Erschienen in: Journal of Visualization ; 23 (2020), 3. - S. 523-537. - Springer
https://dx.doi.org/10.1007/s12650-020-00635-0

Tong Ge · Yonghua Lu · Kecheng Lu · Yunhai Wang · Xin Liu · Zhanglin Cheng · Yi Chen · Oliver Deussen · Baoquan Chen

VEGA: visual comparison of phylogenetic trees for evolutionary genome analysis (ChinaVis 2019)

Abstract In the field of evolutionary genome analysis, biologists seek to identify important genes or chromosome regions by comparing phylogenetic trees and analyzing the mutation at which locus might affect phenotypic traits. Unfortunately, the tree comparison and accompanying analysis are often performed manually. In this paper, we characterize the workflow of evolutionary genome analysis and present a task analysis for the fundamental questions asked by biologists during the analysis procedure. We propose two algorithms to enable quantitative tree comparison. One is to measure the differences between corresponding leaf nodes on two trees, and the other is to compute the classification inconsistency of each leaf node by comparing tree structure with a given biological classification. Configuring with the obtained difference and inconsistency, we present a visual analysis system, visual comparison of phylogenetic trees for evolutionary genome analysis, which not only enables biologists to intuitively explore trees but also identify locus which affects their traits by comparing SNP variants of selected leaf nodes. We conclude with case studies from two biologists who used our system to augment their previous manual analysis workflow and demonstrate that our system can reveal more insight.

Keywords Visual analysis · System · Genome · Phylogenetic tree

Tong Ge and Yonghua Lu assert equal contribution and joint first authorship.

T. Ge (✉) · K. Lu · Y. Wang · B. Chen
Shandong University, Jinan, China
E-mail: tgeconf@gmail.com

Y. Lu
Shenzhen Investigation and Research Institute Co., Ltd, Shenzhen, China

X. Liu
BGI, Shenzhen, China

Z. Cheng
SIAT, Shenzhen, China

Y. Chen (✉)
Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing, China
E-mail: chenyi@th.btbu.edu.cn

O. Deussen
University of Konstanz, Konstanz, Germany

1 Introduction

The rapid development of high-throughput sequencing technologies enables whole-genome sequencing at an unprecedented rate applied to study most of the different organisms. With the obtained whole-genome sequence data, the biologists seek to investigate the genome-wide variation patterns of one species. More specifically, they want to identify the genes which might affect the evolutionary history or response for the significant phenotypic traits changed during the evolution, such as yield, color, size, and others. This identification is facilitated by comparing genomic variation within varieties. Once these genes are identified, they might be used as the basis for future genomic-enabled breeding (Rubin et al. 2010; Li and Zhang 2013) or diagnosing (Xun et al. 2012).

In population genetics, the whole-genome data consist of the DNA sequences of different strains of one species. By organizing each strain as a leaf node, the phylogenetic tree is often used to reveal the evolutionary relationship of different strains. To characterize patterns of genetic variation, biologists often compare the phylogenetic tree generated by one gene or a chromosome region with the tree generated by the whole genome. And they explain differences between two trees with some prior knowledge, such as the biological classification. For example, if some varieties have been misclassified, a further selection of the chromosome region is applied and results in finding the genotype that can explain the phenotypic traits change. Although this phylogenetic tree comparison with accompanying analysis is a common practice in population phylogenetic analysis (Rubin et al. 2010; Xun et al. 2012; Qi et al. 2013), most of the time it is manually operated as far as we know, which largely hinders the progress of science.

A number of metrics (Graham and Kennedy 2010) have been proposed to measure the distance between two trees, whereas there are few works in measuring the distance between leaf nodes with the same label in two trees. To address this issue, we present a novel linear-time phylogenetic tree comparison algorithm which can quantitatively measure the leaf node differences between two trees. Rather than directly measuring the change of the leaf node itself, we define the difference as the path length from the target leaf node to the leaf nodes of its sibling node, where the sibling node is defined on the reference tree. Once obtaining the node differences, we visualize the leaf node differences with treemaps, which facilitates the user to quickly select the nodes with large differences. If biologists are interested in some changes, they can examine how the tree structure is consistent with the known biological classification. Although encoding the classification information to tree edge color can help the user examine the inconsistency, it is hard for the user to quickly identify which leaf nodes have the largest inconsistency. To address this issue, we propose another algorithm to measure the inconsistency between the tree and biological classification, by examining how the label of the target leaf node is similar to the labels of leaf nodes of its sibling node.

To facilitate biologists to identify important genes, we present the first visual analytic system, Visual comparison of phylogenetic trees for evolutionary genetic analysis (VEGA), that enables biologists to explore the relationship between genes and evolutionary history of many populations in one species with the whole-genome data. Our system not only allows the user to intuitively explore tree differences and classification inconsistency but also interactively investigates leaf nodes, which corresponds to individual strains. By visualizing node values with a treemap, the user can quickly get an overview of the differences or inconsistency and easily identify the leaf nodes of interest. Here, the inconsistency refers to the degree of how the classification implied by the phylogenetic tree is different from the biological classification. If the biologist finds some strains with large differences or inconsistencies, the different single nucleotide polymorphisms (SNPs) of these varieties can be visualized that provides the hints to explain the phenotypic traits change. Since our tree comparison algorithm does not place any requirement about the input data, our system can also be used to compare the differences in evolutionary history revealed by different chromosome regions. This facilitates biologists to characterize genes that have not been functionally characterized yet.

We conducted two case studies with two different whole-genome datasets to test the capabilities of VEGA in analyzing whole-genome patterns. Its usefulness has been demonstrated in identifying new important genes or chromosome regions. The main contributions of this paper are summarized as follows:

- We characterize the workflow of evolutionary genome analysis and identify the fundamental questions met at each analysis stage.
- We propose two novel algorithms to perform phylogenetic tree-related comparisons, which facilitate the user to intuitively explore tree differences and tree classification inconsistency;
- We present the first dedicated visual analysis tool which enables intuitive exploration of the relationship between genes and evolutionary history of one species for the biologist.

The rest of the paper is organized as follows. After introducing the related work in Sect. 2, we briefly describe the background of evolutionary genome analysis in Sect. 3. Next, we introduce our two phylogenetic tree-related comparison algorithms in Sect. 4 and present our system in Sect. 5. Finally, we report our case studies in Sect. 6, followed by the conclusions of our work.

2 Related work

Genome Visualization The emergence of extensive genome sequence data has opened the field of genome visualization, which enables the biologist to intuitively explore the data and interactively refine the preliminary automatic analysis result. Nielsen et al. (2010) review existing genomic data visualization tools and discuss their advantages and disadvantages. According to them, genome browsers (Kent et al. 2002; Fiume et al. 2010; Thorvaldsdóttir et al. 2013) are the most commonly used tools in genome visualization. By visually encoding variant attributes and multi-scale annotation information, the recently developed tools, such as cBio (Cerami et al. 2012), MuSiC (Dees et al. 2012) and Variant View (Ferstay et al. 2013), support intuitive exploration of sequence variants. These tools can be taken as a complement to this work, where they can be used to compare the sequences of selected individual strains.

Rather than focusing on individual sequence exploration, evolutionary genetic analysis studies the genetic variation within populations. To characterize genetic variation, the phylogenetic tree (Penny et al. 1992) is often used to reveal the evolutionary relationship between different individual strains.

Tree Construction and Visualization To construct a phylogenetic tree, neighbor-joining algorithm (Saitou and Nei 1987) is most commonly used. It is based on a distance matrix where the distances between different strains are often measured by p-distance. The p-distance is the proportion (p) of nucleotide sites at which two sequences being compared are different. In a rooted phylogenetic tree, leaf nodes correspond to individual strains, while internal nodes represent common ancestors.

Node-link diagrams and treemaps are two main representations of tree visualization. A phylogenetic tree can be drawn by using node-like diagrams with radial, rectangular and circle layouts (Bachmaier et al. 2005; Von Landesberger et al. 2011), whose edge length can be regarded as the evolutionary time. Strains with similar evolutionary histories are grouped together. And in most cases, the grouping revealed by the tree is consistent with the biological classification. This presentation offers an intuitive representation for the user to study the relationship between different nodes. However, it is not space efficient when the number of nodes is large. Many available tools, such as Darwin (Perrier and Jacquemoud-Collet 2006) and Mega (Kumar et al. 2008), visualize phylogenetic trees in this manner. In contrast, treemap (Shneiderman 1998) is a space-efficient technique by recursively laying out child nodes within their respective parent nodes. By encoding the value of a leaf node with box size or color, it is very efficient for the user to select some leaf nodes of interest. However, as demonstrated in Barlow and Neville (2001), the hierarchical structure in the treemap is not as clear as the node-link representation. In this paper, we combine these two representations together where the user selects leaf nodes from treemap and then corresponding subtrees in node-link representation will be highlighted.

Tree Comparison Tree comparison is an important topic in information visualization (Graham and Kennedy 2010; Guerra-Gómez et al. 2013), and we restrict the discussion on this work about phylogenetic trees comparison. TreeJuxtaposer (Munzner et al. 2003) is one of the most closed examples that compares two phylogenetic trees by associating each node in one tree to its most similar node in the other tree. Bremm et al. (2011) developed a set of linked hierarchy views for the comparison of multiple phylogenetic trees with a modified Robinson-Foulds distance. Robinson et al. (2016) proposed an online side-by-side phylogenetic tree comparison tool, Phylo.io, which can automatically find the best corresponding internal node or subtree structure to the user selected ones. However, biologists need to examine all leaf nodes to find those strains with the most change in the evolutionary history or inconsistent with the biological classification. Thus, the comparison methods of computing difference of nodes themselves or locating some specific nodes in the existing tools are not enough to meet the analytical needs. To facilitate biologists to locate genes of interest during analysis, matrices to quantitatively measure the change of strains according to the evolutionary relationship on the tree, and the inconsistency with the biological classification are needed.

DoubleTree (Parr et al. 2004) uses two connected side-by-side phylogenetic trees to highlight topological differences between biological classifications. In our work, the differences of leaf nodes are quantitatively measured and intuitively visualized with a treemap, which helps the user get an overview of the relation differences of all individuals.

3 Background

In this section, we present the first contribution, a characterization of the problem domain. This characterization includes a description of the data structure of whole-genome data, the pipeline that how the population geneticist analyzes the genome sequence and identification of the challenges in genetic analysis. We learn this characterization by closely working with two target biologists for eight months, who have more than 10-year experience in evolutionary genetic analysis.

3.1 Data description

Since each sequence has been aligned to a reference, the input of each genome sequence is the difference between an individual genome and the reference genome, called SNP variants. Each SNP refers to a variation at a single position in a sequence among individuals. Besides four DNA bases A, C, G or T, some SNP variants degenerate bases which are synthesized with multiple bases. For example, the degenerate base represented with the letter *R* corresponds to the mixed position of *A* and *G*. For more details about the representation of degenerate bases, please refer to nucleic acid notation (http://en.wikipedia.org/wiki/Nucleic_acid_notation). In our data, all degenerate bases are diploid, and some of them belong to heterozygous genotype.

Since most SNP variants are within protein-coding regions, each SNP variant can potentially change the amino acid it codes for. If it does not lead to the amino acid change, it is called synonymous SNP otherwise non-synonymous SNP. Non-synonymous SNP potentially alters the function of a protein and consequently changes the phenotypic traits. These SNP variants might be useful in the interpretation of the evolutionary relationship between different strains. They are the targets that our collaborators want to identify.

In our case, *SNP* is the smallest element in the sequence, located at different genes, while one *chromosome* consists of many genes. Since all sequences have been aligned to a reference, the chromosome and genes of each sequence can be visited by using the position information from the reference sequence. Besides the genome, scientists also have their biological classification for each strain, which is used to explain the interesting mutation.

3.2 Evolutionary genome analysis workflow

Advances in next-generation sequencing have enabled the study of genomic evolution and diversity on a whole-genome scale. To date, a few studies have been undertaken to discover important genes and determine their functions from the population of one species (Xu et al. 2012; Qi et al. 2013). After interviewing with the population geneticist, we summarize the following pipeline in population genome studies:

- select multiple strains of one species;
- collect the accessions of each strain and sequence all accessions;
- map the filtered raw sequences to a reference genome sequence and detect the SNP variants;
- investigate SNP variants and identify important genes or mutation site which can explain the change of significant phenotypic traits.

The first three steps prepare the data for the analysis in the last step, where there are multiple open-source software programs, like GATK (McKenna et al. 2010) and SOAP2 (Li et al. 2009) that can be used to help in step 3.

The last step is the most critical to identify important genes. It proceeds in five stages:

1. constructing the whole-genome phylogenetic tree T_r;
2. selecting a genome region to construct the phylogenetic tree T_l;
3. comparing these two trees and identify the tree difference including structure difference and edge length difference;
4. understanding the differences by mapping biological classification to T_l;
5. examining the SNP variants of the leaf nodes in T_l with large inconsistency and checking whether the mutation at the loci of these SNP variants can result in the change of amino acid.

Often the whole genome is used to construct tree T_l. The last four stages iterate until there are no more interesting findings, where the third and fourth stages are interleaved together. When T_l is constructed from the whole genome, biologists are interested in T_r which has large differences with T_l, and then they would find which gene leads to the difference by taking biological classification into consideration. Note that biological classifications are based on phenotypic traits and they are not always consistent with the clusters indicated by the whole-genome phylogenetic tree. Currently, the last three stages are performed manually and this is time-consuming even if the tree is small (dozens of nodes).

3.3 Data questions

We have identified 9 questions that biologists ask to gain insights at different stages of the genome analysis workflow, shown in Table 1. These questions were gathered from interviews with our biologist collaborators about their data analysis methods.

Questions Q1 and Q2 attempt to understand the whole-genome phylogenetic tree. Q3 and Q4 are about the comparison of two phylogenetic trees generated by different genome regions. Q5 through Q10 can lead to insight about the genome region explored in Q3 and Q4. Q5 and Q6 pertain to comparing tree structures with known biological classification. Once the nodes with large consistency are identified, Q7 through Q9 are direct questions about SNP variants of the investigated strains.

Taken all questions as a whole, we can see that evolutionary genome analysis involves four components: phylogenetic tree visualization (Q1 and Q2), phylogenetic tree comparison (Q3, Q4 and Q10), phylogenetic tree classification exploration (Q5 and Q6) and SNPs exploration (Q7 through Q9). Although there are software programs that can do some individual tasks, there is no tool that can support all of them.

We use these questions to motivate and justify our visualization design. These questions provide guidance for what information is required to solve the task at each stage of evolutionary genome analysis.

4 Algorithm

In this section, we describe and discuss new algorithms for computing the leaf node difference between two phylogenic trees from one population and calculating the inconsistency between the phylogenic tree and the given biological classification which group strains into different clusters. With them, Q3 and Q4 which involve the difference and inconsistency of leaf nodes can be quantitatively measured.

4.1 Tree–tree comparison

Two phylogenetic trees to compare are generated by genome regions of one species, and thus, they have the same number of leaf nodes. As each leaf node is labeled by a name, we can easily associate two leaves with the same name. Instead of computing the similarity between two internal nodes (Munzner et al. 2003), our collaborators are more interested in identifying if and how far leaf nodes have been moved. The main reason is that neighboring nodes have similar evolutionary histories. If the path length between two neighboring nodes in the other tree becomes large or small, it indicates that their evolutionary histories become different and some genes may be responsible for this change.

Suppose we have two phylogenetic trees T_l and T_r, where the leaf nodes of them come from the same strains, and T_r is the reference tree. If two nodes are associated with the same label in these two trees, we both name this node as υ_l and υ_r. To measure the change of the leaf node $\upsilon_l \in T_l$, we take the sibling node v_r of υ_r in the reference tree T_r as the reference node. For each υ, the difference is defined as:

$$d(\upsilon) = |L(\upsilon_l, v_l, T_l) - L(\upsilon_r, v_r, T_r)| \tag{1}$$

Table 1 Questions for the evolutionary genome analysis

	Question
Q1	How does the phylogenetic tree correspond to the distance matrix?
Q2	What is the classification label of each leaf node?
Q3	What is the change of the leaf node from one phylogenetic tree to the other tree?
Q4	How large the structural difference between two phylogenetic trees?
Q5	How is the phylogenetic tree consistent with the biologist classification?
Q6	What is the classification inconsistency of one leaf node in the phylogenetic tree generated with one genome region?
Q7	Which SNPs make leaf nodes with inconsistent classifications?
Q8	What is the gene id of the selected SNP variants?
Q9	Is there any SNP variants with the same loci among multiple strains?

where v is the sibling nodes of υ in tree T_r and is a leaf node, and $L(\upsilon, v, T)$ refers to the number of edges traversing from node υ to v on the tree T.

Figure 1a shows an example, where the edges colored in blue are the paths between nodes A and B, and the difference between them is 1. We can see that if two leaf nodes are sibling nodes, they have the same differences.

If v is a internal node, we take its all descendant leaf nodes $S(v) = \{v_1, \ldots, v_m\}$ as the reference nodes defines the distance as

$$d(\upsilon) = \frac{1}{m}\left|\sum_{i=1}^{m} L(\upsilon_l, v_{il}, T_l) - L(\upsilon_r, v_{ir}, T_r)\right|. \tag{2}$$

Figure 1b shows an example, where the sibling node of node C is an internal node in the reference tree and $v = \{A, B\}$. Since the neighboring distance between nodes B and C has not been changed, the difference of node C is less than node B shown in Fig. 1a. Note that, the difference on node D is zero although the leaf nodes of its sibling node B, C have been changed. This is reasonable because it is still grouped with nodes A, B, C together.

To measure the structures difference between T_l and T_r, we define it as the sum of the difference of all leaf nodes $\sum_i d(\upsilon)$. This metric is useful for biologists to select genome regions. They are more interested in the phylogenetic tree which has a large difference with the whole-genome tree, which means the selected genome region might be responsible for the change of the evolutionary history.

4.2 Tree classification comparison

Given a phylogenetic tree, the biologists will check how it is consistent with the given biological classification, which multiple strains are split into different classes, such as wide and cultivated. Often, the tree structure generated by a genome region usually has large inconsistency with such classification. To help the scientist quickly identify the inconsistent nodes, we compute the classification inconsistency of each leaf node μ by measuring the inconsistency between it and its sibling node v:

$$p(\mu) = 1/2^{L(\mu,v,T)-2} * \delta(\mu, v) \tag{3}$$

where $\delta(\mu, v) = 1$ if nodes μ and v have different labels, otherwise $\delta(\mu, v) = 0$. As shown in Fig. 2b, the inconsistency between two leaf nodes with the difference label is 1 according to the classification label shown in Fig. 2a.

If the sibling node v is an internal node, we measure its inconsistency by comparing the labels of its all leaf nodes $S(v) = \{v_1, \ldots, v_m\}$ with the μ:

$$p(\mu) = \sum_{i=1}^{m} 1/2^{L(\mu,v_i,T)-2} * \delta(\mu, v_i). \tag{4}$$

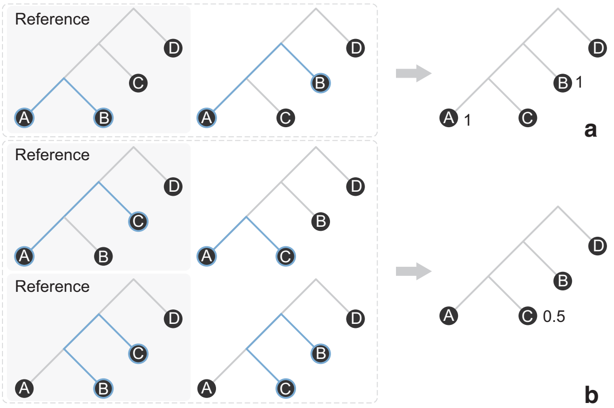


Fig. 1 We show the calculation of differences between leaf nodes on two trees where the left is the reference and the middle is the compared tree. **a** The sibling nodes of nodes A, B are leaf nodes, where the difference of both nodes A and B is 1. **b** The sibling node of node C is an internal node whose leaf nodes are nodes A and B, where the difference of node C is 0.5

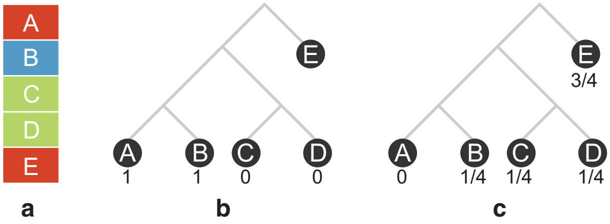


Fig. 2 We show the calculation of inconsistency between leaf nodes on a tree with a classification. **a** The labels of five leaf nodes; **b** the inconsistencies of both nodes A, B are 1 due to different labels, while the inconsistencies of both nodes C, D are 0; **c** the inconsistency of node E is 3/4 where nodes B, C, D contribute 1/4, respectively

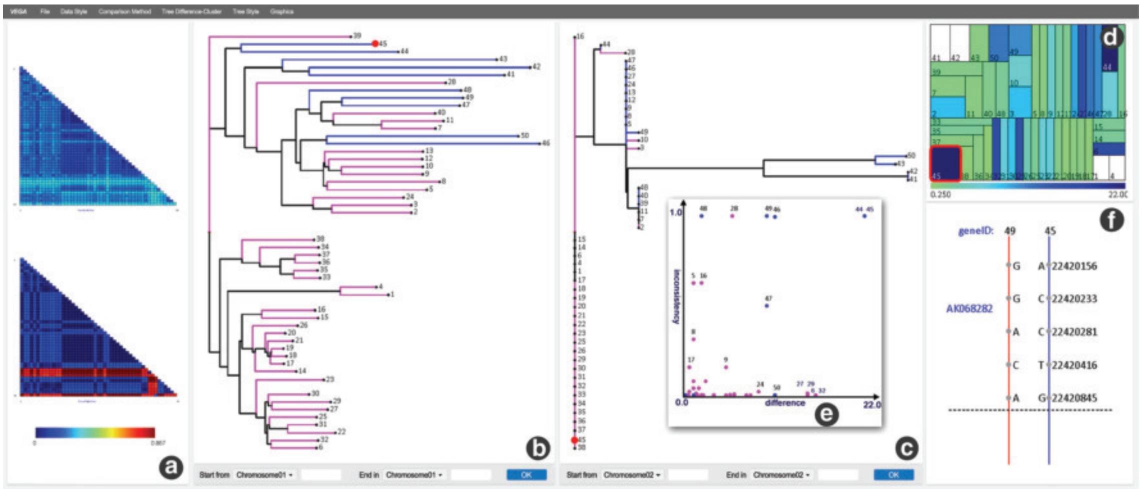

VEGA File Data Style Comparison Method Tree Difference Cluster Tree Style Graphics
a
b
c
d
e
f
Start from Chromosome01 End in Chromosome01 OK
Start from Chromosome02 End in Chromosome02 OK
inconsistency
difference
1.0
0.0
22.0
0.250
geneID: 49 45
AK068282
G A 22420156
G C 22420233
A C 22420281
C T 22420416
A G 22420845

Fig. 3 Investigating the genome-wide variation pattern of 50 strains of cultivated and wild rice by comparing the phylogenetic trees generated by the whole genome and a selected gene, respectively. **a** The side-by-side comparison of the distance matrices with heatmaps; **b**, **c** side-by-side comparison of two phylogenetic trees, where the edge color indicates the classified label of each node (pink is cultivated and blue is wild); **d** treemap shows the difference values of leaf nodes of the tree in **c**; **e** the scatter plot shows the distribution of difference and classification inconsistency within leaf nodes. **f** The SNP variants between the selected leaf nodes, corresponding to the strains 49 and 45, where all variants belong to one gene. When the biologist selects a rectangle with the annotation 45, the corresponding nodes in two trees both are automatically highlighted shown in red

$p(\mu)$ is 1 if all leaf nodes have the different label with node μ. Figure 2c shows how each leaf node contributes to the inconsistency of node E, where each node of B, C, D contributes 1/4.

5 Visual design

Figure 3 shows the interface of our VEGA system, which consists of four components: side-by-side heatmap views (Fig. 3a), side-by-side phylogenetic tree comparison views (Fig. 3b, c), treemap view (Fig. 3d), 2D scatter plot view (Fig. 3e) and SNP variant view (Fig. 3f). Two compared trees can be quickly generated after the user specifies the range of chromosome regions. In the following section, we describe in detail our interactive visualization techniques.

5.1 Visualizing phylogenetic tree

To help the user explore tree structure, we provide four different visualization styles: orthogonal layout, radial layout and both orthogonal and radial layout with unit edge length. As demonstrated by Burch et al. (2011), orthogonal layout is more effective in showing the hierarchy than the radial layout. Since a phylogenetic tree is an unrooted tree, the user is allowed to set different nodes as the root. After re-rooting, the nodes on subtree in the orthogonal layout are moved in the vertical direction, while they are just rotated in

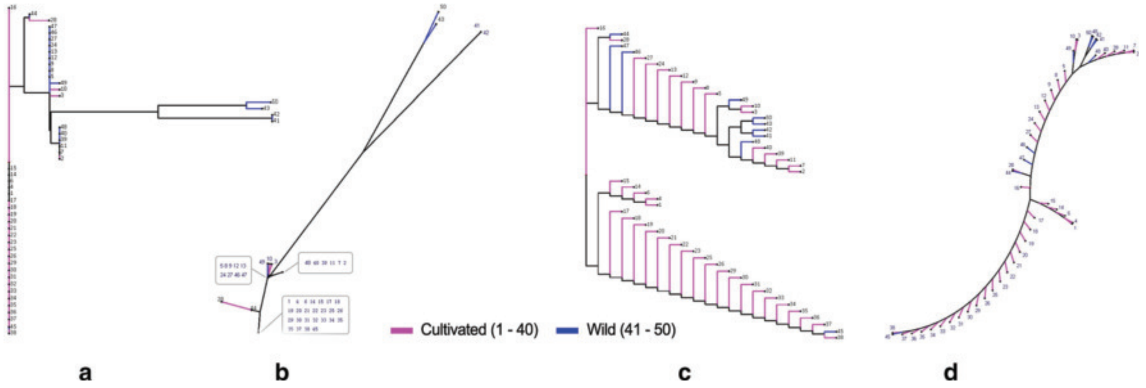


Fig. 4 Four different layouts of phylogenetic tree where all strains have been separated into two classes: wild (blue) and cultivated (red). **a** Orthogonal layout; **b** radial layout; **c** orthogonal layout with unit distance; **d** radial layout with unit distance

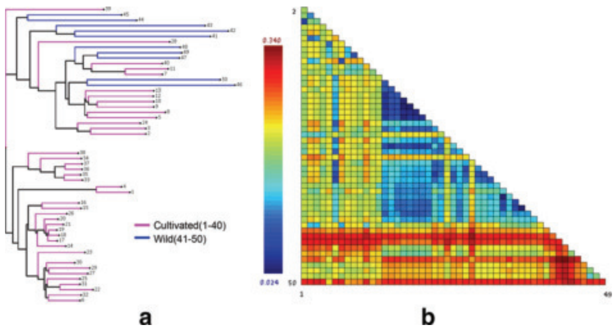


Fig. 5 The phylogenetic tree (**a**) constructed by whole-genome sequence and the corresponding distance matrix (**b**)

radial layout. This difference makes the radial layout better in revealing the cluster structures. Figure 4 shows an example, node 16 has a large distance with the bottom branches in Fig. 4a, but these nodes are grouped to one point in Fig. 4b. To help the user see the grouped nodes, we annotate their names in a nearby box shown in gray in Fig. 4b. In contrast, nodes with the edge length of 0 look like being placed in the same depth of the tree, which could mislead user understanding of the tree hierarchy (Fig. 4a); thus, setting the edge length of all nodes to a unit length enables us to clearly see the hierarchy, as shown in Fig. 4c, d, but this style ignores the original distances between nodes. Thus, we suggest combining these four styles together in exploring tree structure.

Although Darwin and Mega allow the user to manually colorize each node, it cannot directly show the classification on the tree at once. To address this problem, we use the classified label to colorize the edge linked to the leaf node. As shown in Fig. 4, the edge with pink color means the connected leaf node is a wild strain and the edge with blue color means the connected leaf node is a cultivated strain. This color-coding strategy answers question Q2.

We also visualize the distance matrix used for phylogenetic tree construction with a heatmap. This enables the biologists to see how the tree structure corresponds to the distance matrix (Q1) as well as doing a side-by-side comparison of the distance matrices. Figure 5 shows the whole-genome phylogenetic tree (Fig. 5a) and the corresponding distance matrix (Fig. 5b). From Fig. 5b, we can observe three groups: 1–13, 14–40 and 40–50, which correspond to different parts of the tree in Fig. 5a.

5.2 Side-by-side comparison

Besides showing two phylogenetic trees in side-by-side fashion, we also allow the user to make a side-by-side comparison of the distance matrices. To facilitate the comparison, we compute the minimal and maximal from two matrices and then use the same color map to create the heatmaps. As demonstrated in

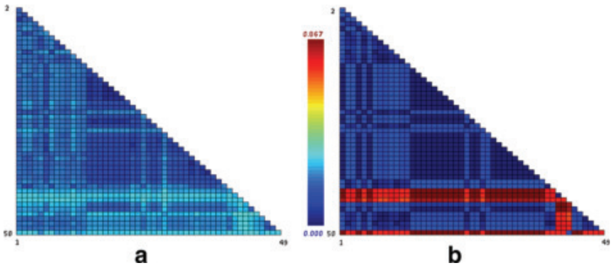


Fig. 6 Side-by-side comparison of distance matrices, which are used to construct the phylogenetic trees in Figs. 4 and 5a, respectively

Fig. 6, the color ranges from blue to red, indicating distances various from small to large, and in Fig. 6b, four nodes with large distances to other nodes are clearly shown which helps in filtering strains of interest during analysis.

5.3 Visualizing node value

By using the measures defined in Sect. 4, we compute two numerical attributes: difference and inconsistency for each leaf node. However, it is hard to encode them into tree node or edge. To help the user explore these values, we provide two views: treemap view and scatter plot view.

Treemap view Treemap is an efficient way in visualizing node values while presenting the hierarchy. We map the obtained numeric attributes to the color of the corresponding rectangular while using tree depth to determine the size of the rectangular. This facilitates the user to perceive the nodes with large values, especially when the number of leaf nodes is large. Note that, the color of the box is white if the value of its corresponding node is zero. By comparing the nodes of the phylogenetic tree in Fig. 4 with the one generated with the whole genome in Fig. 5, the result of difference and inconsistency is shown in Fig. 7a, b, respectively. Together with the scatter plots in Fig. 7c, biologists can easily discover nodes of their interests, for example nodes with both higher difference and inconsistency; then, they can proceed with seeking the genes that cause the inaccuracy of the traits identification during biological classification.

Scatter plot View We provide a 2D scatter plot view, whose axes correspond to difference and inconsistency, respectively. This view helps users to explore nodes with small difference but large inconsistency or the other way around, as illustrated in Fig. 7.

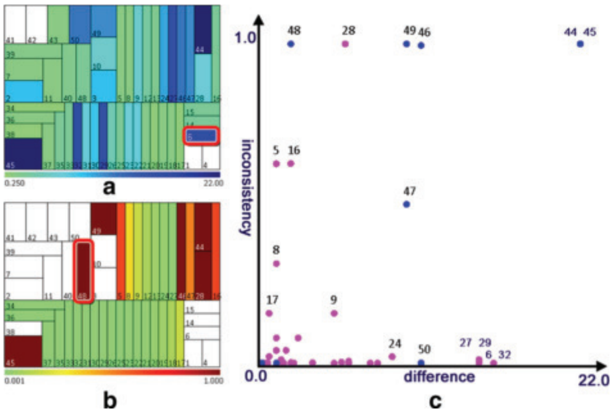
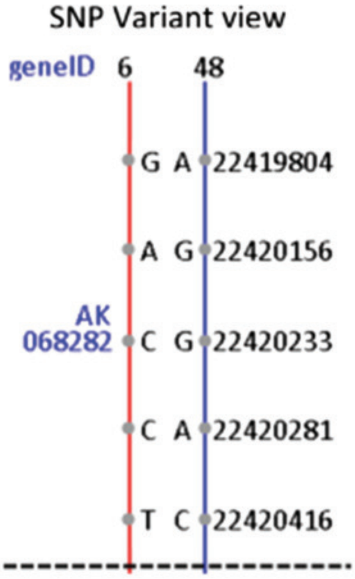


Fig. 7 Visualizing the node values generated by comparing the trees in Fig. 4 with the one in Fig. 5a. **a** Treemap for the value of difference defined on the phylogenetic tree in Fig. 4; **b** treemap for the value of inconsistency defined on the phylogenetic tree in Fig. 4; **c** the 2D scatter plot shows the distributions of the node difference and inconsistency, given in **a**, **b**

5.4 Visualizing SNP variants

After selecting two nodes of interest, the biologists are interested in the differences of the SNP variants in the selected genome region. We provide a light SNP variant view (see figure below) to help the user explore genotype, loci and gene ids of these variants. Since the DNA sequence is stored as a linear array, SNP variants are encoded into paired vertical lines, where the genotypes of two variants and their loci are annotated at each position. To answer Q8, we annotate different genes with a black dashed line (see inset). Since the biologists would like to compare SNP variants of multiple paired strains (Q9), we link the same loci appeared in different pairs with gray solid lines, as shown in Fig. 9e.



5.5 Interaction

Besides supporting basic tree interactions, such as translation, panning and zooming, VEGA supports various advanced interaction.

Local Zooming Although using four styles enables the user to learn the cluster, hierarchy and distance between nodes, style switching is a huge burden to the user. To alleviate this issue, we introduce local zooming. In this manner, the user selects a region in the tree and then the hierarchy with the unit distance will be automatically shown in neighboring empty regions. As demonstrated in Fig. 8, a region is selected by a red circle and then the hierarchy with unit distance is shown in a red box.

Dynamic Classification The biologists usually have more than one kind of classification for one species. To support the analysis with different classifications, we save the classification information with an XML file. With that, the user can dynamically change this file and update the classification. On the other hand, we allow the user to dynamically change the label of nodes. During the analysis, the biologists may change the

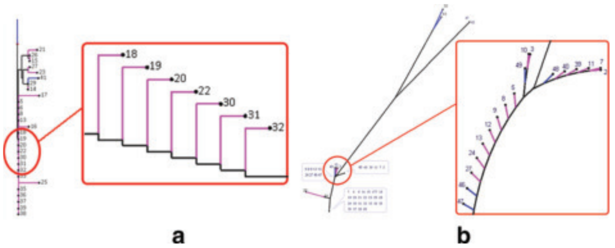


Fig. 8 Local zooming of a selected region in phylogenetic trees with styles: **a** orthogonal and **b** radial

pre-loaded classification if the classification has a large inconsistency with the whole-genome phylogenetic tree. It helps the user to dynamically revise the prediction and classification results made before the analysis.

Brushing and Linking Our system consists of seven views, where the treemap view can be set to display node difference or inconsistency. After selecting a leaf node in one tree, the corresponding leaf node in the other tree will be highlighted. Likely, selecting some boxes in heatmap or treemap views or some points from 2D scatter plot view, the corresponding leaf nodes are highlighted, as demonstrated in Fig. 7.

6 Case studies

In this section, we demonstrate the capabilities of VEGA on two datasets provided by our biologist collaborators, who are active researchers in whole-genome analysis. In these two datasets, the major goals are both to identify genes that have been influenced by artificial selection during domestication, which should be valuable for plant breeding.

6.1 Rice

As the staple food for more than half the world's population, rice (*Oryza sativa* L.) has undergone substantial phenotypic changes in grain size, color, shattering during domestication. The identification of the major genes responsible for these traits should be valuable for advancing rice breeding technology so as to improve rice yield. Our biologist collaborator attempts to identify important genes from the whole-genome sequences of 50 accessions of cultivated and wild rice. After alignment, the obtained number of SNP variants is around 6.5 million, located at 24,209 genes. Our collaborator first identified several chromosome regions which potentially have the genes of interest and then manually compared the evolutionary history of each gene or small chromosome regions and analyze the related SNP variants. In the process of identifying these genes, our collaborator has discovered the evidence to classify the cultivated rice into two groups: indica and japonica.

Our collaborator first would like to see whether our method can quickly identify the genes of interest which have large differences with the whole-genome phylogenetic tree. After computing tree differences in a batch, he found one gene, named as *Os02g0567000*, which has the largest difference but has never been explored before. Figure 4 shows its resulted phylogenetic tree. By inspecting the hierarchical structure with local zooming in orthogonal layouts and comparing this tree with the whole-genome phylogenetic tree in Fig. 5a, he found some interesting structures in this tree that nodes on the bottom branch are grouped together, while other nodes on the above branch have large variations. This indicates that the cultivated strains colored in pink have similar genotypes, while the wild strains have high levels of polymorphisms. One main reason is that this gene in wild strains has not been affected by artificial selection. By further exploring the tree with local zooming, he spotted that strains from 14 to 40 are close to the wild strains 44 and 45 on the tree, while most of the strains from 1 to 13 are close to the wild strains 46, 47, 48 and 49. Since the cultivated strains can be further split into two groups: indica (1–13) and japonica (14–40), he concluded that these two groups might be cultivated from the strain 44 and 45, and 46, 7, 48, 49, respectively. He then verified this observation from the heatmap shown in Fig. 5. In all, he concluded that this gene might play an important role in determining domestication-related traits.

Then, he wondered whether our system can reveal more insight of one known gene, *sh4*, which influences the shattering trait. Figure 9a shows its resulted tree, where the nodes have been classified into four groups: indica (1–13), japonica (14–40), rufipogon (41–45) and nivara (46–50). There are some clear observations from the tree: (1) The group of indica is close to the wild strains; (2) the group of japonica is well separated from wild strains. Comparing this tree with the whole-genome phylogenetic tree in Fig. 5a, he found that all strains seem to have similar evolutionary histories. He then provided the reason that this gene is one of the domestication-related genes (Li et al. 2006). However, by inspecting the SNP variants shown in Fig. 9d, some nodes have large classification inconsistencies, like node 6 which is grouped with 45 (a wild strain). To investigate which loci result in such inconsistency, he further compared SNP variants of these outliers with one cultivated strain, 45. As shown in Fig. 9e, we can find one locus in two pairs of nodes: (6, 22) and (27, 22), two loci in one pair of nodes (47, 22). Thus, he concluded that node 6 is still close to the cultivated strain located at the bottom tree branch, although it is grouped with 45. Meanwhile, he found that the mutation at the loci can lead to the amino acid change. All these messages have not been

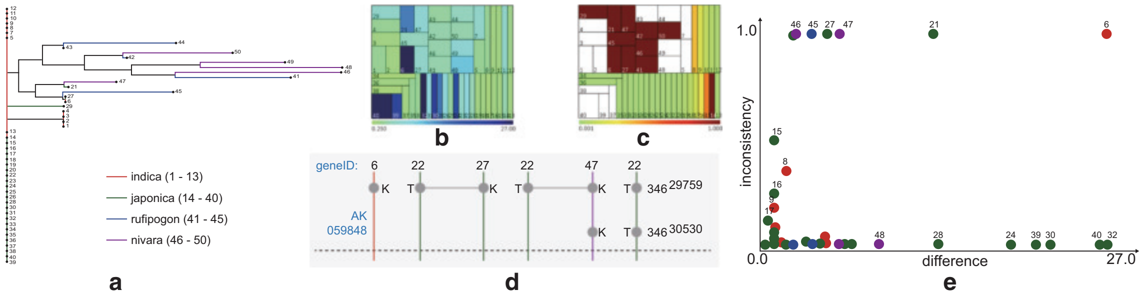


Fig. 9 Visual comparison of the tree shown in **a** with the one Fig. 5a. **a** The phylogenetic tree generated by the gene *sh4*; **b**, **c** treemaps for the values of difference and inconsistency defined on the phylogenetic tree in Fig. 9**a**; **d** the 2D scatter plot shows the distributions of the node difference and inconsistency, shown in **b**, **c**; **e** the SNP variants between the paired strains: 6 and 22, 21 and 22, 47 and 22

revealed by previous studies, and he planned to further combine biological experiments to verify these findings.

6.2 Cucumber

Compared to rice, cucumber has large diversity and its classification is related to the geographic distribution. To gain a comprehensive insight into the genetic basis of domestication, our second collaborator has collected the sequences of 115 cucumber strains sampled from 3342 accessions worldwide. After alignment, these data include around 3.3 million SNP variants within 2336 genes. The collected strains can be classified into 4 geographic groups: East Asian (1–37), Eurasian (38–66), Indian (67–96) and Xishuangbanna (97–115). Among these groups, the Xishuangbanna group uniquely accumulates β-carotene in its fruit. Since cucumber is indigenous to India (Sebastian et al. 2010), biologists assume the Indian group is close to the wild type, while the other three groups belong to the cultivated type.

Our collaborator performed the genome-wide reduction in genetic diversity (Chia et al. 2012) to identify potential chromosome regions and then manually investigated each gene located at these regions to see whether it is related to some domestication-related traits. He has identified several genes and chromosome regions that might be involved in domestication, but he is not sure the functions of them. Hence, he first used our VEGA to verify his finding and then investigated unknown chromosome regions.

He has found the gene, *Csa3G183920*, which encodes a putative β-carotene hydroxylase[33], but he did not understand how this gene affects the trait. By comparing the phylogenetic tree (Fig. 10b) generated by this gene with the whole-genome phylogenetic tree (Fig. 10a), two interesting patterns are revealed: (1) Most of the nodes with large differences belong to the Indian group. This evidence supports that the Indian group is quite different from the other three cultivated groups. (2) The variations of the Xishuangbanna group are relatively small, and all nodes are close to each other. He hypothesized that this gene might be a particular gene of Xishuangbanna group. To further explore this gene, he selected other strains from the other groups to see which locus determines the trait in SNP variant view. After comparing multiple strains with the strains from Xishuangbanna group, we can see all strains from Xishuangbanna have the locus, 12845300 (see Fig. 11a). Combining with genetic codes, he found that the genotype in Xishuangbanna group at this locus affects the aspartic acid. Moreover, the mutation at this site can enrich β-carotene hydroxylase[33], corresponding to a particular trait of yellow fruit with a higher level of β-carotene. In contrast, the genotype at this locus in the other group affects alanine, which is not related to this trait

Through biological experiments, biologists have found that two chromosome regions control the bitterness synthesis and transfer: bi-1 (*bi*) and Bt-1 (*Bt*), respectively. But it is still unknown which region plays a more important role in domestication. To achieve this task, our collaborator first separately compared the trees generated by these two regions with the whole-genome phylogenetic tree (see Fig. 10a). Observing the tree generated with the *Bt* region (see Fig. 10c), we can see that the Indian group has been split into two groups, while the other nodes are close to each other. This indicates that the Indian group has a large variation and cultivated groups are consistent on this gene. In contrast, the tree (Fig. 10d) generated from the *bi* region is closer to the whole-genome tree, although there are some differences. From this comparison, he speculated that the *Bt* region might play a more important role in domestication because the *bi* region of all

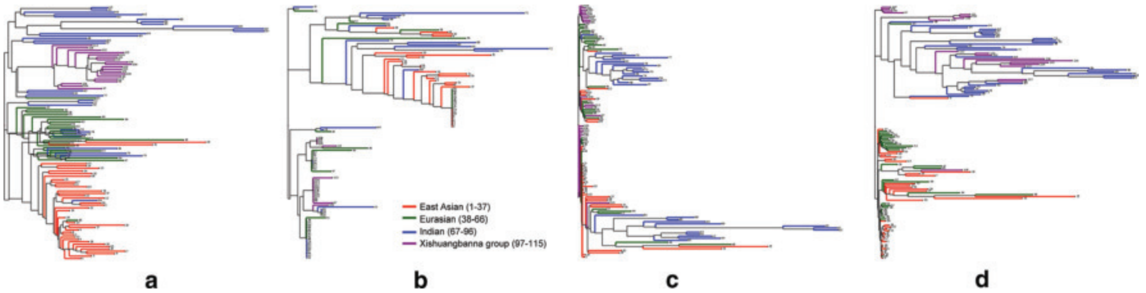


Fig. 10 The phylogenetic trees generated from different genome regions of cucumber. **a** the tree generated by the whole genome; **b** the tree generated by the *Csa3G183920* gene (**b**); **c** The tree generated by the *Bt* region; **d** the tree generated by the *bi* region

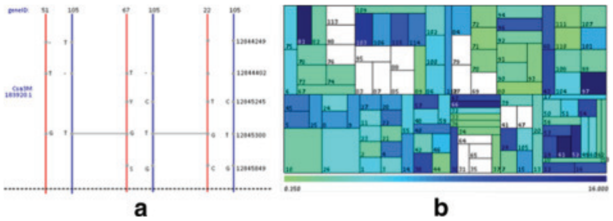


Fig. 11 **a** The SNP variants among the paired strains: 51 and 105, 67 and 105, and 22 and 105; **b** the node differences between two trees in Fig. 10c, d

cultivated groups seems to be selected. To verify this speculation, he further compared the phylogenetic trees generated by these two regions. Figure 11b shows the node differences between these two trees, where we can see most of the nodes with small difference values belong to the Indian group. That means these two trees have different evolutionary histories for most of the cultivated strains. Combining his biology knowledge, he gave the following explanation for this difference. *As we know, the bitter substances are useful in preventing diseases. If these bitter substances of the cultivated groups can be transferred from fruit to foliage, their survival ability would be greatly improved.* He concluded that the *Bt* region is an artificially selected region during domestication and plays a more important role in determining bitterness.

6.3 Expert feedback

Our collaborated biologists have many years of experience in the field of evolutionary genome analysis. However, they did not see any mature software that can allow them to interactively specify a chromosome region and then quickly get a phylogenetic tree. Currently, they first export the SNP variants located at the specified chromosome region into a file, then import this file to Darwin or Mega before drawing a tree and manually compare with another tree drawn by these tools. In contrast, our VEGA not only supports the user to interactively generate a tree and map the classification to the tree but also quantitatively measures the difference with another tree, especially the change of leaf nodes which are desired by them.

Our interactive tool is a comfortable approach for leaf node selection, which enables biologists to quickly locate the nodes of the interest and explore their tree structure. The most important is that our VEGA facilitates them to quickly select strains from tree comparison and compare SNP variants. They told us that this function can potentially accelerate the procedure of gene discovery and they plan to use VEGA in their daily research.

They also provide us some advice that can improve the applicability of the system. Since the genome analysis pipeline involves the chromosome region filtering which is currently done manually, our collaborators suggested us to include this process into the system to enable the biologists perform the whole-genome analysis pipeline with our system.

7 Conclusion and future work

Biologists working in the field of whole-genome analysis are faced with the comparison of phylogenetic trees that requires them to identify the leaf nodes with a large difference and large classification inconsistency. Currently, this challenging task is performed manually, which hinders the progress of important gene discovery. To address this problem, we present a novel characterization of the workflow and identify the questions met at each stage. Guided by this contribution, we propose two algorithms to compare two phylogenetic trees with the same number of leaf nodes and compare one tree with its classification. Configured with obtained node values of difference and inconsistency, our VEGA allows the user to intuitively explore the tree and discover important genes and loci. We present two case studies with biologist collaborators and demonstrate the capabilities of VEGA in assisting biologists to discover important genes.

According to our collaborators' suggestions, it would be interesting future work to integrate the automatic chromosome region filtering method into VEGA so that the biologists can perform the entire genome analysis pipeline in our system. It would also be useful to integrate more variant attributes into SNP variant view, like exons. And we would like to improve the scalability of our tree drawing as well. For sometimes they would analyze sequencing data in large scales and take their collected sequences of 3000 strains of one species for instance. Although our comparison methods and interactive exploring can still perform in real time, the tree construction and drawing process are relatively slow due to the time complexity of neighbor joining and techniques we used to draw the tree. Our collaborators then indicate that the response time is still within tolerable time, and it is not common to perform analysis on such a big dataset. Usually they would regroup the strains and analyze each group separately, which can also improve the efficiency in the following procedure.

Acknowledgements This work is supported by the Grants of NSFC (61772315, 61602273), Shenzhen Science and Technology Program (JSGG20170412170711532) and the Open Research Fund of Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University (BKBD-2017KF02).

References

Bachmaier C, Brandes U, Schlieper B (2005) Drawing phylogenetic trees

Barlow T, Neville P (2001) A comparison of 2-d visualizations of hierarchies. In: IEEE symposium on information visualization. IEEE, pp 131–131

Bremm S, von Landesberger T, Heß M, Schreck T, Weil P, Hamacherk K (2011) Interactive visual comparison of multiple trees. In: 2011 IEEE conference on visual analytics science and technology (VAST). IEEE, pp 31–40

Burch M, Konevtsova N, Heinrich J, Hoeferlin M, Weiskopf D (2011) Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study. IEEE Trans Vis Comput Graph 17(12):2440–2448

Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC et al (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44(7):803–807

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER et al (2012) Music: identifying mutational significance in cancer genomes. Genome Res 22(8):1589–1598

Ethan C, Jianjiong G, Ugur D, Gross Benjamin E, Sumer Selcuk Onur, Aksoy Bülent Arman, Jacobsen Anders, Byrne Caitlin J, Heuer Michael L, Larsson Erik et al (2012) The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data

Ferstay JA, Nielsen CB, Munzner T (2013) Variant view: visualizing sequence variants in their gene context. IEEE Trans Vis Comput Graph 19(12):2546–2555

Fiume M, Williams V, Brook A, Brudno M (2010) Savant: genome browser for high-throughput sequencing data. Bioinformatics 26(16):1938–1944

Graham M, Kennedy J (2010) A survey of multiple tree visualisation. Inf Vis 9(4):235–252

Guerra-Gómez JA, Pack ML, Plaisant C, Shneiderman B (2013) Visualizing change over time using dynamic hierarchies: Treeversity2 and the stemview. IEEE Trans Vis Comput Graph 19(12):2566–2575

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at ucsc. Genome Res 12(6):996–1006

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9(4):299–306

Li Z-K, Zhang F (2013) Rice breeding in the post-genomics era: from concept to practice. Curr Opin Plant Biol 16(2):261–269

Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. Science 311(5769):1936–1939

Li R, Chang Y, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J (2009) Soap2: an improved ultrafast tool for short read alignment. Bioinformatics 25(15):1966–1967

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Res 20(9):1297–1303

Munzner T, Guimbretière F, Tasiran S, Zhang L, Zhou Y (2003) Treejuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. In: ACM transactions on graphics (TOG), vol 22. ACM, pp 453–462

Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T (2010) Visualizing genomes: techniques and challenges. Nat Methods 7(3s):S5

Nucleic acid notation. http://en.wikipedia.org/wiki/Nucleic_acid_notation

Parr CS, Lee B, Campbell D, Bederson BB (2004) Visualizations for taxonomic and phylogenetic trees. Bioinformatics 20:2997–3004

Penny D, Hendy MD, Steel MA (1992) Progress with methods for constructing evolutionary trees. Trends Ecol Evol 7(3):73–79

Perrier X, Jacquemoud-Collet JP (2006) Darwin software

Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Xingfang G et al (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet 45(12):1510

Robinson O, Dylus D, Dessimoz C (2016) Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. Mol Biol Evol 33(8):2163–2166

Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S et al (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464(7288):587

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sebastian P, Schaefer H, Telford IRH, Renner SS (2010) Cucumber (*cucumis sativus*) and melon (*c. melo*) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. Proc Natl Acad Sci 107(32):14269–14273

Shneiderman B (1998) Tree visualization with tree-maps: a 2-d space-filling approach. Technical report

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14(2):178–192

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

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30(1):105

Xun X, Hou Y, Yin X, Bao L, Tang A, Song L, Li F, Tsang S, Kui W, Hanjie W et al (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 148(5):886–895