bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

HiPiler: Visual Exploration of Large Genome Interaction Matrices with Interactive Small Multiples

Fritz Lekschas, Benjamin Bach, Peter Kerpedjiev, Nils Gehlenborg, and Hanspeter Pfister

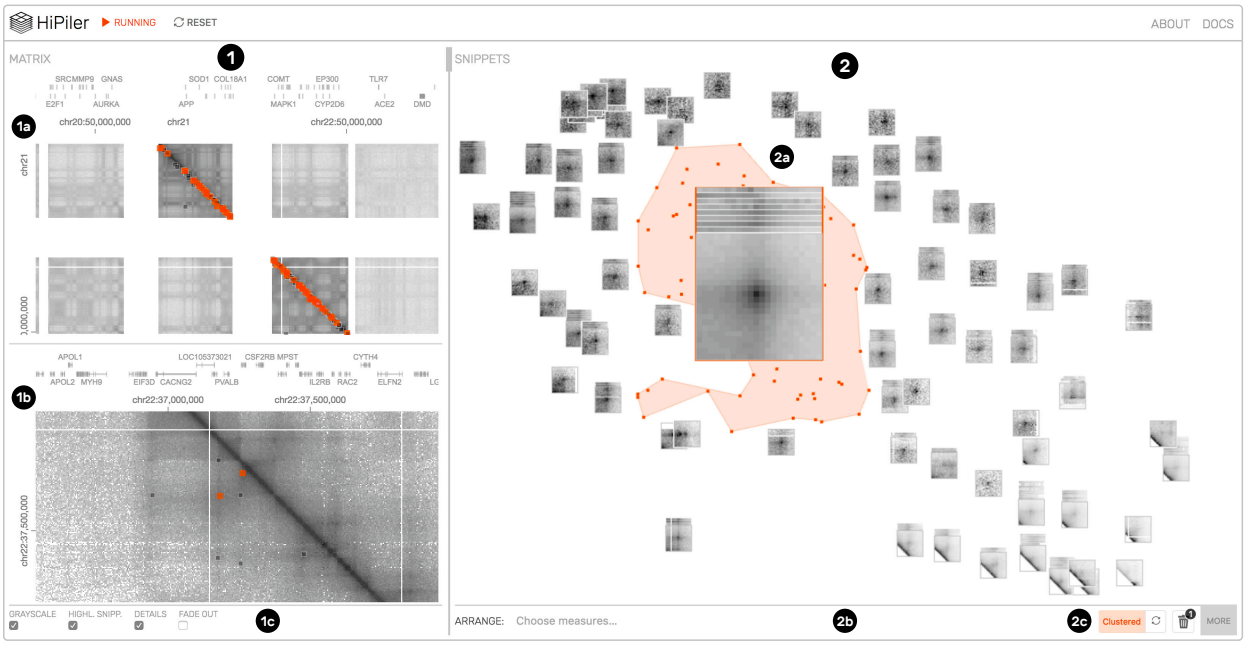

Fig. 1. HiPiler's interface consists of two views. The matrix view (1) contains an overview (1a) and detail (1b) matrix. The snippet view (2) presents regions of the matrix as interactive small multiples. In this example, snippets are arranged with t-SNE (2c) and a well-pronounced pile snippets is highlighted (2a). View menus for operation are located at the bottom (1c and 2b).

Abstract—This paper presents an interactive visualization interface—HiPiler—for the exploration and visualization of regions-of-interest in large genome interaction matrices. Genome interaction matrices approximate the physical distance of pairs of genomic regions to each other and can contain up to 3 million rows and columns with many sparse regions. Traditional matrix aggregation or pan-and-zoom interfaces largely fail in supporting search, inspection, and comparison of local regions-of-interest (ROIs). ROIs can be defined, e.g., by sets of adjacent rows and columns, or by specific visual patterns in the matrix. ROIs are first-class objects in HiPiler, which represents them as thumbnail-like "snippets". Snippets can be laid out automatically based on their data and meta attributes. They are linked back to the matrix and can be explored interactively. The design of HiPiler is based on a series of semi-structured interviews with 10 domain experts involved in the analysis and interpretation of genome interaction matrices. We describe six exploration tasks that are crucial for analysis of interaction matrices and demonstrate how HiPiler supports these tasks. We report on a user study with a series of data exploration sessions with domain experts to assess the usability of HiPiler as well as to demonstrate respective findings in the data.

Index Terms—Interactive Small Multiples, Matrix Comparison, Biomedical Visualization, Genomics

1 Introduction

The human genome is about 2 meters long and tightly folded into the cell nucleus, which has about the same diameter as a human hair. The folded genome forms a dense, fractal-like three-dimensional structure in which genomic regions can be in close proximity that are distant on the genome sequence. The probability of two regions being in proximity to each other can be inferred using modern genome sequencing techniques. The genome interaction matrix is symmetric and contains up to 3 million rows and columns. Known repetitive and hierarchically nested features in the structure of the folded genome appear across multiple scales, ranging from hundreds of millions down to a few thousands of base pairs in size. Interpretation of genome interaction matrices to explore the folding of the genome inside the cell nucleus is essential to identify disease associated with incorrect folding [23, 31, 39].

Heatmaps are currently the de-facto standard for the visualization of genomic interaction matrices. Interactive visualization tools have been developed [44] but are focused on supporting visualization of a single or a small number of views of the matrix and navigation through pan and zoom [13, 25]. However, detailed exploration and comparison of thousands of small regions-of-interest (ROI) is unsupported by current tools yet needed, due to the size and multi-scale nature of the folded genome.

In this paper, we present HiPiler—an interactive visualization tool designed for exploration and analysis of thousands of ROIs extracted

- *Fritz Lekschas, Benjamin Bach, and Hanspeter Pfister are with Harvard University.*
  E-mails: {lekschas, bbach, pfister}@seas.harvard.edu.
- *Peter Kerpedjiev and Nils Gehlenborg are with Harvard Medical School.*
  E-mail: {pkerp, nils}@hms.harvard.edu.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

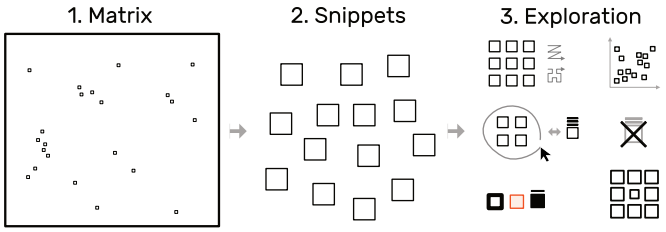

Fig. 2. The HiPiler approach: decompose a large matrix (left) into small snippets (middle) and explore these snippets (right) using different layouts, arrangements, and styles, while maintaining the global context. The small squares within the matrix represent snippet locations.

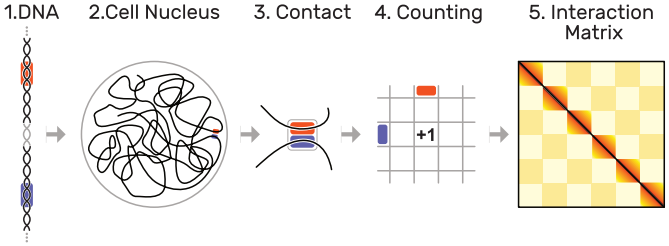

Fig. 3. Hi-C methodology: as the DNA (1) is organized non-arbitrarily in the cell nucleus (2), certain parts (highlighted in orange and blue) are frequently in close contact (3). These contacts are quantified over a set of several hundred million cells(4), leading to interaction matrix of up to 3 by 3 million cells (5). Dark colors indicate more frequent contacts occurrences of two loci.

from one or more genome interaction matrices (Fig. 1). HiPiler takes a set of ROIs, each associated with a pair of genomic locations, and visualizes them as a small heatmaps; called *snippets* hereafter. Each snippet is associated with a set of ordinal and categorical attributes, such as noisiness, size, or source dataset, that are either derived from the matrix data or prior knowledge. Based on this data, HiPiler enables automatic and manual ordering, positioning, grouping, filtering, and visual manipulation to identify patterns present across the set of snippets (Fig. 2). Additionally, the context of snippets in the matrix is maintained through highlighting of snippet locations in the interaction matrix.

HiPiler is based on a set of semi-structured interviews with 10 domain experts from several genomics research labs as well as iterative design sessions over the course of several months. The interviews led to the formulation of six generic and crucial tasks for the exploration of large interaction matrices and ROIs. HiPiler is designed to support four types of scenarios: *i)* retrieval, exploration, and identification of patterns in small ROIs; *ii)* characterization, aggregation, and outlier detection of similar patterns; and *iii)* comparison of ROIs across multiple matrices, e.g., to compare different datasets, experimental conditions, or extraction algorithms. *iv)* correlation of matrix patterns with other genomic attributes, e.g., genes or protein-binding sites.

Through a user study that involved recorded, interactive data exploration sessions with domain experts, we evaluated the usability and appropriateness of HiPiler for the respective data and tasks. Our study showed that HiPiler is easy to learn and use, and that it offers important benefits to scientists who are analyzing and interpreting genome interaction matrices. We conclude with a list of insights and findings from the data exploration sessions, provide a list of requested features that were out of scope for this research, and outline future extensions as well as possible generalizations of our approach.

2 Background—Visual Analysis in Spatial Genome Organization

The genome of eukaryotic cells consists of DNA that is spatially organized around proteins and folded into a compact form within the cell nucleus. It has been shown [18] that the 3D structure of the genome is an important factor for regulation of gene expression, replication, DNA repair, and other biological functions. Biologists are interested in uncovering the mechanisms that drive global and local folding to better understand the vast complex regulatory network. This aids comprehension of the functional diversity of cells and how changes in the spatial conformation of the genome can cause diseases [31]. The DNA is folded hierarchically [19] into nested domains of decreasing size. In contrast to protein structures, there is no single conformation of the DNA; instead, the genome is dynamically reorganized depending on the state of the cell. Smaller domains of the DNA tend to be more flexible compared to larger compartments.

2.1 Hi-C Matrix Analysis

Hi-C [30] is a method to capture genome-wide interactions. It is based on chromosome conformation capture (3C) [10], in which genome segments cross-link when they are spatially close to one another. These cross-linked fragments are extracted, amplified, and mapped to the genome to quantify the interaction per locus (Fig. 3). Except for single-cell Hi-C experiments, a population of millions of different cells is examined at once, producing an average map of contact probabilities. Hi-C datasets are very sparse and the measured contact probabilities follow a power-law decay, i.e., regions that are close to each other on the genome sequence are very likely to be in close contact while regions that are distant from each other on the genome, are expected to have almost no contact.

Despite some breakthroughs, the exact mechanisms that govern the folding of DNA are still unknown. To gain a better understanding, experts typically visualize interaction matrices as a heatmap (Fig. 1.1). Each square represents two genomic locations and the color indicates the contact probability between these regions. Since Hi-C is not comprehensive on its own, other genomic and epigenomic data are often studied alongside the contact probability. Experts usually start exploring the interaction matrix from two angles: (i) finding global patterns or visually confirming computationally-determined patterns and (ii) inspecting various patterns across ROIs to identify variance under different conditions.

2.2 Expert Interviews

In order to identify the current challenges in the analysis of interaction matrices, we conducted a series of semi-structured interviews with ten domain experts (seven postdoctoral researchers and three graduate students). Six of the experts are computer scientists who work on algorithmic tools and pipelines. The other four experts are biologists who mainly focus on analysis of interaction matrices. Each interview lasted one hour and focused on three main parts: (i) long-term goals of genome folding-related research, (ii) workflows and strategies to gain insights, and (iii) current challenges.

The long-term goals for use of genome interaction matrices are to better understand the role of the structural organization of the genome in regards to gene regulatory and other biological processes. In this context, Hi-C analysis is also seen as a complement to existing epigenomic data. Therefore, researchers want to compare multiple conditions or subjects and use the interaction matrix to drive exploration, confirm algorithms, present findings, and generate new ideas. One of the major challenges is the size of the interaction matrix and the high number of relatively small and sparsely distributed ROIs. For example, (i) exploration of long-range interactions is cumbersome with current tools as the context is quickly lost with pan and zoom interactions. Also, (ii) the large number of pattern instances makes it hard to spot subtle differences or outliers. Finally, (iii) the data is very noisy as the folding of the genome is dynamic and not every visual pattern highlights a biological feature but could instead be caused by spatial constraints. Thus, findings needs to be verified by a number of other genomic measures as corroborating evidence.

Over the course of the interviews we learned that domain experts are comfortable with the heatmap visualization of the interaction matrices. Also, the field of interaction matrix analysis is not mature yet. The

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

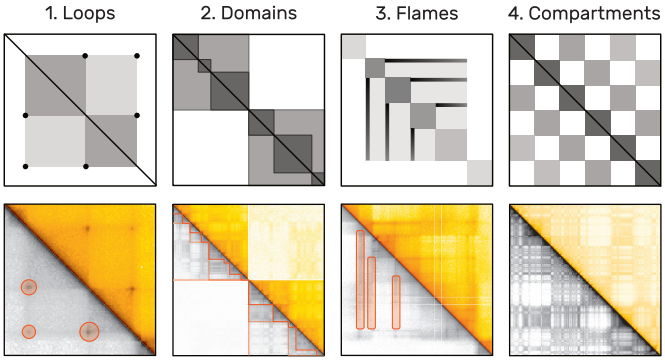

Fig. 4. Frequent patterns in interaction matrices by decreasing size. The upper plots (A) show schematixc illustrations of actual examples (B), taken from Rao et al. [36]. As interaction matrices are symmetric, the lower triangular matrix is displayed in grayscale to highlight the patterns using orange markers. *Loops* (1) appear as dark central dots and span only few cells. *Domains* (2) are darker rectangles that are presumably organized hierarchically. *Flames* (3) are horizontal or vertical lines. Active and inactive compartments of the genome create a global *checkerboard* pattern (4).

number of well-studied patterns that correlate to biological features is limited and there are no community-defined analysis standards yet. Several domain experts mentioned the importance of visualization for their research. For example, one expert stated that p-values are far less important to foster confidence in novel findings compared to visualization of the related matrix patterns. According to another domain expert, when developing feature extraction algorithms for interaction matrices, bioinformaticians try to model what they are seeing with their eyes. The great power of Hi-C comes from the ability to work on averages over millions of cells. At the same time, experts need to carefully check for outliers and false positives to not be fooled by the average.

2.3 Common Hi-C Matrix Patterns

Some of the most common patterns in the analysis of interaction matrices are shown in Fig. 4. *Loop* patterns appear as dark dots in the center and can be seen as an actual loop (Fig. 3) of the DNA. *Domains* are darker rectangles that indicate higher intra-domain than extra-domain contacts. They can be thought of as coils of DNA and are often enclosed by loops. We call loops and domain boundaries (i.e., the location where two domains meet) *point-based* patterns as they normally only span a very limited number of matrix cells. Domains as a whole can be of different sizes and are assumed to be organized hierarchically. *Flames* are horizontal or vertical streaks of darker colors that indicate higher contact probability of one locus with several others. Finally, a well-studied global phenomenon is the *checkerboard* pattern, indicating a fairly strict categorization of the genome into active and inactive compartments with high intra- and low inter-compartment contact probabilities [30]. More details on the specific biological background of the presentation patterns are described in several reviews [8, 9, 17].

2.4 Hi-C Matrix-Analysis Tasks

Based on our interviews, we identified 6 general tasks related to the analysis and exploration of interaction matrices.

- **T1 Search for known patterns:** Some visual patterns are known to have a specific biological meaning. Experts typically look for them first.
- **T2 Discover new patterns:** When studying multiple ROIs, experts often find a variety of recurrent patterns, which have no known biological functions.
- **T3 Study one instance of a pattern:** Once a pattern or ROI has been identified, domain experts are interested in studying the details of this pattern.
- **T4 Compare instances of one pattern type:** The variance of a pattern types and their distributions in interaction matrices are essential for studying biological features.
- **T5 Correlate pattern instances with genomic features:** Snippets are associated with additional attributes describing other genomic features. Experts want to identify correlations between these attributes and the patterns.
- **T6 Compare ROIs across matrices:** Experts want to compare ROI across multiple matrices (e.g., different experimental settings or replicates) to draw causal relationships and assess the stability of patterns.

3 Related Work

Visualizing Genome Interaction Matrices

Advances in high-throughput DNA sequencing have led to a notable increase of available interaction matrices. This sparked development of several specialized software tools for visualization [44]. All of these tools visualize the data in the form of a large matrix, with contact probabilities being translated into color maps, and support pan-and-zoom as a means of navigation. Most applications work offline and integrate 1D tracks, which show various genomic measures in the forms of line graphs or bar charts. HiGlass [25] is a web-based interaction matrix and 1D track viewer that additionally provides seamless view manipulation and view sharing. We integrated HiGlass into HiPiler to provide an overview of the snippet locations, to display 1D tracks, and to select and highlight snippets from the interaction matrix. Pattern-centric visualizations of interaction matrices are currently not supported by any tool. Also, experts use Matplotlib [24] to visualize aggregations of patterns in multiple different ROIs or experimental conditions.

Visualizing Large Matrices

Matrices are a common representation for visualizing networks or graphs [5, 42]. Thus, we briefly overview related visualization techniques that focus on large matrices. Large matrices make it hard to analyze detailed visual patterns and to compare distant parts in the matrix. Common interactions, such as panning and zooming, are *content-agnostic* as they operate entirely on the view level but do not perform any operations on the data itself.

On the other hand, *content-aware* approaches incorporate the data to drive visualization. For example, ZAME [15] aggregates individual cells into higher-level cells as the user zooms out. Zoomed-out cells then show a glyph with the distribution of values grouped by this cell. Such zoom-based interfaces can scale well for very large matrices, but they loose fine-grained visual patterns that are required in interaction matrix analysis. Rather than aggregating cells, Melange [16] allows skipping rows and cells by literally folding them into a third dimension with the effect of reducing the size of the visible matrix. Hence, the remaining non-folded parts are observed in more detail and visual patterns can be compared across long distances. Dinkla et al. [11] present a technique for visually compressing gene regulatory networks and which takes the underlying data into account. The compressed representation works for medium-sized networks but still becomes overwhelmingly large as the network grows. NodeTrix [22] is a hybrid approach that shows only clusters as matrices and visualizes connections between clusters as straight lines.

While all these approaches improve exploration of large matrices, our approach is different in that we do not study exploration of the entire matrix but focus on a set of ROIs of the matrix (snippets). This approach provides enough expressive power to focus on important parts in the matrix only and makes visual analysis dependent on the number of ROIs instead of the actual size of the matrix.

Divide-and-Conquer Approaches

Mining specific patterns in large datasets for display and exploration has been employed successfully in other domains, e.g., image and network analysis. Network motifs [34] refer to subgraphs in a network and are similar to our matrix snippets. Network motifs have frequently been used as first-class objects in network exploration. Visualized in an adjacency matrix, they result in recurring visual patterns [5]. For

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

example, clusters occur as rectangles (similar to *domains in interaction matrices*) and highly connected nodes appear as horizontal and vertical lines (similar to *flames*). Schreiber et al. [38] allow for the selection of a sub-graph (network motif) in a network and consequently retrieve and visualize all occurrences of sub-graphs with similar network topology. Dunne and Shneiderman [12] first detect network motifs such as clusters and fans and then replace their occurrence by specialized glyphs in the node-link diagram of the network. Von Landesberger et al. [41] extract motifs, obtain metrics (e.g., size, density, average degree and others), and visualize the retrieved motifs using a self-organizing-map [27] layout. Except for Cubix [4], which visualizes the evolution of ego-networks through heatmaps similar to adjacency matrices, network motif visualization has so far been focused on node-link representations, and not been seen as an approach for interactive visualization of large matrices. While the approach of von Landesberger et al. [41] could be generalized to adjacency matrices, explicit interactive means for exploration and alternative motif layouts were not the focus of their work. While we can derive inspiration from these approaches, snippets from interaction matrices are very different from network motifs with respect to visual appearance.

Our work is most inspired by MultiPiles [3]—an interface that employs the visual and interactive metaphor of piling adjacency matrices and exploring these piles to visualize time sequences in dynamic networks. We integrate the piling metaphor and some of the exploration features from MultiPiles but heavily extend upon them by introducing linear ordering, multi-dimensional arrangements, clustering, filtering, and grouping approaches for exploring many snippets.

4 Design of HiPiler

The key goal for HiPiler is to enable the exploration of many ROIs in a large matrix via interactive small multiples that represent snippets. Snippets, can be ordered, arranged, filtered, and grouped independently of their neighborhood (Fig. 2).

The development was driven by questions such as *"How can we mindfully limit the number of snippets shown?"*, *"Which interactions are important to efficiently support arrangements?"*, or *"How to effectively link the interaction matrix with the snippets"*. In the following sections we will describe the data model, design, and interactions of HiPiler that are guided by the identified tasks (Sect. 2.4).

HiPiler consists of two main views: the matrix view and the snippets view (Fig. 1.1 and 1.2). The matrix view displays one or more interaction matrices and supports two display modes: *overview* and *detail*. The upper half of the matrix view contains the overview (Fig. 1.1a) and indicates the location of all explored snippets. The detail matrix (Fig. 1.1b) is in the lower half and enables browsing the interaction matrix via pan-and-zoom. Both matrices highlight the snippet location with colored rectangles. A menu at the bottom (Fig. 1.1c) provides options for customization. The snippet view (Fig. 1.2) displays snippets according to user-defined ordering, arrangements, filtering, or grouping. For example, in Fig. 1.2 snippets have been arranged with t-SNE [32](Fig. 1.2c). The center of the plot features a scaled-up pile of multiple snippets (Fig. 1.2a). The original location of the piled-up snippets is indicated by the orange hull drawn in the background. Other means of operating the snippet view are provided via a menu at the bottom (Fig. 1.2c). An overview of all conceptual design aspects of HiPiler is given in Fig. 5.

4.1 Data Model

HiPiler is designed for large matrices with millions of columns and rows. Many graph and network data can be represented as a matrix, where the ⟨i,j⟩ cell contains a correlation measure between node i and node j. In addition, cells can be associated with multiple categorical and ordinal attributes. These attributes can be measures or annotations derived from the matrix or given by prior knowledge. For example, noise, pattern sharpness, or distance-to-diagonal are derived from the matrix and referred to as *data* attributes. Prior knowledge, which we refer to as *meta* attributes, can for example be confidence in the correlation measure, protein-binding sites, or gene expression levels. HiPiler assumes a fixed ordering of rows and columns, which in our case is given by the genome sequence.

We assume, but do not require, that datasets contain many small ROIs that are distributed across the interaction matrix. Here, each ROI is defined as a set of start and end locations.

4.2 Design Aspects

The design of HiPiler is guided by the tasks that we identified during the interviews (Sect. 2.4). During the development of a prototypical implementation, we met with three of the initial and one additional domain expert for 1 -2 hours to collect feedback on our design choices. We additionally presented the prototype to a group of domain experts that haven't been involved in the initial interviews to get unbiased feedback.

Snippet Metaphor: Snippets are the essential building blocks of HiPiler as they help experts to identify known (T1) and unknown patterns (T2) among ROIs. In addition to displaying a part of the matrix, snippets can be associated with categorial and ordinal attributes, which are displayed with additional visual marks (Fig. 6); addressing T3 and T5. The visualization of these attributes has been kept minimal to act as information scent [35] and to avoid distraction from the snippets.

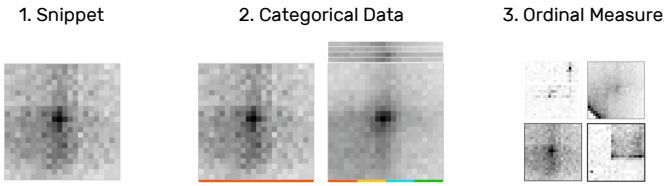

Fig. 6. Snippets are drawn as a heatmap (1), showing the matrix data of their ROI. Categorical attributes can be visualized with color tags (2). An ordinal measure associated with a snippet can be shown via frame width and color (3). For example, a high value is drawn as a dark thick border.

Snippet Layout: Domain experts want to explore several hundreds to thousands of instances of one pattern type simultaneously (**T4**) to identify groups and uncover potential correlations between patterns in the interaction matrix and other genomic features. To support these tasks, HiPiler's layout is entirely data and attribute-driven; allowing for one-dimensional (1D) ordering, two-dimensional (2D) arrangements, or multi-dimensional (MD) clustering via dimensionality reduction (Fig. 7).

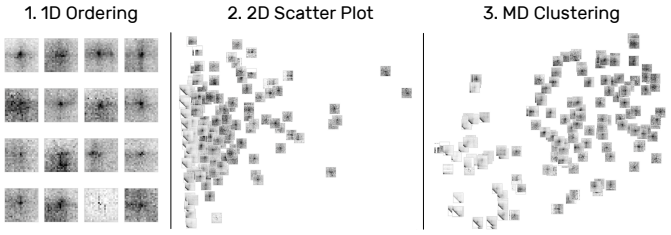

Fig. 7. Snippets can be arrange along various different dimensions. For a single attribute, snippets are laid out in 1D supporting reading direction and Hilbert curves (1). Selecting two attributes creates a scatter plot (2). For more than two attributes HiPiler applies dimensionality reduction algorithms (3) such as t-SNE [32].

Aggregation: To support the exploration of large numbers of snippets, HiPiler applies and extends upon the piling metaphor of Multipiles [3]. Snippets are stacked into a pile featuring a *cover* matrix that shows a summary of the stacked snippets. The cover is calculated by taking the mean or standard deviation of all snippets (Fig. 8.1c and 8.1d). These *cover modes* help experts to assess the average expression and variance of patterns (T4 and T6). To briefly browse piled snippets, HiPiler displays up to eight *pile previews* as 1D heatmaps above the cover (Fig. 8.2a). Previews show the mean column values of their

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

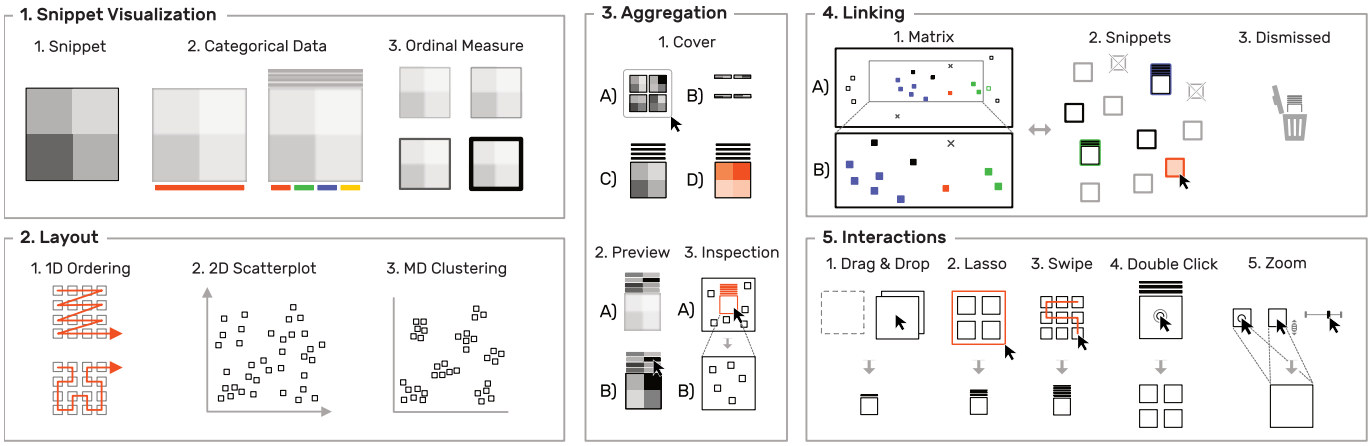

Fig. 5. Overview of the conceptual design aspects of HiPiler.

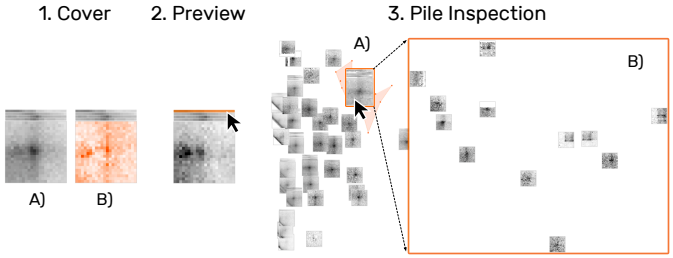

Fig. 8. HiPiler displays a cover matrix of the average (1a) or variance (1b) of snippets on a pile. Additionally, each snippet is displayed as a 1D preview showing a horizontal aggregate of the snippet's rows (2). Moving the mouse cursor over a preview shows the related matrix. Inspecting a pile (3a) temporarily hides all other snippets (3b).

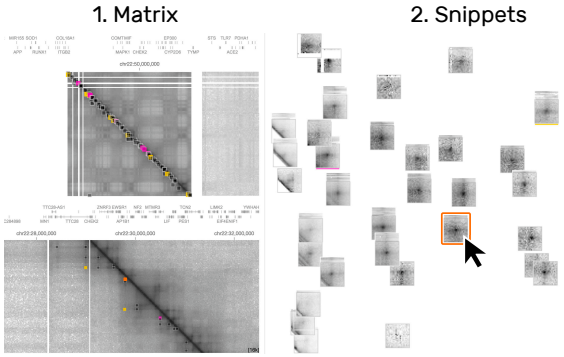

Fig. 9. The location of the snippet in the interaction matrix (1) are highlighted during the exploration to provide context (pink, orange, and yellow rectangles). HiPiler distinguishes between groups of snippets via color tags (2, pink and yellow bar).

underlying matrix data (Fig. 8.2b). Moving the mouse cursor over a preview temporarily displays the related snippet. For a large number of snippets per pile it would be inefficient to limit the exploration to the cover or individual (dispersed) snippets only. Therefore, we have added support for hierarchical *inspection* of piles. When inspecting a pile, only the snippets of the pile are shown and the layout is automatically scaled to accommodate the region occupied by these snippets only (Fig. 8.3).

Linking: In almost all cases the neighborhood of snippets is crucial as the genomic locations associated with snippets act as the *ground-truth* in genome biology. In HiPiler, snippets are therefore interconnected with the matrix view by highlighting their location via colored rectangles (Fig. 9.1). Snippet locations can be shown permanently via color tags or temporarily by selection. To support the investigation of the neighborhood of a snippet, HiPiler implements a *detail matrix* view (Fig. 9.1b). It is possible to fade out snippets that are not visible in the viewport of the detail matrix to provide more focus. Also, HiGlass [25] supports 1D genomic tracks, which enable experts to correlate patterns to many other genomic measures (T5). Finally, the matrix view can host more than one matrix to support comparison across datasets (T6).

4.3 Interaction patterns

We have adopted piling interactions from Multipiles [3], where the user can manually create piles with drag-and-drop (Fig. 5.5.1) or lasso selection (Fig. 5.5.2) and disperse piles via double-clicking (Fig. 5.5.4). In addition, HiPiler implements more fine-grained controls for piling and zooming. For 2D and MD layouts, HiPiler automatically down-scales the size of snippets and piles to fit a larger number of snippets on the screen and to avoid clutter. To make piling in dense layouts easier, we added swipe-based pile selection, where the user can move the mouse cursor over the snippets to be selected while holding down the left mouse button (Fig. 5.5.3). Also, temporary upscaling of individual snippets is supported by steering the scroll wheel while having the mouse cursor placed over the target snippet (Fig. 5.5.5). It is also possible to zoom into the entire snippet view to inspect a sub-region.

5 Use Cases

In the following usage scenario we demonstrate how HiPiler can be used to study the diversity and variance of loop patterns that have previously been reported by Rao et al. [36]. A perfect loop pattern exhibits a dark central dot surrounded by relatively bright areas (Fig. 1.2a). Since genome interaction data is sparse and noisy and since there are no gold standards for pattern extraction yet (Sect. 2), questions that guide our exploration are: *"How do average patterns at extracted locations look?"*, *"Can we compile a set of snippets with well-pronounced loop patterns?"*, *"Are we missing locations which express the same or similar patterns?"*, *"Is there a correlation between patterns and other attributes?"*, and *"Can we see similar patterns at the same locations in other matrices?"*.

Overview

After loading the data, some snippets in the snippet view with loop patterns are immediately visible (T1) (Fig. 11.1a), while others contain parts of the diagonal, are noisy, or appear to be empty (T2) (Fig. 11.1b). Ordering snippets by their distance to the diagonal uncovers more consecutive snippets with similar patterns (Fig. 11.2). Since the number of snippets is too large to get an overview, we group snippets by their pairwise Euclidean distance of the underlying data so that scrolling is avoided (Fig. 11.3). This essentially piles up snippets hierarchically

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

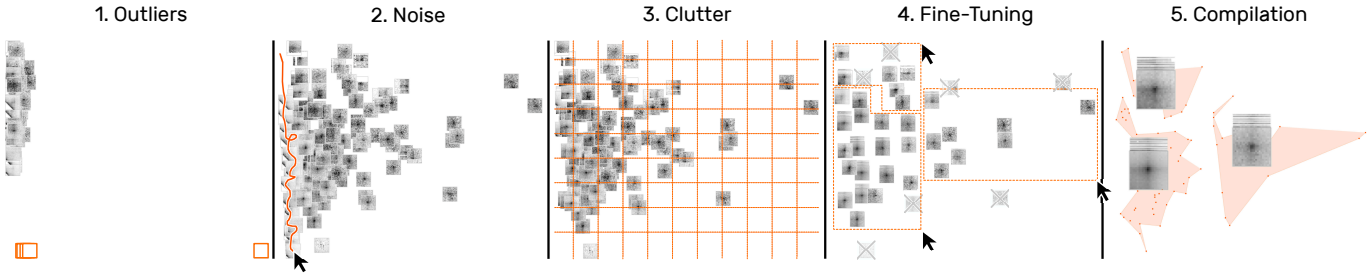

Fig. 10. Snippet curation via filtering. A typical filtering process involves the removal of outliers (1) and noise (2), clutter reduction via automatic piling (3), and manual grouping (4) until a satisfactory collection of snippets is obtained (5).

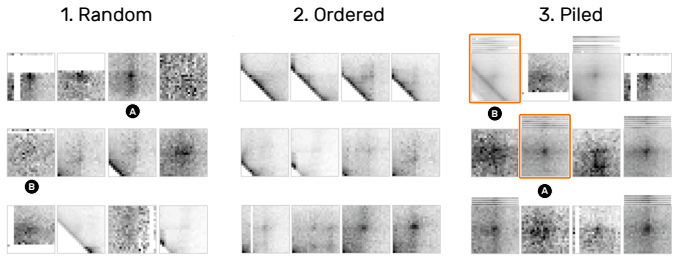

Fig. 11. 1D ordering and similarity piling. (1) Random arrangement of snippets. (2) Snippets ordered by their distance to the diagonal show a progressively emerging pattern. (3) Piling by pairwise similarity highlights similar patterns and outliers. For example, 3a shows a well-pronounced loop pattern while 3b is more noisy due to the diagonals.

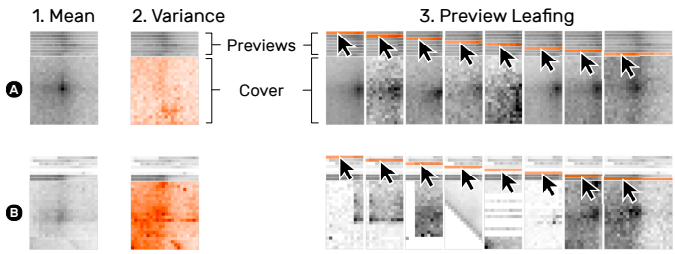

Fig. 12. Snippet aggregation. The default cover shows the mean of snippets on a pile (1). The variance cover mode (2) highlights the deviation of snippets. Moving the mouse cursor over a pile's previews temporarily shows the corresponding snippet or pile (3).

into k clusters, where k is the maximum number of snippets that can be displayed at the current size so that no two snippets or piles overlap. The covers show the mean patterns of piles, indicating that some exhibit a well-pronounced loop pattern (Fig. 11.3a) while others are more diverse (Fig. 11.3b).

The default cover displays the mean signal across all snippets on a pile (Fig. 12.1). Darker or more saturated colors indicate higher values of the underlying data. Changing the cover mode to variance shows the standard deviation of piled-up snippets and supports assessing pattern variance (T4) (Fig. 12.2). The piles that are showing a well-pronounced mean loop pattern do not usually express significant variance, indicated by a relatively flat heatmap (Fig. 12.2a). The variance cover shows significantly darker and more saturated spots for piles containing noisy snippets or outliers (Fig. 12.2b).

To get a better sense of the pile composition (T4), moving the mouse cursor over snippet previews temporarily shows the previewed snippet as a whole on the cover (Fig. 12.3). HiPiler limits the overall amount of previews to a configurable number i to prevent occlusion by high stacks of previews. When a pile consists of more than i snippets HiPiler utilizes k-means clustering [33] to group the snippets. The mean of all clustered snippets is used to represent the preview.

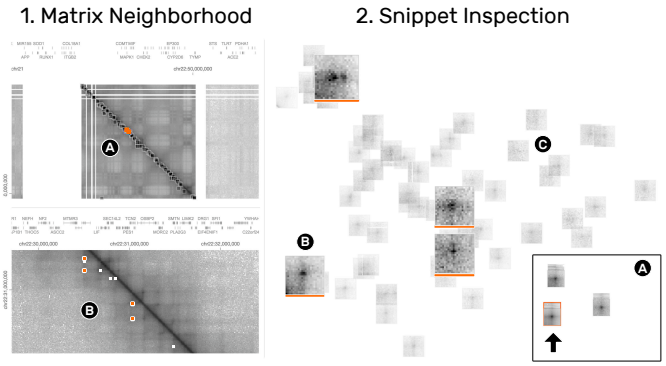

Fig. 13. The matrix (1) and snippets (2) views are highly interconnected to enable the exploration of the spatial neighborhood of snippets. The detail matrix (1b) shows the colored snippets (2b) in context. To focus on the currently visible neighborhood, non-visible snippets are faded-out (2c).

Filtering

One of the initial questions involves dissecting noisy and well-pronounced patterns. Arranging snippets by noise and their distance to the diagonal transforms the 1D ordering into a 2D scatter plot. This spreads out snippets spatially (Fig. 10.1) and supports better differentiation between groups (T1 and T2). First, we notice some outliers, which are completely white and far away from the diagonal (Fig. 10.1 and 1b). Dismissing outliers (i.e., moving them to the trash) re-scales the scatter plot (Fig. 10.2). The *swipe* selection is useful for non-linear fine-grained piling of snippets in dense areas (Fig. 10.2). Moving the mouse over snippets while holding down the left mouse button leaves a trail of which snippets are to be grouped (Fig. 10.2 orange line). Next, to quickly reduce clutter one can auto-pile snippets by grid cells (Fig. 10.3). Manual piling via drag-and-drop or *lasso* selection and dismissing further supports to filter the set of snippets (Fig. 10.4) until a satisfactory collection of well-pronounced loops is obtained. Placing the mouse cursor over a pile displays the location of its piled up snippets given the current arrangement (Fig. 10.5).

Snippet Neighborhood

Having curated a collection of snippets with well-pronounced loop patterns, one task is to study the distribution of the piled-up snippets across the interaction matrix. Clicking on a pile highlights the location of its snippets as orange rectangles in the matrix (Fig. 1.1). *Inspecting* a pile displays only its snippets while hiding any other piles or snippets (Fig. 13.2a). We notice a region which features many highlighted snippets (Fig. 13.1a). Navigating to this region via zoom-and-pan provides spatial context to the snippets (Fig. 13.1b). Being able to explore the neighborhood of snippets is important for correlation of different pattern types, e.g., we can see that the highlighted snippets appear within a dark rectangular area, known as TADs (Sect. 2.4) (Fig. 13.1c). We also find other loop-like patterns that have not been

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

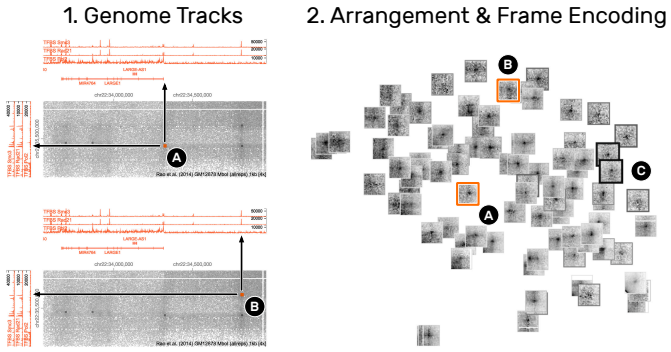

Fig. 14. Correlation of snippets with other attributes. (1) The integration of line graphs next to the interaction matrix show genomic measures. 1a and 1b are two snippets (2a and 2b) that exhibit a loop pattern but are correlated with the expression (peaks in the line graph) of different genomic measures, indicating that the underlying ROIs serve different biological functions. Numerical measures can also be integrated into the snippets view (2) by alternating the frame width and color of snippets (2c). Thicker and darker frames indicate higher values.

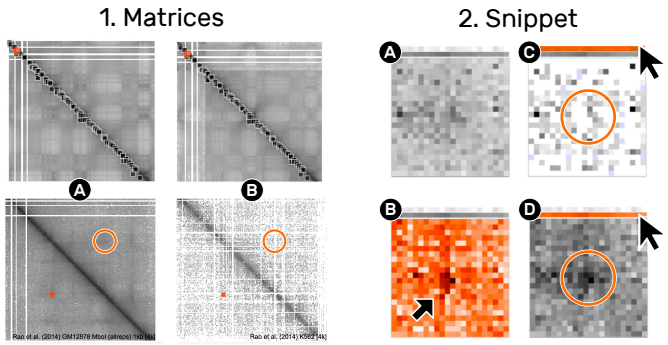

Fig. 15. Comparison of one location in two different matrices. The matrix view (1) shows two different matrices (1a and 1b). The ROI indicated by the small square rectangle is extracted from both matrices, displayed as a pile (2a). The mean cover (2a) does not show a pronounced loop pattern in contrast to the variance cover (2b). The detail matrix view indicates that the right matrix (1b) is less dense than the left matrix. Leafing over the previews reveals that the loop pattern is completely gone in (1c) compared to (1d).

detected by the algorithm. To identify which snippet is currently visible in the detail matrix other non-visible snippets are faded out in the snippet view (Fig. 13.2c). Color tags can be used to permanently highlight snippets (Fig. 13.2b).

Correlation with other Genomic Features

HiPiler supports three ways to investigate correlations between snippets and other attributes (T5): i) integrating additional genomic tracks in the matrix view (Fig. 14.1), ii) visualizing an attribute via frame encoding (Fig. 14.2), or iii) arranging snippets according to meta attributes. For example, the previously colored snippets that show a pronounced loop pattern perfectly align with three other genomic measures displayed above the interaction matrix (Fig. 14.1a and 1b), suggesting that this ROI exhibits a biological function. Frame encoding is useful for integrating attributes into the snippets view and act as information scent [35] for further investigation.

Comparison across Matrices

Comparing ROIs across matrices helps to differentiate the variance of patterns. Extracting the same ROIs in both matrices allows for pairwise comparison of snippets. *Grouping by location* is an operation that automatically piles the pairs of snippets. Activating the variance cover mode highlights the deviation of paired snippet. For example, a pronounced loop pattern uncovers high deviation around the center of the snippets (Fig. 15.2b), meaning that only one of the two snippets contains the loop pattern (Fig. 15.2c and 15.2d). Using this technique in combination with the detail matrix view allows to identify ROIs where the respective loop pattern disappears (Fig. 15.1a and 15.1b).

Further Use Cases

Besides the exploration of loop patterns, HiPiler supports exploration of snippets of any equally-sized squared ROI. Such patterns important for other biological question as well, involving for example: promoter-enhancer pairs, TAD boundaries, telomeric regions, or structural variations. While the pattern types differ, the tasks and resulting interactions are essentially identical.

6 Implementation

HiPiler is implemented as a web application consisting of a front-end interface for the visualizations and a server-side component that provides the data. The front-end is entirely written in JavaScript utilizing Aurelia [14] as its application framework and Redux [2] for fine-grained, history-aware state management. The matrix snippets are visualized with WebGL using Three.js [7] as a middleware. Finally, HiGlass [26] is integrated as a library for displaying the interaction matrix and genomic tracks. The server-side backend serves data to HiGlass and provides the matrix snippets. The backend is implemented in Python and uses Django as its application framework. The contact matrices are accessed through Cooler [1], a Python-based service library for storing and querying of Hi-C data. The front and backend are two separate applications that can be decoupled to load different data types. HiPiler is open source and available on GitHub (`https://github.com/flekschas/hipiler`).

7 User Evaluation

We conducted a qualitative user study with domain experts to investigate the utility and usability of our approach. We obtained agreement from three of the domain experts that we had interviewed earlier as well as from two additional domain experts for an open-ended data exploration session with HiPiler. All experts are computational biologists that work with genome conformation contact data on a daily basis. Two of the domain experts are senior PhD students while the other three are postdoctoral researchers. Their work experience in the field ranges from one to eight years.

We conduced individual sessions that lasted between 1-2 hours. First, we introduced the core concepts of HiPiler and gave a walkthrough of its main functionalities (10-20 minutes) (Sect. 4). Afterwards, each domain expert completed a training session in which they performed specific interactions, e.g., piling up snippets with lasso selection or arranging snippets by some attributes, on a training dataset for 10 minutes. Next, we loaded the expert's own datasets and let them explore the data by themselves (30-90 min). During the exploration, we asked the experts to think aloud and to express their rationale behind operations. The entire study was performed on a pre-configured computer equipped with a 27-inch external monitor and a standard mouse. Keyboard input was not required. We recorded screen content and audio during each session for later analysis.

7.1 Exploratory Sessions

P1, P3, and P4 investigated a set of pairwise enhancer-promoter locations [40], which are assumed to interact structurally [21, 28, 29]. In the interaction matrix, structural interactions appear as dark spots, forming the patterns described in Sect. 2.3. P1, P3, and P4 started browsing the snippets to get an overview: P1 and P3 manually piled snippets that showed a loop-like pattern. P1 ordered snippets by their distance to the diagonal first and then switched to dimensionality reduction with t-SNE to better differentiate between noisy and well-pronounced patterns. P1 and P3 reduced the visual complexity by piling up noisy

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

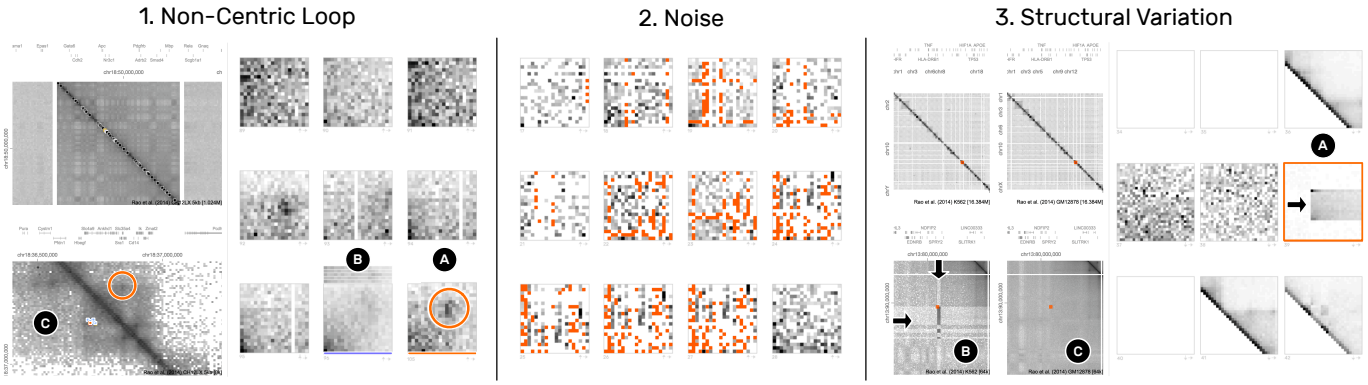

Fig. 16. Three observations made during the data exploration sessions. (1) P1 and P3 individually found a group of pairwise enhancer-promotor interactions that are close to but not directly on a loop (1a and 1b). The loop pattern is indicated with an orange outline. 2a is the snippet with the most pronounced pattern. P1 and P3 wanted to remove the pile (2b) from the same location (C). (2) P4 investigated sparse snippets and visualized low quality cells (highlighted in orange) to identify that most of these snippets are extracted from a low quality region. (3) P5 highlighted a true positive structural DNA deletion (3a) by comparing two datasets (3b and 3c). The deletion (2b) causes the brighter (but not white) columns and rows (2b arrows) and is accompanied by an insertion, indicated by the dark rectangular area next to the highlighted location (3a arrow).

snippets using the swipe selection tool to subsequently dismiss them. P3 found a snippet with a non-centric loop pattern (Fig. 16-1) and investigated its spatial neighborhood using the detail matrix view. P3 noticed that several snippets are located in relative proximity to one another, which is shown by the colored squares in the interaction matrix. P3 decided to keep only one snippet with a strongly pronounced pattern to avoid overrepresentation. Similar to P1, P3 continued with arranging snippets using t-SNE and found more loop-like patterns after examining noisy groups of snippets, indicating that some locations exhibit structural interactions. P4 first checked the overall quality of snippets and noticed high sparsity within the snippets's matrix, which results in salt-and-pepper-like noise. By activating the visualization of low quality cells, P4 found that the large number of low quality cells indicates that these noisy snippets come from a region of low quality (Fig. 16-2). Finally, P4 opened the detail matrix view and navigated to the location of a set of snippets showing a loop-like pattern. They concluded that the patterns are potentially related to another biological feature after finding additional patterns in the matrix view.

P2 explored loop patterns as reported in the literature [36] and wanted to determine the performance of a detection algorithm. P2 started arranging snippets by their distance to the diagonal and noise and identify outliers. To study snippets with a well-pronounced loop pattern, P2 decided to remove noisy snippets first. Finally, P2 tested the t-SNE-based snippet arrangement to further dissect noise from clean patterns and refined the dissection by iteratively applying t-SNE followed by the removal of noisy snippets.

P5 studied structural variations in the genome (e.g., deletions, insertions, or translocations of DNA sequences) and wanted to assess the performance of predicted results from data analysis tools like Delly [37] or Meerkat [43]. They loaded structural deletion sites that are expected to show half empty snippets. Empty snippets can be the result of a i) structural variation or ii) technical limitations of the current technology for generating interaction matrices. To distinguish between them, P5 activated the visualization of low quality cells. P5 piled up and removed some empty (white) snippets and shifted their focus to unexpected non-empty snippets for detailed investigation. P5 browsed the spatial neighborhood of the respective snippet in the detail matrix view and loaded a second interaction matrix for comparison. The second matrix is assumed to have no or less structural variations. P2 found significantly brighter columns and rows in one of the two interaction matrices indicating a true DNA deletion (Fig. 16-3).

7.2 Findings

Snippets are useful for exploring hundreds to thousands of pattern occurrences in interaction matrices. All participants stated that this technique enables them to easily assess the variety and variance of patterns. P1, P3, and P4 pointed out that seeing what an average pattern is composed of is particularly helpful to avoid misinterpretation based on the inclusion of noise or unrelated patterns. The snippets concept further aided P1, P3, and P4 to determine reasonable thresholds of attributes for the exclusion of noisy patterns, i.e., they visually determined at which value they would consider a snippet to be labeled noisy. They note that the snippet view can be used to select promising candidates for further investigation or to build a set of "ground truth" ROIs for evaluating the performance of pattern detection algorithms.

Coordination between the matrix and snippet view is highly appreciated by every participant as many tasks require spatial context. The interplay between the two views enables the participants to explore snippets in new ways, e.g., P2 states that HiPiler enables them to correlate patterns according to prior knowledge while still maintaining context. P1 pointed out that they usually don't know what they can expect to see in a interaction matrix and that it is great to be able to browse the neighborhood of snippets when they spot surprising patterns. P5 noted that for their research questions it is essential to have the matrix view since some biological phenomena lead to patterns that span the entire matrix. During all sessions, the participants spent an equal amount of time on the snippets view, arranging and organizing snippets, and the matrix view, reconfirming findings and further exploring patterns in the neighborhood of snippets.

Users quickly grasp the main operations supported by HiPiler. After the guided 10 minute training session, all participants in our study were able to explore snippets own their own. All participants noted that HiPiler is very easy to learn. P2 said *"the menus are where you would expect them to be"* and P4 stated that HiPiler *"is much more thoughtful than expected"* in comparison to current visualization tools for interaction matrices. However, we acknowledge that all participants are proficient in operating computers and that an initial phase of training is necessary.

HiPiler significantly improves on the state-of-the-art tools for genome interaction matrix exploration. Current tools are currently limited to pan-and-zoom interactions of the entire interaction matrix or require custom, code heavy solutions with Matplotlib [24]. P4 noted *"[HiPiler] takes it from zero to infinity"* to point out that there are no other feature-centric visualization tools for interaction matrices available.

Participants strongly indicated that they will use HiPiler for their research once additional features are implemented. These features include data processing for HiPiler and displaying of various numerical and statistical attributes related to snippets, e.g., translating visual encodings back to their numerical values. P1, P3, and P4 also mentioned the potential for collaborative exploration via shared web-sessions that HiPiler supports.

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

8 Discussion

Additional Features

During the user study (Sect. 7) the domain experts suggested a number of additional features that would make further exploration of ROIs more efficient for analyses on a daily basis. For example, they would like to manually adjust the color intensities of the matrix and snippets in order to emphasize contrast of sub-regions. The domain experts also expressed desire to pick and search patterns manually in the matrix view, for example, as a means to supplement results of a pattern detection algorithm. Yet, this requires image-based pattern detection algorithms similar to Magnostics [6] and provides an interesting set of questions for future research. Some domain experts mentioned that integrating visualizations of other genomic features into snippets would further assist in finding correlations.

In this work, we focus on exploring squared snippets with a specific pattern (e.g., *dots*). In the future, we want to extend HiPiler to support arbitrarily sized snippets of equal ratio, e.g., TADs (Sect. 2.3), which requires first investigating appropriate methods for aggregating snippets of different sizes without destroying patterns in the data. Also, HiPiler supports history and sharing of the state, enabling experts to collaboratively explore matrices. Providing ways to efficiently show the history of exploration and to compare states can foster collaboration and presentation as described by Gratzl et al. [20]. Finally, HiPiler supports comparison of sets of snippets via different snippet covers, frame encoding, and color tags. Further means of visually summarizing and aggregating data attributes of piles could enable experts to more seamlessly transition from context-driven (interaction matrix) to knowledge-driven (data attributes) pattern exploration.

Generalizability

We want to emphasize that the HiPiler approach is not limited to exploration of genome interaction matrices but can theoretically be extended to any graph-based dataset that can be represented in a correlation matrix, which exhibits ROIs with recurrent visual patterns. Large networks are found in biology (e.g., gene regulatory or protein interaction networks), social application (e.g., Facebook), or computer science (e.g., server networks). In other cases, similarity or correlation measures between data objects can be displayed as matrices. The problem of exploring many similar features arises in gigapixel images, too. Here, pixels represent the cells and contain a numerical color value. For example, high content microscopy screening produces very large images of cell cultures, showing the state of cell colonies. Also, in astronomy scientists study very high resolution pictures of star galaxies. In both cases important features (cell bodies and stars) are relatively small compared to the entire image and occur frequently. Eventually, the tasks described in Sect. 2.4 generalize well to different areas with very similar data properties (Sect. 4.1).

Limitations

It is not clear yet how beneficial the piling metaphor would be for snippets of different aspect ratios in terms of aggregation through piling. While it is technically straightforward to visually scale patterns to equal size, domain experts are not sure what the aggregation of differently sized patterns would mean biologically as this would lead to many non-trivial normalization issues. Also, some biological features, e.g., checkerboard pattern (Sect. 2.3), result in patterns too large be visualized and explored as snippets. Visualization of such large-scale patterns requires further specialized visualization techniques.

9 Conclusion

We have introduced HiPiler—a visual interface that enables the exploration of large genome interaction matrices based on many small ROIs through interactive small multiples. In a user study we found that our proposed snippet visualization approach meets the needs of domain experts and complements existing heatmap-based approaches. The tasks identified in our interviews (see Sect. 2.4) prove to be a valid basis for the design of the HiPiler interface and visualizations. Finally, based on our experience with analysis tasks for other large matrices and image data, as well as the successful evaluation of our approach in the user study, we conclude that HiPiler is very likely generalizable and could be applied to other application domains.

Acknowledgments

The authors wish to thank Nezar Abdennur, Burak Alver, Houda Belaghzal, Aafke van den Berg, Job Dekker, Geoff Fudenberg, Johan Gibcus, Anton Goloborodko, David Gorkin, Maxim Imakaev, Yu Liu, Leonid Mirny, Johannes Nübler, Peter Park, Hendrik Strobelt, and Su Wang. This work was supported in part by the National Institutes of Health (U01 CA200059 and R00 HG007583).

References

[1] N. Abdennur et al. A cool place to store your hi-c, 2017. [Online]. Available: https://github.com/mirnylab/cooler (Accessed: 31-March-2017).

[2] D. Abramov et al. Predictable state container for javascript apps, 2017. [Online]. Available: https://github.com/reactjs/redux (Accessed: 31-March-2017).

[3] B. Bach, N. Henry-Riche, T. Dwyer, T. Madhyastha, J.-D. Fekete, and T. Grabowski. Small multipiles: Piling time to explore temporal patterns in dynamic networks. In *Computer Graphics Forum*, vol. 34, pp. 31–40. Wiley Online Library, 2015.

[4] B. Bach, E. Pietriga, and J.-D. Fekete. Visualizing dynamic networks with matrix cubes. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 877–886. ACM, 2014.

[5] M. Behrisch, B. Bach, N. Henry Riche, T. Schreck, and J.-D. Fekete. Matrix reordering methods for table and network visualization. In *Computer Graphics Forum*, vol. 35, pp. 693–716. Wiley Online Library, 2016.

[6] M. Behrisch, B. Bach, M. Hund, M. Delz, L. Von Rüden, J.-D. Fekete, and T. Schreck. Magnostics: Image-based search of interesting matrix views for guided network exploration. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):31–40, 2017.

[7] R. Cabello et al. Javascript 3d library, 2017. [Online]. Available: https://github.com/mrdoob/three.js (Accessed: 31-March-2017).

[8] J. Dekker. Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture. *Epigenetics & chromatin*, 7(1):25, 2014.

[9] J. Dekker, M. A. Marti-Renom, and L. A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403, 2013.

[10] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[11] K. Dinkla, M. A. Westenberg, and J. J. van Wijk. Compressed adjacency matrices: Untangling gene regulatory networks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2457–2466, 2012.

[12] C. Dunne and B. Shneiderman. Motif simplification: improving network visualization readability with fan, connector, and clique glyphs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3247–3256. ACM, 2013.

[13] N. C. Durand, J. T. Robinson, M. S. Shamim, et al. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell Systems*, 3(1):99–101, 2016.

[14] R. Eisenberg et al. The aurelia framework, 2017. [Online]. Available: https://github.com/aurelia/framework (Accessed: 31-March-2017).

[15] N. Elmqvist, T.-N. Do, H. Goodell, N. Henry, and J.-D. Fekete. Zame: Interactive large-scale graph visualization. In *Visualization Symposium, 2008. PacificVIS'08. IEEE Pacific*, pp. 215–222. IEEE, 2008.

[16] N. Elmqvist, N. Henry, Y. Riche, and J.-D. Fekete. Melange: space folding for multi-focus interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1333–1342. ACM, 2008.

[17] J. Fraser, I. Williamson, W. A. Bickmore, and J. Dostie. An overview of genome organization and how we got there: from fish to hi-c. *Microbiology and Molecular Biology Reviews*, 79(3):347–372, 2015.

[18] P. Fraser and W. Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143):413–417, 2007.

[19] J. H. Gibcus and J. Dekker. The hierarchy of the 3d genome. *Molecular cell*, 49(5):773–782, 2013.

[20] S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit. From visual exploration to storytelling and back again. In *Computer Graphics Forum*, vol. 35, pp. 491–500. Wiley Online Library, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/123588; this version posted April 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[21] B. He, C. Chen, L. Teng, and K. Tan. Global view of enhancer–promoter interactome in human cells. *Proceedings of the National Academy of Sciences*, 111(21):E2191–E2199, 2014.

[22] N. Henry, J.-D. Fekete, and M. J. McGuffin. Nodetrix: a hybrid visualization of social networks. *IEEE transactions on visualization and computer graphics*, 13(6):1302–1309, 2007.

[23] D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458, 25 Mar. 2016.

[24] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[25] P. Kerpedjiev, N. Abdennur, F. Lekschas, C. McCallum, K. Dinkla, H. Strobelt, J. M. Luber, S. B. Ouellette, A. Ahzir, N. Kumar, J. Hwang, B. H. Alver, H. Pfister, L. A. Mirny, P. J. Park, and N. Gehlenborg. Higlass: Web-based visual comparison and exploration of genome interaction maps. *bioRxiv*, 2017. doi: 10.1101/121889

[26] P. Kerpedjiev et al. Fast contact matrix visualization for the web, 2017. [Online]. Available: https://github.com/hms-dbmi/higlass (Accessed: 31-March-2017).

[27] T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1):1–6, 1998.

[28] I. Krivega and A. Dean. Enhancer and promoter interactions—long distance calls. *Current opinion in genetics & development*, 22(2):79–85, 2012.

[29] W. Li, D. Notani, and M. G. Rosenfeld. Enhancers as non-coding rna transcription units: recent insights and future perspectives. *Nature Reviews Genetics*, 17(4):207–223, 2016.

[30] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[31] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[32] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[33] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297. Oakland, CA, USA., 1967.

[34] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[35] P. Pirolli, S. K. Card, and M. M. Van Der Wege. The effect of information scent on searching information: visualizations of large tree structures. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 161–172. ACM, 2000.

[36] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[37] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

[38] F. Schreiber and H. Schwöbbermeyer. Mavisto: a tool for the exploration of network motifs. *Bioinformatics*, 21(17):3572–3574, 2005.

[39] L. Seaman, H. Chen, M. Brown, D. Wangsa, G. Patterson, J. Camps, G. S. Omenn, T. Ried, and I. Rajapakse. Nucleome analysis reveals structure-function relationships for colon cancer. *Mol. Cancer Res.*, 3 Mar. 2017.

[40] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2012.

[41] T. von Landesberger, M. Gorner, and T. Schreck. Visual analysis of graphs with multiple connected components. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 155–162. IEEE, 2009.

[42] T. Von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs: state-of-the-art and future research challenges. In *Computer graphics forum*, vol. 30, pp. 1719–1749. Wiley Online Library, 2011.

[43] L. Yang, L. J. Luquette, N. Gehlenborg, R. Xi, P. S. Haseley, C.-H. Hsieh, C. Zhang, X. Ren, A. Protopopov, L. Chin, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153(4):919–929, 2013.

[44] G. G. Yardımcı and W. S. Noble. Software tools for visualizing hi-c data. *Genome Biology*, 18(1):26, 2017.