invis: Exploring High-Dimensional RNA Sequences from In Vitro Selection

Çağatay Demiralp* Stanford University
Eric Hayden† Stanford University
Jeff Hammerbacher‡ Cloudera & Mt. Sinai Medical School
Jeffrey Heer§ Stanford University

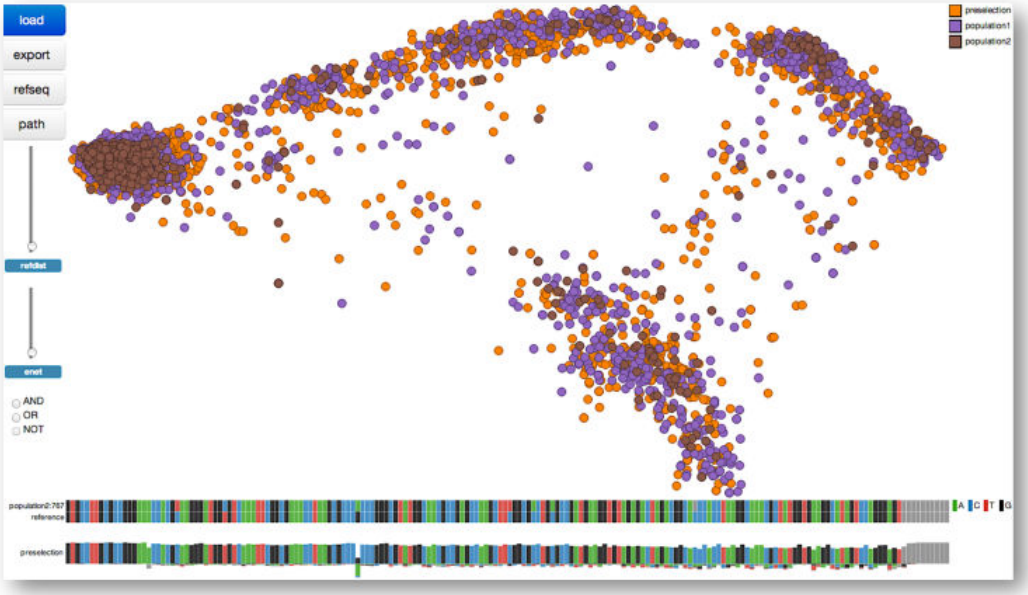


Figure 1: The *invis* interface. The scatterplot shows a projection of high-dimensional RNA sequences; sequence maps enable inspection of nucleotide-level patterns. *invis* is the first tool for interactive exploration of sequences from *in vitro* selection experiments.

ABSTRACT

In vitro selection and evolution is a powerful method for discovering RNA molecules based on their binding and catalysis properties. It has important applications to the study of genetic variation and molecular evolution. However, the resulting RNA sequences form a large, high-dimensional space and biologists lack adequate tools to explore and interpret these sequences. We present *invis*, the first visual analysis tool to facilitate exploration of *in vitro* selection sequence spaces. *invis* introduces a novel configuration of coordinated views that enables simultaneous inspection of global projections of sequence data alongside local regions of selected dimensions and sequence clusters. It allows scientists to isolate related sequences for further data analysis, compare sequence populations over varying conditions, filter sequences based on their similarities, and visualize likely pathways of genetic evolution. User feedback indicates that *invis* enables effective exploration of *in vitro* RNA selection sequences.

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

1 INTRODUCTION

The study of genetic evolution is an integral part of molecular biology research, which seeks to understand biological processes and sources of diversity on Earth. One important question in molecular biology and biochemistry has been whether genetic research can be possible without the use of living organisms [7]. *In vitro* selection is a confirming response to this question, an experimental method for scientists to biochemically (i.e., synthetically) simulate the Darwinian selection and evolution at the molecular level, in "fast-forward", under controlled conditions. With rapid cycles of growth and selection in a cell-free test tube, *in vitro* selection allows scientists to synthetically create large pools of RNA or DNA sequences (libraries) in which they then search for functional molecules. These large libraries can contain as many as 10^{15} unique sequences [36].

Since the sequences from *in vitro* selection are nucleic acid chains, they can be analyzed using high-throughput sequencing (so-called "next generation" sequencing). This enables the detection and genotyping of millions of individual molecules with current technology. Importantly, it allows researchers to assess questions of how functional sequences are distributed in the space of all possible sequences, and how this distribution enables the evolution of novel forms in biology.

However, sequence spaces from *in vitro* selection experiments are large and high dimensional by design, making exploration and interpretation difficult. Visual analysis of these datasets is typically limited to sequence browsing tools and static plots. While useful, these fall short of facilitating basic analysis tasks such as understanding how environmental changes affect selection, exploring the direction of genotypic change due to selection, identifying important mutations and sequences, and exploring possible pathways of evolution from an initiating reference sequence.

In response we present *invis*, the first visual analysis tool that supports interactive exploration of *in vitro* RNA selection sequence spaces (Figure 1). *invis* integrates a novel configuration of coordinated views, enabling simultaneous inspection of global projections of sequence data alongside local regions of selected dimen-

*e-mail:cagatay@cs.stanford.edu
†e-mail:ehayden@stanford.edu
‡e-mail:jeff.hammerbacher@gmail.com
§e-mail:jheer@cs.stanford.edu

IEEE Symposium on Biological Data Visualization 2013
October 13 - 14, Atlanta, Georgia, USA
978-1-4799-1659-7/13/$31.00 ©2013 IEEE

sions and sequence clusters. It supports data analysis by allowing scientists to isolate related sequences for further analysis, compare sequence populations over varying conditions and time, filter sequences based on their similarities, and also visualize likely pathways of sequence evolution.

invis leverages an under-utilized approach to population-scale genomic data visualization. Existing tools focus on visualization of genomic features such as correlations (e.g., linkage disequilibrium), variant histograms and allele frequencies. In contrast, *invis* directly visualizes the sequence space and integrates these views with feature space visualizations. User feedback indicates that *invis* uniquely supports exploration of *in vitro* RNA selection sequences. Crucially, it frees scientists from complex computational work, helping them focus on biological questions.

The rest of the paper is structured as follows. We first review relevant biology to provide context and to help readers without a background in biology. We then discuss related work in visualization of biological sequences and multidimensional data. Next, we describe the design of *invis*, including task analysis, visual encoding, and interaction. We present initial feedback on *invis* through a pair of use cases. Finally, we conclude with a discussion of future applications and larger opportunities for population-scale sequence visualization.

2 Biological Background

Cells are the building blocks of life on earth. Within every cell is a long chain-like molecule called deoxyribonucleic acid (DNA), which encodes the information required for the cell to function. DNA is composed of four types of molecules called nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T). A string of these bases forms a DNA molecule and is called a sequence. Ribonucleic acid (RNA) has a molecular structure similar to DNA and has the same nucleotide bases with the exception of uracil (U) replacing thymine. DNA exerts control on a cell and sustains its function in a continual flow of encoded information: regions of DNA are transcribed into RNA, and RNA is translated to produce proteins. Transcribed RNAs and proteins then perform the work of the cell. Dubbed the *central dogma* by Francis Crick, this flow of information is the fundamental thesis of modern biology and is common to all cells and organisms, despite their great diversity.

DNA sequencing is the process of deducing the order of nucleotides in a DNA molecule. Analysis of DNA sequences is important for understanding biological processes, bases of diseases, and variation between and within species. The first two major sequencing methods were Maxam-Gilbert [22] and Sanger (chain-termination method) [28] sequencing. As sequencing long strands of DNA with these methods has proven impractical, shotgun sequencing methods cut long strands into smaller, overlapping fragments. These overlapping fragments are sequenced using the chain-termination method and reassembled to obtain the sequence of the initial long strand [23]. Shotgun sequencing and related variants enable large-scale genome sequencing, and are critical enabling technologies behind the Human Genome Project [9]. High-throughput RNA sequencing is typically done by first enzymatically copying the RNA sequence to the corresponding DNA sequence and then performing DNA sequencing.

Biochemical methods have played a central role in the molecular biology revolution of the last fifty years. Whether genetic research is possible without the use of living organisms has been an important question in biology and biochemistry [7]. In the 1960s, researchers showed that Darwinian evolution could indeed operate in cell-free tubes, exploiting the fact that RNA both stores genetic information (genotype) like DNA and catalyzes chemical reactions like proteins (phenotype). Note that the primary goal of genetic analysis is to associate genotype and phenotype. Early *in vitro* experiments eventually became practical with the development of *in vitro* replication and the invention of polymerase chain reaction (PCR) combined with improvements in techniques for isolating, directing, and searching molecules in a large collection of sequences.

In vitro selection is essentially an experimental method for discovering rare functional RNA or DNA molecules contained within large pools of sequences. The technique relies on the fact that 1) biological functions change as biological sequences change, 2) rare functions can be found if enough sequences are tested, and 3) rare functional sequences can be isolated and increased in frequency to a more easily detectable level. The process begins by building a pool of sequences. These starting sequences can be constructed by making random nucleotide changes to sequences found in nature, or by synthesizing non-natural or completely random collections of sequences.

Next, sequences are separated based on their ability to catalyze a chemical transformation (*ribozymes*) or bind a specific target molecule (*aptamers*). Several strategies can be used to separate (select) the desired molecules. Two common strategies are self-modification and immobilization. For the self-modification strategy, a functional RNA or DNA molecule that performs the desired task changes in a detectable way. For example, the molecule could become smaller (self-cleavage function) or larger (ligation function), or become attached to a visible molecule such as a fluorescent or radioactive molecule. For the immobilization strategy, typically a target molecule is chemically attached to a solid support, and the pool of RNA or DNA is allowed to bind to this, then rinsed. RNA or DNA that bind to the target molecule stay, and those that cannot are rinsed away. The bound RNA or DNA molecules are then released by changing the rinsing conditions. We refer readers to [36] for further details on *in vitro* selection.

3 Related Work

We now discuss two strands of prior work relevant to *invis*: genomic sequence visualization and multidimensional data visualization.

3.1 Genomic Sequence Visualization

Although sequencing technology has evolved rapidly, the primary visual encoding scheme for DNA sequences has remained the same: linear (or circular) track ideograms.

There are several tools for browsing individual genomes and related data. Both web-based and desktop genome browsers typically use linear track representations of genomes and associated data. Web-based browsers such as the UCSC Human Genome Browser [18] and Ensembl [32] are typically used for locally inspecting, analyzing, and comparing genomic features. These tools often come with curated annotations and they function like archival repositories, allowing researchers to validate and compare their findings with the existing literature. Existing web-based browsers are heavyweight and can be cumbersome for basic viewing interactions. In response, several desktop genome browsers address some of the performance limitations of current web-based browsers. These include IGV [26], GenomeView [2], Savant [12] and Artemis [27].

Comparative analysis is an important part of genomic research. Quantifying what is conserved and what has changed within and across genomes helps researchers understand biological processes as well as sources of normal or abnormal variation. Several tools have been developed using chromosome-wise track representations to provide a global picture of structural variants or rearrangements in genomes [19, 24, 34]. Circos [19] has become increasingly popular in the genetics community, particularly for generating high quality figures for publication; however, the generated visualizations lack interactivity. Cinteny [34] uses chromosome-wise linear ideograms for visualizing synteny: the conservation of genetic loci between chromosomes of two genomes. Neither Circos nor Cinteny supports multi-scale exploration of genomic features. In contrast, MizBee [24] is an interactive, multi-scale genome synteny browser

that uses both circular and linear track representations. For a thorough discussion of genome visualization tools, we refer readers to a review by Nielsen et al. [25].

The work discussed above is a representative set of tools designed for viewing a small number of long sequences (typically only one). There are also visualization tools that support population scale analysis at some level (e.g., [12, 5]). These tools, however, invariably focus on visualization of population feature space, such as linkage disequilibrium, variant histograms, and allele frequencies. Despite the increasing number of large-scale genomic datasets, we lack interactive visual analysis tools that allow users to explore populations in the sequence space. This task is currently performed by computing high-dimensional projections using tools such as Matlab and R, and then looking at static plots.

invis differs from other sequence visualizations by allowing both local and global interactive exploration of sequence populations. Unlike previous work, *invis* directly depicts the sequence space.

3.2 Multidimensional Data Visualization

Genomic sequences are long strings of letters representing nucleotide bases. In this sense, visualization of RNA sequence populations is essentially a large-scale, multidimensional data visualization problem. Multidimensional data are ubiquitous across different domains, from genomics to finance. The general goal is to understand the relations within and between dimensions as well as the structure of the high-dimensional space determined by these relations.

Visualization techniques for multidimensional data include scatterplots [8], scatterplot matrices [10], parallel coordinates [15], radar plots, dense pixel displays [17], and dimensional stacking [20]. Since these visualizations have limited use as static graphics, many tools for multidimensional visualization (e.g., [3, 4, 33, 35]) also include interaction techniques such as filtering, zooming, and brushing & linking.

invis uses a scatterplot to provide a global view of sequence populations. Planar 2D layouts of high-dimensional data can be obtained by interactively choosing two dimensions to show, or by applying dimensionality reduction techniques to project the data to 2D. A number of earlier research projects use low-dimensional projections for exploratory data analysis (e.g., [16, 29, 30]).

invis builds on this prior work in its basic visual representations and interactions. However, it differs in using a novel configuration and coordination of these representations and interactions to facilitate simultaneous exploration of global sequence space along with nucleotide bases of individual and aggregated sequences.

invis also introduces a new dynamic filtering technique that extracts "epsilon" neighborhoods of points based on distances in the original, unprojected space. Our approach has several advantages. First, it allows multi-scale exploration of the data in manner that is more robust to noisy data. Second, it enables the user to assess the accuracy of the dimensional projection. Third, it helps users test hypotheses regarding evolutionary paths between data points under different scales of connectivity.

4 Data Preparation

The initial reference sequences used in *in vitro* selection experiments can be constructed by randomizing sequences found in nature or by synthesizing non-natural sequences. Usually, sequences are selected based on their ability to catalyze a chemical transformation (ribozymes) or bind a specific target molecule (aptamers).

We obtained the data used in this paper from *in vitro* selection of a catalytic activity derived from intronic self-splicing, a natural process. The starting sequence was an RNA intron that catalyzes its own excision from a tRNA in a bacterium. This activity can be selected *in vitro* because RNA molecules capable of catalyzing this reaction can ligate a short piece of a substrate oligonucleotide to their own terminus in a cell-free biochemical reaction. Random

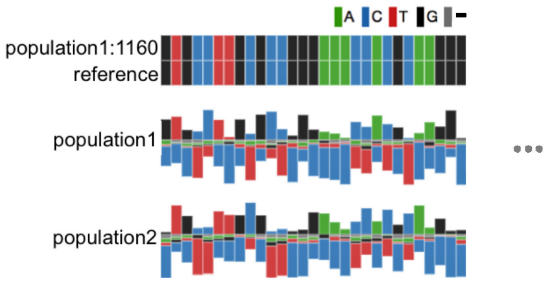


Figure 2: Sequence maps in *invis*. The initial heat map compares a currently selected sequence (top) with a reference sequence (bottom). The other sequence maps use stacked bars to visualize the frequencies of the four nucleotide bases within an RNA population. The frequency of nucleotides matching the reference sequence value at a given position is visualized with a bar stacked above the baseline. Frequencies for the remaining nucleotides are stacked below the baseline.

nucleotide changes were made to the starting sequence to produce a population of sequences with an average of six nucleotide changes per molecule. Samples of this population ($\sim 10^{12}$ molecules) were allowed to react with a substrate oligonucleotide, and functional sequences were retained.

Functional sequences were isolated in three separate reactions that had different concentrations of magnesium (2 mM, 10 mM and 25 mM). Magnesium stabilizes the folded structure of active molecules, and increasing concentrations of magnesium should allow molecules with destabilizing mutations to remain functional. The nucleotide sequence of the retained functional molecules were identified through next generation sequencing (Roche 454 platform). Several thousands of individual molecules (i.e., "reads" or sequences) were obtained for each magnesium concentration.

We aligned these reads with the reference read (wild type) using a pairwise alignment package [13] and filtered out the sequences whose length was outside the range $[l-0.5\sigma, l+0.5\sigma]$, where l is the length of the reference sequence and σ is the standard deviation of the lengths of all the sequences in the experiment. After this initial processing, we had about 19K sequences of length 186 for the four populations (pre-selection, 2 mM, 10 mM and 25 mM) combined. We numerically encoded the sequences as A=1, G=2, T=3, C=4 and gaps (denoted with $-$) with 0. For example, the sequence `ACG--TA-` became a vector of values $(1,4,2,0,0,3,1,0)$. Next, we computed the pairwise Hamming distance between each sequence to use in the filtering interactions discussed in Section 5.2.3. We finally computed a planar projection (embedding) by running principal component analysis (PCA) on the numerically encoded sequence data [14].

We observed that using different numerical encodings, including binary representations where all the bases have the same numerical "weight", does not affect the structure of projection significantly. Similarly, using multidimensional scaling (MDS) [14] on the Hamming distance matrix provides a projection very similar to the projection obtained using PCA. We further discuss the effects of numerical encoding on two-dimensional layout in Section 6.

5 The Design of *invis*

In our collaboration with biologists, we identified four high-level analysis tasks that they would like to perform on the sequenced RNA data from *in vitro* selection experiments:

1. Separate functional from non-functional sequences.
2. Identify individual mutations, groups of mutations, or entire sequences that are frequent within the functional population.
3. Recover unique, independent solutions to the function by subdividing the functional population.
4. Reconstruct evolutionary trajectories between sequences.

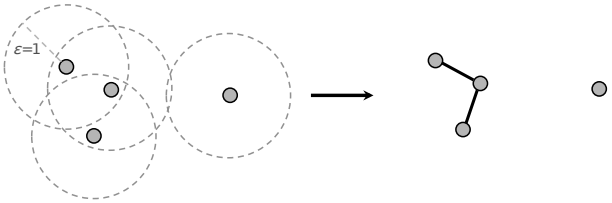


Figure 3: ε-graph construction with $\varepsilon = 1$ (ε_1-graph). An ε-graph of a sequence collection is a graph obtained by putting an edge between every sequence within a distance of ε from each other.

We now discuss the design of visual representations and interaction techniques in *invis* to support these biological data analysis tasks.

5.1 Visual Encoding

We use three basic visual representations in *invis*: scatterplots of projected sequences and two forms of "sequence map". We use heat maps for one-to-one sequence comparison, and stacked bars for showing empirical nucleotide frequencies.

5.1.1 Scatterplots of Sequence Populations

invis uses scatterplots to give a global overview of the sequence space of populations (Figure 1). By using a position encoding, the strongest of visual cues, scatterplots allow viewers to examine both clusters and outliers. Each sequence is drawn as a circular node in the plane and filled with a color identifying its population of origin. The layout is obtained using PCA, which provides a planar approximation to the high-dimensional sequence space and its structures.

5.1.2 Sequence Maps

We developed two types of "sequence map" to convey data at the level of individual nucleotide bases: a heat map for one-to-one sequence comparison and stacked bars for comparing population-level base pair frequencies. The primary function of the heat map is to provide on-demand access to any individual sequence in the populations and enable pairwise comparison with the reference sequence (Figure 2).

invis uses stacked bars to represent difference statistics over populations with respect to the reference RNA sequence (Figure 2). The frequency of nucleotides matching the reference sequence value at a given position is visualized with a bar graph stacked above the baseline. Frequencies for the remaining nucleotides are visualized stacked below the baseline. The goal is to convey both the nature and magnitude of differences between the reference and population sequences. The stacked bars are useful for discovering interesting mutation locations and their variance between sequence populations.

5.2 Interaction Techniques

Although low-dimensional projections can be useful for conveying the structure of the sequence space, by construction they are lossy representations. *invis* compensates for this fact using coordinated heat map views of individual sequences and dynamic filtering. *invis* also provides selection aggregation and shortest path construction methods for exploring patterns and relations among subsets of sequence data.

5.2.1 Brushing & Linking

We use brushing & linking in *invis* to coordinate the contents of the sequences represented in the scatterplot with the data shown in the sequence maps. This is the main mechanism that allows users to inspect global sequence space and individual sequences simultaneously. Hovering over a location on the reference sequence heat map highlights all sequences that are different from the reference at that location. Brushing on nodes in the scatterplot updates the current sequence heat map.

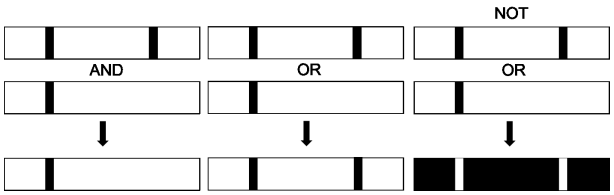


Figure 4: Aggregation of selected sequences. Mutations (indicated with black bars) can be aggregated using the logical operators AND, OR, and NOT. The NOT operator is applied after any AND or OR operations are applied to the current selection.

5.2.2 Selection Aggregation

invis allows users to interactively aggregate mutations of selected sequences using logical operators. This can be useful for discovering mutation patterns persistent across groups of sequences (Figure 4). In our current implementation, the AND operator distinguishes mutation types and can help reveal beneficial (functional) mutations. Conversely, the OR operator is agnostic to the type of mutation, but it is possible to use an additional visual track of stacked bars to show the distribution of the aggregation. Combined with NOT, OR (i.e., NOR) is particularly useful for finding regions conserved across sequences.

5.2.3 Filtering

invis provides two basic dynamic filtering interactions based on Hamming (mutation) distances: distance-to-reference filtering and ε-filtering. Distance-to-reference filtering lets users to show and hide sequences based on their mutation distances to the reference sequence. This allows users to interactively explore level sets of the distance field determined by the distance from the reference sequence (Figure 5).

ε-filtering enables users to explore different "ε-graphs" of the data. An ε-graph is obtained by adding an edge connecting any pair of data points that are within a distance of ε, where ε is an interactively-adjustable parameter (Figure 3). We denote $\varepsilon = k$ also by ε_k. During ε-filtering, *invis* identically colors all sequence points that are part of the same connected component of the current ε-graph. This encoding avoids the visual clutter of drawing explicit edges while still conveying connected clusters.

There are three primary advantages to dynamic ε-filtering: First, it enables users to query the native similarity space of sequences in a multi-scale fashion. When $\varepsilon = 0$ we have a collection of singletons and when ε is sufficiently large then we have a complete graph. The ability to see structures coalescing, splitting, or persisting across scales helps to better understand the structure of the space and discover outliers. Second, it helps to validate the accuracy of the projection used. For example, if two sequences are close to each other in the mutation space they can sometimes be put far apart on the plane due to projection error. Together with distance-to-reference filtering, ε-filtering makes it easier to diagnose such anomalies (Figure 6). Third, ε-filtering helps users test connectivity hypotheses (of evolutionary paths, for example) between data points under different distance scales.

5.2.4 Evolutionary Paths

After identifying functional sequences, an important next task is to understand how those sequences evolve, particularly from the reference sequence. Biologists are particularly interested in single mutations (i.e., ε_1-graphs) because adaptation of beneficial random, single mutations is an important mechanism through which evolution occurs (Figure 7). In *invis*, given an ε-graph of the sequence space, a user can compute shortest paths from the reference sequence to other sequences.

The user can inspect sequences along a selected path through brushing & linking or with a "play-the-path" interaction. The play-

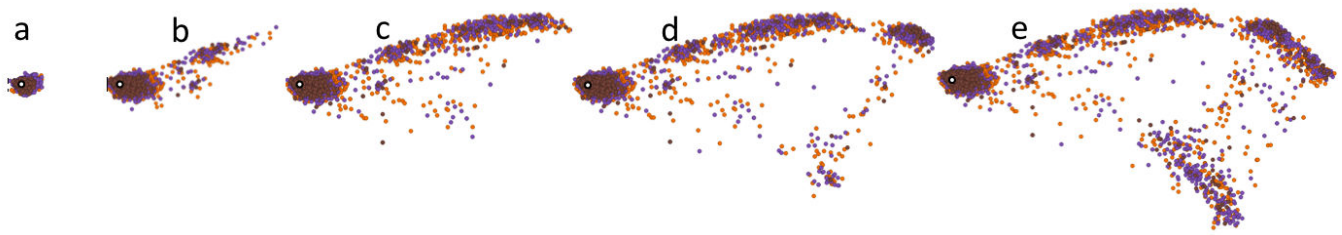


Figure 5: Distance-to-reference filtering allows users to explore the direction of selection. (a-e) Collection of sequences within increasing distances (5, 36, 79, 95, 150) from the reference (the white node).

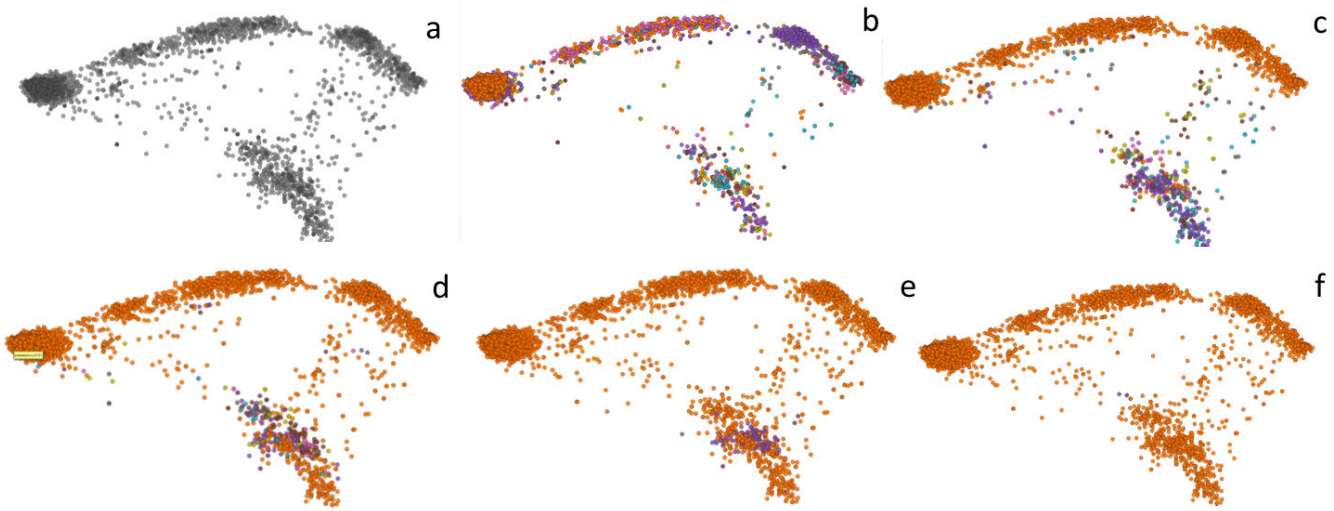


Figure 6: (a-f) ε-graphs of *in vitro* selection populations obtained using increasing ε values (0, 9, 18, 22, 68, 80). Graph edges are not displayed but connected components are colored differentially. It is easy to see the regions where the projection is not accurate. For example, (e) shows two clusters of sequences (in orange and violet, respectively) that overlap in the plane but are disconnected for almost all ε values.

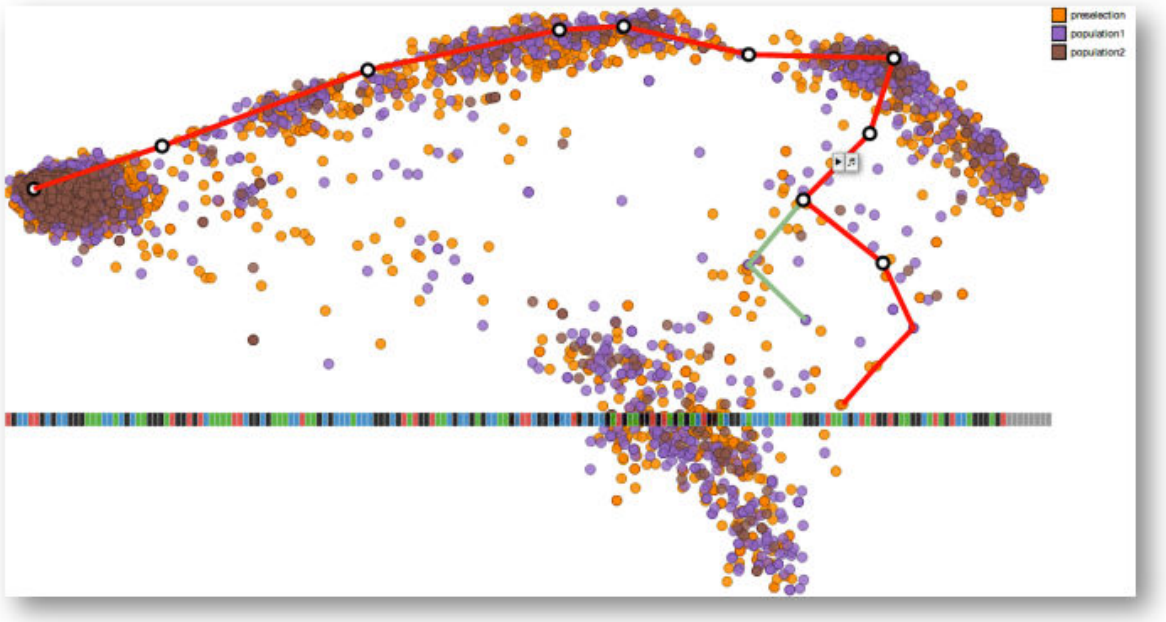


Figure 7: Shortest paths from the reference (the white node) to two sequences in the ε_{20}-graph of the sequence space. In addition to the brushing & linking interaction, a user can inspect sequences on a selected shortest path (highlighted as red) with a "play-the-path" interaction. The play-the-path interaction first brings up the sequence heat map into the user's view and then animates the sequences from start to end, highlighting them in order while updating the heat map with their contents.

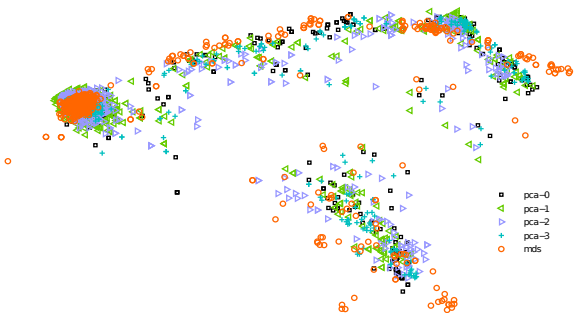


Figure 8: Two-dimensional layouts obtained using PCA and MDS with different numerical encodings of sequence bases. The overall structure of the layouts are very similar.

the-path interaction first brings up the sequence heat map into the user's view and then animates the sequences from start to end along the path, highlighting them in order while updating the heat map with their sequence contents.

6 Numerical Encoding of Sequences

Genomic sequences are often numerically encoded for computational processing. We use the simple arbitrary encoding scheme discussed previously, then apply PCA to obtain a two-dimensional layout of RNA populations. While *invis* is independent of the particular choice of encoding scheme or embedding algorithm, it is important to understand any significant effects of different encoding schemes and embedding algorithms on the two dimensional layout.

Here, we compare layouts computed using different numerical encodings of the sequence population selected with the 25 mM magnesium concentration. In addition to our default numerical encoding, we consider three other encodings obtained by randomly permuting the default. The four encodings that we evaluate are:

0. A=1 G=2 T=3 C=4 gap=0 (default)
1. A=4 G=1 T=0 C=2 gap=3
2. A=0 G=4 T=2 C=3 gap=1
3. A=10000 G=01000 T=00100 C=00010, gap=00001 (binary)

For comparison, we also use MDS to perform the projection. We use Hamming distance as a similarity measure. As Hamming distance does not differ across numerical encodings, we only compute the MDS layout once. Figure 8 shows the five layouts derived from these cases. We align the computed layouts up to a similarity transform (uniform scaling, rotation, and translation). We observe that the overall structure of the projected population space is very similar across the encodings and projection (embedding) methods used.

Next, we statistically verify what we see. We compute a pre-alignment distance matrix for each layout using Euclidean distance in the plane. We then compute Spearman's rank correlation coefficients between the distance matrices and determine their statistical significance with Mantel's permutation test [21]. We also include the distance matrix (D_h) containing Hamming distances between unprojected sequences in the comparison. Results show that all five layouts are highly correlated with each other and with the distance matrix of sequences (Table 1). All correlations are significant at $p < 0.001$.

Why is there no significant effect of numerical encoding on the layout produced? Though answering this question conclusively requires further investigation, one explanation is the correlations in the *in vitro* selection population. To start with, the sequences of the population originate from the same reference and were selected based on having the same functional properties in the *in vitro* experiment. Also, they are aligned to the reference sequence, which further endows structure on their space. Therefore, the structure of the sequence space determined by the correlations of sequence dimensions likely dominates any effects of numerical encoding, especially when the number of data points is large. Conversely, we expect to see more pronounced effects of numerical coding when sequences have uniformly low or no correlations (i.e., uniformly random sequences).

	pca-0	pca-1	pca-2	pca-3	mds	D_h
pca-0	1.000	0.7441	0.7087	0.7831	0.7273	0.7734
pca-1		1.000	0.8145	0.8362	0.7480	0.7689
pca-2			1.0000	0.8379	0.7578	0.7642
pca-3				1.0000	0.8516	0.8097
mds					1.0000	0.7766
D_h						1.0000

Table 1: Correlations among the distance matrices from four different numerical encodings and two embedding methods and the original distance matrix. All the statistics have the same p-value of 0.001.

7 Usage Examples

As a preliminary evaluation of *invis*, we observed multiple users apply the tool for biological data analysis. In the first case, we observed a non-expert in biology to see what they discover using the tool. In the second case, a biologist used *invis* to analyze his experimental data.

7.1 Initial Exploration by a Non-Expert

We asked a (non-biologist) member of our research group to use *invis* to explore sequence data and observed his actions. The user first examined the projection view and familiarized himself with the brushing and linking features. He then examined the projected sequences. Starting from the projection view, the user identified the reference sequence and examined the sequences clustered around that point. He found that the dense cluster around the reference consists of sequence variants with one or more point mutations, but overall good alignment without any nucleotide insertions or deletions.

The user then looked at the "trail" of sequences extending off to the right from the reference, along the top of the display. The user discovered that these points consist of sequences that have a single nucleotide insertion (in addition to point mutations). As one moves further along the trail, the insertion point occurs earlier in the sequence, inducing greater distance from the reference sequence.

Next, the user turned his attention to the group of points at the bottom of the projection view. After hovering over various points and examining the underlying nucleotide sequences, he surmised that this grouping is ill-defined. Both multiple nucleotide inserts and nucleotide deletions appear to occur. The user hypothesized that this grouping actually consists of multiple clusters that have been projected down onto the same region of the plane.

To test this hypothesis, the user enabled ε-filtering. By adjusting the ε value, the user observed a separation among the sequences in the bottom grouping (as in Figure 7). Unlike the other groups in the projection views, the bottom group has sub-collections that are more distant from each other. It appears that the bottom groupings consists of separate clusters that have been projected on top of each other.

Finally, the user spent some time focusing on the population sequence comparisons at the bottom of application window. Examining the sequence views, the user did not discover any major differences between the different RNA populations. He noted some minor variations of potential interest (about two-thirds into the sequence length), but for the most part the populations looked quite similar. This conclusion was further strengthened by examining the projection view and noting that the different colored points representing different populations project onto the overlapping clusters with similar densities.

In each of these explorations, the user made heavy use of the projection view and sequence view in tandem, fluidly switching be-

tween global and local views to make sense of the data.

7.2 Analysis by a Biological Researcher

Our biologist collaborator is interested in discovering functional RNAs that can control gene expression in response to internal and external cellular signals. In a set of recent *in vitro* selection experiments, he wanted to learn how changing environments might affect the distribution of a biological function in sequence space. He used *invis* to explore a dataset collected from these experiments. The dataset (described in Section 4) contained a pre-selection population of reads and three populations that were isolated (selected) using three separate reactions with different magnesium concentrations. Each population had several thousands of individual reads.

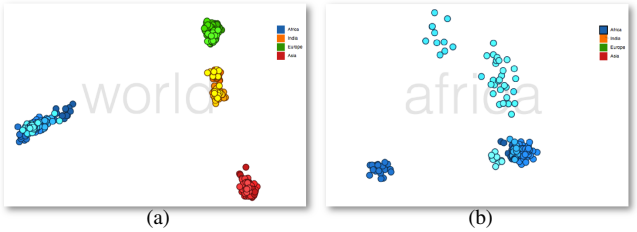
Upon initial loading of the dataset, *invis* provided an immediate qualitative assessment. The collaborator observed that all data sets, including the pre-selection sample, had similar projected distributions. From this, he concluded that the quality of each data set was comparable because none of the samples had a major difference which could be caused by an artifact in a single experiment. For example, short inactive sequences can often escape selection because they are more easily amplified. These types of artifacts can often occur during *in vitro* selection experiments, and detecting them is important. Next, he concluded that the difference between the distribution of sequences under the different conditions must be subtle. This was expected because in this set of experiments, all selected sequences were expected to adopt similar folded structures. This is in contrast to some selections that begin from completely random sets of sequences, which can result in multiple separate solutions to a biological function.

One surprising observation from this qualitative perspective was that the sequence space of the selected molecules overlapped entirely with the pre-selection data. Stated another way, the functional sequences permeate sequence space as far as the experiment explored. This is in agreement with computational studies that have shown that RNA secondary structures form vast neutral networks in sequence space. Further, using ε-filtering, the collaborator explored which sequences could be connected through a series of successive individual mutations. He was excited to discover that many sequences that were far apart could be connected through a series of point mutations. This observation supports the hypothesis that evolution could explore the distant parts of genotype space without losing the required function.

Next, the collaborator searched for patterns in genotype space. He noticed that within close proximity to the reference sequence, the data was rather uniformly dispersed, and occupied a circular region. However at increasing distances, the projection of the data became more sparse, so that a long thin region of sequence space was occupied. He noticed that this was true for both the pre-selection and selection data, which indicated a bias in the randomization procedure. He conjectured that the vast number of sequences possible at higher distances necessarily results in a statistical sampling (bottlenecking) which could have lead to the uneven distribution at higher distances.

The collaborator proceeded to explore the data at the nucleotide level. He found it useful to be able to search for common mutations that occurred in different regions of genotype space. These mutations are expected to have functional significance even if the rest of other parts of the sequence are changed. This process was further facilitated by the ε-filtering, which highlights similar sequences.

Our collaborator found *invis* to have several desirable features. From the perspective of the collaborator, the tool had several desirable features. First, it enabled very easy data exploration. Because populations could be interactively added and removed from the visualization, similarities and differences between populations were easily identifiable. With large numbers of data points, the difference between two data sets was most easily observed by keeping one set visible, and adding and taking away another, repeatedly.



(a) (b)

Figure 9: (a) SNP variation between subjects sampled from 27 populations. This variation is highly correlated with ethnicity and geography. (b) Samples from Africa; substructures of genetic variation appear (see the interactive visualization at `http://graphics.stanford.edu/~cagatay/genome/embed.html`.)

This is easily accomplished with our tool, but not with traditional programs. Additionally, the "sequence space" perspective of the tool enabled comparisons that are normally inaccessible. For example, the collaborator could see if different selection conditions caused selected sequences to become more or less spread out, or if clumps of sequences appeared in different regions. These regions of similarity or difference could then be further analyzed at the individual sequence level, without changing the global perspective. Our collaborator values the ability to maintain a sequence space perspective even while focusing in on individual sequences. Individual sequences could be analyzed while keeping the relative distance to the reference sequence in mind.

8 Discussion

We have presented *invis*, the first visual analysis tool to facilitate exploration of large, high-dimensional *in vitro* RNA selection datasets. *invis* provides a novel configuration of coordinated views, enabling simultaneous inspection of global projections of sequence data alongside local regions of selected dimensions and sequence clusters. *invis* also contributes several interaction methods for aggregating, filtering, and linking sequences of populations based on their similarity. We implemented *invis* in JavaScript using D3 [6]; source code is available at `https://github.com/StanfordHCI/invis`.

invis works directly in the sequence space of RNA populations from *in vitro* selection, yet could be easily combined with additional feature space visualizations. This approach makes it particularly suitable for exploratory visual analysis. While we have focused on *in vitro* selection in this paper, the design of *invis* can be applied to visualize other population-scale genomic datasets. We now discuss two examples.

We are currently working to apply *invis* to visualize the EteRNA player solution space. EteRNA is a web-based game with the goal of uncovering the rules of RNA design [1]. Players are given RNA secondary structures and asked to find the sequence that folds into the given RNA structure. They then submit their solutions, the most promising of which are synthesized and probed *in vitro* using chemical mapping. These experimental results are returned to players, who then modify and resubmit their solutions. This process can be seen as another type of "evolution" for sequences that have a particular phenotype. In this case, it is human-guided instead of *in vitro* selected.

Single nucleotide polymorphisms (SNPs) are used in population genetics to study sources of genetic variation. Figure 9 shows snapshots from our visualization of a SNP dataset [37] containing around 250,000 SNPs genotyped in more than 500 individuals from 27 different populations. We compute PCA on the whole dataset as well as separately on the populations coming from the same continental region. The latter provides a more informative zoom-in when the user would like to see variation within a single continental region. One immediate insight from this visualization is that genetic variation, at least at the SNP level, is strongly correlated with eth-

nicity and geography.

There are also challenges ahead. RNA reads of *in vitro* experiments are short and can be displayed as heat maps without hitting display size limits. On the other hand, visualizing 3 billion DNA base pairs or millions of SNP locations in the same way is not possible. We could still use interactive scatterplots representing low-dimensional projections. However, intelligent filtering and aggregation methods, both at the interaction and data representation levels, are needed to enable flexible exploration of nucleotide-level details.

Motivated by the unprecedented potential of high-throughput genome sequencing, Sidow argued for a data-driven approach to genomic research in his paper titled "Sequence first. Ask questions later." [31], which aptly summarizes the current research trends in genomics. According to Sidow, in-depth comparative analyses that are based on large amounts of sequence data can transform biomedicine. In fact, the last decade has seen developments to test this vision. Initiatives such as The International HapMap Project, The Cancer Genome Atlas (TCGA) Project, The 1000 Genomes Project, The 1000 Plant Genomes Project, and The Genome 10K are generating large collections of genomic data. Neither have these developments been limited to academic research. Now there are companies providing personal sequencing services. In the process, they collect large amounts of genetic data, creating a fertile ground for a new kind of genome-wide association study (e.g., [11]).

Visual analysis tools, however, have not kept pace with this fast influx of new data in the field. For example, if the goal is to browse and compare a few genomes or genomics features in depth, there are many good tools to carry out the task. But if the goal is to explore thousands, if not millions, of genomic data points, especially without knowing in advance where to look and what to look for, existing tools and techniques quickly become limited as they have not been designed for the task at hand. In order to ask meaningful questions "later", we need new visual analysis tools designed for exploring large, population-scale genomic datasets "now"; *invis* can be seen as one step forward in this direction.

ACKNOWLEDGMENTS

We thank Diana MacLean and Stuart K. Card for their feedback on earlier drafts of this paper.

REFERENCES

[1] Eterna project website. `http://eterna.cmu.edu/web/`.

[2] T. Abeel, T. Van Parys, Y. Saeys, J. Galagan, and Y. Van de Peer. Genomeview: a next-generation genome browser. *Nucleic Acids Res.*, 40(2):e12, 2012.

[3] C. Ahlberg. Spotfire: an information exploration environment. *SIGMOD Rec.*, 25(4):25–29, 1996.

[4] C. Ahlberg and B. Shneiderman. Visual information seeking using the filmfinder. In *Proc. CHI*, pages 433–434, 1994.

[5] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.

[6] M. Bostock, V. Ogievetsky, and J. Heer. D^3: Data-driven documents. *IEEE TVCG (Proc. InfoVis)*, 17(12):2301–2309, 2011.

[7] J. M. Burke and A. Berzal-Herranz. In vitro selection and evolution of rna: applications for catalytic rna, molecular recognition, and drug discovery. *The FASEB Journal*, 7(1):106–12, 1993.

[8] W. S. Cleveland and R. McGill. The Many Faces of a Scatterplot. *Journal of the American Statistical Association*, 79:807–822, 1984.

[9] A. Edwards, H. Voss, P. Rice, A. Civitello, J. Stegemann, C. Schwager, J. Zimmermann, H. Erfle, C. Caskey, and W. Ansorge. Automated dna sequencing of the human hprt locus. *Genomics*, 6(4):593 – 608, 1990.

[10] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE TVCG (Proc. InfoVis)*, 14(6):1141–1148, 2008.

[11] N. Eriksson, J. M. Macpherson, J. Y. Tung, L. S. Hon, B. Naughton, S. Saxonov, L. Avey, A. Wojcicki, I. Pe'er, and J. Mountain. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genetics*, 6(6), 2010.

[12] M. Fiume, E. J. M. Smith, A. Brook, D. Strbenac, B. Turner, A. M. Mezlini, M. D. Robinson, S. J. Wodak, and M. Brudno. Savant genome browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Res.*, 40(W1):W615–W621, 2012.

[13] R. S. Harris. *Improved pairwise alignment of genomic DNA*. 2007.

[14] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2013.

[15] A. Inselberg. Multidimensional detective. In *Proc. InfoVis*, pages 100–107, 1997.

[16] H. Jänicke, M. Böttinger, and G. Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *IEEE TVCG (Proc. InfoVis)*, 14(6):1459–1466, 2008.

[17] D. A. Keim and H.-P. Krigel. Visdb: Database exploration using multidimensional visualization. *IEEE CG & A*, 14(5):40–49, Sept. 1994.

[18] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002.

[19] M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 2009.

[20] J. LeBlanc, M. O. Ward, and N. Wittels. Exploring n-dimensional databases. In *Proc. Visualization*, pages 230–237, 1990.

[21] N. Mantel. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, 27(2 Part 1):209–220, 1967.

[22] A. M. Maxam and W. Gilbert. A new method for sequencing dna. *PNAS*, 74(2):560–564, 1977.

[23] J. Messing, R. Crea, and P. H. Seeburg. A system for shotgun dna sequencing. *Nucleic Acids Res.*, 9(2):309–321, 1981.

[24] M. Meyer, T. Munzner, and H. Pfister. Mizbee: A multiscale synteny browser. *IEEE TVCG (Proc. InfoVis)*, 2009.

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

[26] J. T. Robinson, H. Thorvaldsdttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nat. Biotechno.*, 29(1):24–26, 2011.

[27] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.-A. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–945, 2000.

[28] F. Sanger and A. Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *J. Mol. Biol.*, 94(3):441 – 448, 1975.

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

[30] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proc. InfoVis*, pages 65–72, 2004.

[31] A. Sidow. Sequence first. Ask questions later. *Cell*, 111(1):13 – 16, 2002.

[32] J. Stalker, B. Gibbins, P. Meidl, J. Smith, W. Spooner, H.-R. Hotz, and A. V. Cox. The ensembl web site: Mechanics of a genome browser. *Genome Research*, 14(5):951–955, 2004.

[33] C. Stolte and P. Hanrahan. Polaris: A system for query, analysis and visualization of multi-dimensional relational databases. In *InfoVis*, pages 5–14, 2000.

[34] P. Stothard and D. Wishart. Circular genome visualization and exploration using cgview. *Bioinformatics*, 21:537–539, 2005.

[35] D. F. Swayne, D. T. Lang, A. Buja, and D. Cook. Ggobi: evolving from xgobi into an extensible framework for interactive data visualization. *Comput. Stat. Data Anal.*, 43(4):423–444, Aug. 2003.

[36] D. S. Wilson and J. W. Szostak. In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.*, 68(1):611, 1999.

[37] J. Xing, W. S. Watkins, D. J. Witherspoon, Y. Zhang, S. L. Guthery, R. Thara, B. J. Mowry, K. Bulayeva, R. B. Weiss, and L. B. Jorde. Fine-scaled human genetic structure revealed by snp microarrays. *Genome Research*, 19(5):815–825, 2009.