Gene-RiViT: A visualization tool for comparative analysis of gene neighborhoods in prokaryotes

Adam Price* Robert Kosara† Cynthia Gibas‡

University of North Carolina at Charlotte

ABSTRACT

The genomes of prokaryotes are dynamic and shuffling of gene order occurs frequently, along with horizontal transfer of genes from external sources. Local conservation of gene order tends to reflect functional constraints on the genome or on a biochemical subsystem. Comparison of the local gene neighborhoods surrounding a gene of interest gives insight into evolutionary history and functional potential of the gene. The Genomic Ring Visualization Tool (Gene-RiViT) is a high speed, intuitive visualization tool for investigating sequence environments of conserved genes among related genomes. Gene-RiViT allows the user to interact with interconnected global and local visualizations of gene neighborhoods and gene order, through a web-based interface that is easily accessible in any browser. The primary visualization is a wheel of nested rotating circles, each of which represents a single genome. This visualization is similar to common circular genome alignment views, except that the rings can be realigned with each other dynamically based on user selections within the ring view or one of the coordinated views. By allowing the user to dynamically realign genomes and focus on a locally conserved region of interest, and using orthology connections to highlight corresponding structures among genomes, this view provides insight into gene context and preservation of neighbor relationships as genomes evolve. Visualizations are linked into a coordinated multiple view interface to provide multiple selection methods and entry points into the data. These approaches make Gene-RiViT a flexible, unique tool for examining gene neighborhoods that improves on existing methods.

Index Terms: J.3 [Life and Medical Sciences]: Biology and Genetics;D.1.7 [Software]: Programming Techniques—Visual Programming; H.5.m [Information Interfaces and Presentation]: Miscellaneous—Multi-scale visualization

1 INTRODUCTION

It has long been known that multiple independent genes can be coordinately expressed and behave as a single unit in prokaryotic organisms. These units, called operons, encode functionally linked proteins, with a conserved gene order. It has been shown that gene order within operons is often conserved in both closely related and highly divergent organisms, and can therefore be used for making inferences about genes and their functions [9, 13]. For example, if the function of a single gene in a region of conserved gene order for multiple organisms is known, it is possible to infer information about neighboring genes and transfer functional annotation, even when function is known for only a single organism in the comparison [3, 9]. Conversely, if a gene is found in one genome outside of its previously observed functional context, this may be an indication that there is no longer selection pressure acting to keep components of an operon in their functional order.

Several tools currently exist that allow their users to examine prokaryotic genes in the context of gene neighborhoods [6, 7, 26, 1, 11]. These existing tools are limited in a number of ways. Most lack the capacity for users to use their own data, and are confined to pre-loaded data sets. This may be helpful for studying broad concepts, but can be limiting when specific organisms are being examined and data for those organisms has not been pre-computed. It is also limiting if the user wishes to analyze unpublished data. The existing gene neighborhood analysis tools generally make use of visualizations based on linear alignment to a fixed reference genome. The user queries the database using text-based or pulldown based queries, and then a visualization of a local region is generated. It is not possible to seamlessly move back and forth between the whole genome context and the local neighborhood, or to initiate a query in a comparative data set based on observed relationships in the summary visualization rather than by keyword access to a known region of interest.

The Genomic Ring Visualization Tool (Gene-RiViT) is a web-based visual analytic tool that uses emerging web technologies to provide a coordinated multiple view interface to a comparative genomic database. The current version of Gene-RiViT combines a familiar, global 2D dot plot view with an adaptive local neighborhood view, demonstrating the potential of a visual analytic approach for real time exploration of conserved gene neighborhoods and their genomic context.

2 GENE ORDER

Comparison of gene order in closely related or even highly diverged genomes can suggest the biochemical context of a gene in a system, though conserved gene order does not necessarily provide complete biochemical information [27]. When multiple prokaryotic genomes are being compared in order to understand gene content differences among strains that may lead to differences in pathogenicity, in host preference, or in survival in the environment, the first line of inquiry is often simply to examine genomic similarities and differences [22]. For instance, if a component of an operon frequently associated with pathogenicity, such as the Type IV secretion system used by many bacteria to transport proteins and toxins out of the cell, is found as a differentiating feature between two bacterial strains with different levels of pathogenicity, this is potentially of interest. The next question, after we identify those potentially interesting a differences, is whether the complete Type IV secretion system is present or whether that gene is isolated out of its functional context, perhaps due to a horizontal transfer event or a reshuffling of the genome. The number of prokaryotic genome sequences available is growing rapidly, and comparative studies focusing on identifying core genome features common to a set of genomes, or dispensable features that distinguish them, are common. Gene-RiViT builds on an existing genomic comparative analysis platform [5], adding the capability to dynamically examine the genomic context of these gene discoveries, as well as the gene-level effects of observed insertions, deletions, and inversions on a set of genomes.

*e-mail: aprice67@uncc.edu

†e-mail:rkosara@uncc.edu

‡e-mail:cgibas@uncc.edu

Typical sequence alignment visualizations assume that sequences are collinear, and don't adequately display permutations in gene order, especially when multiple genomes are being compared. At the genome level, however, gene order is commonly rearranged and analysis of these permutations cannot be disregarded in a comparison procedure [24]. Analysis of gene order conservation using gapped local alignments of 25 prokaryote genomes has shown that 5-25% of the genes in bacterial and archaeal genomes belong to gene strings that are shared by at least two of the examined genomes, once closely related species were excluded [27]. Gene-RiViT addresses the pervasive permutation problem associated with analysis of gene order, by providing visualizations that make rearrangements in gene neighborhoods obvious and comparable between multiple genomes.

3 Orthology

In order to compare gene order and identify commonalities among different genomes, it is first necessary to determine orthology relationships between genes [24]. Orthologs are defined to be homologous genes that diverged from an ancestral gene in the most recent common ancestor of the species under comparison, while paralogs are genes that are related by a gene duplication event in an ancestral gene [10]. Co-orthology refers to paralogs produced by the duplications of orthologs subsequent to a given speciation event, a phenomenon which is commonly observed between distantly related species [14]. Inparalogs are paralogs in a given lineage that evolved by gene duplications occurring after a given speciation event [25].

Gene-RiViT uses OrthoMCL to cluster genes by orthology, co-orthology, and inparalology. OrthoMCL identifies orthologous groups from the results of all-against-all BLAST comparisons, identifying reciprocal best hits [21]. This method of ortholog clustering is based on the principle that orthologous genes are the most similar among all compared pairs of genes [21, 27]. OrthoMCL has been shown to outperform other clustering methods in terms of efficiency and accuracy [2], however the modular design of the database that supports Gene-RiViT allows for straightforward substitution of other methods for identifying orthologs as new methods evolve. The use of orthoMCL to prepare genomic data for analysis in Gene-RiViT allows any set of genomes to be compared to one another, regardless of whether or not they have been made publicly available or incorporated into existing orthology databases such as Clusters of Orthologous Groups of proteins(COG) [24].

4 Gene-RiViT

Gene-RiViT uses three main modules to process data and provide it to the user: an OLAP database, a custom-built web server, and the client-side visualization. Each of these modules is designed to be scalable and to allow for fast interaction with genome-scale data. Gene-RiViT incorporates multiple coordinated views that use state of the art web technologies to create a dynamic, visually appealing and intuitive interface that provides researchers with the ability to contrast the relationships between genes. The publically available interface requires no setup and is capable of visualizing any combination of prokaryotic genomes. Users can also set up Gene-RiViT to run on local systems with relatively little setup. We currently provide support for importing any genome available in the EMBL database, however, the GenoSets back-end which supports the Gene-RiViT system also supports import and annotation of unpublished genome data.

4.1 Web Server

Gene-RiViT relies on custom-built middleware that functions as a web server and data processing hub between the database and the visualization. This module of Gene-RiViT was developed using Node.js, an efficient and scalable platform for data-intensive, real-time applications [15]. As the amount of genomic data processed can be quite substantial, it was important to address the issue of network latency when designing Gene-RiViT. Node.js achieves high performance when processing large amounts of data, by using the performance optimized javascript V8 engine along with a non-blocking, asynchronous model for data processing and communication [12, 15]. This allows Gene-RiViT to efficiently handle on-the-fly queries on genome-scale data sets and send results back to users over the net at speeds sufficient to allow for seamless interaction.

Figure 1 shows the overall architecture of Gene-RiViT. Of particular interest is the central role of the web server in handling data processing and communication. When a user interacts with a gene of interest in the visualization, the web server will query the database for information about that gene, homologous genes in other organisms that are being examined, and their neighboring genes. The results are then processed into a format that can be recognized by the visualization and are returned over the network as they are processed. The asynchronous processing of node.js allows data to be effectively streamed back to the user, rather than returned as a single, large block. This allows users to interact with large amounts of complex data in real-time.

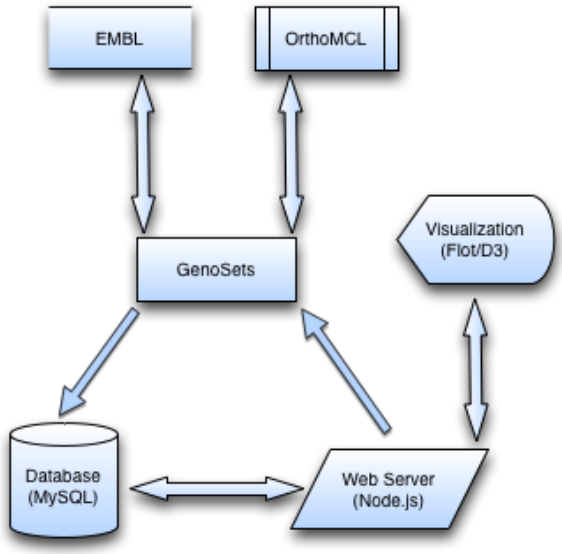

Figure 1: A central web server processes data between the database and the client-side visualizations. New data is generated via GenoSets, which retrieves and clusters data before adding it to the database.

4.1.1 GenoSets

GenoSets is a comparative genomic analysis platform that supports annotation parsing, manages ortholog clustering via orthoMCL, assigns Gene Ontology (GO) terms to genes, and structures data in the database with consistent gene definitions for a set of genomes [5]. GenoSets provides functionality for Gene-RiViT that lets researchers specify any dataset in the EMBL-Bank public repositories [18]. Through the Gene-RiViT interface, researchers can select any combination of genomes from a list of completed microbial projects or EMBL accession id. User requests from the visualization layer are passed to the web server, which uses GenoSets to download the specified genome data, cluster the data using orthoMCL, and load the processed data into the database. When the process is complete, users are notified by e-mail that their data is

ready for viewing. The amount of time required for processing is dependent on the size and number of genomes being loaded. A trial process that was run to cluster and load data of six different E. coli strains took approximately two hours to complete on a desktop computer, however, this is not an innate restriction on the system. Once data is loaded, users can switch between data sets through the Gene-RiViT interface and access any previously loaded data. The entire process of downloading, ortholog clustering, and loading data into the database, however, is a completely invisible process, which contributes to the ease of use of Gene-RiViT.

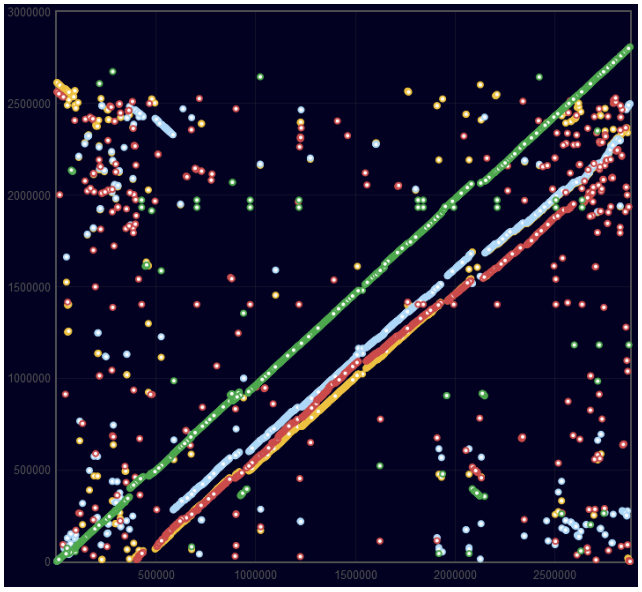

Figure 2: Homologous genes in four strains of staphylococcus with respect to a reference strain. The dot plot view provides a global representation of homologous gene positions of multiple organisms as compared to any selected reference genome. The x axis represents the position of the gene in a selected reference and the y axis shows the position in selected query genomes.

4.2 Database

Gene-RiViT uses a multi-dimensional architecture to support Online Analytical Processing (OLAP): a model typically used in business intelligence software to support real-time, ad hoc querying of data at different levels of granularity [8]. The database employs a star schema, in which source data is partitioned into facts that represent different dimensions of information. This database design facilitates the fast querying capacity that is necessary for interactivity with such a large amount of data. For example, information about a gene's location in the genome is stored separately from information about the ortholog clusters that gene may belong to. Each table in the database stores a different type of information about the gene and all of these dimensions of information are relationally linked through a central table [17]. By keeping the data separated and stored at this level of detail, it is possible for queries to be made for only the information that is necessary on tables that contain only the necessary information. This reduces the overhead that would otherwise slow down the querying process if the database were required to search for and parse out scattered chunks of information from several genomes worth of data.

4.3 Visualization

Gene-RiViT incorporates multiple views that are implemented using javascript-based visualization libraries. The two main libraries used are Flot [19] and D3 [4]. Flot is a javascript-based plotting library that uses the jQuery framework and HTML5 canvas to produce graphical plots on the fly on the client side [16]. The Data Driven Documents library, or D3, is a fast and efficient javascript library for developing interactive web-based visualizations. Both of these readily available technologies ensure compatibility on almost any system with no setup or configuration for the user, in addition to providing fast interfaces that allow users to view and interact with genome-scale data.

4.3.1 Plot View

Gene-RiViT provides researchers with multiple coordinated views of prokaryotic genome data. A global view is provided as a dot plot, in which positions of genes are plotted on a graph with respect to their positions in a reference organism. The reference organism can be selected from a list of any of the genomes that have been loaded into the database. Any number and combination of genomes can be selected for viewing with respect to the selected reference organism, and the reference genome can be changed dynamically. Each organism is represented as a different color in the plot and standard visualization features, such as panning and zooming, are incorporated. Figure 2 shows an example of four strains of Staphylococcus plotted with respect to a fifth strain. The strong main diagonal in this plot indicates that there is high gene-order similarity between these organisms, which is not unexpected as they are very closely related. However, the dot plot makes insertions, deletions and rearrangements easily visible as well as giving access to off diagonal similarities that would simply show up as gaps in a standard linear reference-based genome alignment.

Considering the large amount of data represented in the dot plot, it is necessary to incorporate methods for locating areas that might be of interest. A number of methods for accessing the data presented in the dot plot are implemented. Annotation data stored in the database is provided to the user as a list of Gene Ontology terms. The interface provides a method for selecting GO terms, such as cell surface binding, from a menu, which results in all genes that are either annotated or homologous to genes with the selected function being highlighted in the plot, while all genes without the specified function will be colored grey to make the selected genes more obvious. This functionality can help researchers to identify genes in potentially interesting functional categories as starting points for more detailed exploration, or to improve on the level of detail provided in annotation information.

4.3.2 RiViT View

While the dot plot view provides an overall picture of the organization of genes for selected organisms on a global scale, the RiViT view provides a local context in which to examine relative ordering and reshuffling of genes. The RiViT view consists of a series of nested, rotating semi-circles, each of which represents the ninety genes around a central gene that is aligned with homologous genes in other circles. A gap is incorporated at the nine o'clock position of the view to show that there is a discontinuity between the genes shown in the view and that the circles represent only a local view of a larger volume of information. When the circles rotate clockwise, genes from previously viewed regions of the genome fade into the circle from the gap, while genes entering the gapped region fade out. The opposite is true when rotation is counter clockwise. For cases where an alignment exists, but homologous genes are not found for every organism selected, the genes on an alignment for organisms with no matches are shaded grey to indicate that there is no alignment for that organism. In the opposite case, where multiple possible alignments are found in a single organism, navigation buttons are provided that allow users to rotate the circle for that organism through the set of existing matches.

The number of rings displayed in the RiViT view is dependent

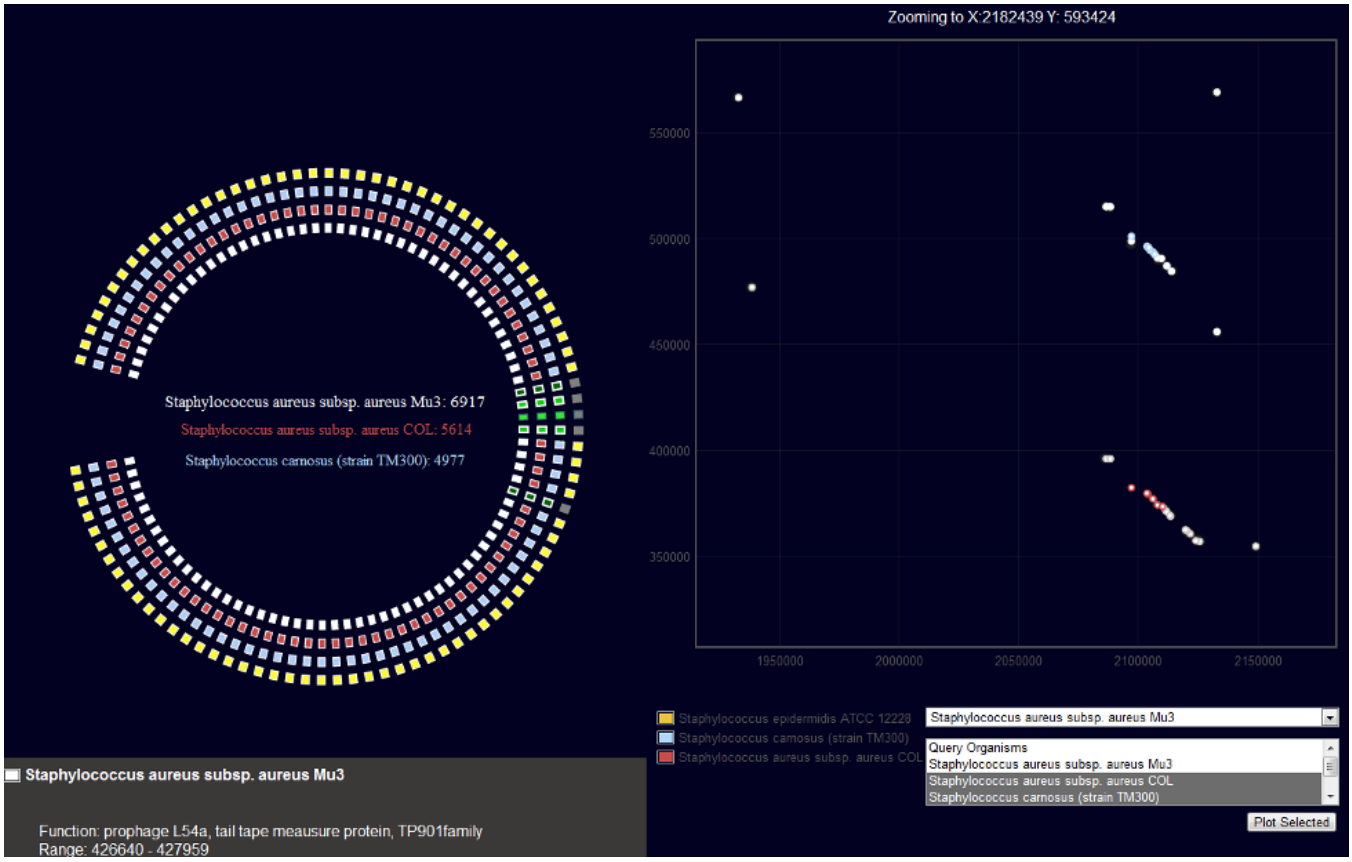

Figure 3: A conserved operon examined using Gene-RiViT. Left(a): The RiViT view provides local context information about gene neighborhoods. Homologous genes are aligned in green. Genes shown in grey show that no homologs were found for an organism. Right(b): A zoomed view of the dot plot view shows a conserved operon for two query species. Five homologous genes are highlighted around a central selected gene. All other points are shaded grey to reduce visibility.

on the number of organisms selected for investigation. The previous example of four staphylococcus strains aligned to a fifth reference strain would result in five color-coded rings representing the selected organisms. When a user interacts with genes in the dot plot view, these rings rotate to align homologous genes for all organisms at the three o'clock angle. A user specified number of neighboring genes around the selected gene are also checked against neighboring genes in the query organisms for homology and color-coded accordingly. The RiViT view, then, provides researchers with information about relative restructuring or conservation that has occurred at a local level between organisms under study without regard to long-range structural changes. This allows genes and gene neighborhoods to be examined and compared in a local context that can provide information about how specific genes might be functionally linked. Figure 3a shows an example of the RiViT view aligning a set of genes from a conserved operon in staphylococcus.

Users are also able to select any gene in the RiViT view to make that gene the new point of reference. In this way, gene neighborhoods can be smoothly explored by allowing users to switch between different organisms that have differing homology relationships, or move in steps along the genome to explore wider ranges of neighborhood information. For example, organisms a, b, and c may have an orthologous match that can be aligned by selecting any of the genes in a specific ortholog cluster. Organisms c and d, may have related genes in the same neighborhood that can be examined by switching focus to organism c. This sort of exploratory analysis could be particularly useful in comparing genomes of species at varying evolutionary distance.

4.3.3 Details View

In addition to the interactive views discussed above, a detailed list of information about genes is provided. This list provides details about each of the genes in the alignment, as well as a list of neighboring genes that can be selected to view details. By default, size, starting and ending positions in the genome, and annotation information are displayed, though users can select additional annotation information to view by selecting options from a menu. Selecting a new gene for investigation will automatically update the list to show information about genes and their neighbors in the new alignment.

5 Discussion

Gene-RiViT is designed to be useful in a variety of research situations. In more recently divergent organisms, Gene-RiViT can be used to examine the effects of gene rearrangement by comparing multiple organisms. In more evolutionarily distant organisms, it can also be used to identify and examine the details of conserved sets of genes and for functional inference by association. A simple example of the utility of Gene-RiViT was performed to illustrate proof of concept.

Figure 3 shows Gene-RiViT in use. In this instance, four strains of Staphylococcus are being examined: *aureus* Mu3, *aureus* COL, *carnosis* TM300, and *epidermitis* ATCC 12228. In this case, strain Mu3 was selected as the reference strain. A region was selected based on a local alignment of genes, visible as a short diagonal,

Table 1: Gene neighborhood alignment annotation information

Distance	Mu3	COL	TM300	ATCC 12228
-2	hypothetical protein	prophage L54a, terminase, large subunit,putative	putative phage terminase, large subunit	No match
-1	hypothetical protein	prophage L54a, Clp protease, putative	putative Clp protease, phage associated	No match
0	hypothetical protein	hypothetical protein	hypothetical protein	No match
1	hypothetical protein	conserved hypothetical protein	conserved hypothetical protein	No match
6	phi PVL ORF 15 and 16 homologue	prophage L54a, tail tape meausure protein, TP901family	truncated phiSLT orf2067-like protein	No match

that was observed in the global dot plot view shown in figure 2. The area was zoomed in and a gene from the center of the aligned region in strain *aureus* COL(red) was selected. Points in the plot that were not homologous to one another in a range of one hundred genes around the selected gene were shaded grey in the plot, showing homologous genes in the region between the reference genome, strain Mu3, and strains TM300(blue) and COL(red). The RiViT view rotated to align the gene selected in the plot with the homologous genes in other organisms. Detail information about the aligned genes was then provided in the detail view. In the case of strain ATCC 12228, no homologous genes were found and the visible genes from this species were shaded grey in the alignment to show that there were no gene neighborhood matches. The alignment, highlighted in green, shows that two genes prior to the selected gene are also homologous to one another, as is the following gene and another gene six gene positions away.

Table 1 shows annotation information for the aligned genes, with the distance column indicating the distance in steps away a gene is from the selected gene, and zero indicating the gene that was selected. In this case, the gene selected was annotated as a hypothetical protein in each of the organisms under comparison. Examination of the surrounding homologous genes, however, shows very similar annotation information between the COL and TM300 strains. Based on the homology information showing known orthology in a conserved range in multiple genomes with a lack of rearrangements, it is possible to infer with reasonable confidence that the function of the hypothetical proteins at the center of the alignment is related to prophage L54a, though this could be experimentally verified to obtain a higher level of confidence and precision. Because the information about the known genes shows that these genes are phage related and have a conserved order spanning multiple genes, most likely these genes exist as a result of a horizontal transfer event [23]. Investigating prophage L54a revealed that the integration of prophage L54a results in a loss of lipase activity in Staphylococcus *aureus* PS54 due to insertion at the 3' end of the lipase structural gene [20]. A researcher using Gene-RiViT could verify this result by navigating through the local context provided by the RiViT view to see if neighboring genes were in fact involved in lipase activity.

Researchers can perform a variety of studies using Gene-RiViT. The previous example illustrates that Gene-RiViT can be used to identify local, functionally significant similarities among genomes even when genomes are not completely collinear. Gene-RiViT is not restricted to only this use. Researchers could use Gene-RiViT, for example, to identify genes associated with pathogenicity in several related species and make observations about their functions and implications in other species who either share the same genes or are specifically lacking them. Gene-RiViT could also be used to make more intelligent decisions about evolutionary distance between closely related species in cases where precision is a factor by allowing researchers to examine the scope of rearrangements that have occurred within a collection of genomes.

6 Conclusion and Future Work

We presented Gene-RiViT, a visual analytic tool for the on-the-fly analysis of gene neighborhoods in bacterial genomes. Gene-RiViT provides a coordinated set of visualizations for examining genomic data at multiple levels of granularity, with particular focus on gene order in local gene neighborhoods. Gene-RiViT uses state of the art web technology to present data as a dynamic and adjustable alignment, rather than the more common presentation of a fixed alignment pegged to a reference genome. The ring visualization in Gene-RiViT allows the user to pick any gene in any of the genomes in the set as the query, upon which the entire genomic alignment rapidly rearranges to bring orthologs in the other genomes into alignment with the query. Highlights on the genome then show orthology relationships in the neighborhood surrounding the query gene, giving insight into the conservation of local genome context and the preservation of functional operons. A second visualization of the genomes being compared as a familiar 2D dot plot allows the user to pinpoint regions of interest based on observed features in the 2D alignment. Selections in the dot plot visualization can be used to control and highlight the dynamic genome ring visualization, and vice versa. Keyword searches and Gene Ontology based searches are also available as entry points to the data. Gene-RiViT has a wide variety of potential uses in comparative genomics studies and will be freely available and easily accessible to the public.

The visualization tools in Gene-RiViT are designed to function as part of a coordinated multiple view interface that includes multiple methods for target gene selection. For instance, we have previously implemented a java Parallel Sets visualization, used in conjunction with a visualization of Gene Ontology categories in a treemap view, to rapidly identify common and differentiating genes in multiple genome data sets and to further subdivide those gene lists by functional category [5]. Gene-RiViT provides an intermediate level of detail between these high level abstractions of the genome data set and the literal linear views to which biologists are accustomed. Integration of Gene-RiViT with Parallel Sets and Gene Ontology hierarchy views is planned, along with incorporation of other feature markup such operon predictions. Further information on the development of Gene-RiViT will be made available through http://gibas-research.uncc.edu.

Acknowledgements

The authors wish to thank Aurora Cain for her contributions to the incorporation of GenoSets. This work was supported in part by the National Science Foundation, award number 1047896.

References

[1] E. J. Alm, K. H. Huang, M. N. Price, R. P. Koche, K. Keller, I. L. Dubchak, and A. P. Arkin. The microbesonline web site for comparative genomics. *Genome Research*, 15(7):1015–1022, 2005.

[2] A. M. Altenhoff and C. Dessimoz. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*, 5(1):e1000262, 01 2009.

[3] L. Aravind. Guilt by association: contextual information in genome analysis. *Genome Research*, 10(8):1074–1077, aug 2000.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

[5] A. Cain, R. Kosara, and C. Gibas. Genosets: Visual analytic methods for comparative genomics. *PLOS*, 2012. Under Review.

[6] T. J. Carver, , K. M. Rutherford, M. Berriman, M.-A. Rajandream, B. G. Barrell, and J. Parkhill. Act: the artemis comparison tool. *Bioinformatics*, 21(16):3422–3423, 2005.

[7] R. R. Chaudhuri, A. M. Khan, and M. J. Pallen. colibase: an online database for escherichia coli, shigella and salmonella comparative genomics. *Nucleic Acids Research*, pages 296–299, 2004.

[8] E. F. Codd, S. B. Codd, and C. Salley. *Providing OLAP (on-line Analytical Processing) to User-analysts: An IT Mandate*, volume 32. Codd and Date Inc., 1993.

[9] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324 – 328, 1998.

[10] W. M. Fitch. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19(2):99–113, 1970.

[11] C. Fong, L. Rohmer, M. Radey, M. Wasnick, and M. J. Brittnacher. Psat: A web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC Bioinformatics*, 9(1):170, Mar. 2008.

[12] Google. V8 javascript engine. http://code.google.com/p/v8/, 2012.

[13] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318 – 356, 1961.

[14] I. K. Jordan, K. S. Makarova, J. L. Spouge, Y. I. Wolf, and E. V. Koonin. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Research*, 11(4):555–565, 2001.

[15] Joyent-Inc. Node.js. http://nodejs.org/, 2006-2012.

[16] jQuery Foundation. The jquery project. http://jquery.org, 2005-2012.

[17] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition)*. Wiley, 2 edition, apr 2002.

[18] T. Kulikova, R. Akhtar, P. Aldebert, N. Althorpe, M. Andersson, A. Baldwin, K. Bates, S. Bhattacharyya, L. Bower, P. Browne, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, G. Hoad, C. Kanz, C. Lee, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, M. P. G. Pastor, S. Plaister, S. Sobhany, P. Stoehr, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. Embl nucleotide sequence database in 2006. *Nucleic Acids Research*, 35(suppl 1):D16–D20, 2007.

[19] O. Laursen. Flot: Attractive javascript plotting for jquery. http://code.google.com/p/flot/, 2009-2012.

[20] C. Y. Lee and J. J. Landolo. Integration of staphylococcal phage l54a occurs by site-specific recombination: Structural analysis of the attachment sites. *Proceedings of the National Academy of Sciences of the United States of America*, 83(15):5474–5478, 1986.

[21] L. Li, C. J. Stoeckert, and D. S. Roos. Orthomcl: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189, 2003.

[22] S. S. Morrison, T. Williams, A. Cain, B. Froelich, C. Taylor, C. Baker-Austin, D. Verner-Jeffreys, R. Hartnell, J. D. Oliver, and C. J. Gibas. Pyrosequencing-based comparative genome analysis of vibrio vulnificus environmental isolates. *PLOS*, 2012. In press.

[23] H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299 – 304, 2000.

[24] I. B. Rogozin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin. Computational approaches for the analysis of gene neighborhoods in prokaryotic genomes. *Briefings in Bioinformatics*, 5(2):131–149, June 2004.

[25] E. L. Sonnhammer and E. V. Koonin. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, 18(12):619 – 620, 2002.

[26] I. Uchiyama, T. Hiuchi, and I. Kobayashi. Cgat: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. *BMC Bioinformatics*, 7(1):472, Oct. 2006.

[27] Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research*, 11(3):356–372, 2001.