# SynMapN: Interactive Visual Comparison for Multiple Genomes

Mingwei Li* Asher Keith Haug-Baltzell† Eric Lyons‡ Carlos Scheidegger§

University of Arizona

**ABSTRACT**

Interactive visualization has become a powerful means to explore syntenic relationships among two genomes, with a variety of available tools for domain scientists to employ. However, these tools do not tend to scale well in the case where *many* genomes are compared against one another. This poster describes ongoing efforts to build techniques and tools to help genomicists understand sets of genomes and their syntenic relationships. Our main contribution is a mechanism that defines *set distances*: this can be used to compare entire genomes to one another, as well as *sets of genomes* to each other. Currently, we use this mechanism to generate dimensionality reduction visualizations. We discuss limitations of this approach, as well as future directions.

**Index Terms:**

## 1 INTRODUCTION

In comparative genomics, one of the visual methods that helps studying the structural relations between two genomes is called the syntenic dotplot [2, 3]. It is a scatterplot that depicts matched genes between two genomes; these matched genes imply *syntenic regions*: those that likely originated from the same ancestor. During genomic analysis, a measure called synonymous mutation rate between two genes (*ks*) is computed. Because many codons translate to the same amino acid (for example, TCT and TCC both translate to serine), it is possible that a mutation does not change the amino acid that is encoded in the gene: this is a "synonymous" mutation. The *ks* value is a rate of such mutations, normalized by gene sizes. Because those mutations are mostly harmless, they tend to not change selection pressure on the genomes. At the same time, they are transmitted hereditarily, and so they can be seen as "biological clocks", and thus can be used to estimate evolutionary relations. The larger the *ks* value is, the longer it has been since these two genomes diverged. In the dotplot, two axes represent gene locations of two genomes respectively. Each dot on the plot represents a match between two genes, and the dot colors usually encode *ks* values.

In some ideal cases, a perfect alignment of two genomes can be observed by seeing a line along the diagonal of a plot, similar to a plot of function $y = x$. This means that the two entities in concern have same genes presented in order. More interesting events like duplication and inversion of genes can be easily spotted from the syntenic dotplot as well (Fig. 3). In other cases, it is also common to see only sparse dots, indicating no significant alignment in the two genomes.

Although tools such as syntenic dotplots and MizBee [8] are widely useful, especially in comparing two genomes, there are not many tools designed to explore more than two genomes at once. Because of projects such as the thousand genomes project [1], there is a demand for understanding the relationships among many genomes. Adding one more axis of genome to make a 3D scatter plot, for example, may work in some (very limited) cases. We would see a line in diagonal in a comparison among human, chimpanzee and gorilla genomes [4]. However, it is not visually intuitive to find out complicated relations in three species in general. In fact, Tory et al. [10] showed that a 3D landscape works not as well as a 2D map in specific tasks such as search and point estimation. This casts doubt on whether three-dimensional techniques are useful for other tasks such as synteny visualiatization. Moreover, a 3D scatterplot clearly does not generalize to more than three genomes.

Traditional scatterplot matrix displays can be readily adapted for comparing multiple genomes, with each subplot being a syntenic dotplot of two of the genomes. However, scatterplot matrices do not scale up well beyond a moderate number of plots []. As the number of genomes increases, it becomes harder to tell the overall relation between the genomes.

In this poster, we describe an ongoing collaboration with domain scientists using multiple-genome comparisons in two use cases. The first use case involves a comparison of about one hundred *arabidopsis thaliana* genomes. In this case, the differences between the genomes are encoded by a set of SNPs (single-nucleotide polymorphisms). The second use case involves 17 genomes, each of a different species of the plasmodium genus. In this case, the genome differences are more complex, often architectural (that is, involving inversions and duplications of entire portions of the genome), and need to be encoded by the entire synteny map.

*e-mail: mwli@email.arizona.edu
†email: ahaug@email.arizona.edu
‡email: elyons.uoa@gmail.com
§e-mail: cscheid@cs.arizona.edu

Figure 1: A SynMapN plot of various species. Note the distant location of Escherichia coli (E.coli), the cluster of chimpanzee, gorilla and human, as well as the cluster of cat, dog and horse

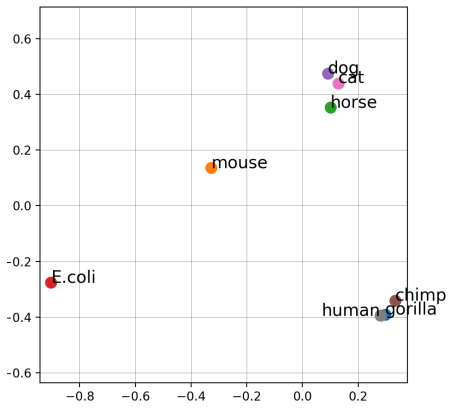


Figure 2: A SynMapN plot of various species. Note the distant location of Escherichia coli (E.coli), the cluster of chimpanzee, gorilla and human, as well as the cluster of cat, dog and horse

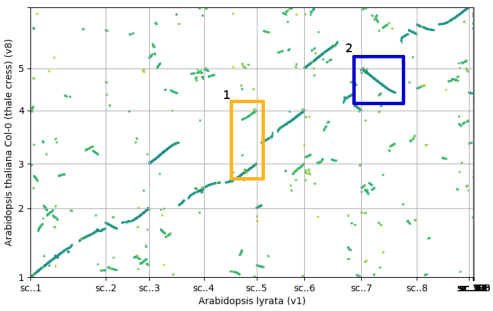


Figure 3: A syntenic dotplot of Arabidopsis lyrata and Arabidopsis thaliana. The orange region 1 shows a duplication in Arabidopsis thaliana with respect to Arabidopsis lyrata. The blue region 2 shows an inversion. The image is regenerated by ks data downloaded from [5].

## 2 Examples

Fig. 2 shows our first attempt of providing an overview of the genomes in concern. We define a similarity measure between two genomes, apply dimensionality reduction (Kernel PCA [9]), and try to visualize the relationships among genomes of interest through a genomic map where distances between genomes encodes their affinity.

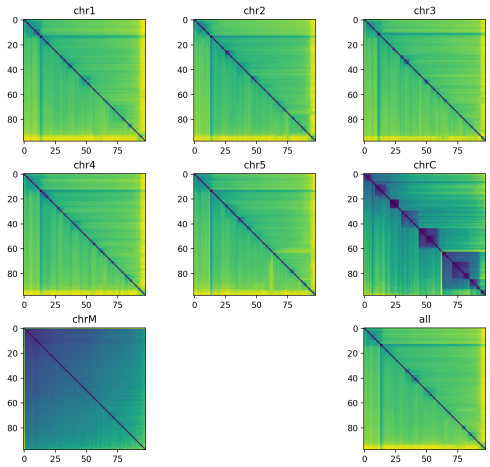


Figure 4: Distance matrices of 97 Arabidopsis thaliana of different ecotypes. Eight subplots are 7 individual chromosomes and a sum of them. Clusters and outliers can be spotted in chromosome C (subplot #6)

Fig. 4 and Fig. 5 show distance matrices and scatterplots (coordinates computed from MDS) of species from another dataset. The genomes are Arabidopsis of different ecotypes. When we break the entire genome into individual chromosomes and try to compute their pairwise edit distances, we see clusters in small chloroplast chromosomes (chrC, subplot #6 in Fig. 4 and Fig. 5). We can also see some common outliers among other plots.

## 3 Method

In the process of drawing a SynMap of two genomes, the program computes a measure called synonymous mutation rate (*ks*) describing a matching score for each pair of aligned genes between two genomes. This score ranges from 0 to $+\infty$, or arbitrary large number in data, with 0 means perfect alignment and infinity indicates no alignment or an error. We want to utilize these measures to build a notion of distance between two genomes.

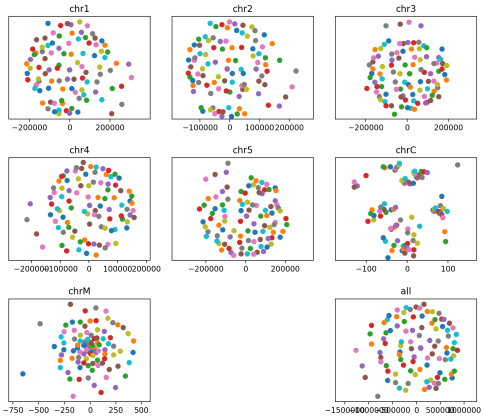


Figure 5: MDS plots of 97 Arabidopsis thaliana of different ecotypes.

For example, when comparing two genomes of human and chimp, we are given IDs of gene pairs and their synonymous mutation rate, denoted by *ks*, which can be seen as a measure of distance (Table 1)

Table 1: Data Input File

| Human gene ID | Chimpanzee gene ID | $ks_i$ |
|---|---|---|
| h1 | c1 | 0.0056 |
| h2 | c2 | 72.6574 |
| ... | ... | ... |

Now we want to find a function of many ks values that describes the distance between two genomes or equivalently, a notion of similarity or kernel between two entities.

$$similarity(Human, Chimp) = g(ks_1, ks_2 ...) \tag{1}$$

Generally the gene IDs in such table may be duplicated, indicating a match of a single gene from one genome with multiple genes in the other genome (duplication). To formalize the computation, we put *ks* values in matrix form

$$K = \begin{bmatrix} ks_{1,1} & ks_{1,2} & \dots & ks_{1,c} \\ \vdots & \vdots & \ddots & \vdots \\ ks_{h,1} & ks_{h,2} & \dots & ks_{h,c} \end{bmatrix}$$

Where $ks_{1,1}$ stores the ks value between the first gene of human and the first gene of chimpanzee, and so on. The indices of rows and columns are in the order of gene locations in ordered chromosomes. $h$ and $c$ are gene counts of human and chimpanzee respectively. One can think of $K$ as an image data of the SymMap plot (Fig. 3).

Now we want a function of $K$ that outputs a scalar value to describe the similarity between two genomes.

$$similarity(Human, Chimp) = f(K) \tag{2}$$

To define a function so that comparison of a genome to itself gives a similarity measure close to 1, we used $f$ to compute kernels between any two genomes, which are later used to plot Fig. 2

$$f(K) = \frac{\sum_{i,j} e^{-\lambda K_{ij}}}{\sqrt{c \times h}} \tag{3}$$

Where $\lambda$ is a scalar constant related to sensitivity of the similarity measure. $c$ and $h$ are the gene counts defined before.

Once we have the pairwise distances among multiple genomes (stored in matrix $M$), we have an implied space and then we project it onto a 2D screen using Kernel PCA. Let $M$ be the matrix of similarities of multiple genomes where each entry is computed by function $f$ defined above. Note that this is a symmetric matrix with diagonal entries set to 1, indicating complete identity in similarity measure of each genomes with respect to itself.

$$M = \begin{bmatrix} f_{human,human} & f_{human,chimp} & f_{human,cat} & \cdots \\ f_{chimp,human} & f_{chimp,chimp} & f_{chimp,cat} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

We then treat it as a kernel matrix and feed it into existing Kernel PCA algorithms and plot the first two principle components, which is shown in Fig. 2

## 4 Other Examples

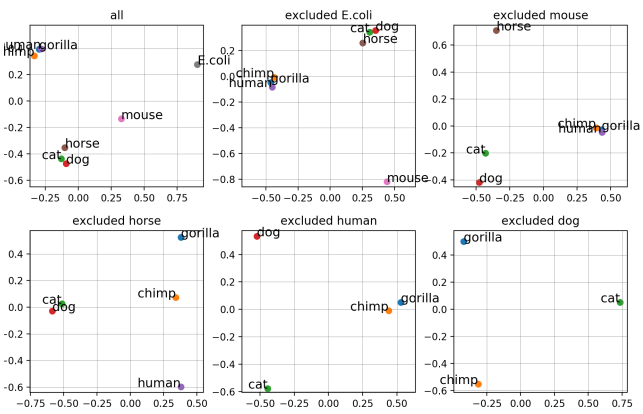


Figure 6: A SynMapN plot with subsets of species concerned. The similarity measure of certain pairs are the same among all plots above so the distance encoded are consistent.

In addition to the example plot in Fig. 2, we explore different subsets of the genomes among the genomes we concerned to show the capacities and constraints of SynMapN.

Plots in Fig. 6 shows subsets of the genomes we concerned with one genome excluded a time. The similarity of each pair of genomes are consistent among all the plots so the distances in the implied high dimensional space remains the same. But note that a 2D Kernel PCA plot reveals more of the true distances wist respect to an outlier through a certain plot, while potentially shrink the pairwise distances among other points. When an outlier is removed from the input of Kernel PCA in the next plot, another outlier shows up, indicating that previous plot did not show the true distance of the later outlier. For example, plot 2 shows affinity of horse, cat and dog, while with mouse excluded in plot 3, horse moves away from the cat-dog cluster.

SynMapN gives an overview of genomes while ignoring the details of the comparison. The detailed patterns within two genomes comparison such as inversions and duplications are not shown in SynMapN. The other restriction lies on the limitation of dimensionality reduction method we use. Kernel PCA and Classical MDS [7] are essentially linear projections that preserves the pairwise distances globally, so the distant dots reflects the dissimilarities of the entities while the closed dots may only be an artifact of the projection, thus misleadingly reflects closeness of entities in similarity. So SynMapN at this stage is very useful in finding one specie that are most dissimilar from the others while the clustering it plots does not always reflect the truth.

## 5 Future Works

In a complete visualization system, we expect to enable users to navigate through various visualizations for different levels of details, from high level SynMapN to very detailed comparison of two specific genomes in SynMap. To further increase the number of levels, we can split genomes into individual chromosomes and compare them in SymMapN. We are also exploring the possibilities of capturing certain features in a SynMap plot, for example, the diagonal alignments or anti-diagonal alignments, using image processing. More improvements on interaction can be done through studying observation level interactions [6], to enable users to specify desired clustering criteria through dragging points in the plot.

## Acknowledgments

## References


[1] 1000 genomes project. `http://www.internationalgenome.org/about`. Accessed: 2017-06-14.

[2] CoGepedia synmap. `https://genomevolution.org/wiki/index.php/SynMap`. Accessed: 2017-06-14.

[3] CoGepedia syntenic dotplot. `https://genomevolution.org/wiki/index.php/Syntenic_dotplot`. Accessed: 2017-06-14.

[4] Synmap 3d. `https://genomevolution.org/r/lf1i`. Accessed: 2017-06-14.

[5] Synmap example. `https://genomevolution.org/r/k7cr`. Accessed: 2017-06-14.

[6] A. Endert, C. Han, D. Maiti, L. House, and C. North. Observation-level interaction with statistical models for visual analytics. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 121–130. IEEE, 2011.

[7] J. B. Kruskal and M. Wish. *Multidimensional scaling*, vol. 11. Sage, 1978.

[8] M. Meyer, T. Munzner, and H. Pfister. Mizbee: a multiscale synteny browser. *IEEE transactions on visualization and computer graphics*, 15(6):897–904, 2009.

[9] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pp. 583–588. Springer, 1997.

[10] M. Tory, D. Sprague, F. Wu, W. Y. So, and T. Munzner. Spatialization design: Comparing points and landscapes. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1262–1269, 2007.