data, and then analyzed microbiome:cytokine data to demonstrate generalizability. For purposes of illustration, we discuss the microbiome:cytokine data first, followed by the microbiome:immune cell data. In both cases, we present findings that suggest associations between opportunistic bacterial infections and immune readouts. We then discuss user responses.

Microbiome:cytokine relationships

Fecal samples provide a non-invasive source of microbiota and proteins generated by immune cells. Here, we describe a study using such samples, and analysis of the resulting data using VOLARE. Fecal samples were collected from study participants who were HIV negative high risk (HR; men having sex with men, n=17) or low risk (LR, n=18). High risk individuals engage in behaviors that put them at increased risk for acquisition of HIV. Biomarkers of inflammation, innate cell activation and intestinal barrier integrity were measured by ELISA either from the feces itself or water extracted from 2 grams of stool. Fecal samples were also analyzed by 16S rRNA sequencing. A representative analytical workflow is illustrated in Figure 5, and described

in detail below. The steps referenced correspond to the generic workflow in Figure 4. The microbe names have been anonymized due to ongoing research. We start by searching the top table for "Mb_6," a microbe of interest in our lab. The filtered table has only one row, showing that Mb_6 is associated with IL.1alpha (Step 1). The table lists both analytes and the F statistic from the partial F-test. A large statistic represents a better model. We click on the mPlot to inspect the detailed plot (Step 2). The x-axis represents the cytokine data (measured in pg/ml) while the y-axis represents the microbiome data (measured in relative abundance, in the range from 0 to 1). Each point represents the values for one sample from one person. Points are color coded to represent the cohort to which the corresponding person belongs. Lines represent the fitted regression model for each cohort. The closer the points are to the line, the better the model. In this case, there is a strong negative association between the bacteria and IL.1alpha for the low risk group (LR in blue), while there is not much of a relationship between Mb_6 and IL.1alpha for the high risk group (HR in red). However, we see that several people in the high risk group have high levels of IL.1alpha, represented by the rightmost points with values around 1,600 and 1,800 pg/ml. We wonder if another bacteria is associated with these high values. This makes IL.1alpha our new analyte of interest (Step 3). Thus, we filter the table for IL.1alpha (Step 1) and drill down on the detailed plots by selecting several of the related bacteria based on mPlots and F statistics (Step 2). We see that our IL.1alpha outliers are indeed associated with high levels of Mb_8, but not with high levels of Mb_12. We speculate that perhaps the Mb_8 is driving an IL.1alpha immune response. Since we have Mb_8 cultures in our lab, we can consider an *in vitro* experiment to recapitulate this association in cells from other study participants.

Microbiome:immune cell relationships

We used VOLARE for the visual analysis of paired microbiome and immune cell data from gut biopsies of 18 volunteers, half of whom were HIV+ and half HIV-. This was a general study population accrued for a diet intervention study. The samples analyzed here were baseline samples prior to any intervention. To show the utility of our approach, we describe a finding in the negative control population that was unrelated to HIV. Inspecting many of detailed plots, we noted that some of the associations were driven by outliers. Two samples, both from HIV- females, showed high levels of *f__Chlamydiaceae.g__Chlamydophila* (between 30 and 40% of the microbial population) associated with low levels of regulatory T cells (Fig. 6). The presence of two outliers having such a high relative abundance made this an interesting finding warranting exploration in the published literature. While chlamydia is the most common sexually transmitted infection worldwide (39), often presenting in the urogenital tract in women, it is common in the gastrointestinal tract of many mammals and birds (40). Evidence suggests that women can be infected in both the anorectal and urogenital tract, and that an infection can be cleared in one tract but not the other (39,40). Evidence further suggests autoinnoculation from the gut to the urogenital tract (39,40). Regulatory T cells tend to suppress immune response, while CD103 is a marker of mucosal homing (41). The relatively low level of CD103+ regulatory T cells for these two samples may suggest that *chlamydia* bacteria in the gut are able to evade a canonical immune response. Findings that support insight about differential microbiome:immune cell relationships based on disease status will be presented by our domain experts in a different venue.

User responses

The investigators observed the domain experts in a work session reviewing the results presented by VOLARE. In this work session, the microbiologist first leveraged the mPlot to categorize the microbes based on evidence of an association in both HIV+ and HIV-, in HIV+ only, and in HIV- only. Using the detailed plot and domain knowledge, she also identified whether or not the relationship was driven by one or more outliers, likely to be contaminants; or by outliers likely to be informative (e.g. potential opportunistic pathogens such as *Clamydia*). The microbiologist listed the microbes by category in a Powerpoint file and included screenshots of the detailed plots. Next, the lead immunologist looked at the detailed plots that the microbiologist had categorized. He used his domain knowledge to add further narrative to the results. Throughout this process, the scientists shared commentary and insights with one another.

To better quantify the usage patterns, we instrumented the system to log the high-level user actions, such as loading files, searching the top table, and generating detailed plots. Any particular analysis session might involve loading the same file several times to fully reset the visual display. Thus, we used the notion of an "analysis pass" to represent all of the activities from loading the file to the last action performed prior to resetting the display. Figure 7 illustrates metrics for 97 analysis passes collected over 20 days coming from five distinct IP addresses. The results show that most passes last ten minutes or less. The most common action in a pass is the generation of detailed plots, with an average of 12 plots per pass. Comparing the number of detailed plots generated per pass to the number of searches, we can identify three main usage scenarios. One

scenario is "big picture" generation of dozens of detailed plots, unaccompanied by searches. Another scenario is a mix of 2-5 searches and generation of 3-15 detailed plots. This may represent a cycle where one set of detailed plot leads the analyst to search for and inspect another set of detailed plots. A third scenario is zero or one searches combined with the generation of 1-5 detailed plots. This may represent a refinement of an earlier analysis, with a goal of generating a specific set of detailed plots for a screen capture, or a quick check of data or functionality. Taken together, these metrics illustrate that users are very interested in "details on demand" and that VOLARE supports a variety of exploration scenarios. Importantly, the on-demand generation of detailed plots is not a feature provided by a static top table.

The domain experts on our team commented that VOLARE has the potential to dramatically change their workflow. It allows them to screen dozens of candidate relationships quickly. The set of candidate relationships is selected by a solid statistical method applied to all possible pairs of data. This contrasts with manual methods in which researchers select a handful of analytes for inspection. VOLARE also allows domain experts to apply their domain knowledge to vetting the candidate relationships, as opposed to relying on a black box algorithm to suggest promising results. Our domain experts noted that the combination of perspectives (network, top table with mPlot, detailed regression plot) is powerful and exciting. They particularly appreciated having both network and regression visualizations in a single application.

4 Discussion

To identify and explore relationships between microbes and immune cell subsets, we applied a linear regression model to microbe:immune cell pairs. This yielded a top table of relationships, ripe for visual analysis. Collaborating with microbiologists and immunologists, we conducted a design study to characterize the problem space and determine appropriate visual encoding. Our resulting application includes a searchable top table, sparkline representations of the regression models, detailed regression plots, and a network summarizing the table. From the top table, we are able to conceptually zoom out to the summary of the table in the network and drill down to detailed regression plots. Using this application, our domain experts were able to quickly prioritize relationships for follow-up research. We repeated the process with microbe:cytokine data, illustrating generalizability of the approach. Our contributions include characterizing the exploration of multiomic data in a cross-disciplinary team science context, and the support of an associated workflow by integrating existing visualization approaches.

There are a number of strengths in our approach. First, the linear regression framework is both well established and flexible, accommodating a wide variety of experimental designs. Second, performing the regressions and permutations off-line yields a fast and responsive web application for visual analysis. Third, the overall presentation of a searchable top table including the mPlot, a detailed plot, and a network summarizing the table gives the researchers a variety of tools for exploring the results of their studies. Fourth, systematically integrating and linking these approaches in one tool allows the domain experts to control the analysis of their results, empowering them to screen dozens of candidate relationships with ease. Fifth, the visualization environment facilitates communication between the domain experts and the bioinformatics experts, allowing both groups to better appreciate the nuances of the data. Sixth, the approach is broadly applicable to a variety of high throughput assay pairs, such as microbe:metabolome, microbiome:proteome, and RNA-Seq:immune repertoire.

There are several limitations to this work. First, as presented here, we have only considered two omes. While more omes could be included by increasing pairwise comparisons, the pairwise approach is self-limiting to a handful of omes. That said, the support for visual analysis of promising regression results from two or more omes is a valuable contribution. Second, the regressions are performed by stand-alone computing resources, with necessary results and underlying details marshaled for the visualization layer. This means that changes to the regression model cannot be made on the fly by end users. However, the regression analysis requires some statistical experience that the domain experts may not have. Thus, this is a natural breakpoint for separating the analysis workflow. In addition, the existing approach of handing off the data to a statistician for analysis has the same limitation. Third, we do not finely tune the regression model for each analyte pair. Instead, we use the same form of the regression model for all pairs, with the results providing a screening mechanism. Domain experts are then able to assess model fit and scientific relevance. Fourth, VOLARE does not provide strong support for extracting and modifying the graphs for presentation and publication. However, its main goal is data exploration rather than presentation. Currently, graphs can be extracted either by screenshot, or by copying svg elements from the document object model of the web page into an svg editor such as Inkscape or Adobe Illustrator.

Our future work includes adding features to VOLARE such as grouping by mPlot, automatically indicating relationships driven by outliers, searching the top table by Boolean expressions of analyte names, and displaying the detailed plot in response to clicking on a network edge. We would like to apply VOLARE to data sets that include different omics platforms, such as paired RNA-Seq and immune cell repertoire, and

paired microbiome and metabolome. The work presented here demonstrates that visual analysis allows biologists to quickly identify and vet top table relationships between two different omic domains, using the information from one platform to better appreciate results from the other platform. Furthermore, the relationships provide insight into possible biological mechanisms. In conclusion, our approach helps domain experts answer two main questions, "What microbes are associated with what immune cell subsets?" and "How do I prioritize the top table results for follow up?"

Acknowledgements

We thank Mike Shaffer for serving as a scribe in structured interviews. We thank Casey Martin for processing the 16S data and Sharon Sen for processing the CyTOF data used in the microbiome:immune cell analysis.

Funding

This work has been supported by NIH Grant 5 R01 LM009254 11.

Conflict of Interest: none declared.

References

1. Sartor RB. Microbial influences in inflammatory bowel diseases. Gastroenterology. 2008 Feb;134(2):577–94.

2. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC, et al. Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature. 2008 Oct 23;455(7216):1109–13.

3. Huang YJ, Boushey HA. The microbiome in asthma. J Allergy Clin Immunol. 2015 Jan;135(1):25–30.

4. Cekanaviciute E, Yoo BB, Runia TF, Debelius JW, Singh S, Nelson CA, et al. Gut bacteria from multiple sclerosis patients modulate human T cells and exacerbate symptoms in mouse models. PNAS. 2017 Oct 3;114(40):10713–8.

5. Chen J, Wright K, Davis JM, Jeraldo P, Marietta EV, Murray J, et al. An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. Genome Medicine. 2016 Apr 21;8:43.

6. Lozupone CA, Li M, Campbell TB, Flores SC, Linderman D, Gebert MJ, et al. Alterations in the gut microbiota associated with HIV-1 infection. Cell Host Microbe. 2013 Sep 11;14(3):329–39.

7. Li SX, Armstrong A, Neff CP, Shaffer M, Lozupone CA, Palmer BE. Complexities of Gut Microbiome Dysbiosis in the Context of HIV Infection and Antiretroviral Therapy. Clin Pharmacol Ther. 2016 Jun;99(6):600–11.

8. Arnolds KL, Lozupone CA. Striking a Balance with Help from our Little Friends – How the Gut Microbiota Contributes to Immune Homeostasis. Yale J Biol Med. 2016 Sep 30;89(3):389–95.

9. Neff CP, Rhodes ME, Arnolds KL, Collins CB, Donnelly J, Nusbacher N, et al. Diverse Intestinal Bacteria Contain Putative Zwitterionic Capsular Polysaccharides with Anti-inflammatory Properties. Cell Host Microbe. 2016 Oct 12;20(4):535–47.

10. Martín R, Bermúdez-Humarán LG, Langella P. Searching for the Bacterial Effector: The Example of the Multi-Skilled Commensal Bacterium Faecalibacterium prausnitzii. Front Microbiol [Internet]. 2018 [cited 2018 Sep 2];9. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00346/full

11. Ivanov II, Atarashi K, Manel N, Brodie EL, Shima T, Karaoz U, et al. Induction of Intestinal Th17 Cells by Segmented Filamentous Bacteria. Cell. 2009 Oct 30;139(3):485–98.

12. Wu H-J, Ivanov II, Darce J, Hattori K, Shima T, Umesaki Y, et al. Gut-residing segmented filamentous bacteria drive autoimmune arthritis via T helper 17 cells. Immunity. 2010 Jun 25;32(6):815–27.

13. Munzner T. A Nested Model for Visualization Design and Validation. IEEE Transactions on Visualization and Computer Graphics. 2009 Nov;15(6):921–928.

14. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May;7(5):335–6.

15. Zakrzewski M, Proietti C, Ellis JJ, Hasan S, Brion M-J, Berger B, et al. Calypso: a user-friendly web-server for mining and visualizing microbiome–environment interactions. Bioinformatics. 2017 Mar 1;33(5):782–3.

16. Beck D, Dennis C, Foster JA. Seed: a user-friendly tool for exploring and visualizing microbial community data. Bioinformatics. 2015 Feb 15;31(4):602–3.

17. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience. 2013 Nov 26;2(1):16.

18. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. PNAS. 2014 Jul 1;111(26):E2770–7.

19. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011 Oct 2;29(10):886–91.

20. Chester C, Maecker HT. Algorithmic Tools for Mining High-Dimensional Cytometry Data. The Journal of Immunology. 2015 Aug 1;195(3):773–9.

21. Breheny P, Burchett W. Visualization of Regression Models Using visreg. The R Journal. 2017;

22. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. Nature Methods. 2010 Mar 1;7(3s):S56.

23. Vehlow C, Kao DP, Bristow MR, Hunter LE, Weiskopf D, Görg C. Visual analysis of biological data-knowledge networks. BMC Bioinformatics. 2015 Apr 29;16:135.

24. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017 Jan 4;45(D1):D183–9.

25. Routy B, Chatelier EL, Derosa L, Duong CPM, Alou MT, Daillère R, et al. Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. Science. 2018 Jan 5;359(6371):91–7.

26. Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, et al. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science. 2018 Jan 5;359(6371):97–103.

27. Wagner BD, Zerbe GO, Mexal S, Leonard SS. Permutation-based adjustments for the significance of partial regression coefficients in microarray data analysis. Genet Epidemiol. 2008 Jan;32(1):1–8.

28. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and Analysis of DNA Microarray Investigations. 2003 edition. New York; London: Springer; 2011. 200 p.

29. Siska C, Bowler R, Kechris K. The discordant method: a novel approach for differential correlation. Bioinformatics. 2016 01;32(5):690–6.

30. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. PNAS. 2014 Mar 4;111(9):3354–9.

31. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.

32. Hutloff A, Dittrich AM, Beier KC, Eljaschewitsch B, Kraft R, Anagnostopoulos I, et al. ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28. Nature. 1999 Jan 21;397(6716):263–6.

33. Chen H, Liakou CI, Kamat A, Pettaway C, Ward JF, Tang DN, et al. Anti-CTLA-4 therapy results in higher CD4+ICOShi T cell frequency and IFN-gamma levels in both nonmalignant and malignant prostate tissues. Proc Natl Acad Sci USA. 2009 Feb 24;106(8):2729–34.

34. Tufte ER. The Visual Display of Quantitative Information. 2nd edition. Cheshire, Conn: Graphics Pr; 2001. 200 p.

35. Munzner T. Visualization Analysis and Design. 1 edition. Boca Raton: A K Peters/CRC Press; 2014. 428 p.

36. Shneiderman B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: The Craft of Information Visualization [Internet]. San Francisco: Morgan Kaufmann; 2003 [cited 2018 Jan 27]. p. 364–71. (Interactive Technologies). Available from: https://www.sciencedirect.com/science/article/pii/B9781558609150500469

37. Ooms J. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:14032805 [cs, stat] [Internet]. 2014 Mar 12 [cited 2018 Jan 29]; Available from: http://arxiv.org/abs/1403.2805

38. Bostock M, Ogievetsky V, Heer J. D3; Data-Driven Documents. IEEE Transactions on Visualization and Computer Graphics. 2011 Dec;17(12):2301–9.

39. Heijne JCM, Liere GAFS van, Hoebe CJPA, Bogaards JA, Benthem BHB van, Dukers-Muijrers NHTM. What explains anorectal chlamydia infection in women? Implications of a mathematical model for test and treatment strategies. Sex Transm Infect. 2017 Jun 1;93(4):270–5.

40. Rank RG, Yeruva L. "Hidden in plain sight:" Chlamydial gastrointestinal infection and its relevance to "persistence" in human genital infections. Infect Immun. 2014 Jan 13;IAI.01244-13.

41. Kreisman LSC, Cobb BA. Glycoantigens Induce Human Peripheral Tr1 Cell Differentiation with Gut-homing Specialization. J Biol Chem. 2011 Mar 18;286(11):8810–8.

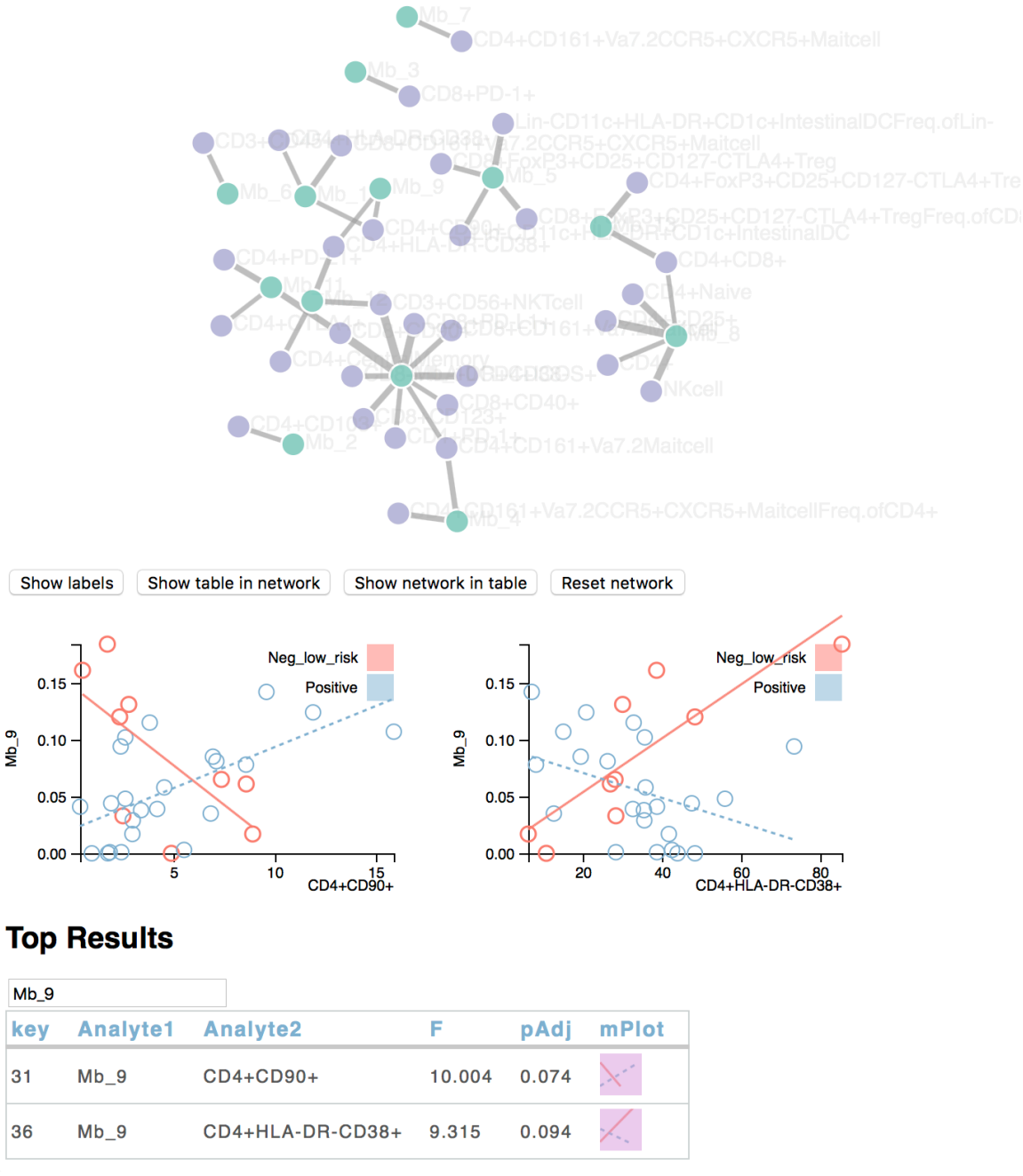

Fig. 1. VOLARE overview: Network at the top, two detailed regression plots below, and top table at the bottom. Buttons add labels to the nodes, synchronize the table with the network, or synchronize the network with the table.

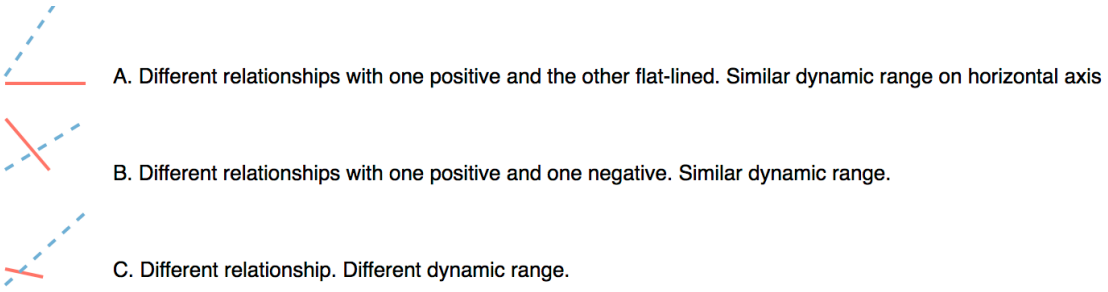

Fig 2. mPlot examples take from actual data. Solid and dotted lines represent different cohorts. Vertical axis represents microbe relative abundance, while the horizontal axis represents the immune readout. Three examples illustrate the different relationships that can be encapsulated in the sparkline-inspired mPlot. (A). The relationship between the microbe and the immune readout exists in one cohort but not the other. This might suggest that the microbe is absent in the "flat line" group. (B). Differences in relationship between the microbe and immune readout across the two cohorts might suggest biological differences across the cohorts. (C). The difference in dynamic range across the cohorts might suggest that the relationship captured by the longer line is driven by an outlier, with high values in both analytes.

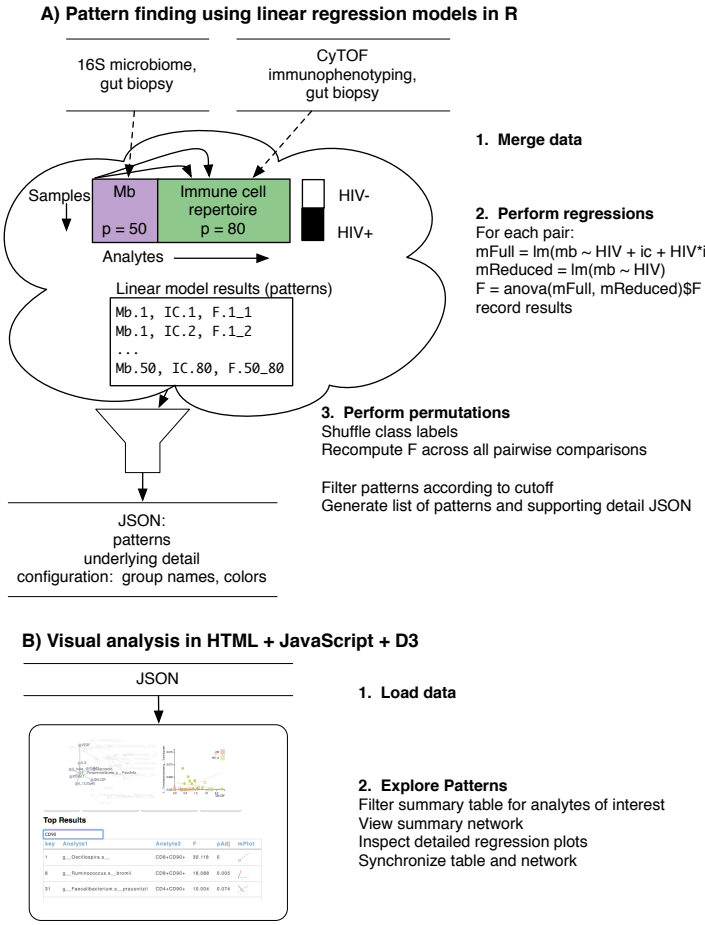

Figure 3. VOLARE implementation architecture. The architecture reflects a separation of concerns in which 16S and CyTOF data, having been processed by domain-specific pipelines, are combined for regression analyis. The resulting patterns are then available for visual analysis. A) Pattern identification is performed in R. P indicates number of analytes in data collection. Results of each regression analysis are recorded, with Mb.1 and IC.1 representing the first microbe and first immune cell respectively. F.1_1 represents the F statistic from the linear model using Mb.1 and IC.1, while F.50_80 represents the F statistic from the model using Mb.50 and IC.80. The R script generates a JSON file that includes a summary top table of the patterns, underlying data, and configuration data for the web application. B) Pattern exploration is performed using a JavaScript web application.

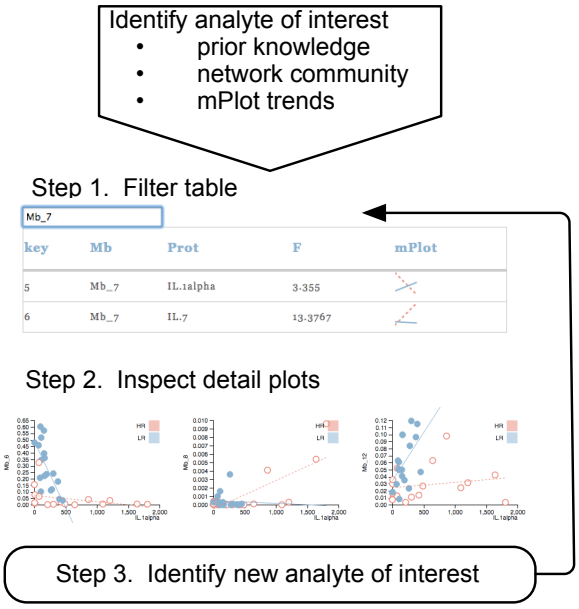

Fig 4. Generic workflow. The researcher identifies an analyte of interest based on a variety of sources. He filters the top table to find relationships including this analyte. He then inspects detailed regression plots. Based on these, he identifies a new analyte of interest and repeats the process.

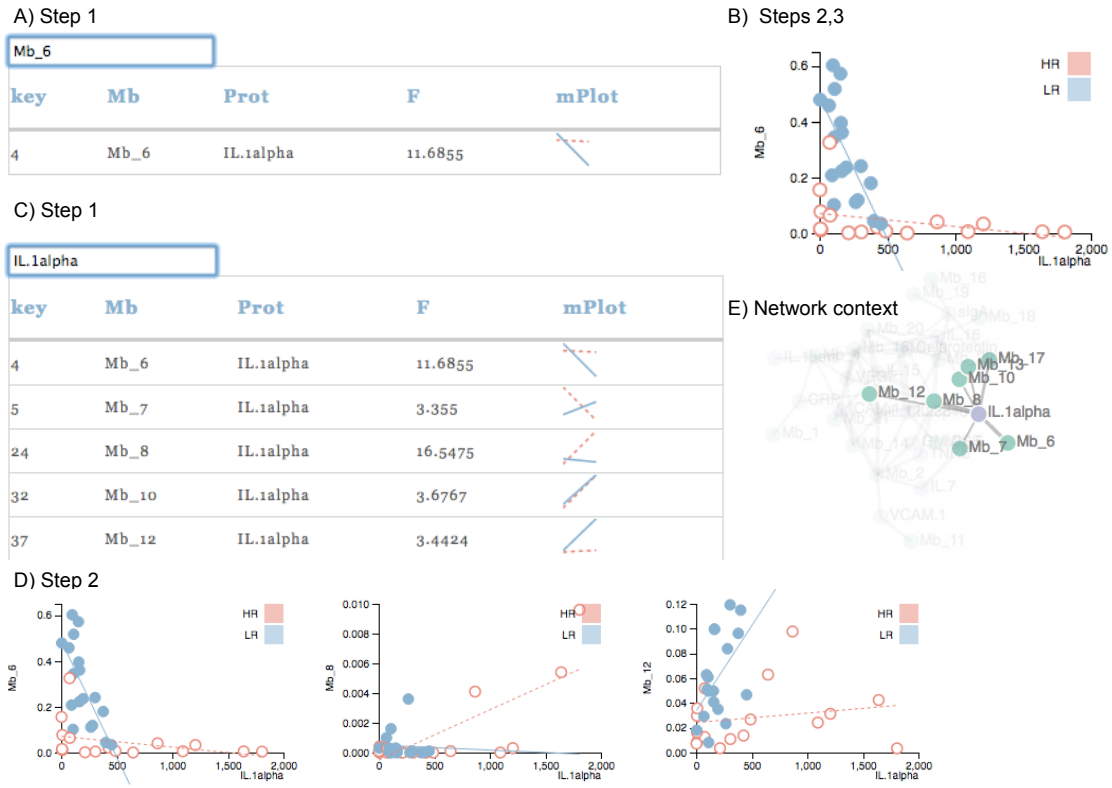

Fig 5. Example workflow. (A) The researcher searched for a microbe of interest in the top table, and then (B) generated and detailed regression plots. She noticed potential outliers, high in IL.1alpha, so (C) searched for this protein. (D) She observed that IL.1alpha high is associated with relatively high levels of Mb_8 but not Mb_12. (E) Network with IL.1alpha as a hub connected to 8 proteins.

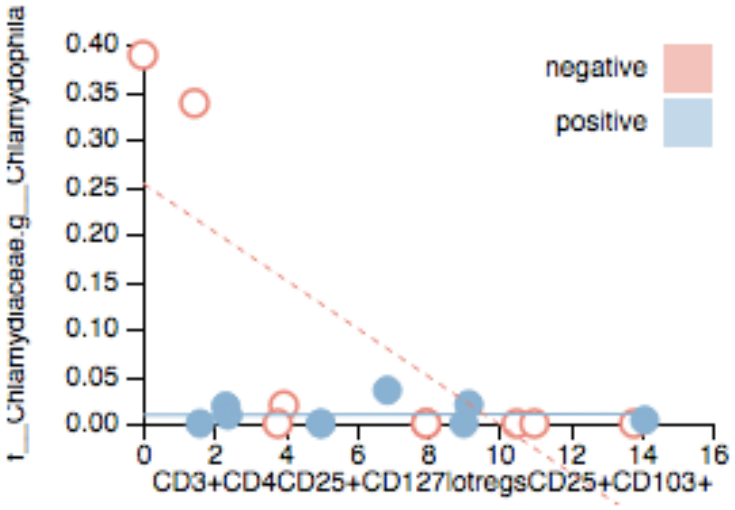

Fig. 6. Two samples from HIV- females show evidence of *chlamydia* infection, coupled with low levels of regulatory T cells.

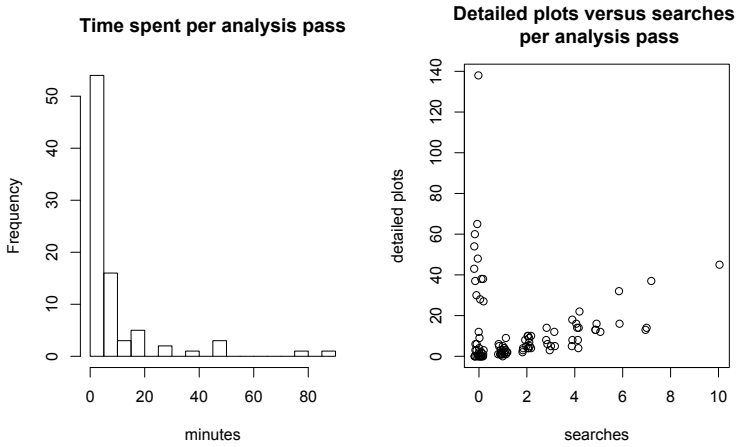

Fig 7. Usage scenarios. An analysis pass consists of loading a file and exploring the data, and lasts until the visual display is reset. (A) Most analysis passes last less than 20 minutes, but have lasted up to 90 minutes. (B) A comparison of the number of detailed plots generated versus the number of searches suggests 3 different analysis scenarios. One scenario is "big picture" generation of dozens of detailed plots, unaccompanied by searches (searches=0, dPlots greater than 20). Another scenario is a mix of 2-5 searches and generation of around 3-20 detailed plots. A third scenario is zero or one single searches combined with the generation of 1-10 detailed plots. Data is jittered on the horizontal axis to reduce overplotting.