Visualizing Virus Population Variability From Next Generation Sequencing Data

Michael Correll* Subhadip Ghosh† David O'Connor‡ Michael Gleicher§

University of Wisconsin, Madison

Abstract

Advances in genomic sequencing techniques allow for larger scale generation and usage of sequence data. While these techniques afford new types of analysis, they also generate new concerns with regards to data quality and data scale. We present a tool designed to assist in the exploration of the genetic variability of the population of viruses at multiple time points and in multiple individuals, a task that necessitates considering large amounts of sequence data and the quality issues inherent in obtaining such data in a practical manner. Our design affords the examination of the amount of variability and mutation at each position in the genome for many populations of viruses. Our design contains novel visualization techniques that support this specific class of analysis while addressing the issues of data aggregation, confidence visualization, and interaction support that arise when making use of large amounts of sequence data with variable uncertainty. These techniques generalize to a wide class of visualization problems where confidence is not known a priori, and aggregation in multiple directions is necessary.

Index Terms: J.3 [Computer Applications]: Life and Medical Sciences—Biology and Genetics;

1 Introduction

New data acquisition techniques that provide large amounts of data provide both opportunities and challenges. While such datasets can enable new kinds of questions to be explored, they also mean that larger-scale analysis and pattern-finding must be sought. Larger data sizes also often bring issues with uncertainty and uneven data quality: large data sets are often merged from multiple sources, acquired over long time periods, and/or use acquisition modalities where quality vs. quantity tradeoffs must be made. To make use of this newly available data, scientists will need tools that explicitly address three challenges: novel questions, large-scale data, and variable data quality. In this paper we consider a specific application that provides a case study where these three challenges must be addressed.

We consider tools for comparing the genetic variability in populations of viruses. Such study demands large amounts of sequence data, sufficient to capture the variability in the populations of viruses collected from multiple samples. Collecting such data has been made practical with the advent of "next generation" high-throughput sequencing technology. While this technology makes large scale data collection practical, it also introduces several sources of uncertainty and variability, meaning that data quality must be considered as part of any analysis. A visualization tool must enable a scientist to find interesting patterns of variations in the genetic populations, across multiple populations to find evidence of phenomena such as a viral populations changing over the course of an infection in response to immune system features and responses. The challenge comes not only from the fact that such patterns comprise small features across large datasets, but also that the data has variable quality such that the confidence in the significance of any finding must be considered.

Our result is a novel tool specifically designed to assist scientists with using next-generation sequence data to study genetic variability in viral populations. We use a color field design (like a heat map) with salience enhancing aggregration and context-preserving zooming that allows for patterns of interest to be rapidly identified and studied. We provide a similar view of the various data confidence factors, allowing the viewer to create a "fog" over the data color field, reducing the salience of less certain data. This combined scheme allows interesting data events to be made salient, but only when viewer-defined confidence criteria are met. Our design uses a number of interactive tools to explore the data set, expose relationships between data value and quality, and highlight features of potential interest. It allows for examination of data confidence factors, including their impact on the results of interest. The design allows for exploring details on demand: once overall patterns are identified specific data, can be examined.

1.1 Contributions

We present a visualization tool that solves a specific problem, comparison of viral population variability, for which no acceptable approaches exist. The tool itself is immediately useful for the domain science, affording new capabilities (at new scales) in an emerging area of research. Scientists can now visualize comparisons of genomic sequence variability data at greater scales, in greater detail, and with greater regard to data quality concerns. Significantly, the population sequence variability comparison has many traits common to many challenging data anaylysis tasks that require seeking patterns in data of variable quality. We believe our approach, and specific visualization methods, applies more generally to such tasks.

We present novel aggregation techniques to make the visual scanning of large sequences tractable while still allowing both general trends and specific outliers to be perceived. In addition, we present the metaphor of dynamic "confidence fog." Our confidence fog technique is a novel solution to the problem of visualizing confidence when there are multiple channels of potential uncertainty but no single a priori method to calculate overall data quality. By using this technique on a variety of data the user can quickly visualize systematic uncertainty as well as the impact of different sources of uncertainty on patterns in data. New interaction techniques to support these visual paradigms reduce the initial complexity involved in learning how to navigate the data.

2 Background

When a virus infects a host organism, it reproduces rapidly, creating many copies of itself. Therefore, at any instant in time there is an entire population of viruses in the host. For RNA viruses, this

*Department of Computer Sciences, e-mail:mcorrell@cs.wisc.edu

†Department of Computer Sciences, e-mail:sghosh@cs.wisc.edu

‡Department of Pathology and Laboratory Medicine, e-mail:david.h.oconnor@gmail.com

§Department of Computer Sciences, e-mail:gleicher@cs.wisc.edu

population is unlikely to be genetically identical: even if the individual is infected by a single virus, or a homogenous population of viruses, errors in replication (mutations) will cause variation. In some viruses, notably HIV/SIV, such variation due to mutation is significant: the viral population adapts to its host environment over the course of the infection. Understanding the variability in viral populations and how they change over time and between different individuals can offer important insight into the viruses and the diseases they cause. For example, observing changes in a viral population over time can suggest which regions of a viral genome are being targeted by the host's immune system.

However, measuring the genetic variability across a viral population necessitates large-scale measurement. A significant enough sample of the population must be sequenced. Until recently, the sequencing of genomes (including large scale projects such as the Human Genome Project) was done using some variation of the original sequential Sanger method [17]. The Sanger method generates "reads" of several hundred base pairs (approx. 800) and then assembles them sequentially. The Sanger method is slow and expensive. The next generation of genome sequencing techniques afford increased parallelization to generate reads that are shorter (less than approx. 400 base pairs per read) but can create greater "depth" of reads (more samples at a particular section of the genome) at less cost [12].

By using these new "deep sequencing" methods it is possible to generate larger amounts of genomics sequence data in shorter amounts of time than previously, opening the door for research efforts such as longitudinal studies of genotype in a population and (eventually) affordable sequencing at the level of the individual[15]. Indeed, entire human genomes can now be sequenced in a single instrument run. Leveraging this techniques to study viral sequences, these methods afford sampling of large heterogenous populations of related viral genomes; it is now possible to quantify genetic variation in viral populations more accurately.

To assess the variation within a viral population,viral nucleic acids (representing many copies of the virus) are taken from a blood plasma sample and deep sequenced. Each one of these reads is aligned with a reference genome of the virus. This produces a dataset with a count of the different read base pairs (A,C,T or G) that were observed at each position relative to the reference. In addition to matching a base pair at a position, the alignment can also include annotations that provide information about data quality. A read may have low-quality sequence at a particular site that cannot be reliably mapped against the reference sequence, and such positions are known as 'n' events. Alternately, a read may differ from the reference sequence by **d**eletion/**i**nsertion **p**olymorphisms (called "dip" events). Therefore, the resulting data set for a viral population (e.g. a specific host at a specific time) is a 6-bin histogram for each position in the reference genome. Comparing populations requires comparing these arrays of histograms. Initially, the histogram may be collapsed to a single number, the percentage of reads that differ from the reference.

2.1 Data Quality

While next generation sequencing is capable of producing more sequence data than Sanger sequencing, this data has a number of quality issues that complicate its use; each sequencing platform has characteristic quality issues that must be considered when analyzing data. Tools to analyze next-generation sequencing need to recognize these different error profiles. In particular, our results draw upon problems characteristic in Roche/454 pyrosequencing. These issues are not unique to next-generation DNA sequencing; they are common to many situations where data volume is growing rapidly, in many domains other than biology. For example, the use of large web repositories of text introduces new issues in vetting and parsing data that are not present in smaller hand-curated text corpora.

One issue with next generation sequencing is that each read tends to be shorter than it would be with other technologies, which can complicate alignment to large vertebrate genomes. Multiple algorithms have been developed to create these alignments in computationally tractable ways [21]. Viral genomes, particularly those of RNA viruses, are orders of magnitude smaller than those of vertebrate genomes. Therefore, the lengths of reads from pyrosequencing are usually sufficient to unambiguously align reads with viral reference genomes even when significant variation exists.

A related issue with current sample preparation methods for viral sequencing is that the reads are not drawn evenly from the length of the sequence [4]. Some parts of the sequence are more represented than others. Because the confidence in extrapolating from the viral population of the sample to the overall viral population is related to number of observations, this means that the confidences will vary over the length of the genome.

Other sources of uncertainty come from the technical limits of the sequencing technology. In addition to random variation, most (if not all) sequencing technologies have some systematic source of read errors. For example, Roche/454 pyrosequencing has issues where repeated base pairs may not be counted precisely. In such cases, the actual base pair may not be identifiable (leading to an "n" value described above). While such misread values clearly should not be viewed as significant genetic variation, they suggest that the particular location may be difficult to read, and suggest some general local uncertainty, as well as reads at neighboring locations (as they can create local alignment ambiguities). Similarly, a dip event also suggests a local alignment ambiguity, and can be considered a source of uncertainty.

2.2 Domain Task

Our goal is to use pyrosequencing data to understand the variation in virus populations, and to compare populations both between individuals and at different times in the infection. Our initial studies involved the simian immunodeficiency virus (SIV, a close relative of HIV). Using a reference genome and multiple animals at different stages of infection, virologists have used deep sequence data to inform research dealing with the genetic variability of SIV as it adapts to particular hosts with particular phenotypes and immune responses [3][4].

Small variations can be important. A single position change can serve to disguise a virus from the host's immune system. Combinations of changes are also significant: a variation in one position may require a compensating variation in another (potentially distant) position [3]. Scientists must sift through the sequence data to find locations and patterns of variation of potential interest.

While the genomes of viruses are relatively small (on the order of tens of thousands of base pairs), there are still too many base pairs to make manual search of the genome practical. Existing strategies for visualizing variation data involve performing filtered queries on a database and then using scatterplots to show the variation at specific sections of the reference genome, with a different data series for each particular sequence (see Fig. 6). Given that the data is on the scale of up to a dozen individuals at multiple time points (requiring glyphs and colors capable of distinguishing on the order of several dozen data series), and that there are on the order of ten thousand base pairs in a single SIV genome, this scatterplot solution leads to overplotting and difficulty separating data series for close analysis. While it would be possible to modify this scatterplot presentation to minimize these issues, we would need to dynamically query the data set and then refine our confidence in certain data, making a static scatterplot a less than ideal choice, even if the technique could scale up to our current level of tens of sequences of tens of thousands of base pairs. Additionally, this approach does not consider confidence in observations, nor does it facilitate looking for patterns and correlations variations.

From our initial survey of the common tasks associated with comparing viral population dynamics using short reads data, as well as the particular technical resources of our researchers, we developed an initial list of requirements for visualization tools in this particular domain:

1) Tools should scale to (at a minimum) on the order of dozens of sequences and tens of thousands of base pairs. This by necessity will require aggregation of base pairs if we are to present an overview of the data. Aggregation, in turn, requires that we provide methods for flagging aggregate items containing items of interest, and drilling down to examine these interesting items.

2) Tools should allow for the creation of dynamic uncertainty maps. The creation of these maps should align to the heuristics already in use by our collaborators (in particular rules of the form "ignore all data where X% or more of the reads are n, and be suspicious of data close to that percentage").

3 Related Work

The problem of visualizing genomic sequencing data is not new, and in fact has received more attention and more urgency as the availability and quantity of sequencing data increases [14]. While the visualization of population variation with sequence data can (and ought to) borrow techniques and visual paradigms from these existing tools, this traditional tools are not sufficient for the unique form of short reads data, where the depth and variation of reads, can be as important as the actual identity of the base pairs at a particular point in a sequence.

3.1 Short Reads Visualization

Previous visualizations of short reads data have mainly focused on the issue of generating alignments and consensus sequences from large numbers of short reads, and using the artifacts that arise during this process to explore and tag genome features. In particular, many (perhaps most) of the views seem to share what Schatz et al. call the "scaffold view" [18]. In this view the short reads are horizontally stacked, creating staggered "towers" of read coverage. In this class of short reads visualization, this scaffold view is the central visual artifact, and the default visual encoding [2][6][13]. Additional information is added by creating multiple visual tracks associated with the scaffold view, presenting aggregate statistical information, and performing semantic zooming to hide details.

While this scaffold view is a natural visual encoding for short reads data, and well suited for visualizing the alignment process and the consensus genome creation process, it is not ideal for the case where our objective is not to create a consensus genome or (necessarily) to discover sequence features, but rather to focus on the variations in reads and the quality of our information about variations. In particular none of the existing solutions merge confidence with data in a robust way. The scaffold solution, since it simply stacks reads atop one another, also does not scale to large numbers of reads. Since viral genomes are compact compared to the genomes of higher organisms, identical numbers of reads will generate much higher read depths than in viral case, requiring an approach that scales beyond the abilities of the scaffold view.

Another drawback of existing solutions (which tend to be tailored to the alignment problem or to the genome feature detection problem) is that they do not support comparison across multiple sequences, let alone multiple sequences at multiple time points, and certainly not at the scale of comparison required for the viral population dynamics problem.

3.2 Confidence Visualization

Tailoring sequence visualization to our particular task required exploration of visualization techniques for data quality and uncertainty. A useful metaphor is to consider uncertainty visualization as a method for merging an "uncertainty map" and a "data map" [11]. Within this paradigm of uncertainty visualization, there is a wide space of blending options, encoding data confidence in a color channel, using glyphs, or position, among other techniques [8]. While uncertainty about data can come from multiple sources, there also multiple modalities of uncertainty (for instance data corruption or error is a qualitatively different sort of uncertainty than, for instance, statistical lack of confidence due to potential sampling error) for which a single static uncertainty map is inappropriate[20].

Even within a single modality of uncertainty, there might be multiple variables associated with data quality. Previous visualization research has dealt with the case of multiple quality variables by presenting the quality metrics as extra data dimensions and then employing standard multivariate data visualizations techniques to assist in selecting particularly high or low quality data while maintaining global quality context [24].

Our use case does not fall neatly into the established problem of confidence visualization; we have multiple simultaneous channels of uncertainty (and thus must blend multiple uncertainty maps together) but no a priori confidence metric (and thus cannot generate uncertainty maps a priori). In addition, while the sources of data quality are important (and should not be hidden from the user, and in fact the user might want to explicitly explore the data quality variables), the actual data tend to be more important (and thus when we display the quality information it should not be given equal visual weight compared to the primary data channels). To add a last wrinkle, since one of the primary issues of sequence visualization is how to filter out irrelevant sequence areas, we do not want to give equal visual weight to uncertain data (in order to allow for the preattentive saliency of data that are both highly certain and semantically significant). Existing strategies for generating dynamic uncertainty annotations (where we overlay the uncertainty map rather than blend to it) [5] preserve data saliency but do not filter out uncertain data, nor do they provide a visual link between multiple uncertainty channels and resulting aggregate uncertainty.

4 Design

The particular nature of our domain task necessitated the development of several new visualization and interaction techniques, as well as the adaptation of existing techniques to fit our unique context. Our data has two primary dimensions: the population or sequence, and the position within the sequence, with a single value (the percentage of reads different from the reference) although drilling into the data will require additional dimensions such as variant type counts. While a multiple line graph may seem an obvious choice, it fails to scale to the datasizes that we consider. The large amount of aggregation required to compress the length of the sequence to the screen size would render small events (which are common and important) invisible, and overlaying dozens of dense line graphs would obscure important data through overplotting.

Our design choice was to consider our primary dimensions spatially, and use color to encode value. Such a heatmap-like representation of multiple series data has been called a Lasagna Plot [19]. Such a color field display of dense sequence data provides opportunities to make significant features and patterns pre-attentively salient [1]. We specifically choose to provide background stripes between rows to aid when the display is scanned sequentially for details.

4.1 Aggregation and Event Striping

While virus genomes are relatively compact, there are still on the order of tens of thousands of base pairs in one genome, far more than can be displayed at once. Even for small datasets, we have more data points than available pixels. Rather than force the user to pan through multiple screens of data, it was important to display a

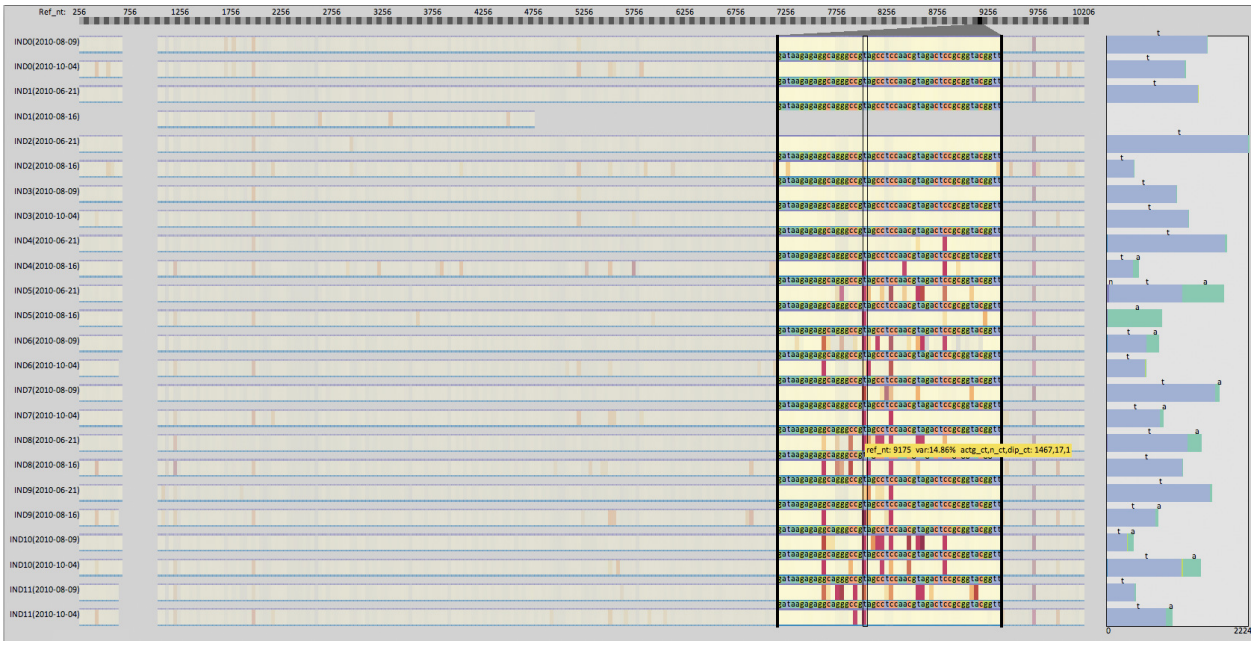

Figure 1: A view of LayerCake, showing 24 sequences, with each block of the color field aggregating 50 individual base pairs. Here the user has zoomed in on one such block, showing all reads for 50 base pairs simultaneously in a zoomed in area. The reference genome is repeated between each row for context. The user can further mouse over a particular column to see the identity of variants (in terms of base pairs) in each sequence in the histogram on the far right. Rows not within the zoom window are slightly "fogged" to decrease their salience during this drilling down task.

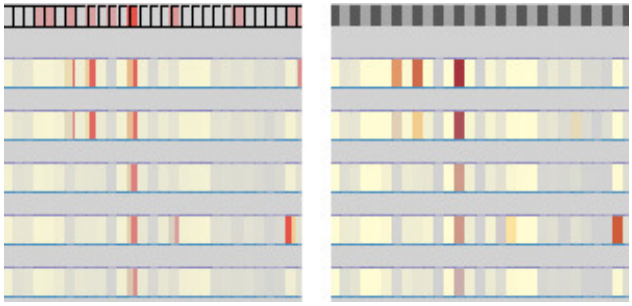

Figure 2: 'Event Striping' on a sample subset of virus mutation dynamics data on the left, compared to the normal view on the right. Red lines denote individual areas where high variation is present that would otherwise be lost in the aggregated data. The histogram allows quick focus onto columns of interest.

lower fidelity of the entire data simultaneously to aid in understanding trends and finding patterns. However, the tool must support filtering and drilling down to see the specific details.

Our aggregation approach uses uniformly sized binning rather than downsampling. While this reduces the ability to precisely localize phenomena, it provides a bounded degree of visual complexity and a visual reminder of the degree of aggregation. In fact, we are better able to provide positional fidelity from important events using "event striping" (below). The observable bins are also convenient for selection, detailed descriptions and zooming (§4.3).

The sequences are divided into bins, each containing an equal number of base pairs. This bin size is always described in terms of base pairs per bin, and is controlled dynamically with a slider, up to a maximum bin size is determined by screen resolution. The number of variants, number of n events, and the number of dip events are all aggregated together and then normalized by the total number

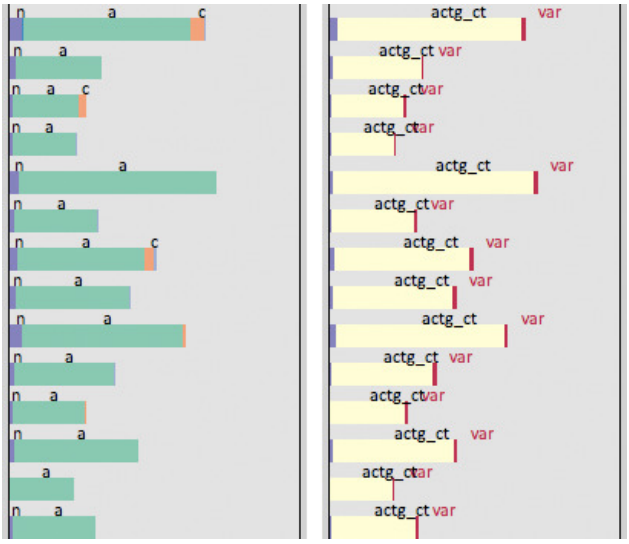

Figure 3: A view of the sidegraph for a sample subset of sequences. When users mouse over a particular column of aggregated blocks, the aggregate depth of reads is presented in a side histogram, along with percent variants. On the left, when the user has zoomed into a particular block, read depth and the precise nature of the variants (in terms of nucleic bases) is shown instead.

of aggregate reads in a bin.

Since most particular base pairs have very few variants in a particular dataset, this aggregation has the effect of dampening the variant percent. While individual locations on a particular sequence might be entirely variants when compared to the reference genome, when aggregated over tens of base pairs, these variant counts tend

to be very small (on the order of 1% of all reads in the bin). We give users the option of visually highlighting these high variant positions in low variant bins by including "event striping," where we create constant width red stripes on values above a certain event threshold which would otherwise be lost in the general trend of the block as a whole [1]. Stripes encode the position of significant events in the data while not breaking the visual metaphor of the aggregated heatmap (see Fig. 2). These bright stripes are highly salient, allowing users to quickly flag bins containing interesting outliers. Event striping highlights areas with high amounts of variance that would otherwise be lost in aggregation. By using event striping, locations in the data where interesting events occur are made visually salient. The user can control this "interest level" dynamically, allowing control over visual complexity. Since there may be many events occurring in particular blocks, and there may be too many sequences to make the selection of particular event-heavy columns feasible without sequential search, when event striping is active we encode the number of events per column (i.e., across all sequences for a particular range of base pairs relative to the reference genome) as a simple color plot overlaid over the horizontal legend. By focusing on areas where the legend is red, the user can quickly find columns where there are many sections with high variability.

4.2 Confidence Fog

Central to the task of analyzing viral population dynamics with pyrosequencing data is the capability to understand the effects of uncertainty on our conclusions from the data. In particular we wish to know what sections of the data can be ignored as too uncertain to provide real insight, where there are systemic issues in collecting data, and the extent to which high values are correlated with low quality data. Complicating the problem is that there is no single uncertainty metric. In fact, to assist in investigating data quality concerns, users may wish to explore a wide space of possible uncertainty metrics. Since we are relying on visual salience to encode informational salience, we also want to reduce the visual salience of items where our confidence is low. This requires a specialized interaction and visualization paradigms.

We have two channels of uncertainty: n events and dip events. In addition, if the read depth at a particular location is low, then even small numbers of n events or dip events are enough to make the resulting variance data questionable. To visualize these uncertainty sources we draw thin "runners" on the top and bottom of every bin (Fig. 4). The top runner is saturated inversely with the percentage of our reads that are n events (so the higher the percentage of n events, the more desaturated). The bottom runner is saturated inversely with percentage of reads that contain dip events. These runners are small enough that they do not interfere with the pre-attentive selection of salient bins, but large enough to be legible once particular bins are investigated. As the total depth of reads in a bin decreases, our runners will desaturate faster. In addition, the user can specific n and dip thresholds that set minimum confidence percents for each uncertainty source, conforming to the existing heuristics where all data with more than X% count of n or dip events are filtered out. Once we have a desaturation value for each uncertainty source, these amounts are multiplied together to get an aggregate confidence for a bin, which we can use to visualize data quality factors dynamically.

Encoding both data value and confidence using color requires a bivariate color map. Creating effective bivariate maps is a known hard problem [22][16], especially if the viewers must be able to distinguish the exact value of both variables [23]. We partially sidestep this problem because instead of generating a "square" color palette where we can perceptually distinguish between every possible combination of x and y for two data variables, we instead wish to produce a "wedge" color palette. That is, as our certainty decreases, we care less about being able to distinguish between different data values. By using a standard ColorBrewer [9] ramp that is monotone in hue, we can represent changes in data by selecting a point on a ramp, and select a confidence by blending to a single "uncertainty color." We introduce the metaphor of "confidence fog:" as data are gauged to become more uncertain, we perform blending in both saturation and brightness until we blend into a background "fog." Fig. 4 illustrates how our data get increasingly foggy as we adjust our confidence metric. We considered incorporating a blurring effect to further strengthen this visual metaphor, however the few sharp features in most regions are generally important, reducing the contrast from bin to bin would make the resulting visualization less clear, and the blur effect might introduce unwanted visual artifacts (such as "bleed over" into surrounding bins). In the resulting visualization there is thus a 1-to-1 mapping from visual salience (bright, highly saturate, red in our final color scheme) to data salience (bins with high numbers of variants, where we are maximally confident).

The four parameters of the color map, minimum and maximum thresholds for value and confidence, can be controlled using an interactive legend (see Fig. 5). By dynamically altering our n or dip confidence thresholds we can even use animation principles to observe the effect of our confidence fog on the data, revealing data quality correlations across sequences that may not be readily apparent from a static confidence picture.

4.3 Focus+Context Zooming

While aggregation is necessary to allow for quick summarization of data, events in a genome still occur at the level of base pairs. In order to analyze the meaning of population dynamics data in genomes, it is necessary to seek specificity in our data, down to the level of base pairs. In addition, to make meaningful use of event striping (see § 4.1) users need to be able to zoom in to see what sort of events have occurred. As a way to support this zooming while maintaining positional data, we allow users to expand specific blocks to provide a per-base pair view of each sequence (see Fig. 1). The portions of a sequence that are not zoomed in are shrunk to make room for the additional space of the zoomed area, and then fogged to provide additional salience to the zoomed in area. We draw the reference genome between each sequence to provide per column context. At all levels of zoom mouseover tooltips provide the exact bin contents in terms of n, dip, and base pair counts.

In order to better enable comparison to understand the differences in a particular column, we provide an optional side graph (see Fig. 3). When enabled, this graph provides details about the column under the mouse, giving a histogram of the different values for each population. The data coloring in our lower levels of zoom gives us information about the location of variants, whereas our highest level of zoom allows us to investigate what these variants actually are, how they change over time, and what effect (if any) they have on gene expression compared to the original reference genome.

4.4 Interaction Techniques

The use of sliders allow users to set the variant threshold (what percentage of our reads must differ from the reference genome before we color the bin maximally red), n threshold (what percentage of our reads must contain n events before we are maximally uncertain), and dip threshold (what percentage of our reads must contain dip events before we are maximally uncertain). This method allows for accuracy, but for novice users it is difficult to map slider functions to the resulting changes in the visualization. To provide visual context for our current data settings, and also to allow for a more intuitive interaction paradigm, we augment our sliders with a color wedge (see Fig. 5). The wedge is both a legend (in that it shows the mapping from data and confidence to color) and an interaction tool (as the user can alter the mappings on the fly).

The combinations of sliders (which allow for precision) and the color wedge (which is visually descriptive) allow fluent interaction

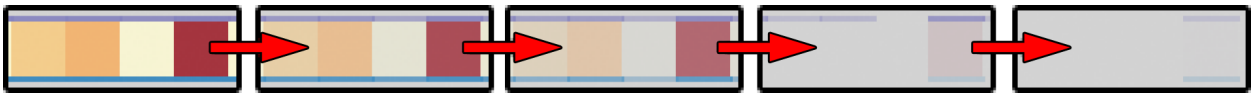

Figure 4: 'Confidence Fog' on a sample subset of virus mutation dynamics data. The top purple runner fades as the number of n events increases, the bottom blue runner fades as the number of dip events increase. As we increase our confidence threshold, this fading happens faster, leading to the visual metaphor of uncertain data receding into the 'fog' of the background. In "Confidence Mode" the data disappear entirely, and the entire block width is devoted to the top and bottom runners.

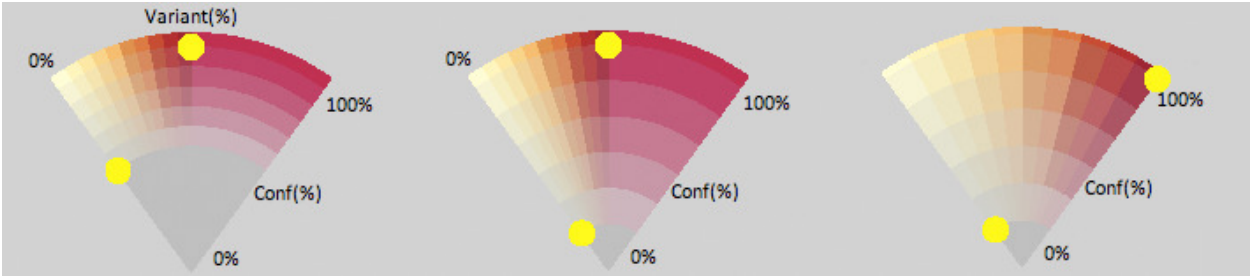

Figure 5: The color wedge tool for modifying confidence and data thresholds. The user can control the four parameters of the color map, the minimum and maximum thresholds for the data and confidence, by moving the yellow dots. In the leftmost wedge the user has set a high confidence threshold (the arc controlled by the lower yellow point), so the aggregate confidence of data must be at least 50% before we begin to "de-fog." The right wedge has a very high data threshold, so data values must be very high before we assign them the maximum (red) color.

with the visual parameters of the tool. The dynamic exploration of confidence and importance reveals interactions between sequences and provides instant feedback on the correlations between data quality factors and data position and value. Figure 7 shows a visual example of the full GUI, with sliders and color wedge visible.

5 Results

We have realized our design as a prototype, called LayerCake because of its striped appearance. LayerCake is built in the Processing[7] library and prototyping environment for Java (with the GUI for Processing library[10]) that allows for cross platform deployment and eventual integration with the existing online data querying systems. This tool is capable of parsing and interactively displaying dozens of genomes with tens of thousands of base pairs, even while running in a browser.

Our initial experiments with the LayerCake tool have been on datasets from SIV sequencing experiments. This virus has a length of approximately 10,000 base pairs. The data sets we have considered we compare up to twelve individuals sampled at two time points (so up to 24 total sequences). Even the initial prototypes proved useful for virologists, who quickly embraced the visual paradigm. Feedback was positive; constructive criticism consisted mainly of feature requests, such as better integration within their workflow, better methods for (vertical) sorting, and making use of additional metadata.

The deployment and use of the tool has already led to gains in the ability to understand viral population dynamics when compared to existing techniques. In particular, the confidence visualization techniques have provided an easy method for detecting and discarding "false positives" (i.e. areas of a viral genome that seem to vary greatly over the course of a population but where these high variant counts are merely artifacts of data quality problems). Manual scanning of the raw data and visualizations that do not consider confidence require additional analytic passes or arbitrary filtering to reduce the dataset. Our approach keeps the entire dataset in context while making confident data salient.

Another immediate use case for our tools is to make formerly intractable comparison problems easier. One category of hypothesis that draws on viral population dynamics data is the assumption that certain elements in the genotype of the infected individual will encourage particular virus variability as an infection proceeds. Twelve individuals were divided into four genotypic groups, and the SIV infection was sequenced at two time points. Without access to LayerCake, verifying hypotheses of this type required painstaking and non-scaleable analysis (see Fig. 6), with multiple sessions of query building, analysis of raw data, and no clear visual paradigms to compare results across many individuals. With LayerCake, (see Fig. 7) users can visualize all three groups of individuals at both time points without the detrimental effects of overplotting. Visual saliency and our side histograms of per-column information allow users to quickly notice patterns between groups and between time points. In particular it is possible to see that as time goes on, the deviation from the reference genome becomes more pronounced (as one would expect as the viral population changes as a result of the idiosyncratic immune response of an individual). By scanning the different genomic groups, it is possible to see that certain groups have changes in variation in different sections of the genomes, at scales where previous displays would impede this sort of analysis.

6 Conclusion

Working in concert, our design elements afford quick perusal of large data sets with the ability for users to rapidly find, and subsequently explore, areas of interest. Even though we do not have a single a priori metric to gauge data quality, by exploring parameter space it is possible to use data quality concerns to guide exploration and ground conclusions. The visualization and interaction paradigm employed by our tool can be naturally extended to other fields where data confidence is dependent on multiple quality channels but there does not exist a single a priori mapping. By tailoring our tool to specifically assist in the research of viral population dynamics, we can supplant existing work flows and make visual analytics easier in a vital and expanding domain. Our tool is not only better at representing variation data than existing solutions, it also includes tools for vertical summarization and exploration of data that make the comparison problem easier.

Our current tools have some limitations. One common problem when dealing with population dynamics from short reads data is visualizing the difference between two sequences taken from the same population at different time points. Currently this is accomplished by juxtaposition, but an explicit computation and visual-

ization of sequence difference would fit within our current visual paradigm with relatively few modifications. This would also allow additional scalability in terms of the number of sequences that can be viewed at once, since juxtaposition is not an effective comparison tool once there are enough data series that scrolling or excessive resizing is required to view all of the data. Our current tool also lacks tools for dynamic changes in vertical ordering, which further complicates the comparison problem. As more data become available this need for scalability in terms of number of sequences will become more pressing.

Another use case currently not supported is that of the "swarm" paradigm of infection. Currently we are assuming that there is a single reference genome but in real world viral infections the actual infecting population may be a distribution of multiple distinct genomes. Rather than comparing a particular section of a genome to a reference we would then need to compare multiple reference genomes with known prevalences to our read data.

Despite these current limitations, the interaction and visualization techniques we developed seem naturally extensible to different problem domains. In many (if not most) domains, the problem of unifying multiple channels of data quality information into a single view is applicable. Outlier preserving aggregation techniques, with vertical summarization and comparison tools, are also a general problem for data field visualizations. By refining and applying the techniques in the Layer Cake tool, we can begin to address a wide class of domains that would benefit from access to information visualization techniques, such as time series comparisons or general sequence comparison.

7 Acknowledgments

This work was supported by NSF awards IIS-0946598 and CMMI-0941013. Related virology research was supported by NIH R01 AI084787.

References

[1] D. Albers, C. Dewey, and M. Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. In *IEEE Transactions on Visualization and Computer Graphics*, volume 17. IEEE, December 2011.

[2] H. Bao, H. Guo, J. Wang, R. Zhou, X. Lu, and S. Shi. MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics (Oxford, England)*, 25(12):1554–5, June 2009.

[3] B. N. Bimber, B. J. Burwitz, S. O'Connor, A. Detmer, E. Gostick, S. M. Lank, D. a. Price, A. Hughes, and D. O'Connor. Ultradeep pyrosequencing detects complex patterns of CD8+ T-lymphocyte escape in simian immunodeficiency virus-infected macaques. *Journal of virology*, 83(16):8247–53, Aug. 2009.

[4] B. N. Bimber, D. M. Dudley, M. Lauck, E. a. Becker, E. N. Chin, S. M. Lank, H. L. Grunenwald, N. C. Caruccio, M. Maffitt, N. a. Wilson, J. S. Reed, J. M. Sosman, L. F. Tarosso, S. Sanabani, E. G. Kallas, A. L. Hughes, and D. H. O'Connor. Whole genome characterization of HIV/SIV intra-host diversity by ultra-deep pyrosequencing. *Journal of Virology*, 84(22):12087–12092, Sept. 2010.

[5] A. Cedilnik and P. Rheingans. Procedural annotation of uncertain information. In *Proceedings of Vis 2000*, pages 77–83. IEEE Computer Society Press, 2000.

[6] M. Fiume, V. Williams, and M. Brudno. Savant: Genome Browser for High Throughput Sequencing Data. *Bioinformatics (Oxford, England)*, 26(16):1938–1944, June 2010.

[7] B. Fry and C. Reas. Processing. http://processing.org.

[8] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *Proceedings of SimVis 2006*, volume 6, pages 143–156, 2006.

[9] M. Harrower and C. A. Brewer. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1):27–37, June 2003.

[10] P. Lager. GUI for Processing. http://www.lagers.org.uk/g4p/index.html.

[11] A. MacEachren. Visualizing uncertain information. *Cartographic Perspective*, 13(3):10–19, 1992.

[12] E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG*, 24(3):133–41, Mar. 2008.

[13] I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, and D. Marshall. Tablet–next generation sequence assembly visualization. *Bioinformatics (Oxford, England)*, 26(3):401–2, Feb. 2010.

[14] C. B. Nielsen, M. Cantor, I. Dubchak, D. Gordon, and T. Wang. Visualizing genomes: techniques and challenges. *Nature methods*, 7(3 Suppl):S5–S15, Mar. 2010.

[15] J. S. Reis-Filho. Next-generation sequencing. *Breast cancer research : BCR*, 11 Suppl 3:S12, Jan. 2009.

[16] P. Rheingans. Task-based color scale design. In *PROC SPIE INT SOC OPT ENG*, volume 3905, pages 35–43, 2000.

[17] F. Sanger, S. Nicklen, and a. R. Coulson. DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology (Reading, Mass.)*, 24(12):104–8, Jan. 1992.

[18] M. C. Schatz, A. M. Phillippy, B. Shneiderman, and S. L. Salzberg. Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome biology*, 8(3):R34, Jan. 2007.

[19] B. J. Swihart, B. Caffo, B. D. James, M. Strand, B. S. Schwartz, and N. M. Punjabi. Lasagna plots: a saucy alternative to spaghetti plots. *Epidemiology (Cambridge, Mass.)*, 21(5):621–5, Sept. 2010.

[20] J. Thomson, B. Hetzler, A. M. MacEachren, M. Gahegan, and M. Pavel. A typology for visualizing uncertainty. *Proceedings of SPIE*, 2005(January):146–157, 2005.

[21] C. Trapnell and S. L. Salzberg. How to map billions of short reads onto genomes. *Nature biotechnology*, 27(5):455–457, 2009.

[22] B. Trumbo. A theory for coloring bivariate statistical maps. *The American Statistician*, 35(4):220–226, Nov. 1981.

[23] H. Wainer and C. M. Francolini. An Empirical Inquiry Concerning Human Understanding of Two-Variable Color Maps. *The American Statistician*, 34(2):81, May 1980.

[24] Z. Xie, S. Huang, M. Ward, and E. Rundensteiner. Exploratory Visualization of Multivariate Data with Variable Quality. *2006 IEEE Symposium On Visual Analytics And Technology*, pages 183–190, Oct. 2006.

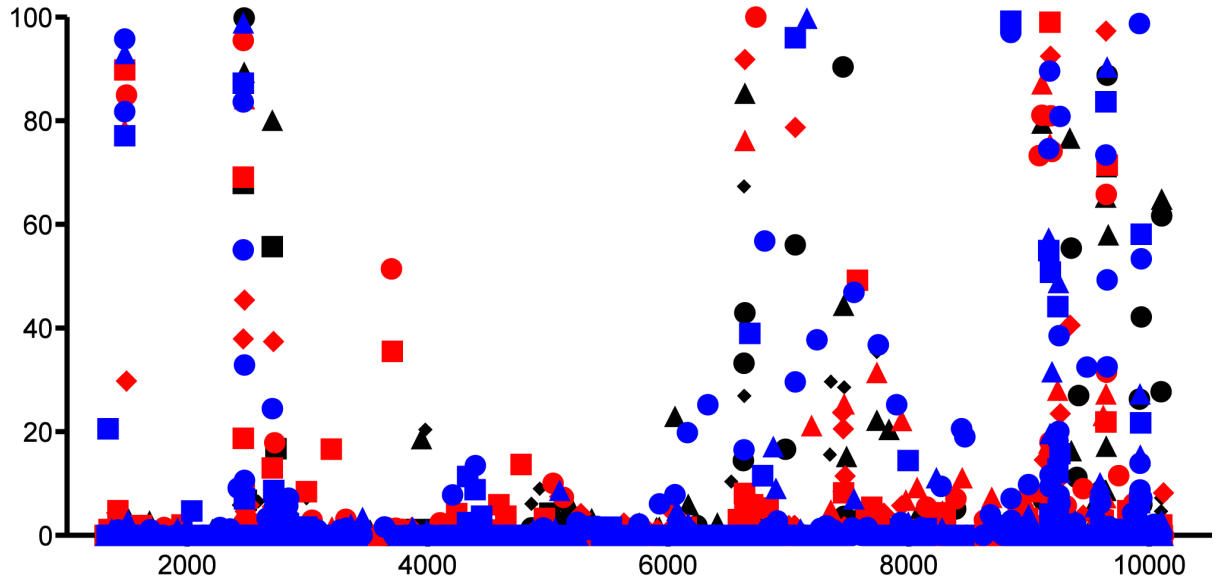

Figure 6: A scatterplot of non-synonymous variant percentage at different areas relative to the reference genome, across 12 individuals in 4 different genotypic groups. There is significant overplotting, and it is difficult to see what effect, if any, genotypic group has on viral population dynamics

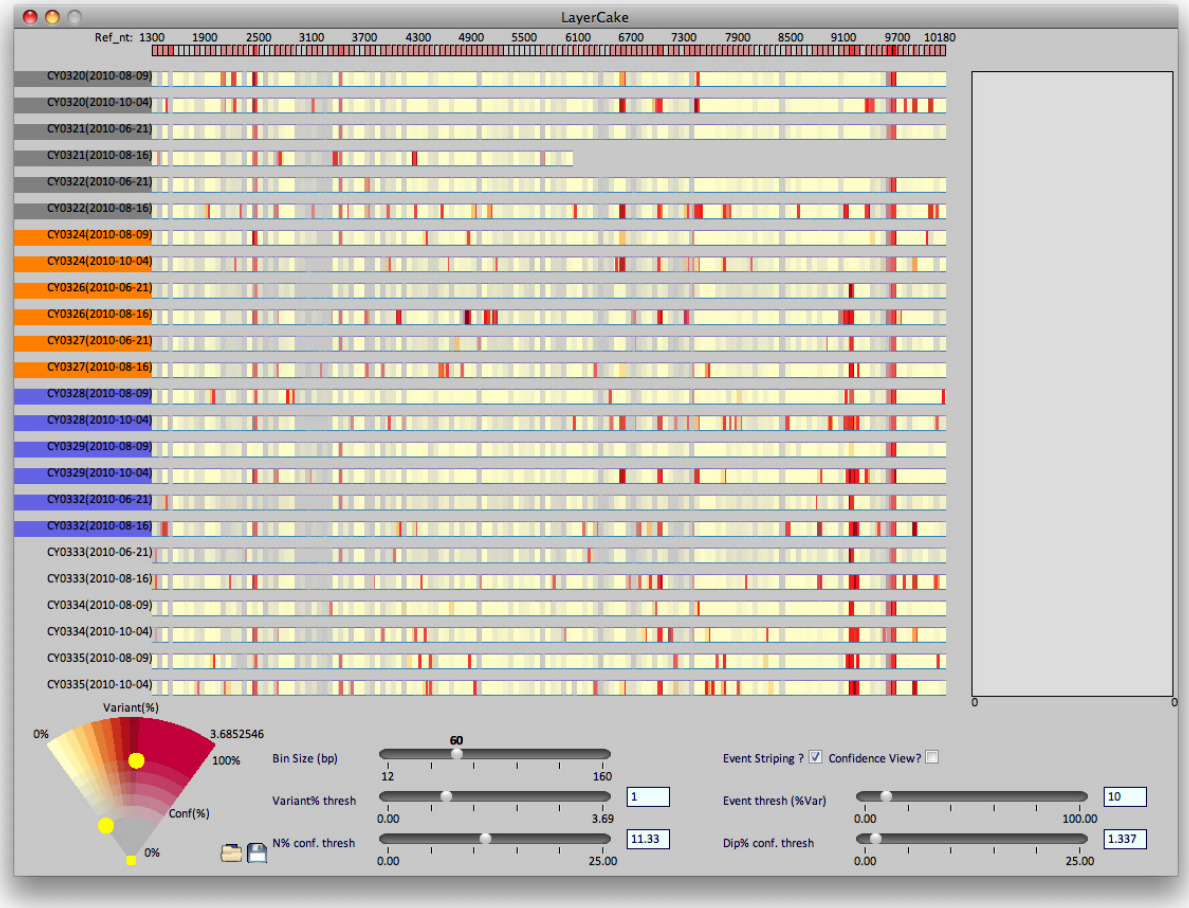

Figure 7: The same data as in Fig. 6, in LayerCake, with full GUI exposed. We can still visually pick out areas with variation “spikes,” but now we have the capability to juxtapose additional time points, and also make per individual and per-group comparisons. Also notice that we can see where there are data gaps or incomplete sequences, (as in the fourth sequence) data that are lost in a scatterplot.