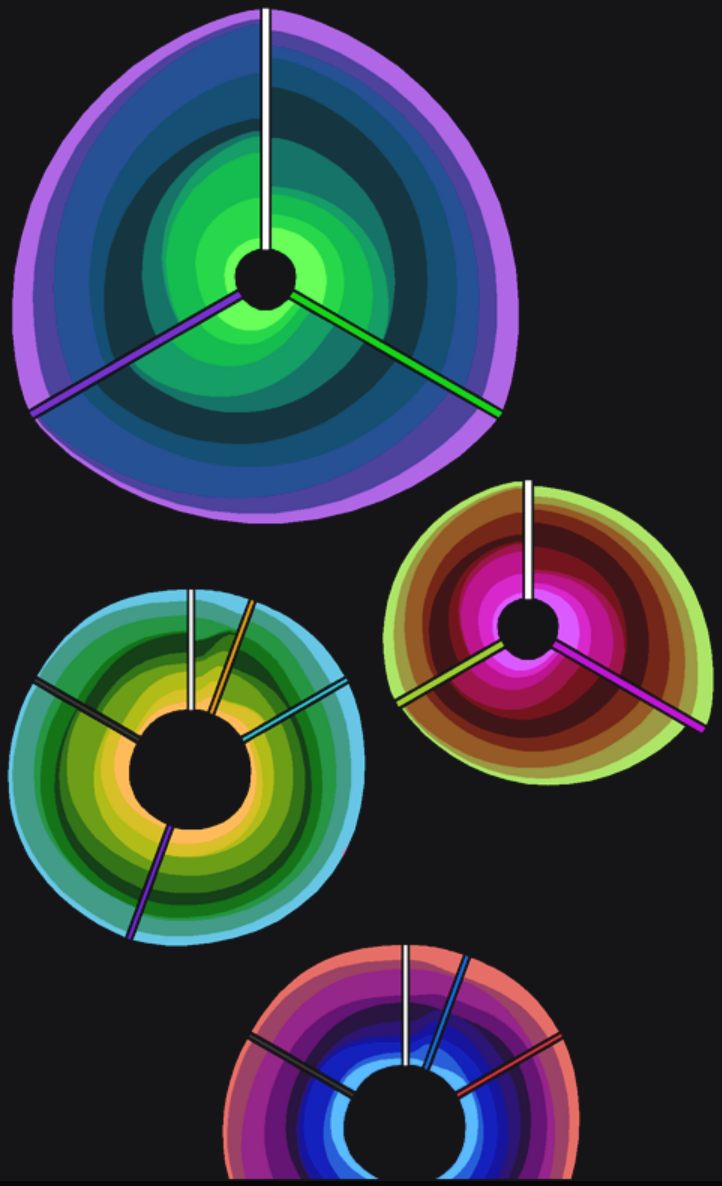
LINEAR LAYOUT FOR VISUALIZATION OF NETWORKS

Martin Krzywinski, Katayoon Kasaian, Olena Morozova, Inanc Birol, Steven Jones, Marco Marra

BC Cancer Research Center, BC Cancer Agency
Vancouver BC Canada

mkweb.bcgsc.ca/linnet



This talk and the visualization software is available at http://mkweb.bcgsc.ca/linnet.
Originally presented at Genome Informatics 2010, Hinxton UK (September 17).

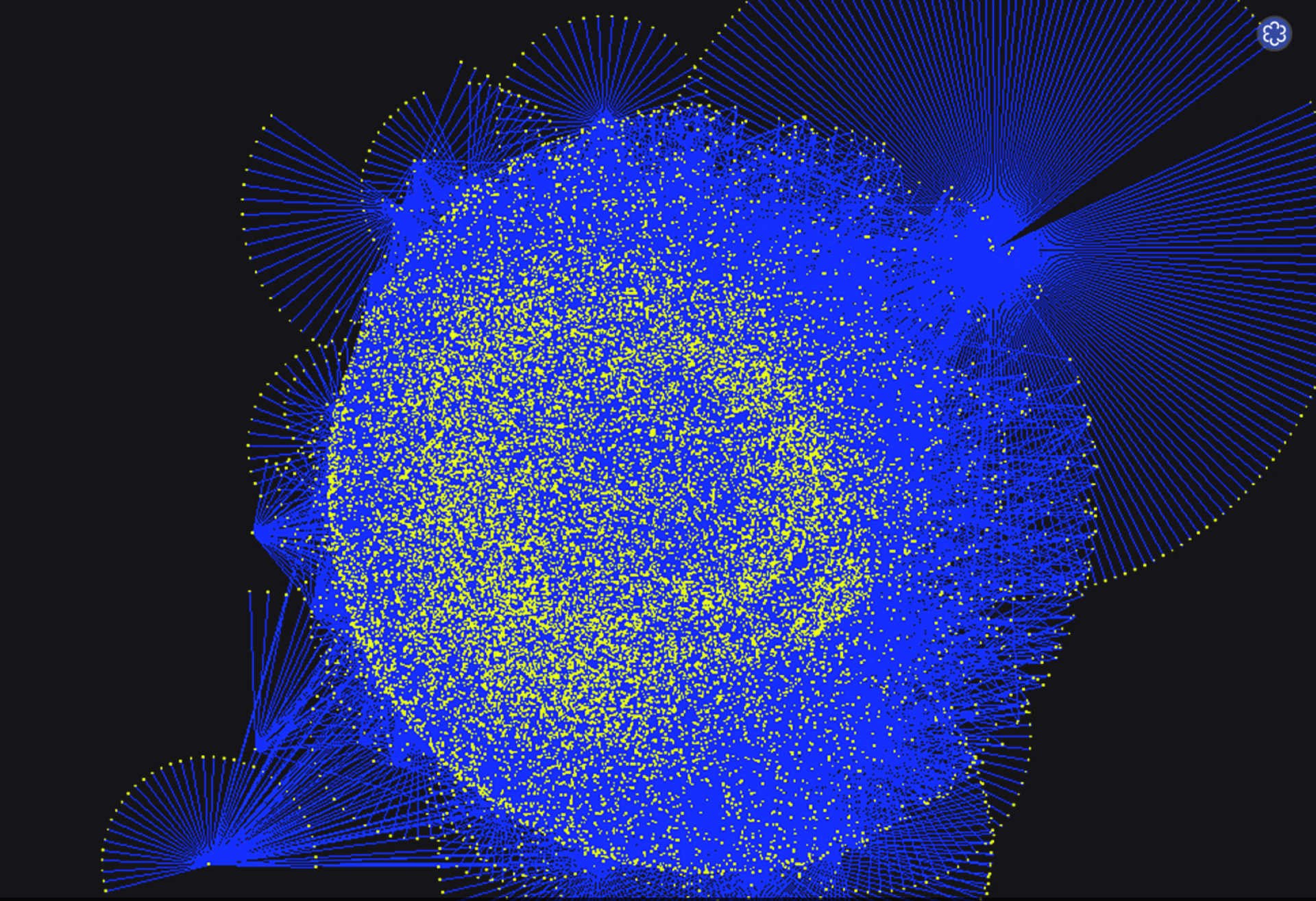
version 2 / 23 Sep 2010

The connectedness within biological systems are described by networks. This data structure naturally captures the interdependency of a system's elements. Many general algorithms are available to productively study the structure of networks. Effectively visualizing large networks remains difficult and requires a case-by-case approach. I propose a method for drawing networks that makes possible finding patterns in connectivity and metadata.

A hairball of the wooly mammoth. It was used as the DNA source for sequencing of the mammoth (Regenerating a Mammoth for 10 Million, NYT 2008). Visual examination of the hairball does not reveal characteristics about the mammoth, except that it is likely a hairy, large animal which can survive low temperatures (the hair is coarse and long). The hairball is therefore an inadequate visualization of the mammoth.

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A et al. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**(7220): 387-390.

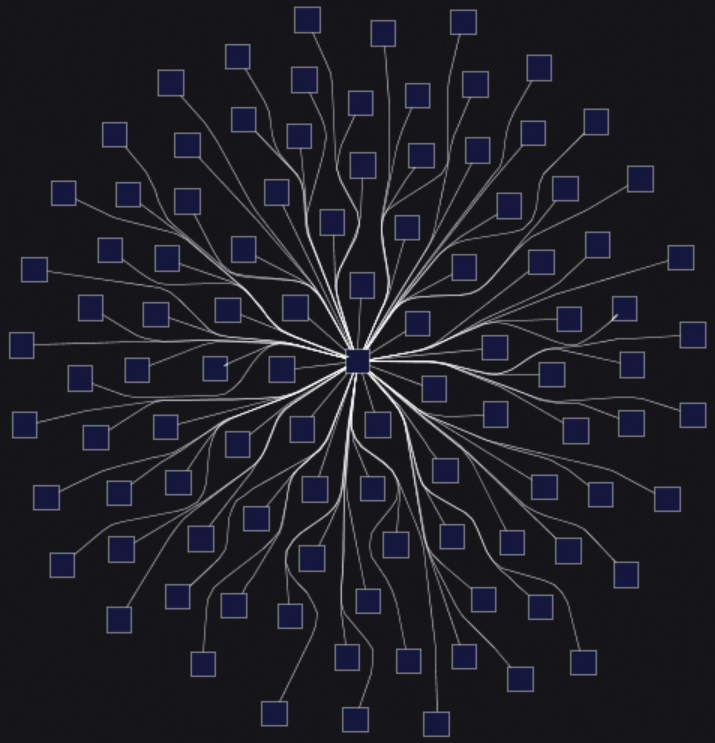
This is an excellent visualization of the mammoth. It communicates shape, size and environment of the mammoth. The aggressive posture suggests the animal is not suitable for immediate domestication. This image is an example of a specific visualization. It required integrated knowledge of the mammoth and artistic skill. The method used to generate this image cannot be automated, nor generalized.

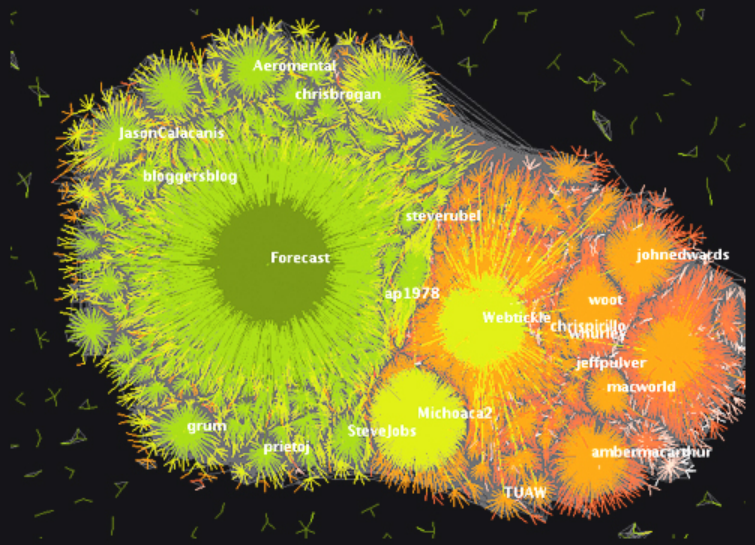


A depiction of the hyperlink network in gene wiki, rendered with Cytoscape (i9606.blogspot.com). It shares properties with the previous two images. Like the image of the mammoth, this visualization explicitly shows the entire object under study. Like the image of the hairball, it is equally unhelpful in understanding the object's properties. You can guess that the network is large and its connectivity is complex, but not more. At best, the visualization is merely decorative.

The genius of Gene Rodenberry allowed him to predict a future in which hairballs run amok. In this episode of Star Trek, Trouble with Tribbles, engineer Scott consults with Kirk and Spock about the hairball crisis. Note the tribble in Kirk's cup and those stuck to the walls. It isn't clear how tribbles, who have no legs, can adhere to a vertical surface.

Star Trek Episode 44, 2nd Season, 29 Dec 1967

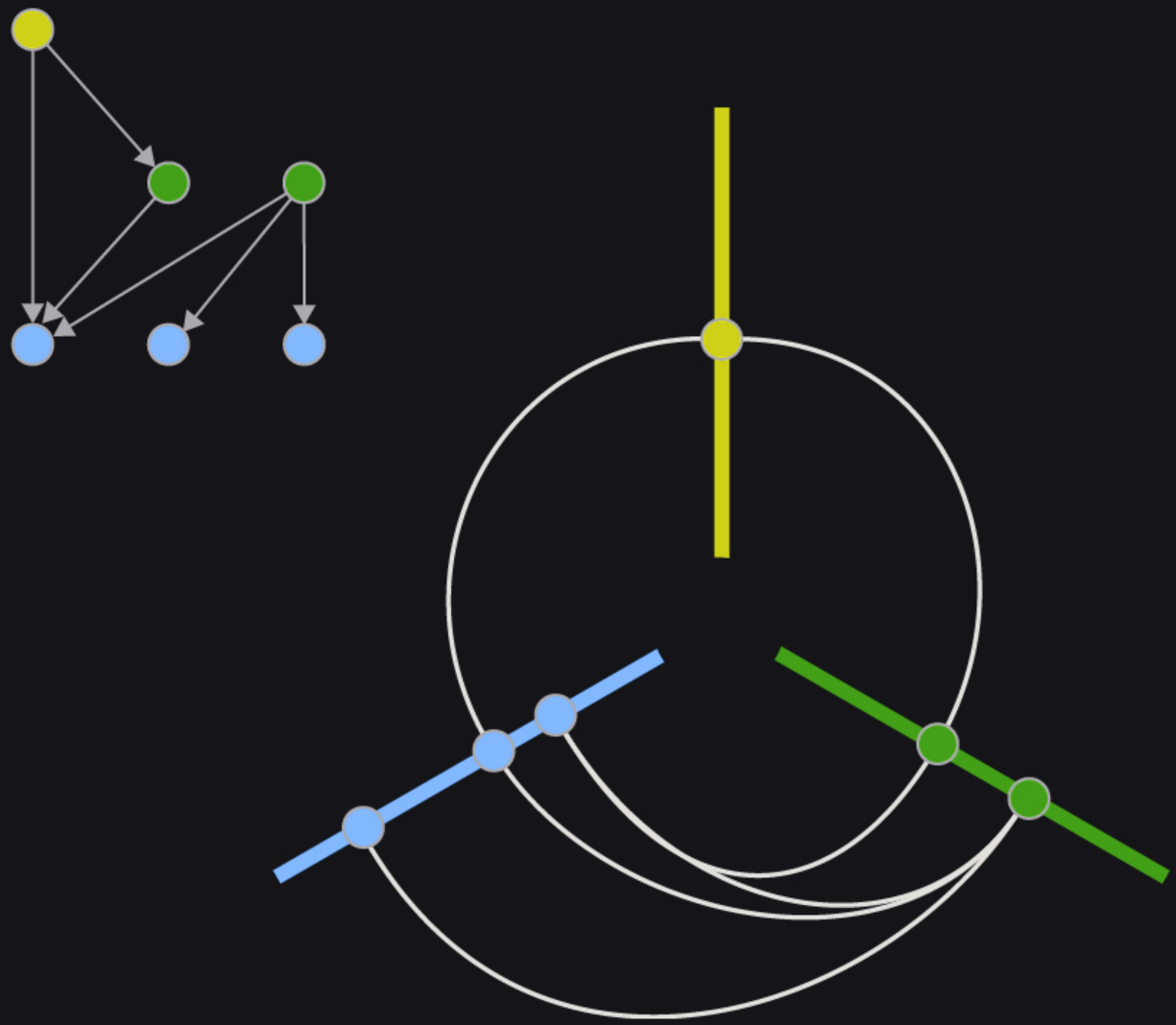




There are many algorithms to draw networks. A network's complexity lends itself to creative exploration. There are many beautiful renditions of large networks. Making these layouts informative is incredibly challenging, requiring manual intervention and the problem must be approach on a case-by-case basis.

y.layout.router Class OrganicEdgeRouter / Large Graph Layout (LGL) / Today by Cada / Mapping the Human Disseaseome (Bloch/Corum NYT 2009)

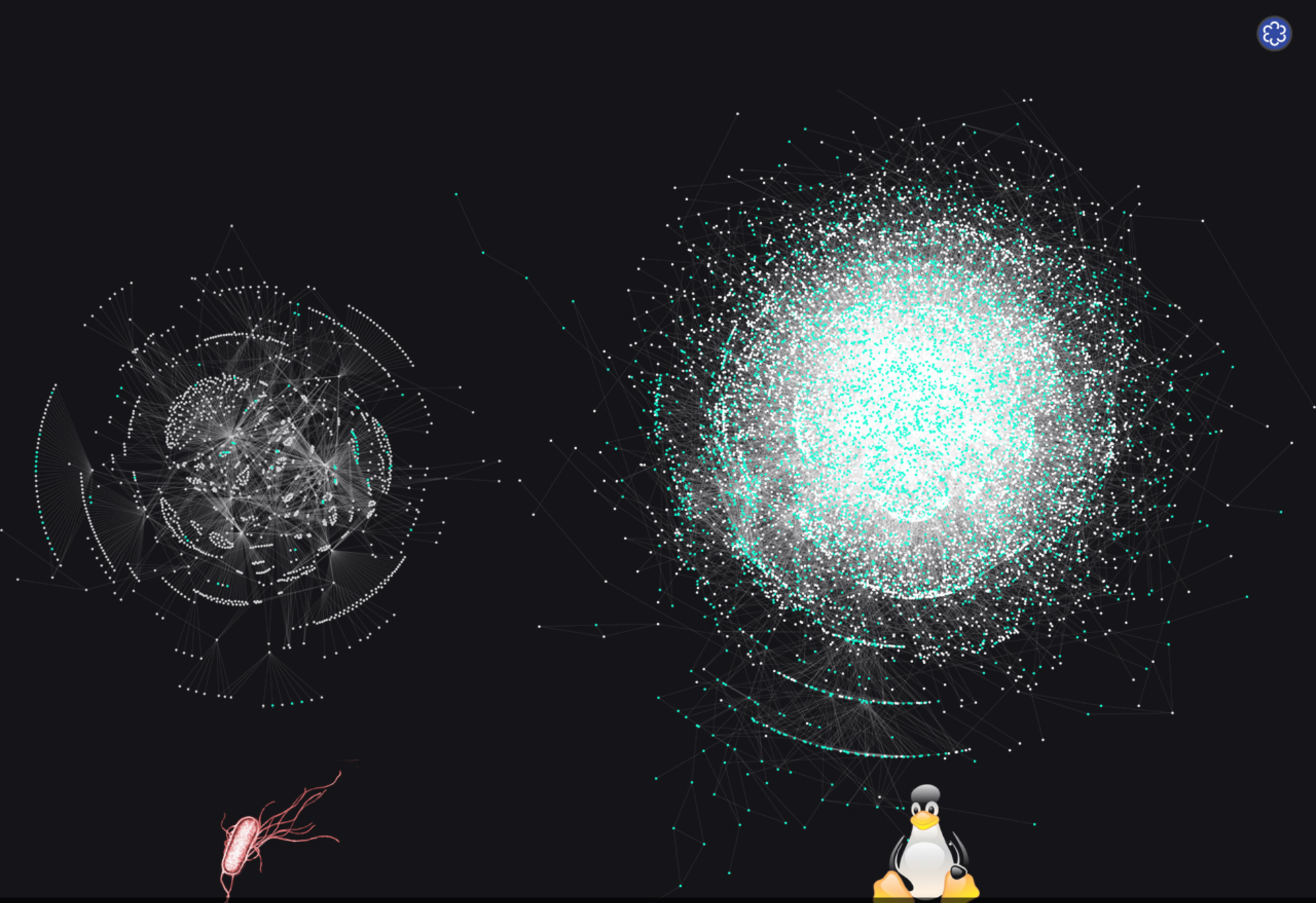
The method that I propose for drawing networks benefits from being applicable to any network, without manual intervention. Aesthetic for node positioning is informed entirely by node connectivity, category and functional annotation. In other words, no attempt is needed to beautify the layout, a step which can be very sensitive to the network's structure and produce significantly different layouts for similar networks, making comparison difficult.



Nodes are placed on linear segments (axes) which are positioned radially, each at a different angle. A node is assigned to an axis based on any relevant criteria, such as topology (as in the example above for a directed graph where nodes are classified as OUT (yellow), IN/OUT (green) and IN (blue)), connectivity (total number of edges), or a functional classification. The node's position on its axis can be based on connectivity, or an independent node property.

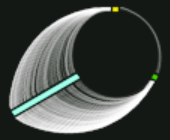
In the next series of slides, the linear visualization method is applied to data taken from [1], where the authors compare the E. coli gene regulatory network to the function call network in the Linux kernel. [1] Yan, K.K., et al., Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.

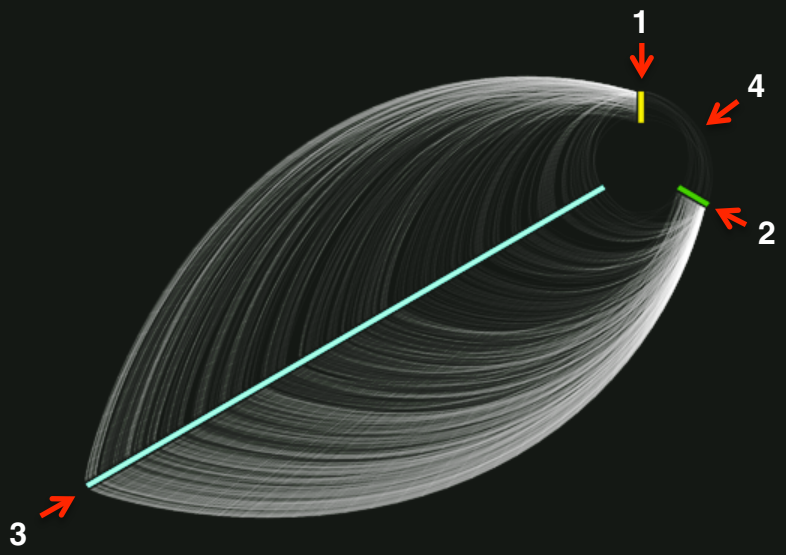
Yan KK, Fang G, Bhardwaj N, Alexander RP, Gerstein M. 2010. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc Natl Acad Sci U S A* **107**(20): 9186-9191.

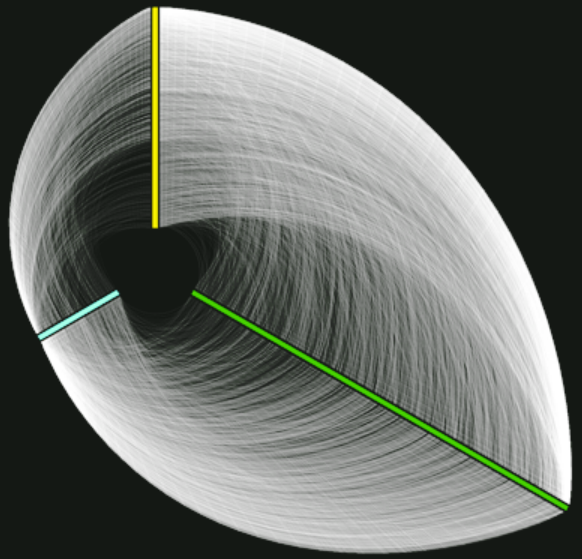


The hairballs of the E. coli and Linux networks are shown here. The size (approximate) of these networks is (nodes/edges): 1,400/3,000 for E. coli and 12,500/33,500 for Linux. The Linux network is 10x larger and has about 2.7 connections per node, than E. coli which has about 2 connections per node. The hairballs here reveal no information about the networks, except that the Linux one is larger. Nodes colored in green are those labeled “persistent” in the paper.

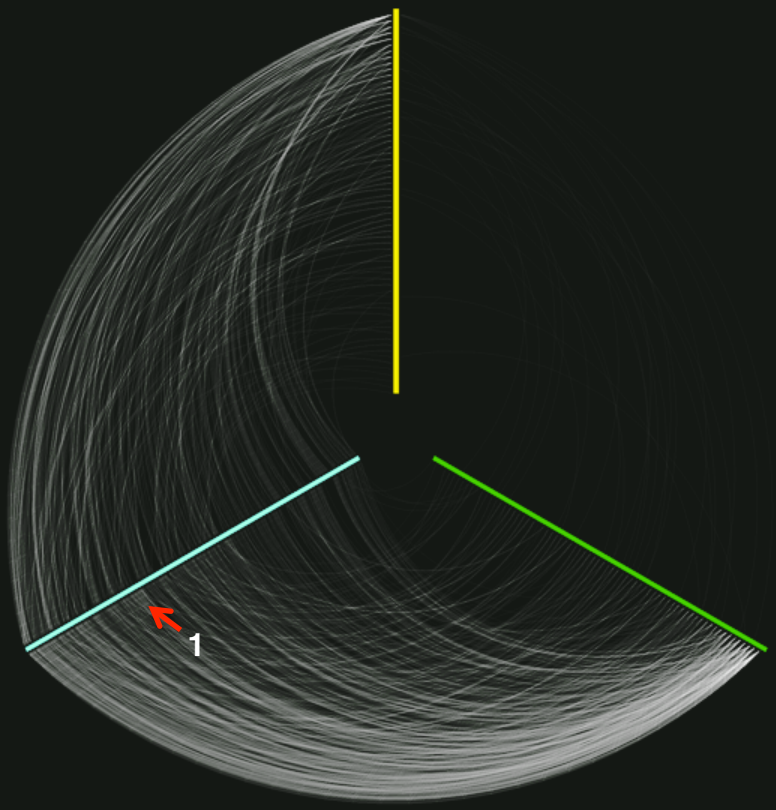
Nodes in the two networks were classified based on their connectivity. The layers were called REGULATOR (out), MANAGER (in/out) and WORKHORSE (in) and color coded yellow, green, and blue, respectively. The bars reveal that the E. coli network is bottom-heavy (many workhorses), whereas Linux is top-heavy (many regulators and managers). Let's see how the linear layout approach can reveal patterns in connectivity between these node groups.

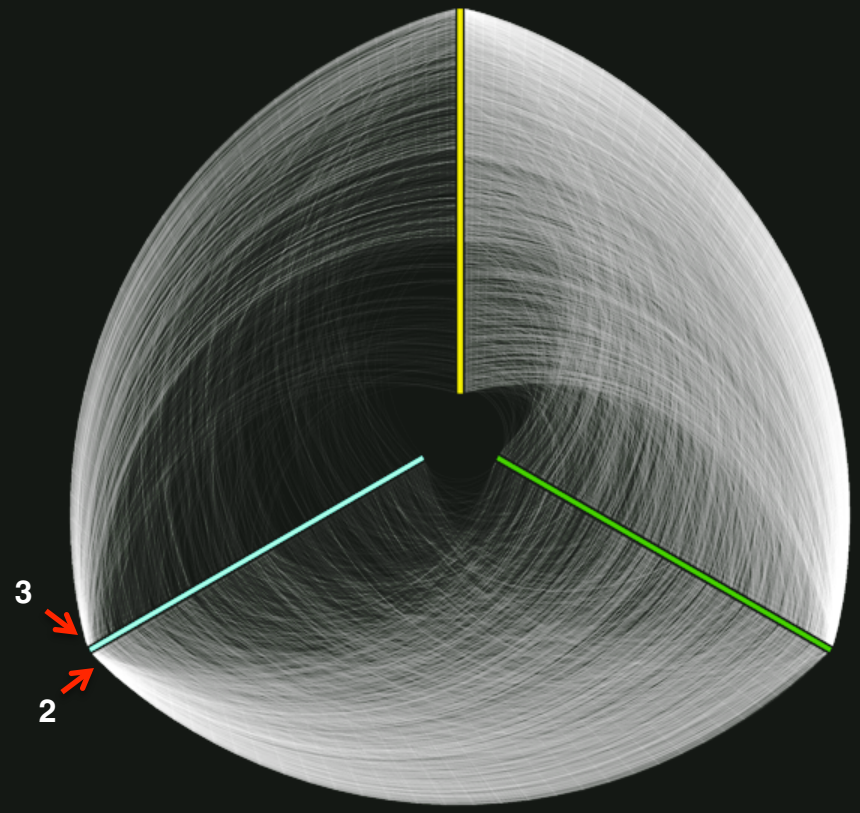




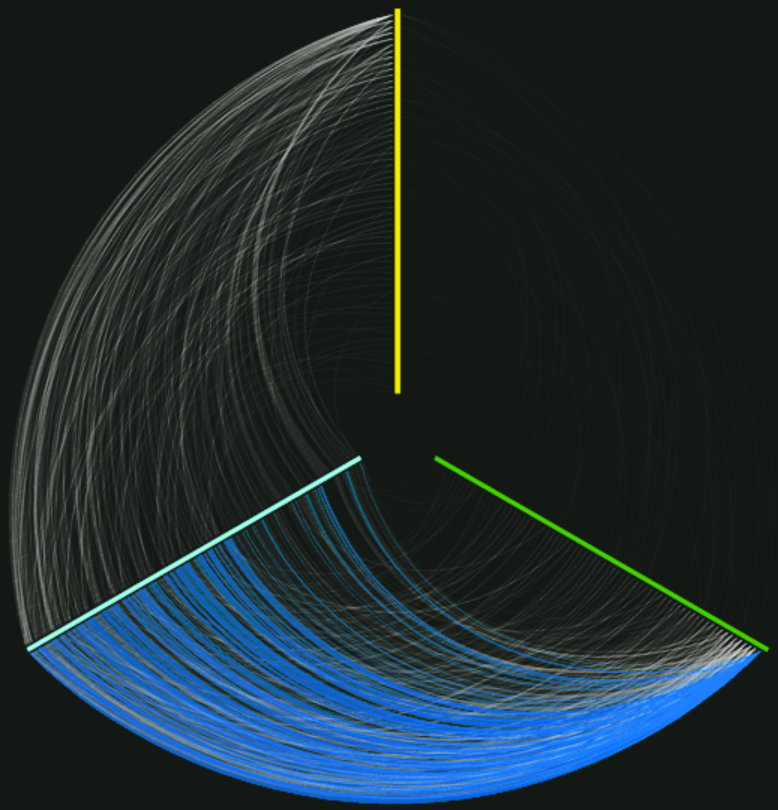


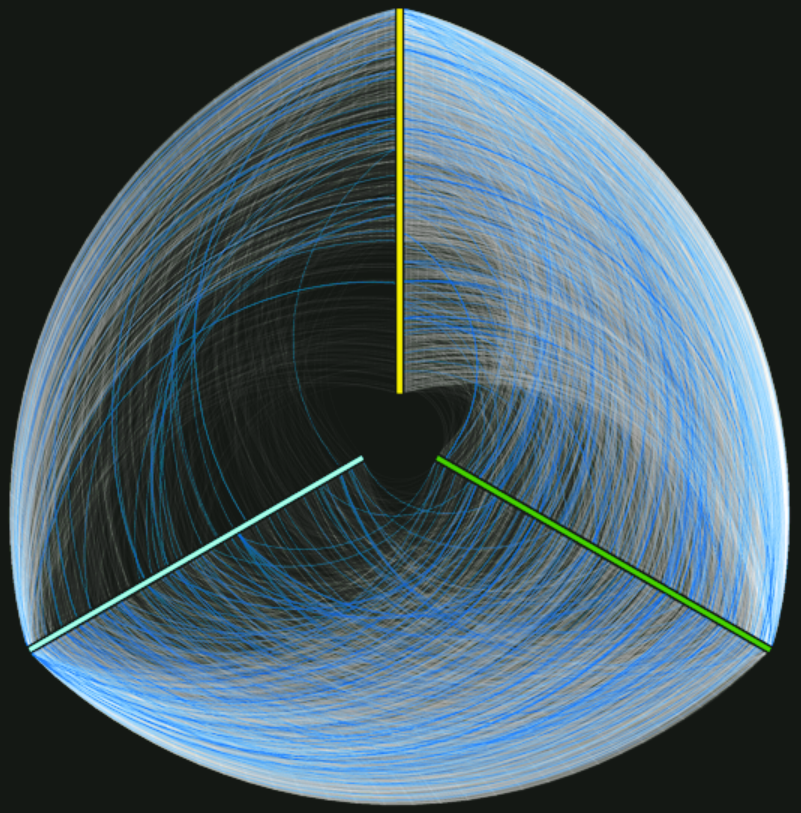
For all views, nodes are assigned to axes clockwise from the top: regulator (out) [1], manager (in/out) [2], and workhorse (in) [3]. The axis length is proportional to number of nodes in category. The E. coli view is magnified 8x (required because it has 10x fewer nodes). Immediately it can be seen that E. coli contains nearly no communication between regulators and managers [4], whereas the Linux figure contains many connections between these node types.



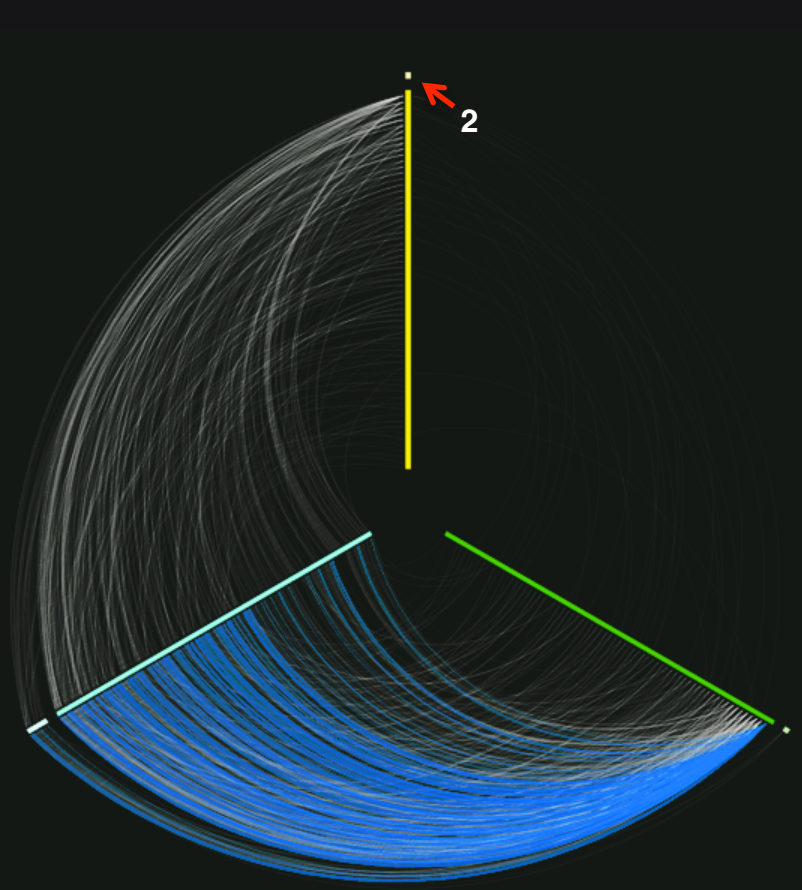


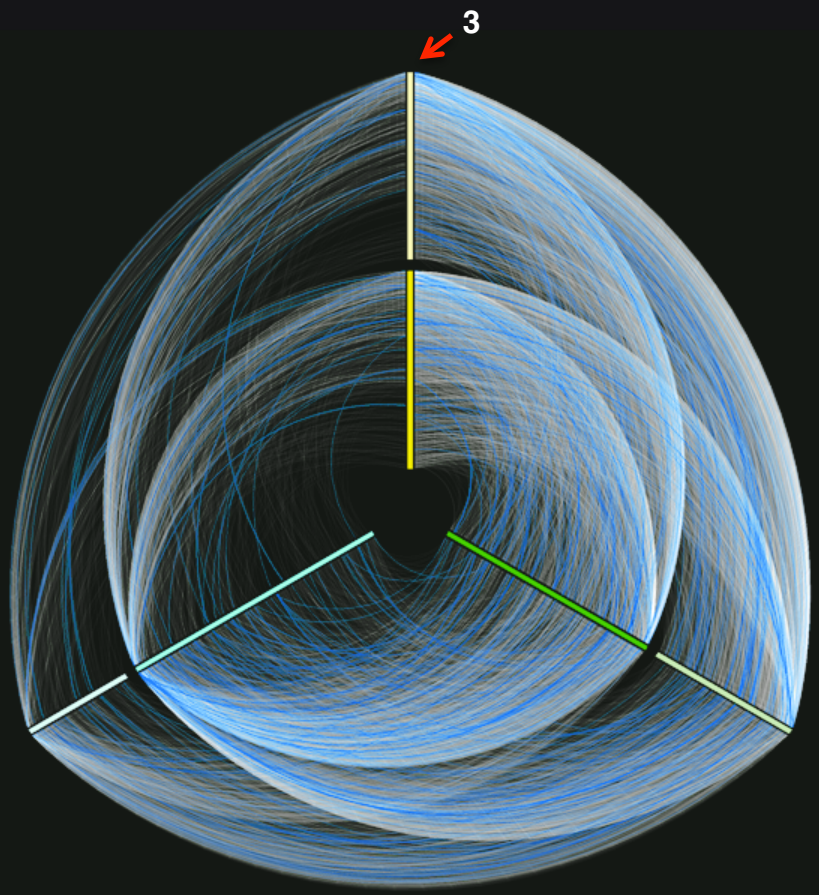
When the axis length is normalized, patterns in connectivity are easier to see. In E. coli, most workhorses have approximately the same number of connections [1], whereas in Linux a large number of managers call a small number of workhorses [2]. This pattern is even more clear in the connections between managers and workhorses [3].



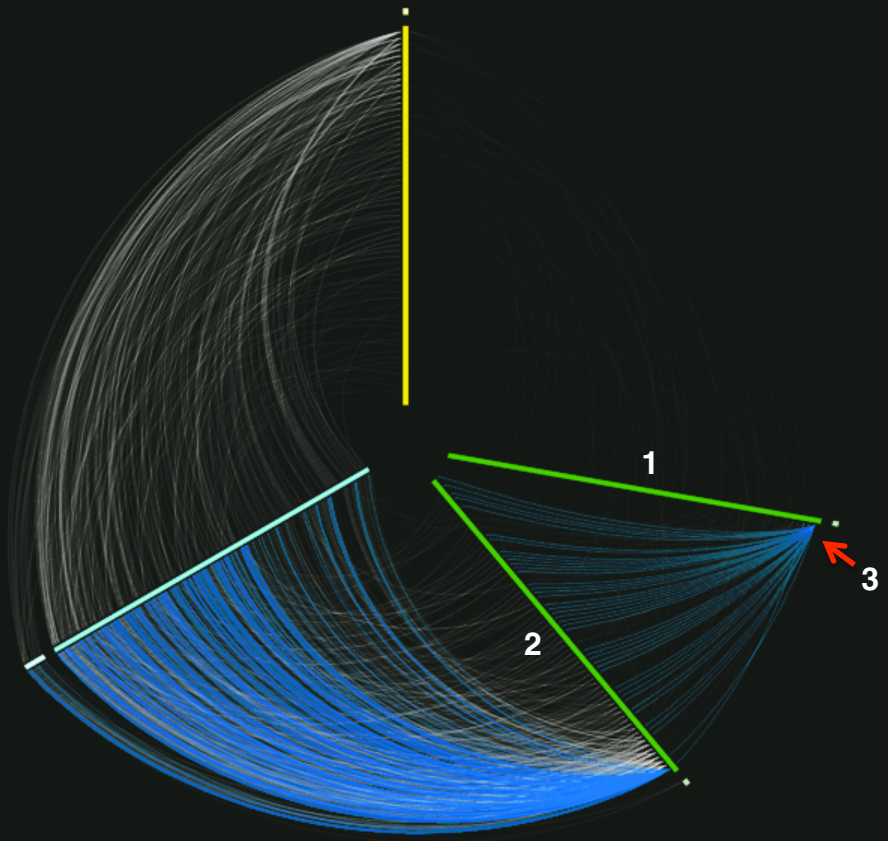


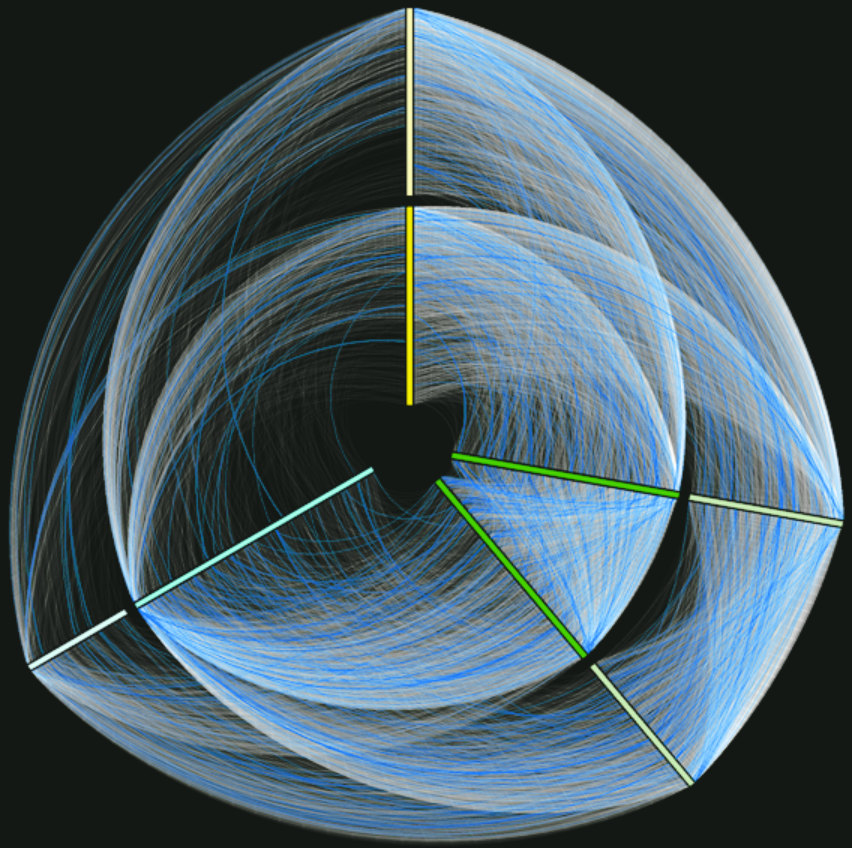
Adding functional information to the view is easy. Edges are colored blue when associated with the CRP gene (E. coli) and set_* functions (linux). Additional markup on the figure is easily accomodated. For example, links can be offset from the axis to allow for a margin between the link end and axis, which can contain information about nodes.



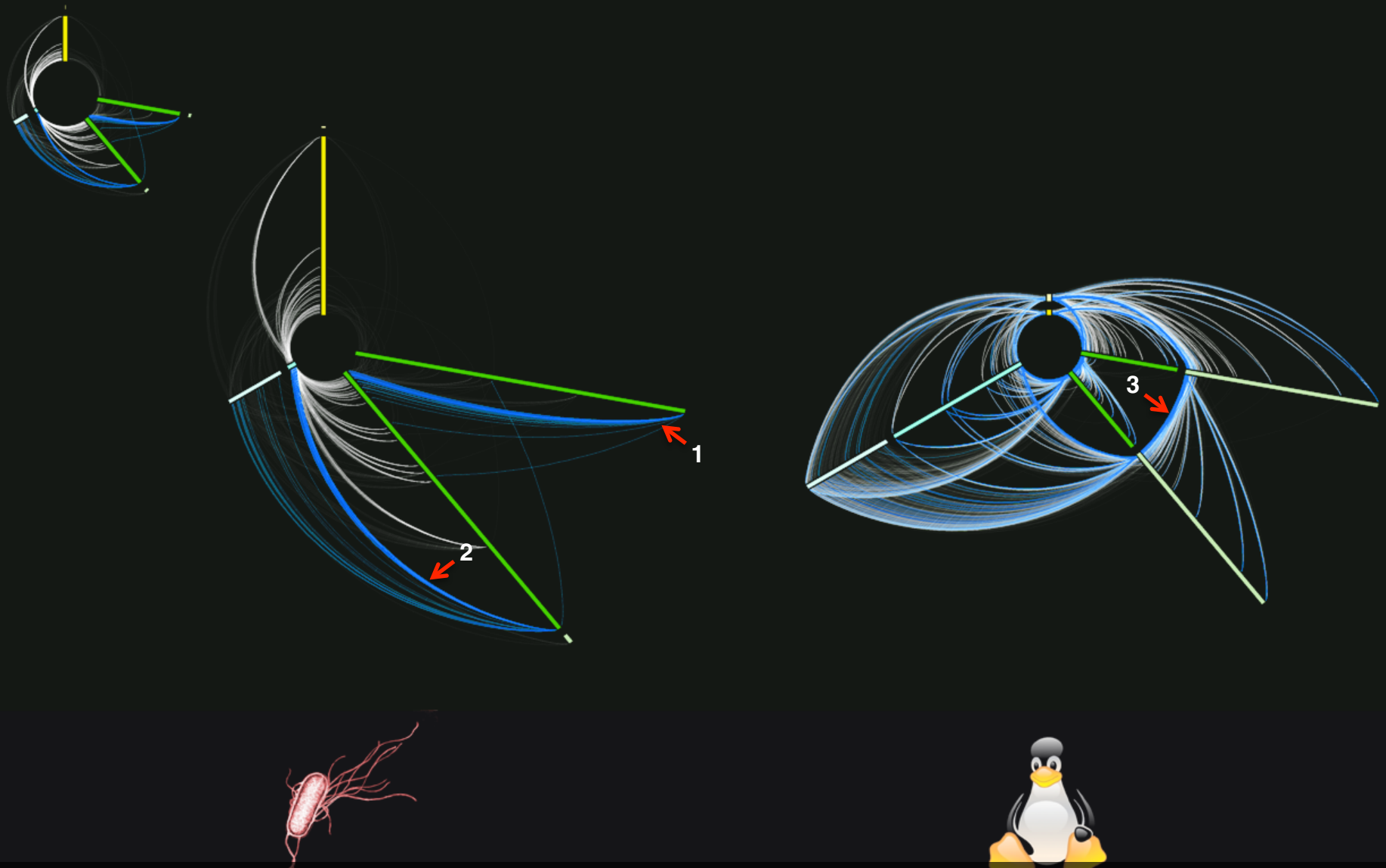


In addition to connectivity, each node in the network had a persistence flag [1]. This property bifurcated the nodes on each axis, which is shown here by the introduction of a new segment [2,3]. The nodes in the inner segment are non-persistent, and those in the outer segment are persistent. It can be immediately seen that E. coli has few persistent nodes [2]. [1] Linux - function was found across kernel versions, E. coli - gene was also found among 200 diverse bacteria).

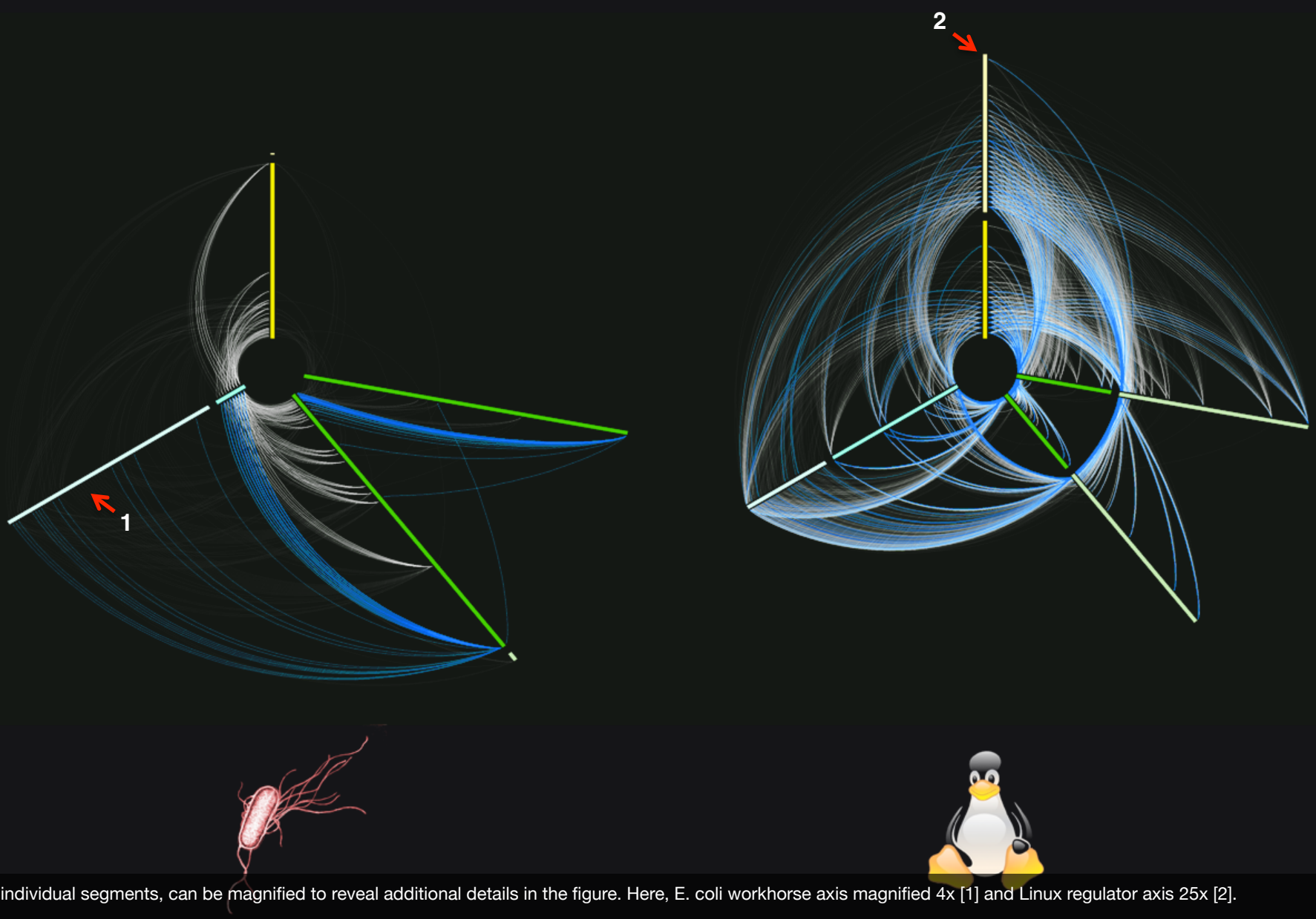




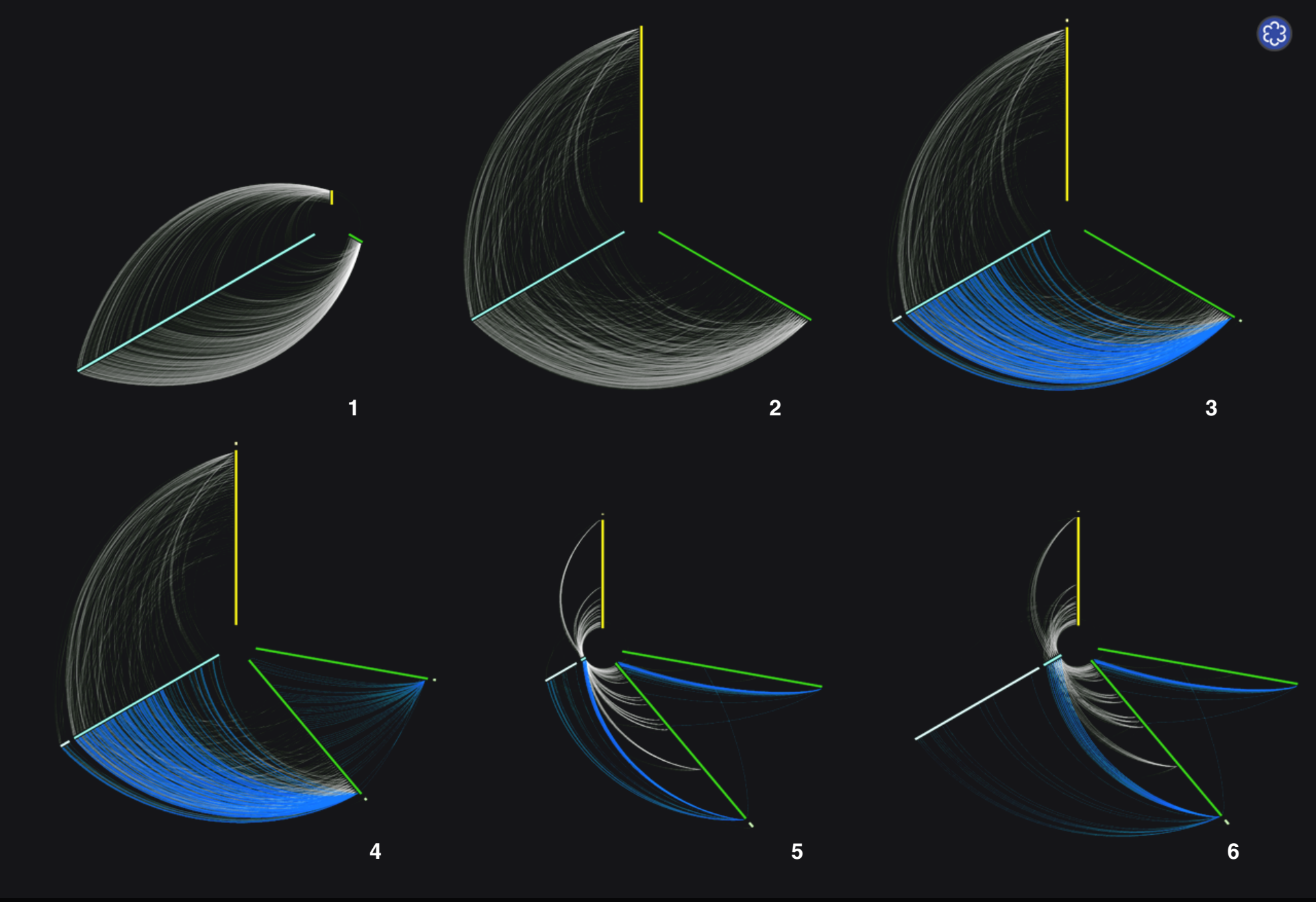
Additional axes can be added to further add texture to the figure. In previous figures, links between manager nodes (in/out) were not shown, because these nodes were all placed on the same axis. Here, the manager axis is internally split into two ([1] manager out, [2] manager in), and used to anchor manager-manager links. The E. coli figure reveals multiple connections from a single manager to many managers [3]. Links between managers and other nodes are drawn as before, from the manager axis closest to the nodes' axis.



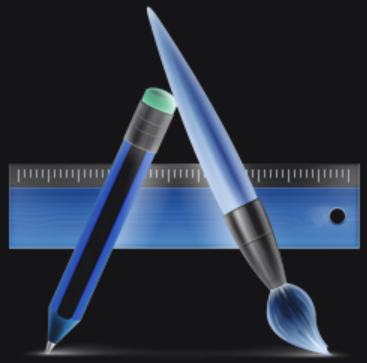
In previous figures, node position was determined using rank order of connectivity. Here, absolute connectivity is used to place nodes (i.e. number of edges at a node). Recall the manager-manager connections from the previous E. coli plot. Here these appear as a multiple links from a node with high connectivity on the manager out axis, to many nodes with low connectivity on the manager in axis [1]. Similar relationship exists between few highly connected manager nodes and workhorses [2]. Note the large number of low connectivity persistent manager-manager connections in Linux [3]. E.coli plot is magnified 4x.



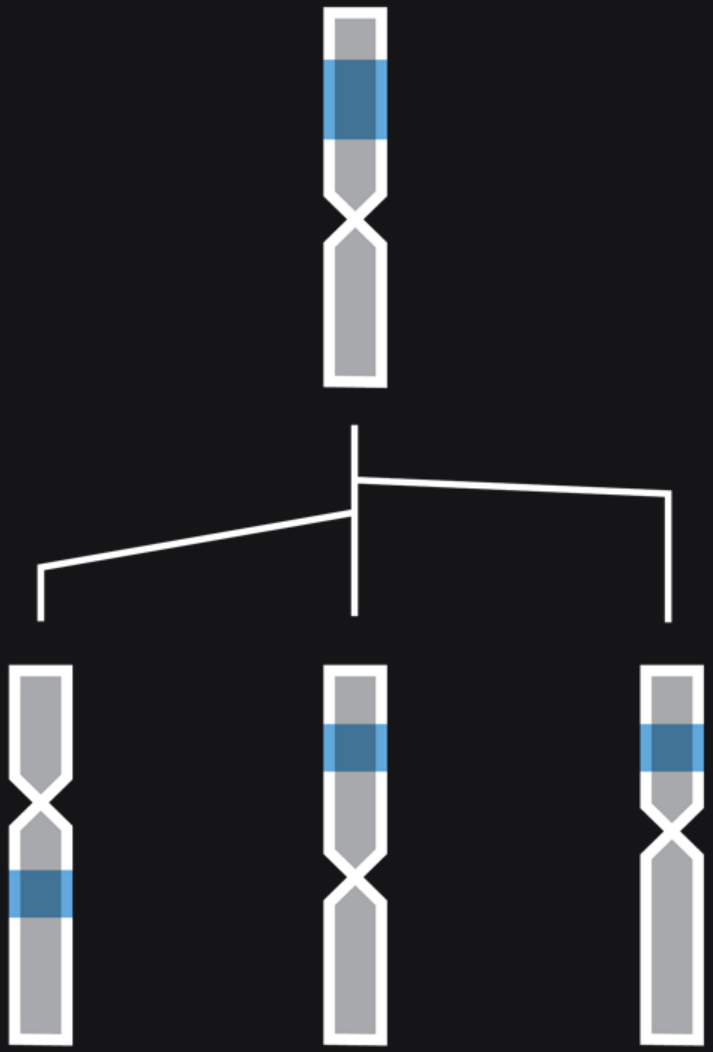
Axes, or individual segments, can be magnified to reveal additional details in the figure. Here, E. coli workhorse axis magnified 4x [1] and Linux regulator axis 25x [2].



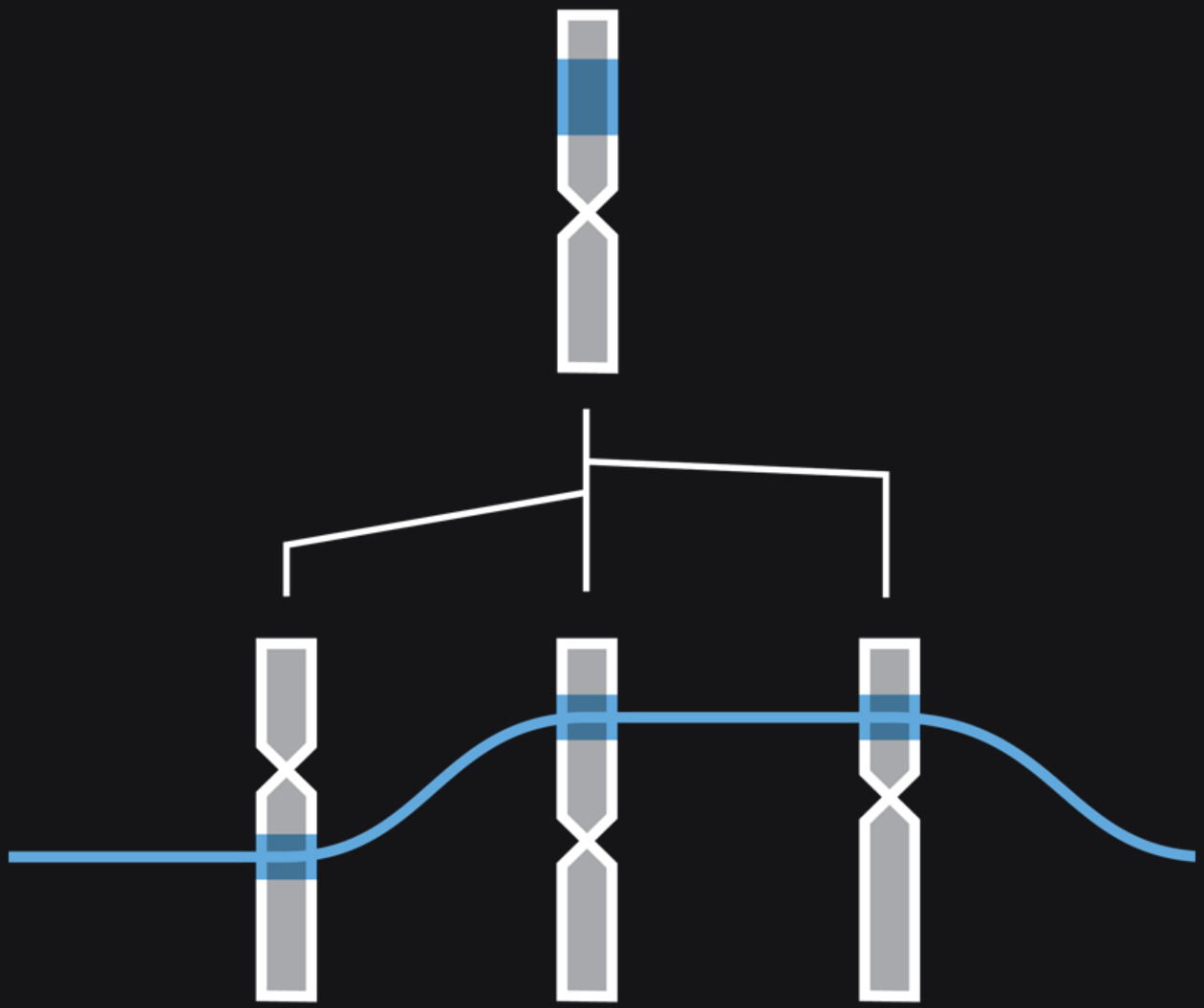
E. coli network figures seen so far. [1] Rank ordered connectivity node position with axes proportional to number of nodes. [2] Axis length normalized. [3] Functional annotation added using color. [4] Links between manager nodes (in-out) shown by splitting manager axis. [5] Node position based on absolute connectivity. [6] Workhorse axis magnified 4x.



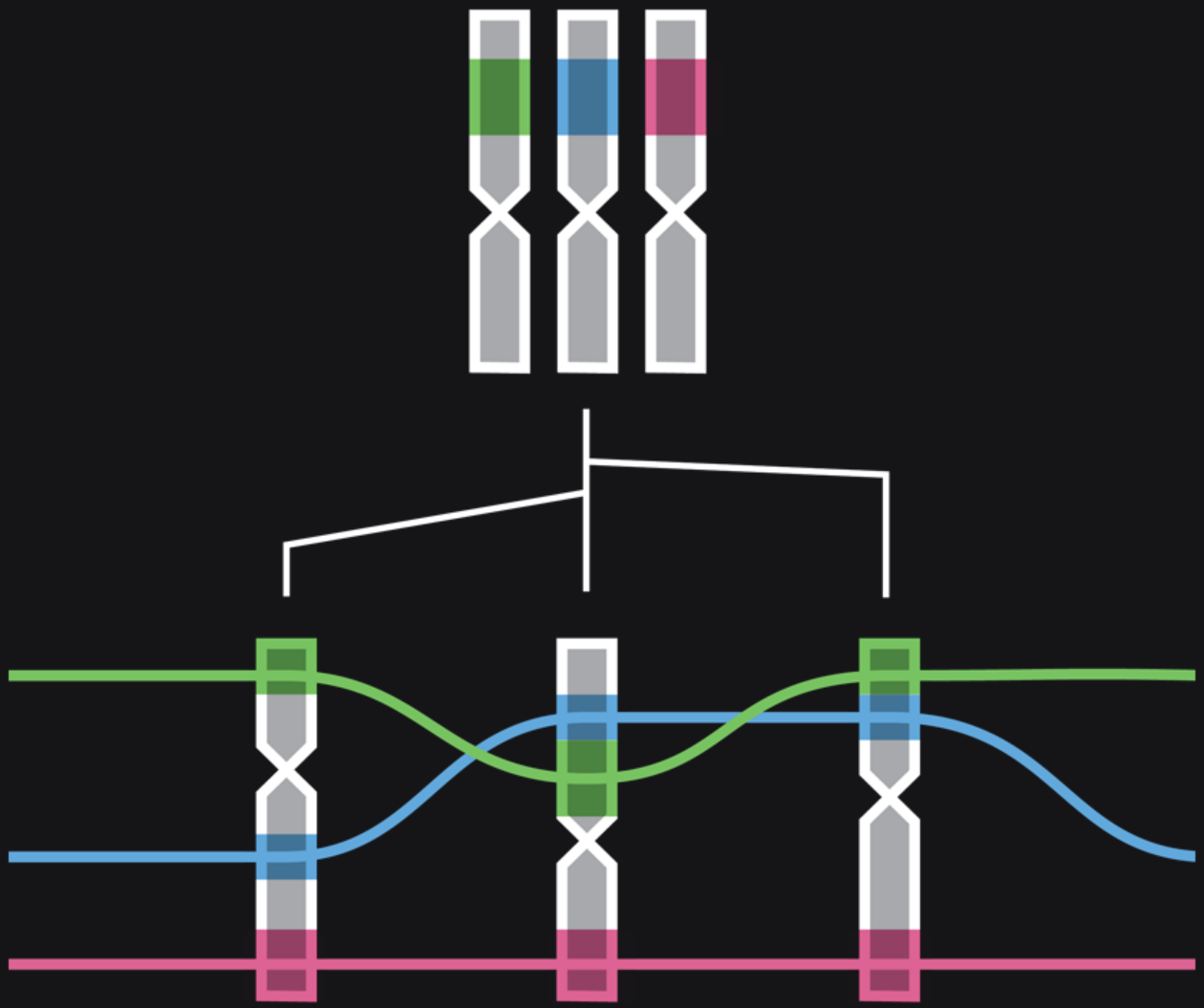
The linear visualization of networks can be applied to a variety of data sets. In the next three examples, I will show its use in showing synteny, tumor expression profiles and assembly quality. In particular, and as we'll see with the assembly quality example, when the links are drawn as ribbons with variable end lengths, a circular stacked bar plot is created.



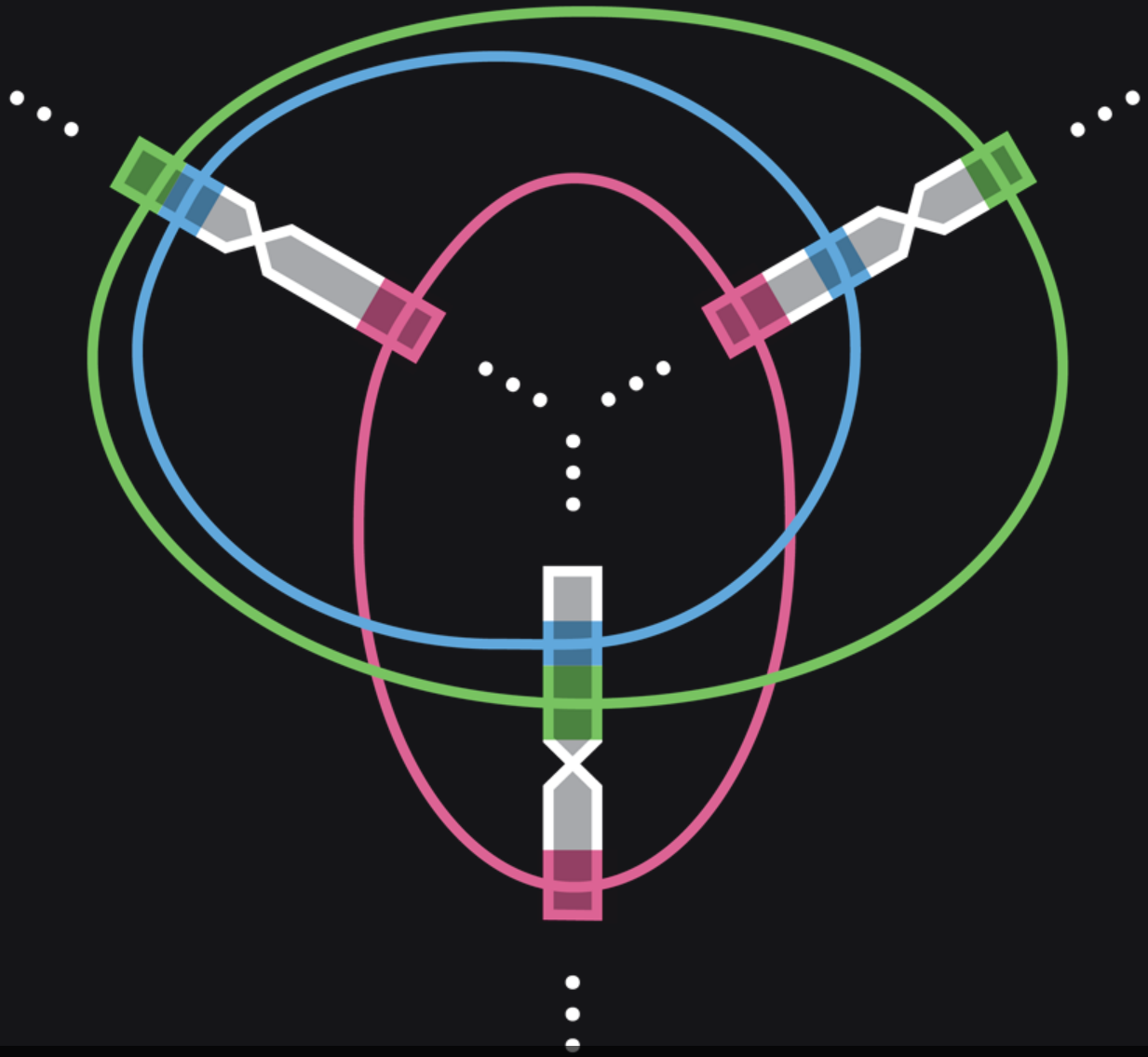
Consider a relationship between an acenstral chromosome (top) and instances of three chromosomes of modern species (bottom). Suppose that a position segment (or some part thereof) on the ancestral chromosome (blue) can be mapped to a location on the modern chromosomes.



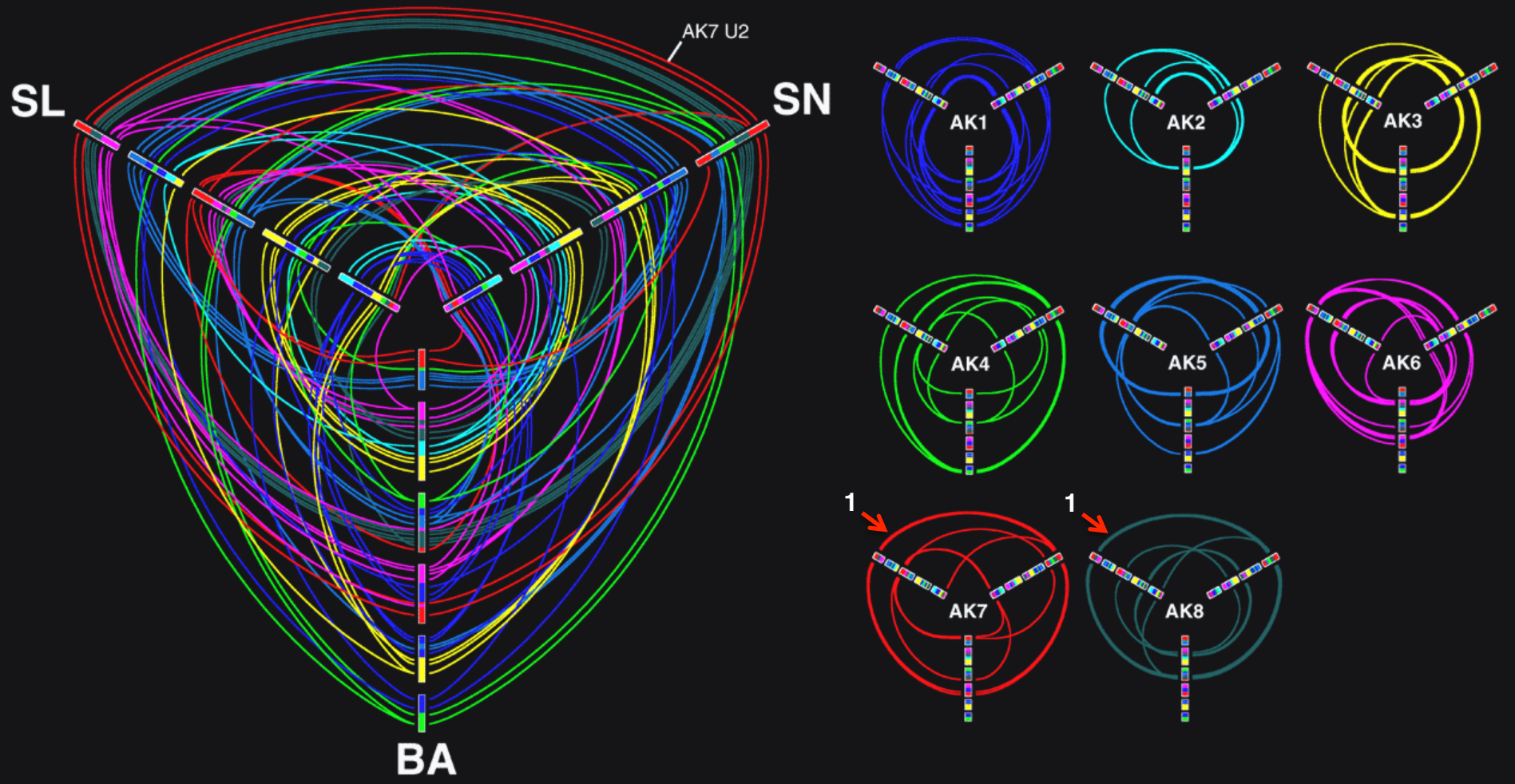
The ancestral-to-modern mapping can be used to connect the modern chromosomes. The relationship between regions on modern chromosomes is defined based on a common point of origin in the ancestral chromosome.



When additional ancestral chromosomes are considered, a synteny map results, in which associates regions on modern chromosomes with the same origin. This kind of relationship between genomic regions (here, modern chromosomes) defined by a shared point of origin (here, ancestral chromosome) can generalize to comparisons between modern species (where the human genome takes place of the ancestral genome), or multiple sequence alignment (where a reference genome takes place of ancestral chromosome and tumor genomes are compared).

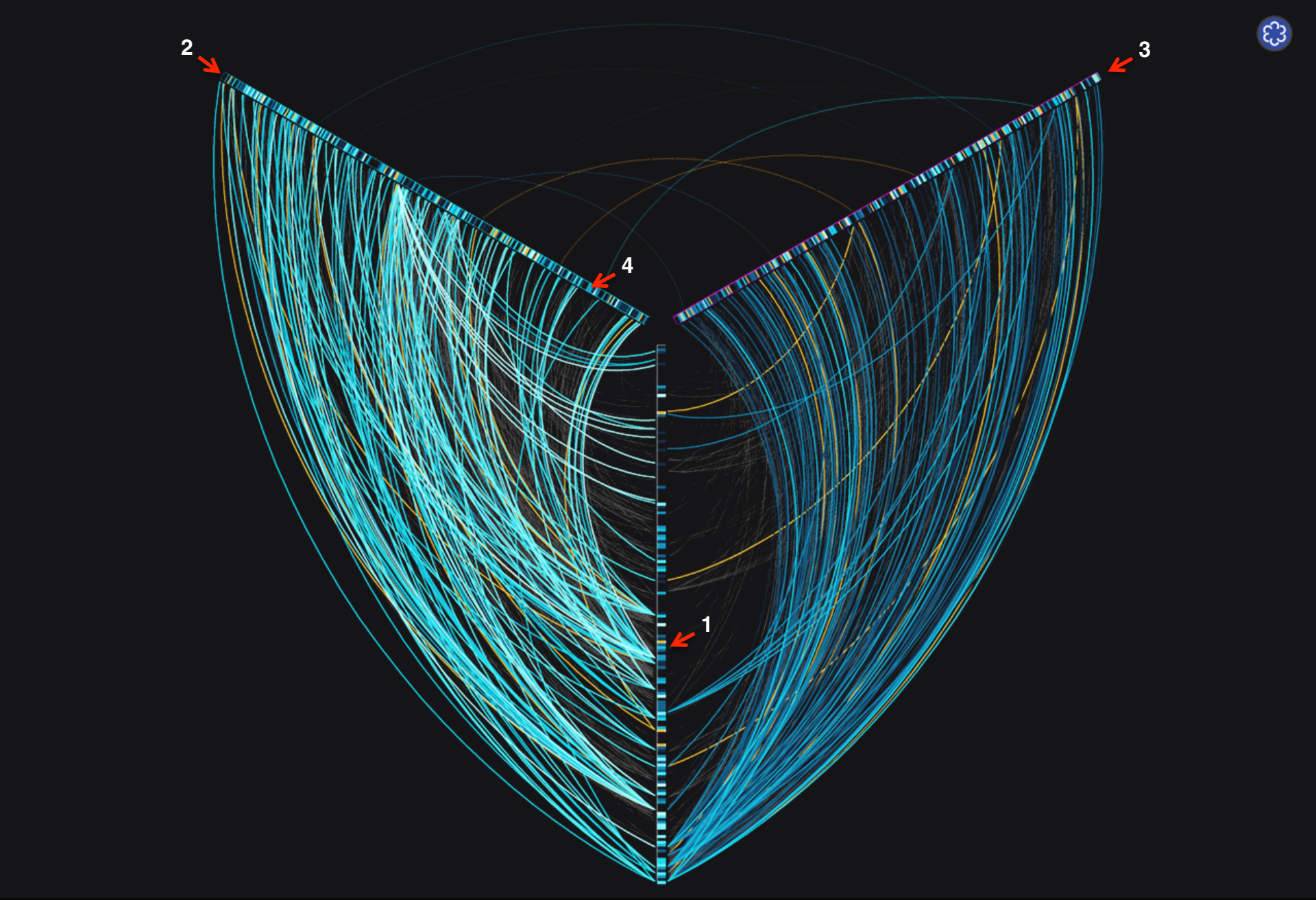


By placing the modern chromosomes along radial axes, the mapping that relates ancestral origin become concentric ovals. The pattern of the ovals describes the contiguity of the syntenic map – the further the ovals are from concentricity, the further the mapping is from identity.



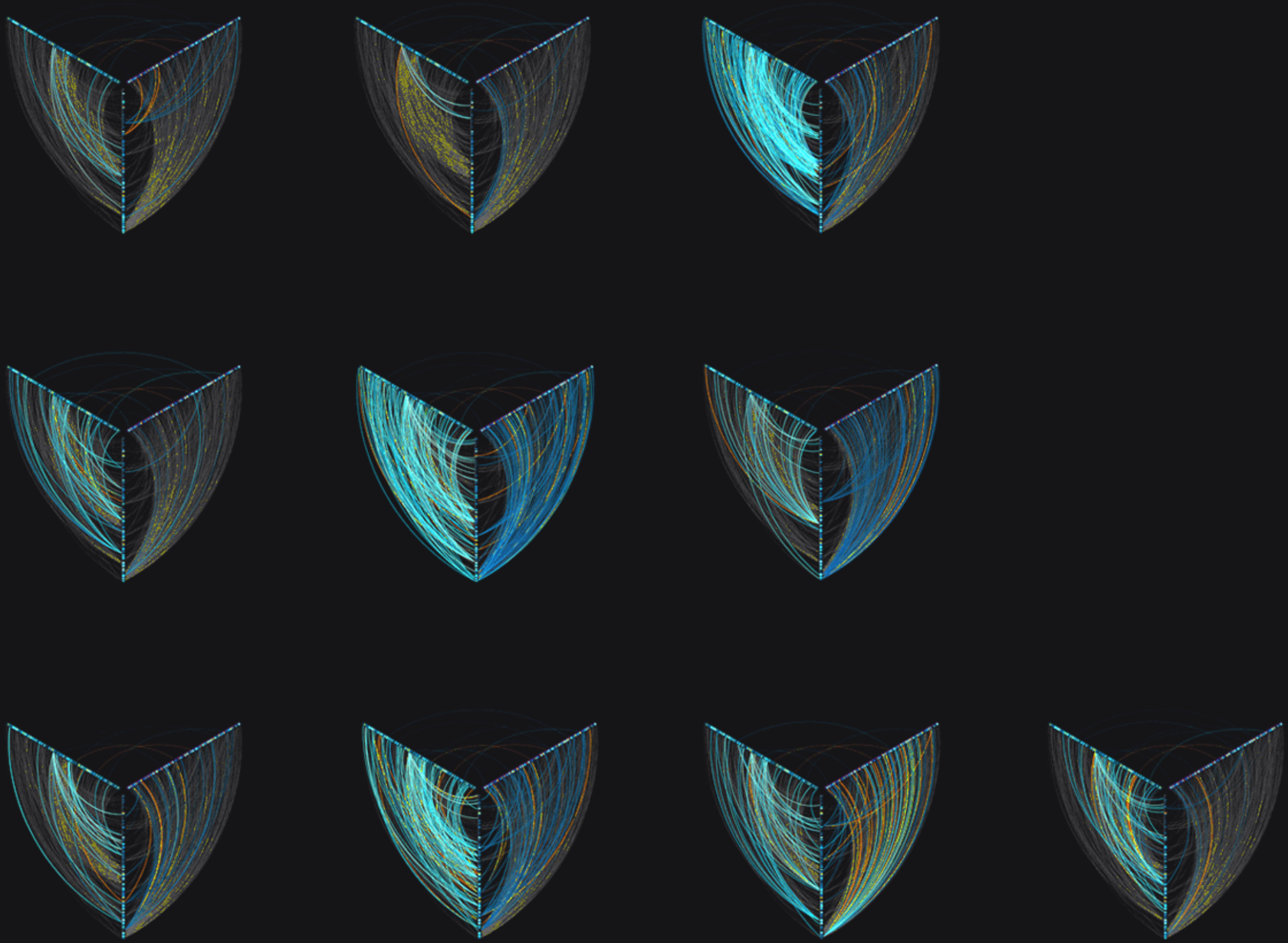
Relationship between three modern crucifer genomes (SL, SN, BA), as defined for regions by common points of origin in ancestral genome (AK). Mapping between each of the eight ancestral chromosomes (AK1 – AK8) and the modern genomes are shown on the left. Regions of AK7 and AK8 are mapped to the same region of SL and SN [1].

Mandakova T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA. 2010. Fast diploidization in close mesopolyploid relatives of Arabidopsis. *Plant Cell* **22**(7): 2277-2290.



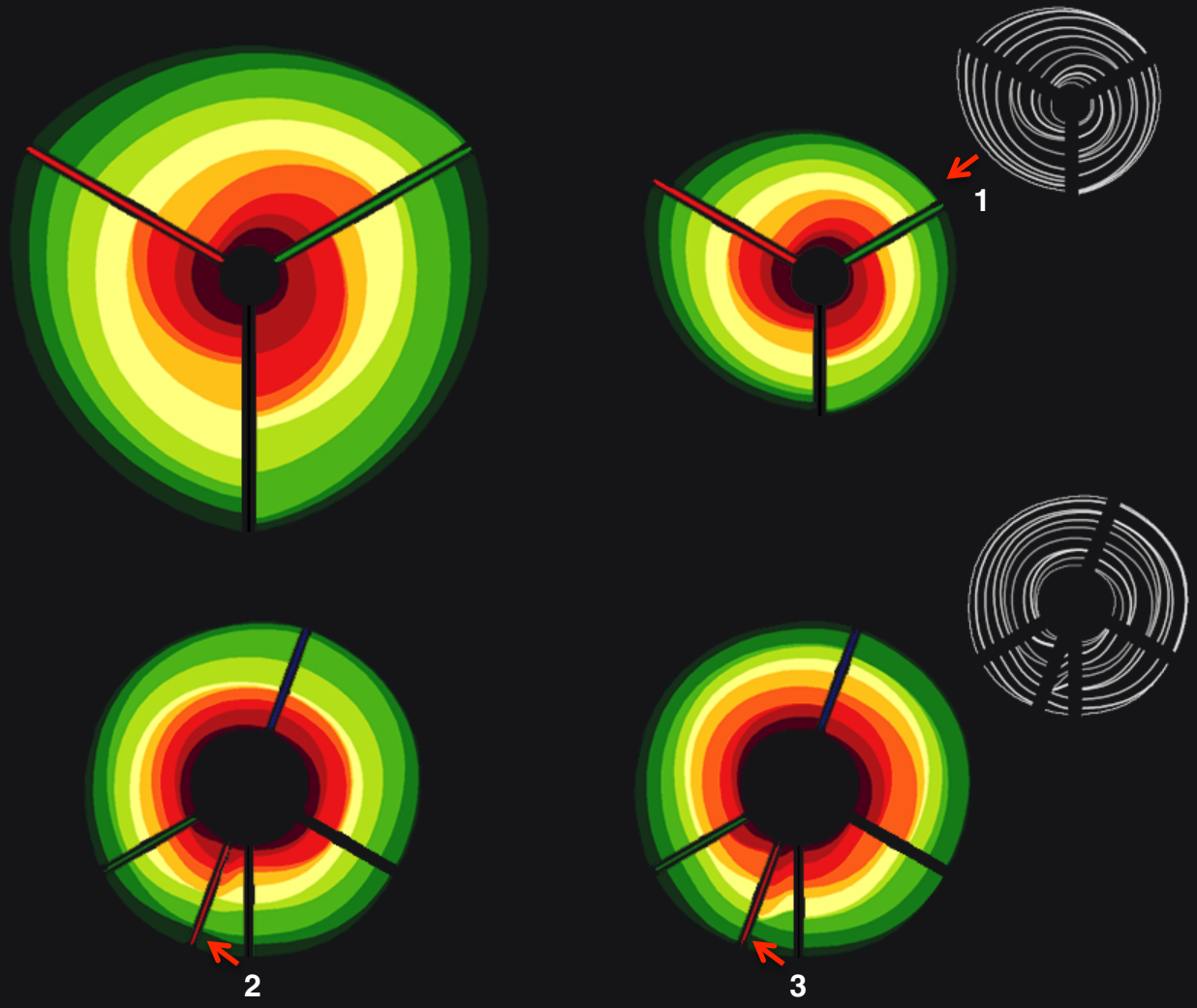
A coexpression profile of a neuroblastoma sample. Genes implicated in cancer were placed on an axis [1], with their coexpressed counterparts (negatively [2] and positively [3]) on the other axes. The observed expression level in the sample, measured by transcriptome sequencing, is represented by a heat map within the axis. Links between coexpressed genes are colored by the departure of their relative expression levels from expectation.

Birol, I. Personal communication.

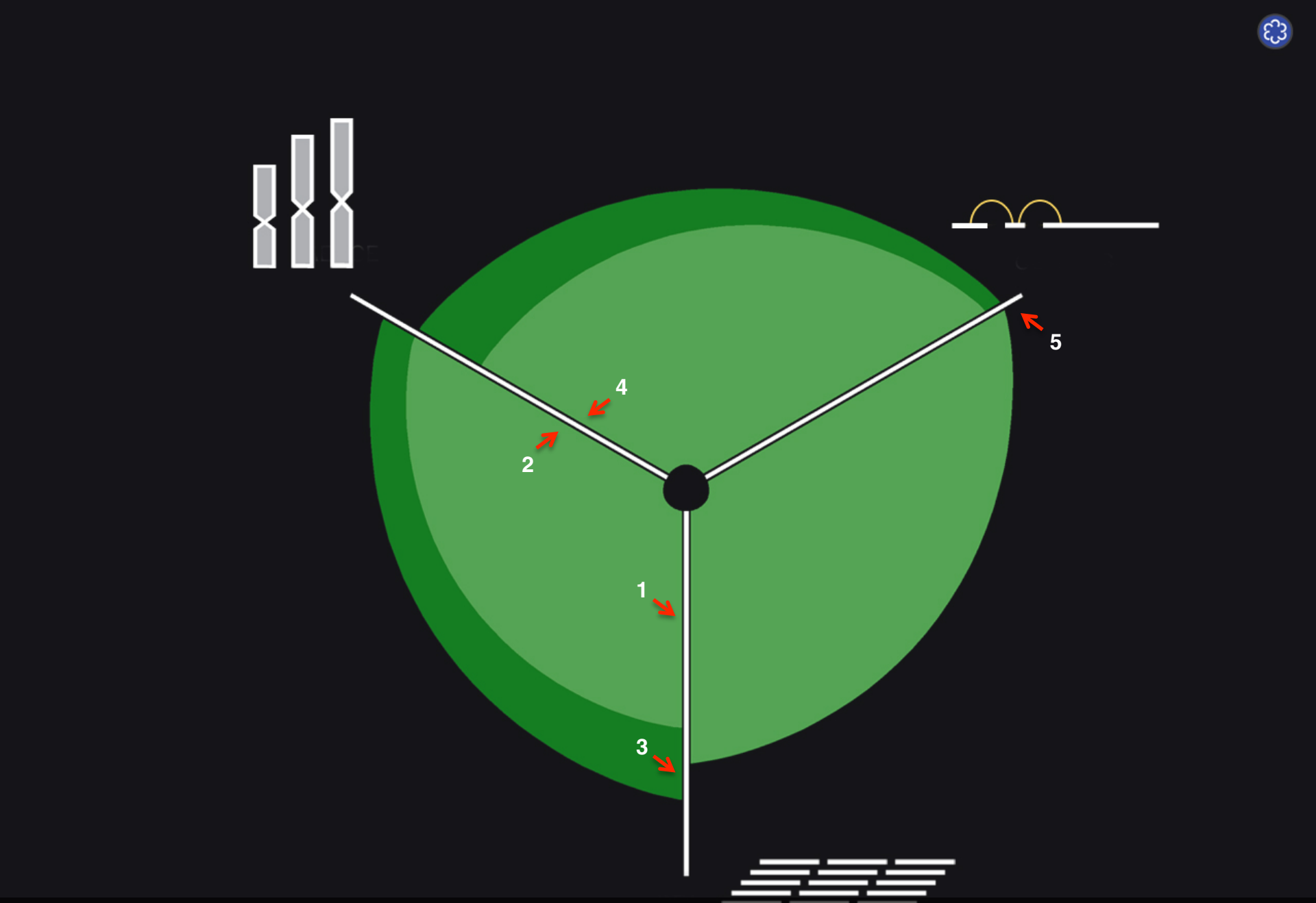


Gene co-expression profiles for 10 neuroblastoma genomes using the layout from the previous slide. Gene position is fixed across profiles, which allows for a direct comparison of profiles. For a sample in which a gene was not observed to be expressed, its links are grey. These images reveal several distinct profile types.

Morozova, O. Personal communication.



When the links that have been previously drawn as lines are rendered as ribbons [1], with variable end sizes, a circular stacked bar plot is obtained (c.f. steamgraph). When three axes are used, three or six stacked bar columns can be shown, depending on whether the values on an axis' sides are the same [2], or allowed to vary [3]. This representation is ideal for demonstrating relative sizes of relationships between three quantities, such as a sequence assembly, explained on the next slide.



A simplified assembly quality report. The axes represent bases in reads (bottom), contigs (right) and reference genome (left). Ribbons indicate fraction of bases aligned between each category pair. For example, approximately 60% of bases in quality reads [1] align to about 80% of the reference [2]. Another category of reads, such as multi-mapped, can be represented by a different ribbon [3]. This view quickly reveals key differences, such as the fraction of reference covered by reads [2] vs by contigs [4], or assembly errors leading to bases in contigs not covered by any reads [5].

LINEAR LAYOUT FOR VISUALIZATION OF NETWORKS

Martin Krzywinski, Katayoon Kasaian, Olena Morozova, Inanc Birol, Steven Jones, Marco Marra

mkweb.bcgsc.ca/linnet

BC CANCER AGENCY

Cydney Nielsen
Shaun Jackman
Rod Docking
Anthony Fejes
Dan Fornika
Jenny Qian

MASARYK UNIVERSITY

Martin Lysak