Eurographics Conference on Visualization (EuroVis) 2016
K.-L. Ma, G. Santucci, and J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

Visualizing Co-occurrence of Events in Populations of Viral Genome Sequences

A. Sarikaya[1], M. Correll[2], J. M. Dinis[3], D. H. O'Connor[4,5], M. Gleicher[1]

[1] Department of Computer Sciences, University of Wisconsin—Madison, USA
[2] Department of Computer Science and Engineering, University of Washington, USA
[3] Department of Pathobiological Sciences, University of Wisconsin—Madison, USA
[4] Department of Pathology and Laboratory Medicine, University of Wisconsin—Madison, USA
[5] Wisconsin National Primate Research Center, USA

Abstract

Virologists are not only interested in point mutations in a genome, but also in relationships between mutations. In this work, we present a design study to support the discovery of correlated mutation events (called co-occurrences*) in populations of viral genomes. The key challenge is to identify potentially interesting pairs of events within the vast space of event combinations. In our work, we identify analyst requirements and develop a prototype through a participatory process. The key ideas of our approach are to use interest metrics to create dynamic filtering that guides the viewer to interesting and relevant correlations of genome mutations, and to provide visual encodings designed to fit scientists' mental map of the data, along with dynamic filtering techniques. We demonstrate the strength of our approach in virology-situated case studies, and offer suggestions for extending our strategy to other sequence-based domains.*

Categories and Subject Descriptors (according to ACM CCS): J.3 [Life and Medical Sciences]: Biology and Genetics—

1. Introduction

Many analytic activities involve understanding *events* in sequences. Events may be significant points in time-series data, locations in text documents, or positions along a genomic sequence. A wide variety of techniques in the visual analytics literature focus on identifying and interpreting events as sparse sets of interesting locations in a sequence. However, the problem of identifying interesting patterns of *co-occurrence* of observations relating events together is much less studied. Examining co-occurrence requires considering a much larger space than with individual events: rather than the one-dimensional space of a sequence, co-occurrence must consider the space of all pairwise relationships. Additionally, analysis must consider incomplete data, as observations may not capture all pairs of events.

In this paper, we present a design study for the problem of the identification and analysis of co-occurrences of mutations within DNA sequence data. In our design study we gather requirements, determine an abstraction of the problem, formulate a strategy based on prior research, evaluate prototypes, and arrive at a final visualization design, driven by participatory design with our collaborators. Through this process, we encountered issues of scale associated with displaying all potential correlations. A key idea in our strategy is to define metrics for quantifying "interestingness," affording a user-driven exploration of the space of correlations. While our motivating application is the population dynamics of viruses and correlation of mutations, we believe the lessons from this design study have broader applicability to discovering correlations in other one-dimensional sequence data.

The specific biological question we consider involves the mutation patterns that a virus makes over the course of its infection in a specific host-individual. When a host is infected with a virus such as HIV or influenza, the virus rapidly makes many copies of itself. Because replication is imperfect, many of the copies of the virus will contain multiple point mutations [S*10]. Some of these variants are advantageous and accumulate within the virus population (e.g., variants that evade the host's immune response). New deep sequencing technologies enable surveillance of viral genomes throughout entire populations. While workflows currently exist for identifying correlation between two genomic positions, the analysis process is a manual effort and prone to errors. Better analysis tools and support are needed to rapidly identify significant co-occurrences of mutations in genomes.

Our contribution is a design study (see Sedlmair *et al.* [SMM12]) of the rapid identification of correlations between mutations in populations of a viral genome, where technology has become available to understand the population dynamics of viruses. We provide a

characterization and abstraction of the problem, allowing us to propose a solution for the generalized problem. We consider standard encodings for sequence and correlation data, and explore their use in an initial prototype. Through a participatory design, we reconcile failures in early prototypes and iterate on our design to better match virologists' needs. We assess this system through two case studies, and end with a discussion of the lessons learned through the problem characterization and the design study.

2. Biological Background

Our work is a part of an established collaboration between virologists and computer scientists to develop better tools for understanding the genetic mechanisms involved in viral infections. Team members from both backgrounds have worked together to build an understanding of problems, and have evolved tools that address them. Here, we describe the general problem of understanding viral population dynamics and the need for new tools to examine co-occurrence in this domain.

For the purposes of this paper, the key biological concept is that the genome replication process of RNA viruses (e.g. HIV, influenza) is highly error-prone, resulting in the incorporation of random mutations of nucleotides throughout the viral genome. In an infected host, HIV and influenza exist as a diverse collection of similar yet distinct viral particles, each with its own genome. While most mutations in RNA viruses are catastrophic to the continued survival of the virus, those mutations that are beneficial to viral fitness continue to propagate. Generally, the longer a virus has infected the host, the more variation in the viral population.

Identifying combinations of mutations (*co-occurrences*) in the viral genome is critical for understanding important biological functions. For example, simultaneous mutations at three or four positions on an external viral protein haemagglutinin (HA) of an avian H5N1 influenza virus permits transmission to mammals [I*12]. Interestingly, these mutations do not confer transmission individually, but rather they need to exist together on the same viral genome (a concept named *epistasis*). Epistatic mutations are co-occurring mutations that, together, can allow a new biological function. Identifying co-occurring mutations from virus populations allows for detailed characterization of genetic diversity and accurate assessments of viral function. A global view of co-occurrence can help understanding of how a virus works at a high-level, and serves to target *in vivo* experimentation of viral activity of larger epistatic interaction.

Nucleotides that mutate can cause the functionality of a virus to change by affecting emitted proteins. Regions of the genome that code for proteins are called open reading frames (ORFs), where a reading frame is a particular sequence of codons, which themselves are triplets of nucleotides. The translation from codons to amino acids (the building blocks of proteins) is degenerate as there are 64 unique codons (4^3) and just 20 amino acids that can be represented by the genetic code. Therefore, a mutation in the genome does not necessarily confer a change in protein coding—these are instances of *synonymous mutations*. Identifying these synomymous mutations as not significant mutations are important to consider (though even synonymous mutations may have RNA structure—and thereby functional—implications).

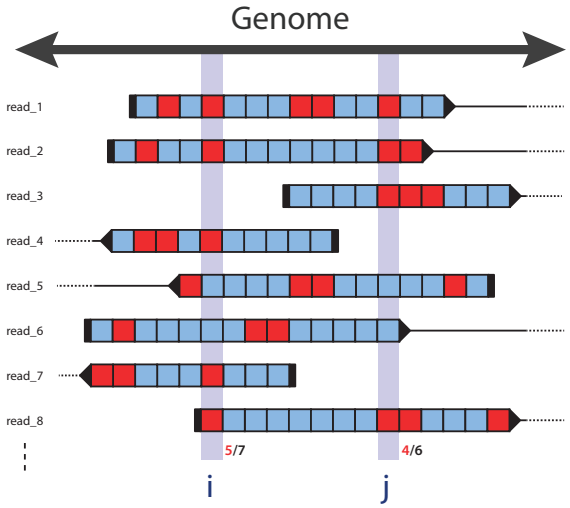

Figure 1: *A visual abstraction of viral genomic data, where red boxes denote nucleotides that do not match the reference genome. Rows are individual reads from NGS, while columns are genomic positions. Two positions (i, j) are checked for mutation co-occurrence.*

The rapid identification of epistasis and characterizing the functionality of sub-populations remains a challenging task. New genomic sequencing technology allows for analysis of the diverse genomic populations and continued disease progression. In particular, Next-Generation Sequencing (NGS) analyzes millions of nucleic acid sequences simultaneously, enabling detailed characterization that captures the proportional presence of viral sub-populations in a sample. The output of the NGS system are aligned sequences of short-read data—see Figure 1 for an abstract representation. These reads (on the order of hundreds of thousands) have associated start points in the global genomic sequence space. Due to limitations in current technology, however, only 300–500 nucleotides can be called for each read, limiting the range of co-occurrences that can be observed in pairwise genomic space. Newer techniques, such as including analysis from "paired reads," can increase this bandwidth, but still represents a hard limit on analyzing distant pairs in the genome.

3. Problem Details and Requirements

The process of discovering these co-occurrences of mutations in viral populations is not well-supported by any existing tool. Current workflows for discovering sub-populations demand either expensive processes examining all potential combinations, manual curation and exploration through the data using tools such as Microsoft Excel, or line-by-line inspection of aligned reads in programs such as Geneious Pro [K*12], CLC Genomics Workbench [CLC], or the Integrated Genome Viewer [R*11].

Our discussions with virologists identified two main analysis goals. The first is an idea of **diversity**: a better understanding of the amount of variation in sequence space. For example, higher variation within a population could indicate there are environmental

pressures (e.g. an effective immune response) that is forcing the virus to diversify to survive. The second insight regards **functionality**, where the population of viruses can be separated into sub-populations that share coordinated mutations. This separation can provide researchers with a vector of attack to characterize the viral sub-population *in vivo* to see if a functionality shift is occurring.

The general problem is to identify pairs of genomic positions where mutations are observed to co-occur together. If we think of the reads (rows in Fig. 1) as observers making measurements about events in a global context (columns in Fig. 1), we can begin to determine how these observers connect these events. To understand the correlation of events, we can gather statistics from pairs of positions that are observed together—we call this *observer continuity*. In contrast, looking at observations without regard of observer continuity reduces to an independent event comparison problem, which is supported by existing visual analysis techniques for time-series data (*cf.* [JME10]) or existing metrics (such as mutual information [S*02]).

From this problem characterization, and through iteration and discussion with our collaborators (see §6.1), we collected a series of analysis tasks. The first (**T1**) is to *identify* significant co-occurrences of mutations. Virologists must be able to *explore in detail* a co-occurrence pair (**T2**), evaluating whether the particular correlation is important and requires further research. Important co-occurrences within the entire genome must be easily *summarized* (**T3**), requiring overview of all significant correlations in the genome.

We collected additional requirements based on the specific task domain. The presentation of the data in the visualization must align with the analysts' existing mental models of genomic data by (1) always presenting data in genomic sequence order (**R1**) and (2) displaying annotations alongside the genome to provide wayfinding for the analyst (**R2**). We found through discussions with virologists that a mental 'map' helped to orient themselves while navigating the viral genome. To be able to discover significant co-occurrences, there needs to be a scaffold to navigate the space of all pairwise correlations (**R3**). Finally, the visualization must scale to the typical dataset size (**R4**): hundreds of genome positions and hundreds of thousands of individual read segments, while remaining interactive to the anaylst in a web-browser-based deployment (which simplifies sharing of datasets).

Our approach to deal with the vast space of correlations is to define interest metrics to aid in filtering. Discussions with stakeholders suggested that there are a variety of factors to consider in developing such metrics. The simplest of these measures is *positive correlation*, which can indicate potential epistatic mutations. The inverse, *negative correlation*, can also be interesting, demonstrating that combinations of mutations can be catastrophic to viral fitness. Secondly, there may be issues with *coverage*, where there may not be enough observations relating two positions to make significant judgments about correlation. Finally, the base rate of mutations at a particular position must be over the error rate of the NGS sequencer to be significant, otherwise spurious correlations that are misaligned may be counted as significant. We elaborate on these metrics in Section 5.

4. Related Work

4.1. Visualizing genomic data

There are many genomic data viewers that support the visualization and analysis of variants (see Nielsen *et al.* [N*10] for a broad survey). The most common of these analysis tools are genome browsers, which juxtapose the raw genomic sequence alongside supplemental data, such as computational predictions and homologies. There are many examples of these tools, each of which are specific either to a particular task (e.g. resolving reads from NGS data [F*10]) or a particular biological domain (e.g. cancer [D*12] or humans [K*02]). Although there are many browsers, most make assumptions that break our model of multiple, competing viral genome populations. For example, the MuSiC system [D*12] contains functionality that identifies statistically-probable correlations of mutations [D*08]. In particular, the use of fixed statistical judgments and sub-sampling methods are not well-suited to analysis of a viral population, as it assumes that non-matching reads are errors instead of an indication of a sub-population.

Specialized genomic visualizations can make visual encoding decisions that directly support particular analysis tasks. These systems either expose trends and relationships between annotations [V*13], between variants and annotations [FNM13, D*13], or between alternative splicing of genes [SAB*16]. Many of these systems directly encode correlation. COMBat [V*13] uses a reorderable matrix view to highlight correlation between annotations, intentionally scrambling the genomic axis. Variant View [FNM13] uses tracks to show overlapping annotations, as well as concise glyphs to convey information on types of mutations at particular positions. DecisionFlow [GS14] allows the analyst to drive exploration through a large electronic health record space and presents health outcomes in Sankey-like diagrams, while Vials [SAB*16] uses a common genomic axis to ground analysis of splice groups. While some of these tools violate several of our initial requirements (e.g. COMBat violates **R1** and **R2**, Variant View doesn't scale to the data scales needed in this application **R4**), they provide precedent for the visual support of our three tasks (**T1–3**).

Many solutions for analyzing viruses, like Alvira [EFBF07], use a 'scaffold view' where sequencing reads are stacked atop one another, mutations are highlighted, and frequency of variants is highlighted by proportional sequence logos. These visual encodings have notorious disadvantages, including inablility to scale and and potentially skewing proportionality judgments (see Maguire *et al.* [M*14] for a discussion), suggesting a more principled ensemble encoding. Similar to our system, LayerCake [C*15] supports finding variants between multiple aligned samples of populations of viral genomes by using color as an ensemble encoding, compressing horizontal space by binning positions together but otherwise maintaining strict sequence order. LayerCake highlights population dynamics only between viral samples, not within a particular sample. Therefore, LayerCake does not support discovering correlations between mutations as there is no notion of observer continuity.

4.2. Visualizing correlation

Visualizing correlation between events is a task of substantial interest in the visual analytics literature. Two primary methods of

visualizing relationships between elements are through node-link and matrix-based visualizations (see Ghoniem *et al.* [GFC04] for a discussion). While node-link visualizations have issues of scale with increasing number of nodes, they are invaluable for analyzing multi-stage connections. On the other hand, matrices excel at larger number of connections, though they suffer at providing aggregate judgments (*cf.* Diaz *et al.* [DPS02]).

Several studies have modified the typical uses of node-link and matrix-based visualizations to uncover trends in combinatorial relationships. Henry and Fekete [HF06] use a matrix view in conjunction with a node-link view to better support analysis tasks of social network connections between individuals. Dunne and Shneiderman [DS13] use aggregate glyphs to represent common visual patterns in node-link diagrams, managing complexity in the number of elements and connections shown. Other visual methods such as parallel sets [BKH05], parallel coordinates [Ins97], and Sankey diagrams show how similar elements relate to one another through many continuous or categorical dimensions. These methods are helpful for conveying a general sense of how a subset interacts with different data dimensions, and we use parallel sets to visually communicate the level of correlation in a co-occurrence pair.

We use general trends found in these works to inform our own design decisions. For example, we anticipated in early designs that a matrix view would be a good way to re-order positions to identify significant co-occurrences. This led us to the requirement of maintaining genome continuity (**R2**) in order to support the virologists' mental models, upon which we elaborate in Section 6.

5. Interest Metrics

To reduce the correlation space that an analyst needs to explore, we identified the following three metrics that capture the intuitions of our audience for what is considered an "interesting" correlation. The first is *coverage*: we must have a sufficient number of the events in order to be confident that the measures we receive are not due to sampling error or noise. The second is *variation*, where each of the two sites must have a sufficient diversity of observations. The third is a *metric of co-occurrence*, which quantifies how unlikely is the relationship between the two sites relative to chance, given what would be expected by the statistics at each individual position under an assumption of independence.

Abstractly, we consider the set of events $\mathbf{E}$ in a data sequence, and observers $\mathbf{O}$ that make observations about those events. Each observer O_k therefore represents a set of observations of the form $\{(i,+),\cdots(j,-)\}$, where each tuple contains a reference to an event in $\mathbf{E}$ (e.g. position i) and a categorical observation (e.g. $+$)—for example, if a mutation is observed at this position or not (a tuple is a square in Figure 1). Throughout our notation, we use Q as a collector of observers that have made a given observation about an event.

These metrics are summarizations of a co-occurrence pair, but do not individually confer a clear indication of significance. In different situations, an analyst may have different considerations. Therefore, we allow the user to dynamically set thresholds for these metrics.

Coverage metric: The *coverage* metric C_i for a particular position i counts the number of observations made about a position and can be used to determine coverage in comparison to other positions. C_i is computed by gathering all observers $B \in O$ that reference the position i and counting the number of observations in the returned set.

$$C_i = |Q_{(i_*)}| = |\{B \in \mathbf{O} \mid (i,*) \in B\}|. \tag{1}$$

We can extend this definition of Q to select sequences that have a particular type of observation at a position. As an example, $Q_{(i_-)}$ would match sequences that have observations at i that are negative.

Variation metric: The *variation* metric V_i can be used to threshold the prior probability for a variant to occur at a position. As an example, V_{i_-} below is the percentage of reads that are mutations at position i in our genomics context:

$$V_{i_-} = \Pr(i_-) = \frac{|Q_{(i_-)}|}{|Q_{(i_*)}|}. \tag{2}$$

Co-occurrence metric: Correlations that are interesting tend to be those where observations regarding one position seem to be conditionally dependent on the observation at another. To quantify this, we first count the observers of both occurrences. We augment Q again, capturing observations about a pair of positions, taking into account observer continuity—that is, an observation is only considered if and only if it contains data about both i and j:

$$Q_{(i_-,j_+)} = \{B \in \mathbf{O} \mid (i,-) \in B \wedge (j,+) \in B\}.$$

Now, we can define a conditional probability. Let us assume that we are interested in the conditional probability that an observation is negative at position j given the negative observation at i:

$$\Pr(j_-|i_-) = \frac{|Q_{(i_-,j_-)}|}{|Q_{(i_-,j_*)}|}.$$

With these formulations, we can define a *co-occurrence metric* M_{i,j_*}.

$$M_{i,j_-} = \Pr(j_-|i_-) - \Pr(j_-|i_+) = \frac{|Q_{(i_-,j_-)}|}{|Q_{(i_-,j_*)}|} - \frac{|Q_{(i_+,j_-)}|}{|Q_{(i_+,j_*)}|}. \tag{3}$$

This metric is similar to metrics such as mutual information (see Steuer *et al.* [S*02]). A key difference is that it takes account of observer continuity, allowing us to use conditional probability in our metric, in contrast to depending on joint probability (a potentially weaker assertion). Our metric also yields values in a fixed domain $[-1,1]$, where -1 identifies strong negative correlation, 1 denotes strong positive correlation, and 0 implies no correlation. This is in contrast to mutual information, which has an unbounded, unsigned domain.

5.1. Interestingness in the virology problem

With Next-Generation Sequencing technology, researchers have the ability to understand the population dynamics of highly varying samples of viruses without the limitations of previous sequencing technology that would implicitly boost only the sequences with the highest occurrence. In engineering our solution, we decided to implement a pre-computation process that would compile counts of

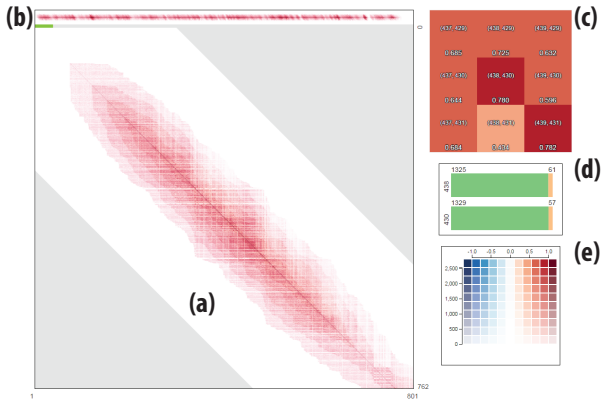

Figure 2: *Our initial prototype to identify pairwise correlations between all positions i (x-axis) and j (y-axis). The matrix view (a) shows these co-occurences, and the overview (b) provides a horizontal overview of the space. The super-zoom window (c) highlights the coordinates and co-occurrence metric currently under the cursor, while the bar chart (d) presents the proportion of reads at a selected pair of positions. The legend (e) presents the 2D color key.*

paired bases (all paired combinations of Q). With these precomputed counts, a front-end visualization permits interactive tuning of interest metrics. To determine mutations at nucleotide positions, we compare the collected data against a reference genome sequence. As our overall goal is to find co-occurrences of mutations, we de-emphasize the common case of reference-to-reference correlation, as this indicates the lack of any sort of epistatic functonality.

6. Visualization Design

Here we will describe our experience designing a visual analytics solution for the given problem, first presenting our early prototype (§6.1) using a matrix-based solution. The failure of this initial prototype prompted us to derive task **T3** (supporting overview), and requirements **R2** and **R3** (wayfinding and tunable filter parameters). We describe the rationale for the designs, and some of our lessons learned (§6.2) in incorporating implicit assumptions of the analyst into requirements for the final design (§6.3).

6.1. Initial prototype: Matrix-based visualization

For our initial prototype, we developed a matrix-based technique for looking at the correlations of mutations between pairs of positions (§4.2, see Figure 2). The design was inspired by previous work that use matrices to communicate relations, which excel at displaying large numbers of relationships in comparison to node-link diagrams.

Each cell in the matrix communicate the level of mutation co-occurrence ($M_{i,j\_}$) at a pair of positions i and j. We use a bi-variate color ramp [Tru81] to communicate the co-occurrence metric (a ColorBrewer red-to-blue diverging ramp [BHH03]) and the coverage (lightness attenuation in Lab color space), together identifying significant co-occurrence. Details are available through a linked "super-zoom" panel, which displays the metrics for a 3×3 area under the current mouse position. A bar chart (below) compares the mutations and reference reads at the two positions, and allows for conditioning on the nucleotide type.

We took advantage of the technological limitations of NGS, where direct correlations are limited to a window in the low hundreds of positions (the maximum read length). This produces a banded matrix about the diagonal, so we thereby limit navigation of the space to a one-dimensional diagonal pan and zoom to prevent getting lost in the data space. To overcome the technical limitation of loading millions of data points to the client and displaying them in a web-based interface, we used WebGL to load the data into buffers in the GPU and to render the matrix interface. Supplemental views such as the super-zoom were implemented using the D3 library [BOH11]. Using the GPU for rendering allowed for real-time navigation of a 20,000 $\times$ 20,000 cell-space, as well as interactive updates by modifying uniform variables sent to shaders (see [ME09] for a discussion).

6.2. Lessons learned from the matrix-based prototype

This early implementation had several problems in practice for exploring NGS data. The visualization was overwhelmed by many false positive results at nearly every pair of positions—many co-occurrence pairs had a saturated color (see Figure 2) but were not significant in practice. Through iteration on this design with stakeholders and a root-cause analysis, we found that although the co-occurrence metric was high in magnitude, the overall proportions of variant reads at those positions were very small (on the order of 1–5%), even though they had very high correlation to other positions. Many of these reads were determined to be misaligned reads by the sequencer. In additional, simultaneously visualizing a third metric (variation) requires a tri-variate color map, which are considered to be impractical [War09]. These constraints suggests an alternative method to filter out task-irrelevant co-occurrences (**R3**).

In order to assist analysts in identifying pairs of positions with significant co-occurrences, we added in a filtering gate to remove co-occurrence pairs where at least one position meets a minimum variant probability. This filtering made the data too sparse in the matrix to identify interesting co-occurrences, suggesting task **T3**: providing overview.

While matrix re-ordering could help to emphasize correlation between positions, reordering the genomic sequence prevents analysts from leveraging their knowledge of particular sections of the genome, such as critical gene-coding regions. A requirement (**R2**) that emerged from discussion with collaborators was to provide a mechanism that exposed *annotations*, or interval identifers of the genome that provide a wayfinding mechanism. They stressed that annotations can provide information on reading frames or identifying regions of interest. The overall difficulty of discovering interesting co-occurrences within the matrix view suggests a guided, interactive approach that does not embed relationships within the full data space.

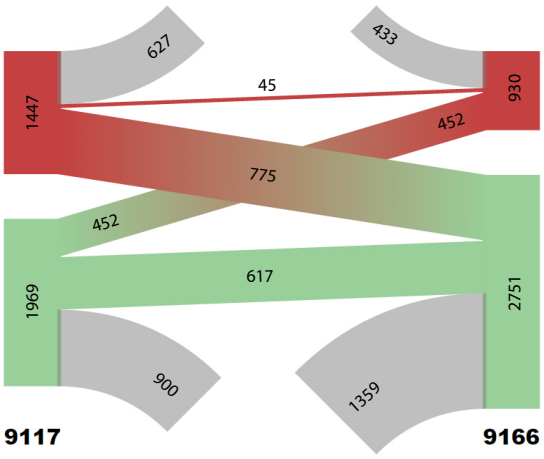

Figure 3: *A close-up of a co-occurrence summary between two positions (counts included for explanation). The positions being compared are mapped to rectangles, with both reference (green) and variant (red) nucleotide types. The links show the correlated proportion of reads between the two positions. The gray arcs represent the proportion of reads that overlap one position but not the other.*

6.3. CooccurViewer visualization

Our experience with the first prototype lead to a second design with revised tasks and requirements to support it (§3). Based on feedback from analysts and brainstorming potential solutions within our team and other virologists, we elected to modify our strategy to be driven by analyst focus (see Shneiderman and Plaisant for a discussion [SP15]). To achieve this, we integrated our three tasks directly into the design (see Figure 4): a one-dimensional map that forms the overview and designates positions where interesting co-occurrence is happening (**T3**), metrics with which to filter the space of correlations (**T1**), and a detail view that describes the correlation between pairs of positions with explicit metrics (**T2**).

6.3.1. Overview

To support user-driven exploration of significant co-occurrence of mutations, we brought filtering to the forefront of the analysis. The virologist has the option to tune parameters of significance (§5), and only those correlations that meet the analyst-defined standards are displayed. The overview of these significant co-occurrences appears at the top of the visualization. Each position displayed has at least one significant co-occurrence with another position. These single positions are connected to their positions on the genomic sequence by gray wedges and are clustered together based on their proximity in genomic space. The overview can support up to 500 positions, but becomes more comfortable with less than 75 individual positions. Virologists using our tool to explore co-occurrences tended to tune the metrics until about 50 positions were visible in the overview.

The CoocurViewer overview includes a linear representation of the genome with annotation data. These annotations are represented by the colored bars above the genome axis (Figure 4(a)), and provide virologists with biological context for positions in the genome. In particular, annotations marked as reading frames are used to determine if mutations within the region are synonymous mutations. Viewers are given the option of supressing synonymous mutations, which treat those mutations as matching the reference genome. In practice, we found that virologists would activate this option to remove synonymous mutations from display, but would also occassionally deactivate the option to identify mutations that could still have conformational implications.

Each position with significant co-occurrence is summarized by the three metrics introduced in Section 5, each color-encoded using separate ramps: coverage (i.e., *read depth*, in green), the base amount of variation at that position (in red), and the magnitude of the co-occurrence metric (in purple). In order to summarize correlations between multiple potential positions, the glyph at each position shows the maximum value of each metric independently. Sliders linked to these metrics (Figure 4(f)) allow the analyst to modify thresholds to filter out less interesting co-occurrences.

6.3.2. Co-occurrence Details

Once the virologist has selected a particular position of interest, the main view populates matching co-occurrences with that position. Through collaborative design, we developed a design to show "flow" between nucleotide types at two positions, similar to a version of parallel sets [BKH05] (see Figure 3). The connecting arcs show the proportion of reads (observations) that are one type at position i and are either the same or opposite type at position j. The gray arcs represent observations that exist at that position, but do not overlap the paired position. Tooltips can present more details on demand such as the number and proportion of nucleotide observations, including whether a particular nucleotide is potentially synonymous. For reasons of screen-space, only two pairs of co-occurrence detail can be shown at once, though all correlations for the current position are shown in a small-multiple display (see Figure 4(e)) and can be brought into full view by selection or through pagination.

6.4. Implementation

CooccurViewer is a system implemented in JavaScript, using the D3 library [BOH11] to map data to shapes on an SVG canvas. We use a pre-processing step to gather the 4×4 contingency tables (nucleotides at each position pair) from SAM files (short sequence read alignments) [L*09] by comparing reads to a given reference sequence and counting paired combinations of bases for each pair of positions. We also compute the co-occurrence metric from these counts (see §5) and compile other data such as annotations. These data are packed into the binary files that are served to the visualization. This allows for minimal transport over the network, and the client-side nature of the visualization entails near-interactive rates for filtering the data shown to the viewer. CooccurViewer and the pre-processing library are open-sourced on GitHub and available at `http://graphics.cs.wisc.edu/Vis/CooccurViewer/`.

7. Case Studies

We present two case studies to demonstrate the utility of our visualization prototype. Through these examples, we illustrate how

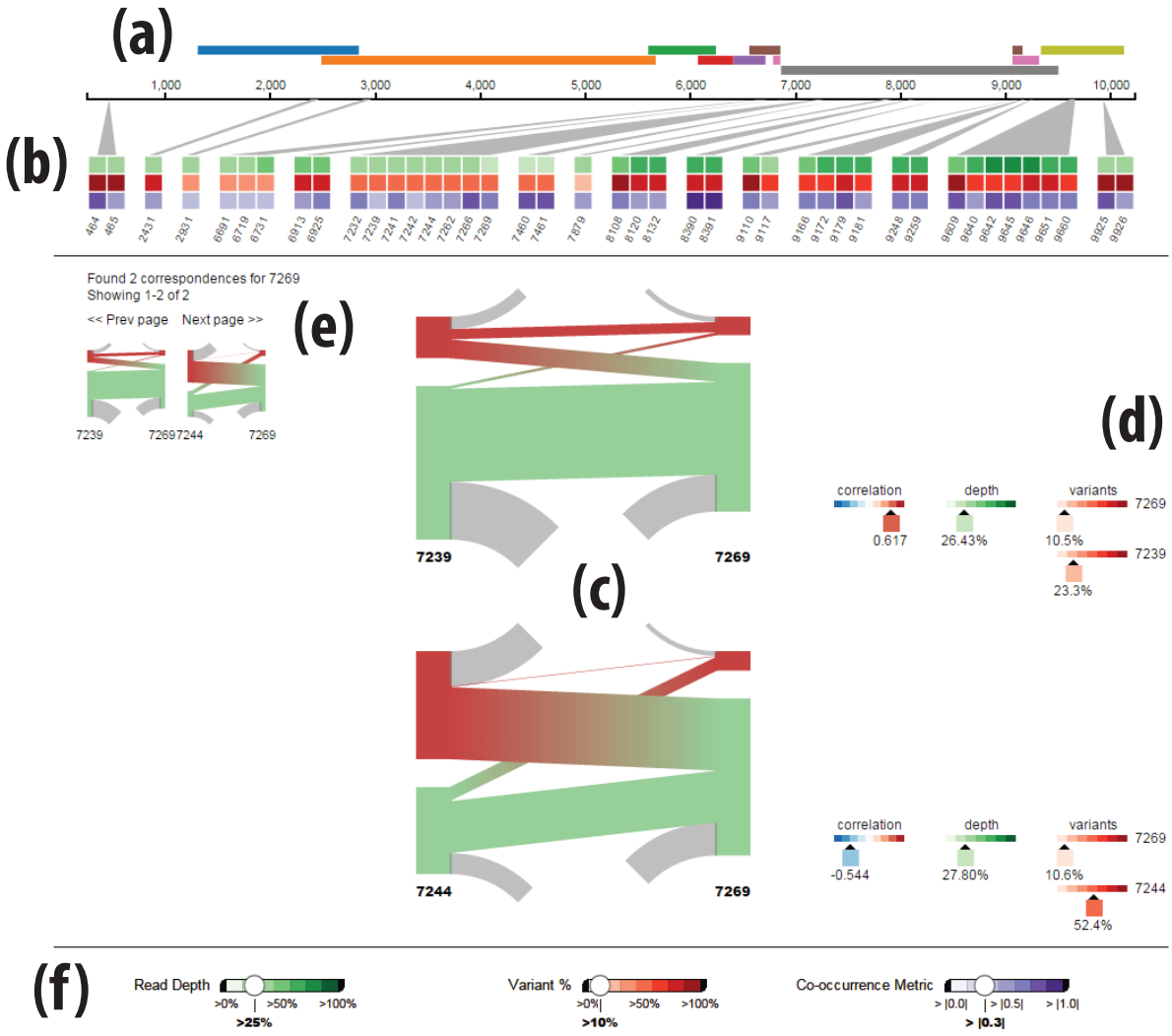

Figure 4: *An overall view of SIV (§7.2) loaded into* CooccurViewer. *Annotations (a) denote regions of the genome that have some biological context, and the overview (b) denotes positions of significant co-occurrence, summarizing the three metrics (§5) using color. The correlation diagrams (c) provide a representation of correlation between pairs of positions, and some details (d) about metric values. The current position's summary of correlations (e) is given on the left, with small-multiple representations. The sliders (f) control the thresholds for the interest metrics and filters the co-occurrences shown in the visualization.*

the visualization can expose significant correlation information. We show how the system is robust to displaying populations of viral genome samples in datasets of millions of pairwise correlations. We also highlight how our visualization design can help reveal new questions and insights about existing datasets. These studies are from two different virology labs, and include virologists beyond the authors of this paper.

7.1. Avian Influenza (H5N1)

In our first case study, different variants of the H5N1 influenza virus are explored. To understand the impact of within-host viral genetic diversity on replication and transmission of avian influenza viruses, Wilker and Dinis, *et al.* [WD*13] used deep sequencing to assess genetic variation from inoculated ferrets and ferrets infected via respiratory droplet transmission [I*12]. The authors reported that sub-populations present at low frequencies ($\sim 6\%$) could transmit via respiratory droplets. Interestingly, they showed that only one to two combinations of co-occurring mutations in the hemagglutinin (HA) gene were detectable early after infection in contact animals. Taken together, this shows that selective forces imposed a significant reduction in influenza genetic diversity during transmission.

We imported reference-based assemblies of the HA gene (1788 base pairs in length) from infected ferrets (six pairs, six samples each) into our pre-processing library. On average, each reference-based assembly contained 140k to 348k sequences (avg. 205k) and individual reads were 100 to 160 base pairs in length (avg. 149).

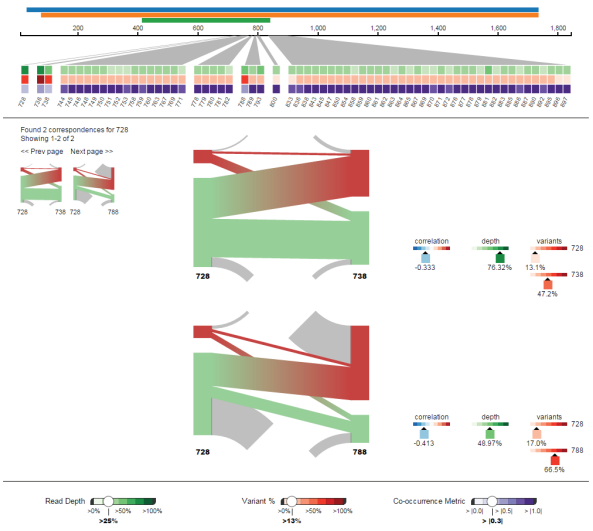

Figure 5: *For this particular sample of an H5N1 viral population, a strong inverse correlation is identified between mutations at 738 to non-variant reads at 728, as well as a inverse correlation between positions 728 and 788, validating the results presented by the reference study [WD*13].*

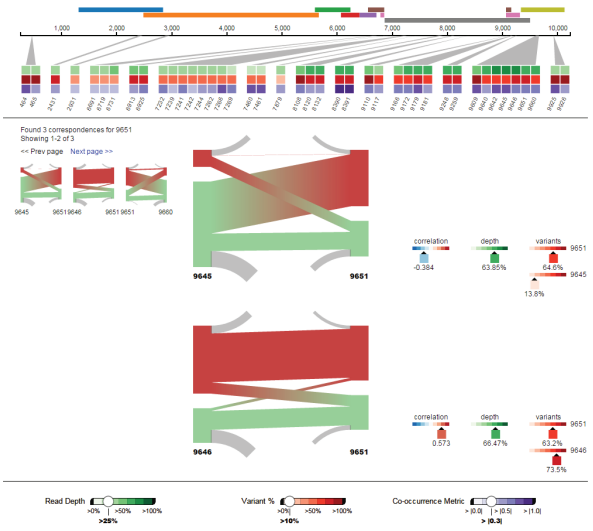

Figure 6: *In this SIV sample, a cluster of correlated mutations appears within the Nef protein (top-right, dark yellow bar), known to harbor viral escape. Variants at positions 9,645 and 9,651 are inversely co-occurring with reference reads (mid-top), while reads at positions 9,646 and 9,651 are positively correlated (mid-bottom).*

Annotations denote regions in the sequence that code for the pre-processed HA protein (blue), a post-processed HA protein that becomes packaged in the viral envelope (orange), and a region on the HA protein that binds to host-cell receptors (green). A single sample's packaged data averages around 42MB.

There is a significant level of nucleotide variation near the receptor-binding domain of H5N1 viruses infecting ferrets. In Figure 5, using sequence data from a directly inoculated ferret sampled five days post-infection, there are a number of significant co-occurrences. Virologists focuseed on two particular positions with relatively higher amounts of nucleotide variability, where the summary glyphs are saturated red. Selecting position 728 (the farthest left summary), a strong inverted correlation is found between non-variant nucleotides at 728 and variants nucleotides at 788— this relationship was identified in the original study.

Through the use of the visualization, potentially interesting co-occurrences were readily identified. This is in contrast to the intensive, manual workflow used to identify co-occurrences in the original work [WD*13], which involved concatenation of all polymorphic sites and tabular exploration through these varying sites to find potential correlation (taking several weeks). The visualization, by contrast, specifically targets the analytical task of rapidly identifying these interesting co-occurrences in the timescale of minutes.

7.2. Simian Immunodeficiency Virus (SIV)

SIV is a commonly-studied virus as an analog to HIV (human immunodeficiency virus). Variants accumulate during an HIV or SIV infection confer resistance to antiretroviral drug treatment or expand the range of cells the virus can productively infect. Understanding epistatic interactions are critical to target antiretroviral treatments. The dataset shown in Figure 6 comes from a macaque monkey 54 weeks after infection with a clonal, pathogenic strain of SIV [O*12]. In this case, we know the exact sequence and composition of the viral sequence (9,973 base pairs) that initiated the infection. The data contains 238k read segments, where each segment length is between 24 to 151 base pairs long (avg. 92). The 2.78 million pairwise count data and associated metadata is minimized to 170MB.

Virologists immediately saw from the summary (see Figure 6) that there is a high amount of variation in this particular SIV sample. Many of these significant correlations are inverse correlations, identified by a strong absolute co-occurrence metric (purple). In particular, virologists observed correlations in this dataset that may merit additional follow-up. First, there are comparatively few correlated variants in the structural proteins of Gag and Pol (the blue and orange regions stretching from positions 1309 to 5666). These are HIV/SIV genes thought to be under the greatest constraints; variation in these genes likely compromises the ability of the virus to replicate. The lack of correlated variants in these genes compared to the accessory and regulatory genes suggests that compensatory variants here are relatively infrequent. Second, they identified a cluster of correlated variants from nucleotides 9,609 to 9,660 that occurs within a known viral sub-population that is recognized by macaque CD8+ T cells. While it is known that the virus can evade detection by immune responses through mutations in this region, the virologists noted that examining the impact of correlated variants within this epitope may resolve sub-structure to the escape variant populations that would be missed with other analytic tools. The ability to foster these global insights demonstrates a remarkable improvement over virologists' previous manual workflows.

8. Discussion

Through this design study, we have learned several key lessons that generalize from our domain problem. Respecting the analysts' mental model of the anaylsis space and providing scaffolds for wayfinding proved to be critical in our design. We use a conjunction of multiple interest metrics to help narrow exploration in the large pairwise space of all pairwise correlations. We have also shown that combining multiple metrics through conjunction can help focus analysis when a single metric is insufficient.

In order to support analytical targeting for our design, we captured discrete components of significant correlations and generated definitions of these components. We quickly discovered that there was no one metric that captured if a co-occurrence between positions was significant or not, and elected to provide a mechanism to allow the analyst to select relevant thresholds dynamically. This interactive exploration affords analysis that can adapt to different analysts and datasets.

In this work, we focused on the problem of discovering co-occurrences of events within one sample of a population of viral genomes, and have shown it to scale to a significant amount of data (e.g. a viral genome 12k positions long with 250k reads leading to a ceiling of nearly 3 million potential co-occurrences). Extending our work to the problem of multi-sample comparison is important future work, though an independent problem. As an example, longitudinal studies of virus mutation usually span multiple timepoints, sometimes under different environmental or transmission conditions. While comparisons can be made implicitly between viral populations by switching the dataset shown in the visualization from one genome to another, it can be difficult to make explicit comparisons of correlation across samples.

The largest dataset we have supported thus far is the SIV dataset, which encodes 2.78 million 4×4 contingency matrices of pairwise correlations into our web-based visualization. We can scale to support the additional data of multi-level correlation (beyond pairwise correlation) and comparison across multiple timepoints by loading data directly to the GPU or offloading computation to a remote server [MHH15]. Applying data management principles such as indexing within the data (such as the *imMens* system [LJH13]) could also increase data retrieval rates.

Finally, we have determined that our viral population dynamics problem is an instance of the abstract problem of understanding partially observed co-occurrences. This abstraction permits us to convey statistics and trends of co-occurrence events in a visual manner. The abstraction also allows us to generalize our work to other domains such as large-scale text analysis and time-series data, although our development of such applications is still in progress.

9. Conclusion

In this work, we have presented a design study for the rapid identification of correlated mutations in populations of a viral genome. Through our characterization of the problem, we have identified requirements that led to metrics used to focus analysis on significant co-occurrences. We have shared our experiences in creating visualization prototypes to support our model task, demonstrated the effectiveness of our prototype design through our case studies, and summarized the lessons we have learned through this work. We hope to extend this work to higher-level correlations, and apply the lessons we have learned through this design study to other sequence-based data domains.

Acknowledgements

We acknowledge the thoughtful reviews, as well as the invaluable conversations with Thomas Fredrich, Shelby O'Connor, Louise Moncla, and Danielle Szafir. We also thank Deidre Stuffer for copy-editing. This work was supported by NIH award 5R01AI077376-07 and NSF award IIS-1162037.

References

[BHH03] BREWER C. A., HATCHARD G. W., HARROWER M. A.: Colorbrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Society 30*, 1 (2003), 5–32. 5

[BKH05] BENDIX F., KOSARA R., HAUSER H.: Parallel sets: Visual analysis of categorical data. In *Proceedings - IEEE Symposium on Information Visualization, INFO VIS* (2005), no. 1, pp. 133–140. 4, 6

[BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D^3: Data-driven documents. *IEEE Trans. Vis. Comp. Graph 17*, 12 (2011), 2301–2309. 5, 6

[C*15] CORRELL M., ET AL.: LayerCake: a tool for the visual comparison of viral deep sequencing data. *Bioinformatics 31*, July (2015), btv407. 3

[CLC] CLC BIO: CLC Genomics Workbench. http://www.clcbio.com. Accessed: 2014-03-28. 2

[D*08] DING L., ET AL.: Somatic mutations affect key pathways in lung adenocarcinoma. *Nature 455*, October (2008), 1069–1075. 3

[D*12] DEES N. D., ET AL.: MuSiC: Identifying mutational significance in cancer genomes. *Genome Research 22*, 8 (2012), 1589–1598. 3

[D*13] DEMIRALP C., ET AL.: *invis*: Exploring high-dimensional RNA sequences from in vitro selection. In *Proc. IEEE Symp. Biological Data Visualization* (2013), pp. 1–8. 3

[DPS02] DÍAZ J., PETIT J., SERNA M.: A survey of graph layout problems. *ACM Computing Surveys 34*, 3 (2002), 313–356. 4

[DS13] DUNNE C., SHNEIDERMAN B.: Motif simplification: improving network visualization readability with fan, connector, and clique glyphs. In *Proc. Conf. Human Factors in Computing Systems* (2013), ACM Press, pp. 3247–3256. 4

[EFBF07] ENAULT F., FREMEZ R., BARANOWSKI E., FARAUT T.: Alvira: Comparative genomics of viral strains. *Bioinformatics 23*, 16 (2007), 2178–2179. 3

[F*10] FIUME M., ET AL.: Savant: Genome browser for high-throughput sequencing data. *Bioinformatics 26*, 16 (2010), 1938–1944. 3

[FNM13] FERSTAY J. A., NIELSEN C. B., MUNZNER T.: Variant view: Visualizing sequence variants in their gene context. *IEEE Trans. Vis. Comput. Graph. 19*, 12 (2013), 2546–2555. 3

[GFC04] GHONIEM M., FEKETE J.-D., CASTAGLIOLA P.: A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations. In *IEEE Symp. Information Visualization* (2004), IEEE, pp. 17–24. 4

[GS14] GOTZ D., STAVROPOULOS H.: DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 1783–1792. 3

[HF06] HENRY N., FEKETE J.-D.: MatrixExplorer: a dual-representation system to explore social networks. *IEEE Trans. Vis. Comput. Graph. 12*, 5 (2006), 677–684. 4

[I*12] IMAI M., ET AL.: Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature 486*, 7403 (2012), 420–428. 2, 7

[Ins97] INSELBERG A.: Multidimensional detective. *Proc. VIZ '97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium* (1997), 1–8. 4

[JME10] JAVED W., MCDONNEL B., ELMQVIST N.: Graphical perception of multiple time series. *IEEE Trans. Vis. Comput. Graph. 16*, 6 (2010), 927–34. 3

[K*02] KENT W. J., ET AL.: The Human Genome Browser at UCSC. *Genome Research* (2002), 996–1006. 3

[K*12] KEARSE M., ET AL.: Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics 28*, 12 (2012), 1647–1649. 2

[L*09] LI H., ET AL.: The Sequence Alignment/Map format and SAMtools. *Bioinformatics 25*, 16 (2009), 2078–2079. 6

[LJH13] LIU Z., JIANG B., HEER J.: imMens : Real-time Visual Querying of Big Data. *Computer Graphics Forum 32*, 3pt4 (2013), 421–430. 9

[M*14] MAGUIRE E., ET AL.: Redesigning the Sequence Logo with Glyph-based Approaches to Aid Interpretation. In *EuroVis - Short Papers* (2014), Elmqvist N., Hlawitschka M., Kennedy J., (Eds.), no. June 2014, The Eurographics Association. 3

[ME09] MCDONNEL B., ELMQVIST N.: Towards utilizing GPUs in information visualization: a model and implementation of image-space operations. *IEEE Trans. Vis. Comput. Graph. 15*, 6 (2009), 1105–12. 5

[MHH15] MORITZ D., HEER J., HOWE B.: Dynamic Client-Server Optimization for Scalable Interactive Visualization on the Web. In *Workshop on Data Systems for Interactive Analysis (DSIA '15)* (2015). 9

[N*10] NIELSEN C. B., ET AL.: Visualizing genomes: techniques and challenges. *Nature Methods 7*, 3 (2010), S5–S15. 3

[O*12] O'CONNOR S., ET AL.: Conditional CD8+ T cell escape during acute simian immunodeficiency virus infection. *Journal of Virology 86*, 1 (2012), 605–609. 8

[R*11] ROBINSON J. T., ET AL.: Integrative genomics viewer. *Nature Biotechnology 29*, 1 (2011), 24–26. 2

[S*02] STEUER R., ET AL.: The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics (Oxford, England) 18 Suppl 2* (2002), S231–S240. 3, 4

[S*10] SANJUÁN R., ET AL.: Viral mutation rates. *Journal of Virology 84*, 19 (2010), 9733–9748. 1

[SAB*16] STROBELT H., ALSALLAKH B., BOTROS J., PETERSON B., BOROWSKY M., PFISTER H., LEX A.: Vials: Visualizing Alternative Splicing of Genes. *IEEE Trans. Vis. Comput. Graph. 22*, 1 (2016), 399–408. 3

[SMM12] SEDLMAIR M., MEYER M., MUNZNER T.: Design study methodology: Reflections from the trenches and the stacks. *IEEE Trans. Vis. Comput. Graph. 18*, 12 (2012), 2431–2440. 1

[SP15] SHNEIDERMAN B., PLAISANT C.: Sharpening Analytic Focus to Cope with Big Data Volume and Variety. *IEEE Computer Graphics and Applications 35*, 3 (May 2015), 10–14. 6

[Tru81] TRUMBO B. E.: A theory for coloring bivariate statistical maps. *The American Statistician 35*, 4 (1981), 220–226. 5

[V*13] VAN BRAKEL R. B. J., ET AL.: COMBat: Visualizing co-occurrence of annotation terms. In *Proc. IEEE Symp. Biological Data Visualization* (2013), IEEE, pp. 17–24. 3

[War09] WARE C.: Quantitative texton sequences for legible bivariate maps. *IEEE Trans. Vis. Comput. Graph. 15*, 6 (2009), 1523–30. 5

[WD*13] WILKER P. R., DINIS J. M., ET AL.: Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. *Nature Communications 4* (2013), 2636. 7, 8