Visual Exploration of Genetic Sequence Variants in Pangenomes

Astrid van den Brandt[1], Eef M. Jonkheer[2], Dirk-Jan M. van Workum[2], Sandra Smit[2] and Anna Vilanova[1]

[1]Eindhoven University of Technology, The Netherlands
[2]Bioinformatics Group, Wageningen University, The Netherlands

Abstract

To study the genetic sequence variation underlying traits of interest, the field of comparative genomics is moving away from analyses with single reference genomes to pangenomes; abstract representations of multiple genomes in a species or population. Pangenomes are beneficial because they represent a diverse set of genetic material and therefore avoid bias towards a single reference. While pangenomes allow for a complete map of the genetic variation, their large size and complex data structure hinder contextualization and interpretation of analysis results. Current visualization strategies fall short because they are created for single references or do not illustrate links to metadata. We present a work in progress version of a novel visual analytics strategy for pangenomic variant analysis. Our strategy is designed through an intensive involvement of genome scientists. The current design uniquely exploits interactive sorting, aggregation, and linkage relations from different perspectives of the data, to help the genome scientists explore and evaluate variant-trait associations in the context of multiple references and metadata.

1. Introduction

Studying genetic variation underlying traits of interest is an important topic in comparative genomics. In plant genomic research, for example, scientists analyze the variation between cultivated and wild types to develop crops with improved resistance to diseases. Traditionally, this analysis is performed by comparing the genome sequences of the studied samples to a single reference genome. Because the number of sequenced genomes is growing rapidly and to avoid the bias toward a single reference genome, the field is shifting towards "pangenomes", abstract data structures to represent all genetic variation in a species or population. While pangenomes allow for a complete picture of the variation, interpreting variants in the biological context remains challenging [FNM13, MMA*18]. Therefore, genome scientists need interactive visual tools supporting exploratory analysis of variants within pangenomes.

Many tools have been developed for exploration of genomes [NHG19]. Genome browsers are the dominant tools for general inspection of the genome sequences and annotations [WPM*09, TRM13]. Furthermore, several visualization tools have been proposed specifically for variant analysis [FNM13]. While these tools provide intuitive views of the sequences, they cannot be used with pangenomes containing multiple references [MMA*18]. For pangenomes, few visualization tools have been proposed [SHZ*17, DBN18]. However, these tools provide (i) only high-level views of pangenome data structure or (ii) do not incorporate links to associated metadata (such as traits). These approaches cannot assist in complex analysis goals requiring visual exploration of genetic sequence variants and metadata within the pangenome. We present a visual analysis strategy that facilitates exploring sequence variants and Traits of Interest (TOIs) in a pangenomic manner. Our strategy uniquely enables genome scientists to get a picture of genetic diversity in a population of interest and identify associations with TOIs, using common representations combined with rich interactions.

2. Workflow and Task Analysis

Our visual analysis strategy is motivated by a general goal of our domain collaborators to interactively explore genetic sequence variants in pangenomes in the context of metadata. This goal was identified through informal interviews and brainstorming sessions with 5 domain experts in breeding and biotechnology companies, as well as 2 PhD researchers working on pangenomic studies. To gain a deeper understanding, we have organized a Creative Visualization Opportunities workshop [KGD*19]. One outcome of this workshop was a shared workflow that allowed us to refine our goal into a set of analysis tasks.

We focus on the workflow of exploring genetic variants that can distinguish the samples into groups, identifying potential biomarkers (i.e., molecular indicators of a biological state). At the start of this workflow, we presume that the user has a list of candidate genes associated with some trait. For example, for *Arabidopsis* plants, there are three genes known to be flowering time regulators [SBW*09, MVMZAB13, LFB*14]. To browse variations in the DNA sequence and trait values, and potentially discover a set of biomarkers linked to the variation, the following four high-level tasks should be supported: **(T1)** Explore genomic sequence context of a gene, **(T2)** Analyze relations between sequences, phylogeny and traits, **(T3)** Define and analyze groups of similar samples, and **(T4)** Explore variant-trait associations within groups.

DOI: 10.2312/evp.20221112

Figure 1: *Overview of the visual analysis strategy containing a **Gene Overview** (A) and **Locus View** (B-I). The Gene Overview visualizes conservation across the whole gene and allows slicing an area for further inspection in the Locus View. The Locus View contains four subviews: (B) a dendrogram or tree of the samples, (I) bipartite graphs connecting tree, sequence and trait orderings, (C) the multiple sequence alignment, (D) traits of interest. Options for filtering, sorting and linking (F). Calculate a new clustering based on positions of interest (E) or inspect the sliced region in current sorting and related to another tree (H). Aggregation of sequences (G).*

3. Visual Analytics Strategy

An overview of our work-in-progress visual analytics tool is shown in Figure 1. The overall design arose from the tasks and various brainstorming sessions with our collaborators. The design has two main views which allow an overview+details exploration of variants in a gene, combined with linked interactions such as sorting, clustering, and selection with aggregation. Below we describe the views, interactions, and insights while inspecting the *rfaL* gene in a *Pectobacterium* pangenome [JBH*21].

Gene Overview Panel A provides an image of the conservation levels of aligned positions and their neighbors (i.e., the ratio of similar nucleotides on a position) across a gene **(T1)**. Any region within a gene can be selected and inspected in the Locus View.

Locus View In panel C, the user can inspect a multiple sequence alignment (MSA) of the selected region **(T1)** by a Heatmap. In panel B, she can examine (evolutionary) relations between conserved or variable positions in the sequences using a phylogenetic tree or customized cluster dendrogram. Panel D shows TOI values with different encodings for each type. Panel I shows a bipartite graph to connect various tree, sequence, and trait orderings **(T2)**.

Interactions and Analytics Our strategy enables interactive sorting, aggregation, and linkage relations from several data perspectives. With Panel E, the user can interactively calculate a clustering based on a selection of positions for inspection of groups of similar sequences **(T3)** with metadata sorted accordingly. For example, panel C shows an insertion of three nucleotides around position 200 (pink arrow). Further sorting can be applied such as by TOIs (D) to discover patterns within groups **(T4)**. For example, after sorting by species, the insertion is shared by half of the *P. aquaticum* species (pink box). It is possible to view the current MSA and metadata sorting linked to another dendrogram or tree (e.g., gene tree, core SNP tree of all genomes or, k-mer tree) derived from this pangenome to compare trees (H). For these interactions, bipartite graphs (I) show connections between tree, sequence, and trait orderings. Lastly, samples can be collapsed through the inner tree nodes to fit more information on the screen and only show detail in relevant parts (G). Regarding biomarkers, we observe some variants around the insertion following the same pattern: positions 192 green vs. blue (c1) and 214 green vs. brown (c2).

4. Conclusion and Future Work

We have developed a novel visual analytics strategy for pangenomic variant exploration. The current tool uniquely supports an interactive exploration of sequences with multiple references and metadata via various sorting and linkage options. We aim to extend interactions by enabling aggregation of samples with similar metadata, and by filtering options based on integrated side views of structural and functional annotations for more detailed inspections.

5. Acknowledgements

This work is funded by TKI TU project TU18034.

References

[DBN18] DING W., BAUMDICKER F., NEHER R. A.: panX: pan-genome analysis and exploration. *Nucleic acids research 46*, 1 (2018), e5. doi:10.1093/nar/gkx977. 1

[FNM13] FERSTAY J. A., NIELSEN C. B., MUNZNER T.: Variant view: Visualizing sequence variants in their gene context. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2546–2555. doi:10.1109/TVCG.2013.214. 1

[JBH*21] JONKHEER E. M., BRANKOVICS B., HOUWERS I. M., VAN DER WOLF J. M., BONANTS P. J., VREEBURG R. A., BOLLEMA R., DE HAAN J. R., BERKE L., SMIT S., DE RIDDER D., VAN DER LEE T. A.: The Pectobacterium pangenome, with a focus on Pectobacterium brasiliense, shows a robust core and extensive exchange of genes from a shared gene pool. *BMC Genomics 22*, 1 (12 2021). URL: /pmc/articles/PMC8045196/ /pmc/articles/PMC8045196/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8045196/, doi:10.1186/s12864-021-07583-5. 2

[KGD*19] KERZNER E., GOODWIN S., DYKES J., JONES S., MEYER M.: A Framework for Creative Visualization-Opportunities Workshops. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 748–758. doi:10.1109/TVCG.2018.2865241. 1

[LFB*14] LI P., FILIAULT D., BOX M. S., KERDAFFREC E., VAN OOSTERHOUT C., WILCZEK A. M., SCHMITT J., MCMULLAN M., BERGELSON J., NORDBORG M., DEAN C.: Multiple FLC haplotypes defined by independent cisregulatory variation underpin life history diversity in Arabidopsis thaliana. *Genes and Development 28*, 15 (2014), 1635–1640. doi:10.1101/gad.245993.114. 1

[MMA*18] MARSCHALL T., MARZ M., ABEEL T., DIJKSTRA L., DUTILH B. E., GHAFFAARI A., KERSEY P., KLOOSTERMAN W. P., MÄKINEN V., NOVAK A. M., PATEN B., PORUBSKY D., RIVALS E., ALKAN C., BAAIJENS J. A., DE BAKKER P. I., BOEVA V., BONNAL R. J., CHIAROMONTE F., CHIKHI R., CICCARELLI F. D., CIJVAT R., DATEMA E., VAN DUIJN C. M., EICHLER E. E., ERNST C., ESKIN E., GARRISON E., EL-KEBIR M., KLAU G. W., KORBEL J. O., LAMEIJER E. W., LANGMEAD B., MARTIN M., MEDVEDEV P., MU J. C., NEERINCX P., OUWENS K., PETERLONGO P., PISANTI N., RAHMANN S., RAPHAEL B., REINERT K., DE RIDDER D., DE RIDDER J., SCHLESNER M., SCHULZ-TRIEGLAFF O., SANDERS A. D., SHEIKHIZADEH S., SHNEIDER C., SMIT S., VALENZUELA D., WANG J., WESSELS L., ZHANG Y., GURYEV V., VANDIN F., YE K., SCHÖNHUTH A.: Computational pan-genomics: Status, promises and challenges. *Briefings in Bioinformatics 19*, 1 (2018), 118–135. doi:10.1093/bib/bbw089. 1

[MVMZAB13] MÉNDEZ-VIGO B., MARTÍNEZ-ZAPATER J. M., ALONSO-BLANCO C.: The Flowering Repressor SVP Underlies a Novel Arabidopsis thaliana QTL Interacting with the Genetic Background. *PLOS Genetics 9*, 1 (1 2013), e1003289. URL: https://doi.org/10.1371/journal.pgen.1003289. 1

[NHG19] NUSRAT S., HARBIG T., GEHLENBORG N.: Tasks, techniques, and tools for genomic data visualization. *Computer Graphics Forum 38*, 3 (2019), 781–805. doi:10.1111/cgf.13727. 1

[SBW*09] SCHWARTZ C., BALASUBRAMANIAN S., WARTHMANN N., MICHAEL T. P., LEMPE J., SUREshkumar S., KOBAYASHI Y., MALOOF J. N., BOREVITZ J. O., CHORY J., WEIGEL D.: Cis-regulatory Changes at FLOWERING LOCUS T Mediate Natural Variation in Flowering Responses of Arabidopsis thaliana. *Genetics 183*, 2 (10 2009), 723–732. URL: https://doi.org/10.1534/genetics.109.104984, doi:10.1534/genetics.109.104984. 1

[SHZ*17] SUN C., HU Z., ZHENG T., LU K., ZHAO Y., WANG W., SHI J., WANG C., LU J., ZHANG D., LI Z., WEI C.: RPAN: Rice pan-genome browser for 3000 rice genomes. *Nucleic Acids Research 45*, 2 (2017), 597–605. doi:10.1093/nar/gkw958. 1

[TRM13] THORVALDSDÓTTIR H., ROBINSON J. T., MESIROV J. P.: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics 14*, 2 (3 2013), 178–192. URL: https://doi.org/10.1093/bib/bbs017, doi:10.1093/bib/bbs017. 1

[WPM*09] WATERHOUSE A. M., PROCTER J. B., MARTIN D. M., CLAMP M., BARTON G. J.: Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics 25*, 9 (2009), 1189–1191. doi:10.1093/bioinformatics/btp033. 1