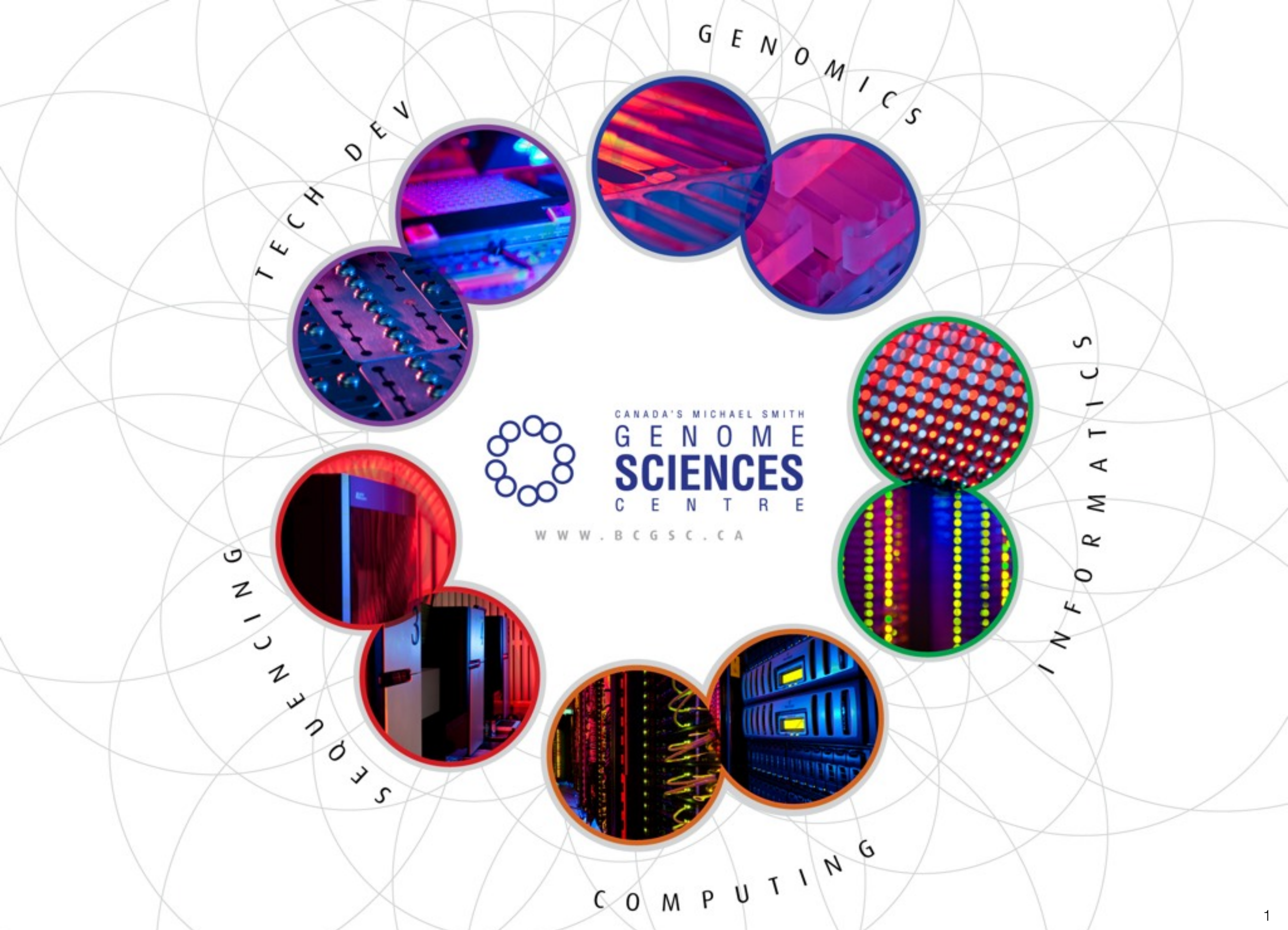


AFFILIATIONS BC Cancer Research Center • BC Cancer Agency • BC Cancer Foundation • Genome BC • Simon Fraser University • University of Britsh Columbia • Genome Sciences Institute

circos & hive plots
challenging visualization paradigms in genomics and network analysis

14.00 - 15.15

MARTIN KRZYWINSKI
Genome Sciences Center
BC Cancer Agency
Vancouver, Canada

PSA ANNUAL MEETING 2011
GENOMICS WORKSHOP
University of Washington
12 July 2011

BC Cancer Agency
CARE + RESEARCH
GENOME SCIENCES CENTRE

CIRCOS

TOOL

circular visualization of relationships and dense data

www.circos.ca

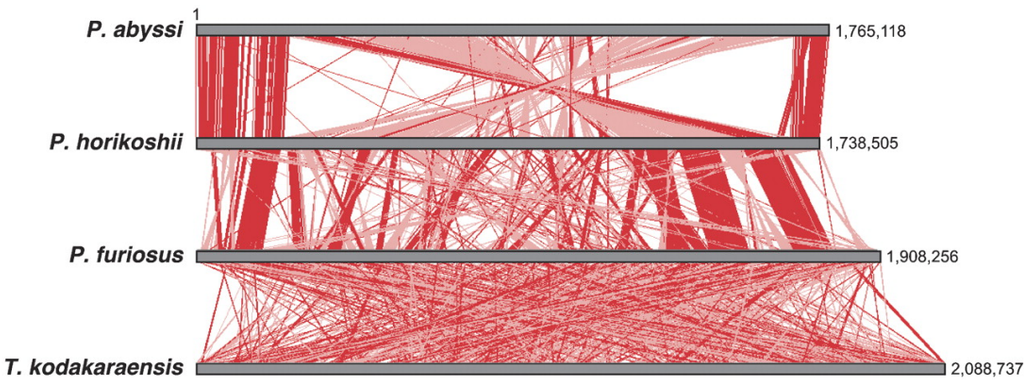
HIVE PLOTS

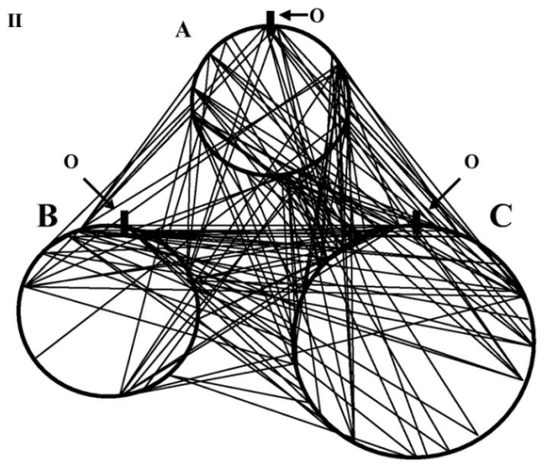
CONCEPT

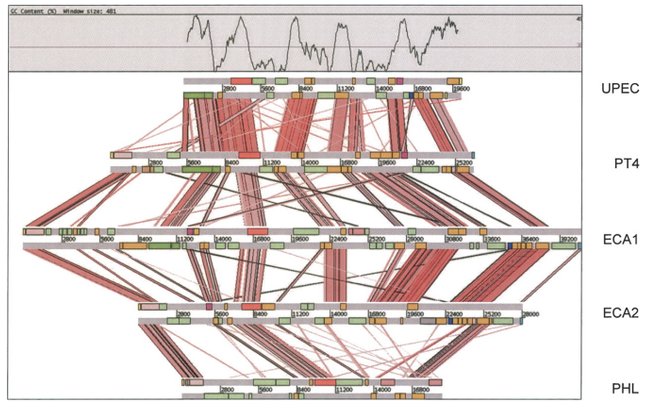
approach for rational, scalable and interpretable visualization of networks

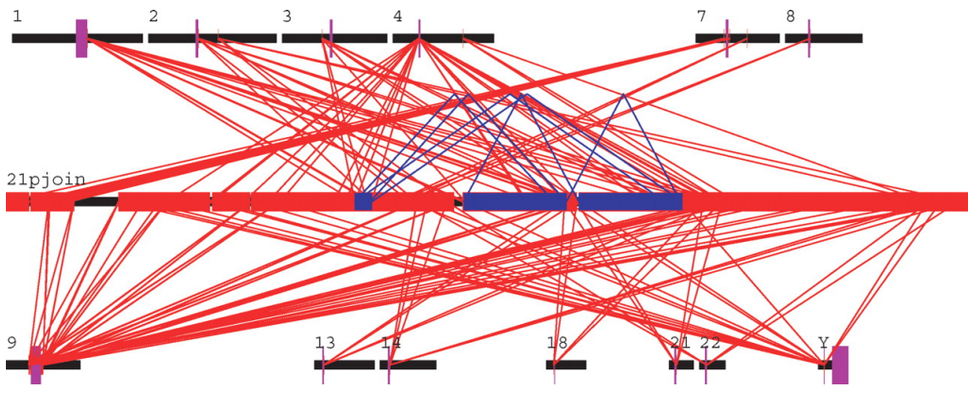
www.hiveplot.com

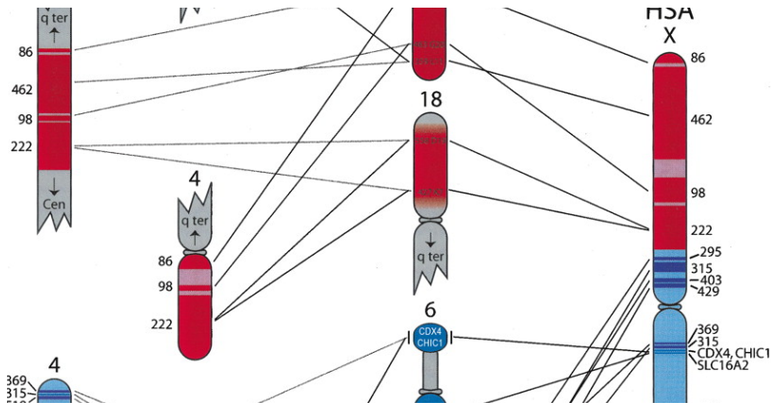
WHAT'S THE PROBLEM?

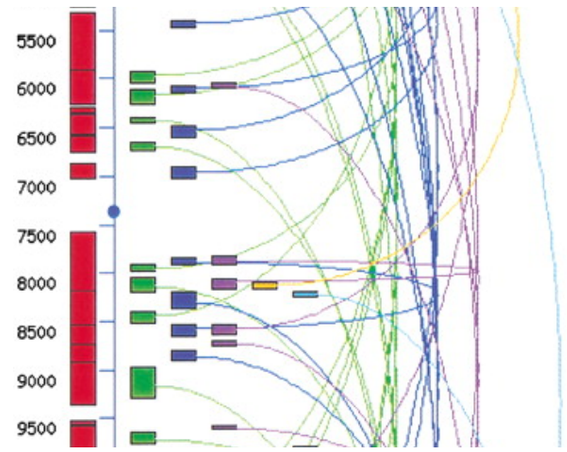


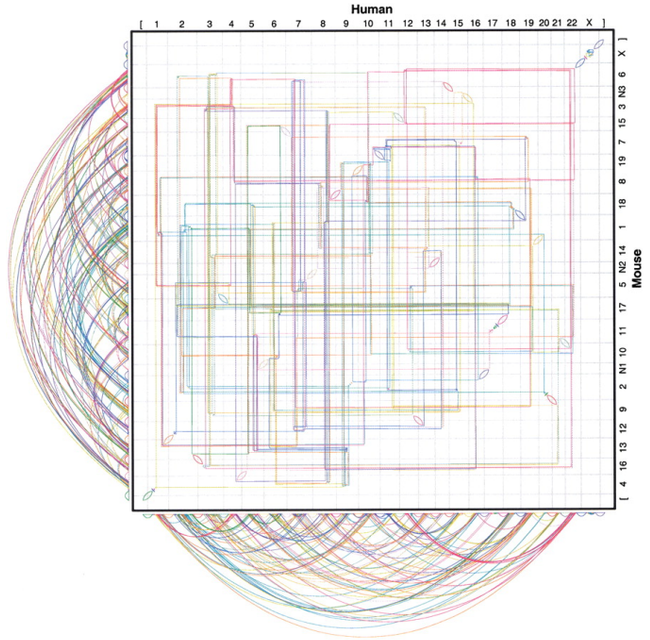


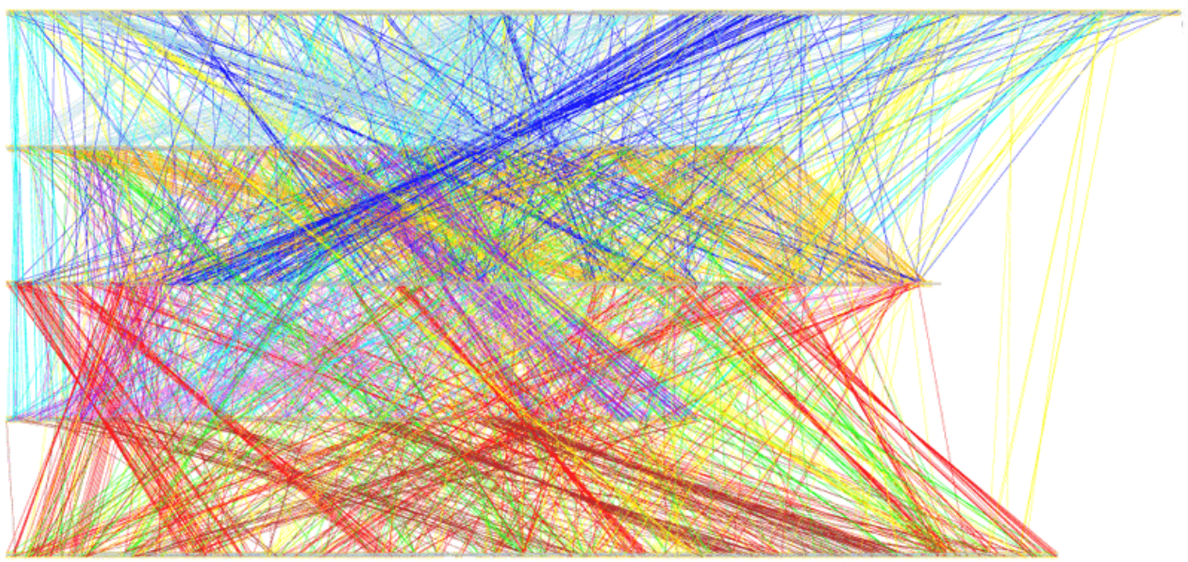


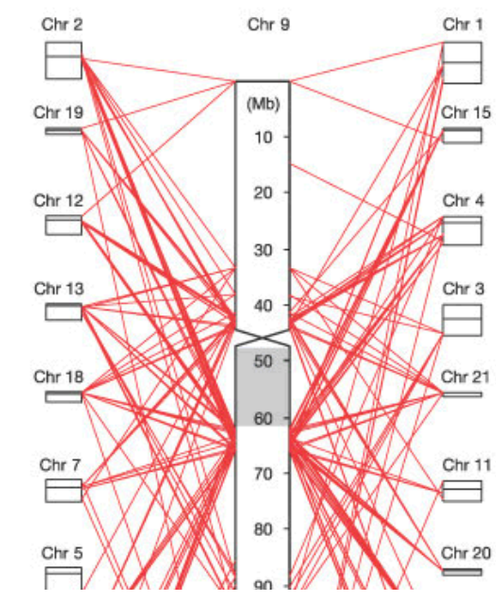






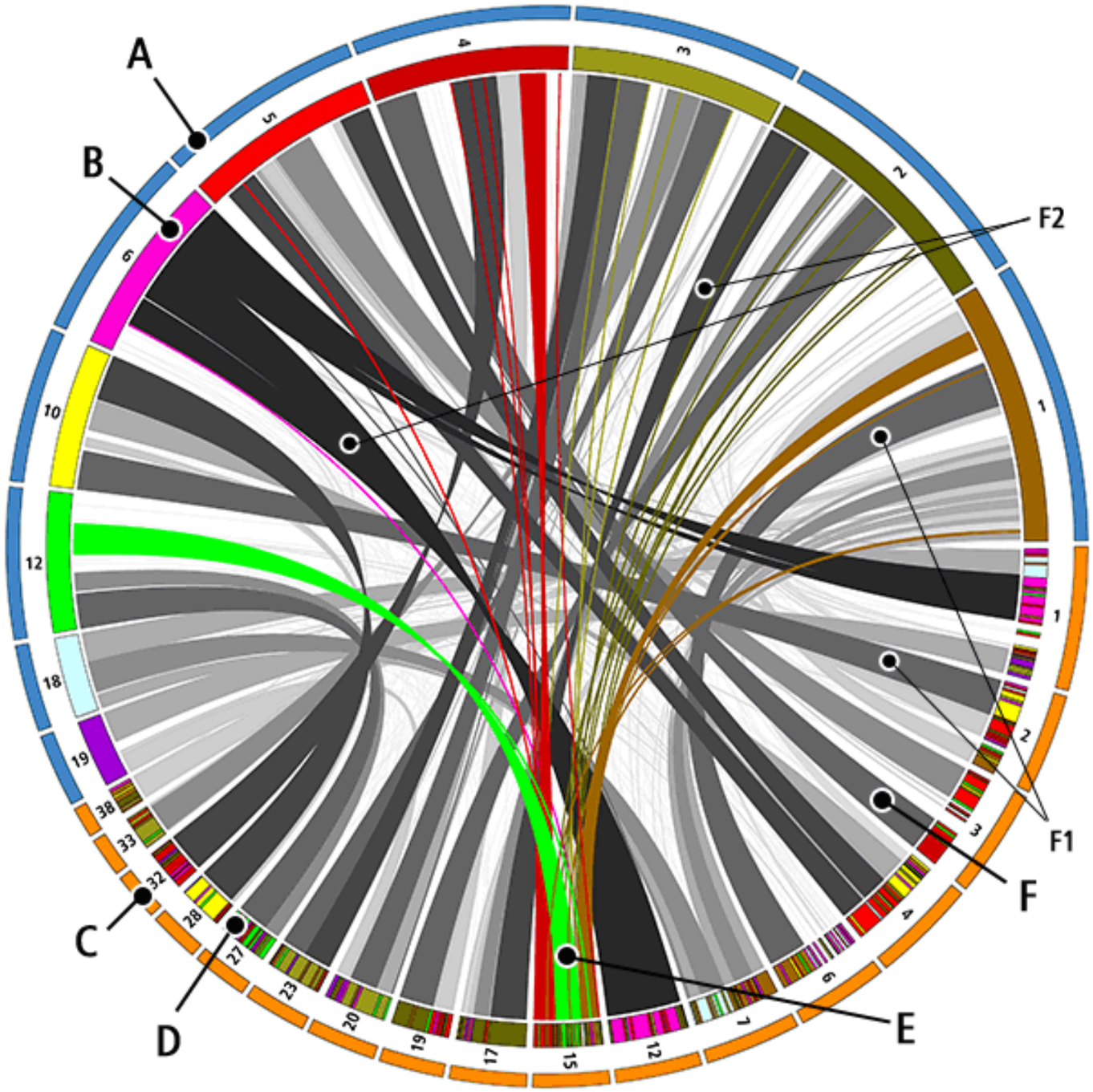






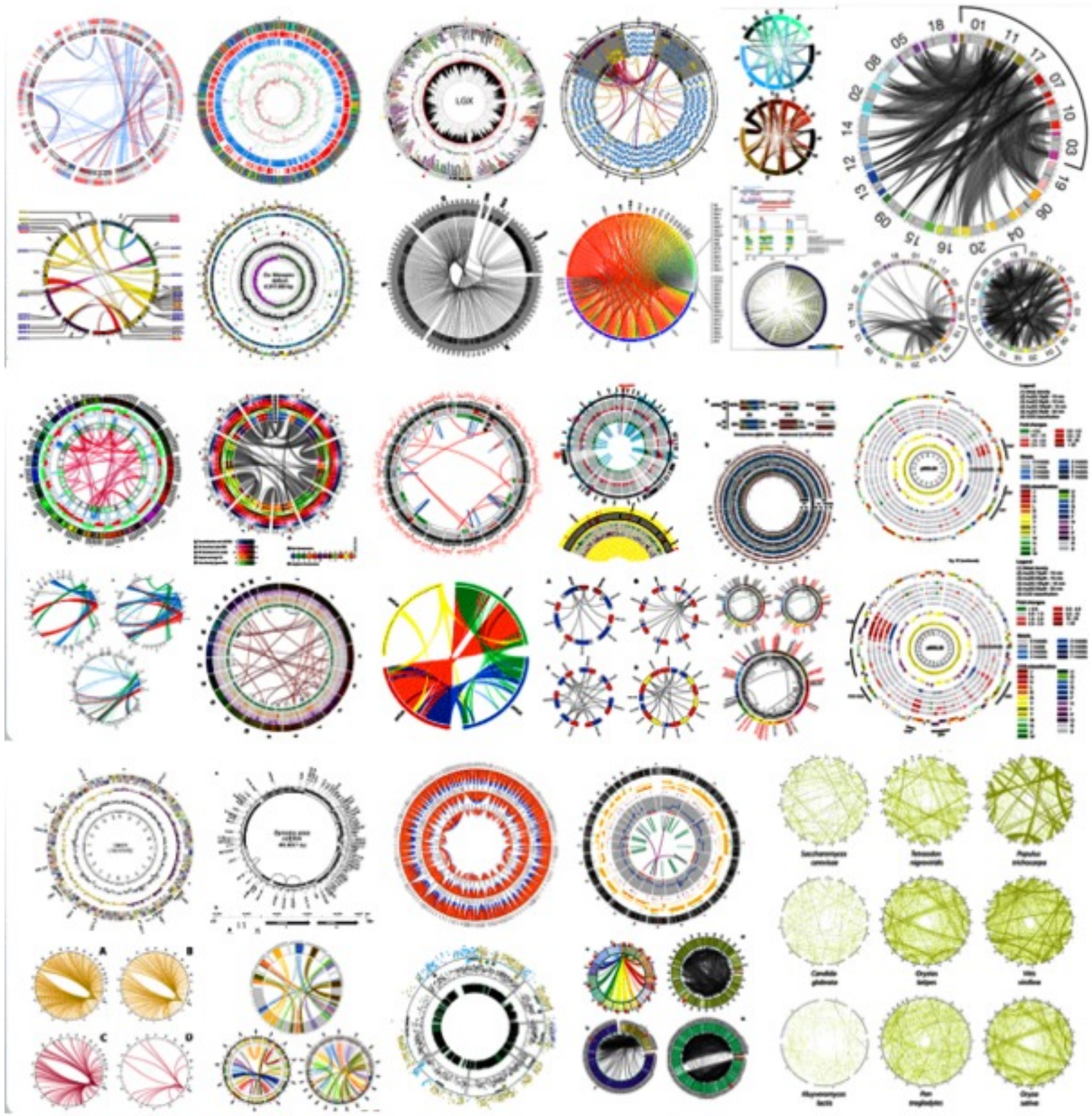
(1) Fukui, T., et al., Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. Genome Res, 2005. 15(3): p. 352-63. (2) Guo, X., et al., Natural genomic design in Sinorhizobium meliloti: novel genomic architectures. Genome Res, 2003. 13(8): p. 1810-7. (3) Thomson, N.R., et al., Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res, 2008. 18(10): p. 1624-37. (4) Lyle, R., et al., Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. Genome Res, 2007. 17(11): p. 1690-6. (5) Veyrunes, F., et al., Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res, 2008. 18(6): p. 965-73. (6) Blanc, G., K. Hokamp, and K.H. Wolfe, A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res, 2003. 13(2): p. 137-44. (7) Pevzner, P. and G. Tesler, Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res, 2003. 13(1): p. 37-45. (9) Humphray, S.J., Oliver K. et al (2004) DNA sequence and analysis of the human chromosome 9. Nature 429(6990): 369-74.

WHAT'S THE SOLUTION?



Regions of similarity between human and dog genomes. (A) human genome. (B) human ideograms. (C) dog genome. (D) dog ideograms, coded by most similar human chromosome. (E,F) link bundles connect similar regions. (F1) rules are used to color bundles by size. (F2) bundles twist when similarity involves opposite strands. American Scientist, Sept-Oct 2007. Cover figure by M Krzywinski.

PARADIGM SHIFT - ROUND IS THE NEW SQUARE



The circle has made its comeback.

CIRCOS IS WIDELY ACCEPTED

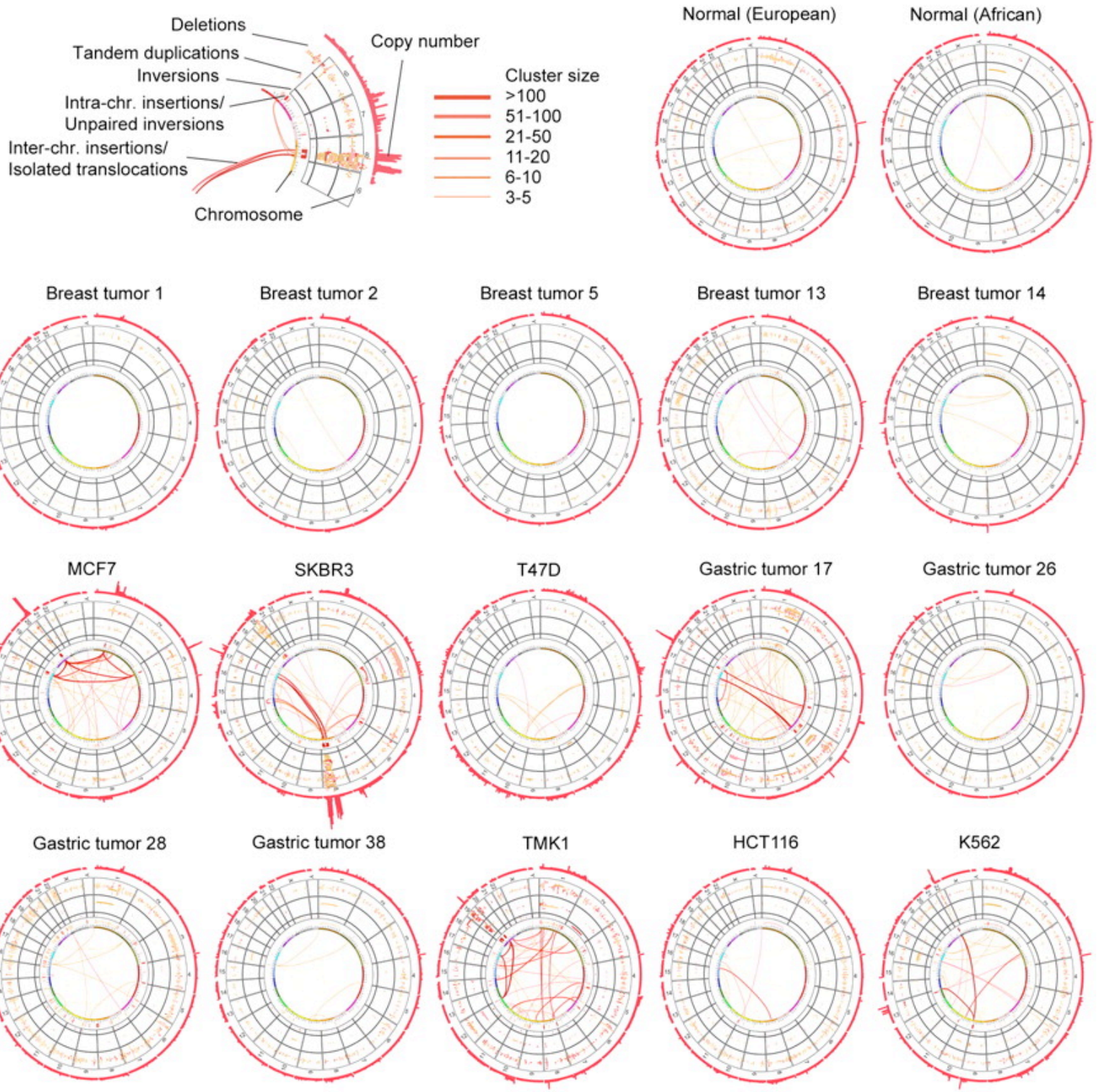
Bioinformatics · Genome Biology · nature · Science · Nucleic Acids Research

American Scientist · Genome Research · Portfolio · WIRED · PNAS

PLoS Public Library of Science · The Plant Cell · The New York Times · Leukemia · SEED

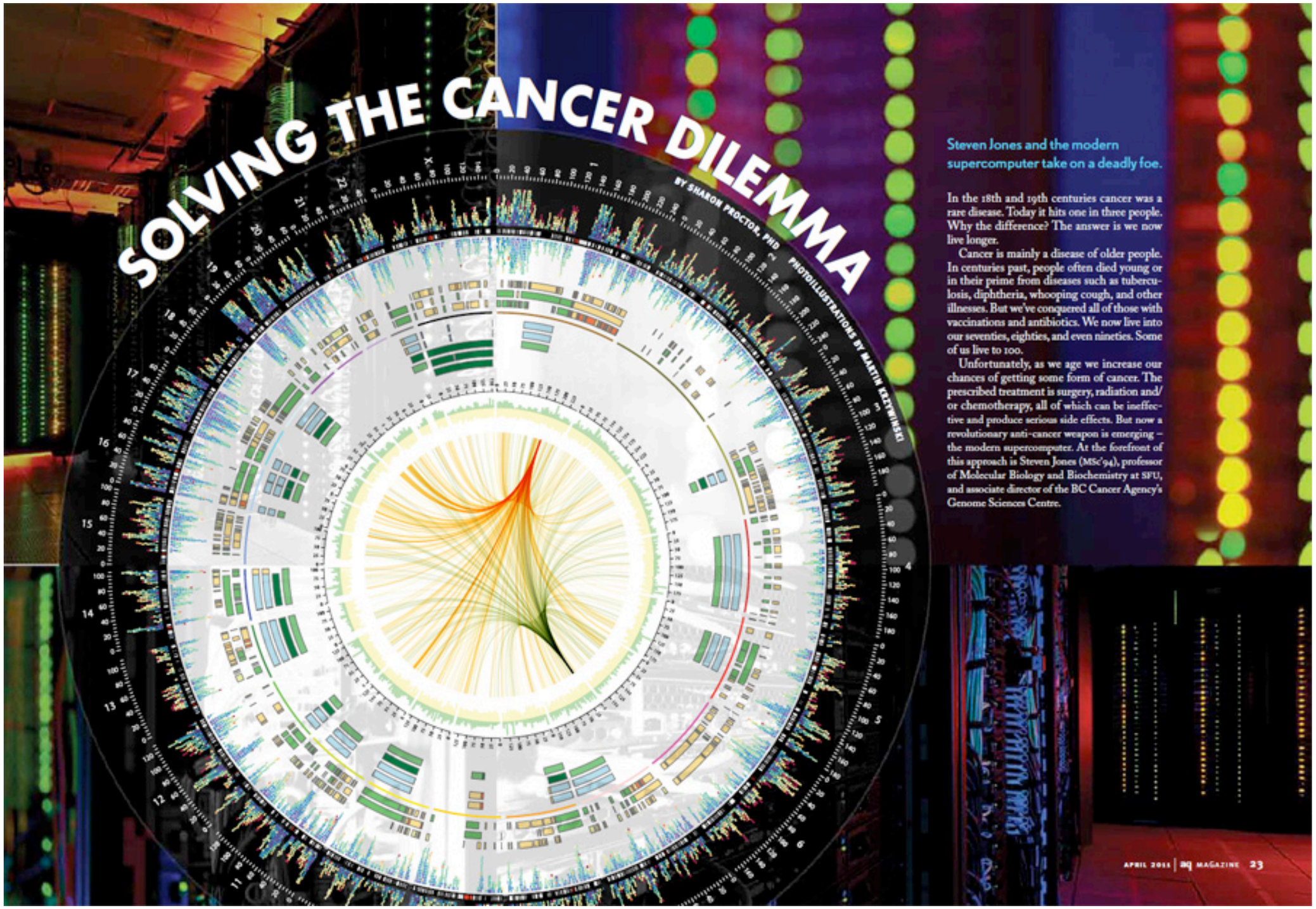
Circos has been accepted by the biological community as a standard for displaying sequence relationships and genome rearrangements.

PRIMARY LITERATURE



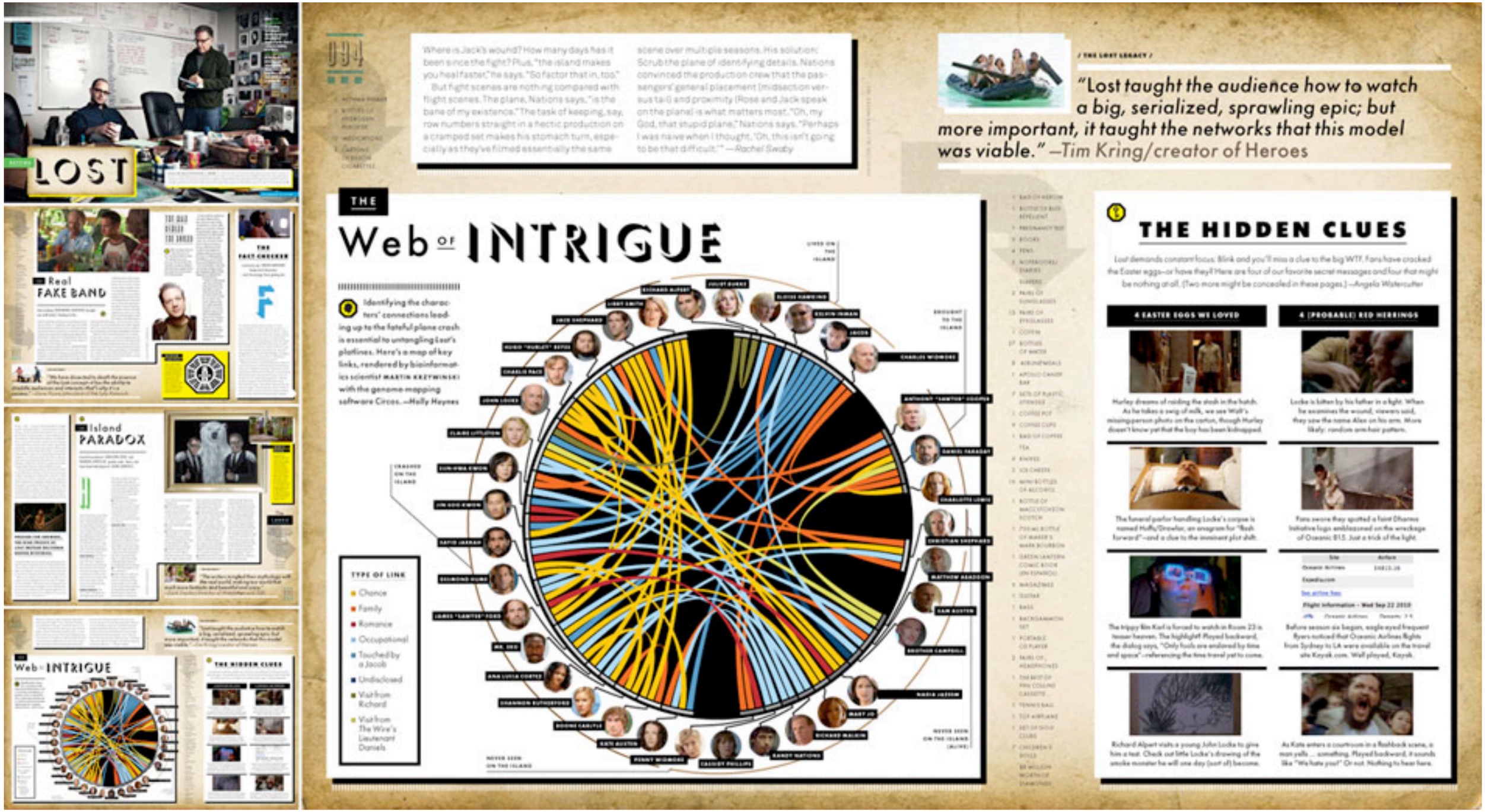
Hillmer AM, Yao F, Inaki K et al. 2011 Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome research 21:665-675.

POPULAR LITERATURE



AQ Magazine, April 2011 (Simon Fraser University). Figure by M Krzywinski.

POPULAR CULTURE



Wired, April 2010. Figure by M Krzywinski.

REVIEW LITERATURE

The CANCER GENOME challenge

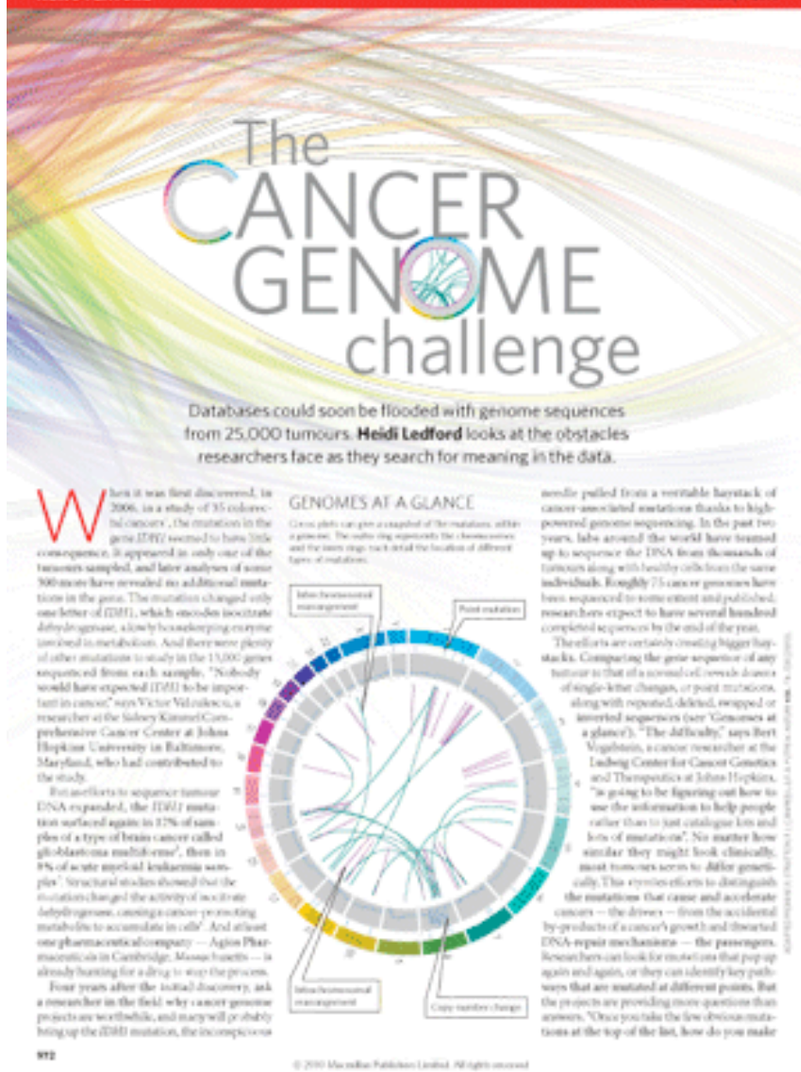
Databases could soon be flooded with genome sequences from 25,000 tumours. **Heidi Ledford** looks at the obstacles researchers face as they search for meaning in the data.

When it was first discovered, in 2006, in a study of 35 colorectal cancers¹, the mutation in the gene *IDH1* seemed to have little consequence. It appeared in only one of the tumours sampled, and later analyses of some 500 more have revealed no additional mutations in the gene. The mutation changed only one letter of *IDH1*, which encodes isocitrate dehydrogenase, a slowly housekeeping enzyme involved in metabolism. And there were plenty of other mutations to study in the 13,000 genes sequenced from each sample. "Nobody would have expected *IDH1* to be important in cancer," says Victor Velculescu, a researcher at the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University in Baltimore, Maryland, who had contributed to the study.

But as efforts to sequence tumour DNA expanded, the *IDH1* mutation surfaced again in 12% of samples of a type of brain cancer called glioblastoma multiforme, then in 8% of acute myeloid leukaemia samples². Structural studies showed that the mutation changed the activity of isocitrate dehydrogenase, causing a cancer-promoting metabolite to accumulate in cells³. And at least one pharmaceutical company — Agios Pharmaceuticals in Cambridge, Massachusetts — is already hunting for a drug to stop the process.

Four years after the initial discovery, ask a researcher in the field why cancer genome projects are worthwhile, and many will probably bring up the *IDH1* mutation, the inconspicuous needle pulled from a veritable haystack of cancer-associated mutations thanks to high-powered genome sequencing. In the past two years, labs around the world have teamed up to sequence the DNA from thousands of tumours along with healthy cells from the same individuals. Roughly 75 cancer genomes have been sequenced to some extent and published, researchers expect to have several hundred completed sequences by the end of the year.

The efforts are certainly creating bigger haystacks. Comparing the gene sequence of any tumour to that of a normal cell reveals dozens of single-letter changes, or point mutations, along with repeated, deleted, swapped or inverted sequences (see 'Genomes at a glance'). "The difficulty," says Bert Vogelstein, a cancer researcher at the Ludwig Center for Cancer Genetics and Therapeutics at Johns Hopkins, "is going to be figuring out how to use the information to help people rather than to just catalogue lots and lots of mutations". No matter how similar they might look clinically, most tumours seem to differ genetically. This stymies efforts to distinguish the mutations that cause and accelerate cancers — the drivers — from the accidental by-products of a cancer's growth and thwarted DNA-repair mechanisms — the passengers. Researchers can look for mutations that pop up again and again, or they can identify key pathways that are mutated at different points. But the projects are providing more questions than answers. "Once you take the few obvious mutations at the top of the list, how do you make sense of the rest of them?" asks Will Parsons, a paediatric oncologist at Baylor College of Medicine in Houston, Texas. "How do you decide which are worthy of follow up and functional analysis? That's going to be the hard part."



Drivers wanted

Because cancer is a disease so intimately associated with genetic mutation, many thought it would be amenable to genomic exploration through initiatives based on the collaborative model of the Human Genome Project. The International Cancer Genome Consortium (ICGC), formed in 2008, is coordinating efforts to sequence 500 tumours from each of 50 cancers. Together, these projects will cost in the order of US$1 billion. Eleven countries have already signed on to cover more than 20 cancers (see map). The ICGC includes two older, large-scale projects: the Cancer Genome Project, at the Wellcome Trust Sanger Institute near Cambridge, UK, and the US National Institutes of Health's Cancer Genome Atlas (TCGA). The Cancer Genome Project has churned out more than 100 partial genomes and roughly 15 whole genomes in various stages of completion, and intends to tackle 2,000–3,000 more over the next 5–7 years. TCGA, meanwhile, wrapped up a three-year, three-cancer pilot project last year, then launched a full-scale endeavour to sequence up to 500 tumours from each of more than 20 cancers over the next five years.

Although the groups collaborate, TCGA has not yet been able to fully join the ICGC owing to differences in privacy regulations governing access to genome data. For now, members of both consortia are sequencing a subset of tumour samples from each cancer type — around 100 — and will follow this by sequencing promising areas in the remaining 400. That's useful, says Joe Gray, a cancer researcher at Lawrence Berkeley National Laboratory in California, but it's just a start. "In the early days, I thought that doing a few hundred tumours would probably be sufficient," he says. "Even at the level of 1,000 samples, I think we're probably not going to have the statistics we want."

What bigger numbers could provide is more driver mutations like the one in *IDH1*. There could, researchers argue, provide the clearest route to developing new cancer therapies. Many scientists have looked for mutations that occur repeatedly in a given type of tumour. "If there are lots and lots of abnormalities of a particular gene, the most likely explanation is often that those mutations have been selected for by the cancers and therefore they are cancer-causing," says Michael Stratton, who co-directs the Cancer Genome Project. This approach has worked well in some cancers. For example, with a frequency of 12%, it is clear that the *IDH1* mutation is a driver in glioblastoma. Such searches should be fruitful for cancers that have fewer mutations overall. The full gene-sequence of acute myeloid leukaemia cells yielded just ten mutations in protein-coding genes, eight of which had not previously been linked with cancer².

Other cancers have proved more challenging. *IDH1* was overlooked at first, on the basis of the colorectal cancer data alone. It was not until the search was expanded to other cancers that its importance was revealed. Moreover, some mutations shown to be drivers haven't turned up as often as expected. "It's very clear now that all the genes have been sequenced in this many tumours, you have drivers that are mutated at very low frequency, in less than 1% of the cancers," says Vogelstein. To find these low-frequency drivers, researchers are sampling heavily — sequencing 500 samples per cancer should reveal mutations that are present in as few as 3% of the tumours. Although they may not contribute to the majority of tumours, they may still have important biological lessons, says Stratton. "We need to know about these to understand the overall genomic landscape of cancer."

> "It's going to take good old-fashioned biology to really determine what these mutations are doing."

Another popular approach has been to look for mutations that cluster in a pathway, a group of genes that work together to carry out a specific process, even if the mutations strike it at different points. In an analysis of 24 pancreatic cancers⁴, for instance, Vogelstein and his colleagues identified 12 signalling pathways that had been altered. Nevertheless, Vogelstein cautions that this approach is not easy to pursue. Many pathways overlap, and their boundaries are unclear. And because many have been defined using data from different animals or cell types, they do not always match what's found in a specific human tissue. "When you layer on top of that the fact that the cancer cell is not wired the same as a normal cell, that raises even further difficulties," says Vogelstein.

How much is enough?

Separating drivers from passengers will become even more difficult as researchers move towards sequencing entire tumour genomes. To date, only a fraction of the existing cancer genomes are complete sequences. To keep costs low, most have covered only the exome, the 1.5% of the genome that does the codes for protein and is therefore the easiest to interpret. Assigning importance to a mutation found in the murky non-protein-coding depths of the genome will be more challenging, especially given that scientists don't yet know what function — if any — most of these regions usually serve. The vast majority of mutations fall here. The full genome sequence of a lung cancer cell line, for example, yielded 22,910 point mutations, only 134 of which were in protein-coding regions (see graphic, left)⁵. Nevertheless, finding them is worth the cost and effort, argues Stratton. "It could be that none of those mutations pertain to the causation of cancer," he says. "But it equally could be that some do, but we'll never find out unless we systematically investigate."

> "Even at the level of 1,000 samples, I think we're probably not going to have the statistics we want."

Not everyone agrees. Some researchers argue that the costs of cancer-genome projects currently outweigh the benefits. Prices are poised to drop dramatically in the next few years as a new generation of sequencing machines comes online, says Ari Melnick, a cancer researcher at Weill Cornell Medical College in New York. "Why not wait for that?" he asks. In the meantime there are lower-hanging fruit to pick, says Stephen Elledge, a geneticist at Harvard Medical School in Boston, Massachusetts. Mutations that affect how many copies of a gene are found in a genome, he argues, are cheaper to assess and provide a more intuitive insight into biological processes. "If you delete something, you can turn a pathway off very efficiently," he says. "And if you amplify something, you can increase flow through the pathway. Making point mutations in genes to activate them is a little dicier."

Changes in gene copy number can be detected using fast, relatively inexpensive array-based technologies, but sequencing can provide a higher-resolution snapshot of those regions, says Elaine Mardis, a sequencing specialist at Washington University in St Louis, Missouri. Sequencing can enable researchers to map the boundaries of insertions and duplications with more precision and to catch tiny duplications or deletions that might have gone undetected by an array. Mardis, along with her colleague Richard Wilson and others, used sequencing to detect overlapping deletions in a breast cancer that had spread to other parts of the body (see page 999)⁶. The deletions spanned the region containing *CDKN2A*, a gene thought to suppress the spread, or metastasis, of cancer.

Meanwhile, cancer genomics is spreading out from under the large, centralized projects. For example, a $65-million, three-year paediatric-cancer genome project headed by researchers at St Jude Children's Research Hospital in Memphis, Tennessee, and Washington University aims to sequence 600 tumours. And more small projects seem poised to pop up. "Pretty much any cancer centre with any interest in the genomics of cancer is now buying those sequencers and using them," says Sam Aparicio, a cancer researcher at the University of British Columbia in Vancouver, Canada.

Part of the reason that cancer-genome proponents don't want to wait for sequencing costs to drop is that the real work starts after the sequencing is over. As Velculescu puts it, "Ultimately it's going to take good old-fashioned biology and experimental analyses to really determine what these mutations are doing." With this in mind, the US National Cancer Institute established two 2-year projects in September last year to develop high-throughput methods to test how the mutations identified by the TCGA pilot project affect cell function. The two centres — one at the Dana-Farber Cancer Center in Boston, and another at Cold Spring Harbor Laboratory in New York — aim to systematize the way that researchers pull other needles like the *IDH1* mutation from the cancer-genomes haystack and make sense of them. The Boston team will systematically amplify and reduce the expression of genes of interest in cell cultures, and the Cold Spring Harbor centre will study cancer-associated mutations using tumours transplanted into mice.

In addition, large-scale projects are being run in parallel with the cancer-sequencing consortia to assess the effects of deleting each gene in the mouse genome, enabling researchers to learn more about the normal function of genes that are mutated in cancer. Sequencing is all very well, researchers have realized, but it won't be enough. "Some people say statistics should get us all the drivers that are worthwhile," says Lynda Chin, an investigator with TCGA at Harvard Medical School. "I don't agree with that. At the end of the day, we need these functional studies to prioritize the list of potential cancer-relevant candidates."

Heidi Ledford is a reporter for *Nature* in Cambridge, Massachusetts.

See also 'Rivers and Views', page 989.

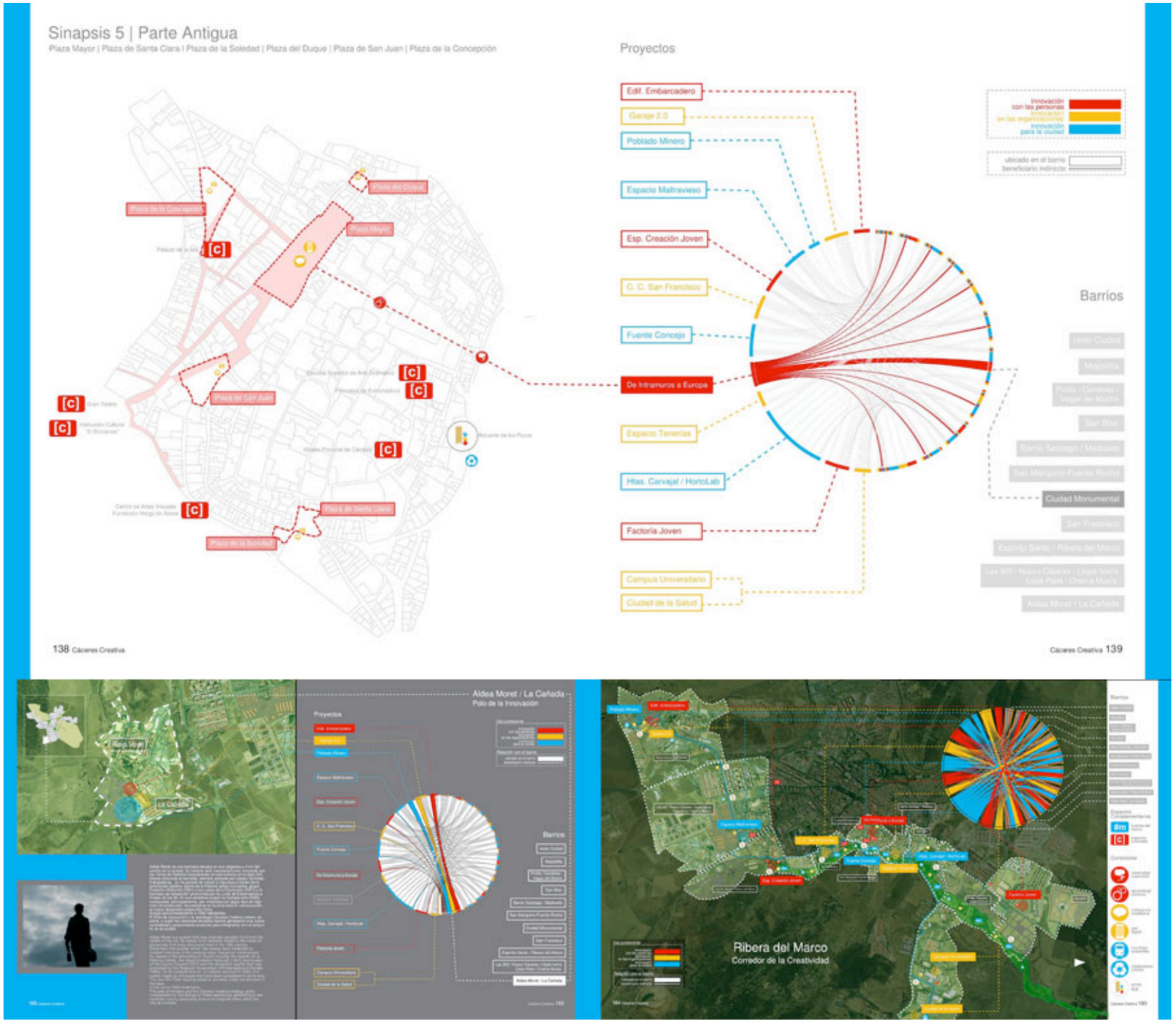
1. Sjöblom, T. et al. Science 314, 268–274 (2006)
2. Parsons, D. W. et al. Science 321, 1807–1812 (2008)
3. Mardis, E. R. et al. N. Engl. J. Med. 361, 1058–1066 (2009)
4. Jones, S. et al. Science 321, 1801–1806 (2008)
5. Pleasance, E. D. et al. Nature 463, 184–190 (2010)
6. Ding, L. et al. Nature 464, 999–1005 (2010)

Ledford H 2010 Big science: The cancer genome challenge. Nature 464:972-974.

BC Cancer Agency CARE • RESEARCH

GENOME SCIENCES CENTRE

URBAN PLANNING

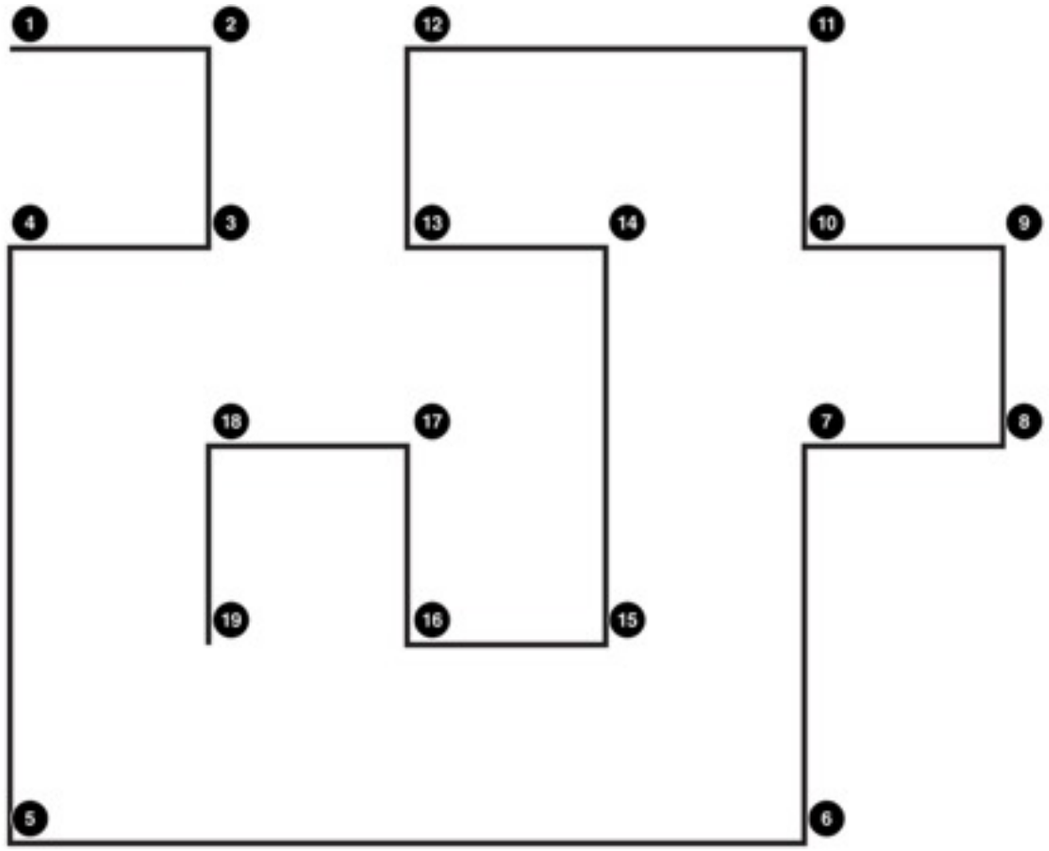


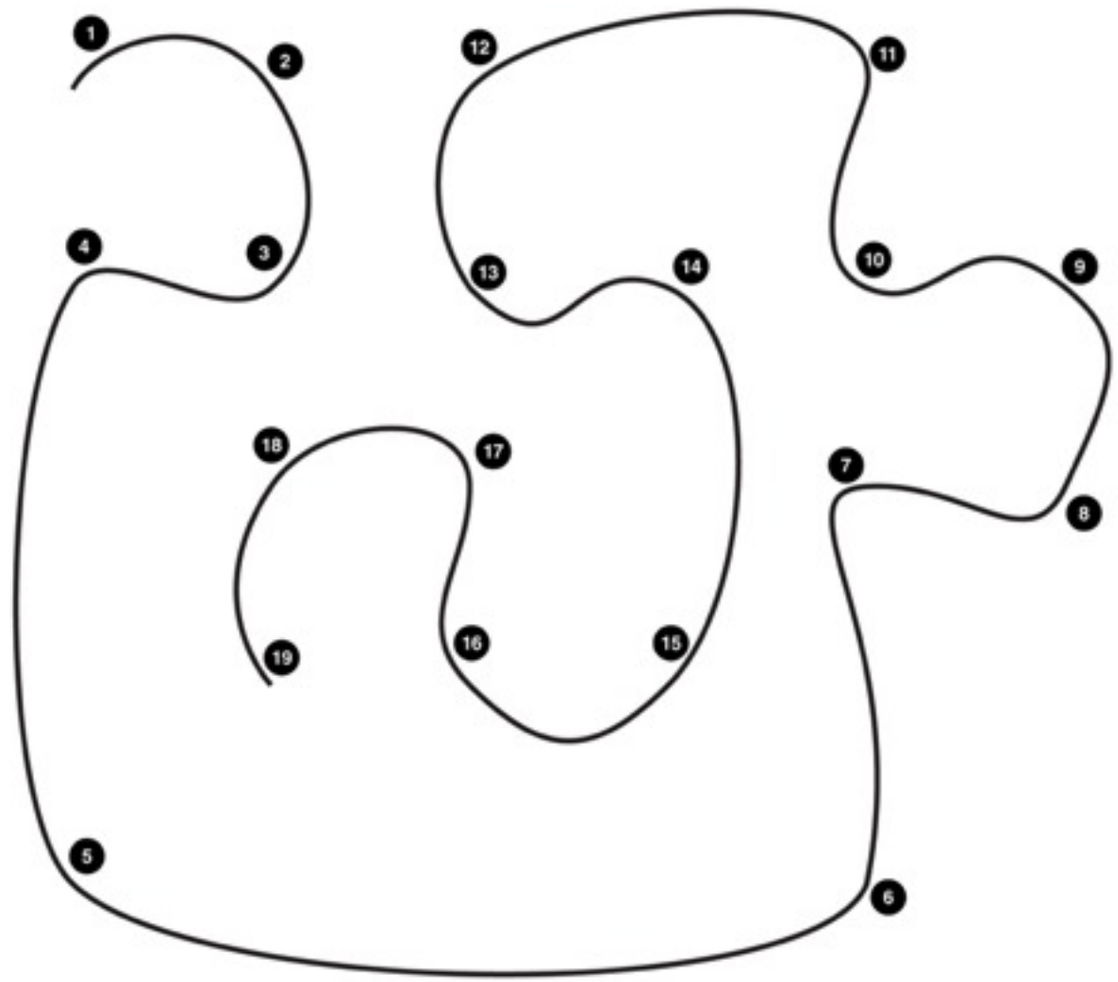
The town of Caceres, Spain, a UNESCO World Heritage Site, used Circos to illustrate the relationships between businesses in their urban planning strategy.

ADVERTISING

.

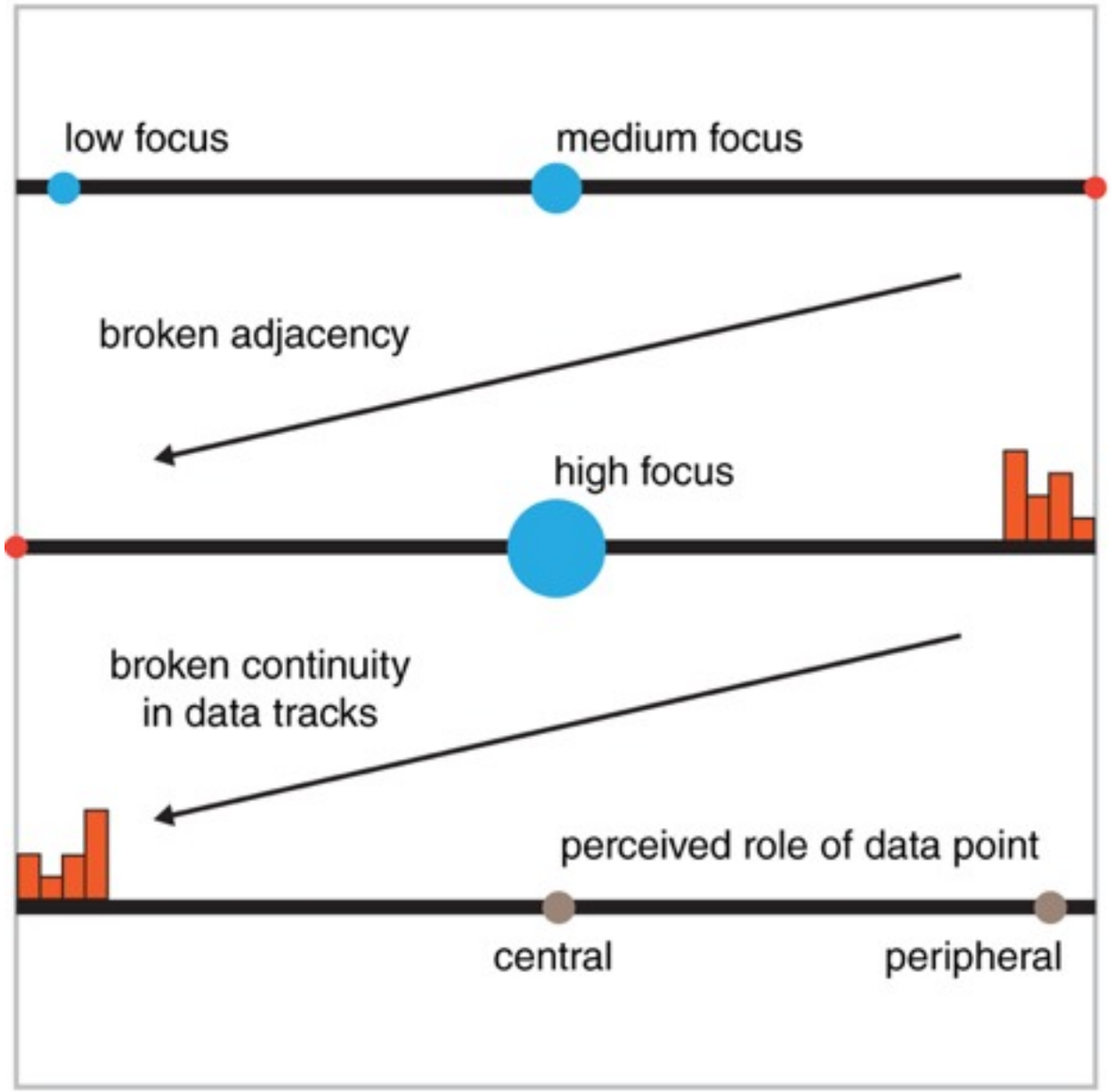
WHY CIRCLES?

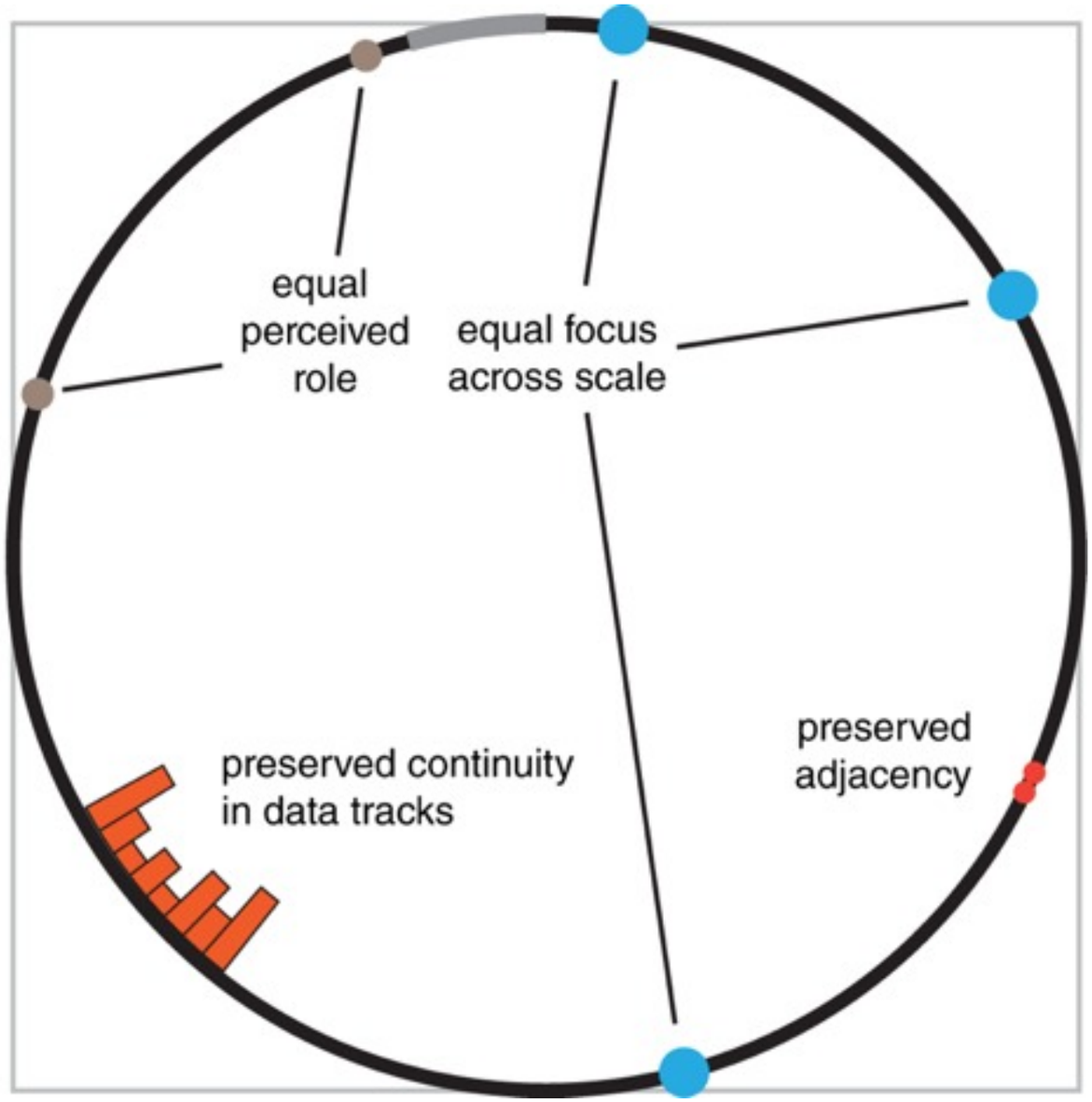




Moving your eye across the curved path is faster and more comfortable.

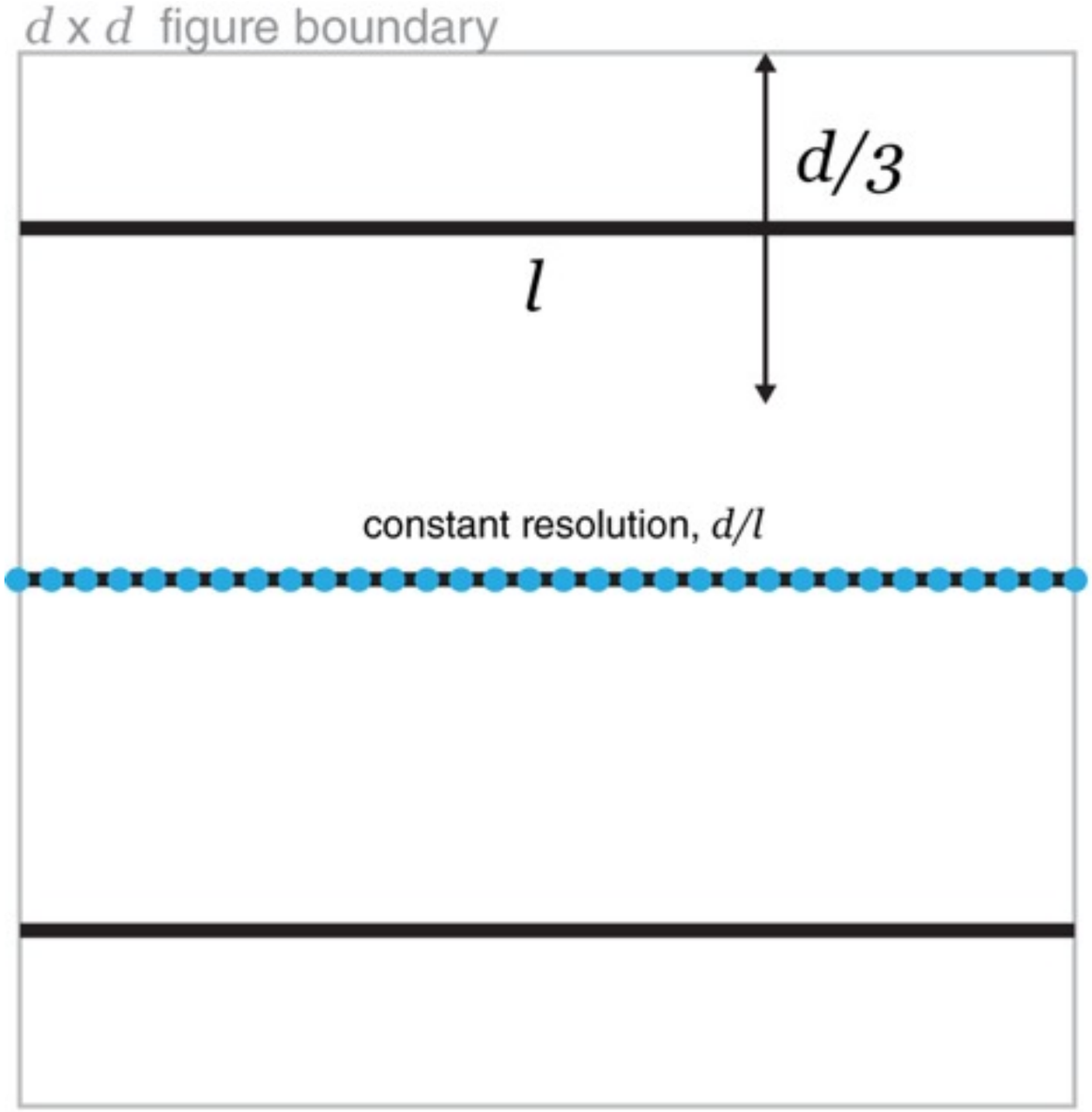
WHY CIRCLES?



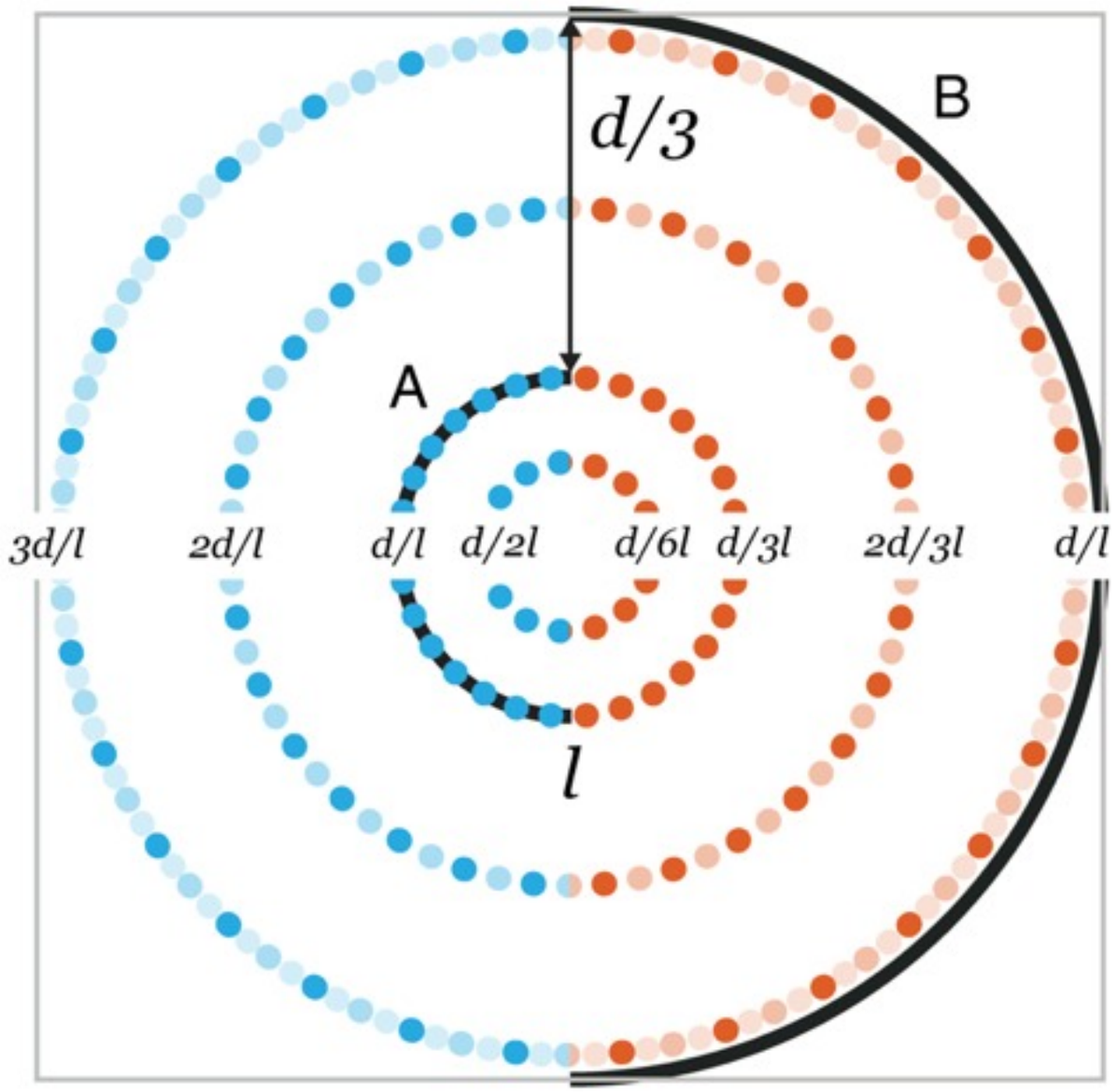


Linear layout of scale has disadvantages of changing focus (regions in the center of the image receive more attention), broken adjacency (neighbouring points on a linear scale are separated), broken continuity (data tracks are difficult to follow from one edge of the figure to another), and non-uniform data emphasis (center and edge of the axis are not perceived uniformly - the edge implies periphery, which may not apply.

WHY CIRCLES?



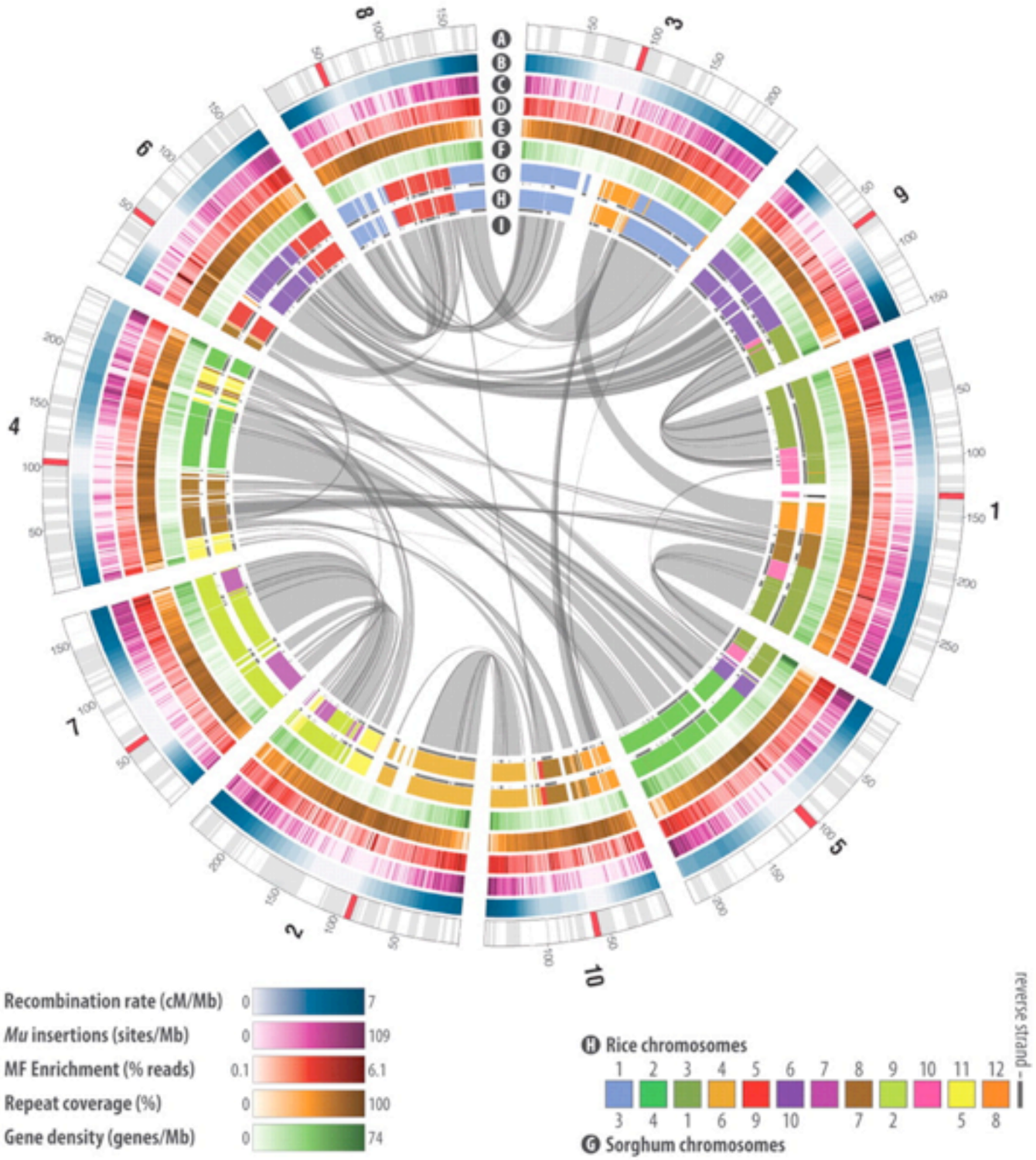
$L = 3l$



$L' = \pi d = 1.05L$

The circular layout accommodates variable resolution.

TYPICAL CIRCOS IMAGE

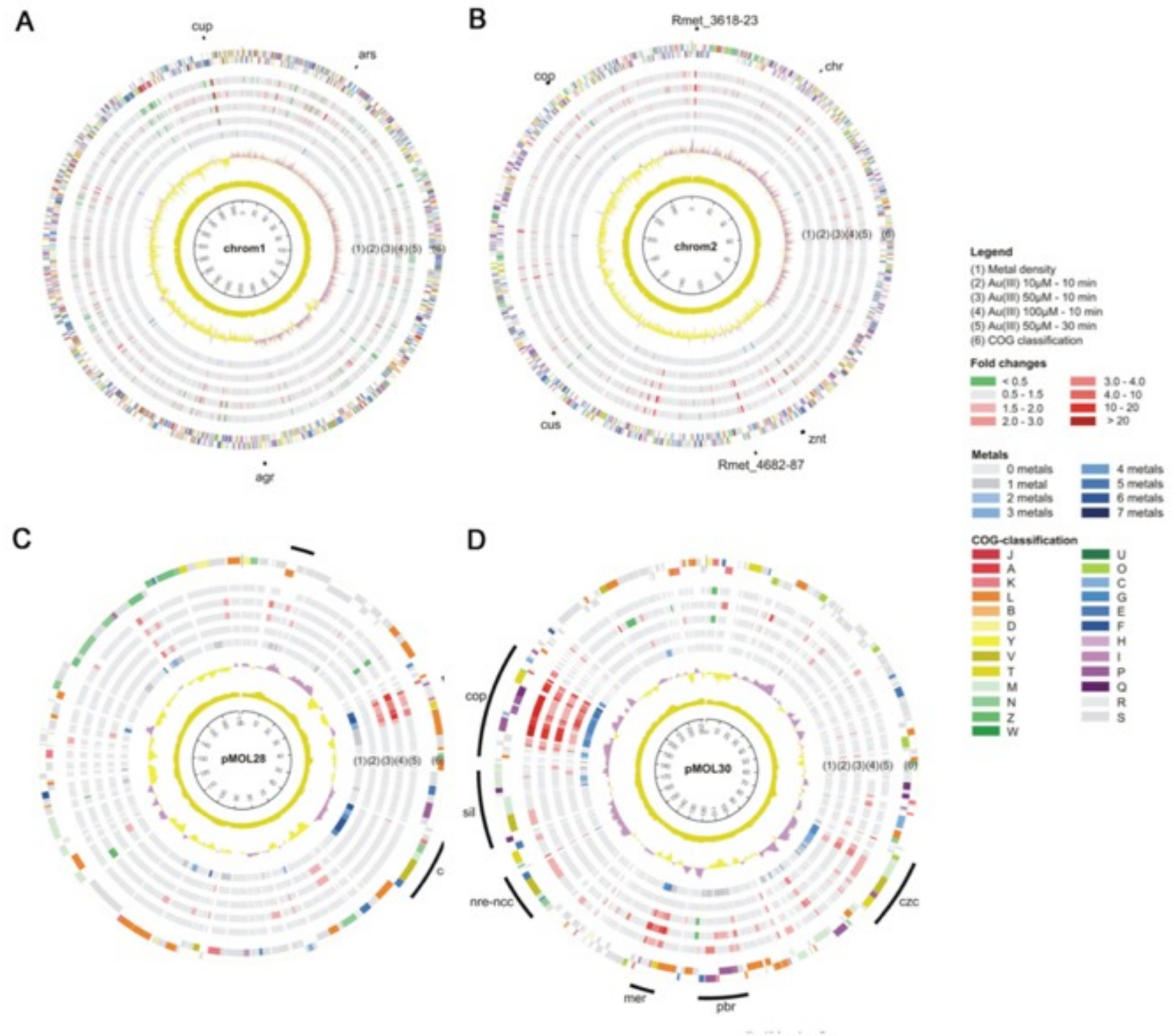


The maize B73 reference genome
B73 RefGen_v1

Concentric circles show aspects of the genome. Chromosome structure (A). Reference chromosomes with physical fingerprint contigs (11) as alternating gray and white bands. Presumed centromeric positions are indicated by red bands (31); enlarged for emphasis. Genetic map (B). Genetic linkage across the genome, on the basis of 6363 genetically and physically mapped markers (14, 19). Mu insertions (C). Genome mappings of nonredundant Mu insertion sites (14, 19). Methyl-filtration reads (D). Enrichment and depletion of methyl filtration. For each nonoverlapping 1-Mb window, read counts were divided by the total number of mapped reads. Repeats (E). Sequence coverage of TEs with RepeatMasker with all identified intact elements in maize. Genes (F). Density of genes in the filtered gene set across the genome, from a gene count per 1-Mb sliding window at 200-kb intervals. Sorghum synteny (G) and rice synteny (H). Syntenic blocks between maize and related cereals on the basis of 27,550 gene orthologs. Underlined blocks indicate alignment in the reverse strand. Homoeology map (I). Oriented homoeologous sites of duplicated gene blocks within maize.

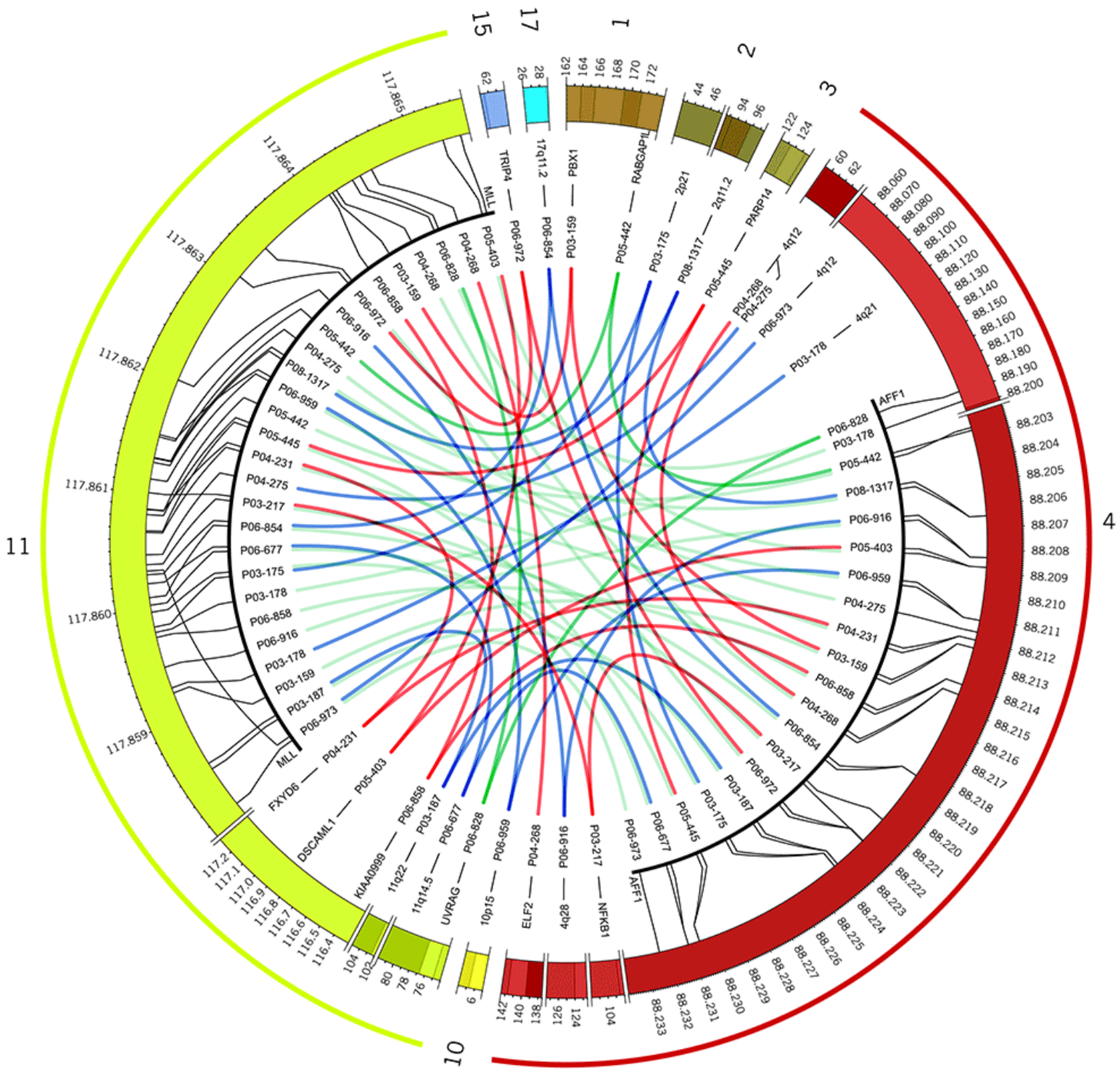
Schnable PS Ware D Fulton RS et al. 2009 The B73 maize genome: complexity, diversity, and dynamics Science 326 1112-1115.

INFORMATION-DENSE, BUT PARSABLE



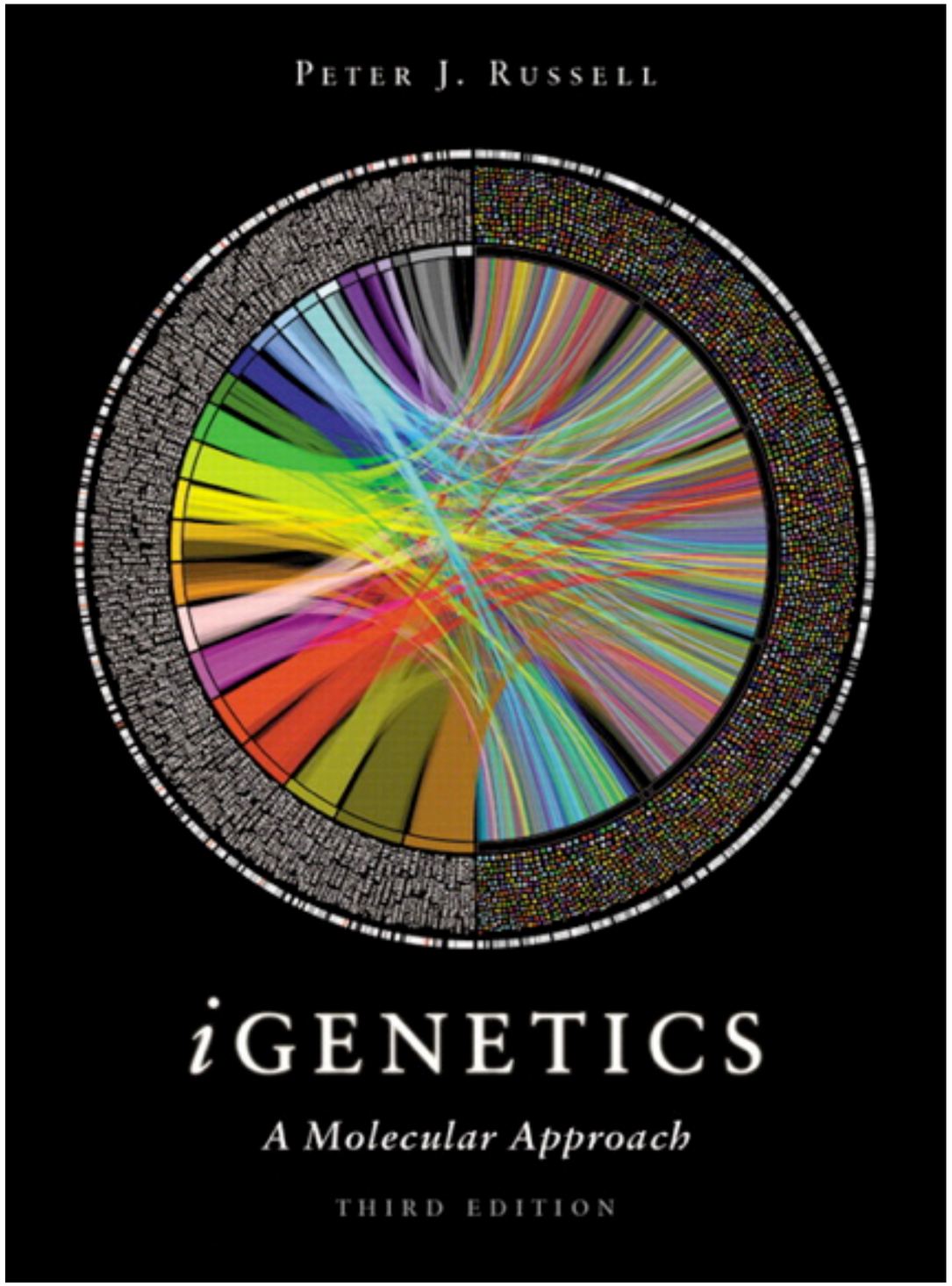
Reith F, Etschmann B, Grosse C et al. 2009 Mechanisms of gold biomineralization in the bacterium Cupriavidus metallidurans Proc Natl Acad Sci U S A 106 17757-17762.

FLEXIBLE IDEOGRAM LAYOUT AND CROPPING



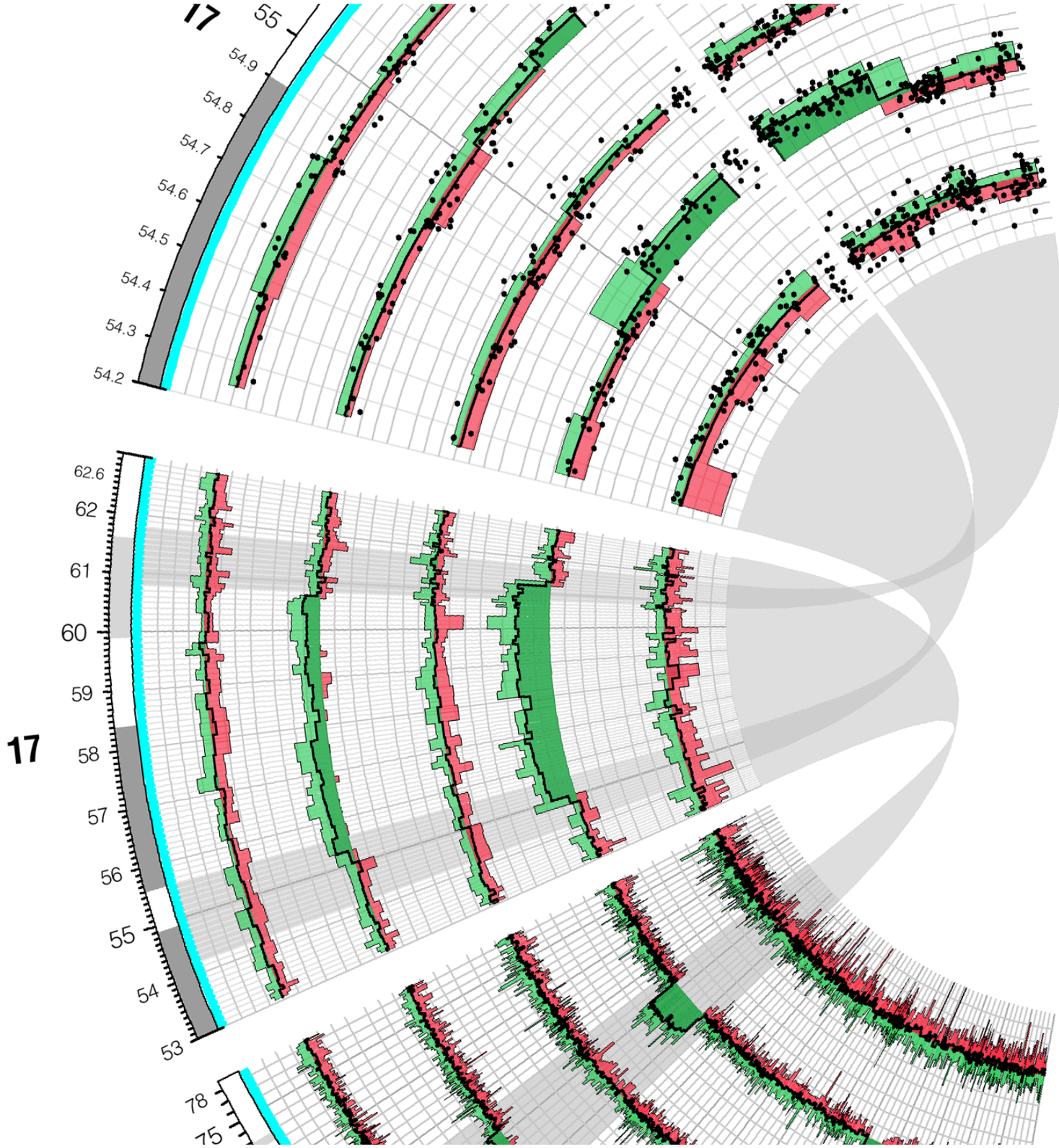
The most frequent complex rearrangements involving MLL and (A) AFF1/AF4. Localization of chromosomal breakpoints and UPN of individual patients are indicated. Colored lines indicate in-frame fusions (green), out-of-frame fusions (red), no partner gene present at the recombination site (blue).
Meyer, C., E. Kowarz, et al. (2009). "New insights to the MLL recombinome of acute leukemias." Leukemia 23(8): 1490-1499. Figure by M Krzywinski.

LINK BUNDLES



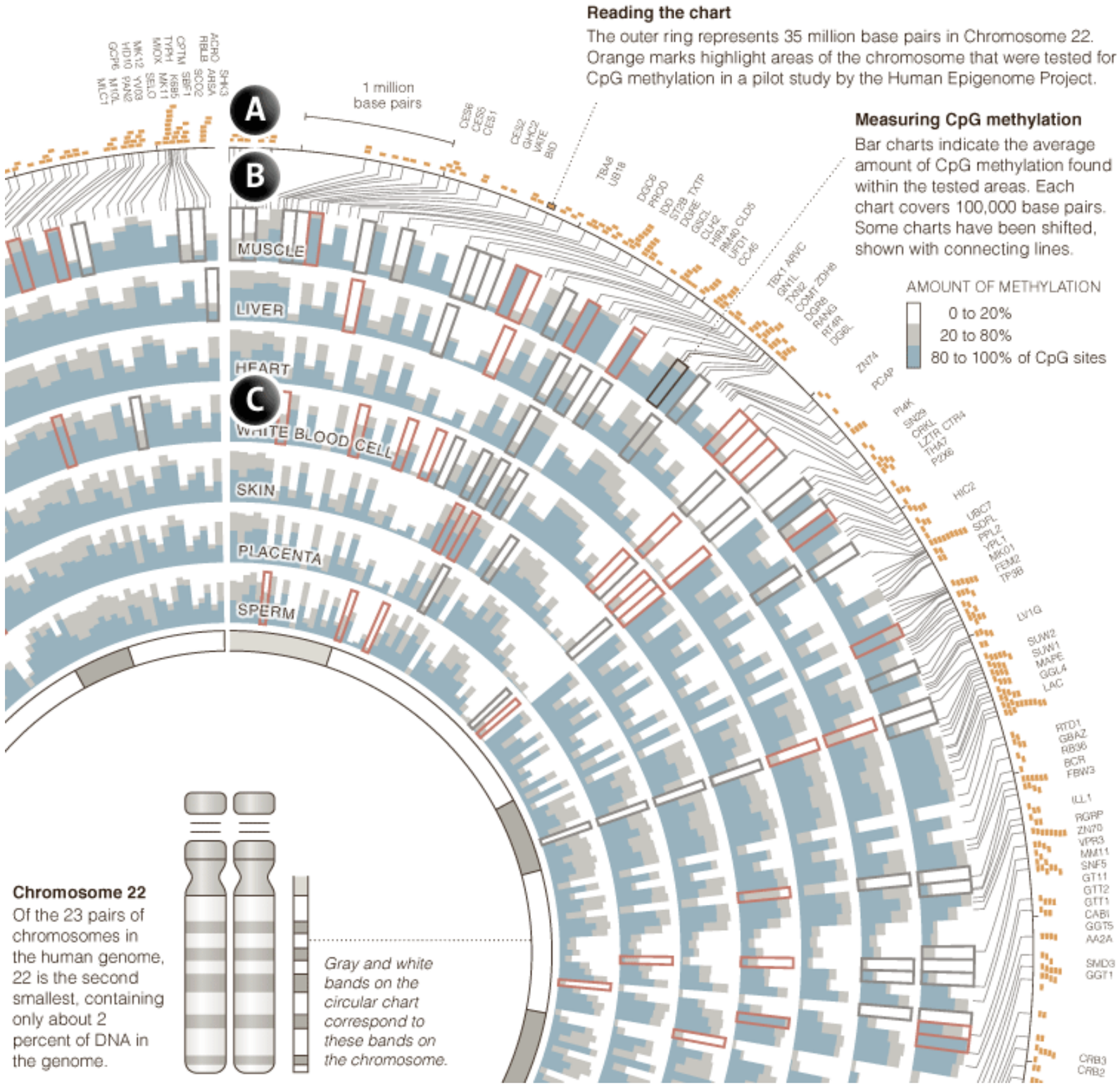
LEFT Regions of similarity between human and dog genomes. American Scientist, Sept-Oct 2007. Figure by M Krzywinski.
RIGHT Similarity between genes in human and fly (D. melanogaster) genomes. Russell, P. J. (2010). iGenetics: A Molecular Approach, Benjamin Cummings. Figure by M Krzywinski.

COMPOUND DATA TRACKS



Various types of data tracks can be stacked. Five instances of a compound track each represent copy number information from a different sample. Using links and highlights, attention is drawn to the progression of scale increase within chr17:53-63Mb. Krzywinski, M., J. Schein, et al. (2009). "Circos: an information aesthetic for comparative genomics." Genome Res 19(9): 1639-1645.

NOT JUST FOR RELATIONSHIPS

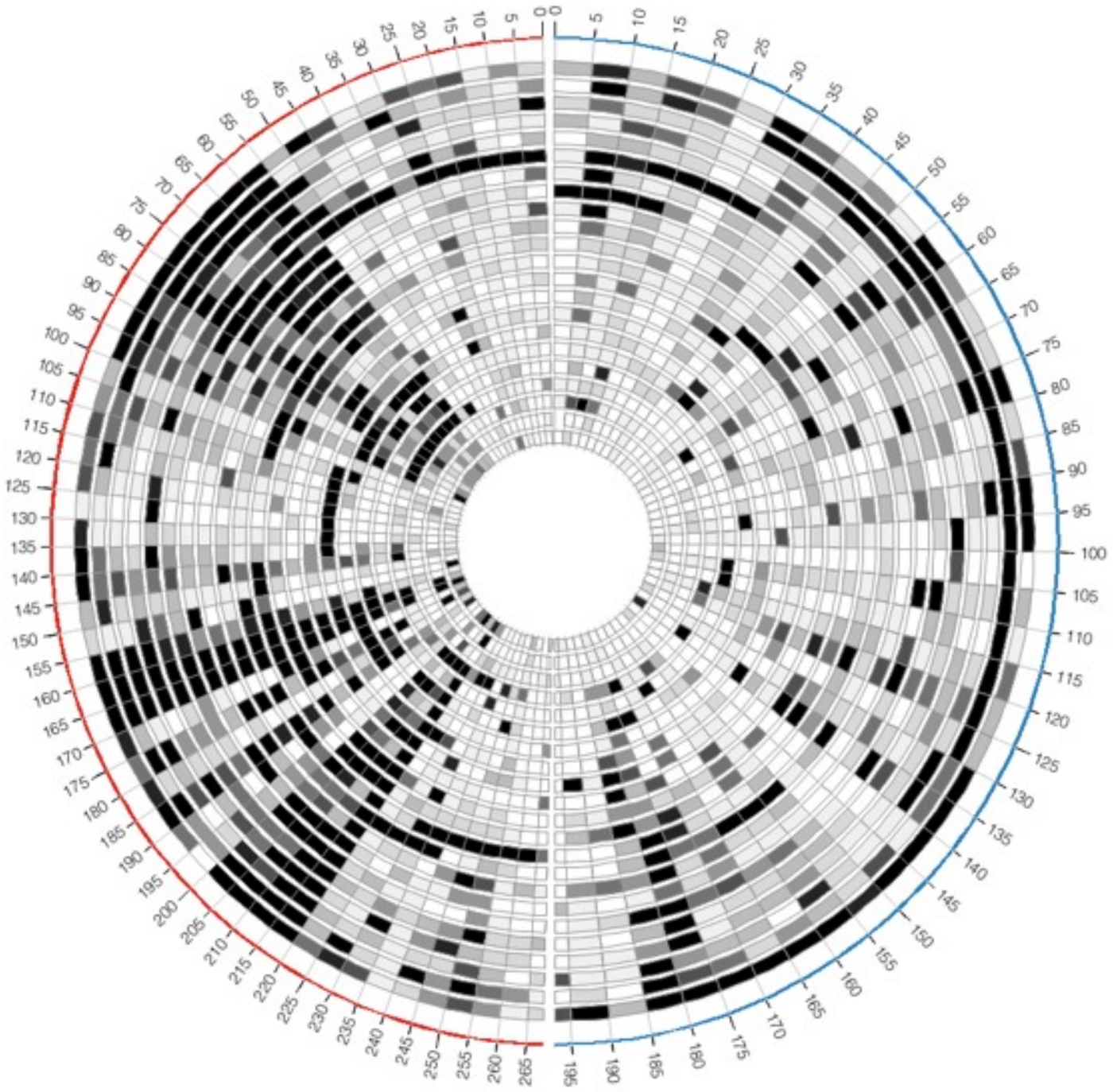


Data sets which do not sample the genome uniformly (A) can be effectively shown by using a connector track (B) to show the remapping onto an index scale (C). Shown in the figure are methylation values (A) for 7 tissues are summarized using stacked histograms (C), whose bins represent statistics for remapped methylation probe positions. Zimmer, C. (2008). Now: The Rest of the Genome. New York Times. Figure by M Krzywinski.

dynamic parameters and rules

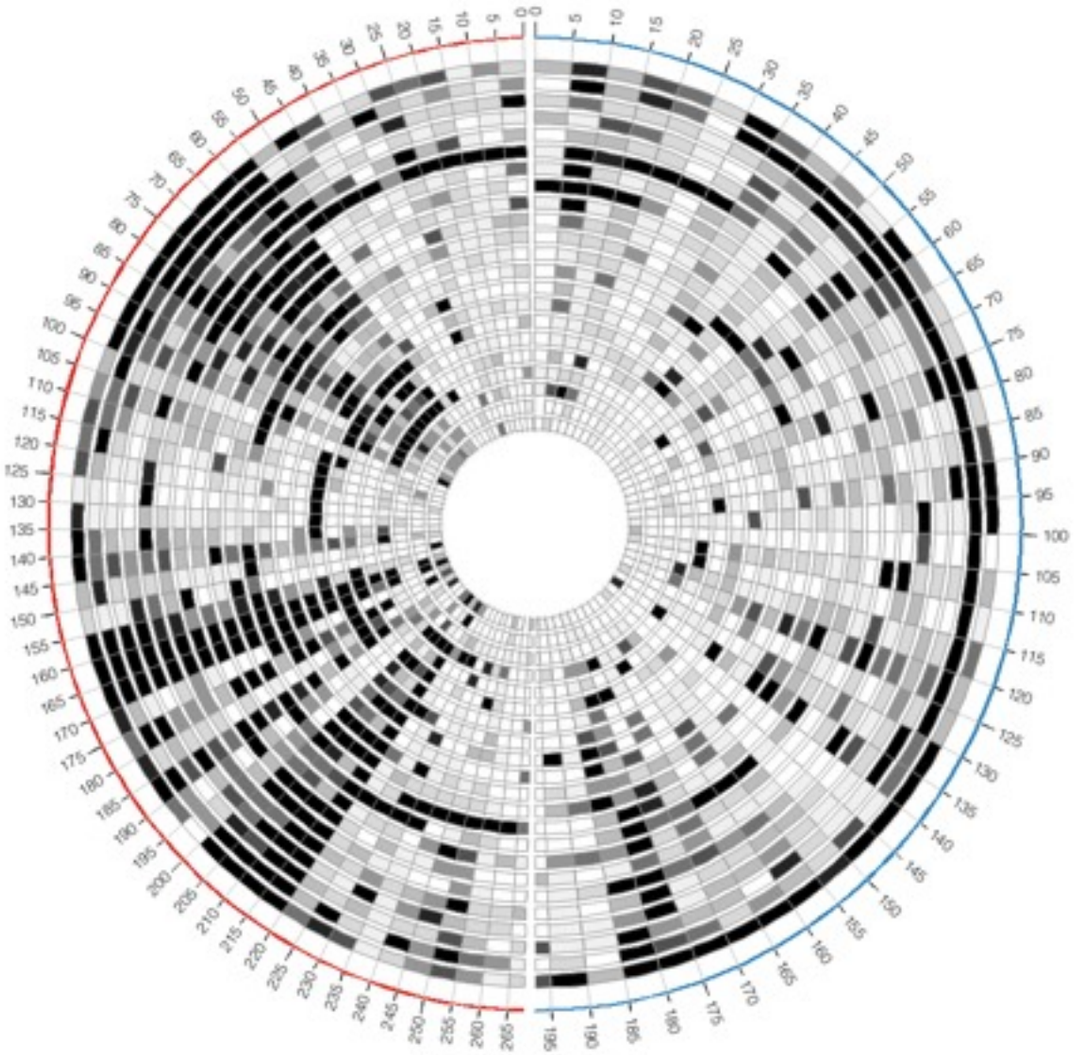
ALTER FORMATTING, NOT DATA

DYNAMIC PARAMETERS

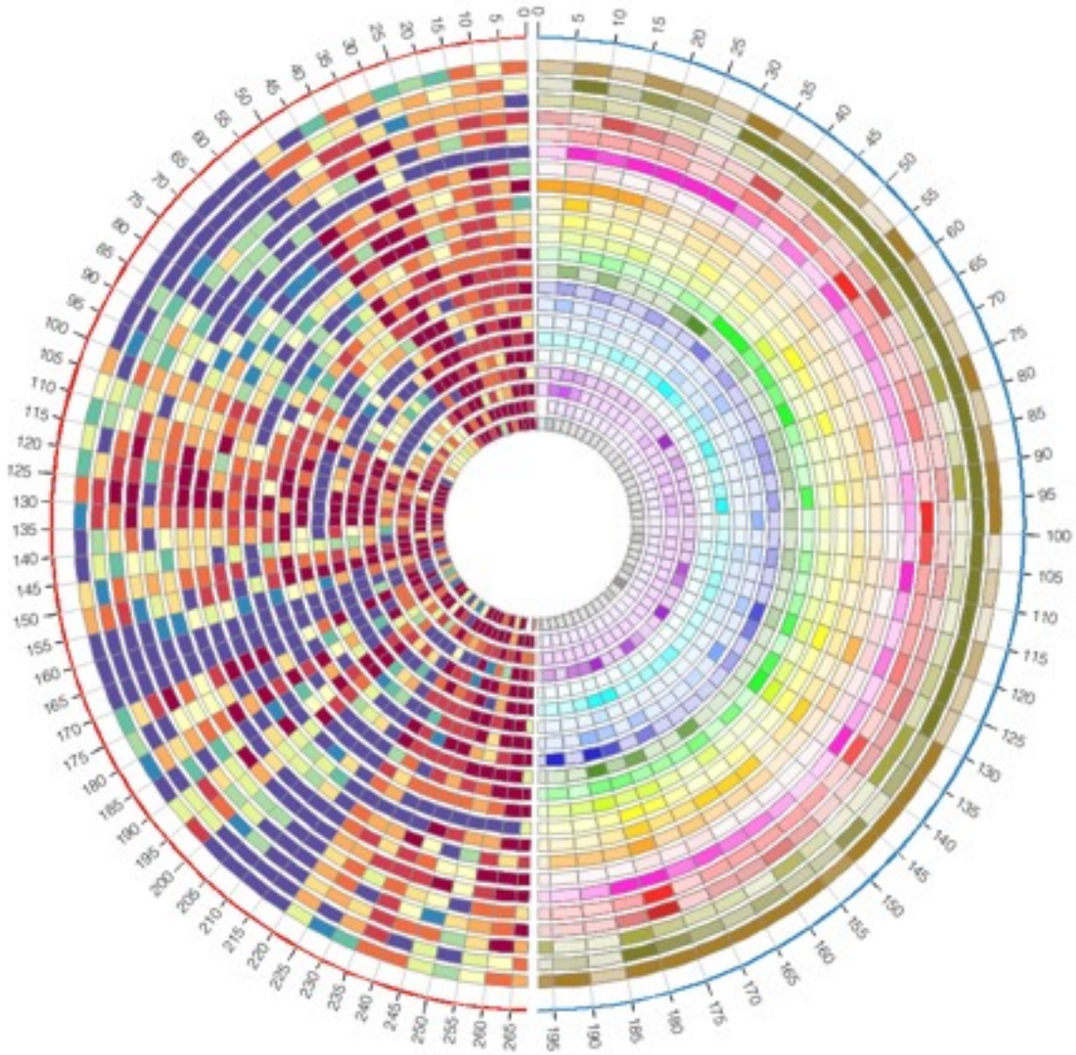


.

DYNAMIC PARAMETERS



```
color = greys-9-seq
```



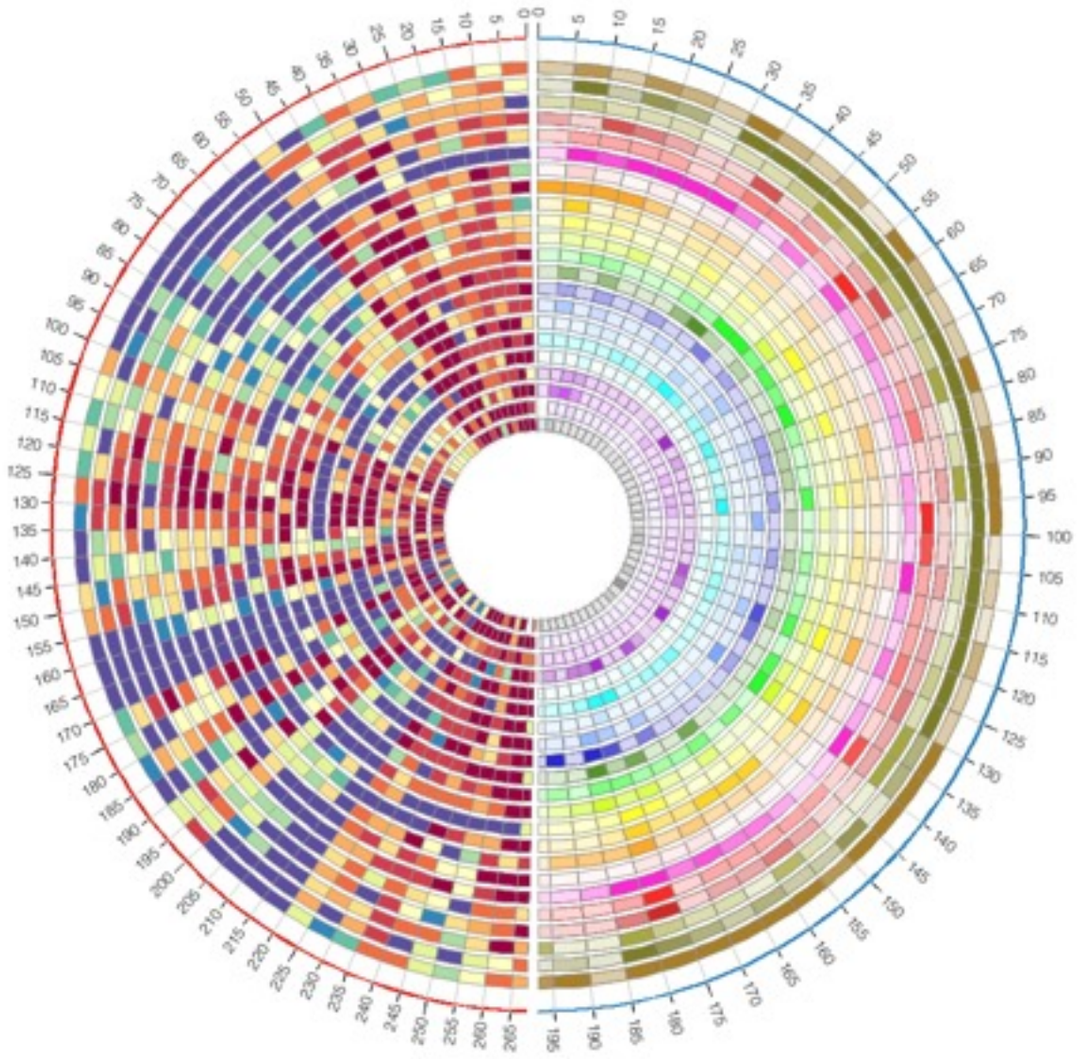
LEFT

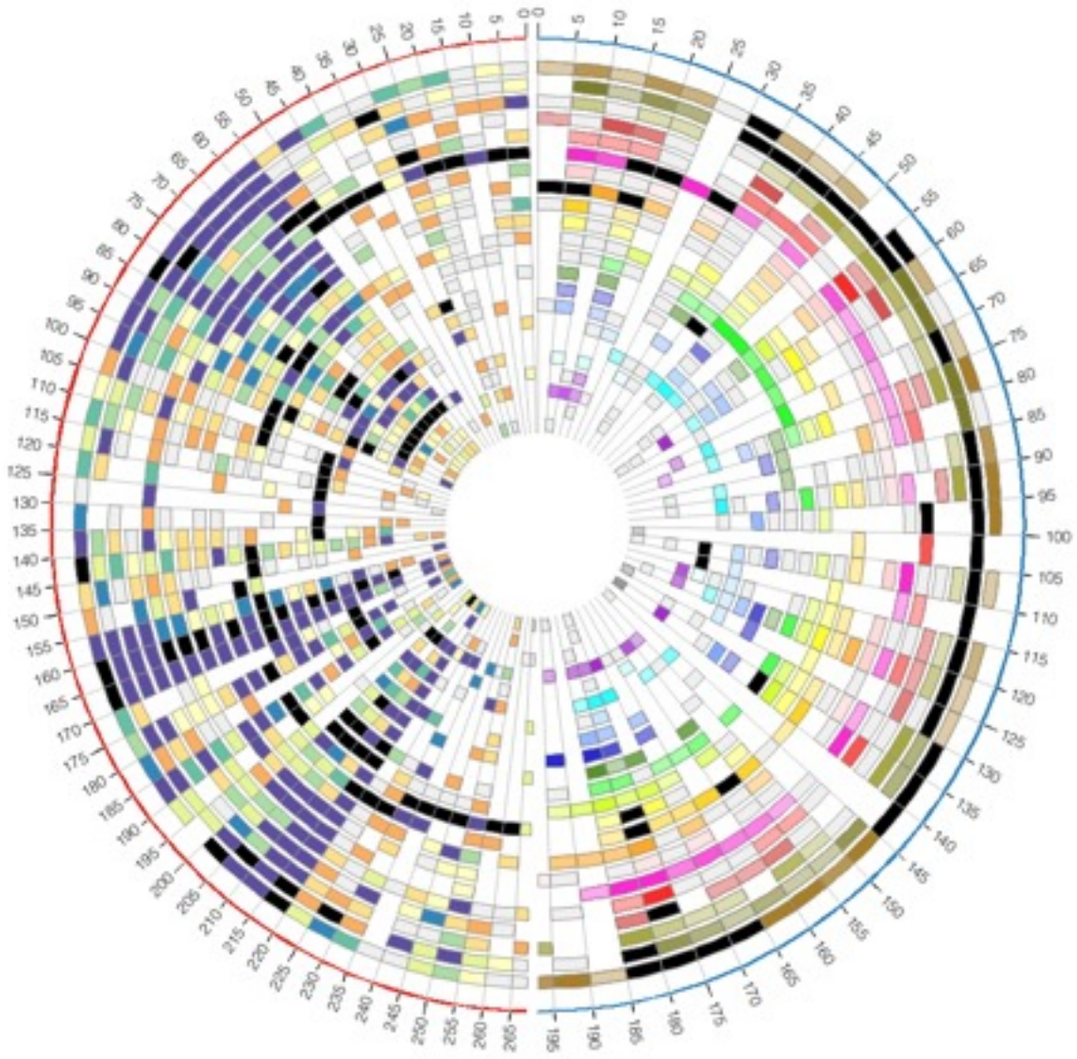
```
color = spectral-11-div
```

RIGHT

```
color = eval(join(",",map { sprintf("chr%d_a
%d",__$CONF{counter}{mmchain}__,$_) }
(5,4,3,2,1) ))
```

DYNAMIC RULES

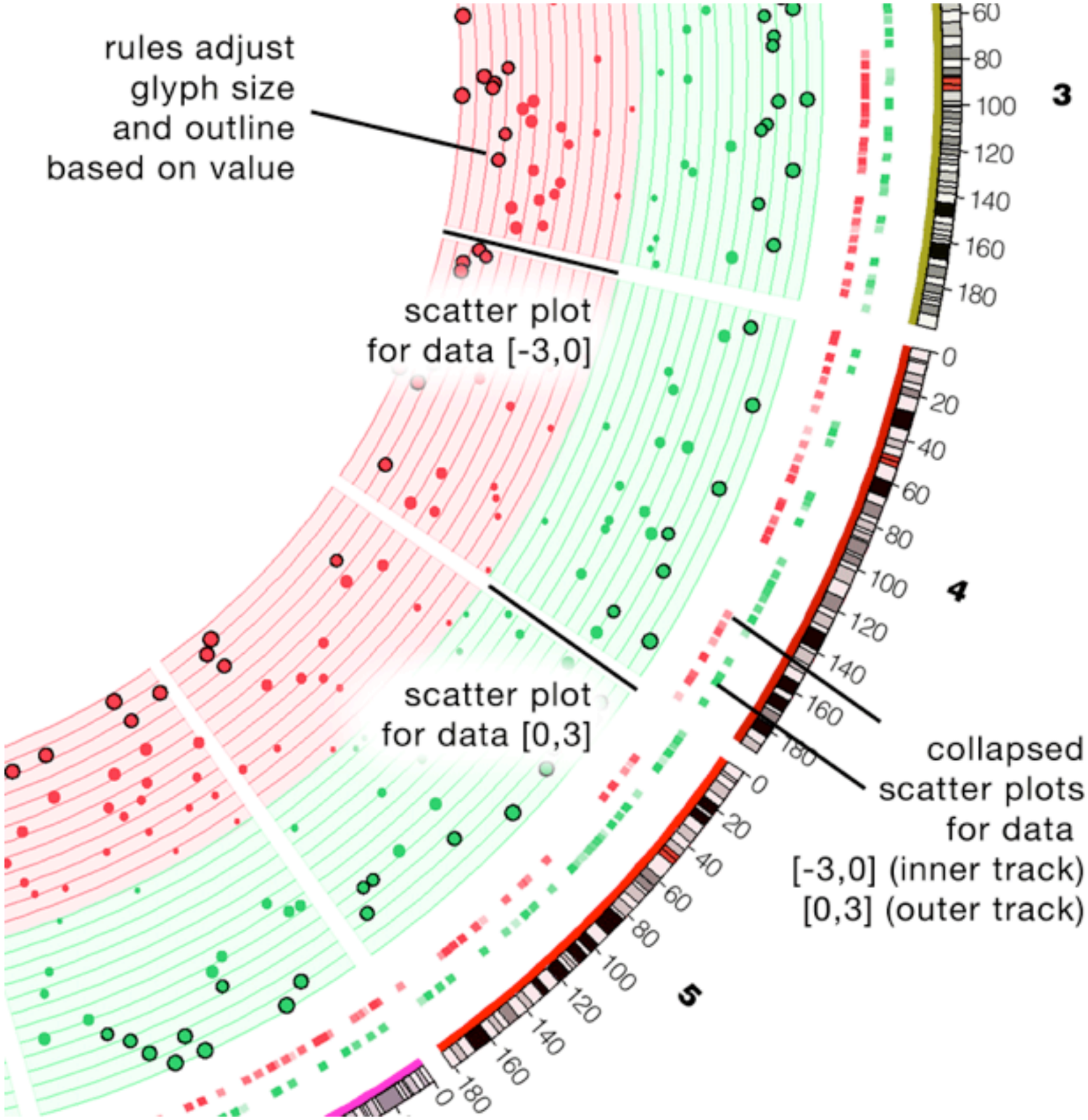




Each data point is tested against a rule chain. When the rule's condition matches, the data point's value, position and formatting can be dynamically adjusted.

```
<rule>
condition  = _VALUE_ < 15000
show       = no
</rule>
<rule>
condition  = _VALUE_ < 20000
color      = vvlgrey
</rule>
<rule>
condition  = _VALUE_ > 100000
color      = black
</rule>
```

DYNAMIC RULES

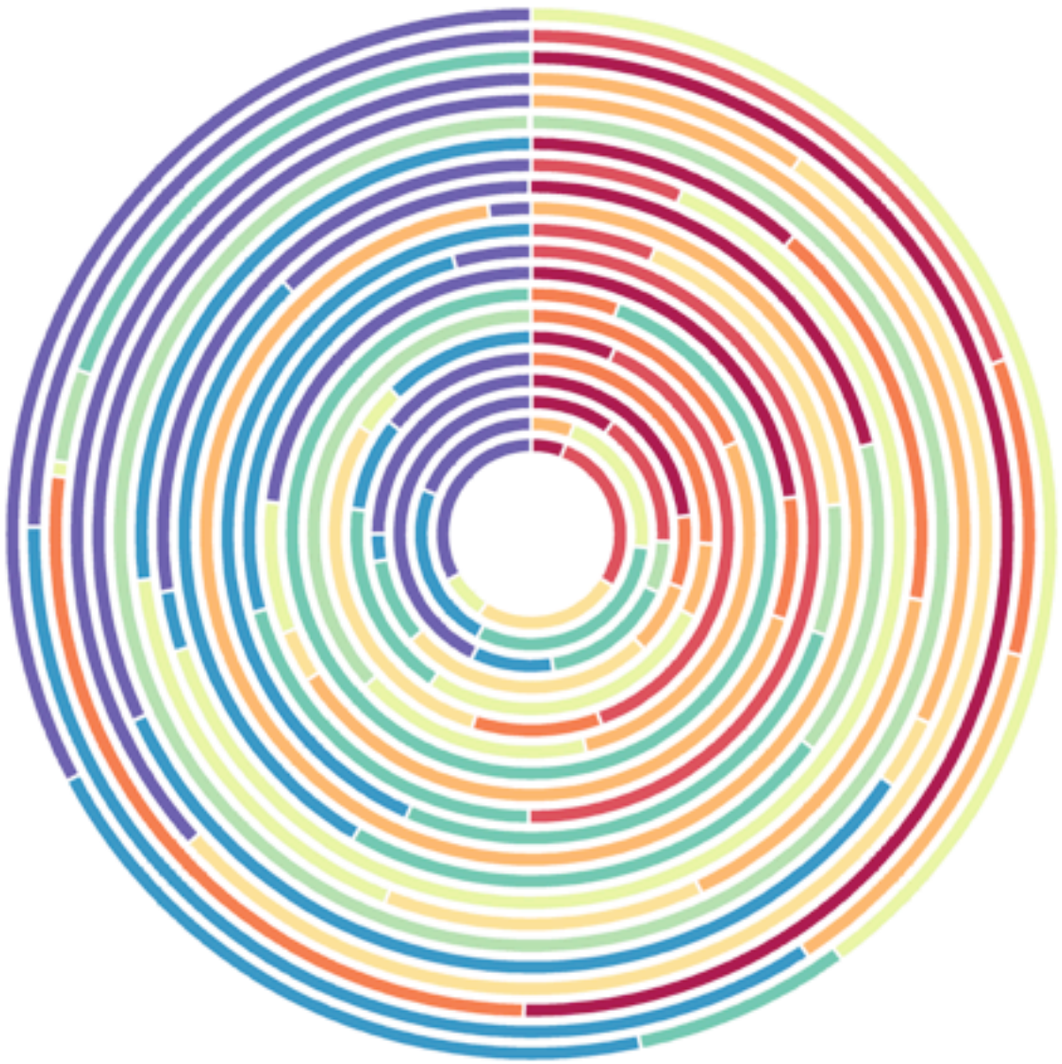


The size and outline of each scatter plot glyph is influenced by the data value. The data value itself can be altered, as see in the two outermost collapsed scatter plots, where the value for each point has been set to 0 to display the glyphs at the same radius.

automation

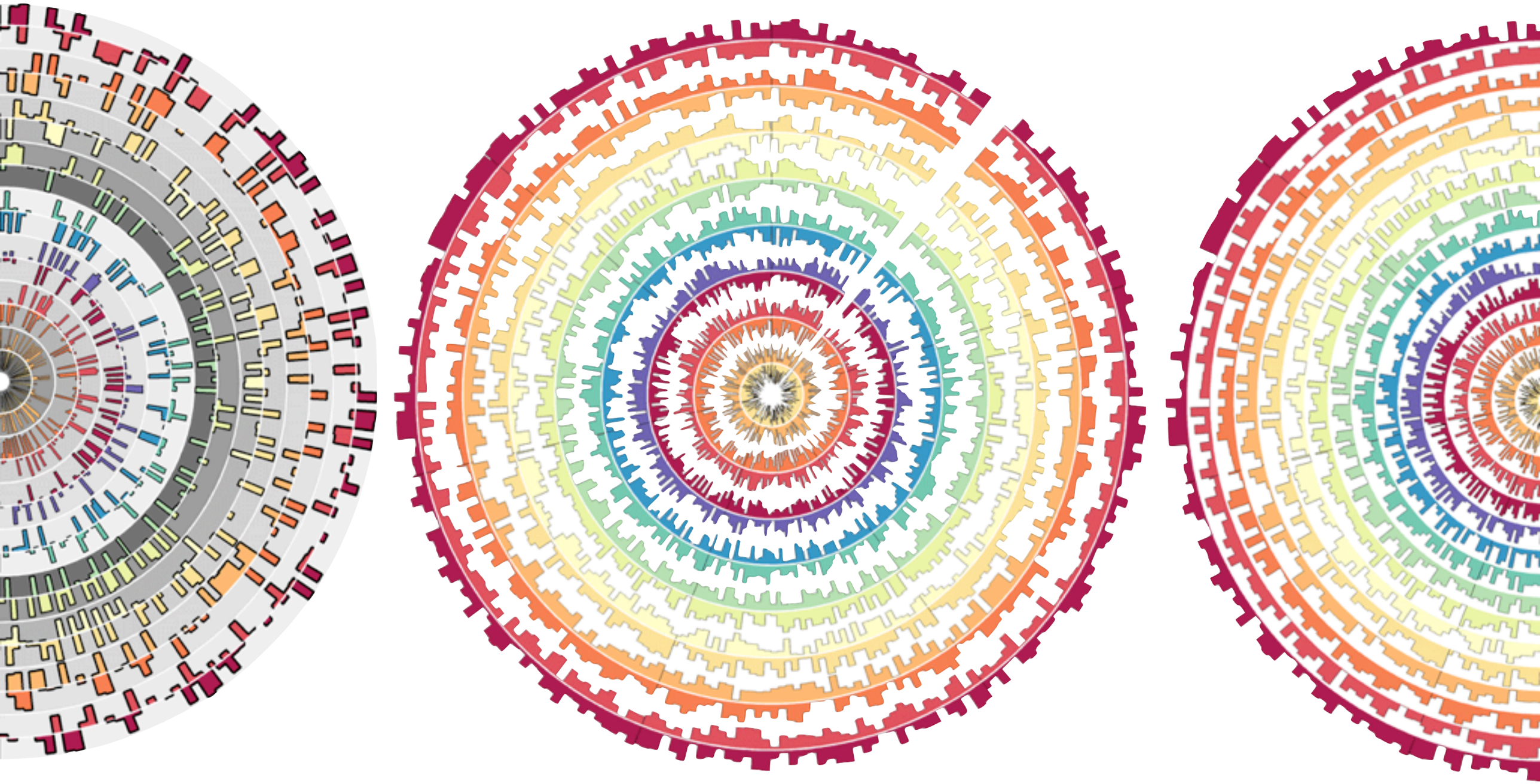
TEMPLATE-DRIVEN TRACKS

TEMPLATE-DRIVEN TRACKS



Each track is associated with several internal counters. The value of the counters are different for each track and can be used to drive track generation from a single template. By referencing the template multiple times, new tracks can be created automatically, without having change the template.

TEMPLATE DRIVEN TRACK PARAMETERS



Properties of each successive track are determined by the track's index. Orientation, color, transparency, background can thus be made to alternate or progressively change.

PLAIN TEXT CONFIGURATION

```
### circos.conf


...
# track definition
<plot>
type  = heatmap
file  = conservation.txt
# track start/end radius
r0    = 0.70r
r1    = 0.75r
# data range
min   = 0.1
max   = 0.9
# color map
color = spectral-11-div
</plot>
...
```

CIRCOS

TOOL

circular visualization of relationships and dense data

www.circos.ca

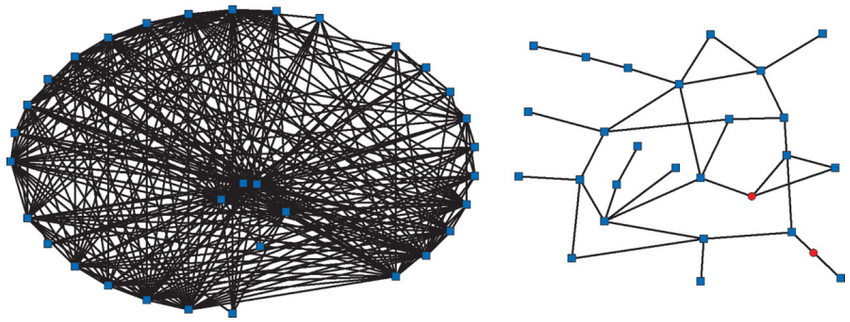
HIVE PLOTS

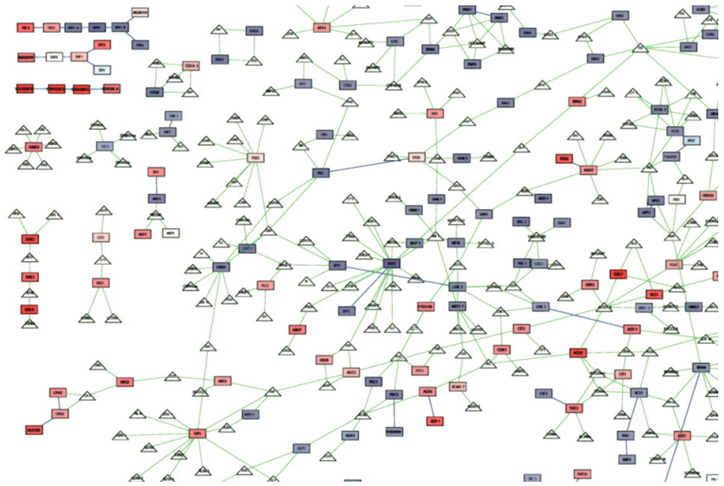
CONCEPT

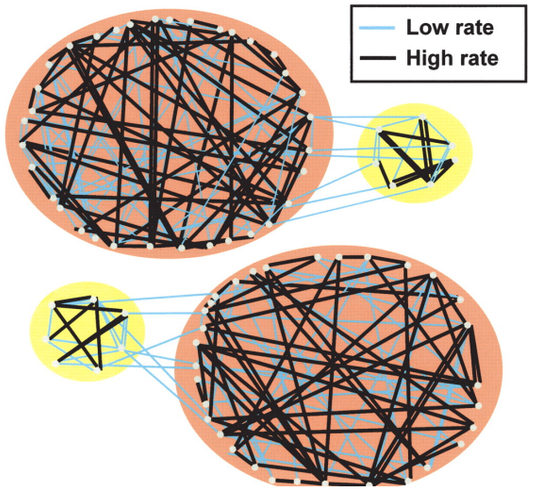
approach for rational, scalable and interpretable visualization of networks

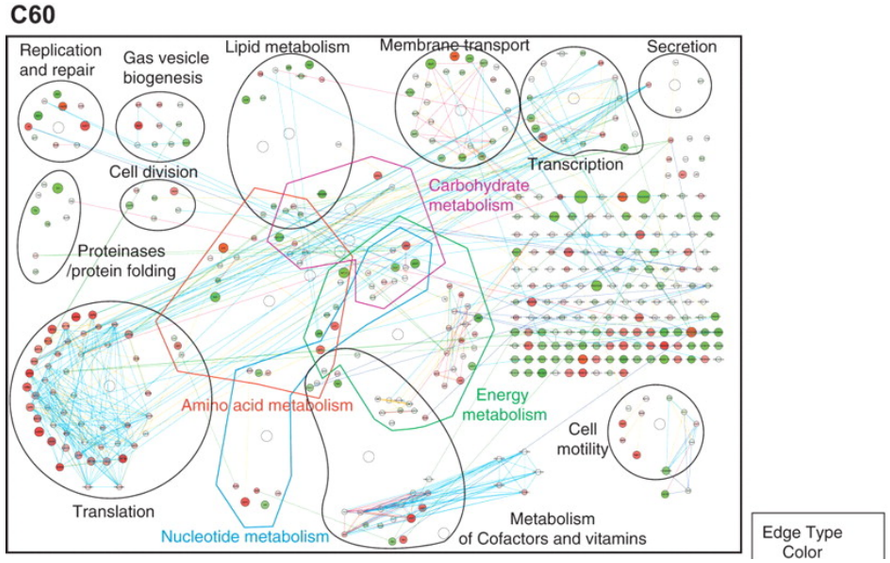
www.hiveplot.com

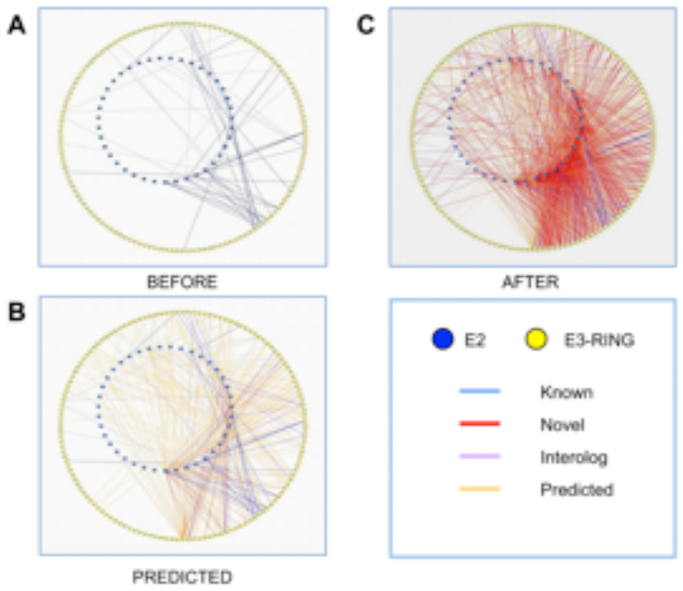
WHAT'S THE OTHER PROBLEM?

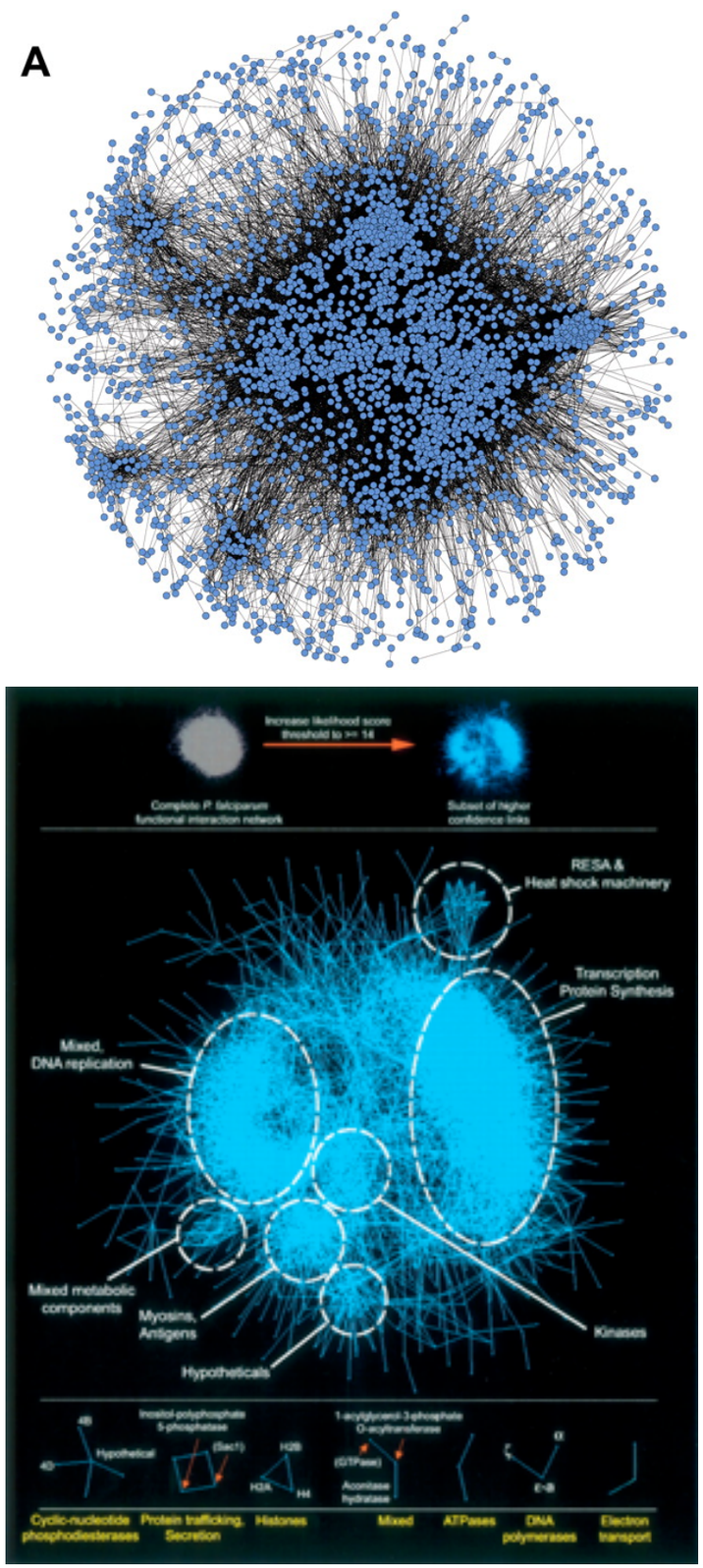






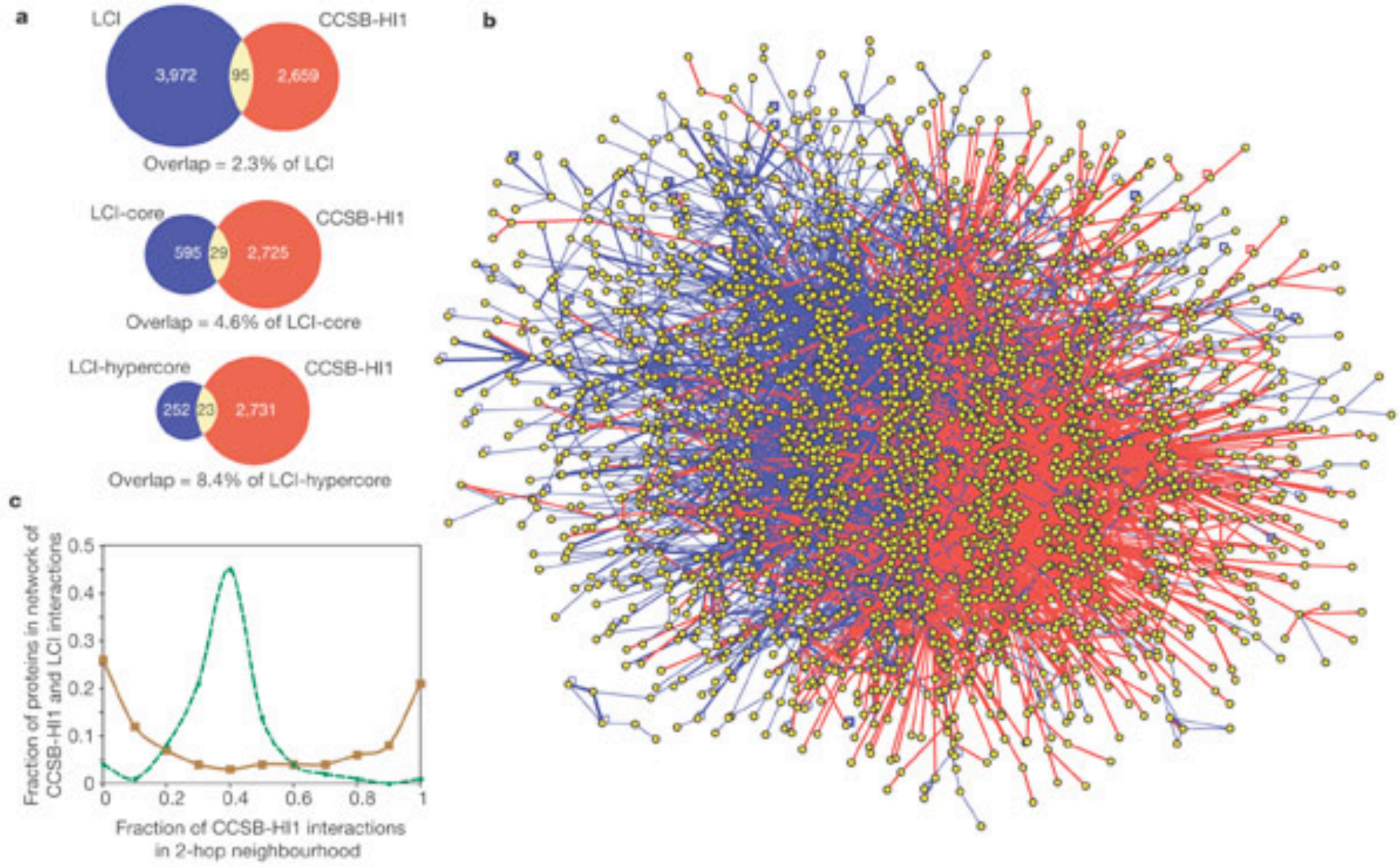






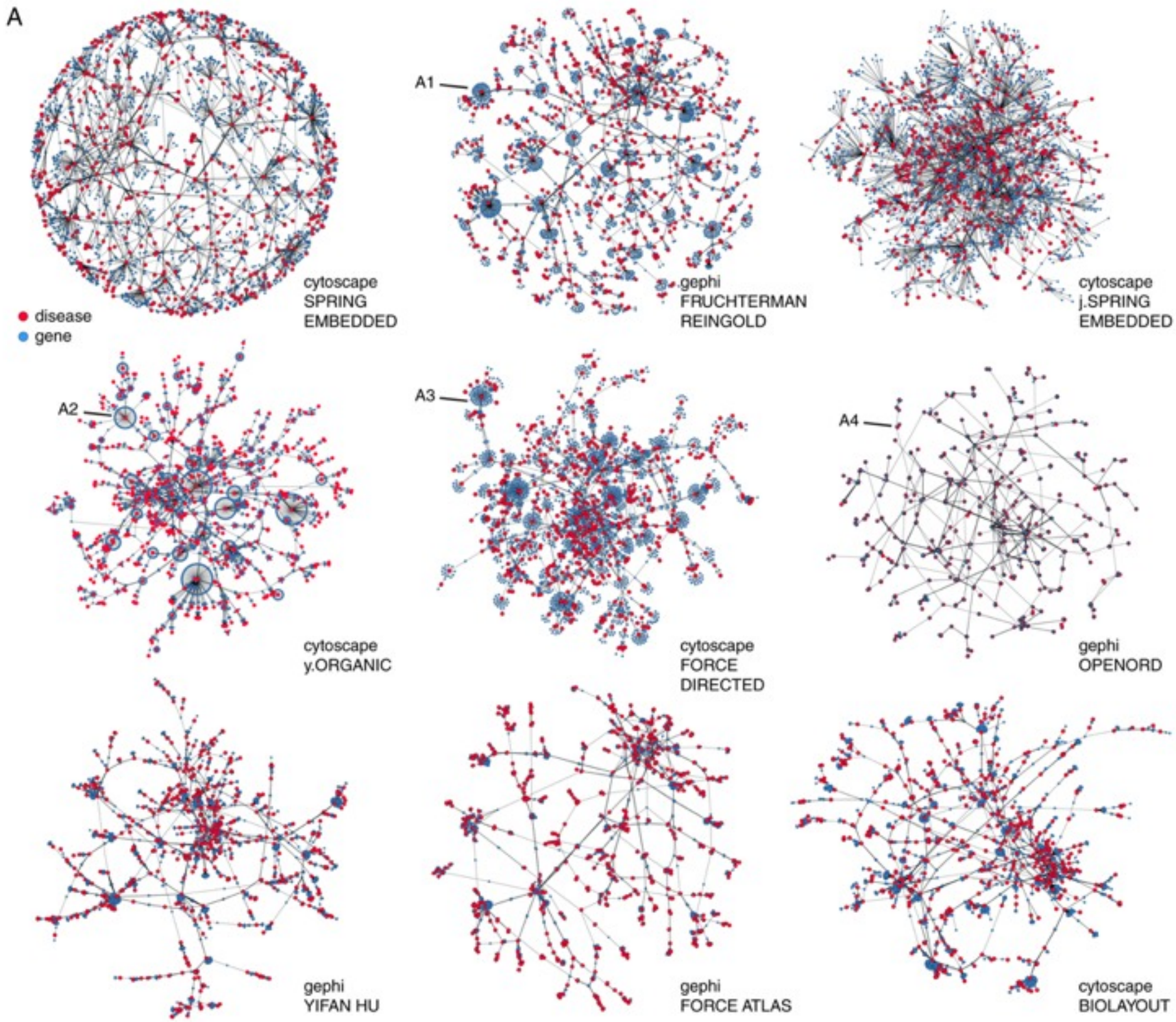
(1) Shakhnovich, B.E. and E.V. Koonin, Origins and impact of constraints in evolution of gene families. Genome Res, 2006. 16(12): p. 1529-36. (2) Prinz, S., et al., Control of yeast filamentous-form growth by modules in an integrated molecular network. Genome Res, 2004. 14(3): p. 380-90. (3) Nayak, R.R., et al., Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. Genome Res, 2009. 19(11): p. 1953-62. (4) Genome Res (5) Genome Res (6) Markson G. et al. Analysis of the human E2 ubiquitin conjugating enzyme protein interaction network. Genome Res. October 2009 19: 1905-1911 (7) Date S.V., Stoeckert Jr., C.J. Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. Genome Res. April 2006 16: 542-549 (8) Formstecher, E., et al., Protein interaction mapping: a Drosophila case study. Genome Res, 2005. 15(3): p. 376-84.

THE MISLEADING HAIRBALL



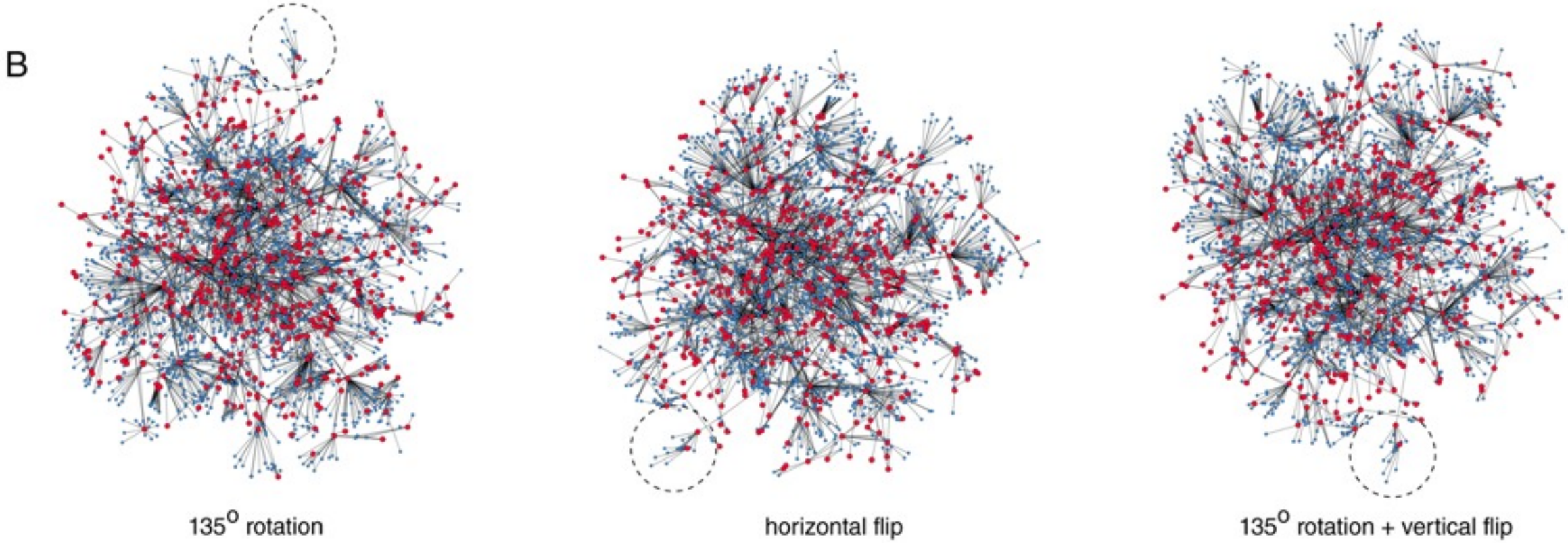
"The apparent banding pattern of the yellow nodes is an artefact of the graph layout algorithm (Supplementary Data). Importantly, the layout algorithm was not informed by type of supporting evidence and therefore does not explain the evident separation of blue and red edges." Figure 2 and caption quote from Rual et al., Nature 437(7062):1173-8.

THE CHAMELEON HAIRBALL



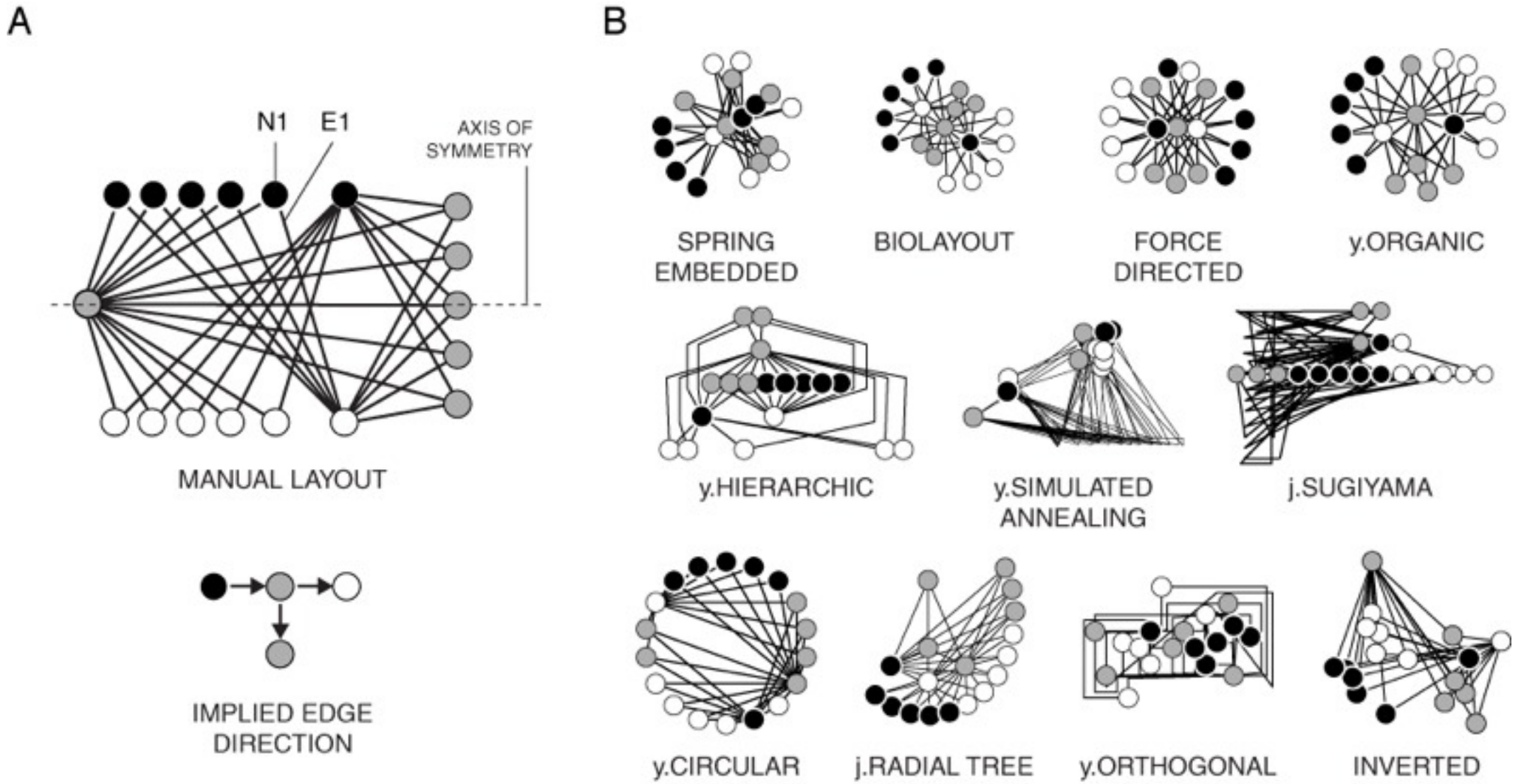
(A) Visualizations of the largest connected component (2,104 gene symbols, 548 diseases, 3,941 edges) of the human disease network (3,823 gene symbols, 1,284 diseases, 6,275 edges) generated with Cytoscape 2.8.1 and Gephi 0.7. Fruchterman Reingold and force directed layouts (A1, A3) render nodes uniformly around dense hubs, whereas the y.organic layout (A2) distributes the nodes at a constant radius around their hub. OpenOrd is efficient for very large networks and distinguishes clusters by collapsing neighbouring nodes (A4).

THE LYING HAIRBALL



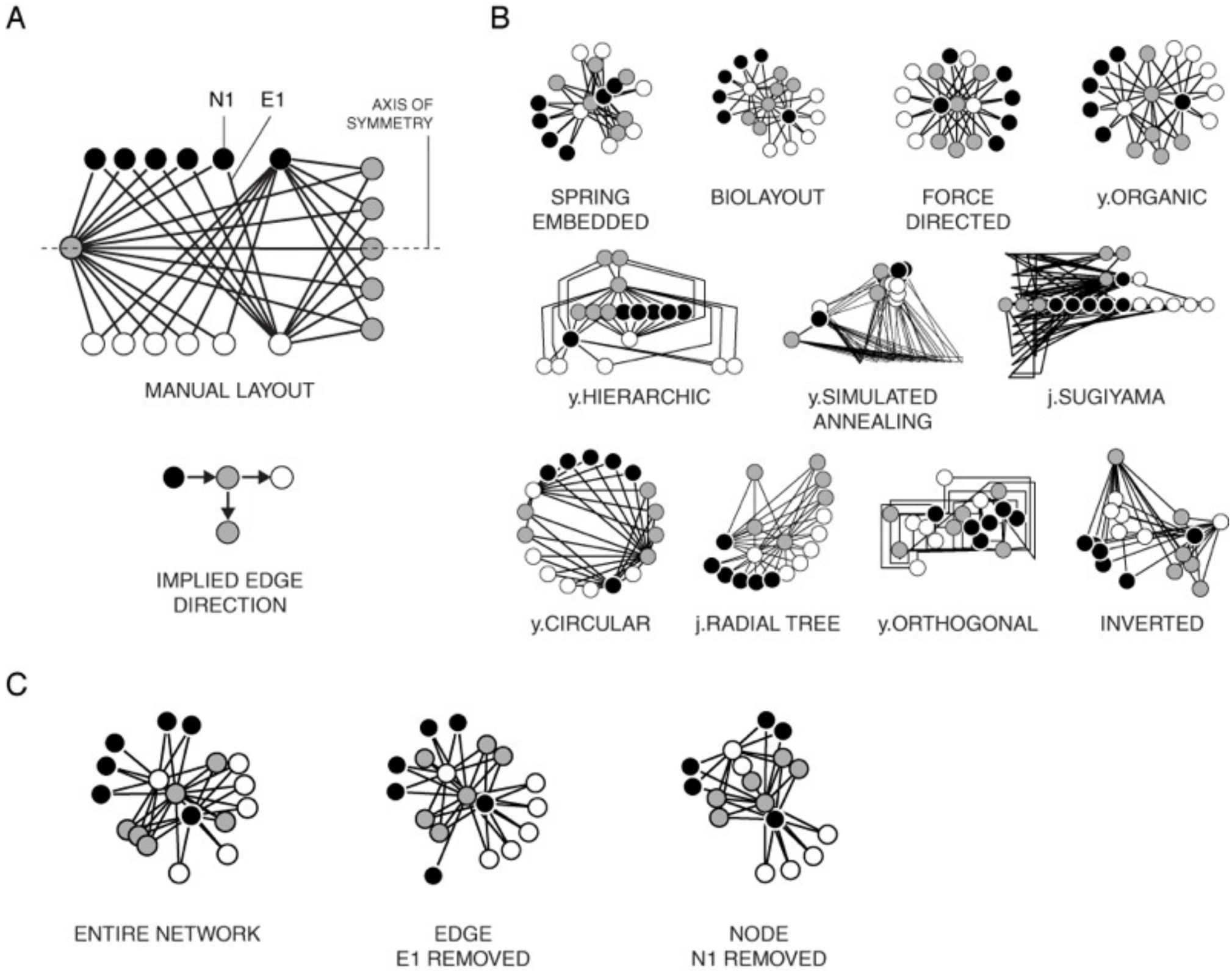
(B) Affine transformations of the j.spring embedded layout from panel A. The same group of nodes is highlighted with a dotted circle for orientation.

THE TRICKSTER HAIRBALL



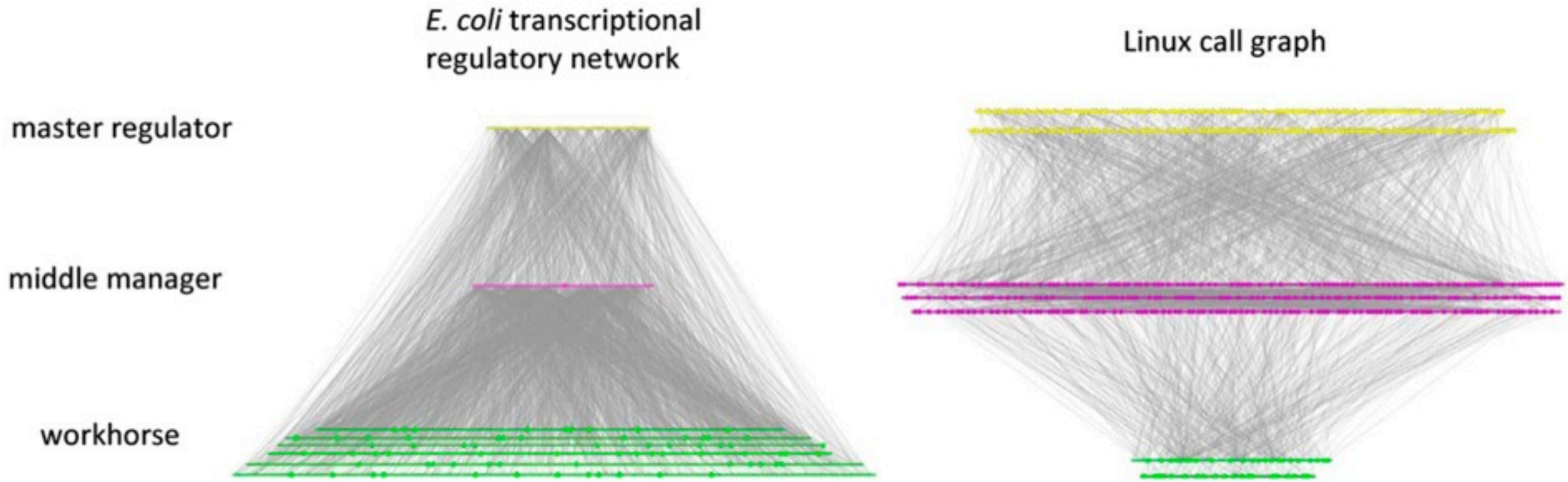
(A) Manual layout of a 15-node symmetric directed network. (B) Automated layouts of (A). (C) Spring embedded layout of instances of (A) with edge E1 and node N1 removed.

THE TRICKSTER HAIRBALL



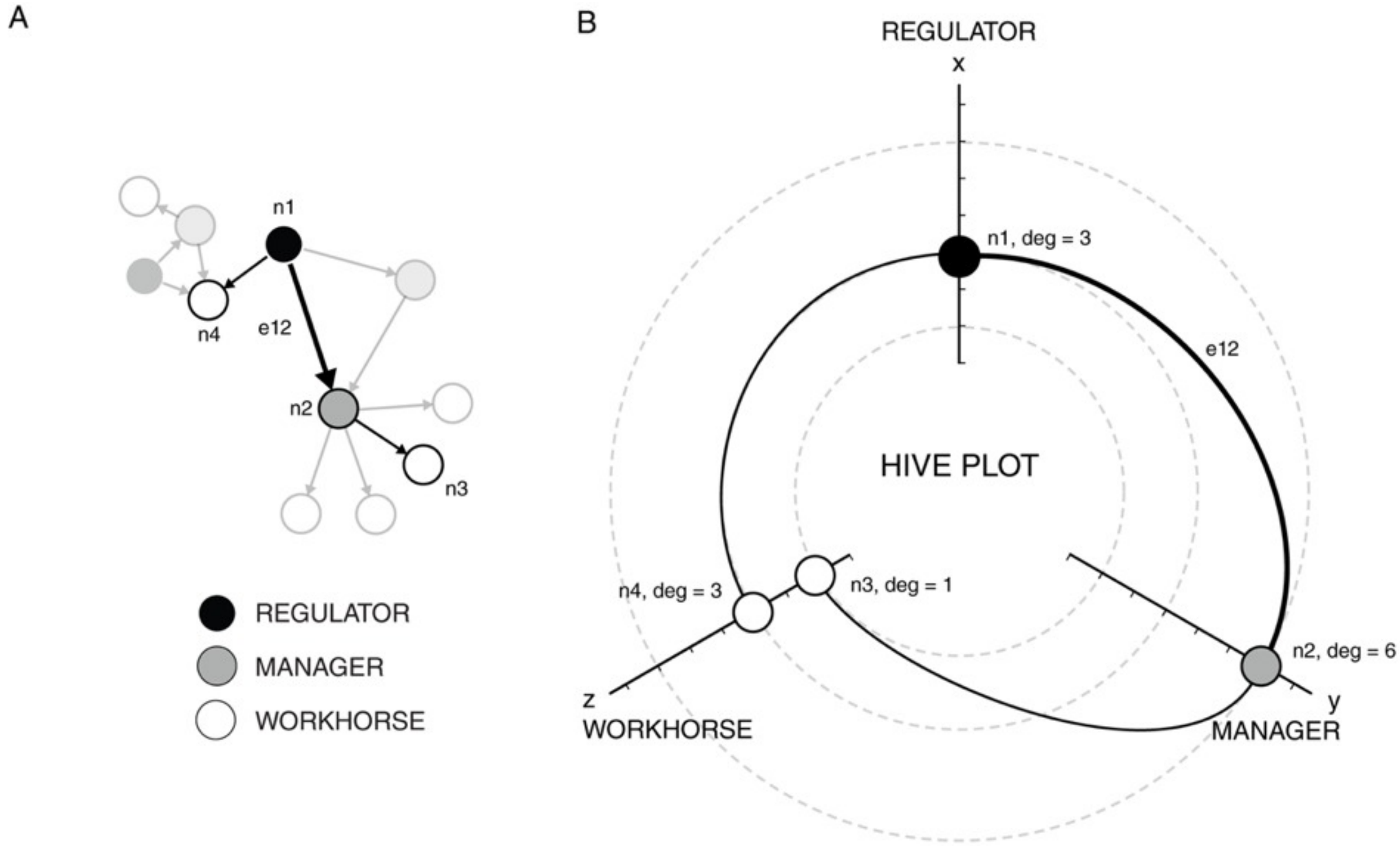
(A) Manual layout of a 15-node symmetric directed network. (B) Automated layouts of (A). (C) Spring embedded layout of instances of (A) with edge E1 and node N1 removed.

TOWARDS A RATIONAL LAYOUT



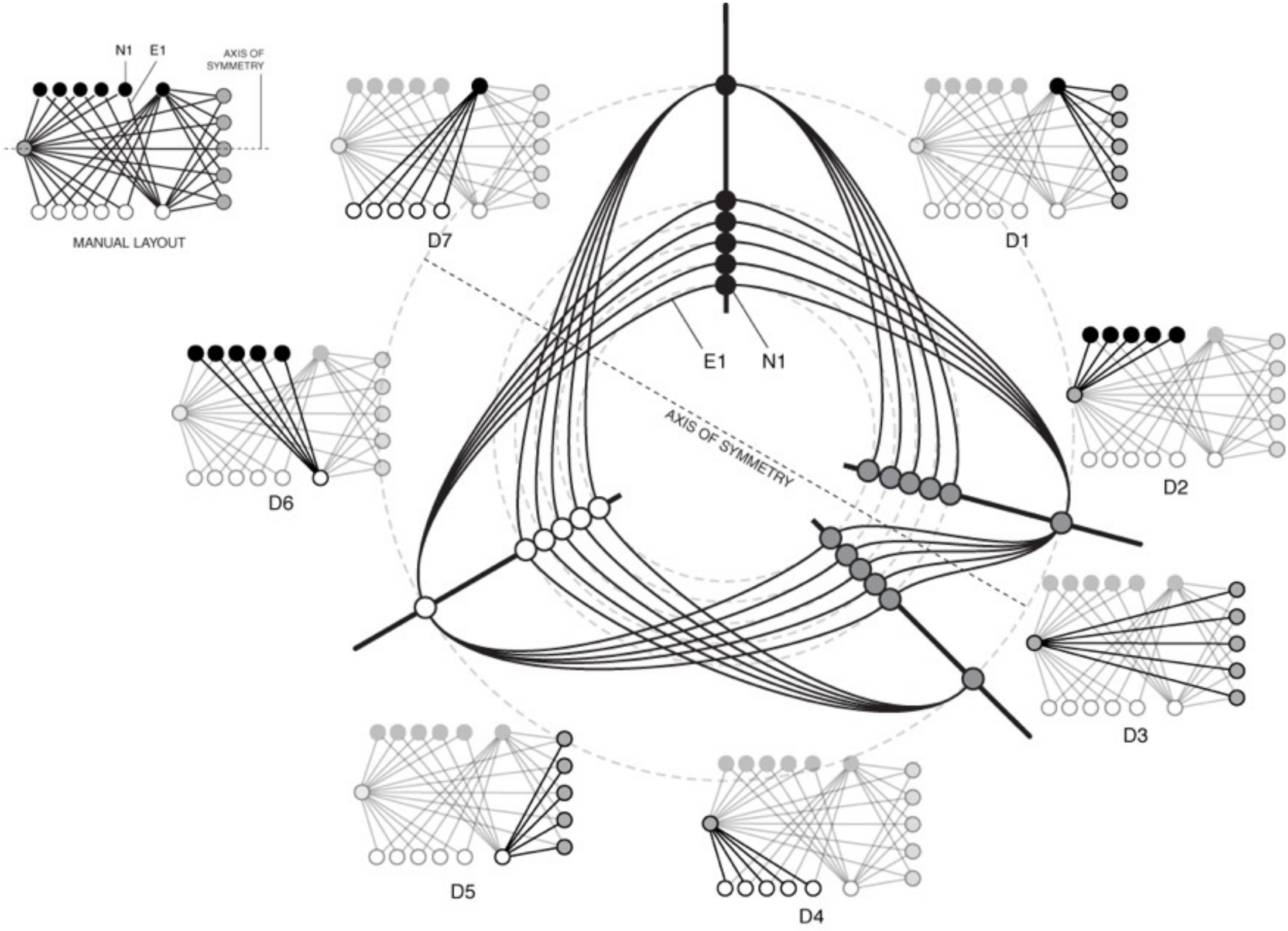
The hierarchical layout of the E. coli transcriptional regulatory network and the Linux call graph. (Left) The transcriptional regulatory network of E. coli. (Right) The call graph of the Linux Kernel. Nodes are classified into three categories on the basis of their location in the hierarchy: master regulators (nodes with zero in-degree, Yellow), workhorses (nodes with zero out-degree, Green), and middle managers (nodes with nonzero in- and out-degree, Purple). Yan KK, Fang G, Bhardwaj N et al.: 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 2010, 107(20):9186-9191.

HIVE PLOT METHOD



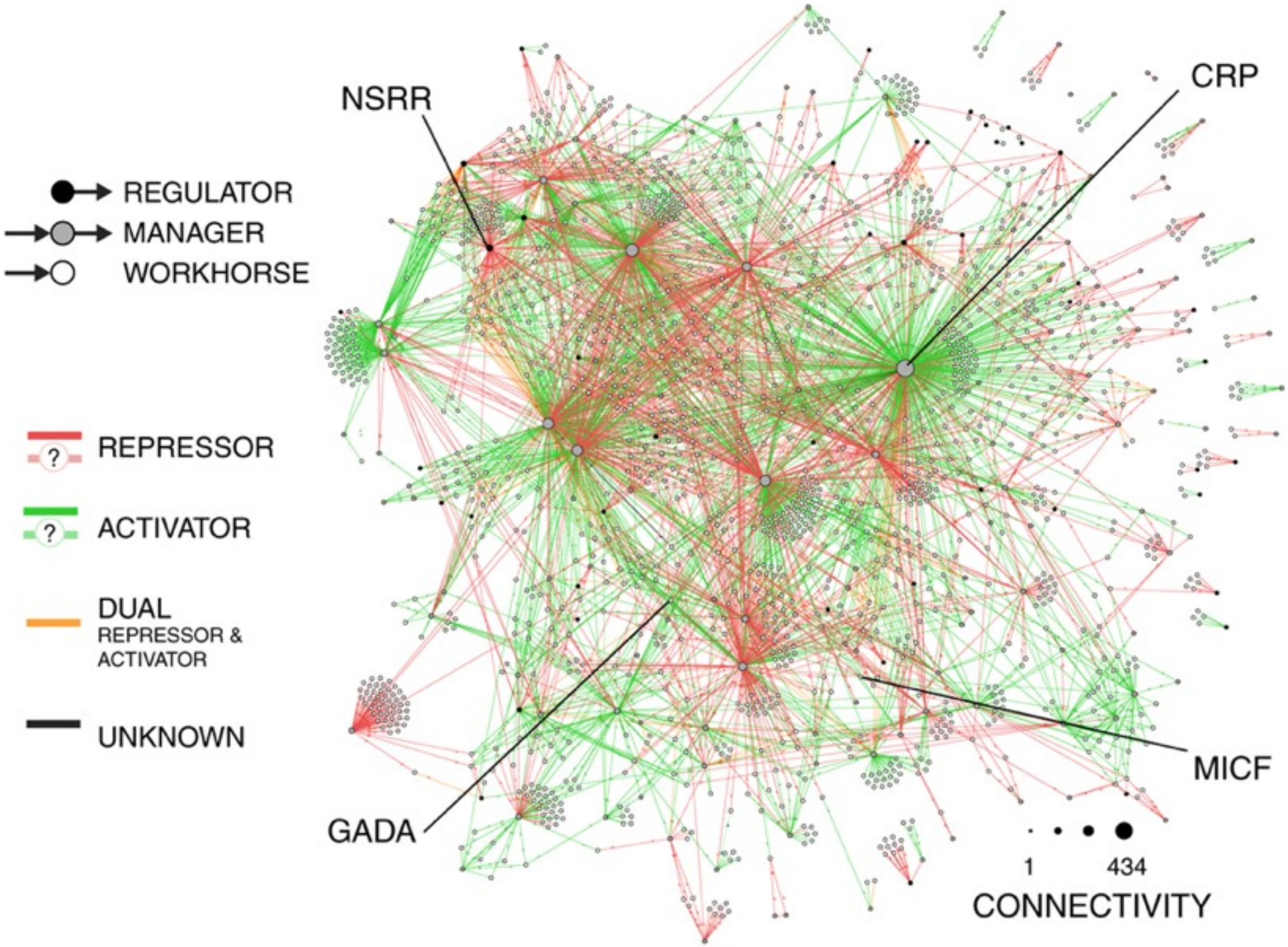
(A) A small directed network, representing gene regulation. (B) 3-axis hive plot (HP) of (A) constructed using role of nodes for axis assignment and connectivity for axis scale.

NAVIGATING THE HIVE PLOT



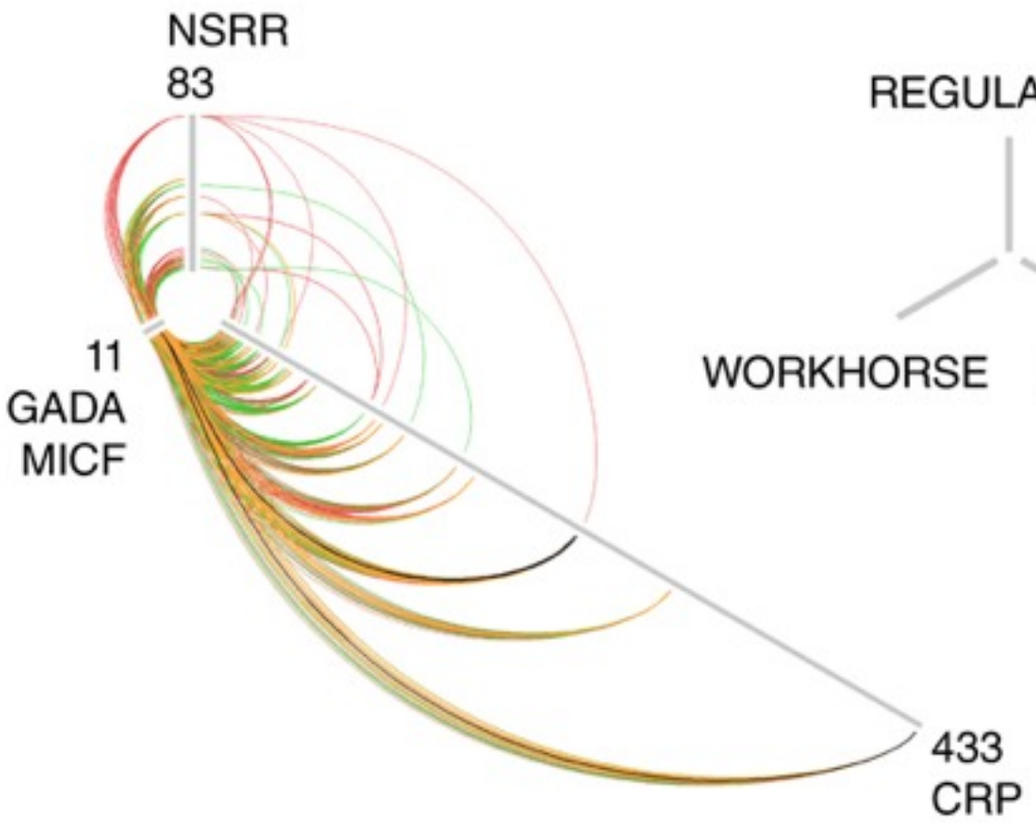
HP of Figure 2A using rules from Figure 3B. Position of nodes on the HP is demonstrated with copies of Figure 2A highlighting the nodes in question. Edge and node elements removed in Figure 2A to generate layouts in Figure 2C are indicated with E1 and L1, respectively.

E COLI REGULATORY HAIRBALL

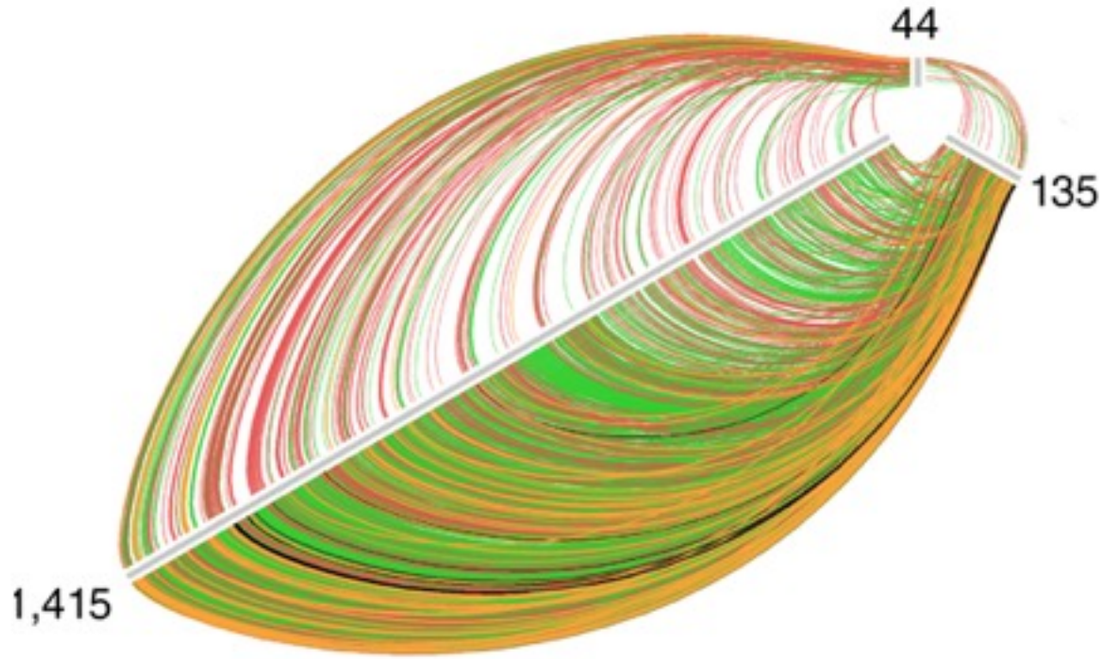


Gama-Castro S, Salgado H, Peralta-Gil M et al.: RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Research 2011, 39:D98-D105.

E COLI REGULATORY HIVE PLOT



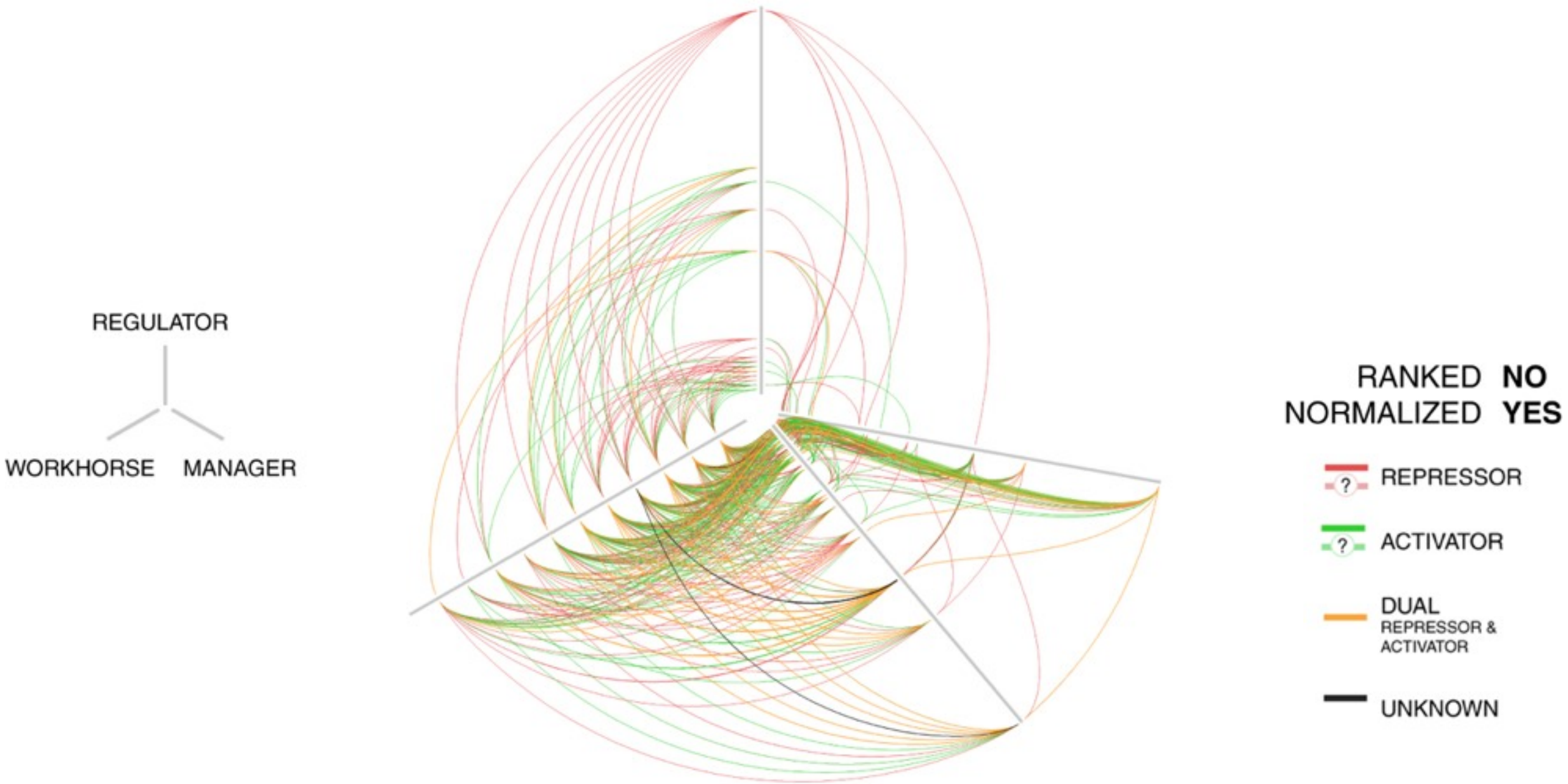
RANKED **NO**
NORMALIZED **NO**



RANKED **YES**
NORMALIZED **NO**

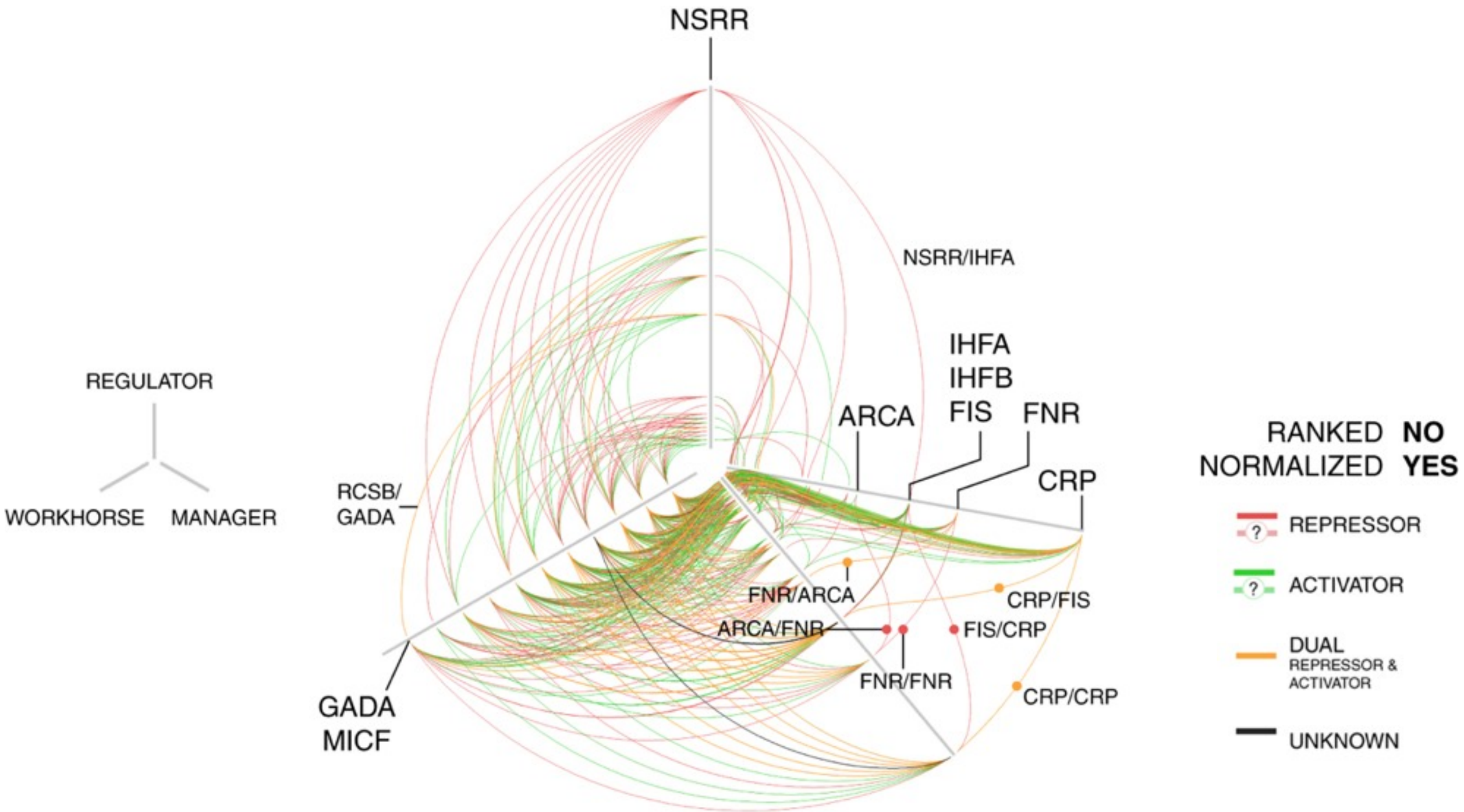
.

E COLI REGULATORY HIVE PLOT



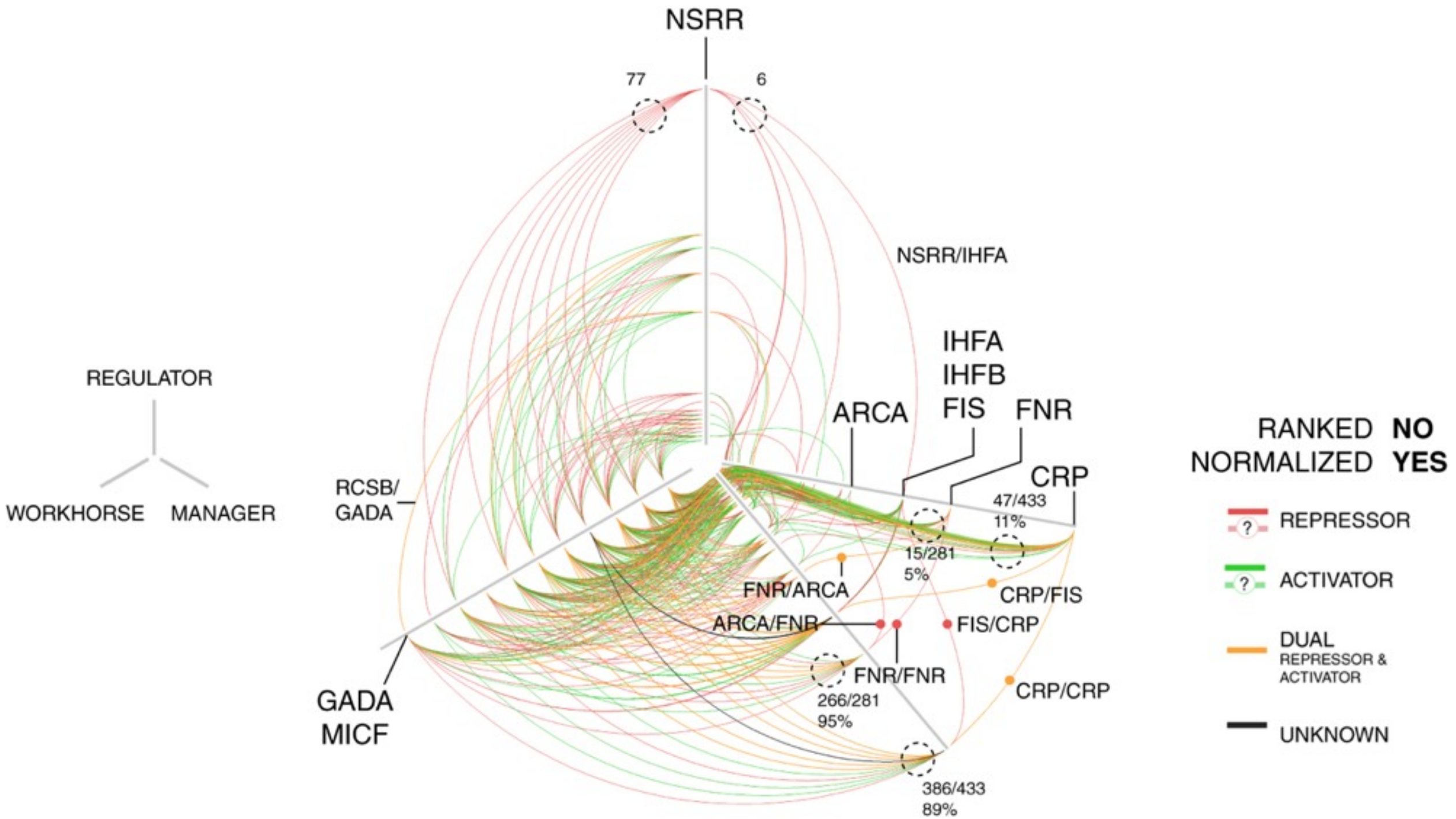
.

E COLI REGULATORY HIVE PLOT



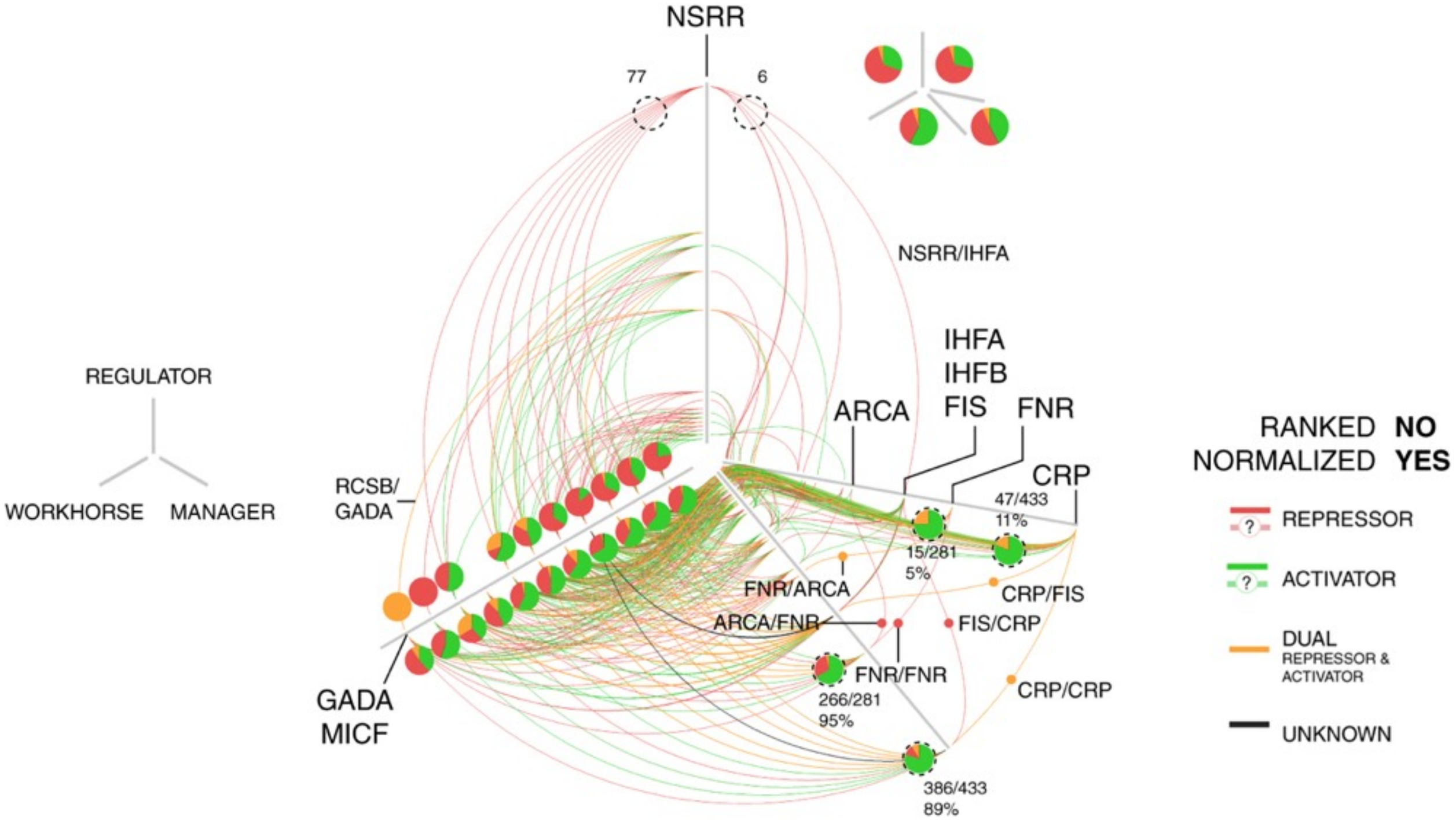
.

E COLI REGULATORY HAIRBALL



.

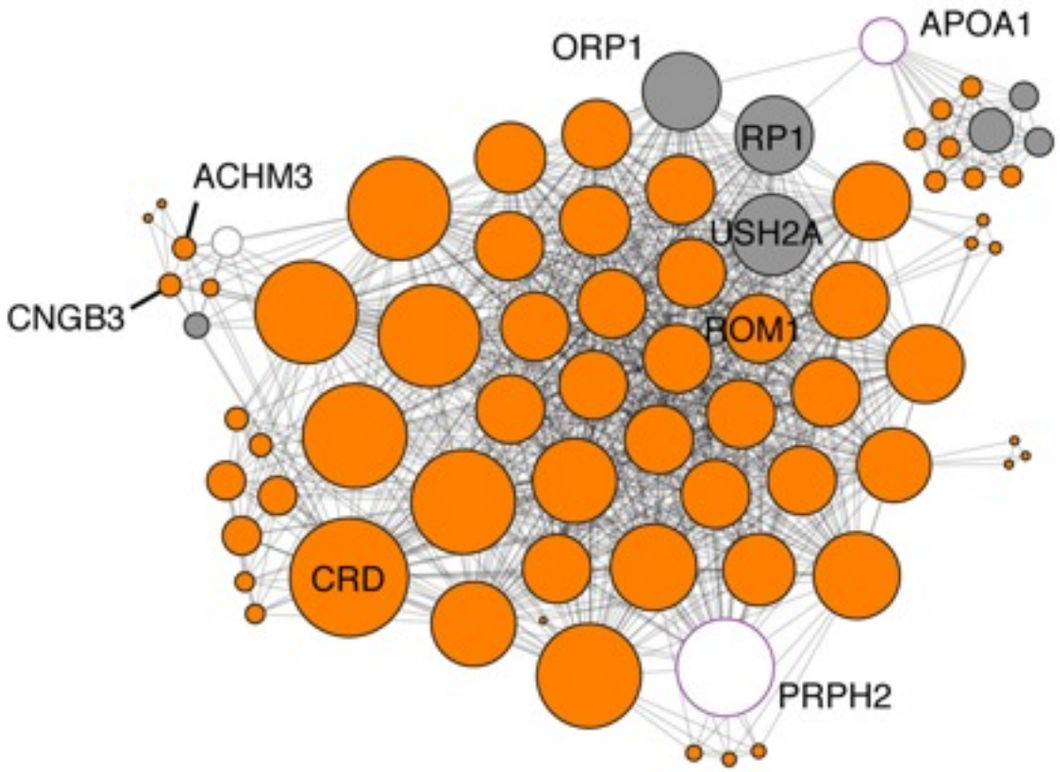
E COLI REGULATORY HAIRBALL



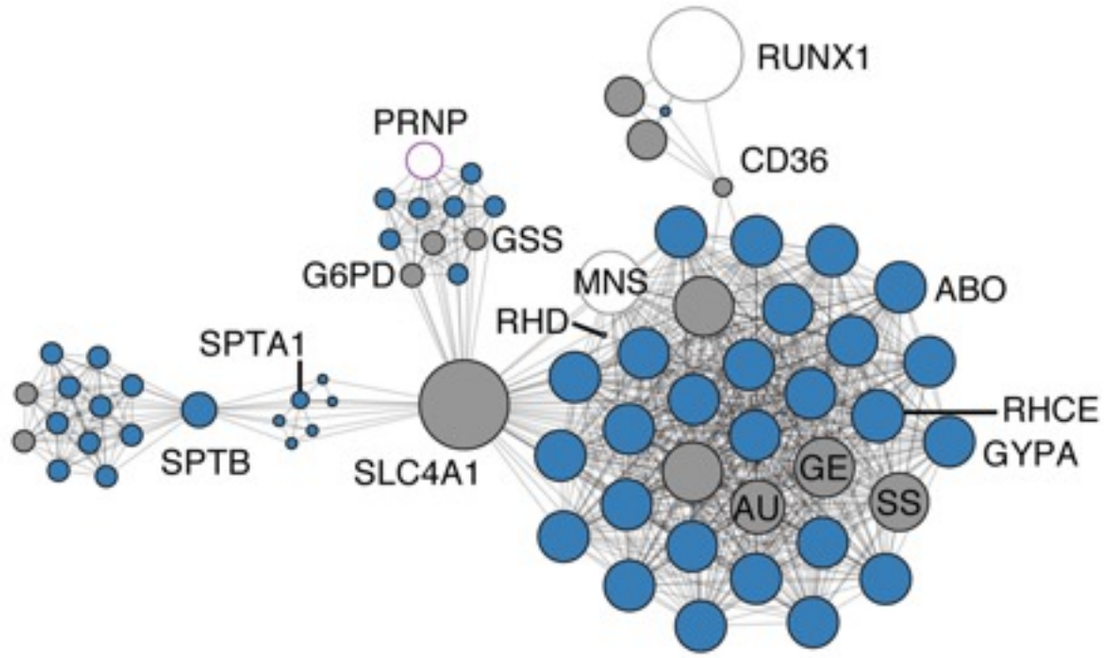
.

COMPARING NETWORKS

OPHTHALMOLOGICAL



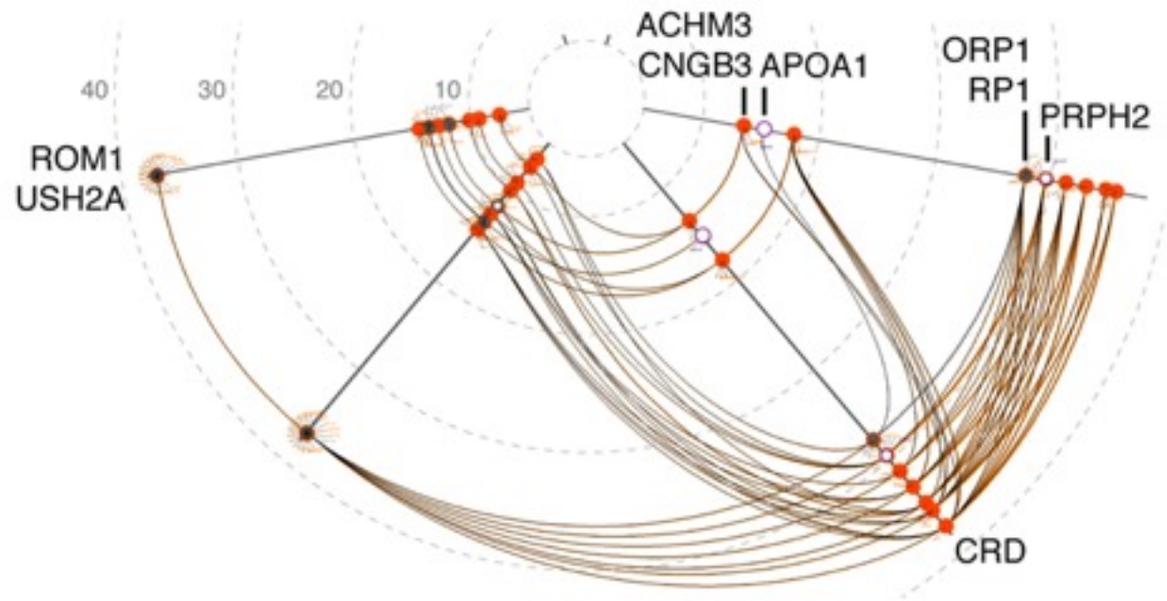
HEMATOLOGICAL



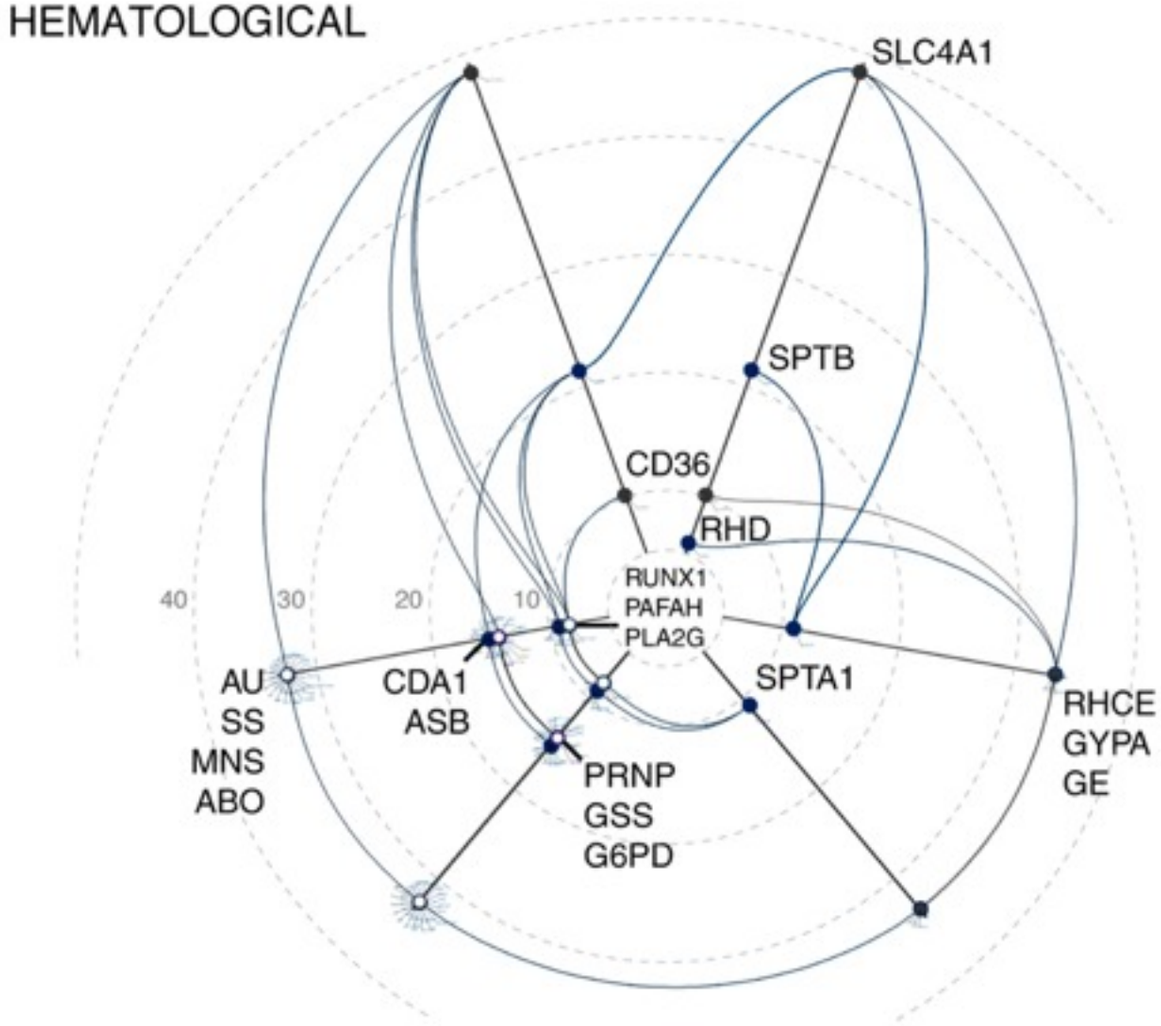
.

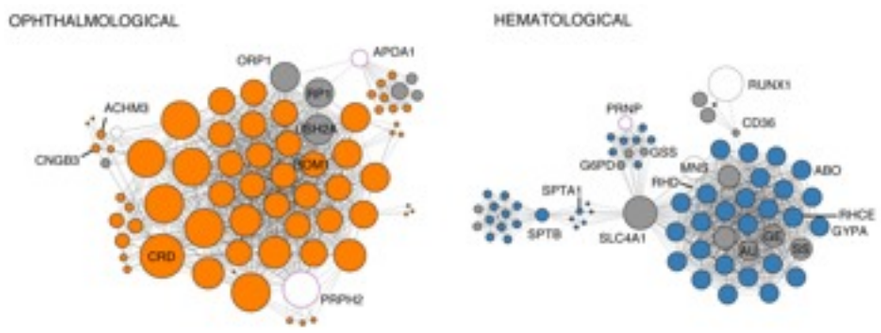
COMPARING NETWORKS

OPHTHALMOLOGICAL



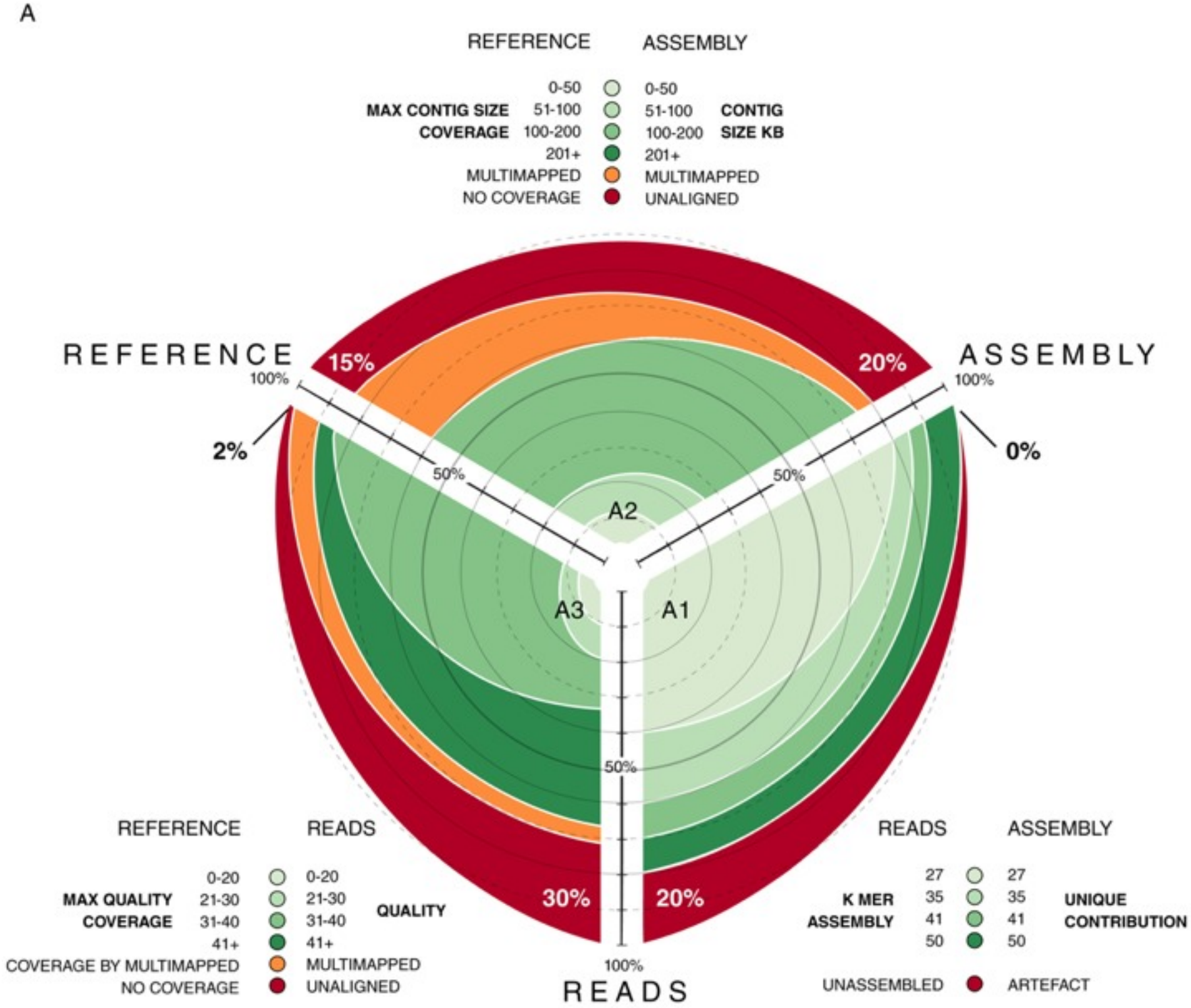
HEMATOLOGICAL





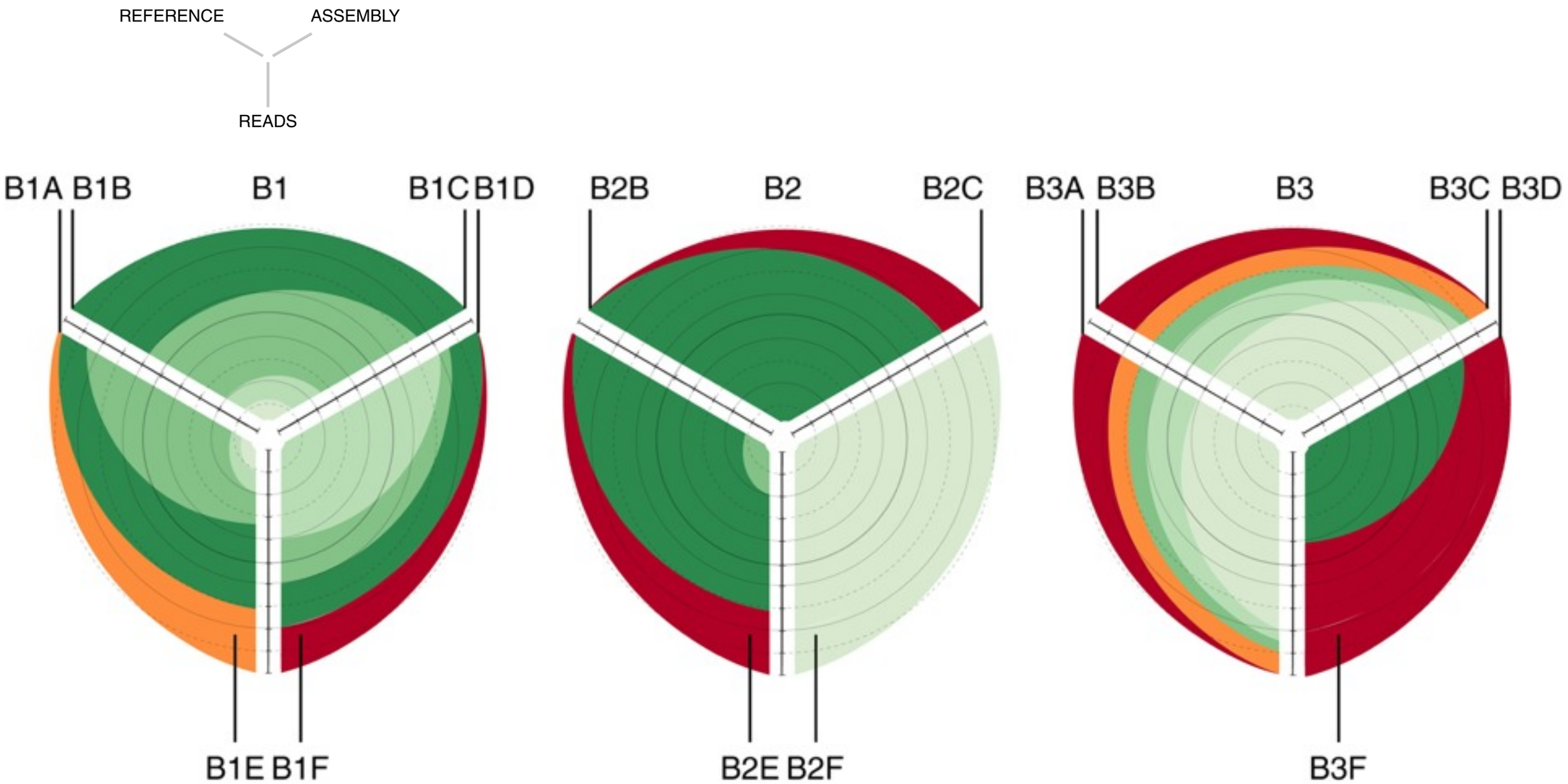
.

VISUALIZING RATIOS - ASSEMBLY QUALITY



Application of HPs to visualizing ratios. (A) Quality of a sequence assembly is visualized by relating (a) the fraction of reads, by assembly parameter, aligning to the assembly, by contribution (b) the fraction of the assembly, by contig size, providing coverage of the reference genome, by contig coverage, and (c) the fraction of reads, by quality, providing coverage of the reference genome, by quality coverage.

VISUALIZING RATIOS - ASSEMBLY QUALITY



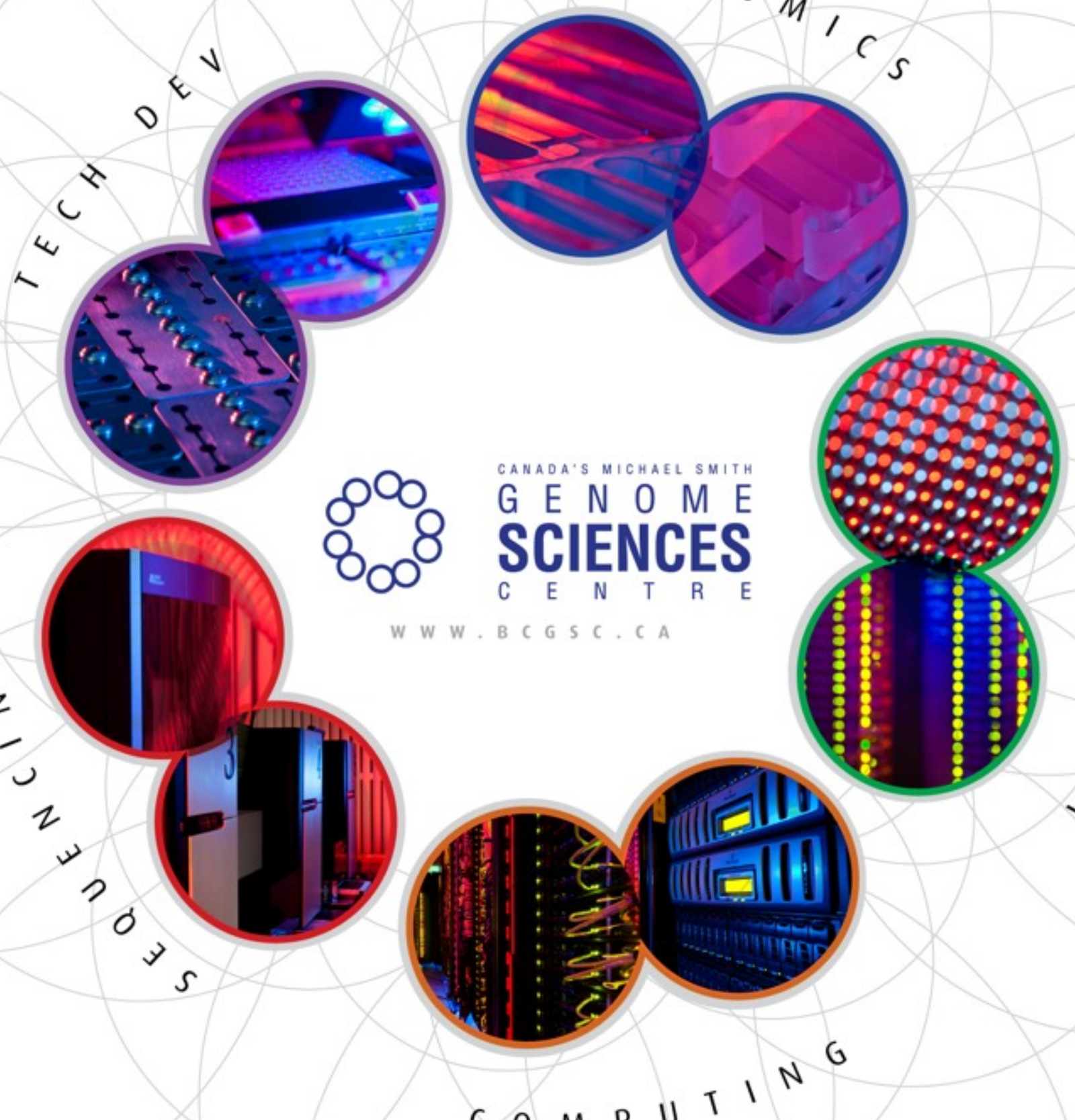
Three assembly scenarios. (B1) complete coverage of the reference genome with some unused reads (B1F) which ambiguously map to the reference (B1E); (B2) complete coverage of the reference with unique sequenced (B2E) and assembled (B2C) content; (B3) poor assembly with large fraction of unassembled reads (B3F), assembly error indicated by regions uncovered by reads (B3D), and incomplete coverage of the reference by both reads (B3A) and contigs (B3B).

CIRCOS

www.circos.ca

HIVE PLOTS

www.hiveplot.com


TECH DEV
GENOMICS
INFORMATICS
COMPUTING
SEQUENCING
CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE
WWW.BCGSC.CA